Patterns of Archaic Hominin DNA in Modern Human Genomes

Benjamin John Hopson Vernot

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Joshua M. Akey, Chair

Philip Green

Jay Shendure

Program authorized to offer degree:

Genome Sciences

Abstract

In this dissertation I describe the development of a method for identifying introgressed archaic haplotypes. I then present the application of this method to several populations. In 15 African hunter-gatherer genomes, I identify signatures of introgression from an unknown archaic hominin with an apparent divergence time with modern humans that is similar to the divergence time of Neanderthals. In a sample of 379 European and 279 East Asian genomes, I identify on average 1/4 of each individual's introgressed Neanderthal sequence, composing a total of 600Mb of the Neanderthal genome. I use characteristics of this sequence to estimate demographic parameters, including ancestral effective population size (Ne), and the complexity of the introgression event. I also present signatures of both purifying selection against Neanderthal sequence in modern humans, and selection for other, beneficial Neanderthal alleles. In a separate analysis, I show that the inferred parameters of the introgression event are not influenced by differing efficiency of selection between Europeans and East Asians, and explore alternative models for the complexity of the introgression event. In a sample of 35 Melanesian individuals from Papua New Guinea, I identify both Neanderthal and Denisovan introgression, and use this map of archaic introgression to identify regions of the genome that are depleted of archaic sequence from two independent introgression events, implying the presence of non-random forces such as selection in the creation of such regions.

Acknowledgements

This thesis would not have been possible without the generous help and support of numerous individuals. I would like to thank everyone in the Department of Genome Sciences, and my committee members: Phil Green, Jay Shendure, and Marshall Horwitz. My advisor, Joshua Akey, has proved to be an unerring font of wisdom, humility and grace, and I am grateful for his support and advice.

I would like to thank my family, including my siblings, all four of my parents, and all eight grandparents – especially my mother Natalie Slater, my father David Vernot, and my grandfather Ed Vernot, who enthusiastically encouraged me to go into science. I would also like to thank my adopted family – all of the friends in Seattle, Pittsburgh, and elsewhere.

Most of all, I would like to thank my partner Louisa Jáuregui, who has contributed not just love and amazing levels of support, but also Illustrator skills and so many mugs of early morning coffee.

Table of Contents

Chapter 1 : Introduction

There has been a longstanding interest in the history of the human species [Darwin, 1871; Dobzhansky, 1962]. Since the discovery of hominin bones in the Neander Valley, and the realization that these bones may have represented a different species [Schaaffhausen, 1858; King, 1864; Howell, 1951], there has been debate that humans and Neanderthals may have co-existed, and possibly interbred [Howell, 1951; Stringer and Gamble, 1993]. When a portion of Neanderthal mitochondrial DNA was sequenced in 1997, and it was shown that the Neanderthal mitochondria diverged prior to all known human mitochondrial sequences, many assumed that interbreeding in fact did not occur [Krings et al., 1997]. However, [Nordborg, 1998] put forth a theoretical model showing that, due to the large variation in coalescent times possible for mitochondrial genomes, the absence of any sharing of mitochondrial DNA between humans and Neanderthals did not in fact exclude the possibility that interbreeding had occurred, and that the mitochondrial results were inconclusive.

The first evidence of interbreeding came with the sequencing of DNA from several Croatian Neanderthal bones to produce a low-coverage draft Neanderthal genome [Green et al, 2010]. The authors of this study observed that non-African individuals shared slightly more derived alleles with the Neanderthal genome than African individuals. They showed that this was consistent with interbreeding between the common ancestor of non-Africans and Neanderthals, resulting in some Neanderthal alleles being introduced into the genomes of the modern human populations, and an average Neanderthal ancestry of 1-3% in each modern non-African. Such alleles are said to have *introgressed* from Neanderthals to modern humans. However, the authors and others noted that this excess of Neanderthal alleles in non-Africans could also be

explained by ancient population structure dating to before the divergence of Neanderthals and modern humans, such that non-Africans and Neanderthals descend from the same ancient portion of the structured population [Green et al, 2010; Eriksson et al, 2012; Yang et al, 2012]. Such a hypothesis predicts that matching alleles would have existed in the human population continuously since prior to the divergence of Neanderthals and modern humans - i.e., more than 400ky. The structure hypothesis was subsequently strongly rejected by an LD-based analysis, showing that non-African Neanderthal alleles date to ~55kya (95% CI 37-86kya) [Sankararaman et al, 2012], consistent with the enrichment in non-Africans deriving from admixture with Neanderthals after the out of Africa event.

Because fossil evidence had shown the existence of Neanderthals in Europe, the sequencing of the Neanderthal genome and evidence of admixture with non-Africans were not unexpected. This was much less the case when a second archaic genome was sequenced from a single pinky bone from the Altai mountains in Russia, and was shown to belong to a second, previously unknown species of archaic hominin [Reich et al, 2010]. Even more surprising was evidence that this second species, "Denisovans," contributed 4-6% of the modern day ancestry of Melanesians and Australian aboriginals, two populations geographically distant from the sequenced Denisovan individual, and seemingly none to Europeans or mainland Asian populations [Reich et al, 2010; Reich et al, 2011; Meyer et al, 2012].

After the publishing of the Neanderthal and Denisovan genomes, several candidate introgressed genes were identified as being similar to Neanderthal in non-African populations [Abi-Rached et al, 2011; Mendez et al, 2012a; Mendez et al, 2012b; Huerta-Sánchez et al, 2014; Vernot and Akey 2014b]. Because most of these were at high frequency, and had previously been implicated as possible targets for selection (e.g., EPAS1 in Tibetans [Beall et al, 2010]), it

was often suggested that these introgressed haplotypes were beneficial to the modern human populations that inherited them. However, no genome-wide scan for introgressed Neanderthal or Denisovan haplotypes had been performed.

In contrast to the study of archaic introgression from Neanderthals and Denisovans through comparison to ancient DNA, it may be possible to produce evidence of archaic introgression in the absence of ancient DNA [Plagnol & Wall, 2006], or even fossils of a candidate archaic species [Hammer et al, 2011]. The S* statistic was developed to identify candidate introgressed haplotypes, and was used to argue for levels of introgression of at least 5%, from Neanderthals to Europeans and from an unknown archaic hominin to Yorubans [Plagnol and Wall, 2006; Hammer et al, 2011]. This statistic relies on discrepancies between two measures of the age of a haplotype - the mutational divergence from other modern human haplotypes, and the haplotype length. For example, haplotypes introgressed into non-Africans from Neanderthals would be expected to have accumulated large numbers of mutations since the split of humans and Neanderthals ~400kya, but because such haplotypes entered the modern human genome much more recently, they have had relatively little time to recombine onto the modern human background. In addition, the S* statistic assumes introgression into one population, but not into a second reference population; e.g., Neanderthal alleles should be present in non-African populations, but largely absent in sub-Saharan Africans. Although S* had been used to argue for the existence of introgression, S* alone was not suitable to genome-wide scans for introgression.

In this thesis, I present a genome-wide method for identifying introgressed haplotypes, based partly on a generalization of the S* statistic. This method can utilize an archaic genome to identify introgression from a specific species, or can be used in the absence of such an archaic

genome, although with reduced power and specificity. I apply this method to several populations with whole-genome sequence data, including 15 African hunter-gatherer individuals in a collaboration Sarah Tishkoff [Lachance, Vernot et al, 2012], 379 European and 279 East Asian individuals from the 1000 Genomes project [Vernot and Akey, 2014a; Vernot and Akey, 2015], and 35 Melanesian individuals from Papua New Guinea [preliminary results]. Additionally, I participated in several projects related to the ENCODE project [e.g., Vernot et al, 2012], but as those publications relate only tangentially to the topic of this thesis, they are not included.

References

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science* 334, 89–94.

Beall, C.M., Cavalleri, G.L., Deng, L., Elston, R.C., Gao, Y., Knight, J., Li, C., Li, J.C., Liang, Y., McCormack, M., et al. (2010). Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *PNAS* 107, 11459–11464.

Darwin, C. R. (1871). *The descent of man, and selection in relation to sex*. London: John Murray.

Dobzhansky, T. (1962). *Mankind evolving: The evolution of the human species*. New Haven, Conn: Yale University Press.

Eriksson, A., and Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *PNAS*.

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

Hammer, M.F., Woerner, A.E., Mendez, F.L., Watkins, J.C., and Wall, J.D. (2011). Genetic evidence for archaic admixture in Africa. *PNAS* 108(37): 15123–15128.

Howell, F.C. (1951). The place of Neanderthal man in human evolution. *Am. J. Phys. Anthropol*. 9, 379–416.

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* advance online publication,.

King, W. (1864). The Reputed Fossil Man of the Neanderthal. *The Quarterly Journal of Science*. 1, 88-97.

Krings, M., Stone, A., Schmitz, R.W., Krainitzki, H., Stoneking, M., and Pääbo, S. (1997). Neandertal DNA Sequences and the Origin of Modern Humans. *Cell* 90, 19–30.

Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.-M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., et al. (2012). Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. *Cell* 150, 457–469.

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*.

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., Filippo, C. de, et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338, 222–226.

Nordborg, M. (1998). On the Probability of Neanderthal Ancestry. *The American Journal of Human Genetics* 63, 1237–1240.

Plagnol, V., and Wall, J.D. (2006). Possible Ancestral Structure in Human Populations. *PLoS Genet* 2, e105.

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

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M.R., Pugach, I., Ko, A.M.-S., Ko, Y.-C., Jinam, T.A., Phipps, M.E., et al. (2011). Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *The American Journal of Human Genetics* 89, 516–528.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. arXiv:1208.2238.

Schaaffhausen, H. Zur Kenntnis der ältesten Rasseschädel. (1858). *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, 453–478.

Stringer, C. and Gamble, C. (1993). *In search of the Neanderthals*. New York: Thames and Hudson.

Vernot, B., and Akey, J.M. (2014a). Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* 343, 1017–1021.

Vernot, B., and Akey, J.M. (2014b). Human Evolution: Genomic Gifts from Archaic Hominins. *Current Biology* 24, R845–R848.

Vernot, B., and Akey, J.M. (2015). Complex History of Admixture between Modern Humans and Neandertals. *The American Journal of Human Genetics* 96, 448–453.

Vernot, B., Stergachis, A.B., Maurano, M.T., Vierstra, J., Neph, S., Thurman, R.E., Stamatoyannopoulos, J.A., and Akey, J.M. (2012). Personal and population genomics of human regulatory variation. *Genome Research* 22, 1689–1697.

Yang, M.A., Malaspinas, A.-S., Durand, E.Y., and Slatkin, M. (2012). Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. *Mol Biol Evol* 29, 2987–2995.

Chapter 2 : Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers

This chapter is an excerpt of a manuscript published in Cell on July 26, 2012. Figure numbers have been left as in the manuscript, for easier comparison.

Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.-M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., et al. (2012). Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. Cell 150, 457–469.

Abstract from original manuscript:

To reconstruct modern human evolutionary history and identify loci that have shaped hunter-gatherer adaptation, we sequenced the whole-genomes of five individuals in each of three different hunter-gatherer populations at > 60x coverage: Pygmies from Cameroon and Khoesan-speaking Hadza and Sandawe from Tanzania. We identify 13.4 million variants, substantially increasing the set of known human variation. We found evidence of archaic introgression in all three populations and the distribution of time to most recent common ancestors from these regions is similar to that observed for introgressed regions in Europeans. Additionally, we identify numerous loci that harbor signatures of local adaptation, including genes involved in immunity, metabolism, olfactory and taste perception, reproduction, and wound healing. Within the Pygmy population, we identify multiple highly differentiated loci that play a role in growth and anterior pituitary function and are associated with height.

Excerpts from original manuscript:

Hunter-gatherer Genomes Possess Signatures of Archaic Admixture

Gene flow between anatomically modern humans and archaic species has been described for European, Melanesian, and African populations (Hammer et al., 2011; Plagnol and Wall, 2006; Wall et al., 2009; Reich et al., 2010). To detect putatively introgressed regions in the Pygmy, Hadza, and Sandawe hunter-gatherer populations, we modified the summary statistic S*, which searches for clusters of population specific SNPs in near complete LD, to be suitable for genome-scale analyses. S* has previously been used to detect archaic admixture in individuals of European and African descent (Hammer et al., 2011; Plagnol and Wall, 2006; Wall et al., 2009). We first verified that our implementation of S* could accurately identify introgressed regions by performing extensive coalescent simulations (Supplemental Information, Figure S7) and analyzing publicly available whole-genome sequences from 9 CEPH and 4 Tuscan individuals sequenced by Complete Genomics. We calculated S* in 50kb sliding windows, and identified the top 350 regions (top ~0.4%) in each population with unusually large values of S* as high-confidence candidates for introgression. T_{MRCA} distributions for these regions are significantly larger than the distribution for all loci ($p < 10^{-16}$), consistent with the hypothesis that they are enriched for introgression (Figure 2A). Moreover, non-African genomic regions with high values of S* were significantly enriched for Neanderthal-specific SNPs ($p < 10^{-16}$, Figure 3B). Thus, S* can robustly detect genomic regions inherited from archaic ancestors.

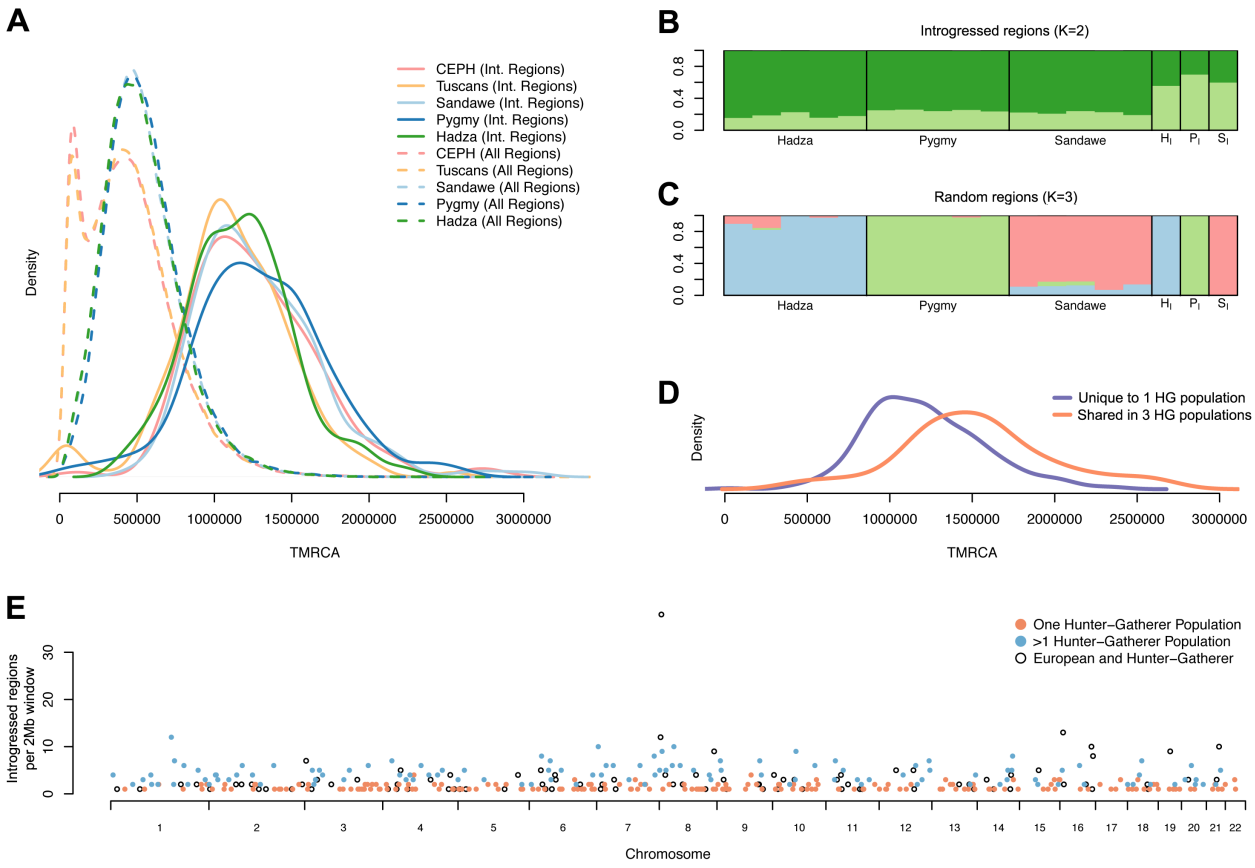

Figure 2. Times until most recent common ancestry and evidence of archaic introgression.

(A) TMRCA of top candidate regions (solid lines), and of all regions (dotted lines) for the Pygmy, Hadza, and Sandawe hunter-gatherer populations and two European populations. Note, TMRCA represents the estimated time of divergence between the anatomically modern human and candidate introgressed sequences (Supplemental Information). TMRCA for top candidate regions is significantly older than random genomic regions (Kruskal-Wallis test, $p < 2.2\times10\text{-}16$), but TMRCA for top candidate regions from each population are not significantly different (Kruskal-Wallis test, p=1). **(B)** and **(C)** STRUCTURE plots showing the proportion of ancestry for each individual based on the most likely number of subpopulations (K = 2 for putatively introgressed regions in Panel B and K =3 for random regions in Panel C). For each population, a 'virtual' genome was constructed by concatenating sequence from individuals containing the putatively introgressed sequence **(B)** or from arbitrary individuals **(C)**. Pi, Hi, and Si denote the virtual genomes constructed for the Pygmy, Hadza, and Sandawe samples, respectively. **(D)** TMRCA of top candidate regions for introgression unique to a single hunter-gatherer population is significantly lower than TMRCA of regions shared between all hunter-gatherer populations (Wilcoxon rank sum test, $p = 2.2\times10\text{-}5$). **(E)** Genomic distribution of the top 350 introgressed regions for the Pygmy, Hadza, and Sandawe populations and two European populations, in 2Mb windows. Colors indicate whether windows contain introgressed regions from a single hunter-gatherer population (orange), multiple hunter-gatherer populations (blue), or hunter-gatherer and European populations (open black circle). Counts are for hunter-gatherer regions only. See also Figure S7.

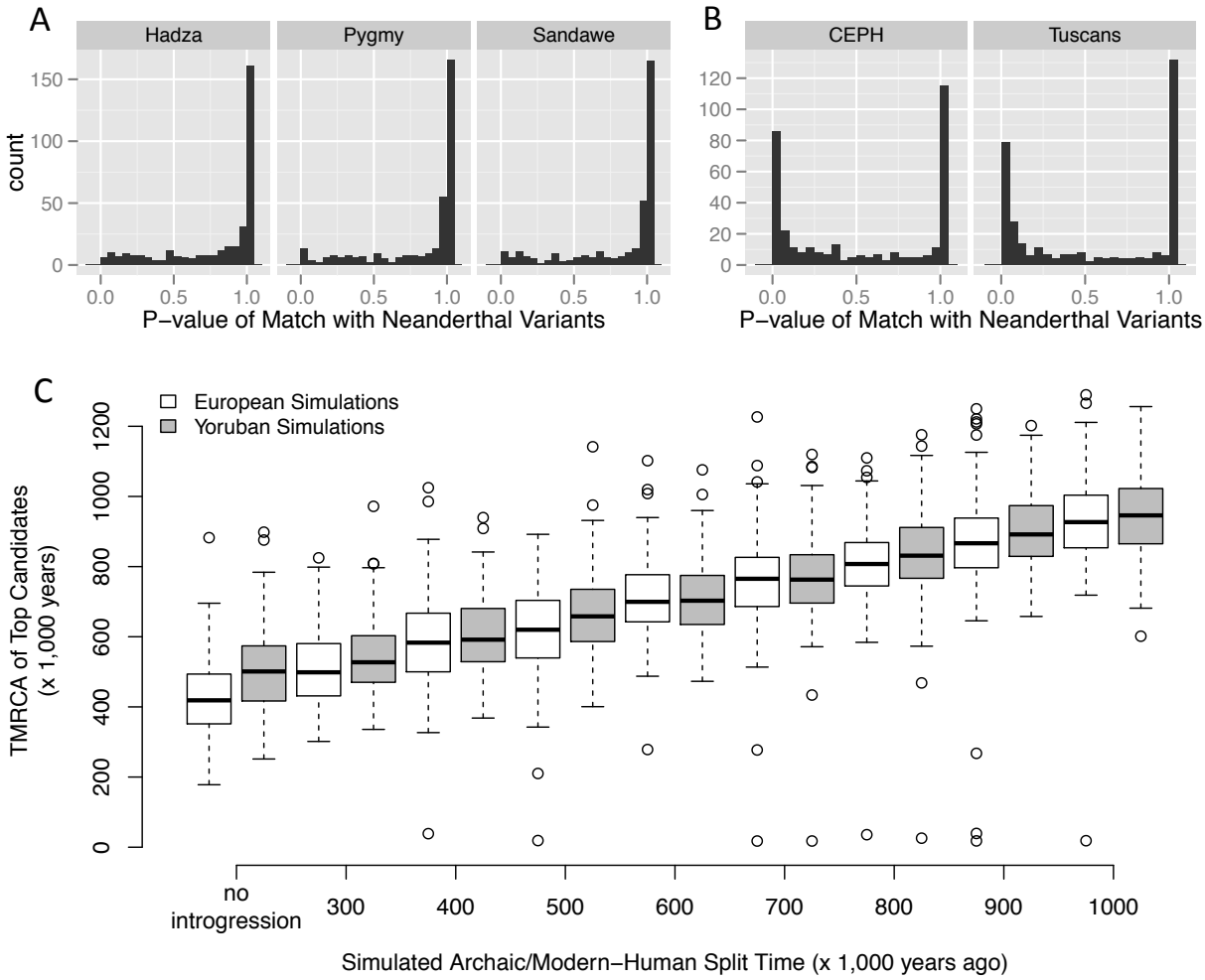

Figure 3. Characteristics of S* in real and simulated data.
(A-B) Neanderthal variants are not enriched in top candidate regions for three hunter-gatherer populations **(A)**, but are enriched in top candidate regions from two European populations **(B)**. **(C)** T_{MRCA} estimates for top 0.5% of 50kb regions in simulated data, varying time of split with the archaic population from 300kya to 1000kya; introgression was simulated into Europeans (white boxes) and Yorubans (gray boxes). See also Figure S7.

We next used S* to identify putatively introgressed regions in the African hunter-gatherer samples. In all three African hunter-gatherer samples, we found evidence of introgression from at least one archaic population. Strikingly, the median T_{MRCA} for putatively introgressed haplotypes in the hunter-gatherer samples is similar to the median T_{MRCA} for introgressed haplotypes in Europeans (1.2-1.3 Mya vs 1.1-1.2 Mya, respectively; Figure 2A), suggesting that the archaic African population diverged from anatomically modern humans in the same time frame as Neanderthals (simulations suggest that relative time of split with archaic populations can be recovered via T_{MRCA}; Figure 3C). Additionally, we performed a STRUCTURE analysis of the putatively introgressed regions and of 350 random regions. If candidate regions identified by unusually large values of S* are enriched for genuine introgressed sequence, then we would expect STRUCTURE to identify two populations, as introgressed regions primarily consist of individuals carrying one archaic and one anatomically modern haplotype. In contrast, we would expect STRUCTURE to identify three populations in the randomly selected regions corresponding to the Pygmy, Hadza and Sandawe populations. Indeed, this is precisely what we find (Figure 2B and 2C), further demonstrating that top ranked S* regions are enriched for putatively introgressed sequence.

There is significant overlap ($p < 10^{-16}$) among putatively introgressed regions in the three hunter-gatherer populations, consistent with either gene flow among the hunter-gatherer populations or introgression events that predate population splitting of these populations. In addition, the T_{MRCA} of introgressed regions shared between all three populations is significantly older compared to introgressed regions observed in only one population (Wilcoxon rank-sum test, $p = 2.2\times10^{-5}$, Figure 2D), consistent with an introgression event predating the divergence of these populations. In contrast, we observed few introgressed regions that overlap with those observed

outside of Africa. One exception is a 2 Mb window on chromosome 8 (Figure 2E, chr8:3Mb-5Mb) that contains introgressed regions in all global populations. However, we note that because the chr8:3Mb-5Mb region is enriched for CNVs, it may be more prone to false positives (Supplemental Information).

Identification of Genomic Regions with Extreme Times to Most Recent Common Ancestry

We scanned the genomes of African hunter-gatherers to identify regions with extremely long or short coalescence times, which are likely to be enriched for targets of natural selection. To this end, we calculated the time to most recent common ancestor (T_{MRCA}) for 50kb sliding windows in the 15 hunter-gatherer genomes. The mean autosomal T_{MRCA} across all windows is 796kya. As expected, windows spanning the HLA region, which exhibits strong signatures of balancing selection (Barreiro and Quintana-Murci, 2010), are the most ancient in the genome, with a maximum T_{MRCA} of 5.1 million years for a 50kb window encompassing *HLA-G*. The oldest genic regions outside of the HLA locus include *NSUN4*, *HCG9*, *MYO3A*, and *APOBEC4*. Conversely, we also found multiple genomic regions with short T_{MRCA} times (< 10 kya) including multiple tripartite motif containing genes (*TRIM53P*, *TRIM64*, and *TRIM64B*), the *SPAG11A* gene that is involved in sperm maturation, and *NCF1*, which is a subunit of neutrophil NADPH oxidase.

Discussion:

Evidence of Archaic Introgression

A striking finding in our dataset is that compelling evidence exists that extant hunter-gatherer genomes contain introgressed archaic sequence, consistent with previous studies (Hammer et al., 2011; Plagnol and Wall, 2006; Reich et al., 2010; Shimada et al., 2007; Wall et al., 2009). We note that unambiguous evidence of introgression is difficult to obtain in the absence of an archaic reference sequence, which currently does not exist and may never be feasible given the rapid decay of fossils in Africa. Although we carefully filtered our data set in an attempt to analyze only high-quality sequences (Supplementary Information), it is possible that unrecognized structural variants or other alignment errors could generate a spurious signature similar to introgression. Encouragingly, we did not see an enrichment of structural variation calls in our candidate introgression regions. Additionally, through extensive simulations and analysis of European whole-genome sequences (Supplementary Information), we have demonstrated that the signatures of introgression we observed are unlikely to be entirely accounted for due to other aspects of population demographic history, natural selection, or sequencing errors. Moreover, we did not find strong evidence that introgressed regions were clustered in the genome more often than expected by chance ($p > 0.05$; Supplemental Information). Nor did we find significant evidence that introgressed regions were enriched in genic regions ($p > 0.05$); rather, genic regions were significantly depleted for introgression in several populations (Supplemental Information). Therefore, the simplest interpretation of these data is that introgressed regions in extant human populations represent neutrally evolving vestiges of archaic sequences. In short, we find that low levels of introgression from an unknown archaic population, or populations, occurred in the three African hunter-gatherer samples examined, consistent with findings of archaic admixture in non-Africans (Reich et al., 2010).

Excerpts from Supplementary Information:

Samples for Whole-genome Sequencing, Quality Control, and Identification of Variants

Prior to sample collection, informed consent was obtained from all research participants, and permits were received from the Ministry of Health and National Committee of Ethics in Cameroon and from COSTECH and NIMR in Dar es Salaam, Tanzania. In addition, appropriate IRB approval was obtained from both the University of Maryland and the University of Pennsylvania. Although the term 'Pygmy' has historically been pejorative, it has recently been used by indigenous groups themselves as well as activist groups working on their behalf (Ballard, 2006; Leonhardt, 2006; Pelican, 2009). Acknowledging this recent trend and the absence of a better term that encompasses the hunting and gathering peoples from Cameroon, we use the word 'Pygmy' to collectively refer to Baka, Bakola, and Bedzan individuals in our study. Hadza samples were collected at sites near Lake Eyasi in the Arusha region and Sandawe samples were collected in the Kondoa district in the Dodoma region of Tanzania. Individuals were chosen to be unrelated based on microsatellite (Tishkoff et al., 2009) and genome-wide SNP (Jarvis et al., 2012) data analyses, including a pi-hat filter of 0.25 using PLINK (Purcell et al., 2007). However, we note that the small population size of the Hadza means that token amounts of relatedness are shared between samples. White blood cells were isolated in the field from whole blood with a salting out procedure modified from (Miller et al., 1988) and DNA was extracted in the lab with a Purgene™ DNA extraction kit (Gentra Systems Inc., Minneapolis, MN). Because DNA was obtained from whole blood, we avoid possible artifacts that can arise from use of cell lines (Maitra et al., 2005).

Hunter-gatherer genomes were sequenced at >60x coverage (Table 1) using the

combinatorial probe-anchor ligation and DNA nanoarray technology of Complete Genomics. The standard Complete Genomics bioinformatics pipeline (Assembly Pipeline version 1.10 and CGA™ Tools 1.4) was used for sequence alignment, read mapping, assembly, and data analysis. This pipeline uses stringent criteria to make variant calls (likelihood ratios >100:1 are required to make homozygous variant calls and likelihood ratios >10,000:1 are required to make heterozygous variant calls), and published error rates are less than 1/100,000bp (Drmanac et al., 2010). Importantly, calls for different individuals were independent (otherwise, our allele frequency distributions would underestimate rare alleles). To assess genotyping accuracy we sequenced two additional genomes as technical replicates (each technical replicate was a duplicate of one of the five sequenced Hadza genomes). Data from our technical replicates revealed low error rates: 26,415 and 28,292 discordant variant calls are found for each pair of technical replicates.

As an additional test of genotyping accuracy we compared calls from whole-genome sequencing and the Illumina1M-duo BeadChip array (of which we had data for 14 of 15 hunter-gatherers). Ignoring A/G and C/T sites (to avoid strand flipping issues in our Illumina1M-duo dataset), calls at total of 743,516 SNPs were compared. For each individual, concordance between platforms was very high (mean=0.999564, range=0.999463 to 0.999635). In practice, only one out of every 2294 variant calls differed between platforms. Median coverage was similar for concordant (48x) and discordant (49x) SNPs, suggesting that errors were not due to poor coverage during whole-genome sequencing. Furthermore, a total of 152 SNPs were classified as highly-discordant (>50%) between genotyping technologies, and 32/152 highly-discordant SNPs were found to be tri-allelic after cross-referencing with dbSNP.

Prior to quality control filters, we observed 15,748,468 variants in hunter-gatherer genomes. We then filtered our data based on "missingness" and departure from Hardy- Weinberg filters. Some genomic regions (such as centromeres and telomeres) are less likely to be successfully sequenced, and we find that sites called successfully in only a subset of individuals are more likely to be discordant (Figure S1). Because of this, we used a "missingness" filter, whereby sites called in <80% of all individuals were excluded from analysis, eliminating 2,311,725 variants (Figure S1). Because genotyping error can yield abnormal proportions of heterozygotes and homozygotes, we also used quality control filter to detect departures from Hardy-Weinberg proportions. We note that even sites under strong selection are expected to pass a departure from Hardy-Weinberg proportion filter (Lachance, 2009a). To avoid artifacts due to population stratification (such as Wahlund effects) we summed Chi-square statistics for each hunter-gatherer population and excluded all sites with Chi-square values ≥13. In practice this involved excluding sites where every individual was heterozygous, and this quality control filter eliminated a further 16,425 variants (Figure S1). We then merged data from pairs of genomes that contained technical replicates. This involved resolving discordant calls, and eliminating an additional 12,801 variants. After all quality control filters, a total of 13,407,517 variants remained.

We note the following additional details: Genomic coordinates used in this paper refer to build37/hg19 of the human genome. For population genetics analyses (including calculation of θ, allele frequencies, and F_{ST}) we treated partially called sites as missing data. We note that the following analyses were restricted to SNPs (as opposed to SNPs and indels): PCA, NJ tree, Neutrality Index calculations, T_{MRCA} scans, archaic introgression, and LSBL scans. In a previous study (Lam et al., 2012), the sequencing technology of Complete Genomics was found to be highly accurate in

detecting indels (22 of 23 successfully amplified indels were validated). However, Lam et al. also note that indel detection by Complete Genomics lacks sensitivity, indicating that the number of indels discovered in hunter-gatherer genomes may be an underestimate.

T_{MRCA} Calculations

TMRCA was calculated using a previously described approach (Hudson, 2007; Thomson et al., 2000), which computes the average coalescent time from nucleotide substitutions assuming that mutations are Poisson distributed. This value is then converted to an estimate of T_{MRCA} in years by computing the divergence between chimpanzee and humans for this region (*D*), and setting the molecular clock to 12My/*D* (assuming the divergence time between humans and chimpanzee is 6 million years). Human/chimpanzee alignments were downloaded from the UCSC Genome Browser (reference versions GRCb37 and panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/syntenicNet/).

We estimated the Time to Most Recent Common Ancestor (T_{MRCA}) for a set of samples for 50kb sliding windows (20kb step) across all autosomes. We use the method of (Thomson et al., 2000), which computes the average T_{MRCA} in nucleotide substitutions for a set of sequences. Using Equation 1 of (Hudson, 2007) we calculated the T_{MRCA} for each 50kb window. This value was then converted to an estimate of T_{MRCA} in years by computing the divergence between chimpanzee and human for this region (*D*), and setting the molecular clock to 12My/*D* (i.e. we normalize by window-specific mutation rates assuming that humans and chimpanzees split 6Mya).

Human/chimpanzee alignments were downloaded from the UCSC Genome Browser, and the more conservative syntenicNet alignments were used (reference versions GRCh37 and panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/syntenicNet/). For each autosomal variant identified using whole-genome sequencing, ancestral and derived states were calculated using a maximum likelihood method. We used DNAmL (Felsenstein, 2005) to evaluate each site, using a single base alignment consisting of the human reference base, the alternative allele found in our dateset, chimpanzee, orangutan, and macaque bases. DNAmL can handle missing data, and missing bases were marked as such. To allow for the possibility of incomplete lineage sorting between chimpanzee and human, three trees were considered for each site, and the highest likelihood tree was selected. Only states with probability > 0.95 were selected. If the ancestral state was unable to be inferred due to alignment difficulties or low probability, then the most parsimonious state was selected based on the 68 unrelated hunter-gatherer and Complete Genomics genomes (i.e. the minor allele was assumed to be the derived allele).

Archaic Introgression

To detect putatively introgressed archaic sequence, we employed a modified version of the S* statistic (Plagnol and Wall, 2006). To mitigate potential confounding effects of mutation rate heterogeneity and paralogous variants, CpGs and repetitive sequences (as defined by RepeatMasker) were removed, and only fully called sites were considered. After filtering, 1.2Gb of sequence was retained. To enable comparisons between windows, we modify S* by dividing by the square of the sequence length after filtering of each window. S* also requires a target

population (in which we search for introgressed sequence) and a reference population. For African populations, we used 13 European genomes (9 CEPH and 4 Tuscan) as the reference, and for non-African populations we used 9 unrelated Yoruban genomes (each genome sequenced by Complete Genomics).

We calculated S* in 50kb windows, using a step size of 20kb. Windows were removed from consideration if they did not contain at least 20kb of unfiltered sequence, leaving ~2Gb of sequence. The top 350 windows (~0.4%) as ranked by S* were considered as high-confidence candidates for introgression, which is likely a very conservative threshold as determined by extensive coalescent simulations and STRUCTURE analyses.

Only fully called autosomal sites were considered for introgression analyses. In addition, CpG sites and regions identified by RepeatMasker were removed from all S*, T_{MRCA} and STRUCTURE analyses (1.2Gb of sequenced retained after filtering). We identified CpG sites by analyzing the human reference genome (hg19), genomes sequenced by Complete Genomics (15 hunter-gatherer genomes and the public data release), and three additional primate genomes (chimpanzee, orangutan, or macaque). To additionally account for ancestral CpGs, we counted any position that is C or G in one genome, and adjacent to a G or C (respectively) in another genome. This CpG metric is less conservative than the metric used for placental mammals, but more conservative than the metric used for primates in (McVicker et al., 2009).

To test the effects of various demographic parameters on our method, we simulated archaic introgression into either European or African populations. As a starting point, we used demographic parameters from (Plagnol and Wall, 2006), which models Europeans and Yorubans. In addition, we also used the following base parameters: a fixed mutation rate of

1.1×10^{-8}, a fixed recombination rate of 1×10^{-8}, 700kya split time between archaic and modern human populations, and 25kya time of introgression. Simulations were then run for a range of parameter values (Figure S7A, varying a single parameter and keeping other parameters at the base value).

S* was computed using a dynamic programming algorithm, with a running time of $O(n \bullet s^2)$, where n is the number of individuals and s is the number of variants in a region. For each parameter set we simulated 50kb windows over a range of archaic-to-modern-human migration parameters (0-2.4%, corresponding to 0-4% introgressed sequence per individual). We then identified the top 0.5% of each simulation, as ranked by S*. These are analogous to the 350 top candidate regions used for several of our introgression analyses. To show that these regions are likely to contain introgressed sequence, we determined the false discovery rate (FDR) for each parameter setting (Figure S7A). For introgression levels above 1%, the FDR for most parameters is close to zero. Higher FDR is seen for very low recombination rates, some hotspot models with heterogeneous recombination rates, and for a recent time of split with the archaic population. We also calculated the distribution of normalized S* for simulated datasets and populations with whole-genome sequence data. In simulated data, the extent of positive skew in the distribution of normalized S* was correlated with the extent of archaic introgression (Figure S7D, simulated data with heterogeneous recombination rate: 80% 0.2×10^{-8}, 20% 4.2×10^{-8}). We see a similar positive skew in each of the studied populations, and patterns were similar for each population (Figure S7E), consistent with broadly similar levels of archaic introgression. However, several demographic parameters can affect the distribution of S*. For this reason, we further investigated the characteristics of putatively introgressed regions.

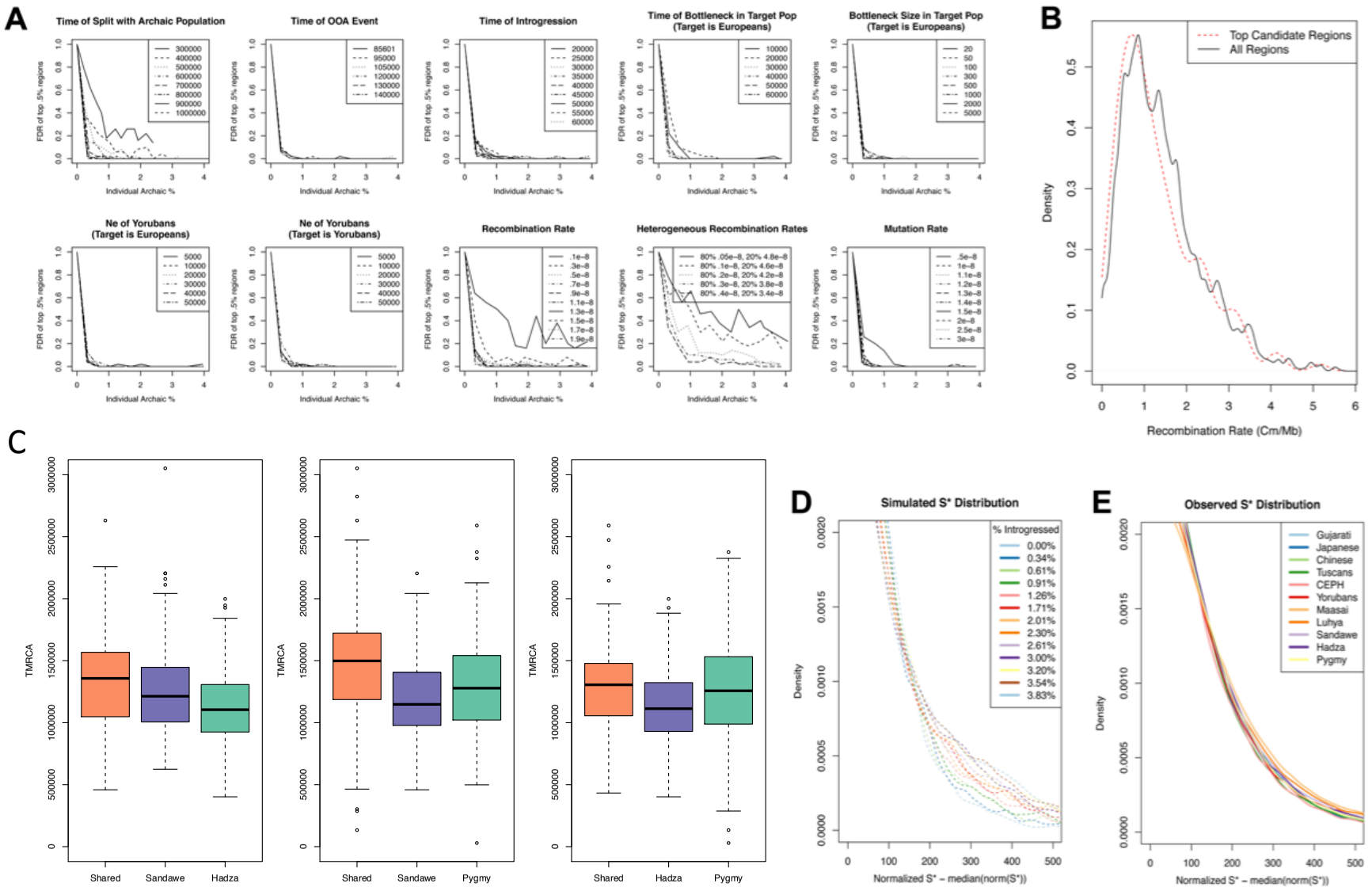
A
Time of Split with Archaic Population
Time of OOA Event
Time of Introgression
Time of Bottleneck in Target Pop (Target is Europeans)
Bottleneck Size in Target Pop (Target is Europeans)
Ne of Yorubans (Target is Europeans)
Ne of Yorubans (Target is Yorubans)
Recombination Rate
Heterogeneous Recombination Rates
Mutation Rate
FDR of top .5% regions
Individual Archaic %
B
Top Candidate Regions
All Regions
Density
Recombination Rate (Cm/Mb)
C
TMRCA
Shared
Sandawe
Hadza
Pygmy
D
Simulated S* Distribution
% Introgressed
0.00%
0.34%
0.61%
0.91%
1.26%
1.71%
2.01%
2.30%
2.61%
3.00%
3.20%
3.54%
3.83%
Density
Normalized S* – median(norm(S*))
E
Observed S* Distribution
Gujarati
Japanese
Chinese
Tuscans
CEPH
Yorubans
Maasai
Luhya
Sandawe
Hadza
Pygmy
Density
Normalized S* – median(norm(S*))

Figure S7. S* statistics are robust at detecting introgression, related to Figures 2 and 3. **(A)** False Discovery Rate (FDR) for top 0.5% of simulated 50kb regions, ranked by S*, under several demographic parameters. Panel **(B)** Recombination rate of top candidate regions (red dashed line), and of all regions, in three hunter gatherer populations. **(C)** T_{MRCA} estimates for shared and unique top candidate regions in Pygmy, Hadza and Sandawe. **(D)** Tail of simulated S* distribution over 0-3.8% introgressed sequence per individual. **(E)** Tail of observed S* distribution for 11 populations sequenced by Complete Genomics.

In addition to determining the false discovery rate (FDR) for top candidate regions, we compared top candidate regions for each population to the draft Neanderthal sequence. Due to the low coverage of the draft Neanderthal sequence (Green et al., 2010), we used a comparison method that is less sensitive to errors in the Neanderthal sequence than direct sequence comparison. Specifically, for each region we performed a 2x2 Chi-square test for enrichment of variants matching the Neanderthal sequence that are unique to the target population, in comparison with variants that are not unique to the target population (i.e., present in the reference population). Variants are only considered if they are covered by at least two Neanderthal reads. Top candidate regions in non-African populations show a clear excess of regions with significant enrichment of Neanderthal variants (Figure 3B); top candidate regions from African populations do not show this enrichment (Figure 3A), demonstrating that our top candidate regions are enriched for archaic introgression.

To increase the chance that analyzed genomic regions include entire introgressed haplotypes, we identified a subset of the 350 top candidate regions in which all introgressed variants for a given 50kb region are found in a single individual. This restricted top-candidate dataset (usually about 50% of the top 350 regions) was used for T_{MRCA} and STRUCTURE analyses of introgressed regions. Note that the T_{MRCA} values given in Figure 2A, Figure 2D, and Figure S7C are calculated on single individuals from this subset (to better estimate the time of divergence of the two haplotypes contained in a single individual), whereas T_{MRCA} estimates in the **T_{MRCA} estimates** section of the main text are calculated on entire populations.

To identify population substructure in putatively introgressed regions, we performed a model based clustering analysis using STRUCTURE (Pritchard et al., 2000). Putatively introgressed regions should primarily consist of a single individual containing one introgressed

haplotype and one modern human haplotype, with the remaining individuals in the population containing entirely modern human haplotypes. In this situation, the differences between archaic and modern sequence should be more pronounced than the differences between Pygmy, Sandawe and Hadza. To test this hypothesis, we analyzed the top putative introgressed regions from each hunter-gatherer population, as well as a similar number of random sequences from each hunter-gatherer population using STRUCTURE (Pritchard et al., 2000). Because introgressed sequences are likely to be at low frequency, no single individual will contain more than a small amount of introgressed sequence, and STRUCTURE is not well suited to identifying subpopulations that have no extant, or representative, individuals. To combat this problem, we created a "virtual genome" for each hunter-gatherer population, where each region is composed of genotypes from a single individual identified as containing the putatively introgressed haplotype (Figure 2). If a region is a candidate introgressed region in multiple populations, each of those populations' virtual genome contains the introgressed region. If a region is not a candidate in a given population, the virtual genome for that region and population is set to missing data. These virtual genomes are expected to consist of roughly 50% archaic and 50% modern human haplotypes.

We then performed a STRUCTURE analysis using standard settings, with a burn-in of 100k and run length of 100k. To reduce run time and the affect of LD, we performed each analysis on a random 10% of the selected sites. For each parameter setting, we computed three runs each of K=1-4, and selected the highest log likelihood of the three runs for each K. For random regions (the negative control), K=3 and K=4 have the highest log-likelihood, but all settings of K produce results where each virtual genome is composed almost entirely of a single population, i.e. there is no evidence of a substantial archaic component in these regions. For the 350 top candidate regions, K = 2 has the highest log-likelihood. However, all settings of K

produce results where each virtual genome is composed of a ~50/50 mixture of two populations, supporting the hypothesis that these regions are significantly enriched for introgressed archaic sequence. As noted above, rare introgressed haplotypes are expected to be heterozygous, fitting with the observed ~50/50 mixture.

Simulations suggest that the relative time of split between archaic and modern human populations can be recovered via a T_{MRCA} analysis. Specifically, we varied the time of split from 300kya to 1000kya, simulated migration levels from 0.01 to 0.024, and selected the top 50 regions (0.5%) from each simulation. Simulated archaic-modern human split time vs. recovered T_{MRCA} is given in Figure 3C. Empty boxes show simulated introgression events involving Europeans and grey boxes show simulated introgression events involving Yorubans.

To test whether putatively introgressed sequences are enriched in coding regions, we compared the distribution of top candidate regions for each population with the distribution of coding sequences, using the hypergeometric distribution. To minimize the effects of overlapping windows, we used every third 50kb window. Gene definitions were obtained from the UCSC Table Browser, RefSeq Genes track, refFlat table. Exons were extended by 2 bp, and overlapping exons were merged using BEDOPS (Neph et al., 2012). Top candidate regions were not enriched for coding sequence compared to the rest of the genome ($p > 0.05$). Rather, for multiple populations we found that top candidate regions were significantly depleted for coding sequence ($p < 0.01$ for Hadza, Yoruban, CEPH, and Tuscani; $p < 0.05$ for Massai, Luhya, Chinese and Japanese). Top candidate regions in Pygmy, Sandawe and Gujarati populations were neither significantly enriched nor depleted for coding sequence, although in all three populations there were fewer overlaps between top candidate regions and coding sequence than expected by chance.

To determine if putatively introgressed sequences are clustered in the genome, we performed permutation tests to obtain the distribution of the expected number of 50kb introgressed windows within a 2Mb region. To minimize the effects of overlapping windows, we used every third 50kb window. Although the actual and expected distributions for each hunter-gatherer population are not significantly different (Wilcoxon rank-sum test, $p > 0.05$), we did see a few regions in each population with more top candidates than expected. In all three hunter gatherer populations, the region with the most top candidates is at chr8:3Mb-5Mb. This could be due to shared ancestral introgression, an increased proclivity for introgression involving this region, or enrichment of false positives due to an excess of CNVs in that region (Shaikh et al, 2009), which could lead to sequencing errors. However, several regions in this window from each non-African population are significantly enriched for Neanderthal variants. These enrichments are unlikely to be due to errors in the Neanderthal sequence caused by the complex structure of this window, as none of the hunter-gatherer regions in this window are enriched for Neanderthal variants.

Overlap between putatively introgressed regions was found for each pair of hunter-gatherer populations, and T_{MRCA} was estimated for introgressed regions found in a single population or shared between two populations (FigureS7). We also examined whether putatively introgressed regions are found in regions of high or low recombination (Figure S7B). While there is a slight shift towards lower recombination rates for top candidate regions, the distributions overlap to a large extent, suggesting that low recombination rates are not a major feature of top candidate regions (recombination rates from (Kong et al., 2002)).

References and Notes:

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Ballard, C. (2006). Strange Alliance: Pygmies in the Colonial Imaginary. World Archaeology *38*, 133-151.

Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet *11*, 17-30.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G.*, et al.* (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78-81.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package). Cladistics *5*, 164-166.

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.

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

Hudson, R.R. (2007). The variance of coalescent time estimates from DNA sequences. J Mol Evol *64*, 702-705.

Jarvis, J.P., Scheinfeldt, L.B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.M., Beggs, W., Hoffman, G.*, et al.* (2012). Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. PLoS genetics *8*, e1002641.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G.*, et al.* (2002). A high-resolution recombination map of the human genome. Nature genetics *31*, 241-247.

Lachance, J. (2009a). Detecting selection-induced departures from Hardy-Weinberg proportions. Genet Sel Evol *41*, 15.

Lam, H.Y., Clark, M.J., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F.E., Habegger, L., Ashley, E.A., Gerstein, M.B., Butte, A.J.*, et al.* (2012). Performance comparison of whole-genome sequencing platforms. Nat Biotechnol *30*, 78-82.

Leonhardt, A. (2006). Baka and the Magic of the State: Between Autochthony and Citizenship. African Studies Review *49*, 69-94.

Maitra, A., Arking, D.E., Shivapurkar, N., Ikeda, M., Stastny, V., Kassauei, K., Sui, G., Cutler, D.J., Liu, Y., Brimble, S.N.*, et al.* (2005). Genomic alterations in cultured human embryonic stem cells. Nature genetics *37*, 1099-1103.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics *5*, e1000471.

Miller, S.A., Dykes, D.D., and Polesky, H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic acids research *16*, 1215.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S.*, et al.* (2012). BEDOPS: High performance genomic feature operations. Bioinformatics.

Pelican, M. (2009). Complexities of Indigeneity and Autochthony: An African Example. American Ethnologist *36*, 52-65.

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

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

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.

Shaikh, T.H., Gai, X., Perin, J.C., Glessner, J.T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L.K., D'Arcy, M.*, et al.* (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome research *19*, 1682-1690.

Shimada, M.K., Panchapakesan, K., Tishkoff, S.A., Nato, A.Q., Jr., and Hey, J. (2007). Divergent haplotypes and human history as revealed in a worldwide survey of X-linked DNA sequence variation. Molecular biology and evolution *24*, 687-698.

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.* (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science *337*, 64-69.

Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J., and Feldman, M.W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A *97*, 7360-7365.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O.*, et al.* (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035-1044.

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

Acknowledgements:

This work was supported by NSF (BCS-0827436) and NIH (R01GM076637, 8 DP1 ES022577-04) grants to S.A.T., an NIH NRSA postdoctoral fellowship (F32HG006648-01) to J.L., Rubicon Grants of the Netherlands Organization of Scientific Research to C.E. and B.F., and support from the Center of Excellence in Environmental Toxicology at the University of Pennsylvania,P30-ES013508-07. We thank J. Hirbo, J. Jarvis, A. Rawlings, L. Scheinfeld, and S. Soi, for their critical feedback and advice, and K. Addya, D. Baldwin, and B. Beggs for assistance in genotyping the samples. We also thank the 15 individuals who graciously supplied their DNA. Data reported in this paper will be available by request and at the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and dbGaP (http://www.ncbi.nlm.nih.gov/gap/) websites.

Chapter 3 : Resurrecting Surviving Neanderthal Lineages from Modern Human Genomes

This chapter was published in Science on February 28, 2014.

Vernot, B., and Akey, J.M. (2014). Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science 343, 1017–1021.

Abstract:

Anatomically modern humans overlapped and mated with Neanderthals such that non-African humans inherit ~1-3% of their genomes from Neanderthal ancestors. We identified Neanderthal lineages that persist in the DNA of modern humans, in whole-genome sequences from 379 European and 286 East Asian individuals, recovering over 15 Gb of introgressed sequence that spans ~20% of the Neanderthal genome (FDR = 5%). Analyses of surviving archaic lineages suggests that there were fitness costs to hybridization, admixture occurred both before and subsequent to divergence of non-African modern humans, and Neanderthals were a source of adaptive variation for loci involved in skin phenotypes. Our results provide a new avenue for paleogenomics studies, allowing substantial amounts of population-level DNA sequence information to be obtained from extinct groups even in the absence of fossilized remains.

Main Text:

Hybridization between closely related species, and the concomitant transfer or introgression of DNA, is widespread in nature (*1,2*). In hominin evolution, the sequencing of Neanderthals (*3*) and their sister lineage, Denisovans (*4,5*), provided evidence for introgression of these lineages into modern humans. Specifically, ~1-3% of each non-African human genome is estimated to have been inherited from Neanderthals (*3,5*). Although initial inferences of introgression between Neanderthals and humans may not have been robust to alternative explanations, most

notably archaic population structure (*3,6*), subsequent analyses have provided evidence for gene flow (*7-9*).

We hypothesized that a substantial amount of the Neanderthal genome may be recovered from the analysis of contemporary humans despite the limited amounts of admixture, as introgressed sequences may vary among individuals (Fig. 1A). Indeed, coalescent simulations for a broad range of admixture models suggest that 35-70% of the Neanderthal genome persists in the DNA of present-day humans (Fig. S1; Fig. S2; *10*). By identifying Neanderthal sequence from a large sample of modern humans, we hope to discover surviving lineages that may come from multiple archaic ancestors (Fig. 1A), allowing population level data to be recovered.

To identify surviving Neanderthal lineages, we developed a two-stage computational strategy (Fig. S3; *10*). First, we identify candidate introgressed sequences using an extension of a previously developed summary statistic referred to as *S** (*11*), which is sensitive to the signatures of introgression (Fig. 1B) and is calculated without using the Neanderthal reference genome. We performed coalescent simulations for a wide variety of demographic scenarios, and found that our implementation of *S** can distinguish introgressed from non-introgressed sequences (Fig. 1C; Fig. S4). Second, we refine the set of candidate introgressed sequences using an orthogonal approach by comparing them to the Neanderthal reference genome and testing whether they match significantly more than expected by chance (*10*). We estimate that using *S** alone, as compared to our two-staged approach, would recover ~30% of Neanderthal lineages at a FDR = 20% (Fig. S5; *10*).

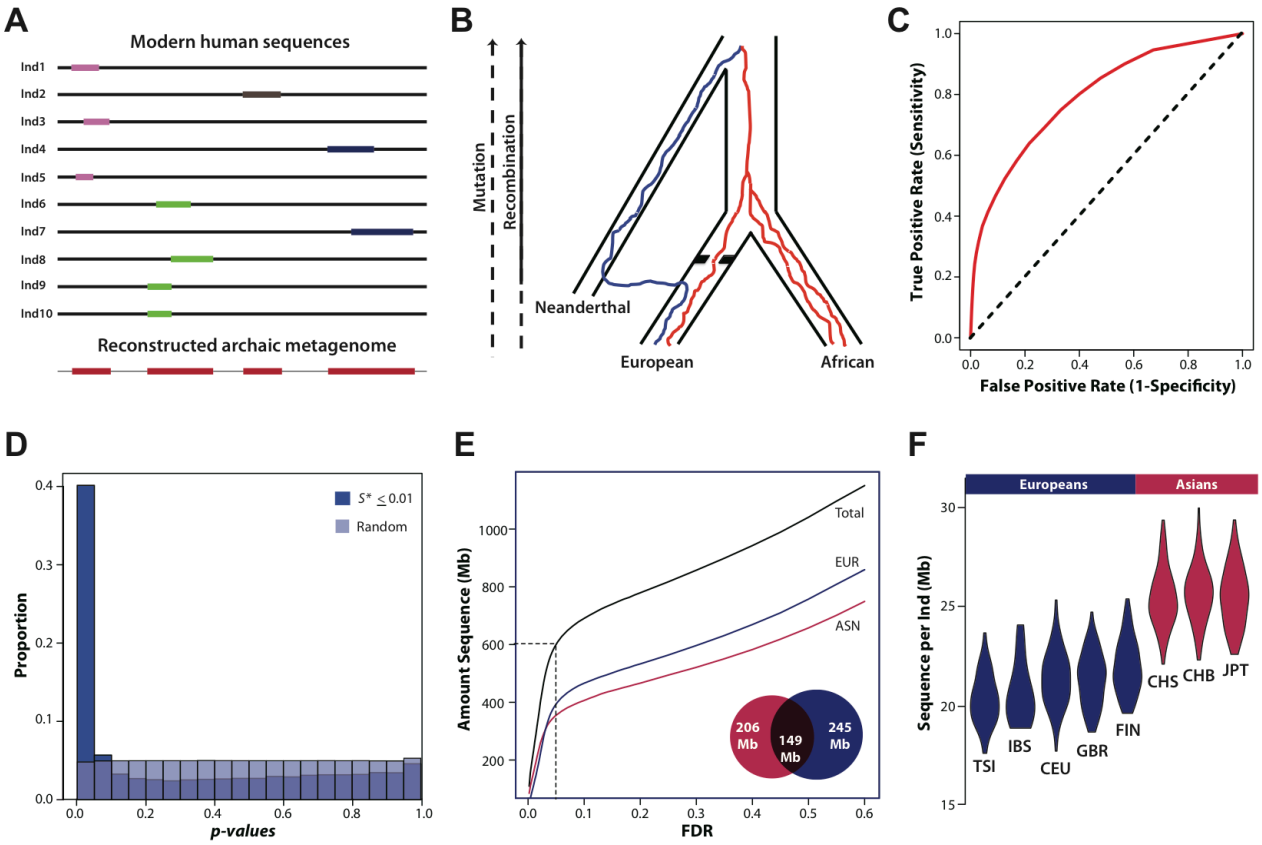

Fig. 1. Recovering Neanderthal lineages from the DNA of modern humans.
(A) Schematic representation illustrating that low levels of introgression may facilitate the recovery of substantial amounts of archaic sequence. Lines represent DNA from contemporary individuals and colored boxes indicate archaic sequences. Different colored boxes represent sequences inherited from distinct archaic ancestors. **(B)** Genealogies of loci in Europeans and Africans in the presence of introgression. The expected signature of an introgressed lineage (blue) that our method exploits is high levels of divergence that persists over relatively long haplotype blocks. **(C)** Receiver operator curve illustrating the performance of S* for detecting introgressed sequence in simulated data (10). The diagonal line represents random predictions. **(D)** Distribution of p-values testing for an enrichment of Neanderthal variants for S* candidate and randomly selected regions. **(E)** Amount of Neanderthal sequence recovered as a function of FDR. The inset Venn diagram shows the amount of sequence overlap between East Asians and Europeans at a FDR = 5%. **(F)** Violin plots showing the distribution of amount of introgressed sequence identified per individual for East Asian and European populations (population abbreviations described in SOM).

We applied this framework to whole-genome sequences from 379 Europeans and 286 East Asians, from the 1000 Genomes Project (Table S1; *12*). Specifically, we calculated S^* in 50kb sliding windows (Tables S2-S8; *10*), and determined statistical significance through coalescent simulations using a computationally efficient approach (Fig. S6; *10*). At an S^* threshold corresponding to *p-value* ≤ 0.01, we identified ~40 Gb of candidate introgressed sequence. Note, S^* *p-values* are robust to demographic uncertainty (Fig. S7). The distribution of Neanderthal match *p-values* for this set of candidate introgressed sequences (Fig. 1D) demonstrates a strong skew towards zero, consistent with the hypothesis that they are strongly enriched for Neanderthal lineages. The distribution of Neanderthal match *p-values* for sequences that do not possess significant evidence of introgression, as revealed by S^*, is approximately uniform (Fig. 1D; *10*), indicating that our statistical approach is able to distinguish between introgressed and non-introgressed lineages (Fig. S8; *10*).

At FDR = 5%, we identified over 15 Gb of introgressed sequence across all individuals, which spans ~20% (600 Mb) of the Neanderthal genome (Fig. 1E; Table S9). Of the 600 Mb of unique sequence, ~25% (149 Mb) was shared between Europeans and East Asians. On average, we found 23 Mb of introgressed sequence per individual (Fig. 1F), with East Asian individuals inheriting 21% more Neanderthal sequence than Europeans. Within subpopulations, we found small but statistically significant variation in the amount of introgressed sequence among Europeans (Kruskal-Wallis rank sum test; *p-value* = 4.2×10^{-12}), but not among East Asians (*p-value* = 0.43).

The average length of introgressed haplotypes was ~57kb (Fig. 2A), and ~26% of all protein-coding genes had one or more exons that overlapped a Neanderthal sequence (Fig. 2B). At a broad scale, the genomic distribution of Neanderthal lineages exhibits marked

heterogeneity, with particular chromosomal arms, such as 8q and 17q, depleted of Neanderthal sequence (Fig. 2A). These qualitative patterns were confirmed by multiple logistic regression, which showed that chromosomal arm was a significant predictor (*p-value* < 10^{-16}) of the odds that a 50kb window possessed introgressed sequence (*10*; Fig. 2C; Fig. S9; Fig. S10). Furthermore, odds ratios were negatively correlated with fixed differences between modern humans and Neanderthal (Fig. 2D; Spearman's $\rho = -0.80$; *p-value* < 5.8×10^{-8}). A strong depletion of Neanderthal lineages spanning ~17 Mb on 7q encompasses the *FOXP2* locus (Fig. 2A), a transcription factor that plays an important role in human speech and language (*13*). The observed negative correlation between odds ratio and divergence remained significant when East Asians and Europeans were analyzed separately (Fig. S11) and when explicitly controlling for the presence of Neanderthal lineages in modern humans (*10*; Fig. S12 Fig. S13). These results suggest that sequence divergence between modern humans and Neanderthals was a barrier to gene flow in some regions of the genome and associated with deleterious fitness consequences (*14*).

We next leveraged the catalog of introgressed sequences in East Asians and Europeans to refine admixture models and infer parameters of gene flow between modern humans and Neanderthals (Fig. S14, Fig. S15). Specifically, using an ABC framework (*10*), we statistically tested a model with a single pulse of introgression into the common ancestor of Europeans and East Asians (*3*) as well as a second model with gene flow both in the common ancestor and a second, smaller pulse into East Asians shortly after the two populations split (Fig. 3A). Consistent with recent inferences *(5,9)*, observed patterns of introgression were incompatible with a one pulse model (Fig. 3B), suggesting that gene flow between Neanderthals and humans occurred multiple times. Although we varied many parameters of each model (*10*; Fig. S14),

only the ratio of ancestral effective population size between East Asians and Europeans ($N_e^{ASN/}N_e^{EUR}$) and the relative amount of introgression between the second and first pulse (m_2/m_1) had appreciable effects on model fit (Fig. 3B). We estimate that $N_e^{ASN/}N_e^{EUR}$ is 1.29 (95% CI of 1.15-1.57), and that East Asians received an additional 20.2% (95% CI of 13.4%-27.1%) more Neanderthal sequence in the second pulse (*10*). We note that additional unexplored models may provide a better fit to the data, and refining demographic models of hominin evolution is an important area of future work.

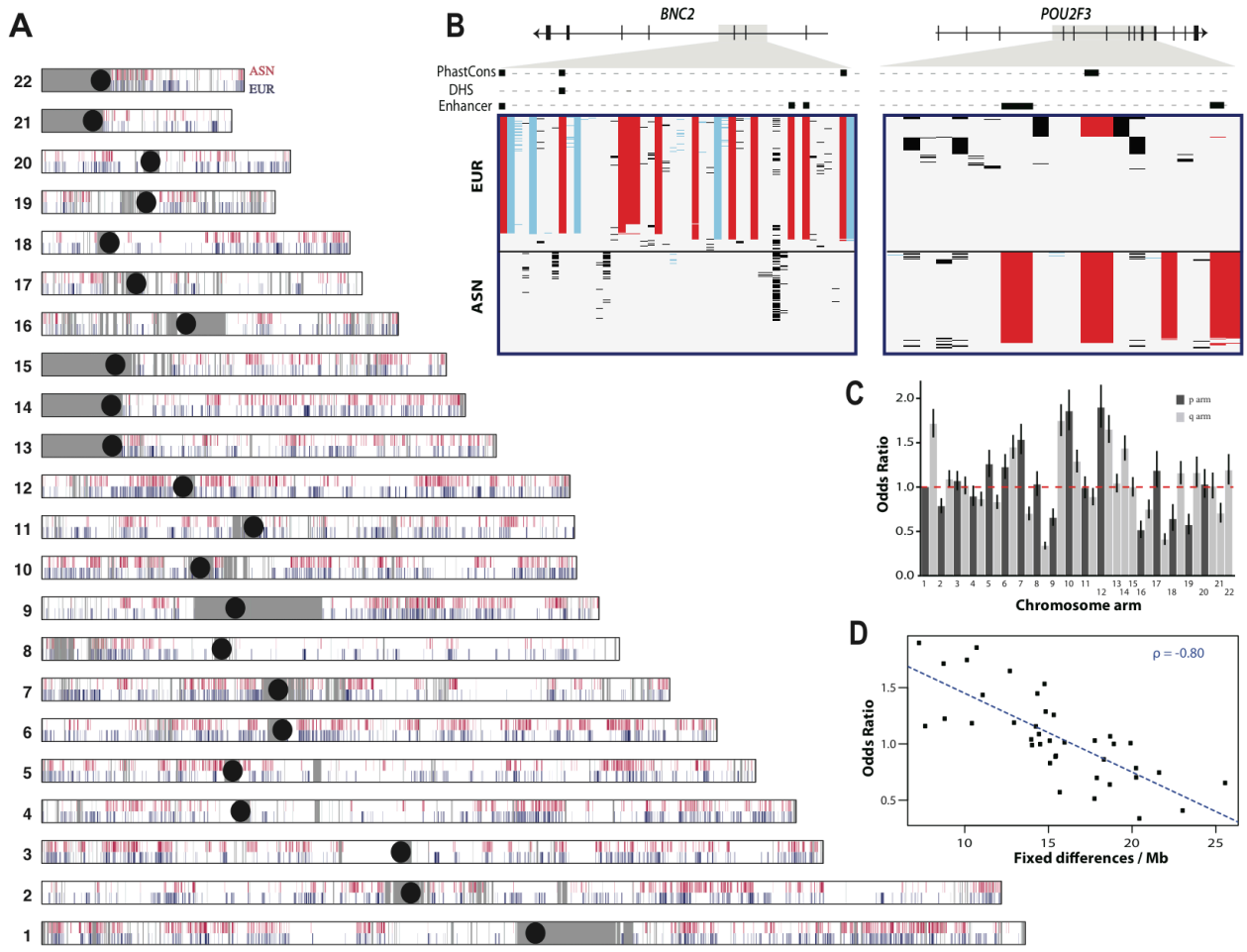

Fig. 2. Genomic distribution of surviving Neanderthal lineages.
(A) Neanderthal lineages identified in East Asians (ASN, red) and Europeans (EUR, blue). Gray shading denotes regions that did not pass filtering criteria (*10*). **(B)** Visual genotype illustrations of introgressed sequence identified in the *BNC2* and *POU2F3* genes. Rows are individuals, columns are variant sites, and rectangles are colored according to genotype (red = predicted Neanderthal variant that matches allele present in the Neanderthal reference genome; blue = predicted Neanderthal variant that does not match allele present in the Neanderthal reference genome; black = other variants). Introgressed variants that overlap a PhastCons conserved element, DNaseI hypersensitive site (DHS), or putative enhancer elements are shown as boxes (*10*). **(C)** Odds of finding an introgressed lineage on each chromosomal arm calculated from a logistic regression model (*10*). Odds ratios (OR) are expressed using chromosome 1p as the baseline level. Horizontal bars represent 95% confidence intervals. **(D)** Relationship between the OR and number of fixed differences per Mb between humans and Neanderthals.

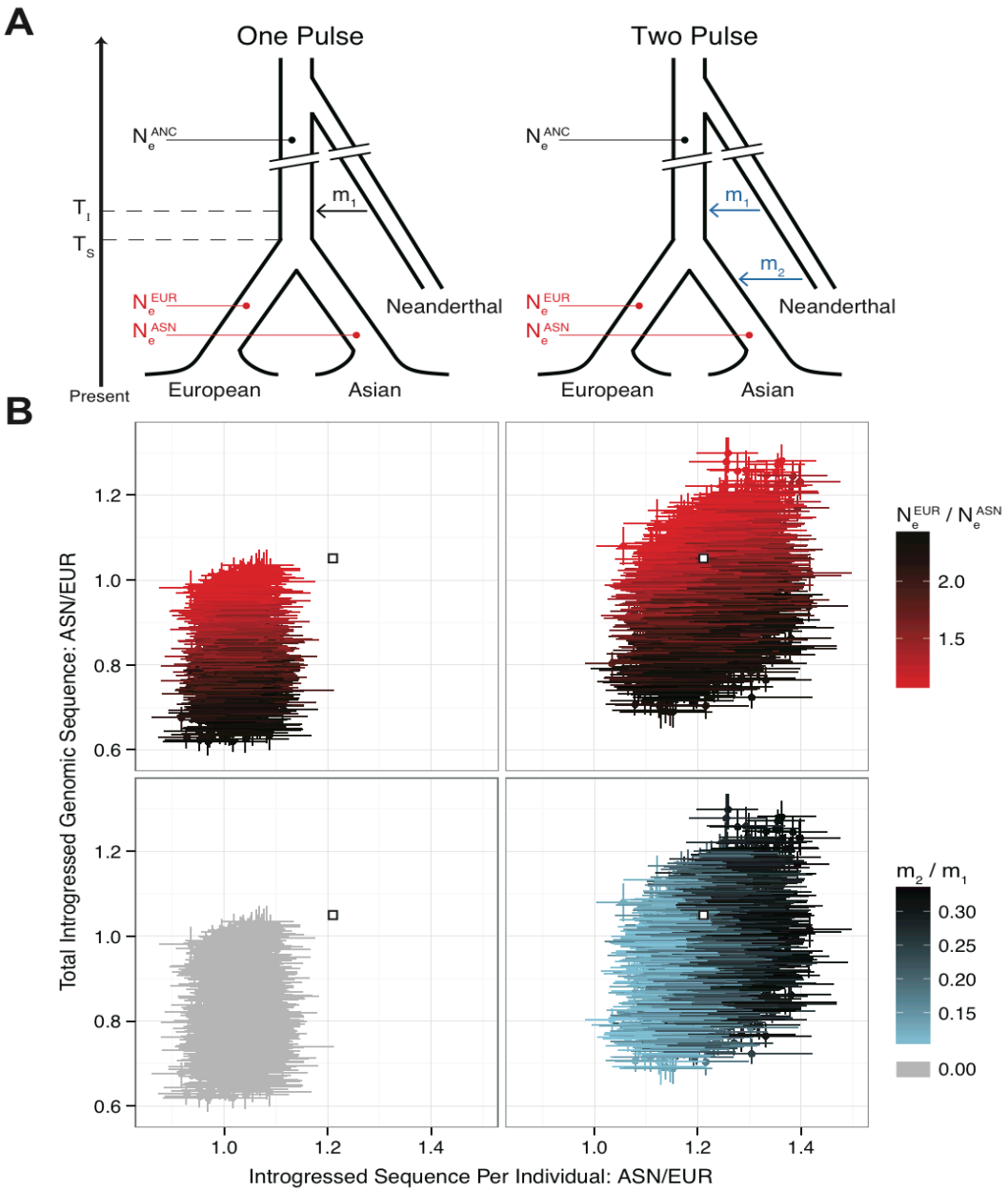

Fig. 3. Organization and characteristics of Neanderthal sequence in Europeans and East Asians suggests at least two admixture events.
(A) Schematic diagrams of the one and two pulse admixture models. N_e^{ASN} and N_e^{EUR} denote effective population sizes of the ancestral, East Asian, and European populations, respectively. In the one pulse model, gene flow *(m_1)* between Neanderthals and the ancestors of Europeans and East Asians occurs at time T_I. In the two pulse model, a second pulse of gene flow (m_2) occurs into East Asians shortly after the divergence of Europeans and East Asians at time TS. **(B)** Values of summary statistics calculated from 2000 simulations under each model (red, blue, and grey points; horizontal and vertical bars denote 95% CIs) show that a single pulse model is incompatible with the observed data (white box, corrected for sample size differences between populations; limits of box denote 95% CI). Simulations that varied $N_e^{ASN/}N_e^{EUR}$ are shown in red and those with variable m_2/m_1 are shown in blue (color bars indicate parameter values).

The collection of surviving Neanderthal lineages that we identified allows us to search for signatures of adaptive introgression (*15,16*). First, we used introgressed variants that exhibit large allele frequency differences between Europeans and East Asians ($F_{ST} > 0.40$; *p-value* < 0.001 by simulation; *10*), to identify four significantly differentiated regions (Fig. 4; Table S10; *10*). Introgressed haplotypes in two of these regions span genes that play important roles in the integumentary system: *BNC2* on chromosome 9 and *POU2F3* on chromosome 11. *BNC2* encodes a zinc finger protein expressed in keratinocytes and other tissues (*17*), and has been associated with skin pigmentation levels in Europeans (*18*). The adaptive haplotype has a frequency of ~70% in Europeans and is completely absent in East Asians (Fig. 2B). *POU2F3* is a homeobox transcription factor expressed in the epidermis and mediates keratinocyte proliferation and differentiation (*19,20*). The adaptive haplotype in East Asians has a frequency of ~66% and is found at less than 1% frequency in Europeans (Fig. 2B). No coding introgressed variants were found in *BNC2* or *POU2F3*, although several highly differentiated introgressed variants were located in functional non-coding elements (*21*; Fig. 2B), suggesting that modern humans acquired adaptive regulatory sequences at these loci. We also searched for shared signatures of adaptive introgression between East Asians and Europeans, identifying six distinct regions that have introgressed haplotype frequencies greater than 40% in both populations (Fig. 4; Table S11; *p-value* < 10^{-4} by simulation; *10*). One of these regions lies in the type II cluster of keratin genes on 12q13 (Table S11), further suggesting that Neanderthals provided modern humans with adaptive variation for skin phenotypes. In total, eight of the ten candidate introgressed regions overlap protein-coding genes (Fig. 4).

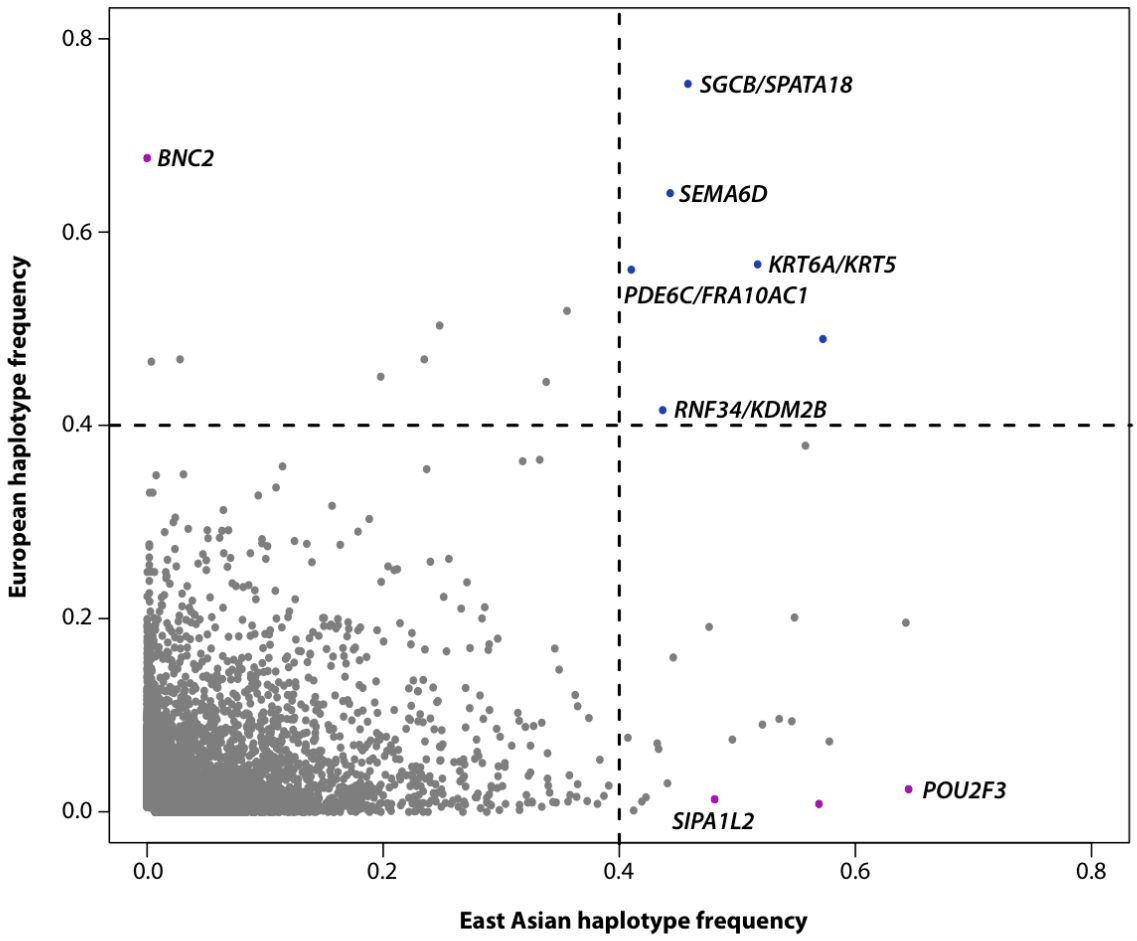

Fig. 4. Signatures of adaptive introgression. Scatter plot of introgressed haplotype frequency in Europeans and East Asians. Significantly differentiated and common shared haplotypes are shown in magenta and blue, respectively. Protein-coding genes that overlap candidate adaptively introgressed loci are shown.

This study shows that the fragmented remnants of the Neanderthal genome carried in the DNA of modern humans can be robustly identified, allowing in aggregate, substantial amounts of Neanderthal sequence to be recovered. In principle, our approach can be used in the absence of an archaic reference sequence, potentially allowing the discovery and characterization of previously unknown hominins that interbred with modern humans (*22-24*). This "fossil free" paradigm of sequencing archaic genomes holds considerable promise to reveal insights into hominin evolution, the population genetics characteristics of archaic hominins, how introgression has influenced extant patterns of human genomic diversity, and to narrow the search for genetic changes that endow uniquely human phenotypes.

References and Notes

1. A. D. Twyford, R. A. Ennos, Next-generation hybridization and introgression. *Heredity*. **108**, 179-189 (2012).
2. D. Zinner, M. L. Arnold, C. Roos . The strange blood: natural hybridization in primates. *Evo Anthropol.* **20**, 96-103 (2011).
3. R. E. Green, et al., A draft sequence of the Neandertal genome. *Science*. **328**, 710-722 (2010)
4. D. Reich D, et al., Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. **468**, 1053-1060 (2010).
5. M. Meyer, et al., A high-coverage genome sequence from an archaic Denisovan individual. *Science*. **12**, 222-226 (2012).
6. A. Eriksson, A. Manica, Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc Natl Acad Sci U S A*. **109**,13956-13960 (2012).
7. M. A. Yang, A. S. Malaspinas, E. Y. Durand, M. Slatkin, Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol Biol Evol.* **29**, 2987-2995 (2012). 8. S. Sankararaman, N. Patterson, H. Li, S Pääbo, D. Reich, The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012).
9. J. D. Wall, et al., Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics*. **194**, 199-209 (2013)
10. Supporting material is available on *Science* Online.
11. V. Plagnol, J. D. Wall, Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
12. 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature*. **491**, 56-65 (2012).
13. W. Enard, et al., Molecular evolution of FOXP2, a gene involved in speech and language. *Nature.* **22**, 869-872 (2002).
14. M. Currat, L. Excoffier, Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proc Natl Acad Sci U S A*. **108**, 15129-15134 (2011).
15. F. L. Mendez, J. C. Watkins, M. F. Hammer, 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 (2012).
16. L. Abi-Rached, et al., The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science.* **334**, 89-94 (2011).
17. A. Vanhoutteghem, P. Djian P, Basonuclins 1 and 2, whose genes share a common origin, are proteins with widely different properties and functions. *Proc Natl Acad Sci U S A*. **103**, 12423-12428 (2006).

18. L. C. Jacobs, et al., Comprehensive candidate gene study highlights *UGT1A* and *BNC2* as new genes determining continuous skin color variation in Europeans. *Hum Genet.* **132**, 147-158. (2013).
19. A. Cabral, D. F. Fischer, W. P. Vermeij, C. Backendorf, Distinct functional interactions of human Skn-1 isoforms with Ese-1 during keratinocyte terminal differentiation. *J Biol Chem.* **278**, 17792-17799 (2003).
20. H. Takemoto, et al., Relation between the expression levels of the POU transcription factors Skn-1a and Skn-1n and keratinocyte differentiation. *J Dermatol Sci.* **60**, 203-205 (2010).
21. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Natur*e. **489**, 57-74 (2012).
22. J. D. Wall, K. E. Lohmueller KE, V. Plagnol, Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol.* **26**:1823-1827 (2009).
23. M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, J. D. Wall, Genetic evidence for archaic admixture in Africa. *Proc Natl Acad Sci U S A.* **108**,15123-15128 (2011).
24. J. Lachance, et al., Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*. **150**:457-469 (2012)
25. W. Fu, et al., Analysis of 6,515 exomes reveals the recent origin of most human proteincoding variants. *Nature*. **493**, 216-220 (2013).
26. S. F. Schaffner, et al., Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576-1583 (2005).
27. J. A. Tennessen et al., Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*. **337**, 64-69 (2012).
28. S. Gravel et al., Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* **108**, 11983-11988 (2011).
29. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. **18**, 337-338 (2002).
30. S.N. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society.* **73**, 3-36 (2011).
31. M. B. Mednikova. A proximal pedal phalanx of a Paleolithic hominin from Denisova Cave, Altai. *Archaeology, Ethnology and Anthropology of EurEast Asia*. **39**, 129-138 (2011).
32. H. Li, et al., The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. **25**, 2078-2079 (2009).
33. J. A. Bailey, et al., Recent segmental duplications in the human genome. *Science.* **297**, 1003- 1007 (2002).
34. H. Li and R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature*. **475**, 493–496 (2011).
35. G. McVicker, et al., Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLoS Genet.* **5**, e1000471 (2009).
36. The International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
37. K. Csilléry, et al., abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* **3**, 475-479 (2012).
38. S. Neph, et al., BEDOPS: high-performance genomic feature operations. *Bioinformatics*. **28**, 1919-1920 (2012).
39. P. Aboyoun, et al., GenomicRanges: Representation and manipulation of genomic intervals. R package version 1.10.7. http://www.bioconductor.org/packages/2.13/bioc/html/GenomicRanges.html
40. B. Paten, et al., Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*. **18**, 1814-1828 (2008)
41. L. R. Meyer, et al., The UCSC Genome Browser database: extensions and updates 2013. *Nucl. Acids Res*. **41**, D64-D69 (2013)
42. K. D. Pruitt, et al., NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501-D504 (2005).
43. A. Karatzoglou, et al., kernlab - An S4 Package for Kernel Methods in R. *J. of Statistical Software* **11**, 1-20 (2004).
44. A. W. Briggs, et al. Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* 325, 318-321 (2009)
45. S. Wang, J. Lachance, S. Tishkoff, J. Hey, and J. Xing. Apparent Variation in Neanderthal Admixture among African Populations Is Consistent with Gene Flow from Non-African Populations. *Genome Biology and Evolution*, doi:10.1093/gbe/evt160 (2013)

46. **Acknowledgments:** We thank members of the Akey laboratory, S. Browning, and B. Browning, and Joan Duffy for critical feedback related to this work, S. Pääbo for providing access to high-coverage Neanderthal sequence data, and L. Jáuregui for help in figure preparation. A description of where sequence data used in our analyses can be found is provided in the SOM. JMA is a paid consultant of Glenview Capital.

Supplementary Materials (Appendix A):
Materials and Methods
Figures S1-S15
Tables S1-S11

Chapter 4 : Complex history of admixture between modern humans and Neanderthals

This chapter was published in The American Journal of Human Genetics on March 5, 2015.

Vernot, B., and Akey, J.M. (2015). Complex History of Admixture between Modern Humans and Neandertals. The American Journal of Human Genetics 96, 448–453.

Abstract:

Recent analyses have found that a substantial amount of the Neanderthal genome persists in the genomes of contemporary non-African individuals. East Asians have, on average, higher levels of Neanderthal ancestry compared to Europeans, which may be due to differences in the efficiency of purifying selection, an additional pulse of introgression into East Asians, or other unexplored scenarios. To better define the scope of plausible models of archaic admixture between Neanderthals and anatomically modern humans, we analyzed patterns of introgressed sequence in whole-genome data of 379 Europeans and 286 East Asians. We find that inferences of demographic history restricted to neutrally evolving genomic regions allow a simple one pulse model to be robustly rejected, suggesting that differences in selection cannot explain the differences in Neanderthal ancestry. We show that two additional demographic models, involving either a second pulse of Neanderthal gene flow into the ancestors of East Asians or a dilution of Neanderthal lineages in Europeans by admixture with an unknown ancestral population, are consistent with the data. Thus, the history of admixture between modern humans and Neanderthals is likely more complex than previously thought.

Main Text:

As modern humans migrated out of Africa and dispersed throughout the world, they encountered and hybridized with Neanderthals[1,2]. The similarly low levels of Neanderthal ancestry found in all modern non-African populations studied to date has been parsimoniously interpreted to be the result of a single pulse of admixture into the population ancestral to all non-Africans. However, recent reports show that East Asians have, on average, inherited ~20% more Neanderthal ancestry than Europeans[3,4,5,6]. Two explanations have been proposed to account for this observation. Sankararaman *et al*.[5] suggested that because Neanderthal lineages appear to be subject to widespread purifying selection in modern humans, differences in the efficiency of purifying selection could account for higher levels of Neanderthal ancestry in East Asians. This hypothesis is supported by previous studies that have shown East Asians have a smaller effective population size compared to Europeans[7]. In contrast, through extensive simulations, Vernot and Akey[4] found that the excess of Neanderthal ancestry in East Asians cannot be explained by a single ancestral introgression event. Rather, the data was better explained by a two pulse model, where introgression occurred in the common ancestor of East Asians and Europeans, followed by additional gene flow into East Asians. However, these simulations did not account for the potential confounding effects of natural selection, and thus ambiguity remains about whether a simple single pulse of admixture between modern humans and Neanderthals can explain the data.

To investigate how patterns of introgressed Neanderthal sequences in East Asians and Europeans are influenced by potential differences in the efficiency of purifying selection between populations, we first partitioned the genome using B-values[8], which measure the degree to which neutral variation has been reduced due to linked selected sites. Specifically, we selected 50kb windows from Vernot and Akey[4] and binned each window according to the minimum B-value at any base in that window. B-values range from zero (all neutral diversity has been

eliminated by background selection) to one (no reduction of neutral diversity by background selection). Next, we calculated values of two summary statistics of Neanderthal introgression[4], R_{ind} and R_{pop} (Figure 1), as a function of B-values. R_{ind} is the ratio of the amount of introgressed Neanderthal sequence per individual in East Asians versus Europeans (Figure 1). R_{pop} is the ratio of genomic bases covered by introgressed Neanderthal sequence in any East Asian versus European individual, corrected for differences in sample size by subsampling sets of 20 individuals from each population (Figure 1). We have previously measured genome-wide values of R_{ind} and R_{pop} of 1.21 and 1.05, respectively[4]. Moreover, R_{ind} has consistently been reported to be greater than one, including values of 1.19[3], 1.20[5], and 1.4[6]. These estimates indicate that although approximately the same amount of the Neanderthal genome survives in Europeans and East Asians, a given Neanderthal haplotype in East Asians is on average at higher frequency (Figure 1).

If elevated levels of Neanderthal ancestry in East Asians are due to differences in the efficiency of purifying selection between East Asians and Europeans, then R_{ind} should vary significantly by B-value. Genomic regions under strong purifying selection would be more strongly depleted in Neanderthal ancestry in the historically larger European population, leading to high R_{ind} at low B-values, and R_{ind} closer to 1 at high B-values. In contrast, we observe that R_{ind} is fairly stable with increasing B-value cutoffs (Figure 2). For example, the estimate of R_{ind} in regions with a minimum B-value of 0.975 (spanning ~106Mb of the genome), which corresponds to a reduction in neutral variation due to background selection of less than 2.5%, is 1.175. Thus, in this more neutral subset of the genome, East Asian individuals have on average 17.5% more introgressed sequence compared to Europeans, closely paralleling genome-wide estimates.

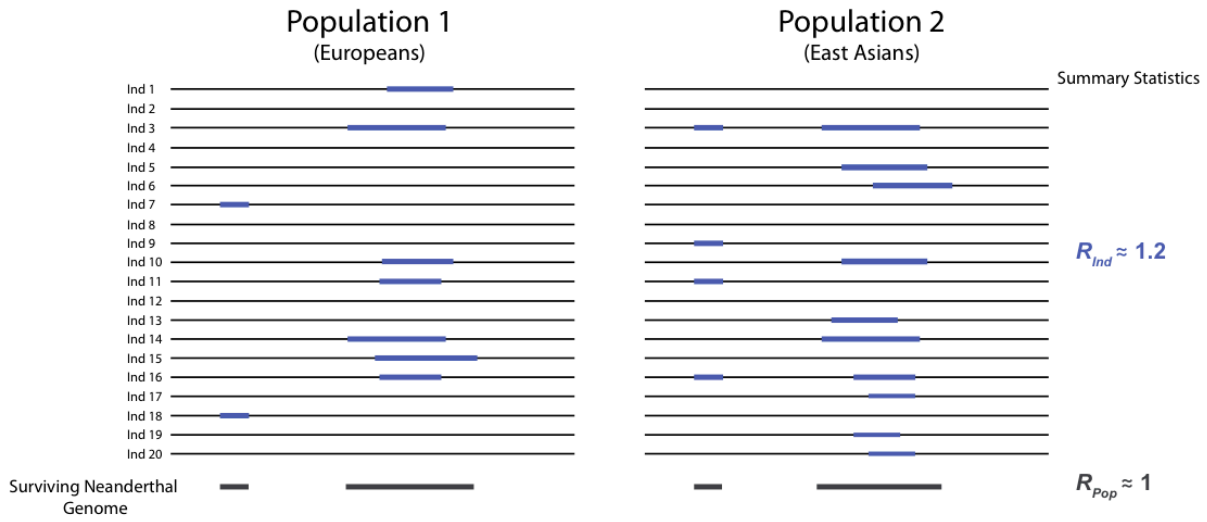

Figure 1. Schematic representation of R_{ind} and R_{pop}.
Two hypothetical populations are shown, with twenty individuals sampled from each population. Each individual's Neanderthal ancestry is shown in blue (top), while the total amount of the Neanderthal genome present in this region in each population is shown in dark gray (bottom). Summary statistic estimates for these hypothetical regions are given (left). Note that the amount of Neanderthal sequence per individual is quite different between populations, with Population 2 containing more Neanderthal sequence per individual, as reflected by an R_{ind} value of ~1.2. However, in each population the same amount of the Neanderthal genome survives, as reflected by an R_{pop} value of ~1.

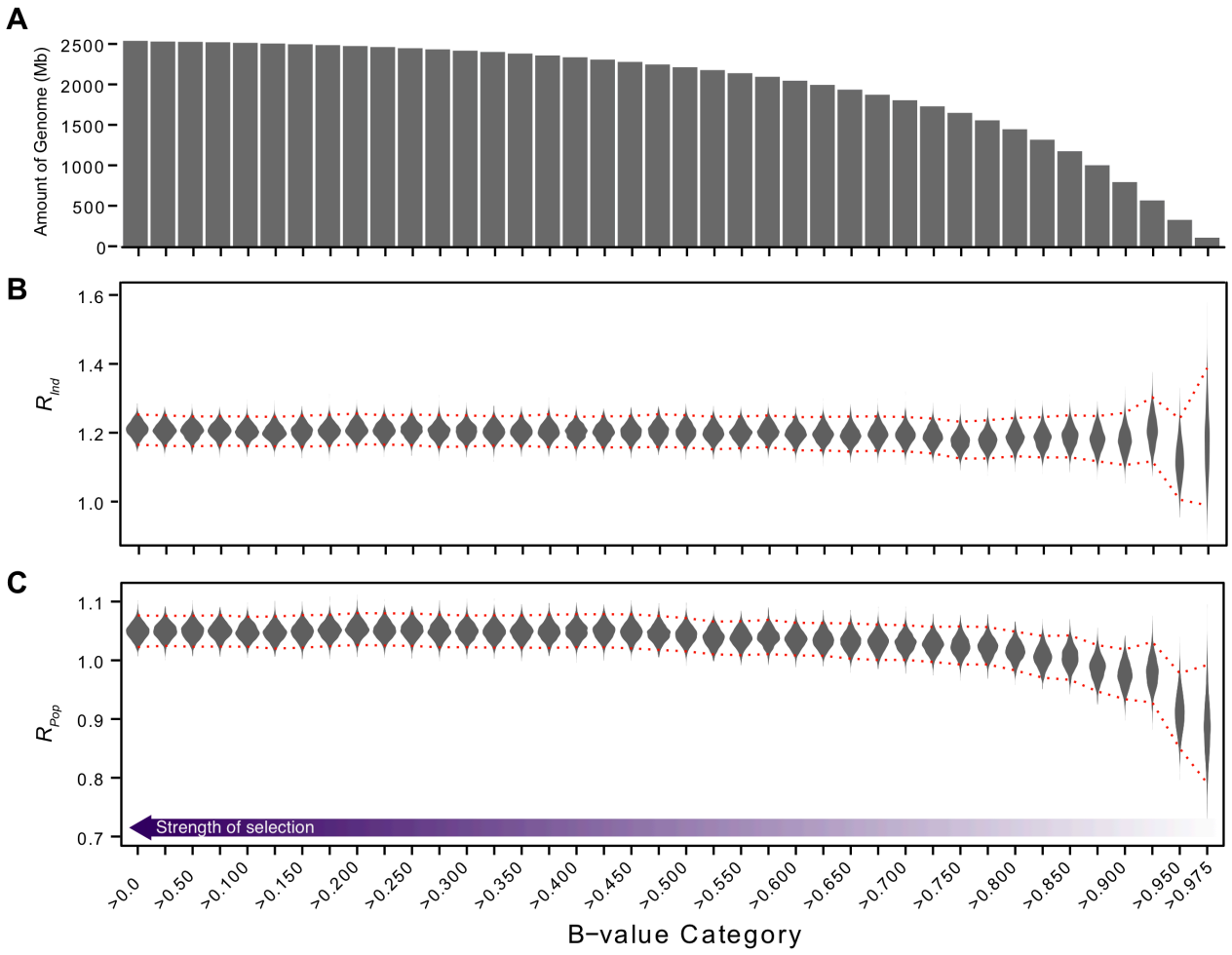

Figure 2. Estimates of Rind and Rpop as a function of B-value cutoffs.
(A) Amount of the genome in each B-value category. Note that each category is a subset of the category to its left (i.e., the category $B \geq 0$ contains the entire genome). Panels **(B)** and **(C)** show Rind and Rpop, respectively, calculated over regions of the genome with progressively higher B-value thresholds (i.e., less functional constraint; see color bar). Violin plots show bootstrap resamples, and red dotted lines denote the 95% CI.

In contrast to stable levels of R_{ind} with increasing B-values, R_{pop} shows a marked decline as the B-value cutoff exceeds 0.850, indicating that in more neutrally evolving genomic regions, less of the Neanderthal genome survives at the population level in East Asians compared to Europeans. This observation is consistent with previous studies, which have found that a smaller ancestral effective population size in East Asians resulted in more intense genetic drift compared to Europeans[7]. Higher genetic drift would result in the loss of low frequency Neanderthal haplotypes; this effect will be strongest in regions where the competing force of purifying selection is weakest[9]. Qualitatively, patterns of R_{ind} and R_{pop} as a function of B-value suggest that the excess of Neanderthal ancestry in East Asians cannot be explained by differences in selective forces alone, and that a model of a single ancestral pulse of Neanderthal introgression is unlikely.

To more formally evaluate demographic models compatible with patterns of Neanderthal ancestry in East Asians and Europeans, we performed Approximate Bayesian Computation (ABC)[10,11] on putatively neutral sequence, by calculating R_{ind} and R_{pop} in genomic regions with a minimum B-value ≥ 0.975. ABC analysis involves simulating data under different demographic models, calculating summary statistics from these simulations, and selecting simulations in a principled manner that best match the observed summary statistics. We first simulated neutral sequence data under: 1) a one pulse model where all introgression from Neanderthals occurs in a single pulse into the common ancestor of East Asians and Europeans (m_1; Figure 3a), and 2) a two pulse model with varying amounts of additional introgression into either Europeans or East Asians (m_2; Figure 3a) after population splitting. Specifically, we performed simulations with m_2/m_1 ranging from -2% to 33% (negative values indicate simulations with additional introgression into Europeans; $m_2 = 0$ is a one pulse model). For summary statistics we used both R_{ind}, and R_{pop}. We have previously shown that these statistics can distinguish between archaic

admixture demographic models[4]. We performed >30,000 simulations and estimated demographic parameters using ABC with the R package 'abc'[12]. Specifically, we used the 'abc' function with a non-linear neural network regression method of correcting accepted parameter values[10]. Results were similar when using local linear regression correction. A complete description of the simulated demographic models can be found in the Appendix at the end of this chapter.

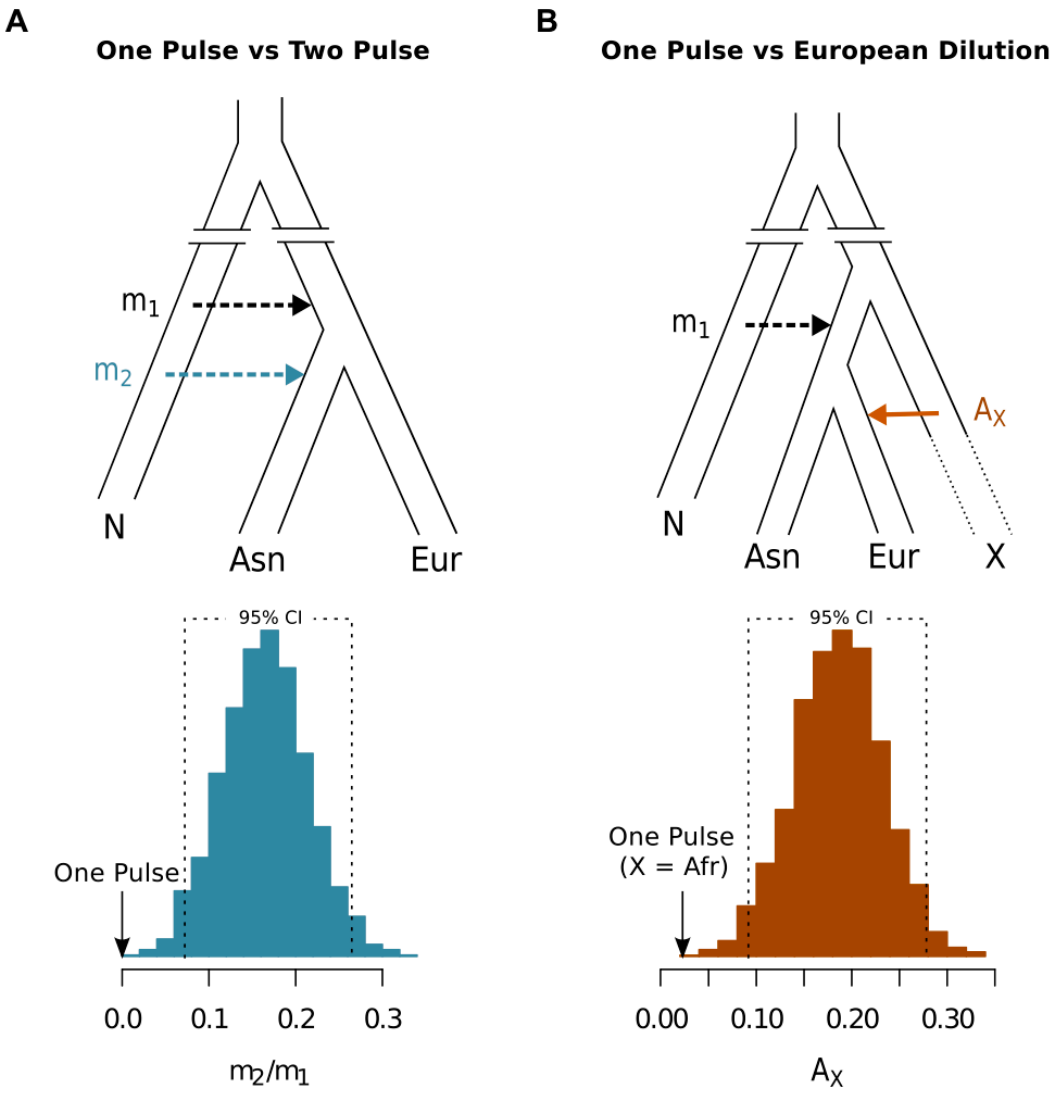

Figure 3. Inference of admixture models from genomic regions with little or no selective constraint.

(A) Schematic illustration of the one and two pulse model of Neanderthal introgression, with the majority of introgression occurring in the common ancestor of all non-Africans (m1), and a smaller additional amount of introgression into East Asians (m2). **(B)** Schematic illustration of the one pulse and European-dilution model, with all Neanderthal introgression occurring in the common ancestor of Europeans and East Asians (m1), and a substantial portion of modern day European ancestry (AX) deriving from a second population (X) that has no Neanderthal ancestry. Below each demographic model are histograms of m2/m1 (left; the proportion of additional introgression into East Asians) and AX (right; the proportion of European ancestry derived from population X) as estimated by 30,000 simulations and ABC analysis (adjusted values of 1500 accepted simulations are shown). Dashed lines demark 95% confidence intervals. Black arrows show the expected values of m2/m1 (0) and AX (0.023)14 under the one-pulse model.

Only simulations with additional introgression into East Asians (two pulse models) were accepted as plausible in the ABC analysis (Figure 3a). We estimate that a second pulse of 15% more introgression into East Asians could explain the observed excess of Neanderthal introgression (95% CI of m_2/m_1 = 6.8% - 26.6%). Under the null hypothesis of a single pulse of admixture, the ratio m_2/m_1 = 0%, which is well outside the preferred range. Given these results, the two pulse model is significantly favored, and we can strongly reject the null hypothesis of a one pulse model ($p < 6.7 \times 10^{-4}$). For completeness, we repeated our ABC analysis using summary statistics from regions with a minimum B-value ≥ 0.950 (spanning 326Mb of the genome), which again significantly favored the two pulse model (Figure S1). Therefore, a one pulse model is rejected both for genome-wide calculations of Neanderthal ancestry[4], and also when regions that may have been subjected to selective constraint are excluded.

It is important to stress that although a two pulse model of admixture explains the empirical data significantly better than the simple one pulse model considered here, it does not necessarily mean the two pulse model is correct. To investigate additional plausible demographic models, we also considered a single ancestral pulse of introgression, but where the European population then later admixes with a third modern human population *that has not interbred with Neanderthals*. In this scenario, the amount of Neanderthal ancestry in Europeans is effectively diluted by admixture with a population not carrying Neanderthal lineages. Such a population could be from Africa, where there is expected to have been no, or little, Neanderthal ancestry[13]. Alternatively, it could be an unknown "ghost" Eurasian population that was entirely absorbed into Europeans.

To determine if a European dilution model could explain the data, we again performed ABC using the summary statistics R_{ind}, and R_{pop} calculated from neutral subsets of the genome

(B-value ≥ 0.975) as described above. Specifically, we simulated sequence data under a one pulse model where all introgression from Neanderthals occurs in a single pulse into the common ancestor of East Asians and Europeans (m_1; Figure 3b), and the ancestral European population then admixes with a third population denoted as X (Figure 3b). We varied the proportion of modern day European ancestry derived from population X (A_X) from 0-35%. We estimate that the observed patterns of Neanderthal introgression are compatible with a European dilution model if on average 18.2% of modern European ancestry (A_X) was contributed by this third population (95% CI of A_X = 9.2%-27.6%). This is significantly larger than current estimates of African ancestry in Europeans (1-3%[14], 2.3%[15]; $p = 5.7 \times 10^{-4}$), suggesting that migration from Africa to Europe cannot explain the excess of Neanderthal ancestry in East Asians compared to Europeans. Tantalizingly, recent work suggests that modern Europeans may be comprised by admixture of three ancestral groups[16]. However, each of these groups is estimated to contain ~2% Neanderthal ancestry[16], and thus could not have diluted the amount of Neanderthal ancestry in modern Europeans enough to account for the differences with East Asians. Thus, based on current evidence, differential migration seems less likely to explain the data, increasing the likelihood of multiple pulse models.

In addition to evaluating models of the interactions between modern humans and Neanderthals, we used these analyses of Neanderthal ancestry to elucidate other aspects of human demographic history. For example, in addition to estimating the parameters m_2/m_1 and A_X, we found that the ratio of ancestral effective population sizes in Europeans and East Asians, N_e^{EUR}/N_e^{ASN}, had a significant effect on the fit of our models. Using the same ABC analyses, we estimate N_e^{EUR}/N_e^{ASN} = 1.93 (95% CI 1.57-2.73) under a two pulse model, and estimate

N_e^{EUR}/N_e^{ASN} = 1.59 (95% CI 1.35-1.89) under the European dilution model (Figure S1). Both of these estimates are consistent with previously accepted values[7,15].

In summary, by focusing on putatively neutral regions of the genome, we have shown that the observed patterns of Neanderthal ancestry in Europeans and East Asians are not consistent with a simple one pulse model of admixture. Thus, differences in the efficiency of purifying selection among populations are unlikely to account for higher levels of Neanderthal ancestry in East Asians compared to Europeans. We have shown that more complex and nuanced models are necessary to explain the data, and furthermore suggested two such models that are consistent with observed patterns of Neanderthal introgression in Europeans and East Asians. Additionally, we have shown that studies of Neanderthal ancestry can be informative about other aspects of human history. In the future, additional studies in geographically diverse populations, combined with ancient DNA analysis of archaic and modern humans, will help narrow the space of plausible demographic models. Such models will provide critical insights into hominin evolutionary history, and the key parameters governing admixture dynamics between modern humans and Neanderthals.

Demographic Models and Simulations (Appendix)

We performed coalescent simulations based on previously inferred demographic models for European, East Asian, and African populations[15,18]. Simulations were performed with ms[17], and coalescent trees were extracted from the output and used to calculate the summary statistics R_{ind} and R_{pop}. It is important to note that sequence variation does not affect these summary statistics, and thus many aspects of these demographic models are essentially "nuisance

parameters" – i.e., they do not have an effect on the final results. The base demographic model is:

a) splitting between modern humans and Neanderthals at 700kya;
b) Neanderthal N_e of 1,500;
c) splitting between Africans and non-Africans at 70kya;
d) African N_e of 14,474 until 5,115 years ago;
e) gradual growth of non-African populations starting at 23kya, ending at 5,115 years ago, to N_e of 8,879 in East Asians and N_e of 9,475 in Europeans, excepting simulations where no growth is required to reach an N_e greater than or equal to 8,879 or 9,475, respectively;
f) rapid growth of all populations starting at 5,115 years ago, to a present day N_e of 424,000 in Africans, 512,000 in Europeans, and 1,370,990 in East Asians;
g) a single 500 year introgression event from Neanderthals to the common ancestor of Europeans and East Asians, at a rate of 0.00075 (0.075% of each generation was sampled from Neanderthal individuals).

Note, although a split time of 700kya between Neanderthals and modern humans is older than the upper bound of 500kya estimated using the draft Neanderthal genome[1], it is within the range estimated from the high-coverage Altai Neanderthal genome[2]. Furthermore, this time was chosen to ensure that any introgressed lineages coalesced before non-introgressed lineages, making it easier to identify introgressed lineages by examining the coalescent trees. It is important to note that this older split time has no effect on the amount of introgressed sequence, as we are considering only the presence and extent of introgressed haplotypes, not variation that has arisen on those haplotypes.

To generate demographic models from this base model, we then sampled the following parameters from uniform distributions (unless otherwise noted):

a) European and East Asian split times between 36kya and 55kya [T(S)],

b) Ancestral Eurasian N_e prior to T(S) between 5,000 and 15,000,

c) European N_e / East Asian N_e between 1 and 2.5 [ne_ratio],

d) European + East Asian N_e after T(S) between 8,000 and 25,000 [ne_sum],

e) N_e for Europeans = ne_ratio * ne_sum

f) N_e for East Asians = ne_sum / (ne_ratio + 1)

g) The time of the first introgression pulse between T(S) and 65kya,

h) Migration rates were set as follows: 1.5×10^{-4} between Africans and the ancestors of Europeans and East Asians, $2.5 \times 10^{-5} \times 23000 / T(S)$ between Africans and Europeans, $7.8 \times 10^{-6} \times 23000 / T(S)$ between Africans and East Asians, and $3.11 \times 10^{-5} \times 23000 / T(S)$ between Europeans and East Asians. The above parameters were adapted from[15, 18], but adjusted for varying divergence times between Europeans and East Asians, and are the rate of migration per generation per chromosome (i.e., the proportion of population A originating in population B per generation).

i) In two-pulse models: a second 500 year introgression event from Neanderthals to East Asians, starting 500 years after T(S), at rate a between 0 and 0.00025. Some simulations instead included additional introgression into Europeans, at a rate between 0 and 0.000015. In our ABC analysis, we frame this as "negative" introgression into East Asians, effectively varying the amount of introgression in this second pulse from -0.000015 to 0.00025. This is equivalent to -2% to 33% more introgression into East Asians.

j) In European-dilution models, Africans were treated as population X, and the rate of migration from population X (m_3) was varied from 1.25×10^{-5} to 1.75×10^{-4} – i.e., we vary this rate from ½ to 7 times the rate estimated in Gravel *et al.*[15]. By multiplying the migration rate by the duration of the migration (T(S)), we can approximate the amount of European ancestry derived from population X ($A_X = m_3 * T(S)$).

An example *ms* command for the two-pulse model is:

```
ms 107 7893 -s 1 -r 14.6197076 50000 -I 4 26 40 40 1 0 -n 4 2.051984e-01 -n 1 58.002735978 -n 2
    70.041039672 -n 3 187.55 -eg 0 1 482.46 -eg 0 2 515.550576 -eg 0 3 720.230000 -em 0 1 2
    0.424271 -em 0 2 1 0.424271 -em 0 1 3 0.132372 -em 0 3 1 0.132372 -em 0 2 3 0.527793 -em 0 3
    2 0.527793 -eg 0.006997264 1 0 -eg 0.006997264 2 0.000000 -eg 0.006997264 3 16.996355 -en
    0.006997264 1 1.98002736 -en 0.031463748 2 1.899590e+00 -en 0.031463748 3 8.013680e-01 -ej
    5.421067e-02 3 2 -en 5.421067e-02 2 1.087825e+00 -em 5.421067e-02 1 2 4.386 -em 5.421067e-02
    2 1 4.386 -ej 9.575923e-02 2 1 -en 9.575923e-02 1 1.980027e+00 -en 0.20246238 1 1 -em
    6.472230e-02 2 4 2.193000e+01 -em 6.540629e-02 2 4 0 -em 5.284268e-02 3 4 3.590305e+00 -em
    5.352668e-02 3 4 0 -ej 9.575923e-01 4 1 -L -T
```

An example *ms* command for the European-dilution model is:

```
ms 107 20000 -s 150 -r 14.6197076 50000 -I 4 26 40 40 1 0 -n 4 2.051984e-01 -n 1 58.002735978 -n
    2 70.041039672 -n 3 187.55 -eg 0 1 482.46 -eg 0 2 502.019361 -eg 0 3 720.230000 -em 0 1 2
    0.7310 -em 0 2 1 0.518573 -em 0 1 3 0.228072 -em 0 3 1 0.228072 -em 0 2 3 0.909364 -em 0 3 2
    0.909364 -eg 0.006997264 1 0 -eg 0.006997264 2 0.000000 -eg 0.006997264 3 5.890454 -en
    0.006997264 1 1.98002736 -en 0.031463748 2 2.088235e+00 -en 0.031463748 3 1.051573e+00 -ej
    5.492750e-02 3 2 -en 5.492750e-02 2 1.614090e+00 -em 5.492750e-02 1 2 4.386 -em 5.492750e-02
    2 1 4.386 -ej 9.575923e-02 2 1 -en 9.575923e-02 1 1.980027e+00 -en 0.20246238 1 1 -em
    6.207387e-02 2 4 2.193000e+01 -em 6.275787e-02 2 4 0 -ej 9.575923e-01 4 1 -L -T
```

References

1. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722.
2. 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.
3. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., Filippo, C. de, et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science *338*, 222–226.
4. Vernot, B., and Akey, J.M. (2014). Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science *343*, 1017–1021.
5. 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.
6. 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.
7. Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet *39*, 1251–1255.
8. McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread Genomic Signatures of Natural Selection in Hominid Evolution. PLoS Genetics *5*, e1000471.
9. Kim, B., and Lohmueller, K. (2015). AJHG, THIS ISSUE
10. Blum, M.G.B., and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. Stat Comput *20*, 63–73.
11. Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. Trends in Ecology & Evolution *25*, 410–418.
12. Csilléry, K., François, O., and Blum, M.G.B. (2012). abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution *3*, 475–479.
13. 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.
14. Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLoS Genet *7*, e1001373.
15. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D., Altshuler, D.L., et al. (2011). Demographic history and rare allele sharing among human populations. PNAS 201019276.
16. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Sudmant, P.H., Schraiber, J.G., Castellano, S., Kirsanow, K., Economou, C., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.
17. Hudson, R.R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.
18. 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. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science *337*, 64–69.

19. Acknowledgments

20. We thank Kirk Lohmueller and members of the Akey laboratory for helpful discussions related to this manuscript. This work was supported by NIH grant 1R01GM110068 to JMA.

Chapter 5 : Neanderthal and Denisovan introgression in 35 Melanesian individuals

This chapter is preliminary work for a manuscript in preparation.

Vernot, B., Tucci, S., Wolf, A., Kelso, J., Prüfer, K., Pääbo, S., and Akey, J.

Although extensive analyses have been performed on Neanderthal introgression into various non-African populations [Sankararaman et al, 2012; Vernot and Akey, 2014; Sankararaman et al, 2014; Vernot and Akey, 2015; Kim and Lohmueller, 2015], to date no genome-wide investigation of Denisovan introgression has been performed. This is despite the publication of a high-coverage Denisovan genome two years before the publication of the high-coverage Neanderthal genome [Meyer et al, 2012; Prüfer et al, 2014], and is likely due to the availability of a large number of European and East Asian genomes [1000 Genomes Consortium, 2012], and a lack of genomes from populations known to harbor Denisovan ancestry. Maps of Neanderthal introgression have provided evidence of adaptive introgression of some Neanderthal haplotypes [Vernot and Akey, 2014; Sankararaman, 2014; others], of purifying selection against others [Vernot and Akey, 2014; Sankararaman, 2014], and have helped to refine demographic models of the introgression events [Vernot and Akey, 2015]. A map of Denisovan introgression into these populations would be of great interest, as it represents the only other known introgression event from an archaic hominin to modern humans, and as such could allow further insight into the characteristics of such events.

To this end, we sequenced 35 Melanesian individuals from Papua New Guinea (Figure 1), and performed a genome-wide scan for both Neanderthal and Denisovan admixture in these individuals. In this chapter, I focus on the application of these introgression maps to the question

of purifying selection against archaic sequence, exemplified by "deserts" of archaic admixture. These deserts were identified in a previously published map of Neanderthal introgression, and consist of multiple-megabase regions of the genome with little to no Neanderthal introgression in 665 European and East Asian individuals [Vernot and Akey, 2014]. The largest such desert, on chromosome 7q, is 15Mb long (Figure 2a).

However, the exact characteristics of the introgression event and demographic history of Europeans and East Asians are not known [Vernot and Akey, 2015], and so it is important to evaluate alternative demographic models. In particular, we consider models where the effective population size (Ne) of the modern human population into which introgression occurred was smaller than what is currently accepted. If a relatively small number of individuals carried Neanderthal sequence, and all of the individuals with Neanderthal sequence in a particular region of the genome did not reproduce, then Neanderthal sequence would be lost in this region. In other words, in a small population, genetic drift is strong [Kimura and Ohta, 1969]. We modified our demographic models to shrink Ne of the ancestors of Europeans and East Asians at the time of introgression, and varied the degree of the reduction in Ne, and the amount of time spent at low Ne. Although some of these models produced large Neanderthal deserts, in those models the number of smaller deserts was dramatically reduced, beyond what is observed in the real Neanderthal desert distribution (Figure 2b). In general, none of the tested models reproduced the observed desert distribution, although many other models are possible, and should be investigated in future work.

An important prediction of neutral models for Neanderthal deserts is that they exist due to random loss of local Neanderthal ancestry. If introgression occurred independently in multiple populations, then each population would represent a new random selection of archaic ancestry.

Although it may be difficult to determine if multiple Neanderthal introgression events occurred, it is safe to say that introgression from Denisovans occurred as an independent event from Neanderthals. Therefore, maps of Denisovan ancestry in Melanesian populations can be used to test the hypothesis that Neanderthal deserts are the product of random processes.

Figure 1. Sample collection locations

Samples were collected from four locations in Papua New Guinea, by Jonathan Friedlaender. 31 from West New Britain, 2 from New Ireland, 1 from New Hannover, 1 from Mussau Island. Historical names are given – e.g., New Ireland is now Latangai Island.

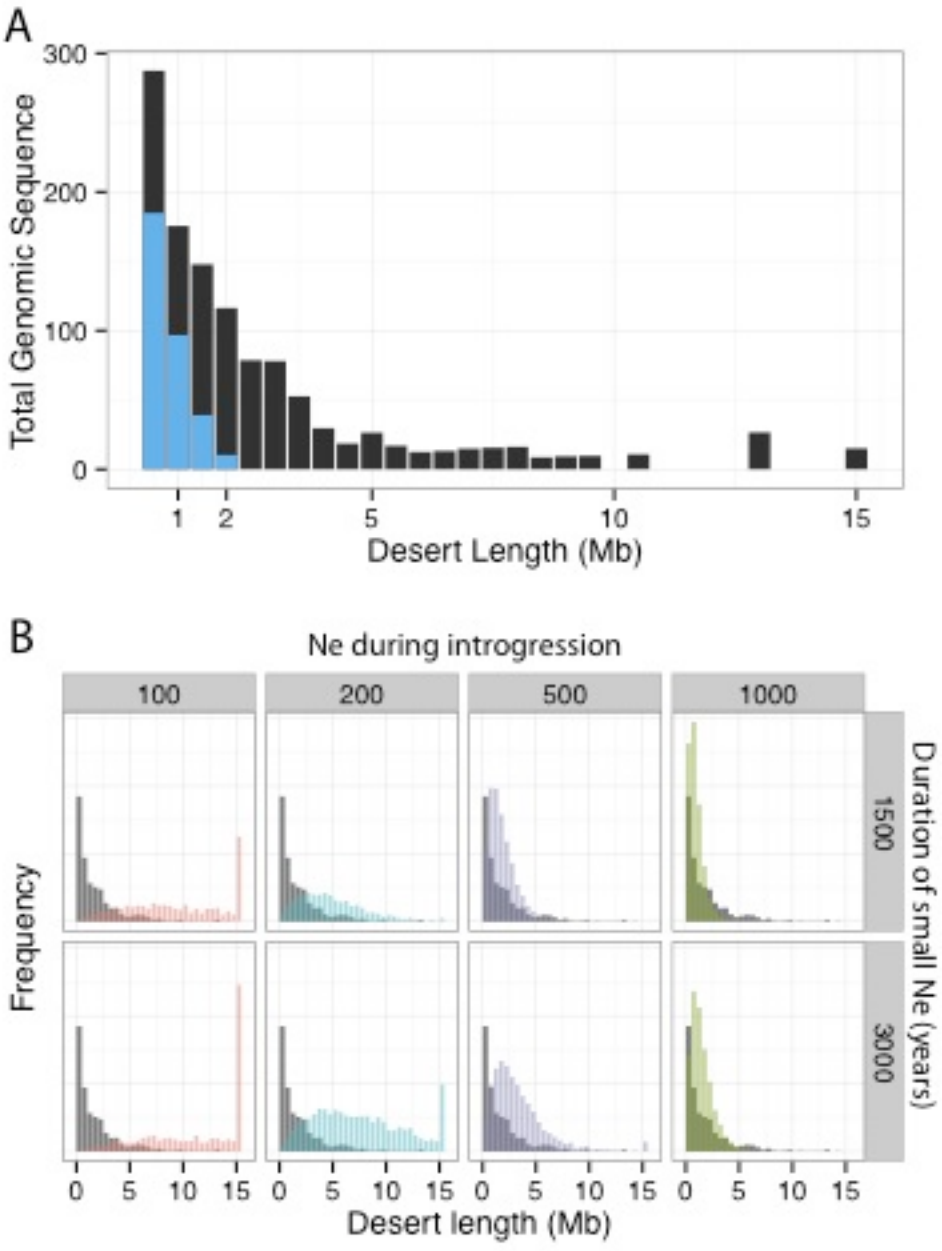

Figure 2. Neanderthal Desert Length Distribution in 665 European and East Asian genomes.

(A) Neanderthal desert lengths (black), and expected Neanderthal desert lengths from simulations (blue) under a standard demographic model. We define a Neanderthal desert as a maximal genomic region such that >95% of all 50kb windows are callable, and less than .3% of all 50kb windows contain introgressed sequence. In all simulations, introgressed fragments were downsampled to match the 25% call rate in [Vernot and Akey, 2014] **(B)** Desert length distribution under alternative demographic models, varying Ne at the time of introgression and the amount of time post-introgression that small Ne was maintained.

To do this, we sequenced 35 Melanesian individuals, and called local Neanderthal and Denisovan ancestry using a method similar to [Vernot and Akey, 2014]. A key difference is that any given Melanesian individual will contain both Neanderthal and Denisovan ancestry, and so it will be useful to separate each predicted introgressed haplotype into one of three categories: likely Neanderthal, likely Denisovan, and ambiguous. As a first approximation, we can identify Neanderthal and Denisovan haplotypes independently, and categorize any haplotype identified by pipelines as "ambiguous." Using this process, we identify 1.8 Gb of Neanderthal sequence at an estimated FDR of 5%, 1.9 Gb of Denisovan sequence at an estimated FDR of 10%, and 1.2 Gb of ambiguous sequence, corresponding to 1.03%, 1.08% and 0.68% of each individual's genome, respectively (Figure 3a).

This is lower than the 2% Neanderthal and 4-6% Denisovan ancestry estimated for Melanesian individuals [Meyer et al, 2012], and in particular we miss roughly 75% of the expected Denisovan ancestry. This is likely due to the larger divergence between the sequenced Denisovan individual and the Denisovan that contributed ancestry to modern Melanesians [Prüfer et al, 2014]. Consistent with this hypothesis, significant S* haplotypes match Denisovan less than Neanderthal, and this affect increases with S* stringency (Figure 3b). However, for the purposes of identifying deserts of archaic sequence, it is less important to distinguish between these categories. For the majority of this chapter, we merge Denisovan and Neanderthal ancestry

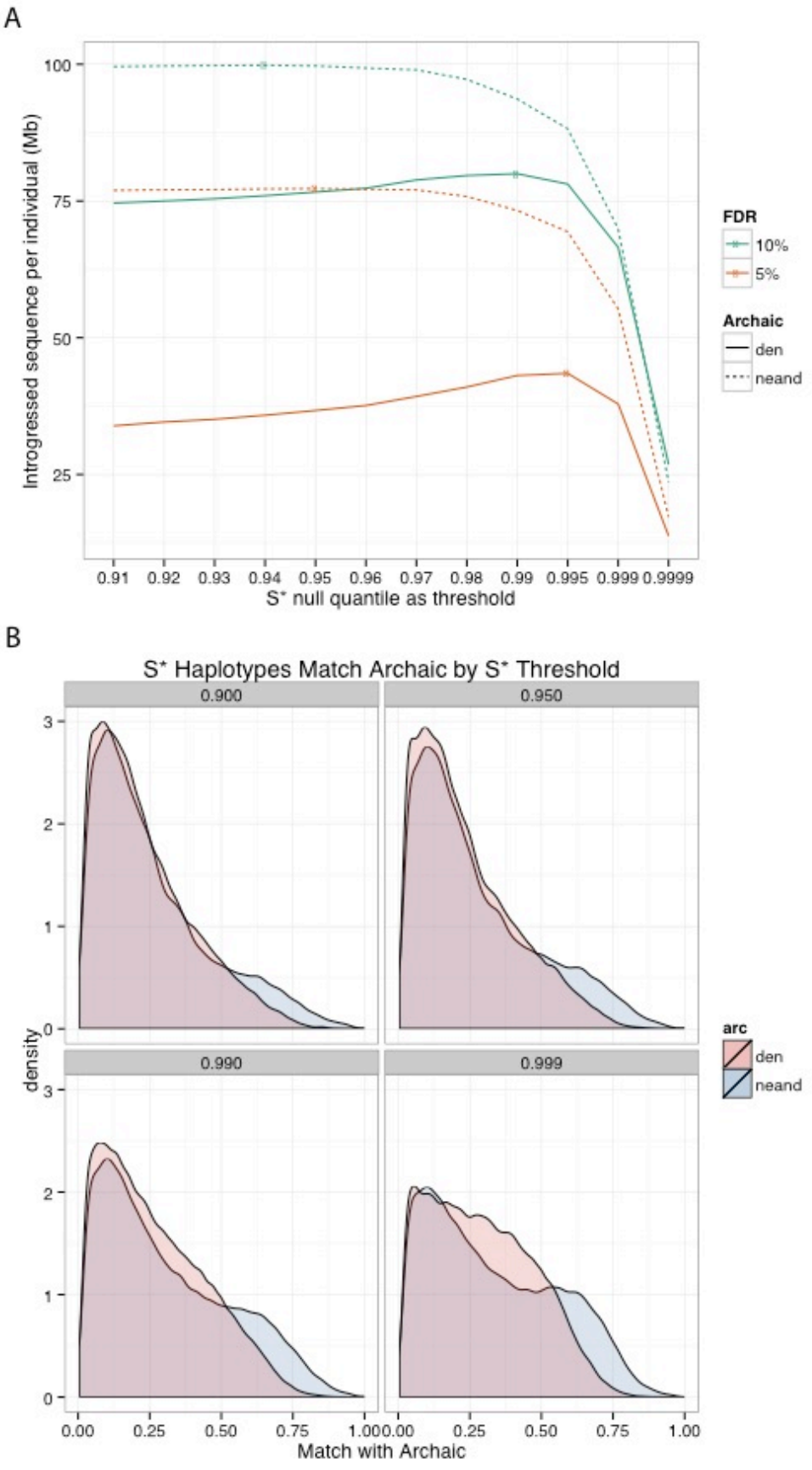
A
100
75
50
25
Introgressed sequence per individual (Mb)
0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 0.995 0.999 0.9999
S* null quantile as threshold
FDR
10%
5%
Archaic
den
neand
B
S* Haplotypes Match Archaic by S* Threshold
0.900
0.950
0.990
0.999
density
0.00 0.25 0.50 0.75 1.00
Match with Archaic
arc
den
neand

Figure 3. Amount of Neanderthal or Denisovan sequence identified at various S* thresholds, and FDR.

(A) As the S* threshold is increased, a decreasing amount of putative introgressed sequence is identified, but this sequence is more enriched for true positives. This interaction differs between Neanderthal (dashed lines) and Denisovan calls (solid line), due to larger evolutionary distance between the introgressing Denisovan population and the sequenced Denisovan individual than is found in Neanderthals [Meyer et al, 2012]. In an attempt to balance the amount of sequence identified, and the false-positive rate in that sequence, we use an S* threshold of 0.95 and FDR of 5% for identifying Neanderthal introgression, and values of 0.99 and 10% for identifying Denisovan introgression. **(B)** Putative introgressed haplotypes as identified by S* on average match Neanderthal better than Denisovan, consistent with [Meyer et al, 2012].

We next compare Neanderthal deserts identified in Europeans and East Asians to maps of Neanderthal and Denisovan ancestry. For example we had previously identified two deserts larger than 3Mb on Chromosome 7 (Figure 4). In the smaller desert region we observe an average of 1.2% archaic sequence per individual, indicating that this desert is not reproduced in these new genomes. In contrast, the larger 15Mb desert has an average of 0.13% archaic sequence per individual, a 20-fold reduction compared to the genome-wide average, consistent with a strong reduction in archaic sequence in Melanesians. Overall, 27/65 Neanderthal deserts representing 131Mb of the genome show a >10-fold reduction in archaic sequence in these 35 Melanesian individuals (Figure 5a).

We have previously observed that local divergence from an archaic species dramatically affects the ability to identify introgressed sequence [Vernot and Akey, 2014]. At the extreme, if there are no differences between Neanderthal and modern humans in a region, then it is impossible to distinguish introgressed and non-introgressed haplotypes. While it is important to consider power to detect archaic introgression in deserts, distance to Neanderthal does not appear to significantly affect the ability to detect introgression in large regions. In genome-wide 5Mb windows, there is no positive correlation between distance to Neanderthal and the amount of archaic introgression detected (Pearson's rho = -.18). Additionally, there is considerable overlap in distance to Neanderthal between Neanderthal deserts and genome-wide 5Mb regions (Figure 5b-c, mean distance of 1.30 x 10-3 and 1.36 x 10-3 per bp respectively).

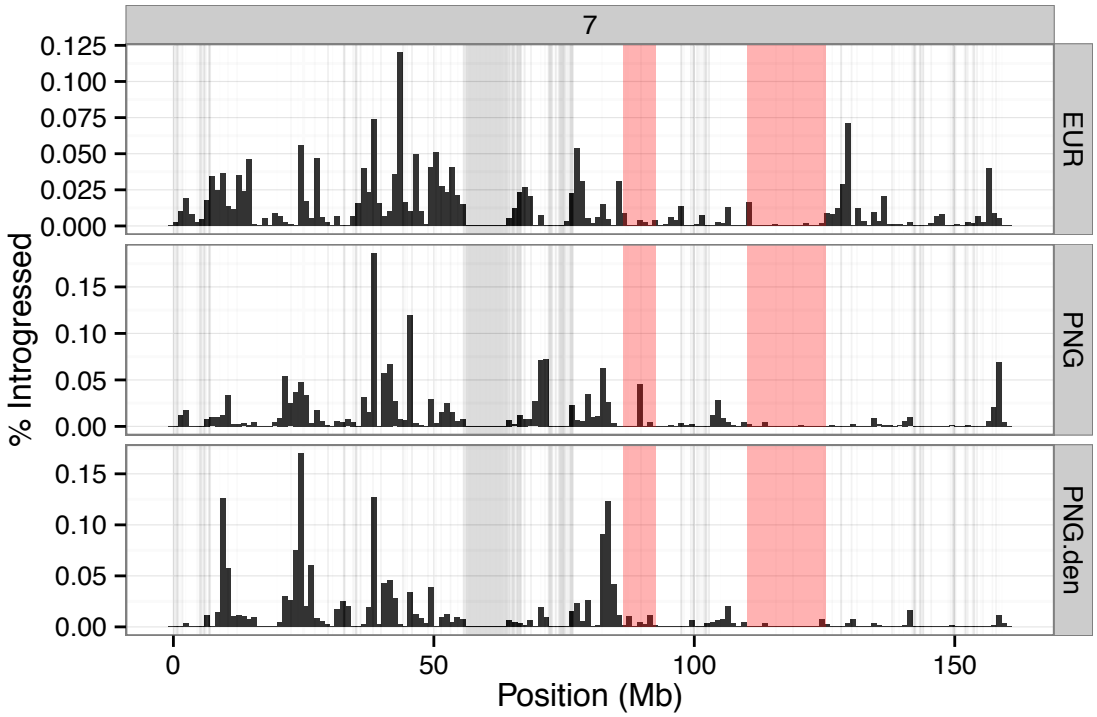

Figure 4. Levels of introgression on Chromosome 7

Average amount of introgression per individual in 1Mb windows, for Neanderthal sequence in Europeans, Neanderthal sequence in Melanesians, and Denisovan sequence in Melanesians. Un-queryable 50kb windows are shown in grey, and deserts identified in [Vernot and Akey, 2014] are shown in red.

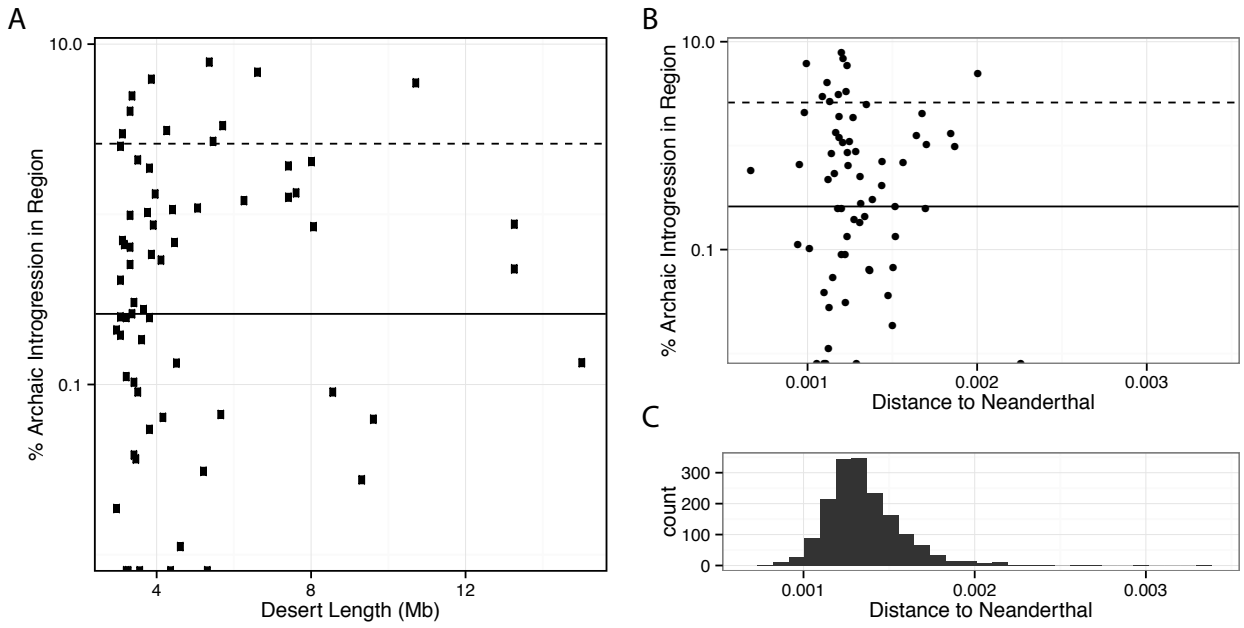

Figure 5. Relationship of desert length to levels of introgression in Melanesians, and distance to Neanderthal.

(A) For each desert identified in [Vernot and Akey, 2014], the amount of Denisovan and Neanderthal introgression in 35 Melanesian individuals. Mean level of introgression (dotted line) and 10-fold reduction in introgression (solid line) are shown in (A) and (B). 27/64 deserts show >10-fold reduction in introgressed sequence. **(B)** Amount of Denisovan and Neanderthal introgression in deserts is not correlated with distance to Neanderthal (similar results hold for comparison to Denisovan). **(C)** Distribution of distance to Neanderthal for genome-wide 5Mb windows.

Given the replication of some Neanderthal deserts in Neanderthal and Denisovan ancestry in Melanesians, and the inability of reduced Ne during the time of introgression to reproduce the desert size distribution, we consider the alternative explanation of purifying selection against archaic sequence. Such purifying selection would occur if incompatibilities existed between archaic sequence in these large regions, and the rest of the modern human genome. So-called hybrid incompatibility loci have been observed in a number of species, including swordtail fishes [Schumer et al, 2014], *A. thaliana* [Chae et al, 2014], and European rabbits [Carneiro et al, 2014]. To investigate the strength of selection require to generate large deserts, we simulate sequence data in a 5Mb region with randomly placed additive deleterious Neanderthal variants, and calculate the probability of observing a desert across the entire 5Mb region (Figure 6) [Hernandez, 2008]. It is immediately obvious that incredibly large amounts of selection are required to generate such large deserts. For example, given two variants with a selection coefficient of .2, a desert is observed 20% of the time. This strength of selection is much higher than other examples of selected loci in modern human populations, e.g. the lactase persistence allele, with estimated s=.01-.03 [Gerbault et al, 2009], and the Duffy-null malarial resistance allele, with estimated s=.066 [Hodgson et al, 2014]. Given these large levels of selective pressures, and the observation that multiple deleterious alleles would be required to generate such deserts, epistatic interactions such as Bateson–Dobzhansky–Muller are a possible explanation [Johnson, 2010].

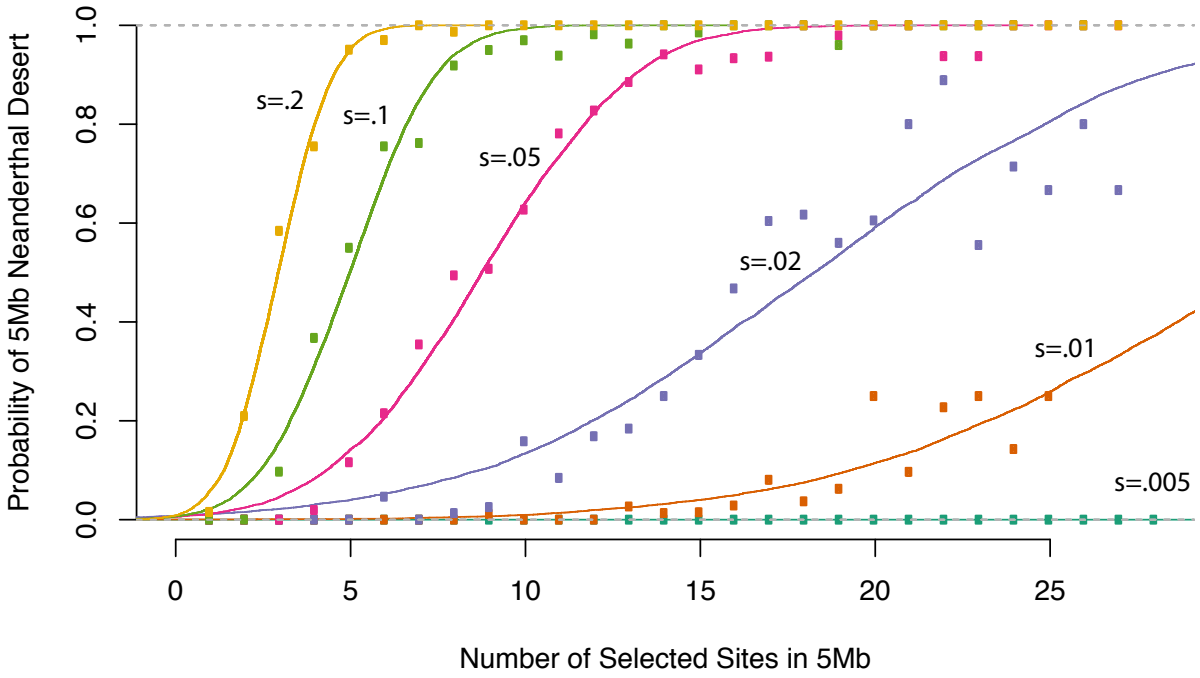

Figure 6. Selective pressure required to produce 5Mb deserts
5Mb regions were simulated with SFS_CODE, 0-30 deleterious archaic sites were randomly placed throughout the regions, and the probability of producing a 5Mb desert was observed. Each color denotes a different selection coefficient for the randomly placed variants, ranging from .005 to .2.

In this chapter I have presented a map of Denisovan and Neanderthal introgression in 35 Melanesian individuals, and used this map to identify regions in the human genome that have resisted archaic introgression in at least two independent introgression events. I have investigated several possible demographic models, all of which failed to reproduce the observed distribution of desert sizes, and have shown that incredibly strong purifying selection against archaic sequence in these regions can produce large deserts of 5Mb. Given the implication that both Neanderthal and Denisovan sequence would be deleterious, it may be that these regions harbor significant modern human specific mutations. Such differences between humans and our archaic hominin cousins may explain phenotypic differences between the species, and in fact may define what it means to be "human". As such, these regions may be of interest for further functional studies. However, this is still an extraordinary claim, and much future work will be required to exclude alternate possibilities, and to identify specific variants that may have generated large selection coefficients. Additional discussion can be found in Chapter 6.

References:

Carneiro, M., Albert, F.W., Afonso, S., Pereira, R.J., Burbano, H., Campos, R., Melo-Ferreira, J., Blanco-Aguiar, J.A., Villafuerte, R., Nachman, M.W., et al. (2014). The Genomic Architecture of Population Divergence between Subspecies of the European Rabbit. PLoS Genet 10, e1003519.

Chae, E., Bomblies, K., Kim, S.-T., Karelina, D., Zaidem, M., Ossowski, S., Martín-Pizarro, C., Laitinen, R.A.E., Rowan, B.A., Tenenboim, H., et al. (2014). Species-wide Genetic Incompatibility Analysis Identifies Immune Genes as Hot Spots of Deleterious Epistasis. Cell 159, 1341–1351.

Gerbault, P., Moret, C., Currat, M., and Sanchez-Mazas, A. (2009). Impact of Selection and Demography on the Diffusion of Lactase Persistence. PLoS ONE 4, e6369.

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

Hernandez, RD. (2008) A flexible forward simulator for populations subject to selection and demography. Bioinformatics, 24:2786-2787

Hodgson, J.A., Pickrell, J.K., Pearson, L.N., Quillen, E.E., Prista, A., Rocha, J., Soodyall, H., Shriver, M.D., and Perry, G.H. (2014). Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proceedings of the Royal Society of London B: Biological Sciences 281, 20140930.

Johnson, N.A. (2010). Hybrid incompatibility genes: remnants of a genomic battlefield? *Trends in Genetics 26*, 317–325.

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. The American Journal of Human Genetics 96, 454–461.

Kimura, M. & Ohta, T. The average number of generations until fixation of a mutant gene in a population. Genetics 61, 763–771 (1969).

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., Filippo, C. de, et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338, 222–226.

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.

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., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. arXiv:1208.2238.

Schumer, M., Cui, R., Powell, D.L., Dresner, R., Rosenthal, G.G., and Andolfatto, P. (2014). High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. eLife Sciences 3, e02535.

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. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 337, 64–69.

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

Vernot, B., and Akey, J.M. (2014a). 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. The American Journal of Human Genetics 96, 448–453.

Chapter 6 : Future Directions

In this chapter, I outline potential future work, and future directions for the study of archaic admixture.

Methods development for analyzing maps of archaic introgression

Although several methods for identifying introgressed sequence in modern humans have been published [Sankararaman et al, 2014; Vernot and Akey, 2014a], there is still a paucity of methods for analyzing such sequence. For example, standard simulations of simple demographic models suggest that introgression into Europeans and East Asians occurred from at least 300 Neanderthal individuals. Because this model assumes all introgression occurred at one time and place, and we know this to be unlikely [Fu et al, 2015; Vernot and Akey, 2015], it is likely that there were many more Neanderthal individuals involved. Thus there is theoretically population-level information about Neanderthals present in the genomes of modern individuals.

It should be possible to develop methods to analyze this population-level information. For example, the distribution of coalescent times between introgressed haplotypes would be informative for the effective population size of the introgressing Neanderthal population – a larger population size would allow deeper coalescences. Methods such as PSMC have been used to estimate historical population sizes from the single high-coverage Neanderthal genome [Prüfer et al, 2014], and these estimates from a single individual could be compared to population estimates to reveal differences between the introgressing population and the sequenced genome – i.e., Neanderthal population structure.

Similarly, patterns of LD between modern human sites that match Neanderthal have been used to estimate the date of introgression at ~55kya [Sankararaman et al, 2012]. However, because this method relied only on matching Neanderthal, and not the more restricted set of probable introgressed haplotypes, its power was low – with 95% CI of 37-86kya. The expected length of introgressed haplotypes is strongly influenced by the date of introgression, with an older introgression event allowing more time for recombination, which results in shorter haplotypes. Thus, a method that explicitly considers the haplotype length distribution would have more power to narrow possible dates of introgression, and potentially to separate the dates of multiple introgression events. However, the observed haplotype length distribution will also be influenced by possible selection for some introgressed haplotypes, which would tend to increase haplotype length, and more importantly by biases in the method for detecting introgressed haplotypes. For example, the methods presented in this thesis have a strong bias towards detecting longer archaic haplotypes – such a bias would have to be considered in methods for analyzing introgression data.

It should also be possible to consider sex-biased gene flow from archaic species into modern humans. There are several examples of sex-biased processes among modern humans [Quintana-Murci et al, 2010; Emery et al, 2010], and so it would be natural to assume a similar process may have occurred between humans and archaic hominins. Although lower rates of introgression have been found on the X chromosome [Sankararaman et al, 2012], this has been explained by increased selective pressure against Neanderthal alleles on the X, rather than sex-biased gene flow, consistent with observations of selective forces acting on X-linked loci [Casto et al, 2010; Lambert et al, 2010]. Alternatively, the depletion of Neanderthal introgression on the X may be explained in part by the fact that it is harder to identify archaic introgression on the

X chromosome, due to decreased divergence between Neanderthal and modern human on the X (on average, 1 difference per 1100-1150bp, vs 1 per 800-900 on autosomal chromosomes). Due to reduced Ne on the X compared to autosomes, there may also be increased variance in the expected amount of Neanderthal introgression, further complicating an analysis of sex-biased gene flow.

Complexity of introgression events untangled by ancient DNA sequencing:

Although some investigation has been done into the timing and demographic model of introgression from Neanderthals into Europeans and East Asians [Sankararaman et al, 2012; Vernot and Akey 2014a; Vernot and Akey 2015; Kim and Lohmueller, 2015], it has so far been difficult to narrow the range of plausible models – e.g., the timing of introgression, whether different populations received archaic introgression during the same event in an ancestral population, or separately after their split, and how many such independent introgression events may have occurred. In the end, however, there are an infinite number of potential models, and it may be difficult to disentangle them from contemporary sequences alone. It is likely that such models will be refined mostly through increased sequencing of individuals who lived within 1-5ky of an introgression event. If the introgression was truly complex, we should observe many differing ages of introgressed haplotypes (as determined by haplotype length distributions). An encouraging example of this is the recently sequenced Oase individual, a modern human found in what is now Romania who lived ~40kya, and who inherited three >50Cm long sections of Neanderthal ancestry. The length of these haplotypes suggests Neanderthal ancestry within 4-6 generations, but additionally this individual carries older Neanderthal ancestry, similar to other

modern humans from this time period [Fu et al, 2015]. This alone confirms that Neanderthal introgression was happening in multiple times and geographic locations.

While ancient DNA from modern humans similar to the Oase individual may be helpful for unraveling the specifics of the introgression event, such sequences will unfortunately be much harder to produce, due to DNA degradation and scarcity of such old bones available for sequencing. For example, while only a handful of individuals older than 30ky have been sequenced [Fu et al, 2013; Fu et al, 2014; Fu et al, 2015], a collection of 83 ancient DNA samples from <10kya was recently announced [Mathieson et al, 2015], and presumably several hundred more samples will be produced in the near future. While these sequences are incredibly useful for modeling recent population movements [Lazaridis et al, 2014] and tracing selected alleles through time [Mathieson et al, 2015; Sams et al, 2015], they will be less informative about the details of the introgression event due to the likely extended separation between introgression and the death of the sequenced individual.

Adaptive introgression of archaic alleles into modern humans:

One of the most prolific fields in the study of archaic admixture has been the identification of putatively positively selected Neanderthal or Denisovan haplotypes in modern humans [Abi-Rached et al, 2011; Mendez et al, 2012a; Mendez et al, 2012b; Huerta-Sánchez et al, 2014; Vernot and Akey, 2014b]. This is an appealing hypothesis – Neanderthals and Denisvans were present in Eurasia for hundreds of thousands of years before the first modern humans arrived, and may have adapted to the local environment in that time frame. Modern humans who interbred with Neanderthals or Denisovans may have then been able to "piggyback" off of

adaptations evolved in the archaic species, and become marginally more suited to their new environment.

The majority of these studies have focused mostly on showing three things: a) that the given haplotype is likely to actually be introgressed from either Neanderthal or Denisovan, b) that the introgressed haplotype covers some gene of interest, and c) that it is at an unusually high frequency in modern humans, suggesting positive selection. In the near future, these studies will extend to showing functional differences between Neanderthal and modern human haplotypes – initially through eQTL-style analyses. Because the Neanderthal and modern human alleles are both present in modern humans, it is possible to find individuals or cell lines that are HH, HN, or NN, and then to look for differences in gene expression in a number of cell types [The GTEx Consortium, 2015]. These analyses may give an indication as to the potential function of the introgressed haplotype, but do not demonstrate an actual phenotype, nor identify the selected variant. It is even possible that if there is a selected variant, it is a modern human variant that happens to be adjacent to an introgressed haplotype, and that the Neanderthal sequence is merely hitchhiking to high frequency. Future work will have to focus on identifying a phenotypic effect, and fine-scale mapping to identify the causative variant.

A different category of study might attempt not to find specific beneficial introgressed alleles, but rather to show the overall effect of archaic introgression into modern humans. Perhaps, for example, Neanderthal alleles are broadly associated with an increased risk for autoimmune diseases, resistance to Eurasian pathogens, or high-latitude associated phenotypes.

Neanderthal deserts, purifying selection against archaic sequence and the search for functional differences between modern humans and archaic hominins

In both previously published genome-wide maps of introgressed sequence [Sankararaman et al, 2014; Vernot and Akey, 2014a], large regions of the genome were observed to have no Neanderthal introgression. The largest such Neanderthal "desert," on chromosome 7q, is 15Mb long and was identified by both studies. Given that Neanderthal ancestry over a large portion of the genome was fairly evenly distributed, and the incredibly low correlation between the location of Neanderthal ancestry in any two given individuals (Fig. S1, Appendix A), it appeared unlikely that such large deserts could occur by chance. Indeed simple demographic models suggested that the largest deserts under neutral models of evolution should be about 2-3Mb (Fig 2a, Chapter 5). An alternative explanation is large-scale purifying selection against the archaic haplotype in the context of modern human genomes. In this case, Neanderthal deserts may harbor essential sequence differences between Neanderthals and modern humans, and may in fact define what it means to be "modern human," making them an intriguing topic for further study.

In Chapter 5, I explored a few modifications to a basic demographic model, and showed that large archaic deserts are unlikely under those models as well. However, additional work is needed to thoroughly exclude non-selective forces as the cause of Neanderthal deserts. This work will include additional simulations of possible demographic models, including simulations of structured populations or spatial models (i.e., Neanderthal admixture into a small wave-front of modern humans entering the Middle East). This type of simulation is difficult to perform, because by definition large regions must be simulated, and each parameter set can take days of CPU time for just 1000 iterations. Spatially structured simulations are even more CPU and memory-prohibitive [Ray et al, 2010]. Any such simulations must be able to reproduce not only

large Neanderthal deserts, but also other known characteristics of Neanderthal introgression, such as Neanderthal haplotype frequency and the full desert size distribution. Some preliminary simulations, for example, produce 15Mb deserts, but cannot also reproduce the distribution of Neanderthal deserts from 1-5Mb.

An additional possible explanation for some Neanderthal deserts is structural variation that has occurred or fixed in the time since the divergence of modern humans and Neanderthals. Specifically, large inversions can depress recombination within the inverted region [Stefansson et al, 2005], potentially leading to low level of introgression. Unfortunately inversions are difficult to identify in modern sequencing data, and may be impossible to detect in the short DNA fragments present in ancient samples (mean fragment length 60-150bp, [Prüfer et al, 2010]). A future study, however, could look at rates of inversions between species of varying divergence, and determine if it is likely to see as many large inversions between Neanderthals and modern humans as would be required to explain the observed ~350Mb of deserts larger than 3Mb.

The most promising avenue for excluding demographic or structural extremes as the cause of Neanderthal deserts is through the comparison of multiple independent introgression events. In the extreme case, this is exemplified by comparing Denisovan to Neanderthal introgression, but it is possible that there were multiple independent introgression events from Neanderthals into modern humans. Because demographic explanations for deserts of archaic sequence rely on random loss of large regions of Neanderthal sequence shortly after introgression, it is unlikely that this process would randomly occur in the same genomic region across multiple introgression events. Similarly, for an inversion to restrict introgression from both species, it would have to have occurred and risen to high frequency in the relatively short

amount of time before Neanderthals and Denisovans split (~100ky), or have occurred and fixed in modern humans.

If alternative explanations for large Neanderthal deserts can be eliminated, then selection remains a plausible choice. Under an additive model of selection, large levels of selective pressure are required to explain large deserts (Figure 6, Chapter 5). In future work, additional investigation will be needed to clarify the possible effects of more complicated models of selection, including epistasis or Dobzhansky-Muller hybrid incompatibility [Johnson, 2010; Schumer et al, 2014].

Ultimately, the goal of investigating Neanderthal deserts may be to identify alleles that differ between Neanderthals and modern humans, and that may have represented barriers to introgression. Such alleles could account for phenotypic differences between the two populations, but a functional effect would need to be determined. This is much more challenging than investigating adaptive introgression, simply because the Neanderthal alleles by definition do not exist in modern humans, barring back mutations.

It may be useful to perform large-scale scans of potential functional sites before exhaustively analyzing any given site. Because most differences between humans and Neanderthals are not coding differences, assays of expression differences may be appropriate – e.g., a reporter assay testing Neanderthal and modern human specific regulatory alleles in a variety of tissue types. Although such a scan could focus attention on the large deserts of archaic sequence, given the levels of selection possibly required to explain such large deserts (Chapter 6, Figure 5), it is plausible that many functional sites could be interspersed with higher levels of introgression, making it important to not restrict any scans for functional variation only to

deserts. After promising candidate sites are identified, it would be interesting to "Neanderthalize" human iPS cells via genome editing technologies, which could then be differentiated into various cell types or organoids and investigated for gene expression differences, and then phenotypic differences.

Any explanation for Neanderthal deserts that invokes large amounts of selection against Neanderthal variants requires fixed functional differences between Neanderthals and modern humans. More specifically, that there are small collections of differing sites that explain large amounts of the difference. It is possible that the majority of phenotypic differences is due to differing allele frequencies at relevant sites, making it more likely to bring collections of functional alleles together in one species than another [Pritchard et al, 2010]. This possibility dramatically increases the number of candidate functional sites, and complicates any search for the genetic underpinnings of phenotypic differences between modern humans and archaic species.

Non-Neanderthal or Denisovan archaic admixture

Given large increases in the ability to sequence DNA from ancient bones [Kircher, 2012], and new methods for quickly identifying the genus or even species of small bone fragments [Brandt et al, 2015], it is increasingly likely that we will find other examples of archaic hominin species that overlapped geographically and temporally with early modern humans, assuming that such species exist. Even in the absence of a new archaic genome, it may be possible to determine if significant introgression from a third species has occurred into any human population simply by looking for putative introgressed haplotypes that do not match Neanderthal or Denisovan.

However, as current ancient-DNA-free methods have false discovery rates from 35-50% [Vernot and Akey, 2014a], significant work remains to improve such methods.

Despite current challenges with identifying introgressed sequence in the absence of a comparative ancient genome, I expect that there will be a lot of interest in this in the next few years. When the S* statistic was initially developed in 2006, it was used to look for signatures of Neanderthal introgression in Europeans, and for introgression from unknown archaic hominins into Yorubans [Plagnol and Wall, 2006]. This study was published four years before the draft of the Neanderthal sequence, and now that it has been confirmed that introgression has happened at least twice, it is fruitful to more fully explore the possibility of detecting other archaic introgression events. The software generated as part of this thesis can interpret both sequencing and simulated data, and thus can be useful for testing variations of methods for identifying introgressed haplotypes. I have hopes that it will be of use to people interested in this field.

Introgression and admixture in non-humans:

Although this thesis has focused on analyses of admixture between humans and other hominins, there are many examples of such admixture in non-human species. These include introgression of adaptive alleles from one species into another [Hedrick et al, 2013], e.g. rodent poison resistance in old world mice [Song et al, 2011]. Examples of hybrid incompatibility loci have been demonstrated in *Arabidopsis thaliana* [Chae et al, 2014], various European rabbit species [Carneiro et al, 2014], and swordtail fishes [Schumer et al, 2014]. As these examples demonstrate, hybridization is common in natural populations, and the effects postulated to have

occurred in modern humans occur in other animal species, with observable functional and evolutionary consequences.

References

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. Science 334, 89–94.

Brandt et al, PLoS ONE, 2015, Species Identification of Archaeological Skin Objects from Danish Bogs: Comparison between Mass Spectrometry-Based Peptide Sequencing and Microscopy-Based Methods

Carneiro, M., Albert, F.W., Afonso, S., Pereira, R.J., Burbano, H., Campos, R., Melo-Ferreira, J., Blanco-Aguiar, J.A., Villafuerte, R., Nachman, M.W., et al. (2014). The Genomic Architecture of Population Divergence between Subspecies of the European Rabbit. PLoS Genet 10, e1003519.

Casto, A.M., Li, J.Z., Absher, D., Myers, R., Ramachandran, S., and Feldman, M.W. (2010). Characterization of X-Linked SNP genotypic variation in globally distributed human populations. Genome Biology 11, R10.

Chae, E., Bomblies, K., Kim, S.-T., Karelina, D., Zaidem, M., Ossowski, S., Martín-Pizarro, C., Laitinen, R.A.E., Rowan, B.A., Tenenboim, H., et al. (2014). Species-wide Genetic Incompatibility Analysis Identifies Immune Genes as Hot Spots of Deleterious Epistasis. Cell 159, 1341–1351.

Emery, L.S., Felsenstein, J., and Akey, J.M. (2010). Estimators of the Human Effective Sex Ratio Detect Sex Biases on Different Timescales. The American Journal of Human Genetics 87, 848–856.

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 advance online publication,.

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

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., and Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. PNAS 110, 2223–2227.

Hedrick, P.W. (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol Ecol 22, 4606–4618.

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 advance online publication,.

Johnson, N.A. (2010). Hybrid incompatibility genes: remnants of a genomic battlefield? Trends in Genetics 26, 317–325.

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. The American Journal of Human Genetics 96, 454–461.

Kircher, M. (2012) Analysis of High-Throughput Ancient DNA Sequencing Data - Springer. B. Shapiro, and M. Hofreiter, eds. (Humana Press)

Lambert, C.A., Connelly, C.F., Madeoy, J., Qiu, R., Olson, M.V., and Akey, J.M. (2010). Highly Punctuated Patterns of Population Structure on the X Chromosome and Implications for African Evolutionary History. The American Journal of Human Genetics 86, 34–44.

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.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Llamas, B., Pickrell, J., Meller, H., Guerra, M.A.R., Krause, J., Anthony, D., et al. (2015). Eight thousand years of natural selection in Europe. bioRxiv 016477.

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.

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.

Plagnol, V., and Wall, J.D. (2006). Possible Ancestral Structure in Human Populations. PLoS Genet 2, e105.

Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology 20, R208–R215.

Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology 20, R208–R215.

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., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., and Green, R.E. (2010). Computational challenges in the analysis of ancient DNA. Genome Biology 11, R47.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P.D., Hoal, E.G., and Behar, D.M. (2010). Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. Am J Hum Genet 86, 611–620.

Ray, N., Currat, M., Foll, M., and Excoffier, L. (2010). SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26, 2993–2994.

Sams, A.J., Hawks, J., and Keinan, A. (2015). The utility of ancient human DNA for improving allele age estimates, with implications for demographic models and tests of natural selection. Journal of Human Evolution 79, 64–72.

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., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. arXiv:1208.2238.

Schumer, M., Cui, R., Powell, D.L., Dresner, R., Rosenthal, G.G., and Andolfatto, P. (2014). High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. eLife Sciences 3, e02535.

Simms, T.M., Wright, M.R., Martinez, E., Regueiro, M., McCartney, Q., and Herrera, R.J. (2013). Y-STR diversity and sex-biased gene flow among Caribbean populations. Gene 516, 82–92.

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M.W., and Kohn, M.H. (2011). Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. Curr Biol 21, 1296–1301.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. (2005). A common inversion under selection in Europeans. Nat Genet 37, 129–137.

The GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660.

Vernot, B., and Akey, J.M. (2014a). Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science 343, 1017–1021.

Vernot, B., and Akey, J.M. (2014b). Human Evolution: Genomic Gifts from Archaic Hominins. Current Biology 24, R845–R848.

Vernot, B., and Akey, J.M. (2015). Complex History of Admixture between Modern Humans and Neandertals. The American Journal of Human Genetics 96, 448–453.

Appendix A: Supplementary Material for Chapter 3

Expected fraction of the genome covered by surviving Neanderthal lineages

To estimate the fraction of an archaic genome that can potentially be recovered from fossil free sequencing as a function of sample size and amount of introgression, we performed coalescent simulations using previously inferred demographic models for European, East Asian, and African populations (*25-28*). Simulations were performed with ms (*29*). The base demographic model is: a) splitting between modern humans and Neanderthals at either 400kya or 700kya (depending on the analysis – see note below), b) Neanderthal N_e of 1,500, c) splitting between Africans and non-Africans at 60kya, d) gradual growth of non-African populations starting at 23kya, ending at 5,115 years ago, to N_e of 8,879 in East Asians and N_e of 9,475 in Europeans, e) African N_e of 14,474 until 5,115 years ago, f) rapid growth of all populations starting at 5,115 years ago, to a modern day N_e of 424,000 in Africans, 512,000 in Europeans, and 1,370,990 in East Asians, g) a single 500 year introgression event from Neanderthals to the common ancestor of Europeans and East Asians, at a rate of 0.0015 (i.e., 0.15% of each generation was sampled from Neanderthal individuals), h) migration rates were set as follows: 1.498975×10^{-4} between Africans and the ancestors of Europeans and East Asians, 2.498291×10^{-5} between Africans and Europeans, 7.794668e-06 between Africans and East Asians, and 3.107874×10^{-5} between Europeans and East Asians. The above parameters were adapted from (*27,28*). Neanderthal N_e estimate from (44). Note: although a Neanderthal / modern human split time of 700kya is older than the upper bound of 500kya estimated by (*3*), this time was chosen to ensure that any introgressed lineages coalesced before non-introgressed lineages, making it

easier to identify introgressed lineages by examining the coalescent trees. This older split time has no effect on the amount of introgressed sequence, as we are considering only the presence and extent of introgressed haplotypes, not variation that has arisen on those haplotypes. *For simulations in analyses where variation is considered, we use a split time of 400kya*.

To generate demographic models from this base model, we then sampled the following parameters from uniform distributions:

a) European and East Asian split times between 36kya and 50kya [T(S)], b) European + East Asian N_e prior to 23kya between 4,000 and 18,353, c) European N_e / East Asian N_e between 1 and 2.5 (additionally, rejecting any parameter sets where European N_e > 9,475 or East Asian N_e > 8,879; N_e of the ancestors of Europeans and East Asians was set to European N_e), d) the time of the first introgression pulse between T(S) and 55kya, e) a second 500 year introgression event from Neanderthals to East Asians, starting 500 years after T(S), at rate a between 0.00015 and 0.0005.

For 40 random models generated from the parameters described above, we then simulated sequence data using ms (*29*) for 200 African, 200 European and 200 East Asian individuals, along with a single Neanderthal chromosome, in 50kb windows with a recombination rate set to 10^{-8} recombinations/bp/generation. We identified introgressed haplotypes by using the –T flag to generate coalescent trees, and identifying trees where the archaic chromosome joins the tree before the simulated join time. The fraction of the genome covered by introgressed lineages as a function of the number of individuals sampled is shown in Fig. S1, and the distribution of the fraction of the genome covered when sampling 200 individuals is shown in Fig. S2.

Calculation of *S** for individuals

Plagnol and Wall (11) developed a simple summary statistic, *S**, that is sensitive to characteristics of introgressed sequences. On average, introgressed haplotypes are expected to have an older TMRCA compared to non-introgressed lineages and therefore exhibit high levels of divergence. Moreover, because admixture occurred relatively recently, the introgressed haplotype will tend to persist over sizeable genomic regions (~50 kb in the case of Neanderthal introgression; 11). Finally, because Neanderthal admixture is expected to have occurred only in non-African populations, variants on the introgressed haplotype should not be found in African individuals (an assumption that can be relaxed to allow for the possibility of gene flow between African and non-African individuals). Thus, S* is designed to detect divergent haplotypes whose variants are in strong linkage disequilibrium and are not found in a "reference" population. It is important to use a reference population that contains a minimal amount of archaic (i.e., Neanderthal) introgressed sequences. We chose 13 randomly selected Yoruban genomes as a reference population; a recent analysis found that levels of Neanderthal variation in Yorubans are not statistically enriched (as measured by the D statistic), as opposed to Sandawe, Maasai, and African Americans (45). An attractive feature of *S** is that it does not require an archaic sequence and is therefore applicable to detecting sequences from unknown archaic groups. However, as originally implemented, *S** is calculated for a genomic region and is designed to test the hypothesis of whether there is significant evidence for introgressed sequence at this locus (11). Therefore, it does not specifically identify archaic sequence or which individuals carry it. We therefore extended *S** such that it is calculated on individuals and then a hypothesis test is

performed to determine whether any individuals have an *S** large enough to reject the null hypothesis that they do not carry introgressed sequence (see below). The statistic for the i^{th} individual in a region is calculated as $S_i^* = \max_{J \subseteq V_i} S(J)$, where:

$$S(J) = \sum_{j \in J} \begin{cases} -\infty, & d(j, j+1) > 5 \\ -10000, & d(j, j+1) \in \{1 \dots 5\} \\ 5000 + bp(j, j+1), & d(j, j+1) = 0 \\ 0, & j = max(J) \end{cases}$$

V_i is the set of all variants in individual *i* in this region, and *J* is a subset of those variants. Variants that are also found in the reference population are not included in this analysis. In the calculation of *S()*, we treat *J* as a list of variants ordered by genomic position. Thus, the variants *j* and *j+1* denote adjacent variants in *J*. *d(j,j+1)* is the genotype distance between two variants, where genotypes are coded as 0, 1, and 2, and the distance between two variants is the sum of the difference between their genotype values in each individual (an example is given below). The term *bp(j,j+1)* is the distance in base pairs between two variants. In the calculation of *S(J)*: -∞ does not allow consecutive variants in *J* to have more than five genotype differences; -10000 is a penalty for consecutive variants with 1-5 genotype differences; variants with no genotype differences (perfect LD) are scored 5000 + the distance between them, which gives a higher score to variants in perfect LD that extend over larger distances; the final line allows the last variant to be added. To calculate *S**, we find the set of variants *J* that maximizes *S(J)*. We developed an efficient dynamic programming algorithm that allows computation of *S** in genome-wide datasets.

Dynamic programming computation details:

For each individual *i*, our goal is to find the subset (*J*) of variants in that individual (V_i) that maximizes *S(J)*; in the original implementation (11), variants were not required to all be present in a given individual. As in the original version, we require the variants in *J* be at least 10 bp apart (to minimize the effects of clusters of variants and/or clusters of bad calls from segregating indels), and remove any variants found in the reference population.

To find the optimal value of *J* (*J* such that *S(J)* is maximized), we could try all possible values of *J*. However, this would be intractable for larger numbers of variants, as the number of subsets is 2^n, where *n* is the number of variants. Instead, we can find the optimal *J* by using a dynamic programing algorithm, thus considering only n^2 pairs of variants.

Recall that in the calculation of *S(J)*, we treat *J* as a list of variants ordered by genomic position. The dynamic programming algorithm depends on the fact that $S(G+y) = S(G-x) + S((x,y))$, where *G* is an ordered set of variants, and *x* is the last variant in *G*. To find *S**, we calculate *S(G+y)* for all possible *x,y* pairs (e.g., Table S4), and select the set *J* that maximizes *S(J)* (e.g., the maximum cell in Table S4).

For clarity, we include a worked example. In this example, we are calculating *S** in a 50kb region for four individuals from a target population (analogous to Europeans or East Asians in our analysis), additionally using two individuals from a reference population (analogous to Yorubans in our analysis). Genotypes for the example individuals are given in Table S2. We show *S** calculations for the first two individuals (individual 1: Tables S3-S5; individual 2: Tables S6-S8), and walk through the calculations for individual 1 below.

In Table S3, we show genotype distance between all snps present in individual 1. These distances are used in the calculation of S(J) (Table S4). In Table S4, we show *S(J)* for all pairs

(x,y) of variants in individual 1,, where *x* and *y* are the last two variants in the set *J*. The columns represent *x*, and the rows represent *y*. To generate this table, we calculate all cells for each row, from top to bottom. We start in the upper left corner, and calculate *S(J)* where *(x,y) = (snp_0, snp_8)*. The genotype distance between these variants is 0 (abs(1-1) + abs(0-0) + abs(0-0) + abs(0-0); shown in Table S5), so *S(J)* = 5000 + *bp(snp_0, snp_8)* = 5000 + 25354 - 2309 = 28045.

The next row (*snp_9*) only contains cells where the genotype distance is greater than 5; thus the values of these cells are -∞. To calculate the exact value for the cell for *(x,y) = (snp_8, snp_9)*, we take the column snp (*snp_8*), find the row for *snp_8*, find the maximum value in that row (i.e., the column snp x that maximizes *S((x,snp_8))*), and add that value to *S((snp_8, snp_9))*. In this case, the maximum value in the row *snp_8* is 28045, but *S((snp_8, snp_9))* = -∞. The same is true for the first two cells in row *snp_10*.

For the third cell, *(x,y) = (snp_9, snp_10)*, the maximum value in row *snp_9* is essentially -∞; thus the best option for this cell is not to take a previous value, but rather to restart the calculation as if the set began with *snp_9* - this is analogous to the top left cell. Thus, the value for this cell is 0 + 8070.

As a final example, consider the cell for *(x,y) = (snp_8, snp_13)*. The best value in the row *snp_8* is 28045 [*(x,y) = (snp_0, snp_8)*], and *S((snp_8, snp_13))* = 16845. Thus, the value in this cell is 44890, which also happens to be the largest *S** for individual 1. The optimal set of variants is *(snp_0, snp_8, snp_13)*. Obviously, the entire table must be calculated, as shown, in order to identify the cell with the largest value. The corresponding tables for individual 2 are given in Tables S6-S8.

Null model coalescent simulations

To determine the expected distribution of *S** under the null model of no introgression, we first simulated sequence data using the program ms (*29*). To simulate East Asian, European and African sequences, we used demographic models given in (*27*) and (*28*). Specifically, we simulated divergence of African and non-African populations at 51kya, divergence of European and East Asian populations at 23kya, and recent growth in all three populations. Tennessen et al. (*27*) used 2,439 exome sequences to re-estimate recent growth parameters from (*28*) in European and African descent populations. To approximate similar recent growth rates in East Asian sequences, we used the model from (*28*), but included exponential population growth similar to that in Europeans. In (*28*), N_e in East Asian populations was inferred to grow at a rate 0.48/0.38 times faster than in Europeans - accordingly, for East Asian populations we used recent growth rates that were equal to 0.48/0.38 times the European growth rates in (*27*). These are the same growth parameters described in the section "Expected fraction of the genome covered by surviving Neanderthal lineages".

*S** values depend on demographic model, recombination rate, and the number of variants in a region. For the population pairs East Asians+Africans and Europeans+Africans, we simulated sequences on a grid for all appropriate values of number of variants and recombination rates; from 30 to 350 variants (step=5), and ln(recomb) from -10.25 to 2.75 (step=0.25) (Fig. S6). We simulated 20,000 50kb regions for 20 non-African and 13 African individuals per grid point. We then calculated *S** on these loci, and recorded the maximum value of *S** for each 50kb

region. From these 20,000 values of *S**, we were able to obtain the expected distribution of *S** under the null model conditional on the number of variants and recombination rate; specifically, we selected the 0.05 and 0.01 quantiles. To estimate null distribution quantiles for arbitrary windows, we fit a generalized linear model to the grid of S* quantiles using the R package mgcv (*30*). An example command is: gam(q99~te(lr, snps, k=10), data=sims, method="GCV.Cp").

Evaluating the performance of *S**

To evaluate the ability of *S** to identify introgressed sequence under a variety of demographic models, we randomly selected 80 demographic models, simulated sequence data, and identified true introgressed haplotypes (as described in the section "Expected fraction of the genome covered by surviving Neanderthal lineages"). We then calculated *S** on all European individuals, using the African individuals as a reference population. For a sliding *S** threshold, we called individuals as introgressed or non-introgressed for each region; we then calculated the fraction of true positives and true negatives (Fig. S4). We estimate that *S** can identify ~30% of all true positives at a low false positive rate, and that this result is largely invariant across demographic models (Fig. S4). Additionally, to determine the robustness of *S** thresholds at *p-value* $\leq$ 0.01 across multiple demographic models, we selected 700 random models and calculated *p-value* $\leq$ 0.01 thresholds for each. We find that these thresholds are largely stable across demographic models (Fig. S7).

Whole-genome sequences and data filtering

We obtained low-coverage whole-genome sequences from 379 individuals of European ancestry and 286 individuals of East Asian ancestry (Table S1) that were sequenced as part of the 1000 Genomes Project (*12*). We also downloaded high-coverage Neanderthal alignments generated from a toe bone discovered in Denisova Cave (*31*) from: http://cdna.eva.mpg.de/neandertal/altai/bam/. Neanderthal variants were called with SAMtools mpileup and BCFtools (*32*), using the flag –C50 to downweight reads with excessive mismatches, and –E for extended BAQ computation. We then removed INDELs, and required the following for SNVs: minimum RMS mapping quality of 25, minimum read depth of 10. We also filtered any variants within 10bp of an INDEL.

For all analyses, we removed CpGs, genomic super dups (*33*), and difficult to map bases (*34*), as described below.

- We identified CpG sites as in (*24*) by considering the human reference genome (hg19), genomes sequenced by Complete Genomics (15 hunter-gatherer genomes [*24*] and the public data release), and three additional primate genomes (chimpanzee, orangutan, and macaque). To account for ancestral CpGs, we counted any position that is C or G in one genome, and adjacent to a G or C (respectively) in another genome. This CpG metric is less conservative than the metric used for placental mammals, but more conservative than the metric used for primates in (*35*).
- Segmental duplications (*33*) were downloaded from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz
- Mappable regions were determined by examining all 35 base long "reads" that overlap a each site. A site is mappable if the majority of overlapping reads are mapped uniquely or without 1-mismatch hits to hg19 (*34*).

For all analyses in which we considered the Neanderthal sequence, we removed any bases that were not "Neanderthal callable" – i.e., we required all bases to have minimum RMS mapping quality of 25, minimum read depth of 10, and be further than 10bp from any INDELs.

Identifying putative Neanderthal lineages with *S**

We calculated S* on the 379 European and 286 East Asian individuals, in 50kb windows at a 20kb step, using the following procedure. First, because *S** was developed to work with small numbers of individuals (*11*), for every 50kb window we randomly subsetted each population into sets of 20 individuals (the last subset for Europeans was 19 individuals, and the remaining six East Asian individuals were excluded), and calculated *S** for each population on those 20 individuals. To reduce the potential influence of consistently grouping some individuals together, different subsets were used for every 50kb window. We used 13 randomly selected Yoruban individuals as the reference population. Additionally, we performed the entire process described above five times.

To determine if an individual had a significant *S** value, we calculated *S** thresholds using a generalized linear model with recombination rate and the number of variants present in the region (African and non-African combined) as parameters (as described in "Null Model Coalescent Simulations"). For each 50kb window, we calculated the average recombination rate over that window from fine-scale recombination data obtained from: http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37 (*36*). Regions where the number of variants or recombination rate fell outside of the simulation limits were discarded. In addition, we required each window to contain at least 25kb of unfiltered bases. We then selected all individuals with *S** greater than the *p-value* ≤ 0.01 threshold. Table S9 shows

the amount of introgressed sequence called at different *S** thresholds that have a Neanderthal enrichment match FDR ≤ 5%. Note, that ~99% of all introgressed sequence called at *S** *p-value* thresholds of 0.0001, 0.0005, and 0.005 are also called at an *S** *p-value* ≤ 0.01. This observation provides confidence that our computationally efficient approach for estimating *p-values* is performing well.

To identify the actual introgressed haplotype on each individual, we examined variants identified by *S** (which we refer to as "tag SNVs"), and identified the computationally estimated haplotype from 1000 genomes data. Singleton, doubleton and tripleton SNVs were not considered, as computational phasing is difficult at such low frequencies. Additionally, at least 80% of all *S** tag SNVs were required to be present on the same haplotype. Reported putative haplotypes span from the farthest bases that pass these filters, and are on the majority haplotype; i.e., given a set of variants on the haplotypes 1,1,2*,2,2,2,2,2,2,2,2,2,2,2,2*,1, the introgressed sequence would be inferred to be on haplotype 2, and span between the two starred variants. A haplotype was considered *S** significant if it was identified in any of the five runs of the genome.

Calculation of Neanderthal match *p-values*.

For every putative Neanderthal haplotype that passed the filtering described above, we then calculated *p-values* of how well that haplotype matched the Neanderthal sequence compared to random similar haplotypes in non-introgressed portions of the genome. Specifically, we calculated the number of variants on that haplotype that were also present in the Neanderthal sequence, divided by the number of variants present on the human haplotype and in the Neanderthal sequence. For example, consider a putative haplotype with 15 variants, in a region where Neanderthal shares 3 of those variants, and has an additional 6 variants. Then M_n =

3/(15+6). For each putative haplotype, we calculate an empirical *p-value* by comparison to similar haplotypes in non-introgressed portions of the genome, conditioned on a) number of variant sites on the haplotype (+/- 1), b) average frequency of variants on the haplotype (+/- 5%), c) standard deviation of frequency of variants on the haplotype (+/- 5%), and d) length of the haplotype (+/- 1000bp). Random haplotypes were drawn only from the portion of the genome not covered by putatively introgressed sequence (i.e., S^* threshold of *p-value* > 0.05, which corresponds to ~1/3 of the genome). On average ~200,000 such similar independent haplotypes are selected for every putative introgressed haplotype.

To further characterize our ability to identify Neanderthal haplotypes in real data, we compared the FDR and recall of two different methods for filtering high-S^* haplotypes into a final call set. For both methods, we start with a call set that consists of all haplotypes with S^* *p-value* ≤ 0.01. As shown in Fig. 1E, we estimate this initial call set to have an FDR of 60%. We estimate FDRs in each method using Neanderthal *p-values* as in Figs. 1D and 1E.

In the first method, we do not use the Neanderthal sequence to filter haplotypes – this method is explicitly agnostic to the Neanderthal sequence, and is analogous to a scenario where we wish to identify archaic sequence from a species with no ancient DNA sequence (23, 24), although an archaic sequence could not be used to test the efficacy of the method. It is likely, however, that characteristics that are effective at distinguishing modern human haplotypes from archaic haplotypes in European or East Asian populations would also be effective in African populations. To filter haplotypes, we trained a bound-constraint SVM using the R package kernlab (43). We used presence/absence of a 50kb window in our final call set from the main text as the dependent variable, and several characteristics of both the putative introgressed haplotypes, 50kb windows in which the haplotypes were identified, and surrounding genomic

region, as explanatory variables (defined below). The SVM was trained on 10k randomly selected haplotypes, and then tested on 100k random haplotypes that did not overlap the training haplotypes. For this test subset of sequences, we recover 74 Mb of sequence at FDR = 20% using the SVM.

The second method explicitly uses the Neanderthal sequence by filtering putative haplotypes on Neanderthal match *p-value*, as was done for Fig. 1E. For comparison to the SVM results, the second method was applied to the subset of the data used in Fig. S5, and these results are shown in Fig. S5. Using the Neanderthal p-value method, we recover 240 Mb at FDR 5%. Thus we estimate that using only S*, if a higher FDR is tolerated, it is possible to identify ~30% (74 Mb / 240 Mb) of the total Neanderthal sequence identified using Neanderthal p-values (Fig. S5).

Explanatory variables in SVM analysis include: adj_500k (the number of putative introgressed sequences within 500kb of this window), remaining_bases (the number of bases in the 50kb window remaining after all filters are applied), recomb (recombination rate), num_snps (number of variants in the 20 target and 13 reference individuals), num_tag_snps (the number of variants identified as putatively introgressed [tag snps]), num_runs (the number of runs [out of five] in which this haplotype was identified as putatively introgressed), num_match_in_subset (the number of individuals [out of 20] in which a similar putatively introgressed haplotype was identified), *S* p-value*, pct_tag_snps_on_hap (percentage of tag snps on the same computationally imputed haplotype), tag_freq (mean tag snp frequency), tag_stdfreq (sd(tag_freq)), freq (frequency of all variants on the putatively introgressed haplotype), stdfreq (sd(freq)), hap_len (length of haplotype).

Bioinformatics analyses

For all analyses using genomic regions, either BEDOPS (38) or the R package GenomicRanges (39) were used. Where applicable, derived and ancestral state were inferred with respect to chimpanzee state in the Ensembl v64 EPO Compara 6 primate alignment (40). Variant annotations were obtained using the SeattleSeq pipeline. Coding and transcribed regions were obtained from UCSC Genome Browser (41) using the RefSeq database and refFlat table (42).

Logistic regression

To formally assess heterogeneity in observed distribution of inferred Neanderthal lineages across the genome, we performed multiple logistic regression. Briefly, the i^{th} 50kb window was coded as y_i = 1 if it overlapped an introgressed lineage and 0 otherwise. We then fit a logistic regression model of the form $y = mean + chromosomal\_arm + \sum covariates + error$. As additional covariates, we considered divergence between modern humans and Neanderthals (i.e., Neanderthal non-reference bases / unmasked bases), total number of unmasked bases, percentage coding (coding bases / unmasked bases), percentage transcribed, and percentage conserved bases (UCSC tracks phastConsElements46way, phastConsElements46wayPlacental, phastConsElements46wayPrimates). Model fits and significance of individual predictors were assessed with likelihood ratio tests. The best model included the predictor terms chromosomal_arm and divergence, although the effect size estimates for the former were very similar regardless of other covariates included in the model. Interestingly, when only percentage coding bases was used, it was a significant negative predictor. However this term was no longer significant with the inclusion of chromosome_arm

and neand_divergence. There was no difference between the models chromosome_arm + neand_divergence and chromosome_arm + neand_divergence + pct_cds. To calculate odds ratios for each chromosomal arm, we used chromosome 1p as a reference category. To be concrete, if the odds ratio is *x*, the interpretation is that the odds of finding introgressed sequence at this chromosomal region is *x* times the odds of finding introgressed sequence on chromosome 1p. Note, the selection of a reference group is arbitrary, and does not influence the estimated statistical significance of the predictor "chromosomal_arm". All regression models were fit in R with the function glm() using the family=binomial("logit") option. To explore the robustness of these results, we repeated the regression analyses using different FDR thresholds for calling introgressed sequence (Fig. S9) and for each population individually (Fig. S10).

Neanderthal fixed differences

Enrichment for Neanderthal sequence is negatively correlated with fixed differences between modern humans and Neanderthal (Fig. 2C), and this observation is consistent in both East Asians and Europeans when the data is analyzed separately (Fig. S11). This relationship is partially a function of the presence of Neanderthal lineages in modern humans. Thus, highly introgressed regions of the genome would be expected to contain fewer fixed differences. To control for the presence of Neanderthal lineages, for each chromosomal arm we also counted fixed differences specifically in non-introgressed sequence, divided by the amount of non-introgressed sequence. By dividing specifically by the amount of non-introgressed sequence, we normalize for the actual amount of non-introgressed sequence in a given region. As expected, non-introgressed sequence contains 6.1x more fixed differences (FD) per Mb than introgressed sequence (17.3 vs 2.8 FD/Mb, respectively). The odds ratio for chromosomal arms remains significantly negatively correlated ($\rho = -0.61$, *p-value* = 2.84×10^{-5}) with fixed differences when

correcting for the presence of Neanderthal sequence (Fig. S12). Additionally, when fixed differences are included in the logistic regression analysis above, they are a significant negative predictor of introgressed status.

To investigate the relationship between introgression and divergence between Neanderthals and modern humans at a finer scale, we also calculated fixed differences for non-introgressed sequence in 5Mb windows. Windows enriched for fixed differences (>17.3 FD/Mb) were significantly depleted for introgression (Fig. S13, Wilcoxon rank-sum test, *p-value* < 2.2 x 10^{-16}).

Inference of demographic history

We observed that East Asians contained more introgressed sequence than Europeans. To test the hypothesis that this imbalance is possible under a single, ancestral pulse of introgression, we simulated 2,000 random demographic models, using the specifications given in "Expected fraction of the genome covered by surviving Neanderthal lineages" – with the exception that we simulated only the first pulse of introgression. We refer to this as a "single pulse" or "one pulse" model. For each set of simulations, we calculated two summary statistics: the ratio of introgressed sequence per individual for East Asians and Europeans (Fig. 3, Fig. S14, x-axis), and the ratio of total introgressed bases "covered" by introgressed sequence for East Asians and Europeans (Fig. 3, Fig. S14, y-axis). By using the ratio of introgressed sequence between populations, as opposed to absolute numbers for each population, we compensate for the fact that we are able to recover only a fraction of the total introgressed sequence per individual.

To compare the simulated values of these summary statistics to the observed values, we first had to correct the observed values for sample size differences. Specifically, for each non-

overlapping 50kb window (60kb step) we randomly selected 20 individuals from each population, and calculated the summary statistics for this set of individuals and windows. We performed this process for 14 sets of 20 individuals (the maximum number of sets of 20 individuals possible without reusing any East Asian individuals), and again shifting all windows +20kb and +40kb (eventually considering all 50kb windows on a 20kb step). This entire process was repeated five times. Confidence intervals were obtained by sampling with replacement from the resulting set of 14*20*3*5 = 4,200 values. The adjusted summary statistics are: ASN/EUR percent introgressed per individual: 1.21; ASN/EUR proportion of genome covered by introgressed sequence: 1.05 (Fig. 3, Fig. S14, white boxes).

When compared to our simulations, the observed summary statistics do not overlap with the simulated statistics from any single pulse model; thus, our observations are incompatible with the single pulse models considered here (Fig. 3, Fig. S14).

We next attempted to a) determine if the observed statistics were compatible with a model in which East Asians received an additional small percentage of Neanderthal gene flow after the split of Europeans and East Asians – a "two pulse" model, and b) if so, determine the amount of additional migration (and Neanderthal sequence received). To this end, we simulated 5,000 random models as before, including a second pulse into East Asians starting 500 years after the split of East Asians and Europeans (details in "Expected fraction of the genome covered by surviving Neanderthal lineages"). Summary statistics were calculated as with the single pulse simulations (Fig. 3, Fig. S14). By comparing the observed summary statistics and the simulated statistics for a two pulse model, we find that there is considerable overlap, suggesting that a two pulse model is consistent with the observed data.

We then identified simulations that most closely matched the summary statistics of the observed data (Fig. S15). The demographic parameters for these simulations can then be used to estimate the true demographic parameters, using Approximate Bayesian Computation (*37*). Although we varied many parameters (Fig. 3, Fig. S14), only two were strongly restricted by the summary statistics we measured: the ratio of ancestral effective population sizes between Europeans and East Asians, and the ratio of amount of introgression between the second and first pulse (Fig. 3, Fig. S15). Using the R package "abc" and the neural network method of adjusting parameter values (*37*) we estimate m_2/m_1 to be 0.202, with a 75% CI of {0.180-0.220} and 95% CI of {0.134-0.271}, and N_e^{EUR}/N_e^{ASN} to be 1.29, with a 75% CI of {1.24-1.36} and 95% CI of {1.15-1.57}.

Detecting signatures of adaptive introgression

To identify putative substrates of adaptive introgression, we pursued two complimentary strategies. First, to identify introgressed sequence that experience geographically restricted selection, we performed a simple genome-wide scan by calculating F_{ST} on *S** tag SNVs (which quantifies allele frequency differences between populations) as: $F_{ST} = 1 - \frac{H_S}{H_T}$, where H_S and H_T are sample size corrected subpopulation and total heterozygosity, respectively. We restricted this analysis to *S** tag SNVs where the introgressed variant and Neanderthal allele were both derived. We identified 4 haplotypes with >3 variants that were outliers with an $F_{ST} > 0.40$ (99.9^{th} percentile of the empirical distribution of F_{ST}). In total, these four haplotypes encompass 28 variants. To determine how unusual these observations were under neutral demographic models, we performed extensive coalescent simulations using 18 different admixture models that most closely match summary statistics of the observed data (see section above "Inference of

demographic history"). The reported *p-values* in the main text correspond to F_{ST} calculated on 3,453,256 *S** tag SNVs identified in simulated introgressed sequences.

Second, to identify putative shared signatures of selection common to both East Asians and Europeans, we identified putatively introgressed regions that were common in both populations (90% trimmed mean of tag variants that match Neanderthal alleles and are derived). We found that a derived allele frequency greater than 40% in both East Asians and Europeans was very unusual in 200 random likely admixture models (*p-value* < 10^{-4}). We functionally annotated variants that occur in regions harboring putative signatures of adaptive introgression by intersecting their positions with conservation and functional genomics. Specifically, we considered PhastCons conserved elements inferred from 46 way alignments of placental mammals (UCSC track "phastCons46wayPlacental"), H3K27Ac marks (UCSC track "Layered H3K27Ac"), DNaseI hypersensitive sites (UCSC tracks "UW DNaseI HS" and "Duke DNaseI HS"), and chromatin state segmentation calls (UCSC track "Broad ChromHMM" calls of "Enhancer").

The following ms command line argument was one of several used to count the number of expected haplotypes at high frequency in both populations, using the length and recombination rate for the region at chr12:52880370-52929370. This command samples very few variants – this is because, for this analysis, we discard all variation, and use only the coalescent trees generated by –T. Additionally, in this command we use a split time of 700kya for Neanderthals and modern humans. Because we discard all variation, this does not affect the analysis (rather, it assists in the detection of introgressed coalescent trees, as discussed in the section "Expected fraction of the genome covered by surviving Neanderthal lineages"):

```
ms 1333 500 -s 15 -r 8.28921759801 44603 -I 4 2 758 572 1 0 -n 4
2.051984e-01 -n 1 58.002735978 -n 2 70.041039672 -n 3 187.55 -eg
0 1 482.46 -eg 0 2 570.18 -eg 0 3 720.23 -em 0 1 2 0.7310 -em 0
2 1 0.7310 -em 0 1 3 0.228072 -em 0 3 1 0.228072 -em 0 2 3
0.909364 -em 0 3 2 0.909364 -eg 0.006997264 1 0 -eg 0.006997264
2 2.089166e+01 -eg 0.006997264 3 3.006376e+01 -en 0.006997264 1
1.98002736 -en 0.031463748 2 7.774282e-01 -en 0.031463748 3
5.820793e-01 -ej 5.453352e-02 3 2 -en 5.453352e-02 2 7.774282e-
01 -em 5.453352e-02 1 2 4.386 -em 5.453352e-02 2 1 4.386 -ej
8.207934e-02 2 1 -en 8.207934e-02 1 1.98002736 -en 0.20246238 1
1 -em 6.566895e-02 2 4 4.386000e+01 -em 6.635294e-02 2 4 0 -em
5.316553e-02 3 4 8.832178e+00 -em 5.384952e-02 3 4 0 -ej
9.575923e-01 4 1 -L -T -seeds 17709 13602 2369
```

Supplementary Tables

Table S1. Summary of 1000 Genomes Project samples.

Population	Description	Number Individuals
TSI	Toscani in Italia	98
IBS	Iberian population in Spain	14
CEU	Utah Residents (CEPH) with Northern and Western European ancestry	85
GBR	British in England and Scotland	89
FIN	Finnish in Finland	93
CHS	Southern Han Chinese	100
CHB	Han Chinese in Bejing, China	97
JPT	Japanese in Tokyo, Japan	89

Tables S2-S8. Worked S* example data – can be found in Supplemental S* Worked Example.

Table S9. Amount of unique sequence covered by Neanderthal lineages as a function of *S threshold.**

***S** p-value*	FDR	ASN Mb	EUR Mb	All Mb	Percent present in ***S** p-value* ≤ 0.01 set[a]
0.0001	0.05	106.1	114.4	184.4	98.8%
0.0005	0.05	163.1	181.3	285.2	99.0%
0.001	0.05	198.5	218.9	343.2	98.9%
0.005	0.05	301.1	334.5	514.6	99.2%
0.01	0.05	354.8	393.9	600.1	NA

[a] This column describes the percent of sequence in the "All Mb" column at this threshold that overlaps with the set of sequences detected at *S** *p-value* ≤ 0.01. For example, there are 184.4 Mb of putatively introgressed sequence detected at *S** *p-value* ≤ 0.0001 and FDR = 5%, of which 181.1 Mb overlap with the set of sequences identified at *S** *p-value* ≤ 0.01 and FDR = 5%.

Table S10. Summary of population specific signatures of adaptive introgression.

Chr	Coordinates	Length (kb)	Number SNPs $F_{ST} \geq 0.40$	Average Freq ASN[a]	Average Freq EUR[a]	Genes[b]
1	232,603,040 - 232,643,331	40.3	9	0.569	0.008	*SIPA1L2*
4	38,424,899-38,548,737	123.8	4	0.580	0.009	*Intergenic*
9	16,720,121-16,786,930	66.8	10	0.00	0.691	*BNC2*
11	120,154,630 - 120,178,414	23.8	5	0.639	0.003	*POU2F3*

[a] Average frequency denotes the mean derived allele frequency for variants with $F_{ST} \geq 0.40$.

[b] A gene name is listed if any of the highly differentiated introgressed variants overlaps a gene.

Table S11. Summary of signatures of adaptive introgression shared between East Asians and Europeans.

Chr	Coordinates	Length (kb)	Genes
4	52,886,169-52,969,221	83.1	*SGCB, SPATA18*
7	110,141,435-110,209,984	68.5	*Intergenic*
10	95,422,244-95,466,847	44.6	*PDE6C, FRA10AC1*
12	52,880,370-52,929,370	49.0	*KRT6A, KRT5*
12	121,842,267-121,908,509	66.2	*RNF34, KDM2B*
15	47,615,219-47,646,349	31.1	*SEMA6D*

Supplementary Figures

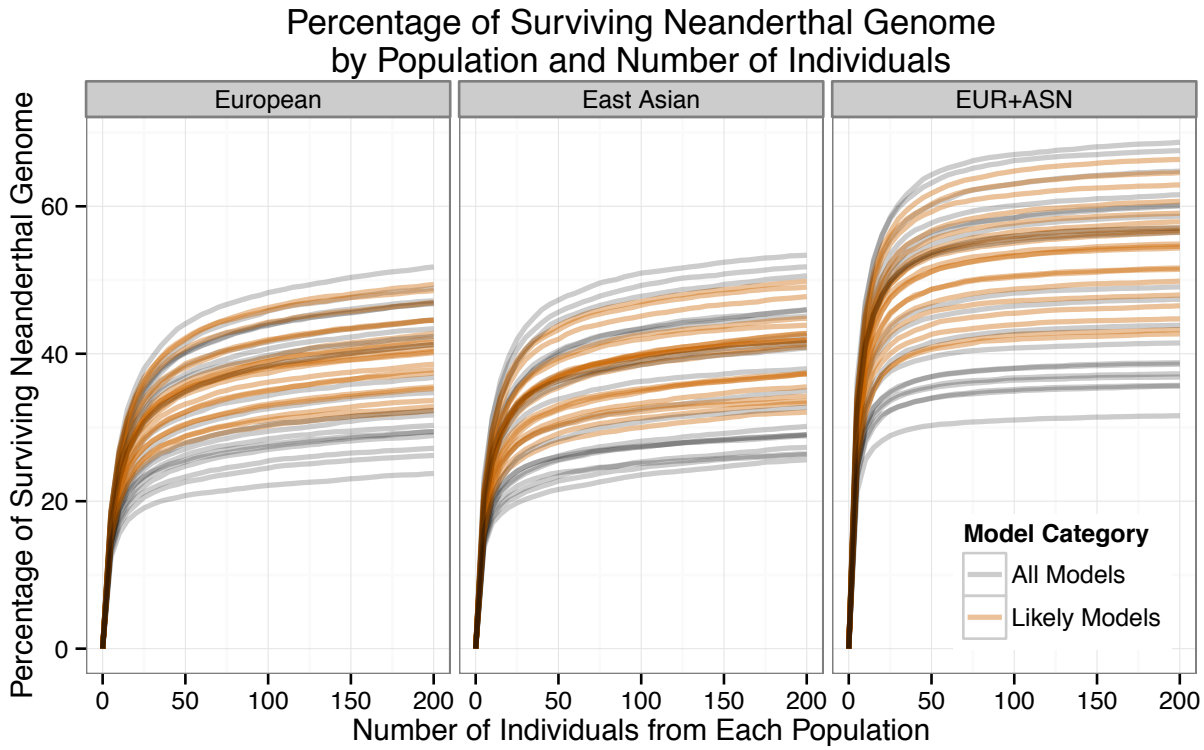

Figure S1. Expected fraction of the genome covered by surviving Neanderthal lineages, as a function of the number of sampled individuals. The expected fraction of the Neanderthal genome present in simulated sequence data for up to 200 Europeans and 200 East Asians, under 40 different admixture models. “Likely” models are those with demographic parameters within the ranges estimated in Figure 3.

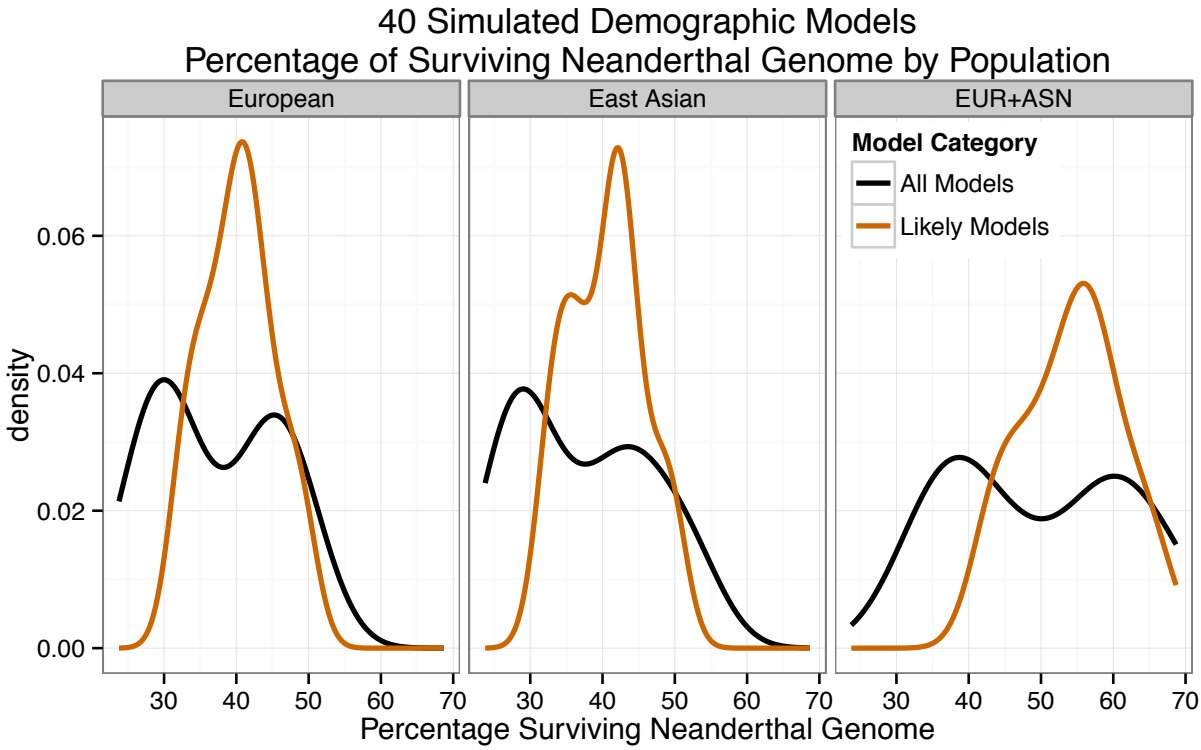

Figure S2. Expected fraction of the genome covered by surviving Neanderthal lineages. The expected fraction of the Neanderthal genome present in simulated sequence data for 200 Europeans and 200 East Asians, under 40 different admixture models. “Likely” models are those with demographic parameters within the ranges estimated in Figure 3.

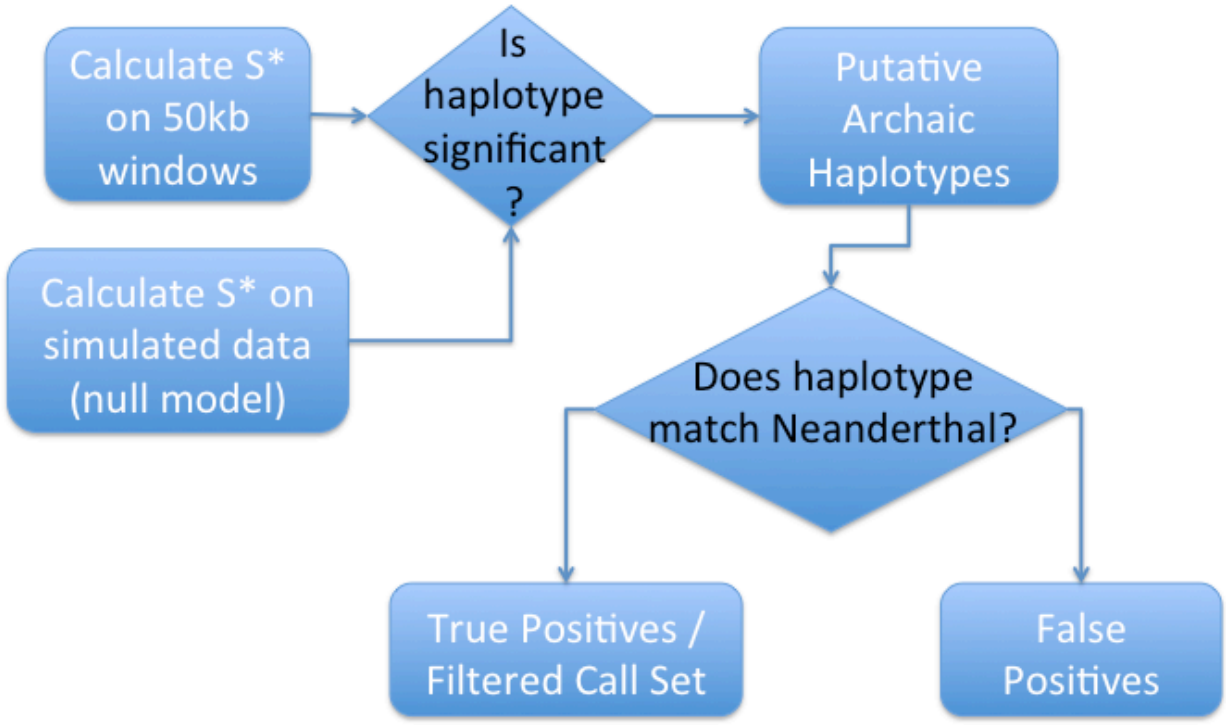

Figure S3. Schematic of computational strategy to identify introgressed lineages.

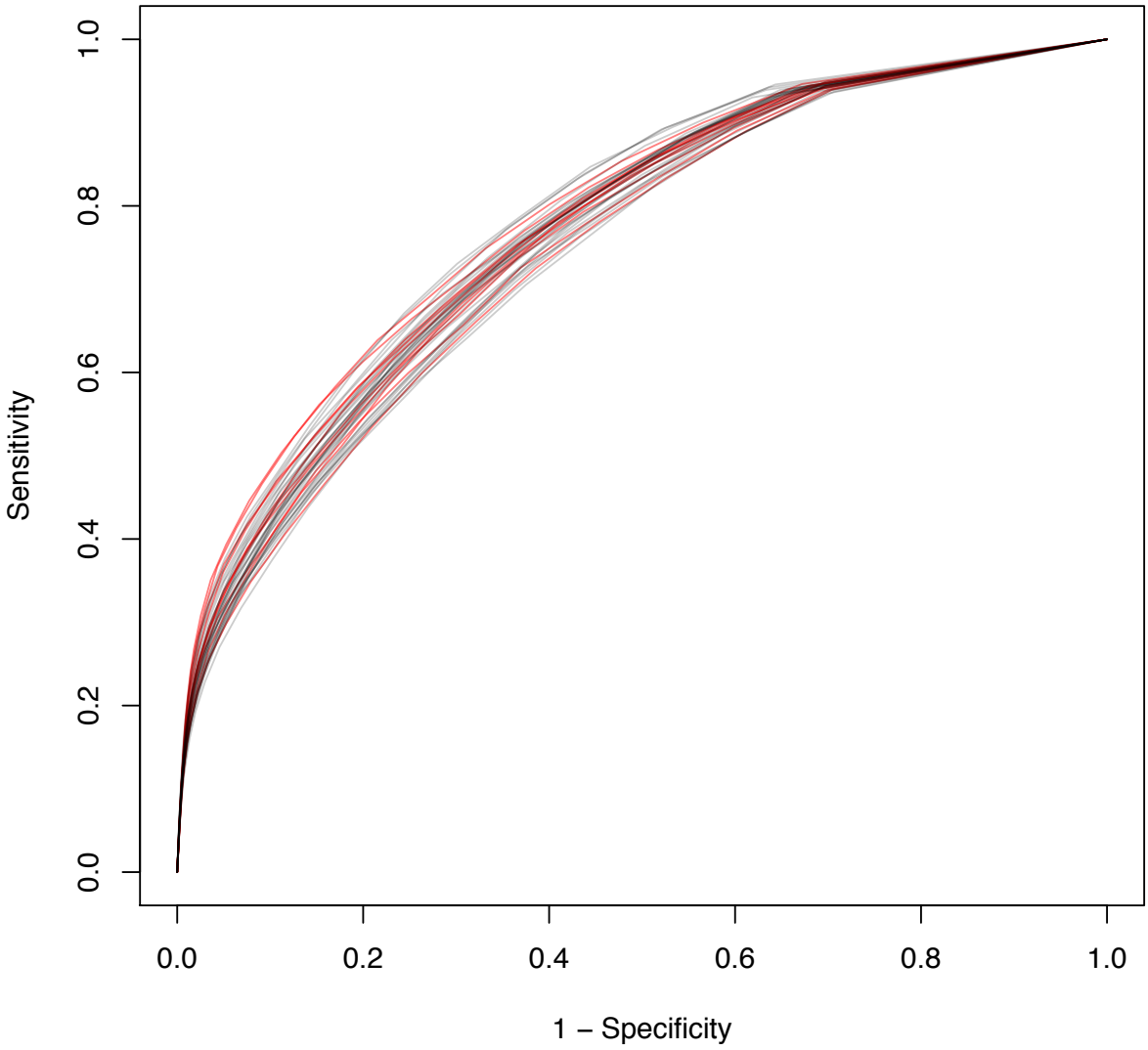

Figure S4. ROC curves for *S for different demographic models.** ROC curves for "Likely" models are shown in red.

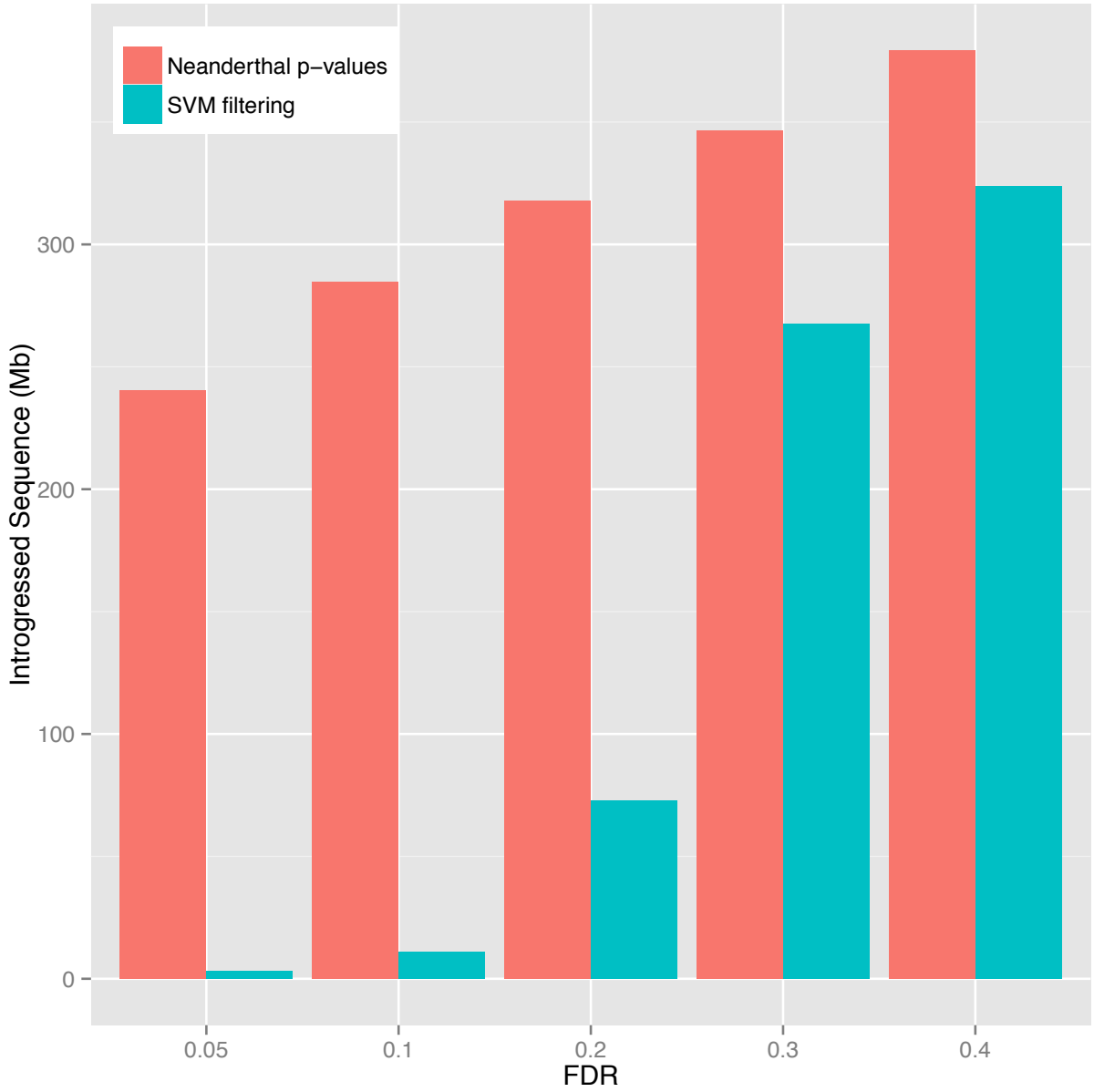

Figure S5. FDR vs Mb of introgressed sequence recovered for two methods of identifying introgressed Neanderthal sequence: an SVM based method that is agnostic to the Neanderthal genome sequence, and our presented method of filtering by Neanderthal *p-values*. Analysis was performed on a subset of *S* p-value* ≤ 0.01 haplotypes. For this test subset of sequences, we recover 74 Mb of sequence at FDR = 20% using the SVM. Using the Neanderthal p-value method, we recover 240 Mb at FDR 5%. Thus we estimate that using only S*, if a higher FDR is tolerated, it is possible to identify ~30% (74 Mb / 240 Mb) of the total Neanderthal sequence identified using Neanderthal p-values

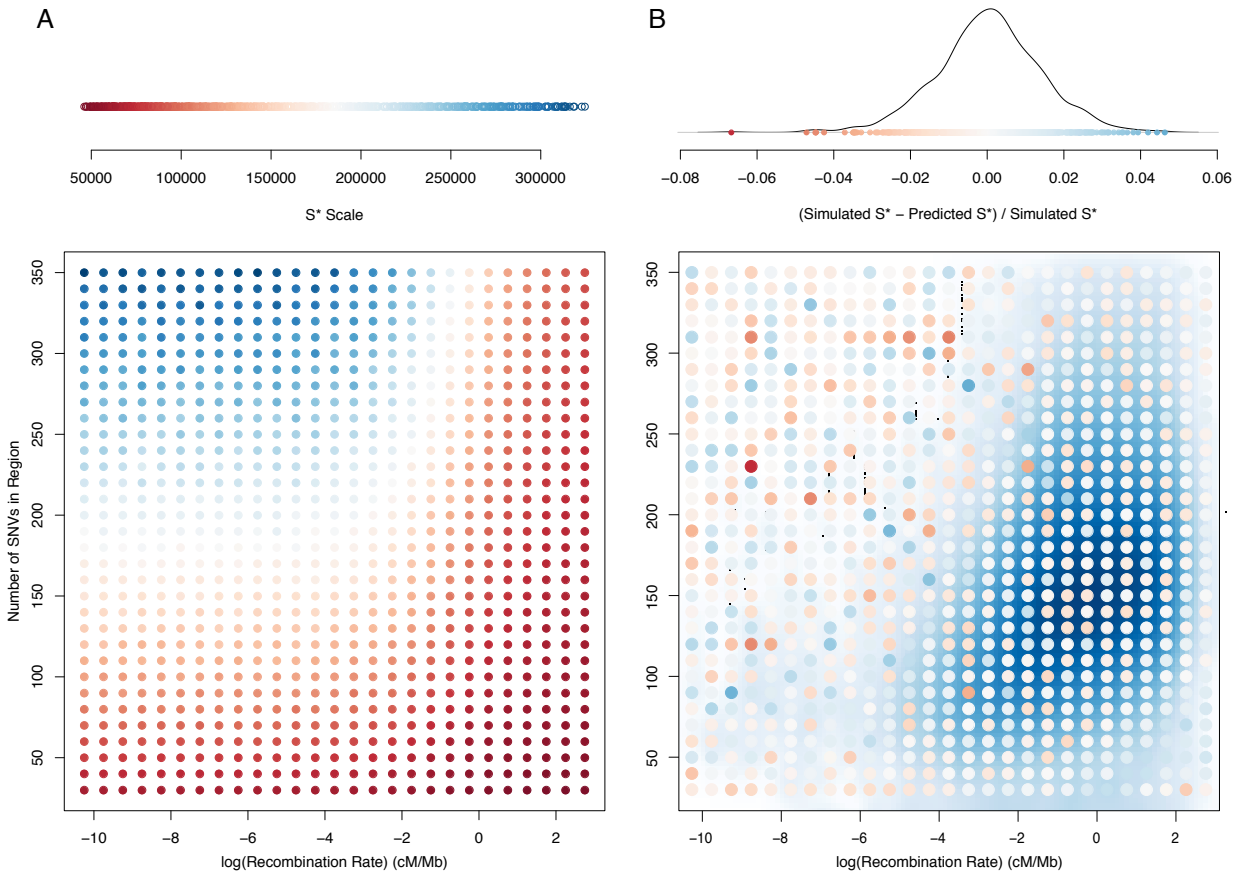

Figure S6. Computationally efficient strategy to estimate *S statistical significance.** A) *S* p-value* $\leq$ 0.01 thresholds as determined by 20,000 simulations for each of 864 grid points. B) Residuals from fitting the thresholds in (A) to a generalized linear model (round points), overlaid on a 2D density plot of the observed recombination rate and number of variants for all 50kb windows (blue cloud background).

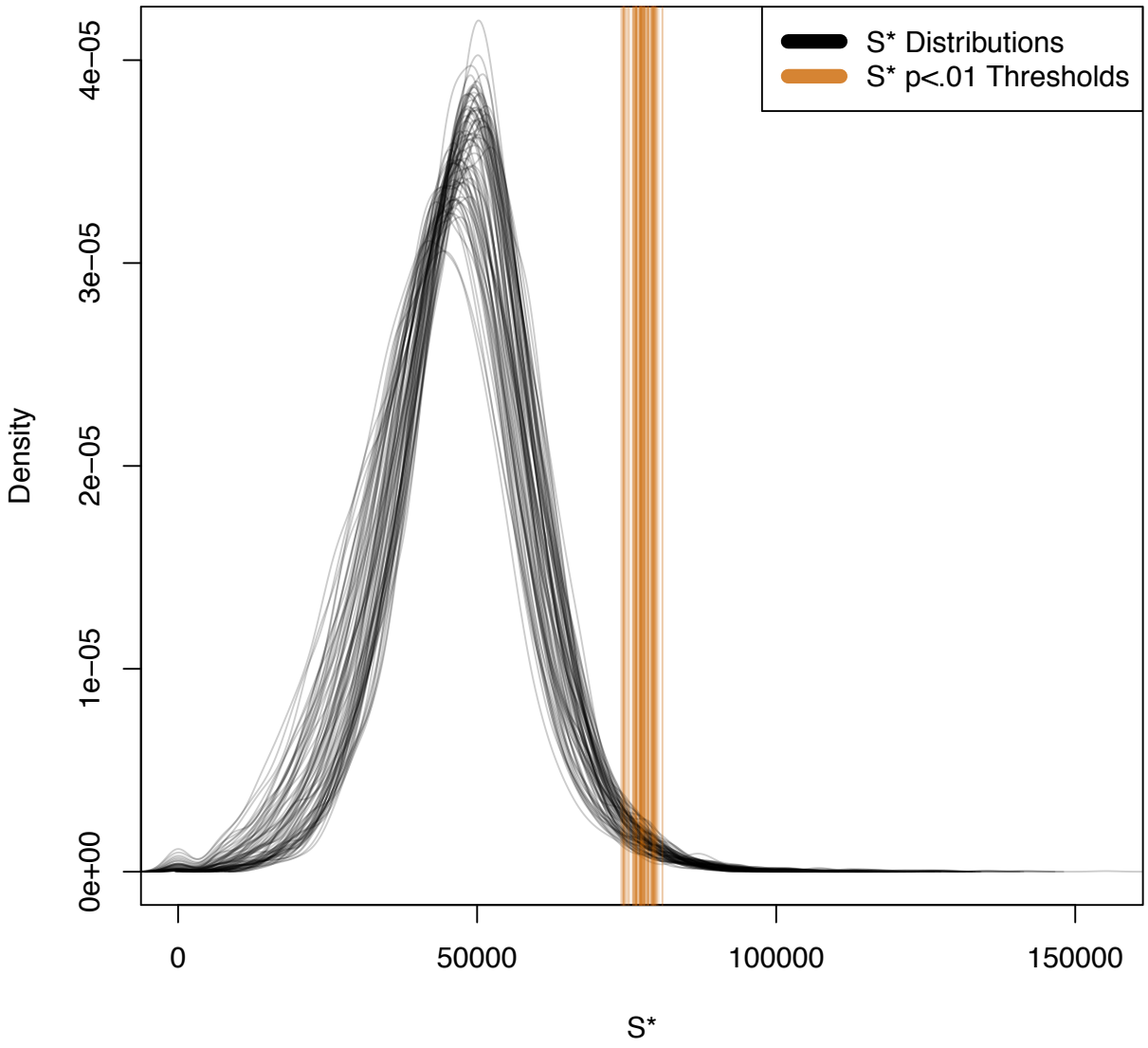

Figure S7. *S* p-values* are robust to demographic uncertainty. *S** null distribution for 10,000 50kb regions of simulated sequence data under ~700 different demographic models of modern human history (black lines). Orange lines represent *S* p-value* ≤ 0.01 thresholds for each model.

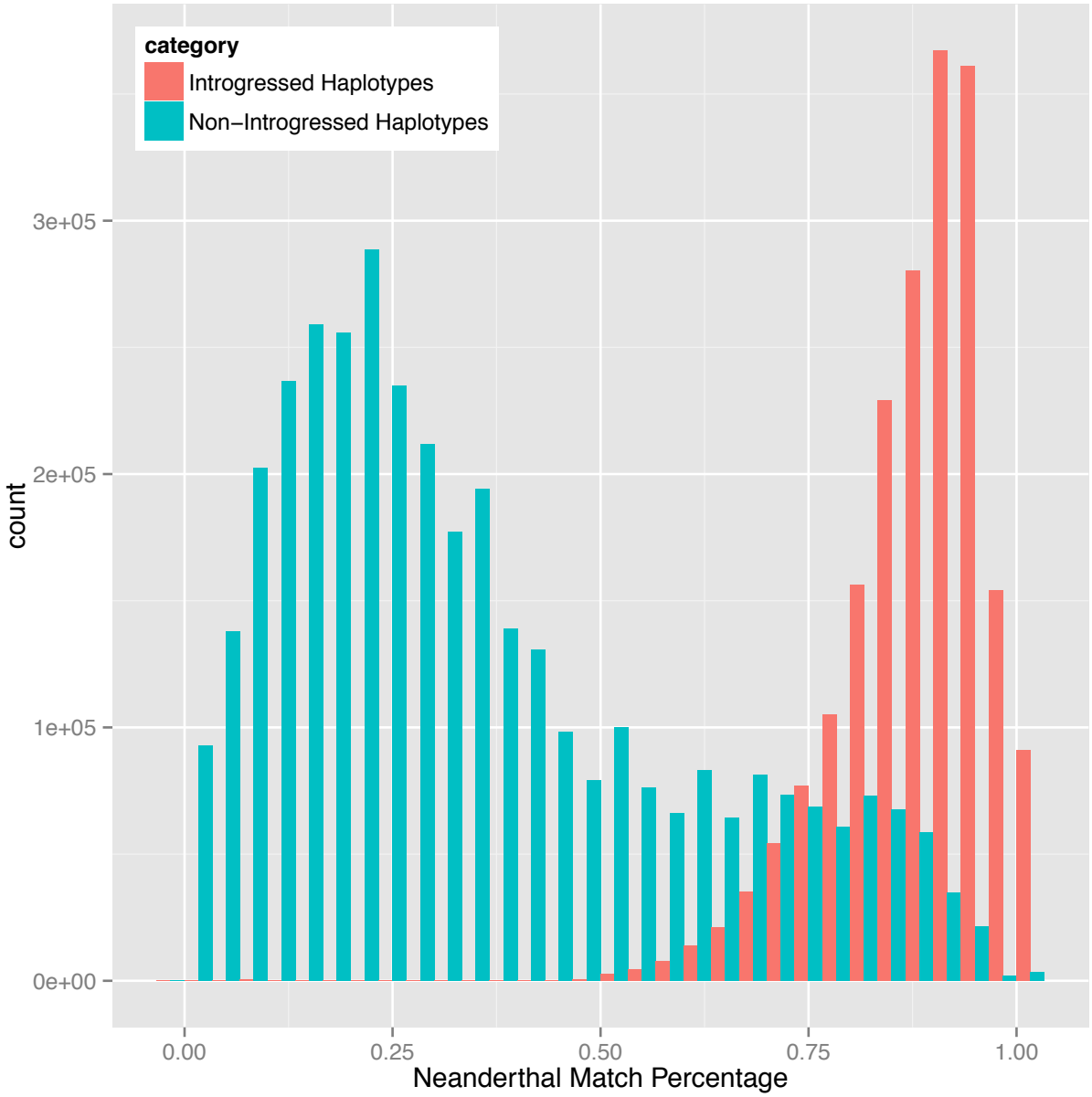

Figure S8. Neanderthal match percentage, used in the calculation of Neanderthal *p-values.* Distribution of match percentage is shown for haplotypes in our FDR 5% call set (red) and all other haplotypes in the *S* p-value* ≤ 0.01 call set (blue).

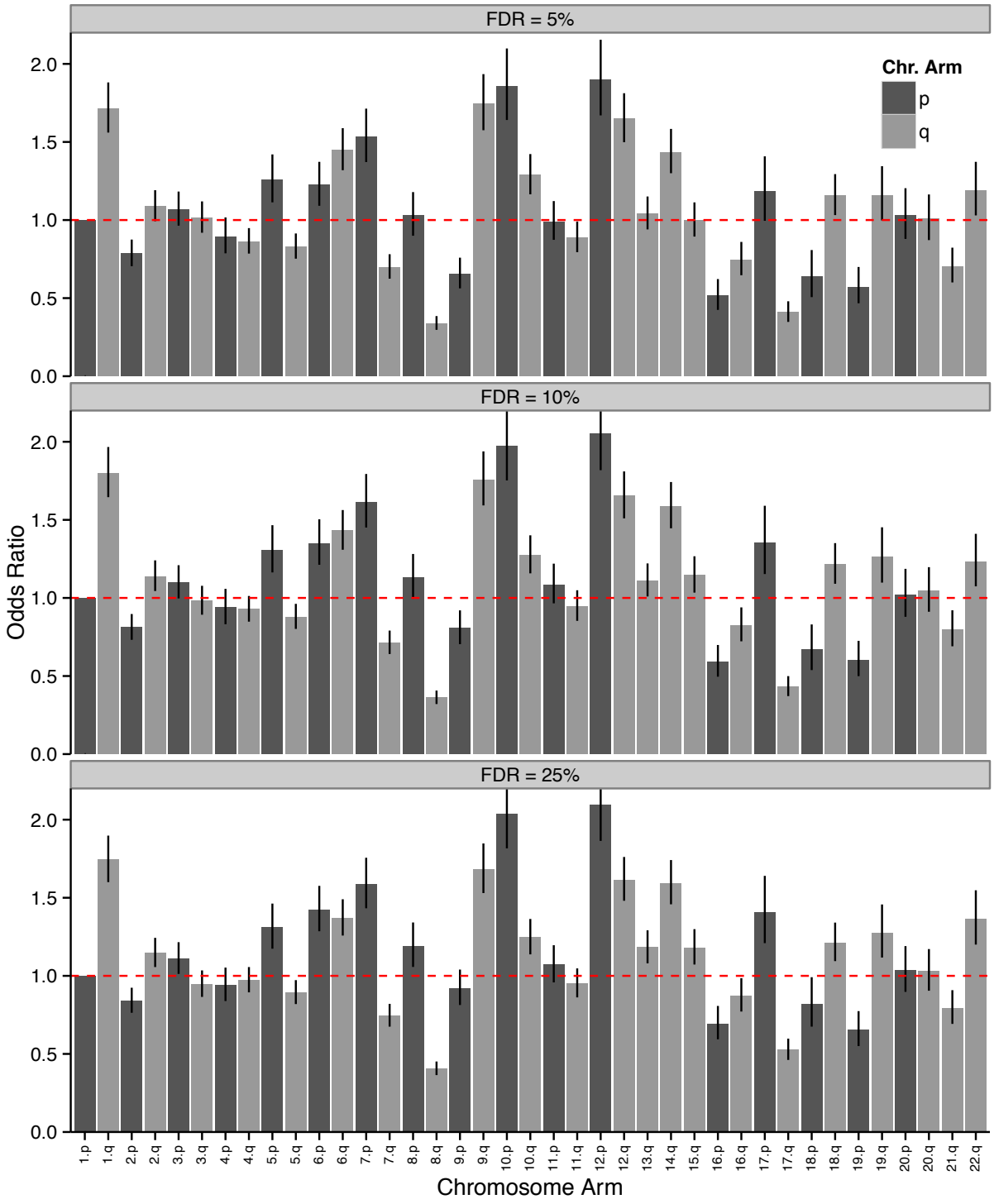

Figure S9. Effect size estimates for chromosomal odds ratios are robust to thresholds used to define introgressed sequence. Odds ratios for FDR = 5%, 10%, and 25% are shown from top to bottom, respectively.

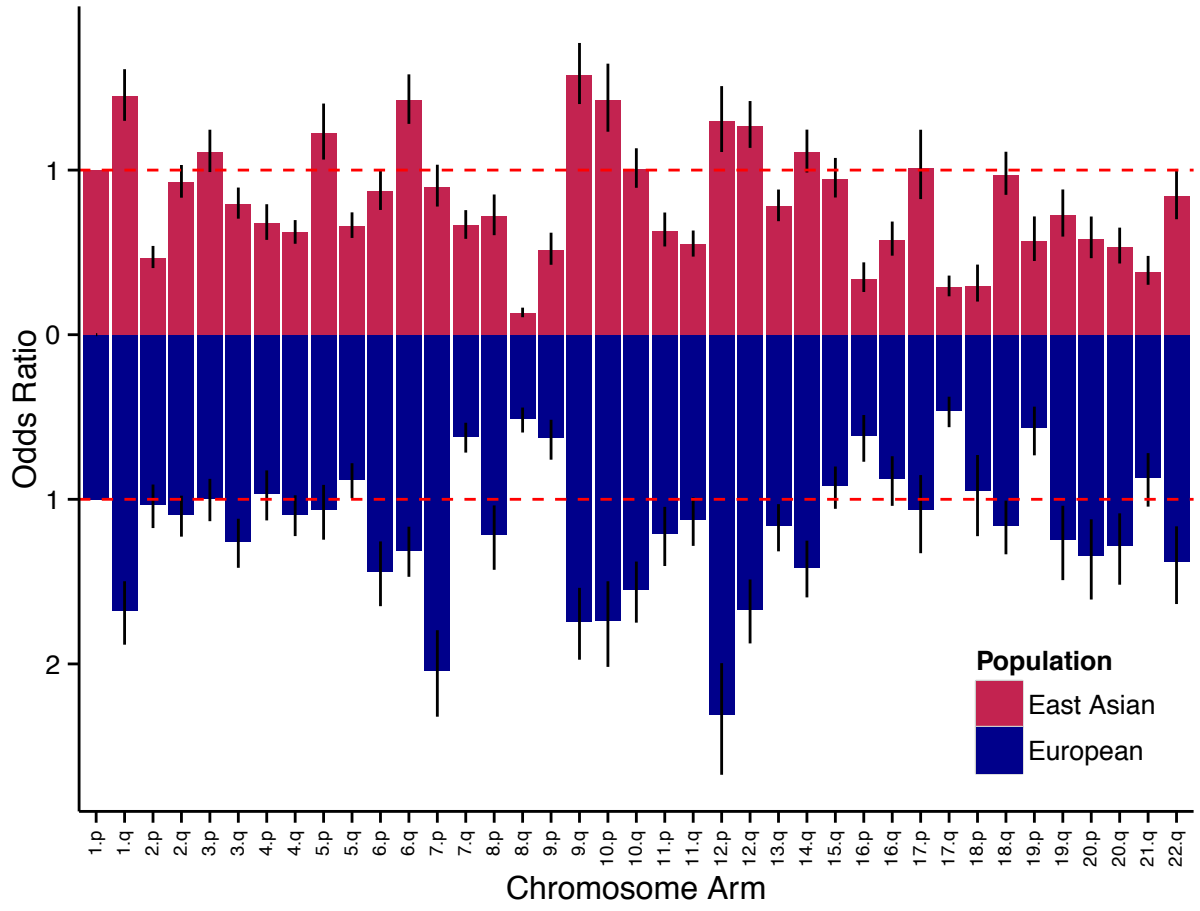

Figure S10. Effect size estimates for chromosomal arms are similar in East Asians and Europeans.

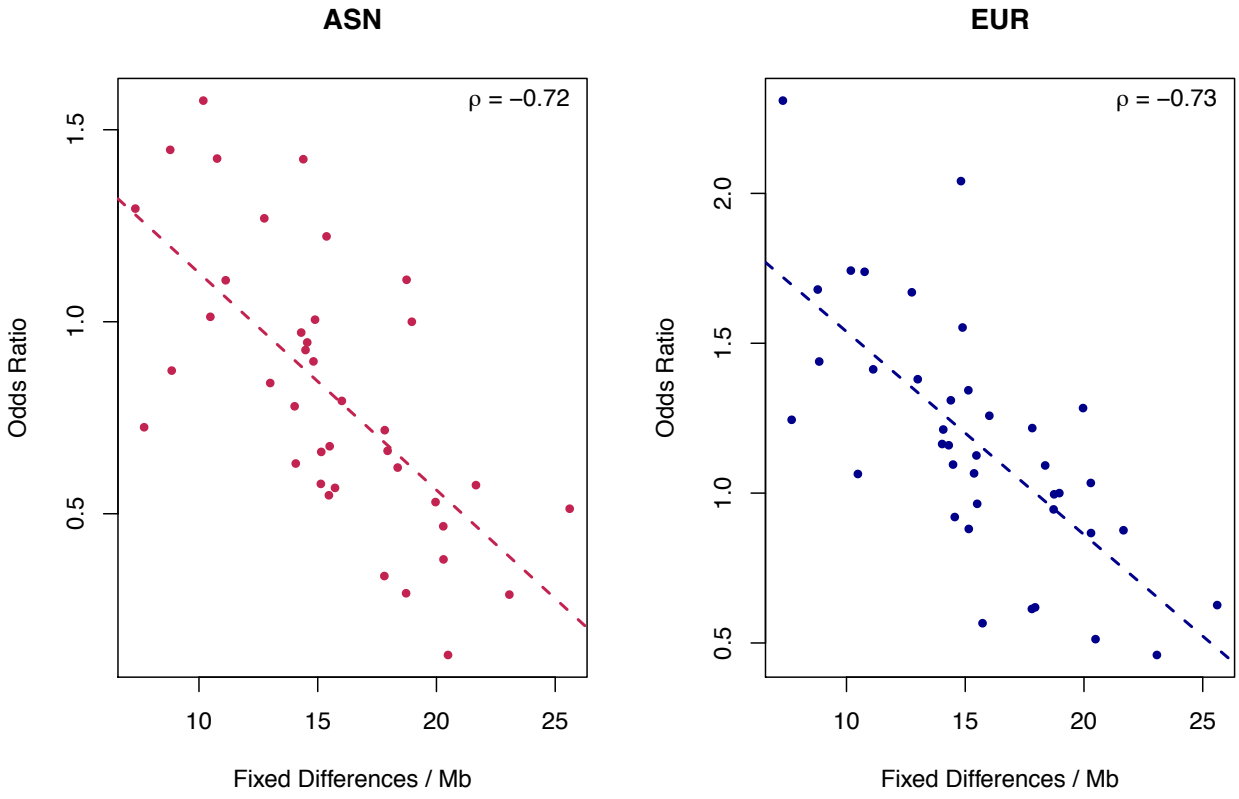

Figure S11. Scatterplots of Odds Ratios (OR) for chromosomal arms versus the number of fixed differences per Mb in East Asians (ASN) and Europeans (EUR). For each population, Spearman rank correlations are shown with a dashed line. *P-values* for all correlation coefficients are less than 10^{-4}.

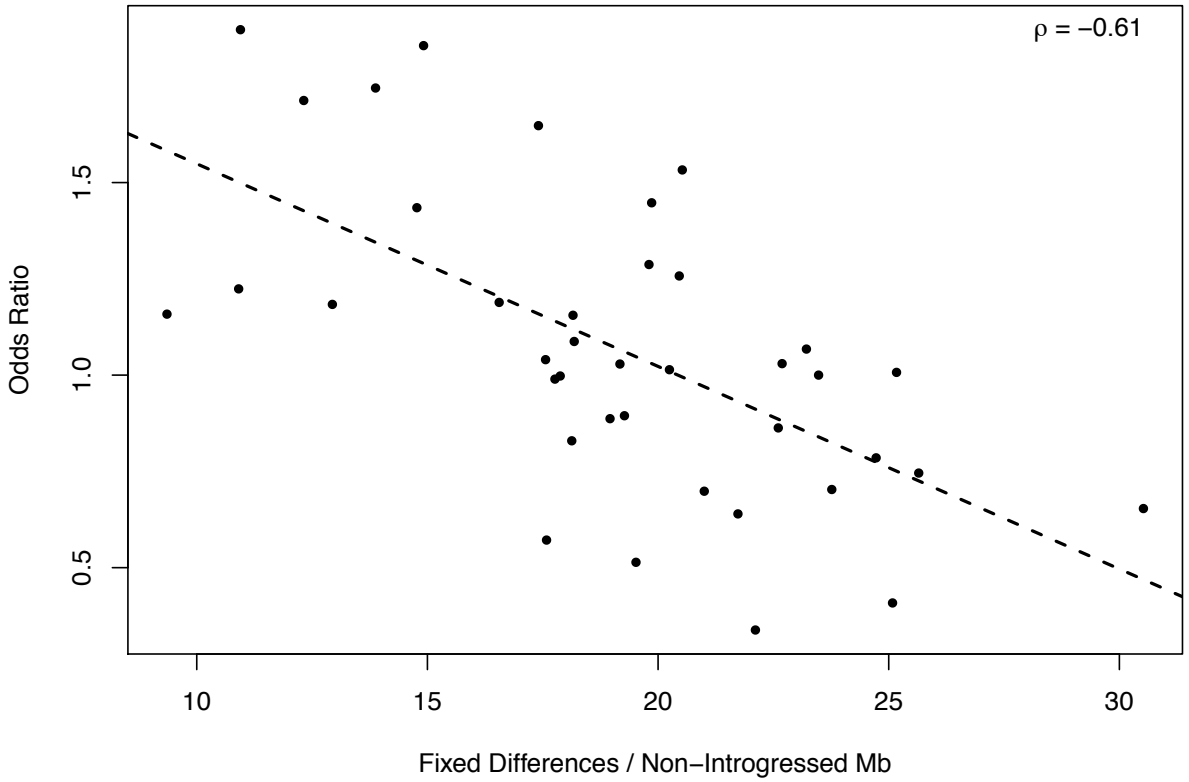

Figure S12. Scatterplot of odds ratio for chromosomal arms versus the number of fixed differences per Mb when correcting for the presence of introgressed lineages. Spearman's rank correlation is shown (dashed line; *p-value* < 10^{-4}).

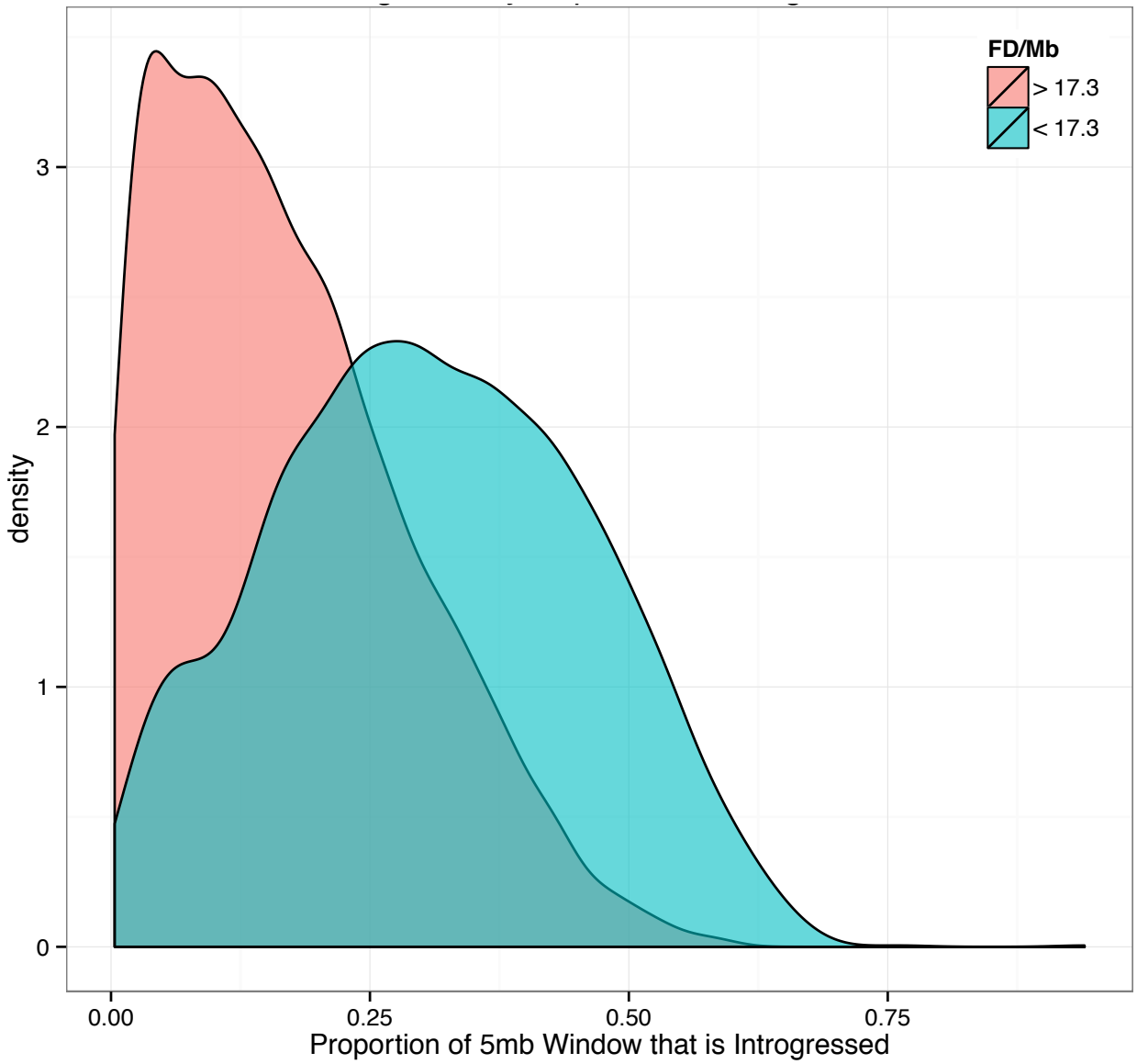

Figure S13. Chromosomal regions with more fixed differences are less likely to carry introgressed sequences. We divided all 5mb windows (1Mb step) into those with above and below average (17.3) fixed differences per Mb. Only non-introgressed portions of each 5mb window were considered, to control for the presence of Neanderthal haplotypes in introgressed portions of the windows. Regions with more fixed differences are significantly depleted for Neanderthal introgression.

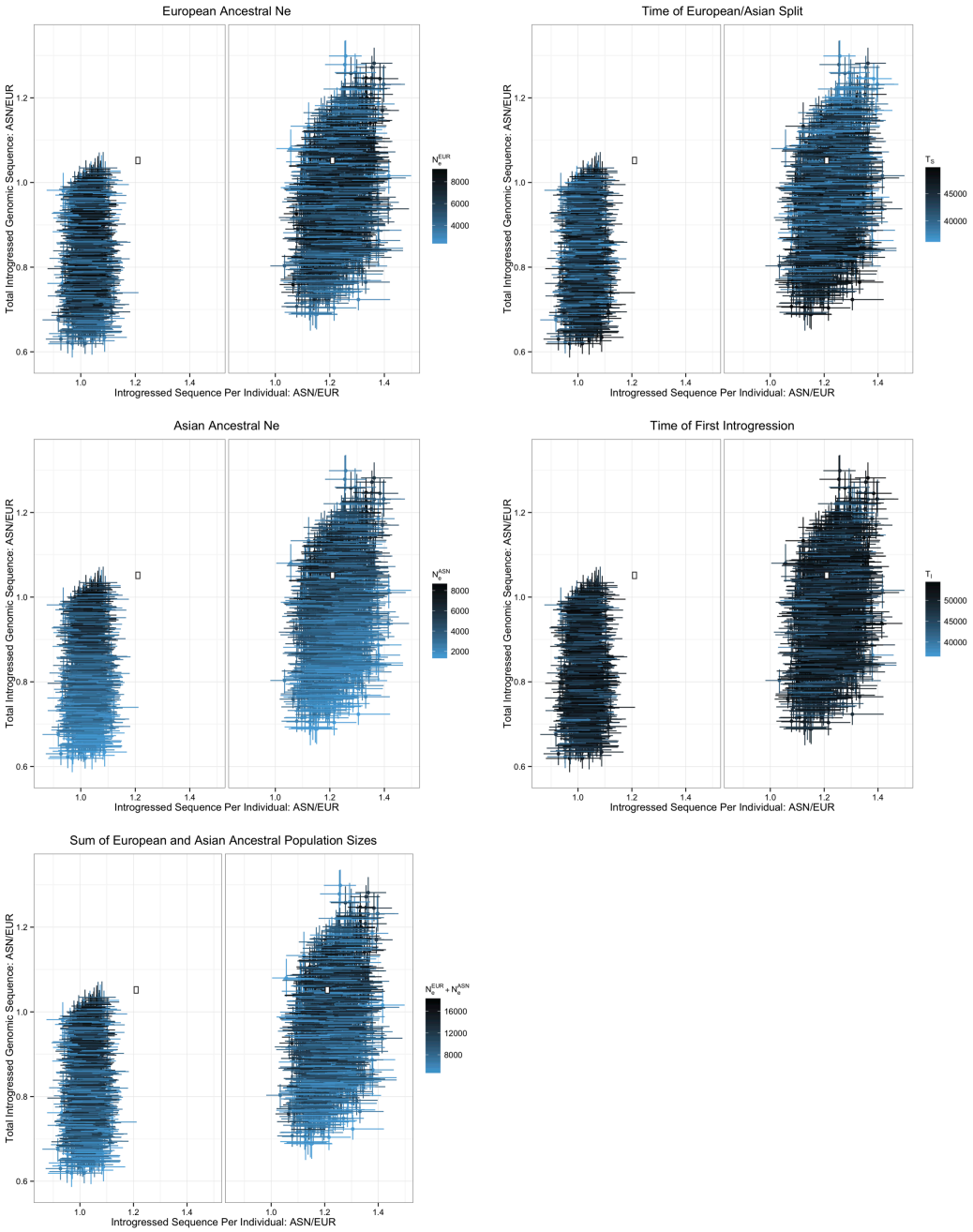

Figure S14. The effects of five different model parameters on the distribution of summary statistics in 2000 simulations. Summary statistics are the ratio of introgressed sequence per individual for East Asians and Europeans (x-axes), and the ratio of total introgressed bases "covered" by introgressed sequence for East Asians and Europeans (y-axes). Both one pulse and two pulse models are considered (left cloud and right cloud in each figure). For each simulation, 95% CI are given. White boxes are the observed values of the summary statistics.

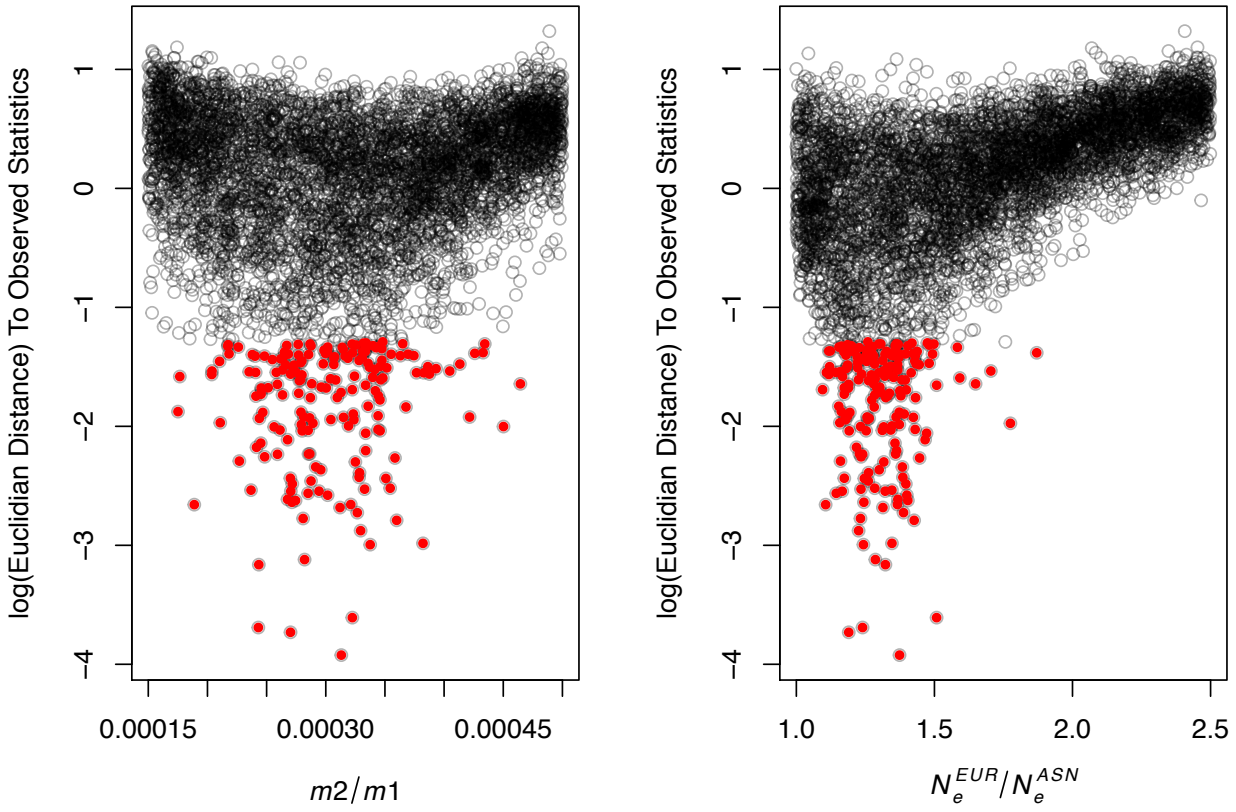

Figure S15. Estimating demographic parameters. Distance of 6,000 simulations to observed summary statistics, by model parameters shown in Fig. 3. The plot on the left corresponds to the ratio of amount of introgression between the second and first pulse and the plot on the right is for the ratio of ancestral effective population sizes between Europeans and East Asians. The top 3% of simulations were selected for estimating the true values of these parameters (red points).