HUMAN GENETICS

Recovering signals of ghost archaic introgression in African populations

Arun Durvasula[1] and Sriram Sankararaman[1,2,3,4]*

While introgression from Neanderthals and Denisovans has been documented in modern humans outside Africa, the contribution of archaic hominins to the genetic variation of present-day Africans remains poorly understood. We provide complementary lines of evidence for archaic introgression into four West African populations. Our analyses of site frequency spectra indicate that these populations derive 2 to 19% of their genetic ancestry from an archaic population that diverged before the split of Neanderthals and modern humans. Using a method that can identify segments of archaic ancestry without the need for reference archaic genomes, we built genome-wide maps of archaic ancestry in the Yoruba and the Mende populations. Analyses of these maps reveal segments of archaic ancestry at high frequency in these populations that represent potential targets of adaptive introgression. Our results reveal the substantial contribution of archaic ancestry in shaping the gene pool of present-day West African populations.

INTRODUCTION

Admixture has been a dominant force in shaping patterns of genetic variation in human populations (*1*). Comparisons of genome sequences from archaic hominins to those from present-day humans have documented multiple interbreeding events, including gene flow from Neanderthals into the ancestors of all non-Africans (*2*), from Denisovans into Oceanians (*3*) and eastern non-Africans (*4*, *5*), as well as from early modern humans into the Neanderthals (*6*). However, the sparse fossil record and the difficulty in obtaining ancient DNA have made it challenging to dissect the contribution of archaic hominins to genetic diversity within Africa. While several studies have revealed contributions from deep lineages to the ancestry of present-day Africans (*7*–*12*), the nature of these contributions remains poorly understood.

RESULTS

We leveraged whole-genome sequence data from present-day West African populations and archaic hominins to compute statistics that are sensitive to introgression in the history of these populations. Specifically, we tabulated the distribution of the frequencies of derived alleles (where a derived allele is determined relative to an inferred human ancestor) in the analyzed African populations at single-nucleotide polymorphisms (SNPs) for which a randomly sampled allele from an archaic individual was observed to also be derived. Theory predicts that this conditional site frequency spectrum (CSFS) is expected to be uniformly distributed when alleles are neutrally evolving under a demographic model in which the ancestor of modern and archaic humans, assumed to be at mutation-drift equilibrium, split with no subsequent gene flow between the two groups (*13*, *14*). This expectation is robust to assumptions about changes in population sizes in the history of modern human or archaic populations. Further, we show that this expectation holds even when there is population structure or gene flow in the history of the archaic population (see Materials and Methods).

We computed $\mathrm{CSFS}_{\mathrm{YRI,N}}$: the CSFS in the Yoruba from Ibadan (YRI) while restricting to SNPs where a randomly sampled allele from the high-coverage Vindija Neanderthal (N) genome was observed to be derived (*15*). In contrast to the uniform spectrum expected from theory, we observe that the $\mathrm{CSFS}_{\mathrm{YRI,N}}$ has a U-shape with an elevated proportion of SNPs with low- and high-frequency–derived alleles relative to those at intermediate frequencies (Fig. 1 and fig. S4). The CSFS is nearly identical when we replace the Vindija Neanderthal genome with the high-coverage Denisova genome (Fig. 1 and fig. S4) (*4*). We observed a similar U-shaped CSFS in each of three additional West African populations [Esan in Nigeria (ESN), Gambian in Western Divisions in the Gambia (GWD), and Mende in Sierra Leone (MSL)] included in the 1000 Genomes Phase 3 dataset (fig. S4).

Mutational biases, errors in determining either the ancestral or the archaic allele, or recurrent mutation could produce the observed CSFS. We confirmed that the shape of the $\mathrm{CSFS}_{\mathrm{YRI,N}}$ was robust to the inclusion of only transition mutations, only transversion mutations, to the exclusion of hypermutable CpG sites (fig. S7), as well as when we computed the spectrum on the Yoruba genomes separately sequenced in the 1000 Genomes Phase 1 dataset (fig. S7).

We verified that this signal was robust to changes in recombination rate and background selection by restricting to regions that are likely to be evolving neutrally (by restricting to sites with estimates of background selection, *B* statistic, >800). We also assessed the effect of biased gene conversion by excluding weak-to-strong and strong-to-weak polymorphisms. We found that the U-shaped signal is robust to variation in recombination rate, background selection, and biased gene conversion (fig. S10). Errors in determining the ancestral allele could make low-frequency ancestral alleles appear to be high-frequency–derived alleles and vice versa and thus could potentially lead to a U-shaped CSFS. However, the shape of the CSFS remains qualitatively unchanged when we used either the chimpanzee genome or the consensus across the orangutan and chimpanzee genomes to determine the ancestral allele (fig. S9). We simulated both ancestral allele misidentification and errors in genotype calling in the high-coverage archaic genome. A fit to the data required both a 15% ancestral misidentification rate and a 3% genotyping error rate in the archaic genome, substantially larger than previous estimates of these error rates [1% for ancestral

[1]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA. [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA.
*Corresponding author. Email: sriram@cs.ucla.edu

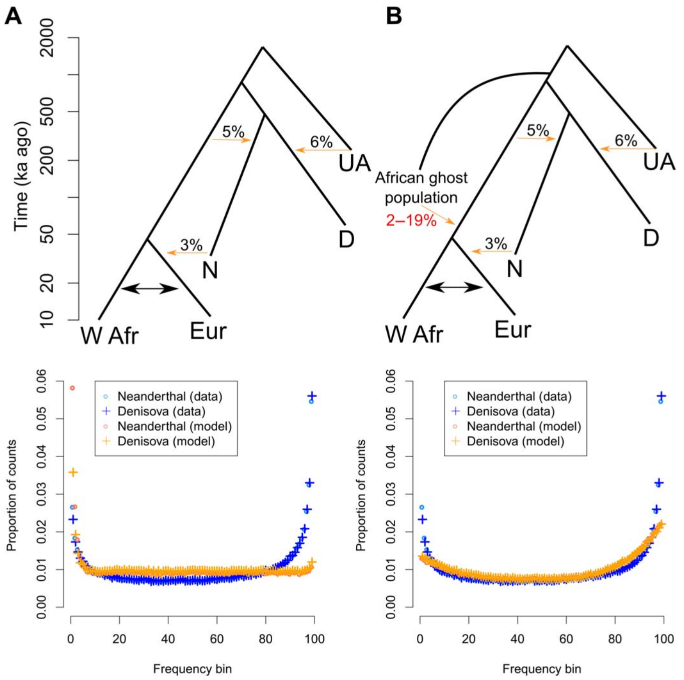

Fig. 1. Demography relating known and proposed archaic lineages to modern human populations. (**A**) Basic demographic model with CSFS fit. W Afr, West Africans; Eur, European; N, Neanderthal; D, Denisovan; UA, unknown archaic [see (*18*)]. Below, we show the CSFS in the West African YRI when restricting to SNPs where a randomly sampled allele from the high-coverage Vindija Neanderthal was observed to be derived [Neanderthal (data)], as well as where a randomly sampled allele from the high-coverage Denisovan genome was observed to be derived [Denisovan (data)]. We also show the CSFS under the proposed model [Neanderthal (model) and Denisova (model)]. Migration between Europe and West Africa introduces an excess of low-frequency variants but does not capture the decrease in intermediate frequency variants and increase in high-frequency variants. (**B**) Newly proposed model involving introgression into the modern human ancestor from an unknown hominin that separated from the human ancestor before the split of modern humans and the ancestors of Neanderthals and Denisovans. Below, we show the CSFS fit from the proposed model, which captures the U-shape observed in the data.

misidentification rate in the Enredo-Pecan-Ortheus (EPO) ancestral sequence (*16*) and 0.6% for the modern human contamination in the Vindija Neanderthal (*15*)] (section S1.1 and fig. S11). To explore the contribution of recurrent mutations, we used forward-in-time simulations that allow for recurrent mutations: The simulated CSFS does not resemble the U-shaped CSFS that we see in data (fig. S43). Together, these results indicate that the U-shaped CSFS observed in the African populations is not an artifact.

To determine whether realistic models of human history can explain the CSFS, we compared the CSFS estimated from coalescent simulations to the observed $\mathrm{CSFS}_{\mathrm{YRI,N}}$ using a goodness-of-fit test (see Materials and Methods and section S2). We augmented a model of the demographic history of present-day Africans (*17*) with a model of the history of Neanderthals and Denisovans inferred by Prüfer *et al.* (*15*) (Fig. 1 and figs. S1 and S16). This model includes key interbreeding events between Neanderthals, Denisovans, and modern human populations such as the introgression from Neanderthals into non-Africans, from early modern humans into Neanderthals (*6*), and into the Denisovans from an unknown archaic population (*18*). The resulting model fails to fit the observed $\mathrm{CSFS}_{\mathrm{YRI,N}}$ [*P* value of a Kolmogorov-Smirnov (KS) test on the residuals being normally distributed $P < 2 \times 10^{-16}$]. Extensions of this model to include realistic variation in mutation and recombination rates along the genome (KS $P < 2 \times 10^{-16}$; fig. S12 and section S1) and low levels of Neanderthal DNA introduced into African populations via migration between Europeans and Africans do not provide an adequate fit (KS $P < 2 \times 10^{-16}$; Fig. 1 and section S1) nor does a model of gene flow between YRI and pygmy populations that has been proposed previously (KS $P < 2 \times 10^{-16}$; fig. S12 and section S1) (*19*). The expectation that the CSFS is uniformly distributed across allele frequencies relies on an assumption of mutation-drift equilibrium in the population ancestral to modern humans, Neanderthals, and Denisovans. We confirmed that violations of this assumption (due to bottlenecks, expansions, and population structure in the ancestral population) were also unable to fit the data (KS $P < 2 \times 10^{-16}$ for all models; section S2, table S3, and fig. S17).

Given that none of the current demographic models are able to fit the observed CSFS, we explored models where present-day West Africans trace part of their ancestry to (A) a population that split from their ancestors after the split between Neanderthals and modern humans, (B) a population that split from the ancestor of Neanderthals after the split between Neanderthals and modern humans, or (C) a population that diverged from the ancestors of modern humans and Neanderthals before the ancestors of Neanderthals and modern humans split from each other (fig. S2 and section S3). Each of these models of admixture (which we refer to as models A, B, and C, respectively) can yield a U-shaped CSFS. The increase in the counts of low derived allele frequency SNPs is largely due to the introduction of the derived allele from the introgressing population at sites that are fixed for the ancestral allele. The increase in the counts of the high-frequency SNPs is largely due to the introduction of the ancestral alleles at sites that are fixed for the derived allele.

A search for the parameters for models A and B that produce the best fit to the CSFS results in a trifurcation, i.e., models in which the introgressing population splits off from the modern human population at the same time as the modern human–Neanderthal. Models A and B fail to fit the observed CSFS even at their most likely parameter estimates (KS $P = 3.3 \times 10^{-15}$ and $P = 5.6 \times 10^{-6}$, respectively; section S3) because of insufficient genetic drift in the African population since the split from the introgressing population (section S4.2). In addition, we show in appendix B that the spectrum for model A is expected to be symmetric, which is not observed in the data (Fig. 1). Model C, on the other hand, is consistent with the data (KS $P = 0.09$), suggesting that part of the ancestry of present-day West Africans must derive from a population that diverged before the split time of Neanderthals and modern humans. In addition to the goodness-of-fit tests, we examined the likelihood of the best-fit parameters for each of the models and found that model C provides a significantly better fit than other models (model C having a higher composite log likelihood than the next best model $\Delta\mathcal{LL} = \mathcal{LL}_{\mathrm{Nextbestmodel}} - \mathcal{LL}_{\mathrm{C}} = -6806$ when we condition on the Vindija Neanderthal genome and $\Delta\mathcal{LL} = -6240$ when we condition on the Denisovan genome; table S4 and Materials and Methods). Our analyses provide support for a contribution to the genetic ancestry of present-day West African populations from an archaic ghost population whose divergence from the ancestors of modern humans predates the split of Neanderthals and modern humans.

We applied approximate Bayesian computation (ABC) to the CSFS to refine the parameters of our most likely demographic model (model C) (section S5). Given the large number of parameters in this demographic model, we fixed parameters that had previously been estimated (*15*) and jointly estimated the split time of the introgressing archaic population from the ancestors of Neanderthals and modern humans, the time of introgression, the fraction of ancestry contributed by the introgressing population, and its effective population size. We determined the posterior mean for the split time to be 625,000 years before the present (B.P.) [95% highest posterior density interval (HPD): 360,000 to 975,000], the admixture time to be 43,000 years B.P. (95% HPD: 6000 to 124,000), and the admixture fraction to be 0.11 (95% HPD: 0.045 to 0.19). Analyses of three other West African populations (ESN, GWD, and MSL) yielded concordant estimates for these parameters (Fig. 2 and table S7). Combining our results across the West African populations, we estimate that the archaic population split from the ancestor of Neanderthals and modern humans 360 thousand years (ka) to 1.02 million years (Ma) B.P. and subsequently introgressed into the ancestors of present-day Africans 0 to 124 ka B.P. contributing 2 to 19% of their ancestry. We caution that the true underlying demographic model is likely to be more complex. To explore aspects of this complexity, we examined the possibility that the archaic population diverged at the same time as the split time of modern humans and Neanderthals and found that this model can also produce a U-shaped CSFS with a likelihood that is relatively high, although lower than that of our best-fit model ($\Delta\mathcal{LL} = -2713$ for the Neanderthal CSFS and $\Delta\mathcal{LL} = -2597$ for the Denisovan CSFS, KS $P \leq 2.9 \times 10^{-6}$). Our estimates of a large

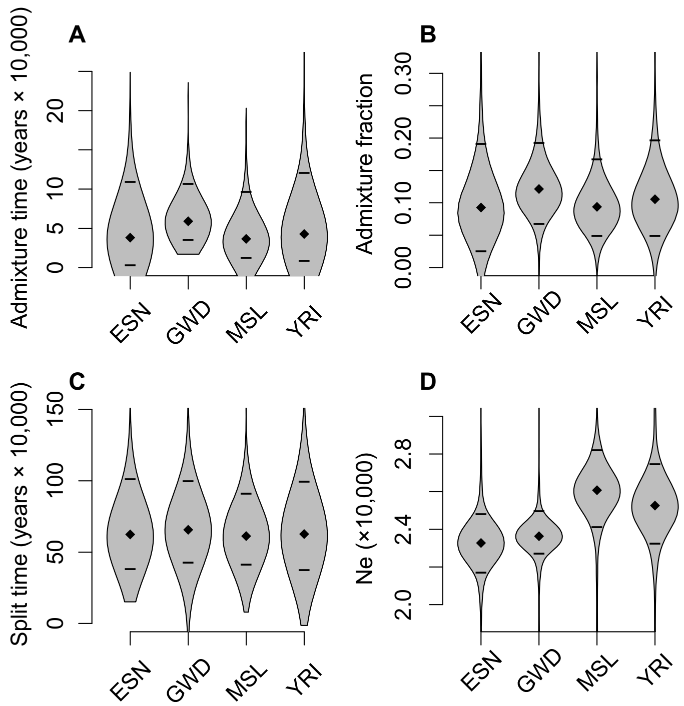

Fig. 2. ABC estimates of the demographic parameters of the archaic ghost population across four West African populations (YRI, ESN, GWD, and MSL). Posterior means are denoted by diamonds, and 95% credible intervals are denoted by lines. (**A**) The admixture time t_a, (**B**) the admixture fraction α, (**C**) the split time of the introgressing population t_s, and (**D**) the effective population size of the introgressing population N_e are shown. The parameter estimates are largely consistent across the African populations: We estimate split times of 360 ka to 1.02 Ma B.P., admixture times of 0 to 124 ka B.P., admixture fractions that range from 0.02 to 0.19, and effective population sizes that range from 22,000 to 28,000.

effective population size in the introgressing lineage (posterior mean of 25,000; 95% HPD: 23,000 to 27,000) could indicate additional structure. We find that the N_e of the introgressing lineage in YRI and MSL is larger than that in the other African populations, possibly due to a differential contribution from a basal West African branch (*20*).

While we have chosen to represent the genetic contribution of the African ghost population as a single discrete interbreeding event, a more realistic model could include low levels of gene flow in a structured population over an extended period of time. Previously proposed models of ancestral structure in Africa do not fit the CSFS [KS $P < 2 \times 10^{-16}$ for the model described in (*21*) and KS $P < 2 \times 10^{-16}$ for the model proposed in (*14*); fig. S18], although we observe that the model of ancestral structure proposed by Yang *et al.* does produce a slight U-shape. We explored additional models of population structure in Africa (*22*) in which a lineage split from the ancestor of the modern humans with split times ranging from 100 to 550 ka B.P. and continued to exchange genes with the modern human population until the present with migration rates ranging from 2.5×10^{-5} to 2×10^{-2} migrants per generation. While these models of continuous gene flow produce a U-shaped CSFS for low migration rates and deep splits, they do not provide an adequate fit to the empirical CSFS over the range of parameters considered (KS $P \leq 2.3 \times 10^{-5}$; section S6 and figs. S14 and S15). We used our ABC framework to explore a more detailed model of continuous migration in which we varied split time, migration rate, and effective population size of the introgressing lineage. Simulations under the best fitting model produce a CSFS that does not adequately fit the data (KS $P = 1.83 \times 10^{-6}$). A possible reason why the continuous migration models that we have explored do not fit the data is that these models can be considered as extensions of model A with multiple admixture events. We have shown that these models can only produce symmetric CSFS, unlike the CSFS that we observe in the data (appendix B). Thus, deep population structure within Africa alone cannot not explain the data (section S6).

Given the uncertainty in our estimates of the time of introgression, we wondered whether jointly analyzing the CSFS from both the CEU (Utah residents with Northern and Western European ancestry) and YRI genomes could provide additional resolution. Under model C, we simulated introgression before and after the split between African and non-African populations and observed qualitative differences between the two models in the high-frequency–derived allele bins of the CSFS in African and non-African populations (fig. S40). Using ABC to jointly fit the high-frequency–derived allele bins of the CSFS in CEU and YRI (defined as greater than 50% frequency), we find that the lower limit on the 95% credible interval of the introgression time is older than the simulated split between CEU and YRI (2800 versus 2155 generations B.P.), indicating that at least part of the archaic lineages seen in the YRI are also shared with the CEU (section S9.2).

We then attempted to understand the fine-scale distribution of archaic ghost ancestry along the genomes of present-day Africans. We used a recently developed statistical method (ArchIE) that combines multiple population genetic statistics to identify segments of diverged ancestry in 50 YRI and 50 MSL genomes without the need for an archaic reference genome (section S7) (*23*). Briefly, the method uses summary statistics computed from present-day genome sequences as input to a logistic regression model to estimate the probability that a haploid segment of an individual genome (defined as a contiguous region of length 50 kilobases) is archaic. While the parameters of the model are estimated by simulating data under a model that closely matches the demographic history relating Neanderthals and non-Africans, we

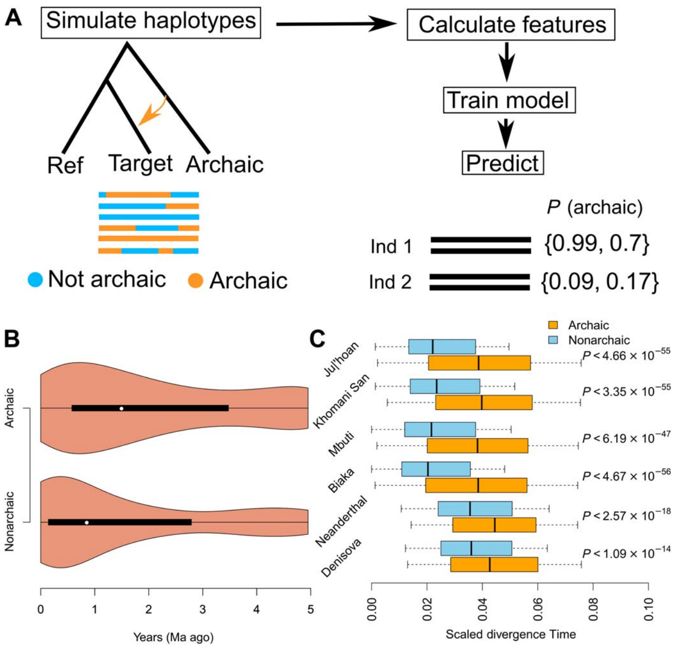

Fig. 3. Analysis of segments of archaic ghost ancestry found in the Yoruba and Mende populations. (**A**) Inference of segments of archaic ancestry was performed with ArchIE. ArchIE proceeds by simulating data under a model of archaic introgression, calculating population genetic summary statistics, and training a model to predict the probability that a 50-kb window in an individual comes from an archaic population. We apply the resulting predictor to genome sequences from the Yoruba and Mende populations. (**B**) Comparison of TMRCA between inferred archaic and nonarchaic segments to the TMRCA of a pair of nonarchaic segments in the Yoruba. On average, archaic segments are 1.69× older than nonarchaic segments. (**C**) Estimates of the divergence times of archaic segments inferred in Yoruba from KhoeSan, Jul'hoan, two modern human pygmy genomes (Mbuti and Biaka), and Neanderthal and Denisovan genomes compared to divergence times of nonarchaic segments. *P* values are computed via block jackknife. Archaic segments are more diverged from all six genomes than nonarchaic segments.

Table 1. Genes harboring a high frequency of archaic segments in the Yoruba and Mende populations. Genes were selected by ranking the union of the set of putative archaic segments by frequency in either the Mende or Yoruba population and selecting the top 10 genes. Genes in bold denote frequencies greater than 50% in the respective population.

Chromosome	Gene name	Frequency (Yoruba)	Frequency (Mende)	Gene type
chr1	RP11-286M16.1	**0.84**	**0.81**	lincRNA
chr4	KCNIP4	**0.73**	**0.69**	Protein coding
chr6	MTFR2	**0.67**	**0.78**	Protein coding
chr8	TRPS1	**0.71**	**0.75**	Protein coding
chr12	RP11-125N22.2	0.12	**0.88**	Pseudogene
chr16	HSD17B2	**0.74**	**0.68**	Protein coding
chr17	NF1	**0.83**	**0.85**	Protein coding
chr17	KRT18P61	**0.84**	0.36	Pseudogene
chr21	MIR125B2	**0.76**	**0.64**	MicroRNA

found that ArchIE has 68% power to detect archaic segments at a false discovery rate of about 7% under our best-fit demographic model, confirming that its inferences are robust and sensitive to archaic introgression in Africa.

On average, ≃6.6 and ≃7.0% of the genome sequences in YRI and MSL were labeled as putatively archaic in ancestry. We sought to test whether the putatively archaic segments identified in YRI and MSL traced their primary ancestry to other African populations (*8*–*10*) or to known archaic hominins such as the Neanderthals or Denisovans. We computed the divergence of these segments to a genome sequence from each of six populations: southern African KhoeSan, Jul'hoan; two Central African pygmy populations (Biaka and Mbuti); and two archaic hominin populations (Neanderthal and Denisovan). We expect segments introgressed from any of these populations to be less diverged relative to nonarchaic segments. On the contrary, the putatively archaic segments are more diverged, consistent with their source not being any of these populations (Fig. 3C and section S7.1). Merging the putatively archaic segments across individual genomes, we obtained a total of 482 and 502 Mb of archaic genome sequence in the YRI and MSL, respectively. We estimated the distribution of the time to the most recent common ancestor (TMRCA) between segments labeled archaic and those labeled nonarchaic using the pairwise mode of multiple sequentially Markovian coalescent (MSMC) (Fig. 3B and section S7.2) (*24*) and observed that the TMRCA is larger for the putatively archaic class of segments. Specifically, we find that the median nonarchaic segment coalescent time is 0.865 Ma ago for both populations, while the median archaic segment coalescent time is 1.51 Ma ago for YRI and 1.15 Ma ago for MSL (1.69- and 1.23-fold increases in age for YRI and MSL, respectively).

We examined the frequencies of archaic segments to investigate whether natural selection could have shaped the distribution of archaic alleles (fig. S40). We found 33 loci with an archaic segment frequency of ≥50% in the YRI (a cutoff chosen to be larger than the 99.9th percentile of introgressed archaic allele frequencies based on a neutral simulation of archaic introgression with parameters related to the time of introgression and admixture fraction chosen conservatively to maximize the drift since introgression; section S7.3 and fig. S40) and 37 loci in the MSL. Some of these genes are at high frequency across both the YRI and MSL, including *NF1*, a tumor suppressor gene (83% in YRI, 85% in MSL), *MTFR2*, a gene involved with mitochondrial aerobic respiration in the testis (67% in YRI, 78% in MSL), *HSD17B2*, a gene involved with hormone regulation (74% in YRI, 68% in MSL), *KCNIP4*, which is a gene involved with potassium channels (73% in YRI, 69% in MSL), and *TRPS1*, a gene associated with trichorhinophalangeal syndrome (71% in YRI, 75% in MSL; Table 1). Three of these genes have been found in previous scans for positive selection in the YRI: *NF1* (*25*, *26*), *KCNIP4* (*27*), and *TRPS1* (*28*). On the other hand, we do not find elevated frequencies at *MUC7*, a gene previously found to harbor signatures of archaic introgression (*29*).

DISCUSSION

Our analyses document introgression in four present-day West African populations from an archaic population that likely diverged before the split of modern humans and the ancestors of Neanderthals and Denisovans. A number of previous studies have found evidence for deeply diverged lineages contributing genetic ancestry to the Pygmy (*8*, *9*) and Yoruba (*7*, *30*) populations. Analyses of ancient African genomes have revealed that stone-age hunter-gatherers from South Africa diverged from other modern-day populations >260,000 years (*31*) B.P. and that present-day West African populations trace part of their ancestry to a basal lineage that diverged before the split of the southern African San (*20*) (although an alternative model consistent with their data includes a complex pattern of isolation by distance between western, eastern, and southern African populations). Placing our results within the context of the complex patterns of deep divergences in the African populations will require the analysis of a diverse set of African populations that include the southern African San populations, as well as the inclusion of ancient African genomes that lack signals of recent admixture that are present in the present-day San populations (*32*).

One interpretation of the recent time of introgression that we document is that archaic forms persisted in Africa until fairly recently (*33*). Alternately, the archaic population could have introgressed earlier into a modern human population, which then subsequently interbred with the ancestors of the populations that we have analyzed here. The models that we have explored here are not mutually exclusive, and it is plausible that the history of African populations includes genetic contributions from multiple divergent populations, as evidenced by the large effective population size associated with the introgressing archaic population. Relatively, recent fossils with archaic features (or combinations of archaic and modern human features) have been found in the fossil record in Africa and the Middle East. While anatomically modern humans appear in the fossil record around 200,000 years ago, fossils with a combination of archaic and modern features can be found across sub-Saharan Africa and the Middle East until as recently as 35,000 years ago (*34*). Examples of these fossils include a cranium from Iwo Eleru (*33*) and human remains from Ishango (*35*) that have been interpreted as being consistent with deep structure and representing a complex history of interaction between modern and archaic hominins in Africa.

The signals of introgression in the West African populations that we have analyzed raise questions regarding the identity of the archaic hominin and its interactions with the modern human populations in Africa. Analysis of the CSFS in the Luhya from Webuye, Kenya (LWK) also reveals signals of archaic introgression, although our interpretation is complicated by recent admixture in the LWK that involves populations related to western Africans and eastern African hunter-gatherers (section S8) (*20*). Non-African populations (Han Chinese in Beijing and Utah residents with northern and western European ancestry) also show analogous patterns in the CSFS, suggesting that a component of archaic ancestry was shared before the split of African and non-African populations. A detailed understanding of archaic introgression and its role in adapting to diverse environmental conditions will require analysis of genomes from extant and ancient genomes across the geographic range of Africa.

MATERIALS AND METHODS

Conditional site frequency spectrum

We define the CSFS, $\mathrm{CSFS}_{\mathrm{YRI,N}}$, as the histogram of the counts of derived alleles in population pop_1 conditional on observing a derived allele in a related outgroup pop_2 (*13*). We define c_k as the number of SNPs at which the derived allele is present on k chromosomes in a sample of n total chromosomes in pop_1, while a single chromosome in the outgroup pop_2 carries a derived allele. $\mathrm{CSFS}_{\mathrm{YRI,N}}$ is the vector of counts c_k for $k \in \{1 \ldots n-1\}$.

Chen *et al.* (*13*) showed that if the ancestor of populations pop_1 and pop_2 is at mutation-drift equilibrium (i.e., the site frequency spectrum in the ancestor is $f(x) \propto \frac{1}{x}$, where $0 < x < 1$ is the derived allele frequency at a polymorphic SNP) and the two populations pop_1 and pop_2 split with no subsequent admixture, then the $\mathrm{CSFS}_{\mathrm{YRI,N}}$ is expected to be uniform, i.e., $\mathrm{CSFS}_{\mathrm{YRI,N}}(k)$ = constant. This result does not depend on any additional aspects of the demographic history of either populations pop_1 or pop_2, except that they are randomly mating. We used the CSFS to study introgression in present-day Africans where we set pop_1 to present-day Africans and pop_2 to an archaic population, i.e., Neanderthal or Denisovan.

One of the complications in applying the CSFS to learn about the history of present-day Africans arises from known departures from a simple model of isolation with no subsequent admixture. However, we considered the possibility of structure in the archaic population. This structure could have several forms that include the ancestral Neanderthal population being structured or it could involve gene flow from early modern humans into Neanderthals (*6*), or as in the case of Denisovans, this could include gene flow from a highly diverged archaic population (*18*). We performed extensive simulations to show that structure in the archaic population continues and also leads to a uniform CSFS (section S1). Further, in appendix A, we show that the CSFS is uniform even if there is structure in the archaic population. However, structure within population the African population (pop_1) since its split from the archaic population (pop_2), e.g., due to admixture, is expected to produce deviations from the uniform CSFS.

Data processing

For our primary analyses of the CSFS, we used the 1000 Genomes Phase 3 dataset (release 20130502) (*36*), the high-coverage Vindija Neanderthal genome (*15*), and the high-coverage Denisovan genome (*4*). We used the annotated ancestral alleles provided by the 1000 Genomes consortium and analyzed only autosomal SNPs. Archaic genotypes (Vindija and Denisovan) come from the pipeline described in (*15*), which used snpAD for SNP calling [see S3 in (*15*)], and required a mapping quality of ≥25 and a mappability filter of 100. We did not apply an additional genotype quality filter for the data presented in fig. S4. However, we tested the sensitivity of the spectrum to the choice of genotype quality filters in the archaic when using a GQ (Genotype Quality) filter of ≥30 and ≥50 and see very little difference in the shape of the spectrum (fig. S8).

In addition, we also computed the CSFS using the chimpanzee genome to polarize the ancestral alleles (fig. S9A) (*37*). We dropped sites in cases where the chimpanzee allele did not match either human allele. As a further check, we also repeated the analysis restricting only to sites where the chimpanzee and orangutan genomes have matching alleles (*38*). These results are reported in fig. S9B. Last, we repeated our analysis filtering out CpG hypermutable sites using the CpG annotations from (*18*).

CSFS from the 1000 Genomes data

We computed $\mathrm{CSFS}_{\mathrm{YRI,N}}$ where pop_1 is a modern human population and pop_2 is an archaic population. Specifically, we chose pop_1, in turn, to be the Yoruba from Nigeria (YRI), MSL, ESN, and GWD, while we chose pop_2 to be either the high-coverage Vindija Neanderthal or the high-coverage Denisovan genome (fig. S4).

We computed the CSFS from the 1000 Genomes phase 3 data (*36*) for each of the four African populations mentioned above (fig. S4), as well as for the CEPH CEU and Han Chinese from Beijing (CHB) (fig. S6).

For all populations, we observed a U-shaped spectrum with an excess of derived alleles at low and high frequencies. In the African populations, we observed that the CSFS from conditioning on the Denisovan is nearly identical to the Vindija Neanderthal except at the lowest-frequency bins, where there is an excess of counts for the Neanderthal CSFS. We interpreted this difference as suggestive of low levels of Neanderthal-related ancestry in these populations consistent with previous studies (*18*). In CEU and CHB, we also observed a U-shaped spectrum for both the Vindija Neanderthal and Denisovan, but with a more pronounced difference between the Neanderthal and Denisovan spectra, i.e., an excess of counts in the low-frequency–derived sites when conditioned on the Vindija Neanderthal relative to the Denisovan. This difference is likely reflective of the Neanderthal introgression event experience by populations outside of Africa around 50,000 years ago (*21*, *39*). Section S8 explores the implication of observing a U-shaped CSFS in African and non-African populations.

To determine the robustness of the shape of the CSFS, we recomputed the CSFS in YRI using only transitions, transversions, and after removing CpG sites. We found very similar U-shaped CSFS across these mutation classes (fig. S7). In addition, we checked whether biased gene conversion could cause this signal by removing weak-to-strong and strong-to-weak polymorphisms. We found that the shape of the CSFS remains without these mutations (fig. S10A). Last, we checked whether the shape of the CSFS was driven by selection or low recombination rates. We used B values from (*40*), which estimate how much background selection has reduced diversity. We restricted to regions of the genome in the top quintile of B values (that is, the top one-fifth of neutral sites; $B \geq 800$) and recomputed the spectrum using YRI individuals. We found that the shape remains the same after this filtering (fig. S10B).

Model comparison

We used coalescent simulations to assess whether a demographic model produces a CSFS that matches the empirical CSFS. To assess the fit of a given demographic model $\mathcal{M}$ to the data, we compared the CSFS computed on the data simulated under $\mathcal{M}$ to that computed on the empirical data. We considered a model in which the empirical CSFS was obtained by sampling from the CSFS computed on the simulated data. For these fits, we modeled the proportion of SNPs that contain a given number k of derived alleles rather than the number of SNPs. To assess the fit of the simulated CSFS under $\mathcal{M}$ ($\boldsymbol{S}_{\mathcal{M}}$) to the observed CSFS ($\boldsymbol{O}$), we used a multinomial composite likelihood

$$L(\mathcal{M}) = P(\boldsymbol{O} \mid \boldsymbol{S}_{\mathcal{M}}) = \prod_{k=1}^{n-1} \left(\frac{S_k}{\sum_k S_k} \right)^{O_k}$$

Here, k indexes the derived allele count, S_k denotes the number of SNPs with k-derived alleles observed in the simulated CSFS, while O_k denotes the number of SNPs with k-derived alleles observed in the empirical CSFS. We caution that L is a composite likelihood that ignores the dependence among SNPs so that comparisons of L must be interpreted with caution. In the results presented here, we reported the log likelihood ($\mathcal{LL}$).

Goodness of fit

We defined a goodness-of-fit statistic that we used to assess whether the CSFS computed under a demographic model explains the major patterns of the empirical CSFS. The goodness-of-fit statistic was defined from the residuals obtained by trying to fit the simulated CSFS to the empirical CSFS. We assumed that the counts of SNPs in each derived allele frequency bin of the empirical CSFS follow a binomial distribution with a mean given by the proportion of SNPs that have the same derived allele frequency in the simulated CSFS. One complication is that the counts across bins of derived allele frequencies are not independent because of linkage disequilibrium. To account for this complication, we attempted to estimate the effective number of independent observations in the observed CSFS (rather than assume that each SNP is an independent observation). We define the residual for bin k as

$$r_k = \sqrt{m_{\text{eff}}} \frac{o_k - s_k}{\sqrt{s_k(1 - s_k)}}$$

Here, m_{eff} is the effective number of independent SNPs, o_k represents the proportion of SNPs with derived allele count k in the empirical CSFS, s_k is the proportion of SNPs with derived allele count k in the simulated CSFS, and k indexes the count of derived allele. These residuals are expected to be approximately normally distributed when the number of observations is large (as is the case with the CSFS where each bin has >1000 observations). m_{eff} is a scaling factor to ensure that the residuals are standardized.

To calculate m_{eff}, we used two replicate whole-genome simulations (3 GB) under the same demographic model and set one as the observed data and one as the simulation. We divided the number of bins n by the sum of the squared residuals

$$m_{\text{eff}} = \frac{n}{\sum_{k=1}^{n} \left(\frac{o_k - s_k}{\sqrt{s_k(1 - s_k)}} \right)^2}$$

A good fit will result in approximately normally distributed residuals, while poor fits will deviate significantly from a normal distribution. To obtain a formal test of fit, we used a KS test comparing the distribution of the residuals to a normal distribution. P values that reject the null hypothesis suggest that the model is a poor fit to the data. We used bins of allele counts ranging from 11 to 90, excluding the lowest- and highest-frequency bins as the counts from these bins are more likely to be affected by unmodeled genotyping errors, leading to false rejections of the null hypothesis. To assess the fit of a class of models (e.g., models A, B, and C), we report the P value of the model with parameter estimates obtained via ABC (sections S3.1 to S3.6).

Last, we expanded the range of derived allele counts in our goodness-of-fit computation from [11, 90] to [6, 95] (table S8). While none of the models fit adequately, model C has substantially higher P values than the other models, indicating that it continues to explain the CSFS better across this range of allele counts. The lack of fit across the expanded range of derived allele counts is likely due to unmodeled complexities in the underlying demographic history, as well as error processes that affect the low- and high-frequency SNPs.

Model fitting

We used ABC to fit a demographic model to the CSFS of each African population using the R package abc (*41*). Using a model relating African and non-African populations with the Neanderthal and Denisovan lineages as a base, we fit the split time, admixture time, admixture fraction, and effective population size of an introgressing lineage (section S5.2). We drew values for each of the parameters from a previous distribution, simulated 300 Mb using ms (*42*), and computed the CSFS for the resulting simulation. We repeated this procedure 75,000 times. We used the "neuralnet" setting in the R package abc to compute posterior distributions over each of the four parameters with a tolerance of 0.005. For the admixture time and split time, we report the posterior distributions in units of years by convolving the posterior generation time with a uniform distribution over [25, 33] to incorporate uncertainty in the generation time.

Local ancestry inference

We used ArchIE (*23*) to infer the segments of the genomes in 50 YRI and 50 MSL individuals who likely trace their ancestry to an archaic population. We trained ArchIE on a model where an archaic population splits 12,000 generations B.P. and introgressed 2000 generations B.P. at a 2% admixture fraction (section S7). We computed the coalescent time for segments we classified as archaic and segments we classified as nonarchaic using the posterior decoding from MSMC using a representative individual from both YRI and MSL (*24*). We also computed the scaled divergence time between archaic and nonarchaic segments with test genomes from hunter-gatherer populations, Central African Pygmy populations, and archaic populations. This scaled divergence was computed as the number of mutations specific to the segment subtracted from the number of mutations shared between the segment and the test genome. We divided this number by the number of segregating sites in the segment to normalize by the local mutation rate.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/7/eaax5097/DC1

Section S1. Current demographic models cannot explain the CSFS
Section S2. The CSFS cannot be explained by departures from panmixia in the ancestor of archaics and modern humans
Section S3. Exploration of models of introgression into the ancestors of present-day Africans
Section S4. Parameter exploration of model A
Section S5. Estimating parameters for the best-fit model of archaic introgression
Section S6. Continuous migration versus a single pulse
Section S7. Local ancestry inference
Section S8. Extended discussion
Section S9. ms command lines
Fig. S1. Demographic model from Prüfer *et al.* (*15*) (see section S1 for details).
Fig. S2. Demographic model topologies for introgression into the ancestors of present-day Africans simulations in figs. S20, S22, S24, S26, S28, and S30.
Fig. S3. Demographic model topologies for mathematical results.
Fig. S4. CSFS from 1000 Genomes Phase 3 data across all African populations included in the dataset.
Fig. S5. CSFS from 1000 Genomes Phase 3 data in the Luhya population.
Fig. S6. CSFS from 1000 Genomes Phase 3 data in the CEU and CHB.
Fig. S7. Robustness of CSFS in YRI across mutation types and the Phase 1 1000 Genomes dataset.
Fig. S8. Robustness of CSFS in YRI to genotype quality thresholds in archaic genomes.
Fig. S9. CSFS in YRI when using alternate sources for the ancestral allele.
Fig. S10. CSFS in YRI when controlling for biased gene conversion and background selection.
Fig. S11. Simulations of the baseline model (section S1) with both ancestral misidentification (e1) and genotyping error in the archaic (e2).
Fig. S12. Mutation rate and recombination rate variation.
Fig. S13. Simulations of the demographic model inferred from Hsieh *et al.* (*19*) relating the Yoruba, Baka, and Biaka populations.
Fig. S14. Simulations of a demographic model with structure and gene flow in Africa.
Fig. S15. Models with continuous migration (*m* in units of migrants per generation) since the introgressing lineages lineage splits.
Fig. S16. Current demographic models from the literature cannot explain the shape of the CSFS observed in fig. S4.
Fig. S17. Models involving structure in the ancestor of modern humans and archaics cannot explain the observed CSFS.
Fig. S18. Models involving ancestral structure from the literature cannot explain the observed CSFS.
Fig. S19. Model A.1: Gene flow from the modern human ancestor branch back into the modern human ancestor before the out-of-Africa event.
Fig. S20. Model sA.1: Simplified model of gene flow from the modern human ancestor branch back into the modern human ancestor before the out-of-Africa event.
Fig. S21. Model A.2: Gene flow from the modern human ancestor branch into the African branch after the out of Africa event.
Fig. S22. Model sA.2: Simplified model of gene flow from the modern human ancestor branch into the African branch after the out of Africa event.
Fig. S23. Model B.1: Gene flow from the archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S24. Model sB.1: Simplified model of gene flow from the archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S25. Model B.2: Gene flow from the archaic branch into the African branch after the out-of-Africa event.
Fig. S26. Model sB.2: Simplified model of gene flow from the archaic branch into the African branch after the out-of-Africa event.
Fig. S27. Model C.1: Gene flow from an unknown archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S28. Model sC.1: Simplified model of gene flow from an unknown archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S29. Model C.2: Gene flow from an unknown archaic branch into the African branch after the out-of-Africa event.
Fig. S30. Model sC.2: Simplified model of gene flow from an unknown archaic branch into the African branch after the out-of-Africa event.
Fig. S31. Simulations of the best-fitting parameters for models A, B, C (section S3).
Fig. S32. Model A.2 with a population size of 0:01 N_a in the introgressing population.
Fig. S33. Model A.2 with a population size of 1×10^{-4} N_a in the introgressing population.
Fig. S34. Model A.2 with a population size of 1×10^{-4} N_a in the introgressing population and migration between CEU and YRI over the last 20 ka B.P.
Fig. S35. Model A.2 with a population size of 1×10^{-5} N_a in the introgressing population, which branches off 200 ka B.P.
Fig. S36. Model A.2 where the introgressing population splits at the same time as the archaic population (550 ka B.P.) with a population size of 0.01 N_a.
Fig. S37. Model A.2 where the introgressing population splits at the same time as the archaic population, 765 ka B.P.
Fig. S38. Parameter estimates using ABC for model A.1 including ancestral misidentification (e1) and genotyping error in the archaic (e2).
Fig. S39. Parameter estimates using ABC for model A.2 including ancestral misidentification (e1) and genotyping error in the archaic (e2).
Fig. S40. Marginalized joint CSFS of YRI and CEU from simulations.
Fig. S41. Distribution of allele frequencies for neutral archaic SNPs from model C with 13% introgression and an introgression time of 42 ka B.P.
Fig. S42. Archaic segment frequency map for MSL and YRI.
Fig. S43. CSFS from the baseline model allowing for recurrent mutations.
Table S1. Description of the models examined in this work.
Table S2. We simulated data from the Prüfer *et al.* (*15*) model and added in ancestral misidentification error and genotyping error in the archaic.
Table S3. Model fits for null models including structure and departures from panmixia in the Modern Human (MH) ancestor.
Table S4. Model fits for alternate models including admixture from other lineages.
Table S5. Model fits for alternate models using a simplified demography.
Table S6. Model fits for variations of model A.
Table S7. Best-fitting parameter values for all populations using ABC.
Table S8. *P* values of a test of goodness of fit for the best-fitting parameters for each class of demographic models.
Appendix A. The CSFS is uniform under structure in the archaic population.
Appendix B. The CSFS is symmetric under model A.
References (*43–55*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky,

S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

2. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
3. D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
4. M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic denisovan individual. *Science* **338**, 222–226 (2012).
5. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell* **173**, 53–61.e9 (2018).
6. M. Kuhlwilm, I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, M. Kircher, Q. Fu, H. A. Burbano, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, T. Marques-Bonet, A. M. Andrés, B. Viola, S. Pääbo, M. Meyer, A. Siepel, S. Castellano, Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
7. V. Plagnol, J. D. Wall, Possible ancestral structure in human populations. *PLOS Genet.* **2**, e105 (2006).
8. M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, J. D. Wall, Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15123–15128 (2011).
9. J. Lachance, B. Vernot, C. C. Elbers, B. Ferwerda, A. Froment, J.-M. Bodo, G. Lema, W. Fu, T. B. Nyambo, T. R. Rebbeck, K. Zhang, J. M. Akey, S. A. Tishkoff, Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).
10. P. H. Hsieh, A. E. Woerner, J. D. Wall, J. Lachance, S. A. Tishkoff, R. N. Gutenkunst, M. F. Hammer, Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Res.* **26**, 279–290 (2016).
11. J. Hey, Y. Chung, A. Sethuraman, J. Lachance, S. Tishkoff, V. C. Sousa, Y. Wang, Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* **35**, 2805–2818 (2018).
12. A. P. Ragsdale, S. Gravel, Models of archaic admixture and recent history from two-locus statistics. *PLOS Genet.* **15**, e1008204 (2019).
13. H. Chen, R. E. Green, S. Pääbo, M. Slatkin, The joint allele-frequency spectrum in closely related species. *Genetics* **177**, 387–398 (2007).
14. M. A. Yang, A.-S. Malaspinas, E. Y. Durand, M. Slatkin, Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol. Biol. Evol.* **29**, 2987–2995 (2012).
15. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, eaao1887 (2017).
16. B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes, E. Birney, Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
17. S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs; The 1000 Genomes Project, C. D. Bustamante, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983–11988 (2011).
18. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
19. P. H. Hsieh, K. R. Veeramah, J. Lachance, S. A. Tishkoff, J. D. Wall, M. F. Hammer, R. N. Gutenkunst, Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res.* 10.1101/gr.192971.115, (2016).
20. P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, M. Hajdinjak, T. Salie, N. Rohland, S. Mallick, A. Peltzer, A. Heinze, I. Olalde, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. I. Cerezo-Román, C. Chiumia, A. Crowther, E. Gomani-Chindebvu, A. O. Gidna, K. M. Grillo, I. T. Helenius, G. Hellenthal, R. Helm, M. Horton, S. López, A. Z. P. Mabulla, J. Parkington, C. Shipton, M. G. Thomas, R. Tibesasa, M. Welling, V. M. Hayes, D. J. Kennett, R. Ramesar, M. Meyer, S. Pääbo, N. Patterson, A. G. Morris, N. Boivin, R. Pinhasi, J. Krause, D. Reich, Reconstructing prehistoric african population structure. *Cell* **171**, 59–71.e21 (2017).
21. S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between neandertals and modern humans. *PLOS Genet.* **8**, e1002947 (2012).
22. B. M. Henn, T. E. Steele, T. D. Weaver, Clarifying distinct models of modern human origins in Africa. *Curr. Opin. Genet. Dev.* **53**, 148–156 (2018).
23. A. Durvasula, S. Sankararaman, A statistical model for reference-free inference of archaic local ancestry. *PLOS Genet.* **15**, e1008175 (2019).
24. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
25. S. Kudaravalli, J.-B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, J. K. Pritchard, Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol. Evol.* **26**, 649–658 (2009).
26. S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili, R. A. Adegbola, R. N. K. Bamezai, A. V. S. Hill, F. O. Vannberg, J. L. Rinn; 1000 Genomes Project, E. S. Lander, S. F. Schaffner, P. C. Sabeti, Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013).
27. International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
28. L. B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40**, 340–345 (2008).
29. D. Xu, P. Pavlidis, R. O. Taskent, N. Alachiotis, C. Flanagan, M. DeGiorgio, R. Blekhman, S. Ruhl, O. Gokcumen, Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. *Mol. Biol. Evol.* **34**, 2704–2715 (2017).

30. J. D. Wall, K. E. Lohmueller, V. Plagnol, Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823–1827 (2009).
31. C. M. Schlebusch, H. Malmström, T. Günther, P. Sjödin, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, M. Jakobsson, Southern african ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).
32. J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, D. Reich, Ancient west eurasian ancestry in southern and eastern africa. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2632–2637 (2014).
33. K. Harvati, C. Stringer, R. Grün, M. Aubert, P. Allsworth-Jones, C. A. Folorunso, The later stone age calvaria from Iwo Eleru, Nigeria: Morphology and chronology. *PLOS ONE* **6**, e24024 (2011).
34. G. P. Rightmire, Middle and later pleistocene hominins in Africa and Southwest Asia. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16046–16050 (2009).
35. I. Crevecoeur, A. Brooks, I. Ribot, E. Cornelissen, P. Semal, Late stone age human remains from ishango (democratic republic of congo): New insights on late pleistocene modern human diversity in africa. *J. Hum. Evol.* **96**, 35–57 (2016).
36. The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
38. D. P. Locke, L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darré, D. Farré, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. D. Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. A. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordoñez, C. López-Otín, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. RoyChoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs, R. K. Wilson, Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).
39. Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, S. Pääbo, An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
40. G. McVicker, D. Gordon, C. Davis, P. Green, Widespread genomic signatures of natural selection in hominid evolution. *PLOS Genet.* **5**, e1000471 (2009).
41. K. Csilléry, O. François, M. G. B. Blum, abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
42. R. R. Hudson, Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
43. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
44. K. Harris, R. Nielsen, Inferring demographic history from a spectrum of shared haplotype lengths. *PLOS Genet.* **9**, e1003521 (2013).
45. M. Petr, S. Pääbo, J. Kelso, B. Vernot, Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U.S.A.*, 1639–1644 (2019).
46. K. Prüfer, snpAD: An ancient DNA genotype caller. *Bioinformatics* **34**, 4165–4171 (2018).
47. B. C. Haller, P. W. Messer, SLiM 2: Flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).
48. J. N. Fenner, Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
49. H. R. Kunsch, The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17**, 1217–1241 (1989).
50. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
51. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. C. Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
52. J. D. Wall, M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, L. S. Stevison, C. Gignoux, A. Woerner, M. F. Hammer, M. Slatkin, Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013).
53. L. Skov, R. Hui, V. Shchur, A. Hobolth, A. Scally, M. H. Schierup, R. Durbin, Detecting archaic introgression using an unadmixed outgroup. *PLOS Genet.* **14**, e1007641 (2018).
54. M. Kimura, Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. U.S.A.* **41**, 144–150 (1955).
55. R. C. Griffiths, The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Popul. Biol.* **64**, 241–251 (2003).

Acknowledgments: We thank K. Lohmueller, N. Patterson, M. Lipson, M. Schumer, P. Moorjani, T. V. Kent, P. Skoglund, and members of the Sankararaman and Lohmueller labs for helpful comments and discussions. **Funding:** A.D. is supported by the NSF Graduate Research Fellowship DGE-1650604, and S.S. is supported, in part, by NIH grants R00GM111744 and R35GM125055, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation. **Author contributions:** A.D. and S.S. designed and carried out the study. A.D. and S.S. wrote the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Sequencing data are available from the 1000 Genomes project website www.internationalgenome.org/data. Local ancestry calls are available at https://sriramlab.cass.idre.ucla.edu/public/. Additional data related to this paper may be requested from the authors.

Submitted 29 March 2019
Accepted 3 December 2019
Published 12 February 2020
10.1126/sciadv.aax5097

Citation: A. Durvasula, S. Sankararaman, Recovering signals of ghost archaic introgression in African populations. *Sci. Adv.* **6**, eaax5097 (2020).

Science Advances

Recovering signals of ghost archaic introgression in African populations

Arun Durvasula and Sriram Sankararaman

Sci Adv **6** (7), eaax5097.
DOI: 10.1126/sciadv.aax5097

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.