Archaic adaptive introgression in *TBX15/WARS2*

Fernando Racimo[1*], David Gokhman[2], Matteo Fumagalli[3], Torben Hansen[4], Ida Moltke[5], Anders Albrechtsen[5], Liran Carmel[2], Emilia Huerta-Sánchez[6], Rasmus Nielsen[1,7*]

Affiliations:
[1] Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA.
[2] Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, Israel.
[3] Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, UK.
[4] The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark.
[5] The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark.
[6] School of Natural Sciences, University of California Merced, Merced, CA 95343, USA.
[7] Department of Statistics, University of California Berkeley, Berkeley, CA 94720, USA.
*Correspondence to: FR (fernandoracimo@gmail.com) and R.N. (rasmus_nielsen@berkeley.edu).

Abstract

A recent study conducted the first genome-wide scan for selection in Inuit from Greenland using SNP chip data. Here, we report that selection in the region with the second most extreme signal of positive selection in Greenlandic Inuit favored a deeply divergent haplotype introgressed from an archaic population most closely related to Denisovans. The region contains two genes, *WARS2* and *TBX15*, and has previously been associated with body-fat distribution in humans. We show that the adaptively introgressed allele has been under selection in a much larger geographic region than just Greenland. Furthermore, it is associated with changes in expression of *WARS2* and *TBX15* in multiple tissues including the adrenal gland and subcutaneous adipose tissue, and is associated with BMI-adjusted waist circumference. We also find that in this region both the Denisovan and individuals homozygous for the introgressed allele display differential DNA methylation.

Introduction

To identify genes responsible for biological adaptations to life in the Arctic, Fumagalli *et al.*[1] scanned the genomes of Greenlandic Inuit using the population branch statistic (PBS)[2], which detects loci that are highly differentiated from other populations. Using this method, they found two regions with a strong signal of selection. One region contains the cluster of *FADS* genes, involved in the metabolism of unsaturated fatty acids. Several of the SNPs with the highest PBS values in this region were shown to be significantly

associated with different phenotypes including fatty acid profile, HDL cholesterol, and height.

The other region contains *WARS2* and *TBX15*, located on chromosome 1. *WARS2* encodes the mitochondrial tryptophanyl-tRNA synthetase. *TBX15* is a transcription factor from the T-box family and is a highly pleotropic gene expressed in multiple tissues at different stages of development. It is required for skeletal development[3] and deletions in this gene cause Cousin syndrome, whose symptoms include craniofacial dysmorphism and short stature[4]. *TBX15* also plays a role in the differentiation of brown and brite adipocytes[5]. Brown and brite adipocytes produce heat via lipid oxidation when stimulated by cold temperatures, making *TBX15* a strong candidate gene for adaptation to life in the Arctic. SNPs in or near both of these genes have also been associated with numerous phenotypes in GWAS studies – in particular waist-hip ratio and fat distribution in Europeans[6] and ear morphology in Latin Americans[7].

Multiple studies have shown extensive introgression of DNA from Neanderthals and Denisovans into modern humans[8-13]. Many of the introgressed tracts have been shown to be of functional importance and may possibly be examples of adaptive introgression into humans, including several genes involved in immunity[14,15], several genes associated with skin pigmentation[11,12], and *EPAS1*, associated with high-latitude adaptation in Tibetans[16]. Archaic humans have been hypothesized to be adapted to cold temperatures[17]. Therefore, in this paper we examine if any of the selected genes in Inuit[1] may have been introduced into the modern human gene pool via admixture from archaic humans, i.e. Neanderthals or Denisovans[8,9]. We will show that the *WARS2/TBX15* haplotype of high frequency in Greenlandic Inuit was likely introgressed from an archaic human population that was closely related to Denisovans. We will also show that the selection affecting this haplotype is relatively old, resulting in a high allele frequency in other New World populations and intermediate allele frequencies in East Asia. Finally, functional genomic analyses suggest that the selected archaic haplotype may affect the regulation of expression of *TBX15* and *WARS2*, and is associated with phenotypes related to body fat distribution.

Results

Suggestive archaic ancestry in Greenlandic Inuit SNP chip data

We first computed two statistics, Patterson's D[14,15] and f_D[18], to assess putative local archaic human ancestry in candidate genes in the highest 99.5% quantile of the PBS genome-wide distribution of Fumagalli *et al.*[1]. Both of these statistics are intended to detect imbalances in allele sharing between a test population panel or genome, and a distantly related archaic population panel or genome, relative to a sister population panel of non-introgressed genomes that are on average more closely related to the test population than the archaic population is. The D statistic was developed to identify excess archaic ancestry in genome-wide analyses[10,19]. Although it has been deployed for assessments of local introgression in some species[20-22], the use of this statistic for detecting local introgression is unreliable, tending to give inflated values in genomic

regions of low diversity[18]. For this reason, we also use the statistic f_D[18], which better controls for local differences in diversity across the genome.

For the test population we used the SNP chip data from Fumagalli *et al.*[1], obtained from 191 Greenlandic Inuits with low European admixture (<5%). When computing D and f_D, we used a Yoruba genome sequenced to high-coverage (HGDP00927)[9] as the non-introgressed genome. For the archaic genome, we used either a high-coverage Denisovan genome[8] or a high-coverage Neanderthal genome[9]. All sites were polarized with respect to the inferred human-chimpanzee ancestral state[23].

The only top PBS locus showing some evidence of introgression is the *WARS2/TBX15* region, which has a high degree of allele sharing with the Denisovan genome (Figures S1, S2, Table S1). For example, in *WARS2*, we find 18.46 sites supporting a local tree in which Denisova is a sister group to Greenlandic Inuit, to the exclusion of Yoruba (ABBA), while we only find 2.5 sites supporting a Yoruba-Denisova clade, to the exclusion of Greenlandic Inuit (BABA) (the numbers are not integers because we are using the panel-version of D, not the single-genome version). However, because we used SNP chip data, we found that most genes had few informative (ABBA or BABA) sites. For genes for which there are 2 or more informative sites, *TBX15* is in the 88% quantile of the genome-wide distribution of f_D, and *WARS2* is in the 87% quantile. For genes for which there are more than 10 informative sites, *TBX15* is in the 88% quantile, and *WARS2* is in the 86% quantile. However, due to the small number of SNPs available for testing in Greenlandic Inuit, we could not assess with confidence whether this region was truly introgressed from Denisovans. We therefore sought to identify the selected haplotype in other samples for which full sequencing data are available.

The alleles with high PBS values and high frequency in Greenlandic Inuit are almost absent in Africa, while present across Eurasia. In Figure 1, we show the geographic distribution of allele frequencies for one of these SNPs (rs2298080) as an example, using data from phase 3 of the 1000 Genomes Project[24] and the Geography of Genetic Variants Browser[25]. The high-frequency alleles tend to match the Denisovan and Altai Neanderthal alleles in this region. For example, rs2298080 has an A allele at a frequency of 45.45% in CHB and at 99.74% frequency in Greenlandic Inuit. This allele is absent or almost absent (<1% frequency) in all African populations, and the Denisovan and Altai Neanderthal genomes are both homozygous for the A allele. We can therefore analyze sequencing data from Eurasians to determine if the selected alleles were truly introgressed from an archaic human population.

Excess archaic ancestry in Eurasians

A particularly useful way to detect adaptive introgression is to identify regions with a high proportion of uniquely shared sites between the archaic source population and the population subject to introgression. This was one of the lines of evidence in favor of Denisovan adaptive introgression in Tibetans at the *EPAS1* locus[16]. We therefore partitioned the genome into non-overlapping 40 kb windows and computed, in each window, the number of SNPs where the Denisovan allele is at a frequency higher than

20% in Eurasians but less than 1% in Africans, using the populations panels from phase 3 of the 1000 Genomes Project[24]. The windows containing *TBX15* and *WARS2* have four and three such sites, respectively, which is higher than 99.99% of all windows in the genome (Figure 2.A).

In each of the same 40 kb windows, we also computed the 95% quantile of Eurasian derived allele frequencies of all SNPs that are homozygous derived in Denisova and less than 1% derived in Africans. This quantile is assigned as the score for each window. This second statistic is designed to detect archaic alleles that are uniquely shared with Eurasians and have risen to extremely high frequencies. Here, *WARS2* and *TBX15* are also strong outliers, with a quantile frequency score above 99.95% of all windows (Figure 2.B). The chance that the region would randomly show such an extreme pattern of both excess number, and high allele frequency, of derived alleles shared with Denisovans and rare or absent in Africa, is exceedingly small (Racimo et al. in prep.) and strongly suggests that selection has been acting on an allele introgressed from archaic humans. We explore the statistical properties of this score elsewhere (Racimo et al. in prep.).

Identifying the introgression tract

We used the Hidden Markov Model (HMM) method of ref. [26] for identifying introgression tracts, using Yoruba as the population without any introgression. Using either Neanderthal or Denisova as the source population, we inferred a clear introgression tract in the middle of the PBS peak, which is especially frequent among East Asians (Figure 3). The inferred tract is more than twice as wide when using Denisova (chr1:119549417-119577202) as the source population (27,784 bp) than when using Neanderthal (13,729 bp) (Figure S3), which suggests the source population may have been more closely related to the Denisovan genome in this region. The frequency of the Denisovan tract is 48.41% in East Asians, 19.78% in Europeans and 12.37% in South Asians.

We further examined the SNPs defining this haplotype and their allelic state in different human population panels (Figure 4). We find a sharp distinction between the two most prevalent haplotypes across Eurasia, with one of them being almost identical to the Denisova genome and the other being highly differentiated from it. As expected, the frequencies of these two haplotypes agree with the frequencies of the inferred introgressed tracts and with the allelic frequencies of the top PBS SNPs. This pattern echoes the pattern observed for another well-known case of adaptive introgression in Tibetans[16], although in the present case the archaic haplotype is more widely distributed across Eurasia.

The selected alleles are the putative introgressed alleles

The archaic haplotype frequencies agree well with the frequencies of the selected alleles in different human populations. We therefore aimed to verify that the selected alleles were the same alleles that were uniquely shared with archaic humans. We focused on the

SNPs in the *TBX15/WARS2* region that are located in the 99.95% quantile of genome-wide PBS scores in the Greenlandic Inuit SNP chip data[1]. Noticeably, 28 out of the 33 top SNPs in the *TBX15/WARS2* region lie in the region where the introgressed haplotype is located. For each of these SNPs, we checked whether the selected allele in Greenlandic Inuit was the same as: a) the alleles present in the Altai Neanderthal genome[9], b) the alleles in the Denisova genome[8], c) the alleles present in a present-day human genome from the 1000 Genomes Project[27] (HG00436) that is homozygous for the introgressed tract, d) the alleles in a present-day human genome (HG00407) that does not contain the introgressed tract, and e) the alleles in 3 modern human genomes obtained from ancient DNA: Ust-Ishim[28] (dated at ~45,000 kya), Stuttgart (dated at ~7,000 kya) and Loschbour[29] (dated at ~8,000 kya).

We find that, in all of the SNPs showing most evidence of selection, the present-day human genomes that are homozygous for the introgression tract are also homozygous for the favored alleles. Additionally, in all of these SNPs, the present-day human genomes lacking the intogression tract are homozygous for the allele that was not favored by selection. Furthermore, in 79% of the SNPs, the selected allele is present in homozygous form in Denisova, while this is only true for 64% of the SNPs when looking at the Neanderthal genome (Table 1). This indicates that the selected alleles are also the introgressed alleles, and that Denisova is the most likely source of archaic introgression. In the 6 SNPs where the introgressed tract carries a different allele than Denisova, the allele in the introgressed tract is derived, suggesting these differences are due to mutations that occurred more recently than the time the introgressing lineage coalesced with the sequenced Denisovan's lineages, and possibly more recently than the introgression event. Finally, we find that in 100% of the SNPs the Ust'-Ishim, Stuttgart and Loschbour genomes are homozygous for the non-introgressed alleles, so they did not carry the introgressed haplotype.

The haplotype also shows selection signatures in Native Americans

To examine whether selection in Greenlandic Inuit on the introgressed haplotype was shared with Native Americans, we performed a scan for positive selection in the latter population by measuring the genetic differentiation against other reference populations. To correct for recent admixture in Latin America, we extracted the individuals showing the highest proportion of Native American ancestry in a 10 Mbp region around the introgressed haplotype (Figure S4).

We observed a local increase of F_{ST} in the proximity of the introgressed haplotype when comparing individuals from a Peruvian (PEL) population against populations of East Asian (CHB), European (TSI) and African (LWK) descent from the 1000 Genomes Project[24] (Figure 5). Such greater levels of genetic differentiation were also observed when considering individuals with Mexican ancestry in Los Angels (MXL) and Colombians individuals (CLM) as proxies for Native Americans but not when considering Puerto Rican individuals (PUR) as such (Figure S5), possibly due to the high level of European and African ancestry in the latter population sample.

To assess whether the observed values of F_{ST} around the introgressed haplotype may be explained by pure genetic drift, we first defined a demographic model for the divergence history between CHB and PEL without gene flow. We estimated demographic parameters for the shared history of CHB and PEL by fitting the observed F_{ST} calculated in the 10 Mbp region around the introgressed haplotype (see Methods). We obtained a divergence time of 29,000 years ago, and effective population sizes of 20,000 and 2,800 for CHB and PEL, respectively. This enabled us to perform coalescent simulations under a realistic demographic model and derive the expected distribution of F_{ST} values under neutral evolution. We found that an F_{ST} equal to 0.49, as observed in the introgressed haplotype between CHB and PEL, is unlikely to occur solely by genetic drift (p<0.001).

Using rs2298080 as a proxy for the selected haplotype, it has an inferred frequency of 0.96, 0.94, and 0.878 in the PEL, MXL, and CLM population panels, respectively. The apparent lack of fixation of the introgressed haplotype in these populations might be explained by the residual non-Native American ancestry (ranging from 8% to 29%) for the analyzed individuals. Clearly, selection on this allele is not unique to Inuit but has affected a large proportion of New World groups. A likely explanation is that selection on this locus has been acting during the early phases of the peopling of the Americas, perhaps in the Beringian ancestors of both modern Native Americans and Greenlandic Inuit.

The source of introgression was more closely related to Denisovans than Neanderthals

The divergence between the haplotype and the Denisova genome (0.0008) is lower than the divergence between the haplotype and the Altai Neanderthal genome (0.0016), suggesting a Denisovan origin. To further examine if the haplotype could be of Neanderthal – rather than Denisovan – origin, we computed the divergence between the Altai Neanderthal genome and a randomly chosen individual that was homozygous for the introgressed tract (HG00436). We compared this divergence to the distribution of divergences between the Altai Neanderthal and the Mezmaiskaya Neanderthal[9], computed across windows of the genome of equal size (Figure S6). If the introgressed haplotype came from Neanderthals, then we would expect the divergence between it and the Altai Neanderthal to be within the distribution of Neanderthal-Neanderthal divergence. To avoid errors due to the lower coverage of the Mezmaiskaya Neanderthal, we only computed divergence on the Altai side of the tree. The observed Altai-haplotype divergence falls in the 94.08% quantile of this distribution, suggesting this is likely not a typical Neanderthal haplotype.

We also obtained the distribution of divergences between the Denisovan genome and a high-coverage Yoruba genome (HGDP00936)[9]. We compared this distribution to the divergence between the introgressed haplotype and Denisova, and observe that this divergence falls towards the left end of the distribution, in the 17.43% quantile (Figure S6). Interestingly though, we observe that the introgressed haplotype is highly diverged with respect to Yoruba, with this divergence falling in the 97.63% quantile of the Denisova-Yoruba divergence distribution, and in the 97.82% quantile of the Neanderthal-Yoruba divergence distribution (Figure S6). All in all, this suggests that both the

introgressed haplotypes and the archaic haplotypes – though being closely related to each other – are both highly diverged relative to the non-introgressed present-day human haplotypes.

Additionally, we simulated two archaic populations that split at different times (100,000, 300,000 or 450,000 years), under a range of effective population sizes (1,000, 2,500, 5,000), based on estimates from ref. [9]. We compared these simulations to the Neanderthal-haplotype divergence, as well as to the Denisova-haplotype divergence and the divergence between the haplotype and the Yoruba genome (Figure S7). In all cases, the divergence that falls closest to the distribution is the Denisova-haplotype distribution, suggesting this is the closest source population for which we have sequence data.

To further understand the relationship between the introgressed haplotype and the archaic and present-day human genomes, we plotted a haplotype network (Figure 6). This network shows the most parsimonious distances among the 20 most common present-day human haplotypes and the 3 archaic haplotypes from Altai Neanderthal and Denisova in the region (Altai Neanderthal is homozygous for the same haplotype). We observe two distinct clusters of present-day human haplotypes: one cluster that is distantly related to both archaic genomes, and another cluster that is closely related to Denisova and contains mostly East Asian and Native American individuals, some European and South Asian individuals and almost no Africans. The Altai Neanderthal haplotype falls at an intermediate position between the Denisovan haplotypes and the first cluster, but shares more similarities with the Denisovan haplotypes. The present-day human cluster that is closest to Denisova (II) has a smaller Hamming distance to Denisova than does the Neanderthal haplotype (9 vs. 15, Figure S8), again suggesting the haplotype was introgressed from individuals more closely related to Denisovans than Neanderthals.

We also compared the distances between the haplotypes in *TBX15/WARS2* to those observed in *EPAS1* (Figure S9), a previously reported case of adaptive introgression in Tibetans from an archaic population closely related to the Denisovan genome[16,30]. We focused on the distances among Neanderthal, Denisova and the haplotype that is most closely related to the archaic genomes, for each of the distinct present-day clusters (II and XIX). We observe that, although the distances among Denisova, Neanderthal and the putatively introgressed haplotypes are similar, the distance between any of these and the non-introgressed haplotype are approximately double of what is observed in the case of *EPAS1* (Figure S9). This further confirms that the archaic and introgressed haplotypes in this region belong to a deeply divergent archaic lineage.

Regulatory differences in archaic and modern humans

We queried the GTEx database[31] to examine if the introgressed SNPs has an effect on human gene expression in various tissues. The sample sizes in the GTEx project do not provide enough power to detect trans-eQTLs, so we could only search for a cis-eQTL relationship with genes within a +/- 1 Mb window of each SNP. We therefore only checked whether the 28 introgressed SNPs that had signatures of positive selection in Native Americans were also eQTLs for *TBX15* or *WARS2*. We find that all these SNPs

are tightly linked and therefore have almost exactly the same p-values in each of the tissues, so we only focus on one of these (rs2298080) below.

Table 2 shows the effect sizes and p-values obtained from GTEx for 41 different tissues. For *TBX15*, we find only one tissue with P < (0.05 / number of tissues tested), in the testis, where the Denisova variant increases expression of the gene (P = 0.00018). When querying the SNP for expression effects on *WARS2*, we find 6 tissues with P < (0.05 / number of tissues tested): subcutaneous adipose, adrenal gland, aorta, tibial artery, esophagus muscularis, skeletal muscle. This suggests expression differences, in the tissues and developmental stages represented in the GTEx database, are more ubiquitous on *WARS2* than *TBX15*. For all significant tissues, we find that the Denisovan variant decreases expression of *WARS2*.

While *WARS2* is expressed in most tissues, the expression of *TBX15* is largely restricted to mesenchymal tissues (see Methods). We therefore also investigated a data set of skin fibroblasts (a mesenchymal tissue), where both genes are expressed, and in which each individual is characterized for methylation, expression and sequence. The data set included 62 individuals, of which two were homozygous for the introgressed allele, 21 were heterozygous, and 39 did not carry the introgressed allele[32]. Analyzing *WARS2* and *TBX15* expression in these individuals revealed that the expression levels change with the number of introgressed alleles (Figure 7A, $P = 0.012$ and 0.029, respectively, ANOVA) with mean expression in heterozygous individuals falling intermediately between the two other genotype groups. *TBX15* in individuals who are homozygous for the introgressed allele is expressed, on average, 39% higher than in individuals who lack the allele. On the other hand, the average expression level of *WARS2* is 58% lower in individuals homozygous for the introgressed allele.

We then examined the euchromatin region between *TBX15* and *WARS2*. For each gene, we looked for clusters of CpG positions whose methylation levels significantly differ between tissues where the gene is expressed and those where it is not expressed. For *WARS2* no such clusters were detected. However, for *TBX15* we found six clusters, containing a total of 357 CpGs ($P < 0.05$; t-test, Bonferroni-corrected; Figure 7B). Of particular interest was cluster 6, which resides within the introgressed region. This cluster presents multiple signatures of regulatory activity: overlap with a large CpG island, DNase I hyper-sensitivity, GC-richness, and the presence of the histone modifications H3K27ac and H3K4me1 across many tissues (See Methods). The 68 CpG positions within the cluster are hyper-methylated in tissues that express TBX15. When we examined whether the methylation level of these CpGs is associated with the level of introgression, we found that individuals who are homozygous for the tract are significantly hyper-methylated, in accordance with their increased expression of *TBX15* ($P = 0.016$, Tukey-Kramer test, Figure 7C). Surprisingly, while heterozygous individuals exhibit increased expression levels compared to individuals who do not carry the introgressed allele, such a trend is absent when analyzing methylation ($P = 0.9987$, Tukey-Kramer test).

Next, we examined archaic methylation patterns by comparing the bone methylation map of the Denisovan to those of the Altai Neanderthal[33] and Ust'-Ishim (see Methods), as well as to 20 additional methylation maps from various modern human tissues. Interestingly, the euchromatin region between *TBX15* and *WARS2* contains four previously reported differentially methylated regions (DMRs) in which the Denisovan methylation patterns differ significantly from those of present-day humans[33]. This makes *TBX15* one of the most DMR-rich genes in the Denisovan genome. These DMRs do not overlap the introgressed region, but rather, they are immediately downstream of it, as they are found in close vicinity to the transcription start site (TSS) of *TBX15* (10kb downstream to 5kb upstream). They also bear active chromatin marks (i.e. DNase I hyper-sensitivity, binding by p300 and H2A.Z, and the active histone modification marks H3K27ac and H3K4me1), suggesting these DMRs might have had a regulatory effect on *TBX15* in the Denisovan individual.

The ability to detect DMRs was previously limited by the fact that archaic methylation maps were compared to a reduced representation bisulfite sequencing (RRBS) map[33], which covers only about 10% of the genome. Hence, we ran an additional DMR-detection analysis in which we compared the Denisovan methylation in this region to that of the Ust'-Ishim modern human individual (who does not carry the introgressed haplotype), instead of the RRBS map. DMRs were defined as regions where the Denisovan clusters with tissues where *TBX15* is expressed, whereas Ust'-Ishim clusters with tissues where the gene is not expressed, or vice-versa. Considering that individuals who carry the introgressed allele exhibit increased expression levels of *TBX15*, we expected the Denisovan genome to bear methylation patterns that are associated with elevated expression, relative to Ust'-Ishim. Surprisingly, we detected the opposite trend; out of the 357 expression-associated CpGs, we found only one in which the Denisovan clusters with tissues where *TBX15* is expressed, whereas we found 86 CpGs in which the Denisovan clusters with tissues where *TBX15* is not expressed ($P < 0.05$, Bonferroni-corrected; see Methods). These include many positions within the introgressed region, where the Denisovan shows hypomethylation compared to Ust'-Ishim, the Neanderthal and the introgressed individuals (Figure 7B).

Lastly, we examined if individuals who carry the introgressed allele present changes in methylation outside the introgressed region. We found no such changes for 4/5 CpG clusters in the region (labeled 1,2,3 and 5 in Figure 7B). However, in cluster 4 we observe a significant decrease in mean methylation level in individuals who carry the introgressed allele ($P = 0.0002$, ANOVA). This is surprising both because the Denisovan displays no changes in methylation in this region, and because decreased methylation in cluster 4 is associated with decreased expression.

Association studies

Because the haplotype is at intermediate frequencies in Europeans, we queried the GIANT consortium GWAS data[34-36], which contains a number of anthropometric traits tested on a European panel. When looking at all p-values in the region, we find that there is a peak of significantly associated SNPs for three phenotypes right where the haplotype

is inferred to be located (Figure 8). These phenotypes are BMI-adjusted waist circumference, waist-hip ratio and BMI-adjusted waist-hip ratio. We then queried the most extreme SNPs that serve to differentiate the archaic from the non-archaic haplotypes (see Methods) and the SNPs that were among the top PBS hits in Greenlandic Inuit (Table 1). We find that, even though some of these SNPs are significantly associated with BMI-adjusted waist-circumference ($P < 10^{-5}$), they are not among the top most significant SNPs in the region for any of the three phenotypes (Figure 8). The Denisovan alleles in the queried SNPs have a positive effect size for all three phenotypes. We also observed that the region overlaps a 100 kb region designated as a mouse QTL for the induction of brown adipocytes (MGI:2149993)[37].

Discussion

We have identified a highly divergent haplotype in the *TBX15/WARS2* region, which was likely introduced into the modern human gene pool via introgression with archaic humans. The most likely source of introgression is the Denisovan population, or a population more closely related to Denisovans than to Neanderthals at this locus. The archaic haplotype is present at higher frequencies in East Asians than in Europeans and South Asians, and at even higher frequencies in Greenlandic Inuit and Native Americans, where it is almost fixed. This suggests there may have been a temporally and geographically extended period of selection for the archaic haplotype in modern human history.

The TBX15/WARS2 region is highly pleiotropic: it has been found to be associated with a variety of traits. These include the differentiation of adipose tissue[5], body fat distribution[6,35,38,39], facial morphology[4,40], stature[4], ear morphology[7,41], hair pigmentation[42] and skeletal development[3,4]. Interestingly, for several of body fat distribution studies, the introgressed SNPs have significant genome-wide associations. The Denisovan alleles tend to be associated with increased waist circumference and waist-hip ratio, after correcting for BMI.

The haplotype is located immediately upstream and partially overlapping the promoter region of *TBX15*, in support of the idea that it affects this gene's regulation. However, when mining the tissues represented in the GTEx database, we find that the Denisovan SNPs only appear to be cis-eQTLs for this gene in the testis. We also find that they are associated with changes of expression of *WARS2* across various tissues. *WARS2* contains one exon that overlaps with the tract. The tissues for which the SNPs are significant eQTLs for WARS2 are subcutaneous adipose, adrenal gland, aorta, tibial artery, esophagus muscularis and skeletal muscle. Some of these, like subcutaneous adipose and skeletal muscle, are consistent with the associations between the region and fat differentiation / skeletal development. Analyzing other data, the introgressed alleles also appear to affect expression of both TBX15 and *WARS2* in skin fibroblasts, although in opposite directions.

An analysis of epigenetic marks in the region suggested that the regulatory function of *TBX15* and *WARS2* is affected by the tract, but also indicated that the relationship

between the introgressed haplotype and its phenotypic effects may be rather complex. Individuals who are homozygous for the tract are significantly hyper-methylated in various CpG clusters in the region and also show significantly increased expression of *TBX15* in fibroblasts. Intriguingly though, the Denisovan genome bears CpG methylation patterns that are consistent with reduced expression, relative to Ust'-Ishim, who does not carry the haplotype (Figure 7B). The Denisovan genome also contains a large number of differentially methylated regions relative to all modern humans, which cover many putative regulatory hotspots in between the two genes, including the promoter of *TBX15*.

Interestingly, whereas hyper-methylation in a promoter region is usually associated with the silencing of a gene[43], *TBX15* presents the opposite trend in all expression-associated CpG clusters upstream to its TSS. That is, hyper-methylation tends to occur in individuals with increased *TBX15* expression. Notably, *TBX15* has an alternative downstream TSS, which raises the intriguing possibility that the hyper-methylation of the upstream promoter indeed silences this TSS, but induces the transcription of the downstream TSS.

Another important observation is that whereas the mean *TBX15* expression in heterozygous individuals in fibroblasts seems to fall intermediately between the other genotype groups (Figure 7), we observe that this is not the case with regard to methylation. We find that heterozygous individuals display a methylation pattern that is very similar to individuals that do not carry the introgressed haplotype. This hints to a recessive-like behavior, where two introgressed alleles are required for changes in methylation to occur. This also suggests that if methylation drives some of the changes in expression, it is not the sole factor, and the introgressed tract possibly affects the regulation of *TBX15* through more than one mechanism. This possibility is also supported by the fact that the tract is linked to changes in the expression of WARS2, but no changes in methylation are observed for this gene, neither in individuals who carry the introgressed allele, nor in the Denisovan genome.

Altogether, our study suggests a complex multi-factorial regulation of *TBX15* and *WARS2*. We show that the introgressed region is associated with regional changes in methylation and expression levels, but our findings also hint to other factors that affect the regulation of these genes and are yet to be elucidated.

Methods

D and f_D statistics in Greenlandic Inuit SNP data

Following refs. [10,19], at site i, let $C_i(ABBA) = ((1 - f_{Yoruba}) \times f_{Greenlandic\ Inuit} \times f_{Archaic\ human})$, where *f* is the derived allele frequency (with respect to the human-chimpanzee ancestor) in either a population panel (for the Greenlandic Inuit) or a diploid genome (for Yoruba and the archaic humans). Furthermore, let $C_i(BABA) = (f_{Yoruba} \times (1 - f_{Greenlandic\ Inuit}) \times f_{Archaic\ human})$. Then, for a set of N sites within a particular region of the genome, we computed D as follows:

$$D = \frac{\sum_{i=1}^{N} C_i(ABBA) - C_i(BABA)}{\sum_{i=1}^{N} C_i(ABBA) + C_i(BABA)}$$

Let S(Yoruba,Greenlandic Inuit,Archaic,Chimpanzee) be the numerator in the D statistic defined above. We computed f_D as follows:

$$f_D = \frac{S(Yoruba, Greenlandic\ Inuit, Archaic, Chimpanzee)}{S(Yoruba, X, X, Chimpanzee)}$$

Here, X is defined – dynamically for each site i – as the population (either Archaic or Greenlandic Inuit) that has the highest derived allele frequency.

Introgressed tracts in Eurasian whole-genome data

We set an admixture proportion of 2%, an admixture time of 1900 generations ago and a constant recombination rate of 2.3×10^{-8} per bp per generation. We used YRI as the population with no introgression. We called tracts if the posterior probability for introgression estimated using the HMM was higher than 90%. We also tried increasing the admixture proportion to 10% and 50%, but did not observe any major differences in the length of the tracts or the proportion of individuals carrying them, except that some of the gaps between tracts in the same chromosome tended to be joined together slightly more often than when using a 2% rate.

Uniquely shared sites

We defined "Eurasian uniquely shared sites" as sites where the Denisovan genome is homozygous and where the Denisovan allele is at low frequency (< 1%) in Africans (AFR, excluding admixed African-Americans), but at high frequency (> 20%) in non-American Eurasians (EUR+EAS+SAS) from phase 3 of the 1000 Genomes Project[27]. Similarly, we defined the "derived shared quantile" statistic as the 95% quantile of all derived allele frequencies in Eurasians, for SNPs where the Denisovan allele is homozygous for the derived allele and where the derived allele is at low frequency (< 1%) in Africans. In both cases, we only used sites that lied in regions with 20-bp Duke mappability equal to 1[44] and that were not in repeat-masked regions[45].

Haplotype clustering

To examine the haplotypes in this region, we computed the number of pairwise differences between every pair of haplotypes in a particular continental panel. Then we ordered the haplotypes based on their number of pairwise distances to the archaic sequence in each continent. Figure 4 is generated using the heatmap.2 function from the gplots package of the statistical computing platform R[46].

Haplotype network

We built a haplotype network based on pair-wise differences using R[46] and the software package pegas[47]. To plot the network, we used the 20 most abundant present-day human haplotypes. To make a fair comparison with published distances for *EPAS1*[30], we looked at the 40 most abundant present-day haplotypes instead, and only counted differences at SNPs that were segregating in present-day humans.

F_{ST} scan in Native Americans

We extracted sequencing data in form of BAM files for a 10 Mbp region surrounding the putative introgressed haplotype from the 1000 Genomes Project data set[24]. We selected all unrelated individuals belonging to 4 population panels: African Luhya (LWK), European Tuscans (TSI), Han Chinese (CHB) and Peruvians (PEL).

To select PEL individuals that would serve as optimal representatives of Native American genetic variation, we calculated admixture proportion among all LWK, TSI, CHB and PEL individuals assuming 4 ancestral populations using NGSadmix[48]. We then extracted the first 30 PEL individuals showing the highest proportion of inferred Native American ancestry (> 0.92). Similarly, we selected the first 30 individuals ranked by African, European and East Asian ancestry, respectively. We analyzed a total of 120 individuals.

As an additional analysis, we repeated the same procedure using either Colombian (CLM), Mexican (MXL) or Puerto Rican (PUR) individuals to represent the Native American component. After calculating the ancestry proportion as described above, 20 individuals for MXL, CLM or PUR were chosen with a threshold on the Native American ancestry component resulting in 0.83, 0.71, and 0.66, respectively.

To visually inspect whether we correctly selected unadmixed individuals in the Latin American cohorts, we performed a multidimensional scaling analysis on matrices of pairwise genetics distances calculated using ngsDist[49], which takes genotype uncertainty into account.

We computed F_{ST}, a measure of population genetic differentiation, between the most unadmixed Latin Americans and other populations using a method-of-moments estimator that incorporates genotype uncertainty[50] implemented in the software ANGSD[51]. Similarly, we calculated allele frequencies for each site using a likelihood-based method that corrects for the uncertainty in genotypes due to low-depth data[52]. To identify signatures of positive selection in Native Americans, we scanned the region around the putatively introgressed haplotype using a sliding-windows approach, with window size of 20kbp and step of 2kbp.

To assess the deviation of the observed F_{ST} from neutral expectations, we inferred a demographic model for the shared history and divergence of PEL and CHB. We used a previously proposed model[53] and refined the divergence time between these two populations assuming no subsequent gene flow by fitting the observed and expected F_{ST} under a given evolutionary model. We used the software *dadi*[54] to compute the expected

F_{ST} from a generic demographic model and estimated parameters using a coarse grid search. Under this model, we ran 100,000 neutral coalescent simulations with neutral evolution using the software *msms*[55] and recorded the F_{ST} value for each repetition.

TBX15 and WARS2 regulation

Data from SNP arrays, methylation arrays and expression arrays for the 62 fibroblast samples were downloaded from Gene Expression Omnibus (GEO), accession number GSE53261. Outliers were removed using an interquartile range (IQR); individuals with expression levels above Q3+1.5IQR or under Q1-1.5IQR were removed from all following analyses. Two individuals were detected as outliers for *TBX15* (GM03075 and WG2193), and four for *WARS2* (WG2193, GM00290, WG3466 and GM02641).

For each gene, we downloaded expression profiles in 20 tissues from EMBL-EBI Expression Atlas (Roadmap and ENCODE samples), and classified the tissues into expressed (>1 RPKM) and not expressed (<= 1 RPKM). Next, we detected expression-associated CpGs for each of the genes. To this end, we ran an overlapping sliding window of 5 CpG positions on chr1: 119,511,126 - 119,577,202. This region includes the euchromatin segment between *TBX15* and *WARS2*, as well as the entire introgressed tract. The euchromatin region was defined based on the ChromHMM annotations[56] and methylation levels. For each CpG position, we used a two-tail t-test (assuming equal variance) to compare its methylation in each of the two groups of tissues. P-values were Bonferroni-corrected. As a sliding window was used during reconstruction of the archaic methylation maps, looked at regional methylation, to allow for a reliable comparison. Expression-associated CpG positions were clustered by requiring each cluster to have a minimum of 10 CpGs, and requiring that adjacent CpGs were not separated by more than 500 bases. This resulted in six clusters, which contained a total of 357 expression-associated CpGs. For *WARS2*, 20 expression-associated CpGs were found, but they could not be clustered.

For each cluster, we removed methylation outliers using the same method as described above. Four individuals lacking the introgressed allele were removed (WG2414, WG3465, WG2275 and GM00940), as well as one heterozygous individual (WG2121). Each cluster was characterized for active chromatin marks (H3K27ac, H3K4me1 and DNase-I), downloaded from an integrative analysis of 111 epigenomes[57].

Ust'-Ishim methylation in the above region was reconstructed as described in Gokhman et al.[33], using a sliding window of 50 CpGs. The 20 modern methylation maps were downloaded from the ENCODE and Roadmap projects, accession numbers: GSE27584, GSE46644 and GSE46644. We used the classification of tissues to expressed and non-expressed (as described above) in order to identify Denisovan-Ust'-Ishim DMRs. A CpG position that in the Denisovan lies within (or is more extreme than) the methylation range of one group of tissues, but in the Ust'-Ishim lies within (or is more extreme than) the other group, was marked as differentially methylated. Out of 357 expression-associated CpG positions, 87 met the above criterion.

Acknowledgements

We thank Montgomery Slatkin and members of the Slatkin and Nielsen labs for helpful advice and discussions. We also thank Amy Ko and Jacob Crawford for their help and advice in inferring admixture tracts. RN is supported by a National Institutes of Health grant (R01HG003229). LC is supported by the Israel Science Foundation FIRST individual grant (ISF 1430/13). MF is supported by a Human Frontier Science Program fellowship (LT00320/2014). E.H.S is supported by UC Merced start-up funds.

References

1 Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343-1347, doi:10.1126/science.aab2319 (2015).

2 Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78, doi:10.1126/science.1190371 (2010).

3 Singh, M. K. *et al.* The T-box transcription factor Tbx15 is required for skeletal development. *Mech Dev* **122**, 131-144, doi:10.1016/j.mod.2004.10.011 (2005).

4 Lausch, E. *et al.* TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome. *Am J Hum Genet* **83**, 649-655, doi:10.1016/j.ajhg.2008.10.011 (2008).

5 Gburcik, V., Cawthorn, W. P., Nedergaard, J., Timmons, J. A. & Cannon, B. An essential role for Tbx15 in the differentiation of brown and "brite" but not white adipocytes. *Am J Physiol Endocrinol Metab* **303**, E1053-1060, doi:10.1152/ajpendo.00104.2012 (2012).

6 Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949-960, doi:10.1038/ng.685 (2010).

7 Adhikari, K. *et al.* A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat Commun* **6**, 7500, doi:10.1038/ncomms8500 (2015).

8 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

9 Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).

10 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

11 Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357, doi:10.1038/nature12961 (2014).

12 Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017-1021, doi:10.1126/science.1245938 (2014).

13 Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet* **16**, 359-371, doi:10.1038/nrg3936 (2015).

14 Abi-Rached, L. *et al.* The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334**, 89-94, doi:10.1126/science.1209202 (2011).

15 Dannemann, M., Andrés, A. M. & Kelso, J. Adaptive variation in human toll-like receptors is contributed by introgression from both Neandertals and Denisovans. *bioRxiv*, doi:http://dx.doi.org/10.1101/022699 (2015).

16 Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-197, doi:10.1038/nature13408 (2014).

17 Steegmann, A. T., Cerny, F. J. & Holliday, T. W. Neandertal cold adaptation: physiological and energetic factors. *Am J Hum Biol* **14**, 566-583, doi:10.1002/ajhb.10070 (2002).

18 Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular biology and evolution*, doi:10.1093/molbev/msu269 (2014).

19 Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* **28**, 2239-2252 (2011).

20 Kronforst, M. R. *et al.* Hybridization reveals the evolving genomic architecture of speciation. *Cell Rep* **5**, 666-677, doi:10.1016/j.celrep.2013.09.042 (2013).

21 Rheindt, F. E., Fujita, M. K., Wilton, P. R. & Edwards, S. V. Introgression and phenotypic assimilation in Zimmerius flycatchers (Tyrannidae): population genetic and phylogenetic inferences from genome-wide SNPs. *Syst Biol* **63**, 134-152, doi:10.1093/sysbio/syt070 (2014).

22 Smith, J. & Kronforst, M. R. Do Heliconius butterfly species exchange mimicry alleles? *Biol Lett* **9**, 20130503, doi:10.1098/rsbl.2013.0503 (2013).

23 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-1828, doi:10.1101/gr.076554.108 (2008).

24 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

25 Marcus, J. H. & Novembre, J. *Geography of Genetic Variants Browser v0.2*, <http://popgen.uchicago.edu/ggv/> (

26 Seguin-Orlando, A. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science*, doi:10.1126/science.aaa0114 (2014).

27 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

28 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).

29 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).

30 Huerta-Sanchez, E. & Casey, F. P. Archaic inheritance: supporting high altitude life in Tibet. *J Appl Physiol (1985)*, jap.00322.02015, doi:10.1152/japplphysiol.00322.2015 (2015).
31 Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).
32 Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol* **15**, R37, doi:10.1186/gb-2014-15-2-r37 (2014).
33 Gokhman, D. *et al.* Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science* **344**, 523-527, doi:10.1126/science.1250368 (2014).
34 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186, doi:10.1038/ng.3097 (2014).
35 Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196, doi:10.1038/nature14132 (2015).
36 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).
37 Xue, B., Coulter, A., Rim, J. S., Koza, R. A. & Kozak, L. P. Transcriptional synergy and the regulation of Ucp1 during brown adipocyte induction in white fat depots. *Mol Cell Biol* **25**, 8311-8322, doi:10.1128/MCB.25.18.8311-8322.2005 (2005).
38 Liu, C. T. *et al.* Multi-ethnic fine-mapping of 14 central adiposity loci. *Hum Mol Genet* **23**, 4738-4744, doi:10.1093/hmg/ddu183 (2014).
39 Liu, C. T. *et al.* Genome-wide association of body fat distribution in African ancestry populations suggests new loci. *PLoS Genet* **9**, e1003681, doi:10.1371/journal.pgen.1003681 (2013).
40 Pallares, L. F. *et al.* Mapping of Craniofacial Traits in Outbred Mice Identifies Major Developmental Genes Involved in Shape Determination. *PLoS Genet* **11**, e1005607, doi:10.1371/journal.pgen.1005607 (2015).
41 Curry, G. A. Genetical and developmental studies on droopy-eared mice. *Development* **7**, 39-65 (1959).
42 Candille, S. I. *et al.* Dorsoventral patterning of the mouse coat by Tbx15. *PLoS Biol* **2**, E3, doi:10.1371/journal.pbio.0020003 (2004).
43 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484-492, doi:10.1038/nrg3230 (2012).
44 Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**, e30377, doi:10.1371/journal.pone.0030377 (2012).
45 Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-3.0* 1996-2010.).
46 Team, R. C. (Vienna, Austria, 2012).
47 Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419-420, doi:10.1093/bioinformatics/btp696 (2010).

48 Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**, 693-702, doi:10.1534/genetics.113.154138 (2013).

49 Vieira, F. G., Lassalle, F., Korneliussen, T. S. & Fumagalli, M. Improving the estimation of genetic distances from Next‐Generation Sequencing data. *Biological Journal of the Linnean Society*, 1095-8312 (2015).

50 Fumagalli, M. *et al.* Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**, 979-992, doi:10.1534/genetics.113.154740 (2013).

51 Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356, doi:10.1186/s12859-014-0356-4 (2014).

52 Kim, S. Y. *et al.* Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**, 231, doi:10.1186/1471-2105-12-231 (2011).

53 Raghavan, M. *et al.* POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884, doi:10.1126/science.aab3884 (2015).

54 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695, doi:10.1371/journal.pgen.1000695 (2009).

55 Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064-2065, doi:10.1093/bioinformatics/btq322 (2010).

56 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).

57 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

Figures

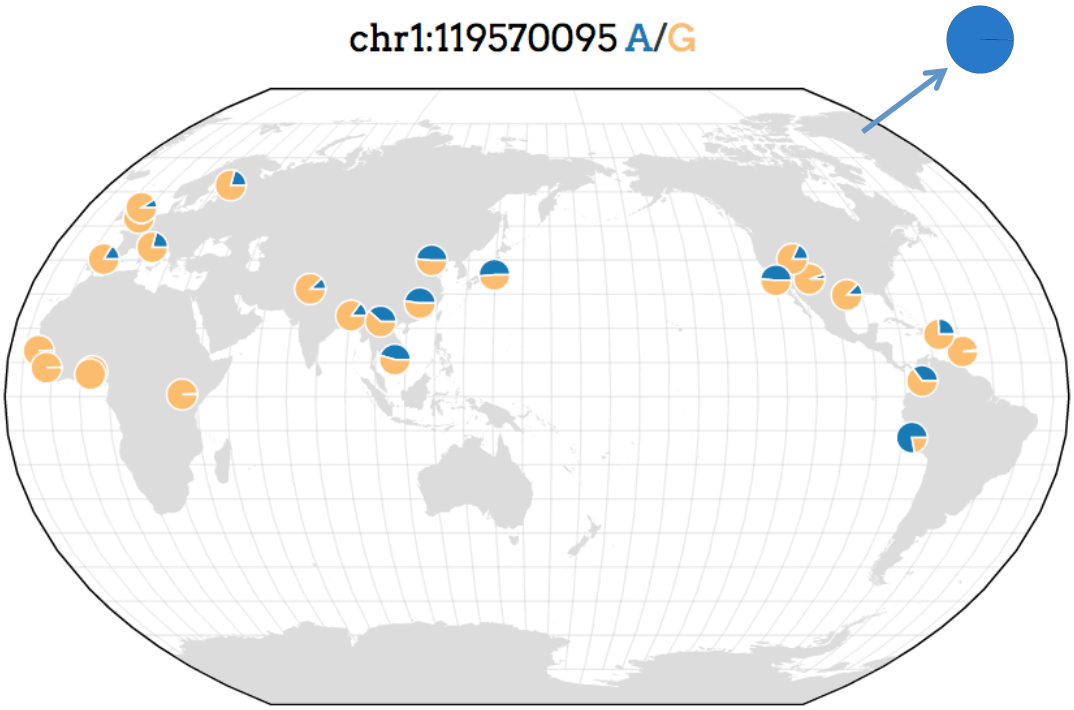

Figure 1. Geographic distribution of rs2298080 in different 1000 Genomes populations. The color blue corresponds to the archaic allele in this SNP. For comparison, the allele frequency in Greenlandic Inuit is also shown. This figure was made using the Geography of Genetic Variants browser v.0.1, by J. Novembre and J.H. Marcus: http://popgen.uchicago.edu/ggv/

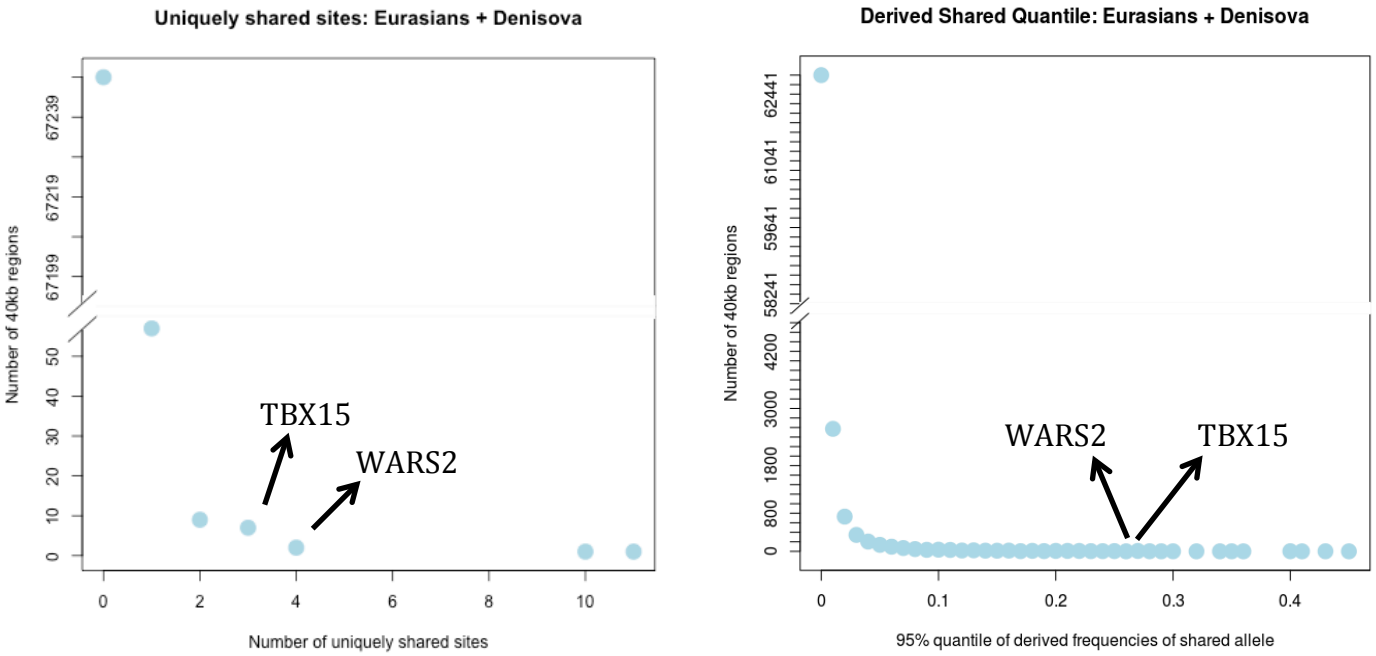

Figure 2. A) Genome-wide counts of uniquely shared sites where the Denisovan allele is at less than 1% frequency in Africans (AFR, excluding admixed African-Americans) and at more than 20% frequency in Eurasians (EUR + SAS + EAS). The counts were computed in non-overlapping 40 kb regions of the genome. The y-axis is truncated, as the vast majority of regions have 0 uniquely shared sites. *TBX15* and *WARS2* are among the few regions that have 3 and 4 uniquely shared sites, respectively. B) In each of the same 40 kb windows, we also computed the 95% quantile of Eurasian derived allele frequencies of all SNPs that are homozygous derived in Denisova and less than 1% derived in Africans. The 95% quantile scores for *TBX15* and *WARS2* are 0.27 and 0.26, respectively.

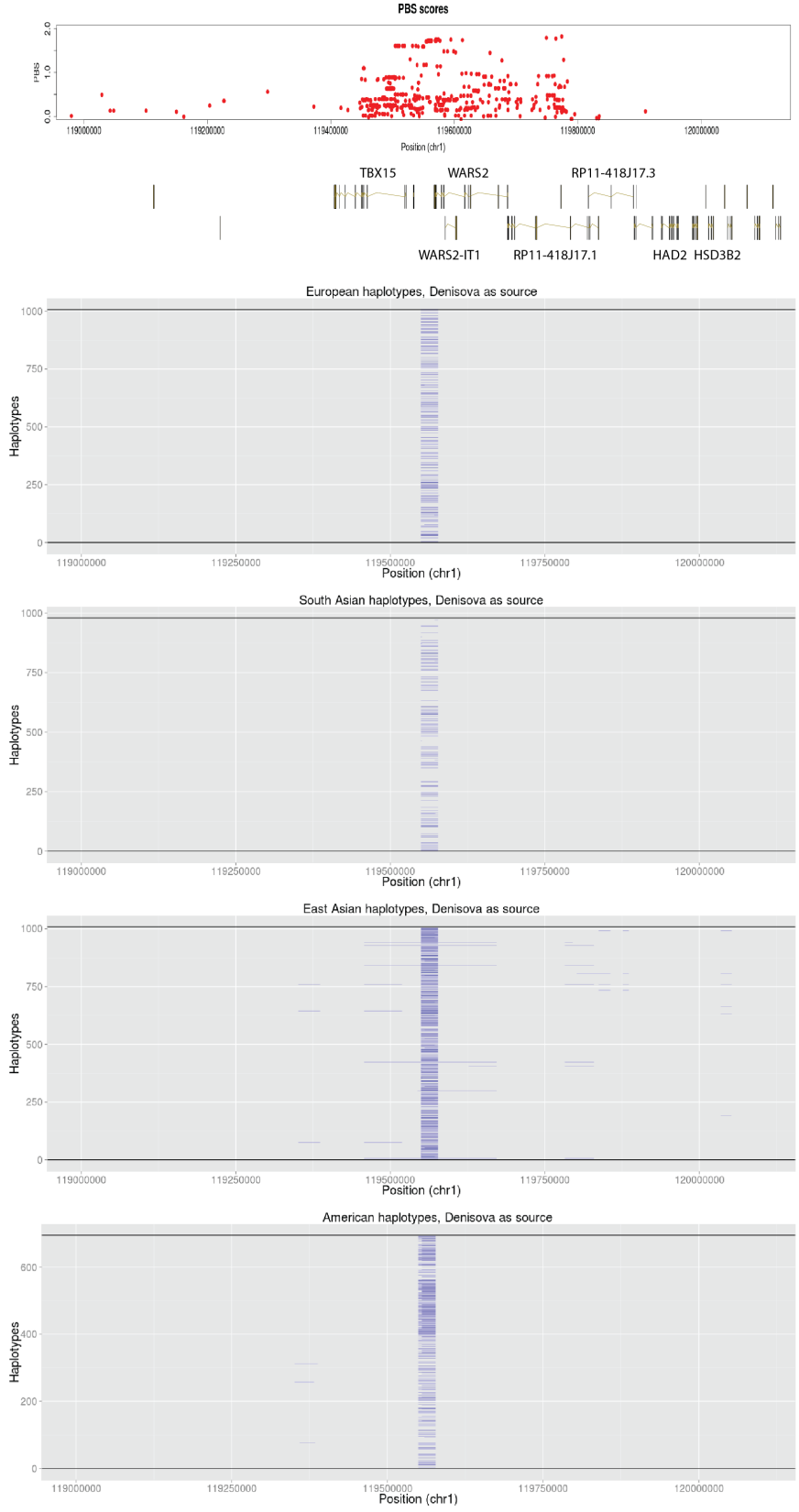

Figure 3. East Asian, South Asian, American and European phased chromosomes from the 1000 Genomes Project are inferred to contain archaic haplotypes (blue tracts), which likely came from an individual that was more closely related to the Denisovan than to the Altai Neanderthal in this region. The frequency of the archaic haplotype is higher among East Asians than among South Asians and Europeans.

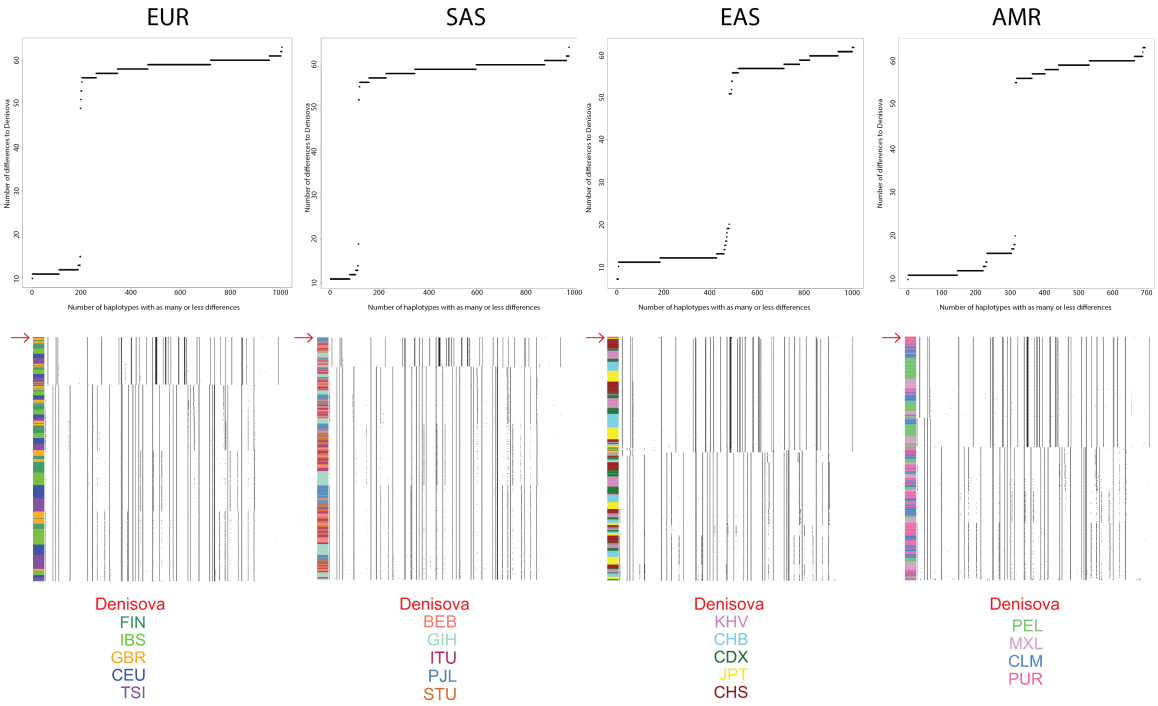

Figure 4. Top panel: We counted the differences between the Denisovan genome and all non-African haplotypes in each 1000G population (EUR, SAS, EAS and AMR). The plot shows the cumulative number of haplotypes that have as many or less differences to the Denisovan genome than specified in the x-axis. Bottom panel: Haplotype plots for all 1000G non-African haplotypes in the introgressed tract, clustered by similarity to the Denisovan genome (red arrow, at the top of each panel). The color codes at the bottom refer to 1000G sub-populations to which the different haplotypes belong, as indicated by the right color column in each panel. The haplotypes were clustered by decreasing similarity to the Denisovan sequence, as described in the main text.

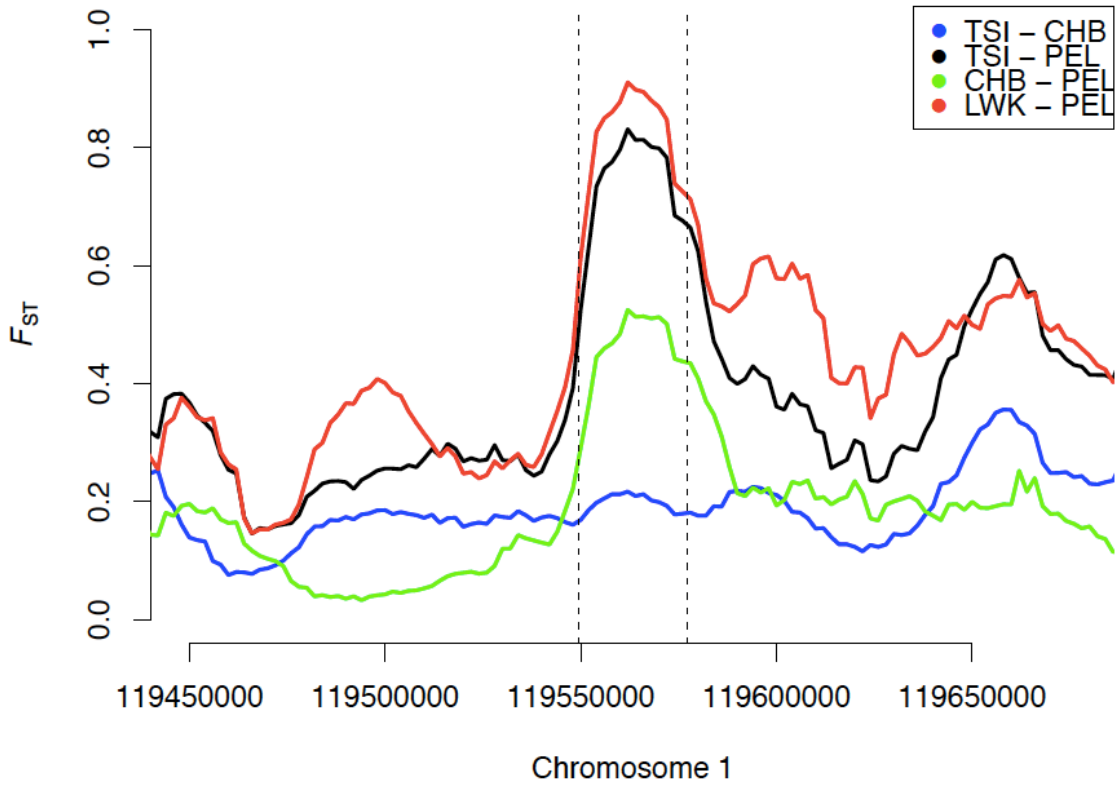

Figure 5. We computed local F_{ST} values around the putatively introgressed haplotype using a sliding-windows approach, with window size of 20kbp and step of 2kbp. We used different pairs of populations, and find a local increase in F_{st} when using PEL against different African (LWK) and Eurasian (TSI, CHB) populations, which is located exactly where the introgressed haplotype is inferred to be (dotted lines). As a comparison, we found no increase in F_{ST} when comparing Europeans and East Asians (blue line).

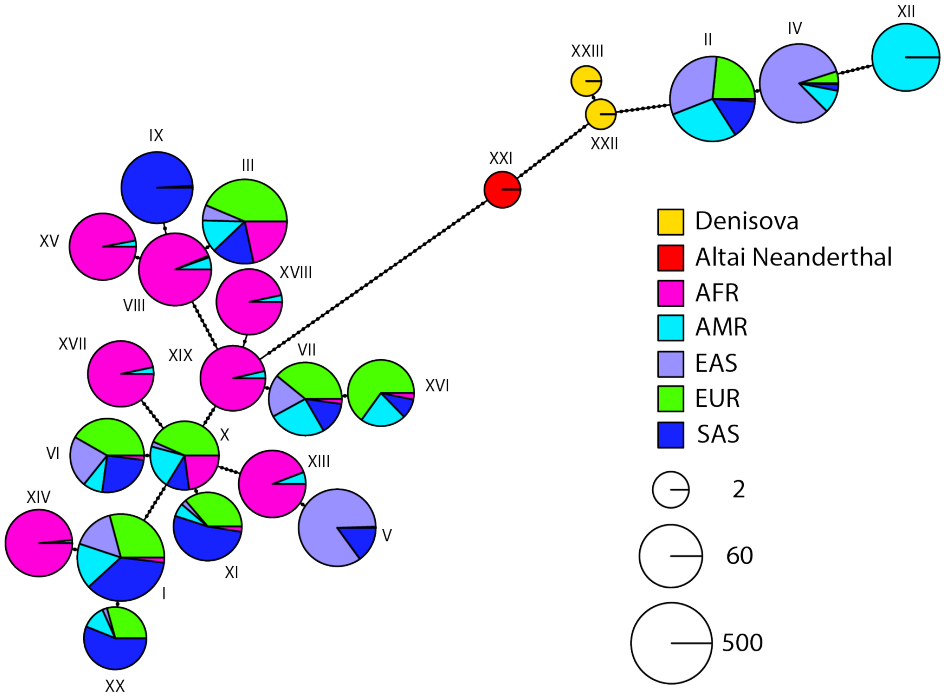

Figure 6. Network of archaic haplotypes and 20 most common present-day human haplotypes from the 1000 Genomes Project. Each pie chart is a haplotype, and the dots along each line represent the number of differences between each haplotype. The size of each pie chart is proportional to the log base 2 of the number of individuals in which that haplotype appears, and the colors refer to the proportion of those individuals that come from different continental populations. AFR: Africans. AMR: Americans. EAS: East Asians. EUR: Europeans. SAS: South Asians.

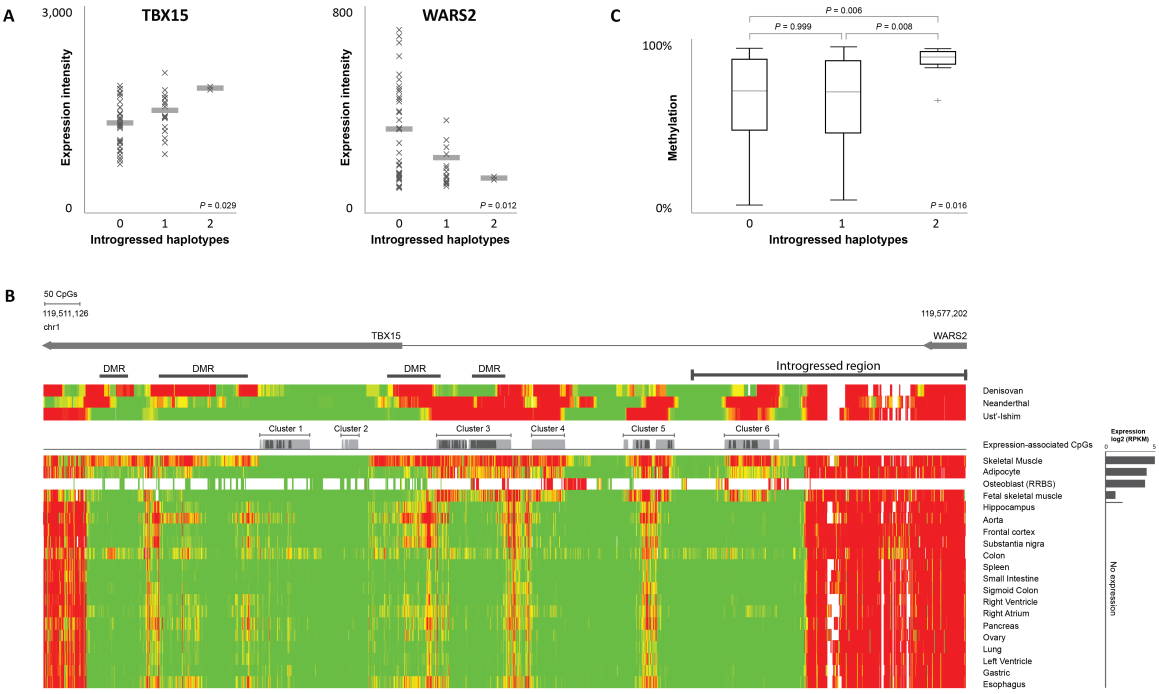

Figure 7. The regulatory effects of the introgressed haplotype on *TBX15* and *WARS2*. A. Expression intensity vs. number of introgressed haplotypes for *TBX15* and *WARS2* in skin fibroblasts. X's Mark individuals and horizontal lines mark the mean methylation of each group. B. Methylation maps for the euchromatin region between *TBX15* and *WARS2*. The top panel shows the reconstructed methylation maps of the Denisovan, Neanderthal and the Ust'-Ishim anatomically modern human, as well as previously reported DMRs. Methylation level is color coded from green (unmethylated) to red (methylated). The middle panel displays CpG clusters that are significantly associated with the expression of *TBX15*, marked by grey bars. Out of them, the 87 differentially methylated CpG positions between the Denisovan and Ust'-Ishim are marked by dark grey bars. The Denisovan displays a regulatory pattern that is associated with decreased expression. The bottom panel shows 20 modern tissue methylation maps. The osteoblast map includes fewer positions as it was produced using a reduced representation bisulfite sequencing (RRBS) protocol. Expression levels are displayed on the right. C. Box plot of methylation vs. number of introgressed haplotypes for CpG cluster 6.

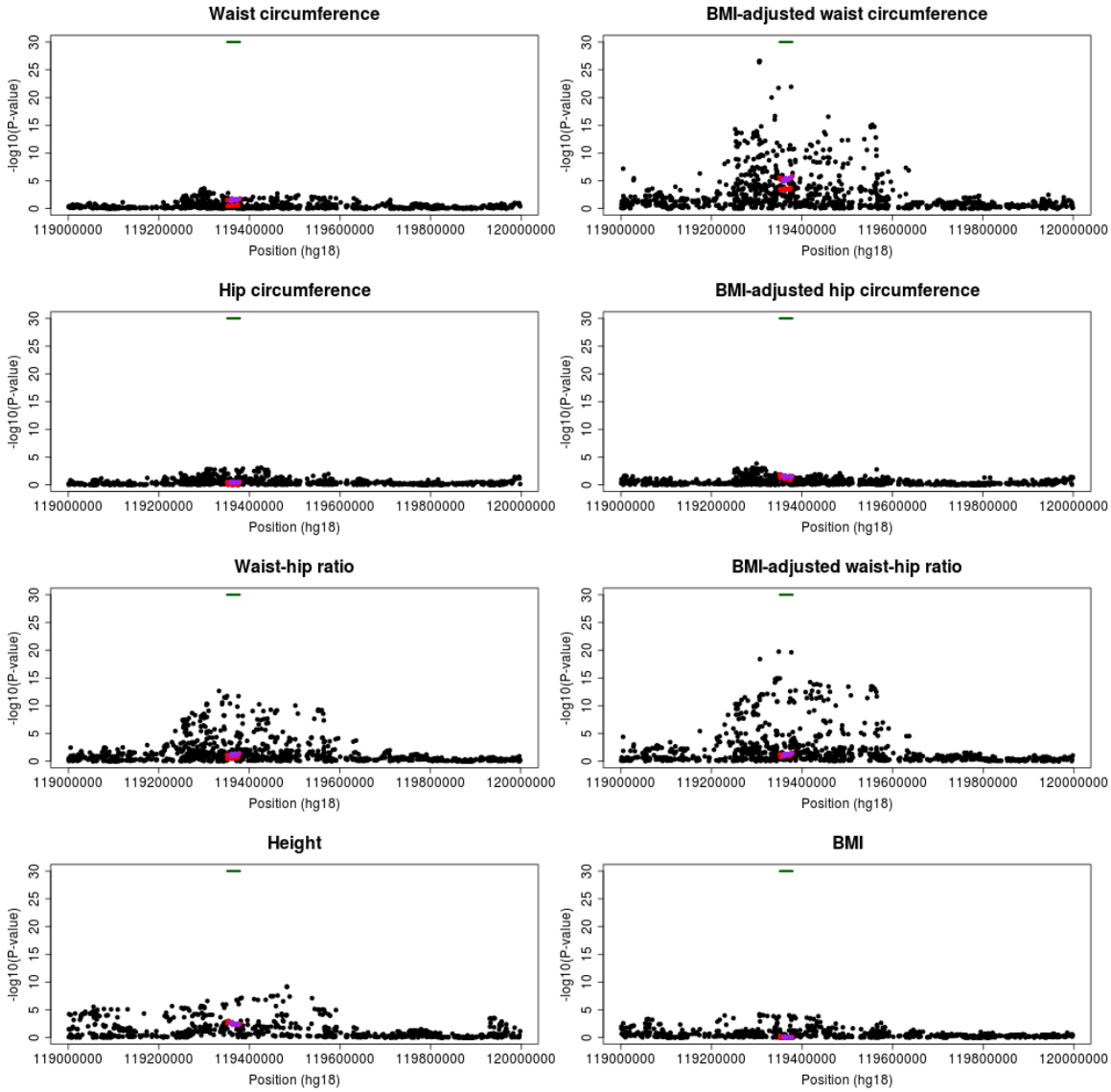

Figure 8. We queried the GIANT data and found that the introgressed region (top green tract) lies in the middle of an association peak for BMI-adjusted waist-circumference, waist-hip ratio and BMI-adjusted waist-hip ratio. Purple dots represent SNPs that are top PBS hits in Greenlandic Inuit (Table 1). Red dots represent additional SNPs where: a) the allele frequency difference between the closest non-African haplotypes to the Denisovan (defined as having less than 15 differences to the Denisovan SNPs) and the 20 non-African haplotypes that are farthest (in Hamming distance) from the Denisovan is >= 95%, b) where the allele that is at high frequency in the introgressed haplotype is the same as the allele observed in the Denisovan genome, and c) where the Denisovan allele is also at low frequency (< 1%) in Yoruba. The Denisovan alleles have positive effect sizes on these traits, but are not among the top significant SNPs.

Tables

CHR	POS	SNP ID	REF	ANC	NONSEL	SEL	GR FREQ	CHB FREQ	CEU FREQ	NEA	DEN	HG00436	HG00407	UST'-ISHIM	STUTTGART	LOSCHBOUR
1	119570095	rs2298080	G	A	G	A	0.997382	0.4545	0.1833	2	2	2	0	0	0	0
1	119571463	rs10923735	C	T	C	T	0.997382	0.4545	0.1833	2	2	2	0	0	0	0
1	119574289	rs12030972	G	G	G	A	0.997382	0.4545	0.1833	0	0	2	0	0	0	0
1	119574537	rs12026409	G	G	G	A	0.997382	0.4545	0.1833	0	0	2	0	0	0	0
1	119557151	rs10923726	A	A	A	G	0.997382	0.4659	0.1833	0	2	2	0	0	0	0
1	119557772	rs4659140	T	T	T	C	0.997382	0.4659	0.1833	0	0	2	0	0	0	0
1	119558112	rs12027501	C	C	C	G	0.997382	0.4659	0.1833	0	0	2	0	0	0	0
1	119558126	rs12027524	C	T	C	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119559155	rs12410034	A	T	A	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119559260	rs4659141	T	T	T	G	0.997382	0.4659	0.1833	0	0	2	0	0	0	0
1	119559297	rs4658995	C	C	C	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119559607	rs1325932	C	T	C	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119559793	rs1325931	A	C	A	C	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119559796	rs1325930	A	A	A	C	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119560089	rs10923730	C	C	C	G	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119560512	rs963171	A	T	A	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119561611	rs3951792	G	G	G	A	0.997382	0.4659	0.1833	0	2	2	0	0	0	0
1	119562014	rs12024063	G	G	G	A	0.997382	0.4659	0.1833	0	0	2	0	0	0	0
1	119562180	rs12031562	C	C	C	G	0.997382	0.4659	0.1833	0	2	2	0	0	0	0
1	119562274	rs12031597	C	C	C	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119562382	rs12021830	T	C	T	C	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119562460	rs10923733	A	A	A	G	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119562557	rs10494218	A	T	A	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119562716	rs10494219	T	C	T	C	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119562920	rs12402563	G	G	G	C	0.997382	0.4659	0.1833	0	2	2	0	0	0	0
1	119562932	rs12405154	C	C	C	T	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119563172	rs12025128	G	A	G	C	0.997382	0.4659	0.1833	2	2	2	0	0	0	0
1	119563638	rs113389819	G	C	G	C	0.997382	0.4659	0.1833	2	2	2	0	0	0	0

Table 1. The 28 SNPs in the 99.95% highest PBS quantile (testing for positive selection in Greenlandic Inuit) that lie in the introgressed tract. CHR = chromosome. POS = position (hg19). SNP ID = dbSNP rs ID number. REF = reference allele. ANC = human-chimpanzee ancestor allele (based on EPO alignments). NONSEL = non-selected allele. SEL = selected allele. GR FREQ = frequency of the selected allele in Greenlandic Inuit. CHB FREQ = frequency of the selected allele in CHB (Chinese individuals from Beijing). CEU FREQ = frequency of the selected allele in CEU (Individuals of Central European descent living in Utah). NEA = selected allele counts in the Altai Neanderthal genome. DEN = selected allele counts in the Denisova genome. HG00436 = selected allele counts in a present-day human genome that is homozygous for the introgressed tract. HG00407 = selected allele counts in a present-day human genome that is homozygous for the absence of the tract. UST'-ISHIM = selected allele counts in the Ust'-Ishim genome. STUTTGART = selected allele counts in the Stuttgart genome. LOSCHBOUR = selected allele counts in the Loschbour genome.

Tissue	TBX15 P-Value	TBX15 Effect Size	WARS2 P-Value	WARS2 Effect Size
Adipose - Subcutaneous	0.12	0.089	0.000098	-0.31
Adipose - Visceral (Omentum)	0.18	0.11	0.04	-0.2
Adrenal Gland	0.19	0.22	0.000066	-0.53
Artery - Aorta	0.22	0.12	0.00021	-0.33
Artery - Coronary	0.58	-0.077	0.0063	-0.35
Artery - Tibial	0.093	0.12	0.000035	-0.36
Brain - Amygdala	0.4	-0.18	0.16	-0.26
Brain - Anterior cingulate cortex (BA24)	0.67	-0.071	0.2	-0.26
Brain - Caudate (basal ganglia)	0.34	-0.15	0.044	-0.35
Brain - Cerebellar Hemisphere	0.84	0.039	0.081	-0.36
Brain - Cerebellum	0.89	0.021	0.0034	-0.47
Brain - Cortex	0.86	-0.026	0.094	-0.3
Brain - Frontal Cortex (BA9)	0.12	-0.23	0.012	-0.46
Brain - Hippocampus	0.99	-0.0019	0.15	-0.26
Brain - Hypothalamus	0.008	-0.36	0.19	-0.25
Brain - Nucleus accumbens (basal ganglia)	0.19	-0.25	0.099	-0.25
Brain - Putamen (basal ganglia)	0.83	0.032	0.93	-0.021
Brain - Substantia nigra	0.43	-0.25	0.17	-0.33
Breast - Mammary Tissue	0.22	-0.083	0.028	-0.25
Cells - EBV-transformed lymphocytes	0.089	0.14	0.47	0.098
Colon - Transverse	0.14	0.19	0.29	-0.12
Esophagus - Muscularis	0.17	0.13	0.00077	-0.31
Heart - Atrial Appendage	0.1	0.17	0.04	-0.25
Heart - Left Ventricle	0.24	0.12	0.029	-0.16
Liver	0.55	0.086	0.13	-0.19
Lung	0.009	-0.25	0.074	-0.16
Muscle - Skeletal	0.24	-0.054	0.00092	-0.22
Nerve - Tibial	0.22	0.06	0.23	-0.12
Ovary	0.18	0.29	0.15	-0.21
Pancreas	0.31	0.16	0.023	-0.23
Pituitary	0.57	-0.098	0.0013	-0.58
Prostate	0.0073	0.52	0.76	-0.064
Skin - Not Sun Exposed (Suprapubic)	0.87	0.0094	0.027	-0.27
Skin - Sun Exposed (Lower leg)	0.52	0.024	0.0016	-0.28
Spleen	0.58	-0.13	0.14	-0.32
Stomach	0.12	0.16	0.035	-0.17
Testis	0.00018	0.45	0.43	-0.11
Thyroid	0.22	0.078	0.016	-0.24
Uterus	0.25	0.23	0.9	-0.023
Vagina	0.64	-0.088	0.06	-0.34
Whole Blood	0.67	-0.031	0.61	-0.032

Table 2. Cis-eQTL p-values and beta (effect size) values retrieved from the test performed between rs2298080 and either *TBX15* or *WARS2*, in the GTEx pilot data, across different tissues. P-values in red are < 0.05 / number of tissues (Bonferroni multiple test correction). Effect sizes are with respect to the selected allele in Greenlandic Inuit.

Supplementary Figures

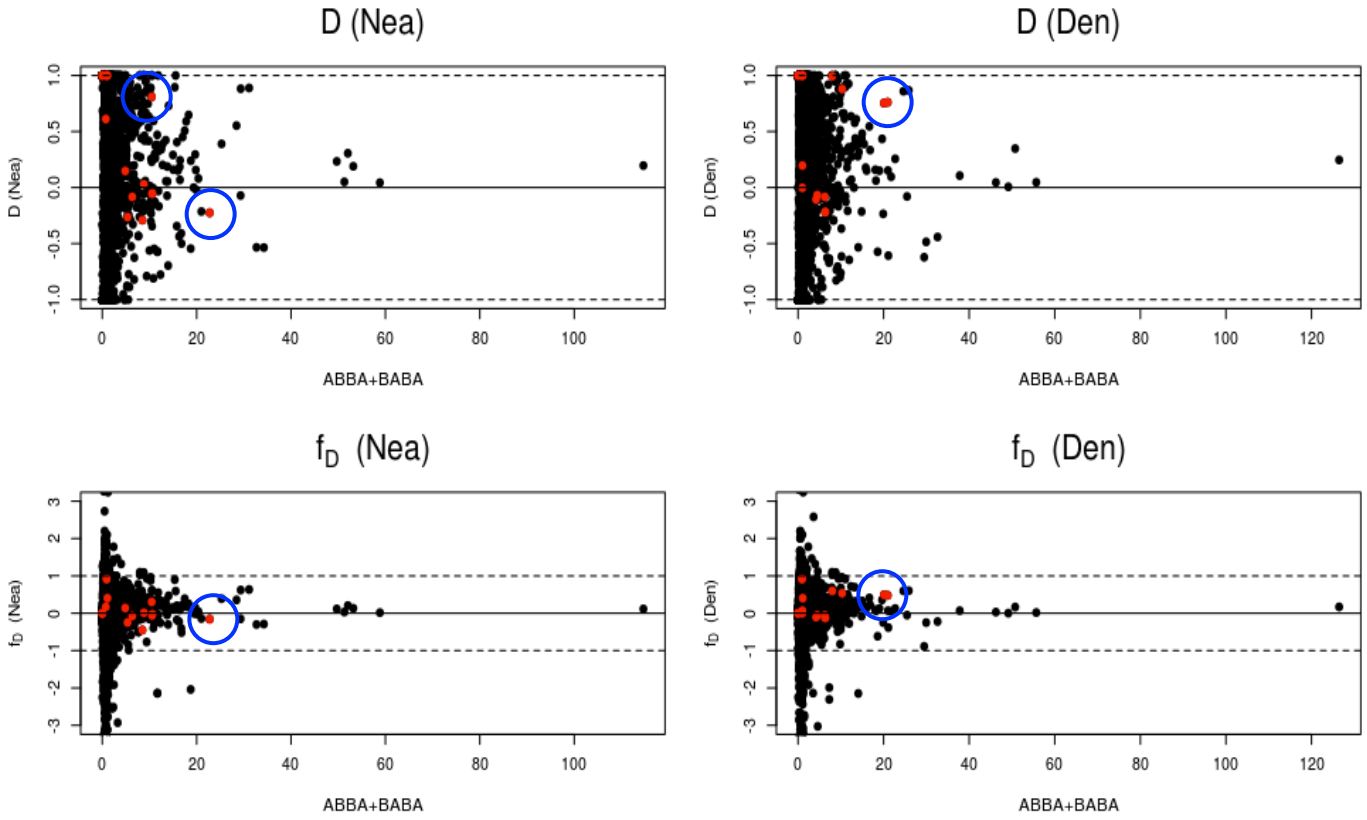

Figure S1. D and f_D plotted as a function of the denominator in the D statistic, (ABBA+BABA), for all genes with coverage in SNP chip data, plus a 20kbp region on either side of the gene. Points in red are the candidate genes for selection in Fumagalli et al (2015). The blue circles denote the TBX15 and WARS2 genes. The solid and dashed lines were drawn to denote the location in the plot where D and f_D are equal to -1, 0 and 1. We only show values of f_D that are larger than -3 and smaller than 3. D(X) = D(Yoruba,GI,X,Chimpanzee) and f_D (X) = f_D(Yoruba,GI,X,Chimpanzee).

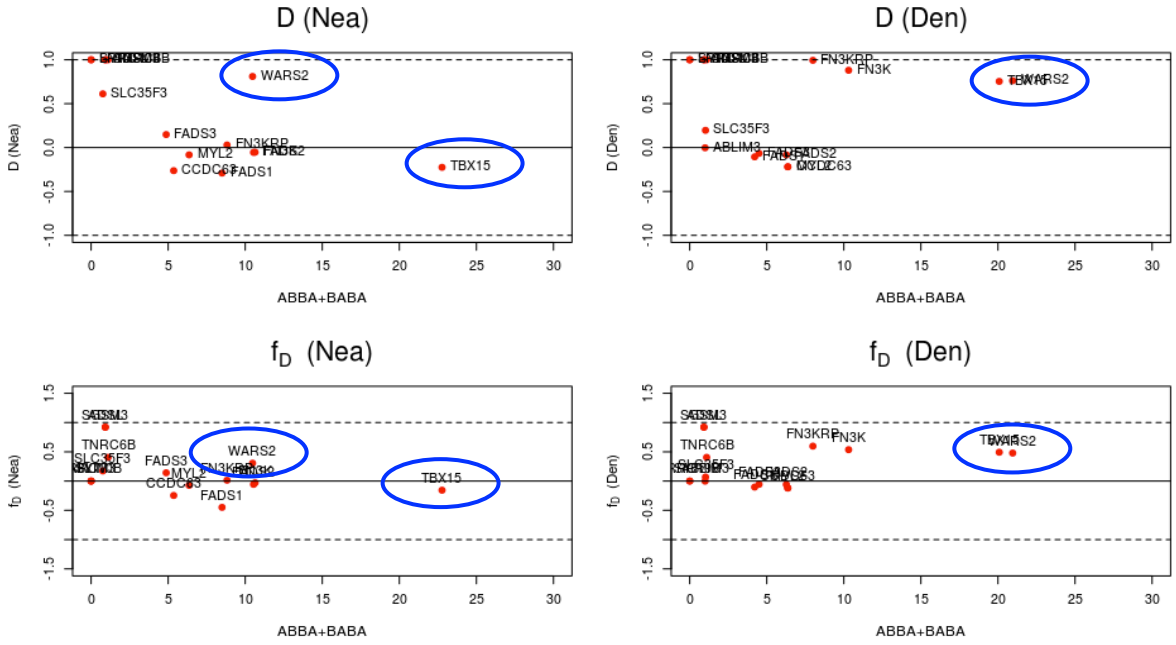

Figure S2. D and f_D plotted as a function of the denominator in the D statistic, (ABBA+BABA), for all candidate genes for selection in the SNP chip data of Fumagalli et al. (2015), plus a 20kbp region on either side of the gene. The blue circles denote the TBX15 and WARS2 genes. The solid and dashed lines were drawn to denote the location in the plot where D and f_D are equal to -1, 0 and 1.

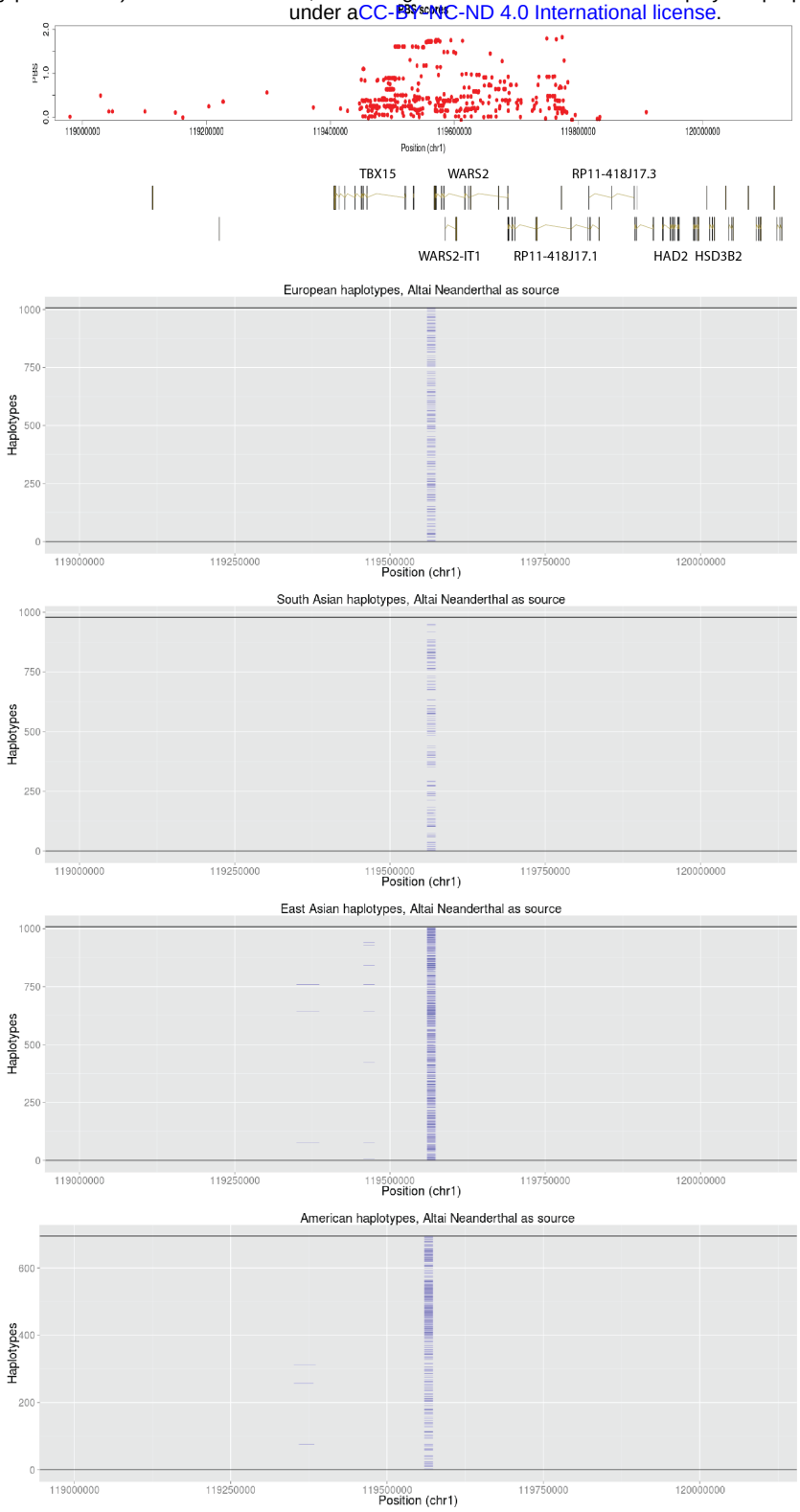

Figure S3. East Asian, South Asian, American and European phased chromosomes from the 1000 Genomes Project are inferred to contain archaic haplotypes (blue segments). Here, instead of using Denisova as the archaic source, we used the Altai Neanderthal, which results in archaic fragments of shorter size.

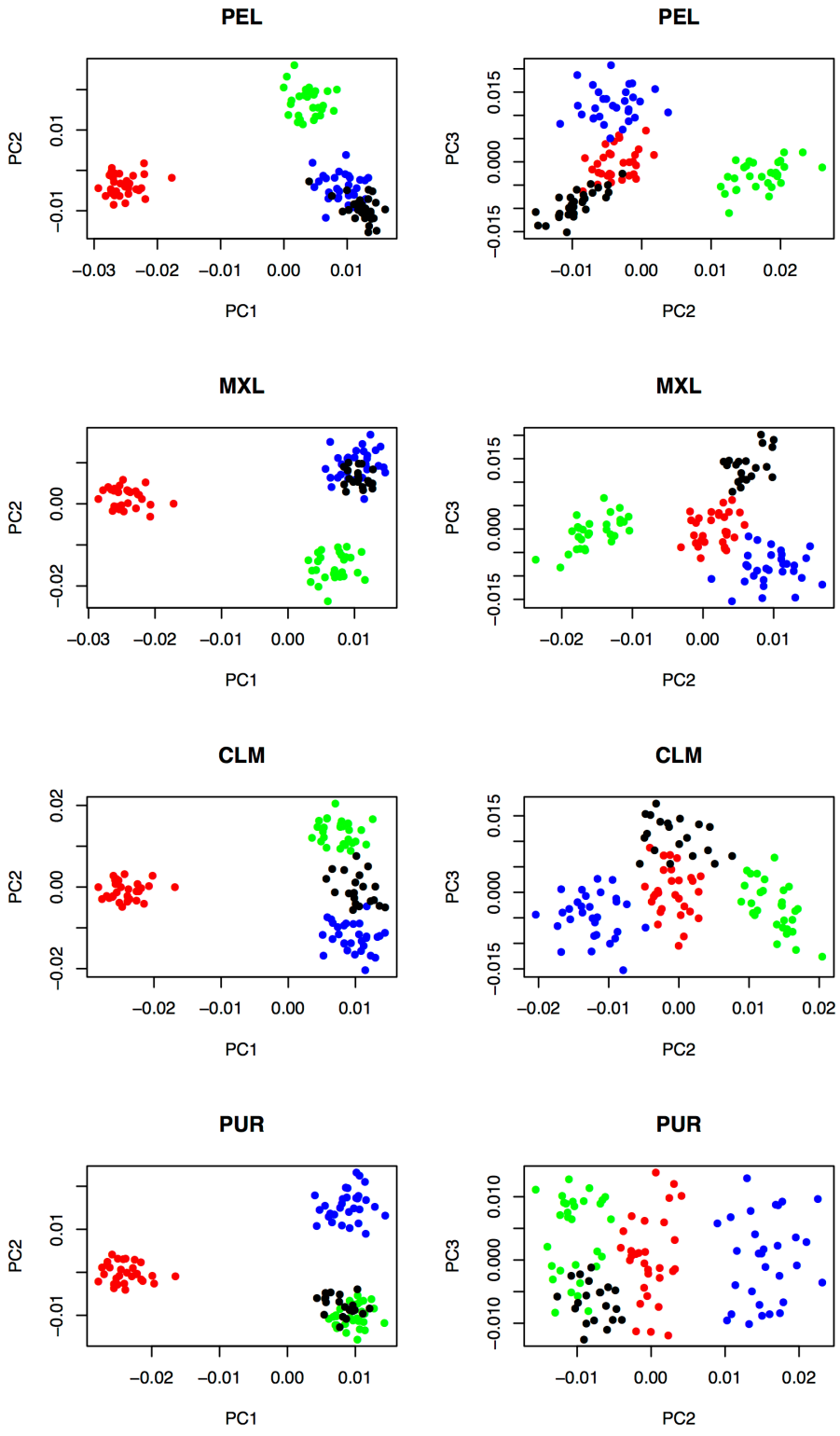

Figure S4. Multidimensional scaling plot for individuals of African (red), European (green), East Asian (blue) and Native American (black) descent used in this study. To represent Native Americans, we separately analyzed unadmixed individuals from Peru (PEL), Mexico (MXL), Columbia (CLM) and Puerto Rico (PUR). The first two components separate Native Americans from Africans and Europeans, while the third component separates them from East Asians.

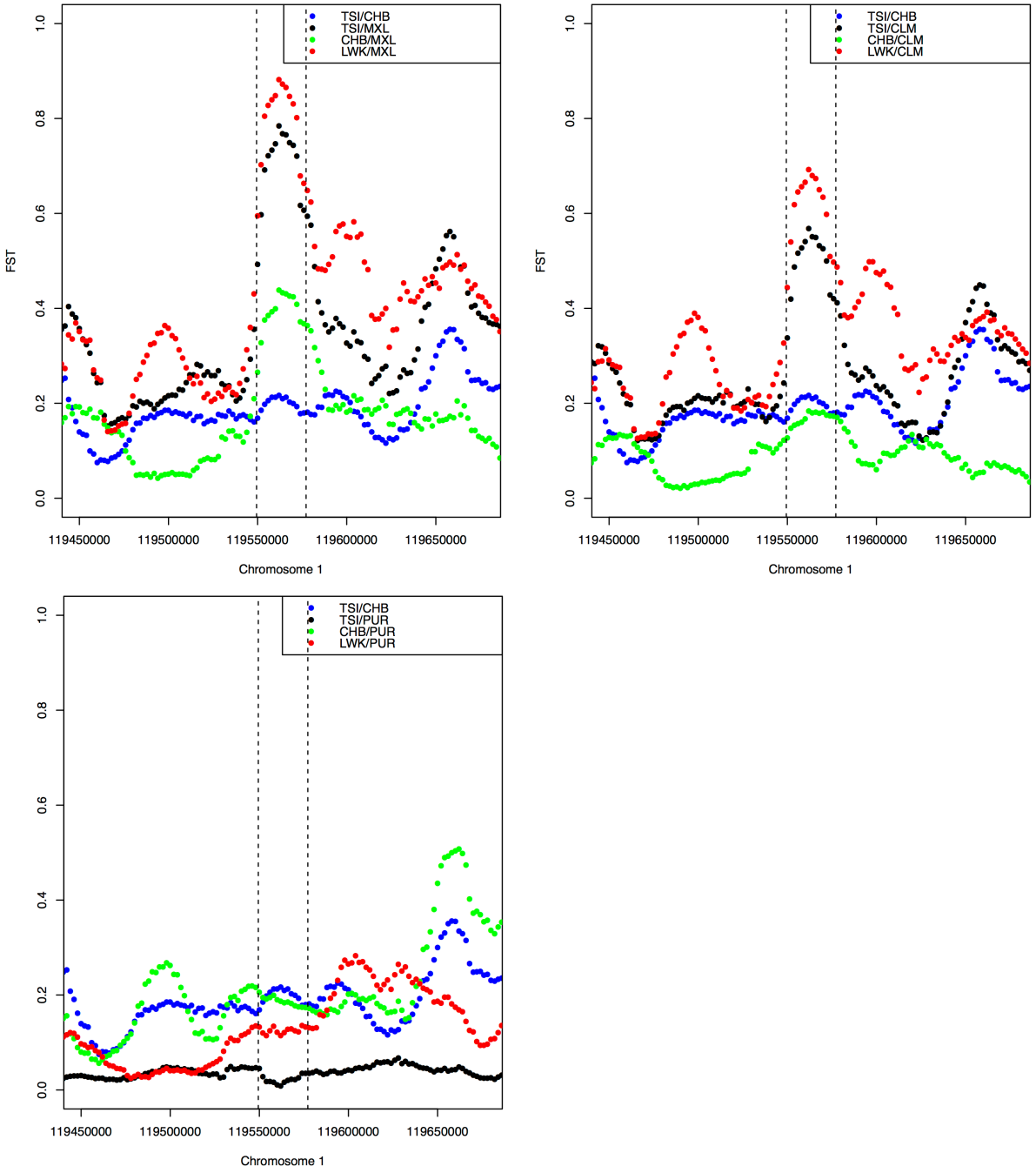

Figure S5. We computed local F_{st} values around the putatively introgressed haplotype using a sliding-windows approach, with window size of 20kbp and step of 2kbp. We tested Mexicans (MXL, left), Colombians (CLM, center) and Puerto Ricans (PUR, right) against different African and Eurasian populations (TSI, CHB, LWK). For MXL and CLM, we find a local increase in F_{st}, which is located exactly where the introgressed haplotype is inferred to be (dotted lines).

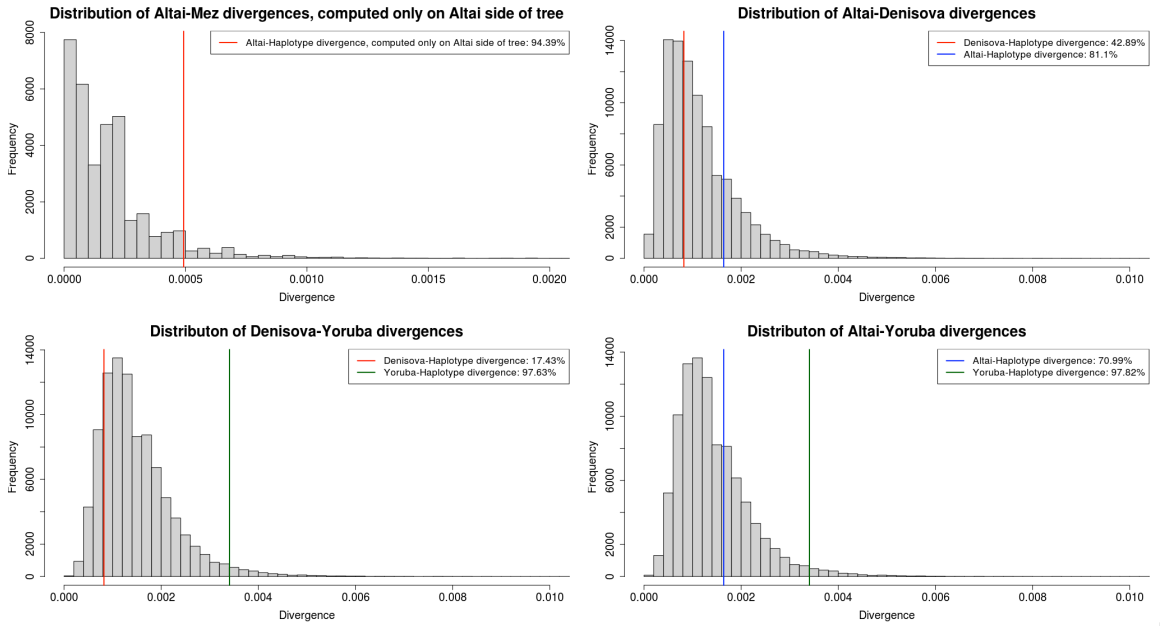

Figure S6. Upper-left panel: Genome-wide distribution of divergences between the Altai Neanderthal genome and the Mezmaiskaya Neanderthal genome, computed only on the Altai side of the tree, for regions of the genome of equal size as the introgressed haplotype. The red line shows the divergence between a genome that is homozygous for the introgressed haplotype (HG00436) and the Altai Neanderthal genome, computed only on the Altai side of the tree, and the quantile in which it falls within the distribution. Upper-rightl pane: genome-wide distribution of divergences between the Denisova genome and the Altai Neanderthal genome. The red and blue lines denote the divergence between the introgressed haplotype and Denisova and Altai Neanderthal, respectively. Lower-left panel: genome-wide distribution of divergences between the Denisovan genome and a high-coverage Yoruba genome (HGDP00936). The red and green lines show the divergence between the introgressed haplotype and the Denisovan and Yoruba genomes, respectively. Lower-right panel: genome-wide distribution of divergences between the Altai Neanderthal genome and HGDP00936. The blue and green lines show the divergence between the introgressed haplotype and the Altai Neanderthal and Yoruba genomes, respectively.

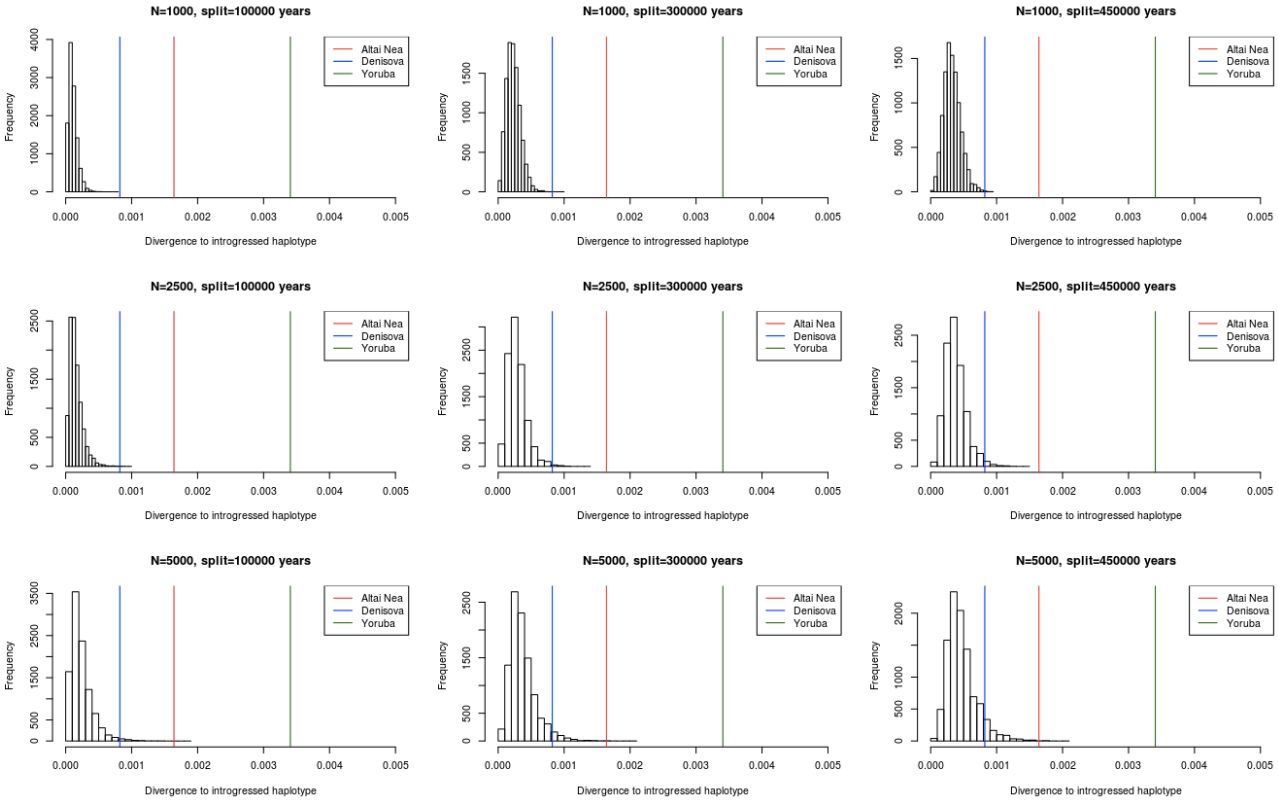

Figure S7. We simulated two populations that split at different times (100,000, 300,000 or 450,000 years), under a range of effective population sizes (1,000, 2,500, 5,000), based on estimates for the Neanderthal/Denisova population sizes from Prüfer et al. (2014). We compared these simulations to the divergence between the Altai Neanderthal genome and a human genome that is homozygous for the introgressed tract (HG00436) (red line), as well as to the divergence between the same human genome and the Denisovan genome (blue line), and the divergence between the same human genome and a Yoruba genome (HGDP00936) (green line).

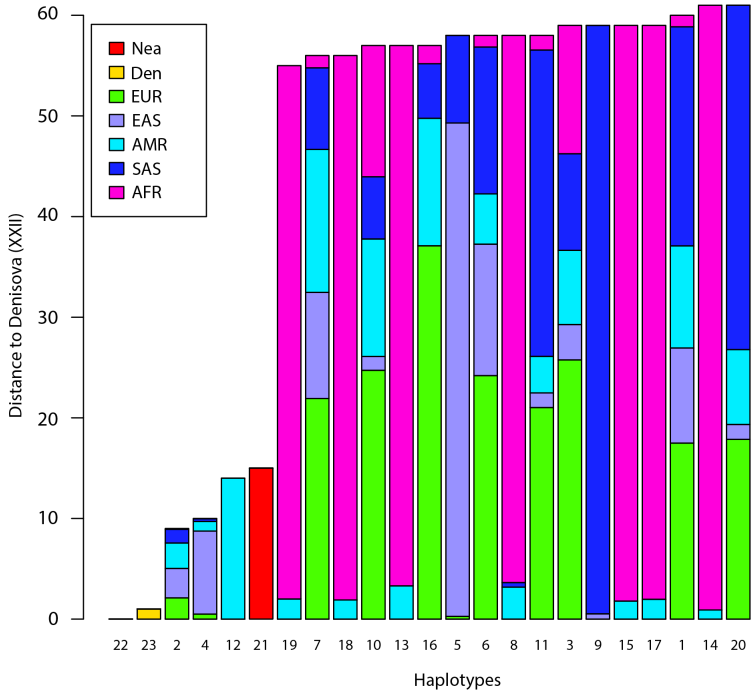

Figure S8. Hamming distance between Denisova haplotype XXII and other archaic and present-day human haplotypes, including the other Denisova haplotype (XXIII), the Altai Neanderthal haplotype (XXI) and the 20 most abundant present-day human haplotypes from the 1000 Genomes Project. The colors indicate the proportion of individuals from each human population that carries a given haplotype. AFR: Africans. AMR: Americans. EAS: East Asians. EUR: Europeans. SAS: South Asians.

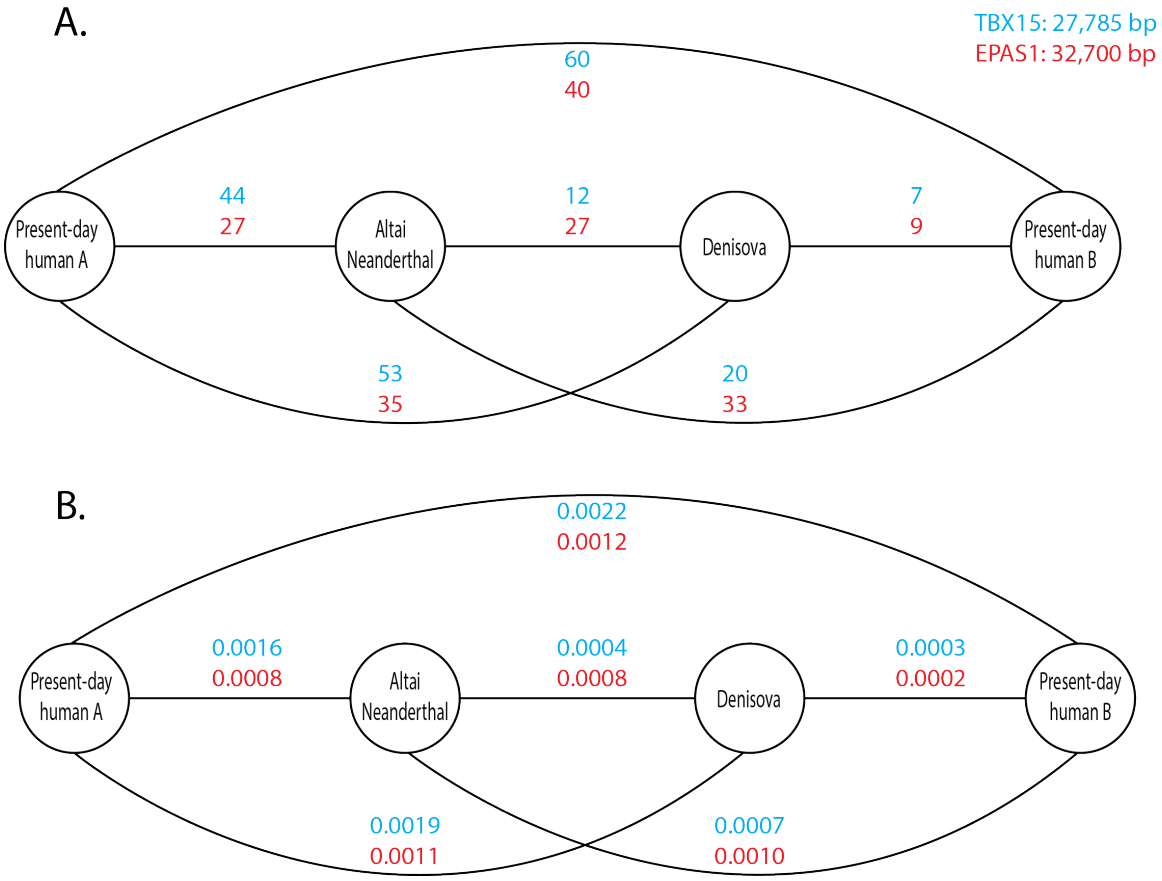

Figure S9. We compared the divergences between Denisova, Altai Neanderthal and the closest haplotype from each of the clusters in the *TBX15/WARS2* network to their corresponding values in *EPAS1*. A. Absolute number of differences. B. Sequence divergence (accounting for differences in length between the introgressed tracts in the two regions). We only used sites that were segregating in present-day humans, to make a fair comparison with published data for *EPAS1*.

Supplementary Tables

chr	start	end	gene	ABBA (Nea)	ABBA (Den)	BABA (Nea)	BABA (Den)	D (Nea)	D (Den)	f_D (Nea)	f_D (Den)
1	119395669	119562179	*TBX15*	8.83	17.61	13.94	2.47	-0.22	0.75	-0.16	0.49
1	119543839	119713294	*WARS2*	9.48	18.46	1	2.5	0.81	0.76	0.31	0.48
1	234010679	234490262	*SLC35F3*	0.61	0.61	0.15	0.41	0.61	0.2	0.17	0.07
4	10411498	10489034	*ZNF518B*	0	0	0	0	NA	NA	NA	NA
5	148491046	148670105	*ABLIM3*	0	0.5	0	0.5	NA	0	0	0
6	7511808	7616950	*DSP*	0	0	0	0	NA	NA	NA	NA
8	143923772	143991262	*CYP11B1*	0	0	0	0	NA	NA	NA	NA
11	61246272	61308482	*LRRC10B*	0	0	0	0	1	1	0	0
11	61252785	61378620	*SYT7*	0	0	0	0	1	1	0	0
11	61530452	61664826	*FADS2*	4.98	2.87	5.56	3.4	-0.05	-0.08	-0.06	-0.06
11	61537099	61626790	*FADS1*	3.01	1.89	5.48	2.33	-0.29	-0.11	-0.45	-0.1
11	61610991	61689523	*FADS3*	2.8	2.1	2.07	2.4	0.15	-0.07	0.14	-0.06
12	14926506	15089520	*C12orf60*	0	0	0	0	NA	NA	NA	NA
12	111254573	111375339	*CCDC63*	1.98	2.49	3.38	3.88	-0.26	-0.22	-0.25	-0.12
12	111318623	111388526	*MYL2*	2.92	2.49	3.44	3.88	-0.08	-0.22	-0.07	-0.12
17	80644559	80718204	*FN3KRP*	4.54	7.97	4.28	0.03	0.03	0.99	0.02	0.6
17	80663451	80739073	*FN3K*	5.04	9.7	5.6	0.62	-0.05	0.88	-0.03	0.54
19	10173014	10243472	*ANGPTL6*	0	0	0	0	NA	NA	NA	NA
19	16577122	16662163	*C19orf44*	0	0	0	0	NA	NA	NA	NA
19	16598700	16683341	*CHERP*	0	0	0	0	NA	NA	NA	NA
20	44432449	44501914	*SNX21*	0	0	0	0	NA	NA	NA	NA
22	40410821	40761811	*TNRC6B*	1.1	1.1	0	0	1	1	0.41	0.41
22	40712507	40794823	*ADSL*	0.92	0.92	0	0	1	1	0.92	0.92

Table S1. Differential archaic ancestry statistics calculated, using the SNP chip data, on the candidate genes for selection in Fumagalli et al. (2015). D(X) = D(Yoruba,Greenlandic Inuit,X,Chimp). f_D = f_D(Yoruba,Greenlandic Inuit,X,Chimp). Values with "NA" denote genes where a specific statistic has a denominator equal to 0.