Long-range regulatory effects of Neandertal DNA in modern humans

Danat Yermakovich[1], Vasili Pankratov[1], Urmo Võsa[1], Bayazit Yunusbayev[1,2], Estonian Biobank Research Team[1], Michael Dannemann[1,*]

[1] Institute of Genomics, University of Tartu, Tartu, Estonia

[2] ITMO University, SCAMT Institute, Saint-Petersburg, Russia

[*] Correspondence: michael.dannemann@ut.ee

Abstract

The admixture between modern humans and Neandertals has resulted in ~2% of the genomes of present-day non-Africans still being composed of Neandertal DNA. Association studies have shown that introgressed DNA still significantly influences skin and hair traits, immunity and behavioral phenotypes in people today. Several of the phenotype-associated archaic variants had links to regulatory effects as well. In general, analyses of allele-specific expression, regulatory sequence composition and cis-eQTL have demonstrated a significant contribution of this introgressed DNA to the transcriptomic landscape in people today. However, little is known about the impact of Neandertal DNA on trans-eQTLs - long-range regulatory effects that have been shown to explain ~20% of expression variation.

Here we used blood eQTL results from >30,000 individuals from eQTLGen Consortium. The cohort size allowed for a robust identification of trans-eQTLs and in addition enabled quantifying the role of transcription factors (TF) in mediating long-range regulatory effects. In our study we used this information to (i) annotate trans-eQTLs that are linked to Neandertal variants and (ii) predict long-range regulatory effects that are induced by Neandertal DNA by screening for the predicted target genes of TFs that are cis-eQTLs linked to Neandertal variants. We show that both trans-eQTL-associated Neandertal variants and those predicted to have long-range regulatory effects affect genes in genomic regions devoid of Neandertal DNA. In addition, both types of variants included candidates for local adaptation and show associations with autoimmune disorders, a severe Covid-19 phenotype, blood cell type composition and anthropometric measures.

Our results suggest that the regulatory reach of Neandertal DNA goes beyond the 40% of genomic sequence that it still covers in present-day non-Africans and that via this mechanism Neandertal DNA additionally influences the phenotypic variation in people today.

Introduction

The genome sequences of archaic humans such as Neandertals and Denisovans have shown that modern humans admixed with both archaic groups ~55,000 years ago [1–3]. As a result of these admixtures ~2% of the genomes of present-day non-Africans are composed of Neandertal DNA and an additional 2-5% of the genomes of Oceanian are derived from Denisovans [4]. Several studies have used information from genome-wide association studies to link archaic DNA to their potential phenotypic effects. These studies have shown that Neandertal DNA significantly influences skin and hair traits, immunity and behavioral phenotypes [5–8]. Many of the phenotype-associated archaic variants have been associated with regulatory effects as well, including several instances of Neandertal DNA with additional evidence for local adaptation [9,10]. In general, it has been demonstrated that both Neandertal and Denisovan DNA have significantly shaped the transcriptomic landscape in people today [8,11–14]. The analyses of the regulatory impact of introgressed archaic DNA have been based on cis-eQTLs, allele-specific expression and presence of archaic variants in regulatory sequence motifs. At the same time it has been demonstrated that regulatory sequences in modern human genomes also have been subject to negative selection against archaic DNA [15,16]. Both observations indicate that one major mechanism through which archaic DNA is influencing modern human phenotype variation is by its impact on gene expression regulation. These observations are perhaps not surprising given that the heritability of many phenotypes including disease is often mediated by gene expression [17].

However, previous studies have also shown that among eQTLs, cis-eQTLs only explain ~6% of gene expression variation [18]. In contrast, trans-eQTLs, i.e., long-distance regulatory effects, have been shown to explain with ~20% a substantially larger proportion of gene expression differences. The underlying mechanisms of the long-range impact of trans-eQTLs include transcription factor (TF) activation and chromatin-chromatin interactions that can modulate the expression of large sets of genes across the genome [19,20]. Given the polygenic nature of many complex traits and the growing evidence that many of those traits are regulated by complex gene expression networks, trans-eQTLs and their ability to impact many genes simultaneously might provide an important mechanism in understanding the molecular bases of polygenic phenotypes [21].

Large consortia like GTEx have revolutionized our understanding of the genetic basis of gene expression regulation [22]. Thousands of cis-eQTLs have been annotated in a diverse set of 49 tissues. However, little is known about the genetic bases of long-range regulatory effects. In fact, only 163 trans-eQTLs were detected in GTEx across all 49 tissues combined. This

discrepancy between the number of cis and trans-eQTLs can be explained by both the smaller effect sizes of trans-eQTLs compared to cis-eQTLs and the substantially larger number of variant-gene pairs tested in a genome-wide trans-eQTL screen, that also require higher statistical significance thresholds to account for multiple testing [18]. Among the 163 trans-eQTLs none of the underlying genetic variants show evidence of being of Neandertal ancestry. However, the limited number of annotated trans-eQTLs in GTEx makes it difficult to evaluate the significance of this observation. Recently, the eQTLGen consortium has published blood expression data from 31,684 samples, a dataset with a substantially increased association power compared to other existing datasets [23]. The authors provide a genome-wide cis-eQTL maps and a trans-eQTL screen for 10,317 trait-associated variants and show that 47% of significant trans-eQTLs could be linked to direct and indirect transcription factor (TF) activity.

In this study we particularly focussed on the trans-eQTL screen in eQTLGen and explored it for significant trans-eQTLs associated with introgressed Neandertal DNA. We investigated associated Neandertal variants for their molecular effects. Using this information, we attempted to reconstruct the underlying long-range regulatory mechanism with a particular focus on the role of TFs in facilitating these effects. Furthermore, we scanned GTEx and eQTLGen Neandertal-DNA-associated cis-eQTLs for those linked to TFs expression and used this information to predict potential trans-eQTL effects. Using TF target gene information, we then explored the potential genomic range and targets of TFs of eQTLGen and predicted trans-eQTLs that are associated with Neandertal DNA. We then sought to establish a connection of long-range eQTL variants to various disease and non-disease phenotypes. Finally, we tested for evidence for positive selection for our candidates in present-day human populations.

Materials and methods

eQTL datasets

We downloaded cis and trans-eQTL summary statistics from eQTLGen [23] and GTEx [22] [v8]. We annotated significant cis and trans-eQTLs based on an FDR cutoff of 0.05.

Annotation of introgressed Neandertal variants

We identified Neandertal variants in the eQTLGen, GTEx and UK and Japan Biobank cohorts based on a list of previously annotated introgressed Neandertal marker variants, referred to as aSNPs [24]. These aSNPs have been annotated in the 1,000 genomes project dataset (v5a) [24,25] and were defined based on the following conditions: (i) one allele is fixed in the African Yoruba population (ii) the second allele is present in a homozygous state in at least in one of the genomes of the Altai Neandertal, Vindjia Neandertal or Denisovan [2,3,26] and (iii) the allele in (ii) being present in at least one non-African 1,000 Genomes individual. We detected 616,647 autosomal aSNPs in the 1,000 Genomes and 151,388 aSNPs in 15 Papuans from the Simons Genome Diversity Project (SGDP) cohort [27]. In addition, we inferred archaic haplotypes for candidate trans-eQTL aSNPs as genomic regions between the left- and right-most aSNPs showing linkage disequilibrium of r^2>0.8 with the candidate aSNP in the 1,000 Genomes dataset. We required these inferred haplotypes to overlap with previously annotated haplotypes that have been shown to exceed a length that is not consistent with incomplete lineage sorting (ILS), a genomic phenomenon that can also result in allele sharing properties described in (i-iii).

Sequence similarity of inferred archaic haplotypes with archaic humans

We compared the sequences of haplotypes in each archaic candidate region in all 1,000 Genomes individuals and the 15 SGDP Papuans by calculating the nucleotide differences to the genome sequences of each the Altai Neandertal, Vindija Neandertal and Denisovan. We restricted this analysis to all biallelic positions for which the genomes of the three archaic humans showed a homozygous genotype.

Putative molecular consequences

We functionally annotated aSNPs associated with haplotypes for the 13 trans-eQTL candidates based on computational and experimental information. First, we used ENSEMBL's variant effect predictor and RegulomeDB to assign potential molecular consequences of the 446 aSNPs residing on the candidate haplotypes (RegulomeDB; [28], VEP, [29]). Information on

aSNPs with predicted regulatory or protein sequence changing effects can be found in Tables S1 and S2. In addition, we leveraged information from two studies that tested for expression-modulating effects of aSNPs using massively parallel reporter assays [30,31]. These variants included 5,353 high-frequency Neandertal variants, from which 2,548 showed potential as cis-regulatory elements (CREs) and 292 altered the regulatory potential (emVar) in one study [31] while the other included 613 variants (including 290 aSNPs) with 327 CREs and 20 emVars. We intersected our set of 446 aSNPs with the set of reported variants (Table S3, S4).

TF prediction databases

We leveraged TF target gene prediction information from seven databases. The target gene information in these databases is based on computational prediction algorithms (MotifMap [32], Human TFTG [33], JASPAR [34], TRANSFAC predicted [35]) and experimental information (ENCODE [36], CHEA [37], TRANSFAC curated [35]). The numbers of evaluated TFs and predicted targets varies substantially between these databases (Table S5). In order to ensure comparability between information from these databases we equalized gene names from all databases to common GeneSymbols, collapsed duplexes and limited the analysis to protein coding genes. Human TFs without target gene information in any of these databases were annotated based on AnimalTFDB3.0 [38].

Phenotype databases

We extracted phenotypic information for both trans-eQTL associated aSNPs in eQTLGen and predicted ones from GTEx cis-eQTL associated TFs from EBI GWAS catalog (https://www.ebi.ac.uk/gwas/) [39], Immunobase, a blood cell traits study [40] and the Biobank Japan for associations with candidate aSNPs (Table S6, S7). We considered all phenotype associations with $P<5x10^{-8}$.

Definition of archaic deserts

We used information on shared archaic deserts of Neandertal and Denisovan introgression from the studies of Sankararaman et al. and Vernot et al. [1,4]. We considered all autosomal deserts reported by both studies and collapsed overlapping deserts into one by using the range of the combined regions. That resulted in five deserts that were defined by the following hg19 coordinates: chr1:99Mb-114.9Mb, chr3:76.5Mb-90.5Mb, chr7:108Mb-128Mb, chr8:54.5Mb-65.4Mb and chr13:49Mb-61Mb. We annotated genes as desert genes if they were located within the desert's borders.

Transcription factor enrichment analysis for hub trans-eQTLs and archaic desert genes

We tested for an over-representation of TF target genes sets within the sets of genes for each of the two hub trans-eQTL candidates (*rs72973711* and *rs13098911/rs13063635*) and archaic desert genes using prediction information from all seven TF databases. We performed two-sided Fisher's Exact tests for all combinations and retrieved Odds ratios and P values. For these analyses we used all available protein-coding target genes for a given database as a background gene set. We corrected the obtained P values for multiple testing using the Benjamini–Hochberg method.

Evidence for natural selection acting on trans-eQTL aSNPs

To estimate changes in the allele frequency over the last 500 generations and to test whether these changes might have happened under natural selection we applied CLUES [41] to 16 significant trans-eQTL aSNPs. CLUES for each SNP was run as in Marnetto et al. [42]. Briefly, we started with building local trees by applying Relate (version 1.1.4)[43] to whole genome sequences of 2,420 Estonian Biobank participants (dataset is described in [44,45]). When running Relate for tree building we used the strict callability mask, recombination map and the reconstructed human ancestral genome all generated based on the 1,000 Genomes Project (GRCh37), a mutation rate of 1.25×10^{-8} and an effective population size of 30,000. Next, we extracted subtrees corresponding to a subset of 1,800 individuals by removing individuals that were related (relatives up to 3rd degree), PCA or singleton count outliers and individuals with excessive IBD sharing (≥166.2 cM) with 2 or more other individuals in the dataset. This step was implemented to run CLUES on a more homogeneous dataset and to reduce run time. Coalescent rate over time for the Estonian population needed for CLUES was estimated based on 100 randomly sampled individuals. When running CLUES we extracted the local tree with the SNP of interest and re-sampled its branch length 200 times. We then removed the first 100 samples and took every 5th tree (20 in total) for importance sampling. If a trans-eQTL aSNP was not passing the applied mask, was not mapped to a local tree or was flipped by the Relate algorithm we used a proxy aSNP with the highest r^2 to the trans-eQTL choosing the physically closest SNP if there were more than one with the same r^2 value. To account for the variation in the results coming from the uncertainty in branch length estimation we ran the analysis (starting from sampling branch length) twice for each aSNP and averaged frequencies and log likelihood ratios. For the three pairs of SNPs in strong linkage we ran CLUES for each SNPs and then averaged the estimates within each pair. In addition we also ran CLUES as described above for

2,829 (out of 3,853) trans-eQTLs being present on our local trees and having MAF >= 5% in 2,420 Estonians once for each SNP.

Results

Trans-eQTLs in eQTLGen with a link to Neandertal variants

A total of 3,853 of the 10,317 trait-associated SNPs that have been assessed for trans-eQTL effects in eQTLGen showed a significant long-range regulatory effect on one or multiple genes [23]. We found that 16 of these 3,853 genetic variants were linked to SNPs that are of likely Neandertal ancestry. We will refer to these SNPs as aSNPs throughout the manuscript. Such aSNPs have previously been annotated in 1,000 Genomes Project individuals based (i) on their allele-sharing patterns of variants that are absent in the African Yoruba population and found in both Neandertals and non-Africans and (ii) their link to a haplotype that is not compatible with incomplete lineage sorting, another genomic phenomena that can lead to a similar sharing pattern (Materials and methods). Among the 16 Neandertal-linked trans-eQTLs were 3 pairs of SNPs, showing high levels of linkage disequilibrium among them (r^2>0.5 between the two aSNPs for all three pairs, Materials and methods). We have collapsed these three pairs, reducing the numbers of independent Neandertal trans-eQTL associations to 13 (Table S8). Ten of these trans-eQTLs were associated with a significantly altered expression of one target gene, while the three other trans-eQTLs were linked to expression changes of multiple genes with 2, 27 and 34 target genes respectively. Notably, the archaic alleles for the two trans-eQTLs with 27 and 34 target genes were associated with directional impacts on expression (Table S8, Figure 2A-B). In the presence of the archaic allele at the given loci, 67% (18/27, P=0.12, Binomial Test) of genes associated with *rs72973711* subjected to an expression-increasing effect, while 78% (28/34, P=2.0×10^{-4}, Binomial Test) of genes related to *rs13098911/rs13063635* region show lower expression levels.

Archaic source of introgression for trans-eQTL aSNPs

With the close genetic relationship of Neandertals and Denisovans, it is possible that trans-eQTL aSNPs represent introgressed haplotypes from either of the two archaic groups. To assess the likely archaic source population we first defined introgressed haplotype based on the region that other aSNPs span that are in LD of r^2>0.8 with a given trans-eQTL aSNP and then analyzed the sequence similarity of those haplotypes to the genome sequences of both Neandertals and the Denisovan (Materials and methods). We found that haplotypes for all 13 trans-eQTL candidates showed a closer sequence relationship with Neandertals (Figure S1). Six of the haplotypes showed a similar sequence to both the Altai and Vindija Neandertal individuals, six haplotypes showed a closer relationship to the Vindija individual while one had a

more Altai Neandertal-like sequence. This observation is consistent with a closer genomic sequence relationship of the introgressing Neandertal population with the Vindija individual [2].

There are two potential factors that could lead to a mis-classification of Denisovan haplotypes as Neandertal: 1) incomplete lineage sorting among archaic humans; 2) the lower sequences similarity between the sequenced Denisovan and the introgressing Denisovan population compared to the sequenced and introgressing Neandertals [1]. We therefore explored two additional features of our 13 candidate haplotypes to further explore their origin. First, we studied the presence of these haplotypes in different present-day populations. We found that all haplotypes were present (frequency>1%) in the majority of European and South Asian 1,000 Genomes populations (Figure 3). Similarly, all but two haplotypes could also be detected in East Asians. With Denisovan DNA in present-day people being mostly restricted to Oceanians and low levels in some Asians [46], the wide spread of candidate haplotypes in diverse Eurasian populations is another indicator of their Neandertal origin.

Finally, we have explored the genome sequences of 16 Papuan individuals from the Simons Genome Diversity Project (SGDP) for evidence of additional Denisovan haplotypes in the 13 trans-eQTL regions. We found evidence for archaic haplotypes in three regions and no evidence for any difference between the aSNP composition of those haplotypes compared to the reconstructed haplotypes in the 1,000 Genomes dataset.

These results suggest that archaic haplotypes for all our trans-eQTL aSNPs are of likely Neandertal origin.

Impact of trans-eQTL associated aSNP on gene expression regulation and protein sequence

The sparse trans-eQTL screen provided by eQTLGen prohibits a robust assessment of whether the archaic variants at a given locus are the likely driver of the trans-eQTL signal or whether any non-archaic variation in linkage disequilibrium is showing substantially stronger association results. And even in the case of archaic variants in a given region showing the top association wouldn’t necessarily imply causality [47]. Another factor that is particularly pronounced for introgressed DNA from archaic humans is their extended haplotype structure. Due to the time of admixture ~55,000 years ago Neandertal variants are found on haplotypes with tens or even hundreds of aSNPs in high LD, spanning over several tens of kilobases [48]. In order to (a) provide additional evidence that trans-eQTL aSNPs or aSNPs in LD with them show regulatory potential, and (b) narrow down the aSNPs on a given haplotype that are the most likely candidates for such regulatory effects we used computational effect prediction tools

and experimental data from reporter assays to annotate aSNPs on Neandertal haplotypes that are linked to our 13 trans-eQTL candidates.

We first used ENSEMBL's variant effect predictor (VEP) [29] to annotate the effects of trans-eQTL-associated aSNPs on the gene and protein sequence as well as on regulatory regions. We found that among the 446 aSNPs that are associated with archaic haplotypes of our 13 trans-eQTL candidates were 130 variants with a potential regulatory function, including 10 aSNPs that modify the TF binding site (TFBS) sequence, 14 that are located in 3' and 5' UTRs and 106 aSNPs in other regulatory regions. In general, 11 Neandertal haplotypes associated with our 13 candidates carry at least one of these putative regulatory aSNPs, providing additional evidence for links between trans-eQTL associated aSNPs and different molecular processes (Table S1). In addition, we found four aSNPs that altered the amino acid sequence, suggesting that some of the molecular consequences of the trans-eQTL-associated Neandertal variants also affect the protein sequence. Three of those aSNPs were linked to the same trans-eQTL haplotype associated with *rs2066807/rs2066819* on chromosome 12. The aSNPs associated with that haplotype modified the protein sequence of *ANKRD52* (*rs59626664*) and STAT2 (*rs2066807*) as well as the start codon for *COQ10A* (*rs60542959*) (Figure 1A-B). An additional aSNP (*rs13079478*) linked to the trans-eQTL on chromosome 3 (*rs13063635/rs13098911*) affected the amino acid sequence of *FYCO1*.

The results from VEP were consistent with those we obtained from RegulomeDB [28]. A total of 42 of the 446 trans-eQTL Neandertal haplotype aSNPs were categorized as 'likely to affect binding', 17 of which with an additional link to the expression of a target gene. These 17 aSNPs were associated with four haplotypes that were also classified to carry putative regulatory variants using VEP. An additional 46 aSNPs associated with these four haplotypes and four additional trans-eQTL candidate regions showed evidence for binding affinity as well, but to a lesser degree than the 42 above mentioned aSNPs (Table S2).

Finally, we explored two studies that have tested both a set of cis-regulatory aSNPs and aSNPs that have been associated with a severe Covid-19 phenotype, respectively for their expression-modulating potential in an immune cell line [30,31]. A total of 151 of the 446 aSNPs that were associated with our trans-eQTL haplotypes have been included in the experimental setup by those studies. The 151 aSNPs were distributed across six of our 13 candidate haplotypes with 126 of them being linked aSNPs on chromosome 3 (*rs13063635/rs13098911*) and an additional 20 of them were linked to aSNPs of the trans-eQTL on chromosome 12 (*rs2066807/rs2066819*, Figure 1B). Sixtynine of the 151 tested aSNPs have been assigned as active cis-regulatory variants in the tested cell lines and for six of them the archaic allele has

been shown to also significantly alter the expression level. Three of these SNPs fall onto the archaic haplotypes on chromosomes 3 and 12, respectively (Table S3, S4).

Regulatory mechanism of trans-eQTL aSNPs

Next, we sought to explore whether we can reconstruct the putative mechanisms through which the long-range regulatory effects of our trans-eQTLs are modulated. It has been shown that in eQTLGen up to 47% of the trans-eQTL activity can be explained by direct and indirect regulation involving a TF and additional co-regulatory effects [23]. Indeed, three of our 13 candidates have been assigned to such mechanisms by the study, including one of the trans-eQTL with the largest number of target genes (tag aSNPs *rs13098911/rs13063635*), which have been linked to be co-regulated by *CXCR6* and *CCR3*, two receptors that play a role in immune response among other functions [49,50]. Another candidate trans-eQTL (aSNP: *rs16997087*) associated with the modified expression of *ITGB3BP* has been linked to a complex indirect mechanism involving a co-regulation via *MARCOD2*, which itself has been modulated by a TF that is co-regulated with a gene that is in close proximity to the trans-eQTL aSNP (Table S9). Finally, two aSNPs in high LD (*rs2066807/rs2066819*) have been linked to an indirect regulation of the TF *STAT6* and a subsequent effect on the reported trans-eQTL genes.

In an attempt to further uncover potential regulatory mechanisms for our Neandertal-linked trans-eQTL candidates, we explored TF cis-eQTLs that were associated with aSNPs residing on trans-eQTL candidate loci haplotypes. We then tested for a direct connection between the TF and the trans-eQTL target genes using TF target prediction information from seven databases (CHEA, ENCODE, Human TFTG, JASPAR, MotifMap, TRANSFAC curated and predicted; Materials and methods). These databases report TF target information from different techniques, including computational binding motif-based approaches or experimental data such as Chip-Seq to generate their TF target predictions. The number of TFs with available information varies widely between databases and lies between 111 (JASPAR) and 671 (Human TF-TG). In total these databases provide target gene information for 866 TFs, which is almost half of the 1,791 annotated TF in the human genome (Materials and methods). Due to the different algorithms the sensitivity between these databases is expected to differ as well. In total we found 17 TFs that were in testable distance for the eQTLGen cis-eQTL screen from our 13 trans-eQTL candidates. Four of these TFs (*KLF3*, *BLZF1*, *STAT2* and *STAT6*) showed a significant cis-eQTL association (referred to as cTFs from hereon) with aSNPs that were also associated with three trans-eQTL candidates (*KLF3*: *rs6531656*, *BLZF1*: *rs10919070/rs10919071* and *STAT2/STAT6*: *rs2066807/rs2066819*, FDR<0.05, Table S10).

Unfortunately none of the prediction databases provided target gene information for *KLF3* or *BLZF1*, therefore not allowing us to assess whether trans-eQTL genes were putatively targeted by these cTFs. However, for our third locus (*rs2066807/rs2066819*) we found evidence for a potential regulatory link between both cTFs *STAT2* and *STAT6* and *IFI16* - one of the two trans-eQTL genes (ENCODE, Figure 1D). This observation is also consistent with the inference of Võsa & Claringbould et al. who report *STAT6* as being an indirectly affected TF. Nevertheless, using the TF target prediction data provides additional complexity. First, *IFI16* is a target of both *STAT2* and *STAT6*. Second, *STAT2* and *STAT6* are also predicted to target each other (TRANSFAC and MotifMap). Notably, *STAT2* showed significantly lower expression in the presence of the archaic alleles ($P=3.3x10^{-122}$, Z-Score=-23.5, Table S8), while *STAT6* showed the opposite regulatory direction ($P=2.6x10^{-7}$, Z-Score=5.2). Both TFs have been shown to function as transcription activators. The lower expression of *IFI16* in the presence of the archaic alleles suggests that a direct regulation by *STAT2* is more consistent with these observations. However, the potential connectedness between these three genes would therefore also be compatible with other more complex interaction scenarios.

In addition, trans-eQTL loci with large numbers of affected genes provide a second inroad to test for potential links with TFs that might be involved in the regulatory network underlying the trans-eQTL. Two of our candidates showed large numbers of trans-eQTL genes (27 and 34 trans-eQTL genes of *rs72973711* and *rs13098911/rs13063635*). Again, using information from the seven TF target prediction databases we tested for an enrichment of these two gene sets among the predicted target genes of TFs. The 27 genes linked to *rs72973711* showed the strongest enrichment results among predicted target genes of *ISX* ($P=9.42x10^{-5}$, FDR=0.20, OR=5.4, Human TF-TG, Figure 2C). Unfortunately, *ISX* is only tested in one of the seven databases, hindering us from collecting additional evidence for the robustness of this inference. A second notable TF was *NANOG*, which was tested in three databases (ENCODE, CHEA, Human TF-TG) and showed a consistently larger overlap of predicted targets and the 27 trans-eQTL genes (ORs>1 for all tested databases, Materials and methods). The results in two of those databases showed significant enrichment P values as well (OR=4.4, P=0.02 for Human TF-TG and OR=3.5, P=0.002 for ENCODE). Trans-eQTL genes linked to the second multi-gene trans-eQTL *rs13098911/rs13063635* showed the most significant overlap with target genes of *STAT6* (OR=12.3, P=5.7x10-6, FDR=0.01, CHEA) - the only result with FDR<0.05 among all tested sets of TFs. Target gene information for *STAT6* was also available in three other prediction databases (curated version of TRANSFAC, MotifMap, Human TF-TG) but showed no significant enrichment results in any of those (P>0.5 in all and OR>1 in ⅔ databases). However,

it was noteable to observe the recurrent link to *STAT6* for this trans-eQTL gene set, a TF which was already shown to be a relevant component for the regulatory mechanism of the previously discussed trans-eQTL candidates associated with *rs2066807/rs2066819*.

Prediction of Neandertal-linked trans-eQTL effects

The limited number of SNPs tested for trans-eQTL effects in eQTLGen and the reduced power to test for trans-eQTLs in smaller datasets like GTEx prevents us from directly associating Neandertal variants with long-range regulatory effects on a genome-wide scale. However, the observation of the involvement of TFs to initiate and facilitate long-range regulatory effects allows us to use this information to predict potential genomic regulatory reach of Neandertal DNA. In this study, we focussed on one particular mechanism: the effect of Neandertal variants on nearby TFs expression and on the predicted targets of the affected TF. Overall, up to 9% of the trans-eQTL effects observed by Võsa & Claringbould et al. can be linked to this particular mechanism, with another 38% involving TFs at some other stage in the regulatory chain. In order to identify TFs that show evidence for a regulatory link to nearby Neandertal DNA we first scanned the eQTLGen dataset for cTFs that were linked to aSNPs. We found that a staggering 441 of the 1301 tested TFs showed a significant cis-eQTL (FDR<0.05, Table S11) that was linked to an aSNP. That number was in stark contrast to aSNP-linked cTFs in GTEx (v8, Materials and methods). Here, we found that across 49 diverse tissues a total of 65 TFs were aSNPs with significant cis-eQTL in at least one tissue (FDR<0.05, Table S12). Most of these TFs were found to be significantly regulated in only one (46 TFs) or two (11 TFs) tissues. Conversely, five cTFs were found in more than 6 tissues. *ZNF143* and *ZNF189* showed significant cis-eQTLs in 17 and 16 tissues respectively, including diverse sets of tissues such as the heart, arteries, blood, skin, adipose tissues, intestine and the brain (Table S12). For both genes we observed a consistently lower expression in the presence of the archaic allele. Among the other three TFs with cis-eQTL in multiple tissues was *STAT2* with a total of seven tissues with cis-eQTLs linked to aSNPs. The affected tissues with such regulatory effects included two adipose tissues, liver, skin, nerve, cerebellar and artery tissues, with five of these tissues showing a higher expression of *STAT2* in the presence of the archaic alleles (Figure 1E). In general, at least one significantly Neandertal-linked cTF was found in 40 of the 49 tissues. The tissue with the most active TFs was skeletal muscle (8 TFs) followed by lung, adipose (subcutaneous), skin, thyroid (all 7 TFs) and blood (6TFs).

Thirtythree of the 65 Neandertal-linked GTEx cTFs overlapped with those found in eQTLGen. The overlapping cTFs showed significantly larger effect sizes among eQTLGen cTFs

(Mann-Whitney-U test, $P=3.3x10^{-7}$) with 28 of them being in the top half in effect size, and 13 of those even within the top 10%. In general, the correlation of effect sizes between GTEx and eQTLGen was high (Spearman's $\rho=0.66$, $P=1.3x10^{-13}$) and highly consistent (78.0% showed the same expression direction, Figure 4). The eQTLGen data also included data from GTEx blood expression data. When removing GTEx blood cTFs from the analysis the correlation remains comparable (Spearman's $\rho=0.66$, $P=2.8x10^{-13}$, 79% shared expression direction), suggesting that the correlation is not primarily driven by the partially shared data in eQTLGen and GTEx.

These results are consistent with the higher statistical power in eQTLGen to detect more subtle regulatory effects than is possible in a smaller cohort like GTEx. The results also suggest that Neandertal DNA can be linked to regulatory effects on a substantial number of TFs and therefore potentially be involved in a wide range of downstream regulatory effects.

The regulatory reach of eQTLGen and predicted trans-eQTLs

We next analyzed predicted target genes of Neandertal-linked cTFs in eQTLGen and GTEx. Target gene prediction information across seven prediction databases was available for 30 of 65 GTEx cTFs and 185 of 441 cTFs in eQTLGen (Table S5). However, prediction information varied widely between databases (Table S13). For example, while the 30 GTEx cTFs were present in at least one of the seven prediction databases, cTFs in individual databases ranged from just four (TRANSFAC predicted targets) to 20 (Human TF-TG). A similar discrepancy was also observable for prediction information of cTFs in eQTLGen (Table S5). Consequently, also the number of predicted target genes for these sets of TFs showed substantial differences between databases and ranged between 7,330 and 17,323 for eQTLGen TFs and 1,648 and 16,696 for GTEx TFs.

Interestingly, some of the predicted target genes were located in genomic regions that have previously been reported to be devoid of Neandertal and Denisovan ancestry [1,4]. Among the 257 protein-coding genes located in the five autosomal deserts reported by both studies 239 were within the predicted target gene sets. All those genes were predicted to be regulated by at least one of the 185 eQTLGen cTFs in at least one database (Table S5). Also, the smaller set of 30 GTEx cTFs was predicted to affect almost all (237/239) of those desert genes (Figure 5C). Two genes that have been associated with modern human-specific biology, *FOXP2* and *ROBO2*, were among the cTF targets [51,52]. Our results suggest that they might nevertheless be influenced by introgressed archaic DNA. However, it remains unclear how many of the predicted target genes are in fact regulated by these TFs and in which tissue or developmental stage these effects are relevant. We found one empirical case in eQTLGen where the trans-

eQTL gene *TCEA1* (associated with aSNP *rs4805834*) is located within a desert on chromosome 8 54.5-65.4MB. Another example was *UTP14A*, which was linked to trans-eQTL aSNP *rs12603526*. *UTP14A* is located on chromosome X within one of the multiple deserts on that chromosome (Figure 5D).

Both predicted targets of cTFs and trans-eQTL information in eQTLGen suggest that genes in deserts are not out of regulatory reach of Neandertal DNA. However, we were seeking to further assess the significance of the role of Neandertal DNA on desert genes. For this purpose, we approached this question from a different angle. We first scanned for TFs with predicted target genes that were over-represented in archaic deserts. We again leveraged information from the seven TF target prediction database to screen for such instances. If prediction information for a TF was available for more than one database, we required all odds ratios to be larger than 1 and at least two enrichment P values (Fisher's exact test) to be smaller than 0.05, at least one of which remaining at that significance after multiple testing correction. TFs with prediction information in only one database were required to have a more restricted FDR threshold of 10^{-4}. We found 18 TFs passing that criterion, 17 in GTEx and 14 in eQTLGen. Interestingly, this list included with *ASXL1*, *JUN*, *PRDM5* and *SMARCB1* - four Neandertal-linked cTFs. All four candidates were cTFs in eQTLGen, with *JUN* and *PRDM5* also being cTFs in GTEx (*JUN*: Brain cortex; *PRDM5*: Frontal cortex BA9 and spleen). While predicted target genes for *JUN* and *SMARCB1* were evenly distributed across all five deserts, targets for *ASXL1* and *PRDM5* showed substantial differences in their prevalence between individual deserts (Figure 5A-B). A total of 31 of the 34 desert target genes for *ASXL1* were found in two desert regions (24 of 97 desert genes on chr1 and 7 of the 45 genes in desert on chr13) and all 28 *PRDM5* desert target genes were found on chromosome 7 (18 of 68) and 8 (10 of 35).

The impact of Neandertal-linked trans-eQTLs on modern human phenotype variation

Furthermore, we sought to investigate links between eQTLGen and predicted trans-eQTL aSNPs and their potential phenotypic effects. The initial selection of SNPs tested for trans-eQTL effects by Võsa & Claringbould et al. included variants that showed significant associations in GWAS ($P<5x10^{-8}$) in the EBI GWAS catalog [39], Immunobase and a blood trait GWAS [40] (Materials and methods). By this selection a link between these trans-eQTLs and phenotypic effects has already been established. We have explored these databases and additional association data from Biobank Japan to annotate significant phenotype associations for our Neandertal-linked trans-eQTL aSNPs (Table S6). We found that most of our candidates (7/13) were showing associations with various blood cell composition measures, including the

two trans-eQTLs with 27 and 33 target genes, respectively. These results are consistent with the fact that most of the tested and trans-eQTL-associated variants by Võsa & Claringbould et al. were linked to this group of phenotypes. However, one of these two multi-gene trans-eQTLs (*rs13098911/rs13063635*) showed associations not related to blood cells. Two of those associations showed a decreased risk for mouth ulcers and other dental issues. The archaic alleles for this risk locus have previously also related to increased risks for a severe Covid-19 phenotype [53] and Celiac disease, an autoimmune disorder, where dietary gluten intake causes inflammation in the small intestine at gluten intake [54]. Notably, another trans-eQTL locus (*rs2066807/rs2066819*) was also associated with an increased risk in autoimmune disease. The archaic alleles at that locus were more prevalent in individuals with psoriasis. And also this locus was not limited to one association, but showed multiple additional phenotype links, including multiple anthropometric measures related to increased height and body mass and several pulmonary functional measurements. Two other trans-eQTL loci (aSNPs *rs17331332* and *rs7811653*) were associated with anthropometric and pulmonary measures as well. The directional effects for the phenotypes included increasing and decreasing height, weight and pulmonary measures. The remaining phenotype associations for Neandertal-linked trans-eQTL variants were related to increased risk for colorectal cancer (*rs12603526*), electrocardiogram measurements (*rs10919070/rs10919071*), hypospadias (*rs7811653*) and brain connectivity measurements (*rs16997087*).

In order to evaluate this observation in a dataset that was not pre-selected for GWAS-associated variation we explored whether aSNPs linked to any of the 65 GTEx cTFs showed phenotype associations ($P<5x10^{-8}$) in any of the previously used cohorts as well. We found that 14 cTF loci showed aSNPs with significant associations, including the region on chromosome 12 (cis-eQTL TF: *STAT2*) which was also detected as a trans-eQTL in eQTLGen and described in the previous section (Tables S5, S6). The aSNPs for seven additional cTFs (*ATOH7*, *CCDC88A*, *FOXC1*, *HOXA13*, *SMARCA4*, *ZNF592* and *ZKSCAN4*) showed associations with height measures. This phenotype was also among the ones associated with the *STAT2* aSNPs on chromosome 12 as well (Figure 1F), making this by far the most frequently associated phenotype among TF-associated aSNPs. Other non-height phenotype associations included cholesterol levels, coronary artery disease, neuroma, breast cancer, grip strength, body fat measures, varicose veins, and abnormal red blood cell volumes.

Evidence for local adaptation

One of the candidate haplotypes on chromosome 12 overlaps a genomic region that previously has been linked to a Neandertal DNA with signals of positive selection in Papuans [55]. We therefore re-evaluated the frequency distribution of this and all other eQTLGen trans-eQTL candidate regions in present-day non-African populations for evidence of positive selection. We quantified the frequency of our candidate aSNPs and compared it to the frequency distribution of all detected aSNPs within 15 Eurasian populations from the 1,000 Genomes cohort and 16 SGDP Papuans [25,27]. As previously reported by Mendez et al. we found that while the region on chromosome 12 showed frequencies of <10% in all 1,000 Genomes Eurasians, the frequency was, with 57%, substantially higher in Papuans (Figure 1C, 3). This frequency in Papuans puts these aSNPs within the top 5% of all aSNPs in that population. This observation is even more remarkable for a Neandertal haplotype, given that in Papuans many aSNPs are derived from Denisovans and expected to have an on average higher frequency compared to their Neandertal counterparts. In total we found that eight other candidate regions reached frequencies in at least one population that were among the top 5% among aSNPs in the given population. Particular outliers here were the regions associated with the trans-eQTL aSNPs *rs13063635/rs13098911* and *rs6531656*, which were found at archaic allele frequencies in the top 1% among aSNPs (34-38% in South Asians for *rs13063635/rs13098911* and 43-51% in three East Asian populations for *rs6531656*).

Additionally, we sought to explore whether any of the 13 trans-eQTL candidate aSNPs show evidence for recent directional allele frequency changes, possibly as a result of selective pressures. For that we used a dataset of around 2,400 whole genome sequences from the Estonian Biobank as by leveraging such a big sample from a rather homogeneous population we can gain power in detecting frequency changes in the recent past. We first generated local genealogical trees for this dataset using Relate [43] and then applied CLUES [41], an approximate full-likelihood method for inferring allele frequency trajectories to our trans-eQTL candidate aSNPs (Materials and methods). This analysis revealed trans-eQTL aSNPs that experienced either an increase or a decrease in the frequency of the Neandertal allele over the last 500 generations. Most of those changes are characterized by low log likelihood ratios (logLR) suggesting little evidence for rejecting neutrality (Figure 6, Table S14). Nevertheless, two trans-eQTL candidates, that also showed exceptional frequencies in some populations before (*rs13063635*/*rs13098911* and *rs2066807*/*rs2066819*), were predicted to have increased in frequency by more than two-fold over the last 500 generations in the population ancestral to

Estonians with a strong support for non-neutrality. The corresponding average logLRs for each of the two trans-eQTL candidates aSNP pairs were high compared to ~2,900 significant trans-eQTL SNPs from eQTLGen, reaching the 93rd (*rs13063635*/*rs13098911* and 97th percentile for only *rs13063635*) and 97th (*rs2066807*/*rs2066819*) percentile of the trans-eQTL SNP logLR distribution (Materials and methods).

Both analysis on the frequencies of trans-eQTL-associated aSNPs in populations and recent frequency changes in Estonian genomes suggest that particularly the candidate haplotypes on chromosomes 3 and 12 show signals of positive selection.

Discussion

In this study we explored the eQTLGen blood gene expression dataset for evidence of associations between introgressed Neandertal DNA and gene expression variation. The sample size of more than 30,000 individuals for this dataset provides the statistical power to annotate trans-eQTLs, long-range regulatory SNP associations, which are typically smaller in effect size compared to cis-eQTLs, and harder to confidently identify in smaller cohorts such as GTEx. Among the ~10,000 preselected and phenotype-associated SNPs tested for trans-eQTL effects in eQTLGen were 16 that had a link to 13 unique Neandertal haplotypes. Several studies have previously shown evidence for a significant contribution of Neandertal DNA to gene expression regulation including a directional allele-specific expression in the brain and testes [12], an impact on regulatory genomic motifs [13,16] and larger number of significant cis-eQTL associations compared to frequency-matched non-archaic variants [8,11]. However, the sparseness of the set of tested trans-eQTL SNPs prevents the application of previously used methods to robustly assess whether the proportion of Neandertal associations among all significant trans-eQTLs is unexpectedly higher or lower. Another shortcoming of the sparse coverage of the trans-eQTL screen is the lack of information on the association strength of genetic variation that is in linkage disequilibrium with the tested SNPs. It is conceivable that some of our aSNP associations are in high linkage disequilibrium with non-archaic variants that show even stronger effect sizes. While the strongest effect size doesn't necessarily translate to actual molecular consequences [47] it likely represents a higher probability for causality.

Nevertheless, we show that archaic haplotypes linked to 11 of our 13 trans-eQTL aSNPs are carrying aSNPs that alter regulatory sequences. One of those haplotypes is located on chromosome 12 and linked to the trans-eQTL aSNPs *rs2066807/rs2066819*. Other aSNPs linked to this haplotype and the trans-eQTL aSNPs on chromosome 3 (*rs13063635*/*rs13098911*) showed in addition to their computationally identified links to modify regulatory sequences also expression changing effects, as determined in reporter assays in an immune cell line. Cis-eQTL data in eQTLGen and GTEx as well as previous CRISPR experiments [31] associate the variants on chromsome 12 also with expression-modulating effects of *STAT2*. This observation is particularly relevant given that one of the two trans-eQTL genes associated with the trans-eQTL aSNPs on chromosome 12, *IFI16*, is linked to the same haplotype and a predicted target of *STAT2*. We found multiple additional instances where our trans-eQTL aSNPs were cis-eQTLs of TFs as well.

Võsa & Claringbould et al have demonstrated that a substantial amount of trans-eQTLs can be traced back to direct or indirect effect on TFs that can, in turn, affect the expression of

the genes in distal loci [23]. In this study we explored what implications such TF-related processes have in the context of introgressed Neandertal DNA. We focussed on one specific mechanism where Neandertal variant has a cis-regulatory effect on local TF and determined which putative target genes of this TF could be under distal effect from the same cis-regulatory variant We identified 65 such cTFs across various tissues in GTEx with aSNP cis-eQTL associations. Tissues with most such cis-eQTLs were skeletal muscle (8 cTFs), lung, adipose (subcutaneous), skin, thyroid (all 7 cTFs) and blood (6 cTFs). Based on the fossil record previous studies presented evidence for differences in skeletal features and fat distribution between Neandertals and modern humans [56]. Furthermore, it has repeatedly been shown that introgressed Neandertal DNA significantly influences skin and hair traits [5–7]. Some of these studies have also demonstrated that Neandertal DNA showed an over-proportional number of associations with behavioral traits such as sleep [5–7,57], which are partially controlled by hormones produced in the thyroid. Our results may indicate that some of these phenotypic effects and differences may be linked to regulatory networks involving TFs in related tissues, like adipose tissue, skeletal muscle and skin. We then explored seven TF target gene prediction databases and found target gene information for 30 of the 65 GTEx cTFs and 185 of 441 eQTLGen cTFs (Table S5). The number of predicted target genes across these databases with 17,610 and 17,807 in GTEx and eQTLGen respectively comprised a large fraction of the transcriptome. Among these predicted target genes were more than 90% locate in genomic regions that have been shown to be devoid of Neandertal and Denisovan introgressed DNA. In addition, we show that among TFs that show an enrichment of their predicted target gene among genes in these archaic deserts were four TFs that show significant cis-eQTLs with aSNPs. Two of these TFs that were found in GTEx show their significant cis-eQTLs in brain tissues. This is particularly notable given that genes in these deserts have previously been shown to be enriched for brain-expressed genes [1]. This expression information for desert genes has previously been attributed to biological systems that might have differed between modern and archaic humans. Our results imply that the assumption of genes in these regions not being influenced by archaic DNA may be too simplified and that while only ~40% of the Neandertal genome is still recoverable in people today, its reach across the genome is likely beyond that percentage.

We note that the different prediction algorithms we used in our study are based on different prediction strategies, including computational frameworks and inferences based on experimental data. Consequently, the accuracy of these algorithms varies [58] and is hard to exactly quantify as it often remains unclear in which tissue or developmental stage the predicted

effects are functionally relevant. Nevertheless, our results indicate the potential regulatory reach of introgressed Neandertal DNA just by this one mechanism alone. And given that this mechanism is likely to explain substantially less trans-eQTLs when compared to indirect TF-mediated mechanisms [23] it is likely that the long-distance reach of Neandertal DNA goes way beyond what we have shown in our study. It remains challenging to computationally reconstruct regulatory mechanisms and therefore fully reconstruct the regulatory role of Neandertal DNA through its interaction with TFs. However, the results of our study and other previous work that investigated the impact of Neandertal DNA on the modification of TF binding motif sequences [59,60] have demonstrated the importance of studying the interaction of Neandertal DNA and TF activity in the quest to complete the picture of the impact of Neandertal introgression on the transcriptomic landscape of people today.

Finally, we explored trans-eQTL loci in eQTLGen and GTEx cTF loci that are linked to introgressed Neandertal DNA for phenotype associations with their corresponding aSNPs in various GWAS cohorts. Eight of the 13 trans-eQTL loci showed associations with blood cell type measures. The only other group of phenotypes with more than one locus linked to it were autoimmune diseases and height, each having two loci linked to it (Table S6). A total of 13 of the 65 GTEx cTF loci showed significant phenotype associations, too (Table S7). Notably, seven of these loci had associations with height and other body measures. Another four loci showed links to blood cell type composition. No other phenotype had associations with more than one of the 13 cTF loci. Given that trans-eQTL SNPs that were tested in eQTLGen were selected based on preexisting phenotype associations it remains unclear how representative these results are. However, given that blood cell type composition and height were also the most prevalent groups with associations to cTF loci suggests that particularly these groups of phenotypes may be linked to TF activation in the context of Neandertal introgression. We note that given the biases that allele frequencies introduce to association studies with gene expression and phenotypes and that the number of significant genetic loci highly various between phenotypes will likely influence these observations. Consequently, it is hard to quantify whether these observations represent a significant deviation from the expectation given these factors. Nevertheless, it is intriguing to observe that highly polygenic phenotypes such as height were often linked to regulatory activity of TFs who themselves have the ability to influence expression levels of large sets of genes and therefore represent a polygenic equivalent on the transcriptomic level. Another observation that is noteworthy in that context is that aSNPs for eight of the 13 trans-eQTL loci in eQTLGen are found at frequencies that have reached the top 5% among all Neandertal variants in at least one 1,000 Genomes population. And while this

observation is again not surprising given the elevated detection power in eQTL studies that comes with higher allele frequencies it is still informative to observe the regulatory network that is linked to these putative candidates for adaptive introgression. Here, particularly the trans-eQTLs associated with *rs13063635*/*rs13098911* and *rs2066807*/*rs2066819* stood out. Both sets of aSNPs showed both exceptionally high frequencies in some present-day populations and evidence for a more than two-fold recent increase in frequency in a European population. Both candidates also had associations with an increased risk of autoimmune disorders, consistent with an increase of the prevalence of such disorders in recent human evolution [61]. Our results suggest that Neandertal DNA may have contributed to that process.

In our study we provide new insights into the regulatory activity of Neandertal DNA, its potential phenotypic consequences and its role in recent human evolution. Future studies with genome-wide trans-eQTL information for various tissue types will help to refine and complete the impact of Neandertal DNA on the transcriptomic landscape of present-day people and help to decipher regulatory mechanisms that are modulated by introgressed DNA.

Figures

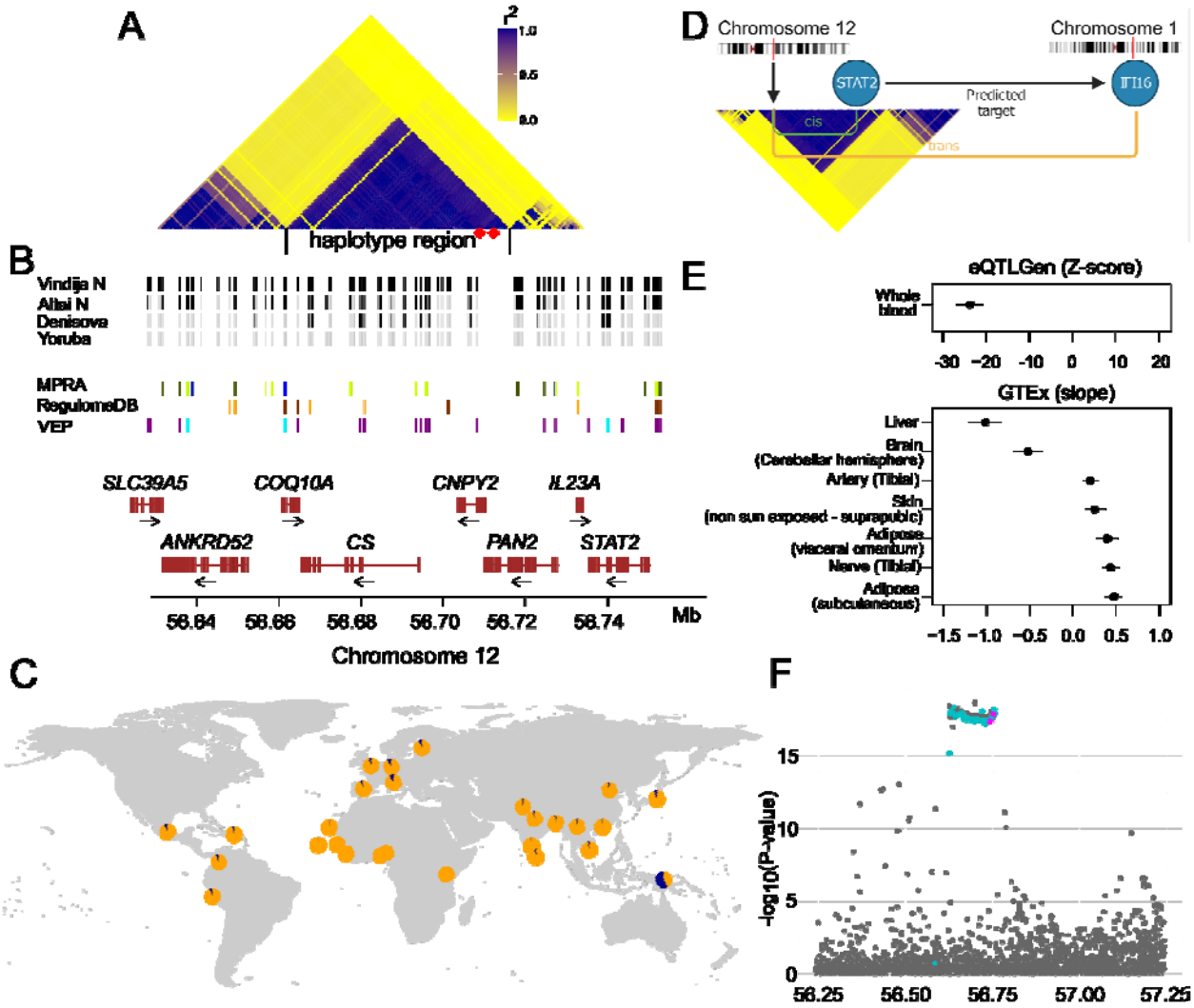


Figure 1: Neandertal-linked regulatory effect modulated by *STAT2*

(A) Linkage disequilibrium plot showing r^2 for aSNP surrounding *rs2066807/rs2066819* (location highlighted in red) on chromosome 12 in the 1,000 Genomes dataset. Inferred haplotype region encompassing aSNPs with r^2>0.8 is shown below LD plot (chr12:56,627,300-56,753,822). **(B)** Each bar represents an aSNP within the inferred Neandertal haplotype on chromosome 12 (A). Upper part of the panel shows genotypes across these aSNPs in the Vindija and Altai Neandertals, the Denisovan and the Yoruba 1,000 Genomes population (hg19 references=black, alternative=light gray and heterozygous=gray). Middle part of the panel displays molecular inferences for aSNPs based on massively parallel reporter assay information (tested = dark green, cis regulatory activity = light green, expression modulating = blue), RegulomeDB (likely to affect binding and linked to expression of target gene = light orange, likely to affect binding = dark orange) and ENSEMBL's variant effect predictor (missense variant

= light blue, potential regulatory effect = purple). Lower part of the panel displays gene models for protein coding genes in the haplotype region (coordinates are displayed at the bottom of the panel). **(C)** Frequency distribution of *rs2066807* (dark blue) in 1,000 genomes populations and SGDP Papuans. **(D)** Illustration of putative regulatory mechanism for candidate region with a cis-eQTL link of candidate aSNPs to *STAT2* and a trans-eQTL link for the same aSNPs to *IFI16* - a predicted target of *STAT2*. **(E)** Effect sizes with 95% confidence intervals for significant cis-eQTLs (FDR<0.05) between trans-eQTL candidate aSNPs and STAT2 in GTEx tissues (y-axis, effect size shown in the form of slope estimates) and eQTLGen blood (effect size shown as Z-score). **(F)** Manhattan plot displaying association P values (-$\log_{10}$-transformed, y-axis) with height (UK Biobank) of SNPs in and around the haplotype region. Candidate aSNPs are shown in pink and other aSNPs in light blue.

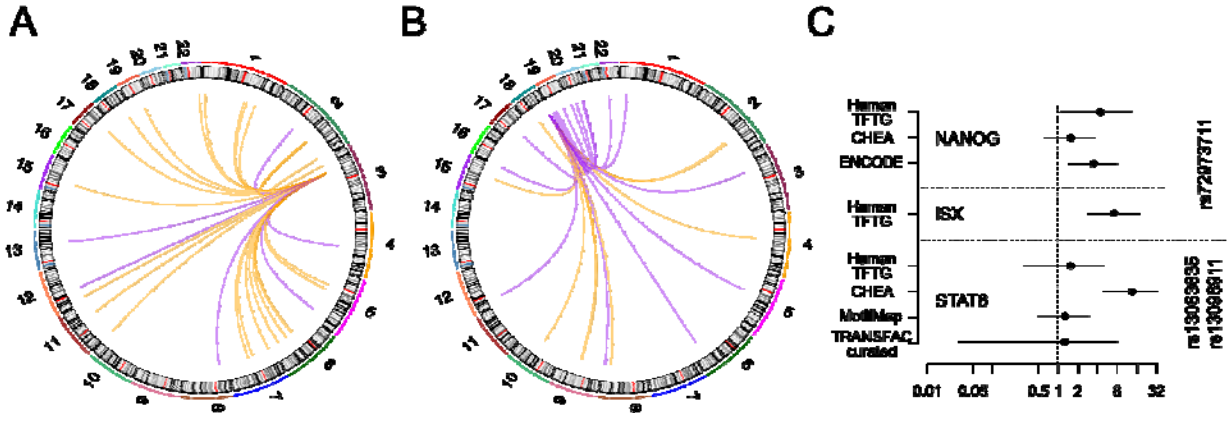
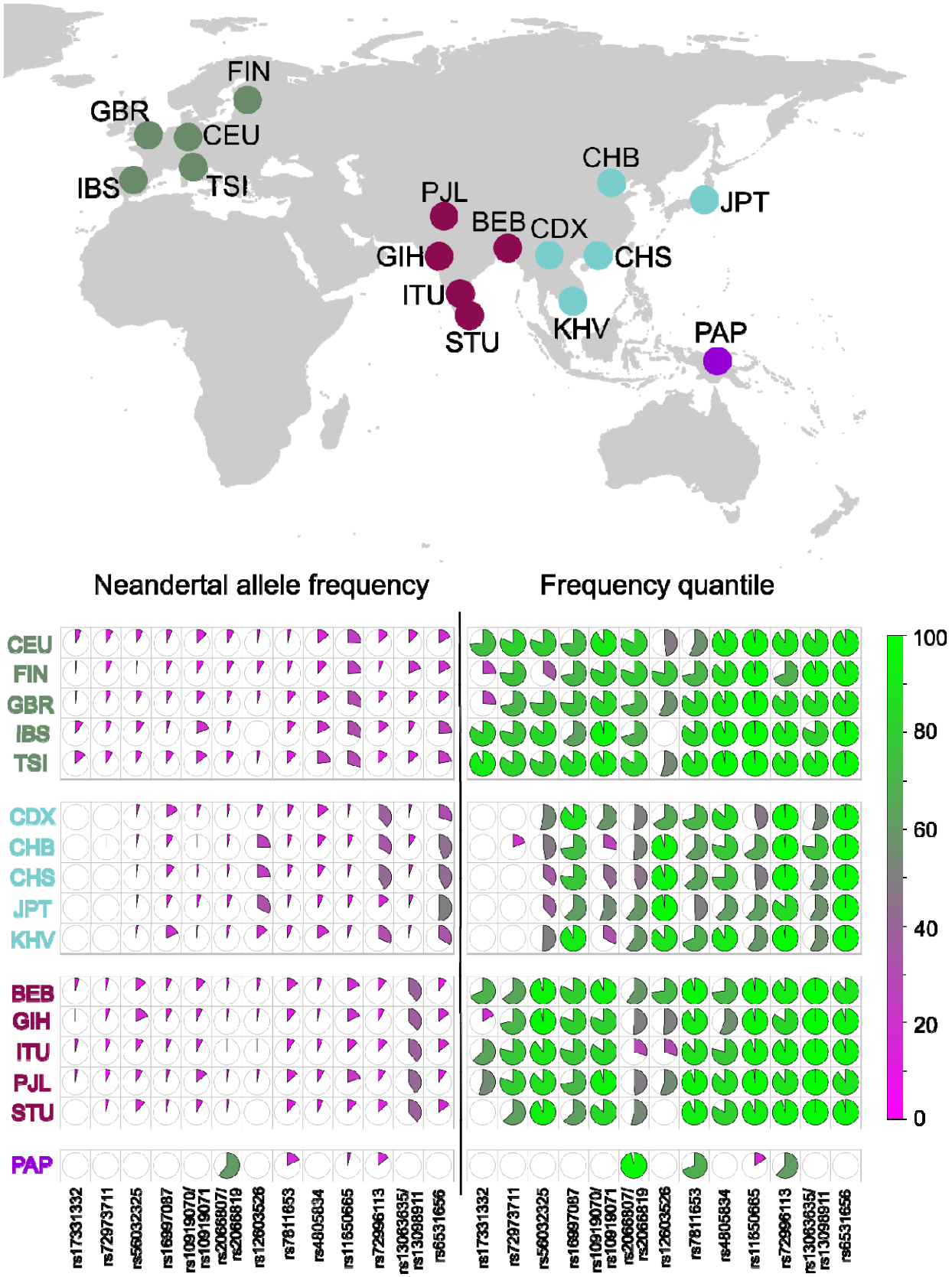


Figure 2: Transcription factors with target genes that significantly overlap with two aSNP trans-eQTL gene sets

Circo plots for two hub trans-eQTL aSNP loci [**(A)** *rs13063635/rs13098911*; **(B)** *rs72973711*) displaying the chromosomal links between trans-eQTL aSNPs and trans-eQTL genes. Genes with both a higher and lower expression in the presence of the archaic allele are shown in orange and purple, respectively. **(C)** Odds ratios with 95% confidence intervals representing the proportional overlap of trans-eQTL genes for **(A, B)** and sets of transcription factor target genes are displayed.


FIN
GBR
CEU
IBS
TSI
CHB
JPT
PJL
BEB
CDX
GIH
CHS
ITU
KHV
STU
PAP
Neandertal allele frequency
Frequency quantile
CEU
FIN
GBR
IBS
TSI
CDX
CHB
CHS
JPT
KHV
BEB
GIH
ITU
PJL
STU
PAP
100
80
60
40
20
0
rs17331332
rs72973711
rs56032325
rs16997087
rs10919070/ rs10919071
rs2066807/ rs2066819
rs12603526
rs7811653
rs4805834
rs11650665
rs72996113
rs13063635/ rs13098911
rs6531656

691

Figure 3: The frequency of trans-eQTL aSNPs in Eurasian populations

Pie charts illustrating the frequencies (lower panel, left) and quantiles among all aSNPs in a given population (lower panel, right) for 13 trans-eQTL aSNPs (x axis) in 15 Eurasian 1,000 Genomes populations and Papuans from the Simons Genome Diversity Project (geographic distribution of populations shown in upper panel). Pie charts are color-coded based on frequency and quantile values.

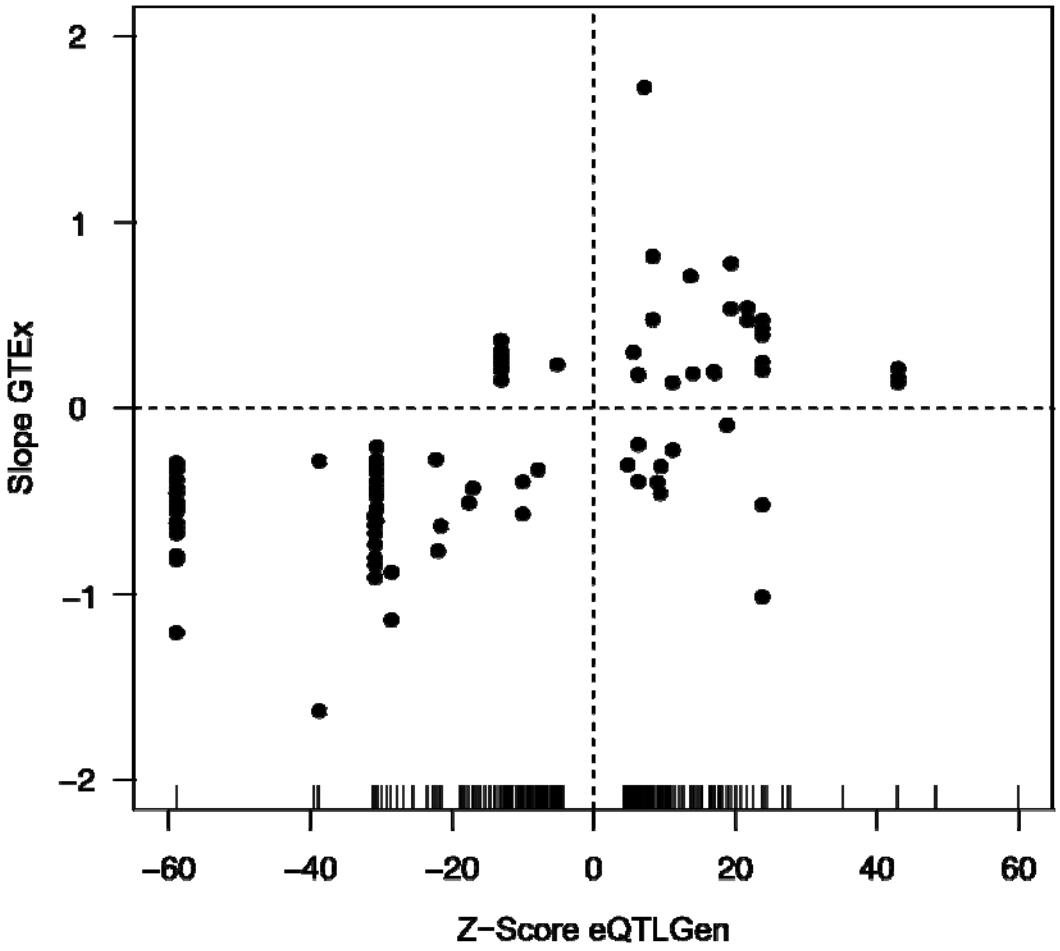


Figure 4: Correlation between directional effects between GTEx and eQTLGen cTFs

Scatterplot showing correlation (Spearman's ρ=0.66; P = 1.3×10^{-13}) between directional summary statistics for transcription factors with a significant (FDR<0.05) aSNP cis-eQTL in both eQTLGen (Z-scores, x axis) and GTEx (slopes, y axis). Bars at the x-axis indicate eQTLGen Z-scores for significant cis-eQTLs (FDR) that are linked to aSNPs and TFs.

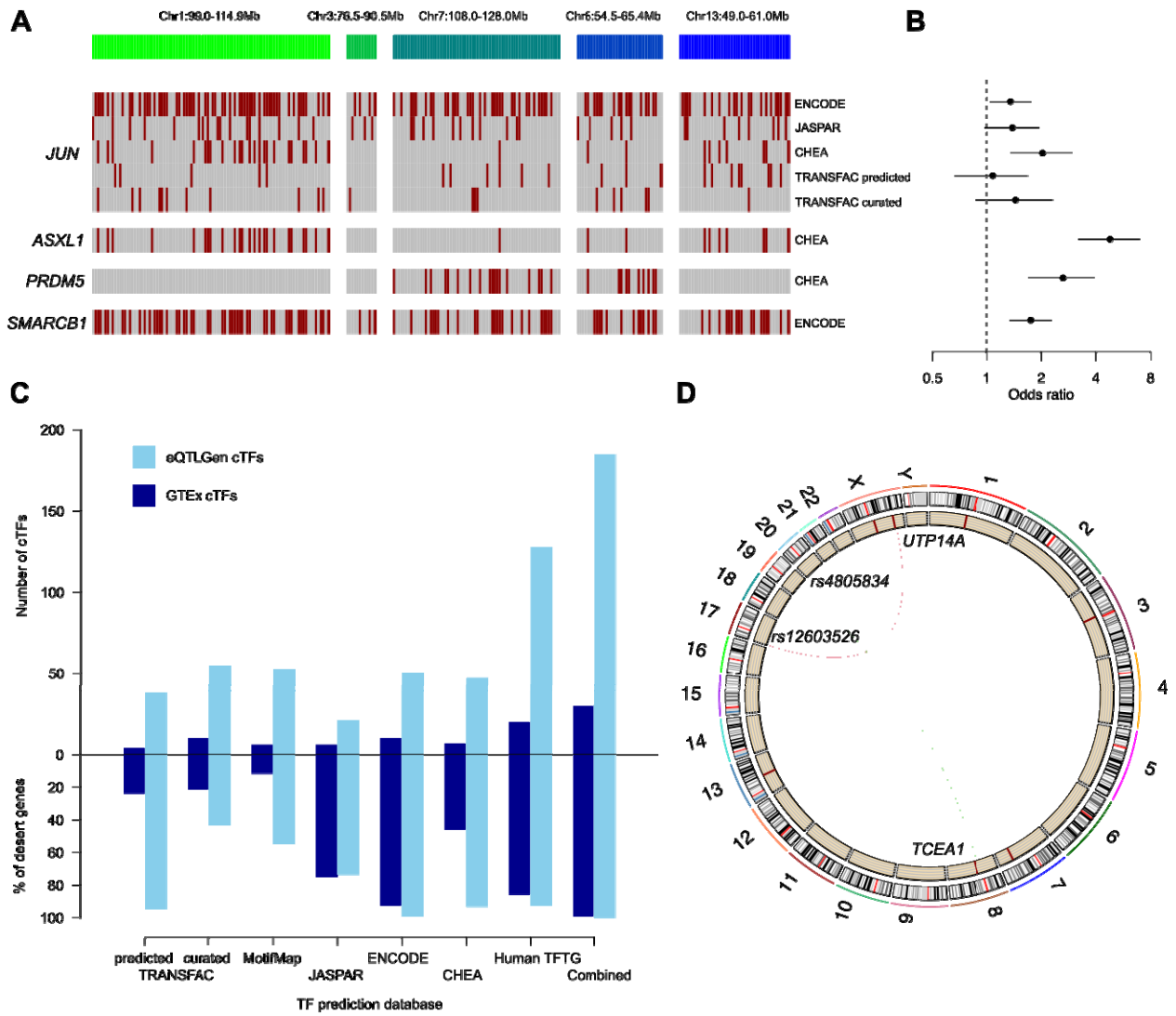


Figure 5: The regulatory Impact of Neandertal DNA on archaic deserts

(A) Predicted TF target genes for four cTFs (*JUN*, *ASXL1*, *PRDM5*, *SMARCB1*) with an over-proportional overlap with genes within autosomal archaic deserts. Each tile represents a desert gene (aligned in chromosomal order) and predicted targets for a given database are colored in red. **(B)** Odds ratios with 95% confidence intervals representing the proportional overlap between genes in archaic deserts and target genes for a given cTF and prediction database, as displayed in **(A)** are shown. **(C)** For each TF target gene database, the number of cTFs (light blue: eQTLGen; dark blue: GTEx) with prediction information (upper part) and the proportion of predicted archaic desert genes (lower part) is displayed. **(D)** Circo plot showing the connection between genomic coordinates for two Neandertal trans-eQTL aSNPs and their predicted desert target genes. Archaic deserts are highlighted in the inner band in red.

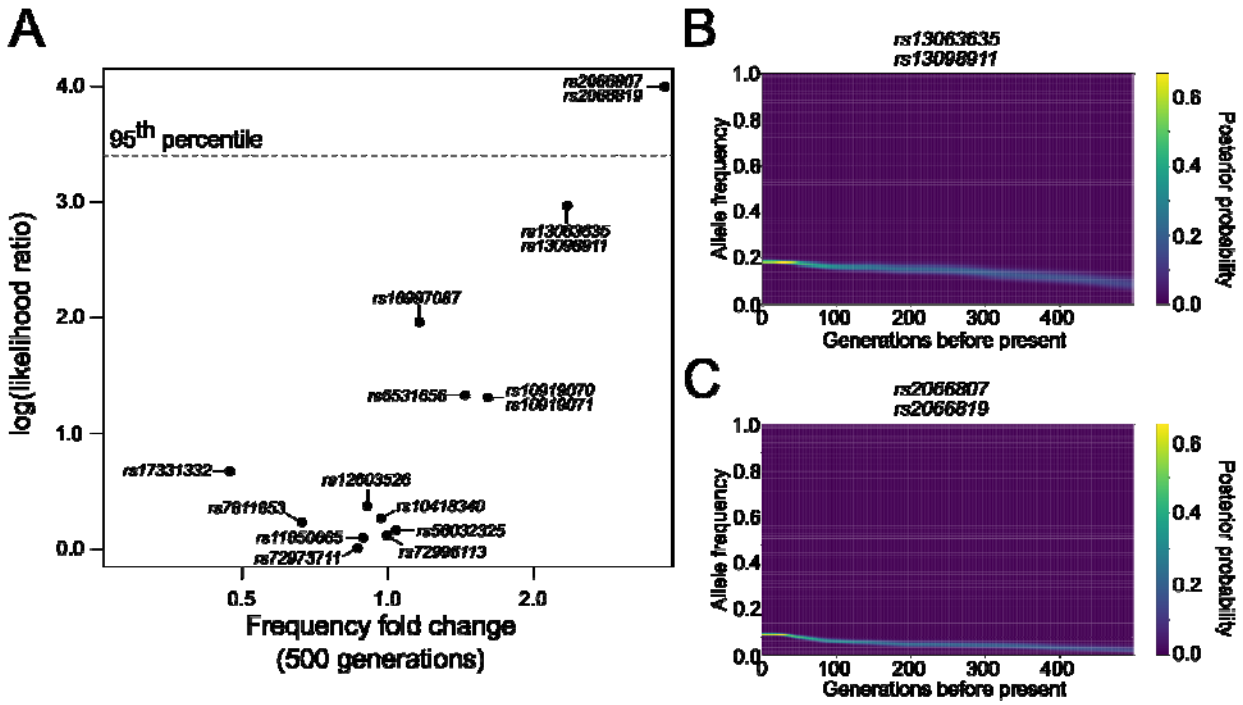


Figure 6: Frequency trajectory estimates for trans-eQTL associated aSNPs

(A) For each significant trans-eQTL aSNP locus the average frequency change in the Estonian population over the last 500 generations (x-axis) and the corresponding average log likelihood ratio (y axis) based on two runs of CLUES are displayed. The 95th percentile of the logLR distribution of eQTLGen trans-eQTLs is shown as a dotted line. **(B-C)** The estimated posterior probability for the Neandertal allele frequency for two trans-eQTL aSNP candidates over the last 500 generations in the Estonian population are shown.

Acknowledgments

We would like to thank Mayukh Mondal for his comments on the manuscript. Figure 1D was generated using Biorender.com. Some of the analyses were carried out with the facilities of the High-Performance Computing Center of the University of Tartu.

Funding

D.Y., V.P. and M.D. were supported by the European Union through Horizon 2020 Research and Innovation Program under Grant No. 810645 and the European Union through the European Regional Development Fund Project No. MOBEC008. U. V. was supported by the European Regional Development Fund, the Mobilitas Pluss program (MOBTP108) and through the Estonian Research Council grant PUT (PRG1291). B. Y. was supported by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.15-0012) and the European Regional Development Fund (Project No. 2014-2020.4.01.16-0030).

Conflict of interest disclosure

The authors declare they have no conflict of interest relating to the content of this article. Michael Dannemann is a recommender for PCI Genomics.

Consortia

Estonian Biobank Research Team:
Andres Metspalu, Mari Nelis, Lili Milani, Reedik Mägi & Tõnu Esko

Data, script and code availability

Script and code availability

Code related to the analyses presented in this manuscript is available at:

- https://github.com/SillySabertooth/Neandertal_trans-eQTLs
- https://myersgroup.github.io/relate/
- https://github.com/35ajstern/clues

Data availability:

Summary statistics for eQTLs:

- eQTLGen: https://www.eqtlgen.org/
- GTEx: https://gtexportal.org/home/datasets

Transcription factor target prediction information:

- https://maayanlab.cloud/Harmonizome/dataset/
- http://tfbsdb.systemsbiology.net/download

GWAS summary statistics:

- Immunobase: https://genetics.opentargets.org/immunobase
- GWAS catalog: https://www.ebi.ac.uk/gwas/

Molecular effects predictions

- VEP: https://www.ensembl.org/info/docs/tools/vep/index.html
- RegulomeDB: https://regulomedb.org/

Genotype data

- Neandertal and Denisovan genomes: http://ftp.eva.mpg.de/neandertal/ and http://ftp.eva.mpg.de/denisova
- 1,000 Genomes: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
- Simons Genome Diversity Project: https://www.simonsfoundation.org/simons-genome-diversity-project/, respectively
- Estonia Biobank WGS: https://genomics.ut.ee/en/access-biobank (Estonian Biobank ethics approval number nr 1.1-12/2859)

Supplementary information

Tables

Table S1: Predicted molecular effects (VEP) for trans-eQTL associated aSNPs with predicted consequences on gene expression regulation and modification of the protein sequence.

Table S2: Predicted regulatory potential (RegulomeDB) for trans-eQTL associated aSNPs with likely binding effects.

Table S3: MPRA regulatory information for trans-eQTL associated aSNPs with previous evidence for positive selection.

Table S4: MPRA regulatory information for trans-eQTL associated aSNPs overlapping a severe Covid-19 phenotype risk locus.

Table S5: Transcription factor target information.

Table S6: Significant phenotype associations for trans-eQTL associated aSNPs.

Table S7: GWAS summary statistics for GTEx cTF aSNPs.

Table S8: Summary statistics for significant trans-eQTLs associated with aSNPs in eQTLGen.

Table S9: Predicted trans-eQTL mechanism (Võsa & Claringbould et al.) for trans-eQTL aSNPs.

Table S10: Cis-eQTL summary statistics for transcription factors that co-localize with trans-eQTL-associated aSNPs.

Table S11: Summary statistics for aSNPs of cTFs in GTEx.

Table S12: Summary statistics for aSNPs of cTFs in eQTLGen.

Table S13: Overlap in available transcription factors with prediction information between databases.

Table S14: Frequency trajectory estimates for trans-eQTL aSNPs.

Figures

Figure S1: Sequence comparison between candidate haplotypes and genomes of Neandertals and Denisovan

Histograms illustrating the sequence similarity of haplotypes in the 1,000 Genomes cohort with the genome sequence of the Altai (1) and Vindija (2) Neandertals and the Denisovan (3) for each trans-eQTL candidate genomic region (Methods and materials). Haplotypes carrying the archaic allele for a given trans-eQTL aSNP are colored in red and other haplotypes in blue.

References

1. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science. 2016;352:235–9.

2. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358:655–8.

3. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.

4. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr Biol. 2016;26:1241–7.

5. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, et al. The phenotypic legacy of admixture between modern humans and Neandertals. Science. 2016;351:737–41.

6. Dannemann M, Kelso J. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. Am J Hum Genet. 2017;101:578–89.

7. McArthur E, Rinker DC, Capra JA. Quantifying the contribution of Neanderthal introgression to the heritability of complex traits. Nat Commun. 2021;12:4481.

8. Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell. 2016;167:643–56.e17.

9. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Curr Biol. 2016;26:3375–82.

10. Dannemann M, Racimo F. Something old, something borrowed: admixture and adaptation in human evolution. Curr Opin Genet Dev. 2018;53:1–8.

11. Dannemann M, Prüfer K, Kelso J. Functional implications of Neandertal introgression in modern humans. Genome Biology. 2017;18(1):61.

12. McCoy RC, Wakefield J, Akey JM. Impacts of Neanderthal-Introgressed Sequences on the

Landscape of Human Gene Expression. Cell. 2017;168:916–27.e12.

13. Silvert M, Quintana-Murci L, Rotival M. Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. Am J Hum Genet. 2019;104:1241–50.

14. Vespasiani DM, Jacobs GS, Brucato N, Cox MP, Romero IG. Denisovan introgression has shaped the immune system of present-day Papuans. bioRxiv. 2021.

15. Petr M, Pääbo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. Proc Natl Acad Sci U S A. 2019;116:1639–44.

16. Telis N, Aguilar R, Harris K. Selection against archaic hominin genetic variation in regulatory regions. Nat Ecol Evol. 2020;4:1558–66.

17. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet. 2020;52:626–33.

18. Ouwens KG, Jansen R, Nivard MG, van Dongen J, Frieser MJ, Hottenga J-J, et al. A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. Eur J Hum Genet. 2020;28:253–63.

19. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

20. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat Methods. 2016;13:366–70.

21. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169:1177–86.

22. Consortium TG, The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318-1330.

23. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10.

24. Dannemann M. The Population-Specific Impact of Neandertal Introgression on Human Disease. Genome Biol Evol. 2021;13(1):evaa250

25. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

26. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.

27. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.

28. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.

29. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

30. Jagoda E, Marnetto D, Montinaro F, Richard D, Pagani L, Capellini TD. Regulatory dissection of the severe COVID-19 risk locus introgressed by Neanderthals. bioRxiv. 2021

31. Jagoda E, Xue JR, Reilly SK, Dannemann M, Racimo F, Huerta-Sanchez E, et al. Detection of Neanderthal Adaptively Introgressed Genetic Variants that Modulate Reporter Gene Expression in Human Immune Cells. Mol Biol Evol. 2021;msab304

32. Xie X, Rigor P, Baldi P. MotifMap: a human genome-wide map of candidate regulatory motif sites. Bioinformatics. 2009;25(2):167-74.

33. Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. Cell Systems. Cell Press; 2016;3:172–86.

34. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014;42(Database issue):D142-7.

35. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34:D108-10.

36. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011;9(4):e1001046

37. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010;26(19)2438–2444.

38. Hu H, Miao Y-R, Jia L-H, Yu Q-Y, Zhang Q, Guo A-Y. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. Oxford Academic; 2018;47:D33–8.

39. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45:D896–901.

40. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167:1415–29.e19.

41. Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. 2019;15:e1008384.

42. Marnetto D, Pankratov V, Mondal M, Montinaro F, Pärna K, Vallini L, et al. Ancestral contributions to contemporary European complex traits. bioRxiv. 2021

43. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. Nat Genet. 2019;51:1321–9.

44. Kals M, Nikopensius T, Läll K, Pärn K, Sikka TT, Suvisaari J, et al. Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. bioRxiv. 2019

45. Pankratov V, Montinaro F, Kushniarevich A, Hudjashov G, Jay F, Saag L, et al. Differences in local population history at the finest level: the case of the Estonian population. Eur J Hum Genet. Nature Publishing Group; 2020;28:1580–91.

46. Qin P, Stoneking M. Denisovan Ancestry in East Eurasian and Native American Populations. Mol Biol Evol. 2015;32:2665–74.

47. Battle A, Montgomery SB. Determining causality and consequence of expression quantitative trait loci. Hum Genet. 2014;133:727–35.

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

49. Deng HK, Unutmaz D, KewalRamani VN, Littman DR. Expression cloning of new receptors used by simian and human immunodeficiency viruses. Nature. 1997;388:296–300.

50. Uguccioni M, Mackay CR, Ochensberger B, Loetscher P, Rhis S, LaRosa GJ, et al. High expression of the chemokine receptor CCR3 in human blood basophils. Role in activation by eotaxin, MCP-4, and other chemokines. J Clin Invest. 1997;100:1137–43.

51. Peyrégne S, Boyle MJ, Dannemann M, Prüfer K. Detecting ancient positive selection in humans using extended lineage sorting. Genome Res. 2017;27:1563–72.

52. Maricic T, Günther V, Georgiev O, Gehre S, Curlin M, Schreiweis C, et al. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. Mol Biol Evol. 2013;30:844–52.

53. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature. 2020;587:610–2.

54. Caio G, Volta U, Sapone A, Leffler DA, De Giorgio R, Catassi C, et al. Celiac disease: a comprehensive current review. BMC Med. 2019;17:142.

55. Mendez FL, Watkins JC, Hammer MF. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. Am J Hum Genet. 2012;91:265–74.

56. Venner SJ. A New Estimate for Neanderthal Energy Expenditure. 2018.

57. Dannemann M, Milaneschi Y, Yermakovich D, Stiglbauer V, Friese MA, Otte C, et al. Neandertal introgression dissects the genetic landscape of neuropsychiatric disorders and associated behavioral phenotypes. medRxiv. 2021.

58. Jayaram N, Usvyat D, R Martin AC. Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics. 2016;17:547.

59. Barker HR, Parkkila S, Tolvanen MEE. Evolution is in the details: Regulatory differences in modern human and Neanderthal. bioRxiv. 2020.

60. Findley AS, Zhang X, Boye C, Lin YL, Kalita CA, Barreiro L, et al. A signature of Neanderthal introgression on molecular mechanisms of environmental responses. bioRxiv. 2021.

61. Quintana-Murci L. Human Immunology through the Lens of Evolutionary Genetics. Cell. Cell Press; 2019;177:184–99.