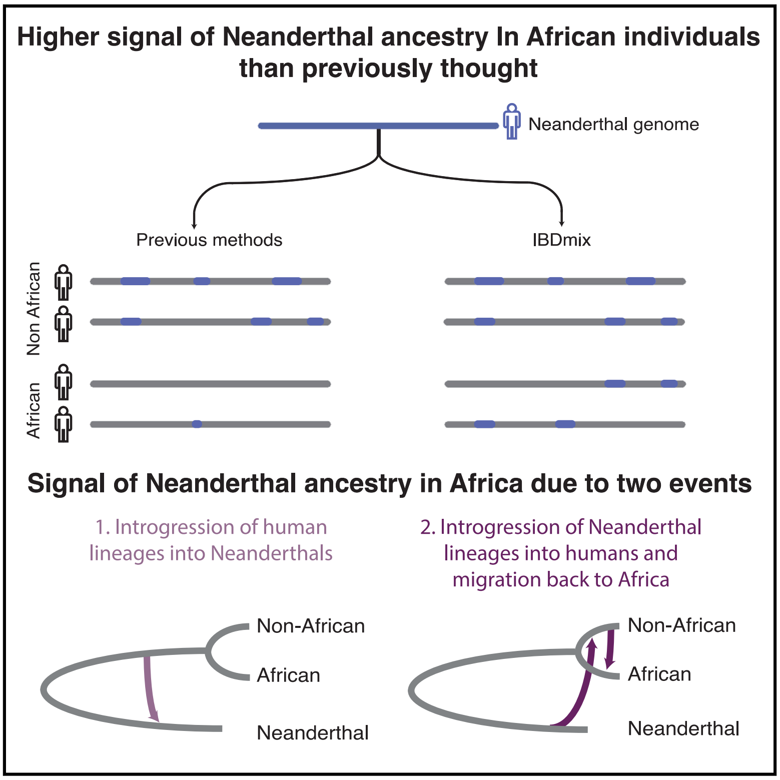
Cell

Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals

Graphical Abstract


Higher signal of Neanderthal ancestry In African individuals than previously thought
Neanderthal genome
Previous methods
IBDmix
Non African
African
Signal of Neanderthal ancestry in Africa due to two events
1. Introgression of human lineages into Neanderthals
2. Introgression of Neanderthal lineages into humans and migration back to Africa
Non-African
African
Neanderthal
Non-African
African
Neanderthal

Authors

Lu Chen, Aaron B. Wolf, Wenqing Fu, Liming Li, Joshua M. Akey

Correspondence

jakey@princeton.edu

In Brief

Detecting archaic introgression in modern humans without using an unadmixed reference panel reveals higher Neanderthal ancestry in African individuals than previously seen and suggests that back-to-Africa migrations contributed to this signal.

Highlights

- IBDmix detects archaic ancestry without using a modern human reference population
- African individuals have a stronger Neanderthal ancestry signal than previously thought
- Evidence of back-to-Africa migrations contributing to Neanderthal ancestry in Africans
- Variation in non-African Neanderthal ancestry has been overestimated

Chen et al., 2020, Cell *180*, 1–11
February 20, 2020 © 2020 Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.01.012

CellPress

Please cite this article in press as: Chen et al., Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals, Cell (2020), https://doi.org/10.1016/j.cell.2020.01.012

Cell

Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals

Lu Chen,[1,4] Aaron B. Wolf,[1,2,4] Wenqing Fu,[3] Liming Li,[1] and Joshua M. Akey[1,5,*]

[1]The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[3]1 Microsoft Way, Redmond, WA 98052, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: jakey@princeton.edu

https://doi.org/10.1016/j.cell.2020.01.012

SUMMARY

Admixture has played a prominent role in shaping patterns of human genomic variation, including gene flow with now-extinct hominins like Neanderthals and Denisovans. Here, we describe a novel probabilistic method called IBDmix to identify introgressed hominin sequences, which, unlike existing approaches, does not use a modern reference population. We applied IBDmix to 2,504 individuals from geographically diverse populations to identify and analyze Neanderthal sequences segregating in modern humans. Strikingly, we find that African individuals carry a stronger signal of Neanderthal ancestry than previously thought. We show that this can be explained by genuine Neanderthal ancestry due to migrations back to Africa, predominately from ancestral Europeans, and gene flow into Neanderthals from an early dispersing group of humans out of Africa. Our results refine our understanding of Neanderthal ancestry in African and non-African populations and demonstrate that remnants of Neanderthal genomes survive in every modern human population studied to date.

INTRODUCTION

Studies of ancient DNA are transforming our understanding of human evolutionary history and, in particular, how admixture has shaped past and present patterns of human genomic variation (Nielsen et al., 2017; Pääbo, 2014; Vattathil and Akey, 2015; Vernot and Pääbo, 2018). Of particular interest has been the discovery that admixture with archaic hominins occurred multiple times throughout human history (Green et al., 2010; Meyer et al., 2012; Prüfer et al., 2014; Reich et al., 2010). In particular, approximately 2% of all non-African ancestry is derived from Neanderthals (Green et al., 2010; Meyer et al., 2012; Prüfer et al., 2014; Sankararaman et al., 2016; Vernot et al., 2016; Wall et al., 2013), with Oceanic populations having an additional 2%–4% of ancestry attributable to gene flow with Denisovans (Browning et al., 2018; Mallick et al., 2016; Sankararaman et al., 2016; Vernot et al., 2016).

The ability to identify introgressed hominin sequence in the genomes of modern humans enables inferences about the functional, evolutionary, and phenotypic significance of archaic admixture. For example, the genomic distribution of surviving Neanderthal and Denisovan lineages has been influenced by purifying selection (Harris and Nielsen, 2016; Juric et al., 2016), which has purged introgressed sequence that was deleterious in modern humans. Indeed, some exceptionally large regions depleted of archaic ancestry (also referred to as “archaic deserts”) have been identified and may be due to selection (Sankararaman et al., 2014; Sankararaman et al., 2016; Vernot and Akey, 2014; Vernot et al., 2016). There is also strong evidence that some Neanderthal and Denisovan sequences were beneficial (Dannemann et al., 2016; Huerta-Sánchez et al., 2014; Mendez et al., 2012a, 2012b; Racimo et al., 2017; Racimo et al., 2015) and were rapidly driven to high frequency in modern human populations by a process known as adaptive introgression (Dannemann et al., 2017; Gittelman et al., 2016; McCoy et al., 2017; Simonti et al., 2016). In general, however, the functional impacts of introgressed sequences, how they have been shaped by selection, and how they have influenced modern human health and disease are only beginning to be explored.

Moreover, a consistent observation in all studies of archaic hominin admixture is that East Asian populations have approximately 20% more Neanderthal ancestry compared to Europeans (Nielsen et al., 2017; Sankararaman et al., 2014; Sankararaman et al., 2016; Vernot and Akey, 2014; Vernot et al., 2016; Wall et al., 2013). Numerous models have been invoked to explain this difference, including the interaction of demography and selection (Kim and Lohmueller, 2015; Lazaridis et al., 2016; Sankararaman et al., 2014), dilution by non-admixed populations (Lazaridis et al., 2016; Meyer et al., 2012), or additional population-specific admixture events (Kim and Lohmueller, 2015; Vernot and Akey, 2015; Villanea and Schraiber, 2019). Accurately determining variation in Neanderthal ancestry among non-African populations has important implications for refining our understanding of admixture between modern human ancestors and Neanderthals.

Despite the methodological progress that has been made to identify introgressed hominin sequence, opportunities for further development of statistical tools abound and may result in novel

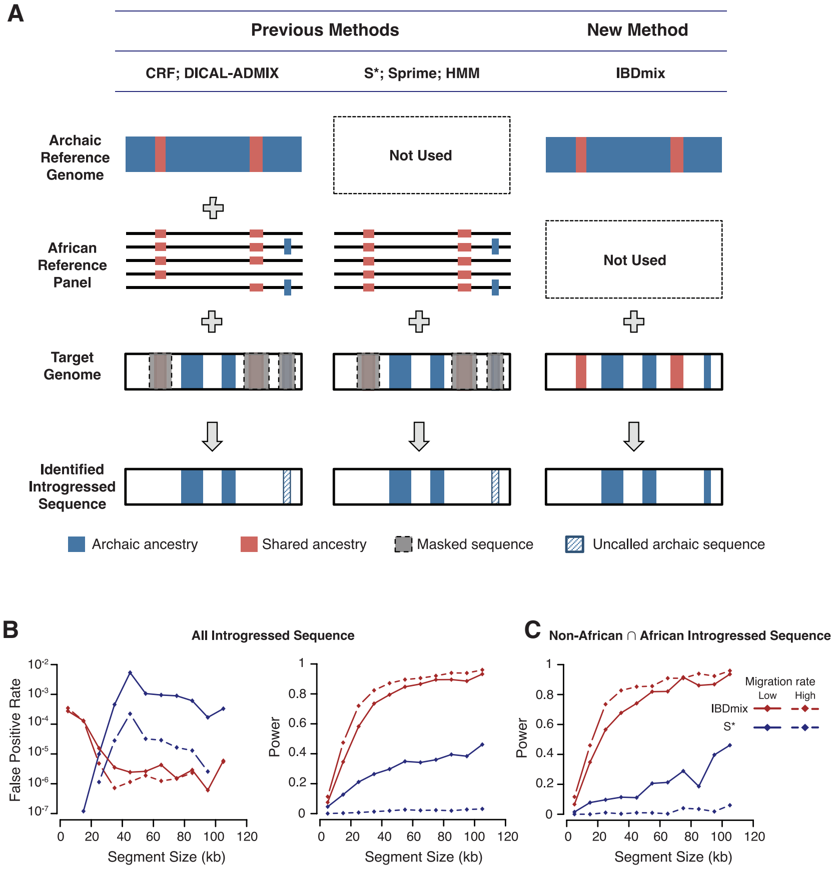


Figure 1. Evaluation of IBDmix Performance and Comparison to Previous Methods

(A) Summary of IBDmix workflow compared to previous methods for identifying introgressed archaic sequences in modern human genomes.

(B and C) Comparison of IBDmix performance to S^* using simulated data generated from models with a low back-migration rate (1.7×10^{-5}/generation) and high back-migration rate (5×10^{-4}/generation). In (B), power and false-positive rates are calculated for all simulated Neanderthal segments in non-Africans. In (C), we show the power to detect a Neanderthal segment in non-Africans conditional on it also being present in Africans.

insights. For example, a recent extension of the *S** framework revealed two waves of Denisovan admixture in East Asian populations that were not previously detectable (Browning et al., 2018). To this end, we describe a novel method for detecting Neanderthal ancestry in modern humans that does not require an unadmixed reference human panel, which we refer to as IBDmix. We apply IBDmix to genotype data from a large set of modern human individuals from Eurasia, America, and Africa. We make novel discoveries regarding Neanderthal ancestry in Africans and re-examine the relative levels of Neanderthal ancestry in Eurasian populations. We also replicate, extend, and discover new instances of adaptive introgression that may offer insight into human evolution and phenotypic variation in modern humans.

RESULTS

Evaluating the Power and Robustness of IBDmix

Methods that identify introgressed Neanderthal lineages in modern humans must differentiate between sequences shared with Neanderthals because of ancient hybridization or because of a shared common ancestor. Previous approaches, such as *S** (Plagnol and Wall, 2006; Vernot and Akey, 2014), CRF (Sankararaman et al., 2014), diCal-admix (Steinrücken et al., 2018), and HMM (Skov et al., 2018), use an "unadmixed" modern reference panel, commonly an African population such as Yoruba (YRI), to control for false positives due to shared ancestry by "masking" putative archaic sequence present in the reference panel and the target sample. If the reference panel carries introgressed Neanderthal sequence, this will result in missing Neanderthal sequence in the target sample (Figure 1A). Our new method IBDmix, which is based on identity by descent (IBD), does not use a modern reference panel (Figure 1A). IBDmix calculates the probabilities that a variant site in a modern individual is and is not shared IBD with a reference archaic genome, while accounting for genotyping errors in the reference archaic and modern human sequences (STAR Methods; Table S1). The ratio of these probabilities is used to construct a single-site LOD score, where higher values indicate a greater likelihood that a modern individual's genotype is shared IBD with the reference archaic genome. IBDmix uses a dynamic programming algorithm to sum together single-site LOD scores and maximize this score in order to identify introgressed segments (STAR Methods). The false-positive rate for IBDmix is controlled by the LOD score threshold and length of introgressed segments considered. Unlike existing methods that require phased sequence data, IBDmix works on unphased genotype data, making it more computationally tractable by avoiding time-consuming preprocessing and inaccuracies caused by phasing errors. It should be noted, however, that accurate estimates of allele frequency are required to calculate the probability of IBD, and so IBDmix cannot be used on individual genomes or in small sample sizes. In practice, we found that a minimum of ten individuals is sufficient for robust inferences (STAR Methods; Table S2).

We evaluated IBDmix's performance and operating characteristics using simulated data generated from a previously inferred realistic demographic model and compared it to results using *S** (STAR Methods; Figure S1). As expected, IBDmix's false-positive rate decreases and power increases as the introgressed segment size increases (Figure 1B). Compared to *S**, IBDmix has a lower false-positive rate and higher power for all introgressed segment sizes >30 kb (Figure 1B). Specifically, for introgressed segment sizes >30 kb, the power of IBDmix is >60%

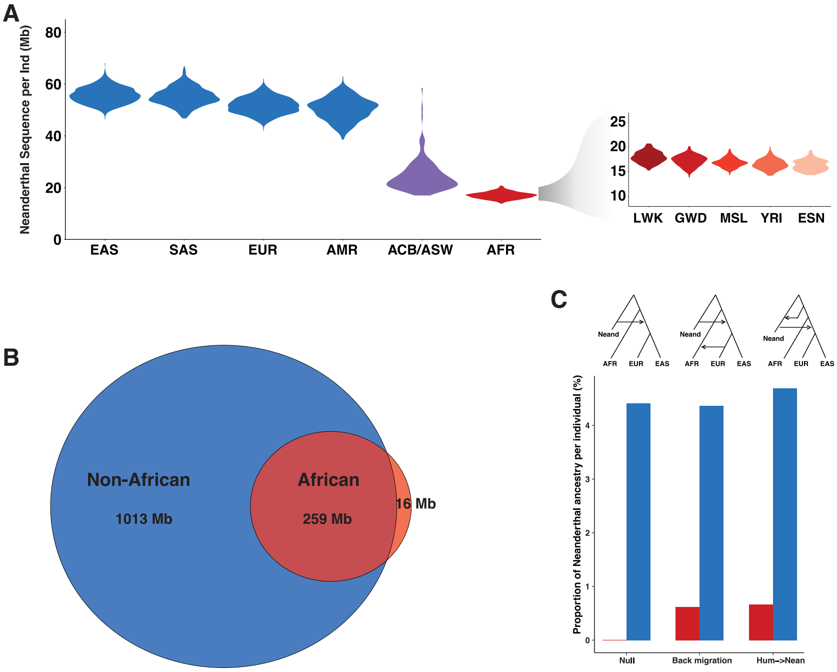


Figure 2. Neanderthal Introgressed Sequence Detected in 1000 Genomes Project Populations

(A) Violin plots showing the amount of Neanderthal sequence called per individual across geographically diverse populations from the 1000 Genomes Project. Non-African, African admixed, and African populations are shown in blue, purple, and red, respectively. The inset figure shows the amount of Neanderthal sequence per individual for five African subpopulations.

(B) Venn diagram showing the amount of overlap in identified Neanderthal sequence in non-African and African populations.

(C) Bar plot showing the proportion of Neanderthal ancestry per individual in non-African (blue) and African (red) populations in different simulated models.

with an FDR ≤10% (Figures 1B and S1B). Note that the power and FDR of IBDmix in non-African populations are not influenced by gene flow from non-Africans into Africans, whereas they do have a large effect on *S** (Figures 1B and 1C). The power to detect introgressed sequence in non-African populations is particularly low for *S** when this sequence is also found in the reference population (Africans), whereas IBDmix maintains power (Figure 1C). This observation implies that biases may arise in methods that use a modern human reference panel, as the power to detect introgressed sequence will be a function of its presence in the reference panel.

We also tested the impact of genetic variation and mis-specification of recombination rates on IBDmix using simulated data. The performance of IBDmix improved overall with higher mutation rates (Figure S1C). As expected, we observed a noticeable improvement for shorter segments (FPR, FDR, and power; Figure S1C). In testing the effect of recombination rate on IBDmix performance, we used data generated from a model with no Neanderthal introgression. We evaluated the FPR of IBDmix under models with a recombination rate equal to the genome-wide average (1cM/Mb) and models 1/10th that rate (0.1cM/Mb). For larger segments (≥40 kb), we observed marginally higher false-positive rates in situations with the reduced recombination rate (Table S3).

Previous studies have identified the introgressing Neanderthal population as a sister clade of the sequenced Altai Neanderthal (Malaspinas et al., 2016; Prüfer et al., 2017). We therefore tested how IBDmix would perform when the reference archaic genome is distantly related to the introgressing archaic. We simulated models with two Neanderthal lineages representing an introgressing lineage and a sampled reference lineage (non-introgressing lineage) and varied the split time between these two populations (STAR Methods). We observed a small decrease in power and FPR using the non-introgressing Neanderthal as the reference genome, but overall performance measures remained consistent (Figure S1D).

In summary, IBDmix has higher power and lower FDR compared to *S** and is robust to reference population biases. In the following, unless otherwise noted, we used a LOD score threshold of 4 and a minimum segment size of 50 kb, which provides a reasonable tradeoff between power and false-positive rate (Figure S1B).

IBDmix Reveals Substantial Amounts of Neanderthal Signal in Africans and Nearly Uniform Levels in Non-African Populations

We applied IBDmix to samples from the 1000 Genomes Project (Auton et al., 2015), collected from geographically diverse populations, and used the Altai Neanderthal reference genome (Prüfer et al., 2014) to identify introgressed Neanderthal sequence in these individuals. After filtering (STAR Methods), we identified 110.98 Gb of Neanderthal sequence among 2,504 modern individuals. When overlapping introgressed segments are merged, this equates to 1.29 Gb of unique Neanderthal sequence.

Because IBDmix does not use a putatively unadmixed modern reference population, we were able to robustly identify regions of apparent Neanderthal sequence in African populations for the first time (Figure 2A). Surprisingly, we identified on average 17 Mb of Neanderthal sequence per individual in the African samples analyzed, and this value was similar across the mostly northern African subpopulations represented in the dataset (ranging from 16.4 Mb/individual in ESN to 18.0 Mb/individual in LWK; Figure 2A; Table S4). Furthermore, we observed a significant overlap of sequence identified in Africans with that in non-Africans (Figure 2B). Specifically, of the Neanderthal sequence identified in African samples, more than 94% was shared with non-Africans.

We also recovered a substantial amount of Neanderthal sequence in non-African samples across populations. Notably, we found similar levels of Neanderthal ancestry in Europeans (51 Mb/individual), East Asians (55 Mb/individual), and South Asians (55 Mb/individual) (Figure 2A; Table S4). Surprisingly, we observed only a modest enrichment (8%) of Neanderthal ancestry in East Asian compared to European individuals. This contrasts with previous reports that have indicated ~20% enrichment of Neanderthal ancestry in East Asians compared to Europeans (Sankararaman et al., 2014; Sankararaman et al., 2016; Vernot and Akey, 2014; Wall et al., 2013). The observed level of East Asian enrichment was even smaller (~3%) when we were less conservative in our filtering methods (Table S5). We compared the Neanderthal sequences in non-African individuals identified by IBDmix (merged regions) to those identified by previous methods, including *S**, diCal-admix, and CRF, for individuals shared in all these studies. Approximately 80% of the sequences overlapped between the IBDmix callset and the other callsets (Figure S2).

Back-Migration with Non-Africans and Pre-out-of-Africa Human-to-Neanderthal Gene Flow Contribute to Apparent Neanderthal Ancestry in Africans

Given the unexpectedly large amounts of Neanderthal sequence identified in African individuals, we next performed analyses to understand their origins. To rule out systematic biases, we first called Denisovan sequence in African individuals using IBDmix (STAR Methods) and only identified 1.2 Mb/individual of Denisovan sequence in African samples (Table S6). This is similar to the amount of Denisovan sequence called in non-African individuals (~1Mb/individual) and considerably lower than the amount of Neanderthal sequence identified by IBDmix in African individuals. We also performed extensive simulations and found that the signal of Neanderthal ancestry in Africans was unlikely to be explained by false positives due to shared ancestry (Figure 2C).

We next considered two demographic models that could plausibly generate signals of Neanderthal ancestry in Africans that are detectable by IBDmix. Specifically, we studied models where non-African individuals, who carry Neanderthal sequences inherited from hybridization, migrated back to Africa and models of human-to-Neanderthal gene flow due to an early pre-out-of-Africa (pre-OOA) dispersal of modern humans (Hubisz et al., 2019; Kuhlwilm et al., 2016). We found that IBDmix is sensitive to both back migrations and pre-OOA gene flow from modern humans to Neanderthals (Figure 2C).

We therefore explicitly tested whether putative Neanderthal sequences identified in Africans were more likely to be explained by back-migration from non-Africans into Africa or by pre-OOA human-to-Neanderthal gene flow. To differentiate these scenarios, we compared the empirical data to simulated data, analyzing a variety of sequence characteristics (Figure 3). Specifically, we simulated genotype data under a series of demographic models that included Neanderthal admixture into non-Africans, increasing levels of back-migration from Europeans into Africans, and gene flow from a pre-OOA human lineage into Neanderthals at varying time points. We then identified introgressed sequence for these models using IBDmix. We compared the empirical and simulated data across features including introgressed segment length, frequency of introgressed segments in the African population that are shared with non-Africans, and the ratio of East Asian Neanderthal ancestry to European Neanderthal ancestry before and after masking Neanderthal sequence shared between Africans and non-Africans.

In the empirical data, segments identified in Africans (YRI) that are shared with non-Africans (EAS and EUR) have a distribution of segment sizes more similar to that of non-African calls and also occur predominantly at high frequency (>10%) in the African population (Figure 3). As noted previously, there is only a small enrichment (<10%) for Neanderthal ancestry in East Asians compared to Europeans without masking sequence shared with Africans. When shared sequence is masked, however, this enrichment increases to ~18% (Figure 3).

These features are not replicated in either models with back-migration or human-to-Neanderthal gene flow alone. Specifically, while features like the distribution of segment lengths and the frequency of African segments in the African population are replicated in models with human-to-Neanderthal gene flow, only models with back-migration rates elevated in comparison to standard demographic estimates (5×10^{-5}/generation) can replicate the enrichment of East Asian Neanderthal ancestry when masking shared African sequence. A model that combines both of these events, elevated back migration and human-to-Neanderthal gene flow, matches the empirical data best across all features. In summary, these data indicate that both pre-OOA human-to-Neanderthal gene flow and elevated historic back-migration contribute to the signal of Neanderthal ancestry detected in Africans.

Back-Migration from European Ancestors Introduced Neanderthal Sequence into African Populations

To further confirm the role of back-migration in introducing Neanderthal sequence into African populations, we examined the rate of overlap between called Neanderthal segments and non-African ancestry tracks in African samples. We hypothesized that if the Neanderthal sequence in Africans was introduced by back-migration from ancestors of contemporary Europeans, then there should be enrichment for overlap of Neanderthal segments and European ancestry segments in African samples. To test this hypothesis, we compared data from chromosome 1 for all 504 African samples in our analysis. For each individual, we identified tracks of European and East Asian ancestry using *RFMix* (Maples et al., 2013) and measured the rate of overlap with identified Neanderthal segments in the same individual (Figure 4A). We averaged these rates of overlap to calculate empirical rates of overlap for European ancestry and East Asian ancestry separately (Figure 4B). We found the rate of overlap with European ancestry to be highly significant (permutation $p < 0.0001$), while the rate of overlap with East Asian ancestry was not (permutation $p > 0.05$) (Figure 4B). These data are consistent with the hypothesis that back-migration contributes to the signal of Neanderthal ancestry in Africans. Furthermore, the data indicate that this back-migration came after the split of Europeans and East Asians, from a population related to the European lineage.

Previously Inferred Differences in Neanderthal Ancestry Between East Asians and Europeans Were Biased due to Unaccounted-for Back-Migration

Previous methods that have relied on unadmixed modern reference populations, like *S**, have reported >20% enrichment of Neanderthal sequence in East Asians compared to Europeans

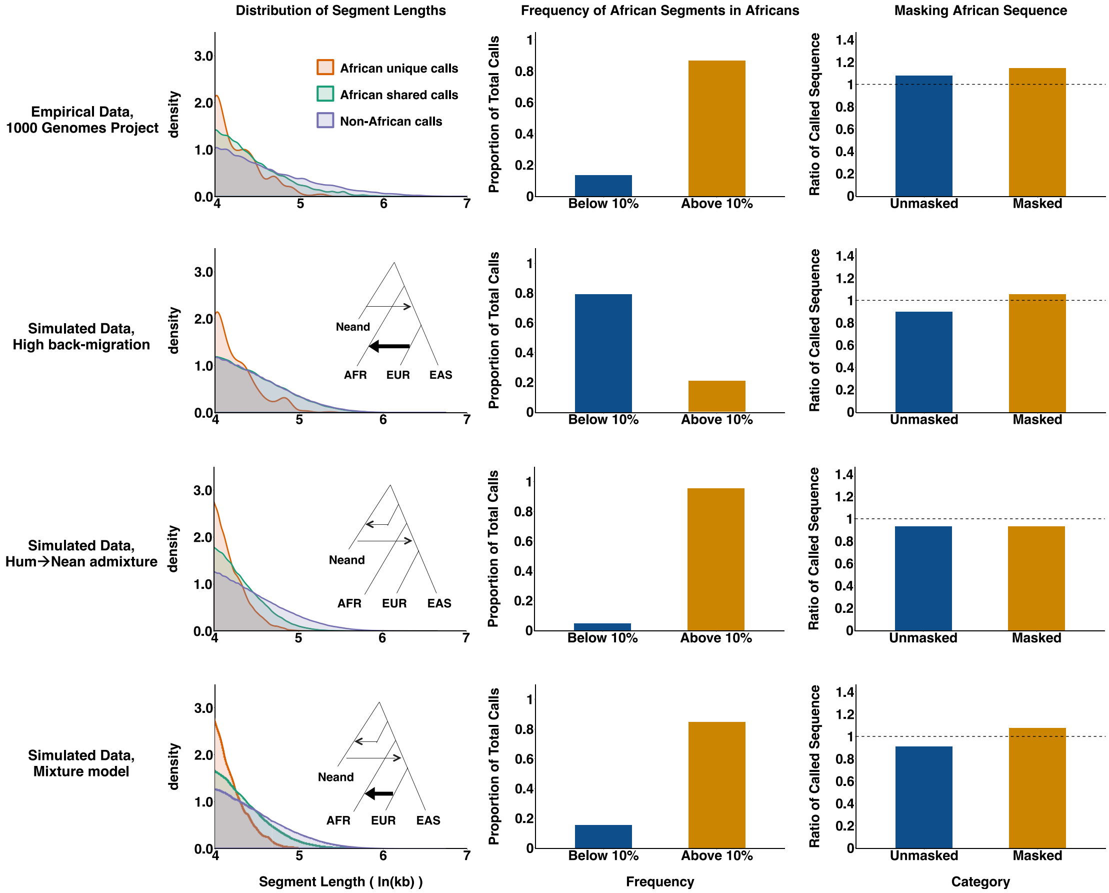


Figure 3. Neanderthal Segments Identified in Africans Are a Consequence of Back-Migration and Human-to-Neanderthal Gene Flow

Features of the empirical data were compared to data simulated under a model of back-migration, human-to-Neanderthal gene flow, and a mixture of both models (see the STAR Methods). From left to right, the distribution of Neanderthal segment lengths, frequency of segments in Africans that segregate in Africans and non-Africans, and the ratio of East Asian to European Neanderthal ancestry with and without masking sequence shared with Africans.

(Figure 5A). However, results from IBDmix show only 8% enrichment of Neanderthal sequence in East Asians compared to Europeans (Figure 5A). This level of enrichment is robust to changes in the segment size cutoff (30 kb, 40 kb, 50 kb) used for IBDmix calling (Table S5). To better understand the discrepancy between IBDmix and previous inferences, we first removed Neanderthal sequence called by IBDmix in Europeans and East Asians that was shared with Africans (YRI) and replicated an 18% enrichment of Neanderthal ancestry in East Asians compared to Europeans (Figure 5A). This result shows that our observation of similar levels of Neanderthal ancestry in Europeans and East Asians is due to no longer masking Neanderthal sequence shared with Africans.

In the IBDmix callset for Africans, Europeans, and East Asians, there is a large enrichment of Neanderthal sequence shared exclusively between Africans and Europeans compared with the sequence shared exclusively between Africans and East Asians (Figure 5B). As a proportion of the total amount of Neanderthal sequence for each population, 7.2% of European sequence is shared exclusively with Africans, which is substantially higher than the 2% of East Asian sequence shared exclusively with Africans (Figure 5B). The disproportionate level of sharing between Africans and Europeans is consistent even after down-sampling the recovered Neanderthal segments in Europeans to match the total coverage of Neanderthal sequence in East Asians (STAR Methods). This imbalance in the proportion of exclusively shared sequence between African and non-African populations directly contributes to the biased Neanderthal ancestry estimates in previous methods that use an African reference panel.

We also examined how the reference panel size for *S** affects Neanderthal ancestry estimates by bootstrap resampling the Yoruba samples in 1000 Genomes Project data (n = 108) and re-analyzing chromosome 1 for Europeans and East Asians (Figure 5C). We generated multiple reference panels based on different sample sizes and re-called Neanderthal sequence for European and East Asian individuals using the *S**-pipeline and the new reference panels. We compared the total *S**-sequence called for each sample to the average amount of *S**-sequence called for samples using a reference panel of 1 individual. Increasing the reference panel size showed a significant reduction ($p < 2 \times 10^{-16}$) in the amount of Neanderthal sequence

Please cite this article in press as: Chen et al., Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals, Cell (2020), https://doi.org/10.1016/j.cell.2020.01.012

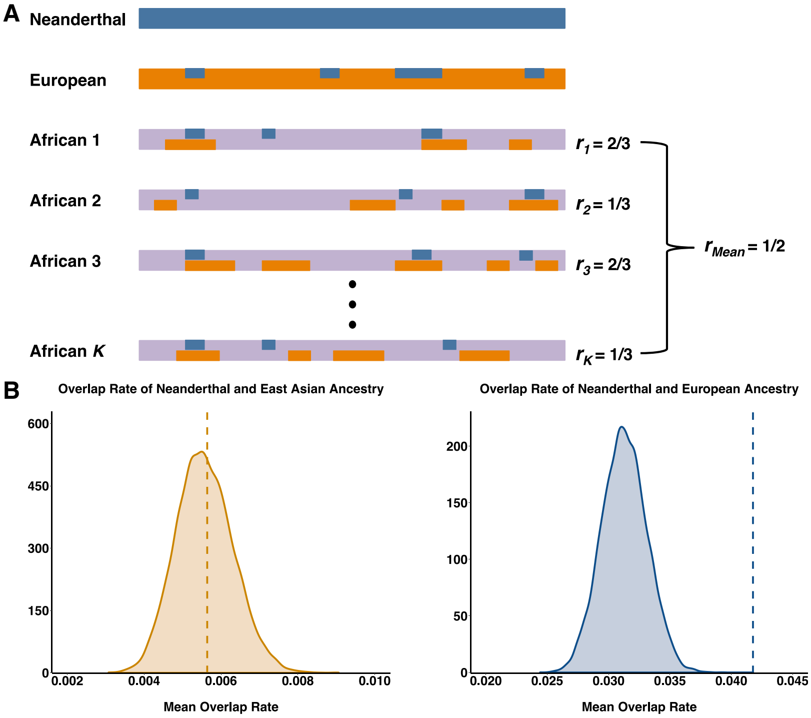


Figure 4. Enrichment in Overlap of Neanderthal Segments and European Ancestry Segments in African Individuals

(A) Schematic of how an enrichment of European ancestry overlap was assessed. For each African individual, data from chromosome 1 were analyzed for tracks of Neanderthal and European ancestry. For each individual, the rate of overlap between Neanderthal segments and European segments was calculated, and the mean across all African individuals was taken as the empirical value.

(B) Distributions of the mean rate of overlap from permuted data for European ancestry and East Asian ancestry, with the empirical values demarcated as dashed lines. The rate of overlap for European ancestry is highly significant ($p < 0.0001$), while the rate of overlap for East Asian ancestry is not ($p > 0.05$).

called per individual. In addition, when comparing the amounts of Neanderthal sequence identified in Europeans and East Asians, increasing the reference panel size decreased the amount detected for both populations, but there was a greater loss in Europeans than in East Asians. Using a reference sample larger than 10 led to an apparent 20% enrichment of Neanderthal ancestry in East Asians compared to Europeans, as previously reported. Simulations of European to African back-migration using rates consistent with standard demographic models also generate a significant enrichment of Neanderthal ancestry in East Asians compared to Europeans when the data are analyzed with S^*, so long as back-migration occurs after the split of European and East Asian lineages ($p < 8 \times 10^{-7}$; Figure S3). Collectively, these results show that Neanderthal ancestry estimates in East Asians and Europeans were biased due to unaccounted for back-migrations from European ancestors into Africans.

IBDmix Reveals Novel Insights into Signatures of Adaptive Introgression

Admixture with Neanderthals may have provided a mechanism for modern humans to acquire novel adaptive variation. Previous analyses have reported population-specific high-frequency introgressed Neanderthal haplotypes, which may be instances of adaptive introgression (Dannemann et al., 2017; Gittelman et al., 2016; Racimo et al., 2015; Simonti et al., 2016) or the reintroduction of alleles lost in the modern human lineage (Rinker et al., 2019). We examined our IBDmix callset for similar findings. We leveraged population-level derived allele frequencies of variants that overlapped calls made by IBDmix and matched the Neanderthal allele, in order to detect Neanderthal haplotypes with unusually large differences in frequency between populations.

Specifically, for variants that intersected identified Neanderthal segments, we calculated the differences in the derived allele frequencies between Europeans and East Asians, Africans and Europeans, and Africans and East Asians. We then took an outlier approach to identify loci with allele frequency differences in the 99th percentile. We further filtered on loci where the derived allele matched the Neanderthal allele. Overall, we identified 38 non-African-specific high-frequency haplotypes and 13 African-specific high-frequency haplotypes (Table S7). We compared these identified high-frequency haplotypes with previously identified high-frequency haplotypes (Gittelman et al., 2016) and the presence of previously reported GWAS SNPs.

Of the 38 non-African-specific high-frequency Neanderthal haplotypes we identified, 19 were previously reported by Gittelman et al. (2016), including well-known targets of adaptive introgression like *WDR88*, *POU2F3*, and *TLR1/6/10* (Figure 6A and 6B). Intriguingly, we also identified 31 high-frequency haplotypes shared by Africans and Europeans, including *TRIM55* (Figure 6C; Table S7). These haplotypes would have been undetected in previous methods that relied on unadmixed reference human panels. Furthermore, we were for the first time able to detect African-specific high-frequency Neanderthal haplotypes (Figure 6D; Table S7). The 13 African-specific high-frequency Neanderthal haplotypes we identified show enrichment for genes involved in immunological function (e.g., *IL22RA1* and *IFNLR1*) and ultraviolet-radiation sensitivity (e.g., *DDB1* and *IL22RA1*) (Keeney et al., 1993; Kim et al., 2017). While some high frequency Neanderthal-like variants in Africans may derive from human-to-Neanderthal gene flow, only one of the high-frequency haplotypes shared by Africans and Europeans (chr3:89,587,868–90,134,709) overlaps a locus previously identified as introgressed from modern humans into the Altai Neanderthal (Kuhlwilm et al., 2016), and none of our detected African-specific high-frequency haplotypes do. These novel findings provide insight into the evolutionary history of these populations, the selective pressures they faced, and current variation in health and disease.

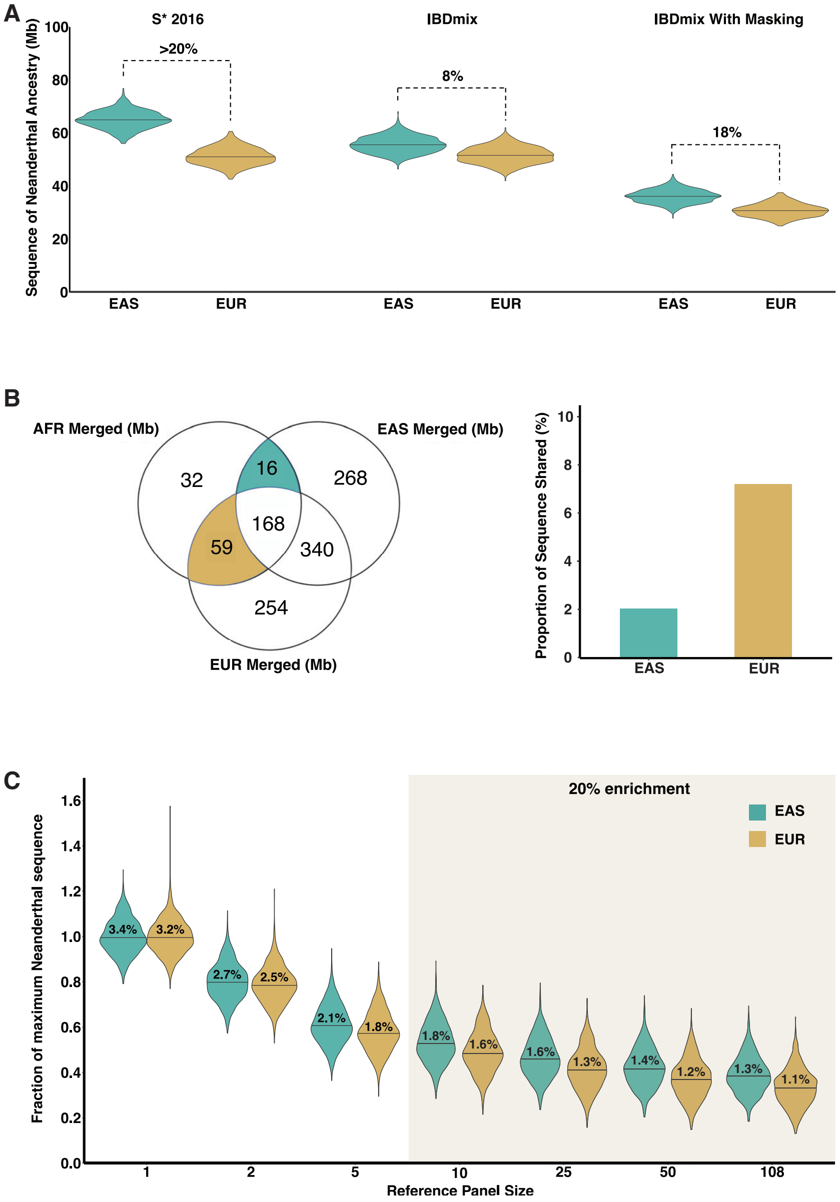


Figure 5. Disproportionate Sharing of Neanderthal Sequence Differentially Biases Estimates of Neanderthal Ancestry

(A) Violin plots showing enrichment of Neanderthal ancestry in East Asians compared to Europeans for *S** and for IBDmix with and without masking Neanderthal sequence shared with Yoruba.

(B) Venn diagram illustrating the amount of sequence shared among Africans and non-Africans. The bar plot shows the amount of exclusively shared sequence between Africans and non-Africans as a proportion of the total amount of sequence for each population.

(C) Violin plot showing the decreasing amount of Neanderthal sequence identified in East Asian and European individuals by *S** with increasing African reference-panel size.

IBDmix Refines Loci Depleted of Neanderthal Ancestry

Previous analyses have identified large (>10 Mb) autosomal regions of the genome that are significantly depleted of Neanderthal ancestry in all non-African populations (Sankararaman et al., 2014, 2016; Vernot and Akey, 2014; Vernot et al., 2016). These large "deserts" of archaic introgressed sequence appear at frequencies greater than expected under neutral models. We analyzed our IBDmix call set to see if we could replicate previous findings or determine if deserts were a function of previous methodological biases. Following previously described methods to identify archaic deserts, we analyzed our IBDmix callset from both African and non-African samples (STAR Methods). We replicated 4 of the 6 previously reported deserts of Neanderthal sequence, including the deserts that contain *FOXP2* (chr7) and *ROBO1* and *ROBO2* (chr3) (Table S8; Figure S4). Moreover, the four replicated deserts are the same regions previously shown to also be significantly depleted of Denisovan ancestry. Thus, depletions of archaic ancestry seem to be a general feature of the data and are not likely due to methodological issues in identifying introgressed sequence. It is noteworthy that including all African samples, a subset (YRI), or none does not dramatically change the distribution of the frequencies of large deserts. This is consistent with the observation that the African Neanderthal sequence is predominantly a subset of non-African segments.

DISCUSSION

We developed a novel approach to identify an introgressed hominin sequence that persists in the genomes of modern humans, and we show that it performs well compared to existing methods. The main novelty of IBDmix is that compared to previous methods, it does not use an unadmixed reference panel. As such, we were able to make unbiased inferences about signals of Neanderthal ancestry in African populations, which are a combination of genuine introgressed Neanderthal sequences and human sequences present in the Neanderthal genome. We also demonstrate that back-migrations to Africa confounded previous estimates of variation in Neanderthal ancestry among non-African populations. Furthermore, we confirmed and refined genomic regions significantly depleted of Neanderthal ancestry, as well as putative targets of adaptive introgression, including several loci that were previously not detectable when using an African reference population.

It is important to note, however, that IBDmix has several limitations. In particular, IBDmix requires an archaic reference

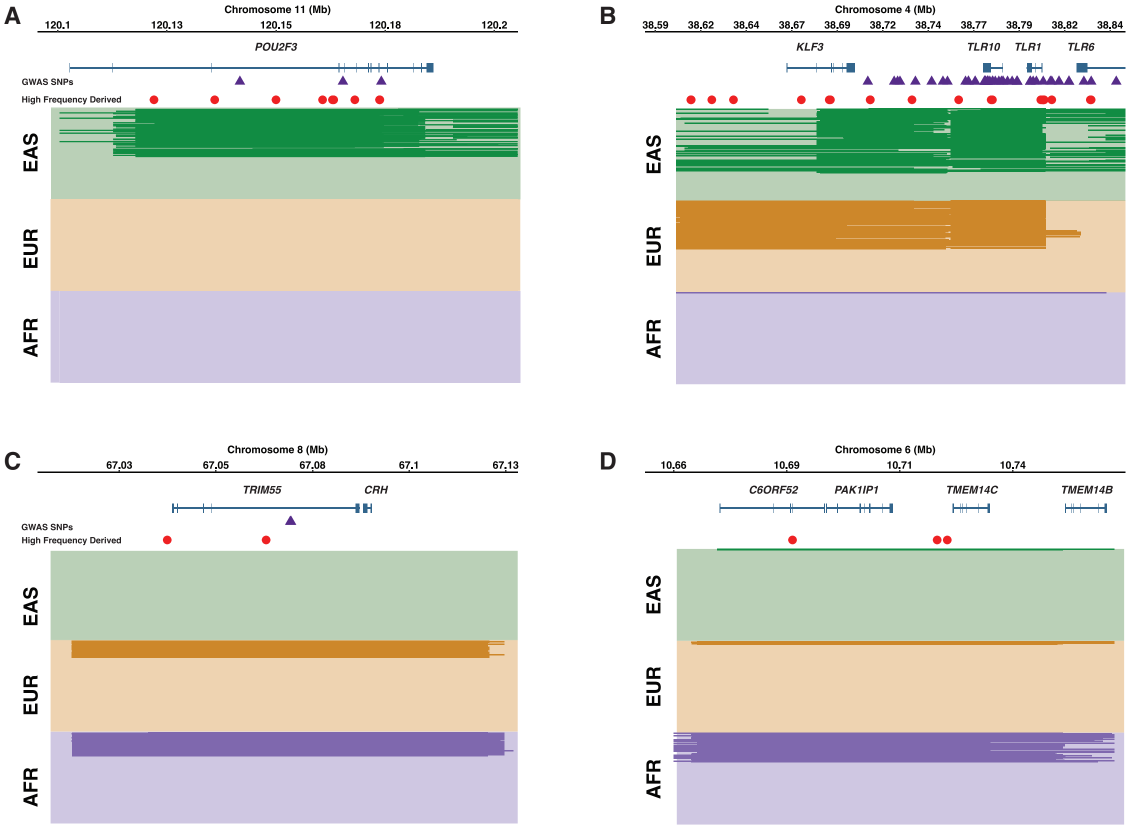


Figure 6. Population-Specific High-Frequency Introgressed Segments

In all plots, each row is an individual and is organized by population. Neanderthal segments called by IBDmix are plotted in dark green (EAS), orange (EUR), or purple (AFR). GWAS SNPs are shown as purple triangles and populations-specific high-frequency-derived alleles (DAF > 40%) that match the Altai reference genome are shown as red circles. In (A) and (B), examples of high-frequency introgressed segments detected in East Asian and European populations are shown for the *POU2F3* and the *TLR1/6/10* cluster.

(C) An example of a high-frequency Neanderthal segment shared between Europeans and Africans at *TRIM55*. This haplotype, identified by IBDmix, is missed by methods that mask sequence shared by African and non-African populations.

(D) Example of an African-specific high-frequency haplotype that spans multiple genes.

genome and therefore is not suitable for discovering introgressed sequence from unknown or unsequenced hominin lineages. IBDmix also requires that populations be analyzed separately, and that a sufficiently large sample size be used, in order to robustly estimate population allele frequencies, assign LOD scores, and determine IBD (simulations suggest a minimum of ten individuals; Table S2). Additionally, recombination rate heterogeneity across the genome and between populations can influence IBDmix segment size cutoffs. Consequently, it will be difficult to apply IBDmix to individual genomes or ancient human samples, where the sample size is limited and estimates of allele frequencies and recombination rates are imprecise. As such, IBDmix complements existing approaches for identifying introgressed sequences in modern humans.

Applying IBDmix to geographically diverse populations revealed two unexpected observations. First, we discovered a stronger than expected signal of Neanderthal ancestry among African individuals. Specifically, among the 1000 Genomes African populations, we identified approximately 17 Mb of putative Neanderthal sequence per individual (Figure 2; Table S4), whereas previous inferences found considerably less than a megabase (ranging from 0.026 Mb in Esan to 0.5 Mb in Luhya) (Vernot et al., 2016). Accordingly, African individuals have approximately 33% as much detected sequence compared to non-African individuals. The higher signal of Neanderthal ancestry in African individuals is not entirely unexpected, as recent studies have suggested that assumptions about Neanderthal ancestry in Africans may have led to underestimates (Lorente-Galdos et al., 2019; Petr et al., 2019). Moreover, even early estimates of Neanderthal ancestry in non-Africans noted that there was likely some amount of Neanderthal sequence in Africans (Green et al., 2010; Sánchez-Quinto et al., 2012; Wang et al., 2013), albeit not at the magnitude we find. Furthermore, it is increasingly recognized that gene flow occurred among structured populations across the African continent (Scerri et al., 2018; Schlebusch et al., 2012; Skoglund et al., 2017), and Eurasian ancestry is found across Africa (Pickrell et al., 2014). Even early diverging groups like the Khoe-San have up to 30% ancestry from recent admixture with East Africans and Eurasians (Schlebusch et al., 2017). Therefore, it will not be surprising if Neanderthal ancestry, due to back-migrations, is present at varying levels across the African continent.

Our results also provide strong evidence that human sequence in the Neanderthal genome also contributes to the signal of the Neanderthal ancestry we detect in Africans. Previous studies have noted the genetic contribution of a pre-out-of-Africa gene-flow event from humans into Neanderthals (Hubisz et al., 2019; Kuhlwilm et al., 2016). The timing of this event, however, has been under debate, with estimates being revised from ~100 ka (Kuhlwilm et al., 2016; Prüfer et al., 2017) to ~150 ka (Kuhlwilm et al., 2016; Prüfer et al., 2017), and now perhaps as early as 250 ka (Hubisz et al., 2019). Our own data are most consistent with models of human-to-Neanderthal gene flow between 100 and 150 ka, as IBDmix does not detect any signal in simulations with earlier gene flow. However, our results do not preclude earlier instances of gene flow, only that IBDmix is not powered to detect them. Thus, it is tempting to speculate that perhaps there were multiple waves of pre-OOA dispersals and admixture between modern humans and Neanderthals, although additional data are needed to make more definitive inferences.

The second major insight afforded by IBDmix is that levels of Neanderthal ancestry among non-African populations are more uniform than previous estimates. Specifically, as opposed to the 20% enrichment of Neanderthal sequence previously found in East Asians compared to Europeans (Kim and Lohmueller, 2015; Lazaridis et al., 2016; Meyer et al., 2012; Vernot and Akey, 2015), we only find an approximately 8% enrichment (Figure 5A; Table S4). We show that the reason for this discrepancy is that previous inferences using an African reference population underestimated the amount of Neanderthal sequence in Europeans. Due to historical back-migrations preferentially from ancestral European populations, Neanderthal sequence has been disproportionately under-called in present-day Europeans compared to East Asians. We believe the modest 8% enrichment of Neanderthal sequence found by IBDmix is most parsimoniously explained by a single wave of Neanderthal admixture occurring after the out-of-Africa dispersal. Variation in Neanderthal ancestry could be attributable to later dilution by unadmixed populations (Lazaridis et al., 2016). In particular, present-day European populations are thought to be a mixture of three ancestral groups, one of which had ancestry from a Basal Eurasian lineage that had little or no Neanderthal ancestry (Lazaridis et al., 2014). Previous studies found that dilution could not explain Neanderthal ancestry differences as large as 20% (Kim and Lohmueller, 2015; Vernot and Akey, 2015) but can readily account for the modest differences we now find. Note that, however, our data do not preclude the possibility of additional, population-specific admixture events with Neanderthals. Numerous instances of admixture events are known from ancient human samples, even though these individuals did not contribute genetically to contemporary human populations (Fu et al., 2015; Yang et al., 2017). Nonetheless, the majority of Neanderthal ancestry can likely be explained by a single wave of admixture in the population ancestral to all non-Africans.

In summary, our data show that out-of-Africa and in-to-Africa dispersals must be accounted for when interpreting archaic hominin ancestry in contemporary human populations. It is notable that Neanderthal sequences have been identified in every contemporary modern human genome analyzed to date. Thus, the legacy of gene flow with Neanderthals likely exists in all modern humans, highlighting our shared history.

STAR★METHODS

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

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - ○ Details of the IBDmix Algorithm
  - ○ Simulation Study
  - ○ Whole Genome Sequence Data
- QUANTIFICATION AND STATISTICAL ANALYSES
  - ○ Refining Neanderthal Callset by Using Denisovan Sequences as a Negative Control
  - ○ Replicating Regions Significantly Depleted of Neanderthal Introgressed Sequence
  - ○ Comparing Simulated Data to Empirical Data
  - ○ Reference Panel Size Effect on S^* Admixture Estimates
  - ○ Identifying High-Frequency Introgressed Haplotypes From IBDmix Data
  - ○ Calculating the Rate of Overlap Between Neanderthal Calls and European Ancestry in African Samples
  - ○ Calculating rate of exclusively shared sequence between African and non-African populations
  - ○ Comparing callsets from different methods in shared individuals
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.01.012.

ACKNOWLEDGMENTS

Research reported in this publication was supported by NIH grant R01 GM110068 to J.M.A. We would like to thank current and past members of the Akey Lab for their feedback and contributions to this work, especially Selina Vattathil, Troy Comi, and Lance Parsons.

AUTHOR CONTRIBUTIONS

J.M.A. and W.F. planned and J.M.A. directed this study. W.F. derived the analytical theory and wrote the software. W.F., L.C., A.B.W., and L.L. developed the methods and conducted the analyses. L.C., A.B.W., and J.M.A. wrote the manuscript. All authors contributed to editing the manuscript.

DECLARATION OF INTERESTS

J.M.A. is a paid consultant of Glenview Capital. W.F., L.C., A.B.W., and L.L. have no competing interests to declare.

Received: June 13, 2019
Revised: November 13, 2019
Accepted: January 7, 2020
Published: January 30, 2020

REFERENCES

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.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., and Akey, J.M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell *173*, 53–61.e59.

Dannemann, M., Andrés, A.M., and Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. Am. J. Hum. Genet. *98*, 22–33.

Dannemann, M., Prüfer, K., and Kelso, J. (2017). Functional implications of Neandertal introgression in modern humans. Genome Biol. *18*, 61.

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.

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.

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.

Harris, K., and Nielsen, R. (2016). The Genetic Cost of Neanderthal Introgression. Genetics *203*, 881–891.

Hubisz, M.J., Williams, A.L., and Siepel, A. (2019). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. bioRxiv.

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.

Keeney, S., Chang, G.J., and Linn, S. (1993). Characterization of a human DNA damage binding protein implicated in xeroderma pigmentosum E. J. Biol. Chem. *268*, 21293–21300.

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.

Kim, Y., Lee, J., Kim, J., Choi, C.W., Hwang, Y.I., Kang, J.S., and Lee, W.J. (2017). The pathogenic role of interleukin-22 and its receptor during UVB-induced skin inflammation. PLoS ONE *12*, e0178567.

Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de la Rasilla, M., et al. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. Nature *530*, 429–433.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424.

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

Lorente-Galdos, B., Lao, O., Serra-Vidal, G., Santpere, G., Kuderna, L.F.K., Arauna, L.R., Fadhlaoui-Zid, K., Pimenoff, V.N., Soodyall, H., Zalloua, P., et al. (2019). Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biol. *20*, 77.

Malaspinas, A.S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., et al. (2016). A genomic history of Aboriginal Australia. Nature *538*, 207–214.

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.

Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.

McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell *168*, 916–927.e12.

Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2012a). Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. Mol. Biol. Evol. *29*, 1513–1520.

Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2012b). A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. Am. J. Hum. Genet. *91*, 265–274.

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.

Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. Nature *541*, 302–310.

Pääbo, S. (2014). The human condition-a molecular approach. Cell *157*, 216–226.

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. Proc. Natl. Acad. Sci. USA *116*, 1639–1644.

Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. USA *111*, 2632–2637.

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

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.

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

Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sánchez, E., and Nielsen, R. (2017). Archaic Adaptive Introgression in TBX15/WARS2. Mol. Biol. Evol. *34*, 509–524.

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.

Rinker, D.C., Simonti, C.N., McArthur, E., Shaw, D., Hodges, E., and Capra, J.A. (2019). Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations. bioRxiv.

Sánchez-Quinto, F., Botigué, L.R., Civit, S., Arenas, C., Avila-Arcos, M.C., Bustamante, C.D., Comas, D., and Lalueza-Fox, C. (2012). North African populations carry the signature of admixture with Neandertals. PLoS ONE *7*, e47765.

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.

Scerri, E.M.L., Thomas, M.G., Manica, A., Gunz, P., Stock, J.T., Stringer, C., Grove, M., Groucutt, H.S., Timmermann, A., Rightmire, G.P., et al. (2018). Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? Trends Ecol. Evol. *33*, 582–594.

Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G., et al. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science *338*, 374–379.

Schlebusch, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Vicente, M., Steyn, M., Soodyall, H., et al. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science *358*, 652–655.

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., Thompson, J.C., Prendergast, M.E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., et al. (2017). Reconstructing Prehistoric African Population Structure. Cell *171*, 59–71.e21.

Skov, L., Hui, R., Shchur, V., Hobolth, A., Scally, A., Schierup, M.H., and Durbin, R. (2018). Detecting archaic introgression using an unadmixed outgroup. PLoS Genet. *14*, e1007641.

Steinrücken, M., Spence, J.P., Kamm, J.A., Wieczorek, E., and Song, Y.S. (2018). Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. Mol. Ecol. *27*, 3873–3888.

Tennessen, J.A., Bigham, A.W., O’Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Vattathil, S., and Akey, J.M. (2015). Small Amounts of Archaic Admixture Provide Big Insights into Human History. Cell *163*, 281–284.

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., and Pääbo, S. (2018). The Predecessors Within. Cell *173*, 6–7.

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.

Villanea, F.A., and Schraiber, J.G. (2019). Multiple episodes of interbreeding between Neanderthal and modern humans. Nat. Ecol. Evol. *3*, 39–44.

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.

Wang, S., Lachance, J., Tishkoff, S.A., Hey, J., and Xing, J. (2013). Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from Non-African populations. Genome Biol. Evol. *5*, 2075–2081.

Yang, M.A., Gao, X., Theunert, C., Tong, H., Aximu-Petri, A., Nickel, B., Slatkin, M., Meyer, M., Pääbo, S., Kelso, J., and Fu, Q. (2017). 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. Curr. Biol. *27*, 3202–3208.e9.

Please cite this article in press as: Chen et al., Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals, Cell (2020), https://doi.org/10.1016/j.cell.2020.01.012

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
1000 Genomes project data, phase 3 version 5a	1000 Genomes project (Auton et al., 2015)	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
Altai Neanderthal, Altai Denisovan genomes	Kay Prufer (Prüfer et al., 2014)	http://cdna.eva.mpg.de/neandertal/altai/
IBDmix calls for 1000 Genomes populations	This paper	https://drive.google.com/drive/folders/1mDQaDFS-j22Eim5_y7LAsTTNt5GWsoow?usp=sharing
Software and Algorithms		
IBDmix for detection of Neanderthal introgressed sequence	This paper	https://github.com/PrincetonUniversity/IBDmix
S* for detection of Neanderthal introgressed sequence	Benjamin Vernot (Vernot et al., 2016)	https://github.com/bvernot/freezing-archer
Msprime coalescent-based simulation software	Jerome Kelleher (Kelleher et al., 2016)	https://github.com/tskit-dev/msprime
R	The R Project for Statistical Computing	https://www.r-project.org/
RFMix for detecting non-African ancestry	B.K. Maples (Maples et al., 2013)	https://github.com/slowkoni/rfmix.git

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Joshua Akey (jakey@princeton.edu). This study did not generate unique reagents.

METHOD DETAILS

Details of the IBDmix Algorithm

Overview

As an input, IBDmix requires format-converted genotype data from whole genome sequencing for one archaic reference individual and a group of modern humans as the target genome. IBDmix is distinct from previous methods because it does not use a modern human unadmixed reference population to control for ILS between the archaic and modern human populations.

IBDmix is developed based on identity by descent (IBD), the principal that an identical sequence of alleles is shared by two individuals and inherited from a common ancestor. Proceeding site-by-site, IBDmix operates on one pair of archaic and modern human genomes at a time. At each position that passes variant filtering (described below), IBDmix estimates the probability of IBD between the archaic and modern sample based on allele frequencies and summarizes this as a LOD score. In order to identify putatively introgressed archaic segments in the modern genome, IBDmix applies a scanning algorithm based on dynamic programming to maximize the sum of LOD scores across a region above a pre-set threshold. Under this dynamic program, variants are added consecutively to calculate the sum of the LOD scores; expanding the interval until the sum of the LOD scores become a negative value. The region with the maximized LOD score (above the pre-set LOD threshold) is called as a putative introgressed segment in the modern individual. Scanning restarts from the next variant after the putative introgressed segment.

At completion, the output from IBDmix is a list of putatively introgressed segments and the probability of IBD between the archaic and modern human sample summarized as a maximized LOD score. Greater positive LOD scores reflect a higher probability of IBD across the specified region.

IBDmix LOD Score Calculation

Our IBDmix method is based on summing single site IBD LOD scores. We define the IBD LOD score for an allele to be the base 10 logarithm of the IBD likelihood divided by the non-IBD likelihood. Positive scores indicate evidence for IBD and negative scores indicate evidence against IBD. We use the alternative allele frequencies to compute the likelihood of the IBD model in which the modern human individual and archaic Neanderthal share one IBD allele, and of the non-IBD model in which they do not share any IBD allele. Approximate IBD and non-IBD likelihoods and their ratios under a model with independent errors in alleles are summarized in Table S1, and the derivations of these likelihoods and ratios are presented below (see Likelihood Estimation with Allele Error).

The scores in Table S1 are applied to variants that pass filtering (see Variant Filtering). To be conservative, we do not use excluded variants to determine the evidence for or against IBD. However, we impute the genotype data for modern humans where they are missing and the archaic genome is heterozygous or homozygous for the alternative allele. In particular, discordant homozygotes provide significant evidence against IBD, which adds important information without increasing the false-positive IBD detection rate.

For each pair of samples (one archaic and one modern human), we report all autosomal segments for which the sum of LOD scores within the interval reaches a maximum. We identify these segments by using a scanning algorithm based on dynamic programming. Because we are working on the log scale, summing IBD LOD scores corresponds to multiplying likelihood ratios.

Allele Error Rates in IBDmix Calculation

In IBDmix, we use an error model in which allele errors are independent. For archaic genomes we set $\eta = 0.01$ as allele error rate. For modern human genomes, the probability ϵ of incorrectly calling an allele depends on the minor allele frequency (MAF). For an allele with observed minor allele frequency f_B, the allele error rate is $\epsilon = \min\{\sigma, \rho f_B\}$. σ is the maximum allele error rate and ρ is the ratio between allele error rate and minor allele frequency. In our analyses, we set $\sigma = 0.002$ and $\rho = 2$. Accordingly, the allele error rate for human genomes is σ for higher frequency variants and is proportional to the observed minor allele frequency for lower frequency variants.

IBD Likelihood Estimation with Allele Error

In this section we derive estimates for the likelihood of observed genotypes for IBD and non-IBD modes under Hardy-Weinberg equilibrium when each allele for human genomes is observed incorrectly with $\epsilon \geq 0$ and each allele for archaic genomes is observed incorrectly with $\eta \geq 0$ and errors are independent. Under IBD model $P_o(\cdot|IBD)$ and $P(\cdot|I)$ where one archaic individual and one modern human share an allele identical by descent and under non-IBD model ($P_o(\cdot|nonIBD)$ and $P(\cdot|nI)$) where no alleles are identical by descent, individuals are ordered (first is archaic and second is modern human) and genotypes are unordered. $P_o(\cdot|IBD)$ denote the probabilities of a pair of observed genotypes (with error) while $P(\cdot|I)$ denote the corresponding probability for the true genotypes (without error). We assume that variants are biallelic, with reference allele A and alternative allele B. p_A and p_B are the observed frequency of each allele in the target modern population.

$$
\begin{aligned}
& P_o(AA,AA|IBD) = (1-\eta)^2(1-\epsilon)^2 P(AA,AA|I) \\
& + (1-\eta)^2\epsilon(1-\epsilon)P(AA,AB|I) \\
& + (1-\eta)^2\epsilon^2 P(AA,BB|I) \\
& + \eta(1-\eta)(1-\epsilon)^2 P(AB,AA|I) \\
& + \eta(1-\eta)\epsilon(1-\epsilon)P(AB,AB|I) \\
& + \eta(1-\eta)\epsilon^2 P(AB,BB|I) \\
& + \eta^2(1-\epsilon)^2 P(BB,AA|I) \\
& + \eta^2\epsilon(1-\epsilon)P(BB,AB|I) + \eta^2\epsilon^2 P(BB,BB|I) \\
& = (1-\eta)^2(1-\epsilon)^2 p_A + (1-\eta)^2\epsilon(1-\epsilon)p_B \\
& + \eta(1-\eta)(1-\epsilon)^2 p_A \\
& + \eta(1-\eta)(1-\epsilon)\epsilon + \eta(1-\eta)\epsilon^2 p_B \\
& + \eta^2\epsilon(1-\epsilon)p_A + \eta^2\epsilon^2 p_B \\
& = [(1-\eta)(1-\epsilon) + \eta\epsilon][(1-\epsilon)p_A + \epsilon p_B]
\end{aligned}
$$

$$
\begin{aligned}
&P_o(AA,AA|nonIBD)| = (1-\eta)^2(1-\epsilon)^2P(AA,AA|nI)\\
&+(1-\eta)^2\epsilon(1-\epsilon)P(AA,AB|nI)\\
&+(1-\eta)^2\epsilon^2P(AA,BB|nI)\\
&+\eta(1-\eta)(1-\epsilon)^2P(AB,AA|nI)\\
&+\eta(1-\eta)\epsilon(1-\epsilon)P(AB,AB|nI)\\
&+\eta(1-\eta)\epsilon^2P(AB,BB|nI)\\
&+\eta^2(1-\epsilon)^2P(BB,AA|nI)\\
&+\eta^2\epsilon(1-\epsilon)P(BB,AB|nI)\\
&+\eta^2\epsilon^2P(BB,BB|nI)\\
&=(1-\eta)^2(1-\epsilon)^2p_A^2\\
&+2(1-\eta)^2\epsilon(1-\epsilon)p_Ap_B+(1-\eta)^2\epsilon^2p_B^2\\
&+(1-\eta)^2(1-\epsilon)^2p_A^2\\
&+2\eta(1-\eta)(1-\epsilon)\epsilon p_Ap_B+\eta(1-\eta)\epsilon^2p_B^2\\
&+\eta^2(1-\epsilon)^2p_A^2+2\eta^2\epsilon(1-\epsilon)p_Ap_B+\eta^2\epsilon^2\\
&=[1-\eta(1-\eta)](1-\epsilon)^2p_A^2+2[1-\eta(1-\eta)]\epsilon(1-\epsilon)p_Ap_B+[1-\eta(1-\eta)]\\
&=[1-\eta(1-\eta)][(1-\epsilon)p_A+\epsilon p_B]^2
\end{aligned}
$$

$$
\begin{aligned}
&P_o(AA,AB|IBD)=(1-\eta)^2\left[(1-\epsilon)^2+\epsilon^2\right]P(AA,AB|I)+2(1-\eta)^2\epsilon(1-\epsilon)P(AA,AA|I)\\
&+2(1-\eta)^2\epsilon(1-\epsilon)P(AA,BB|I)+\eta(1-\eta)\left[(1-\epsilon)^2+\epsilon^2\right]P(AB,AB|I)+2\eta(1-\eta)\\
&\epsilon(1-\epsilon)P(AB,AA|I)+2\eta(1-\eta)\epsilon(1-\epsilon)P(AB,BB|I)+\eta^2\left[(1-\epsilon)^2+\epsilon^2\right]P(BB,AB|I)+2\eta^2\epsilon\\
&\Big(1\epsilon)P(BB,AA|I)+2\eta^2\epsilon(1-\epsilon)P(BB,BB|I)=(1-\eta)^2\left[(1-\epsilon)^2+\epsilon^2\right]p_B+2(1-\eta)^2\epsilon(1-\epsilon)p_A\\
&+\eta(1-\eta)+\eta^2\left[(1-\epsilon)^2+\epsilon^2\right]p_A+2\eta^2(1-\epsilon)\epsilon p_B=[1-2(1-\epsilon)\epsilon][\eta p_A+(1-\eta)p_B]\\
&+2(1-\epsilon)\epsilon[(1-\eta)p_A+\eta p_B]\ =(\epsilon+\eta-2\epsilon\eta)f_A+[1-(\epsilon+\eta-2\epsilon\eta)]f_B
\end{aligned}
$$

$$
\begin{aligned}
&P_o\Big(AA,AB|nonIBD)=(1-\eta)^2\left[(1-\epsilon)^2+\epsilon^2\right]P(AA,AB|nI\Big)\\
&+2(1-\eta)^2\epsilon(1-\epsilon)P(AA,AA|nI)+2(1-\eta)^2\epsilon\\
&(1-\epsilon)P\Big(AA,BB|nI)+\eta(1-\eta)\left[(1-\epsilon)^2+\epsilon^2\right]P(AB,AB|nI\Big)\\
&+2\eta(1-\eta)\epsilon(1-\epsilon)P(AB,AA|nI)+2\eta(1-\eta)\epsilon(1\\
&-\epsilon)P\Big(AB,BB|nI)+\eta^2\left[(1-\epsilon)^2+\epsilon^2\right]P(BB,AB|nI\Big)+2\eta^2\epsilon(1\\
&\epsilon)P\big(BB,AA|nI)+2\eta^2\epsilon(1-\epsilon)P(BB,BB|nI\big)=2\Big[(1-\epsilon)^2\\
&+\epsilon^2\Big]\left[(1-\eta)^2+\eta(1-\eta)+\eta^2\right]p_Ap_B+2\epsilon(1-\epsilon)\Big[(1-\eta)^2\\
&+\eta(1-\eta)+\eta^2\Big]p_A^2+2\epsilon(1-\epsilon)\left[(1-\eta)^2+\eta(1-\eta)+\eta^2\right]p_B^2\\
&=2[1-\eta(1-\eta)]\left[\epsilon(1-\epsilon)+(1-2\epsilon)^2p_Ap_B\right]\\
&=2[1-\eta(1-\eta)]f_Af_B
\end{aligned}
$$

$$\begin{gathered}
P_o\Bigl(AA,BB|IBD)=(1-\eta)^2(1-\epsilon)^2P(AA,BB|I\Bigr)+(1-\eta)^2 \\
\epsilon(1-\epsilon)P\Bigl(AA,AB|I)+(1-\eta)^2\epsilon^2P(AA,AA|I\Bigr)+\eta(1-\eta) \\
(1-\epsilon)^2P(AB,BB|I)+\eta(1-\eta)\epsilon(1-\epsilon)P(AB,AB|I)+\eta \\
(1-\eta)\epsilon^2P\Bigl(AB,AA|I)+\eta^2(1-\epsilon)^2P(BB,BB|I\Bigr)+\eta^2\epsilon \\
(1-\epsilon)P\bigl(BB,AB|I)+\eta^2\epsilon^2P(BB,AA|I\bigr)=(1-\eta)^2\epsilon(1-\epsilon)p_B \\
+(1-\eta)^2\epsilon^2p_A+\ \eta(1-\eta)(1-\epsilon)^2p_B+\eta(1-\eta)\epsilon(1-\epsilon)\epsilon \\
+\eta(1-\eta)\epsilon^2p_A+\eta^2(1-\epsilon)^2p_B+\eta^2\epsilon(1-\epsilon)p_A=[(1-\eta)\epsilon \\
+\eta(1-\epsilon)][(1-\epsilon)p_B+\epsilon p_A]
\end{gathered}$$

$$\begin{gathered}
P_o\Bigl(AA,BB|nonIBD)=(1-\eta)^2(1-\epsilon)^2P(AA,BB|nI\Bigr) \\
+(1-\eta)^2\epsilon(1-\epsilon)P(AA,AB|nI)+(1-\eta)^2\epsilon^2P(AA,AA|nI) \\
+\eta(1-\eta)(1-\epsilon)^2P(AB,BB|nI)+\eta(1-\eta)\epsilon(1-\epsilon)P(AB,AB|nI) \\
+\eta(1-\eta)\epsilon^2P\Bigl(AB,AA|nI)+\eta^2(1-\epsilon)^2P(BB,BB|nI\Bigr) \\
+\eta^2\epsilon(1-\epsilon)P\bigl(BB,AB|nI)+\eta^2\epsilon^2P(BB,AA|nI\bigr)=[1 \\
-\eta(1-\eta)][(1-\epsilon)p_B+\epsilon p_A]^2
\end{gathered}$$

$$\begin{gathered}
P_o\Bigl(AB,AA|IBD)=\Bigl[(1-\eta)^2+\eta^2\Bigr](1-\epsilon)^2P(AB,AA|I\Bigr) \\
+\Bigl[(1-\eta)^2+\eta^2\Bigr]\epsilon(1-\epsilon)P(AB,AB|I) \\
+\Bigl[(1-\eta)^2+\eta^2\Bigr]\epsilon^2P\bigl(AB,BB|I)+2\eta(1-\eta)(1-\epsilon)^2P(AA,AA|I\bigr) \\
+P(BB,AA|I)+2\eta(1-\eta)\epsilon(1-\epsilon)P(AA,AB|I)+P(BB,AB|I) \\
+2\eta(1-\eta)\epsilon^2P(AA,BB|I)+P(BB,BB|I)=(1-\epsilon)^2p_A \\
+\ \epsilon(1-\epsilon)+\epsilon^2p_B=f_A
\end{gathered}$$

$$\begin{gathered}
P_o\Bigl(AB,AA|nonIBD)=\Bigl[(1-\eta)^2+\eta^2\Bigr](1-\epsilon)^2P(AB,AA|nI\Bigr) \\
+\Bigl[(1-\eta)^2+\eta^2\Bigr]\epsilon(1-\epsilon)P(AB,AB|nI)+\Bigl[(1-\eta)^2 \\
+\eta^2\Bigr]\epsilon^2P\Bigl(AB,BB|nI)+2\eta(1-\eta)(1-\epsilon)^2P(AA,AA|nI\Bigr) \\
+P(BB,AA|nI) \\
+2\eta(1-\eta)\epsilon(1-\epsilon)P(AA,AB|nI)+P(BB,AB|nI)+2\eta \\
(1-\eta)\epsilon^2P(AA,BB|nI)+P(BB,BB|nI)=[1+2\eta(1-\eta)][(1 \\
-\epsilon)p_A+\epsilon p_B]^2=[1+2\eta(1-\eta)]f_A^2
\end{gathered}$$

$$\begin{gathered}
P_o\Bigl(AB,AB|IBD)=\Bigl[(1-\eta)^2+\eta^2\Bigr]\Bigl[(1-\epsilon)^2+\epsilon^2\Bigr]P(AB,AB|I\Bigr) \\
+2\Bigl[(1-\eta)^2+\eta^2\Bigr]\epsilon(1-\epsilon)P(AB,AA|I)+2\Bigl[(1-\eta)^2 \\
+\eta^2\Bigr]\epsilon(1-\epsilon)P\Bigl(AB,BB|I)+2\eta(1-\eta)\Bigl[(1-\epsilon)^2+\epsilon^2\Bigr]P(AA,AB|I\Bigr) \\
+P(BB,AB|I)+4\eta(1-\eta)\epsilon(1-\epsilon)P(AA,AA|I)+P(BB,AA|I) \\
+4\eta(1-\eta)\epsilon(1-\epsilon)P(AA,BB|I)+P(BB,BB|I)=\Bigl[(1-\eta)^2 \\
+\eta^2\Bigr]\ \Bigl[(1-\epsilon)^2+\epsilon^2\Bigr]+2\Bigl[(1-\eta)^2+\eta^2\Bigr]\epsilon(1-\epsilon)p_A+2\Bigl[(1-\eta)^2 \\
+\eta^2\Bigr]\epsilon(1-\epsilon)p_B+2\eta(1-\eta)\Bigl[(1-\epsilon)^2+\epsilon^2\Bigr]+4\eta(1-\eta) \\
\epsilon(1-\epsilon)p_A+4\eta(1-\eta)\epsilon(1-\epsilon)p_B\ =1
\end{gathered}$$

$$
\begin{aligned}
&P_o(AB,AB|nonIBD)\\
&=\Big[(1-\eta)^2+\eta^2\Big]\Big[(1-\epsilon)^2+\epsilon^2\Big]P(AB,AB|nI)+2\Big[(1-\eta)^2\\
&+\eta^2\Big]\epsilon(1-\epsilon)\ P(AB,AA|nI)+2\Big[(1-\eta)^2\\
&+\eta^2\Big]\epsilon(1-\epsilon)P(AB,BB|nI)+2\eta(1-\eta)\Big[(1-\epsilon)^2\\
&+\epsilon^2\Big]P(AA,AB|nI)\\
&+P(BB,AB|nI)+4\eta(1-\eta)\epsilon(1-\epsilon)P(AA,AA|nI)+P(BB,AA|nI)\\
&+4\eta(1-\eta)\epsilon(1-\epsilon)P(AA,BB|nI)+P(BB,BB|nI)=2[1\\
&+2\eta(1-\eta)]\ \Big[(1-\epsilon)^2+\epsilon^2\Big]p_Ap_B+2[1+2\eta(1-\eta)]\epsilon(1-\epsilon)p_A^2\\
&+2\eta(1-\eta)\epsilon(1-\epsilon)p_B^2=2[1+2\eta(1-\eta)]f_Af_B
\end{aligned}
$$

$$
\begin{aligned}
&P_o\Big(AB,BB|IBD)=\Big[(1-\eta)^2+\eta^2\Big](1-\epsilon)^2P(AB,BB|I\Big)\\
&+\Big[(1-\eta)^2+\eta^2\Big]\epsilon(1-\epsilon)P(AB,AB|I)\\
&+\Big[(1-\eta)^2+\eta^2\Big]\epsilon^2P\Big(AB,AA|I)+2\eta(1-\eta)(1-\epsilon)^2P(AA,BB|I\Big)\\
&+2\eta(1-\eta)\epsilon(1-\epsilon)P(AA,AB|I)+P(BB,AB|I)+2\eta(1\\
&-\eta)\epsilon^2P(AA,AA|I)+P(BB,AA|I)=\Big[(1-\eta)^2+\eta^2\Big](1-\epsilon)^2p_B\\
&+\Big[(1-\eta)^2+\eta^2\Big]\epsilon(1-\epsilon)+\Big[(1-\eta)^2+\eta^2\Big]\epsilon^2p_A+2\eta(1\\
&-\eta)(1-\epsilon)^2p_B+2\eta(1-\eta)\epsilon(1-\epsilon)+2\eta(1-\eta)\epsilon^2p_A\\
&=(1-\epsilon)^2p_B+\epsilon(1-\epsilon)+\epsilon^2p_A=f_B
\end{aligned}
$$

$$
\begin{aligned}
&P_o\Big(AB,BB|nonIBD)=\Big[(1-\eta)^2+\eta^2\Big](1-\epsilon)^2P(AB,BB|nI\Big)\\
&+\Big[(1-\eta)^2+\eta^2\Big]\epsilon(1-\epsilon)P(AB,AB|nI)+\Big[(1-\eta)^2\\
&+\eta^2\Big]\epsilon^2P\Big(AB,AA|nI)+2\eta(1-\eta)(1-\epsilon)^2P(AA,BB|nI\Big)\\
&+2\eta(1-\eta)\epsilon(1-\epsilon)P(AA,AB|nI)+P(BB,AB|nI)+2\eta(1\\
&-\eta)\epsilon^2P(AA,AA|nI)+P(BB,AA|nI)=[1+2\eta(1-\eta)]f_B^2
\end{aligned}
$$

$$
\begin{aligned}
&P_o\Big(BB,AA|IBD)=(1-\eta)^2(1-\epsilon)^2P(BB,AA|I\Big)+(1-\eta)^2\epsilon(1\\
&-\epsilon)P\Big(BB,AB|I)+(1-\eta)^2\epsilon^2P(BB,BB|I\Big)+\eta(1-\eta)(1\\
&-\epsilon)^2P(AB,AA|I)+\eta(1-\eta)\epsilon(1-\epsilon)P(AB,AB|I)+\eta(1\\
&-\eta)\epsilon^2P\Big(AB,BB|I)+\eta^2(1-\epsilon)^2P(AA,AA|I\Big)+\eta^2\epsilon(1\\
&-\epsilon)P\Big(AA,AB|I)+\eta^2\epsilon^2P(AA,BB|I\Big)=(1-\eta)^2\epsilon(1-\epsilon)p_A\\
&+(1-\eta)^2\epsilon^2p_B+\ \eta(1-\eta)(1-\epsilon)^2p_A+\eta(1-\eta)\epsilon(1-\epsilon)\epsilon\\
&+\eta(1-\eta)\epsilon^2p_B+\eta^2(1-\epsilon)^2p_A+\eta^2\epsilon(1-\epsilon)p_B=[(1-\eta)\epsilon\\
&+\eta(1-\epsilon)][(1-\epsilon)p_A+\epsilon p_B]
\end{aligned}
$$

$$\begin{aligned}
&P_o\Big(BB, AA|nonIBD) = (1-\eta)^2(1-\epsilon)^2 P(BB, AA|nI\Big)\\
&+(1-\eta)^2\epsilon(1-\epsilon)P\Big(BB, AB|nI) + (1-\eta)^2\epsilon^2 P(BB, BB|nI\Big)\\
&+\eta(1-\eta)(1-\epsilon)^2 P(AB, AA|nI) + \eta(1-\eta)\epsilon(1-\epsilon)P(AB, AB|nI)\\
&+\eta(1-\eta)\epsilon^2 P\Big(AB, BB|nI) + \eta^2(1-\epsilon)^2 P(AA, AA|nI\Big)\\
&+\eta^2\epsilon(1-\epsilon)P\big(AA, AB|nI) + \eta^2\epsilon^2 P(AA, BB|nI\big) = (1-\eta)^2(1\\
&-\epsilon)^2 p_A^2 + 2(1-\eta)^2\epsilon(1-\epsilon)p_A p_B + (1-\eta)^2\epsilon^2 p_B^2 + \eta(1\\
&-\eta)(1-\epsilon)^2 p_A^2 + 2\eta(1-\eta)\epsilon(1-\epsilon)p_A p_B + \eta(1-\eta)\epsilon^2 p_B^2\\
&+\eta^2(1-\epsilon)^2 p_A^2 + 2\eta^2\epsilon(1-\epsilon)p_A p_B + \eta^2\epsilon^2 p_B^2\\
&= [1-\eta(1-\eta)][(1-\epsilon)p_A + \epsilon p_B]^2
\end{aligned}$$

$$\begin{aligned}
&P_o\Big(BB, AB|IBD) = (1-\eta)^2\Big[(1-\epsilon)^2 + \epsilon^2\Big]P(BB, AB|I\Big)\\
&+2(1-\eta)^2\epsilon(1-\epsilon)P\Big(BB, AA|I) + 2(1-\eta)^2\epsilon(1-\epsilon)P(BB, BB|I\Big)\\
&+\eta(1-\eta)\Big[(1-\epsilon)^2 + \epsilon^2\Big]P(AB, AB|I) + 2\eta(1-\eta)\epsilon(1\\
&-\epsilon)P(AB, AA|I) + 2\eta(1-\eta)\epsilon(1-\epsilon)P(AB, BB|I) + \eta^2\Big[(1-\epsilon)^2\\
&+\epsilon^2\Big]P(AA, AB|I)\\
&+2\eta^2\epsilon(1-\epsilon)P\big(AA, AA|I) + 2\eta^2\epsilon(1-\epsilon)P(AA, BB|I\big) = (1-\eta)^2[(1\\
&-\epsilon)^2 + \epsilon^2\Big]p_A + 2(1-\eta)^2\epsilon(1-\epsilon)p_B + \eta(1-\eta) + \eta^2\Big[(1-\epsilon)^2\\
&+\epsilon^2\Big]p_B + 2\eta^2(1-\epsilon)\epsilon p_A = (1-\eta)^2(1-\epsilon)[(1-\epsilon)p_A + \epsilon p_B]\\
&+(1-\eta)^2\epsilon[\epsilon p_A + (1-\epsilon)p_B] + \eta(1-\eta) + \eta^2(1-\epsilon)[\epsilon p_A\\
&+(1-\epsilon)p_B] + \eta^2\epsilon[(1-\epsilon)p_A + \epsilon p_B] = (1-\eta)(1-\epsilon)f_A\\
&+(1-\eta)\epsilon f_B + \eta\epsilon f_A + \eta(1-\epsilon)f_B
\end{aligned}$$

$$\begin{aligned}
&P_o\Big(BB, AB|nonIBD) = (1-\eta)^2\Big[(1-\epsilon)^2 + \epsilon^2\Big]P(BB, AB|nI\Big)\\
&+2(1-\eta)^2\epsilon(1-\epsilon)P(BB, AA|nI) + 2(1-\eta)^2\epsilon(1\\
&-\epsilon)P\Big(BB, BB|nI) + \eta(1-\eta)\Big[(1-\epsilon)^2 + \epsilon^2\Big]P(AB, AB|nI\Big)\\
&+2\eta(1-\eta)\epsilon(1-\epsilon)P(AB, AA|nI) + 2\eta(1-\eta)\epsilon(1\\
&-\epsilon)P\Big(AB, BB|nI) + \eta^2\Big[(1-\epsilon)^2 + \epsilon^2\Big]P(AA, AB|nI\Big) + 2\eta^2\epsilon(1\\
&-\epsilon)P\big(AA, AA|nI) + 2\eta^2\epsilon(1-\epsilon)P(AA, BB|nI\big) = 2\Big[(1-\epsilon)^2\\
&+\epsilon^2\Big]\Big[(1-\eta)^2 + \eta(1-\eta) + \eta^2\Big]\ p_A p_B + 2\epsilon(1-\epsilon)\Big[(1-\eta)^2\\
&+\eta(1-\eta) + \eta^2\Big]\ p_A^2 + 2\epsilon(1-\epsilon)\Big[(1-\eta)^2 + \eta(1-\eta) + \eta^2\Big]\ p_B^2\\
&= 2[1-\eta(1-\eta)]\Big[\epsilon(1-\epsilon) + (1-2\epsilon)^2 p_A p_B\Big] = 2[1-\eta(1\\
&-\eta)f_A f_B
\end{aligned}$$

$$\begin{aligned}
&P_o\Big(BB, BB|IBD) = (1-\eta)^2(1-\epsilon)^2 P(BB, BB|I\Big) + (1-\eta)^2\epsilon(1\\
&-\epsilon)P\Big(BB, AB|I) + (1-\eta)^2\epsilon^2 P(BB, AA|I\Big) + \eta(1-\eta)(1\\
&-\epsilon)^2 P(AB, BB|I) + \eta(1-\eta)\epsilon(1-\epsilon)P(AB, AB|I) + \eta(1\\
&-\eta)\epsilon^2 P\Big(AB, AA|I) + \eta^2(1-\epsilon)^2 P(AA, BB|I\Big) + \eta^2\epsilon(1\\
&-\epsilon)P\big(AA, AB|I) + \eta^2\epsilon^2 P(AA, AA|I\big) = (1-\eta)^2(1-\epsilon)^2 p_B\\
&+(1-\eta)^2\epsilon(1-\epsilon)p_A + \eta(1-\eta)(1-\epsilon)^2 p_B + \eta(1-\eta)(1-\epsilon)\epsilon\\
&+\ \eta(1-\eta)\epsilon^2 p_A + \eta^2\epsilon(1-\epsilon)p_B + \eta^2\epsilon^2 p_A = [(1-\eta)(1-\epsilon)\\
&+\eta\epsilon][(1-\epsilon)p_B + \epsilon p_A]
\end{aligned}$$

$$
\begin{aligned}
&P_o\Big(BB, BB|nonIBD) = (1-\eta)^2(1-\epsilon)^2 P(BB, BB|nI\Big)\\
&+(1-\eta)^2\epsilon(1-\epsilon)P\Big(BB, AB|nI) + (1-\eta)^2\epsilon^2 P(BB, AA|nI\Big)\\
&+\eta(1-\eta)(1-\epsilon)^2 P(AB, BB|nI) + \eta(1-\eta)\epsilon(1-\epsilon)P(AB, AB|nI)\\
&+\eta(1-\eta)\epsilon^2 P\Big(AB, AA|nI) + \eta^2(1-\epsilon)^2 P(AA, BB|nI\Big)\\
&+\eta^2\epsilon(1-\epsilon)P\big(AA, AB|nI) + \eta^2\epsilon^2 P(AA, AA|nI\big) = (1-\eta)^2(1\\
&-\epsilon)^2 p_B^2 + 2(1-\eta)^2\epsilon(1-\epsilon)p_A p_B + (1-\eta)^2\epsilon^2 p_A^2 + \eta(1\\
&-\eta)(1-\epsilon)^2 p_B^2 + 2\eta(1-\eta)(1-\epsilon)\epsilon p_A p_B + \eta(1-\eta)\epsilon^2 p_A^2\\
&+\eta^2(1-\epsilon)^2 p_B^2 + 2\eta^2\epsilon(1-\epsilon)p_A p_B + \eta^2\epsilon^2 p_A^2 = [1-\eta(1\\
&-\eta)][(1-\epsilon)p_B + \epsilon p_A]^2
\end{aligned}
$$

Variant Filtering for Empirical Genotype Data Prior to IBDmix Calculation

For the empirical genotype data we filtered out multi-allelic SNVs and indels from the archaic genome. We also eliminated all variants with one or fewer minor allele counts in the target sample. Singletons are more likely than other variants to be genotype-calling artifacts or very recent mutations and are therefore not helpful for IBD estimation.

Sites that are missing in the archaic genome are not considered for analysis. Sites that are present in the archaic genome but are missing in the modern human genomes are only included in the analysis if the archaic sample carries at least one alternative allele, in which case the modern human genotypes are "imputed" as homozygous for the reference allele. IBDmix introduces allele error rates into the genome data for both archaic and modern humans, so including a greater number of variants leads to better performance.

Test for Population Size Effect on IBDmix Calculation

IBDmix estimates allele frequencies for the modern samples from the empirical data and uses these for the calculation of the IBD LOD score. For accurate IBDmix calls, a minimum sample size is required to ensure the accuracy of allele frequency estimates. We tested the effect of sample size on IBDmix using the CEU (Utah Residents with Northern and Western European Ancestry) subgroup from 1000 Genomes Project. We used bootstrap resampling of the entire CEU subgroup (n = 99) to generate multiple target samples of sizes n = [10, 20, 50, 70, 90, 99]. We then re-called Neanderthal introgressed sequence for these individuals using IBDmix. We found the average amount of Neanderthal sequence called for this population stabilized when sample size was larger than 10 (Table S2) while more than 99.9% of introgressed regions that were called at the size of 10 overlapped the result of a full population size. We repeated this test on East Asian (Han Chinese in Beijing, CHB) and African subgroups from the 1000 Genomes Project, and found similar results regarding the minimum population size to stabilize IBDmix estimates of archaic ancestry. We therefore recommend that IBDmix be used with sampled human populations of 10 individuals or more. We recognize as well, that the accuracy of allele frequency estimates will be sensitive to population structure, and so the exact minimum population sample size for IBDmix may vary in some cases.

Simulation Study

IBDmix Performance

We used msprime (Kelleher et al., 2016) to simulate sequence data and to call introgressed segments in simulated European, East Asian, and African modern individuals. Our simulations comprised 100 replicates of 15 Mb, sampling 100 diploid genomes each for African, European, and East Asian lineages, and 1 Neanderthal diploid genome. We used the coalescent trees from the simulations to identify the true introgressed haplotypes in the human populations. We simulated a mutation rate of 1.25×10^{-8} per bp per generation. We used a recombination rate of 10^{-8} per bp per generation (1cM/Mb). The parameters for our demographic model were based on published estimates and assume a generation time of 25 years and a haploid ancestral effective population size of 7310. The split between the ancestors of Neanderthals and modern humans was set to 28,000 generations ago. The out-of-Africa human migration occurred 3,920 generations ago. The rate of migration between the African and out-of-Africa populations was 2×10^{-4} haploid individuals per generation, which corresponds to a cumulative Eurasian admixture into Africa over 2,400 generations of 2.4%. The rate of back-migration from the modern European to the African population was 1.7×10^{-5} haploid individuals per generation. We allowed for Neanderthal introgression to occur between 2,200 to 2,230 generations ago at a rate of 0.1% per generation, for an overall admixture proportion of 3%. We allowed for rapid growth of ~2% per generation in all human populations starting 200 generations ago, simulating the development of agriculture. See Figure S1A for the schematic of our simulated model. We also used a model with a higher migration rate (5×10^{-4}) between African and Eurasian lineages to evaluate IBDmix and S^* performance under different demographic scenarios.

We randomly introduced sequence error to the genotype data created from msprime and therefore allowed sequence errors in both archaic and modern human genotypes in the simulation model. We tuned the parameters for IBDmix (LOD cutoff, archaic sequence error, maximum sequence error in modern human, sequence error as a function of MAF in modern human) using the simulated data. We evaluated the performance of IBDmix on simulated data, assessing metrics such as false positive rate, power, false discovery rate, precision and recall (Figure S1B).

We also simulated models with higher mutation rates, 2x, 5x, and 10x the default value (1.25x10^{-8} per bp per generation). We evaluated IBDmix performance under these models (Figure S1C).

To investigate the impact of recombination rate on IBDmix calling, we simulated null models using the genome-wide average (10^{-8} per bp per generation) and 1/10th that rate (10^{-9} per bp per generation). These models did not include Neanderthal introgression. We evaluated FPR of IBDmix under these null models (Table S3).

We simulated models with two Neanderthal lineages representing an introgressing lineage and a sampled lineage. We tested several models varying the split time between these two lineages (70 kya, 100 kya, 145 kya). We called introgressed sequence using IBDmix with the sampled Neanderthal lineage as the reference genome, rather than the introgressing Neanderthal, and evaluated IBDmix performance (Figure S1D).

Because determining the precise endpoints of introgressed segments for any method remains difficult, when evaluating IBDmix performance we required IBDmix identified segments to overlap a call made using the coalescent trees by > 1bp in order to be determined a true positive. Any introgressed segment called by IBDmix that does not overlap a call from the coalescence tree is considered a false positive. We calculated power as: (the counts of true positives) / (the counts of true segments from coalescence tree). We calculated FDR as: (the counts of false positives) / (the counts of false positives + the counts of true positives). We calculated FPR as: (total bp of false positives) / (15 Mb – total bp of true segments from coalescence trees).

Simulations of Demographic Models with Back-Migration and pre-out-of-Africa gene-flow

To analyze the effects of back-migration and pre-OOA gene-flow on the level of Neanderthal ancestry in Africans we compared empirical data from IBDmix calls made on 1000 Genomes samples in EUR (n = 503), EAS (n = 504), and YRI (n = 108) populations to simulated data from msprime. Our simulations consisted of 1000 replicates of 15MB chromosomes with diploid sample sizes matching those of the empirical data and including a sampled Neanderthal lineage (n = 1). We used the same demographic model as was used for IBDmix performance evaluation, and kept a recombination rate of 1×10^{-8} per bp per generation, a mutation rate of 1.25×10^{-8} per bp per generation, and a generation time of 25 years per generation. We included a single pulse of admixture from the Neanderthal into the non-African lineage 2,000 generations ago, at a level of 5% per generation for a single generation. To test the effect of back-migration, we included a single migration parameter from either the ancestral Eurasian population into Africans, which stopped after the split of Europeans and East Asians, or from Europeans into Africans after the split with East Asians until the present. We specified the migration to occur only in one direction (from non-Africans into Africans) and tested a range of migration rates (Figure S3) that included levels established in previous demographic models (Tennessen et al., 2012). In order to test the effect of pre-OOA gene-flow from humans to Neanderthals, we added a single migration parameter from the ancestral human lineage into the Neanderthal lineage at a level of 10% per generation for a single generation, and specified this admixture to occur at 4×10^3, 6×10^3, or 10×10^3 generations ago. For reference, African and non-African lineages split in our model at 3.92×10^3 generations ago.

Sequence data from the simulations were collected in vcf format and analyzed separately using IBDmix and the S^* pipeline (Vernot et al., 2016) in order to identify Neanderthal introgressed segments in simulated human individuals. As well, we collected the true introgressed segments from the simulated data using the coalescent trees. For IBDmix, we used a threshold of LOD > 4 and removed segments < 50kb in order to create a final call set of introgressed segments. In order to identify introgressed segments using S^*, we calculated S^*-scores and Neanderthal-match-percent in 50kb windows at 10kb overlapping steps. We determined statistically significant S^*-scores and match-percent levels using 10,000 replicates of a null simulation. We required that windows have S^* p value < 0.01 and Neanderthal-match-percent p value < 0.05 to be considered Neanderthal-introgressed. Overlapping statistically significant introgressed windows were merged to produce full Neanderthal introgressed segments.

Whole Genome Sequence Data

We analyzed whole-genome sequence data from all populations from the 1000 Genomes phase 3 data. The populations analyzed were East Asians, Europeans, South Asians, Americans, and Africans, consisting of 26 geographically diverse subgroups and 2504 individuals in total. We first removed multi-allelic SNVs and indels from archaic genome. We then removed the sites that are not biallelic SNVs in the entire 1KG dataset. High coverage archaic genomes for the Altai Neanderthal and Altai Denisovan (Prüfer et al., 2014) were obtained from http://cdna.eva.mpg.de/neandertal/altai/.

All analyses were performed on autosomes. We performed archaic ancestry detection in each subgroup (e.g., CEU, CHB, YRI) rather than continental populations to avoid potential effects of population structure.

We applied the following filters to the empirical data (1000 Genomes, Altai Neanderthal and Altai Denisovan genomes):

- CpGs were masked as in (Prüfer et al., 2014).
- Mappable regions were determined by examining all 35 base long "reads" that overlap each site. A site is mappable if the majority of overlapping reads are mapped uniquely or without 1-mismatch hits to hg19 (Li and Durbin, 2011).
- Segmental duplications (Bailey et al., 2002) were removed and downloaded from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz

- Sites within 5bp of indels were removed.
- The 1000 Genomes accessibility mask was applied, downloaded from: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20140520.strict_mask.autosomes.bed
- We also applied the Altai and Denisovan minimal filter mask (Prüfer et al., 2014), downloaded from: https://bioinf.eva.mpg.de/altai_minimal_filters

QUANTIFICATION AND STATISTICAL ANALYSES

Refining Neanderthal Callset by Using Denisovan Sequences as a Negative Control

We adopted a conservative approach to filtering our callset in order to maximize our signal of detected Neanderthal ancestry. After initially calling Neanderthal and Denisovan sequences using IBDmix, we refined the Neanderthal callset by masking any regions that were called as Denisovan sequence in Africans and also present as Neanderthal sequence in any population. Such regions represent either ILS shared in all hominins from a deep coalescent event, or true Neanderthal sequence mis-assigned as Denisovan sequence. After filtering, the average amount of Neanderthal ancestry in each population decreased by several Mb, but maintained the same patterns and relative proportions as discussed in the paper (Tables S4 and S5). Furthermore, we observed some regions with a high proportion of derived alleles in the Neanderthal genome that also shared an unusually high proportion of derived alleles in some or all modern-human populations. These regions may contain exceptional local genetic features, and may exhibit more complex evolutionary and recombination histories than other genomic regions. To be conservative, we also provide a callset removing regions where the proportion of derived alleles in the Neanderthal genome for a given window fell in the upper 99.9th-percentile. This further reduced the amount of detected Neanderthal ancestry in all populations, however relative levels of Neanderthal ancestry for different populations were still robust (Table S4).

For our callset of identified Denisovan introgressed segments, we introduced additional filters to refine the initial callset. We masked any regions that were both detected as Neanderthal and Denisovan sequence for all populations, removing mis-assigned sequence and ILS. We further controlled for ILS by removing from all populations segments that were called as Denisovan in Africans at a frequency $\geq$ 30%, accounting for 10% of detected Denisovan segments in Africans. The average amounts of detected Denisovan sequence in all populations are reported in Table S6.

As discussed, it is necessary to re-parameterize IBDmix when applied to other archaic hominins since the approach in this study is focused on maximizing Altai Neanderthal signal.

Replicating Regions Significantly Depleted of Neanderthal Introgressed Sequence

We have previously described a method for identifying regions significantly depleted of Neanderthal sequence identified by S^* in non-African populations (Vernot et al., 2016). In summary, we break the genome into windows of varying size (8-15Mb) at 100kb overlapping steps, requiring that a window be composed of > 70% unfiltered bases. We then determine, for a given window, the average number of Neanderthal introgressed bases across all individuals. We perform this measure for all windows that meet the filtering requirements in order to generate a distribution for the average level of Neanderthal ancestry across the genome. Windows that are in the lower 99th-percentile for average amount of introgressed sequence are considered significantly depleted and are merged with overlapping windows to define depleted regions. The final list of depleted regions is determined by merging the significant regions of all window sizes. We applied the same analysis to Neanderthal introgressed calls made with IBDmix and compared these sets of depletions to those identified using the S^*-callset (Figure S4; Table S8) (Vernot et al., 2016).

Comparing Simulated Data to Empirical Data

In cases where we compared simulated data to empirical data (Figure 3) we filtered the simulated IBDmix calls to replicate filtering for empirical data, removing segments < 50kb. To analyze the distribution of segment lengths for calls made in African and non-African populations, we used unmerged calls from all African individuals (LWK, GWD, MSL, YRI, ESN), and all non-African individuals, except for ASW and ACB. Calls made by IBDmix in African samples that overlapped any non-African call by 1bp were categorized as “African shared calls” (n = 95032), and those that did not overlap any non-African calls were categorized as “African unique calls” (n = 900).

To analyze the frequency within the African population of segments identified as African and shared with non-Africans, we limited our analysis to calls made in YRI that overlapped by 1bp with calls made in Europeans or East Asians (n = 19333). We then counted for each call the number of other African individuals who carried an overlapping call, and assigned each call as either “Below 10%,” where < 11 YRI individuals carried an overlapping segment (n = 2586), or “Above 10%,” where $\geq$ 11 other YRI individuals carried an overlapping segment (n = 16747). We measured the number of calls in each category as a proportion of the total number of calls in YRI that intersected calls made in Europeans or East Asians.

We measured the ratio of Neanderthal sequence in East Asians compared to Europeans with and without masking overlapping YRI calls. Eurasian calls were removed if they overlapped a YRI call by 1bp. We summed together the total amount of sequence called for each population separately, and the ratio between the East Asian and European populations was obtained.

Reference Panel Size Effect on S* Admixture Estimates

We examined how reference panel size for S^* affects Neanderthal ancestry estimates by bootstrap resampling the Yoruba 1000 Genomes Project samples and reanalyzing chromosome 1 for Europeans and East Asians. We bootstrap sampled Yoruba (YRI, n = 108) individuals from the 1000 Genomes Project to generate multiple reference panels of sizes n = [1, 2, 5, 10, 25, 50, 75, 108]. We then re-called Neanderthal introgressed sequence on chromosome 1 for European (n = 503) and East Asian (n = 504) individuals using the S^*-pipeline (Vernot et al., 2016) and the new reference panel, requiring S^* p value < 0.01 and Neanderthal match-percent p value < 0.05. We performed 10 replicates of this analysis resampling the YRI reference panel for each replicate and calculated the mean level of S^*-sequence identified per sample.

The mean total S^*-sequence called for each sample across the 10 replicates was compared to the average amount of S^*-sequence called for samples using a reference panel of YRI = 1. We used this normalized mean to test for significant difference (t test) between the amount of S^*-sequence called in EUR and EAS for different reference panel sizes. In addition, for each reference panel size, an average admixture proportion was calculated for each population across replicates by dividing the mean S^*-sequence for all 10 replicates by the total amount S^*-queryable sequence.

Identifying High-Frequency Introgressed Haplotypes From IBDmix Data

We used derived allele frequencies calculated from 1000 Genomes Project to identify population specific high-frequency introgressed haplotypes. To do this, we identified sites that had extreme differences in derived allele frequency between populations, intersected Neanderthal segments identified by IBDmix, and matched the Altai Neanderthal reference alleles.

We began by removing 1000 Genomes Project variants that we masked during the IBDmix analysis. We then intersected the remaining variants with Neanderthal calls made by IBDmix in EUR, EAS, and AFR populations. For variants that intersected identified Neanderthal segments, we calculated the differences in the derived allele frequencies between EUR and EAS, AFR and EUR, and AFR and EAS. We identified the lower and upper 1% values for the differences in derived allele frequencies as part of an outlier approach. For example, in the comparison of EUR and EAS sites, we retained sites where the absolute difference in the derived allele frequency between EUR and EAS was > 40%. We further filtered on the derived allele matching the Neanderthal allele, and in the case of EUR and EAS calls, that the AFR derived allele frequency was < 1%. To maximize our ability to identify population-specific high-frequency haplotypes, we required that, for EUR-specific calls, the EUR derived allele frequency be > 40% and the EAS derived allele frequency be < 10%; for EAS-specific calls, the EUR derived allele frequency be < 10% and the EAS derived allele frequency be > 40%; for AFR-specific calls, the EUR and EAS derived allele frequencies both be < 5%. We also required that for a given allele, the number of individuals in a population who carry the Neanderthal sequence at that locus be greater than 5. By intersecting the alleles that met these filtering criteria with the merged Neanderthal callsets for EUR, EAS, and AFR, we identified a final set of distinct high-frequency introgressed haplotypes (Table S7). We compared our haplotypes with previously identified high-frequency haplotypes (Gittelman et al., 2016), and the presence of previously reported GWAS SNPs pulled from UCSC Genome Browser with reported $p \leq 1x10^{-5}$.

Calculating the Rate of Overlap Between Neanderthal Calls and European Ancestry in African Samples

Under the model that back-migration from Europeans to Africans accounts for a substantial amount of Neanderthal ancestry in Africans, we hypothesized that we should find an enrichment for Neanderthal ancestry in Africans at loci that also show evidence of European ancestry. To test this hypothesis, we compared for chromosome 1 the rate of overlap of Neanderthal segments identified by IBDmix with tracks of European and East Asian ancestry identified by RFMix (Maples et al., 2013) on a per individual basis for all 504 African individuals analyzed in our study.

We began by taking the phased genotype data for chromosome 1 and processing these with vcftools and custom scripts to retain only bi-allelic, completely phased sites that could be mapped to genomic coordinates. After processing, we retained 245,126 sites for analysis with RFMix.

We used RFMix to analyze the ancestry of each African individual separately. Specifically, we adopted a leave-one-out approach, in which each African individual was analyzed against a reference panel composed of the remaining 503 African samples, 503 European samples, and 504 East Asian samples. We recoded the ancestry tracks determined by RFMix from genomic positions into base-pair coordinates, and merged tracks of European or East Asian ancestry that were within 10kb of similar ancestry tracks. The median track length for European ancestry is 142kb, and for East Asian ancestry is 132kb. The average level of European and East Asian ancestry per individual is 2.2% and 0.45%, respectively.

Next, we compared the rate of overlap of Neanderthal calls with either European or East Asian ancestry tracks on a per individual basis,$r_{emp} = (\#\text{ of Neand segments overlapping EUR or EAS ancestry }/\text{Total }\#\text{ of Neand segments})$

and took the average across all 504 African individuals to calculate empirical values for the average rate of overlap of Neanderthal and European ancestry, and the rate of overlap for Neanderthal and East Asian ancestry. To test the significance of these empirical values, we performed permutation tests, analyzing an individual's Neanderthal calls against a random individual's European and East Asian ancestry tracks. We performed 10,000 replicates of this analysis, averaging the rate of overlap for all 504 Africans in each replicate. When we compared the empirical average rate of overlap for East Asian ancestry to the null distribution, we found 4495/10000 replicates equaled or exceeded the empirical value. When we repeated this with the European ancestry data, we found 0/10000 replicates equaled or exceeded the empirical value.

Calculating rate of exclusively shared sequence between African and non-African populations

In Europeans, Neanderthal sequence covers 821Mb across 503 individuals, and in East Asians, Neanderthal sequence covers 792Mb across 504 individuals. We took the intersection of unmerged Neanderthal sequence in Africans and Europeans, e.g., segments in Africans that overlapped segments in Europeans by > 1bp, and merged the genomic coverage as African-European shared sequence. We then subtracted Neanderthal sequence from this shared collection that was also present in East Asians. This defined the collection of "exclusively shared sequence between Africans and Europeans." We used the same approach to identify exclusively shared sequence between Africans and East Asians. In the observed data our reported values are 59Mb of African-European exclusively shared sequence, and 16Mb of African-East Asian exclusively shared sequence.

After assessing the level of exclusively shared sequence in the empirical data, we also randomly sampled unmerged European segments to generate 792Mb of merged sequence, matching the overall coverage for East Asians. We then re-calculated the amount of exclusively shared sequence with Africans across 10 replicates. After down-sampling, we still observed ~57Mb of European-African exclusively shared sequence versus ~17Mb of East Asian-African exclusively shared sequence.

Comparing callsets from different methods in shared individuals

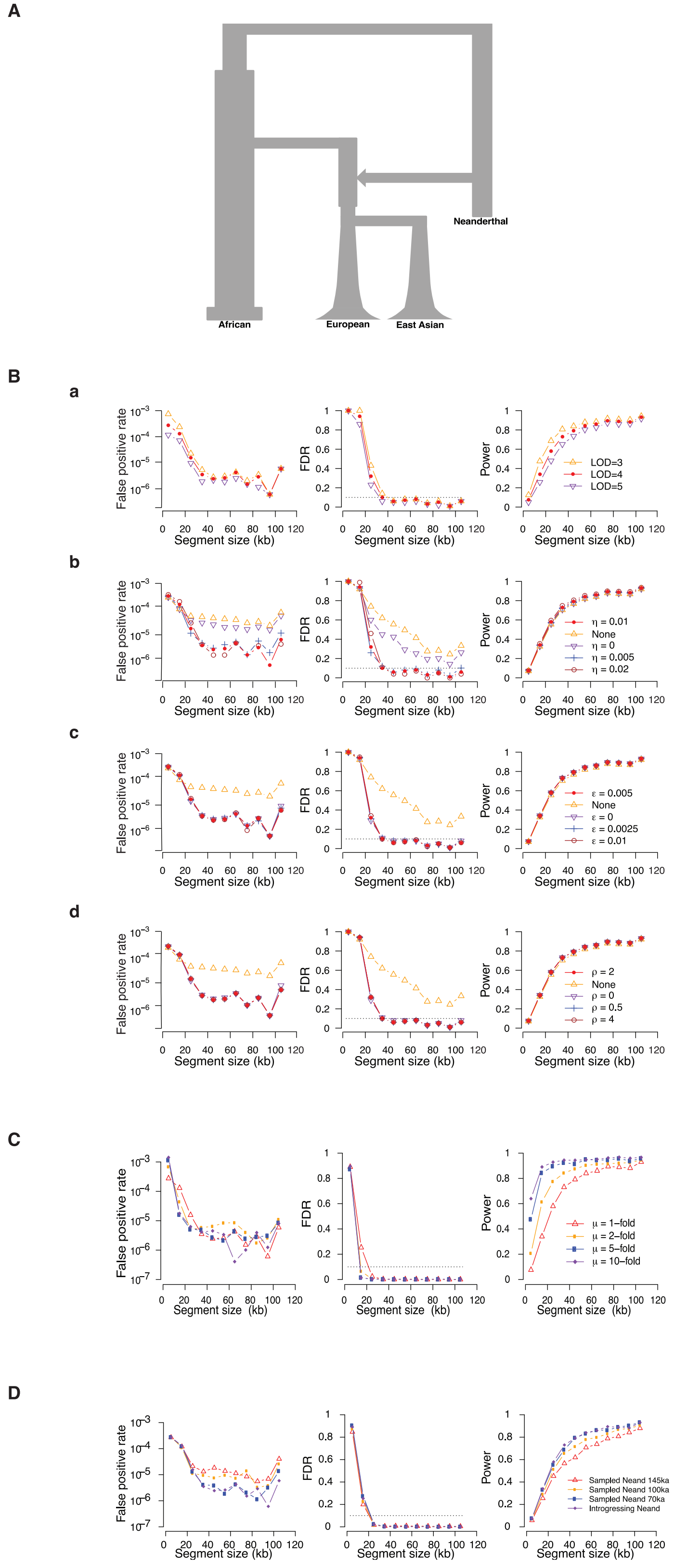
Since IBDmix, CRF, diCal-admix, and *S** used different versions of population data from 1000 Genomes Project, we first picked out the shared non-African individuals among these callsets and only worked on the introgressed sequence in these individuals. We then merged the sequence from one callset and compared the genomic coverage to each other.

DATA AND CODE AVAILABILITY

The code for IBDmix software is available online at https://github.com/PrincetonUniversity/IBDmix.

The segments of introgression detected in 1000 Genomes data using IBDmix are available here: https://drive.google.com/drive/folders/1mDQaDFS-j22Eim5_y7LAsTTNt5GWsoow?usp=sharing

Supplemental Figures



(legend on next page)

Figure S1. Simulated Model and Performance Evaluation for IBDmix, Related to Figure 1 and STAR Methods

(A) Simplified schematic of the demographic model used for simulations evaluating the performance of IBDmix. (B) Optimizing IBDmix function parameters under the basic simulation model (A): (a) LOD score, (b) Archaic sequence error, (c) maximum sequence error in modern human, and (d) sequence error as a function of MAF in modern human. (C) Impact of genetic variation on IBDmix performance under the basic simulation model (A). IBDmix performance (FPR, FDR and Power) under the simulation models with mutation rates 2x, 5x, and 10x the default value ($1.25x10^{-8}$ per bp per generation). (D) Evaluation of IBDmix performance under the simulation models using a reference archaic genome distantly related to the introgressing archaic. In different models, the sampled reference lineage diverges from the introgressing archaic at 70 kya (blue), 100 kya (yellow), and 145 kya (red). For comparison, IBDmix performance using the introgressing archaic genome (purple) is shown.

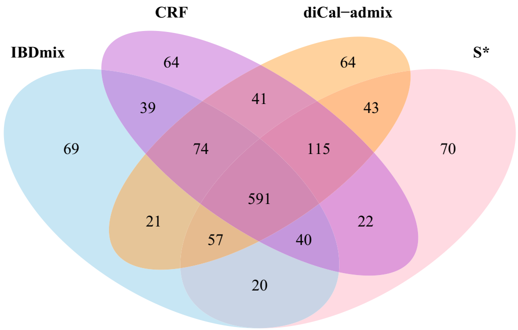


Figure S2. Comparing the Genomic Coverage of Neanderthal Sequence Detected by Different Methods, Related to STAR Methods
The intersections of merged callsets (Mb) from IBDmix (blue), CRF (purple), diCal-admix (yellow), and *S** (pink) are shown.

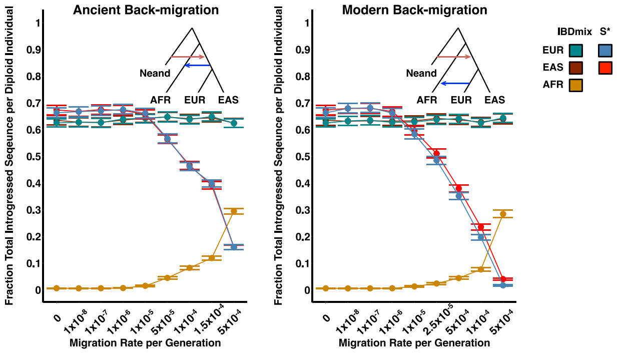


Figure S3. Back-Migration Can Bias Amount of Recovered Neanderthal Sequence in *S, But Not IBDmix, Related to STAR Methods**

Back-migration from ancestral Eurasians (left) reduces the amount of Neanderthal sequence recovered by *S**, but does not produce the apparent enrichment in East Asians when compared to Europeans, as seen in migration from ancestral Europeans (right). IBDmix is robust to both the rate and timing of migration. The level of Neanderthal ancestry is reported as an average for the population with the corresponding 95% confidence interval.

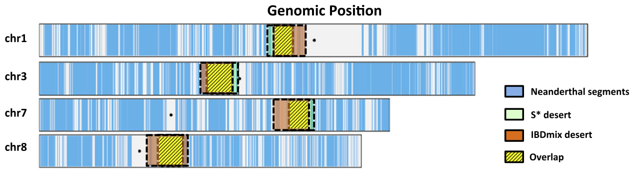


Figure S4. Visualization of *S and IBDmix Identified Desert Regions and Their Overlap, Related to STAR Methods**