Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation

Laura B. Scheinfeldt[a,1,2], Sameer Soi[a,b,1], Charla Lambert[a,3], Wen-Ya Ko[a,4], Aoua Coulibaly[a], Alessia Ranciaro[a], Simon Thompson[a], Jibril Hirbo[a,5], William Beggs[a], Muntaser Ibrahim[c], Thomas Nyambo[d], Sabah Omar[e], Dawit Woldemeskel[f], Gurja Belay[f], Alain Froment[g], Junhyong Kim[h], and Sarah A. Tishkoff[a,h,6]

[a]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [b]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA 19104; [c]Department of Molecular Biology, Institute of Endemic Diseases, University of Khartoum, Khartoum, Sudan; [d]Department of Biochemistry, St. Joseph University College of Health Sciences, Dar es Salaam, Tanzania; [e]Kenya Medical Research Institute, Center for Biotechnology Research and Development, Nairobi, Kenya; [f]Department of Biology, Addis Ababa University, Addis Ababa, Ethiopia; [g]UMR 208, Institut de Recherche pour le Développement-Muséum National d'Histoire Naturelle, Musée de l'Homme, 75116 Paris, France; and [h]Department of Biology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104

Contributed by Sarah A. Tishkoff, January 5, 2019 (sent for review October 15, 2018; reviewed by Rob J. Kulathinal and Mark D. Shriver)

Anatomically modern humans arose in Africa ∼300,000 years ago, but the demographic and adaptive histories of African populations are not well-characterized. Here, we have generated a genome-wide dataset from 840 Africans, residing in western, eastern, southern, and northern Africa, belonging to 50 ethnicities, and speaking languages belonging to four language families. In addition to agriculturalists and pastoralists, our study includes 16 populations that practice, or until recently have practiced, a hunting-gathering (HG) lifestyle. We observe that genetic structure in Africa is broadly correlated not only with geography, but to a lesser extent, with linguistic affiliation and subsistence strategy. Four East African HG (EHG) populations that are geographically distant from each other show evidence of common ancestry: the Hadza and Sandawe in Tanzania, who speak languages with clicks classified as Khoisan; the Dahalo in Kenya, whose language has remnant clicks; and the Sabue in Ethiopia, who speak an unclassified language. Additionally, we observed common ancestry between central African rainforest HGs and southern African San, the latter of whom speak languages with clicks classified as Khoisan. With the exception of the EHG, central African rainforest HGs, and San, other HG groups in Africa appear genetically similar to neighboring agriculturalist or pastoralist populations. We additionally demonstrate that infectious disease, immune response, and diet have played important roles in the adaptive landscape of African history. However, while the broad biological processes involved in recent human adaptation in Africa are often consistent across populations, the specific loci affected by selective pressures more often vary across populations.

African hunter-gatherers | African diversity | population genetics | natural selection | human evolution

Genetic, archaeological, and linguistic evidence reflect a complex demographic history for populations in Africa. Anatomically modern humans emerged in Africa ∼300 kya (1–3) and lived in Africa for tens of thousands of years before a subset migrated out of Africa 80–40 kya (4). Many studies have focused on when, where, and how modern humans colonized the rest of the globe, but relatively few have characterized prehistoric demography within Africa during the Late Pleistocene 70–10 kya (4). This is likely because the archaeological and paleo-biological record is incomplete during that time period (5), linguistic reconstruction does not extend much beyond 10 kya (6), and recent demographic events, such as historical migrations, complicate genomic signatures of older population movements and interactions.

Relatively more is known about population histories in Africa during the recent past due to linguistic reconstructions. One of the most striking recent demographic events in Africa was the expansion of Bantu peoples (speakers of Bantu languages) from West Africa accompanying agricultural innovation in the Neolithic ∼5 kya (7). This expansion, commonly referred to as the "Bantu expansion," significantly impacted the landscape of genetic and cultural diversity in Africa (8, 9). While Bantu languages, which belong to the Niger-Congo (NC) language family, are widely spoken across Africa, languages belonging to two additional language families, Nilo-Saharan (NS) and Afro-Asiatic (AA), are spoken by populations primarily located in central, eastern, and northern Africa who practice pastoralism and agriculture (10). A fourth language family, Khoisan, which contains click phonemes, includes several languages spoken by hunter-gatherer populations in southern Africa, as well as two languages spoken by hunter-gatherer populations in eastern Africa, the Hadza and Sandawe (11). While the Sandawe language has been identified as linguistically more similar to the Khoisan languages spoken in southern Africa than it is to the Hadza language, the inclusion of the latter two languages within the Khoisan language family is generally contentious, arguably

Significance

African populations have been underrepresented in human genomics research yet are important for understanding modern human origins and the genetic basis of adaptive traits. Here we analyze a genome-wide dataset in 840 ethnically and geographically diverse Africans. We find that geographically distant hunter-gatherer populations from East Africa share unique common ancestry and we see strong signatures of local adaptation near genes that play a role in immune response, as well as lipid and glucose metabolism.

Author contributions: L.B.S., J.K., and S.A.T. designed research; L.B.S., S.S., W.-Y.K., A.R., S.T., J.H., M.I., T.N., S.O., D.W., G.B., A.F., and S.A.T. performed research; L.B.S., S.S., C.L., A.C., W.B., and S.A.T. analyzed data; and L.B.S., S.S., J.K., and S.A.T. wrote the paper.

Reviewers: R.J.K., Temple University; and M.D.S., Pennsylvania State University.

The authors declare no conflict of interest.

Data deposition: Genotype data from this study have been deposited in the NIH dbGAP repository, https://www.ncbi.nlm.nih.gov/gap (accession no. phs001780.v1.p1).

[1]L.B.S. and S.S. contributed equally to this work.

[2]Present address: Coriell Institute for Medical Research, Camden, NJ 08103.

[3]Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

[4]Present address: Department of Life Sciences and Institute of Genome Sciences, National Yang Ming University, Taipei City 112, Taiwan.

[5]Present address: Division of Genetic Medicine, Vanderbilt University Medical Center, Vanderbilt University, Nashville, TN 37232.

[6]To whom correspondence should be addressed. Email: tishkoff@pennmedicine.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1817678116/-/DCSupplemental.

Published online February 19, 2019.

because the relationships between the eastern and southern African Khoisan languages are older than 10 kya (12).

Several languages spoken throughout Africa remain unclassified and are considered "language isolates." One such example is the Shabo language, also referred to as Mikeyir, spoken by the people of Ethiopia who self-identify as Sabue (also known as Sabu). While proto-Shabo is thought to be an early branch of the NS languages, the classification of Shabo into any linguistic family is unresolved (13, 14). The language spoken by the Dahalo, also referred to as Sanye, of Kenya is another such example. Some linguists classify Dahalo as AA, but it also shares a dental click phoneme with Khoisan languages (12). The shared presence of clicks has led linguists to hypothesize that the Dahalo share recent common ancestry with the Hadza and Sandawe or that ancestors of the Dahalo, speaking a proto-Dahalo language, came into contact with and subsequently borrowed linguistic features from individuals speaking a proto-Khoisan language in East Africa (15). The archaeological evidence for a common ancestry of East African hunting-gathering (HG) populations and the Khoisan-speaking populations of southern Africa is debated: while there has been some evidence of a genetic connection, the archaeological data are not conclusive (16–19). However, there have been no prior genetic studies of the Sabue or Dahalo. Here, we examine the genetic relationships of the Sabue, Dahalo, and the Khoisan-speaking populations of eastern Africa to shed light on the history of East African HG populations.

In addition to the linguistic diversity found in Africa—over 2,000 languages are spoken in the continent—African populations practice diverse subsistence strategies (11). As previously noted, agricultural technologies spread throughout sub-Saharan Africa with the Bantu expansion 5–3 kya. Before that, pastoralism spread from northeastern Africa southward into central and eastern Africa 6–3 kya (10). Populations speaking Khoisan languages, including the Hadza and Sandawe, engage, or until recently engaged, in an HG subsistence strategy. The Dahalo, Boni, El Molo, Yaaku, Sengwer, and Ogiek populations living in Kenya, and the Wata from Ethiopia, also practice an HG lifestyle, as do the Sabue of Ethiopia. Anthropologists have debated whether these East African HG populations represent distinct groups or whether they represent descendants of communities that were displaced due to past political, economic, and social phenomena (20). Other African populations traditionally practicing HG include the western (e.g., Biaka, Baka, Bakola, Bedzan) and eastern (e.g., Mbuti) rain forest hunters and gatherers (WRHG and ERHG, respectively), commonly referred to as "Pygmies," who have adopted the languages of neighboring populations, and the Khoisan-speaking San from southern Africa. Here, we analyze the genetic diversity of 16 ethnic groups in Africa that practice a foraging life-style to better understand their relationships with each other and with neighboring populations.

Taken together, linguistic, archaeological, and genetic data have led to a proposed wide range for Khoisan-speaking HG populations throughout southern and eastern Africa, extending from Ethiopia to southern Africa (6, 15, 21, 22). However, this hypothesis remains contentious, and the origins of East African HG populations remain unknown largely because of the limits of linguistic reconstruction, archaeological data, and sparse sampling of genomic diversity in East Africa (12, 16). To explore this question further, we have genotyped 724 individuals from 46 diverse ethno-linguistic populations living in central and eastern Africa with the Illumina 1M-Duo SNP array (Fig. 1*A*), including all of the eastern African and WRHG populations described above. We merged these data with publicly available data from population samples including Mbuti ERHG living in the Democratic Republic of Congo, San living in Namibia and South Africa, Mandenka living in Senegal, and Mozabite living in Algeria (23, 24). In total, the merged dataset is comprised of 840 individuals sampled from 50 populations living throughout sub-Saharan Africa and genotyped for a set of ~621,000 markers present on all platforms (*SI Appendix*, Table S1).

Results

Genome-Wide Patterns of Diversity. To characterize genome-wide patterns of diversity in Africa, we employed principal components analysis (PCA) of individuals at 621,000 biallelic SNPs (25, 26) (Fig. 1 *B* and *C* and *SI Appendix*, Fig. S1). The first principal component (PC1), which explains 2.11% of the genotypic variance, is well predicted by a linear model with latitude, longitude, and linguistic affiliation variables ($R^2 = 0.86$; $P < 1.0 \times 10^{-16}$) (*SI Appendix*, Fig. S2*A*). On one extreme end of the PC1 axis are North African Mozabite (Algeria) individuals, and on the other end of the axis are Mbuti ERHG (Democratic Republic of Congo) and San (southern Africa) individuals (Fig. 1*B*). We also observed a good fit between PC2, which explains 0.91% of the genotypic variance, and a linear model with latitude, longitude, and linguistic affiliation variables (adjusted $R^2 = 0.58$, $P < 1.0 \times 10^{-16}$), albeit less strongly than PC1 (*SI Appendix*, Fig. S2*B*). Individuals speaking NC languages are represented at one end of the PC2 axis and San individuals at the other end. Thus, geography and language are significantly correlated with patterns of genetic variation in Africa.

To explore our hypothesis of a possible common ancestry of the Hadza, Sandawe, Sabue, and Dahalo, heretofore referred to as the eastern HG (EHG), we tested whether they cluster more closely to each other in the PCA compared with other populations. We observed that they cluster significantly closer to each other than to any other populations on PC1 and PC2 based on a comparison of Euclidean distances among EHG and among EHG and non-EHG individuals (Wilcoxon rank-sum test: $W = 52{,}704{,}648$; $P < 1.0 \times 10^{-16}$) (*SI Appendix*, Fig. S3*A*). In addition, the Sabue, Hadza, and Dinka individuals significantly cluster together at one extreme of PC3 (Wilcoxon rank-sum test, $W = 30{,}247.5$; $P < 1.0 \times 10^{-16}$) (Fig. 1*C* and *SI Appendix*, Fig. S3*B*). These observations are consistent with possible shared ancestry between the Hadza and Sabue, and some evidence for shared ancestry of these populations with the Dinka (NS language) (27), as well as linguistic evidence supporting a relationship between the Shabo language and proto-NS (13). PCs explaining a smaller proportion of the genetic variance in the data are presented in *SI Appendix*, Fig. S1.

We explored patterns of population structure in African population samples using STRUCTURE analysis (28) (Fig. 1*D*) with a set of 20,000 SNPs, pruned to reduce linkage disequilibrium (LD). We also used haplotype clusters inferred by BEAGLE (29) as a *k*-allele system at the same 20,000 loci (*SI Appendix*, Fig. S4). We found that $K = 9$ was the number of ancestral allele clusters (AAC) that consistently produced the highest data likelihoods across runs for both genotypes (*SI Appendix*, Fig. S5*A*) and haplotypes (*SI Appendix*, Fig. S5*B*) without producing multiple modes (i.e., inferring different ancestral allele clusters across runs). Additionally, we observed lower variance in likelihood scores at $K = 9$ compared with higher values of *K*. At $K = 9$ (Fig. 1*D*), we find that individuals from populations that speak languages belonging to the same language family have significantly similar AAC proportions (Mantel test: $M = 0.473$; $P = 0.001$) (*SI Appendix*). However, several population samples are distinguished by unique AACs at $K = 9$. These include the North African Mozabite (Fig. 1*D*, dark blue) who have the greatest proportion of Saharan ancestry compared with other populations (Wilcoxon rank-sum test: $W = 191$; $P < 1.0 \times 10^{-16}$). The other AACs at $K = 9$ distinguish HG populations: San (Fig. 1*D*, light green), WRHG and ERHG (Fig. 1*D*, dark green), Hadza (Fig. 1*D*, yellow), Dahalo (Fig. 1*D*, pink), and Sabue (Fig. 1*D*, light blue) populations, respectively. Distinct AACs corresponding to HG populations may be explained by genetic drift caused by isolation of these populations or persistently small effective population sizes (N_e) (30). Unlike other EHG populations, the Sandawe are not enriched for a particular AAC at $K = 9$; rather, they have considerable AA (Fig. 1*D*, dark purple) and NC (Fig. 1*D*, orange) ancestry: 37.1% and 26.4% on average, respectively. In contrast, the Elmolo, Yaaku, Boni, Wata, Ogiek, and Sengwer from East Africa share ancestry with neighboring agriculturalist or

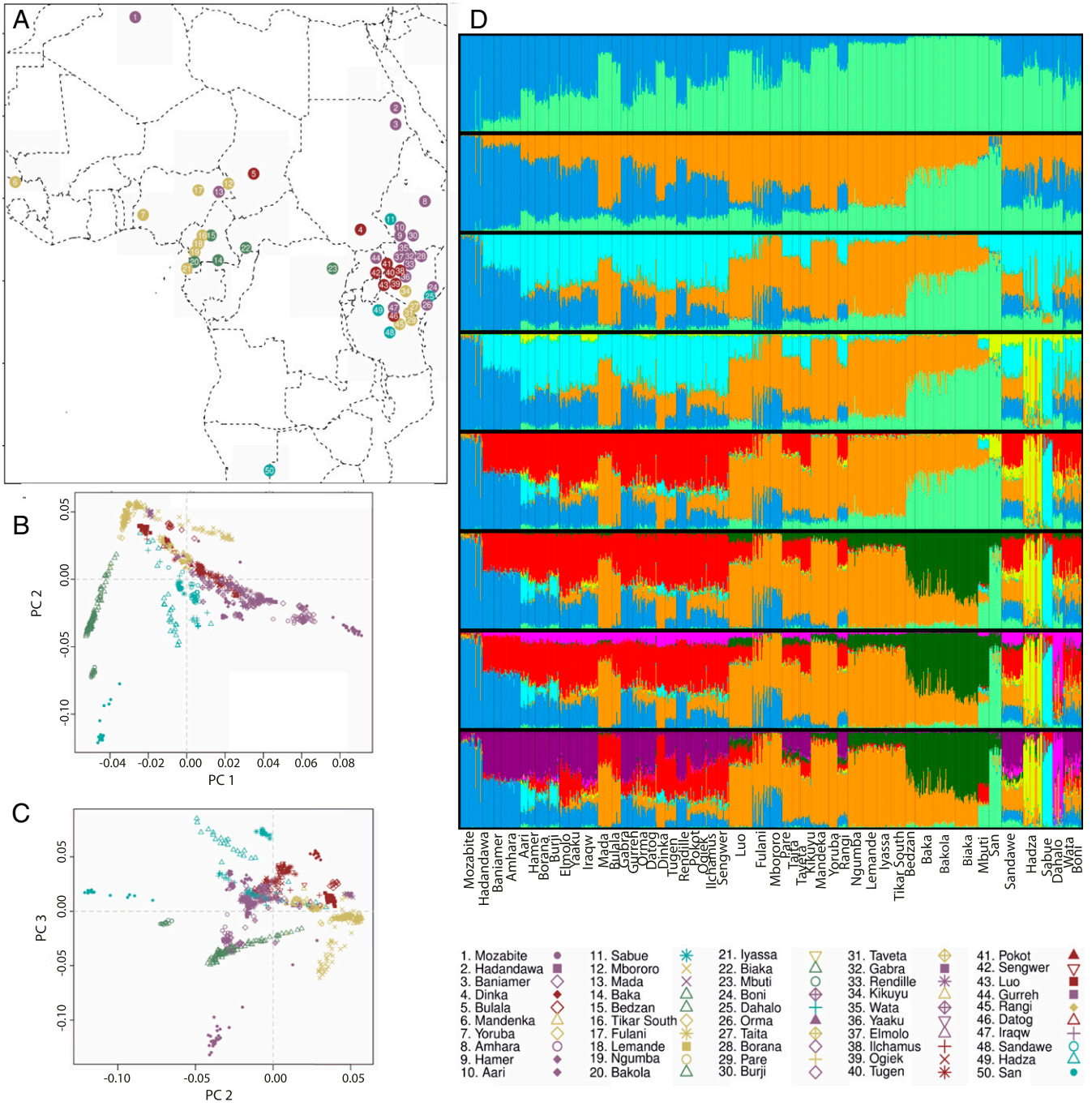

Fig. 1. Geographic distribution of populations studied and summaries of population structure. (*A*) The geographic distribution of populations included in the study presented on a map of Africa. The legend indicates the colors assigned to each language family and the number and unique combination of color and symbol for each ethno-linguistic population. (*B*) PCA was performed using individuals' genotypes; PC1, which explains 2.11% of the genotypic variance and shows a North–South cline, was plotted against PC2, which explains 0.91% of the genotypic variance and separates individuals with NC ancestry. (*C*) Hadza and Sabue individuals cluster at one extreme end of PC3, which explains 0.73% of variance in individuals' genotypes; NS-speaking individuals are also found clustering near the Hadza and Sabue. (*D*) Population structure was inferred using the STRUCTURE software using 20,000 unlinked loci; results are shown from $K = 2$ to $K = 9$, the latter of which was identified as having the best, most stable fit to the data. The STRUCTURE analysis revealed $K = 9$ AAC. Supporting the PCA, two AAC's corresponded to NC ancestry (orange); that is, correlated with the Bantu expansion, and North African ancestry (blue). In addition, the other AACs identify structure between HG populations: San (light green), WRHG (dark green), Hadza (yellow), Dahalo (light purple), and Sabue (light blue). Results from $K = 2$ to $K = 8$ are discussed in *SI Appendix*.

pastoralist populations. Patterns of clustering at lower AACs, which support ancient common ancestry between the rain forest HG and San and between the Hadza and Sabue, as well as the genetic relationship between AACs based on the inferred ancestral allele frequencies, are described in the *SI Appendix*.

Historical changes in population size contributes to contemporary genetic variation; therefore, we used patterns of LD decay to estimate N_e in population samples with at least 10 individuals (*SI Appendix*, Figs. S6 and S7*A*) (31). Several of the EHG, the Hadza, Dahalo, and Sabue, have relatively low estimates of N_e

(~9,000–11,000), consistent with their relatively smaller census population sizes (~1,000–3,000) (32). In contrast, the Sandawe and WRHG have maintained relatively higher N_e (on the order of 17,000 and 19,000, respectively), consistent with their larger census sizes (~30K) (33, 34). The estimates of N_e from LD in the Hadza, WRHG, and Sandawe are consistent with estimates of N_e based on levels of genetic diversity from whole-genome sequence data in the same populations (35). The largest N_e estimates are for agriculturalist and pastoralist populations, which is also consistent with prior studies (35, 36).

To examine the influence of demographic history on patterns of haplotype sharing within populations, we examined sharing of identity-by-descent (IBD) regions, which are stretches of DNA between individuals inherited from a common ancestor, and runs of homozygosity (ROH), which are stretches of DNA that are identical between the two haploid chromosomes of an individual. For each population, we calculated the average of the total IBD between all pairs of individuals, which we refer to as cumulative IBD (cIBD) and cumulative ROH (cROH) within individuals. Comparing cROH with cIBD for each population (*SI Appendix*, Fig. S7*B*), we find that the Hadandawa-Beja have the greatest cROH (195 cM) but are only the 15th highest for cIBD (45 cM); the high cROH is consistent with the documented practice of consanguineous marriages in this population (37). In contrast, the Hadza have the greatest cIBD (398 cM) as well as the second greatest cROH (158 cM). The presence of elevated cIBD and cROH in the Hadza is consistent with a small census size (~1,000), a low N_e and long-term endogamy (24, 35).

Historical Relationships Among African Populations. We reconstructed a population tree using pairwise estimates of genetic distance based on the F_{ST} statistic (Fig. 2*A*). We assessed statistical support for internal nodes in the neighbor-joining (NJ) population tree by bootstrapping loci with 1,000 replicates (*SI Appendix*). Broadly, the tree reflects geographic residence and linguistic affiliation as observed in the previous results. In addition, four geographically dispersed EHG populations—the Hadza, Sabue, Sandawe, and Dahalo—form a clade. The Hadza form a subclade with the Sabue with 97% bootstrap support. Support for inclusion of the Dahalo and Sandawe in the EHG clade is lower; however, examination of the bootstraps shows that this is because >80% of the replicates show the Dahalo cluster with the Boni, a neighboring population with whom they share recent contact (see IBD results below). As noted above, the linguistic relationships among these populations are unclear and contentious. While evidence for recent common ancestry between the neighboring Hadza and Sandawe has previously been shown (17, 24, 27), our results represent genetic evidence for a uniquely shared common ancestry of these populations with the Dahalo and Sabue from Kenya and Ethiopia, respectively. It is also noteworthy that other HG populations from central and southern Africa cluster with high bootstrap support: the San and Mbuti form a clade despite being geographically isolated from each other, and both form a clade with the WRHG, supporting results from PCA (Fig. 1 *B* and *C*) and STRUCTURE analyses (Fig. 1*D*) (18, 36, 38, 39). In contrast, other HG populations from East Africa (i.e., Ogiek, Dorobo, and so forth) cluster together with neighboring agriculturalist or pastoralist populations.

In addition to using F_{ST} to infer relationships between populations, we examined the distribution of the number and length of IBD tracts between individuals across populations to identify recent shared ancestry (40, 41). We explored the possibility that the signal of EHG common ancestry represents shared gene flow with Cushitic- and Bantu-speaking populations who expanded into East Africa within the past 5 kya (42). We used a distance measure, based on the ratio of IBD tracts (≥2 cM) found within and between populations, to construct a population tree (Fig. 2*B*). We compared F_{ST}- and IBD-based distances between populations (*SI Appendix*), as the latter measure is more sensitive to recent gene flow originating 25–50 generations ago (43). Unlike the F_{ST}-based tree, the EHG populations do not form a clade in the IBD-based tree, instead clustering with geographically proximate populations (Fig. 2*B*), indicating an increase in interactions between the EHG and neighboring agriculturalist and pastoralist populations in the recent past (18). Notably, the Boni and Dahalo, who neighbor each other, cluster together on the IBD tree, which

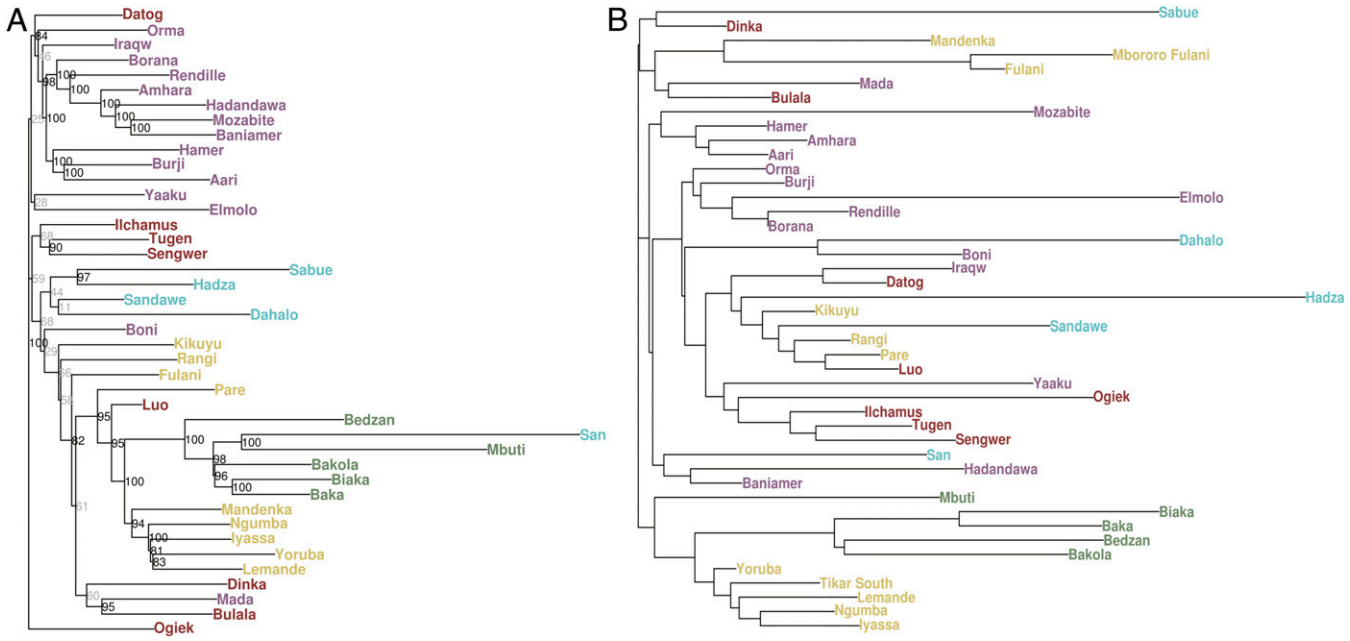

Fig. 2. Population trees. (*A*) An NJ population tree was inferred using estimates of pairwise genetic distances between populations based on F_{ST} values scaled by N_e. Populations largely cluster by geography or language affiliation, with the notable exceptions of the clade consisting of the Hadza, Sabue, Sandawe, and Dahalo and the clade consisting of the WRHG, ERHG, and San, whose populations cluster together despite being geographically distant. (*B*) An NJ population tree based on pairwise distances based on the ratio of within-population to between-population haplotype sharing (i.e., IBD); this statistic is more sensitive to recent demographic events, such as gene flow than F_{ST}. The EHG cluster most closely with neighboring populations.

explains the frequency with which they appear together in bootstrap replicates of the F_{ST}-based tree. Furthermore, in the IBD tree, the Dinka form a clade with the Sabue, consistent with the genotypic PCA (Fig. 1*C*), suggesting recent gene flow and/or shared ancestry. These observations are consistent with a model suggested by some linguists in which the Shabo language is classified with the NS language family (13), although other linguists (14) find inadequate evidence for a connection between the Shabo language and proto-NS. This uncertainty could be due to the age of the linguistic relationship between Shabo and proto-NS, which may predate the upper bound for linguistic reconstruction (~10 kya) (44).

The lack of clustering of the EHG in the IBD-based distance tree indicates that the pattern of shared ancestry among the EHG populations observed in the reconstructed F_{ST} -based population tree (Fig. 2*A*) is not due to recent events. We used an approach introduced by Fearnhead and Prangle (45) to construct summary statistics for approximate Bayesian computation (ABC) inference based not only on allele frequency differences between populations, but also on patterns of LD and admixture LD (i.e., LD weighted by differences in allele frequencies between populations) to infer divergence times between pairs of EHG populations (31, 46–51) (*SI Appendix*). The demographic model we employed (*SI Appendix*, Fig. S8) included changes in N_e, gene flow from populations speaking NC, NS, or AA languages, and ascertainment bias due to SNPs on the Illumina 1M array that were identified primarily in non-African populations.

The maximum a posteriori estimate and 95% credible interval for pairwise divergence time estimates are shown in Fig. 3. The maximum a posteriori divergence time estimates for the Hadza and Sandawe were 13 or 22 kya when accounting for different primary sources of admixture based on STRUCTURE analysis (Fig. 1*D*) (NC or AA, respectively); these estimates overlap with previous studies (17). The Hadza split times with other populations were older; the divergence time estimates with the Sabue (NC or AA gene flow) were 44 or 61 kya, respectively, and with the Dahalo (NC or AA gene flow) were 55 or 61 kya, respectively. Sandawe population divergence time estimates with the Sabue (NS or AA admixture) were 30 or 52 kya, respectively, and with the Dahalo (NS or AA gene flow) were 50 or 57 kya, respectively. The estimated times of divergence of the Sabue and Dahalo (NC or AA gene flow) were 63 or 72 kya, respectively. These results are consistent with a model in which population divergence between the Dahalo and Sabue and the ancestors of the Sandawe and Hadza occurred >30 kya, whereas the Hadza and Sandawe divergence was more recent. Whole-genome sequence analyses will be informative for more accurately resolving the time of population divergence among EHGs.

Genome-Wide Patterns of Adaptation. Over the past two decades, several genome-wide scans for selection have been developed and applied to worldwide human genetic data (52–54). Fewer studies, however, have focused on variation within Africa (24, 35, 36, 38, 55–57). These studies have tended to focus on specific regions (e.g., Southern Africa or Ethiopia) in Sub-Saharan Africa (24, 36, 55) or specific populations of interest living in Africa (35, 38, 57). Thus, the pattern and distribution of adaptive candidate loci among geographically and culturally diverse African populations is not well understood. We used three complementary statistical tests of neutrality to characterize African genome-wide signatures of adaptation. We combined individuals into larger population groupings based on shared ethno-linguistic affiliation and on shared ancestry, as inferred from PCA clustering (*SI Appendix*, Fig. S9) for the subsequent analyses.

Shared Adaptive Signals. Given the wide range of diverse populations sampled in the study, we were interested in studying the distribution of adaptive candidate genes within and among population groupings. We first employed the D statistic, an extension of the locus-specific branch length statistic that includes more than three population samples (58, 59), to identify signatures of regionally restricted adaptation within population groupings. We identified genes near (within 100 kb) SNPs in the top 0.1% of the empirical distribution of results for the D statistic test (expected to be enriched for targets of natural selection) for each population grouping, and we performed pathway-enrichment analyses of these adaptive candidate genes (*SI Appendix*, Table S3). Because the D statistic identifies SNPs with allele frequencies that are unusual in one sample relative to all others in the analysis (58), it was not surprising that the majority of top (0.1%) candidate genes (93%) occur in only a single population grouping (*SI Appendix*, Fig. S10 and Dataset S1).

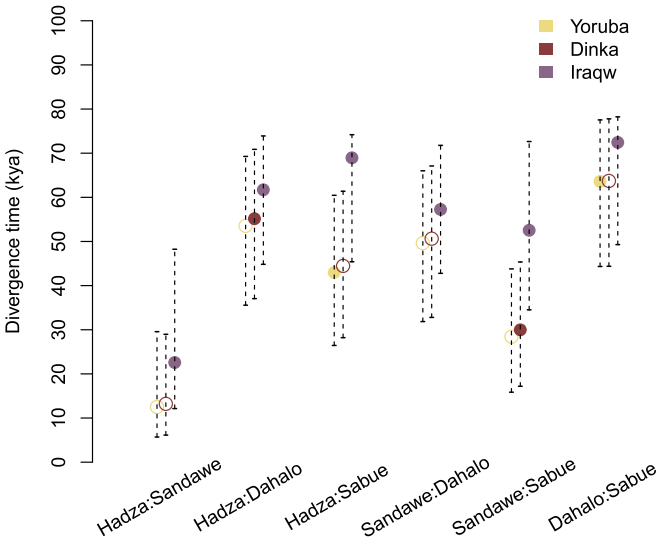

Fig. 3. Divergence time estimates. The maximum a posteriori estimates and 95% credible intervals for pairwise divergence time estimates are displayed for each set of population samples. Estimates incorporated shared gene flow with the Yoruba, Iraqw, and Dinka, representing NC, AA, and NS source populations, and are color-coded as yellow, purple, and red, respectively. The closed circles represent population combinations for which we believe the included source population likely contributed migrants to either HG population in the past.

We next employed the integrated haplotype score [iHS; a within-population statistic (54)] to identify relatively recent signatures of selective sweeps within population groupings based on extended haplotype homozygosity, and a cross-population composite likelihood ratio test [XP-CLR; a between-population statistic using the NC-west grouping as the reference population (60)] to identify older signatures of adaptation and signatures of selection from standing variation (genes near SNPs in the top 0.1% of the empirical distributions are shown in Datasets S2 and S3 for iHS and XP-CLR, respectively). When we looked at the degree to which top iHS and XP-CLR candidate genes were shared across population groupings, we found that the majority occur in only a single grouping (57% and 66%, respectively) (*SI Appendix*, Figs. S11 and S12), and this prevalence of population-specific signatures is significantly more than would expected to occur by chance (bootstrap $P < 1e\text{-}06$).

As expected, we identified the *MCM6* locus upstream of lactase (*LCT*), which contains SNPs associated with regulating lactase gene expression (61, 62) in the top 0.1% of candidate loci identified by all three tests in several pastoral population groupings: Eastern-Cushitic, Beja, Datog, Southern-Nilotic, and Fulani (as well as in populations that have experienced recent gene flow with pastoralists). This result supports previous work demonstrating the *MCM6* region to have one of the strongest signals of adaptation in East African pastoralists (62, 63), and validates the sensitivity of our chosen methods for detecting adaptation.

In addition, we identified a number of immune-related candidate loci that show shared signatures of selection in several population groupings. Seven of the 52 candidate loci identified

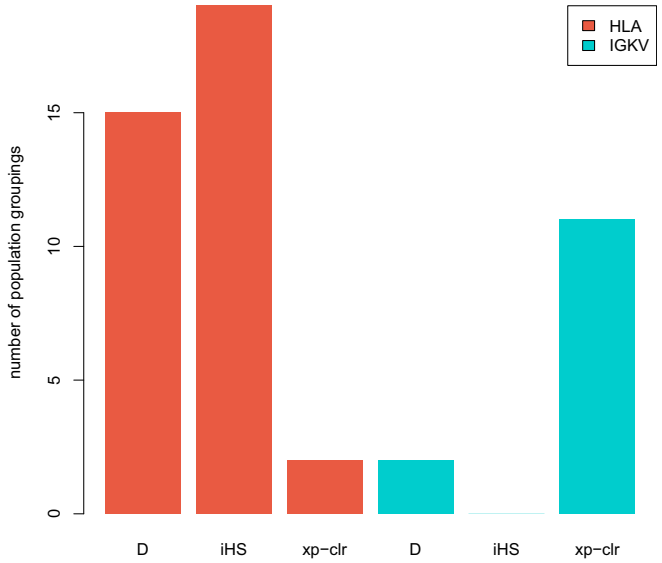

Fig. 4. Signatures of selection shared among population groupings. These shared signals among population groupings include candidate loci in the top 0.01% of the empirical distribution of each neutrality test statistic (D, iHS, XP-CLR, respectively). The HLA and IGKV gene families are displayed along the *x* axis for each neutrality test. The *y* axis displays the number of population groupings that share signatures of selection at these loci.

using iHS candidate genes that are present in many population groupings (≥10) belong to the histocompatibility complex (HLA) gene family, which is known to be critical to immune function (64), and 14 of the 32 XP-CLR candidate genes that are present in at least 10 population groupings belong to the Igκ chain variable (IGKV) gene cluster, which is known to have been subjected to positive selection in humans (65) (Fig. 4). Because many of the adaptive candidate genes across the population groupings in the study are involved in immune function, we more formally tested for enrichment of gene ontology (GO) immune system process terms (GO:0002376) (66). We found significant enrichment in all three sets of results (*Methods*): XP-CLR (P < 10e-05), iHS (P < 10e-05), and D (P < 10e-05).

Given the significant enrichment of GO immune system process genes in each set (XP-CLR, iHS, D) of adaptive candidate loci pooled across population groupings, we were interested in testing whether particular environmental variables have impacted the degree to which immune function genes are overrepresented in adaptive candidate genes among population groupings. Because our study includes populations living in diverse environments with a range of malaria endemicities and practicing a wide range of subsistence strategies, we tested whether these two variables were associated with immune function enrichment (*Methods*). We found that the degree to which adaptive candidate genes identified with iHS (which is sensitive to the most recent signals of adaptation relative to the D statistic and XP-CLR) are enriched for immune function genes is significantly correlated with both subsistence and malaria endemicity (R^2 = 0.59, P = 0.021) (*Methods*). This result is also significant for adaptive candidate genes identified with the D statistic (R^2 = 0.52, P = 0.038), but is not significant for adaptive candidate genes identified with XP-CLR (R^2 = 0.29, P = 0.42). One possible explanation for the lack of XP-CLR significance is that this test is more sensitive to older adaptive signatures (60) that may predate the emergence of malaria as a strong selective pressure in Africa.

Adaptive Signals Present Within Population Groupings. Given the extent of population-specific signals of adaptation in the data, we explored the genes near (within 100 kb) SNPs in the extreme tails (top 100 loci) of the population-specific results in more detail (Dataset S1). As noted above, many of the strongest signals of population-specific adaptation are involved in immune function (Table 1). These include genes involved in innate and adaptive immune function, which have been shown to be important in resistance to malaria and other infectious diseases (67–75). More specifically, we have identified genes involved in the production and regulation of B and T cells (76–80), genes involved in resistance to malaria and viral infections (including HIV-1) (81–85), genes involved in resistance to bacterial infection (86, 87), and genes involved in inflammatory response (88, 89). We additionally observed significant pathway enrichment of inflammation mediated by chemokine and cytokine signaling in the El Molo population grouping (*SI Appendix*, Table S3).

In addition, we observed candidate loci that may play a role in adaptation to diverse diets and climates (Datasets S1–S3). For example, the D and the XP-CLR statistics identified loci near *CISH* and *DOCK3* on chromosome 3, which are highly differentiated in WRHG and were previously identified as targets of selection and associated with stature in the same population, thought to be an adaptation to a tropical environment (38). The D and XP-CLR statistics identified a cluster of taste receptor loci on chromosome 12, and XP-CLR identified the amylase gene cluster, which plays a role in starch digestion (90), as targets of selection in the WRHG. Additionally, several of the strongest candidates for selection we identified encode proteins involved in insulin resistance (91–95), hypoglycemia (96), lactate dehydrogenase B deficiency (97), as well as lipid metabolism, transfer, and storage (98–100) (Table 1). Additionally, we observed significant enrichment of the cholesterol biosynthesis pathway in the southern-Nilotic population grouping living in Kenya, who are predominantly pastoralists (*SI Appendix*, Table S3).

Table 1. Signatures of adaptation within population groupings

Biological role and locus	Population grouping
Innate and adaptive immune function	
MYLK	Fulani
TRAF3	Amhara
IL6	Bulala
TRAF3IP2	Hadza
RAG2	Niger-Congo–east
NFX1	Eastern-Cushitic
IL2RA	Sandawe
LGALS3	Elmolo
NCAM1	Bulala
MAVS	Eastern-Cushitic
GAB2	Niger-Congo–east
ISCU	Dinka
ICAM1	Bulala
CD46	Sabue
FCGR3A	Southern-Nilotic
FCGR2B	Southern-Nilotic
IFNGR1	Eastern-Cushitic
COLEC11	Ogiek
ORM1	Sabue
TFCP2	Ogiek
Digestion and metabolism	
SLC2A10	Boni
PPARGC1A	Iraqw
IDE	Luo
PSMB9	Mada
ALMS1	Niger-Congo–west
FBP1	Amhara
LDHB	Iraqw
PNPLA2	Yaaku
LPIN2	Hadza
PLTP	Southern Nilotic

These include genes within 100 kb of candidate loci in the most extreme 100 D test statistic results.

Discussion

In this study we have characterized genomic variation in sub-Saharan populations representing a breadth of cultural and geographic diversity. The results of the study support the influence of geographic proximity, as well as cultural affiliation (e.g., language and subsistence patterns), in defining the complex relationships among populations. In particular, the Hadza, Sandawe, Dahalo, and Sabue live relatively far apart from each other in Tanzania, Kenya, and Ethiopia; however, we show that there is a closer genetic relationship among these populations than would be expected based on their geographic residences alone. They all either currently, or until very recently, have employed hunting and gathering as a primary subsistence strategy, and three of the languages spoken by these populations contain click consonants. Our results indicate that these HG populations, like the San and rain forest HG, are not impoverished agriculturalists or pastoralists who have lost their land or livestock; instead, they likely have remained relatively isolated for an extended period of time and have only come into contact with other populations in the more recent past. On the other hand, other East African populations who practice an HG lifestyle and speak AA or NS languages, appear to be genetically similar to neighboring non-HG populations. This could either be due to the loss of domestication or may reflect older ancestral subsistence patterns (20).

These relationships are consistent with a demographic history in which structure among EHG populations emerged before the Last Glacial Maximum (~21 kya). This period has been identified as one of increased aridity and reduced temperatures in East Africa; these climatic conditions were accompanied by shifts in vegetation, particularly reduced forest coverage (101, 102), and these environmental changes are thought to have triggered human dispersals into environmental refugia (103). Thus, we have uncovered a connection among geographically disparate HG populations in East Africa, consistent with a broad geographic distribution of their ancestors in the late Pleistocene before 30 kya.

Our analysis of signals of positive selection in geographically and ethnically diverse African population samples highlights the degree to which recent, regionally restricted positive selection has shaped patterns of variation in contemporary Africans. We have identified candidate loci that may be targets of natural selection; future in vitro and in vivo studies will be necessary to determine functional impact. We found that the majority of genes near SNPs showing the strongest signals of positive selection occurred in only one of the population groupings included in the analysis. This result is consistent with previous work that found a minority of overlapping signals of adaptation across continental groups (44–12%) (104). Because of the ascertainment strategy used for the Illumina 1M-Duo SNP array—common variants were prioritized, and these variants tend to be older (arose before the out of Africa migration) and may exclude population-specific SNPs—future studies based on high coverage whole-genome sequencing are likely to uncover additional loci that play a role in adaptation to diverse diets, climates, and infectious diseases across sub-Saharan Africa. Given our results, we argue that the common practice of using only one or a handful of population samples to represent an entire continent is inadequate, and this is especially true for sub-Saharan Africa, which harbors the largest proportion of human genetic variation relative to other regions across the world.

Our study includes populations living in highly diverse environments, with variable pathogen exposure, and practicing a wide range of subsistence strategies. Therefore, we were able to explore whether this diversity has had an impact on the ways in which adaptation has shaped variation in Africa. The loci that were identified as putative targets of selection are significantly enriched for genes that play a role in immune function. It is not especially surprising that loci that play a role in response to infectious disease have had such a large impact on variation among African genomes, given that infectious disease mortality is one of the strongest selective pressures identified in contemporary populations (105). Additionally, we identified candidate adaptive loci that play a role in cholesterol and glucose metabolism, taste perception, and starch digestion, many of which are specific to population groupings. Loci that may be adaptive in indigenous environments could be associated with disease in urban environments (106); therefore, it is critical to include diverse populations in studies of human adaptation, especially when the results have implications for human health and disease.

Conclusion

Human demographic history in Africa involves a complex tapestry of population movements, admixture, and adaptations to diverse environments that have shaped the genomic landscape of Africa. We have used patterns of genomic variation to investigate the demographic history of HG populations living in East Africa, demonstrating ancient common ancestry. Changes in environment and subsistence within Africa have resulted in novel and distinct selective pressures. While these biological pressures appear consistent across African populations, the specific genetic regions affected by selective pressures often vary across populations. These combined results demonstrate the importance of including ethnically diverse sub-Saharan African populations in human genetic studies to improve our understanding of complex population histories. Finally, these data demonstrate the critical importance of including African populations in biomedical studies to best encompass the full range of human diversity.

Methods

Sample Acquisition and Genotyping. Institutional Review Board approval for this project was obtained from the University of Maryland at College Park and the University of Pennsylvania. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions before sample collection: COSTECH (the Tanzania Commission for Science and Technology) and the National Institute of Medical Research in Dar es Salaam, Tanzania; the Kenya Medical Research Institute in Nairobi, Kenya; the University of Khartoum in Sudan; the Nigerian Institute for Research and Pharmacological Development, Abuja, Nigeria; the Ministry of Health and National Committee of Ethics, Cameroon; the University of Addis Ababa and the Federal Democratic Republic of Ethiopia Ministry of Science and Technology National Health Research Ethics Review Committee. A total of 816 samples were genotyped on the Illumina 1M-Duo Bead Array SNP chip. We removed individuals with <95% successfully genotyped SNPs. We also removed related individuals as inferred by PLINK ($\hat{\pi} > 0.25$). A total of 697 individuals passed these filters; this sample was then merged with data from Li et al. (23) and Henn et al. (107), resulting in 840 individuals with genotypes available at ~621,000 SNPs used for further analyses.

Principal Components Analysis. The *smartpca* program provided in EIGENSOFT 4.2 was used to calculate principal components of the sample genotype matrix of all 840 individuals at all SNPS; to account for LD, the regress option of *smartpca* was utilized.

Bayesian Clustering. For analysis with STRUCTURE, the full complement of SNPs was pruned to a smaller set of 20,000 SNPs using PLINK with the goal of minimizing LD. The model was run at K values from two through nine; each chain was run 10 times. Results from different runs were aligned using the CLUMPP software; the modal configuration for ancestry was identified visually and presented using the DISTRUCT software. Haplotypes were phased using the algorithm implemented in the BEAGLE 3.3.2 software suite (29). We inferred phase and haplotype clusters using all SNPs and then reran the k-allele STRUCTURE analysis with haplotype clusters at the same sites as with the biallelic STRUCTURE analysis.

LD Decay and N_e. LD decay was calculated by sampling pairs of SNPs within 20 kbp of each other and calculating the genotypic correlation, which approximates r^2. SNPs were placed into bins based on their distance: SNPs 0–1 kbp apart were placed into one bin; SNPs 1–2 kbp were placed into another bin, and so forth. The r^2 values of pairs of SNPs within bins were then averaged to obtain $E[r^2]$. The relationship between $E[r^2]$ and N_e derived by Tenesa et al. (46) was used to estimate N_e via nonlinear least squares.

Population Tree. The F_{ST} statistic as defined by Weir (108) was implemented in R and calculated for population samples with ≥10 individuals using the same set of 20,000 SNPs used for STRUCTURE analysis. The NJ algorithm was

used to estimate a population tree from pairwise distances between populations. The pairwise distance employed between populations i and j was defined as follows: $t_{ij} = -2N_e' \log(1 - F_{ST})$; here, N_e' is the harmonic mean of the N_e estimates for populations i and j, which helps mitigate the potential for long-branch attraction due to bottlenecks or population expansions and concomitant changes in allele frequencies. The bootstrap support of the tree was estimate by resampling SNPs as well as individuals.

Identity-By-Descent. Haplotypes were phased using the algorithm implemented in the BEAGLE 3.3.2 software suite. To infer IBD tracts between pairs of individuals, we used the GERMLINE v2.2 software. The lengths and number of IBD tracts between pairs of individuals were used to calculate a distance based on the model of Huff et al. (109). A statistic F_{IBD}, analogous to F_{ST}, was calculated between populations by averaging the IBD-based distances between pairs of individuals within populations and between populations. The NJ algorithm was used to reconstruct a population tree from the IBD-based distance matrix.

Inference of Divergence Time. We employed the ABC approach to infer the time of divergence between EHG populations; specifically, we used rejection sampling with local linear regression adjustment (110). For simulations, we utilized a realistic demographic model representative of four contemporary populations: two EHG populations, an agriculturalist or pastoralist (A/P) population, and a non-African population. The parameters of the demographic history of the A/P population and the non-African population were based on previous results (111, 112). The unknown parameters in this model included not only time of divergence, but also gene flow rates from the A/P population to the two simulated EHG populations as well as EHG N_e, the population size of the population ancestral to the EHG, and finally the population size of the simulated ancestral African population. Each of these parameters was sampled from our prior distributions. Gene flow from the population representing the A/P was introduced into the EHG populations 100–200 generations in the past, approximately the time populations in Neolithic populations in African began expanding (8, 10). For the EHG population, likely sources of gene flow from A/P population (i.e., representative NC, AA, or NS populations) were identified from STRUCTURE results. We also fixed parameters regarding the evolution of N_e in the A/P and non-African population. In addition, the A/P population diverged from other populations 4,500 generations in the past and the non-African population diverged 3,500 generations in the past. To simulate the effect of SNP ascertainment bias, only SNPs with a frequency >5% in the non-African population were retained from the simulated African populations (113–115). In addition, we accounted for the ascertainment bias introduced by choosing tag SNPs: we removed SNPs from analysis if they were in high LD ($r^2 > 0.70$) with other SNPs in the simulated non-African population. The demographic model was simulated using the coalescent framework (116); for each simulation (10,000 replicates), a total of 200 regions, 50 kbp in length, were generated. The mutation rate was fixed for each region (1.1×10^{-8} mutations per base pair per generation). The recombination rate was allowed to vary; we matched the average local recombination rates in 200 randomly selected 50-kbp regions in the deCode recombination map (117).

We constructed summary statistics using the approach of Fearnhead and Prangle (45). We proposed an initial set of summary statistics $S(X_{sim})$ based on the f_2 distances between the EHG and between each and A/P population, LD decay, and admixture LD decay (*SI Appendix*) (46, 50, 51, 118–120). We simulated these summary statistics in a pilot stage of 10,000 simulations. We estimated the functional relationship between each of the seven parameters and corresponding summary statistics: that is, $\theta_p \sim \hat{g}_p(S(X_{sim}))$ using gradient boosting machines, an ensemble method that constructs a functional approximation by iteratively combining regression trees while minimizing the squared error with respect to the target function (121–125). We then ran a second stage of simulations (10,000 replicates); summary statistics were transformed using the functional approximations obtained by gradient boosting machines in the pilot stage, $\hat{g}_p(S(X_{obs}))$. We used ABC with local linear regression adjustment to draw samples from the posterior distribution $f(\theta \mid \hat{g}_p(S(X_{obs})))$.

Selection Scan Population Groupings. When we grouped two or more population samples, we used shared ethno-linguistic affiliations among the included population samples to refer to these population groupings in the text (*SI Appendix*, Table S2). In particular, we grouped the Gabra, Gurreh, and Rendille into an eastern-Cushitic population grouping; the Baniamer and Hadandawa into a Beja population grouping; the Cameroon Fulani, Nigeria Fulani, and Mbororo Fulani into a Fulani population grouping; the Lemande, Ngumba, southern Tikar, and Yoruba into an NC-west population grouping; the Pare, Taita, and Taveta into an NC-east population grouping; the Pokot and Sengwer into a southern-Nilotic population grouping; and the Aari and Hamer into an Omotic population grouping. Because the Baka, Bakola, and Bedzan are thought to have adopted languages that belong to the NC language family, we refer to them as the WRHG grouping in place of linguistic affiliation. All other ethno-linguistic populations are referred to individually.

Genome-Wide Tests of Neutrality. We utilized three complementary statistics for identifying regions of the genome that deviate from neutral expectations: the D statistic (58), XP-CLR (60), and the iHS (54). The D statistic leverages information across all of the included population groupings so that the SNPs with the most extreme values will have allele frequencies that are distinct only in our reference population sample; this method therefore, is designed to identify regions of the genome that are highly differentiated in one population sample. XP-CLR leverages pairwise populations sample comparisons to identify regions of the genome that contain highly differentiated regions of LD. XP-CLR has also been shown to be sensitive to selection from standing variation. The iHS complements the other strategies by identifying regions that contain extended haplotype homozygosity within a given population, a classic signature of selective sweeps.

Following Akey and colleagues (58, 59), we calculated pairwise F_{ST} among all of the 23 population groupings using the method described in Weir (108). We then calculated the D statistic, and identified the top 0.1% of SNPs and genes100 kb up and downstream as our candidate regions (to account for regulatory SNPs that are typically within 100 kb of genes which they regulate; we refer to these as "candidate genes") (Dataset S1).

We employed the (iHS) test of neutrality within each of the 23 groupings as described previously (38). Briefly, we used the software package BEAGLE v3.3.2 to infer phase (29), and we generated a fine-scale recombination map relevant to the African populations with LDhat v2.1 (126). Individuals used to generate the recombination map were 100 unrelated samples, 25 males and 25 females, each from two populations in HapMap3 Release 2: the Yoruba from Ibadan, Nigeria (YRI) and the Luhya from Webuye, Kenya (LWK) (127). We estimated a genetic map in Morgan units of r from $\rho = 4Ner$ units using an N_e of 15,700, consistent with the estimation in Myers et al. (128). We used genome-wide sequence data from several nonhuman primates (chimpanzee, orangutan, and rhesus macaque) downloaded from the University of California, Santa Cruz Genome Browser website (129) to establish the ancestral allele for each of the SNPs included in our iHS analysis. Approximately 5% of the SNPs in our data could not be assigned an unambiguous ancestral state and were removed before our iHS analysis. In addition, SNPs with minor allele frequencies less than 5% in population samples were removed from the phased dataset used in the iHS analysis, in agreement with other publications (e.g., ref. 54). Finally, we removed SNPs containing missing data. The unstandardized scores returned by the iHS binary executable were adjusted such that all scores had zero means and unit variances with respect to SNPs with similar derived allele frequencies (for iHS, as described in ref. 54). We considered all of the results (for iHS we took the absolute values) in the top 0.1% of the distribution to be the top candidates (Dataset S2).

We additionally performed XP-CLR because it has been shown to be robust to ascertainment bias and because it has been shown to be sensitive to detecting selection from standing variation (60). Using the recombination map described above, we ran the XP-CLR software package (60) with 0.005-cM sliding windows and a between-window distance of 5 kb. Previous work has shown that many if not all of the included population groupings, have experienced recent gene flow resulting from the Neolithic expansion of peoples, technologies, and Bantu languages, often referred to as the Bantu expansion (8, 18, 130). Therefore, we wanted to minimize the effects of this gene flow on the XP-CLR results. Thus, in this analysis we used the NC-west grouping as our comparison population for each of the other population groupings to highlight regions of the genome that are unusually structured between the NC-west agriculturalists and the other diverse population groupings included in the study. We considered all of the results in the top 0.1% of the distribution to be the top candidates (Dataset S3).

Pathway Enrichment. We tested for significantly overrepresented Panther biological pathways (131) in our top candidate regions for each of the three genome-wide tests of neutrality. For each genome-wide scan of selection we generated a list of genes [annotated with Biomart (132)] within 100 kb of a top 0.1% SNP and tested whether this list contained more Panther pathway genes that would be expected by chance using a χ^2 test. We corrected the pathway results for multiple testing with a Bonferroni correction. We used a range of 100 kb because we were interested in retaining potential *cis*-regulatory variants in our analysis.

Bootstrap Analysis. To test the null hypothesis that the number of candidate genes that are present among population groupings could be explained just by chance, we randomly sampled 1,000 SNPs from our empirical data for 24 population groupings, including all genes 100 kb up and downstream of the SNPs, and then quantified the overlap among population groupings. We assessed the significance of this result with 1,000,000 bootstraps and found that all bootstrap runs resulted in a minority (<34%) of candidate genes occurring in a single population grouping ($P < 1e\text{-}06$).

Enrichment of GO Immune Function Genes. Given the prevalence of immune-related genes in our top adaptive candidate genes, we more formally tested whether this enrichment was statistically significant. We used the set of immune system process terms (GO:0002376) defined by the GO website (66), and tested for overrepresentation in each set of unique candidate genes identified with a given statistic (XP-CLR, iHS, D) across population groupings. We assessed statistical significance with 1,000,000 bootstrap runs, all of which resulted in lower levels of enrichment than our empirical results for candidate genes identified with each of the three statistics (XP-CLR, iHS, D) ($P < 1e\text{-}06$).

We were also interested in any variability in the degree to which sets of candidate genes were enriched for particular biological processes among population groupings. For this analysis we tested whether environmental variables (malaria endemicity and subsistence strategy) had any impact on the degree to which a given set of adaptive candidate genes identified with a given statistic (XP-CLR, iHS, D) in a given population was enriched for GO immune system genes. We used linear modeling with the χ^2 measure of enrichment as our dependent variable and malaria endemicity (estimated from information available through the Malaria Atlas Project, https://map.ox.ac.uk/) (133) and subsistence strategy as our explanatory variables ($\chi^2 \sim$ malaria_endemicity + subsistence_strategy + malaria_endemicity × subsistence_strategy). Because the residuals of the linear model were not normally distributed, we bootstrapped both malaria endemicity and subsistence strategy 1,000 times to generate statistical significance.

ACKNOWLEDGMENTS. We thank Joseph Lachance for helpful comments and discussion; and the African volunteers for samples. Genotyping services were provided by Hakon Hakonarsson of the Center for Applied Genomics at the Children's Hospital of Philadelphia. This research was funded by National Science Foundation Grants BCS-0196183 and BCS-0827436 and National Institutes of Health Grants 8DP1ES022577, 5-R01-GM076637, 1R01DK104339, and 1R01GM113657 (to S.A.T.).

1. McDougall I, Brown FH, Fleagle JG (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 433:733–736.
2. McDermott F, et al. (1996) New Late-Pleistocene uranium–thorium and ESR dates for the Singa hominid (Sudan). *J Hum Evol* 31:507–516.
3. Hublin JJ, et al. (2017) New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* 546:289–292.
4. Scheinfeldt LB, Soi S, Tishkoff SA (2010) Colloquium paper: Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proc Natl Acad Sci USA* 107:8931–8938.
5. Mcbrearty S, Brooks AS (2000) The revolution that wasn't: A new interpretation of the origin of modern human behavior. *J Hum Evol* 39:453–563.
6. Nurse D (1997) The contributions of linguistics to the study of history in Africa. *J Afr Hist* 38:355–391.
7. Philipson D (1975) The chronology of the Iron Age in Bantu Africa. *J Afr Hist* 16: 321–342.
8. de Filippo C, et al. (2011) Y-chromosomal variation in sub-Saharan Africa: Insights into the history of Niger-Congo groups. *Mol Biol Evol* 28:1255–1269.
9. Patin E, et al. (2017) Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356:543–546.
10. Bower J (1991) The pastoral neolithic of East Africa. *J World Prehist* 5:49–82.
11. Ehret C (2000) Language and history. *African Languages: An Introduction*, eds Heine B, Nurse D (Cambridge Univ Press, Cambridge, UK), pp 272–297.
12. Guldemann T, Stoneking M (2008) A historical appraisal of clicks: A linguistic and genetic population perspective. *Annu Rev Anthropol* 37:93–109.
13. Blench R (2006) *Archaeology, Language, and the African Past* (Altamira Press, Lanham, MD).
14. Ehret C (1992) Do Krongo and Shabo belong in Nilo-Saharan. *Proceedings of the Fifth Nilo-Saharan Linguistics Colloquium, Nice* (Rudiger Koppe Verlag, Cologne, Germany), pp 169–193.
15. Nurse D (1986) Reconstruction of Dahalo history through evidence from loanwords. *Sugia: Sprache und Geschichte in Afrika* (Rudiger Koppe Verlag, Cologne, Germany), Vol 7, pp 267–305.
16. Morris AG (2003) The myth of the East African 'Bushmen'. *S Afr Archaeol Bull* 58:85–90.
17. Tishkoff SA, et al. (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol* 24:2180–2195.
18. Tishkoff SA, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.
19. Edwards A, Cavalli-Sforza L (1963) Analysis of Human Evolution. *Genetics Today. Proceedings, 11th International Congress of Genetics, The Hague* (Pergamon Press, Oxford), pp 923–933.
20. Stiles D (1992) The hunter-gatherer 'revisionist'debate. *Anthropol Today* 8:13–17.
21. Ambrose SH (1982) Archaeology and linguistic reconstructions of history in East Africa. *The Archaeological and Linguistic Reconstruction of African History*, eds Ehret C, Posnansky M (Univ of California Press, Berkeley, CA), pp 104–157.
22. Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* 70:265–268.
23. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
24. Henn BM, et al. (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA* 108:5154–5162.
25. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.
26. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genet* 5:e1000686.
27. Pickrell JK, et al. (2012) The genetic prehistory of southern Africa. *Nat Commun* 3:1143.
28. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
29. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81:1084–1097.
30. Rosenberg NA, et al. (2002) Genetic structure of human populations. *Science* 298: 2381–2385.
31. Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* 2:125–141.
32. Marlowe FW (2004) Mate preferences among Hadza hunter-gatherers. *Hum Nat* 15: 365–376.
33. Bahuchet S (2006) Languages of African rainforest "Pygmy" hunter-gatherers: Lanugage shifts without cultural admixture. *Hunter-Gatherers and Linguistic History: A Global Perspective*, eds Güldemann T, McConvell P, Rhodes R (Cambridge Univ Press, Cambridge, UK).
34. Newman JL (1970) *The Ecological Basis for Subsistence Change Among the Sandawe of Tanzania* (National Academies, Washington, DC).
35. Lachance J, et al. (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150:457–469.
36. Schlebusch CM, et al. (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338:374–379.
37. Saha N, et al. (1978) A study of some genetic characteristics of the population of the Sudan. *Ann Hum Biol* 5:569–575.
38. Jarvis JP, et al. (2012) Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet* 8:e1002641.
39. Lopez M, et al. (2018) The demographic history and mutational load of African hunter-gatherers and farmers. *Nat Ecol Evol* 2:721–730.
40. Ralph P, Coop G (2013) The geography of recent genetic ancestry across Europe. *PLoS Biol* 11:e1001555.
41. Gusev A, et al. (2012) The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol* 29:473–486.
42. Ehret C, Keita SO, Newman P (2004) The origins of Afroasiatic. *Science* 306:1680; author reply 1680.
43. Browning SR, Browning BL (2012) Identity by descent between distant relatives: Detection and applications. *Annu Rev Genet* 46:617–633.
44. Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC (2005) Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072–2075.
45. Fearnhead P, Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J R Stat Soc Series B Stat Methodol* 74:419–474.
46. Tenesa A, et al. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17:520–526.
47. McEvoy BP, Powell JE, Goddard ME, Visscher PM (2011) Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21:821–829.
48. Chakraborty R, Smouse PE (1988) Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. *Proc Natl Acad Sci USA* 85:3071–3074.
49. Pfaff CL, et al. (2001) Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet* 68:198–207.
50. Moorjani P, et al. (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* 7:e1001373.
51. Loh P-R, et al. (2013) Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233–1254.
52. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* 16:980–989.
53. Nielsen R, et al. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res* 15:1566–1575.
54. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.
55. Pagani L, et al. (2012) Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am J Hum Genet* 91:83–96.

56. Granka JM, et al. (2012) Limited evidence for classic selective sweeps in African populations. *Genetics* 192:1049–1064.
57. Scheinfeldt LB, et al. (2012) Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol* 13:R1.
58. Akey JM, et al. (2010) Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci USA* 107:1160–1165.
59. Shriver MD, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics* 1:274–286.
60. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Res* 20:393–402.
61. Enattah NS, et al. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30:233–237.
62. Tishkoff SA, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31–40.
63. Ranciaro A, et al. (2014) Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet* 94:496–510.
64. Gras S, et al. (2012) A structural voyage toward an understanding of the MHC-I-restricted immune response: Lessons learned and much to be learned. *Immunol Rev* 250:61–81.
65. Sitnikova T, Nei M (1998) Evolution of immunoglobulin kappa chain variable region genes in vertebrates. *Mol Biol Evol* 15:50–60.
66. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25:25–29.
67. Finkelman FD, Vercelli D (2007) Advances in asthma, allergy mechanisms, and genetics in 2006. *J Allergy Clin Immunol* 120:544–550.
68. Dhiman N, et al. (2008) Associations between cytokine/cytokine receptor single nucleotide polymorphisms and humoral immunity to measles, mumps and rubella in a Somali population. *Tissue Antigens* 72:211–220.
69. Jones SA (2005) Directing transition from innate to acquired immunity: Defining a role for IL-6. *J Immunol* 175:3463–3468.
70. Klareskog L, Padyukov L, Rönnelid J, Alfredsson L (2006) Genes, environment and immunity in the development of rheumatoid arthritis. *Curr Opin Immunol* 18:650–655.
71. Liang HE, et al. (2002) The "dispensable" portion of RAG2 is necessary for efficient V-to-DJ rearrangement during B and T cell development. *Immunity* 17:639–651.
72. Simmons WA, et al. (1997) Novel HY peptide antigens presented by HLA-B27. *J Immunol* 159:2750–2759.
73. Tsoi LC, et al.; Collaborative Association Study of Psoriasis (CASP); Genetic Analysis of Psoriasis Consortium; Psoriasis Association Genetics Extension; Wellcome Trust Case Control Consortium 2 (2012) Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* 44:1341–1348.
74. Xu Y, Cheng G, Baltimore D (1996) Targeted disruption of TRAF3 leads to postnatal lethality and defective T-dependent immune responses. *Immunity* 5:407–415.
75. Stevenson MM, Riley EM (2004) Innate immunity to malaria. *Nat Rev Immunol* 4:169–180.
76. Ocklenburg F, et al. (2006) UBD, a downstream element of FOXP3, allows the identification of LGALS3, a new marker of human regulatory T cells. *Lab Invest* 86:724–737.
77. Bochud PY, Bochud M, Telenti A, Calandra T (2007) Innate immunogenetics: A tool for exploring new frontiers of host defence. *Lancet Infect Dis* 7:531–542.
78. Chtanova T, et al. (2004) T follicular helper cells express a distinctive transcriptional profile, reflecting their role as non-Th1/Th2 effector cells that provide help for B cells. *J Immunol* 173:68–78.
79. Hidalgo LG, Einecke G, Allanach K, Halloran PF (2008) The transcriptome of human cytotoxic T cells: Similarities and disparities among allostimulated CD4(+) CTL, CD8 (+) CTL and NK cells. *Am J Transplant* 8:627–636.
80. Nishida K, et al. (1999) Gab-family adapter proteins act downstream of cytokine and growth factor receptors and T- and B-cell antigen receptors. *Blood* 93:1809–1816.
81. Milet J, et al. (2010) Genome wide linkage study, using a 250K SNP map, of *Plasmodium falciparum* infection and mild malaria attack in a Senegalese population. *PLoS One* 5:e11616.
82. Favre N, et al. (1999) Role of ICAM-1 (CD54) in the development of murine cerebral malaria. *Microbes Infect* 1:961–968.
83. Nam DH, Ge X (2013) Development of a periplasmic FRET screening method for protease inhibitory antibodies. *Biotechnol Bioeng* 110:2856–2864.
84. Niederer HA, et al. (2010) Copy number, linkage disequilibrium and disease association in the FCGR locus. *Hum Mol Genet* 19:3282–3294.
85. Saifuddin M, et al. (1997) Human immunodeficiency virus type 1 incorporates both glycosyl phosphatidylinositol-anchored CD55 and CD59 and integral membrane CD46 at levels that protect from complement-mediated destruction. *J Gen Virol* 78:1907–1911.
86. Decker T, Müller M, Stockinger S (2005) The yin and yang of type I interferon activity in bacterial infection. *Nat Rev Immunol* 5:675–687.
87. Devuyst O, Dahan K, Pirson Y (2005) Tamm-Horsfall protein or uromodulin: New ideas about an old molecule. *Nephrol Dial Transplant* 20:1290–1294.
88. van Dijk W, et al. (1991) Inflammation-induced changes in expression and glycosylation of genetic variants of alpha 1-acid glycoprotein. Studies with human sera, primary cultures of human hepatocytes and transgenic mice. *Biochem J* 276:343–347.
89. Randall CN, et al. (2009) Cluster analysis of risk factor genetic polymorphisms in Alzheimer's disease. *Neurochem Res* 34:23–28.
90. Perry GH, et al. (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39:1256–1260.
91. Deng GY, Muir A, Maclaren NK, She JX (1995) Association of LMP2 and LMP7 genes within the major histocompatibility complex with insulin-dependent diabetes mellitus: Population and family studies. *Am J Hum Genet* 56:528–534.
92. Farris W, et al. (2003) Insulin-degrading enzyme regulates the levels of insulin, amyloid beta-protein, and the beta-amyloid precursor protein intracellular domain in vivo. *Proc Natl Acad Sci USA* 100:4162–4167.
93. Scheepers A, Joost HG, Schürmann A (2004) The glucose transporter families SGLT and GLUT: Molecular basis of normal and aberrant function. *JPEN J Parenter Enteral Nutr* 28:364–371.
94. Yoneda M, et al. (2008) Association between PPARGC1A polymorphisms and the occurrence of nonalcoholic fatty liver disease (NAFLD). *BMC Gastroenterol* 8:27.
95. Scheinfeldt LB, et al. (2009) Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. *Mol Biol Evol* 26:1357–1367.
96. Moon S, et al. (2011) Novel compound heterozygous mutations in the fructose-1,6-bisphosphatase gene cause hypoglycemia and lactic acidosis. *Metabolism* 60:107–113.
97. Maekawa M, et al. (1993) Detection and characterization of new genetic mutations in individuals heterozygous for lactate dehydrogenase-B(H) deficiency using DNA conformation polymorphism analysis and silver staining. *Hum Genet* 91:163–168.
98. Fischer J, et al. (2007) The gene encoding adipose triglyceride lipase (PNPLA2) is mutated in neutral lipid storage disease with myopathy. *Nat Genet* 39:28–30.
99. Kirschning CJ, et al. (1997) Similar organization of the lipopolysaccharide-binding protein (LBP) and phospholipid transfer protein (PLTP) genes suggests a common gene family of lipid-binding proteins. *Genomics* 46:416–425.
100. Reue K, Zhang P (2008) The lipin protein family: Dual roles in lipid biosynthesis and gene expression. *FEBS Lett* 582:90–96.
101. Felton AA, et al. (2007) Paleolimnological evidence for the onset and termination of glacial aridity from Lake Tanganyika, Tropical East Africa. *Palaeogeogr Palaeoclimatol Palaeoecol* 252:405–423.
102. Hetherington R, et al. (2008) Climate, African and Beringian subaerial continental shelves, and migration of early peoples. *Quat Int* 183:83–101.
103. Carto SL, Weaver AJ, Hetherington R, Lam Y, Wiebe EC (2009) Out of Africa and into an ice age: On the role of global climate change in the late Pleistocene migration of early modern humans out of Africa. *J Hum Evol* 56:139–151.
104. Pickrell JK, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826–837.
105. Fumagalli M, et al. (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet* 7: e1002355, and erratum (2011) 7.
106. Neel JV (1999) The "thrifty genotype" in 1998. *Nutr Rev* 57:S2–S9.
107. Henn BM, et al. (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* 8:e1002397.
108. Weir BS (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data* (Sinauer Associates, Sunderland, MA).
109. Huff CD, et al. (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res* 21:768–774.
110. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.
111. Schaffner SF, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15:1576–1583.
112. Emery LS, Felsenstein J, Akey JM (2010) Estimators of the human effective sex ratio detect sex biases on different timescales. *Am J Hum Genet* 87:848–856.
113. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 15:1496–1502.
114. Lohmueller KE, Bustamante CD, Clark AG (2009) Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182:217–231.
115. Li S, Jakobsson M (2012) Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. *BMC Genet* 13:22.
116. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
117. Kong A, et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467:1099–1103.
118. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. *Am J Hum Genet* 69:1–14.
119. Wright S (1943) Isolation by distance. *Genetics* 28:114–138.
120. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: Defining, estimating and interpreting F(ST). *Nat Rev Genet* 10:639–650.
121. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann Stat* 29:1189–1232.
122. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. *Front Neurorobot* 7:21.
123. Lin K, Li H, Schlötterer C, Futschik A (2011) Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics* 187:229–244.
124. Aeschbacher S, Beaumont MA, Futschik A (2012) A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics* 192:1027–1047.
125. Lin K, Futschik A, Li H (2013) A fast estimate for the population recombination rate based on regression. *Genetics* 194:473–484.
126. McVean GA, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.
127. Altshuler DM, et al.; International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.
128. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310:321–324.
129. Fujita PA, et al. (2011) The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* 39:D876–D882.
130. Salas A, et al. (2002) The making of the African mtDNA landscape. *Am J Hum Genet* 71:1082–1111.
131. Thomas PD, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* 13:2129–2141.
132. Kasprzyk A (2011) BioMart: Driving a paradigm change in biological data management. *Database (Oxford)* 2011:bar049.
133. Gething PW, et al. (2011) A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J* 10:378.