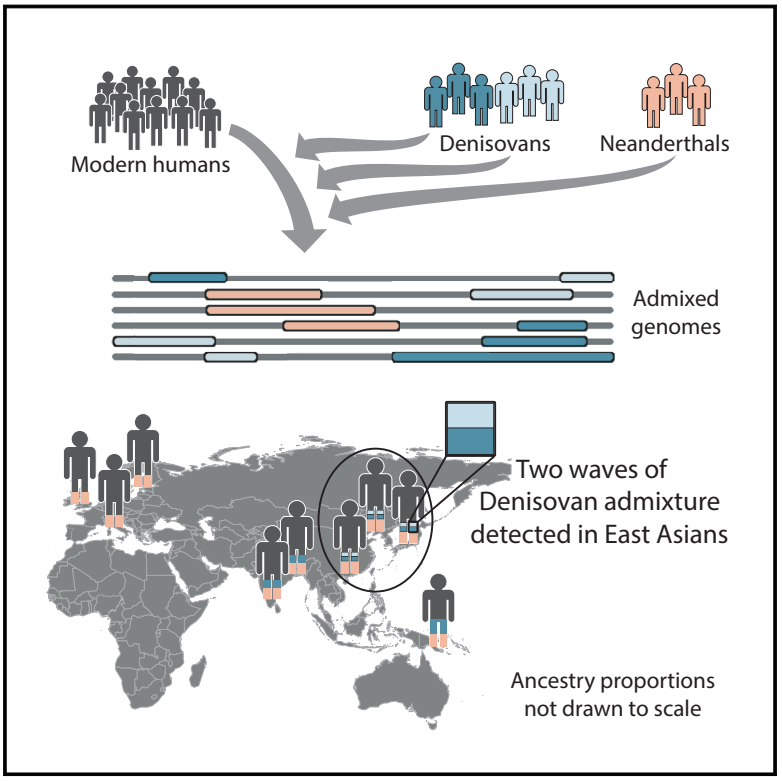
Article

Cell

Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture

Graphical Abstract



Authors

Sharon R. Browning, Brian L. Browning, Ying Zhou, Serena Tucci, Joshua M. Akey

Correspondence

sguy@uw.edu

In Brief

Two waves of Denisovan ancestry have shaped present-day humans.

Highlights

- Asian genomes carry introgressed DNA from Denisovans and Neanderthals
- East Asians show evidence of introgression from two distinct Denisovan populations
- South Asians and Oceanians carry introgression from one Denisovan population

Browning et al., 2018, Cell *173*, 53–61
March 22, 2018
https://doi.org/10.1016/j.cell.2018.02.031

CellPress

Cell

Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture

Sharon R. Browning,[1,5,*] Brian L. Browning,[2] Ying Zhou,[1] Serena Tucci,[3] and Joshua M. Akey[3,4]
[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
[2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA
[3]Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
[4]The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[5]Lead Contact
*Correspondence: sguy@uw.edu
https://doi.org/10.1016/j.cell.2018.02.031

SUMMARY

Anatomically modern humans interbred with Neanderthals and with a related archaic population known as Denisovans. Genomes of several Neanderthals and one Denisovan have been sequenced, and these reference genomes have been used to detect introgressed genetic material in present-day human genomes. Segments of introgression also can be detected without use of reference genomes, and doing so can be advantageous for finding introgressed segments that are less closely related to the sequenced archaic genomes. We apply a new reference-free method for detecting archaic introgression to 5,639 whole-genome sequences from Eurasia and Oceania. We find Denisovan ancestry in populations from East and South Asia and Papuans. Denisovan ancestry comprises two components with differing similarity to the sequenced Altai Denisovan individual. This indicates that at least two distinct instances of Denisovan admixture into modern humans occurred, involving Denisovan populations that had different levels of relatedness to the sequenced Altai Denisovan.

INTRODUCTION

Sequencing the Neanderthal genome (Green et al., 2010; Prüfer et al., 2014), the Denisovan genome (Reich et al., 2010), and several early modern human genomes from Eurasia (Fu et al., 2014, 2015) has confirmed that archaic hominins left their mark in the genomes of modern humans (Plagnol and Wall, 2006; Sankararaman et al., 2014; Vernot and Akey, 2014; Vernot et al., 2016). Present-day individuals in Eurasia inherited ~2% of their genome from Neanderthals (Green et al., 2010), and individuals from Oceania inherited ~5% of their genome from Denisovans (Reich et al., 2010). Suggestive evidence indicates that admixture from other unidentified hominin species occurred in Africa (Hammer et al., 2011; Hsieh et al., 2016; Lachance et al., 2012; Plagnol and Wall, 2006; Wall et al., 2009).

To understand the functional, phenotypic, and evolutionary consequences of archaic admixture, it is necessary to identify the specific haplotypes and alleles that were inherited from archaic hominin ancestors (Huerta-Sánchez et al., 2014; Juric et al., 2016; Sankararaman et al., 2014; Simonti et al., 2016; Vernot and Akey, 2014). Approaches to identifying introgressed haplotypes include methods that specifically incorporate reference archaic hominin genome sequences and reference-free methods that do not utilize such information. An example of the former category is the method of Sankararaman et al. (2014), which identifies archaic haplotypes by comparing modern human haplotypes to a reference archaic sequence. The latter category of methods include the S* statistic (Plagnol and Wall, 2006), which searches for the mutational signature that ancient admixture leaves in the genomes of present-day humans.

The S* approach is powerful for finding introgressed haplotypes in the absence of an archaic reference genome because it leverages the unusual mutational characteristics of introgressed haplotypes. Because of the long divergence time between Neanderthals and modern humans, Neanderthals carry many alleles that are specific to their lineage. Such alleles are present on introgressed haplotypes but are absent or rare in African genomes. Further, based on the recent timing of admixture, introgressed haplotypes are expected to be maintained without recombination over distances of approximately 50 kb on average (Sankararaman et al., 2012), resulting in high levels of linkage disequilibrium (LD) between Neanderthal-specific alleles in non-African human genomes.

In this study, we develop an S*-like method that has increased power and is suitable for large-scale genome-wide data. We apply the method to large sets of sequenced data from Eurasia and Oceania and identify putative archaic-specific alleles. We examine the rate at which these alleles match the sequenced archaic genomes and the role of the genes containing these alleles, to obtain insights into the history of the admixture events and their impact on modern human genomes.

RESULTS

Simulation Results

We first verify the accuracy of our method and compare it with two previous versions of S* designed for windowed analysis of

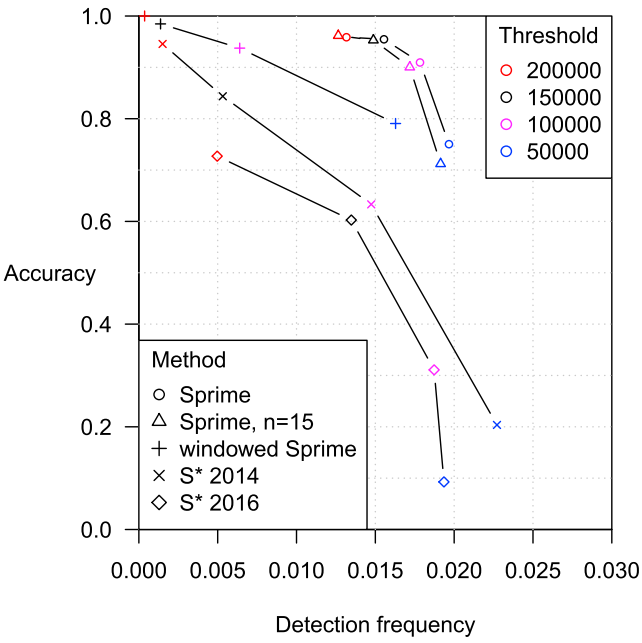


Figure 1. Comparison of Our Method (Sprime) with a Previous Method (S*) on Simulated Data

Detection frequency is the proportion of haplotypes with detected introgression after removing false-positive results and has a maximum possible value of 0.03 (the simulated admixture proportion). Accuracy is the proportion of the putative introgressed alleles that are truly introgressed. We use data simulated with a constant recombination rate ($\rho = 10^{-8}$) and constant mutation rate ($\mu = 1.2 \times 10^{-8}$) so that the application of a score threshold is equivalent to application of a p value threshold. The full simulation model can be found in the STAR Methods. We show results for a range of score thresholds for each method. The default threshold for our method is 150,000 (black data points). Analyses were applied to 100 simulated target individuals and 100 outgroup (African) individuals, except as otherwise noted. Simulated regions were 10 Mb long. Analyses with S* used a sliding window of 50 kb, moving by 10 kb each step. The standard analysis with our method analyzes the full region without a sliding window, but we also applied it with the 50-kb sliding window for comparison. The 2014 variant of S* analyzes target individuals in subsets of 20 and uses only 13 of the outgroup individuals. For comparison we also show results for Sprime with 15 target individuals. The 2016 variant of S* analyzes target individuals one by one (utilizing only the two haplotypes within an individual to determine LD) and uses all 100 outgroup individuals.
See also Figures S1 and S2.

genome-wide data (Vernot and Akey, 2014; Vernot et al., 2016). The main differences between these two versions of S* is in the number of target individuals analyzed simultaneously. The 2014 version of S* analyzes subsets of 20 individuals (Vernot and Akey, 2014), similar to the original gene-based S* (Plagnol and Wall, 2006), whereas the 2016 version analyzes one individual at a time; doing so avoids potential effects of population structure (Vernot et al., 2016). For both methods, we use a sliding window of 50 kb, with a step size of 10 kb. Previous analyses using these methods also used sliding windows of size 50 kb and step sizes of 10 or 20 kb (Vernot and Akey, 2014; Vernot et al., 2016). We see that our method (Sprime) has a much better trade-off of detection frequency to accuracy than these previous versions (Figure 1). Several factors contribute to Sprime's superior performance. One factor is that our method avoids windowing, which allows it to build up power across larger regions of tiled introgression (Figure 2). When we apply our method in 50-kb windows, its performance drops considerably but still remains higher than that of the previous S* versions. Another factor is our simultaneous analysis of a larger number of individuals, although the difference in performance between analyzing 100 individuals and analyzing 15 individuals with our method is not very large. Other likely contributing factors for the difference in performance include our method's allowance for migration from introgressed populations to the simulated African outgroup, and our method's different scoring function.

We verified that our method is robust to variability in mutation rate, allele frequency, and demographic history. Accuracy remains high across a range of mutation rates, provided the sample size is at least 15 (Figure S1). The accuracy of reported putative archaic-specific alleles is over 93% across the range of allele frequencies, with highest accuracy (over 98%) for the lower frequency alleles (frequency < 0.02). Detection frequency varies with mutation rate. With a constant mutation rate of 1.2×10^{-8} per base pair per meiosis, approximately half of the introgressed material is detected. The remaining introgressed material cannot be confidently identified because the introgressed segments are too short and/or the local mutation rate is too low. Across a wide range of demographic histories, with differing archaic-human split times and admixture times, accuracy remains at least 93% (Figure S2).

Detection of Putative Archaic Introgression in Human Populations

We analyzed each non-African population from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015) (Table 1). Across the 19 European, Asian, and American populations, we find 1.36 Gb of the genome covered by putative introgressed segments. In individual populations, coverage ranges from 382 Mb in Peruvians (PEL) to 655 Mb in Bengalis (BEB), and the average proportion of haplotypes carrying a detected segment at a position ranges from 0.80% in Puerto Ricans (PUR) to 1.23% in Han Chinese (CHB and CHS) (Figure 3). These detection rates are around half of the estimated introgression proportions obtained using f_4-ratio statistics (Prüfer et al., 2017). This is in line with the simulation results, in which half the introgressed material can be detected, whereas the other introgressed segments are too short for confident detection. East Asian populations have higher introgression detection rates than do European populations, consistent with previous reports of higher Neanderthal introgression rates in East Asians than in Europeans (Meyer et al., 2012; Sankararaman et al., 2014; Vernot and Akey, 2014; Wall et al., 2013). The South Asian populations and European populations have similar rates of detected introgression, a fact that has been previously reported (Vernot et al., 2016).

In the UK10K study (UK10K Consortium et al., 2015), we find 304 Mb of the genome covered by one or more detected segments, and the average proportion of haplotypes carrying a detected segment at a position is 0.63%. This is lower than in the 1000 Genomes European populations. The lower detection rate in the UK10K may reflect characteristics of the methods used to generate this dataset.

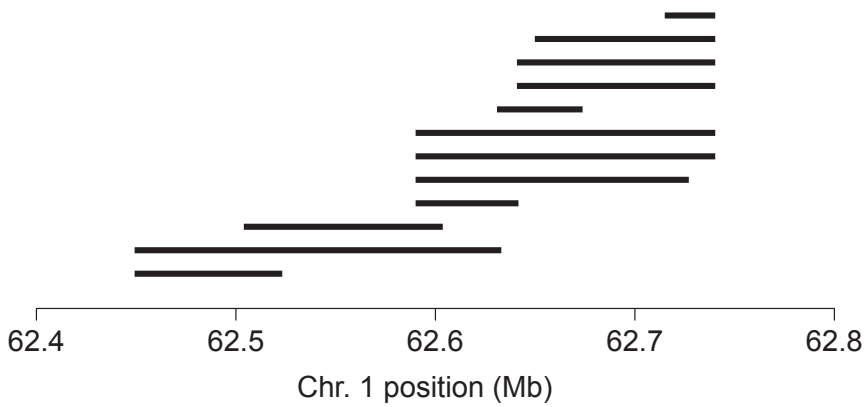


Figure 2. Tiling of Introgressed Haplotypes
This example is from the analysis of Utah residents (CEU) from the 1000 Genomes Project. A single inferred segment (i.e., tiling of overlapping putatively introgressed haplotypes) is shown. This segment includes 141 putatively archaic-specific variants, with a match rate of 94% to the Altai Neanderthal genome. Each horizontal line represents a introgressed haplotype, with the line running from the start of inferred introgression to the end of inferred introgression. Many haplotypes share the same start or end point because of inheritance of that part of the segment from a common ancestor.

Papuans, in addition to Neanderthal ancestry, harbor significant amounts of Denisovan ancestry. In the Papuans from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016), we find 239 Mb of the genome covered by one or more detected segments, and the average proportion of haplotypes carrying a detected segment at a position is 1.48%.

In the 1000 Genomes Eurasian populations, the detected putative introgressed haplotypes have median lengths ranging from 59 kb in Bengalis (BEB) to 71 kb in Finns (FIN). Due to tiling across individuals, the full segments that our method reports can be much longer (Figure 2). In the Eurasian 1000 Genomes populations, the median segment length varies from 205 kb in Iberians (IBS) to 239 kb in Telugus (ITU). The longest detected segment is 7.9 Mb.

Comparison to Sequenced Archaic Genomes

Our method infers putative archaic-specific alleles. If an archaic reference sequence exists, we can determine the proportion of putative archaic alleles that match the reference sequence. Some putative archaic-specific alleles cannot be compared to an archaic genome because of the masking filters that we applied (see the STAR Methods) to eliminate questionable regions due to factors such as low coverage or poor mappability. The match rate that we report is the proportion of matches for unmasked alleles.

In the 1000 Genomes European populations the overall match rate to the sequenced Altai Neanderthal genome is 0.719 (Figure 3). By considering the larger UK10K sample, we can investigate the effect of allele frequency in detail. In the UK10K analysis, the rate of matching of the detected alleles to the Altai Neanderthal is fairly constant across the full range of allele frequencies, with an overall rate of 0.743. In contrast, randomly selected alleles that, like the putatively archaic-specific alleles, are at frequency < 0.01 in the West African outgroup have a very low rate (0.034) of matching to the Altai Neanderthal (Figure S3). This demonstrates that the match rate achieved by our method is much higher than would be found if a high proportion of the putative archaic-specific alleles were false positive. The match rate to the Altai Neanderthal and Altai Denisovan genomes is lower in the American populations than in the other 1000 Genomes populations (Figure 3). This is likely because the American populations are admixed and thus have higher background levels of LD that could cause false-positive results.

Two Waves of Denisovan Ancestry

In order to look more closely at the Neanderthal and Denisovan ancestry in present-day humans, we plot two-way densities of match rate to the Altai Neanderthal and Altai Denisovan genomes for segments with at least ten positions that can be compared to the Altai Neanderthal and at least ten positions that can be compared to the Altai Denisovan (Figure 4). In each population, we see a large cluster of segments with high matching to the Altai Neanderthal and low matching to the Altai Denisovan. This cluster corresponds to segments introgressed from Neanderthals. In each population the mode of matching to the Altai Neanderthal for this cluster is approximately 0.8, whereas the mode of matching to the Altai Denisovan genome is approximately 0.2. Thus approximately 20% of the archaic-specific variants introgressed from Neanderthals are also carried by the Altai Denisovan due to the relatedness of the Neanderthal and Denisovan populations, whereas 80% of the archaic-specific variants introgressed from Neanderthals are present in the Altai Neanderthal. In each population we also see a small cluster of segments with almost no matching to the Altai Neanderthal or to the Altai Denisovan; these are likely to be false-positive results that do not correspond to archaic introgression. In the Asian and Papuan populations we see a third cluster of segments. The segments in this third cluster have high matching to the Altai Denisovan and low matching to the Altai Neanderthal. This cluster corresponds to segments introgressed from Denisovans and confirms the previous finding of Denisovan admixture in Papuans and in Asians (Prüfer et al., 2014; Qin and Stoneking, 2015; Sankararaman et al., 2016; Skoglund and Jakobsson, 2011). Figures 4 and 5 also indicate that several other populations may carry a small proportion of segments introgressed from Denisovans. These include the Finns, who are estimated to have obtained around 7% of their ancestry from East Asia (Sikora et al., 2014), and admixed American populations whose Native American ancestors are related to East Asians (Gutenkunst et al., 2009).

In the Japanese and Chinese (Dai, Beijing, and Southern Han) populations we see that the Denisovan cluster of segments has a wide and bimodal distribution of match rates to the Altai Denisovan genome (Figure 4). A test for two distinct components of Denisovan ancestry (see the STAR Methods) is statistically significant ($p < 0.05$ after adjusting for multiple testing) in each

Table 1. Samples Analyzed

Identifier	Description	1000 Genomes Region	Number of Individuals
BEB	Bengali from Bangladesh	South Asia	86
CDX	Chinese Dai in Xishuangbanna, China	East Asia	93
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	Europe	99
CHB	Han Chinese in Beijing, China	East Asia	103
CHS	Southern Han Chinese	East Asia	105
CLM	Colombians from Medellin, Colombia	America	94
FIN	Finnish in Finland	Europe	99
GBR	British in England and Scotland	Europe	91
GIH	Gujarati Indian from Houston, TX, USA	South Asia	103
IBS	Iberian Population in Spain	Europe	107
ITU	Indian Telugu from the United Kingdom	South Asia	102
MXL	Mexican ancestry from Los Angeles, CA, USA	America	64
JPT	Japanese in Tokyo, Japan	East Asia	104
KHV	Kinh in Ho Chi Minh City, Vietnam	East Asia	99
PEL	Peruvians from Lima, Peru	America	85
PJL	Punjabi from Lahore, Pakistan	South Asia	96
PUR	Puerto Ricans from Puerto Rico	America	104
STU	Sri Lankan Tamil from the United Kingdom	South Asia	102
TSI	Toscani in Italia	Europe	107
YRI	Yoruba in Ibadan, Nigeria	Africa	108
UK10K	TwinsUK and Avon Longitudinal Study of Parents and Children (ALSPAC)	–	3,781
Papuans	From SGDP. Sampling location is 4S 143E (East Sepik province of Papua New Guinea)	–	15
SGDP Africans	From SGDP	–	44

of these four populations (Table 2) but is not significant in the other 1000 Genomes populations. The fitted two-component mixture has approximately one-third of the Denisovan segments in the Japanese and Chinese populations coming from the component with higher affinity to the Altai Denisovan genome. The putative archaic-specific alleles in the high-affinity component have a match rate of around 80% to the Altai Denisovan genome, which is similar to the match rate of putative archaic-specific alleles in Neanderthal introgressed segments with the Altai Neanderthal, whereas the putative archaic-specific alleles in the other (moderate-affinity) component have a match rate of around 50% to the Altai Denisovan genome.

To check that the moderate-affinity component is not due to segments that are a mosaic of Neanderthal and Denisovan ancestry, we reran the two-component mixture test excluding segments containing any Neanderthal-specific alleles (putative archaic-specific alleles matching the Neanderthal genome but not the Denisovan genome). We find that the same four populations (the three Chinese populations and the Japanese population) still have statistically significant p values for a two-component mixture after adjusting for multiple testing ($p < 0.0026$), and the estimated mixture parameters are essentially unchanged.

Based on the mode of matching to the Denisovan genome, most of the Denisovan ancestry in the South Asian and Papuan populations is from the archaic component with moderate affinity to the Altai Denisovan (Figure 4). This is consistent with previous work that noted that the Altai Denisovan is significantly more distantly related to the introgressing Denisovans compared to the relationship between the Altai Neanderthal and the introgressing Neanderthals (Prüfer et al., 2014).

To facilitate further analyses, we extracted subsets of segments based on their affinity to the Altai Neanderthal and to the Altai Denisovan (see the STAR Methods). We performed several analyses to check for possible confounders of match rate to the Denisovan genome. We checked whether the divergence between the Altai Neanderthal and Altai Denisovan differs between regions covered by the moderate-affinity Denisovan introgression and the high-affinity Denisovan introgression in case such differences could account for the two components. In the East Asian data, the mean relative divergence (number of homozygous discordances between the Altai Neanderthal and Altai Denisovan divided by the number of 1000 Genomes variants) per segment was 1.65 (SE 0.26) for high-affinity Denisovan segments and 2.51 (SE 1.00) for moderate-affinity Denisovan segments. The difference is not statistically significant ($p > 0.05$). We also investigated the average density of putative archaic-specific variants in segments attributed to the different components. We adjusted for length of the detected segments, because the power to detect segments increases with both length and the density of archaic-specific variants. In the East

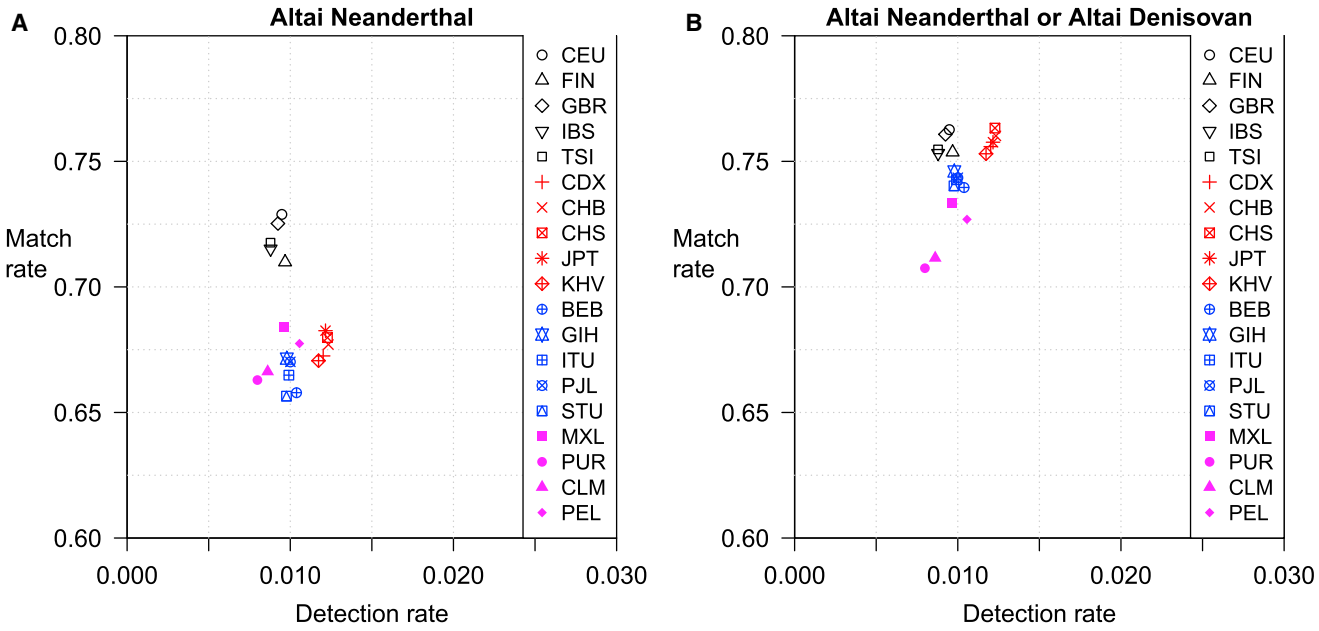


Figure 3. Detection and Match Rates in 1000 Genomes Populations

(A) Match rates are for the Altai Neanderthal.

(B) Match rates are for the Altai Neanderthal or Altai Denisovan. Population codes can be found in Table 1. Populations are colored by region (European in black; East Asian in red; South Asian in blue; and American in magenta). Match rate is the rate at which putative archaic-specific alleles match the sequence of the archaic genomes. Detection rate is the average proportion of each haplotype inferred to be introgressed.

See also Figure S3.

Asian data, the adjusted mean inverse density (bp per archaic-specific variant) was 103 (SE 440) for the high-affinity Denisovan segments, 395 (SE 464) for the moderate-affinity Denisovan segments, and 1,164 (SE 72) for the Neanderthal segments. The difference is not statistically significant ($p > 0.05$). Thus we do not find confounding by divergence or by density of archaic-specific alleles.

We investigated the lengths of haplotypes within segments attributed to the different components in order to investigate potential differences in admixture time between components. We analyzed haplotype lengths in units of centimorgans (cM) rather than base pairs because centimorgan distances reflect recombination and are thus less variable. We adjusted for frequency and overall segment length because high frequency and high segment length increase power to detect a segment and are correlated with haplotype length. In the East Asian data, the mean adjusted haplotype length was 0.066 (SE 0.014) cM for Neanderthal segments, 0.19 (SE 0.13) cM for high-affinity Denisovan segments, 0.072 (SE 0.13) cM for moderate-affinity Denisovan segments, and 0.13 (SE 0.06) cM for Denisovan segments overall. These are not significantly different. We also checked for differences in Europeans, in South Asians, in Asians overall (East and South), and in Papuans, again finding no significant differences. While it is probable that the Neanderthal admixture and the two waves of Denisovan admixture occurred at distinct times, there is insufficient information in the data to determine the ordering of these events.

Overall, East Asians and South Asians carry similar amounts of detected Denisovan ancestry, while Papuans carry much more detected Denisovan ancestry (Figure 5). Approximately one-third of the Denisovan ancestry segments in the East Asians are from the high-affinity component (Table 2), whereas very little of the Denisovan ancestry in the South Asians and Papuans is from the high-affinity component (Figure 4). A possible scenario consistent with this pattern would have the high-affinity component introgressing into East Asia after the split between East and South Asia. Because the Papuans have a much higher frequency of the moderate-affinity Denisovan component than other populations, it may be that this component was primarily introgressed into the ancestors of Papuans after they split from Asia, and arrived in Asia via migration from the ancestors of Papuans; however, other scenarios are also possible (Prüfer et al., 2014; Sankararaman et al., 2016).

Lack of Evidence for Multiple Waves of Neanderthal Ancestry

The frequency of Neanderthal introgression is substantially higher (~30%) in East Asians than in Europeans (Meyer et al., 2012; Wall et al., 2013). This difference cannot be explained by differential effects of selection, but could be due to an additional Neanderthal admixture event into the ancestors of East Asians after the Europe-Asia split (Kim and Lohmueller, 2015; Vernot and Akey, 2015). Another possible explanation would be dilution of Neanderthal admixture in Europe due to migration from a population without Neanderthal admixture (Meyer et al., 2012; Vernot and Akey, 2015).

In our results, the Neanderthal-introgressed segments in East Asians and in Europeans show indistinguishable levels of

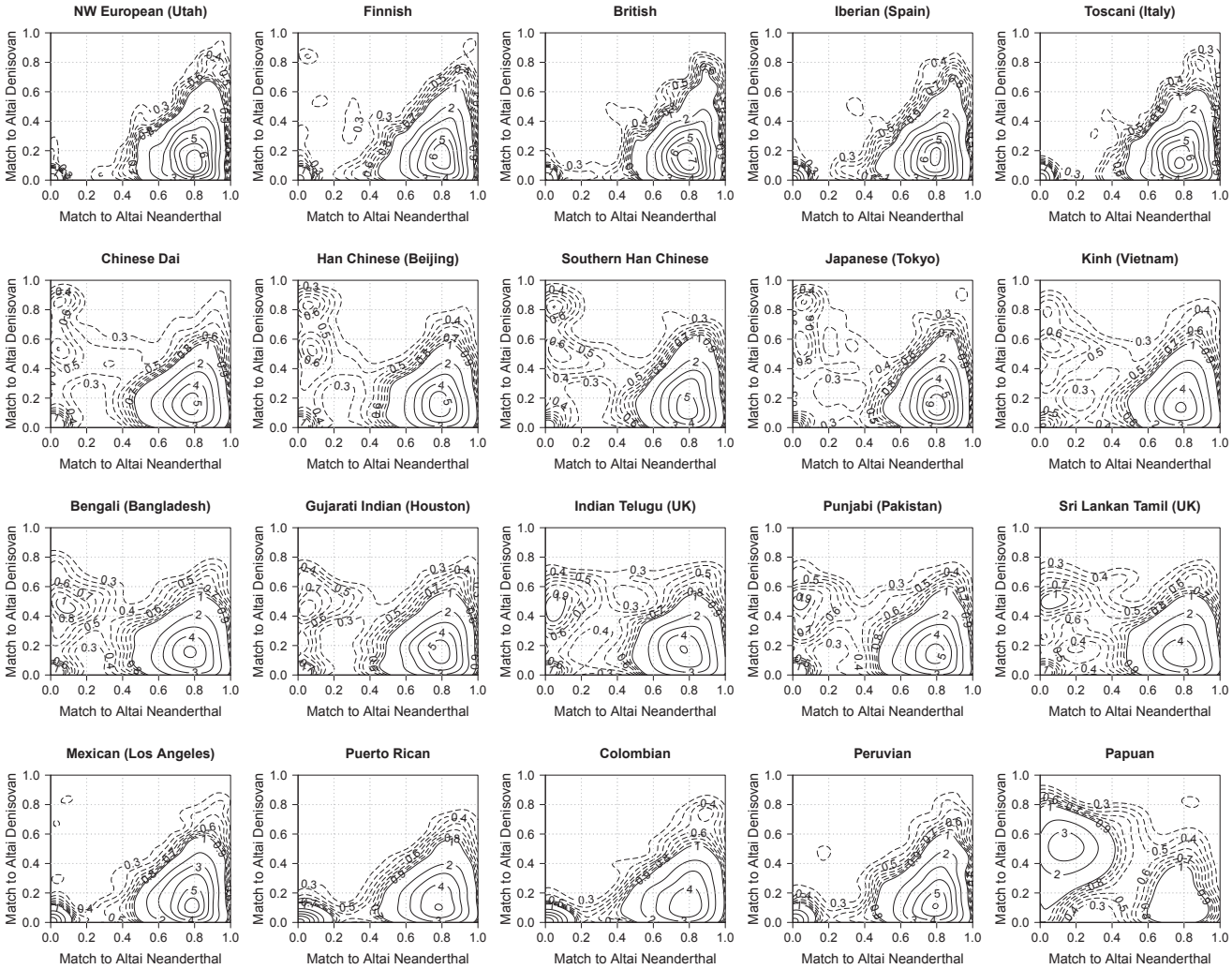


Figure 4. Contour Density Plots of Match Proportion of Introgressed Segments to the Altai Neanderthal and Altai Denisovan Genomes

The match proportion is the proportion of putative archaic-specific alleles in a segment that match the given archaic genome, excluding alleles at positions masked in the archaic genome sequence. Segments with at least ten variants not masked in the Neanderthal genome and at least ten variants not masked in the Denisovan genome are included. Numbers inside the plots indicate the height of the density corresponding to each contour line. Contour lines are shown for multiples of 1 (solid lines). In addition, contour lines for multiples of 0.1 between 0.3 and 0.9 (dashed lines) are shown for additional detail. European populations are given in the first row, East Asian populations in the second row, South Asian populations in the third row, and American and SGDP Papuan populations in the final row. Additional information about the populations can be found in Table 1.
See also Figure S4.

similarity to the Altai Neanderthal (Figure 4). There is also no clear difference between East Asians and Europeans in the similarity of their Neanderthal-introgressed segments to the Vindija 33.19 Neanderthal (Figure S4). Thus, if the ancestors of East Asians received a large pulse of Neanderthal admixture after splitting from Europeans, then the original (shared Eurasian) and additional (East Asian-specific) admixing populations must have been closely related.

Signals of Positive Selection

We looked for introgressed segments with highest frequency in 1000 Genomes populations. Specifically we found in each population the two regions of highest frequency that had high matching to the Altai Neanderthal or Altai Denisovan genome (see the STAR Methods). Table S1 lists the regions. All these regions appear to have been introgressed from Neanderthals rather than Denisovans. Several of the positively selected regions have been described previously, including *BNC2*, *POU2F3*, and *KRT71*, which are involved in skin and hair traits (Sankararaman et al., 2014; Vernot and Akey, 2014). Genomic regions introgressed from Neanderthals and under positive selection have been shown to be enriched for genes involved in pigmentation and immunity (Deschamps et al., 2016; Gittelman et al., 2016; Racimo et al., 2015; Sankararaman et al., 2014, 2016; Vernot and Akey, 2014; Vernot et al., 2016).

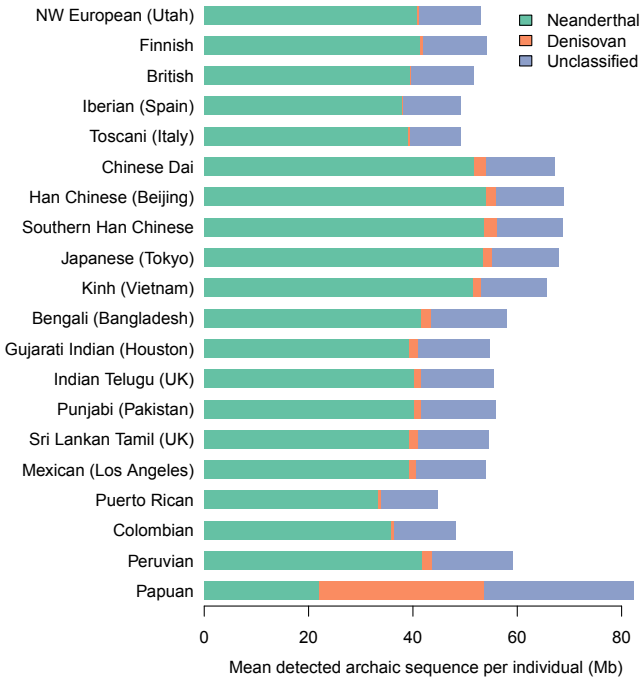


Figure 5. Mean Amounts of Detected Introgressed Material per Individual, Classified by Affinity to the Altai Neanderthal and Altai Denisovan Genomes

Definitions of the affinity groups are given in the STAR Methods. Unclassified material includes segments that are too short to be confidently classified into an affinity group, as well as longer segments that have low levels of affinity to the archaic genomes.

Table 2. Two-Component Mixtures for Denisovan-Related Introgression

Population	μ_1	μ_2	σ_1	σ_2	p	p value
Southern Han Chinese	0.82	0.46	0.08	0.12	0.42	0.00002
Han Chinese (Beijing)	0.84	0.50	0.08	0.14	0.36	0.00021
Chinese Dai	0.86	0.52	0.04	0.18	0.20	0.00069
Japanese (Tokyo)	0.86	0.52	0.06	0.18	0.26	0.00143
Finnish	0.84	0.50	0.04	0.14	0.22	0.00348
Punjabi (Pakistan)	0.82	0.48	0.10	0.12	0.10	0.04589

Populations with a p value < 0.05 for a two-component mixture are shown, ordered by p value. p values should be compared to a Bonferroni-corrected threshold of $0.05/19 = 0.0026$ since 19 non-African populations from the 1000 Genomes Project were tested. μ_1 and μ_2 are the estimated mean per-segment matching of putative archaic-specific variants to the Altai Denisovan genome for the two components; σ_1 and σ_2 are the estimated standard deviations of per-segment matching for the two components; p is the proportion of segments attributed to the first (high-affinity) component; and the p value is for a likelihood ratio test for the two-component Gaussian mixture versus a single Gaussian distribution.

In addition to the regions that have been extensively described in previous studies of positively selected archaic introgression, our results include two immunity-related regions, which we highlight here. The first of these immunity-related regions is on chromosome 3p21.31. This region was included in a supplementary table of high-frequency introgressed haplotypes in Gittelman et al. (2016), but was not discussed in that work. The introgressed alleles at this locus are at high frequency in South Asia (0.38). The region contains *CCR9* (C-C motif chemokine receptor 9) and *CXCR6* (C-X-C motif chemokine receptor 6), which are chemokine receptors involved in immunity (Papadakis et al., 2000; Paust et al., 2010; Zlotnik and Yoshie, 2000).

The second of these immunity-related regions is on chromosome 14q32.33. The introgressed alleles in this region are at very high frequency throughout Eurasia. This region is located in the immunoglobulin heavy locus, which contains multiple genes that code for antibodies (Schroeder and Cavacini, 2010). Immunoglobulin heavy genes contained within the high-frequency region are *IGHA1*, *IGHG1*, and *IGHG3.* The most highly conserved introgressed position is rs10144746 (PhyloP score 4.1) and is an expression quantitative trait locus (eQTL) for IGHG4 and several other immunoglobulin heavy genes in various tissues including esophagus and liver. The high-frequency introgression is in a region with significant masking of the Altai Neanderthal and Altai Denisovan genomes due to poor quality sequence. For example, for the segment found in the Southern Han Chinese (CHS) population, 119 of the 145 putatively introgressed alleles are filtered in the Altai Neanderthal genome (see the STAR Methods). Of the 26 unfiltered alleles, 22 match the Altai Neanderthal genome. Thus this region appears to be derived from Neanderthal admixture, but would be difficult to find using a reference-based approach.

DISCUSSION

We applied a new method for detecting archaic introgressed segments to worldwide non-African populations from the 1000 Genomes project, Papuans from the SGDP and individuals from the UK10K project. Our method is reference-free, which means that it can detect introgression from archaic admixing populations without a reference sequence. We show that when a reference sequence exists, comparison of the detected segments to the reference genome can lead to new insights into population history.

We found evidence that Asians carry Denisovan introgression, confirming previous reports that used alternative methods (Prüfer et al., 2014; Qin and Stoneking, 2015; Sankararaman et al., 2016; Skoglund and Jakobsson, 2011). Further, we found evidence for two waves of Denisovan admixture, one from a population closely related to the Altai Denisovan individual, and one from a population more distantly related to the Altai Denisovan. The component closely related to the Altai Denisovan is primarily present in East Asians, whereas the component more distantly related to the Altai Denisovan forms the major part of the Denisovan ancestry in Papuans and South Asians. The East Asian populations are the only populations with relatively equal and non-negligible contributions from both components, and it is in these populations that the two waves of Denisovan admixture are most evident.

In contrast, we did not find evidence for two or more waves of Neanderthal admixture from diverged Neanderthal populations. The higher rates of Neanderthal introgression in East Asians

relative to Europeans may be due to dilution of Neanderthal admixture in Europeans as a result of migration from a population without Neanderthal admixture (Meyer et al., 2012; Vernot and Akey, 2015). If there was an additional pulse of Neanderthal admixture into East Asians after the Europe-Asia split, then it was from a population closely related to the primary admixing Neanderthals.

We found a number of high-frequency introgressed haplotypes that appear to have been subject to positive selection. Two of these regions are involved in immunity, containing the immunoglobulin heavy locus and a cluster of chemokine receptors. These regions, in addition to previous reports of positively selected introgressed haplotypes in histocompatibility leukocyte antigen (HLA) genes (Abi-Rached et al., 2011), Toll-like receptors (Deschamps et al., 2016), and many other immunity genes (Abi-Rached et al., 2011; Deschamps et al., 2016; Quach et al., 2016; Racimo et al., 2015) underscore the crucial role that Neanderthal introgression played in adapting the human immune system to the pathogenic landscape of Eurasia.

Our results were obtained using a new S*-like algorithm for genome-wide reference-free introgression detection. Our method is implemented in the freely available software package Sprime and is computationally efficient for analysis of large sequenced datasets. For example, genome-wide analysis of nearly 4,000 UK10K individuals required only 4 hr of computing time on a 2.6 GHz CPU. As the number of whole-genome sequences continues to grow, computationally efficient methods, such as the one described here, will be essential for constructing a map of all surviving archaic hominin sequences in present-day human populations.

Our method reports the set of putative archaic-specific alleles in each introgressed segment. Direct identification of the putative archaic-specific alleles is useful for downstream analyses. Rates of matching of these alleles to a reference archaic genome indicate the degree of divergence between the introgressing and sequenced archaic individuals. The usefulness of these match rates is shown in this study, where they reveal two pulses of Denisovan admixture.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Details of the Sprime algorithm
  - Simulation study
  - Whole genome sequence data
- QUANTIFICATION AND STATISTICAL ANALYSES
  - Estimation of match rate probability densities
  - Test for two distinct Denisovan components
  - Tests for differences between subsets of segments based on their affinity to the archaic genomes
  - Adaptive introgression
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and two tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.02.031.

A video abstract is available at https://doi.org/10.1016/j.cell.2018.02.031#mmc2.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of General Medical Sciences of the NIH under Award Number R01GM110068. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

S.R.B. planned and directed this study, developed the methods, and wrote the manuscript. B.L.B. wrote the software. S.R.B., Y.Z., and S.T. conducted the analyses. J.M.A. provided advice for the planning and direction of this study. All authors contributed to editing the manuscript.

DECLARATION OF INTERESTS

J.M.A. is a paid consultant of Glenview Capital.

Received: October 4, 2017
Revised: November 21, 2017
Accepted: February 12, 2018
Published: March 15, 2018

REFERENCES

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science *334*, 89–94.

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

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

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature *524*, 216–219.

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R.A., Sing, C.F., Clark, A.G., and Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. Proc. Natl. Acad. Sci. USA *111*, 757–762.

Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M., and Akey, J.M. (2016). Archaic hominin admixture facilitated adaptation to out-of-Africa environments. Curr. Biol. *26*, 3375–3382.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D.; 1000 Genomes Project (2011).

Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983–11988.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695.

Hammer, M.F., Woerner, A.E., Mendez, F.L., Watkins, J.C., and Wall, J.D. (2011). Genetic evidence for archaic admixture in Africa. Proc. Natl. Acad. Sci. USA *108*, 15123–15128.

Hsieh, P., Woerner, A.E., Wall, J.D., Lachance, J., Tishkoff, S.A., Gutenkunst, R.N., and Hammer, M.F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res. *26*, 291–300.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature *512*, 194–197.

Juric, I., Aeschbacher, S., and Coop, G. (2016). The strength of selection against Neanderthal introgression. PLoS Genet. *12*, e1006340.

Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.

Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. *12*, e1004842.

Kim, B.Y., and Lohmueller, K.E. (2015). Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. Am. J. Hum. Genet. *96*, 454–461.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., et al. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell *150*, 457–469.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

Papadakis, K.A., Prehn, J., Nelson, V., Cheng, L., Binder, S.W., Ponath, P.D., Andrew, D.P., and Targan, S.R. (2000). The role of thymus-expressed chemokine and its receptor CCR9 on lymphocytes in the regional specialization of the mucosal immune system. J. Immunol. *165*, 5069–5076.

Paust, S., Gill, H.S., Wang, B.Z., Flynn, M.P., Moseman, E.A., Senman, B., Szczepanik, M., Telenti, A., Askenase, P.W., Compans, R.W., and von Andrian, U.H. (2010). Critical role for the chemokine receptor CXCR6 in NK cell-mediated antigen-specific memory of haptens and viruses. Nat. Immunol. *11*, 1127–1135.

Plagnol, V., and Wall, J.D. (2006). Possible ancestral structure in human populations. PLoS Genet. *2*, e105.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija cave in Croatia. Science *358*, 655–658.

Qin, P., and Stoneking, M. (2015). Denisovan ancestry in East Eurasian and Native American populations. Mol. Biol. Evol. *32*, 2665–2674.

Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell *167*, 643–656.e17.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. *16*, 359–371.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature *468*, 1053–1060.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. PLoS Genet. *8*, e1002947.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr. Biol. *26*, 1241–1247.

Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. *13*, 745–753.

Schroeder, H.W., Jr., and Cavacini, L. (2010). Structure and function of immunoglobulins. J. Allergy Clin. Immunol. *125* (2, Suppl 2), S41–S52.

Siepel, A., Pollard, K.S., and Haussler, D. (2006). New methods for detecting lineage-specific selection. RECOMB. Proceed. *3909*, 190–205.

Sikora, M., Carpenter, M.L., Moreno-Estrada, A., Henn, B.M., Underhill, P.A., Sánchez-Quinto, F., Zara, I., Pitzalis, M., Sidore, C., Busonero, F., et al. (2014). Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. PLoS Genet. *10*, e1004353.

Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbring, S.J., Jarvik, G.P., Kullo, I.J., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. Science *351*, 737–741.

Skoglund, P., and Jakobsson, M. (2011). Archaic human ancestry in East Asia. Proc. Natl. Acad. Sci. USA *108*, 18301–18306.

UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017–1021.

Vernot, B., and Akey, J.M. (2015). Complex history of admixture between modern humans and Neandertals. Am. J. Hum. Genet. *96*, 448–453.

Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., et al. (2016). Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science *352*, 235–239.

Wall, J.D., Lohmueller, K.E., and Plagnol, V. (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol. Biol. Evol. *26*, 1823–1827.

Wall, J.D., Yang, M.A., Jay, F., Kim, S.K., Durand, E.Y., Stevison, L.S., Gignoux, C., Woerner, A., Hammer, M.F., and Slatkin, M. (2013). Higher levels of neanderthal ancestry in East Asians than in Europeans. Genetics *194*, 199–209.

Zlotnik, A., and Yoshie, O. (2000). Chemokines: a new classification system and their role in immunity. Immunity *12*, 121–127.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
1000 Genomes project data, phase 3 version 5a	1000 Genomes project (1000 Genomes Project Consortium, 2015)	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
Simons Genome Diversity Project data	David Reich (Mallick et al., 2016)	https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public/
UK10K data	UK10K project (UK10K Consortium et al., 2015) available from the European Genome-Phenome Archive https://ega-archive.org/	accessions EGAD00001000740 and EGAD00001000741
Altai Neanderthal, Altai Denisovan and Vindija 33.19 genomes	Kay Prüfer (Prüfer et al., 2017)	http://cdna.eva.mpg.de/neandertal/Vindija/
HapMap genetic map	International HapMap Consortium (International HapMap Consortium, 2007)	http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/
Single tissue expression QTLs	GTEx project	https://www.gtexportal.org/home/
Single nucleotide variant annotation	dbSNP	https://www.ncbi.nlm.nih.gov/projects/SNP/
PhyloP100wayAll conservation scores	UCSC Genome Browser	https://genome.ucsc.edu/
Sprime scores for 1000 Genomes populations and SGDP Papuans	This paper	https://doi.org/10.17632/y7hyt83vxr.1
Software and Algorithms		
Sprime software for detection of segments of archaic introgression	This paper	http://faculty.washington.edu/browning/sprime.html
Msprime coalescent-based simulation software	Jerome Kelleher; Kelleher et al., 2016	https://github.com/jeromekelleher/msprime
R	The R Project for Statistical Computing	https://www.r-project.org/
MASS (R package)	The Comprehensive R Archive Network (CRAN)	https://cran.r-project.org/web/packages/MASS/index.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sharon Browning (sguy@uw.edu).

METHOD DETAILS

Details of the Sprime algorithm

Overview

The input to our method consists of whole genome sequence genotypes of the target and outgroup individuals, and an LD-based (i.e., fine scale) genetic map. The outgroup is a population that is closely related to the ancestors of the target population, but that is not expected to have experienced admixture from the archaic population that contributed to the target population. However, the outgroup population may contain a small amount of introgressed sequence from the archaic population due to migration from an introgressed population. The 108 Yoruban individuals (YRI) from the 1000 Genomes dataset are used as the outgroup to detect Neanderthal introgression, except as otherwise noted, because they are thought to have no direct admixture from Neanderthals and to have received limited genetic material from Eurasian populations.

The output of our method is a list of detected introgressed segments and the putative archaic-specific alleles that comprise those segments. Using this list and the original genotype data, one can reconstruct the introgression status of the target individuals along their genomes. The putative archaic-specific alleles are mutations that our method infers to have arisen on the archaic lineage after the split with the ancestor of modern humans.

The first step of the algorithm is to read in the genotype data, considering each allele at each position of variation and grouping the alleles into three classes: those common in the outgroup, those not seen in the outgroup, and those uncommon but present in the outgroup. The threshold used to distinguish common versus uncommon variants is 0.01 for the analyses presented here. The common alleles (those with frequency > 0.01 in the outgroup) are excluded from further consideration, because it is unlikely that they are mutations that arose on the archaic lineage.

The next step is to find segments (i.e., sets of alleles) which maximize the score function (described below). We process one chromosome at a time, and we first find the set of alleles with the highest score. These comprise the first segment. We remove those alleles from consideration and proceed to find the next set of alleles with the highest score: these comprise the second segment. We remove these alleles and continue until the set of alleles with the highest score, among those that remain, has a score that is less than the user-specified threshold (150,000 in this study, except where otherwise noted).

S*-type scoring

For a given set of alleles, $a_1, a_2, \ldots, a_J$, ordered by position on the chromosome, the score S is defined to be the sum of adjacent pairwise scores T. That is,

$$S(a_1, a_2, \ldots, a_J) = \sum_{j=1}^{J-1} T(a_j, a_{j+1})$$

Due to the pairwise definition of the score S, it can be maximized over all possible sets of alleles using a dynamic programming algorithm. This algorithm was described by Plagnol and Wall (2006), and is reported here for completeness.

Consider all alleles $v_1, v_2, \ldots, v_V$ that are absent or uncommon in the outgroup, ordered by genomic position. Define $S^*_1 = 0$, and recursively define $S^*_{j+1} = \max\left\{0, \max_{k=1,2,\ldots,j} S^*_k + T(v_k, v_{j+1})\right\}$. S^*_j is the maximal score for a segment that ends with allele v_j. The highest scoring segment has score equal to $\max_{j=1,2,\ldots,V} S^*_j$. The putative archaic-specific alleles that comprise this segment are those alleles that were used to obtain this maximal score.

Much of the description so far is essentially the same as that of Plagnol and Wall (2006), with the exceptions being the application of a non-zero allele frequency threshold to the outgroup, and the iterative procedure to find multiple segments across a chromosome rather than just one segment in a small region. However, our definition of the pairwise score function T, given below, differs significantly in its details from the pairwise score used by Plagnol and Wall, although it is similar in spirit. Our pairwise score is designed to enable analysis of arbitrarily large target samples and to adjust for local mutation and recombination rates.

Accounting for number of target samples in the score

Consider a pair of alleles v_1 and v_2 for which we want to calculate the score $T(v_1, v_2)$. Let i index individuals in the target sample. Write X_{i1} for the dose (number of copies) of v_1 in individual i, and similarly X_{i2} for v_2. If the two alleles are truly archaic-specific and derive from a single introgressed haplotype, we expect to see $X_{i1} = X_{i2}$, unless individual i has a change in introgression status between the positions of these two alleles due to recombination between introgressed and non-introgressed haplotypes on the lineage of one of individual i's haplotypes. Define $D = \sum_i |X_{i1} - X_{i2}|$, which is the count of such apparent recombination events. Note that when $X_{i1} = X_{i2} = 1$, it is possible that there have been two recombination events, so that while individual i's maternal chromosome was introgressed at the first position, it is individual i's paternal chromosome that is introgressed at the second position, or vice versa; however we ignore this possibility since it is expected to have low frequency. We expect nearby introgressed archaic-specific alleles to be in high LD (thus have small D) when they are close together on a chromosome.

We require that adjacent alleles in a segment are not too far apart (so that the method doesn't bridge distinct segments) or too close (so that the method doesn't double count multi-nucleotide mutations). If the two alleles are more than 20 kb apart, or less than 10 bp apart, the score T is $-\infty$. We also require that there be some overlap in carrier individuals for the two alleles. If there is no overlap, i.e., $\sum_i X_{i1} X_{i2} = 0$, then the score T is $-\infty$.

In the original S*, the score T takes a positive value if D is zero, a negative values if D is between 1 and 5, and is $-\infty$ if D is greater than 5. Because D is often greater than zero in high frequency introgressed regions when the sample size is large, we take a more nuanced approach. We form the score as the sum of a positive part that depends on the local mutation and recombination rates and a negative part that penalizes the score when values of D are large relative to the introgression frequency.

Let n be the smaller of the number of haplotypes carrying the first allele and the number of haplotypes carrying the second allele, i.e., $n = \min(\sum_i X_{i1}, \sum_i X_{i2})$. Each introgressed haplotype can end, resulting in a contribution to D, when it experiences a recombination. Thus when n is large (the introgressed haplotype is common), we expect to observe a proportionately larger number of recombination events, i.e., a larger value of D, compared to when n is small. Thus, in the pairwise score we work with the normalized value D/n.

Accounting for migration, and for local mutation and recombination rates

In the original S*, the terms in the score T did not depend on the local mutation rate or recombination rate. Also, the presence of low levels of introgression in the outgroup due to migration was not allowed for. We modify the positive term our score to adjust for local mutation and recombination rates, and to allow for a low frequency of introgression in the outgroup.

We want to consider alleles that have non-zero outgroup frequency f provided they are uncommon in the outgroup $(f \leq 0.01)$; however we slightly down-weight such alleles with $0 < f \leq 0.01$ relative to alleles with $f = 0$ because alleles with $f > 0$ are more likely to be of non-archaic origin. Let the weight w take value 1 if $f = 0$ and value 0.8 if $0 < f \leq 0.01$, where f is the frequency of the second allele in the outgroup.

We need to account for local differences in mutation rate and recombination rate. If the mutation rate is high and the recombination rate is low in a region, we will find a lot of high-LD pairs of alleles in the region, although many may not be due to introgression. On the other hand, if the mutation rate is low and the recombination rate is high, we will not find many high-LD pairs of alleles, archaic-specific or otherwise. The ratio of mutation rate to recombination rate determines the rates at which we expect find high-LD pairs of alleles, in both the introgressed and non-introgressed settings. Let m represent the local rate of mutation per centiMorgan (cM) per meiosis. We describe how we estimate m below. We score pairs of high-LD alleles more highly if they are in a region with a low rate of mutation per cM, however we don't want to score them arbitrarily highly when m is extremely low, as this would add excessive variability. Thus we adjust for the local rate of mutation per cM using the function $(1 - \exp(-0.01/\mathrm{m}))/(1 - \exp(-1))$. This function is approximately proportional to $1/(100\mathrm{m})$ for large m, and has a maximum value of approximately 1.6.

The scoring function

Putting these pieces together, the positive term in T is proportional to $w(1 - \exp(-0.01/\mathrm{m}))/(1 - \exp(-1))$, and the negative term is proportional to D/n. We chose the constants of proportionality that gave high accuracy in simulated data, obtaining the following score for a pair of alleles:

$$6000w\frac{1-\exp(-0.01/m)}{1-\exp(-1)} - 25000\frac{D}{n}.$$

We now give a procedure to estimate m, the local rate of mutation per cM per meiosis.

$$\mathrm{m} = (\textit{local rate of mutations per bp per meiosis}) \times (\textit{local bp per cM rate})$$

For the local rate of mutation we can utilize the local rate of assayed variation, or we can use the number of differences between the human reference sequence and a primate reference sequence such as macaque. We investigated both options and obtained similar results (data not shown), so we decided to proceed with the local rate of assayed variation based on the input genotype data file. Hence

$$\textit{local rate of mutations per bp per meiosis} = \frac{\textit{local variant density}}{\textit{global variant density}} \times \mu_g$$

where μ_g is the genome-wide mutation rate, for which we use 1.2×10^{-8} mutations per bp per meiosis for analysis of human data (Scally and Durbin, 2012), and the true simulated mutation rate for simulated data. The local and global variant densities are obtained from the input VCF file as the number of variant positions in the region divided by the number of basepairs in the region. For the local rate of recombination, we utilize the HapMap genetic map, which is a fine scale map based on patterns of LD (International HapMap Consortium, 2007). Although recombination rates differ between populations, at distances of 10 kb and greater the rates between diverse human populations are highly correlated (Kong et al., 2010).

Over small distances, the estimated bp per cM rate and the estimated local mutation rate are expected to be somewhat inaccurate, as well as potentially varying between populations. We thus combine estimates from the small region of interest with estimates from slightly larger surrounding regions in a conservative manner, preferring to over-estimate rather than under-estimate m. Suppose the region of interest runs from position p_1 bp to position p_2 bp. To estimate the local mutation rate, we calculate the local rate of assayed variation for the region from p_1 bp to p_2 bp, then for the region from $p_1 - 5000$ bp to $p_2 + 5000$ bp, then for the region from $p_1 - 10000$ bp to $p_2 + 10000$ bp, considering regions with at least 6 assayed variants and stopping once we have a region with at least 10 assayed variants. We then take the maximum local mutation rate over these regions. Similarly, for local bp per cM rate we obtain the local rate (using interpolation as needed) from the HapMap map for the region from p_1 bp to p_2 bp, then for the region from $p_1 - 5000$ bp to $p_2 + 5000$ bp, then for the region from $p_1 - 10000$ bp to $p_2 + 10000$ bp, stopping once we have a region of genetic length at least 0.01 cM. We then take the maximum local bp per cM rate over these regions.

We tuned the parameters for our algorithm using simulated data and using the UK10K data with reference to the Altai Neanderthal genome. Unless otherwise specified, we use a score threshold of 150,000, which we found to give a good tradeoff between power and accuracy.

Simulation study

We used msprime (Kelleher et al., 2016) to simulate data and to determine introgression status in the simulated European individuals. Our simulation code is given in Table S2. Our simulations comprised 100 replicates of 10 Mb, each with 100 present-day African individuals, 4,000 present-day European individuals, and 100 Neanderthal individuals from 2,000 generations ago. The purpose of the Neanderthal individuals is to provide true introgression status for the simulated European individuals. All individuals are diploid. We simulated a mutation rate of 2.4×10^{-8} per bp per meiosis and performed some analyses using all simulated variants and other

analyses using a subset of the variants to obtain a lower mutation rate. We used a recombination rate of 10^{-8} per bp per meiosis. The parameters for our demographic model were based on published estimates (Gazave et al., 2014; Gravel et al., 2011; Keinan and Clark, 2012; Scally and Durbin, 2012; Vernot and Akey, 2014). The split between the ancestors of Neanderthals and modern humans was set to 16,000 generations ago (400,000 kya, assuming a generation time of 25 years). Effective population sizes were set to 15,000 for the ancestral human population and 1,500 for Neanderthals. The out-of-Africa human migration occurred 2,400 generations ago (60,000 kya), and the size of the out-of-Africa/European population was 2,000. The rate of migration between Africa and the out-of-Africa/European population was 10^{-5} per individual per generation, which corresponds to a cumulative Eurasian admixture into Africa over 2,400 generations of 0.024. We allowed for introgression to occur between 1,980 to 2,000 generations ago at a rate of 0.0015, for an overall admixture proportion of 0.03. We allowed for rapid growth of 2% per generation in all human populations starting 200 generations ago with the development of agriculture.

To determine true introgression status for comparison with inferred introgression status, we checked coalescence times between each European haplotype and the 200 Neanderthal haplotypes. If a European haplotype coalesced with a Neanderthal haplotype more recently than the Neanderthal-human split time, we inferred that the European haplotype was an introgressed Neanderthal haplotype.

We use two metrics in assessing the performance of our method on simulated data. The first is accuracy. Our method calls putative introgressed alleles, and accuracy is the proportion of these alleles that are introgressed. That is, such alleles should be found only on introgressed haplotypes. The second metric is detection frequency, which is the proportion of haplotypes in the population that are both truly introgressed and carry introgressed alleles found by our method. This can be compared to the proportion actually carrying an introgressed haplotype (detected or not), which is 3%.

Whole genome sequence data

We ran our statistical framework on 5,639 whole genome sequences from individuals in populations in Europe, Asia and Oceania. See Table 1 for a detailed description of the datasets used in this study. The 1000 Genomes data (1000 Genomes Project Consortium, 2015) are phase 3 version 5a downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The Simons Genome Diversity Project (SGDP) data (Mallick et al., 2016) were downloaded from https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public/. The UK10K data (UK10K Consortium et al., 2015) were obtained from the EGA (EGAD00001000740 and EGAD00001000741).

High coverage reference genomes for the Altai Neanderthal, Altai Denisovan and Vindija 33.19 Neanderthal (Prüfer et al., 2017) were obtained from http://cdna.eva.mpg.de/neandertal/Vindija/. Most of the analyses in this study were conducted prior to the publication of the Vindija 33.19 genome, and hence most Neanderthal-based analyses use the Altai Neanderthal rather than the Vindija 33.19 Neanderthal. These three genomes have been called with snpAD, an ancient DNA damage-aware genotyper (Prüfer et al., 2017). When comparing putative introgressed alleles with an archaic genome, we applied the masks provided with these data and downloaded from http://cdna.eva.mpg.de/neandertal/Vindija/FilterBed/ which filter sites with coverage depth below 10, or mapping quality below 25, or that are within tandem repeats or indels, or that have poor mappability (Li and Durbin, 2011). For sites that pass these filters, we report a match if the archaic genotype includes the putative archaic-specific allele and a mismatch otherwise. The match rate is calculated as the number of matches divided by the number of compared sites (matches and mismatches). Sites that do not pass the filters do not contribute to the match rate.

All analyses were performed on autosomes. We performed introgression detection in individual populations rather than continental groups to avoid potential effects of population structure. For analyses of the 1000 Genomes non-African populations, we used Yorubans as the outgroup. We also used the 1000 Genomes Yorubans as the outgroup when analyzing the UK10K data. Since the UK10K sample is large, rare introgressed variants may be present in the UK10K but not present in the 1000 Genomes study due to the much smaller European sample size in that study. So for the UK10K analysis, we assumed that the Yorubans are monomorphic for the UK10K major allele for variants called in the UK10K but not called in the 1000 Genomes project, which avoids filtering the minor alleles of such variants. In our analysis of the SGDP Papuans, we used the SGDP African samples as the outgroup.

QUANTIFICATION AND STATISTICAL ANALYSES

Estimation of match rate probability densities

Two-dimensional probability densities for the contour density plots were estimated using the function kde2d from the MASS package in R with default settings but restricting to the range of interest.

Test for two distinct Denisovan components

We consider segments with at least 30 putatively introgressed alleles that can be compared with the Altai Neanderthal genome, at least 30 putatively introgressed alleles that can be compared with the Altai Denisovan genome, a match rate to the Altai Neanderthal genome of less than 0.3, and a match rate to the Altai Denisovan genome greater than 0.3. This should remove most false positive segments and segments that are of Neanderthal rather than Denisovan ancestry. Then, within a single population, we consider the distribution of match rates to the Denisovan genome. The match rate of a segment from a single introgression component should have an approximately normal distribution since the segments are relatively long (on average 300 kb with 150 variants) after the

filtering steps listed above. We fit either a mixture of two Normal distributions or a single Normal distribution, in either case truncated below at 0.3. We perform a likelihood ratio test to compare these two models. We also performed a second likelihood ratio test using a different set of filters, as described in the main text.

Tests for differences between subsets of segments based on their affinity to the archaic genomes

In order to compare the properties of introgressed haplotypes from different archaic sources, we extracted segments that are clearly of Neanderthal origin, or of Denisovan origin, considering only segments with at least 30 putatively introgressed alleles that can be compared with the Altai Neanderthal genome, and at least 30 putatively introgressed alleles that can be compared with the Denisovan genome. The Neanderthal segments are those with match rate higher than 0.6 to Altai Neanderthal and less than 0.4 to the Altai Denisovan genome. The Denisovan segments are those with match rate higher than 0.4 to the Altai Denisovan genome, and less than 0.3 to the Altai Neanderthal genome. From the latter group, we also extracted segments with higher or lower affinity to the Denisovan genome. The high-affinity segments are those with match rate higher than 0.7 to the Denisovan genome, while the moderate-affinity segments are those with match rate less than 0.6 (but higher than 0.4) to the Denisovan genome.

To examine divergence between the Altai Neanderthal and Altai Denisovan genomes for subsets of segments, we calculated for each segment the number of homozygous discordances between the Altai Neanderthal and the Altai Denisovan divided by the number of 1000 Genomes variants in the region. We calculated the mean of these values, and a standard error based on bootstrap resampling over chromosomes. From this we obtained approximate Wald test statistics as $(\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2 + s_2^2}$ to test for differences in average divergence for regions covered by moderate-affinity Denisovan segments and regions covered by high-affinity Denisovan segments.

We performed similar analyses to test for differential density of putative archaic-specific variants or differential introgressed haplotype lengths. For density of putative archaic-specific variants, we calculated the inverse density (bp per putative archaic-specific variant) for each segment, and the length of the segment in bp. We used least-squares regression to fit a linear model for inverse density as a function of the length and the square root of the length since the relationship with length was not linear. The intercept of this fitted model is the inverse density adjusted for length. We calculated approximate standard errors for the adjusted inverse density by bootstrapping over chromosomes, refitting the linear model for each bootstrap sample. For haplotype length, we first determined introgressed haplotype lengths as the lengths of the subintervals over which individuals carry putatively introgressed alleles, and averaged these lengths for each segment. We also determined the average frequency of the putative archaic-specific alleles in each segment, and the length of the segment in bp. We used least-squares regression to fit a linear model for average haplotype length as a function of the average archaic-specific allele frequency of the segment, the segment length, and the square root of the segment length. The intercept of this fitted model is the average haplotype length adjusted for frequency and segment length. We calculated approximate standard errors for the adjusted average haplotype length by bootstrapping over chromosomes, refitting the linear model for each bootstrap sample.

Adaptive introgression

Following previous work (Gittelman et al., 2016; Racimo et al., 2015; Sankararaman et al., 2014, 2016; Vernot and Akey, 2014), we report genomic regions containing putative introgressed alleles with high frequency in our analysis of the 1000 Genomes Eurasian populations.

We selected two high frequency regions in each population as follows: We first removed all alleles with frequency less than 0.3. Then for each introgressed segment we removed alleles with frequency less than 0.2 below the maximum introgressed allele frequency in the segment. After this frequency pruning we removed segments that didn't have either at least 10 variants that could be compared to the Altai Neanderthal genome and over 50% matching to that genome, or at least 10 variants that can be compared to the Denisovan genome and over 40% matching to that genome. Then we selected the two segments containing the highest frequency alleles. The region boundaries are the first and last introgressed alleles in the segment that have frequency no more than 0.2 below the highest introgressed allele frequency in the segment. When regions for two populations overlapped, we took the intersection. We report the selected high frequency regions in Table S1.

For each putatively introgressed variant site within those regions, we retrieved basewise conservation phyloP scores (Pollard et al., 2010; Siepel et al., 2006) for multiple alignments of 99 vertebrate genomes to the human genome (phyloP100wayAll) available at the UCSC Genome Browser Conservation track. We further checked whether the variants with the highest PhyloP scores are gene expression QTLs using the GTEx Portal, and we checked whether these variants are coding variants using dbSNP.

DATA AND SOFTWARE AVAILABILITY

Sprime software: http://faculty.washington.edu/browning/sprime.html
Segments of introgression detected in 1000 Genomes and SGDP Papuan data: https://doi.org/10.17632/y7hyt83vxr.1

Cell

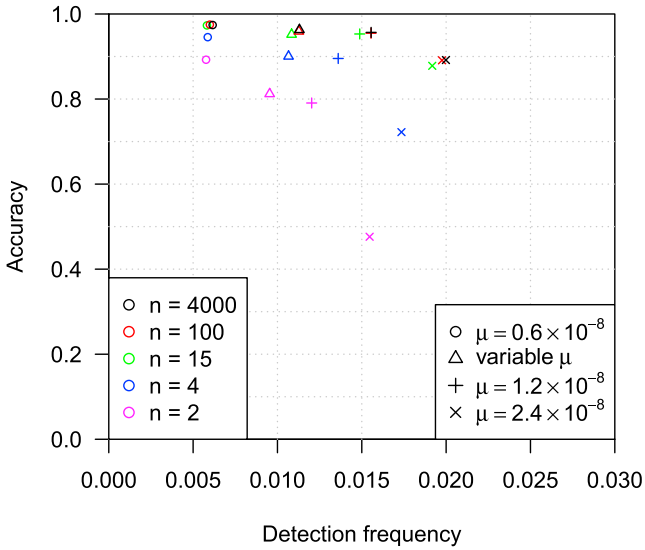


Figure S1. Power and Accuracy in Simulated Data as a Function of Mutation Rate and Sample Size, Related to Figure 1

The simulated mutation rate (μ, per bp per meiosis) varies from low (0.6×10^{-8}) to high (2.4×10^{-8}), as well as variable (uniformly distributed between 0 and 2.4×10^{-8}, with mean 1.2×10^{-8}, and a new rate randomly selected every 10 kb). Sample size varies from 2 to 4000 individuals. Accuracy is the proportion of alleles estimated to be introgressed that are truly introgressed. Detection frequency is the proportion of haplotypes with detected introgression after removing false positive results, and has a maximum possible value of 0.03 (the simulated admixture proportion).

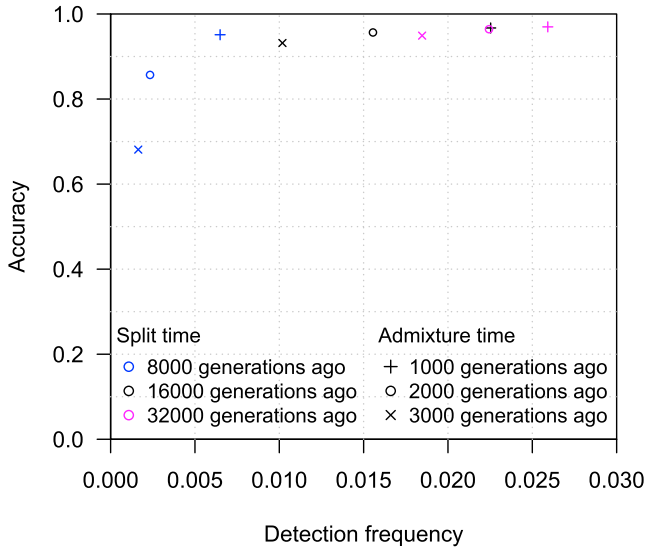


Figure S2. Detection Frequency and Accuracy in Simulated Data as a Function of Population Split and Admixture Times, Related to Figure 1

The time of the split of the archaic population from the human population and the time of admixture from the archaic population into the out-of-Africa population are varied. When the admixture time is 3000 generations ago, we move the out-of-Africa event time to 3100 generations ago (the baseline value is 2400 generations ago). The baseline simulation used in other results in the main paper is the one represented by the black circle (split time 16000 generations ago and introgression time 2000 generations ago). Accuracy is the proportion of putative introgressed alleles that are truly introgressed. Detection frequency is the proportion of haplotypes with detected introgression after removing false positive results, and has a maximum possible value of 0.03 (the simulated admixture proportion).

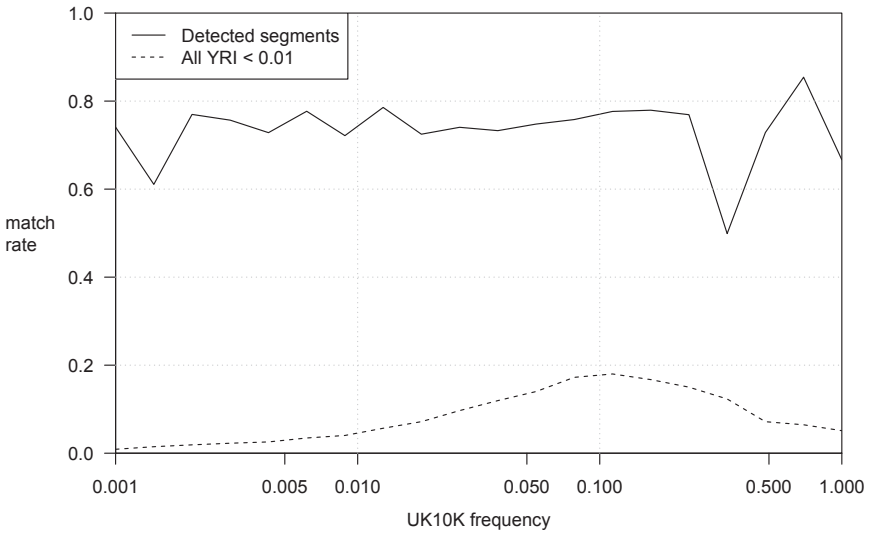


Figure S3. Rate of Matching to the Altai Neanderthal Genome as a Function of Allele Frequency in the UK10K Data, Related to Figure 3

The y axis shows the proportion of alleles that match the Altai Neanderthal genome, and the x axis gives the allele frequency in the UK10K (plotted on a log scale). All alleles included in these analyses have frequency < 0.01 in the West African Yoruban (YRI) outgroup. The solid line is for putative archaic-specific alleles from our analyses, while the dashed line is for all alleles in the data with frequency < 0.01 in YRI.

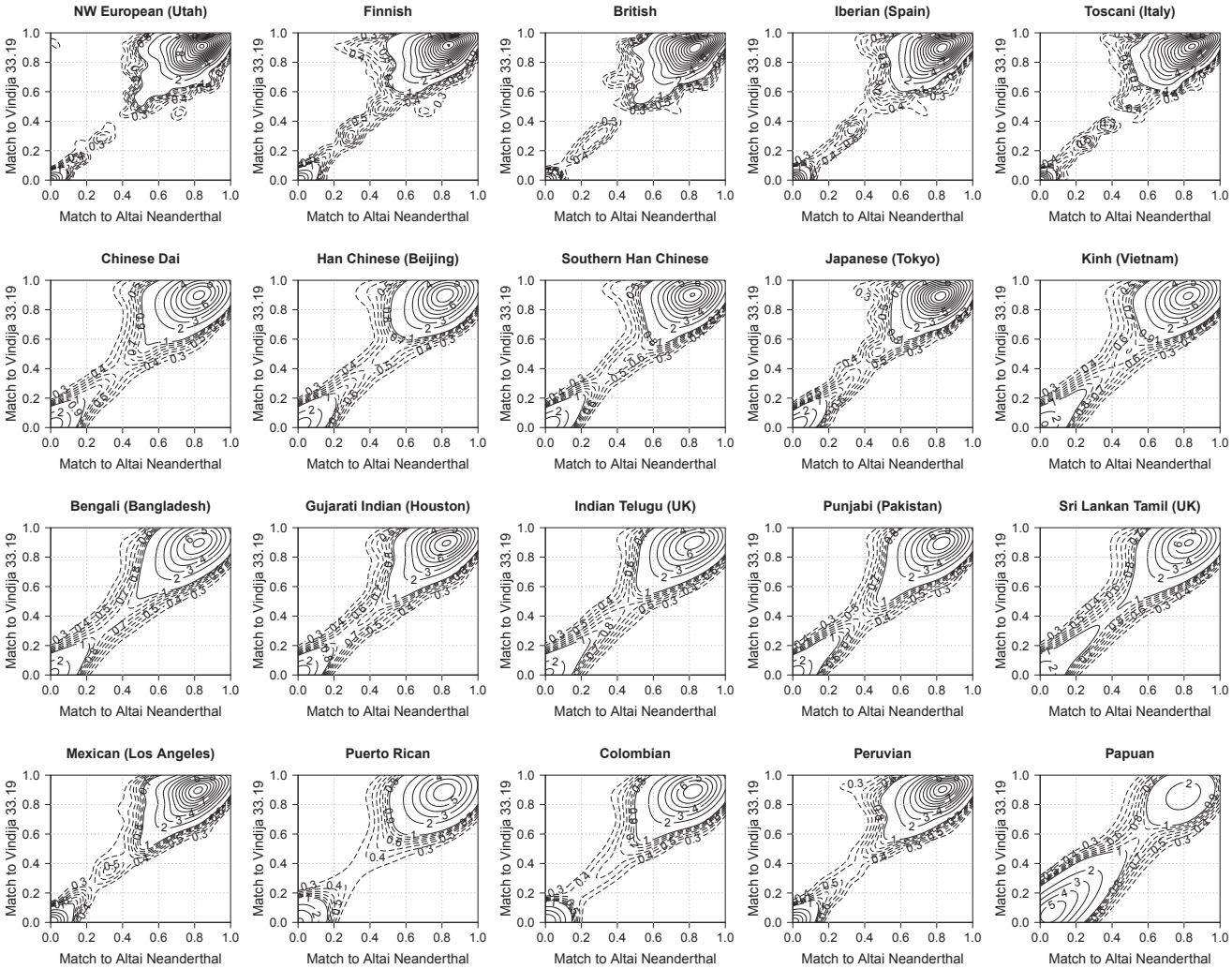


Figure S4. Contour Density Plots of Match Proportion of Introgressed Segments to the Altai Neanderthal and Vindija 33.19 Neanderthal, Related to Figure 4

The match proportion is the proportion of putative archaic-specific alleles in a segment that match the given archaic genome, excluding alleles at positions masked in the archaic genome sequence. Segments with at least 10 variants not masked the Altai Neanderthal genome and at least 10 variants not masked in the Vindija 33.19 genome are included. Numbers inside the plots indicate the height of the density corresponding to each contour line. Contour lines are shown for multiples of 1 (solid lines). In addition, contour lines for multiples of 0.1 between 0.3 and 0.9 (dashed lines) are shown for additional detail. European populations are given in the first row, East Asian populations in the second row, South Asian populations in the third row, and American and SGDP Papuan populations in the final row. Additional information about the populations can be found in Table 1.