www.sciencemag.org/content/356/6337/543/suppl/DC1

Supplementary Materials for

Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America

Etienne Patin,* Marie Lopez, Rebecca Grollemund, Paul Verdu, Christine Harmant, Hélène Quach, Guillaume Laval, George H. Perry, Luis B. Barreiro, Alain Froment, Evelyne Heyer, Achille Massougbodji, Cesar Fortes-Lima, Florence Migot-Nabias, Gil Bellis, Jean-Michel Dugoujon, Joana B. Pereira, Verónica Fernandes, Luisa Pereira, Lolke Van der Veen, Patrick Mouguiama-Daouda, Carlos D. Bustamante, Jean-Marie Hombert, Lluís Quintana-Murci*

*Corresponding author. Email: epatin@pasteur.fr (E.P.); quintana@pasteur.fr (L.Q.-M.)

Published 5 May 2017, *Science* **356**, 543 (2017)
DOI: 10.1126/science.aal1988

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/356/6337/543/DC1)

Tables S1 to S12 (as Excel file)

Materials and Methods

Population samples. A large-scale sampling campaign was conducted in Gabon in 2003, by linguists specializing in Bantu languages. Blood samples were collected from 19 well-characterized Bantu-speaking populations (BSPs) (*26, 27*). Additional samples were collected to complete population coverage in western central Africa and the ports historically involved in the transatlantic slave trade, in Cameroon, Gabon, Angola, Benin and Ivory Coast. The geographic location, linguistic affiliation, lifestyle and sample sizes of all populations are presented in table S1. Informed consent was obtained from all participants. This study was approved by the institutional review board of Institut Pasteur, France (2011-54/IRB/6).

Genotyping and quality-controls. We genotyped 1,318 individuals from 35 populations with the Illumina HumanOmniExpress-12 array. Genotype clustering was performed with Genome Studio v.2011.1 (Illumina) with the manifest v.1.1 B. SNPs presenting low GenTrain scores (<0.35) were inspected and their genotype clusters were manually redefined, when applicable. Of the 719,665 SNPs genotyped, we excluded 1,204 unmapped SNPs, 16,092, 1,372 and 463 SNPs on the X chromosome, Y chromosome and pseudo-autosomal region, respectively, 2,122 C/G or A/T SNPs, 7,410 SNPs with call rate <95%, 2 SNPs with no matched position in dbSNP build 138, 3 duplicated SNPs, and 258 SNPs in Hardy-Weinberg disequilibrium (*i.e.*, *P*-value below 10^{-5} in one population, or below 0.05 in half of the populations). These filters yielded a final set of 690,739 SNPs. The rate of genotype concordance between 18 duplicate samples was 99.9995%.

We excluded 13 individuals with a genotyping call rate <95%, 18 technical duplicates, and 28 individuals with a heterozygosity at least 3 standard deviations (SD) lower or higher than the mean for the population, suggestive of inbreeding or DNA contamination, respectively. We also excluded 79 individuals because of cryptic relatedness. We considered two individuals to be cryptically related if they presented a first-degree or second-degree relationship, as inferred by KING (*28*). Because rainforest hunter-gatherer (RHG) populations are small isolated communities, and RHG individuals are typically related to many other RHG individuals, we implemented a specific algorithm to remove cryptically related individuals and maximise sample size. A score was first estimated for each individual, which equals the mean kinship coefficient of the individual with all other individuals from the population, divided by its genotyping call rate. At each iteration, the algorithm removes the individual with the highest score, and updates the score of all the remaining individuals accordingly. The algorithm stops when there are no more pairs of related individuals. The newly generated dataset comprised data for a final number of 1,180 filtered individuals, with an average call rate of 99.56%. The detailed number of samples excluded per population is reported in table S1.

Merging with other datasets. We merged our newly-generated dataset with six publically available datasets for African populations (table S2) (*7, 12, 29-32*). Each dataset was filtered for the individual and SNP quality-control filters described above, when necessary (fig. S1). When merging datasets, we retained only the SNPs common to all datasets and performed additional checks and filters. We accounted for misaligned SNPs by removing all C/G or A/T SNPs, and SNPs with highly discordant allele frequencies between closely-related populations that were present in different datasets. These SNPs were defined as SNPs associated ($P<10^{-7}$) with a binary phenotype corresponding to the datasets of origin of individuals, using the logistic regression model implemented in PLINK (*33*). If individuals were common to different datasets (*e.g.*, 74 individuals were common to this study and a previous study (*31*)), we

excluded SNPs with a discordant rate >5%. We also removed cryptically related individuals present in different datasets, using the algorithm described above. Such related individuals were observed between the RHG population samples collected here and those investigated in two previous studies (*31, 32*).

Two datasets were generated for subsequent analyses. The first dataset, *GabonDiv_PGSRP*, was used in most analyses (fig. S1), and comprised data for 2,055 individuals and 548,055 SNPs. It included datasets from five previous studies (*7, 12, 30-32*). The second dataset, *GabonDiv_PGSJ*, included an additional study of western central Africans (*29*), and was used in analyses focusing on the fine-scale genetic structure of western central Bantu-speaking populations (fig. S1). We did not use this latter dataset systematically, because 118,083 SNPs were lost when it was merged with *GabonDiv_PGSRP*.

We also merged our dataset with data from four cohorts of African American individuals (table S2). These cohorts were selected because genotype data was available from dbGAP for general medical research, and because they were genotyped with the Illumina 1M array, which includes a large number of SNPs common to our dataset. We removed individuals with a call rate <95% from each cohort, and those who did not declare themselves to be African Americans. We excluded cryptically related individuals in each cohort, or present in the GENEVA and CIDR Multiethnic cohorts, using the procedure described above. The final dataset, *GabonDiv_PGSRAA*, comprised data for 5,244 African Americans and 414,714 SNPs (fig. S1).

Linkage disequilibrium and runs of homozygosity. To provide clues on differences in effective population sizes between Bantu-speaking populations, we used linkage disequilibrium decay and runs of homozygosity. In each wBSP, the r^2 statistic was estimated between every possible pair of SNPs in a 1-Mb sliding window, using the Expectation-Maximization (EM) method implemented in Haploview (*34*). As r^2 estimates are sensitive to sample size, we randomly sampled 10 individuals from each population and repeated resampling and r^2 estimation ten times. We avoided bias due to SNP ascertainment by restricting this analysis to SNPs with a population frequency greater than 5% in every population subsample. The genetic distances between all possible pairs of SNPs were retrieved from the HapMap phase 2 interpolated genetic map (*35*). For each population, all pairwise r^2 values were averaged over replicates and bins of increasing genetic distance.

Runs of homozygosity (ROH) were detected with the sliding-window approach implemented in PLINK (*33*). We minimized the bias due to SNP ascertainment by including only SNPs with a minor allele frequency exceeding 5% in every population. The genome of each individual was explored by a sliding window of 20 consecutive SNPs. If more than 18 SNPs were homozygous in the individual considered, the window was considered to be homozygous. We allowed for five missing genotypes in each homozygous window. ROH regions were defined as regions of at least 500 kb in which all SNPs were included in at least one homozygous window. The number and cumulative length of ROH regions were then determined in each individual.

ADMIXTURE analyses. ADMIXTURE (*11*) estimates for each individual the proportions of the genome originating from *K* ancestral populations, *K* being specified *a priori*. The method assumes linkage equilibrium between SNPs. For this analysis, we pruned the datasets for SNPs in linkage disequilibrium (LD), using PLINK (*33*), removing SNPs presenting LD r^2>0.5 in 50-SNP windows. All ADMIXTURE analyses were run five times with different random seeds, for each *K* value, to ensure

maximum support for the results and check their stability. We retained the results providing the lowest cross-validation (CV) error among iterations (fig. S3 and S6).

Phasing and fineSTRUCTURE analyses. Datasets were phased with SHAPEIT2 (*36*), using 500 states, 50 MCMC main steps, 10 burnin and 10 pruning steps. These parameters have been shown to minimize switch error rates (*36*). Recombination maps were interpolated from the HapMap phase 2 genetic map (*35*). Phased datasets were then analysed with ChromoPainter/fineSTRUCTURE. This approach models LD with the Li & Stephens's model (*13*). It "paints" phased chromosomes of a given recipient individual with the haplotypes of all other individuals in the dataset. The N_e and θ parameters were estimated by ChromoPainter with five EM steps. We performed PCA on the co-ancestry matrix obtained with fineSTRUCTURE, with the *eigen* R function.

Tests for admixture. We tested for the occurrence of admixture in the studied populations with two different methods. The first approach, 3-pop ALDER, tests if a population is admixed between two parental sources, and estimates the time since admixture and admixture proportions using admixture LD decay with distance (*17*). All possible triplets of populations in the data were tested. This method assumes a single pulse of admixture and linkage equilibrium between SNPs. The second method, GLOBETROTTER (*14*), is based on ChromoPainter haplotype "painting" results (see above). It estimates the probability, with distance, that two contiguous alleles on a phased chromosome are inherited from the same ancestral population, and fits one or two exponential curves on the data to identify the best-fit model of admixture (*i.e.*, one-way single pulse, multiple-way single pulse, one-way two pulses, or multiple-way two pulses), and date the admixture event(s). GLOBETROTTER also estimates the respective contribution of donor populations to the genome of the tested recipient population with a linear mixed model (*37*). The N_e and θ parameters of recipient and donor populations were estimated by ChromoPainter with five EM steps. The populations chosen as recipients or donors are summarized in tables S4 and S11. Confidence intervals of admixture dates and of genetic contributions from donor populations were estimated by bootstrapping. Namely, for each bootstrap iteration, we generated *n* "pseudo-individuals", *n* being the sample size of the recipient population, by randomly sampling the 22 chromosomes of each "pseudo-individual" from independent recipient individuals. GLOBETROTTER was then run on the new "chunklengths.out" files obtained for the bootstrapped recipient population and original donor populations. We repeated this procedure 100 times and computed the 95% quantiles of the sampling distribution of estimated parameters, across the 100 bootstrap re-samples.

Identical-by-descent segment sharing. We used GERMLINE (*38*) to infer Identical-by-descent (IBD) segments shared between all individuals of different BSPs. The "-w_extend -bits 80 -err_het 0 -err_hom 1 -min_m 3" arguments were used, and only IBD segments of length 3 cM or longer were considered. Because GERMLINE is prone to false positive IBD assignment, we filtered out genomic regions that showed a number of overlapping IBD segments exceeding 10% of all possible pairs of individuals analysed (*i.e.*, 8865 pairs). We then counted the number of IBD segments shared between eBSPs (or seBSPs) and wBSPs from northern (Cameroon) or southern (Angola) western central Africa. Counts were then divided by the total number of possible pairs of individuals from the two compared populations.

Linguistic distance matrix. Linguistic data and cognate encoding for Wide- and Narrow-Bantu languages were retrieved from a previous study (*15*). Niger-Congo languages (including Yacouba, Bariba, Ahizi and Yoruuba) were added to the published database, which yielded a total of 652 languages belonging to the following subgroups: Mande, Gur, Kru, Dogon, Kwa, Ijoid, Kainji, Plateau, Wide- and

Narrow-Bantu languages. We included 100 words from the basic vocabulary, corresponding to the following words in English: *all, animal, arm, ashes, bark, bed, belly, big, bird, bite, blood, bone, breast (woman), burn, child, cloud, come, count, die, dog, drink, ear, eat, egg, elephant, eye, face, fall, fat, oil, feather, fingernail, fire, fire-wood, fish, five, fly, four, give, goat, ground, soil, hair (on head), head, hear, heart, horn, house, hunger, intestine, kill, knee, knife, know, leaf, leg, liver, louse, man, moon, mountain, hill, mouth, name, navel, neck, night, nose, one, person, rain, road/path, root (big), salt, sand, see, send, shame, sing, skin, sky, sleep, smoke, snake, spear, steal, stone, sun, tail, ten, three, tongue, tooth, tree, two, urine, village, vomit, walk, war, water, wind, woman*. For each word, we looked for cognate sets (similar words with a similar meaning, suggesting potential relatedness between languages). Each cognate set was coded as a discrete binary character. In total, 10,716 cognate sets were identified.

Distances between languages were obtained by estimating a distance matrix between the 652 languages, based on the Hamming distance, which counts the number of differences between two sequences/characters. We used a neighbor-joining algorithm to construct a phylogenetic network based on the linguistic distance matrix, and sequential agglomeration: taxa were combined into progressively larger and larger overlapping clusters. With the SplitsTree program (*39*), we then selected the languages spoken by the 30 Niger-Congo-speaking populations included in this study, from the 652 languages, and extracted a new linguistic distance matrix based on the sub-network.

Impact of linguistic and geographic distances on the genetic differentiation of wBSPs. To test different hypotheses on the dispersal of BSPs across sub-Saharan Africa, we first evaluated if Non-Bantu, Wide-Bantu and Narrow-Bantu-speaking populations of western central Africa form a structured population, and whether this genetic structure results from geography or linguistic barriers. Haplotype-based PCA (*i.e.*, eigendecomposition of the fineSTRUCTURE covariance matrix) separated on PC1 populations speaking Non-Bantu, Wide-Bantu, Bantu zone A from those speaking Bantu zone B and H (accounting for 10.61% of the variance; fig. S7). PC2 (5.16% of the variance) reflected genetic structure within Bantu zone A speakers, separating the Fang (A75) from the Ngumba, Nzime and Badwee (A81 and A84). PC3 (3.16% of the variance) separated wBSPs of Gabon and Angola into North Western B10-B30-B40 Bantu speakers (Galoa, Tsogo, Eviya ad Eshira) and West Western BSPs (i.e., B50-B70 and H Bantu speakers). The genetic distinctiveness of the Eviya accounted for only 0.96% of the total variance in the haplotype-based PCA (fig. S7), while it was highlighted at K=3 by ADMIXTURE (fig. S5). Increased genetic drift may explain these different results, as shown by the high levels of LD and the large cumulative length of ROH (fig. S22) observed in the Eviya, relative to all other wBSPs.

We estimated the contribution of linguistic and geographic distances to the genetic differentiation observed between Non-Bantu, Wide-Bantu and Narrow-Bantu-speaking populations of western central Africa. We computed the matrix of genetic distances between the 30 populations included in the *GabonDiv_PGSJ* dataset with AMOVA-based F_{ST} values, computed with a home-made perl script (available upon request). The matrix of geographic distances between these populations was obtained with the *distm* function of the *geosphere* R package. The three distance matrices were normalized to vary between 0 and 0.02. We estimated the respective effects of linguistic and geographic distances on the matrix of genetic distances with partial Mantel tests (*40*), with 10^6 permutations. A strong correlation of genetic distances with geography was observed, conditioning on linguistics (r=0.56; P=3.7x10^{-5}). A strongly significant, although weaker, correlation was also detected with linguistics, conditioning on geography (r=0.39; P=3.4x10^{-4}).

Genome-wide scans for recent positive selection. Candidate genomic regions of recent positive selection in wBSPs, eBSPs and seBSPs were detected with an outlier approach that considers six neutrality statistics: intrapopulation |ΔiHH| (*41*) and |iHS| (*42*), interpopulation PBS (*43*) scores with two different outgroups, and interpopulation $|\Delta iHH_D|$ and XP-EHH (*44*). Interpopulation statistics require a reference population and PBS statistics an outgroup population. The reference population for scans of positive selection in wBSPs and in eBSPs/seBSPs were the Niger-Congo non-Bantu-speaking populations of western Africa and wBSPs, respectively. The outgroup populations for the two PBS statistics were Europeans and the Ju/'hoansi San for wBSPs and eBSPs, and Europeans and the western RHG Baka for seBSPs (because seBSPs are admixed with the Ju/'hoansi San). For this analysis, wBSPs were represented by all Narrow-Bantu-speaking populations of western central Africa; eBSPs by the Bakiga and the Luhya; seBSPs by the Sotho and the S.E. Bantus from HGDP; and swBSPs were not analysed because of their low sample size (n=8).

The derived allele of each SNP was defined based on the 6-EPO alignment in 98.5% of the cases. Statistics based on extended haplotype homozygosity were computed in 2-cM windows, with home-made scripts (available upon request). Only SNPs with a derived allele frequency (DAF) between 10% and 90% were analysed further. All statistics except PBS were normalized in 40 separate bins of derived allele frequency. We combined the genome-wide rank of the six statistics for each SNP with a Fisher score (*19*), equal to the sum, over the six statistics, of $-\log_{10}$(rank of the statistic/number of SNPs), referred to as F_{SC}. An outlier SNP was defined as a SNP with a F_{SC} among the highest 1% of the genome. A putatively selected genomic region was defined as a 100-SNP window presenting a proportion of outlier SNPs among the highest 0.5% of the genome. We combined the six statistics because it has been shown that combining different neutrality statistics into a single score increases the power to identify signals of recent positive selection (*19, 41*). The rationale behind this approach is that neutrality statistics are expected to be more correlated for positively-selected variants, relative to neutral variants (*41*). Consequently, candidate variants for positive selection should show outlier values for several statistics included in the combined score.

In wBSPs, eight genomic regions were considered to be putative targets of positive selection (Fig. 3A, fig. S11 and table S7). Genome-wide, the strongest proportion of selection signals was observed at the *HLA* locus, reaching 50.5% in the *HLA-D* gene region (Fig. 3, A and D). The best-candidate SNP at the locus was located in the *HLA-DOA* gene (rs3129302, F_{SC} = 12.7, P_{emp} = 2.9x10^{-5}) and displayed high levels of extended haplotype homozygosity (XP-EHH=-5.30, iHS=-3.43; fig. S12 and table S7), relative to non-Bantu-speaking western African populations. The second strongest proportion encompassed the *CD36* gene (27.7%), associated with *Plasmodium falciparum* malaria (*18*). The putatively selected variant was located in *CD36*, had a frequency of up to 25% in wBSPs, and presented both high levels of population differentiation and extended haplotype homozygosity (rs3211881; F_{SC}=15.4, P_{emp}=5.8x10^{-6}; fig. S11E-F and S12).

In eBSPs, five genomic regions were considered to be putative targets of positive selection (Fig. 3B, fig. S13 and table S8). The genome-wide highest proportion of selection signals was again observed at the *HLA* locus, reaching 62.4% in the *HLA-D* gene region (Fig. 3, B and E, and fig. S13, A and B). The best candidate SNP at this locus was located close to *HLA-DOA* (rs6907291, F_{SC}=11.4, P_{emp}=6.9x10^{-5}) and had high levels of extended haplotype homozygosity, relative to wBSPs (XP-EHH=-3.78, iHS=-3.22; fig. S14 and table S8). The comparison of positive selection signals at the *HLA* locus in wBSPs and eBSPs revealed a positive correlation among individual SNP selection scores (Pearson's r=0.384, P<0.00001).

This suggested that wBSPs and eBSPs have in common a recent positive selection event at *HLA* genes, which has left a relatively stronger selection signature in eBSPs than in wBSPs. The second strongest proportion was found close to the region regulating the *LCT* gene (28.7%; Fig. 3, B and E, and table S8). The best candidate variant at this locus had the genome-wide strongest individual SNP score in eBSPs (rs4954204; F_{SC}=25.65; P_{emp}=5.7x10^{-6}), owing to both strong genetic differentiation and high levels of extended haplotype homozygosity, relative to wBSPs (figs. S13, C and D, S14, and table S8). This variant is in linkage disequilibrium (r^2=0.69 in Kenyan Maasai) with the lactase persistence mutation C-14010, which appears to have originated in eastern Africa (*20, 45*).

In seBSPs, six genomic regions were considered to be putative targets of positive selection, although the proportions of selection signals were weaker than in wBSPs and eBSPs (Fig. 3C, fig. S15 and table S9). This possibly reflects the different demographic and adaptive history of seBSPs, limited sample size and population coverage, or genotyping issues identified in the study that generated the data (*12*). The genome-wide highest proportion of outlier SNPs was observed on chromosome 3, reaching 23.8% close to the *GPR156* and *FSTL1* genes (Fig. 3F). The best candidate SNP at this locus was located in *GPR156* (rs979683, F_{SC}=14.9, P_{emp}=2.3x10^{-5}; fig. S15A), and presented a high degree of extended haplotype homozygosity (XP-EHH=-2.69, iHS=-4.20), relative to wBSPs. The *GPR156* gene encodes a protein of unknown function, and the FSTL1 protein has an important role in wound repair (*46*).

Evidence for adaptive introgression in Bantu-speaking populations. We used genome-wide scans for positive selection and deviations in local ancestry to identify candidate variants for adaptive introgression in BSPs. The genomes of BSPs were decomposed into segments of different ancestries with RFMix v.1.5.4 (*47*), including two EM steps. Local ancestry was excluded within the 2 Mb from the telomeres of each chromosome (*25*). We identified most likely introgressed alleles as those showing the genome-wide highest association, using the linear regression model implemented in PLINK, with the inferred local ancestry in genomic regions displaying deviations from the genome-wide average. Two loci, the *HLA* locus in wBSPs and the *LCT* region in eBSPs, displayed signals of both positive selection and increased local ancestry from western RHG and eastern Africans, respectively. At the *HLA* locus in wBSPs, a remarkable excess of ancestry from RHG was observed: 38% RHG ancestry, 6.74 standard deviations from the genome-wide average of 16% (Fig. 3A). This excess was confidently reproduced in a dataset obtained from a single genotyping platform, excluding a methodological bias due to the alignment of alleles from different SNP arrays (fig. S11A). As long-range LD can also confound scans for deviations in local ancestry (*48*), and because *HLA* haplotypic structure may be too complex to be correctly modelled by RFMix, we performed phasing and local ancestry inference on a restricted dataset in which the classical *HLA* region (*i.e.*, all *HLA* genes) was excluded. Consistently, a strong excess of RHG ancestry was observed at the locus (fig. S11B), supporting the view that the signal is true and extends beyond the complex *HLA* locus. The variant most strongly associated with introgressed western RHG haplotypes was located near the *HLA-DMB* gene (rs149392, P=7.7x10^{-40}); the rs149392 derived allele had a frequency of <5% in non-Bantu-speaking western Africans, 52% in the Bakoya western RHG, and 18% in wBSPs.

To provide further support to these analyses, we tested with simulations if the current allele frequency of introgressed variants could be explained under neutrality. It is expected that an introgressed allele under positive selection in an admixed population will reach a frequency that is too high given its frequency in the parental populations and their respective contribution to the admixed population (*49, 50*). However, strong genetic drift in the admixed population, or continuous gene flow from parental populations since admixture, may also increase the frequency of a neutral allele in the admixed population. We thus used

the Wright-Fisher model to simulate the fate of a neutral allele in an admixed population, under strong genetic drift and/or a model of continuous gene flow. Specifically, an admixed population, *adm*, was generated by an admixture event between two parental populations, *pop1* and *pop2*, that occurred *t* generations in the past, *t* being the admixture time estimated by GLOBETROTTER. The initial admixture proportions from *pop1* and *pop2* were those estimated from GLOBETROTTER, except when continuous gene flow from *pop1* to *adm* was simulated. In this case, we tuned initial admixture proportions so that the *present-day* admixture proportions in *adm* were equal to GLOBETROTTER estimates. The frequency of the simulated allele in the parental populations at time *t* was considered to be equal to that observed in present-day best-matching parental populations. We let vary two parameters in the model: the *adm* population size N_{adm}, and the migration rate m_{adm} from *pop1* to *adm*, assumed to be unidirectional, continuous and constant from generation *t* to the present. The m_{adm} parameter was modeled as a proportion of *adm* individuals that migrated from *pop1* at each generation. To test if genetic drift alone could explain our results, we simulated 50 different models where N_{adm} varied between 100 and 10,000. To test if continuous gene flow could explain our results, we simulated 50 different models where m_{adm} varied between 0.01% and 1%. Each model was replicated 10,000 times, and the 95% confidence interval of the current frequency of the neutral allele in *adm* was computed.

Simulations showed that the frequency of the *HLA-DMB* rs149392 introgressed allele was unexpectedly high in wBSPs ($P<0.0001$), except when we assumed an unrealistic scenario where the population size of wBSPs was 1,000 since admixture with RHGs ($P>0.05$; fig. S16 and table S10), far below recent estimations (*31, 51, 52*). However, despite its high allele frequency in wBSPs, the variant did not have a strong selection score ($F_{SC}=7.65$), owing to its low iHS value (iHS=0.457). Furthermore, the variant for which the evidence for positive selection at *HLA* was strongest in wBSPs did not segregate with the introgressed haplotypes ($P=0.02$). These results support the view that the strong signal of positive selection close to *HLA-D* genes observed in wBSPs does not result from introgression from western RHG populations. At the *LCT* region in eBSPs, the genetic variant most strongly associated with introgressed eastern African haplotypes was rs4954204 ($P=9.3\text{x}10^{-21}$), which was also the variant with the highest selection score genome-wide (table S8). Its derived allele had a frequency of 30% in the eastern Bantu-speaking Bakiga of Uganda, yet it was virtually absent from wBSPs (fig. S13D). Simulations showed that the frequency of this introgressed allele was unexpectedly high in the Bakiga ($P<0.0001$), in all scenarios explored (fig. S17 and table S10). By contrast to what was found for *HLA* in wBSPs, these results demonstrate that eBSPs have acquired the lactase persistence allele C-14010 – which is in linkage disequilibrium with rs4954204 – via introgression from eastern African populations, with this variant subsequently subjected to positive selection in Bantu speakers.

Tests for direct contribution of western RHGs to the slave trade. The ~4.8% of RHG ancestry detected in the African genome of African Americans (AAs) (Fig. 4 and fig. S19) can result from two alternative scenarios: RHGs were embarked during the slave trade, or the RHG component observed in the genome of AAs derives from the substantial contribution of wBSPs, who themselves have ~16% western RHG ancestry (Fig. 4 and fig. S2). To test these hypotheses, we computed *D*-statistics (*53*), in the forms *D*(AAs, wBSP; western RHG, San) and *D*(western RHG, wBSP; AAs, Yoruba). All *D*-statistics were significantly negative (*Z*-score<-4; table S12), suggesting no greater genetic affinities between western RHGs and AAs, relative to wBSPs and AAs. Furthermore, western RHGs were also included as a source population for AAs in the GLOBETROTTER analysis (Fig. 4), which, by accounting for the RHG

ancestry observed in wBSPs, estimated that western RHGs did not directly contributed to the African genome of African Americans.

Genome-wide scans for natural selection in African Americans. To test for the occurrence of changes in selective pressure in AA genomes, we compared the minor allele frequency of each SNP in a parental population with its frequency in chromosomes inherited from this parental population in the admixed AA population of interest, identified on the basis of local ancestry. The genomes of admixed populations were decomposed into segments of different ancestries with RFMix v.1.5.4 (*47*), with no EM step, to reduce computation time. Local ancestry was excluded within the 2 Mb from the telomeres of each chromosome (*25*). We tested for significant differences in allele frequencies between a parental and an admixed population using the model-based method developed by Bhatia and colleagues (*49*). For analyses in which AAs were considered to derive from multiple African parental populations, we assumed that the allele frequency in the parental population was equal to its frequency in each parental source, weighted by the local ancestry proportion from each parental source in the admixed population.

African parental populations and admixture mapping. To test the impact of the choice of African parental populations on admixture mapping (*54*), we compared the African local ancestry profiles of AA genomes using, as the parental African populations: the Yoruba only (considered as a good proxy of the African ancestry of AAs in most admixture mapping studies), or a more diverse, realistic set of African populations representing the historical regions of Senegambia, the Windward Coast, the Bight of Benin and western central Africa (table S11). The genomes of AAs were decomposed into segments of African and European ancestries using RFMix. We observed limited differences in local ancestry between the two sets of parental populations (maximal difference of 0.8%; fig. S21). The genes within the genomic regions displaying the largest differences in local ancestry between the two analyses were counted and annotated as being associated or not with complex traits in African Americans, using the NHGRI-EBI GWAS catalog, downloaded on 02/08/2016. We then resampled the same number of genes a million times from all known protein-coding genes, and genes associated by genome-wide association studies (GWAS) with complex traits in AAs were identified for each resampled set. A resampling *P*-value, which tests for an enrichment in associated genes among regions with the largest differences in local ancestry, was then obtained, by dividing the number of resampled sets with more associated genes than in the observed data by the total number of resampled sets. The genomic regions with the largest differences in African local ancestry between the two sets of African parental populations were not enriched in GWAS hits identified in AAs (resampling $P>0.05$). These results suggest a negligible influence of African parental populations on admixture mapping results.

Detailed acknowledgements

eMERGE Vanderbilt: Funding support for the Vanderbilt Genome-Electronic Records (VGER) project was provided through a cooperative agreement (U01HG004603) with the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences (NIGMS). The dataset and samples used for the VGER analyses were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant UL1RR024975 from NCRR/NIH. Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000188.v1.p1.

eMERGE Northwestern University: Samples and data used in this study were provided by the NUgene Project (www.nugene.org). Funding support for the NUgene Project was provided by the Northwestern University's Center for Genetic Medicine, Northwestern University, and Northwestern Memorial Hospital. Assistance with phenotype harmonization was provided by the eMERGE Coordinating Center (Grant number U01HG04603). This study was funded through the NIH, NHGRI eMERGE Network (U01HG004609). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000237.v1.p1.

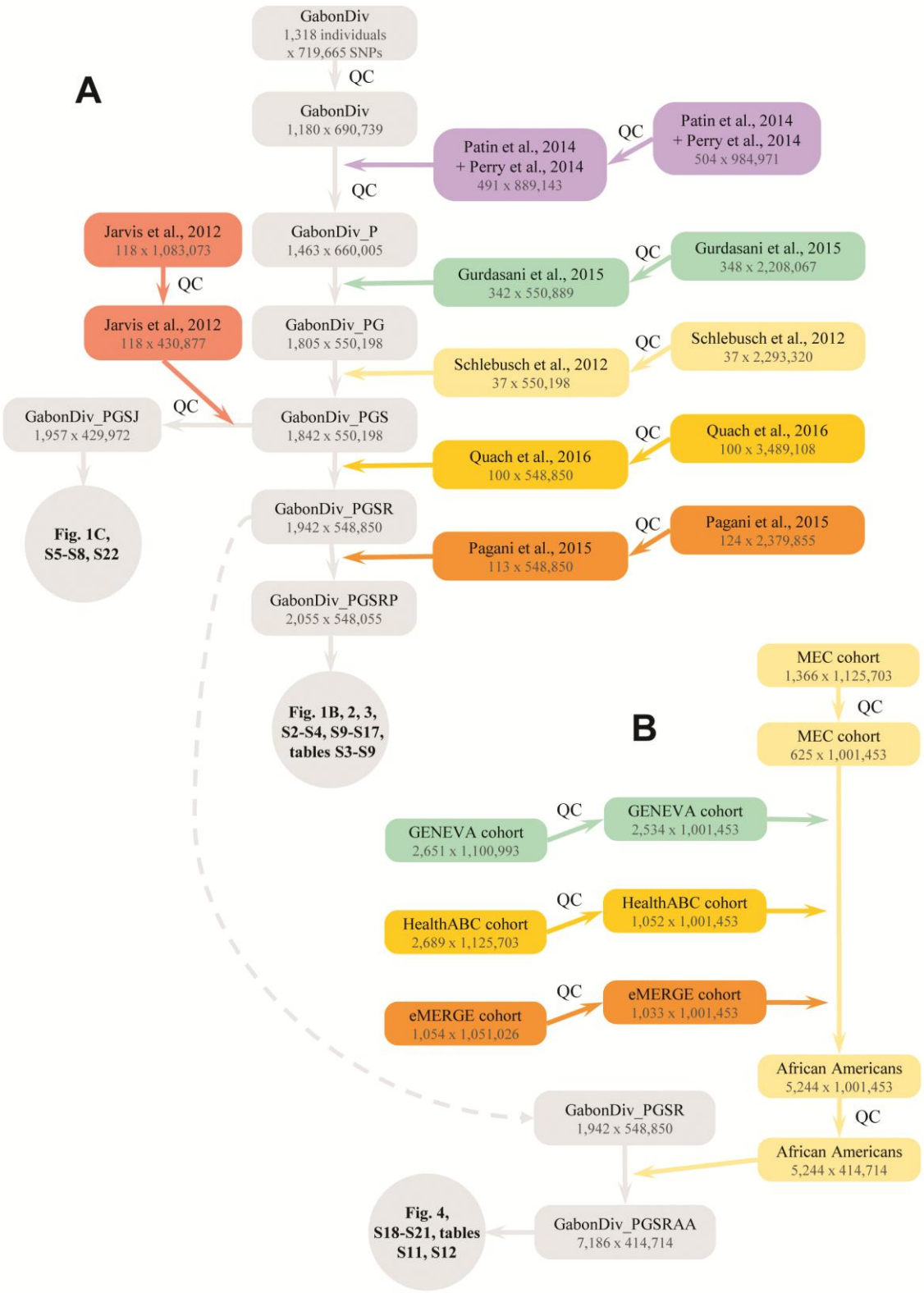

Fig. S1. Datasets used in this study.

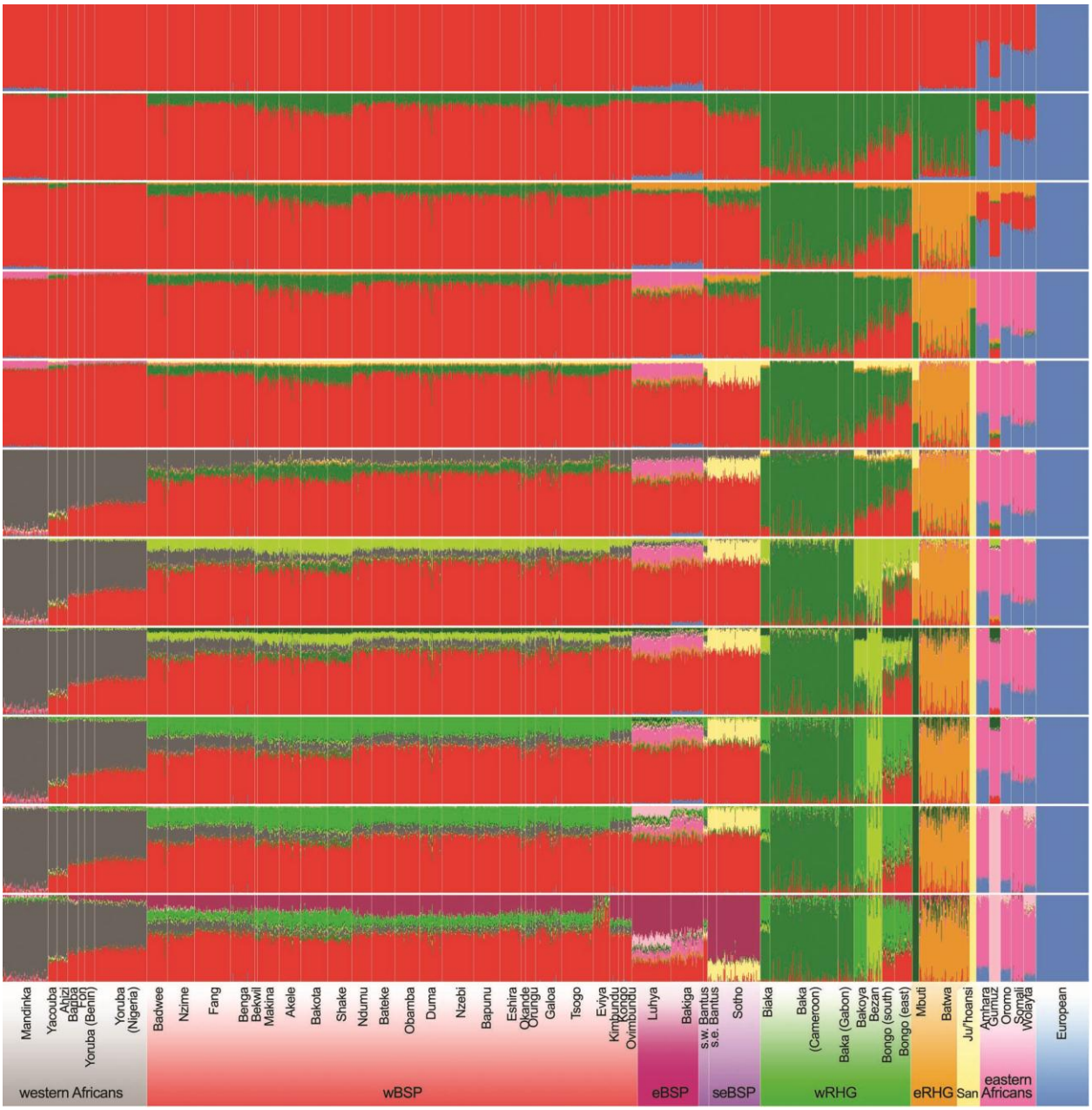

Fig. S2. Genetic structure of African populations. Clustering analysis was performed on 2,055 individuals and 406,798 independent SNPs with ADMIXTURE. Five runs were performed with different random seeds for each *K* value, which ranged from 2 to 12. Results of the runs providing the lowest CV error are shown in fig. S3.

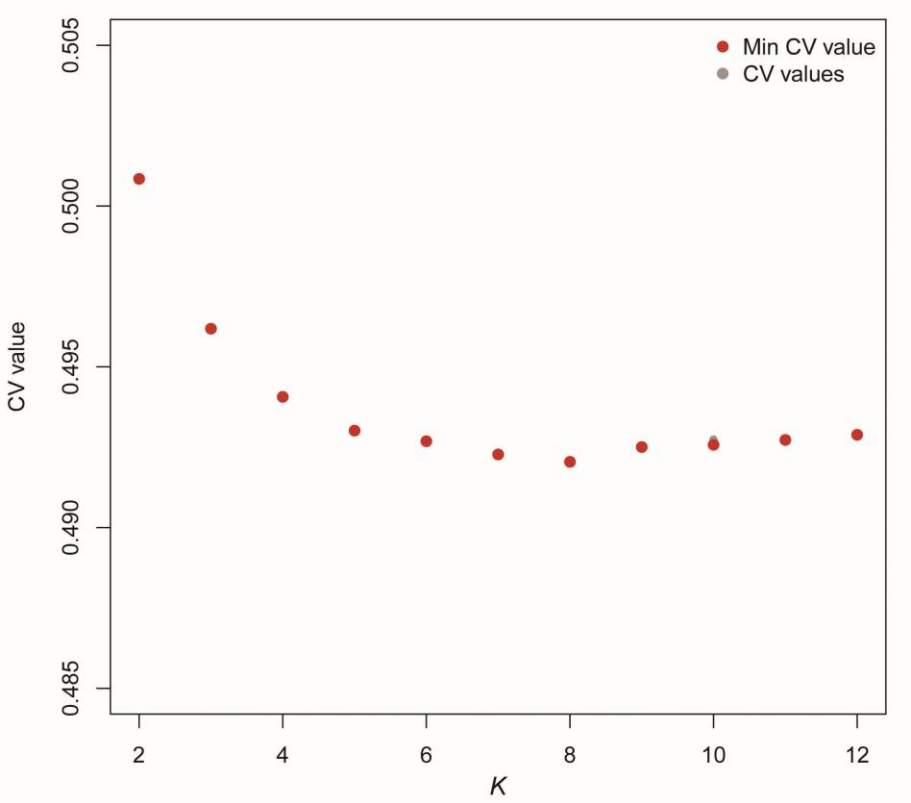

Fig. S3. Error estimation for varying numbers of genetic clusters in sub-Saharan African populations. The error rate was estimated by cross-validation (CV) values, for each *K* value in the ADMIXTURE analysis shown in fig. S2. For each *K* value, the red point indicates the minimum CV value obtained across five independent runs of ADMIXTURE, and gray points indicate the four other CV values.

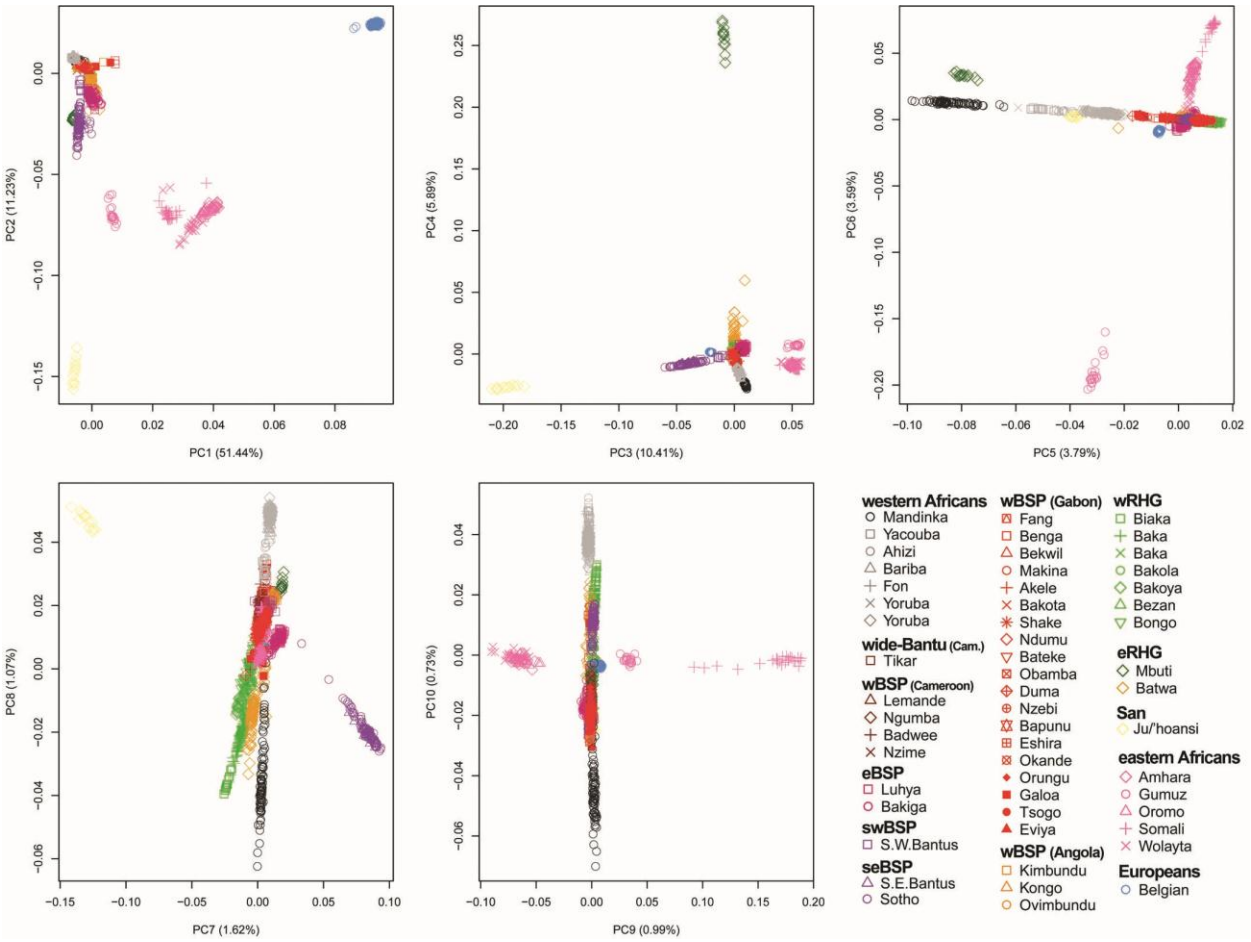

Fig. S4. High-resolution genetic structure of sub-Saharan African populations, obtained by haplotype-based Principal Component Analysis (PCA). Haplotype painting and PCA were performed on 2,055 individuals and 548,055 genome-wide SNPs, with ChromoPainter and fineSTRUCTURE. The ten first principal components of a PCA decomposing the fineSTRUCTURE genetic covariance matrix are shown. The proportion of variance explained by each PC is indicated in brackets.

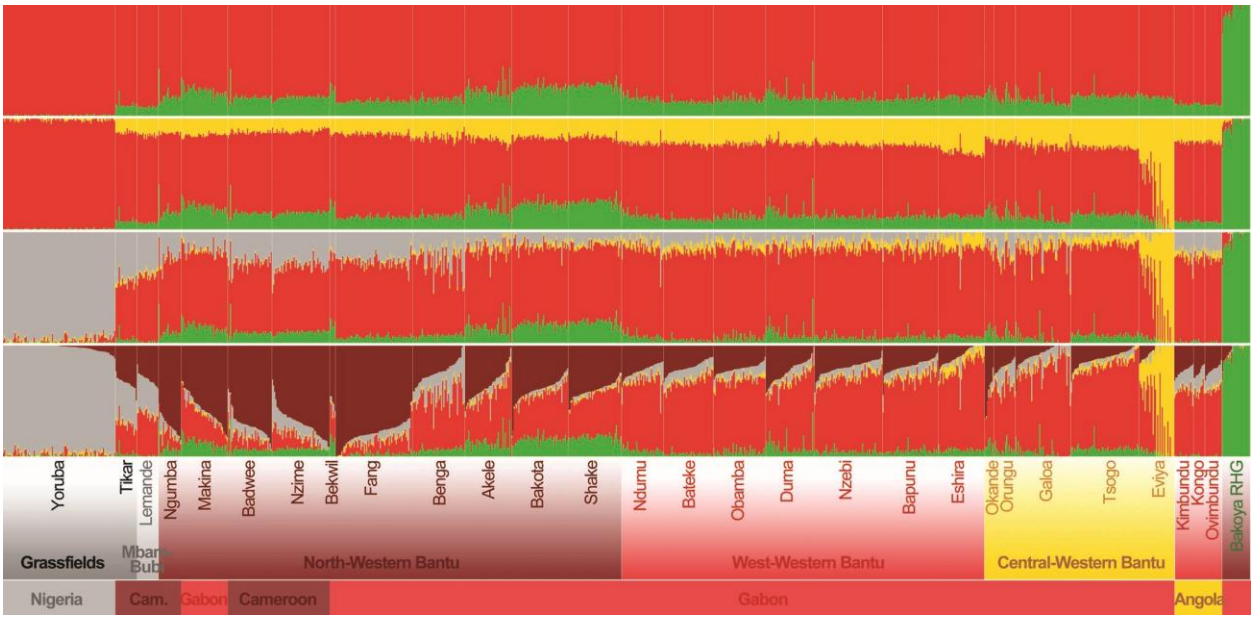

Fig. S5. Genetic structure of western central Bantu-speaking populations. Clustering analysis was performed on 1,139 individuals and 337,056 independent SNPs with ADMIXTURE. Five runs were performed with different random seeds for each *K* value, which ranged from 2 to 5. Results of the runs providing the lowest CV error are shown in fig. S6. Coloured squares at the bottom indicate the language classification (first row) and the country of origin (second row) of the populations.

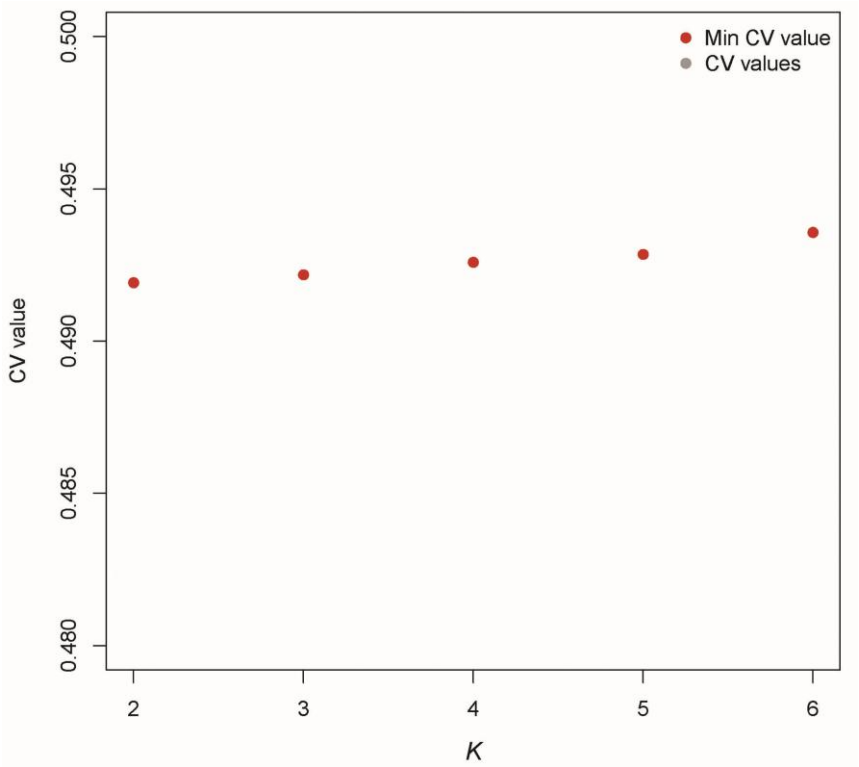

Fig. S6. Error estimation for varying numbers of genetic clusters in western central Bantu-speaking populations. The error rate was estimated by cross-validation (CV) values, for each *K* value in the ADMIXTURE analysis shown in fig. S5. For each *K* value, the red point indicates the minimum CV value obtained across five independent runs of ADMIXTURE, and gray points indicate the four other CV values.

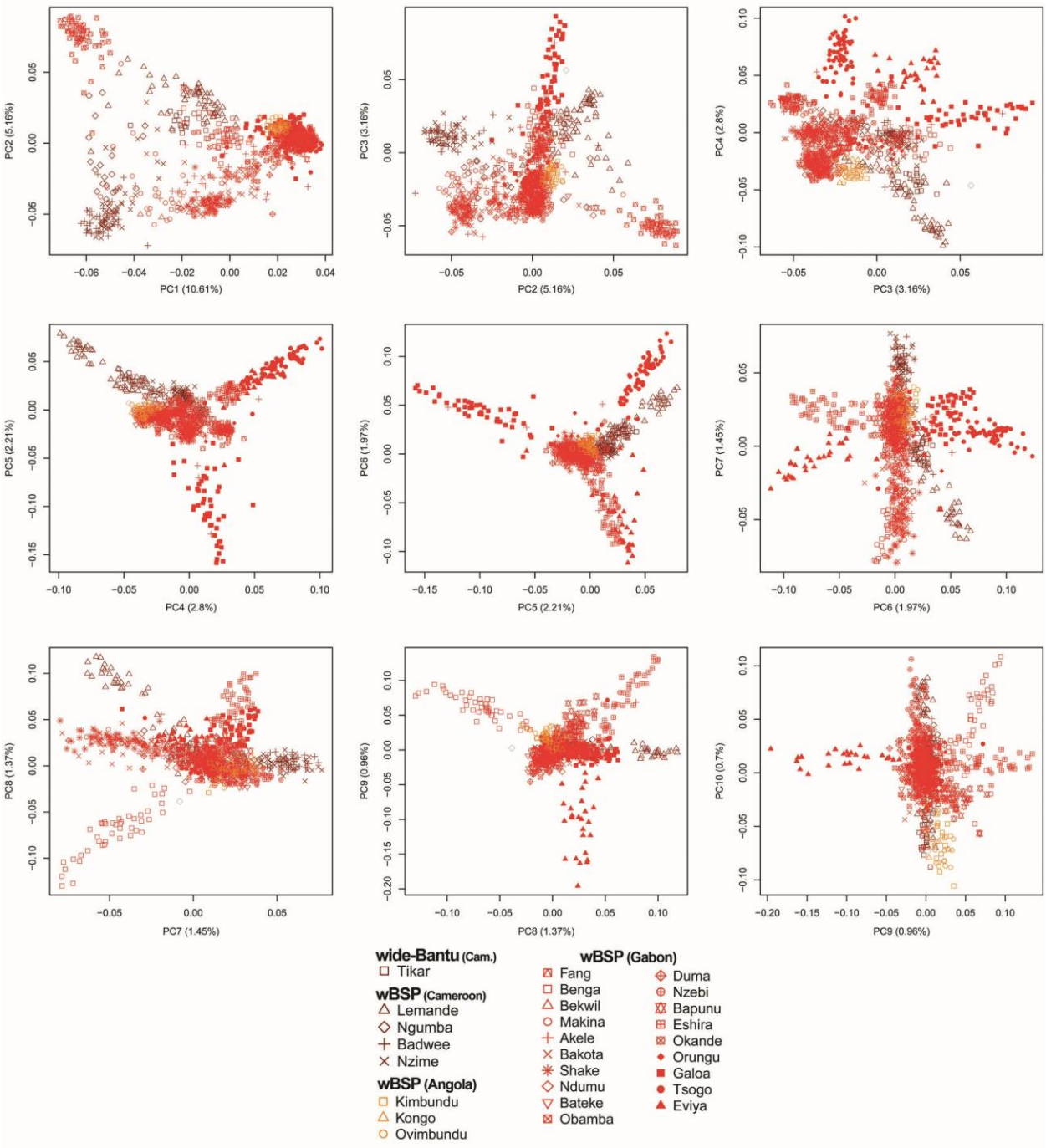

Fig. S7. High-resolution genetic structure of western central Bantu-speaking populations, obtained by haplotype-based Principal Component Analysis (PCA). Haplotype painting and PCA were performed on 1,015 individuals and 429,972 SNPs, using ChromoPainter and fineSTRUCTURE. The 10 first principal components of a PCA decomposing the fineSTRUCTURE genetic covariance matrix are shown. The proportion of variance explained by each PC is indicated in brackets.

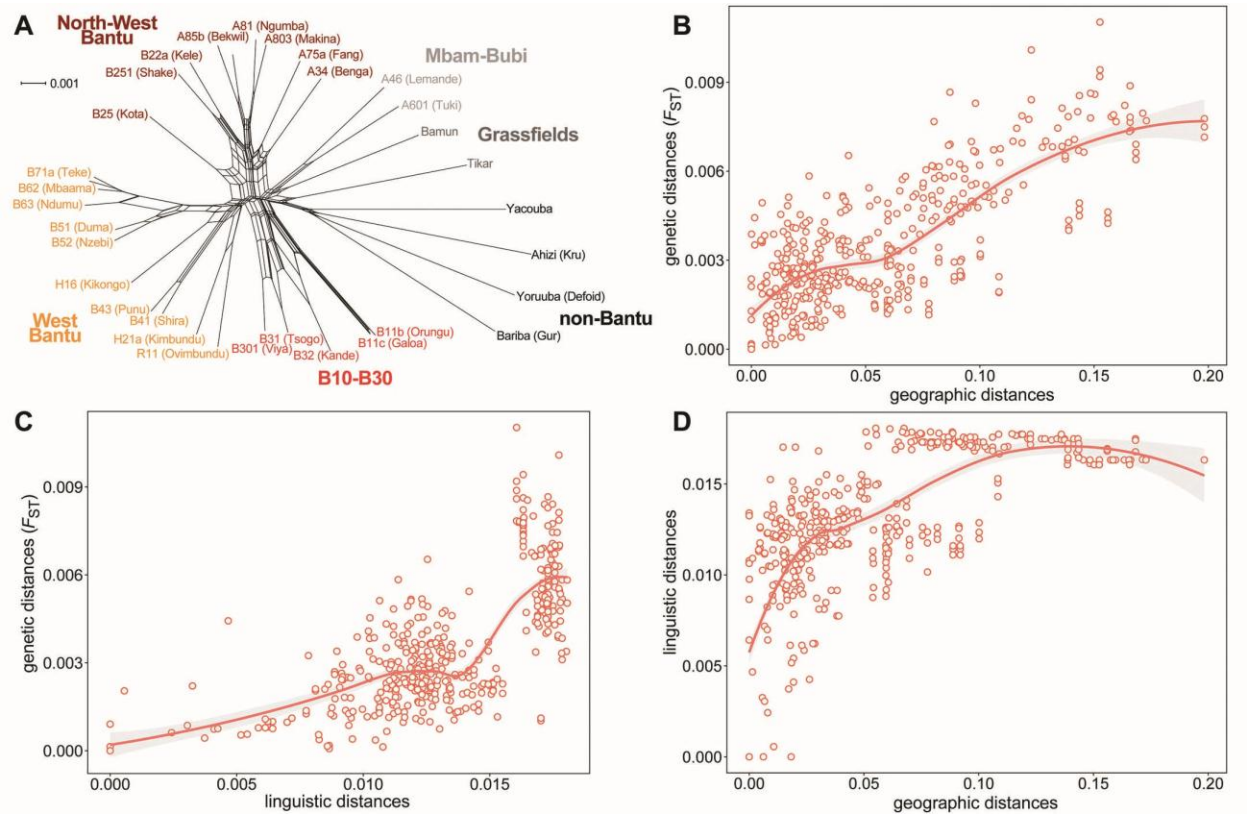

Fig. S8. Contribution of geographic and linguistic distances to the genetic differentiation of western central Bantu-speaking populations. (**A**) Network of linguistic distances among 30 Non-Bantu-, Wide-Bantu- and Narrow-Bantu-speaking populations of western and western central Africa (Materials and Methods). (**B-D**) Relationships between geographic, linguistic and genetic distances. Regression lines were fitted using local multinomial regression.

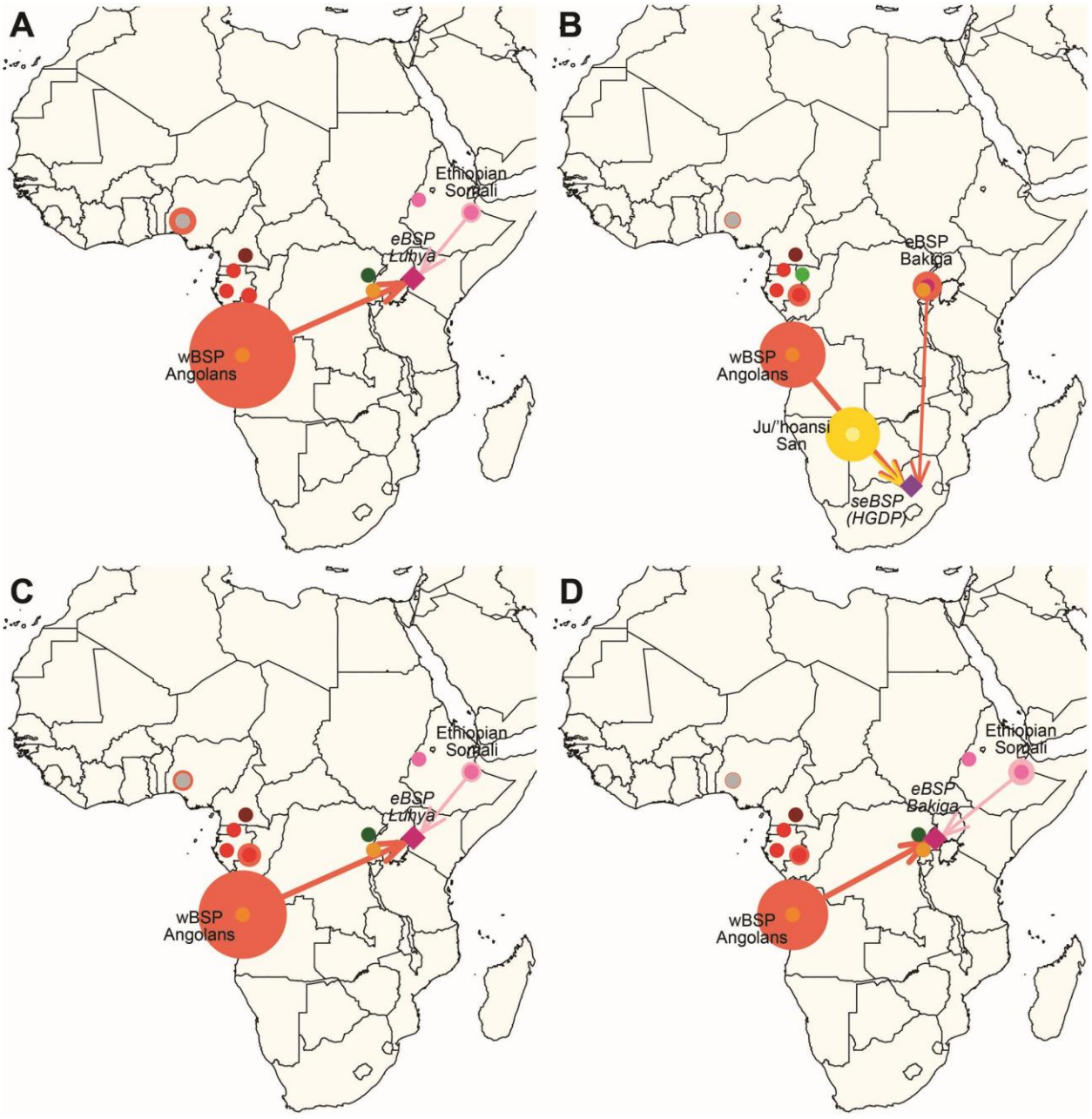

Fig. S9. Haplotype-based inference of the genetic origins of (**A**) Luhya eastern Bantu-speakers (eBSP) and (**B**) southeastern Bantu-speakers from HGDP (seBSP). GLOBETROTTER estimated that the coancestry curves of eBSPs (fig. S10, A and B) were best explained by a two-pulse admixture model (table S4). For the Luhya eBSP, the oldest (**A**) and most recent (**C**) admixture events are shown. (**D**) Haplotype-based inference of the genetic origins of the Bakiga eBSP. Only the most recent admixture event is represented. Estimates for the oldest admixture event are shown in Fig. 2A. (**A-D**) The names of the tested admixed populations are shown in italics. Circle sizes are proportional to the relative genetic contribution of the parental populations to the admixed populations, estimated by GLOBETROTTER.

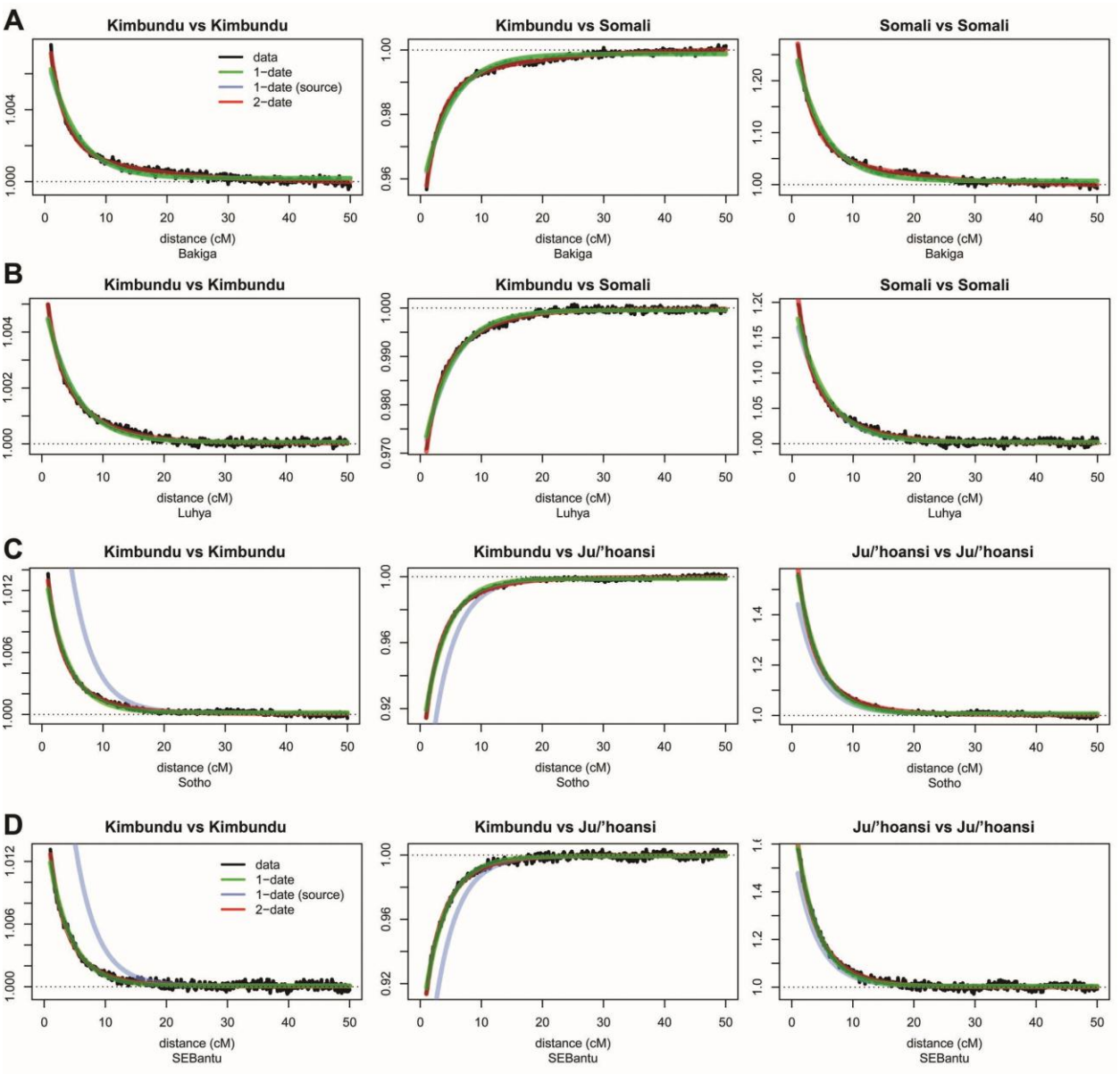

Fig. S10. Examples of observed coancestry curves (in black) and fitted exponential decay curves expected after one (in green) or two (in red) admixture events in the (**A**) Bakiga eBSP, (**B**) Luhya eBSP, (**C**) Sotho seBSP, and (**D**) seBSP from HGDP. The best-matching parental populations are the Kimbundu wBSP of Angola, the Afroasiatic-speaking Ethiopian Somali of eastern Africa and the Ju/'hoansi hunter-gathering San of Namibia. Lines show the relative probability, with genetic distance, of jointly copying two ancestry blocks from the two population sources indicated in the title of each plot. The good fit between observed and expected curves strongly support an admixture model in eBSPs and seBSPs, with estimated parameters shown in table S4.

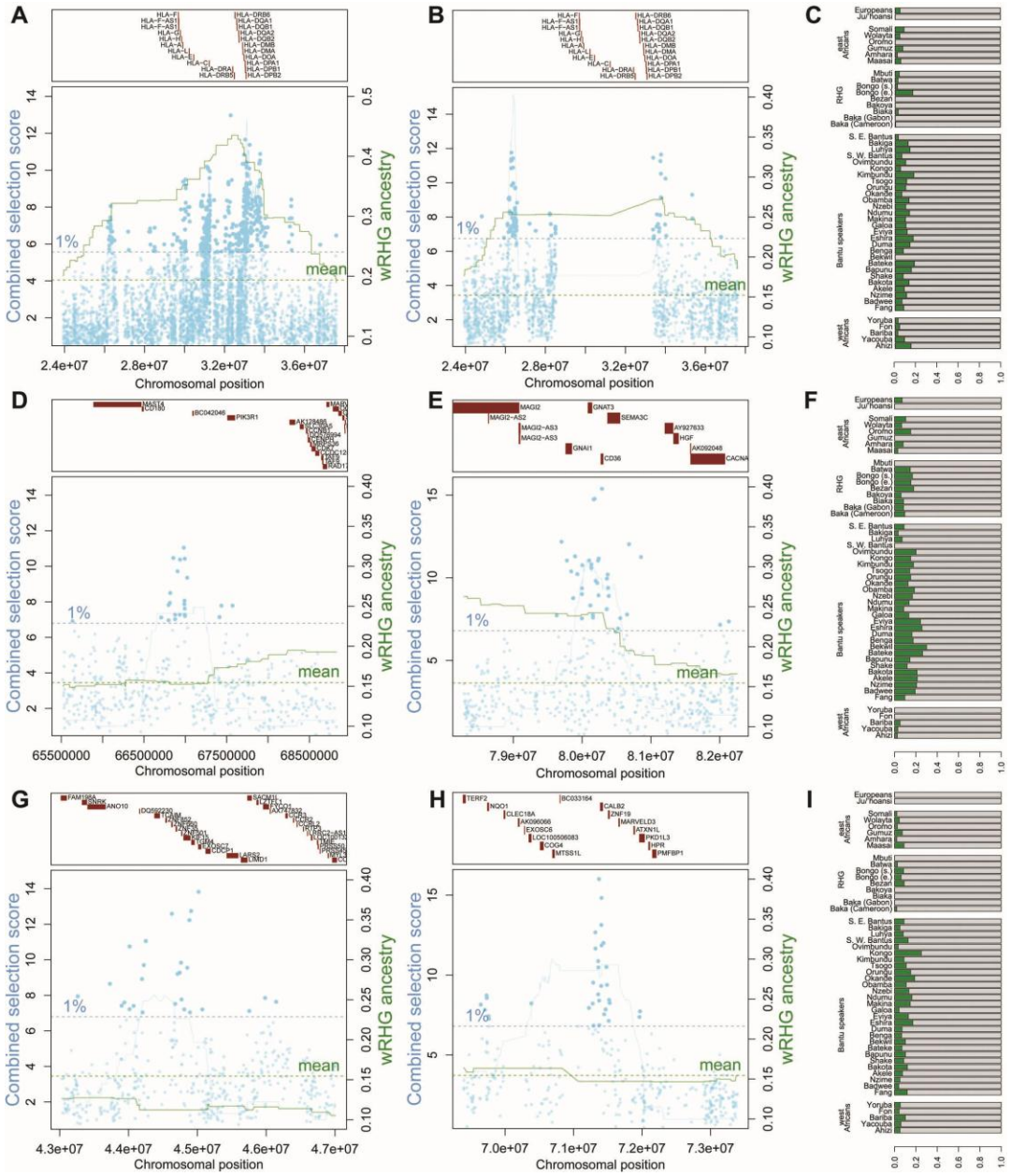

Fig. S11. Strongest signatures of recent positive selection in western central Bantu speakers (wBSP). (**A**) Local signatures in the *HLA* region, in our primary dataset genotyped with the OmniExpress array only. (**B**) Local signatures in the *HLA* region, when removing from all analyses (i.e., phasing, local ancestry inference and positive selection scan) the *HLA* classical region. Analyses in (**A,B**) showed that the signal of increased wRHG local ancestry and recent positive selection observed in Fig. 3 are unlikely to result from from the incorrect modeling of the complex *HLA* haplotype structure by RFMix or misalignments of alleles in the different SNP arrays. (**C**) Allele frequencies of the best candidate SNP in the *HLA* region, rs3128924, in different populations. Local signatures in the (**D**) *CD180*, (**E**) *CD36*, (**G**) *EXOSC7* and (**H**) *CALB2* gene regions. Blue points indicat the positive selection score of individual SNPs, which combines the empirical *P*-values of 6 different neutrality statistics. Blue solid lines indicate the proportion, in 100-SNP windows, of SNPs showing outlier neutrality statistics. (**F**) Allele frequencies of the best candidate SNP in the *CD36* region, rs3211881, in different populations. (**I**) Allele frequencies of the best candidate SNP in the *EXOSC7* gene region, rs11130038, in different populations.

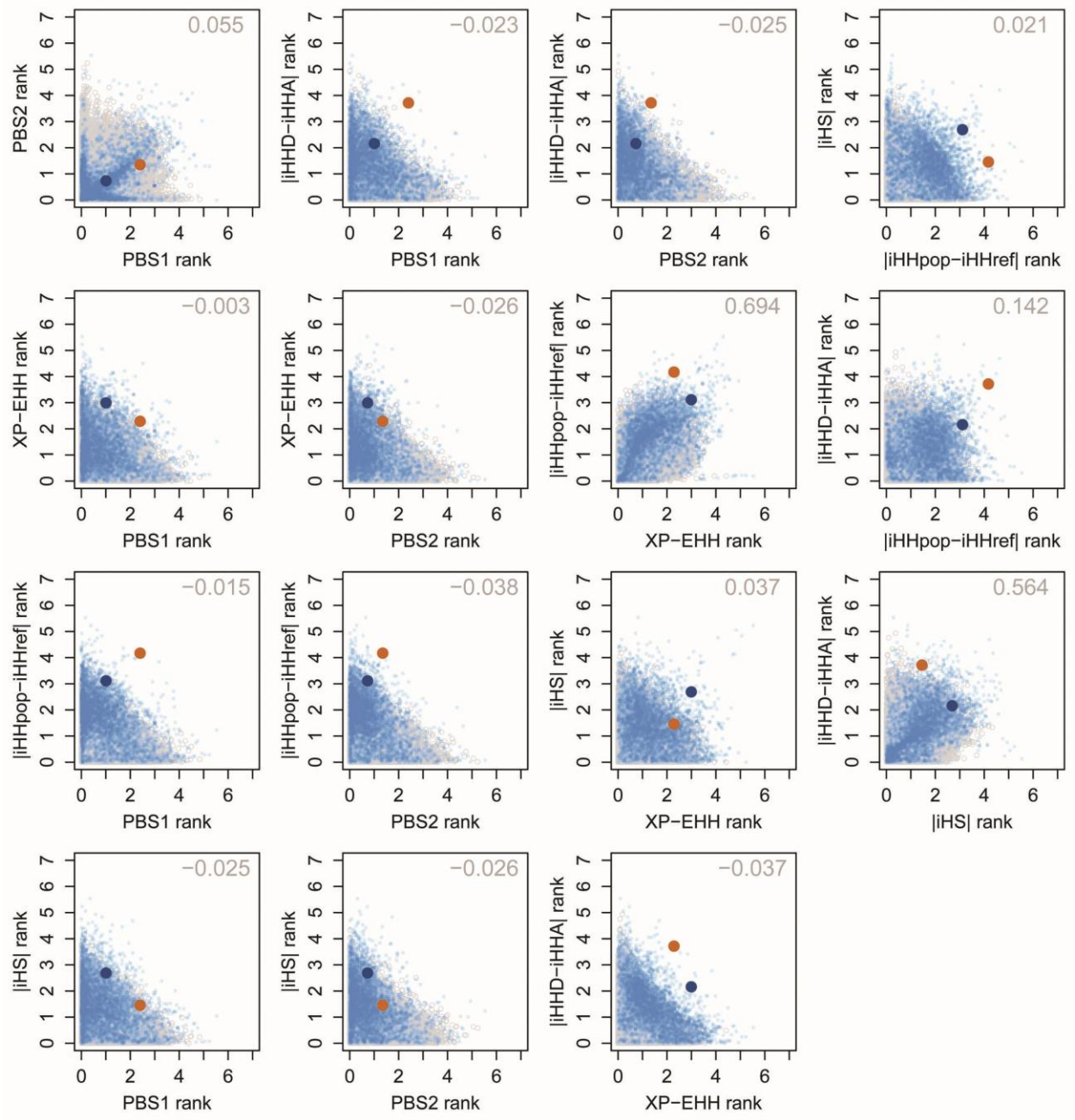

Fig. S12. Relationships between the ranks of combined neutrality statistics in western central Bantu-speaking populations (wBSP). Recent and strong selective sweeps were detected with a Fisher score that combines six neutrality statistics, including two based on interpopulation differentiation (PBS1 and PBS2), interpopulation differences in haplotype homozygosity (XP-EHH and $|iHH_{pop}\text{-}iHH_{ref}|$) and intrapopulation differences in haplotype homozygosity between alleles (|iHS| and $|iHH_D\text{-}iHH_A|$). Small gray and blue points indicate genome-wide SNPs and SNPs with the 1% lowest Fisher scores of the genome, respectively. No strong correlation was observed between the three categories of neutrality statistics (Pearson's coefficient in gray). Large points indicate candidate variants for positive selection (*HLA* locus in dark blue; *CD36* locus in orange; table S7).

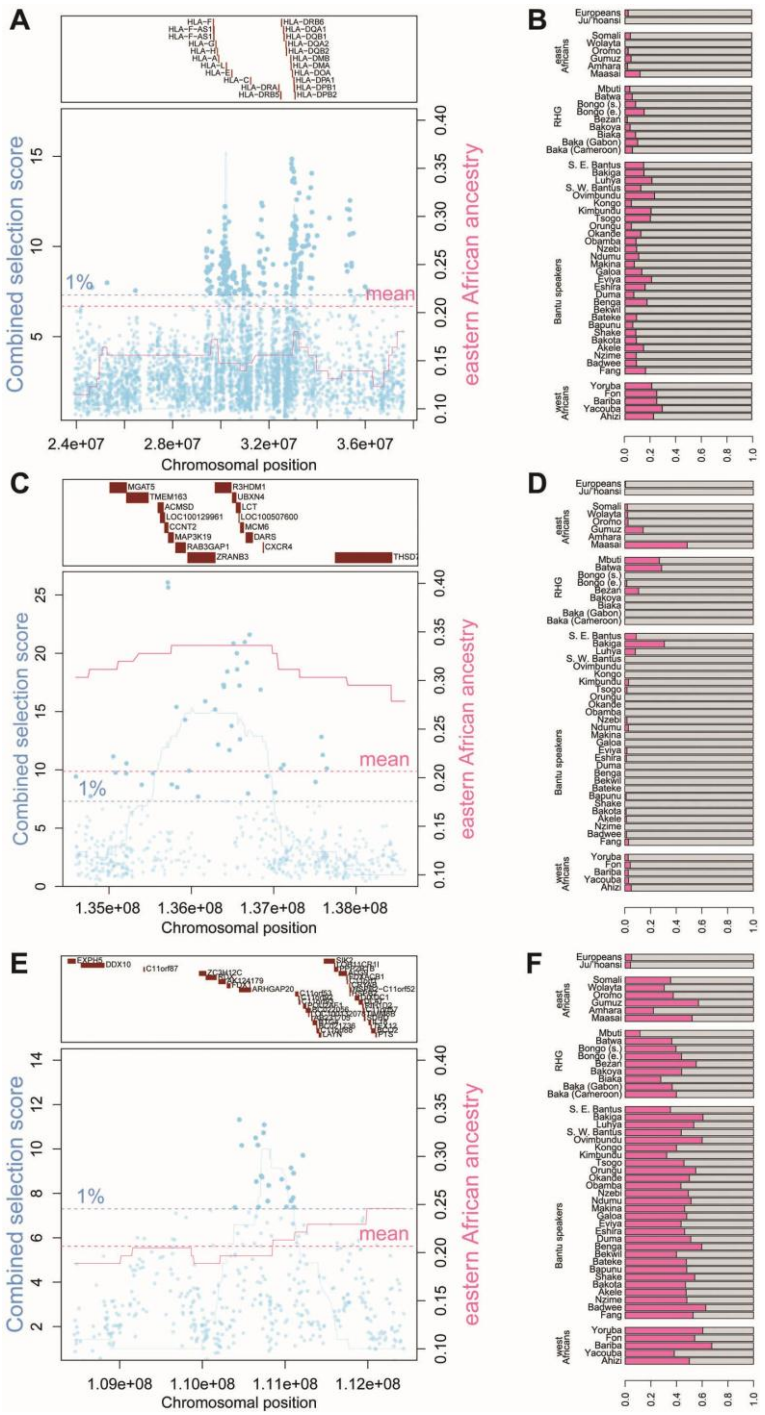

Fig. S13. Strongest signatures of recent positive selection in eastern Bantu speakers (eBSP). Local signatures in the (**A**) *HLA*, (**C**) *LCT* and (**E**) *ARHGAP20* gene regions. Blue points indicate the positive selection score of individual SNPs, which combine the empirical *P*-values of 6 different neutrality statistics. Blue solid lines indicate the proportion, in 100-SNP windows, of SNPs showing outlier neutrality statistics. (**B**) Allele frequencies of the best candidate SNP in the *HLA* region, rs2057726, in different populations. (**D**) Allele frequencies of the best candidate SNP in the *LCT* gene region, rs4954204, in different populations. (**F**) Allele frequencies of the best candidate SNP in the *ARHGAP20* gene region, rs11213629, in different populations.

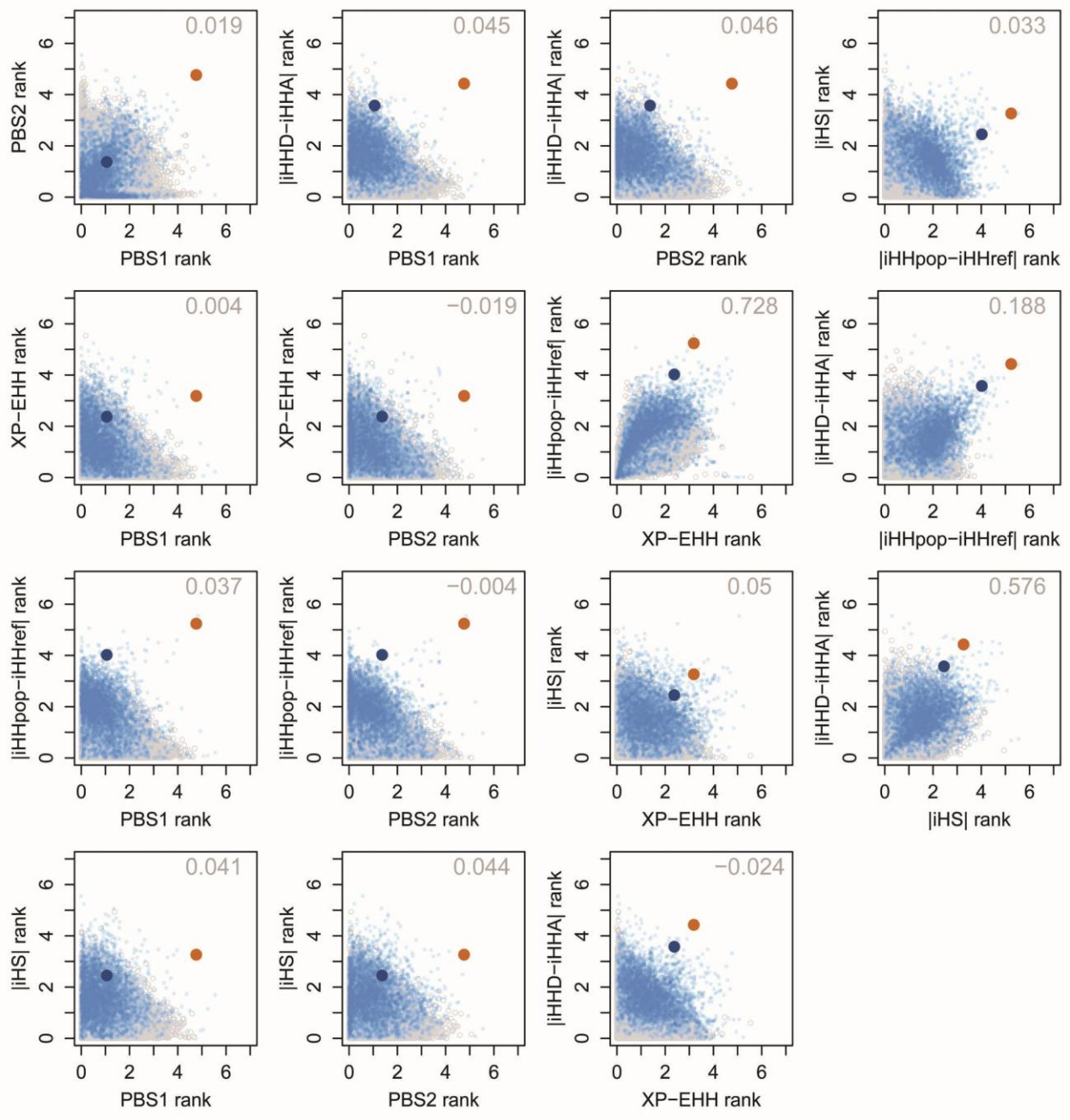

Fig. S14. Relationships between the ranks of combined neutrality statistics in eastern Bantu-speaking populations (eBSP). Recent and strong selective sweeps were detected with a Fisher score that combines six neutrality statistics, including two based on interpopulation differentiation (PBS1 and PBS2), interpopulation differences in haplotype homozygosity (XP-EHH and $|iHH_{pop}\text{-}iHH_{ref}|$) and intrapopulation differences in haplotype homozygosity between alleles (|iHS| and $|iHH_D\text{-}iHH_A|$). Small gray and blue points indicate genome-wide SNPs and SNPs with the 1% lowest Fisher scores of the genome, respectively. No strong correlation was observed between the three categories of neutrality statistics (Pearson's coefficient in gray). Large points indicate candidate variants for positive selection (*HLA* locus in dark blue; *LCT* locus in orange; table S8).

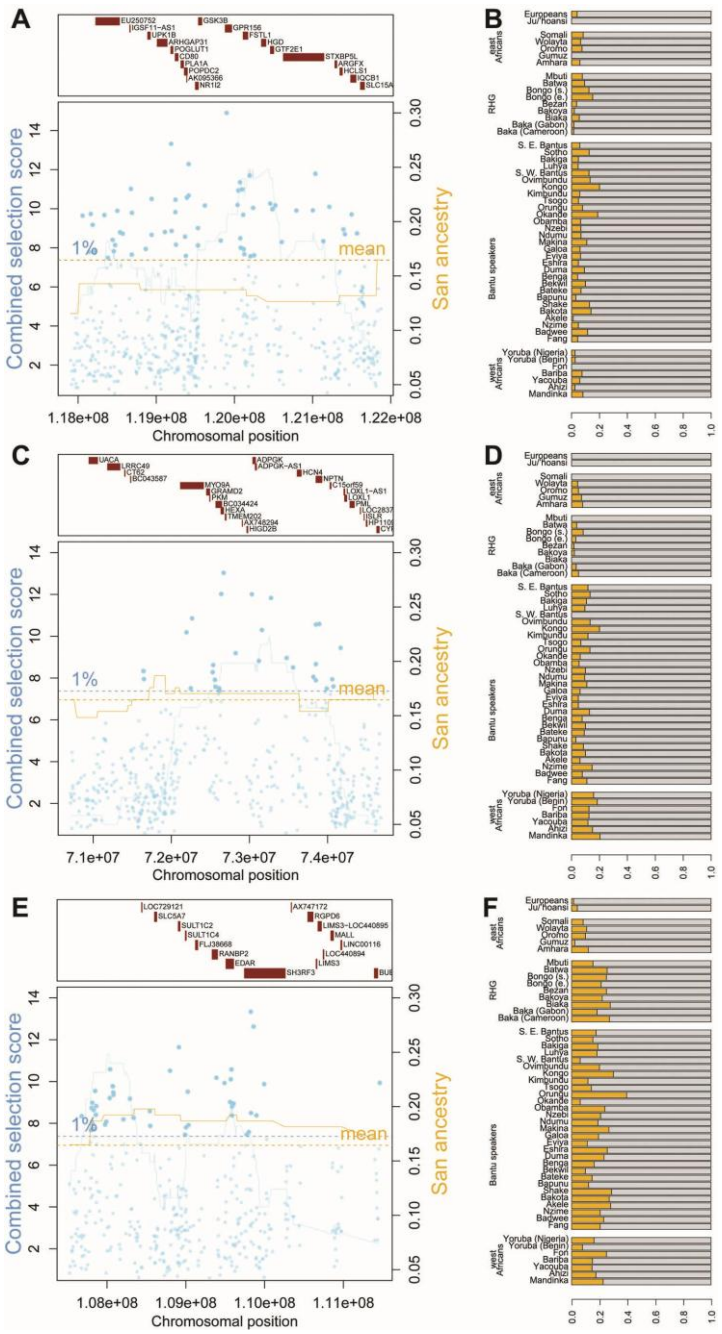

Fig. S15. Strongest signatures of recent positive selection in southeastern Bantu speakers (seBSP). Local signatures in the (**A**) *GPR156*, (**C**) *HEXA* and (**E**) *EDAR* gene regions. Blue points indicate the positive selection score of individual SNPs, which combine empirical *P*-values of 6 different neutrality statistics. Blue solid lines report the proportion, in 100-SNP windows, of SNPs showing outlier neutrality statistics. (**B**) Allele frequencies of the best candidate SNP in the *GPR156* gene region, rs979683, in different populations. (**D**) Allele frequencies of the best candidate SNP in the *HEXA* gene region, rs1800430, in different populations. (**F**) Allele frequencies of the best candidate SNP in the *EDAR* gene region, rs3769775, in different populations.

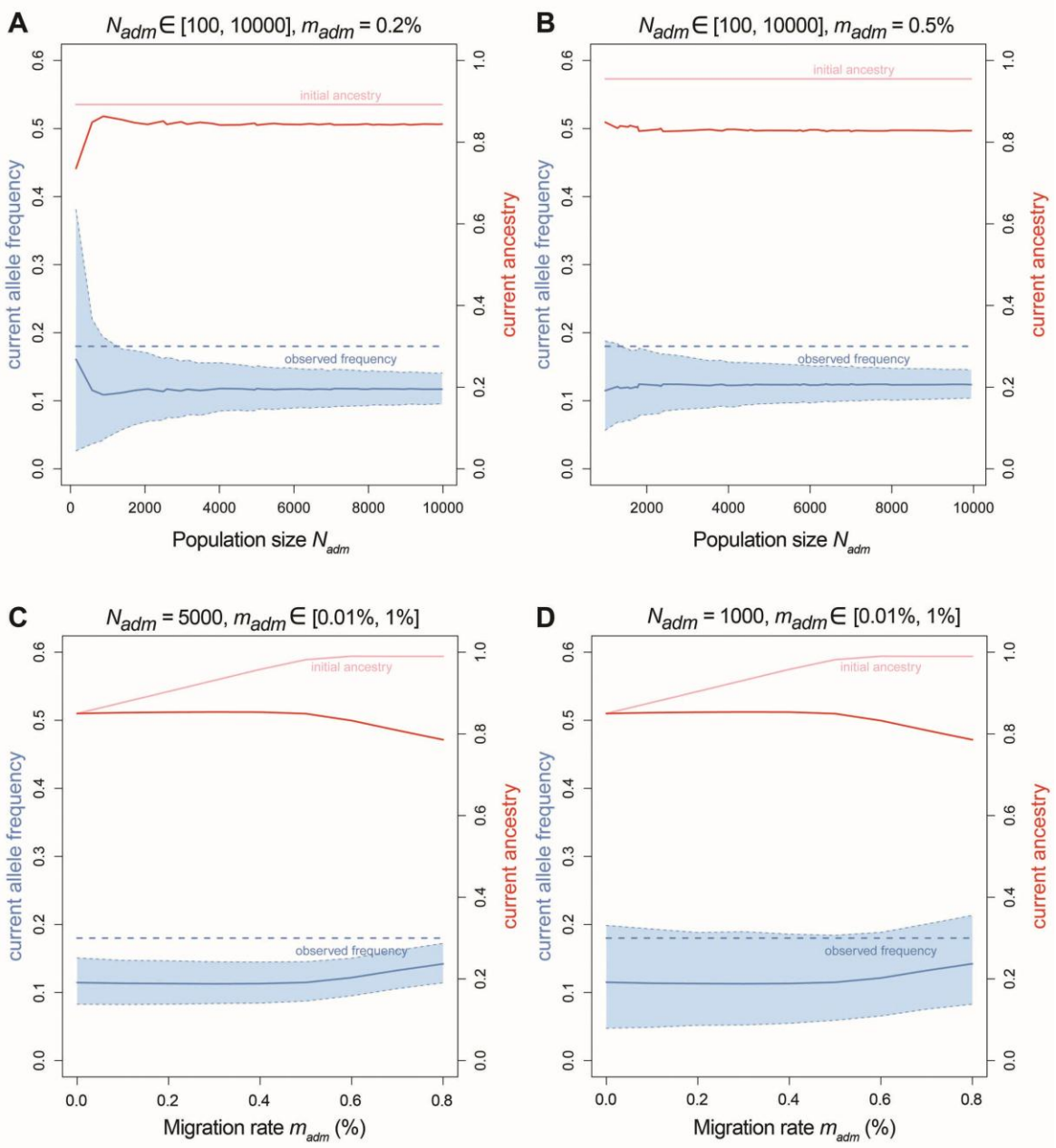

Fig. S16. Simulation-based evidence for positive selection on the *HLA* introgressed allele in western central Bantu-speaking populations (wBSP). We performed 10,000 Wright-Fisher simulations of an admixed population with (**A**, **B**) varying population size N_{adm} and (**C**, **D**) varying continuous, constant migration rate m_{adm} from the local, parental population. The dashed blue line shows the observed allele frequency in wBSPs, and the blue area the 95% quantile of the distribution of simulated allele frequencies. The light pink line shows the initial, simulated BSP ancestry in the admixed population, and the red line the final BSP ancestry (in the last generation). We tuned the initial BSP ancestry so that the final ancestry fits that estimated by GLOBETROTTER. When m_{adm} was high, the initial ancestry was set to 99%, in order to provide a realistic final BSP ancestry proportion. Higher m_{adm} were not simulated because they resulted in unrealistic final BSP ancestry proportions.

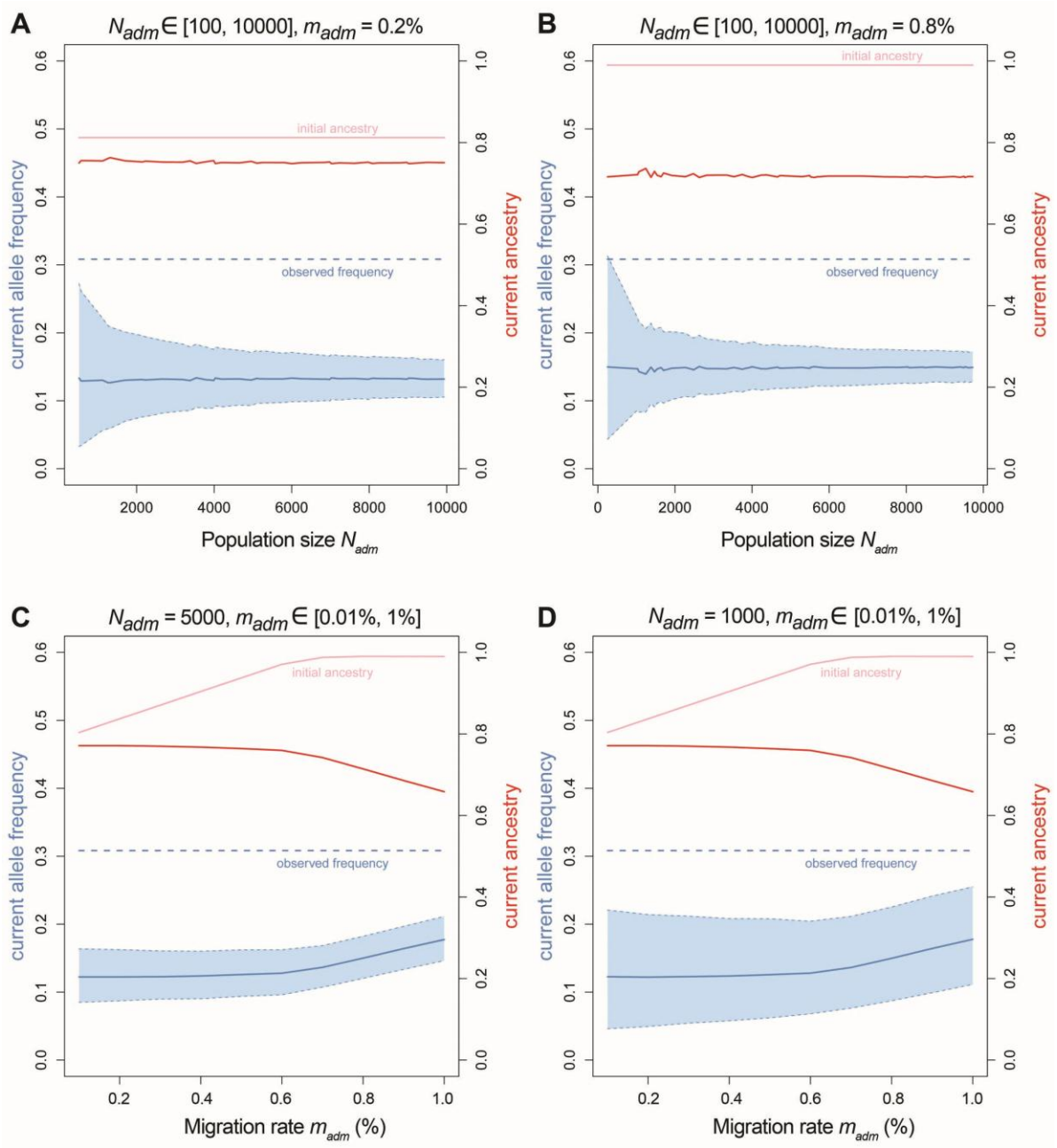

Fig. S17. Simulation-based evidence for positive selection on the *LCT* introgressed allele in eastern Bantu-speaking populations (eBSP). We performed 10,000 Wright-Fisher simulations of an admixed population with (**A**, **B**) varying population size N_{adm} and (**C**, **D**) varying continuous, constant migration rate m_{adm} from the local, parental population. The dashed blue line shows the observed allele frequency in wBSPs, and the blue area the 95% quantile of the distribution of simulated allele frequencies. The light pink line shows the initial, simulated BSP ancestry in the admixed population, and the red line the final BSP ancestry (in the last generation). We tuned the initial simulated BSP ancestry so that the final ancestry fits that estimated by GLOBETROTTER. When m_{adm} was high, the initial ancestry was set to 99%, in order to provide a realistic final BSP ancestry proportion. Higher m_{adm} were not simulated because they resulted in unrealistic final BSP ancestry proportions.

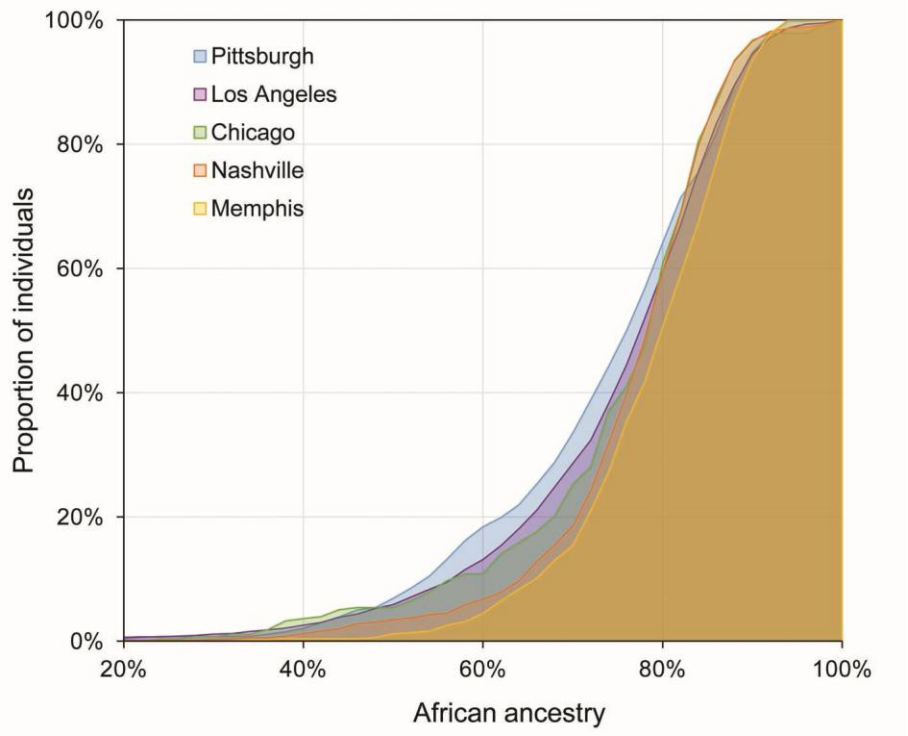

Fig. S18. Cumulative distribution of the African genetic ancestry of African Americans (AA) from five different cities of North America. Population structure analyses were performed on 5,244 AAs and 323,510 independent SNPs (fig. S19) with ADMIXTURE.

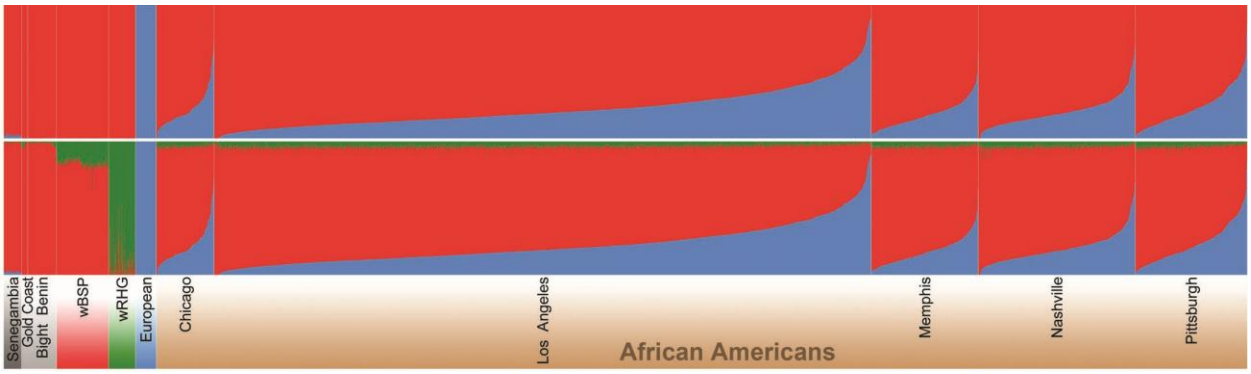

Fig. S19. Population structure of African Americans (AAs) from five different cities of North America. Population structure analyses were performed with ADMIXTURE on 5,244 AA and 323,510 independent SNPs. Five iterations with different random seeds were performed for K=2 and K=3. The iterations with the lowest CV error are shown.

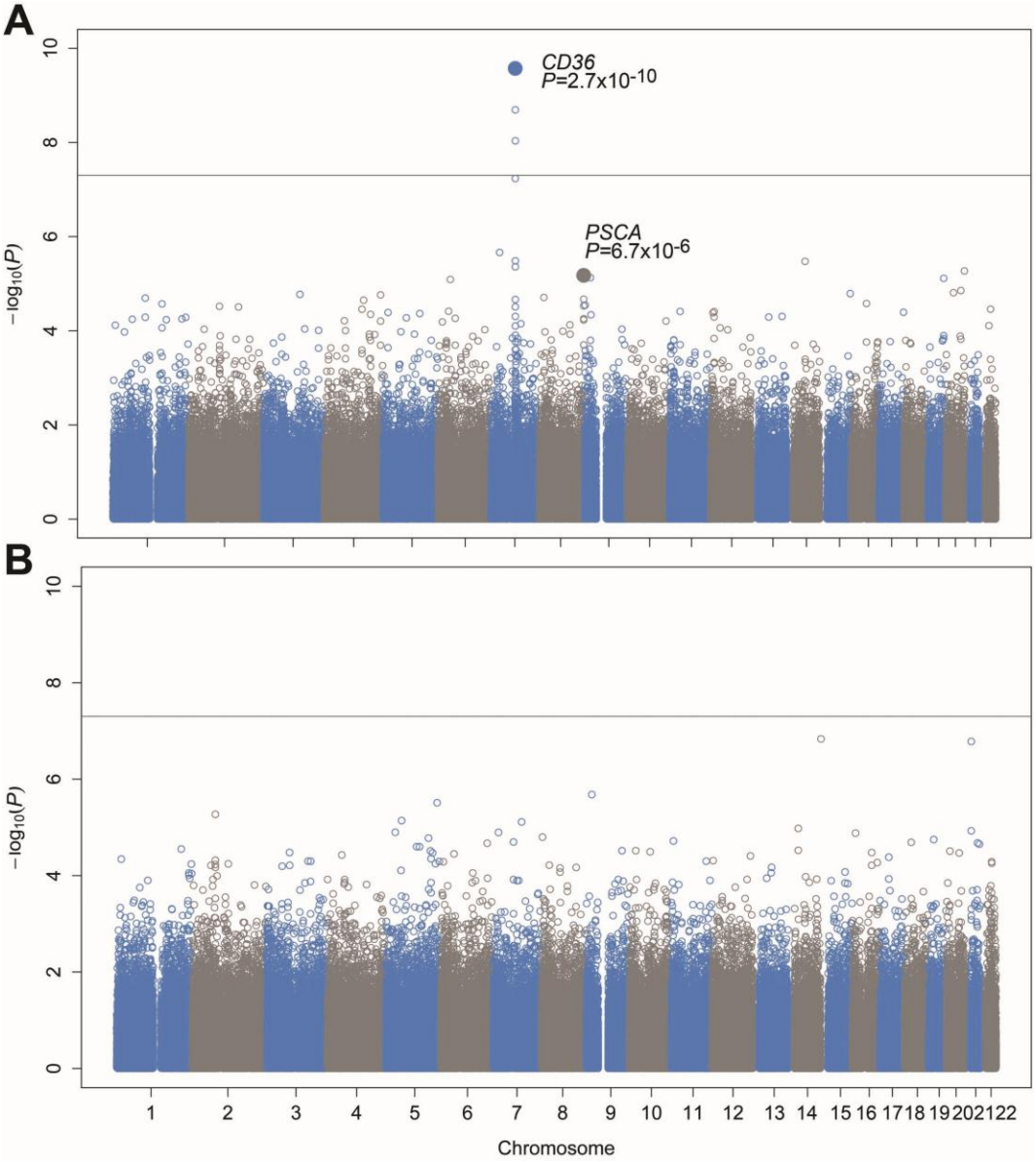

Fig. S20. Genome-wide scan for post-admixture selection in 5,244 African Americans. At each SNP, we tested if allele frequencies in the parental African population and the African genome of AAs were significantly different, following a previous study (*25*). Two sets of African parental populations were used: (**A**) the classically-used Yoruba of Nigeria, or (**B**) a more realistic, diverse set of African populations from Senegambia, the Windward Coast, the Bight of Benin and western central Africa (Fig. 4 and table S11). The genome-wide significant signal at *CD36* is lost when using the diverse population dataset.

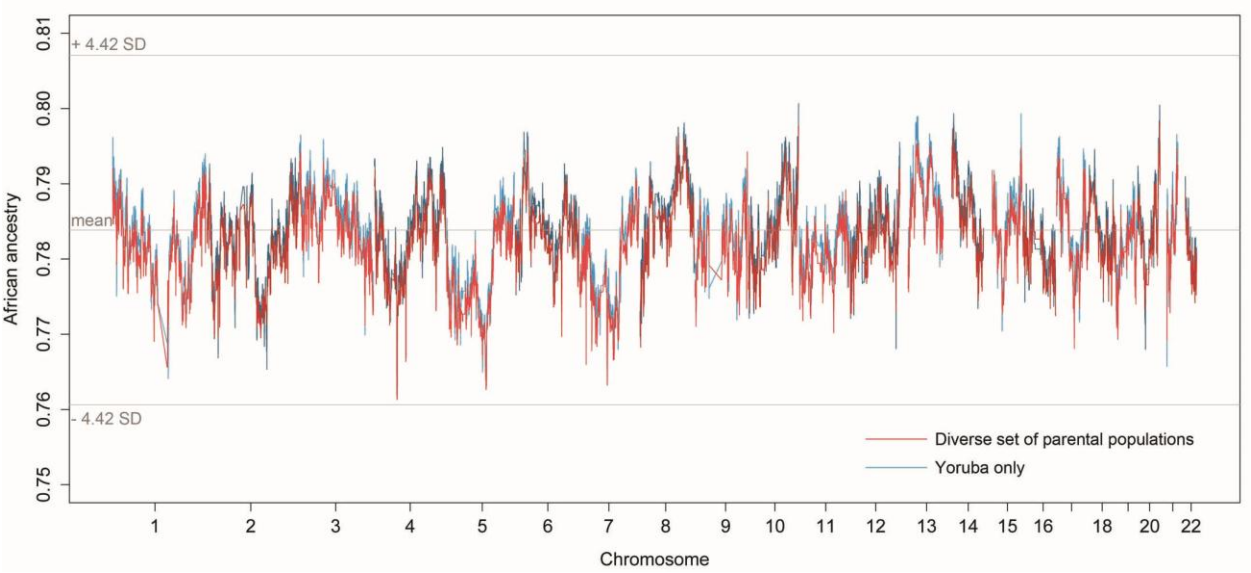

Fig. S21. African local ancestry along the genome of African Americans. The genome of 5,244 AAs was decomposed into African and European ancestry blocks with RFMix. Two sets of African parental populations were used: the classically-used Yoruba of Nigeria (in blue), or a more realistic, diverse set of African populations (in red) from Senegambia, the Windward Coast, the Bight of Benin and western central Africa (Fig. 4 and table S11). Limited differences are observed between the two analyses. Furthermore, no region reaches the significance threshold of 4.42 standard deviations from the mean (*25*).

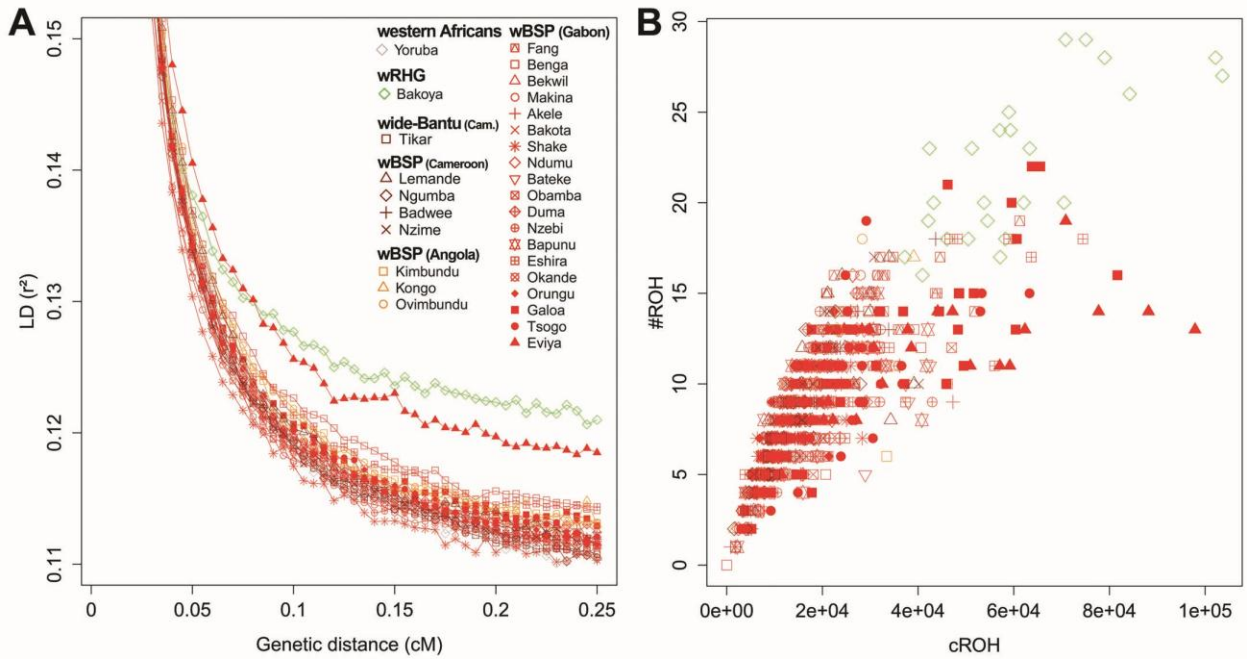

Fig. S22. Signals of decreased effective population size in the Eviya wBSP. (**A**) Linkage disequilibrium (LD) decay, with genetic distance, in western central African Bantu-speaking populations (wBSPs). For each population, LD levels estimated with r^2 were averaged over 50 bins of genetic distances and over five random resamplings of 10 individuals. (**B**) Runs of homozygosity (ROH) in wBSPs. The x-axis shows the cumulative length of ROHs of all analysed individuals (cROH). The y-axis shows the number of ROHs detected per individual (#ROH). The Bakoya RHG and the Yoruba of Nigeria were also included, for comparison purposes.

Supplementary Tables

Table S1. Geographical location, language and lifestyle of the newly genotyped populations, as a separate Excel file

Table S2. Additional included populations and cohorts from previously published datasets, as a separate Excel file

Table S3. Genetic differentiation between sub-Saharan Africans, estimated by the average AMOVA-based F_{ST} for 406,798 independent SNPs, as a separate Excel file

Table S4. Admixture events in African Bantu-speaking populations estimated by GLOBETROTTER, as a separate Excel file

Table S5. Number of shared Identical-By-Descent (IBD) segments among Bantu speakers, as a separate Excel file

Table S6. Admixture events in African hunter-gathering and Bantu-speaking farming populations detected by 3-pop ALDER, as a separate Excel file

Table S7. Candidate genomic regions for recent positive selection in western central Bantu-speaking populations, as a separate Excel file

Table S8. Candidate genomic regions for recent positive selection in eastern Bantu-speaking populations, as a separate Excel file

Table S9. Candidate genomic regions for recent positive selection in southeastern Bantu-speaking populations, as a separate Excel file

Table S10. Simulation-based evidence for adaptive introgression at the *HLA* locus in wBSPs and at the *LCT* locus in eBSPs, as a separate Excel file

Table S11. Estimated parameters of admixture in African Americans, as a separate Excel file

Table S12. Tests for direct contribution of western African rainforest hunter-gatherers to African Americans, as a separate Excel file

References

1. D. W. Phillipson, *African Archaeology* (Cambridge Univ. Press, ed. 2, 1993).

2. G. B. Busby, G. Band, Q. Si Le, M. Jallow, E. Bougama, V. D. Mangano, L. N. Amenga-Etego, A. Enimil, T. Apinjoh, C. M. Ndila, A. Manjurano, V. Nyirongo, O. Doumba, K. A. Rockett, D. P. Kwiatkowski, C. C. A. Spencer; Malaria Genomic Epidemiology Network, Admixture into and within sub-Saharan Africa. *eLife* **5**, ••• (2016). doi:10.7554/eLife.15266 Medline

3. C. de Filippo, K. Bostoen, M. Stoneking, B. Pakendorf, Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc. Biol. Sci.* **279**, 3256–3263 (2012). doi:10.1098/rspb.2012.0318 Medline

4. S. Li, C. Schlebusch, M. Jakobsson, Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proc. Biol. Sci.* **281**, 20141448 (2014). doi:10.1098/rspb.2014.1448

5. S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, S. M. Williams, The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009). doi:10.1126/science.1172257 Medline

6. J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Güldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P.-R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, B. Pakendorf, The genetic prehistory of southern Africa. *Nat. Commun.* **3**, 1143 (2012). doi:10.1038/ncomms2140 Medline

7. C. M. Schlebusch, P. Skoglund, P. Sjödin, L. M. Gattepaille, D. Hernandez, F. Jay, S. Li, M. De Jongh, A. Singleton, M. G. B. Blum, H. Soodyall, M. Jakobsson, Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379 (2012). doi:10.1126/science.1227721 Medline

8. C. Jeong, G. Alkorta-Aranburu, B. Basnyat, M. Neupane, D. B. Witonsky, J. K. Pritchard, C. M. Beall, A. Di Rienzo, Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat. Commun.* **5**, 3281 (2014). Medline

9. D. Eltis, D. Richardson, *Atlas of the Transatlantic Slave Trade* (Yale Univ. Press, 2010).

10. See the supplementary materials.

11. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi:10.1101/gr.094052.109 Medline

12. D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, K. Hatzikotoulas, S. Karthikeyan, L. Iles, M. O. Pollard, A. Choudhury, G. R. S. Ritchie, Y. Xue, J. Asimit, R. N. Nsubuga, E. H. Young, C. Pomilla, K. Kivinen, K. Rockett, A. Kamali, A. P. Doumatey, G. Asiki, J. Seeley, F. Sisay-Joof, M. Jallow, S. Tollman, E. Mekonnen, R. Ekong, T. Oljira, N. Bradman, K. Bojang, M. Ramsay, A. Adeyemo, E. Bekele, A.

Motala, S. A. Norris, F. Pirie, P. Kaleebu, D. Kwiatkowski, C. Tyler-Smith, C. Rotimi, E. Zeggini, M. S. Sandhu, The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015). doi:10.1038/nature13997 Medline

13. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLOS Genet.* **8**, e1002453 (2012). doi:10.1371/journal.pgen.1002453 Medline

14. G. Hellenthal, G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, S. Myers, A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014). doi:10.1126/science.1243518 Medline

15. R. Grollemund, S. Branford, K. Bostoen, A. Meade, C. Venditti, M. Pagel, Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13296–13301 (2015). doi:10.1073/pnas.1503793112 Medline

16. K. Bostoen, B. Clist, C. Doumenge, R. Grollemund, J.-M. Hombert, J. K. Muluwa, J. Maley, Middle to Late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of Western Central Africa. *Curr. Anthropol.* **56**, 354–384 (2015). doi:10.1086/681436

17. P. R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, B. Berger, Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013). doi:10.1534/genetics.112.147330 Medline

18. D. P. Kwiatkowski, How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77**, 171–192 (2005). doi:10.1086/432519 Medline

19. M. Deschamps, G. Laval, M. Fagny, Y. Itan, L. Abel, J.-L. Casanova, E. Patin, L. Quintana-Murci, Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.* **98**, 5–21 (2016). doi:10.1016/j.ajhg.2015.11.014 Medline

20. S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, P. Deloukas, Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007). doi:10.1038/ng1946 Medline

21. K. Bryc, A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser, S. Williams, A. Froment, J.-M. Bodo, C. Wambebe, S. A. Tishkoff, C. D. Bustamante, Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 786–791 (2010). doi:10.1073/pnas.0909559107 Medline

22. K. Bryc, E. Y. Durand, J. M. Macpherson, D. Reich, J. L. Mountain, The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015). doi:10.1016/j.ajhg.2014.11.010 Medline

23. F. Montinaro, G. B. J. Busby, V. L. Pascali, S. Myers, G. Hellenthal, C. Capelli, Unravelling the hidden ancestry of American admixed populations. *Nat. Commun.* **6**, 6596 (2015). doi:10.1038/ncomms7596 Medline

24. W. Jin, S. Xu, H. Wang, Y. Yu, Y. Shen, B. Wu, L. Jin, Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* **22**, 519–527 (2012). doi:10.1101/gr.124784.111 Medline

25. G. Bhatia, A. Tandon, N. Patterson, M. C. Aldrich, C. B. Ambrosone, C. Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, N. Caporaso, G. Casey, S. L. Deming, W. R. Diver, S. M. Gapstur, E. M. Gillanders, C. C. Harris, B. E. Henderson, S. A. Ingles, W. Isaacs, P. L. De Jager, E. M. John, R. A. Kittles, E. Larkin, L. H. McNeill, R. C. Millikan, A. Murphy, C. Neslund-Dudas, S. Nyante, M. F. Press, J. L. Rodriguez-Gil, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. A. Tucker, J. K. Wiencke, J. S. Witte, X. Wu, Y. Yamamura, K. A. Zanetti, W. Zheng, R. G. Ziegler, S. J. Chanock, C. A. Haiman, D. Reich, A. L. Price, Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *Am. J. Hum. Genet.* **95**, 437–444 (2014). doi:10.1016/j.ajhg.2014.08.011 Medline

26. G. Berniell-Lee, F. Calafell, E. Bosch, E. Heyer, L. Sica, P. Mouguiama-Daouda, L. van der Veen, J.-M. Hombert, L. Quintana-Murci, D. Comas, Genetic and demographic implications of the Bantu expansion: Insights from human paternal lineages. *Mol. Biol. Evol.* **26**, 1581–1589 (2009). doi:10.1093/molbev/msp069 Medline

27. L. Quintana-Murci, H. Quach, C. Harmant, F. Luca, B. Massonnet, E. Patin, L. Sica, P. Mouguiama-Daouda, D. Comas, S. Tzur, O. Balanovsky, K. K. Kidd, J. R. Kidd, L. van der Veen, J.-M. Hombert, A. Gessain, P. Verdu, A. Froment, S. Bahuchet, E. Heyer, J. Dausset, A. Salas, D. M. Behar, Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1596–1601 (2008). doi:10.1073/pnas.0711467105 Medline

28. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W.-M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010). doi:10.1093/bioinformatics/btq559 Medline

29. J. P. Jarvis, L. B. Scheinfeldt, S. Soi, C. Lambert, L. Omberg, B. Ferwerda, A. Froment, J.-M. Bodo, W. Beggs, G. Hoffman, J. Mezey, S. A. Tishkoff, Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLOS Genet.* **8**, e1002641 (2012). doi:10.1371/journal.pgen.1002641 Medline

30. L. Pagani, S. Schiffels, D. Gurdasani, P. Danecek, A. Scally, Y. Chen, Y. Xue, M. Haber, R. Ekong, T. Oljira, E. Mekonnen, D. Luiselli, N. Bradman, E. Bekele, P. Zalloua, R. Durbin, T. Kivisild, C. Tyler-Smith, Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96**, 986–991 (2015). doi:10.1016/j.ajhg.2015.04.019 Medline

31. E. Patin, K. J. Siddle, G. Laval, H. Quach, C. Harmant, N. Becker, A. Froment, B. Régnault, L. Lemée, S. Gravel, J.-M. Hombert, L. Van der Veen, N. J. Dominy, G. H. Perry, L. B. Barreiro, P. Verdu, E. Heyer, L. Quintana-Murci, The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat. Commun.* **5**, 3163 (2014). doi:10.1038/ncomms4163 Medline

32. G. H. Perry, M. Foll, J.-C. Grenier, E. Patin, Y. Nédélec, A. Pacis, M. Barakatt, S. Gravel, X. Zhou, S. L. Nsobya, L. Excoffier, L. Quintana-Murci, N. J. Dominy, L. B. Barreiro, Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-

gatherers. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3596–E3603 (2014). doi:10.1073/pnas.1402875111 Medline

33. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi:10.1086/519795 Medline

34. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005). doi:10.1093/bioinformatics/bth457 Medline

35. K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart; International HapMap Consortium, A second generation human

haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007). doi:10.1038/nature06258 Medline

36. O. Delaneau, J.-F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013). doi:10.1038/nmeth.2307 Medline

37. S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E. C. Royrvik, B. Cunliffe, D. J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, W. Bodmer; Wellcome Trust Case Control Consortium 2; International Multiple Sclerosis Genetics Consortium, The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015). doi:10.1038/nature14230 Medline

38. A. Gusev, J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, I. Pe'er, Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009). doi:10.1101/gr.081398.108 Medline

39. D. H. Huson, SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998). doi:10.1093/bioinformatics/14.1.68 Medline

40. P. E. Smouse, J. C. Long, R. R. Sokal, Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Syst. Zool.* **35**, 627–632 (1986). doi:10.2307/2413122

41. S. R. Grossman, I. Shlyakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, E. S. Lander, S. F. Schaffner, P. C. Sabeti, A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010). doi:10.1126/science.1183863 Medline

42. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLOS Biol.* **4**, e72 (2006). doi:10.1371/journal.pbio.0040072 Medline

43. X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, P. Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, J. Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, J. Wang, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010). doi:10.1126/science.1190371 Medline

44. P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D.

Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart; International HapMap Consortium, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007). doi:10.1038/nature06250 Medline

45. A. Ranciaro, M. C. Campbell, J. B. Hirbo, W.-Y. Ko, A. Froment, P. Anagnostou, M. J. Kotze, M. Ibrahim, T. Nyambo, S. A. Omar, S. A. Tishkoff, Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am. J. Hum. Genet.* **94**, 496–510 (2014). doi:10.1016/j.ajhg.2014.02.009 Medline

46. G. M. Sundaram, J. E. A. Common, F. E. Gopal, S. Srikanta, K. Lakshman, D. P. Lunny, T. C. Lim, V. Tanavde, E. B. Lane, P. Sampath, 'See-saw' expression of microRNA-198 and FSTL1 from a single transcript in wound healing. *Nature* **495**, 103–106 (2013). doi:10.1038/nature11890 Medline

47. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013). doi:10.1016/j.ajhg.2013.06.020 Medline

48. A. L. Price, M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, D. B. Goldstein, D. Reich, Long-range LD can confound

genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135, author reply 135–139 (2008). doi:10.1016/j.ajhg.2008.06.005 Medline

49. G. Bhatia, N. Patterson, B. Pasaniuc, N. Zaitlen, G. Genovese, S. Pollack, S. Mallick, S. Myers, A. Tandon, C. Spencer, C. D. Palmer, A. A. Adeyemo, E. L. Akylbekova, L. A. Cupples, J. Divers, M. Fornage, W. H. L. Kao, L. Lange, M. Li, S. Musani, J. C. Mychaleckyj, A. Ogunniyi, G. Papanicolaou, C. N. Rotimi, J. I. Rotter, I. Ruczinski, B. Salako, D. S. Siscovick, B. O. Tayo, Q. Yang, S. McCarroll, P. Sabeti, G. Lettre, P. De Jager, J. Hirschhorn, X. Zhu, R. Cooper, D. Reich, J. G. Wilson, A. L. Price, Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet.* **89**, 368–381 (2011). doi:10.1016/j.ajhg.2011.07.025 Medline

50. J. A. Hodgson, J. K. Pickrell, L. N. Pearson, E. E. Quillen, A. Prista, J. Rocha, H. Soodyall, M. D. Shriver, G. H. Perry, Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. *Proc. Biol. Sci.* **281**, 20140930 (2014). doi:10.1098/rspb.2014.0930 Medline

51. C. Aimé, G. Laval, E. Patin, P. Verdu, L. Ségurel, R. Chaix, T. Hegay, L. Quintana-Murci, E. Heyer, F. Austerlitz, Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. *Mol. Biol. Evol.* **30**, 2629–2644 (2013). doi:10.1093/molbev/mst156 Medline

52. G. Laval, E. Patin, L. B. Barreiro, L. Quintana-Murci, Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLOS ONE* **5**, e10284 (2010). doi:10.1371/journal.pone.0010284 Medline

53. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). doi:10.1534/genetics.112.145037 Medline

54. D. Shriner, A. Adeyemo, E. Ramos, G. Chen, C. N. Rotimi, Mapping of disease-associated variants in admixed populations. *Genome Biol.* **12**, 223 (2011). doi:10.1186/gb-2011-12-5-223 Medline