Genetic Ancestry of Hadza and Sandawe Peoples Reveals Ancient Population Structure in Africa

Daniel Shriner[1], Fasil Tekola-Ayele[2], Adebowale Adeyemo[1], and Charles N. Rotimi[1,*]

[1]Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Maryland

[2]Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland

*Corresponding author: E-mail: rotimic@mail.nih.gov.

Accepted: March 5, 2018

Abstract

The Hadza and Sandawe populations in present-day Tanzania speak languages containing click sounds and therefore thought to be distantly related to southern African Khoisan languages. We analyzed genome-wide genotype data for individuals sampled from the Hadza and Sandawe populations in the context of a global data set of 3,528 individuals from 163 ethno-linguistic groups. We found that Hadza and Sandawe individuals share ancestry distinct from and most closely related to Omotic ancestry; share Khoisan ancestry with populations such as ≠Khomani, Karretjie, and Ju/'hoansi in southern Africa; share Niger-Congo ancestry with populations such as Yoruba from Nigeria and Luhya from Kenya, consistent with migration associated with the Bantu Expansion; and share Cushitic ancestry with Somali, multiple Ethiopian populations, the Maasai population in Kenya, and the Nama population in Namibia. We detected evidence for low levels of Arabian, Nilo-Saharan, and Pygmy ancestries in a minority of individuals. Our results indicate that west Eurasian ancestry in eastern Africa is more precisely the Arabian parent of Cushitic ancestry. Relative to the Out-of-Africa migrations, Hadza ancestry emerged early whereas Sandawe ancestry emerged late.

Key words: Africa, ancestry, migration, population structure.

Introduction

The Hadza and Sandawe populations in present-day Tanzania speak click languages thought to be distantly related to southern African Khoisan languages (Ehret 2000; Güldemann and Vossen 2000; Heine and Nurse 2000). (Throughout, we use "Khoe-San" to refer to people and "Khoisan" to refer to both language and ancestry, without implying identity.) Eastern and southern African hunter-gatherer groups have been genetically separated for at least 30,000 years (Tishkoff et al. 2007). Herding and cultivating Cushitic speakers reached northern Tanzania ~4,000 years ago, followed by pastoralist Nilo-Saharan speakers, and then followed by agricultural Niger-Congo speakers ~2,500 years ago (Newman 1995).

In the Hadza population, the distribution of Y chromosomes includes mostly B2 haplogroups, with a smaller number of E1b1a haplogroups, which are common in Niger-Congo-speaking populations, and E1b1b haplogroups, which are common in Cushitic populations (Tishkoff et al. 2007). In the Sandawe population, E1b1a and E1b1b haplogroups are more common, with lower frequencies of B2 and A3b2 haplogroups (Tishkoff et al. 2007). Using autosomal data, Tishkoff et al. (2009) concluded that the Hadza population had ~72% ancestry distantly related to Khoisan and Pygmy ancestries, with ~22% Niger-Congo ancestry and ~6% Cushitic ancestry. Similarly, the Sandawe population had ~73% ancestry distantly related to Khoisan and Pygmy ancestries, with ~18% Niger-Congo ancestry and ~9% Cushitic ancestry (Tishkoff et al. 2009). Henn et al. (2011) concluded that 1) the Hadza and Sandawe populations share ancestry with the South African ≠Khomani population but distinct from Pygmy ancestry, 2) the Hadza and Sandawe populations share substantial amounts of eastern African ancestry with the Maasai population in Kenya, 3) the Hadza and Sandawe populations share ancestry with Niger-Congo-speaking populations such as Yoruba from Nigeria and Luhya from Kenya, and 4) the Sandawe population shares a small amount of ancestry with Europeans (represented by Tuscans from Italy). Using whole-genome sequence data, Lachance et al. (2012) concluded that Khoisan-speaking populations diverged first, followed by divergence of Pygmies, and then followed by divergence of the ancestors of the Hadza and Sandawe populations. Pickrell et al. (2012) also

inferred that the Hadza and Sandawe populations shared ancestry with Khoisan-speaking populations, with gene flow around 3,000 years ago of west Eurasian ancestry into eastern Africa (Pickrell et al. 2014).

The recent origin of modern humans in sub-Saharan Africa involves a basal divergence event such that one lineage includes Khoisan ancestry in south Africa; Pygmy ancestry in central Africa; Niger-Congo ancestry across west, east, and south Africa; and Cushitic, Nilo-Saharan, and Omotic ancestries in east Africa (Shriner et al. 2014). The other lineage includes Berber ancestry in north Africa; Indian and Kalash ancestries in south Asia; Chinese, Japanese, and southeast Asian ancestries in east Asia; Siberian ancestry in north Asia; Native American ancestry in the Americas; Melanesian ancestry in Oceania; southern and northern European ancestries; and Arabian and Levantine-Caucasian ancestries in the Middle East and the Caucasus (Shriner et al. 2014). These ancestries reflect shared history at a scale bigger than tribes or ethno-linguistic groups but smaller than continents. The divergence of ancestries is mainly due to random genetic drift following serial founder effects as modern humans peopled the world (Li et al. 2008). A notable exception is Cushitic ancestry, which did not form by a splitting event but rather by a mixing event between Arabian ancestry and Nilo-Saharan or Omotic ancestry (Shriner et al. 2016). We previously described integration of genotype data from 12 human diversity projects, yielding 3,528 unrelated individuals from around the world (Shriner et al. 2014). To more precisely identify west Eurasian ancestry and to investigate the origins and phylogenetic relationships of Hadza and Sandawe ancestries in the global context, we merged samples from these two populations (Henn et al. 2011) into our data set. Using cluster analyses and analysis of ancestry-specific allele frequencies, we provide greater detail about the history of the Hadza and Sandawe populations as well as novel insights into the ancestries of modern humans.

Cluster Analyses in a Global Context

We integrated 13 Hadza and 25 Sandawe individuals into our global data set of 3,528 individuals. The process of merging genotype data from different data sets and genotyping platforms left 19,206 SNPs. To address the possible effect of SNP ascertainment bias on F_{ST} estimation, we compared pairwise estimates for the samples from the 1000 Genomes Project (Auton et al. 2015) based on our panel of SNPs versus the whole genome sequences. The median difference was 0.0031 (95% confidence interval $[-0.0002, 0.0177]$), indicating that F_{ST} estimation was not significantly biased by either SNP ascertainment or the number of SNPs (supplementary table S1, Supplemenatry Material online).

We first performed semisupervised clustering. In this type of cluster analysis, the goal is to describe the Hadza and Sandawe samples in terms of predefined allele frequencies while allowing the samples to update the allele frequency estimates. Thus, for each of the 19 ancestries we previously described (Shriner et al. 2014), we generated pseudo-samples by identifying which individuals had the highest percentage of that ancestry, regardless of the sample of origin. After denoising and renormalization, seven of the pseudo-samples were ancestrally homogeneous (supplementary table S2, Supplemenatry Material online). We then estimated the ancestral composition of the Hadza and Sandawe individuals using these pseudo-samples as reference training data. At the individual level, 12 Hadza individuals had ancestry corresponding to the Omotic pseudo-sample, nine had ancestry corresponding to the Khoisan pseudo-sample, eight had ancestry corresponding to the Niger-Congo pseudo-sample, and one had contributions corresponding to the Cushitic, Nilo-Saharan, and Pygmy pseudo-samples. Thus, the predominant ancestry of the Hadza individuals was most closely related to Omotic ancestry. At the sample level, Niger-Congo ancestry was not significant because of large variance between individuals (table 1). There was no evidence for ancestry corresponding to any of the Asian or European pseudo-samples. At the individual level, all 25 Sandawe individuals had contributions from Cushitic, Khoisan, Niger-Congo, and Omotic pseudo-samples. Additionally, one Sandawe individual had ancestry corresponding to the Arabian pseudo-sample and one had ancestry corresponding to the Pygmy pseudo-sample. Compared with the Hadza sample, the Sandawe sample had larger contributions from the Niger-Congo and Cushitic pseudo-samples (table 1). As with the Hadza individuals, there was no evidence in the Sandawe individuals for

Table 1

Mean Ancestry Proportions in the Hadza and Sandawe Samples in Semi-Supervised Clustering Analysis, After Denoising and Rescaling

Pseudo-Sample	Hadza	Sandawe
Amerindian	0	0
Arabian	0	0.001
Berber	0	0
Chinese	0	0
Cushitic	0	0.258
Indian	0	0
Japanese	0	0
Kalash	0	0
Khoisan	0.073	0.088
Levantine-Caucasian	0	0
Niger-Congo	0	0.392
Nilo-Saharan	0	0
Northern European	0	0
Oceanian	0	0
Omotic	0.862	0.261
Pygmy	0.066	0
Siberian	0	0
Southeastern Asian	0	0
Southern European	0	0

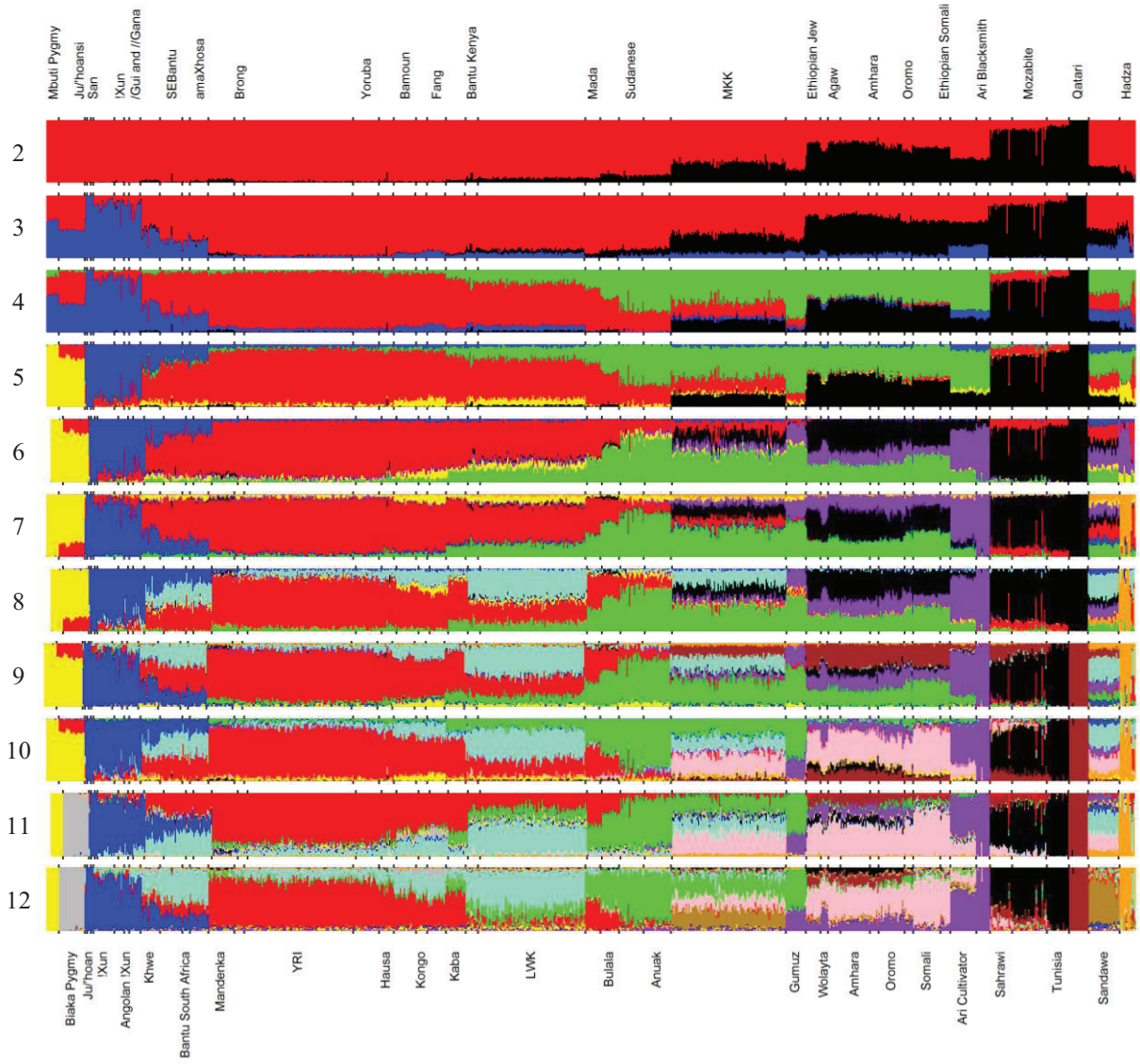

Fig. 1.—Unsupervised clustering analysis. The 47 samples are labeled alternating across the top and bottom. The numbers of ancestries are labeled in the left margin. In the plot with 12 ancestries, the ancestries from left to right are eastern Pygmy (yellow), western Pygmy (gray), Khoisan (blue), eastern and southern Bantu-speaking (aquamarine), Western Niger-Congo (red), Nilo-Saharan (green), Cushitic (pink), Omotic (purple), Berber (black), Arabian (brown), Sandawe (dark goldenrod), and Hadza (orange).

ancestry corresponding to any of the Asian or European pseudo-samples. Taken together, the semi-supervised clustering analysis revealed a combination of Cushitic, Khoisan, Niger-Congo, Nilo-Saharan, Omotic, and Pygmy ancestries in the Hadza and Sandawe individuals.

We then performed unsupervised clustering, to allow for the possibility of distinct ancestries not captured by our reference panel. Given that semisupervised clustering revealed no Asian or European ancestries in either the Hadza or Sandawe samples, we filtered our reference set to exclude samples with these ancestries. The first split separates sub-Saharan African ancestry from all other ancestries (fig. 1). Subsequent splits define Khoisan ancestry ($K = 3$ ancestral components), eastern African ancestry ($K = 4$), and Pygmy ancestry ($K = 5$). Eastern African ancestry is split into Nilo-Saharan and Omotic ancestries at $K = 6$ (which is the value of K with the lowest cross-validation error). The ancestry of the Hadza sample is predominantly explained by this Omotic ancestry, consistent with the semisupervised analysis. The ancestry at $K = 7$ corresponds to the majority of the Hadza and a minority of the Sandawe, replacing the Omotic component in the former but not as much in the latter. Niger-Congo ancestry in western Africa separated from eastern and southern Bantu-speaking ancestry at $K = 8$, followed by Berber ancestry

Table 2
Pairwise Distances between Ancestries

Ancestry	Arabian	Berber	Cushitic	Eastern and Southern Bantu	Eastern Pygmy	Hadza	Khoisan	Nilo-Saharan	Omotic	Sandawe	Western Niger-Congo	Western Pygmy
Arabian	0	0.023	0.039	0.067	0.089	0.129	0.090	0.064	0.059	0.064	0.067	0.076
Berber	0.023	0	0.026	0.052	0.074	0.114	0.075	0.049	0.044	0.049	0.052	0.061
Cushitic	0.039	0.026	0	0.063	0.085	0.125	0.086	0.060	0.055	0.060	0.063	0.072
Eastern and Southern Bantu	0.067	0.052	0.063	0	0.033	0.099	0.034	0.017	0.039	0.043	0.008	0.020
Eastern Pygmy	0.089	0.074	0.085	0.033	0	0.122	0.028	0.039	0.062	0.064	0.033	0.022
Hadza	0.129	0.114	0.125	0.099	0.122	0	0.122	0.096	0.101	0.110	0.099	0.108
Khoisan	0.090	0.075	0.086	0.034	0.028	0.122	0	0.040	0.062	0.065	0.033	0.023
Nilo-Saharan	0.064	0.049	0.060	0.017	0.039	0.096	0.040	0	0.036	0.042	0.017	0.026
Omotic	0.059	0.044	0.055	0.039	0.062	0.101	0.062	0.036	0	0.044	0.039	0.048
Sandawe	0.064	0.049	0.060	0.043	0.064	0.110	0.065	0.042	0.044	0	0.043	0.051
Western Niger-Congo	0.067	0.052	0.063	0.008	0.033	0.099	0.033	0.017	0.039	0.043	0	0.020
Western Pygmy	0.076	0.061	0.072	0.020	0.022	0.108	0.023	0.026	0.048	0.051	0.020	0

($K = 9$), Cushitic ancestry ($K = 10$), western versus eastern Pygmy ancestry ($K = 11$), and Sandawe ancestry ($K = 12$, fig. 1). Six of the 13 Hadza individuals were ancestrally homogeneous for Hadza ancestry (fig. 1). Of the remaining seven Hadza individuals, all had Hadza ancestry, ranging from 4.5% to 78.3%; six had Niger-Congo ancestry, ranging from 10.2% to 81.8%; three had Cushitic ancestry, ranging from 6.7% to 24.5%; one had 7.9% Omotic ancestry; and one had 6.8% Nilo-Saharan ancestry. All Sandawe individuals had multiple ancestries but with similar ancestral fractions across individuals (fig. 1). Comparing $K = 11$ to $K = 12$, Sandawe ancestry consisted of 39.1% eastern and southern African Bantu-speaking ancestry, 28.7% Cushitic ancestry, 12.6% Omotic ancestry, 10.4% Hadza ancestry, and 9.3% Khoisan ancestry (fig. 1). In addition to these common ancestries, six individuals had eastern Pygmy ancestry, ranging from 4.5% to 7.2%, and four individuals had Arabian ancestry, ranging from 4.1% to 5.6%.

Early Divergence of Hadza Ancestry

To place Hadza ancestry into context, we used TreeMix to estimate pairwise $\hat{F}_{ST}$ values based on the ancestry-specific allele frequencies inferred from the unsupervised clustering analysis conditional on $K = 12$ (table 2). By basing this analysis on ancestry-specific allele frequencies rather than sample-based allele frequencies, the effects of recent admixture or migration in the samples were removed. To allow for ancient admixture and gene flow, we incorporated migration events between ancestries in the TreeMix model. We found evidence for four migration events: unstable placement of Sandawe ancestry and excess covariance between Nilo-Saharan and Cushitic ancestries, between Khoisan and Hadza ancestries, and between Khoisan and Omotic ancestries.

F_{ST} can be written as a composite function of effective population sizes and divergence time. Consequently, the branching orders based on TreeMix migration graphs do not necessarily correspond to chronological branching orders. Therefore, we obtained estimates of effective populations sizes based on whole genome sequence data (supplementary table S3, Supplemenatry Material online) (Lachance et al. 2012; Mallick et al. 2016). The effective population size estimated from Hadza whole genome sequence data indicate a reduction in diversity of 24% compared with other sub-Saharan Africans (Lachance et al. 2012), smaller in magnitude than the reduction associated with the Out-of-Africa migration(s) that occurred ~76,000–55,000 years ago (Fu et al. 2013; Poznik et al. 2016; Rieux et al. 2014). We estimated divergence times of ~98,000 to ~96,000 years for Hadza ancestry from Eastern Pygmy and Khoisan ancestries, respectively, followed by divergence times of ~89,000 and ~88,000 years for Western Pygmy and Sandawe ancestries, respectively, and then followed by divergence times of ~81,000 to 76,000 years for Arabian, Berber, eastern and southern Bantu-speaking, Nilo-Saharan, and Western Niger-Congo ancestries (table 3). These divergence times are all before Out-of-Africa, and therefore support early divergence of Hadza ancestry. In contrast, we estimated divergence times for Sandawe ancestry of ~55,000–34,000 years (table 3). These divergence times are after Out-of-Africa but before the ancestral split of present-day speakers of Niger-Congo and Nilo-Saharan languages.

Discussion

We have performed genetic analyses to better understand the history of the Hadza and Sandawe populations in Tanzania. Using a combination of semi-supervised and unsupervised

Table 3
Hadza and Sandawe Divergence Time Estimates

Ancestry	Hadza Divergence Time (Years)	Sandawe Divergence Time (Years)
Arabian	81,400	41,900
Berber	75,900	33,900
Eastern and Southern Bantu	80,300	37,000
Eastern Pygmy	97,500	54,500
Hadza	NA	87,800
Khoisan	96,400	54,000
Nilo-Saharan	76,600	35,100
Sandawe	87,800	NA
Western Niger-Congo	79,500	36,500
Western Pygmy	89,400	44,700

clustering analysis with a large global reference panel, we better defined ancestral composition. In the context of the 19 ancestries we previously detected (Shriner et al. 2014), we found that the Hadza and Sandawe populations shared a distinct ancestry that we eponymously named Hadza ancestry (because six Hadza individuals were homogeneous for this ancestry). We also found that genotype and sequence data support an early divergence model for Hadza ancestry.

We detected low levels of mixed ancestry among a subset of Hadza individuals. Specifically, we detected Niger-Congo ancestry in 6 of 13 Hadza individuals. We also detected Cushitic ancestry in 3 of 13 Hadza individuals and at lower levels than Niger-Congo ancestry. Additionally, we detected Nilo-Saharan ancestry in one Hadza individual. These results are consistent with the presence of both E1b1a Y chromosome haplogroups, common in populations with Niger-Congo ancestry, and E1b1b Y chromosome haplogroups, common in populations with Nilo-Saharan and Cushitic ancestries, in the Hadza population (Tishkoff et al. 2007). Collectively, the autosomal data and Y chromosome data provide evidence for the presence of Hadza, Niger-Congo, Cushitic, and Nilo-Saharan ancestry in the Hadza population. Within the Hadza sample, the simultaneous presence of ancestrally heterogeneous individuals with large amounts of Niger-Congo ancestry and individuals homogeneous for Hadza ancestry is consistent with very recent admixture.

We detected a more complex mixture of ancestries in the Sandawe individuals than in the Hadza individuals. Whether this finding reflects more inter-mating in the Sandawe population or more loss of lineages in the Hadza population is unknown. We identified Niger-Congo (more specifically, eastern and southern Bantu-speaking) ancestry, Cushitic ancestry, Omotic ancestry, Hadza ancestry, and Khoisan ancestry in all the Sandawe individuals. Additionally, we identified Arabian ancestry and eastern Pygmy ancestry in a minority of Sandawe individuals. Compared with the Y chromosomal haplogroup frequencies in the Hadza population, the Sandawe population has more E1b1a and E1b1b and less B2 (Tishkoff et al. 2007), consistent with our autosomal findings of higher amounts of Niger-Congo and Cushitic ancestry in the Sandawe individuals.

Tishkoff et al. (2009) reported Hadza and Sandawe ancestries but did not include samples of speakers of Omotic languages. Our results are consistent with Hadza ancestry having formed by a splitting process. In contrast, our results indicate that Sandawe ancestry reflects a mixture of eastern and southern African Bantu-speaking, Cushitic, Omotic, Hadza, and Khoisan ancestries.

Pickrell et al. (2012) inferred the presence in both the Hadza and Sandawe populations of ancestry shared with Khoisan-speaking peoples. Based on the clustering analyses, we found no Khoisan ancestry in the Hadza individuals and low levels of Hadza, Khoisan, and Omotic ancestries in the Sandawe individuals. However, analysis of ancestral allele frequencies revealed a migration event between Khoisan and Hadza ancestries and a migration event between Khoisan and Omotic ancestries. Thus, ancestry in the Hadza and Sandawe populations shared with Khoisan-speaking populations could reflect the distant common ancestor of Hadza, Khoisan, and Omotic ancestries or these more recent migration events.

Mitochondrial DNA provides uniparental information about maternal lineages. The divergence of L0d and L0a'b'f'k haplogroups occurred ~119,000 [100,100–138,200] years ago, the divergence of L0k and L0a'b'f haplogroups occurred ~98,700 years [82,300 to 115,400] ago, and the divergence of L0a occurred ~42,400 [33,000–52,000] years ago (Rito et al. 2013). The! Xun have >50% L0d and ~25% L0k, whereas the Hadza population has 5% L0a and the Sandawe population has 26% L0a, L0d, and L0f (Tishkoff et al. 2007). The L0k haplogroup was not observed in either the Hadza or Sandawe populations (Tishkoff et al. 2007). These data are consistent with the early divergence of Hadza ancestry, that is, before the divergence of L0k, and a more recent acquisition of L0a. However, the absence of L0k could have resulted from loss in the Hadza and Sandawe populations due to random genetic drift. Y chromosome DNA provides uniparental information about paternal lineages. The emergence of B-M181 105,800 years ago and B2-M182 100,600 years ago (Poznik et al. 2016) are consistent with an early divergence of Hadza ancestry.

Collectively, autosomal, Y, and mitochondrial DNA support early divergence of Hadza ancestry (Knight et al. 2003; Tishkoff et al. 2007). Lachance et al.'s (2012) conclusion of late divergence was based on a neighbor-joining tree; the assumption of treeness or bifurcation is violated by admixture and gene flow, thus invalidating their conclusion. An early divergence of Hadza and Khoe-San peoples is consistent with the grouping of Hadza and Khoisan languages (Knight et al. 2003). On the other hand, evidence for gene flow between 7.5 and 20 thousand years ago is consistent with the

hypothesis that click sounds are a recent addition to Hadza and Sandawe languages (Rito et al. 2013). Our results suggest a third possibility. The semi-supervised analysis revealed that Hadza ancestry is closer to Omotic ancestry than to Khoisan ancestry. Also, Omotic ancestry does not cluster with Arabian, Berber, or Cushitic ancestries, consistent with the hypothesis that Omotic languages are not part of the Afroasiatic language family (Theil 2006). Taken together, our genetic findings support a phylolinguistic hypothesis that Omotic and Hadza languages form a language family (Elderkin 1982). Furthermore, if both Cushitic and Niger-Congo ancestries in the Sandawe sample are comparatively recently acquired, then the core of the Sandawe sample is predominantly Omotic, supporting a phylolinguistic hypothesis that the Sandawe language also belongs with Omotic and Hadza languages. The hypothesis that Hadza and Sandawe peoples are not Khoe-San peoples is supported by previous osteological and serogenetic studies (Morris 2002).

West Eurasian ancestry (closely related to southern European or Levantine populations) has been described throughout eastern Africa and southern African Khoe-San populations (Pickrell et al. 2014). The results of our semisupervised clustering analysis directly exclude southern European or Levantine-Caucasian ancestries in both the Hadza and Sandawe samples. We previously found that Cushitic ancestry formed by sex-biased gene flow, with female ancestry closely related to Arabian and male ancestry closely related to Nilo-Saharan or Omotic ancestry (Shriner et al. 2016). We found the largest amounts of Cushitic ancestry in Somalia and Ethiopia, with smaller amounts across northern Africa and the Middle East (Shriner et al. 2014). In Tanzania, we detected more Cushitic than Arabian or Nilo-Saharan ancestry (Shriner et al. 2014). Across southern Africa, we detected more Nilo-Saharan ancestry than Cushitic ancestry and no Arabian ancestry (Shriner et al. 2014). The Nama sample was the only southern African sample in which we detected Cushitic ancestry, likely identical to the East African ancestry shared between the Nama and the Maasai (Schlebusch et al. 2012). This result is consistent with the high frequency in the Nama of the lactase persistence trait, which is associated with pastoralists more so than with foragers or agriculturists, and the derived allele –14010*C, thought to have originated in eastern Africa (Coelho et al. 2009; Macholdt et al. 2014) or more specifically within individuals with Cushitic ancestry (Breton et al. 2014; Ranciaro et al. 2014). This result is also consistent with the distribution of the Y chromosomal haplogroup E3b1f-M293, proposed to have spread from eastern to southern Africa before the Bantu Expansion (Henn et al. 2008). Taken together, we infer that west Eurasian ancestry reflects the Arabian parentage of Cushitic ancestry.

In summary, our autosomal data provide evidence for ancient population structure in Africa. We found that Hadza ancestry diverged early, rather than late. We found evidence for contributions of Cushitic and Niger-Congo ancestries in Tanzania, consistent with the movements of herding and cultivating Cushitic speakers ~4,000 years ago and agricultural Niger-Congo speakers ~2,500 years ago (Newman 1995). However, we did not find evidence of a substantial contribution of Nilo-Saharan ancestry that might have resulted from movement of pastoralist Nilo-Saharan speakers (Newman 1995). We also identified west Eurasian ancestry in eastern and southern African populations more precisely as the Arabian parent of Cushitic ancestry. Finally, our ancestry analyses support the hypothesis that Omotic, Hadza, and Sandawe languages group together, rather than Omotic languages belonging to the Afroasiatic family and Hadza and Sandawe languages belonging to the Khoisan family.

Materials and Methods

Ethics

This project consisted of genotype data in the public domain and was determined to be excluded from IRB Review by the National Institutes of Health Office of Human Subjects Research Protections (OHSRP ID# 17-NHGRI-00282).

Materials

We obtained genotype data for 542,263 markers from 17 Hadza and 28 Sandawe individuals, with 1st degree relatives already excluded (Henn et al. 2011). SNPs with a call rate-$<$10%, a minor allele frequency $<$0.5%, or not in Hardy–Weinberg equilibrium ($P < 0.001$ per sample) were excluded (Henn et al. 2011). We removed four Hadza individuals and three Sandawe individuals identified as 2nd degree or closer relatives using the –genome function in PLINK (Purcell et al. 2007). All individuals had sample call rates $>$95%. We then integrated these data into our global data set of 3,528 unrelated individuals genotyped at 19,372 diallelic autosomal SNPs (Shriner et al. 2014), yielding a merged data set of 3,566 unrelated individuals genotyped at 19,206 SNPs. We did not include the ≠Khomani individuals from Henn et al. (2011) due to excessive relatedness to the ≠Khomani individuals from Schlebusch et al. (2012) already present in our global data set. Data management and quality control were performed using PLINK version 1.07 (Purcell et al. 2007). Graphical and statistical analyses were performed using R (R Core Team 2013).

SNP Ascertainment Bias

To investigate possible SNP ascertainment bias, we used the –weir-fst-pop function in VCFtools, version 0.1.12b (Danecek et al. 2011). We compared pairwise F_{ST} estimates based on our panel of SNPs to pairwise estimates based on whole genome sequences (Auton et al. 2015).

Clustering Analysis

Semi-supervised and unsupervised clustering analyses were performed using ADMIXTURE version 1.22 (Alexander et al. 2009). Analyses were performed in triplicate with different starting seeds and five-fold cross-validation. Standard errors were estimated using 200 bootstrap replicates. For the semi-supervised analysis, we generated pseudo-samples by identifying individuals with the highest proportions of each of the 19 previously defined ancestries from a global reference panel of 3,528 individuals (Shriner et al. 2014) as training data. We required an ancestry proportion of $\geq$50%, regardless of the sample to which the individual belonged. For each ancestry, we sorted all individuals by ancestry proportion and identified the top 20, except for Oceanian ancestry, for which only seven individuals met our minimum ancestry proportion criterion. The labeled training data set for the semi-supervised analysis thus comprised genotype data for 367 individuals. We then analyzed the Hadza and Sandawe sample data given these training data. Given individual estimates, sample means were estimated using inverse variance weights. Sample means not significantly different from zero were zeroed out. Sample means were rescaled to sum to 1. Semi-supervised analysis is called supervised analysis in ADMIXTURE (Bansal and Libiger 2015) and can be performed by invoking the option –supervised. Supervised analysis based on predefined allele frequencies that are not allowed to be updated by the sample genotype data is called projection analysis in ADMIXTURE and can be performed by invoking the option –P (Shringarpure et al. 2016). Supervised analysis is not recommended if there are ancestries missing from the panel of predefined allele frequencies.

For the unsupervised analysis, we filtered our reference set to exclude samples with Asian and/or European ancestry. This filtering step resulted in a data set of 881 individuals from 47 samples, including/Gui and//Gana, ! Xun (two samples), Agaw, amaXhosa, Amhara (two samples), Angolan! Xun, Anuak, Ari Blacksmith, Ari Cultivator, Bamoun, Bantu from Kenya, Bantu from South Africa, Biaka Pygmy, Brong, Bulala, Ethiopian Jews, Fang, Gumuz, Hadza, Hausa, Ju/'hoansi (two samples), Kaba, Khwe, Kongo, Luhya, Maasai, Mada, Mandenka, Mbuti Pygmy, Mozabite, Oromo (two samples), Qatari Arab, Sahrawi, San, Sandawe, SEBantu (Sotho, Tswana, and Zulu), Somali (two samples), Sudanese, Tunisia, Wolayta, and Yoruba (two samples) (Shriner et al. 2014). Unsupervised analysis was performed in ADMIXTURE's default mode.

Migration Analysis

We reformatted the ancestry-specific allele frequencies from the unsupervised clustering analysis for migration analysis using TreeMix (Pickrell and Pritchard 2012). To do this, we estimated the effective sample size for each ancestry by summing the mixture proportions across individuals from ADMIXTURE's Q matrix. We then multiplied these effective sample sizes by two to estimate the effective number of alleles. Finally, we multiplied the effective number of alleles by the ancestry-specific allele frequencies to arrive at ancestry-specific allele counts. We defined a root by coding two copies of the ancestral allele at each position. We set the number of migration events from 0 to 8. Conditional on the number of migration events, we generated 100 bootstrap replicates. Our stopping rule was the number of migrations events at which the range of residuals stopped decreasing.

Estimation of Divergence Time

Each pairwise distance estimated from TreeMix involves the distance from a terminal tip to an internal node plus the distance from that internal node to a second terminal tip and thus is an estimate of $2\hat{F}_{ST}$, assuming equal sample size (Weir and Hill 2002). We estimated divergence time using the estimators $\hat{N}_e = \frac{\hat{\theta}}{4\hat{\mu}}$ and $1 - \hat{F}_{ST} = \left(1 - \frac{1}{2\tilde{N}_e}\right)^t$, in which t is generations, $\hat{\mu}$ is mutations per generation per site, $\hat{F}_{ST} =$ is half of the pairwise distance from TreeMix, and $\tilde{N}_e$ is the harmonic mean of the estimated effective population sizes $\hat{N}_e$ (Weir and Hill 2002), assuming that $\hat{F}_{ST} = 0$ at $t = 0$ (Hartl 2000). We estimated $\hat{N}_e$ using the mlrho autosomal heterozygosities $\hat{H}$ reported in the Simons Genome Diversity Project (Mallick et al. 2016) and the relationship $\hat{\theta} = \frac{\hat{H}}{1-\hat{H}}$. The Simons Genome Diversity Project did not include individuals representing Cushitic or Omotic ancestral majorities. We estimated $\hat{N}_e$ for Hadza and Sandawe by scaling the Western Pygmy estimate (Lachance et al. 2012). For $\hat{\mu}$, we used the weighted average (% non-sub-Saharan ancestry) $\times$ 1.17 $\times$ 10^{-8} mutations per generation per site + (% sub-Saharan African ancestry) $\times$ 0.97 $\times$ 10^{-8} mutations per generation per site (1000 Genomes Project Consortium 2010), based on supervised clustering analysis of the Simons Genome Diversity Project data (Mallick et al. 2016) and our reference panel of ancestries (Baker et al. 2017). To convert generations into years, we assumed a generation interval of 28 years (Fenner 2005; Moorjani et al. 2016).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Data Availability

The global reference data are available at http://crggh.nih.gov/resources.cfm, last accessed March 13, 2018. The Tanzania data are available at http://www-evo.stanford.edu/repository/paper0002/, last accessed March 13, 2018.

Acknowledgments

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official view

of the National Institutes of Health. This research was supported by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of Diabetes and Digestive and Kidney Disease, the Center for Information Technology, and the Office of the Director at the National Institutes of Health (1ZIAHG200362).

Literature Cited

1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9):1655–1664.
Auton A, et al. 2015. A global reference for human genetic variation. Nature 526(7571):68–74.
Baker JL, Rotimi CN, Shriner D. 2017. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. Sci Rep. 7(1):1572.
Bansal V, Libiger O. 2015. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. BMC Bioinformatics 16:4.
Breton G, et al. 2014. Lactase persistence alleles reveal partial east African ancestry of southern African Khoe pastoralists. Curr Biol. 24(8):852–858.
Coelho M, Sequeira F, Luiselli D, Beleza S, Rocha J. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. BMC Evol Biol. 9:80.
Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.
Ehret C. 2000. Language and history. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 272–297.
Elderkin ED. 1982. On the classification of Hadza. Sprache Und Geschichte in Afrika 4:67–82.
Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 128(2):415–423.
Fu Q, et al. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol. 23(7):553–559.
Güldemann T, Vossen R. 2000. Khoisan. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 99–122.
Hartl DL. 2000. A primer of population genetics. Sunderland (MA): Sinauer Associates, Inc.
Heine B, Nurse D. 2000. African languages: an introduction. Cambridge: Cambridge University Press.
Henn BM, et al. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci U S A. 105(31):10693–10698.
Henn BM, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A. 108(13):5154–5162.
Knight A, et al. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol. 13(6):464–473.
Lachance J, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150(3):457–469.
Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104.
Macholdt E, et al. 2014. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. Curr Biol. 24(8):875–879.
Mallick S, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538(7624):201–206.
Moorjani P, et al. 2016. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proc Natl Acad Sci U S A. 113(20):5652–5657.
Morris AG. 2002. Isolation and the origin of the Khoisan: late Pleistocene and early Holocene human evolution at the southern end of Africa. Hum Evol. 17(3-4):231–240.
Newman J. 1995. The peopling of Africa. New Haven: Yale University Press.
Pickrell JK, et al. 2012. The genetic prehistory of southern Africa. Nat Commun. 3:1143.
Pickrell JK, et al. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A. 111(7):2632–2637.
Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLOS Genet. 8(11):e1002967.
Poznik GD, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 48(6):593–599.
Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Ranciaro A, et al. 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet. 94(4):496–510.
Rieux A, et al. 2014. Improved calibration of the human mitochondrial clock using ancient genomes. Mol Biol Evol. 31(10):2780–2792.
Rito T, et al. 2013. The first modern human dispersals across Africa. Plos One 8(11):e80031.
Schlebusch CM, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338(6105):374–379.
Shriner D, Tekola-Ayele F, Adeyemo A, Rotimi CN. 2016. Ancient human migration after Out-of-Africa. Sci Rep. 6:26565.
Shriner D, Tekola-Ayele F, Adeyemo A, Rotimi CN. 2014. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. Sci Rep. 4:6055.
Shringarpure SS, Bustamante CD, Lange K, Alexander DH. 2016. Efficient analysis of large datasets and sex bias with ADMIXTURE. BMC Bioinformatics 17:218.
Theil R. 2006. Is Omotic Afroasiatic? Proceedings from the David Dwyer Retirement Symposium, Michigan State University, East Lansing, Michigan, 21 October 2006. Available online: http://www.uio.no/studier/emner/hf/iln/LING2110/v07/THEIL+Is+Omotic+Afroasiatic.pdf
Tishkoff SA, et al. 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol. 24(10):2180–2195.
Tishkoff SA, et al. 2009. The genetic structure and history of Africans and African Americans. Science 324(5930):1035–1044.
Weir BS, Hill WG. 2002. Estimating F-statistics. Annu Rev Genet. 36:721–750.

Associate editor: Partha Majumder