The genetic landscape of Ethiopia: diversity, intermixing and the association with culture

Authors: Saioa López[1,2]*, Ayele Tarekegn[3]*, Gavin Band[4], Lucy van Dorp[1,2], Tamiru Oljira[5], Ephrem Mekonnen[6], Endashaw Bekele[7], Roger Blench[8,9], Mark G. Thomas[1,2], Neil Bradman[10], Garrett Hellenthal[1,2]

Affiliations:

1) Research Department of Genetics, Evolution & Environment, University College London, Darwin Building, Gower Street, London, WC1E 6BT, United Kingdom
2) UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, United Kingdom
3) Department of Archaeology and Heritage Management, College of Social Sciences, Addis Ababa University, New Classrooms (NCR) Building, Second Floor, Office No. 214, Addis Ababa University, P. O. Box 1176, Ethiopia
4) Wellcome Centre for Human Genetics, University of Oxford, United Kingdom
5) Bioinformatics & Genomics Research Directorate (BGRD), Ethiopian Biotechnology Institute (EBTi), Ethiopia
6) Institute of Biotechnology, Addis Ababa University, Addis Ababa, Ethiopia
7) College of Natural and Computational Sciences, Addis Ababa University, Ethiopia
8) McDonald Institute for Archaeological Research, University of Cambridge, United Kingdom
9) Department of History, University of Jos, Nigeria
10) Henry Stewart Group, London WC1A 2HN, United Kingdom
(* = contributed equally)

Summary

The rich linguistic, ethnic and cultural diversity of Ethiopia provides an unprecedented opportunity to understand the level to which cultural factors correlate with -- and shape -- genetic structure in human populations. Using primarily novel genetic variation data covering 1,268 Ethiopians representing 68 different ethnic groups, together with information on individuals' birthplaces, linguistic/religious practices and 31 cultural practices, we disentangle the effects of geographic distance, elevation, and social factors upon shaping the genetic structure of Ethiopians today. We provide examples of how social behaviours have directly -- and strongly -- increased genetic differences among present-day peoples. We also show the fluidity of intermixing across linguistic and religious groups. We identify correlations between cultural and genetic patterns that likely indicate a degree of social selection involving recent intermixing among individuals that have certain practices in common. In addition to providing insights into the genetic structure and history of Ethiopia, including how they correlate with current linguistic classifications, these results identify the most important cultural and geographic proxies for genetic differentiation and provide a resource for designing sampling protocols for future genetic studies involving Ethiopians.

Introduction

Ethiopia is one of the world's most ethnically and culturally diverse countries, with over 70 different languages spoken across more than 80 distinct ethnicities (www.ethnologue.com). Its geographic position and history (briefly summarised in SI section 1) motivated geneticists to use blood groups and other classical markers to study human genetic variation (Mourant et al., 1974, Harrison, 1976). More recently, the analysis of genomic variation in the peoples of Ethiopia has been used, together with information from other sources, to test hypotheses on possible migration routes at both 'Out of Africa' and more recent 'Migration into Africa' timescales (Pagani et al., 2015, Gallego-Llorente et al., 2015). The high genetic diversity in Ethiopians facilitates the identification of novel variants, and this has led to the inclusion of Ethiopian data in studies of the genetics of elite athletes (Rankinen et al., 2016, Ash et al., 2011, Scott et al., 2005), adaptation to living at high elevation (Huerta-Sanchez et al., 2013, Stobdan et al., 2015, Simonson, 2015, Ronen et al., 2014), milk drinking (Liebert et al., 2017, Jones et al., 2013, Ingram et al., 2009) and drug metabolising enzymes (Creemer et al., 2016, Browning et al., 2010, Sim et al., 2006). While the relationships of Beta Israel with other Jewish communities have been the subject of focused research following their migration to Israel (Non et al., 2011, Behar et al., 2010, Thomas et al., 2002), studies involving genomic analyses of the history of wider sets of Ethiopian groups have been more limited (van Dorp et al., 2015, Tadesse et al., 2014, Poloni et al., 2009). Although as early as 1988 Cavalli-Sforza et al. (1988) drew attention to the importance of bringing together genetic, archaeological and linguistic data, there have been few attempts to do so in studies of Ethiopia (Boattini et al., 2013, Pagani et al., 2012, Gallego-Llorente et al., 2015). Generally, studies have been limited by analysing data from single autosomal loci, non-recombining portion of the Y chromosome and mitochondrial DNA (Semino et al., 2002, Kivisild et al., 2004, Poloni et al., 2009, Boattini et al., 2013, Messina et al., 2017) and/or relatively few ethnic

groups (Scheinfeldt et al., 2012, Pagani et al., 2012, Pagani et al., 2015, Scheinfeldt et al., 2019, Gopalan et al., 2019), which has severely limited inferences that can be drawn. Furthermore, hitherto there has been little exploration of how genetic similarity is associated with shared cultural practices (see however van Dorp et al., 2015) despite the considerable variation known to exist in cultural practices, particularly in the southern part of the country (The Council of Nationalities, Southern Nations and Peoples Region, 2017). For example, Ethiopian ethnic groups have a diverse range of religions, social structures and marriage customs, which may impact which groups intermix, and hence provide an on-going case study of socio-cultural selection (see for example Levine (2000), Freeman & Pankhurst (2003), The Council of Nationalities, Southern Nations and Peoples Region, (2017)) that can be explored using DNA.

Here we analyse autosomal genetic variation information at 591,686 single nucleotide polymorphisms (SNPs) in 1,268 Ethiopian individuals that include 1,144 previously unpublished samples and 124 samples from Lazaridis et al., 2014 and Gurdasani et al., 2015. Our study includes people from 68 distinct self-reported ethnicities (9-81 individuals per ethnic group) that comprise representatives of many of the major language groups spoken in Ethiopia, including Nilo-Saharan speakers and three branches (Cushitic, Omotic, Semitic) of Afroasiatic speakers, as well as languages of currently uncertain classification (Shabo, and the speculated, possibly extinct language of the Negede Woyto) (www.ethnologue.com) (Fig 1a, Fig S1, Extended Table 1, SI Section 2). Each of the 1,144 newly genotyped individuals were selected from a larger collection on the basis that their self-reported ethnicity, and typically birthplace, matched that of their parents, paternal grandfather, maternal grandmother, and any other grandparents recorded, analogous to recent studies of population structure in Europe (Leslie et al., 2015, Byrne et al., 2018). For these individuals we also recorded their reported religious affiliation (four categories), first language (66 total classifications) and/or

second language (40 total classifications) (Table S1). Furthermore, some of the authors of this study (A. Tarekegn, N. Bradman) translated into English and edited a compendium (originally published in Amharic) that documented the oral traditions and cultural practices of 56 ethnic groups of the Southern Nations, Nationalities and Peoples' Region (SNNPR) of Ethiopia through interviews with members of different ethnic groups (The Council of Nationalities, Southern Nations and Peoples Region, 2017). From this new resource, we compiled a list of 31 practices that were reported as cultural descriptors by members of 47 different ethnic groups out of the 68 in this study (see Methods). These practices included male and female circumcision, and 29 different pre-marital and marriage customs, including arranged marriages, polygamy, gifts of beads or belts, and covering the bride in butter.

Our study has four principal aims:

(1) assess the extent to which reported ethnic identity, and linguistic and religious affiliation are associated with genetic similarity, while accounting for the co-dependencies between them,
(2) assess the extent to which current linguistic categories are concordant with genetic distances,
(3) determine whether shared cultural or social factors are associated with intermixing among groups, and
(4) elucidate which Ethiopian groups share recent ancestry and the timescales over which various groups have become isolated or intermixed with one another.

Results

Overview of methods

We compared SNP patterns in each present-day Ethiopian to those in all other present-day Ethiopians and the 4,500 year-old Ethiopian sample "Mota" (Gallego-Llorente et al., 2015), plus an additional 252 labeled groups encompassing 2,812 non-Ethiopian individuals (average group sample size = 11,

range: 2-100), including 8 unpublished Kotoko individuals from Cameroon, 20 Moroccan Berbers, 13 individuals from Senegal and 11 Chagga individuals from Tanzania (Fig 1b, Table S2). We focus on inferring patterns of haplotype sharing among individuals, which has increased resolution over commonly-used allele-frequency based techniques (Alexander et al., 2009, Price et al., 2006) when identifying latent population structure and inferring the ancestral history of peoples sampled from relatively small geographic regions, such as within a country (Hellenthal et al., 2014, Leslie et al., 2015, van Dorp et al., 2019). In particular we applied CHROMOPAINTER (Lawson et al., 2012) to identify the sampled individuals to whom each Ethiopian shares strings of matching SNP patterns (i.e. haplotypes) in common. In general, a person shares a higher proportion of matching haplotypes along their genome with individuals they share more recent ancestry with. For a set of pre-defined groups (see below), we tabulated the total proportion of genome-wide DNA that each Ethiopian shares with individuals from that group, as inferred by CHROMOPAINTER. We then calculated the total variation distance (TVD) (Leslie et al., 2015) between two Ethiopians as the summed absolute difference between their inferred proportions across groups. We report 1-TVD, which is on a 0-1 scale, as a haplotype-based measure of genetic similarity between them (see Methods).

Mimicking van Dorp et al. (2015), we performed two CHROMOPAINTER analyses in order to infer the broad time periods over which individuals became isolated from one another (Fig 2a). The first, which we call "Ethiopia-internal," compares haplotype patterns in each Ethiopian to those in all other sampled individuals. The second, which we call "Ethiopia-external," instead compares patterns in each Ethiopian only to those among individuals in non-Ethiopian groups. A key conceptual difference between the two analyses is that the "Ethiopia-external" analysis mitigates genetic differences between individuals attributable predominantly to recent isolation, e.g. endogamy, that can cause a groups' individuals to match large proportions of their haplotype patterns to each other (van Dorp et

al., 2015; van Dorp et al., 2019). Average genetic similarity among members of different ethnicities are shown for both analyses in Fig 2bc and Fig S2.

The "Ethiopia-external" analysis can also be used to highlight admixture from outside sources that have contributed genetic ancestry to some Ethiopians but not others. To learn about the admixture history of Ethiopians under this analysis, we first used fineSTRUCTURE (Lawson et al., 2012) to assign the 1,268 Ethiopians into 78 clusters of relative genetic homogeneity (Fig 3a, Fig S3, Extended Table 2). We infer the admixture history of each cluster separately, rather than e.g. inferring the admixture history of each ethnic group, because individuals from the same ethnic label may have disparate ancestries and hence confuse inference. Nonetheless, these 78 clusters largely correspond to ethnic labels (Fig S3, Extended Table 2). Therefore, below we report results from analysing clusters as reflecting the admixture history of the particular ethnic groups that comprise the majority of individuals in that cluster, unless otherwise noted. The clustering of individuals along ethnic lines demonstrates a degree of endogamy within ethnicities, though on average Ethiopians are more genetically similar to other Ethiopians than they are to the non-Ethiopians included in this study (FigS1B-C, Fig S2, Extended Tables 3-4). We applied SOURCEFIND (Chacón-Duque et al., 2018) to each of the 78 Ethiopian clusters to infer the proportion of ancestry that the clusters' individuals share most recently with each non-Ethiopian population. We also applied GLOBETROTTER (Hellenthal et al., 2014) to each cluster to test whether its individuals' genetic patterns are consistent with them descending from an admixture event where two or more different source populations intermixed over a brief time period(s) (see Methods). In such cases, GLOBETROTTER also infers when the putative admixture occurred, and the average proportion of DNA inherited from each admixing source. Importantly, if two clusters infer admixture events with matching dates, sources and proportions, a parsimonious explanation is that the ethnic groups contained in these clusters were

a single combined population when this admixture occurred, hence placing an upper bound on their split time (van Dorp et al., 2015, van Dorp et al., 2019). Conversely, if two clusters show different admixture inferences, they were likely not a combined population when the most recent admixture occurred.

SOURCEFIND ancestry inference under the "Ethiopia-external" analysis (Fig 3) showed matching primarily to Mota and six major geographic regions to which at least one Ethiopian cluster matched >5% of their DNA: Egypt, East Africa, North Africa, Somalia, West Africa and West Eurasia (Fig 1b). Broadly, ancestry inference places the Ethiopian clusters on a northeast to southwest cline corresponding to increasing ancestry related to West African groups and decreasing ancestry related to present-day Egyptian groups, consistent with geography (Fig 3). This ancestry cline was also observed and used in the sampling strategy adopted by Browning et al. (2010) and Creemer et al. (2016) in analysing variation in Drug Metabolising Enzyme genes in Ethiopians. Contributions from Mota are seen primarily in southern Ethiopian groups, and overall range from 2.5-54.7% in all but seven groups across Ethiopia (Fig 3, Extended Table 5). When excluding Mota as a possible source, this fraction of ancestry is replaced by portions matching to samples included from all sampled geographic regions except Egypt and Somalia (Fig S4). GLOBETROTTER infers admixture events in 69 of the 78 Ethiopian clusters with dates ranging from ~100 to 4000 years ago, involving some combination of mixture between sources primarily related to East-Africans (including Mota), West-Africans, Egyptians and/or Somalians (SI section 3, Fig 3b, Extended Table 5).

Social constructs have increased genetic differences over short time periods in Ethiopia.

Comparing results under the "Ethiopia-internal" and "Ethiopia-external" analyses can reveal

important insights into ancestry sharing. For example, under the former analysis Ari and Wolayta people who work as cultivators or weavers are more genetically similar to members of other ethnicities on average than they are to people from their own ethnicities who work as potters, blacksmiths and tanners (top left of squares in Fig 2c). However, under the "Ethiopia-external" analysis, Ari and Wolayta are more genetically similar to members of their own ethnicities on average, regardless of occupation (bottom right of squares in Fig 2c). Furthermore, under the "Ethiopia-external" analysis GLOBETROTTER and SOURCEFIND infer very similar sources and dates of admixture in independent analyses of distinct clusters that correspond to the occupational groups within the Ari, with the same trend observed in analyses of clusters corresponding to Wolayta occupational groups (Fig 3b, Extended Table 5). This is consistent with the ancestors for each the Ari and Wolayta being a single population when these admixture events occurred, regardless of the present-day occupational status of their descendants. Inferred admixture dates for the Ari blacksmiths (cluster 16 in Fig 3), Ari cultivators (cluster 17) and Ari potters (cluster 24) have overlapping 95% confidence intervals spanning 49-146 generations, suggesting that ancestors of different Ari occupational workers became isolated from one another only within the past ~146 generations (<4200 years, assuming 28 years per generation), spanning the time period during which iron working is believed to have first appeared in Ethiopia (Phillipson, 2005). Analogously, for the Wolayta, GLOBETROTTER infers similar admixture events for a cluster containing cultivators and weavers (cluster 43) and a separate cluster containing blacksmiths, potters and tanners (cluster 52), with dates suggesting genetic isolation occurred within the last 19 generations (<600 years).

Anthropological studies have documented the existence of present-day societal divisions based on caste-like occupation in Ethiopia. For example, Ari and/or Wolayta individuals who work as farmers or weavers avoid intermarrying with individuals from the same ethnicities who work as blacksmiths,

tanners or potters (Freeman & Pankhurst, 2003). There has been an ongoing debate among anthropologists about whether these divisions reflect recent marginalisation of certain groups based on craft (Biasutti, 1905; Pankhurst, 1999) or whether the occupational groups descend from different, distantly related sources (Lewis, 1962; Todd, 1978; Freeman & Pankhurst, 2003). Our analyses strongly supports the former theory, corroborating a genetic study considering a subset of the Ari occupational groups considered here (van Dorp et al., 2015).

We generated an interactive map to graphically display which labelled groups are relatively more genetically similar under each of the "Ethiopia-internal" and "Ethiopia-external" analyses (https://www.well.ox.ac.uk/~gav/work_in_progress/ethiopia/v5/index.html), with averages summarised in Fig 2b and Extended Tables 3-4. This webpage can be used to test anthropological theories such as those outlined above, and can also be used as a resource to design sampling strategies for e.g. genotype-phenotype association studies involving Ethiopian participants.

Genetic similarity decays over spatial distance between both present-day and ancient genomes

Under both the "Ethiopia-internal" and "Ethiopia-external" analyses, we found significant associations (p-val < 0.05) between genetic distance and each of geographic distance, elevation difference, ethnicity and first language after controlling each factor for the others (Fig 4ab, Fig S5-S6, Table S3-S5). In contrast, we found no significant association (p-val > 0.2) between genetic distance and each of religion and second language (Fig 4ab, Fig S5, Table S3-S5). However, we found evidence (p-val < 0.05) of genetic isolation between Christians and either Muslims or people practicing traditional religions within seven of 16 groups for which we sampled at least five individuals from each religion (Table S6). Strikingly, the SOURCEFIND-inferred proportion of

DNA that each Ethiopian cluster matches to the 4,500-year-old Ethiopian Mota also showed a significant (linear regression p-value < 0.0001) decrease with increasing spatial distance between the average location of individuals in that cluster (based on birthplace) and where Mota was discovered (Fig 4cd, Table S7). This is consistent with the preservation of substantial population structure over the past 4.5k years in the region of the Gamo Highlands where Mota was found (Gallego-Llorente et al., 2015, Gopalan et al., 2019).

Genetics broadly correlates with linguistic classifications but supports pervasive recent intermixing among ethnic groups speaking diverged languages

Our study contained individuals from four different branches (second tier of classifications at www.ethnologue.com) within the Afroasiatic (AA) and Nilo-Saharan (NS) language families: the NS Satellite-Core (193 individuals), AA Cushitic (390 individuals), AA Omotic (565 individuals) and AA Semitic (95 individuals) branches. Reflecting a fluidity of genetics across these linguistic classifications, several fineSTRUCTURE-inferred clusters contain individuals from different ethnic groups that represent multiple language categories (Fig 3b). For example, the AA Cushitic-speaking Agew cluster genetically with the AA Semitic-speaking Amhara. In addition, the AA Omotic-speaking Shinasha, the AA Cushitic-speaking Qimant and the AA Semitic-speaking Beta Israel show very similar inferred ancestry and admixture dates to clusters predominantly containing AA Semitic-speakers (Fig 3b), with the Qimant and Beta Israel having been reported previously to be related linguistically to the Agew (Appleyard, 1996).

The observed blending of groups' genetics across language classifications may be attributable in part to ethnic groups adopting their current languages relatively recently and/or to pervasive recent intermixing among distinct Ethiopian groups that reside near one another. To test for the latter, we

applied GLOBETROTTER to each of the 78 Ethiopian clusters under the "Ethiopia-internal" analysis, which includes Ethiopians as surrogates for admixing sources and hence can identify intermixing that has occurred among Ethiopian groups. GLOBETROTTER inferred admixture events in 70 clusters in this analysis, 59 (84.3%) of which had estimated dates <30 generations ago (<900 years ago) (Extended Table 6). Across clusters, inferred dates under the "Ethiopia-internal" analysis typically are much more recent than those inferred under the "Ethiopia-external" analysis (Fig 5a) that does not allow Ethiopian populations as surrogates for the admixing source. This indicates that the "Ethiopia-internal" analysis is identifying recent intermixing among Ethiopian groups rather than relatively older admixture involving non-Ethiopian sources; otherwise dates under the two analyses would be similar. Supporting the possibility that geographically nearby Ethiopians are intermixing, the GLOBETROTTER "Ethiopia-internal" analysis infers intermixing among Ethiopian clusters that reside more geographically near to each other than expected by chance (p-value < 0.00002, Fig 5b).

Nonetheless, we also explored whether individuals from the same linguistic category are more genetically similar on average, which would reflect a general tendency to share more recent ancestry despite groups changing language and/or some degree of recent intermixing among groups (SI section 4). As an example, on average, individuals from the AA Cushitic, AA Omitic, AA Semitic, and NS classifications, as well as individuals from separate sub-branches within each of these categories, are genetically distinguishable from each other under both the "Ethiopia-internal" and "Ethiopia-external" analyses (p-val < 0.01; Fig S7; Extended Tables 7-8, see Methods). Indeed, individuals from five of 14 NS-speaking ethnic groups are more genetically similar to the NS-speaking Dinka from Sudan than they are to Ethiopians from AA-speaking ethnic groups, with two of these five (Anuak, Nuer) most genetically similar to the Dinka overall (Fig S2, Extended Tables 3-4). These observations suggest that the first three tiers of Ethiopian language classifications at

www.ethnologue.com are genetically -- in addition to linguistically -- separable on average, and that these genetic differences may not solely be attributable to recent isolation. Consistent with this, individuals from different language categories sometimes display substantially different inferred ancestry. For example, among clusters that predominantly consist of NS speakers, GLOBETROTTER infers admixture events typically occurring <50 generations (<1450 years) ago between EastAfrican/Somali-related sources and SSAfrican-related sources that carry some West/Central African-like ancestry (Fig 3b, Extended Table 5). Markedly different to this, all clusters that predominantly consist of AA Semitic speakers show similar inferred ancestry proportions and admixture events occurring ~57-97 generations (~1600-2800 years) ago between a source related to present-day SSAfricans/Somalians and a source related to present-day Egyptians/W.Eurasians (Fig 3b, Extended Table 5). Meanwhile AA Cushitic and AA Omotic speakers display a wide range of inferred dates and admixture proportions that fall between those inferred for NS and AA Semitic speakers (Fig 3b, Extended Table 5).

The Shabo, a hunter-gatherer group and linguistic isolate, show the strongest overall degree of genetic differentiation from other ethnic groups, consistent with the relatively high degree of isolation that has been previously suggested (Gopalan et al 2019, where the Shabo were referred to as Chabu). Under the "Ethiopia-external" analysis, both the Shabo and the AA Omotic-speaking Karo show similar genetic patterns to those of individuals in clusters of predominantly NS speakers near which they both reside (Fig 2b, Fig 3). The genetic similarity between the Shabo and NS speakers supports some conclusions based on linguistics (Blench, 2006; Ehret, 1992) and has been suggested in a study of genetic data from other Shabo individuals (Scheinfeldt et al., 2019, where the Shabo were referred to as Sabue). Our analysis finds the Shabo to be significantly most genetically similar to the NS-speaking Mezhenger relative to any other ethnic group considered (Fig S2a); for example, merging

with the Mezhenger prior to all other NS groups in the fineSTRUCTURE-inferred tree (Fig S3). The Shabo also share very similar inferred admixture events, dated to 300-900 years ago, and ancestry proportions, with the Mezhenger (Fig 3b, Extended Table 5), whom have been suggested to share origins with the Shabo (Dira and Hewlett 2017). These inferences are consistent with a high degree of intermarrying among the Shabo and Mezhenger, as has been proposed (Gopalan et al 2019; Anbessa & Unseth, 1989), and/or Shabo speakers having split from some other NS speakers more recently than 900 years ago, with some degree of current genetic differences between NS speakers and the Shabo attributable to recent isolation (e.g. due to social marginalisation of the Shabo).

The other group with uncertain linguistic classification in our study, the NegedeWoyto, are significantly differentiable (p-val < 0.001) from all other ethnic groups under the "Ethiopian-internal" analysis, and cluster separately (Fig 2b, Fig S2a, Fig S3). However, under the "Ethiopia-external" analysis they are not significantly distinguishable from multiple ethnic groups representing all three AA branches (Fig 2b, Fig S2b) and AA Semitic speakers as a whole (p-val > 0.01; Fig S7), consistent with the NegedeWoyto sharing ancestry most recently with AA speakers. Their relatively high amount of Egyptian ancestry (34%, Extended Table 5) is consistent with the group's own origin narrative of a migration from Egypt by way of the Abay river. Scholars have also proposed possible genealogical relationships with the Beta Israel and/or Agaw (Legesse, 2013); we observe a high genetic similarity between Negede Woyto and each of these groups (Fig S2b) though with some degree of recent isolation among them (Fig S2a).

Recent intermixing among different groups is associated with shared culture

For all pairwise combinations of 47 ethnic groups from the SNNPR where we had more detailed cultural information, we also calculated cultural distance as the number of 31 reported cultural

practices, primarily related to marriage traditions, that were shared between each pair, with and without weighting by the relative rarity of each practice (see Methods). In each case, we found a significant association between genetic similarity and reporting of shared cultural traits under the "Ethiopia-internal" analysis (Mantel-test p-value < 0.01; Fig 6a, Table S8), which remained after accounting for geographic or elevation distance (partial Mantel-test p-value < 0.025; Table S8) or language group (partial Mantel-test p-value < 0.025; Table S8). This association disappeared under the "Ethiopia-external" analysis (Fig 6b, Table S8), suggesting that ethnicities sharing cultural traits can be ancestrally quite different.

As an illustration, we find strong genetic similarity -- beyond that explained by spatial distance -- between the Suri, Mursi, and Zilmamo, the only three Ethiopian ethnic groups that share the practice (not included among the 31) of wearing decorative lip plates, in a manner consistent with recent shared ancestry between these three groups and recent isolation from others (Fig 6c-d, Table S9). Among the 31 cultural traits, six out of the 20 reported by more than one ethnic group exhibited significantly higher (p-value < 0.05) genetic similarity among ethnic groups participating in the practice relative to those who did not participate or whose participation in the practice was unknown (Fig 7). These six cultural traits were male circumcision, female circumcision and four different marriage practices: arrangement by parents, abduction, sororate/cousin, and belt-giving (see SI section 5 for details).

Assuming that individuals have not changed ethnic labels, these findings are consistent with some groups that share cultural practices splitting from one another relatively more recently and/or recently intermixing. To distinguish between these two possibilities, we determined whether two individuals from different groups showed evidence of sharing recent ancestors in a manner consistent with recent genetic exchange between the groups. In particular, only in the scenario where two groups have recently intermixed do we expect some pairings of individuals from the two groups to share a most

recent common ancestor (MRCA) for atypically long segments of DNA relative to those shared among all other pairings of individuals from the two groups. Therefore, to test for evidence of recent intermixing between two groups, we assess whether some pairings of individuals, one from each group, have average inferred MRCA segments that are over 2.5 centimorgans (cM) longer than the median length of average inferred MRCA segments across all such pairings of individuals from the two groups (Fig 7). We see such a trend in 134 (12.9%) of 1035 (= 46 choose 2) total pairings of groups considered in this analysis, versus in 11 (20%) of 55 pairings involving male circumcision and 7 (23.8%) of 21 pairings involving Sororate/Cousin marriages. We also see this trend in the AA Cushitic-speaking Dasanech and NS-speaking Nyangatom, who share practices of arranged and abduction marriages (Fig 7). These two groups belong to different language families and show different genetic patterns on average under the "Ethiopia-external" analysis (Fig 2b, Fig S2b, Fig 3b), which is inconsistent with them having recently split. However, they reside near each other (Fig 3a), which can be conducive to intermixing and sharing culture, and occasionally cluster together (Fig S3, Extended Table 2), suggesting some individuals from the two groups are genetically inseparable by our approach. In addition, under the "Ethiopia-internal" analysis, GLOBETROTTER infers reciprocal admixture events 15 generations ago (95% CI: 11-19gen) between the cluster containing the majority of Dasanech (cluster 14) and the cluster containing the majority of Nyangatom (cluster 10) (Fig 5b, Extended Table 6).

Discussion

Here we apply new statistical analyses to a large-scale Ethiopian cohort densely sampled across ethnicities, geography and annotated for cultural practices (SI section 2), which enable us to disentangle factors contributing to the substantial genetic structure of Ethiopians. In particular we

show a strong concordance between genetic differences and geographic distance among individuals (Fig 4), similar to that shown previously among peoples sampled from European (Novembre et al., 2008, Leslie et al., 2015) and worldwide countries (Li et al., 2008). In Ethiopia, strikingly this correlation holds true even when evaluating the degree of genetic similarity between present-day Ethiopians and an ancient farmer individual (Gallego-Llorente et al., 2015) whose body was found in the Gamo Highlands of present-day Ethiopia 4,500 years ago (Fig 4cd). This indicates a substantial preservation of population structure in some regions of Ethiopia. We also show a correlation between genetic similarity and elevation difference, even after correcting for genetic similarity over geographic distances. We caution that these analyses assume that the relationships between genetic similarity and spatial distance can be broadly explained by simple linear or exponential relationships, which the data seem to support (Fig 4), and that the association between genetic and elevation distance adheres to a linear structure, which is less clear (Fig S6c-d). Larger sample sizes may reveal deviations from these assumptions. A genuine relationship between genetics and elevation may reflect variation in peoples' adaptations to different environments. Consistent with this, we performed genome-wide scans to identify selected loci related to elevation using two different approaches (Szpiech & Hernandez, 2014; Coop et al, 2010; Günther and Coop, 2013; see Methods), for which the top hits (Fig S10) included genes associated with body temperature regulation (*TRPV1*), heat-related cell stress (*DJAJC1*) and a vision disorder that shows increased prevalence among populations living at high elevations (*ABCA4*; Klein et al., 1999).

We show that individuals from the same ethnic group or speaking the same first language are more genetically similar on average, with a clear lack of increased genetic similarity among people of the same religious affiliation or speaking the same second language. The general concordance between genetics and shared language replicates previous findings in Ethiopians (Pagani et al., 2012) while

also accounting for potential confounding due to geographic distance. However, we note that individuals from ethnicities belonging to different language groups can show strong genetic similarity, which possibly in part illustrates the fluidity of some cultural features (e.g. language) as a consequence of government policy and economic activity. We also show that ethnic groups who report practicing the same cultural activities are on average more genetically similar than expected by chance, even after accounting for genetic similarity attributable to geographic distance and language classification (Fig 6, Fig 7, Table S8). We provide evidence that these patterns may be explained in part by the recent intermixing of individuals from groups sharing the same cultural practices, e.g. between the Dasanech and Nyangatom and between the Suri and Zilmamo (Fig 7), pointing towards shared culture sometimes facilitating or accompanying genetic exchange. Though this observation could also be explained by individuals reidentifying with different ethnic labels at some point in time, our criterion of including individuals whose parents/grandparents reported the same ethnicity (wherever possible) reduces the likelihood of such an explanation in this sample. This criterion mimics a recent study of the UK (Leslie et al., 2015), and suggests that the genetic patterns we have inferred reflect genetic patterns in Ethiopia approximately two generations prior to the present-day. This plausibly underrepresents genetic similarity and intermixing among ethnic groups that would be observable in a random sample. Nonetheless, our results do support widespread recent intermixing among ethnic groups (Fig 5), while also suggesting that the participation in certain customs like male/female circumcision and some marriage customs may have acted as a barrier to intermixing with groups that do not practice these customs.

We also show how comparing patterns of genetic variation in Ethiopians to those in different worldwide populations can elucidate different layers of history, in particular by mitigating the effects of recent isolation and endogamy in our “Ethiopia-external” analysis (van Dorp et al., 2015). For

example, we have shown how particular Ari and Wolayta occupational groups exhibit high genetic distances to other occupational groups from the same ethnicity under the "Ethiopia-internal" analysis that is sensitive to endogamy, matching observations from applying other commonly-used genetic distance measures such as F_ST (Weir & Cockerham, 1984, Pagani et al., 2012). However, the genetic differences inferred between different occupational groups of the same ethnicity disappears under the "Ethiopia-external" analysis that greatly diminishes the effects of endogamy. Indeed, this analysis supports that the Ari/Wolayta individuals from the same ethnic group typically are more recently related to each other, regardless of occupation, than they are to members of other ethnic groups. Analogously, minority discriminated groups such as the Negede Woyto (Teclehaimanot, 1984; Legesse, 2013), Shabo (Dira & Hewlett, 2017), Manjo from Kefa Sheka (Freeman & Pankhurst, 2003), and Manja from Dawro (Dea, 2007) exhibit patterns of relatively low genetic similarity to other Ethiopians under the "Ethiopia-internal" analysis that disappear under the "Ethiopia-external" analysis (Fig 2b), in addition to showing higher degrees of genetic homogeneity (Fig S8a). Given this consistent pattern across discriminated groups, we argue it is most likely that discriminatory practices are directly responsible for having increased the genetic differentiation between discriminated peoples and other Ethiopians. In contrast, Ethiopia's two largest ethnicities, Amahara and Oromo, exhibit the lowest levels of homozygosity (Fig S8a), and we observe a significant (p-val<0.001) decrease of homozygosity versus increasing population census size (census from 2007 (The Council of Nationalities, Southern Nations and Peoples Region, 2017)) across ethnic groups in the SNNPR (Fig S8b). A caveat to the interpretation that groups with similar inference under the "Ethiopian-external" analysis share similar recent ancestry is that this analysis will have decreased (or have no) power to discriminate between Ethiopian groups that indeed have separate ancestral sources if we have not included relevant non-Ethiopian groups to represent these separate sources. The large number of non-Ethiopian groups included in this sample, particularly those surrounding Ethiopia,

diminishes this possibility, but more samples from other sources, in particular DNA from ancient individuals like Mota found in present-day Ethiopia, may increase precision in identifying ancient ancestral differences between Ethiopians using these techniques.

Under the "Ethiopia-external" analysis, GLOBETROTTER infers admixture in 69 out of 78 Ethiopian clusters (Fig 3b, Extended Table 5), with results indicating intermixing between a source represented by (a) Sub Saharan African groups (often including Mota) and another source represented by (b) W.Eurasian (related primarily to present-day Saudis, Yemenites and Iranians; Fig 1b), Egyptian and/or N.African groups (Fig 3b, Fig S9, Extended Table 5). Notably, Somalia differs among clusters in that it acts as a surrogate to source (a) in north/northeastern clusters with higher amounts of inferred ancestry related to Egyptian groups (type "a" clusters), while it acts as a surrogate to source (b) in west/southwest clusters with higher amounts of inferred ancestry related to East/West Sub-Saharan Africa (type "b" clusters) (Fig 3, Fig S9b-c, Table S10, Appendix B). Overall these patterns suggest intermixing between Sub-Saharan and Egyptian/West Eurasian sources starting around 100-125 generations ago (~2800-3500 years ago) as represented by type "a" clusters, and intermixing between a E.African/W.Eurasian source and West/Central African sources starting around 120-150 generations ago (~3400-4200 years ago) as represented by type "b" clusters (Fig 3). This inference is broadly consistent with previous reports of West-Eurasian admixture in Ethiopian groups (Pickrell et al., 2014; Pagani et al., 2012), but elucidates multiple waves of migrations involving different sources. However, we also highlight how more recent intermixing among different Ethiopian groups has made these older admixture signals more difficult to decipher (Fig 5). Taken at face value, the timing and sources of admixture related to Egypt/W.Eurasian-like sources (type (a) clusters) is consistent with significant contact and gene flow between the peoples of present day Ethiopia and northern Africa even before the rise of the kingdom of D'mt and interactions with the

Saba kingdom of southern Yemen which traded extensively along the Red Sea (Currey, 2014; Phillipson, 2012) It is also consistent with evidence of trading ties between the greater Horn and Egypt dated back only to 1500 BCE, when a well-preserved wall relief from Queen Hateshepsut's Deir el-Bahari temple shows ancient Egyptian seafarers heading back home from an expedition to what was known as the Land of Punt (SI Section 1A). On the other hand, the increased amount of ancestry related to West and Central African groups (type (b) clusters) suggest independent recent DNA contributions (<4200 years ago) from individuals carrying ancestry related to the migrations of Bantu speaking peoples from West and Central Africa into East and South Africa, which started around 5,000 KYA but are not believed to have reached present-day Ethiopia during the initial migrations (Vansina, 1995; Ashley, 2010; Clist, 1987). Clusters predominantly consisting of Nilo-Saharan speakers are either in type (b) clusters or unclear, with each showing a relatively high degree of ancestry related to West/Central Africans and an inferred admixture date more recent than 1600 years ago, perhaps reflecting these groups recently intermixing with other non-Nilo-Saharan speakers in Ethiopia.

Overall our findings illustrate how genetic data provide a rich additional source of information that can either corroborate or conflict with records from other disciplines (linguistics, geography, archaeology, social anthropology, sociology and history) while adding further details or novel insights and directions for future investigation. Our results also indicate the importance of dense sampling based on topographical and cultural factors, in particular language, ethnicity and in some cases occupation, when studying Ethiopian genetics. Our interactive map, which illustrates the genetic distances between ethnic and linguistic groups, can provide a resource for designing sampling strategies for such future studies. Similar strategies may also be necessary to capture the genetic structure of peoples in some other African countries that also exhibit relatively high levels of genetic diversity (Tishkoff et al., 2009, Busby et al., 2016). Finally, our analyses illustrate how cultural

practices can act as both a barrier and facilitator of gene flow among groups and consequently act as an important factor in human diversity and evolution.

Figures

Figure 1. Sample collection (a) Locations of sampled Ethiopians based on birthplace (in some cases slightly moved due to overlap), with landscape colors showing elevation and colored symbols depicting the language category (plus unclassified languages Negede Woyto, Shabo) of each individual's ethnic identity (the legend for the symbols is provided in Fig S1d). **(b)** Non-Ethiopian populations, plus the 4.5kya Ethiopian Mota (Gallego-Llorente et al., 2015), that DNA patterns of Ethiopians in (a) were compared to under the "Ethiopian-external" analysis. Open circle populations contributed to describing <1% of the ancestry of any Ethiopian cluster using SOURCEFIND. Filled circles are darkened according to their maximal SOURCEFIND-inferred contribution to any Ethiopian cluster relative to that of all other populations from that major geographic region; the legend gives this maximum contribution of a single population from each major geographic region.

Figure 2. Genetic similarity between ethnic groups, before and after mitigating recent isolation effects. (a) Schematic of two CHROMOPAINTER analyses. In the "Ethiopa-internal" analysis, haplotype patterns in Ethiopians *A* and *B* ("Eth.A","Eth.B") are matched to those of non-Ethiopians and other Ethiopians, giving the paintings below the black line, which are then compared to give the genetic similarity (1-TVD) between *A* and *B* (red square). In the "Ethiopia-external" analysis, *A* and

B are matched only to non-Ethiopians, typically giving a higher genetic similarity between their respective paintings (blue square) by substantially mitigating effects of recent isolation. **(b)** Average pairwise genetic similarity (1-TVD) between individuals from different Ethiopian labeled groups (colored on axis by language category -- see Fig 1a), under the "Ethiopia-internal" analysis (top left, red color scale) versus the "Ethiopia-external" analysis (bottom right, blue color scale). Dots denote pairs of groups for which we cannot reject the null hypothesis that they are genetically homogeneous at a per-test threshold of $p<0.01$ (green) or $p<0.001$ (black). Black boxes highlight Ari and Wolayta (Wol) occupational groups, with **(c)** their TVD values given in the boxes at right, illustrating how the "Ethiopia-external" analysis shows increased similarity between groups of the same ethnicity relative to that seen under the "Ethiopia-internal" analysis (C = cultivator, P = potter, S = blacksmith, T = tanner, W = weaver).

Figure 3. Inferred ancestral composition and recent admixture events in each Ethiopian cluster. **(a)** FINESTRUCTURE-inferred genetically homogeneous clusters of Ethiopians, with location placed on the map by averaging the latitude/longitude of each cluster's individuals, colored by the language category most represented across the ethnicities contained in that cluster (see Fig 1a for key) and numbered according to Fig 3b. The cyan diamond provides the sampling site of Mota Cave. **(b)** Bottom: SOURCEFIND-inferred ancestry proportions for each Ethiopian cluster under the "Ethiopia-external" analysis. The number of individuals per cluster are given in parentheses. Clusters are ordered according to increasing "Egypt" minus "W_African" proportions. Middle: Barplots give the proportion of individuals from ethnic groups assigned to each language category within each cluster. Top: GLOBETROTTER inferred dates (generations from present; dot=mean, line=95% CI) when testing for admixture within each Ethiopian cluster, with the major geographic region of the single

best-fitting group representing each admixing source in the primary event depicted by the colored squares at bottom (minor contributing source on top). The labels at top give the type of inferred event ("1" = a single date of admixture between two sources, "M" = a single date of admixture between more than two sources, "2" = multiple dates of admixture, "U" = unclear admixture), with colors/font denoting whether Somalia is inferred to represent the same admixing source as other Sub-Saharan Africans (green italics) versus Egypt/West-Eurasians (red).

Figure 4. Genetic similarity decays with spatial (geographic, elevation) distance and correlates with shared ethnicity and language among Ethiopians. (a) Fitted model for genetic similarity (1-TVD; under the "Ethiopia-internal" analysis) between pairs of individuals versus geographic distance, with points depicting the average genetic similarity within 25km bins, for all individuals (black; dots) or restricting to individuals who share group label (green; diamonds), speak the same first language (orange; open squares), speak the same second language (blue; triangles), have the same religious affiliation (purple; asterisks), or whose ethnicities are from the same language group (red; closed squares). Labels at right give permutation-based p-values when testing the null hypothesis of no increase in genetic similarity among individuals sharing the given trait (see Methods). **(b)** Analogous fitted model for genetic similarity versus elevation distance, with points depicting averages within 100km bins. **(c)-(d)** SOURCEFIND-inferred proportions of DNA matching to Mota versus **(c)** geographic distance or **(d)** elevation difference from Mota for each Ethiopian cluster, with black squares the average proportions within 50 km bins and p-value from linear regression at top right. Numbers/colors correspond to Fig 3b; dots represent clusters that did not infer admixture events under GLOBETROTTER. Coordinates for Mota are approximated to within 5km of Mota Cave (Gallego-Llorente et al., 2015).

Figure 5. Evidence of recent intermixing among nearby Ethiopian groups. (a) GLOBETROTTER-inferred dates (generations from present) for each Ethiopian cluster inferred to have a single date of admixture under each of the "Ethiopia-external" and "Ethiopia-internal" analyses. Inferred dates typically are more recent under the latter, indicating this analysis is picking up relatively more recent intermixing among sources represented by present-day Ethiopian clusters. **(b)** GLOBETROTTER-inferred ancestry sources under the "Ethiopia-internal" analysis. Each cluster/region *X* has a corresponding color (outer circle); lines of this color emerging from *X* indicate that *X* was inferred as the best surrogate for an admixing source contributing ancestry to each other cluster it connects with. The thickness of lines is proportional to the contributing proportion. Ethiopian clusters, with labels colored by language category as above, are ordered so that geographically close groups are next to each other (i.e. ordered by the first component of a principal components analysis applied to the distance matrix between groups). The "geographic proximity score" gives the average ordinal distance between an admixture target and the surrogate that best represents the source contributing a minority of the admixture, with the p-value testing the null hypothesis that admixture occurs randomly between groups (i.e. independent of the geographic distance between them) based on permuting cluster labels around the circle.

Figure 6 Genetic distance versus cultural similarity. (a)-(b) Fitted linear models (dotted blue lines) of genetic versus cultural similarity between pairwise comparisons of 46 Ethiopian groups (black dots) for the "Ethiopia-internal" and "Ethiopia-external" analyses, respectively, with Mantel-test p-values given at top right. Red dots give average genetic similarity at each cultural similarity score. **(c)-(d)** Average genetic similarity between the Mursi/Suri/Zilmamo (enclosed with black circles), the only ethnicities in our dataset observed to wear decorative lip plates, and individuals from each labeled group (other circles, colored by the group's language classification as in Fig 1a) under the

"Ethiopia-internal" and "Ethiopia-external" analyses, respectively. Circles are placed at the average of each group's individuals' locations, with Suri and Zilmamo slightly shifted (as they have the same such average). Note these three ethnic groups show a relatively high genetic similarity to each other under the "Ethiopia-internal" analysis (Table S9ab), while their similarity is not higher than that between these three and other Nilo-Saharan speakers under the "Ethiopia-external" analysis (Table S9cd). These observations are consistent with these three groups' relatively high genetic similarity to one another being primarily attributable to recently separating from one another and/or recently intermixing with each other.

Figure 7. Sharing of cultural practices is associated with genetic similarity. Boxplots depict the pairwise genetic similarity (under the "Ethiopia-internal" analysis) among ethnic groups that reported practicing ("Y"), not practicing ("N") or gave no information about practicing ("U") each of six different cultural traits (labeled above heatmaps, with text colors matching the boxplots' colors). Stars above the boxplots denote whether there is a significant increase (p-val < 0.05) in genetic similarity among groups in (black) "Y" versus "U" or (green) "Y" versus "N". The bottom right of each heatmap shows the increase (red) or decrease (blue) in average genetic similarity relative to that expected based on the ethnicities' language classifications (key in bottom right heatmap), after accounting for the effects of spatial distance, between every pairing of ethnicities who reported practicing ("Y") the given trait (axis labels colored by language group as in Fig 1a; group labels given in Table S1). Green squares in the top left portion of the heatmaps indicate whether >=1 pairings of individuals from different ethnic groups share atypically long DNA segments relative to all other comparisons of people from the two groups (see Methods). Such a pattern is indicative of recent intermixing between the two groups.

Materials and Methods

Samples

DNA samples from the 1,144 Ethiopians whose autosomal genetic variation data are newly reported in this study were collected in several field trips from 2000-2010, through a long-standing collaboration including researchers at University College London and Addis Ababa University, and with formal consent granted by the Ethiopian Science and Technology Commission and National Ethics Review Committee and by the UK ethics committee London Bentham REC (formally the Joint UCL/UCLH Committees on the Ethics of Human Research: Committee A and Alpha, REC reference number 99/0196, Chief Investigator MGT).

Buccal swab samples were collected from anonymous donors over 18 years of age, unrelated at the paternal level. For all individuals we recorded their, their parents', paternal grandfather's and maternal grandmother's village of birth, language, cultural ethnicity and religion. In order to mitigate the effects of admixture from recent migrations that may be causing any genetic distinctions between ethnic groups to blur, analogous to Leslie et al. (2015), where possible we genotyped those individuals whose grandparents' birthplaces and ethnicity were coincident. However, for a few ethnic groups (Bana, Meinit, Negede Woyto, Qimant, Shinasha, Suri), we did not find any individuals fulfilling this birthplace condition; in such cases we randomly selected from individuals whose grandparents had the same ethnicity. In these cases, the geographical location was calculated as the average of the grandparents' birthplaces (see SI section 2). Information about elevation was obtained using the geographic coordinates of each individual in the dataset with the "Googleway" package. All the Ethiopian individuals included in the dataset are classified into 75 groups based on self-reported ethnicity (68 ethnic groups) plus occupations (Blacksmith, Cultivator, Potter, Tanner, Weaver) within the Ari and Wolayta ethnicities.

Table S1 shows the number of samples from each population and ethnic group that passed genotyping

QC and were used in subsequent analyses. Fig 1a shows the geographic locations (i.e. birthplaces) of the Ethiopian individuals, though jittered to avoid overlap. In addition to these we also included 52 novel individuals from four non-Ethiopian groups: Cameroon (Kotoko), Morocco (Berbers), Senegal and Tanzania (Chagga). DNA samples were genotyped using The Affymetrix Human Origins SNP array, which targets 627,421 SNPs. These samples were merged with the Human Origin dataset published by Lazaridis et al. (2014) and Lazaridis et al. (2016), excluding their haploid samples (some ancient humans and primates). To these data we added Iranian and Indian individuals from Broushaki et al., 2016 and Lopez et al., 2017, Malawian samples from Skoglund et al. (2017), the African genomes published by Gurdasani et al. (2015), and the high coverage ancient sample GB20 'Mota' from Ethiopia (Gallego-Llorente et al., 2015). For curation of the dataset we removed those individuals and SNPs with a genotyping missing rate higher than 0.05 using PLINK v1.9 (Chang et al., 2015). IBD analysis was also performed to identify and remove duplicated individuals (a total of 29). The total number of samples in the merge was 4,081, analyzed at 591,686 autosomal SNPs (Figure 1, Fig S1a). We performed a principal-components-analysis (PCA) on the SNP data using PLINK v1.9 to check the consistency of the merge of our study's samples with Lazaridis and Gurdasani datasets (Fig S1b-c).

Using chromosome painting to evaluate whether genetic differences among ethnic groups are attributable to recent or ancient isolation

To quantify relatedness among individuals, we first used SHAPEIT (Deleneau et al., 2012) to jointly phase all 4,081 individuals, using default parameters and the linkage disequilibrium-based genetic map build 37 available on the 1000 Genomes Project website (http://ftp.1000genomes.ebi.ac.uk/vol1/). We then employed a "chromosome painting" technique, implemented in CHROMOPAINTER (Lawson et al., 2012), that identifies strings of matching SNP

patterns (i.e. shared haplotypes) between a phased target chromosome and a set of phased donor chromosomes. By modelling correlations among neighboring SNPs (i.e. "haplotype information"), CHROMOPAINTER has been shown to increase power to identify genetic relatedness over other commonly-used techniques such as ADMIXTURE and PCA (Lawson et al., 2012, Hellenthal et al., 2014, Leslie et al., 2015). In brief, at each position of a target individual's genome, CHROMOPAINTER infers the probability that a particular donor chromosome is the one which the target shares a most recent common ancestor (MRCA) relative to all other donors. These probabilities are then tabulated across all positions to infer the total proportion of DNA for which each target chromosome shares an MRCA with each donor. We can then sum these total proportions across donors within each of *K* pre-defined groups.

Following van Dorp et al. (2015), we used two separate CHROMOPAINTER analyses that differed in the *K* pre-defined groups used:

1. "Ethiopian-internal", which matches DNA patterns of each sampled individual to that of all other sampled people, including all other Ethiopians, from *K*=328 groups (defined in Tables S1-S2 plus Mota), and

2. "Ethiopia-external", which matches DNA patterns of each sampled individual to that of non-Ethiopians from *K*=252 groups only (shown in Fig 1b).

Relative to our genetic similarity score (1-TVD, described below) under the "Ethiopia-internal" analysis, our score under the "Ethiopia-external" analysis mitigates the effects of any recent genetic isolation (e.g. endogamy) that may differentiate a pair of Ethiopians. This is because individuals from groups subjected to such isolation typically will match relatively long segments of DNA to only a subset of Ethiopians (i.e. ones from their same group) under analysis (1). However, this isolation will

not affect how the same individuals match to each non-Ethiopian under analysis (2), for which they typically share more temporally distant ancestors. Consistent with this, in our sample the average size of DNA segments that an Ethiopian matches to another Ethiopian is 0.64cM in the "Ethiopia-internal" analysis, while the average size that an Ethiopian matches to a non-Ethiopian is only 0.22cM in the "Ethiopia-external" analysis, despite the latter analysis matching to substantially fewer donors overall and hence having a higher *a priori* expected average matching length per donor.

Following López et al., 2017, van Dorp et al., 2019, and Broushaki et al., 2016, for each analysis (1) and (2) we estimated the CHROMOPAINTER algorithm's mutation/emission (Mut, "-M") and switch rate (Ne, "-n") parameters using ten steps of the Expectation-Maximisation (E-M) algorithm in CHROMOPAINTER applied to chromosomes 1, 4, 15 and 22 separately, analysing only every ten of 4,081 individuals as targets for computational efficiency. This gave values of {192.966, 0.000801} and {339.116, 0.000986} for {Mut, Ne} in CHROMOPAINTER analyses (1) and (2), respectively, after which these values were fixed in a subsequent CHROMOPAINTER run applied to all chromosomes and target individuals. The final output of CHROMOPAINTER includes two matrices giving the inferred genome-wide total expected counts (the CHROMOPAINTER ".chunkcounts.out" output file) and expected lengths (the ".chunklengths.out" output file) of haplotype segments for which each target individual shares an MRCA with every other individual.

Inferring genetic similarity among Ethiopians under two different CHROMOPAINTER analyses

Separately for each of the "Ethiopia-internal" and "Ethiopia-external" CHROMOPAINTER analyses, for every pairing of Ethiopians *i,j* we used total variation distance (TVD) (Leslie et al., 2015) to measure the genetic differentiation (on a 0-1 scale) between their *K*-element vectors of

CHROMOPAINTER-inferred proportions, i.e:

$$TVD_{ij} = 0.5 \sum_{k=1}^{K} \quad | f^{i}{}_{k} - f_{k}{}^{j} |, \qquad (1)$$

where $f^{i}{}_{k}$ is the total proportion of genome-wide DNA that individual i is inferred to match to individuals from group k. Throughout we report $1 - TVD_{ij}$, which is a measure of genetic similarity. When calculating the genetic similarity between two groups, we average $(1 - TVD_{ij})$across all pairings of individuals *(i,j)* where the two individuals are from different groups. We note an alternative approach to measure between-group genetic similarity is to first average each $f^{i}{}_{k}$ across individuals from the same group, and then use (1) to calculate TVD between the groups by replacing each $f^{i}{}_{k}$ with its respective average value. However, we instead use our approach of averaging $(1 - TVD_{ij})$ across individuals because of the considerable reduction in computation time when performing large numbers of permutations.

We define people from two group labels *A* and *B* to be genetically differentiable if, on average, two individuals that are either both from *A* or both from *B* are significantly (p-val < 0.001) more genetically similar to one another than an individual from *A* is to an individual from *B*. To test whether individuals from group *A* are more genetically similar on average to each other than an individual from group *A* is to an individual from group *B*, we repeated the following procedure 100K times. Let n_{A} and n_{B} be the number of sampled individuals from *A* and *B*, respectively, with $n_{X} = min(n_{A}, n_{B})$. First we randomly sampled $floor(n_{X}/2)$ individuals without replacement from each of *A* and *B* and put them into a new group *C*. If $n_{X}/2$ is a fraction, we added an additional unsampled individual to *C* that was randomly chosen from *A* with probability 0.5 or otherwise randomly chosen from *B*, so that *C* had n_{X} total individuals. We then tested whether the average genetic similarity, $\sum_{i,j} \frac{1 - TVD_{ij}}{(n_{X}\ choose\ 2)}$, among all$(n_{X}\ choose\ 2)$pairings of individuals (*i,j*) from

C is greater than or equal to that among all $(n_X \; choose \; 2)$pairings of n_X randomly selected (without replacement) individuals from group Y, where $Y \in \{A,B\}$ (tested separately). We report the proportion of 100K such permutations where this is true as our one-sided p-value testing the null hypothesis that an individual from group Y has the same average genetic similarity with someone from their own group versus someone from the other group (Fig S2, Extended Tables 3-4). To test whether groups A and B are genetically distinguishable, we take the minimum such p-value between the tests of $Y=A$ and $Y=B$ (Fig 2b), which accounts for how some groups have more sampled individuals that therefore may decrease their observed average genetic similarity. Overall this permutation procedure tests whether the ancestry profiles of individuals from A and B are exchangeable, while accounting for sample size and avoiding how some permutations may by chance put an unusually large proportion of individuals from the same group into the same permuted group.

For each Ethiopian group A, in Figure S2 and Extended Tables 3-4 we also report the other sampled group $A_{max}$with highest average pairwise genetic similarity to A. To test whether A_{max} is significantly more similar to group A than sampled group B is, we permuted the group labels of individuals in $A_{max}$and B to make new groups $A^p_{max}$and B^pthat preserve the respective sample sizes. We then found the average genetic similarity between all pairings of individuals where one in the pair is from $B^p$and the other from A, and subtracted this from the average genetic similarity among all pairings of individuals where one is from $A^p_{max}$and the other is from A. Finally, we found the proportion of 100K such permutations where this difference is greater than that observed in the real data (i.e. when replacing $B^p$with B and $A^p_{max}$with A_{max}), reporting this proportion as a p-value testing the null hypothesis that individuals from group $A_{max}$and group B have the same average genetic similarity to individuals from group A. For each A, any group B where we cannot reject the null hypothesis at the 0.001 type I error level is enclosed with a white rectangle in Figure S2 and reported

in Extended Tables 3-4.

Classifying Ethiopians into genetically homogeneous clusters

We used the software fineSTRUCTURE (Lawson et al 2012) to classify 1,268 Ethiopians into clusters of relative genetic homogeneity based on the CHROMOPAINTER ".chunkcounts.out" matrix from the "Ethiopia-internal" analysis. We used default parameters, with the fineSTRUCTURE normalisation parameter "c" estimated as 0.20245. To focus on the fine-scale clustering of Ethiopians rather than clustering individuals from the rest of the world, we fixed all non-Ethiopian samples in the dataset as seven super-individual populations (Africa, America, Central Asia Siberia, East Asia, Oceania, South Asia and West Eurasia) that were not merged with the rest of the tree. We performed 2,000,000 sample iterations of Markov-Chain-Monte-Carlo (MCMC), sampling an inferred clustering every 10,000 iterations. Following Lawson et al. (2012), we next used fineSTRUCTURE to find the single MCMC sampled clustering with highest overall posterior probability. Starting from this clustering, we then performed 100,000 additional hill-climbing steps to find a nearby state with even higher posterior probability. This gave a final inferred number of 180 clusters containing Ethiopians. Results were then merged into a tree using fineSTRUCTURE's greedy algorithm. We used a visual inspection of this tree to merge clusters, starting at the bottom level of 180 clusters, that had small numbers of individuals of the same ethnicity, as shown in Fig S3. After these mergings, we ended up with a total of 78 Ethiopian clusters.

Describing the genetic make-up of Ethiopians as a mixture of recent ancestry sharing with other groups

We applied SOURCEFIND (Chacón-Duque et al., 2018) to describe genetic variation patterns among individuals in each Ethiopian cluster as a mixture of those in 253 reference groups (consisting of Mota and the 252 non-Ethiopian populations described in Table S2), using data from Lazaridis et al., 2014, Lazaridis et al., 2016, Skoglund et al., 2017, and novel samples described in Table S2 (Fig 1b). Briefly, SOURCEFIND identifies the reference groups for which each Ethiopian cluster shares most recent ancestry, and at what relative proportions, while accounting for potential biases in the CHROMOPAINTER analysis attributable to sample size differences among the reference groups. To do so, first each reference group and Ethiopian cluster k is described as a vector of length 252, where each element i in the vector for group k contains the total amount of genome-wide DNA that individuals from k are, on average, inferred to match to all individuals in donor group i under the "Ethiopia-internal" CHROMOPAINTER analysis. (These elements are proportional to the ${f^i}_k$ described in the section "Inferring genetic similarity among Ethiopians under two different CHROMOPAINTER analyses" above.) SOURCEFIND then uses a Bayesian approach to fit the vector for each Ethiopian cluster as a mixture of those from the 253 reference groups, inferring the mixture coefficients via MCMC (Chacón-Duque et al., 2018). In particular SOURCEFIND puts a truncated Poisson prior on the number of non-Ethiopian groups contributing ancestry to that Ethiopian cluster. We fixed the mean of this truncated Poisson to 4 while allowing 8 total groups to contribute at each MCMC iteration, otherwise using default parameters. For each Ethiopian cluster, we discarded the first 50K MCMC iterations as "burn-in", then sampled mixture coefficients every 5000 iterations, averaging these mixture coefficients values across 31 posterior samples. In Extended Table 5 and Fig 3b, we report the average mixture coefficients as our inferred proportions of ancestry by which each Ethiopian cluster relates to the 253 reference groups grouped by major geographic region. For comparison, we also repeated this analysis while excluding the genome of the 4,500-year-old Ethiopian Mota (Fig S4).

Identifying and dating admixture events in Ethiopia

Under each of the "Ethiopia-internal" and "Ethiopia-external" analyses, we applied GLOBETROTTER (Hellenthal et al., 2014) to each Ethiopian cluster to assess whether its ancestry could be described as a mixture of genetically differentiated sources who intermixed (i.e. admixed) over one or more narrow time periods. GLOBETROTTER assumes a "pulse" model whereby admixture occurs instantaneously for each admixture event, followed by the random mating of individuals within the admixed population from the time of admixture until present-day.

When testing for admixture in each Ethiopian cluster under the "Ethiopia-internal" analysis, we used all other Ethiopian clusters (excluding 11 clusters – those with the suffix 'adm' in Extended Table 2 – that each consisted of small numbers of individuals from multiple ethnic groups) and all 252 non-Ethiopian populations shown in Table S2 as potential surrogates to describe the genetic make-up of the admixing sources. In contrast, under the "Ethiopia-external" analysis, we included only the 252 non-Ethiopian populations, plus the 4.5kya Ethiopian Mota, as potential surrogates. GLOBETROTTER requires two paintings of individuals in the target population being tested for admixture: (1) one that is primarily used to identify the genetic make-up of the admixing source groups (used as "input.file.copyvectors" in GLOBETROTTER), and (2) one that is primarily used to date the admixture event (used as the "painting_samples_filelist_infile" in GLOBETROTTER). For both the "Ethiopia-external" and "Ethiopia-internal" analyses, we used the respective paintings described in "Using chromosome painting to evaluate whether genetic differences among ethnic groups are attributable to recent or ancient isolation" above to define the genetic make-up of each group for painting (1). For (2), following Hellenthal et al. (2014), we painted each individual in the target cluster against all other individuals except those from the target cluster. For the "Ethiopia-

external" analysis, by design the painting in (2) is the same as the one used in (1). For the "Ethiopia-internal" analysis, we had to repaint each individual in the target group for step (2); to do so we used the previously estimated CHROMOPAINTER {Mut, Ne} parameters of 192.966 and 0.000801. In each analysis, for painting (2) we used ten painting samples inferred by CHROMOPAINTER per haploid of each target individual.

Using these inputs, for each analysis we ran GLOBETROTTER for five mixing iterations (with each iteration alternating between inferring mixture proportions versus inferring dates) and performed 100 bootstrap re-samples of individuals to generate confidence intervals around inferred dates. We report results for null.ind = 1, which attempts to disregard any signals of linkage disequilibrium decay in the target population that is not attributable to genuine admixture when making inference (Hellenthal et al 2014). All GLOBETROTTER results, including the inferred sources, proportions and dates of admixture, are provided in Extended Tables 5-6 and summarized in Fig 3b and Fig 5 -- see SI Section S3 for more details. To convert inferred dates in generations to years in the main text, we used years ~= 1950 - 28 x (generations + 1), which assumes a generation time of 28 years and an average birthdate of 1950 for sampled individuals.

Testing for associations between genetic similarity and spatial distance, shared group label, language and religious affiliation

To test for a significant association between genetic similarity and spatial distance, we used novel statistical tests that are analogous to the commonly-used Mantel test (Mantel, 1967) but that account for the non-linear relationships between some variables and/or adjust for correlations among more than three variables. We calculated genetic similarity (G_{ij}) between individuals i

and j as $G_{ij} = 1\text{-}TVD_{ij}$, geographic distance ($d_{ij}$) using the haversine formula applied to the individuals' location information, and elevation distance (h_{ij}) as the absolute difference in elevation between the individuals' locations. We assessed the significance of associations between G_{ij} and d_{ij} and between G_{ij} and h_{ij} using 1000 permutations of individuals' locations. When using distance bins of 25km, we noted that the mean genetic similarity across pairs of individuals showed an exponential decay versus geographic distance in the "Ethiopia-internal" analysis (Fig 4a). Therefore, we assumed

$$G_{ij} = \alpha + \beta \exp(- \lambda d_{ij}) + e_{ij} . \qquad (2)$$

To infer maximum likelihood estimates (MLEs) for (α,β,λ), we first used the "Nelder-Mead" algorithm in the optim function in R to infer the value of λ that minimizes the sum of e_{ij}^2 across all pairings of individuals i,j when α=0 and β=1, and then found the MLE for α and β under simple linear regression using this fixed value of λ. As the main observed signal of association between genetic and spatial distance is the increased G_{ij} at small values of d_{ij}, (e.g. d_{ij}=0, which is not always accurately fit via the Nelder-Mead algorithm), our reported p-values are the proportion of permutations for which the mean G_{ij} among all (i,j) with permuted d_{ij} < 25km is greater than or equal to that of the (unpermuted) real data (Table S4a).

In contrast, we noted a linear relationship between mean G_{ij} and d_{ij} in the "Ethiopia-external" analysis (Fig S5b) and between mean G_{ij} and h_{ij} when using 100km elevation bins under both analyses (Fig 4b, Fig S5cd). Therefore for these analysis we assumed:

$$G_{ij} = \gamma + \delta x_{ij} + \varepsilon_{ij} , \qquad (3)$$

where $x_{ij} = d_{ij}$ or h_{ij}. Separately for each analysis, we found the MLEs for (γ,δ) using the lm function in R. When testing for an association with elevation, we only included individual pairs (i,j) whose elevation distance was less than 2500km, which occurred in 800,540 (99.7%) of 803,278 total comparisons, to avoid undue influence from outliers. As we expect (and observe) the change in genetic similarity δ to be negative as spatial distance increases, our reported p-values are the proportion of permutations for which the MLE of δ in the 1000 permutations is less than or equal to that of the real data (Table S4b-d).

As d_{ij} and h_{ij} are correlated (r = 0.22, Fig S6cd), we also assessed whether each was still significantly associated with G_{ij} after accounting for the other. To test whether geographic distance was still associated with genetic similarity after accounting for elevation difference, we assumed:

$$d_{ij} = \eta + \theta\, h_{ij} + \kappa_{ij}\,, \tag{4}$$

and used the lm function in R to infer maximum likelihood estimates for (η,θ). Then to test for an association between genetic similarity and geographic distance after accounting for elevation, we used the above equations (2) and (3) for the "Ethiopia-internal" and "Ethiopia-external" analyses, respectively, but replacing G_{ij} with the fitted residuals $\varepsilon_{ij} = G_{ij} - \gamma - \delta h_{ij}$ from equation (3) and replacing d_{ij} with the fitted residuals $\kappa_{ij} = d_{ij} - \eta - \theta\, h_{ij}$ from equation (4). We then repeated the procedure described above to calculate permutation-based p-values, first shifting κ_{ij} to have a minimum of 0 to do so for the "Ethiopia-internal" analysis (Table S4a,c). Similarly, to test for an association between genetic similarity and elevation difference after accounting for geographic distance, we replaced x_{ij} in equation (3) with the fitted residuals from an analogous model to (4) that instead regresses elevation

on geographic distance, and replaced G_{ij} in equation (3) with the fitted residuals $e_{ij} = G_{ij} - \alpha - \beta \exp(-\lambda d_{ij})$ from equation (2) or $\varepsilon_{ij} = G_{ij} - \gamma - \delta d_{ij}$ from equation (3) for the "Ethiopia-internal" and "Ethiopia-external" analyses, respectively. We used the same permutation procedure described above to generate p-values (Table S4b,d).

We then tested whether sharing the same (A) self-reported group label, (B) language category of reported ethnicity, (C) self-reported first language, (D) self-reported second language, or (E) self-reported religious affiliation were significantly associated with increased genetic similarity after accounting for geographic distance and/or elevation difference. For (A), we used 75 group labels for (A) (Table S1), 66 first languages for (C), and 40 second languages for (D). For (B), we used the four labels in the second tier of linguistic classifications at www.ethnologue.com for which we have data (i.e. Afroasiatic Nilotic, Afroasiatic Semitic, Afroasiatic Cushitic, Nilo-Saharan Core-Satellite), excluding the Negede-Woyto and Shabo as they have not been classified into any language family. For (E), we compared genetic similarity across three religious affiliations (Christian, Jewish, Muslim), excluding religious affiliations recorded as "Traditional" as practices within these affiliations may vary substantially across groups.

To test whether each of these factors are associated with genetic similarity, we repeated the above analyses that use equations (2)-(4) while restricting to individuals (including permuted individuals) that share the same variable Y, separately for Y={A,B,C,D,E}. Our reported p-values give the proportion of permutations for which genetic similarity among permuted individuals sharing the same Y is greater than or equal to that of the real (un-permuted) data. For the "Ethiopia-internal" analysis when testing genetic similarity against geographic distance, this was defined as having a larger estimated α (i.e. the fitted value of G_{ij} at $d_{ij}=\infty$) from equation (2). When testing genetic

similarity against geographic distance under the "Ethiopia-external" analysis, or testing genetic similarity against elevation difference under either analysis, this was instead defined as having any fitted value of G_{ij}, at 48 equally-spaced bins of $d_{ij} \in$ {12.5,1187.5km} or 25 equally-spaced bins of $h_{ij} \in$ {50,2450m}, greater than or equal to that of the observed data. These analyses were repeated with or without first adjusting geographic and elevation distance for each other as described above.

As group label, language and religion can also be correlated with spatial distance and with each other (e.g. see Fig S6ab), we performed additional permutation tests where we fixed each of (A)-(E) when carrying out the permutations described above. For example, when fixing (A), we only permuted birthplaces and each of (B)-(E) across individuals within each group label, hence preserving the effect of group label on G_{ij}. Applying this permutation procedure for each of (A)-(E), we repeated all tests described above, reporting p-values in Table S4.

For each of geographic distance, elevation difference, and (A)-(E), our final p-values reported in the main text and Fig 4 and Fig S5 that test for an association with genetic similarity are the maximum p-values across the six permutation tests that permute all individuals freely or fix each of (A)-(E) while permuting (i.e. the maximum values across rows of Table S4), with the following two exceptions. First, relative to the distances between birthplaces among all individuals, Ethiopians who share the same group label or who share the same first language live near each other (Table S5), so that permuting birthplaces while fixing group label or first language do not permute across large spatial distances. Therefore, we ignore those permutations when reporting our final p-values for geographic distance and elevation difference (i.e. in the main text and Fig 4, Fig S5). Secondly, the high correlation between group label, first language, and first language group (e.g. Fig S6ab) makes

accounting for one challenging (in terms of loss of power) when testing the others. Therefore, under the "Ethiopia-external" analysis, which we expect to have more difficulty distinguishing between Ethiopian individuals' ancestry relative to the "Ethiopia-internal" analysis due to excluding Ethiopian reference populations, we excluded permutations fixing group and fixing first language when testing all variables when reporting our final p-values in the main text and Fig S5. In general disentangling whether there is significant evidence of independent effects for each of language, group label and spatial distance while also mitigating effects of recent isolation, though suggested in Table S4c-d, will require larger sample sizes than those considered here.

Permutation test to assess significance of genetic similarity among individuals from different linguistic groups

To test whether individuals from language classification *A* are more genetically similar to each other than an individual from classification *A* is to an individual from classification *B*, we followed an analogous procedure to that detailed above to test for genetic differences between group labels *A* and *B*. Again let n_A and n_B be the number of sampled individuals from *A* and *B*, respectively, with $n_X = min(n_A, n_B)$. For each of 100K permutations, we first randomly sampled $floor(n_X/2)$ individuals without replacement from each of *A* and *B* and put them into a new group *C*. If $n_X/2$ is a fraction, we added an additional unsampled individual to *C* that was randomly chosen from *A* with probability 0.5 or otherwise randomly chosen from *B*, so that *C* had n_X total individuals. We then tested whether the average genetic similarity, $\sum_{i,j} \frac{1 - TVD_{ij}}{(n_X \, choose \, 2)}$, among all $(n_X \, choose \, 2)$ pairings of individuals (*i,j*) from *C* is greater than or equal to that among all $(n_X \, choose \, 2)$ pairings of n_X randomly selected (without replacement) individuals from group *Y*, where $Y \in \{A, B\}$ (tested separately). Individuals from the same ethnic/occupation label (i.e. those

listed in Table S1) are often substantially genetically similar to one another (Fig 2), which may in turn drive similarity among individuals within the same language classification. Therefore, whenever a language classification contained more than two different ethnic/occupation labels, we restricted our averages to only include pairings (*i,j*) that were from different ethnic/occupation labels (including in permuted group *C* individuals). We report the proportion of 100K such permutations where this is true as our one-sided p-value testing the null hypothesis that an individual from language classification *Y* has the same average genetic similarity with someone from their own language group versus someone from the other language group (Fig S7, Extended Tables 7-8). To test whether classifications *A* and *B* are genetically distinguishable, we take the minimum such p-value between the tests of *Y=A* and *Y=B* (Fig S7), which accounts for how some linguistic classifications include more sampled individuals and/or more sampled ethnic groups that therefore may decrease their observed average genetic similarity.

Genetic diversity and homogeneity

To assess within-group genetic homogeneity in the Ethiopian ethnic groups, we followed three different approaches. First, we computed the observed autosomal homozygous genotype counts for each sample using the --het command in PLINK v1.9 (Chang et al., 2015) and averaged them across groups. Second, we detected runs of homozygosity (ROH), an algorithm designed to find runs of consecutive homozygous SNPs within groups that are identical-by-descent. Prior to this we pruned SNP data based on linkage disequilibrium (--indep-pairwise 50 5 0.5), which left us with 359,281 SNPs. ROH were identified in PLINK v1.9 using default parameters.

We performed an additional analysis to infer relatedness coefficients using FastIBD (Browning and Browning, 2011), implemented in the software BEAGLE v3.3.2, to find tracts of identity by descent

(IBD) between pairs of individuals. For each population, we inferred the pairwise IBD fraction between each pairing of these individuals. For each population and chromosome, fastIBD was run for ten independent runs using an IBD threshold of 10^{-10}, as recommended by Browning and Browning (2011), for every pairwise comparison of individuals.

We assessed whether the degree of genetic diversity in Ethiopian ethnic groups was associated with census population size, by comparing different measures of genetic diversity described above (homozygosity, IBD and ROH) with the census population size using standard linear regression. As population census are not always available and can be inaccurate, we limited this analysis to ethnic groups in the SNNPR, for whom census information was recently reported in the SSNPR book (The Council of Nationalities, Southern Nations and Peoples Region, 2017).

Genetic similarity versus cultural distance

Between each pairing of 46 sampled SNNPR ethnic groups, we calculated a cultural similarity score as the number of practices, out of 31 reported in the SSNPR book (The Council of Nationalities, Southern Nations and Peoples Region, 2017), that both groups in the pair reported. Despite the SSNPR book also containing information about the Ari, we did not include them among these 46 because of the major genetic differences among caste-like occupational groups (Fig 2c). For the Wolayta, we included individuals that did not report belonging to any of the caste-like occupational groups. We included the 31 practices described in SI Section 5.

We also calculated a second cultural similarity score whereby practices shared by many groups contributed less to a pair's score than practices shared by few groups. To do so, if a practice was

reported by *H* ethnic groups in total, any pair of ethnicities that shared this practice added a contribution of *1.0/H* to that pair's cultural similarity score, rather than a contribution of 1 as in the original cultural similarity score.

Genetic similarity, geographic distance and elevation difference between two ethnic groups *A*, *B* were each calculated as the average such measure between all pairings of individuals where *i* is from *A* and *j* from *B*. We then applied a mantel test using the *mantel* package in the *vegan* library in R with 100,000 permutations to assess the significance of association between genetic and cultural similarity across all pairings of ethnic groups (Fig 6ab, Table S8). We also used separate partial mantel tests, using the mantel.partial function in R with 100,000 permutations, to test for an association between genetic and cultural similarity while accounting for one of (i) geographic distance, (ii) elevation difference, or (iii) shared language classification. To account for shared language classification, we used a binary indicator of whether *A,B* were from the same language branch: AA Cushitic, AA Omotic, AA Semitic, NS Satellite-Core.

For each of the 31 cultural practices, all 46 ethnic groups were classified as either (i) reporting participation in the practice, (ii) reporting not participating in the practice or (iii) not reporting whether they participated in the practice. For cultural practices where at least two of (i)-(iii) contained >=2 groups, we tested the null hypothesis that the average genetic similarity among groups assigned to category X was equal to that of groups assigned to Y, versus the alternative that groups in X had a higher average genetic similarity to each other. To do so, we calculated the difference in mean genetic similarity among all pairs of groups assigned to X versus that among all pairs assigned to Y. We then randomly permuted ethnic groups across the two categories 10,000 times, calculating p-values as the proportion of times where the corresponding difference between permuted groups assigned to X

versus Y was higher than that observed in the real data. For 16 of 31 cultural practices, we tested X=(i) versus Y=(iii). For one cultural practice, we tested X=(ii) versus Y=(iii). For three cultural practices, we tested {X=(i) versus Y=(ii)}, {X=(i) versus Y=(iii)}, and {X=(ii) versus Y=(iii)}. Six practices gave a p-value < 0.05 for one of the above permutation tests (Fig 7). We calculated the average genetic similarity between all ethnic groups sharing these six practices after accounting for the effects of spatial distance and language classification. To account for spatial distance, we used equations (2)-(4) above, first adjusting geographic distance out of each of genetic similarity and elevation difference, and then regressing the residuals from the genetic similarity versus geographic distance regression against the residuals from the elevation difference versus geographic distance regression. We take the residuals for individuals *i,j* from this latter regression as the adjusted genetic similarity between individuals *i* and *j* (denoted G^*_{ij}). In each of the above regressions, we fit our models using all pairs of Ethiopians that were NOT from the same language classification at the branch level (i.e. AA Cushitic, AA Omotic, AA Semitic, NS Satellite-Core), in order to account for only spatial distance effects that are not confounded with any shared linguistic classification. We calculate the average spatial-distance-adjusted genetic similarity between each ethnic group *A,B* as the average G^*_{ij} between all pairings of individuals where *i* is from *A* and *j* from *B*. Then to adjust for language classification, we calculated the expected spatial-distance-adjusted genetic similarity for each pairing of language branches *C,D* as the average adjusted genetic similarity across all pairings of ethnic groups *A, B* where *A* is from *C* and *B* is from *D*. For each pair of ethnic groups that share a reported cultural trait shown in Fig 7, we show the adjusted genetic similarity between that pair minus the expected spatial-distance-adjusted genetic similarity based on their language classification. This therefore illustrates the genetic similarity between the two groups after adjusting for that expected by their spatial distance from each other and their respective languages (lower right triangles of heatmaps in Fig 7).

For each of these six cultural practices shown in Fig 7, we also assessed whether there was evidence of recent intermixing among people from pairs of groups that both reported the given practice. To do so, we exploit the fact that if two groups have recently intermixed, it is expected that some -- but not all -- pairings of individuals from the two groups will share a MRCA for atypically long stretches of DNA. Therefore, to test for evidence of recent intermixing between two groups, we assess whether at least some pairings of individuals, one from each group, have average inferred MRCA segments (inferred by CHROMOPAINTER under the "Ethiopia-internal" analysis) that are >2.5cM longer than the median length of average inferred MRCA segments across all such pairings of individuals (upper left triangles of heatmaps in Fig 7).

Identifying genetic loci with evidence of selection related to adaptation

We performed two different selection scans to identify SNPs associated with elevation. For both of these scans, we excluded the 7 (out of 14) Beta Israel individuals taken from Lazaridis et al 2014, because their geographic coordinates resulted in elevation values (2414 meters) were drastically different from those of the Beta Israel newly genotyped in this study (940 meters) that we can more readily verify. In the first selection scan, we applied XPEHH in SELSCAN v.1.2.0 (Szpiech & Hernandez, 2014) with default parameters, dividing our samples into two "populations": one containing 119 individuals living at the 90th percentile of elevation (2302-3362 meters) versus the other containing 126 individuals living at the 10th percentile of elevation (0-563 meters). We annotated SNPs with the highest XPEHH scores (>=5) using ANNOVAR, which gives the nearest gene to each SNP. Annotation for the most significant hit produced the gene *TRPV1* (transient receptor potential cation channel subfamily V member 1, OMIM 602076), whose main function is the detection and regulation of body temperature (Fig S10a). The second most significant hit region contains the genes *DNAJC1* (DnaJ Heat Shock Protein Family (Hsp40) Member C1, OMIM 611207)

and *SPAG6* (Sperm Associated Antigen 6, OMIM 605730). Intriguingly, *DNAJC1* is a heat shock protein, involved in the protection of cells from stressful conditions like thermal stresses or UV light.

For our second selection scan, we applied Bayenv2, a Bayesian method to estimate the empirical pattern of covariance in allele frequencies between the 75 sampled Ethiopian groups (Coop et al, 2010; Günther and Coop, 2013). As recommended by the authors, we initially used PLINK to remove SNPs within a 50Kb window size and a r2 threshold of 0.001 (i.e. "indep-pairwise 50 5 0.001"), in order to mitigate the effects of linkage disequilibrium (LD). We then estimated a covariance matrix among the 75 groups using the remaining 9408 SNPs and 100,000 iterations. Fixing this matrix, we estimated Bayes factors (using 100,000 iterations) for each of 344,955 SNPs remaining after LD-based pruning using "indep-pairwise 50 5 0.5", testing each SNP for an association between allele frequencies and elevation. For this analysis, from ANNOVAR the most significant hit contains the locus *ABCA4* (ATP Binding Cassette Subfamily A Member 4, OMIM 601691) (Figure S10b). *ABCA4* is expressed in retina photoreceptor cells that play a role in photoresponse and dark adaptation, i.e. how the eye recovers its sensitivity in the dark after exposure to intense lights (Yang et al., 2015). Interestingly, among the factors that can affect the process of dark adaptation, hypoxia at high altitudes has been reported as a significant one (Yang et al., 2015). *ABCA4* has also been linked to age-related macular degeneration (AMD), whose prevalence has been shown to be higher among populations living at high altitudes, like the Tibetans, compared to populations living at lower altitudes like the Uighur or the Han (Klein et al., 1999).

Acknowledgements

This work is funded by BBSRC (Grant Number BB/L009382/1). GH is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number

098386/Z/12/Z) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. We thank David Reich and the Children's Hospital of Philadelphia for genotyping the samples on the Human Origins array.

The authors declare the following competing interests:

GH is a founding member of GenSci.

References

Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. **19**, 1655-1664 (2009).

Anbessa, T. & Unseth, P. Toward the classification of Shabo (Mikeyir). In: *Topics in Nilo-Saharan linguistics* (Helmut Buske, Hamburg, 1989).

Appleyard, D. *Preparing a Comparative Agaw Dictionary In: Cushitic & Omotic Languages* (ed. Griefenow-Mewis & Voigt, 1996)

Ash, G. I. et al. No association between ACE gene variation and endurance athlete status in Ethiopians. *Med. Sci. Sports Exerc*. **43**, 590-597 (2011).

Ashley, C. Z. Towards a socialised archaeology of ceramics in Great Lakes Africa. *Afr. Archaeol. Rev*. **27**, 135–163 (2010)

Behar, D. M. et al. The genome-wide structure of the Jewish people. *Nature* **466**, 238-242 (2010).

Biasutti, R. Pastori, agricoltori e cacciatori nell 'Africa Orientale'. *Bolletino del Reale Societa geografica italiana* **6**, 155–179 (1905).

Blench, R. *Archaeology, Language, and the African Past* (Rowman Altamira, 2006)

Boattini, A. et al. mtDNA variation in East Africa unravels the history of Afro-Asiatic groups. *Am. J. Phys. Anthropol.* **150**, 375-385 (2013).

Broushaki, F. et al. Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499-503 (2016).

Browning, S. L., Tarekegn, A., Bekele, E., Bradman, N. & Thomas, M. G. CYP1A2 is more variable than previously thought: a genomic biography of the gene behind the human drug-metabolizing enzyme. *Pharmacogenet. Genomics* **20**, 647-664 (2010).

Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703-714 (2011).

Busby, G. et al. Admixture into and within sub-Sarahan Africa. *eLife* **5**, e15266 (2016).

Byrne, R. P. et al. Insular Celtic population structure and genomic footprints of migration. *PLoS Genet.* **14**, e1007152 (2018).

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 6002-6006 (1988).

Chacón-Duque, J. C. et al. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).

Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4** (2015).

Clist, B. A critical reappraisal of the chronological framework of the early Urewe Iron Age industry. *Muntu* **6**, 35–62 (1987)

Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. Genetics **185**, 1411-1423 (2010).

Creemer, O. J. et al. Contrasting exome constancy and regulatory region variation in the gene encoding CYP3A4: an examination of the extent and potential implications. *Pharmacogenet. Genomics* **26**, 255-270 (2016).

Currey, J. *Foundations of an African Civilisation: Aksum and the northern Horn* (Oxford, 2014).

Dea, D. *Rural Livelihoods and Social Stratification Among The Dawro, Southern Ethiopia*. (Addis Ababa: Social Anthropology Dissertation Series, No 14, Series Editor Gebre Yntiso, Department of Sociology and Social Anthropology, Addis Ababa University, 2007).

Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).

Dira, S. J. & Hewlett, B. S. The Chabu hunter-gatherers of the highland forests of Southwestern Ethiopia. *Hunt. Gatherer Res.* **3**, 323-352 (2017).

Ehret, C. Do Krongo and Shabo belong in Nilo-Saharan? In: *Nilo-Saharan Linguistic Analysis and Documentation* (Rüdiger Köppe Verlag, Cologne, 1992)

Freeman, D. & Pankhurst, A. *Peripheral People: The Excluded Minorities of Ethiopia* (Hurst and Company, London, 2003).

Gallego-Llorente, M. et al. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820-822 (2015).

Gopalan, S. et al. Hunter-gatherer genomes reveal diverse demographic trajectories following the rise of farming in East Africa. BioRxiv. doi: https://doi.org/10.1101/517730 (2019).

Günther, T & Coop, G. Robust identification of local adaptation from allele frequencies. Genetics

195, 205-220 (2013).

Gurdasani, D. et al. The African genome variation project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

Harrison, G. A. Genetic and anthropological studies in the human adaptability section of the International Biological Programme. *Philos Trans. R. Soc. Lond. B Biol. Sci.* **274**, 437-445 (1976).

Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).

Huerta-Sanchez, E. et al. Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. Mol. Biol. Evol. **30**, 1877–1888 (2013).

Ingram, C. J. et al. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *J. Mol. Evol.* **69**, 579-88 (2009).

Jones, B. L. et al. Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. *Am. J. Hum. Genet.* **93**, 538-544 (2013).

Kivisild, T. et al. Ethiopian mitochondrial DNA heritage: Tracking gene flow across and around the gate of tears. *Am. J. Hum. Genet.* **75**, 752–770 (2004).

Klein, R., et al. Prevalence of age-related macular degeneration in 4 racial/ethnic groups in the multi-ethnic study of atherosclerosis. *Ophthalmology* **113**, 373-380 (1999).

Lawson, D., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**: e1002453 (2012).

Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature* **513**, 409–413 (2014).

Lazaridis, I. et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419-424 (2016).

Legesse, W. *Religion and Everyday Life: Negede-Woyto Muslim Community of Bahir Dar Town Ethiopia* (MA Thesis, Addis Ababa University, 2013).

Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309-314 (2015).

Levine, D. N. *Greater Ethiopia: The Evolution of A Multi-ethnic Society, Second Edition* (Chicago and London: The University of Chicago Press, 2000).

Lewis, H. Historical Problems in Ethiopia and the Horn of Africa. *Ann. New York Acad. Sci.* **96**, 504–511 (1962).

Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008).

Liebert, A. et al. World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Hum Genet.* **136**, 1445-1453 (2017).

López, S. et al. The Genetic Legacy of Zoroastrianism in Iran and India: Insights into Population Structure, Gene Flow, and Selection. *Am. J. Hum. Genet.* **101**, 353-368 (2017).

Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27,** 209–220 (1967).

Messina, F. et al. Linking between genetic structure and geographical distance: Study of the maternal gene pool in the Ethiopian population. *Ann. Hum. Biol.* **44**, 53-69 (2017).

Mourant, A. E. et al. The blood groups and haemoglobins of the Kunama and Baria of Eritrea, Ethiopia. *Ann. Hum. Biol.* **1**, 383-392 (1974).

Non, A. L., Al-Meeri, A., Raaum, R. L., Sanchez, L. F. & Mulligan, C. J. Mitochondrial DNA reveals distinct evolutionary histories for Jewish populations in Yemen and Ethiopia. *Am. J. Phys. Anthropol.* **144**, 1-10 (2011).

1 Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).

2 Pagani, L. et al. Tracing the route of modern humans out of Africa by using 225 human genome
3 sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* 96, 986-991 (2015).

4 Pagani, L. et al. Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex
5 Influences on the Ethiopian Gene Pool. *Am. J. Hum. Genet.* **91**, 83–96 (2012).

6 Pankhurst, A. 'Caste' in Africa: The Evidence from South-Western Ethiopia Reconsidered. *Africa*
7 **69**, 485–509 (1999).

8 Phillipson, D. W. *African Archaeology* (Cambridge University Press, 2005)

9 Phillipson, D. W. Foundations of and African Civilisation: Aksum & The Northern Horn 1000 BC-
10 AD 1300 (Suffolk & Rochester: James Currey, 2012)

11 Pickrell, J. K. et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad.*
12 *Sci. U. S. A.* **111**, 2632–2637 (2014).

13 Poloni, E. S. et al. Genetic evidence for complexity in ethnic differentiation and history in East Africa.
14 *Ann. Hum. Genet.* **73**, 582-600 (2009)

15 Price A. L. et al. Principal components analysis corrects for stratification in genome-wide association
16 studies. *Nat. Genet.* **38**, 904–909 (2006).

17 Rankinen, T. et al. No Evidence of a Common DNA Variant Profile Specific to World Class
18 Endurance Athletes. *PLoS One* **11**, e0147330 (2016).

19 Ronen, R., Zhou, D., Bafna, V. & Haddad, G. G. The genetic basis of chronic mountain sickness.
20 *Physiology (Bethesda)* **29**, 403-412 (2014).

21 Scheinfeldt, L. B. et al. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.*
22 **13**: R1 (2012).

Scheinfeldt, L. B. et al. Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. *Proc. Natl. Acad. Sci. U. S. A.* pii: 201817678 (2019).

Scott, R. A. et al. Mitochondrial DNA lineages of elite Ethiopian athletes. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **140**, 497-503 (2005).

Semino, O., Santachiara-Benerecetti, A. S., Falaschi, F., Cavalli-Sforza, L. L. & Underhill, P. A. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.* **70**, 265–268 (2002).

Sim, S. C. et al. A common novel CYP2C19 gene variant causes ultrarapid drug metabolism relevant for the drug response to proton pump inhibitors and antidepressants. *Clin. Pharmacol. Ther.* **79**, 103-113 (2006).

Simonson, T.S. Altitude Adaptation: A Glimpse Through Various Lenses. *High Alt. Med. Biol.* **16**, 125-137 (2015).

Skoglund, P. et al. Reconstructing Prehistoric African Population Structure. *Cell* **171**, 59-71 (2017).

Stobdan, T. et al. Endothelin receptor B, a candidate gene from human studies at high altitude, improves cardiac tolerance to hypoxia in genetically engineered heterozygote mice. *Proc. Natl. Acad. Sci. U. S. A.* **112**: 10425-10430 (2015).

Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multi-threaded program to calculate EHH-based scans for positive selection. Mol. Biol. Evol. **31**, 2824-2827 (2014).

Tadesse, L., Tafesse, F. & Hamamy, H. Communities and community genetics in Ethiopia. *Pan. Afr. Med. J.* **18**: 115 (2014).

Teclehaimanot, G. S. *The Wayto of Lake Tana: An Ethno-History* (MA Thesis in History, School of

Graduate Studies, Addis Ababa University, 1984).

The Council of Nationalities (CoN), Southern Nations and Peoples Region (SNNPR). *A Profile of the Nations, Nationalities and Peoples of the Southern Region* (Awassa, Ethiopia, 2017).

Thomas, M. G. et al. Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am. J. Hum. Genet.* **70**, 1411-1420 (2002).

Tishkoff, S. A. et al. The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044 (2009).

Todd, D. The Origins of Outcastes in Ethiopia: Reflections on an Evolutionary Theory. *Abbay* **9**, 145–158 (1978).

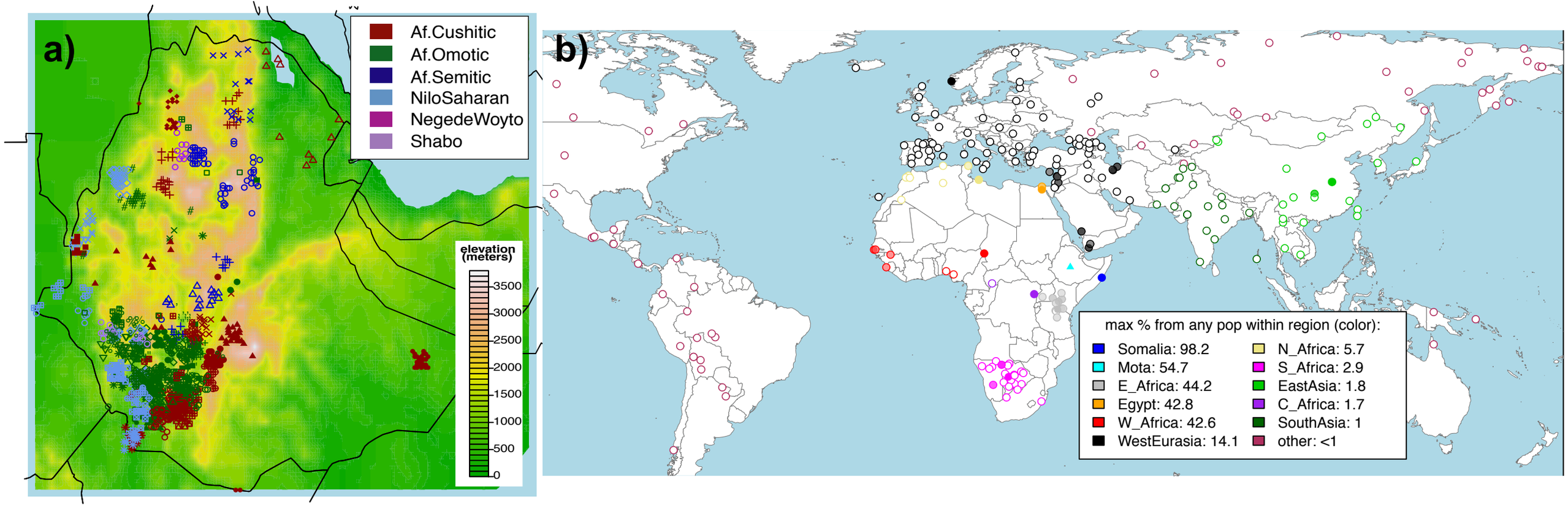
van Dorp L. et al. Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: Lessons for clustering-based inference. *PLoS Genet.* **11**, e1005397 (2015).

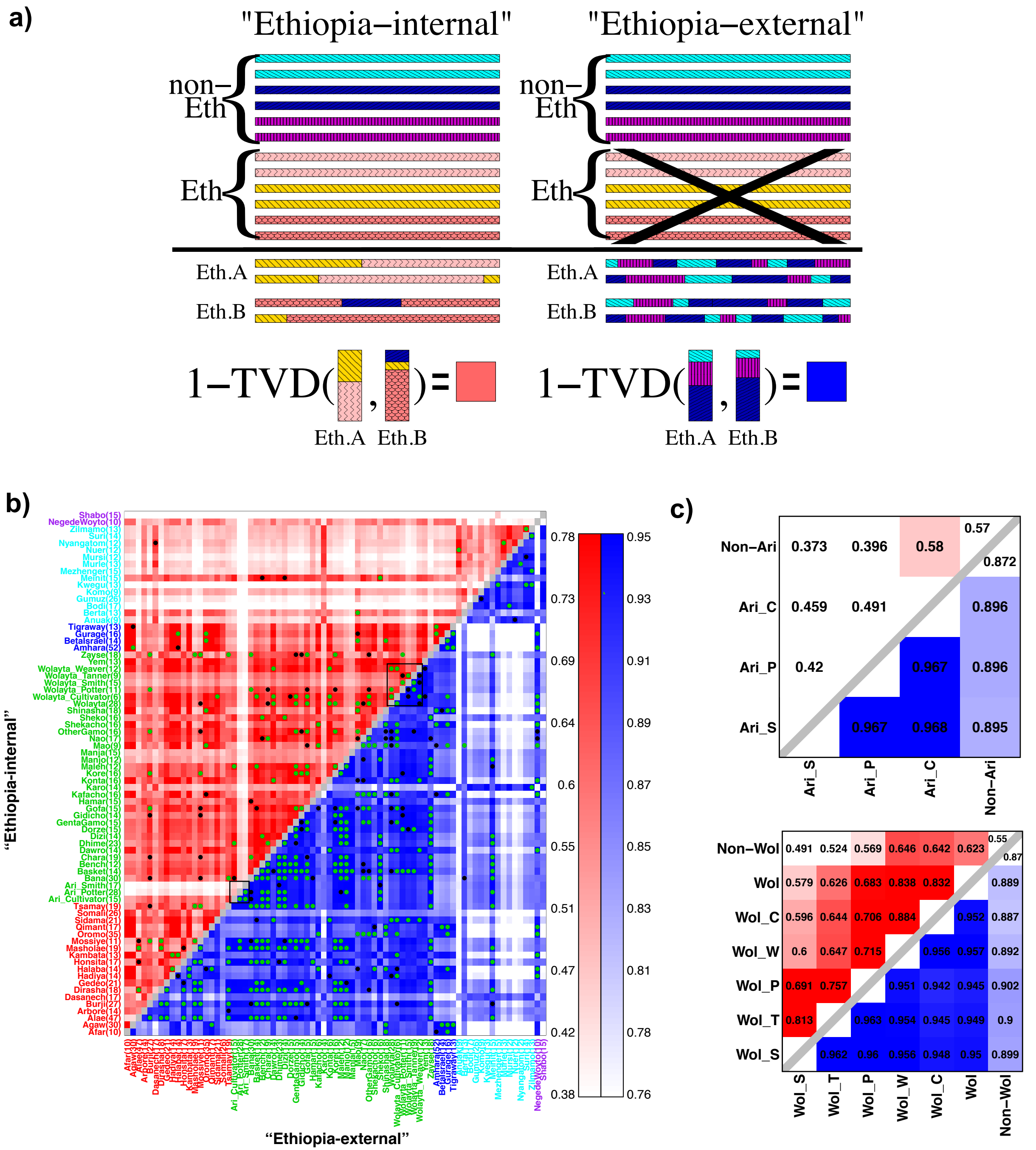
van Dorp, L. et al. Genetic legacy of state centralization in the Kuba Kingdom of the Democratic Republic of the Congo. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 593-598 (2019).

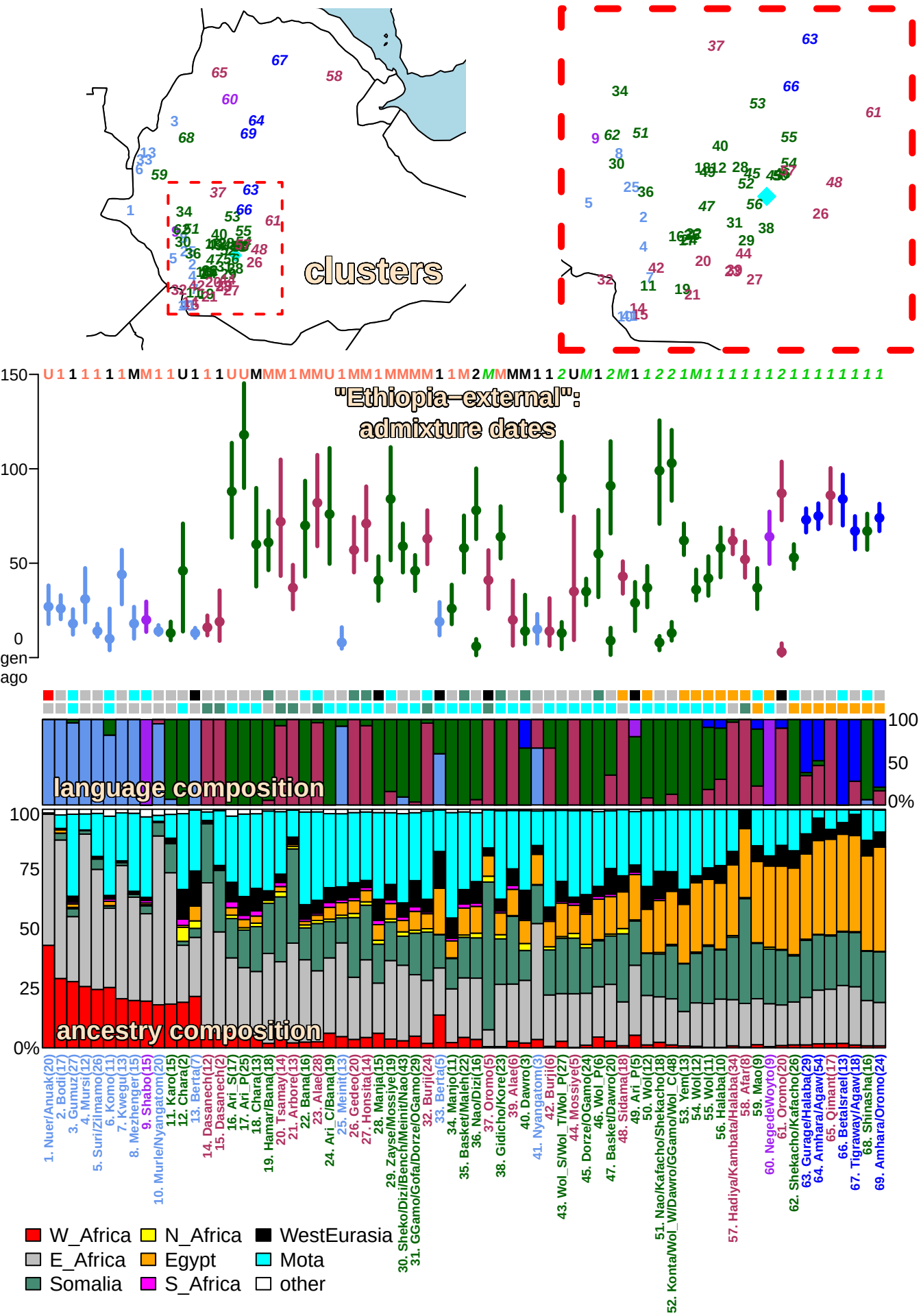
Vansina, J. New Linguistic Evidence and 'the Bantu Expansion'. *J. Afr. Hist.* **36**, 173–195 (1995).

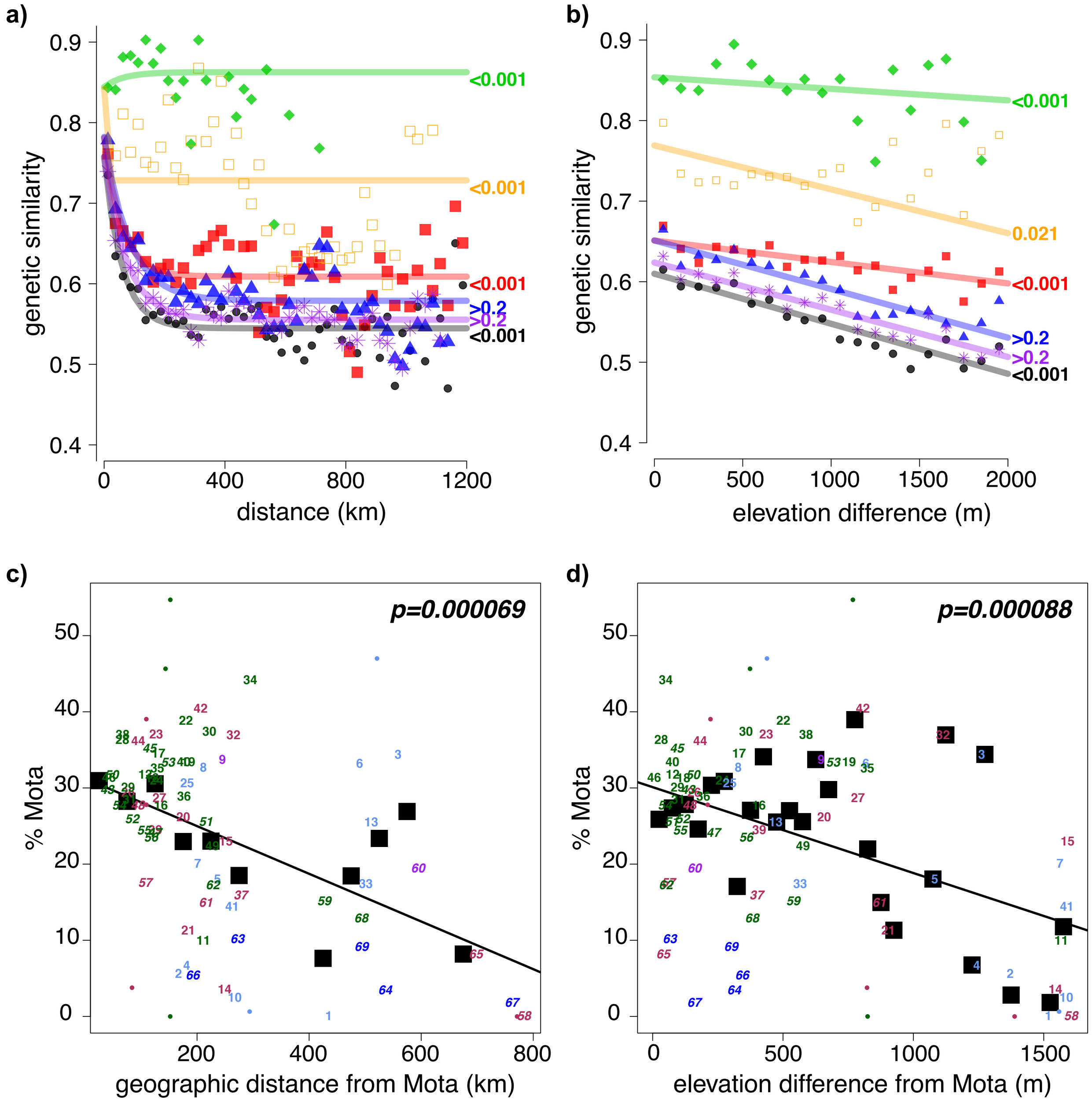
Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

Yang, G., Chen, T., Tao, Y. & Zhang, Z. Recent advances in the dark adaptation investigations. *Int. J. Ophthalmol.* **8**, 1245-1252 (2015).

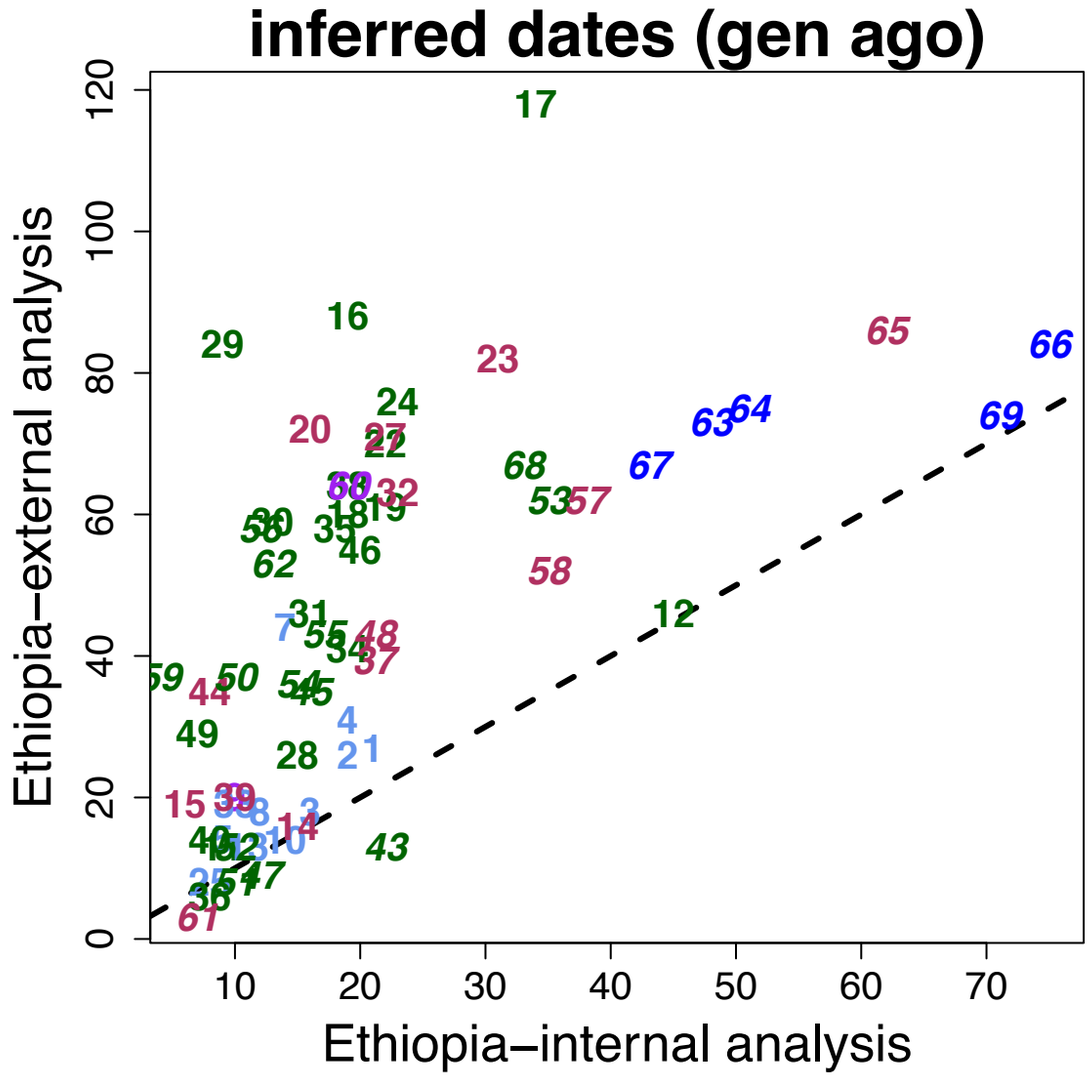

a)
Af.Cushitic
Af.Omotic
Af.Semitic
NiloSaharan
NegedeWoyto
Shabo
elevation (meters)
3500
3000
2500
2000
1500
1000
500
0
b)
max % from any pop within region (color):
Somalia: 98.2
Mota: 54.7
E_Africa: 44.2
Egypt: 42.8
W_Africa: 42.6
WestEurasia: 14.1
N_Africa: 5.7
S_Africa: 2.9
EastAsia: 1.8
C_Africa: 1.7
SouthAsia: 1
other: <1


a)
"Ethiopia–internal"
"Ethiopia–external"
non–Eth
Eth
Eth.A
Eth.B
1–TVD(,)=
Eth.A Eth.B
b)
"Ethiopia-internal"
"Ethiopia-external"
0.78
0.73
0.69
0.64
0.6
0.55
0.51
0.47
0.42
0.38
0.95
0.93
0.91
0.89
0.87
0.85
0.83
0.81
0.78
0.76
c)
Non–Ari 0.373 0.396 0.58 0.57
Ari_C 0.459 0.491 0.896 0.872
Ari_P 0.42 0.967 0.896
Ari_S 0.967 0.968 0.895
Ari_S Ari_P Ari_C Non–Ari
Non–Wol 0.491 0.524 0.569 0.646 0.642 0.623 0.55
Wol 0.579 0.626 0.683 0.838 0.832 0.889 0.87
Wol_C 0.596 0.644 0.706 0.884 0.952 0.887
Wol_W 0.6 0.647 0.715 0.956 0.957 0.892
Wol_P 0.691 0.757 0.951 0.942 0.945 0.902
Wol_T 0.813 0.963 0.954 0.945 0.949 0.9
Wol_S 0.962 0.96 0.956 0.948 0.95 0.899
Wol_S Wol_T Wol_P Wol_W Wol_C Wol Non–Wol

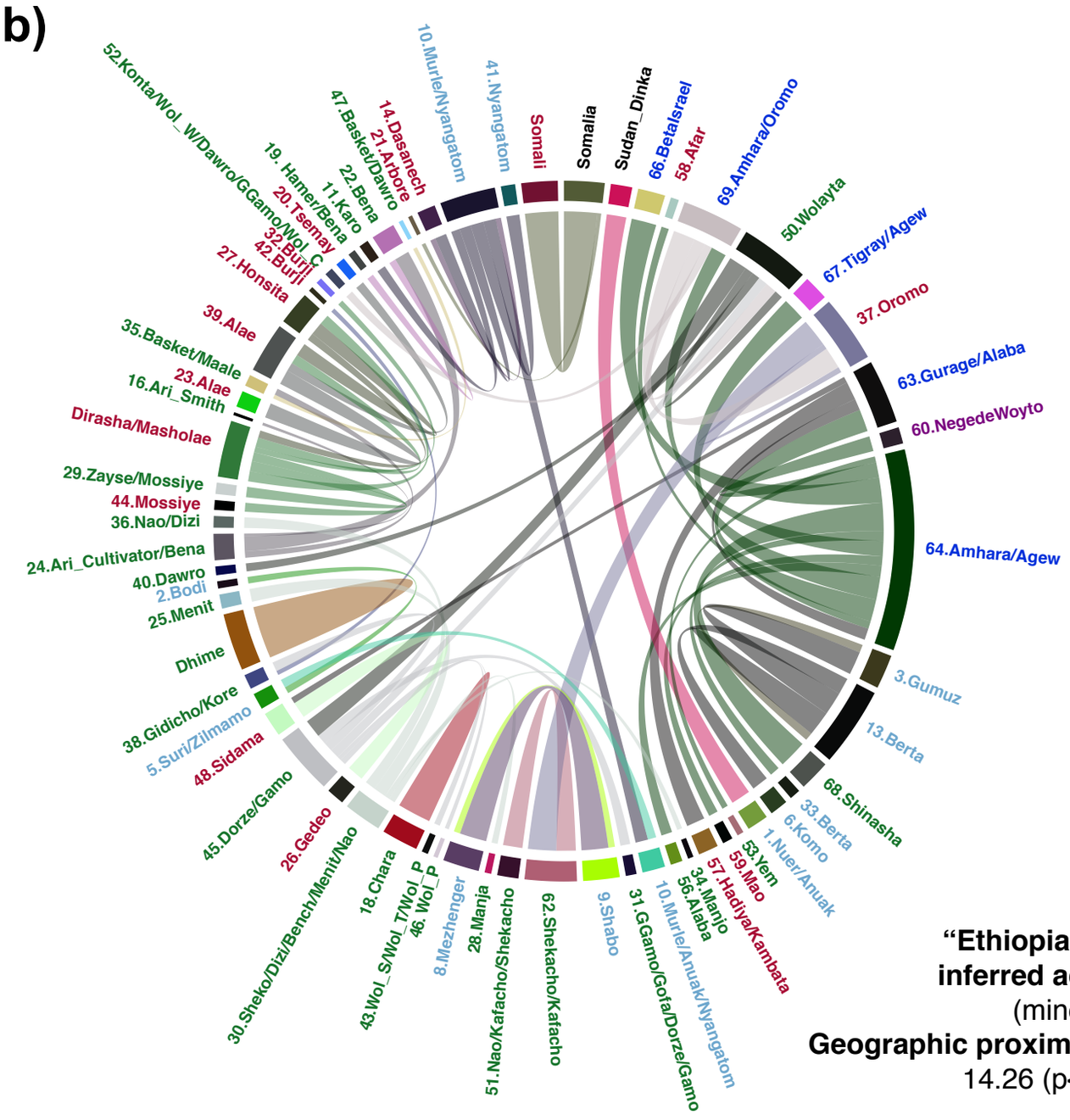

clusters
"Ethiopia-external":
admixture dates
150
100
50
0
gen ago
language composition
100
50
0%
ancestry composition
100
75
50
25
0%
1. Nuer/Anuak(20)
2. Bodi(17)
3. Gumuz(27)
4. Mursi(12)
5. Suri/Zilmamo(26)
6. Komo(11)
7. Kwegu(13)
8. Mezhenger(15)
9. Shabo(15)
10. Murle/Nyangatom(20)
11. Karo(15)
12. Chara(2)
13. Berta(7)
14. Dasanech(12)
15. Dasanech(2)
16. Ari_S(17)
17. Ari_P(25)
18. Chara(13)
19. Hamar/Bana(18)
20. Tsamay(14)
21. Arbore(13)
22. Bana(16)
23. Alae(28)
24. Ari_C/Bana(19)
25. Meinit(13)
26. Gedeo(20)
27. Honsita(14)
28. Manja(15)
29. Zayse/Mossiye(19)
30. Sheko/Dizi/Bench/Meinit/Nao(43)
31. GGamo/Gofa/Dorze/OGamo(29)
32. Burji(24)
33. Berta(5)
34. Manjo(11)
35. Basket/Maleh(22)
36. Nao/Dizi(16)
37. Oromo(5)
38. Gidicho/Kore(23)
39. Alae(6)
40. Dawro(3)
41. Nyangatom(3)
42. Burji(6)
43. Wol_S/Wol_T/Wol_P(27)
44. Mossiye(5)
45. Dorze/OGamo(24)
46. Wol_P(6)
47. Basket/Dawro(20)
48. Sidama(18)
49. Ari_P(5)
50. Wol(12)
51. Nao/Kafacho/Shekacho(18)
52. Konta/Wol_W/Dawro/GGamo/Wol_C(48)
53. Yem(13)
54. Wol(12)
55. Wol(11)
56. Halaba(10)
57. Hadiya/Kambata/Halaba(34)
58. Afar(8)
59. Mao(9)
60. NegedeWoyto(9)
61. Oromo(20)
62. Shekacho/Kafacho(26)
63. Gurage/Halaba(29)
64. Amhara/Agaw(54)
65. Qimant(17)
66. Betalsrael(13)
67. Tigraway/Agaw(18)
68. Shinasha(16)
69. Amhara/Oromo(24)
W_Africa
N_Africa
WestEurasia
E_Africa
Egypt
Mota
Somalia
S_Africa
other


a)
genetic similarity
distance (km)
<0.001
<0.001
<0.001
>0.2
>0.2
<0.001
b)
genetic similarity
elevation difference (m)
<0.001
0.021
<0.001
>0.2
>0.2
<0.001
c)
p=0.000069
% Mota
geographic distance from Mota (km)
d)
p=0.000088
% Mota
elevation difference from Mota (m)

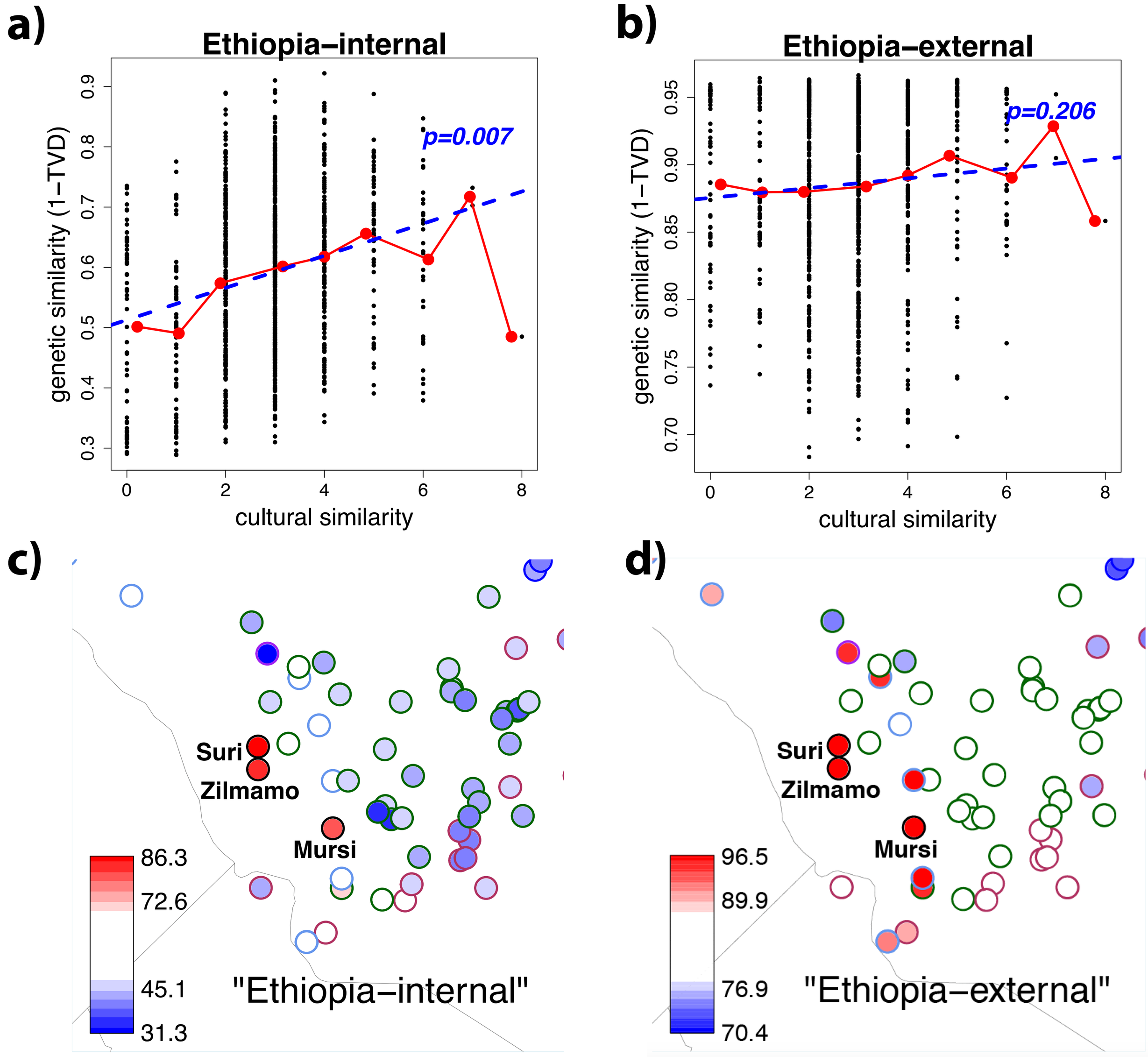
a)

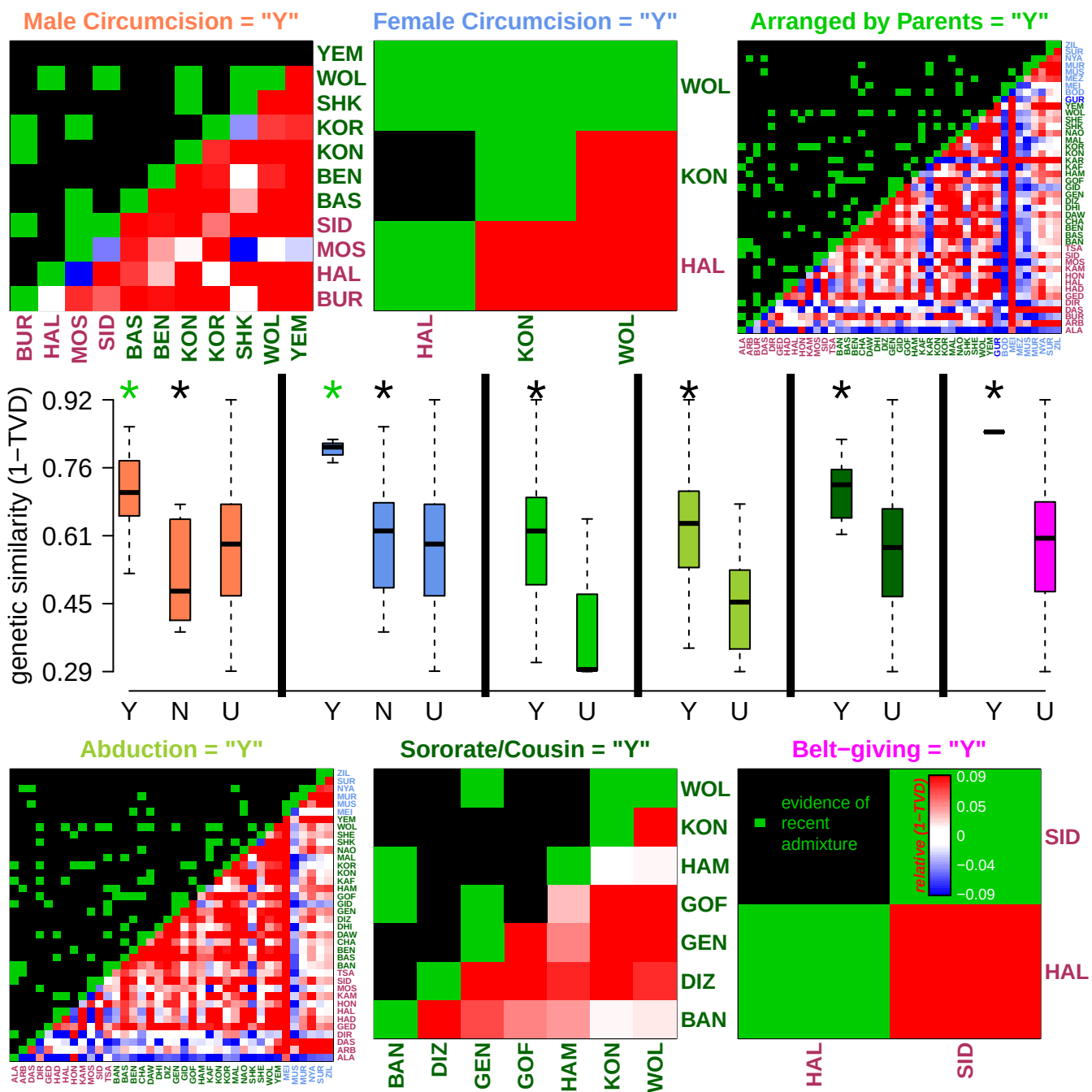


b)



"Ethiopia-recent": inferred admixture
(minor source)
Geographic proximity score:
14.26 ($p<0.00002$)


a)
Ethiopia–internal
p=0.007
genetic similarity (1−TVD)
cultural similarity
b)
Ethiopia–external
p=0.206
genetic similarity (1−TVD)
cultural similarity
c)
Suri
Zilmamo
Mursi
86.3
72.6
45.1
31.3
"Ethiopia–internal"
d)
Suri
Zilmamo
Mursi
96.5
89.9
76.9
70.4
"Ethiopia–external"


Male Circumcision = "Y"
YEM
WOL
SHK
KOR
KON
BEN
BAS
SID
MOS
HAL
BUR
BUR HAL MOS SID BAS BEN KON KOR SHK WOL YEM
Female Circumcision = "Y"
WOL
KON
HAL
HAL KON WOL
Arranged by Parents = "Y"
genetic similarity (1−TVD)
0.92
0.76
0.61
0.45
0.29
Y N U
Y N U
Y U
Y U
Y U
Y U
Abduction = "Y"
Sororate/Cousin = "Y"
WOL
KON
HAM
GOF
GEN
DIZ
BAN
BAN DIZ GEN GOF HAM KON WOL
Belt−giving = "Y"
evidence of recent admixture
relative (1−TVD)
0.09
0.05
0
−0.04
−0.09
SID
HAL
HAL SID