SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
GENETIC VARIATION
ANTHROPOLOGY

Received
5 March 2014

Accepted
28 July 2014

Published
13 August 2014

Correspondence and requests for materials should be addressed to C.N.R. (rotimic@mail.nih.gov)

Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry

Daniel Shriner, Fasil Tekola-Ayele, Adebowale Adeyemo & Charles N. Rotimi

Center for Research on Genomics and Global Health, National Human Genome Research Institute, Building 12A, Room 4047, 12 South Drive, Bethesda, Maryland 20892 USA.

We investigated ancestry of 3,528 modern humans from 163 samples. We identified 19 ancestral components, with 94.4% of individuals showing mixed ancestry. After using whole genome sequences to correct for ascertainment biases in genome-wide genotype data, we dated the oldest divergence event to 140,000 years ago. We detected an Out-of-Africa migration 100,000–87,000 years ago, leading to peoples of the Americas, east and north Asia, and Oceania, followed by another migration 61,000–44,000 years ago, leading to peoples of the Caucasus, Europe, the Middle East, and south Asia. We dated eight divergence events to 33,000–20,000 years ago, coincident with the Last Glacial Maximum. We refined understanding of the ancestry of several ethno-linguistic groups, including African Americans, Ethiopians, the Kalash, Latin Americans, Mozabites, Pygmies, and Uygurs, as well as the CEU sample. Ubiquity of mixed ancestry emphasizes the importance of accounting for ancestry in history, forensics, and health.

Several diversity projects have been performed to investigate the ability of genetic data to reveal the migratory history and geographical structuring of modern human populations. The recent origin of modern humans is widely thought to reflect migration(s) from sub-Saharan Africa, with gene flow estimated to have ended anywhere from 140,000 to 12,000 years ago[1–4]. Li *et al.*[5] focused on continental-level ancestry, identifying seven ancestral components: sub-Saharan Africa, the Middle East, Europe, south and central Asia, east Asia, Oceania, and (Native) America. Following a more detailed characterization of the genetic history of African peoples[6], these results were refined into 14 ancestral components: Fulani, Cushitic, Nilo-Saharan, Chadic-Saharan, Niger-Kordofanian, Southern African/Khoesan/Mbuti, western Pygmy, Hadza, and Sandawe ancestral components in Africa and Oceanian, European, Indian, Native American, and East Asian ancestral components in the rest of the world.

Here, we meta-analyzed ancestry from 12 global and regional diversity projects[5,7–17]. We collected genome-wide genotype data for 3,528 unrelated individuals from 163 samples from around the world (Fig. 1). Our analysis revealed 19 ancestral components, providing greater resolution of ancestry worldwide. Our inferred African ancestral components were largely consistent with the earlier results for sub-Saharan Africa[6], with the notable addition of Omotic-speaking peoples in Ethiopia. Using whole genome sequence data, we corrected for ascertainment biases in chip-based genotype data in estimation of genetic differentiation and heterozygosity. We then estimated the divergence times of the ancestral components and compared these divergence times to historical records. We observed that multiple divergence events coincided with the Last Glacial Maximum. The oldest divergence event dated to ~140,000 years ago.

Results and Discussion

Unsupervised ancestry analysis of 19,372 autosomal single nucleotide polymorphisms genotyped for 3,528 individuals from 163 samples revealed 19 ancestral components (Fig. 2, Table 1, and Supplementary Fig. 1). The 19 identified ancestral components were Click Speaker in south Africa; Pygmy in central Africa; Niger-Congo across west, east, and south Africa; Lowland East Cushitic, Nilo-Saharan, and Omotic in east Africa; Berber in north Africa; Indian and Kalash in south Asia; Chinese, Japanese, and southeast Asian in east Asia; Siberian in north Asia; Native American in the Americas; Melanesian in Oceania; southern and northern European; and

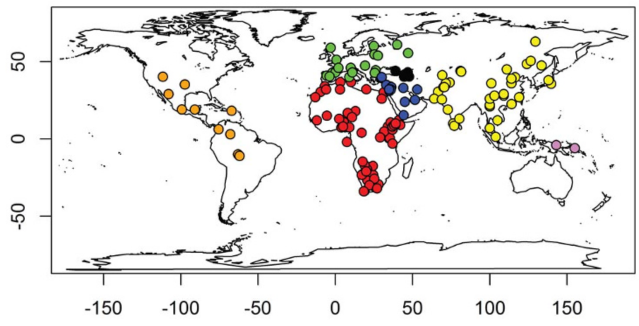

Figure 1 | Global distribution of samples. Red represents Africa, orange represents the Americas, yellow represents Asia, black represents the Caucasus, green represents Europe, blue represents the Middle East, and violet represents Oceania. The map was drawn using the R library maps.

Arabian and Levantine-Caucasian in the Middle East and in the Caucasus (Fig. 2 and Fig. 3). Consistent with prior findings[6], 94.4% of individuals had mixed ancestry, independent of self-identified ethno-linguistic group labels. Based on the estimated standard errors, our analysis was powered to detect an ancestral component present at a proportion of at least 2.5%.

Traditional analysis of F_{ST} between samples is complicated by recent admixture. In contrast, ancestral components are constructed to be ancestrally homogeneous and consequently unaffected by recent admixture. Therefore, we analyzed F_{ST} between ancestral components. Using hierarchical clustering analysis, the six sub-Saharan ancestral components clustered together; the south Asian ancestral components clustered with the European, Middle Eastern, Caucasian, and Berber ancestral components; and the east Asian ancestral components clustered with the north Asian, Native American, and Oceanic ancestral components (Fig. 4). To assess

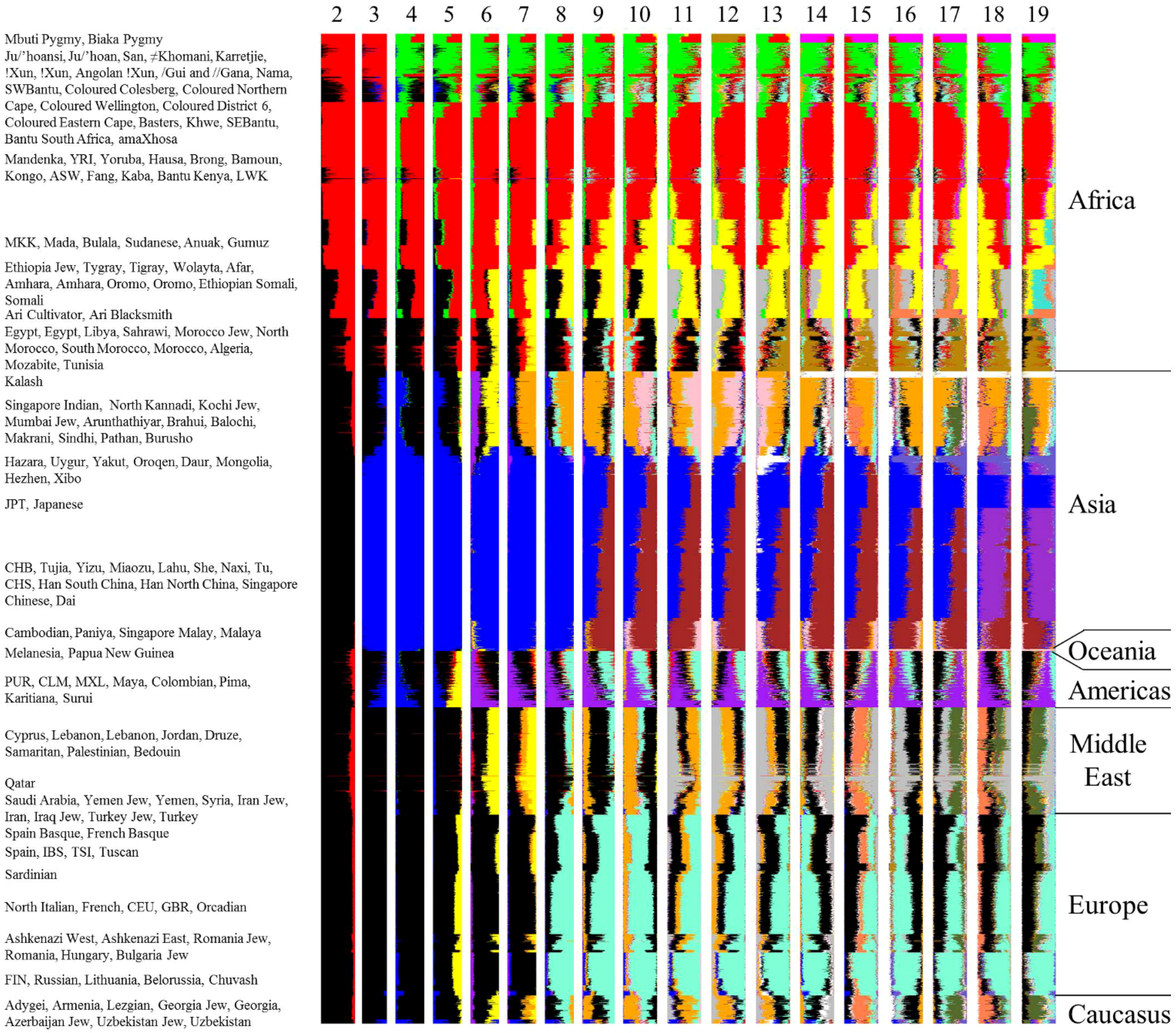

Figure 2 | Ancestry analysis of the global data set. The 163 samples are labeled in the left margin, the numbers of ancestral components are labeled in the top margin, and the geographical sample origins are labeled in the right margin. In the plot with 19 ancestral components, the ancestral components from top to bottom are Pygmy (magenta), Click Speaker (green), Niger-Congo (red), Nilo-Saharan (yellow), Lowland East Cushitic (turquoise), Omotic (coral), Berber (dark goldenrod), Kalash (white), Indian (orange), Siberian (slate blue), Japanese (blue), Chinese (dark orchid), Southeast Asian (brown), Melanesian (pink), Native American (purple), Levantine-Caucasian (dark olive green), Arabian (gray), Southern European (black), and Northern European (aquamarine).

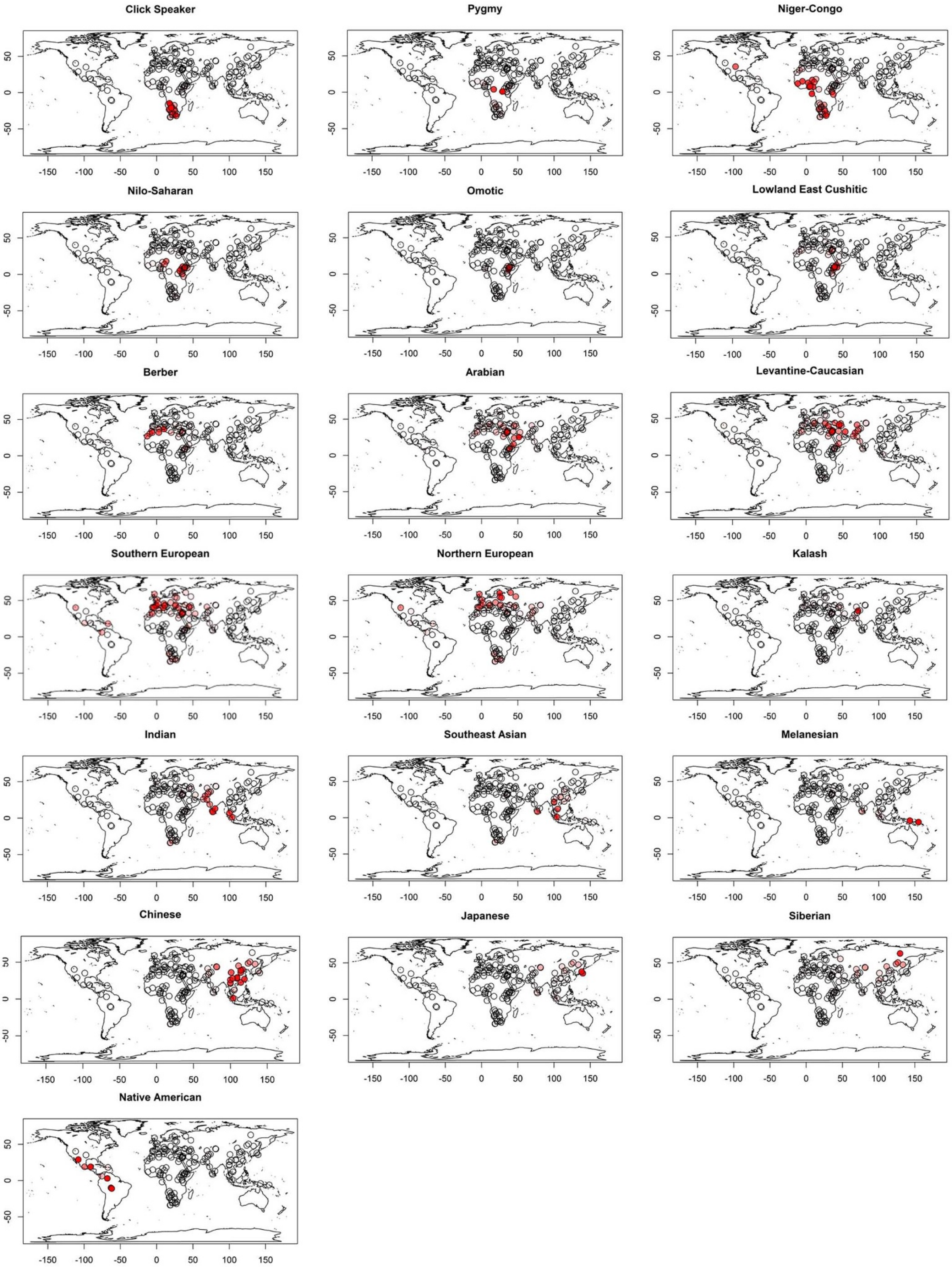

Figure 3 | **Global distribution of the ancestral components.** Each sample is represented by a circle. The intensity of the red color is directly proportional to the percentage for the ancestral component in the sample. Maps were drawn using the R library maps.

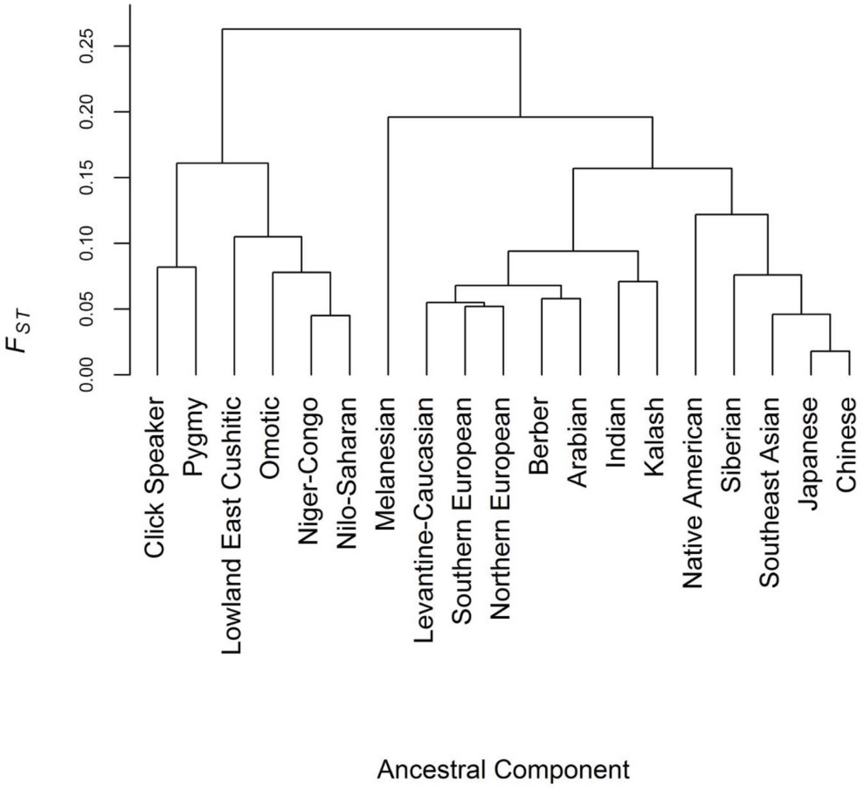

Figure 4 | Dendrogram of ancestral components by F_{ST}. The plot was drawn using hierarchical cluster analysis with complete linkage.

ascertainment bias in these F_{ST} estimates resulting from the use of chip-based genotype data, we used the 1000 Genomes sequence data. Since the 1000 Genomes samples showed heterogeneous ancestry, we limited this comparison to the JPT and YRI samples, both of which had only one ancestry (Japanese and Niger-Congo, respectively) and the FIN sample, which was the least ancestrally heterogeneous sample from Europe; that is, the FIN, JPT, and YRI samples and the Northern European, Japanese, and Niger-Congo ancestral components represented the closest matches between sequenced samples and ancestral components. F_{ST} values for the FIN/YRI, FIN/JPT, and JPT/YRI pairs were 0.0754, 0.0524, and 0.0879, respectively. In comparison, F_{ST} values for the Northern European/Niger-Congo, Northern European/Japanese, and Japanese/Niger-Congo pairs were 0.163, 0.121, and 0.177, respectively. Thus, we estimated that pairwise F_{ST} values between ancestral components were inflated by an average of 2.16-fold. To account for this inflation, we divided all pairwise F_{ST} values between ancestral components by 2.16.

To estimate divergence times from F_{ST}, we need estimates of the effective population size, N_e. Given allele frequencies per marker per ancestral component, we first estimated heterozygosity for each ancestral component. Heterozygosity estimates ranged from 0.255 to 0.327 (Table 2), similar to the range of 0.20 to 0.31 for the 52 samples in the Human Genome Diversity Project[5]. To assess ascer-

Table 1 | Ancestral components and proxy samples

Ancestral Component	Exemplar[a]	Proportion[b]
Arabian	Qatari	91.5%
Berber	Tunisia	79.8%
Chinese	She	79.0%
Click Speaker	Ju/'hoan	96.5%
Indian	Arunthathiyar	75.5%
Japanese	JPT	86.3%
Kalash	Kalash	93.9%
Levantine-Caucasian	Georgia	50.0%
Lowland East Cushitic	Somali	56.2%
Melanesian	Melanesian	100%
Native American	Surui/Karitiana	100%
Niger-Congo	Yoruba	86.8%
Nilo-Saharan	Anuak	78.3%
Northern European	FIN	74.6%
Omotic	Ari Blacksmith	94.9%
Pygmy	Mbuti Pygmy	99.1%
Siberian	Yakut	87.3%
Southeast Asian	Singapore Malay	65.3%
Southern European	Sardinian	64.1%

[a] "Exemplar" refers to the sample with the highest proportion of the given ancestral component.
[b] "Proportion" refers to the percentage of the ancestral component in the exemplar.

Table 2 | Heterozygosity by ancestral component

Ancestral Component	Heterozygosity
Arabian	0.314
Berber	0.324
Chinese	0.299
Click Speaker	0.269
Indian	0.321
Japanese	0.300
Kalash	0.308
Levantine-Caucasian	0.320
Lowland East Cushitic	0.327
Melanesian	0.255
Native American	0.275
Niger-Congo	0.318
Nilo-Saharan	0.316
Northern European	0.317
Omotic	0.318
Pygmy	0.289
Siberian	0.296
Southeast Asian	0.297
Southern European	0.314

Table 3 | Heterozygosity and effective population size (N_e) estimates based on whole genome sequence data

Sample	Heterozygosity for Polymorphic Sites	Heterozygosity for All Sites	N_e based on θ_W[a]	N_e based on θ_{Het}[a]
ASW	0.150	0.000952	30,859	21,649
CEU	0.190	0.000717	17,084	16,312
CHB	0.190	0.000672	15,577	15,280
CHS	0.190	0.000671	15,462	15,254
CLM	0.167	0.000769	22,443	17,486
FIN	0.195	0.000719	16,404	16,364
GBR	0.188	0.000721	17,179	16,394
IBS	0.264	0.000706	19,125	16,052
JPT	0.198	0.000673	15,289	15,311
LWK	0.142	0.000966	30,090	21,982
MXL	0.172	0.000736	20,451	16,750
PUR	0.167	0.000784	23,278	17,830
TSI	0.179	0.000723	17,795	16,446
YRI	0.154	0.000954	27,834	21,708

[a] θ_W refers to Watterson's estimator of the population-scaled mutation rate θ, which is frequency-independent. θ_{Het} refers to the estimator based on heterozygosity, which is frequency-dependent.

tainment bias in our heterozygosity estimates, we again used the 1000 Genomes sequence data. Across all 14 of the 1000 Genomes samples, ascertainment for common variation compared to all variation resulted in slight overestimation of heterozygosity, with heterozygosity for polymorphic markers ranging from 0.142 to 0.264 (Table 3). Ascertainment for variation resulted in massive overestimation of heterozygosity, with heterozygosity for all sites ranging from 0.000671 to 0.000966 (Table 3). Rather than attempting to correct the heterozygosity for the ancestral components in light of these ascertainment biases, we estimated the inbreeding effective population size based on heterozygosity for the 1000 Genomes samples (Table 3). We used the average N_e of 21,780 from the ASW, LWK, and YRI samples for the Click Speaker, Lowland East Cushitic, Niger-Congo, Nilo-Saharan, Omotic, and Pygmy ancestral components, the average N_e of 15,281 from the CHB, CHS, and JPT samples for the Chinese, Japanese, Melanesian, Native American, Siberian, and southeast Asian ancestral components, and the average N_e of 16,446 from the CEU, FIN, GBR, IBS, and TSI samples for the Arabian, Berber, Indian, Kalash, Levantine-Caucasian, northern European, and southern European ancestral components. If these N_e values are too small for any ancestral component, then divergence times will be underestimated. Conversely, if these N_e values are too large, then divergence times will be overestimated.

After correcting the pairwise F_{ST} values between ancestral components for ascertainment bias as described above, we estimated divergence times using the three sequence-based N_e values. Mean divergence times for the ancestral components ranged from 256 generations to 4,664 generations (Fig. 5), corresponding to ~7,700 to ~140,000 years ago, assuming a generation time of 30 years[18,19]. Note that the order of appearance of ancestral components in the

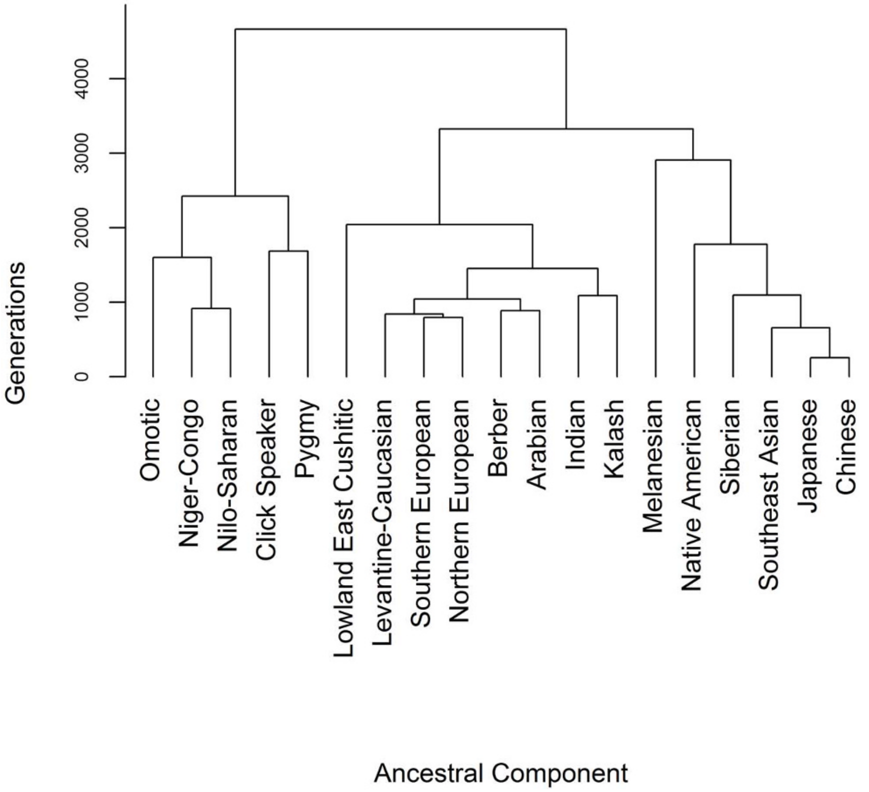

Figure 5 | Dendrogram of ancestral components by generations since divergence. The plot was drawn using hierarchical cluster analysis with complete linkage.

ADMIXTURE analysis (Fig. 2) reflects a composite of individuals' ancestry proportions and ancestry-specific allele frequencies, the order of divergence of ancestral components by F_{ST} (Fig. 4) reflects a composite of ancestry-specific allele frequencies and time, and the order of divergence of ancestral components by time (Fig. 5) reflects only time.

At the global scale, the oldest divergence event dated to 4,664 generations or ~140,000 years ago (Fig. 5). This time is consistent with estimates of the coalescence time for the major haplogroups of the Y chromosome of 138,000 years ago[20] and 142,000 years ago[21] as well as an estimate of ~140,000 years ago for African *vs*. Eurasian divergence based on multilocus resequencing[4]. The next divergence occurred 3,326 generations or ~100,000 years ago giving rise to the cluster of east and north Asian, Native American, and Oceanic ancestral components (Fig. 5). A separate divergence event occurred 2,041 generations or ~61,000 years ago giving rise to Caucasian, European, Middle Eastern, and south Asian ancestral components (Fig. 5). We detected two Out-of-Africa migrations principally due to the inclusion of samples allowing for the inference of a Lowland East Cushitic ancestral component (Supplementary Fig. 2). If we assume an African origin for the Lowland East Cushitic ancestral component, then these results are consistent with an Out-of-Africa migration giving rise to east and north Asian/Native American/Oceanic ancestral components, followed by another Out-of-Africa migration giving rise to Caucasian/European/Middle Eastern/south Asian ancestral components, followed by back migration to north Africa giving rise to the Berber ancestral component. Alternatively, if we assume a non-African origin for the Lowland East Cushitic ancestral component, then these results are consistent with an Out-of-Africa migration giving rise to east and north Asian/Native American/Oceanic ancestral components, followed by back migration into Africa, followed by an Out-of-Africa migration giving rise to Caucasian/European/Middle Eastern/south Asian ancestral components, followed by another back migration to north Africa giving rise to the Berber ancestral component. The former interpretation is more parsimonious.

The rate of admixture of archaic lineages into modern humans has been estimated to be higher in East Asians than in Europeans[22]. Furthermore, the maximum-likelihood estimates of the times of admixture of archaic lineages are 55,100 years ago for Europeans and 75,800 years ago for East Asians[23]. We detected an Out-of-Africa migration 100,000–87,000 years ago, leading to peoples of the Americas, east and north Asia, and Oceania. We also detected another migration 61,000–44,000 years ago, leading to peoples of the Caucasus, Europe, the Middle East, and south Asia. Taken together, these results suggest that introgression of archaic lineages occurred at two different times and places: an older event in East Asia involving migrants from the first Out-of-Africa migration and a more recent event in the Middle East before dispersal of migrants from the second Out-of-Africa migration into the Caucasus, Europe, and south Asia.

Africa. Sub-Saharan ancestral components diverged 2,426 generations or ~73,000 years ago (Fig. 5). The Pygmy ancestral component diverged 1,686 generations or ~51,000 years ago from the Click Speaker ancestral component (Fig. 5). The Mbuti Pygmy sample had 99.1% ± 3.1% (mean ± standard error) Pygmy ancestry (Supplementary Table 1), indicating ancestral homogeneity and implying a lack of admixture. The Biaka Pygmy sample showed evidence of admixture, with 77.9% ± 3.3% Pygmy ancestry and 21.6% ± 2.9% Niger-Congo ancestry (Supplementary Table 1). These results are consistent with a higher level of gene flow between western Pygmies (*e.g.*, Biaka Pygmies) and agricultural populations than between eastern Pygmies (*e.g.*, Mbuti Pygmies) and agricultural populations[24]. In contrast, using the Yoruba and San samples and assuming two-way admixture, Loh *et al.*[25] inferred that the Mbuti Pygmy sample showed evidence of admixture ~ 28 generations ago with ~15.9% Yoruba-related ancestry and that the Biaka Pygmy sample showed evidence of admixture ~ 38 generations ago with ~28.8% Yoruba-related ancestry. Use of divergent reference samples for parental populations of admixed samples leads to estimation of admixture proportions that are biased towards equal proportions for all referent samples and estimation of generations since admixture that are upward biased. We also detected small amounts of Pygmy ancestry in multiple samples throughout central and south Africa (Fig. 3, Supplementary Fig. 3, and Supplementary Table 1). The Click Speaker ancestral component was the major ancestral component in several Khoesan samples from south Africa (Fig. 3, Supplementary Fig. 3, and Supplementary Table 1). The Ju/'hoan sample had 96.5% ± 2.2% Click Speaker ancestry and the San sample had 94.8% ± 2.0% Click Speaker ancestry and 5.2% ± 2.0% Pygmy ancestry, whereas the other Khoesan samples had ≤75.0% Click Speaker ancestry and various amounts of other ancestries, most notably Niger-Congo ancestry (Supplementary Table 1).

The Omotic ancestral component diverged from the sub-Saharan cluster 1,602 generations or ~48,000 years ago (Fig. 5). The Omotic ancestral component showed a distribution mostly limited to Ethiopia (Fig. 3 and Supplementary Fig. 3). The majority of the ancestry of the Ari Blacksmith and Ari Cultivator samples was Omotic (Supplementary Table 1). The Omotic ancestral component was also the largest component in the Wolayta sample (Supplementary Table 1).

The Niger-Congo ancestral component included non-Bantu speakers from Senegambia and Nigeria as well as Bantu speakers from east and south Africa (Fig. 3, Supplementary Fig. 3, and Supplementary Table 1). Several samples from South Africa, such as amaXhosa, showed mixed ancestry between Click Speaker and the Niger-Congo components, consistent with linguistic evidence that isiXhosa is a language in the Niger-Congo family with ~15% Khoekhoe vocabulary [http://www.ethnologue.com/language/xho]. The Niger-Congo and Nilo-Saharan ancestral components diverged 917 generations or ~28,000 years ago (Fig. 5), possibly reflecting expansion of the Sahara around the time of the Last Glacial Maximum[26]. The Nilo-Saharan ancestral component was the major component in the Anuak, Sudanese, Gumuz, and Bulala samples across Chad, South Sudan, and Ethiopia (Fig. 3, Supplementary Fig. 3, and Supplementary Table 1). The clustering of the Niger-Congo and Nilo-Saharan ancestral components is consistent with grouping in the Niger-Congo family Kordofanian languages that are spoken in the Nuba Mountains in what is presently the Republic of the Sudan.

The Lowland East Cushitic ancestral component was the major ancestral component in Somali from Ethiopia and Somalia (Fig. 3, Supplementary Fig. 3, and Supplementary Table 1), but it may be capturing some Central Cushitic ancestry if the Afar sample is actually Agaw (the sample was collected from the Wag Hemra Zone and the language was listed as Xamtan[8]). Lowland East Cushitic ancestry diverged from the Caucasian/European/Middle Eastern/south Asian cluster 2,041 generations or ~61,000 years ago (Fig. 5). The MKK (Maasai in Kinyawa, Kenya) sample showed mostly Nilo-Saharan ancestry, some Lowland East Cushitic ancestry, and smaller amounts of Niger-Congo and Click Speaker ancestry, whereas the LWK (Luhya in Webuye, Kenya) and BantuKenya samples showed predominantly Niger-Congo ancestry, some Nilo-Saharan ancestry, and a small amount of Pygmy ancestry, but no Lowland East Cushitic ancestry (Supplementary Fig. 3 and Supplementary Table 1).

All of the north African samples showed significant amounts of Berber ancestry (Fig. 3, Supplementary Fig. 3, and Supplementary Table 1), presumably reflecting Imazighen peoples. The Berber and Arabian ancestral components diverged 888 generations or ~27,000 years ago (Fig. 5). This divergence time is ~21,000 years before the E-M81 or E1b1b1b Y chromosome haplogroup (referred to as the

Berber marker) originated in north Africa[27,28]. The Berber ancestral component clustered with the Caucasian/European/Middle Eastern/south Asian ancestral components, not with the sub-Saharan ancestral components (Fig. 5). We detected Niger-Congo ancestry (7.6%) but no European ancestry in the Mozabite sample (Supplementary Table 1), inconsistent with admixture between individuals with ancestry similar to the YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah Residents with Northern and Western European Ancestry) ~ 100 generations ago[29].

Our data set included five samples of South African Coloureds, one from the Eastern Cape, two from the Northern Cape, and two from the Western Cape. Whereas all five samples showed European ancestry, the samples from the Western Cape showed more Indian, Melanesian, and southeast Asian ancestry whereas the samples from the Eastern and Northern Capes showed more Click Speaker and Niger-Congo ancestry (Fig. 3 and Supplementary Table 1). Our data set also included the admixed African American sample ASW (Americans of African Ancestry in SW USA). Niger-Congo ancestry represented the major African ancestry in the ASW, but we also detected a significant amount of Pygmy ancestry (Supplementary Table 1). No Pygmy ancestry was detected in either sample of Yoruba individuals (Supplementary Table 1), indicating that the Yoruba and YRI samples are not adequate proxies of African ancestry in the ASW sample and therefore possibly inadequate for other samples of African Americans. In our data set, there is no single sample that might serve as a better proxy; therefore, we suggest adding Western Pygmies (*e.g.*, the Biaka Pygmy sample) as an additional parental population for ancestry analysis of African Americans.

The Amhara, Oromo, and Wolayta samples from Ethiopia had Lowland East Cushitic, Nilo-Saharan, Omotic, and Arabian ancestry, and the Tygray sample also had a small amount of Levantine-Caucasian ancestry (Supplementary Table 1). These samples of Ethiopians had no Niger-Congo or European ancestry (Supplementary Table 1). These results indicate that the YRI and CEU samples are not optimal choices as proxies for the parental populations of Ethiopians. Furthermore, these Ethiopian samples have four or five ancestries and therefore should not be modeled by two-way admixture. As with the Mozabite sample, use of the YRI and CEU samples as proxies for the parental populations for the Ethiopians will lead to reconstruction of excessively short haplotypes, estimation of excessively long times since admixture began, and poor estimates of admixture proportions.

Previously, nine ancestral components were identified among Africans: Chadic-Saharan, Cushitic, Fulani, Hadza, Niger-Kordofanian, Nilo-Saharan, Sandawe, Southern African/Khoesan/Mbuti, and western Pygmy[6]. In comparison, we identified seven ancestral components: Berber, Click Speaker, Lowland East Cushitic, Niger-Congo, Nilo-Saharan, Omotic, and Pygmy. Our data set lacked samples of Chadic speakers, Fulani, Hadza, and Sandawe but included samples of Berbers and Omotic speakers. Our data set included more Khoesan samples, revealing divergence between Click Speaker and Pygmy ancestral components, implying a more recent divergence of eastern *vs*. western Pygmy[24]. The other ancestral components appear directly comparable.

The Americas, Asia, and Oceania. Ancestral components in Asia grouped into two clusters: one in south Asia containing the Indian and Kalash ancestral components and the other in east and north Asia containing the Native American, Melanesian, Siberian, Southeast Asian, Chinese, and Japanese ancestral components (Fig. 5). The south Asian ancestral components diverged from the Caucasian/European/Middle Eastern ancestral components 1,452 generations or ~44,000 years ago (Fig. 5). Kalash and Indian ancestral components subsequently diverged 1,090 generations or ~33,000 years ago (Fig. 5). The Kalash ancestral component predominantly identified the Kalash sample and appeared in small amounts (<10%) in any other sample (Fig. 3, Supplementary Fig. 4, and Supplementary Table 1). This result is consistent with the Kalash people representing a population isolate. We detected no evidence of Arabian or southern European ancestry (Supplementary Table 1), indicating that the Kalash people are not of Arab or Greek origin. The Indian ancestral component was detected in several samples throughout central and south Asia, the Middle East and the Caucasus, and South Africa (Fig. 3, Supplementary Figs. 3, 4, and 5, and Supplementary Table 1).

The Melanesian ancestral component diverged 2,907 generations or ~87,000 years ago (Fig. 5). The Melanesian ancestral component was the major component in the two samples from Oceania and was present in small amounts in samples from Singapore, India, and South Africa (Fig. 3, Supplementary Figs. 3 and 4, and Supplementary Table 1), suggesting some degree of representation of island southeast Asia as well as Oceania. The Native American ancestral component diverged 1,777 generations or ~53,000 years ago (Fig. 5). This divergence time predates most estimates of the time(s) of the crossing of Beringia, consistent with isolation in Beringia prior to migration to the Americas. The Native American ancestral component was the major component in several samples from the Americas and was undetected in all east Asian and European samples (Fig. 3, Supplementary Figs. 4 and 5, and Supplementary Table 1)[30]. The Siberian ancestral component was the next to diverge, 1,095 generations or ~33,000 years ago (Fig. 5). The Siberian ancestral component was predominant in the Yakut sample, with a significant presence in several samples from Manchuria, Mongolia, and north China (Fig. 3, Supplementary Fig. 4, and Supplementary Table 1).

The southeast Asian, or perhaps more precisely mainland southeast Asian, ancestral component diverged 658 generations or ~20,000 years ago (Fig. 5). Wangkumhang *et al.*[31] also identified one major ancestral component common to four Thai populations. Chinese and Japanese ancestral components diverged 256 generations or ~7,700 years ago (Fig. 5). The Chinese ancestral component was the major ancestral component in several samples from both south and north China (Fig. 3, Supplementary Fig. 4, and Supplementary Table 1). The Japanese ancestral component was the major component only in the two samples from Japan (Fig. 3, Supplementary Fig. 4, and Supplementary Table 1).

The Uygur sample showed highly heterogeneous ancestry: 20.8% Chinese, 18.0% Siberian, and 9.7% Japanese; 9.4% Indian and 4.6% Kalash; and 15.4% Levantine-Caucasian and 12.3% northern European (Supplementary Fig. 4 and Supplementary Table 1). These proportions indicate south Asian and Middle East/Caucasus ancestry in addition to east Asian and European ancestry, consistent with trade on the Silk Road. The non-Jewish Uzbekistan and Hazara samples showed similar ancestry to the Uygur sample (Supplementary Figs. 4 and 5 and Supplementary Table 1).

The CLM (Colombians from Medellín, Colombia), MXL (Mexican Ancestry from Los Angeles, USA), and PUR (Puerto Ricans from Puerto Rico) samples from the Americas all showed mixtures of predominantly Native American and European ancestry with <10% Niger-Congo ancestry (Supplementary Fig. 4 and Supplementary Table 1). Additionally, the PUR sample showed a significant amount of Berber ancestry, which likely did not derive from a Spanish parental population as none of the three Spanish samples (Spain_Basque, IBS (Iberian population in Spain), and Spain) showed significant amounts of Berber ancestry (Supplementary Fig. 4 and Supplementary Table 1)[32]. Furthermore, the CLM and PUR samples showed more Arabian ancestry plus Berber ancestry than the MXL sample (7.9% and 10.8% *vs.* 5.3%, respectively). Given that Arabian and Berber ancestral components cluster with European ancestral components, divergence of the "Latino-specific European component" from the presumed Iberian parental populations may reflect imprecise usage of "European ancestry"[32].

Europe. The Levantine-Caucasian and European ancestral components diverged 842 generations or ~25,000 years ago (Fig. 5). Northern and southern European ancestral components subsequently diverged 795 generations or ~24,000 years ago (Fig. 5). The northern European ancestral component was the major ancestral component in samples from Finland, Lithuania, Russia, and Belorussia (Fig. 3, Supplementary Fig. 5, and Supplementary Table 1). The northern European ancestral component clustered with the Caucasian/European/Middle Eastern/south Asian ancestral components (Fig. 5), inconsistent with an origin of northern European ancestry in north Asia. However, Siberian ancestry was detected in the Russian and FIN (Finnish in Finland) samples (6.1% and 4.2%, respectively, Supplementary Table 1), consistent with a small amount of westward migration from Siberia to north Europe. The Spanish and Italian samples showed southern and northern European ancestry with varying amounts of Levantine-Caucasian, Arabian, and Berber ancestry (Supplementary Fig. 5 and Supplementary Table 1). In contrast, the Basque samples showed only southern and northern European ancestry (Supplementary Fig. 5 and Supplementary Table 1), consistent with genetic isolation. Also, we detected more Arabian than Berber ancestry in Spain and Italy[33]. The oft-used CEU sample showed northern European, southern European, and Levantine-Caucasian ancestry, similar to the GBR (British in England and Scotland) and French samples (Supplementary Fig. 5 and Supplementary Table 1).

The Middle East and the Caucasus. Arabian and Levantine-Caucasian ancestral components diverged 1,044 generations or ~31,000 years ago (Fig. 5). The Arabian ancestral component was the major ancestral component in the Qatari and Bedouin samples (Supplementary Fig. 5 and Supplementary Table 1). The Arabian ancestral component had a decreasing presence westward across north Africa (Fig. 3).

The Levantine-Caucasian ancestral component was the major ancestral component in only the Georgia sample, but held a plurality in several samples across the Middle East and the Caucasus and was detected in south Asian samples (Fig. 3, Supplementary Figs. 4 and 5, and Supplementary Table 1). The sample of Ethiopian Jews lacked Levantine-Caucasian ancestry but had ancestry similar to the Amhara (Supplementary Table 1), consistent with conversion of indigenous Ethiopians to Judaism. Similarly, the Kochi Jews and Mumbai Jews had large amounts of Indian ancestry (Supplementary Table 1), consistent with conversion. In contrast, Moroccan Jews differed from the other samples from Morocco by having Levantine-Caucasian ancestry but less Berber ancestry (Supplementary Table 1), consistent with migration of Jewish people. Paired analysis of Jews and non-Jews were available for seven countries: Georgia, Iran, Morocco, Romania, Turkey, Uzbekistan, and Yemen. Compared to non-Jews, Jews had more Southern European ancestry (21.9% *vs.* 13.8%), Arabian ancestry (18.9% *vs.* 10.8%), Levantine-Caucasian ancestry (33.5% *vs.* 27.8%), and Lowland East Cushitic ancestry (4.4% *vs.* 2.5%)[34].

To contextualize these findings, six points should be kept in mind. One, markers were not ascertained for ancestry informativeness. However, markers were ascertained for common polymorphisms. Using whole genome sequence data, we estimated and corrected for the effects of ascertaining for (1) common *vs.* lower frequency polymorphisms and (2) segregating sites. Two, genetic history revealed by autosomal markers need not be identical to genetic histories of uniparentally inherited markers (the Y chromosome or mitochondria). Three, estimated times since divergence of ancestral components assumed the absence of gene flow. These times more likely reflect the recent past than the distant past. Four, genetics and self-identified ethno-linguistic labels do not perfectly correlate. Five, unsupervised ancestry analysis does not require the investigator to choose external reference samples to serve as proxies of parental populations for putative admixed samples and is amenable as-is for analysis of multi-way ancestry. Importantly, unsupervised ancestry analysis takes advantage of ancestry across the entire data set, increasing confidence by increasing the effective sample size by ancestral component. This can be seen by noting that the average number of individuals per sample was 21.6 whereas the average number of individuals per ancestral component was 185.7. However, unsupervised ancestry analysis does not allow for exact identification of parental populations in terms of real-world samples. Six, the time period of history revealed by our data set is the Late Pleistocene. That is, our conclusions are unaffected by recent population growth during the Holocene. Furthermore, the inbreeding effective population size captures the effects of bottlenecks.

In summary, we showed that ancestry of modern humans covered 140,000 years of history, with two major Out-of-Africa migrations. Eight divergence times occurred between ~33,000 to ~20,000 years ago, coinciding with the Last Glacial Maximum. We recommend that ancestry analyses should be globally comprehensive, even if interest is regional, because redefining an existing ancestral component or defining a new ancestral component will impact the definitions of other ancestral components. Characterization of human ancestry is ongoing as sampling of some ancestries is poorer than others. To name a few examples, the Melanesian ancestral component has the lowest effective sample size, Chadic- and Cushitic-speaking peoples are not well represented in our data set, and Polynesian samples are absent. In contrast, some ancestries are well sampled, including Chinese. We anticipate that most unsampled lineages reflect recent divergence events. However, it is possible that an unsampled lineage could reflect a divergence event older than 140,000 years. Also, the limited density of markers precludes accurate dating of potential admixture events because many ancestral switches will be missed. Our findings strongly inform control for population stratification in genetic association studies and inference of local ancestry in admixed individuals. Shared ancestry provides another layer of insight into human evolution, particularly with respect to migrations.

Methods

We collected genome-wide genotype data on autosomal single nucleotide polymorphisms (SNPs) from publicly available human genomic diversity projects. The global data set included 916 individuals from the Human Genome Diversity Project[5], 1,092 individuals from the 1000 Genomes Project[7], 222 individuals from east Africa[8], 268 individuals from the Singapore Genome Variation Project[9], 75 individuals from Lebanon[10], 145 individuals from north Africa and the Basque Country[11], 323 individuals from south Africa[12,13], 18 Arabs from Qatar[14], 106 individuals from west and central Africa[15], 133 Maasai from the International HapMap Project[16], and 462 individuals from a study of the Jewish Diaspora[17]. Data management and quality control were performed using PLINK version 1.07[35]. Graphics were generated using R[36]. Maps were drawn using the R libraries maps and plotrix.

Individuals or markers with genotyping call rates $<$ 95% were excluded. We also removed individuals identified as identical samples, 1st degree relatives, or 2nd degree relatives. After quality control, the global data set comprised 3,528 individuals from 163 samples. The mutual intersection of all data sets yielded 19,372 diallelic, autosomal SNPs with experimentally determined genotypes (*i.e.*, no imputation of missing genotypes was performed). The genotyping call rate in the remaining individuals was 99.8%. The average distance between markers was 142.8 kb (135.4 kb excluding centromeres). Due to very small sample sizes for some samples, no additional pruning of markers based on linkage disequilibrium was performed.

Principal components analysis was first performed on the cleaned data set of 3,528 individuals and 19,372 SNPs to confirm the expected continental-level structure (Supplementary Fig. 6)[37]. We then performed unsupervised ancestry analysis using ADMIXTURE[38] with the number of ancestral components *K* ranging from 1 to 30. The optimal value of *K* was determined by five-fold cross-validation, averaged over three runs with different starting seeds. For each ancestral component, the sample with the largest proportion of that ancestral component was identified as an exemplar. Conditioned on the optimal value of *K*, ADMIXTURE analysis was repeated with the addition of 200 bootstrap replicates to obtain standard errors for the proportions of ancestral components for each individual. Average ancestry proportions and 95% confidence intervals for each sample were calculated accounting for both within and between individual variance. Average proportions for which the 95% confidence intervals included 0 were zeroed out (Supplementary Table 1).

ADMIXTURE produces two files: the .P file contains an estimated allele frequency for each marker for each ancestral component and the .Q file contains an estimated proportion for each individual for each ancestral component. Heterozygosity for each marker within each ancestral component was estimated from the .P file. The mean heterozygosity for each ancestral component was estimated by averaging heterozygosity across all markers. ADMIXTURE reports pairwise divergence between each ancestral component as assessed by F_{ST} but without accompanying confidence intervals (Supplementary Table 2). These confidence intervals require estimates of the variances of the allele frequencies for each ancestral component.

To account for ascertainment biases in F_{ST} and heterozygosity estimated from chip-based genotype data, we estimated F_{ST} and heterozygosity using the 1000 Genomes sequence data[7] (a total of 36,820,992 variable sites across a total of 2,881,033,286 sites). We estimated pairwise F_{ST} using the definition $F_{ST} = \frac{H_T - H_S}{H_T}$, in which H_T is the mean of the expected heterozygosity across samples and H_S is the mean of the observed heterozygosity across samples[39]. This estimator of F_{ST} is robust to the proportion of polymorphic sites because H_T and H_S scale identically. We estimated the effective population size N_e within samples two different ways. One, we used the estimators $\theta = \frac{S}{a_1}$ and $N_e = \frac{\theta}{4\mu}$[39], in which S is the number of segregating sites, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$ for a sample size n, and $\mu = 1.1 \times 10^{-8}$ mutations/generation/site[40]. Two, we used the estimators $\theta = \frac{H}{1-H}$ and $N_e = \frac{\theta}{4\mu}$[39], in which H is the mean of the observed heterozygosity within the sample. Note that S does not make use of allele frequencies whereas H does. Pairwise divergence times between ancestral components were estimated using the relationship $1 - F_{ST} = \left(1 - \frac{1}{2N_e}\right)^t$[39], with t in generations and N_e being the harmonic mean for the two ancestral components being compared, assuming that $F_{ST} = 0$ at $t = 0$.

Ethics. This project was determined to be excluded from IRB Review by the National Institutes of Health Office of Human Subjects Research Protections, Protocol #12183.

1. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
2. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
3. Harris, K. & Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* **9**, e1003521 (2013).
4. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
5. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
6. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
7. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
8. Pagani, L. *et al.* Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* **91**, 83–96 (2012).
9. Teo, Y. Y. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19**, 2154–2162 (2009).
10. Haber, M. *et al.* Genome-wide diversity in the Levant reveals recent structuring by culture. *PLoS Genet.* **9**, e1003316 (2013).
11. Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **8**, e1002397 (2012).
12. Petersen, D. C. *et al.* Complex patterns of genomic admixture within southern Africa. *PLoS Genet.* **9**, e1003309 (2013).
13. Schlebusch, C. M. *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379 (2012).
14. Hunter-Zinck, H. *et al.* Population genetic structure of the people of Qatar. *Am. J. Hum. Genet.* **87**, 17–25 (2010).
15. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. USA* **107**, 786–791 (2010).
16. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
17. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010).
18. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
19. Tremblay, M. & Vézina, H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.* **66**, 651–658 (2000).
20. Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013).
21. Cruciani, F. *et al.* A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am. J. Hum. Genet.* **88**, 814–818 (2011).
22. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013).
23. Lohse, K. & Frantz, L. A. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics* **196**, 1241–1251 (2014).
24. Patin, E. *et al.* Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.* **5**, e1000448 (2009).
25. Loh, P.-R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013).
26. Clark, P. U. *et al.* The Last Glacial Maximum. *Science* **325**, 710–714 (2009).
27. Arredi, B. *et al.* A predominantly Neolithic origin for Y-chromosomal DNA variation in North Africa. *Am. J. Hum. Genet.* **75**, 338–345 (2004).
28. Cruciani, F. *et al.* Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.* **74**, 1014–1022 (2004).
29. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
30. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
31. Wangkumhang, P. *et al.* Insight into the peopling of mainland southeast Asia from Thai population genetic structure. *PLoS ONE* **8**, e79522 (2013).
32. Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9**, e1003925 (2013).
33. Botigué, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl. Acad. Sci. USA* **110**, 11791–11796 (2013).
34. Elhaik, E. The missing link of Jewish European ancestry: contrasting the Rhineland and the Khazarian hypotheses. *Genome Biol. Evol.* **5**, 61–74 (2013).
35. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
36. R Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing: Vienna, Austria, 2013).
37. Shriner, D. Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity* **107**, 413–420 (2011).
38. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
39. Hartl, D. L. *A Primer of Population Genetics*. (Third edn, Sinauer Associates, Inc.: Sunderland, Massachusetts, 2000).
40. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

Acknowledgments

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official view of the National Institutes of Health. This research was supported by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, the Center for Information Technology, and the Office of the Director at the National Institutes of Health (Z01HG200362).

Author contributions

D.S. conceived and designed the study. D.S., F.T.-A. and A.A. collected the data. D.S. performed the analyses and wrote the manuscript. D.S., F.T.-A., A.A. and C.N.R. interpreted the data, discussed the results, and commented on the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Shriner, D., Tekola-Ayele, F., Adeyemo, A. & Rotimi, C.N. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. *Sci. Rep.* **4**, 6055; DOI:10.1038/srep06055 (2014).