Denisovan introgression has shaped the immune system of present-day Papuans

Davide M. Vespasiani[1,2], Guy S. Jacobs[3], Nicolas Brucato[5], Murray P. Cox[6‡], Irene Gallego Romero[1,2,7*]

1 Melbourne Integrative Genomics, University of Melbourne, Royal Parade, Parkville, 3010 Victoria, Australia

2 School of Biosciences, University of Melbourne, Royal Parade, Parkville, 3010 Victoria, Australia

3 Department of Archaeology, Downing Street CB2 3DZ, Cambridge, UK

5 Department of Evolutionary Medicine, University of Toulouse III Paul Sabatier, 31330 Toulouse, France

6 Statistics and Bioinformatics Group, School of Fundamental Sciences, Massey University, Palmerston North 4410, New Zealand

7 Center for Stem Cell Systems, University of Melbourne, Royal Parade, Parkville, 3010 Victoria, Australia

* To whom correspondence should be addressed: irene.gallego@unimelb.edu.au

Abstract

Modern humans have substantially admixed with multiple archaic hominins. New Guineans, in particular, owe up to 5% of their genome to Denisovans, a sister group to Neanderthals, whose remains have only been identified in Siberia and Tibet. Unfortunately, the biological and evolutionary significance of these events remain poorly understood. Here we investigate the function of archaic alleles of both Denisovan and Neanderthal ancestry characterised within a previously published set of 72 genomes from individuals of Papuan genetic ancestry living in the island of New Guinea. By comparing the distribution of archaic and modern human variants, we are able to assess the consequences of archaic admixture across a multitude of different cell types and functional elements. We find that archaic alleles are often located within cis-regulatory elements and transcribed regions of the genome, suggesting that they are actively involved in a wide range of cellular regulatory processes. We identify 39,954 high-confidence Denisovan variants that fall within annotated cis-regulatory elements and have the potential to alter the affinity of multiple transcription factors to their cognate DNA motifs, highlighting a likely mechanism by which introgressed DNA can impact phenotypes in present-day humans. Additionally, we detect a consistent signal across Denisovan variants of strong involvement in immune-related processes. Lastly, we show how such regulatory effects might underlie some of the observed gene expression differences between multiple Indonesian populations carrying varying amount of Denisovan DNA. Together, these data provide support for the hypothesis that, despite their broadly deleterious nature, archaic alleles actively contribute to modern human phenotypic diversity and might have facilitated early adaptation to non-African environments.

Introduction

Modern humans are known to have interbred with Neanderthals [1], Denisovans [2] and possibly other archaic hominins [3]. While genetically similar populations of Neanderthals are thought to have contributed approximately 2% to non-African genomes, Denisovan introgression has been observed to be more variable [4]. Particularly, Denisovan ancestry accounts for up to 5% of the genomes of the indigenous peoples of New Guinea and Australia [4]. In addition, these components exhibit a deep divergence from the reference Altai Denisovan genome, providing strong evidence for the occurrence of multiple Denisovan introgression events across time and space [5, 6].

At the genomic level, these introgressed archaic alleles are mostly observed outside protein-coding sequences, distributed over non-functional and regulatory regions [7, 8]. Enhancers, in particular, are amongst the top targeted elements of introgression. Here, archaic alleles are thought to drive phenotypic differences by altering gene pre- and post-transcriptional regulatory processes [7]. Furthermore, both Neanderthal and Denisovan variants seem to preferentially affect enhancers in a tissue-specific manner, with highly pleiotropic elements showing a depletion of archaic variation [9].

However, other than general agreement that introgressed archaic DNA has mainly been deleterious and negatively selected from coding sequences and conserved non-coding elements [10–12], the actual phenotypic consequences of this variation are not well understood. Several lines of evidence highlight associations between archaic DNA and risk for disease traits, including autoimmune diseases [12, 13], or with traits of possible evolutionary advantage for early non-Africans [14, 15]. For example, Neanderthal variants within immune genes and immune-related cis-regulatory elements (CREs) have been associated with differential responses to viral infections among present-day Europeans [16, 17].

Unfortunately, two main factors limit our interpretation of the biology of these alleles. First, we still lack a detailed characterisation of world-wide levels of human genetic diversity, vital to identifying differential archaic hominin contributions across modern populations [18, 19]. Second, and more challenging, most of these alleles lie within non-coding sequences where, despite their acknowledged contributions to human evolutionary history [20, 21], an understanding of their actual biological functions remains elusive.

To gain insights into the consequences of archaic introgression in Papuans, we have analysed a previously published dataset of 72 genomes sampled across the island of New Guinea, representing the largest collection of samples from this region to date [5]. By comparing the distribution of archaic single nucleotide polymorphisms (aSNPs) and non-archaic SNPs (naSNPs) segregating within these populations, across multiple genomic elements and cell types, we find that aSNPs are enriched within functional cis-regulatory and transcribed elements, particularly those active within immune-related cells. We also observe that the presence of archaic alleles within these elements results in disruption of the binding sites of transcription factors, with Denisovan variants being strongly involved during immunological processes.

Results

Building a high-confidence set of archaic variants and SNPs annotation

To curate list of high-confidence archaic SNPs to analyse for possible functional contributions to present-day Papuans, we took advantage of a recently described dataset of both archaic and non-archaic haplotypes segregating within 72 individuals of Papuan ancestry [5]. Due to incomplete lineage sorting and recombination, a large fraction of variants included in these archaic haplotypes is not expected to be of archaic origin [22]. To account for this while enriching for variants of putative evolutionary interest we applied a series of filtering steps to our starting list of 8,337,067 SNPs [5]. Briefly, we began by grouping variants into putative aSNPs and naSNPs based on the ancestry of the inferred haplotypes they segregated with. We then removed all SNPs (both aSNPs and naSNPs) found segregating also within the African continent [23] and further discarded all putative aSNPs shared between archaic and non-archaic haplotypes. Finally, to analyse the patterns of archaic introgression across SNPs frequency ranges we grouped variants into low frequency (i.e., SNP frequency < 0.05) and common-to-high-frequency (i.e., SNP frequency ≥ 0.05), on the basis of derived allele frequency (DAF) for naSNPs and on the major introgressed archaic allele frequency (MIAF) for aSNPs (see Methods for additional details). Applying these stringent criteria resulted in subset of 228,511 high-confidence aSNPs (2.74% of the original set), which we assigned to either Denisovan (143,453) or Neanderthal (85,058) ancestries, and 1,110,762 (13.32% of the original set) naSNPs, which were used for all downstream analyses (Supp Table 1).

Next, to assess the putative biological functions of these variants we began by examining their distribution across multiple genomic elements. Due to the intensity of purifying selection as well as the small fraction of the human genome with protein coding functions, introgressed aSNPs have been extensively reported as lying outside coding regions [11, 22, 24]. Using the Ensembl Variant Effect Predictor (VEP) tool [25], we confirmed that, on average, almost all (98.1%) of our aSNPs and naSNPs fall within non-coding genomic elements. In particular, we observed that introns, regulatory and intergenic regions together contained the vast majority of our variants, regardless of the ancestry or the allele frequency bin that SNPs were grouped into (Supp. Fig. S1). As expected, compared to naSNPs we found a significant depletion of aSNPs from exonic sequences (binomial test: n = 30,790 naSNPs, n = 3,479 Denisovan and n = 1,900 Neanderthal aSNPs. $p < 2.2 \times 10^{-16}$ for both Denisovan and Neanderthal aSNPs).

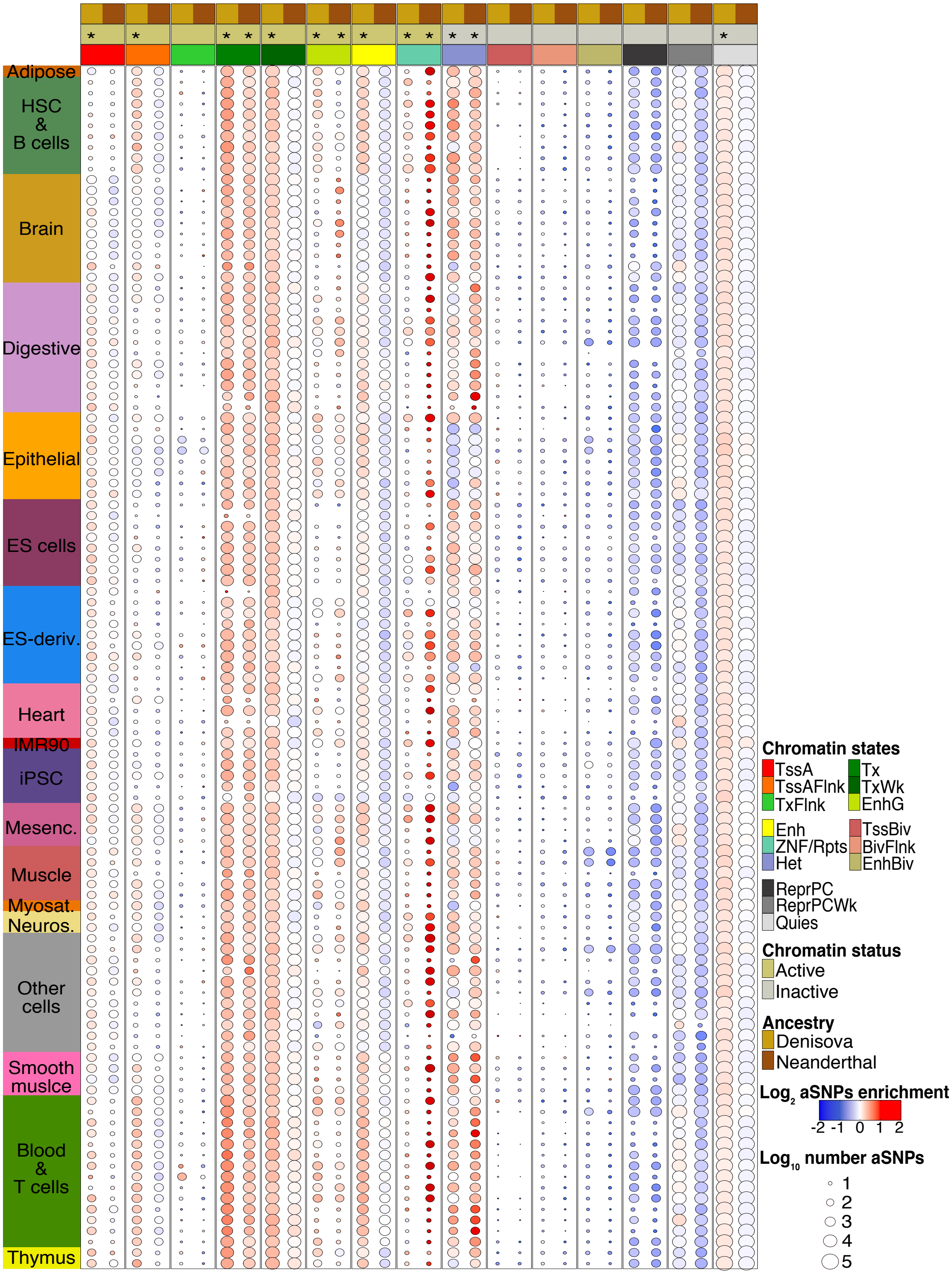

Fig 1. aSNP enrichment across the Roadmap Epigenomics data. For only common-to-high frequency variants we show the $\log_2$-transformed aSNP enrichment values for every chromatin state cell type combination computed over the mean of the aSNP:naSNP across all chromatin states and cell types (see main text). Top column annotations report (from top to bottom): ancestry information (i.e., Denisovan, Neanderthal); chromatin state functional activity (i.e., active = khaki, inactive = grey) as detailed in [26]; the 15 different chromatin states. Left column annotation shows the 18 different tissues. Colours within the heatmap cells represent the $\log_2$-transformed aSNP enrichment scores. Circle diameters represent the number of Denisovan or Neanderthal aSNPs annotated within any particular cell type/chromatin state combination. Asterisks indicate whether the $\log_2$-transformed aSNP enrichment was significantly higher than zero across all 111 cell types (FDR adjusted one-sample t-test p value: $* = < 0.01$, for detailed results, see Supp Table 2-Supp Table 3).

Distribution of archaic SNPs across chromatin states

Due to the large number of non-coding alleles within our dataset, we decided to compare the distribution of aSNPs and naSNPs over multiple chromatin functional states and human cell types. This allowed a deeper investigation into their possible biological functions *in vivo*, which would be otherwise hard to achieve from genomic positional information alone. Therefore, we retrieved the distribution of 15 different chromatin states identified across 111 cell types from the Roadmap Epigenomics Project [26], and merged this information with our list of SNPs.

Because Denisovan and Neanderthal aSNPs are respectively 12.9% and 7.6% of the total number of naSNPs, we first verified whether these differences were broadly maintained across all functional states and cell types. Thus, for each ancestry we first counted the number of aSNPs and naSNPs that fell within each chromatin state/cell type combination. For each of these, we then calculated a) the ratio of the number of aSNPs to that of naSNPs (henceforth aSNP:naSNP); b) the mean across all these values (0.12 and 0.073 for Denisovan and Neanderthal aSNPs, respectively) and used them as a reference to then compare the incidence of aSNPs within each chromatin state for all 111 cell types. Overall, we found the aSNP:naSNP to be consistent across all chromatin states. However, we noted that Denisovan and Neanderthal aSNPs both had a higher incidence than expected within quiescent and weakly transcribed chromatin states (Quies and TxWk, respectively) and within promoters and genic enhancer states (TssAFlnk and EnhG, respectively) for Denisovan aSNPs only (Supp. Fig. S2).

Given these differences, we asked whether the aSNP:naSNP varied significantly among cell types, chromatin states and/or relative to allele frequencies. Specifically, for each cell type/chromatin state combination we computed an aSNP enrichment score by dividing the aSNP:naSNP by the mean ratio across all 111 × 15 combinations. Then, for each chromatin state we asked whether the aSNP:naSNP ratios across all cell types for either ancestry were significantly higher than the sample-wide expected mean value. We found that aSNPs were generally depleted from inactive chromatin states (not associated with expressed genes as defined by Roadmap Epigenomics) [26]. For instance, Polycomb and weakly Polycomb repressed (ReprPC, ReprPCWk, respectively) or bivalent chromatin states (TssBiv, EnhBiv and BivFlnk) showed a consistent depletion of aSNPs across all cell types (Supp. Fig. S3 and Supp Table 2-Supp Table 3). On the other hand, we found that Quies and heterochromatin (Het) states contained a significant excess of both Denisovan and Neanderthal aSNPs. Removal of low frequency variants resulted in stronger depletion signals within all inactive states except for Quies and Het states. In the latter state we detected a significant excess of both Denisovan and Neanderthal aSNPs segregating at MIAF/DAF ≥ 0.05. We also found that only

Denisovan aSNPs were significantly enriched within Quies state whereas Neanderthal aSNP:naSNP closely matched its expected mean value (Fig. 1, FDR-corrected p-values for all tests are provided in Supp Table 2-Supp Table 3).

When considering active chromatin states (associated with expressed genes), Denisovan and Neanderthal aSNPs showed markedly different patterns. While both were significantly enriched amongst transcribed and zinc-finger protein-genes and repeats chromatin states (Tx and ZNF/Rpts, respectively), we noted a significant enrichment exclusively for Neanderthal aSNPs within genic enhancers (EnhG), whereas Denisovan aSNPs were significantly enriched for weakly transcribed regions, transcription start sites and enhancers (TxWk, TssA and Enh, respectively) (Supp. Fig. S3, all FDR-corrected p-values are available in Supp Table 2-Supp Table 3). These results were recapitulated when analysing common-to-high frequency variants. In particular, while no differences were observed for Neanderthal aSNPs, Denisovan variants, in addition to the states above described, were also significantly overrepresented amongst promoters and genic enhancers (TssAFlnk and EnhG, respectively).

To confirm whether these signals were caused by SNPs occurring within highly constitutive or cell-specific functional elements, we counted the number of different tissues within the Roadmap Epigenomics data across which any given aSNP or naSNP-containing element was consistently annotated to a particular chromatin state. This yielded an estimate of the potential pleiotropic effect of each functional element covered by our set of SNPs. As expected, we found that the vast majority of SNPs included within constitutive states such as Quies, Tx and TxWk [26], despite the frequency range considered, fell within highly pleiotropic elements (Supp. Figs. S4-S5). Nevertheless, aSNPs annotated within these states occurred within elements that have significantly higher pleiotropic activities than those carrying naSNPs (Supp Table 4).

Conversely, across the remaining chromatin states the vast majority of SNPs, again regardless of allele frequency, were highly cell-specific, showing consistent functional annotations in only one tissue (Supp. Figs. S4-S5). In line with Telis *et al.* [9] we found that Denisovan aSNPs within Enh, EnhG and EnhBiv all occurred within elements of significantly reduced pleiotropy compared to naSNPs (mean pleiotropy calculated across the three states for Denisovan aSNPs = 5.3, naSNPs = 5.6; Supp Table 4). However, when analysing common-to-high frequency variants we found this difference to be statistically significant solely for Denisovan aSNPs within EnhG and EnhBiv. Instead, only EnhBiv elements carrying Neanderthal aSNPs had significantly lower pleiotropic activities than naSNPs whereas variants annotated within both Enh and EnhG, regardless of allele frequency, occurred within higher pleiotropic elements although the magnitude of this difference is small (mean pleiotropy calculated across the two states for Neanderthal aSNPs = 6.1, naSNPs= 6.0; Supp Table 4).

Taken together, these findings suggest that aSNPs are likely to occur and be maintained within constitutive states that are functionally inert (Quies), perhaps as a consequence of purifying selection being less intense in those regions. At the same time, our results indicate that a non-trivial fraction of introgressed variants still segregating within Papuans today, might have functional consequence, particularly for gene regulatory processes, in a limited number of cell types.

Active aSNPs are highly enriched within immune cells

To identify the cell types showing a significantly excess of aSNPs across each of the 15 chromatin states, we normalised the aSNP:naSNP for each cell type/chromatin state combination by the mean ratio for the pertinent chromatin state (e.g. all 111 EnhG observations were normalised by the mean EnhG aSNP:naSNP ratio). We then grouped similar cell types into 18 different tissues, as in the original Roadmap Epigenomics data, and compared the distribution of aSNP enrichment scores. We found that, while the aSNP:naSNP within Quies was consistent among all tissues and between SNP frequency ranges for both Denisovan and Neanderthal variants, many of the remaining chromatin states (e.g. ZNF/Rpts, Het, TssA, TssAFlnk, Enh, EnhG, as well as TxFlnk, a state associated with both promoter and enhancer epigenetic marks) showed evidence for tissue-specific enrichment of aSNPs, seemingly driven by immune cells (Supp. Figs. S6-S9).

To better quantify these patterns, we asked whether the enrichment of aSNPs within immune T-cells was significantly higher than in the other tissues. Indeed, immune T cells contained a significant excess of both Denisovan and Neanderthal aSNPs within TssAFlnk, Tx, TxWk, Enh, ZNF/Rpts, Het and ReprPCWk; there is also a significant excess of Denisovan, but not Neanderthal, aSNPs within TxFlnk and EnhG (Supp. Figs. S6 - S8 and Supp Table 5). Analysis of only common-to-high-frequency variants showed, immune T cells contained significant enrichments of both Denisovan and Neanderthal aSNPs within TxWk, Enh, ZNF/Rpts and Het states (FDR adjusted one-tailed t-test p: 0.0013, 2.98×10^{-15}, 0.035 and 0.034, for Denisovan aSNPs respectively and p: 1.43×10^{-5}; 7.22×10^{-13}, 4.59×10^{-8} and 9.42×10^{-4}, for Neanderthal aSNPs respectively). In addition, Denisovan, but not Neanderthal, aSNPs were significant overrepresented within TssA, TssAFlnk, TxFlnk, Tx, EnhG, and ReprPCWk (FDR adjusted one-tailed t-test p: 0.019 and 1.4×10^{-18}, 0.029 and 5.48×10^{-11}, 1.33×10^{-6} and 0.0012, respectively) (Fig. 2, Supp. Figs. S7-S9; see Supp Table 5 for full results).

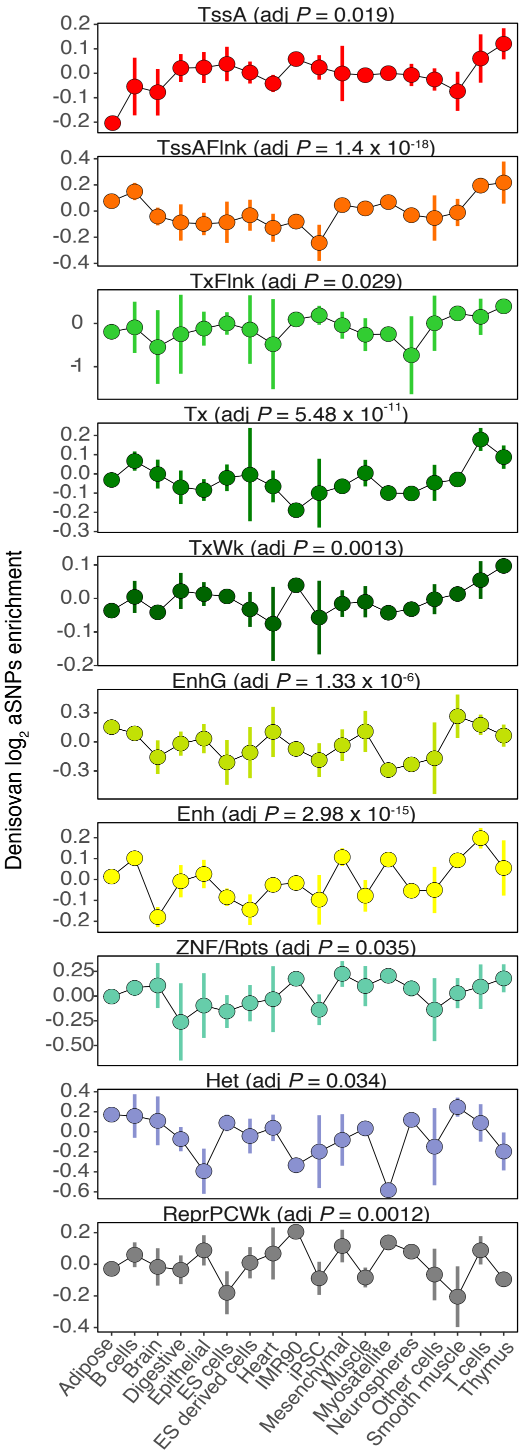

Fig 2. Relative enrichment of Denisovan aSNPs across the 18 tissues and selected chromatin states. The figure shows the $\log_2$-transformed Denisovan aSNP enrichment values across 18 different tissues for common-to-high frequency variants. Only chromatin states where Denisovan aSNPs are significantly overrepresented within immune T cells compared to all other tissues are shown. Dots and cross bars respectively represent the mean and the standard deviation of the $\log_2$-transformed aSNP enrichment scores calculated across the sorted cell types. Chromatin state abbreviations are as in Fig 1. Numbers within parentheses report one-tailed t-test FDR-adjusted p values. Results for all 15 states are available as Supp Fig. S5. For further details, see Supp Table 5.

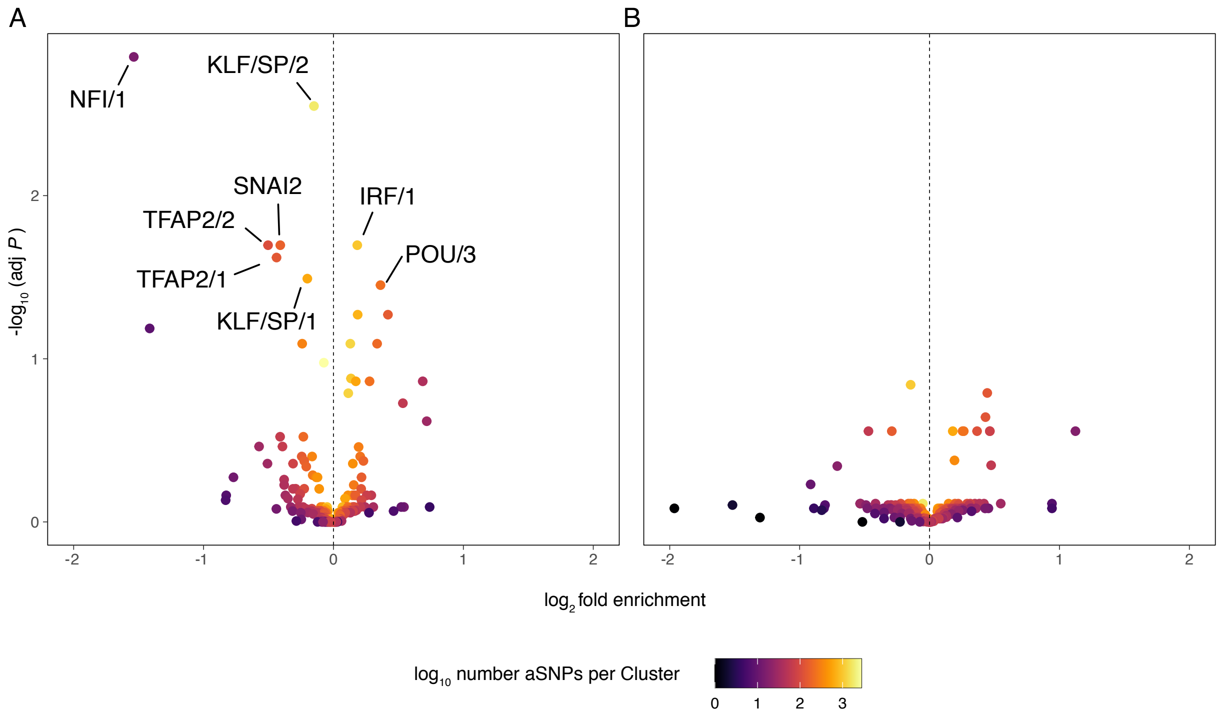

Fig 3. Distribution of TFBS-disrupting aSNPs amongst transcription factor clusters. For each TF cluster we show the $\log_2$-transformed aSNP enrichment only for common-to-high frequency A) Denisovan and B) Neanderthal TFBS-disrupting variants relative to naSNPs against the $-\log_{10}$ FDR-corrected p values resulting from Fisher's exact test. Dots are coloured based on the number of TFBS-disrupting aSNPs annotated within the given cluster. Only clusters with a significant excess of aSNPs are labelled (FDR-adjusted $p \leq 0.05$).

Archaic SNPs impact transcription factor binding sites

We next sought to characterise the potential for variants within our set of aSNPs to alter gene regulation, and the functional mechanisms by which this may occur. Given the observed aSNP enrichment within states commonly associated with cis-regulatory element (CREs) (i.e., Enh, EnhG, TssAFlnk, TxFlnk and TssA), we investigated whether any of our SNPs disrupted known transcription factor binding sites (TFBSs). Thus, we subset our list of aSNPs and naSNPs into those that fell within either CREs-associated or actively transcribed states (i.e., Tx and TxWk) in at least one cell type, retaining 113,567 Denisovan, 64,646 Neanderthal aSNPs and 877,046 naSNPs (79.16%, 76% and 78.95% of the high-confidence subset of variants, respectively). We then analysed these variants using motifbreakR [27], to assess their impact on position weight matrices (PWMs) from 690 different transcription factors drawn from Jaspar 2018 [28] and HOCOMOCO v.11 [29]. After removing duplicated variants predicted to disrupt the same TFBS across the two databases (see Methods for further details), we found that 39,954 Denisovan, 22,756 Neanderthal aSNPs and 310,055 naSNPs (35.24% of the subset of regulatory variants, averaged across all ancestries) were predicted to disrupt at least one DNA motif (Supp File 1-Supp File 3). Looking at the distribution of these variants across the genome we found that, regardless of ancestry, the vast majority of the SNPs were predicted to occur within intronic and intergenic elements. Only 6.0% and 1.6% of SNPs, on average, were respectively located within enhancer or promoter regions (Supp Fig. S8).

To understand whether any of our variants had already been associated with regulatory activities, we intersected the set of TFBS-disrupting SNPs with the list of significant eQTLs from version 8 of the Genotype-Tissue Expression (GTEx) project [30]. As expected given the heavy bias towards individuals of European ancestry within GTEx, only 36 Denisovan and 75 Neanderthal aSNPs and 525 Papuan naSNPs (0.09%, 0.33% and 0.17% of the total number of TFBS-disrupting SNPs, respectively) had already been identified as cis-eQTLs by GTEx. However, when comparing allele frequencies within GTEx and our Papuan dataset, we observed a significant increase in New Guinea for Neanderthal aSNPs and naSNPs (FDR adjusted pair-wise Wilcoxon rank-sum test W = 624, $p = 2.5 \times 10^{-5}$ and W = 31.902, $p < 1.55 \times 10^{-26}$, respectively), whereas no significant difference was observed for Denisovan aSNPs (FDR adjusted pair-wise Wilcoxon rank-sum test W = 360, $p = 0.68$) (Supp. Fig. S9a). While evolutionary forces including drift and local adaptation might explain these frequency differences for Neanderthal and non-archaic variants between Papuans and other human populations, the scarcity of non-European individuals within the GTEx sample limits the inclusion of Denisovan variants, making interpretation of these findings complex.

We then asked whether aSNPs preferentially disrupted TFBS associated with a specific TF relative to naSNPs. To avoid redundancy due to the similarities between predicted motifs across closely related TFs, we took advantage of the work from Vierstra *et al.* [31], which groups PWMs on the basis of sequence similarity into 286 distinct clusters (see Methods). To identify clusters containing a significant excess of aSNPs relative to naSNPs, we calculated the aSNP fold enrichment score by dividing the aSNP:naSNP within a cluster by the mean value of this ratio across all clusters, similar to analyses above. We found that Denisovan aSNPs were significantly enriched within multiple clusters associated with immune-related TFs. Among these, the IRF1, IRF2 and NFAC2 clusters (FDR adjusted Fisher's exact test: $p = 0.01$, 0.01 and 0.036, respectively) contain PWMs respectively recognised by 7, 8 and 1 TFs and are known to regulate cellular responses towards viral infections [32] and to be involved during T cell activation and differentiation [33] (Supp. Fig. S12 and Supp Table 6). No Neanderthal specific enrichment was observed within any cluster.

Removal of low frequency variants resulted in Denisovan aSNPs being significantly enriched within IRF1 and POU3 clusters (FDR adjusted Fisher's exact test: $p = 0.02$ and 0.035, respectively), with the latter containing 7 different homeobox genes including POU2F2 known to be involved in B cell differentiation [34]. Again, no Neanderthal specific enrichment was observed (Fig. 3 and Supp Table 6).

Denisovan TFBS-disrupting aSNPs target genes are involved in immune-related processes

We reasoned that introgressed alleles that confer a selective advantage to carriers should be segregating at medium or high frequency within our sample. Thus, to identify potential instances of adaptive introgression in Papuans, we subset our set of TFBS-disrupting SNPs to those variants with MIAF/DAF ≥ 0.3, and further required that they were predicted to cause a 'strong' disruption on TFBS by motifbreakR (see Methods) [27]. We then focused on those variants annotated within active cell-lineage specific CREs associated chromatin states (TssAFlnk, TxFlnk, Enh and EnhG), arriving at a final set of 952 Denisovan, 1,080 Neanderthal and 13,997 TFBS-disrupting SNPs (2.4%, 4.7% and 4.5% of the total number of TFBS-disrupting SNPs, respectively).

As above, we computed the aSNP enrichment score for these subset of SNPs, by dividing the aSNP:naSNP ratio calculated within each cell type by the mean value across all 111 cell types. We then grouped all cell types into the 18 different tissues and separately compared the distribution of Neanderthal and Denisovan TFBS-disrupting aSNPs to that of naSNPs. Overall, we noted that Neanderthal and Denisovan TFBS-disrupting aSNPs were both significantly depleted from pluripotent or multipotent tissues, but had a broadly similar impact to naSNPs within derived tissues (Supp. Fig. S13). The main exception to these trends were brain and immune T and B cells. Both archaic lineages were significantly overrepresented amongst TFBS-disrupting SNPs within immune T cells (FDR adjusted one-sample t-test: $n = 277$, $p = 1.06 \times 10^{-8}$ and $n = 240$, $p = 1.34 \times 10^{-6}$, respectively for Denisovan and Neanderthal aSNPs). Denisovan TFBS-disrupting aSNPs were also overrepresented within immune B cells, where Neanderthal aSNPs were significantly depleted (FDR adjusted one-sample t-test: $n = 369$, $p = 0.0101$ and $n = 307$, $p = 0.0014$, respectively; Fig. 4a and Supp Table 7). However, Neanderthal TFBS-disrupting aSNPs were significantly enriched within sites active in brain cells, and Denisovan ones significantly depleted (FDR adjusted one-sample t-test: $n = 318$, $p = 0.013$ and $n = 240$, $p = 8.1 \times 10^{-4}$, respectively; Supp. Fig. S13 and Supp Table 7).

In addition, we found that transversions (Tv), which have been predicted to have greater impact on TFBS than transitions (Ts) [35], occurred at approximately similar rates to transitions on CREs active within both immune T cells (FDR adjusted one-sample Wilcoxon rank-sum test: $n = 277$, $p = 8.6 \times 10^{-4}$; n

$= 240$, $p = 4.7 \times 10^{-5}$, for Denisovan and Neanderthal aSNPs, respectively) and B cells (FDR adjusted one-sample Wilcoxon rank-sum test: $n = 369$, $p = 8.6 \times 10^{-4}$; $n = 307$, $p = 6.25 \times 10^{-4}$, for Denisovan and Neanderthal aSNPs, respectively). Similar patterns were also observed for other derived tissues including brain cells, where Neanderthal aSNPs had Ts:Tv significantly lower than 2. These findings are independent of our decision to assess only aSNPs with a predicted 'strong' effect on the surrounding DNA motifs according to motifbreakR, as the equivalent set of naSNPs had an average Ts:Tv of 2.36 (Fig. 4b, Supp. Fig. S11 and Supp Table 8), close to the genome-wide expected 2:1 (Ts:Tv) ratio [36].

Next, given the above observations we examined the biological functions of the set of genes whose regulation may be altered by this subset of TFBS-disrupting SNPs. From the list of common-to-high frequency, strong TFBS-disrupting variants, we retained only those that overlapped CREs active either within immune B or T cells, resulting in 439 Denisovan, 382 Neanderthal aSNPs and 4,875 naSNPs (respectively 46.1%, 35.4% and 34.8% of the filtered set of TFBS-disrupting SNPs). We used GREAT [37] to link these variants with possible target genes, and performed GO enrichment analysis for each ancestry separately. While genes regulated by Neanderthal aSNPs did not show significant enrichment for any specific GO terms associated with immune function, Denisovan aSNP-target genes were strongly enriched for multiple immune related processes (FDR adjusted Hypergeometric test $\leq 0.01\%$). In particular, cellular responses and signalling pathways for cytokines, interferon and interleukins were amongst the top enriched terms (Fig. 5b; Supp Table 9). Interestingly, among the set of putative genes regulated by Denisovan TFBS-disrupting aSNPs, we noted the presence of *TNFAIP3*, *OAS1*, *OAS2* and *OAS3*, all of which have been repeatedly identified as harbouring archaic hominin contribution that impact immune responses to pathogens [38–40].

Taken together, these results point to a substantial contribution of archaic hominins to immune-specific regulatory elements with biological differences between Neanderthals and Denisovan variants, with the latter having contributed to active immune responses against pathogenic infections.

Denisovan TFBS-disrupting aSNPs regulate genes that are differentially expressed across the Indonesian Archipelago

To begin validating the potential influence these TFBS-disrupting SNPs might have on their target genes, we focused on a set of 3,996 genes found to be differentially expressed (DE) at a FDR $\leq 0.01\%$ between 19 Korowai individuals, a genetically Papuan population living on the Indonesian side of New Guinea Island, and 48 individuals of Austronesian ancestry sampled in Mentawai, a small island located off the western coast of Sumatra, in western Indonesia, from [41]. These two populations are indicative of the extremes of the cline of Papuan and Denisovan ancestry that is observed across the Indonesian archipelago [5, 42].

We thus asked how many of the genes associated with the GO terms and whose regulation was predicted to be affected by our subset of TFBS-disrupting variants within immune T and B cells were differentially expressed between Mentawai and Korowai in [41]. Overall, we found that 31, 26 and 426 genes predicted to be regulated by Denisovan, Neanderthal and non-archaic TFBS-disrupting variants (31.0%, 22.4% and 26.1% of those associated with all GO terms) were differentially expressed. There was no significant enrichment for DE genes amongst the subset of Neanderthal aSNPs-target genes (Hypergeometric test: $n = 26$, $p = 0.099$), but we detected a significant over-representation of DE genes within the list of genes regulated by Denisovan and non-archaic TFBS-disrupting variants (Hypergeometric test: $n = 31$, $p = 5.6 \times 10^{-4}$ and $n = 426$, $p = 9.9 \times 10^{-15}$, respectively), suggesting that the set of TFBS-disrupting variants we have identified may impact expression of these genes.

Finally, to understand whether these expression differences might be related to the diverse genetic background of these human populations, we retrieved allele frequencies in western Indonesia for the TFBS-disrupting SNPs from Jacobs *et al.* [5] (we note that this sample does not contain the majority of individuals from Natri *et al.* [41], but it does include nearby groups, and allele frequencies are broadly similar amongst these populations). We focused only on 124 Denisovan and 95 Neanderthal TFBS-disrupting aSNPs associated by GREAT with genes that were DE between Mentawai and Korowai (respectively 28.2% and 24.8% of those considered in the previous step), and asked how many of these were also segregating within western Indonesia. We found that 75 (78.9%) Neanderthal TFBS-disrupting aSNPs were also segregating within western Indonesians, but 104 (83.9%) Denisovan aSNPs are uniquely Papuan, and do not segregate within western Indonesia (Fig. 6c). Notably, among these Denisovan TFBS-disrupting aSNPs 5 (rs370899715, rs372433785, rs146859513, rs375463218, rs372139279) were predicted by GREAT to regulate *OAS2* and *OAS3*, both of which are differentially expressed between Korowai and Mentawai (Fig. 6a,b) [41]. In all five cases, archaic alleles segregate at frequencies between ≥ 0.33 and 0.39 in Papuans, while west Indonesians are fixed for the reference, non-archaic alleles. In addition, alignment with 6 different non-human primates indicated that for 4 of these 5 Denisovan aSNPs the alternative introgressed alleles are derived [5].

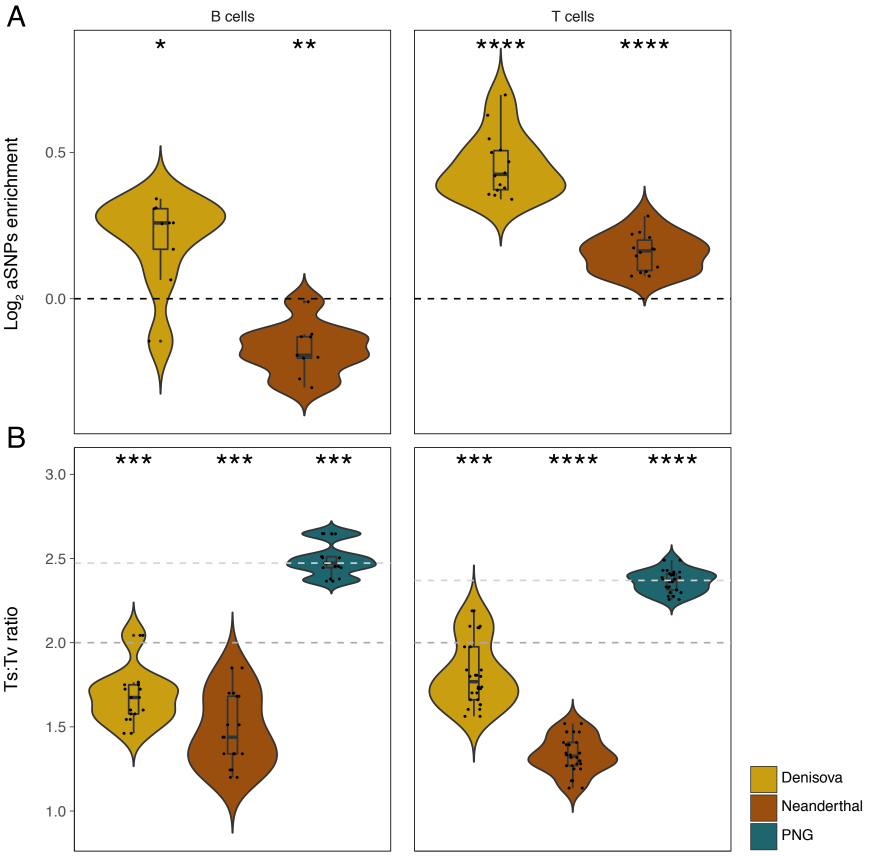

Fig 4. Impact of TFBS-disrupting aSNPs within immune T and B cells. For the subset of TFBS-disrupting variants segregating at frequencies ≥ 0.3 within Papua New Guinea and predicted to cause a strong disrupting of the DNA motif by motifbreakR, we show A) the $\log_2$-transformed aSNP enrichment over naSNPs and B) their Ts:Tv ratios calculated for either tissue . Dashed lines in B) report the expected genome-wide 2:1 Ts:Tv (dark grey) and the median naSNP Ts:Tv calculated across the cell types (light grey). Violin plots represent the full distribution of A) the $\log_2$-transformed aSNP enrichment and B) the Ts:Tv ratios for the sorted cell types. Lower and upper hinges of the boxplots correspond to the first and third quartiles of the distribution, whiskers extend to a maximum of 1.5 × IQR beyond the box. Asterisks indicate whether for the given tissue A) the aSNP enrichment was significantly different from its expected mean value calculated across all tissues and B) the Ts:Tv ratio was significantly different from an expected value of 2 (FDR-adjusted one-sample A) t-test and B) Wilcoxon rank-sum test: $**** = p \leq 0.0001$; $**\ p \leq 0.01$ and $* = p \leq 0.05$). For further details see Supp Table 7-Supp Table 8.

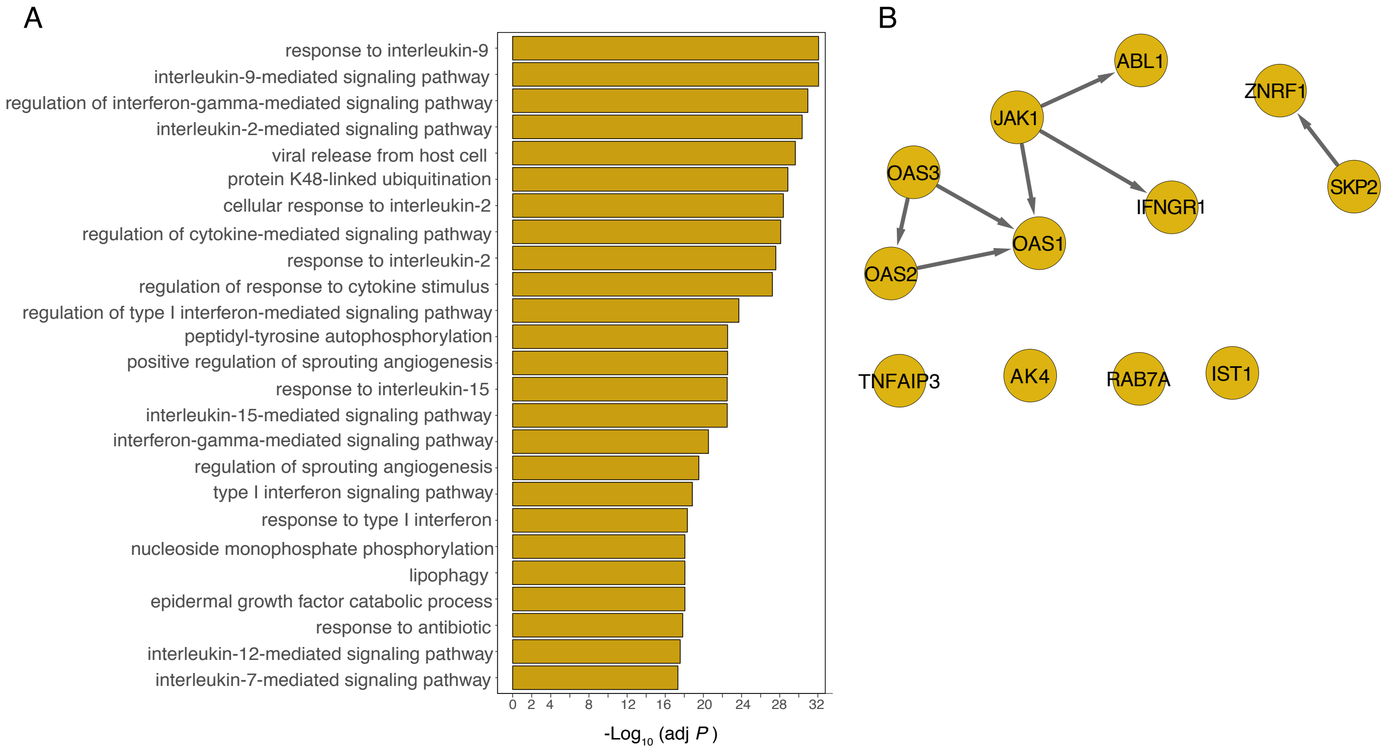

Fig 5. Denisovan TFBS-disrupting aSNPs impact immune processes. For the genes predicted to be regulated by the subset of common-to-high-frequency Denisovan TFBS-disrupting aSNPs active within immune cells we show A) their top 25th mostly enriched GO terms along with their relative -log$_{10}$ Benjamini-Hochberg (BH) adjusted *p* values; B) protein-protein interactions for the set of genes also predicted to be regulated by Denisovan TFBS-disrupting aSNP within the most significant GO terms.

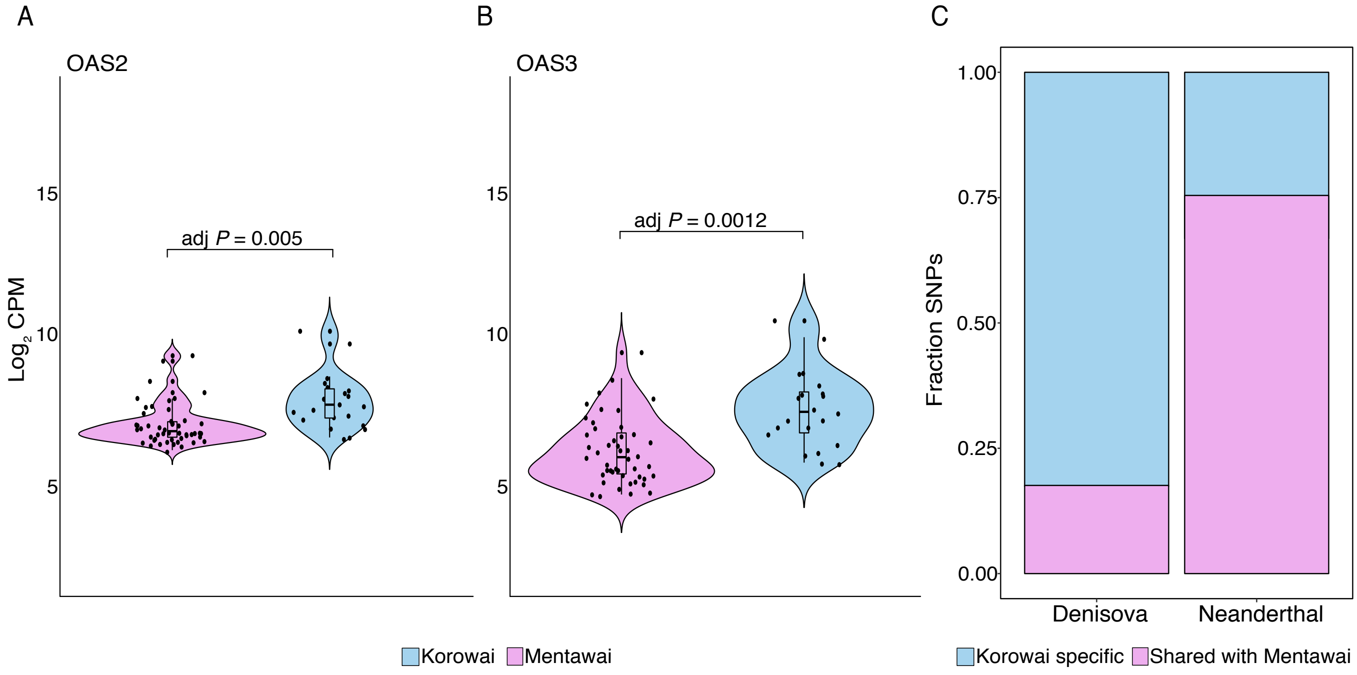

Fig 6. *OAS* gene expression between western Indonesian and New Guinean populations. Log_2 RNA-seq counts per million (CPM) in whole blood for a) *OAS2* and b) *OAS3* between Korowai (New Guinea) and Mentawai (western Indonesia), from [41]. Violin plots represent the full distribution of the $\log_2$ CPM. Lower and upper hinges of the boxplots correspond to the first and third quartiles of the distribution, whiskers extend to a maximum of 1.5 × IQR beyond the box. Reported FDR-adjusted p values were taken from [41]. c) Fraction of TFBS-disrupting aSNPs associated with DE genes between West Indonesia and New Guinea that were either shared with Mentawai or observed only in Korowai.

Discussion

There is significant interest in understanding the functional consequences of archaic introgression. Evidence indicates that both Denisovan and Neanderthal aSNPs, especially those within protein coding and conserved non-coding elements, are mostly deleterious and negatively selected in modern humans [11, 24]. Similar findings have been recently reported for highly pleiotropic enhancers, where aSNPs are depleted likely as a consequence of their potential to perturb gene expression across multiple tissues [9]. Nevertheless, out of the substantial number of archaic variants still segregating within present-day populations, a large fraction still falls within genomic regions that show strong evidence of functional activity. Studies conducted primarily on Neanderthal introgressed DNA have suggested a non-negligible contribution to gene expression variation in modern humans [7, 8, 17], with repeated examples of Neanderthal archaic variants falling within regulatory elements or the seed region of mature micro-RNAs predicted to affect transcriptional and post-transcriptional regulatory processes [7, 43].

In this study, we have taken advantage of a recently published dataset [5] to investigate the landscape of archaic introgression in individuals of Papuan ancestry, the functional consequences of which remained poorly understood. This has allowed us to characterise the putative contribution of Denisovan DNA, which is known to account for up to 5% of the genome of present-day Papuans [4, 44], while also comparing it to that of Neanderthal DNA. We specifically analysed all of our variants across multiple cell types and functional chromatin states aiming to account for the strong dependency on the cells' 1) chromatin landscapes [26]; 2) developmental stages [45] and 3) experienced environmental stimuli [46], when assessing the activities of introgressed alleles.

First, we confirm previous reports that aSNPs mostly occur within non-coding sequences and are significantly depleted from protein coding elements [22, 24]. At the same time, we have shown that a large number of aSNPs lie within regions of the genome that are highly constitutive and/or functionally inert, perhaps a consequence of purifying selection being less powerful in these elements. However, we find that the remaining aSNPs are overrepresented within active chromatin with notable differences between the two archaic ancestries. Indeed, while Neanderthal and Denisovan aSNPs are both enriched within actively transcribed regions, only Denisovan variants are consistently overrepresented within CREs-associated states. Moreover, aSNPs annotated within these states, and in particular within enhancers and genic enhancers, often target highly cell-specific functional elements. Therefore, our results support the current hypothesis of a potential archaic contribution to pre-and post-transcriptional regulatory processes [7, 43], with important differences among tissues between Neanderthal and Denisovan variants.

Second, we have extensively characterised the impact of introgressed variants on transcription factor binding sites, a well-known mechanism with clear potential to impact phenotypic variation [47]. At least 27.3% of archaic SNPs (averaged between the ancestries) falling within transcribed regions and CRE-associated states are predicted to modify interactions between transcription factors and their cognate binding sites. However, not all families of TFs are equally affected and, compared to non-archaic SNPs, Denisovan aSNPs excessively target clusters whose motifs are recognised by immune-related TFs while Neanderthal variants are not enriched for any specific family. Therefore, these findings while indicating that admixture with archaic hominins might have facilitated adaptation by expanding modern humans to new environments [14], this contribution likely differs between Neanderthal and Denisovan among present-day populations.

Third, we consistently find evidence for a sizeable contribution of Denisovan archaic DNA to Papuan regulatory variation, especially within immune-related traits. Particularly, compared to modern human variants, we found a significant excess of Denisovan TFBS-disrupting variants segregating at frequencies ≥ 0.3 within Papuans which are are predicted to impact gene expression within both immune B and T cells. Again highlighting that different archaic hominins made very different contributions to present-day Papuans, only the set of genes predicted to be regulated by these Denisovan aSNPs was strongly associated with immune-related processes. This finding further suggests that Denisovan alleles contribute to Papuan phenotypic variation in immune physiological responses actively mounted against pathogenic infections, mirroring similar findings made for Neanderthal DNA in Europeans [16, 17, 48].

Notably, *OAS1*, *OAS2* and *OAS3* were among the genes whose regulation might be affected by Denisovan TFBS-disrupting aSNPs (Fig. 5). These genes belong to a family of pattern-recognition receptors involved in innate immune responses against both RNA and DNA viruses, with *OAS3* considered to be essential in reducing viral titers during Chikungunya, Sindbis, influenza or vaccinia viral infections [49]. Furthermore, at least two previous studies have shown the presence of both Neandertal and Denisovan archaic haplotypes segregating at this locus respectively within European [39] and Papuan [40] individuals. In addition, Sams *et al.* found two variants (rs10774671, rs1557866) within these Neanderthal haplotypes which are respectively associated with the codification of different *OAS1* splicing isoforms and with a reduction in *OAS3* expression levels, this latter only upon viral infection [39].

Our previous work has found that both *OAS2* and *OAS3* are differentially expressed between western Indonesian and Papuan populations [41]. Here we report a cluster of five SNPs (rs370899715, rs372433785, rs146859513, rs375463218 and rs372139279), located roughly 43 kb upstream of *OAS2* and *OAS3*, all of which are predicted to strongly alter the ability of different transcription factors, including *IRF4*, *IRF8*, *SPI1* and *SPIB*, to bind to their underlying DNA motifs. Interestingly, in all cases, the reference allele is fixed within western Indonesian populations, whereas in Papuans the alternative archaic alleles segregate at frequencies between ≥ 0.33 and 0.39. Given the high pathogenic load that characterises the coastal areas of New Guinea Island, this and the other findings strongly argue that admixture with archaic hominins might have shaped the immune-related phenotypes of early modern humans in the region, favouring adaptation to the local environment [17, 48].

Overall, we have shown multiple lines of evidence that argue for a functional contribution of Denisovan DNA to present-day Papuans, especially at loci implicated in immune response. While further studies, including *in vitro* validation of these findings will be vital to characterise the actual molecular functions of such introgressed alleles, and to establish causative links with any specific trait, our work indicates that admixture with archaic hominins other than Neanderthals might have been a fundamental event in the evolutionary history also for non-European populations. This is essential if we are to improve our understanding of how past events contributed to the phenotypic differences observable among humans today.

Acknowledgements

We would like to thank all members of the Gallego Romero and McCarthy groups as well as Laura E. Cook (University of Melbourne) and Ashlee Hutchinson (University of Melbourne) for their helpful discussions and comments on the manuscript. We also wish to acknowledge all of the study participants who generously consented to genome sequencing in the original study, and the leadership of Herawati Sudoyo and Chelzie Crenna Darusalam (Eijkman Institute for Molecular Biology) in generating these datasets. This work was supported by an award from the Leakey Foundation and by Australian Research Council Discovery Project DP200101552, both to I.G.R.

Author contributions

D.M.V. and I.G.R. designed the study. D.M.V. performed the analyses. G.S.J., N.B. and M.P.C. provided raw data. D.M.V. and I.G.R. wrote the paper with input from all other authors. All authors approved the manuscript before submission.

Materials and Methods

Samples

All human genome data in this study was taken from Jacobs *et al.* [5]. All collections followed protocols for the protection of human subjects established by institutional review boards at the Eijkman Institute (EIREC #90) and Nanyang Technological University (IRB-2014-12-011); the analyses in this publication were additionally approved by University of Melbourne's Human Ethics Advisory Group (1851585.1). Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology (RISTEK). All individuals gave their full informed consent to participate in the study.

Variant filtering and selection

In order to curate a high confidence set of regional SNPs, we started by analysing two BED files, one per archaic ancestry source (i.e., Denisovan or Neanderthal), each containing the genomic location (hg19) of all 8,337,067 SNPs identified in an alignment of 72 individuals of Papuan ancestry, the Altai Neanderthal and Altai Denisovan, with heterozygotes in the archaic genomes masked (see [5] for detailed methods and sample information). Variants were initially labelled as aSNPs or naSNPs depending on whether they fell within archaic haplotypes (Supp Table 1). To avoid potential confounding errors we first removed all singletons found across the two files and then merged their information. We then relabelled as naSNPs only those SNPs that consistently fell outside both sets of archaic haplotypes and for which alignment with 6 different non-human primates provided the ancestral information. The remaining variants were sorted into Denisovan aSNPs or Neanderthal aSNPs based on the inferred ancestry of the haplotype of origin. aSNPs in common between the two archaic haplotypes were assigned, where possible, to one of the two ancestries by considering whether the state of the main introgressed allele in Papuans resembled the homozygous state of the Altai Neanderthal [50] or of the Altai Denisovan [51] reference genomes. Variants that could not be disambiguated in this manner were labelled as ambiguous and not included in the downstream analyses (the number of SNPs retained after each step is detailed in Supp Table 1).

We then calculated allele frequencies for aSNPs and naSNPs by considering the fraction of Papuan individuals carrying the variant within the haplotype of the respective ancestry. In particular, for naSNPs we computed the derived allele frequency (DAF) by dividing the number of observations of the derived allele in individuals carrying the modern haplotype over the total number of chromosomes at that locus (generally 144). Likewise, for each aSNP we calculated the major introgressed allele frequency (MIAF), defined as the number of observations of the major allele in individuals carrying an archaic haplotype, divided by the total number of chromosomes. In addition, for each aSNP we also calculated the frequency of the major introgressed allele in individuals carrying a modern human haplotype, which we later used to filter our variants (see below). All SNPs were subsequently sorted into two main frequency ranges within our Papuan sample: low frequency (i.e., MIAF/DAF frequency < 0.05) and common-to-high frequency (i.e., MIAF/DAF frequency ≥ 0.05).

We then performed a series of filtering steps designed to refine our set of aSNPs. For every aSNP we first computed the absolute difference between the MIAF and the frequency of the same allele within modern haplotypes. To control for incomplete lineage sorting between humans, Neanderthals and Denisovans [2] we discarded all aSNPs where this difference was ≤ 0.25. Following annotation with the Ensembl Variant Effect Predictor (VEP) tool (see below), we further removed all aSNPs for which the main introgressed allele is segregating at frequencies ≥ 0.005 within sub-Saharan African populations from the 1000 Genomes Project [23]. We also applied this filtering criteria to the non-archaic derived alleles to ensure a fairer comparison between naSNPs and aSNPs (Supp Table 1).

SNP annotation and external data integration

SNPs were annotated using the Ensembl VEP tool [25] to assess both their distribution across multiple genomic elements, and their allele frequencies across populations in the 1000 Genomes Project [23]. In addition, we downloaded 15-state hg19 mnemonics BED files containing cell type-specific segmented chromatin state predictions for 111 cell types from the Roadmap Epigenomics project [26], `https://egg2.wustl.edu/roadmap/web_portal/index.html`. We then intersected our SNPs with this chromatin state information to define their potential activity across every cell type.

Enrichment of aSNPs across chromatin states and cell types

Following SNP annotation with the Roadmap Epigenomics data, we quantified the impact of our sets of SNPs across each chromatin state/cell type combination in a manner similar to Silvert *et.al* [7]. Specifically, for each archaic ancestral component we independently calculated the ratio of aSNPs to naSNPs (aSNP:naSNP) within each chromatin state/cell type combination. This yields a matrix of 111 × 15 values, one for each combination, from which we defined a genome-wide expectation of aSNP distributions by calculating the mean of the data. We then computed an aSNP enrichment score for each cell type/state combination by normalising the aSNP:naSNP by the overall mean value for the relevant archaic ancestry. Finally, we tested whether any chromatin state showed a consistent enrichment of aSNPs across all cell types by computing one-tailed t-tests on the $\log_2$-transformed normalised aSNP enrichment values after correcting p values for multiple hypothesis testing (False Discovery Rate-FDR) (Supp Table 2-Supp Table 3).

To test whether aSNPs in any chromatin states had excessive impacts on immune cells, we normalised the aSNP:naSNP ratio calculated for each cell-type/state combination by the mean aSNP:naSNP ratio for the relevant chromatin state (not averaged across the entire data set). We then grouped similar cell types into the 18 broad tissue categories, according to the Roadmap Epigenomics [26], and performed a one-tailed t-test on the $\log_2$-transformed normalised aSNPs enrichment values to assess whether the aSNP enrichment within immune T cells was larger than that for all the other tissues combined. Resulting p values were FDR-corrected for multiple hypothesis (Supp Table 5).

Finally, to assess whether SNPs fell within constitutive or cell-specific functional elements, we again grouped the 111 cell types into the 18 tissues, and counted across how many of these any SNP, were annotated to a particular chromatin state. This returned a measurement of the potential pleiotropic effect of each SNP within our dataset. For each chromatin state, For each chromatin state, we then independently tested whether the pleiotropic activities of the subset of elements carrying aSNPs significantly differed from those of naSNPs-containing elements, by performing two-tailed t-tests and FDR-adjusting the resulting p values (Supp Table 4).

Variant impact on transcription factor binding sites

To characterise the impact of our sets of SNPs on transcription factor binding sites (TFBSs), we focused on variants that fell within transcribed and weakly transcribed states (Tx and TxWk), transcription start sites (TssA), enhancers (Enh), genic enhancers (EnhG), promoters (TssAFlnk) or promoter/enhancer-like associated states (TxFlnk) in at least one Roadmap Epigenomics cell type. We then used the motifbreakR R package (v 2.2.0) [27], retrieving the collection of DNA motifs from Jaspar 2018 [28] and HOCOMOCO v.11 [29] databases, to assess whether any of our SNPs overlapped any TFBS and to calculate the score of the position weight matrices for the sequences containing the two alleles. Motifs were matched to the hg19 genome using a p-value threshold of 1^{-5}, and the allelic impact on each identified motif was quantified via the built-in sum of log-probabilities algorithm, setting a background probability of 0.25 for each nucleotide in any given genomic sequence. When a SNP was predicted to disrupt the same binding site across both databases, we counted these as a single observation. The final set of SNPs, together with their affected motifs and their predicted disruptive potential, are listed in Supp File 1 -Supp File 3.

To avoid redundancy between TFs due to similar binding preferences across motifs, we retrieved motif clustering information from Vierstra *et al.* [31] and grouped all our TFs accordingly, again counting instances where a SNP was predicted to disrupt the same genomic location across multiple motifs in the same cluster as a single observation. To identify clusters containing an excess of aSNPs we calculated the aSNP:naSNP for each cluster and divided this value by the mean computed across all clusters. For each ancestry we then performed a Fisher's exact test, and FDR-adjusted the resulting p values, to assess whether aSNPs and naSNPs were equally disrupting TFBSs associated to each given cluster (Supp Table 6).

Comparison of TFBS-disrupting SNPs with known cis-eQTLs

To begin understanding the potential regulatory activity of our TFBS-disrupting SNPs we downloaded the list of V8 expression quantitative trait loci (eQTLs) and their eGenes from the GTEx web portal `https://www.gtexportal.org/home/` [30]. We then used liftOver from the rtracklayer R package (v 1.48.0) [52] to convert files from hg38 to hg19-related coordinates, retaining only significant eSNP-eGene pairs (FDR q-value ≤ 0.05). We then merged this list of variants with our set of TFBS-disrupting SNPs and for those where both the reference and alternative alleles matched between the two datasets we compared the reported GTEx minor allele frequency (maf) with the observed allele frequency (i.e.,MIAF/DAF) within our Papuan sample across all three ancestries. For each ancestry, we finally quantified significant difference in the allele frequencies between the two datasets by performing a paired Wilcoxon rank-sum test followed and FDR-adjusting the resulting set of p values.

Assessing the impact of TFBS-disrupting SNPs across tissues

To highlight variants of potential functional importance, we applied multiple filters to our set of TFBS-disrupting archaic and non-archaic SNPs. First, we removed all variants with frequency (MIAF/DAF) < 0.3 within our Papuan sample and retained only SNPs where the allelic change was predicted to cause a 'strong' disruption on the TFBS by motifbreakR (i.e. delta PWM ≥ 0.7), which reflects the difference in the score of the position weight matrices when the sequences contain either the reference or the alternative allele. We further focused on variants that fell within chromatin states associated with tissue-specific cis-regulatory elements (CREs) (i.e., TssAFlnk, TxFlnk, Enh and EnhG) [26]. From this subset of variants we then counted the number of variants annotated within every cell type and re-calculated the aSNP enrichment by dividing the TFBS-disrupting aSNP:naSNP ratio within every cell type by the mean value calculated across all cell types. As above, we then grouped similar cell types into the 18 different tissues and, for each archaic ancestry, performed a one-tailed t-test on the resulting $\log_2$-transformed aSNPs enrichment values testing any eventual statistical difference from an expected value of zero and FDR-correcting the resulting set of p values (Supp Table 7).

In parallel, for both aSNPs and naSNPs we independently calculated the transition over transversion (Ts:Tv) ratio within every cell type. We then grouped together similar cell types into the 18 broader tissues and performed one-tailed Wilcoxon rank-sum tests, FDR-adjusting the resulting set of p values, to quantify any significant deviations from an expected 2:1 Ts:Tv ratio (Supp Table 8).

Gene Ontology enrichment analysis and protein-protein interactions

To understand the biological processes our variants might be involved in, we analysed the set of genes whose expression could be affected by our set of TFBS-disrupting SNPs. In particular, from the refined list of TFBS-disrupting variants described above we focused on those active in at least one immune B and/or T cell type. We then used the rGREAT R package (v 1.20.0) [53] to assign these variants to the regulatory domain of the two nearest genes located within 1 megabase of distance in either direction from the genes' transcription start sites. For each archaic ancestry, we then independently performed a Gene Ontology (GO) enrichment analysis [54], using as a background set the test set itself as well as the equivalent set of TFBS-disrupting naSNPs-associated genes. This approach allowed us to account for independence among the multiple introgression events, highlighting pathways the two archaic variants might be differentially involved into (Supp Table 9).

For visualisation purposes, we retrieved the top 25th Denisovan GO enriched terms as well as their associated genes that were also predicted to be regulated by our set TFBS-disrupting SNPs. We used the stringApp v 1.3.0 [55] in Cytoscape v 3.7.1 [56] to visualise whether they might participate in common protein interaction networks.

Differences in gene expression between Papuans and West Indonesians

To assess whether any SNP contained in the subset of TFBS-disrupting variants annotated within immune T and B cell are likely to modulate the expression of their predicted target genes, we retrieved whole blood gene expression (RNA-seq) estimates across three Indonesian populations with varying degrees of Papuan ancestry [41]. We then retained 3,996 genes originally identified as DE between 48 Mentawai (West Indonesia, no Papuan ancestry) individuals and 19 Korowai (New Guinea Island, genetically Papuan) individuals (FDR $\leq 1\%$) and intersected this list with the set of genes predicted to be regulated by our

TFBS-disrupting SNPs. Genes were assigned to the three different ancestral components based on the ancestry of the associated regulatory variants. For each ancestry, we then tested whether the regulated set of genes contained an excess of DE instances by performing a hypergeometric test. Finally, we assessed whether the archaic alleles underlying these TFBS-disruption events were equally shared across the two ends of the Indonesian archipelago, by examining their allele frequency data in west Indonesian populations as reported by Jacobs *et al.* [5].

Code availability

All analyses were performed using R software v 4.0.0. The complete set of scripts used for the analyses is available at `https://github.com/dvespasiani/Archaic_introgression_in_PNG`

Supplementary materials

Supplementary materials include 11 figures and 9 tables and 3 files.

Supp Table 1 Table reporting the number of aSNPs and naSNPs found in PNG

Supp Table 2 Table reporting the aSNPs enrichment across each chromatin states/cell type combination

Supp Table 3 Table reporting p values related to the aSNPs enrichment for each chromatin states/cell type combination

Supp Table 4 Table reporting p values for the pleiotropic activities of functional elements carrying aSNPs

Supp Table 5 Table reporting p values related to the aSNPs enrichment within immune T cells for each chromatin states compared to all other tissues

Supp Table 6 Table reporting the $\log_2$ fold change and p values for the aSNPs enrichment across each cluster of TFs.

Supp Table 7 Table reporting the $\log_2$-transformed archaic enrichment and p values for the refined subset of TFBS-disrupting SNPs across each cell type/tissues

Supp Table 8 Table reporting the Ts:Tv ratio and related p values calculated across each cell type

Supp Table 9 Table reporting all the significant GO terms for Denisovan, Neanderthal aSNPs and non-archaic variants

Supp File 1 File reporting all Denisovan TFBS-disrupting aSNPs

Supp File 2 File reporting all Neanderthal TFBS-disrupting aSNPs

Supp File 3 File reporting all modern TFBS-disrupting naSNPs

References

1. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328(5979):710–723. doi:10.1126/science.1188021.

2. Reich D, Green RE, Kircher Mt. Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature. 2010;468(7327):1053–1060. doi:10.1038/nature09710.

3. Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LFK, Arauna LR, et al. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biology. 2019;20(1):77. doi:10.1186/s13059-019-1684-5.

4. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. American Journal of Human Genetics. 2011;89(4):516–528. doi:10.1016/j.ajhg.2011.09.005.

5. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell. 2019; p. 1010–1021. doi:10.1016/j.cell.2019.02.035.

6. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018;173(1):53–61.e9. doi:10.1016/j.cell.2018.02.031.

7. Silvert M, Quintana-Murci L, Rotival M. Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. American Journal of Human Genetics. 2019;104(6):1241–1250. doi:10.1016/j.ajhg.2019.04.016.

8. McCoy RC, Wakefield J, Akey JM. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell. 2017;168(5):916–927.e12. doi:10.1016/j.cell.2017.01.038.

9. Telis N, Aguilar R, Harris K. Selection against archaic DNA in human regulatory regions. bioRxiv. 2019; p. 708230. doi:10.1101/708230.

10. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current Biology. 2016;26(9):1241–1247. doi:10.1016/j.cub.2016.03.037.

11. Petr M, Pääbo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. Proceedings of the National Academy of Sciences of the United States of America. 2019;116(5):1639–1644. doi:10.1073/pnas.1814338116.

12. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507(7492):354–357. doi:10.1038/nature12961.

13. Dannemann M, Kelso J. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. American Journal of Human Genetics. 2017;101(4):578–589. doi:10.1016/j.ajhg.2017.09.010.

14. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Current Biology. 2016;26(24):3375–3382. doi:10.1016/j.cub.2016.10.041.

15. Dolgova O, Lao O. Evolutionary and medical consequences of archaic introgression into modern human genomes. Genes. 2018;9(7). doi:10.3390/genes9070358.

16. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. American Journal of Human Genetics. 2016;98(1):5–21. doi:10.1016/j.ajhg.2015.11.014.

17. Quach H, Rotival M, Pothlichet J, Loh YHE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell. 2016;167(3):643–656.e17. doi:10.1016/j.cell.2016.09.024.

18. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nature Reviews Genetics. 2018;19(3):175–185. doi:10.1038/nrg.2017.89.

19. Popejoy Alice B, M FS. Genomics is failing on diversity. Nature. 2016;538:161–164.

20. Franchini LF, Pollard KS. Human evolution: The non-coding revolution. BMC Biology. 2017;15(1):1–12. doi:10.1186/s12915-017-0428-9.

21. Gallego Romero I, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics. 2012-07;13(7):505–516. doi:10.1038/nrg3229.

22. Dannemann M, Prüfer K, Kelso J. Functional implications of Neandertal introgression in modern humans. Genome Biology. 2017;18(1):1–11. doi:10.1186/s13059-017-1181-7.

23. Auton A, Abecasis GR, Altshuler DM, Durbin RMt. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.

24. Harris K, Nielsen R. The genetic cost of neanderthal introgression. Genetics. 2016;203(2):881–891. doi:10.1534/genetics.116.186890.

25. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17(1):1–14. doi:10.1186/s13059-016-0974-4.

26. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–329. doi:10.1038/nature14248.

27. Coetzee SG, Coetzee GA, Hazelett DJ. MotifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics. 2015;31(23):3847–3849. doi:10.1093/bioinformatics/btv470.

28. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research. 2017;46(D1):D260–D266. doi:10.1093/nar/gkx1126.

29. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46(D1):D252–D259.

30. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv. 2019;doi:10.1101/787903.

31. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, et al. Global reference mapping and dynamics of human transcription factor footprints. bioRxiv. 2020; p. 2020.01.31.927798. doi:10.1101/2020.01.31.927798.

32. Ren G, Cui K, Zhang Z, Zhao K. Division of labor between IRF1 and IRF2 in regulating different stages of transcriptional activation in cellular antiviral activities. Cell & Bioscience. 2015;5(1):17. doi:10.1186/s13578-015-0007-0.

33. Müller MR, Rao A. NFAT, immunity and cancer: a transcription factor comes of age. Nature Reviews Immunology. 2010;10(9):645–656. doi:10.1038/nri2818.

34. Schubart K, Massa S, Schubart D, Corcoran LM, Rolink AG, Matthias P. B cell development and immunoglobulin gene transcription in the absence of Oct-2 and OBF-1. Nature Immunology. 2001;2(1):69–74. doi:10.1038/83190.

35. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, et al. Transversions have larger regulatory effects than transitions. BMC Genomics. 2017;18(1):1. doi:10.1186/s12864-017-3785-4.

36. Collins DW, Jukes TH. Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence; 1994.

37. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology. 2010;28(5):495–501. doi:10.1038/nbt.1630.

38. Zammit NW, Siggs OM, Gray PE, Horikawa K, Langley DB, Walters SN, et al. Denisovan, modern human and mouse TNFAIP3 alleles tune A20 phosphorylation and immunity. Nature Immunology. 2019;20(10):1299–1310. doi:10.1038/s41590-019-0492-0.

39. Sams AJ, Dumaine A, Nédélec Y, Yotova V, Alfieri C, Tanner JE, et al. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. Genome Biology. 2016;17(1):1–15. doi:10.1186/s13059-016-1098-6.

40. Mendez FL, Watkins JC, Hammer MF. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. Molecular Biology and Evolution. 2012;29(6):1513–1520. doi:10.1093/molbev/msr301.

41. Natri HM, Bobowik KS, Kusuma P, Crenna Darusallam C, Jacobs GS, Hudjashov G, et al. Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the Indonesian archipelago. PLOS Genetics. 2020;16(5):1–21. doi:10.1371/journal.pgen.1008749.

42. Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, et al. Complex Patterns of Admixture across the Indonesian Archipelago. Molecular Biology and Evolution. 2017;34(10):2439–2452. doi:10.1093/molbev/msx196.

43. Lopez-Valenzuela M, Ramírez O, Rosas A, García-Vargas S, De La Rasilla M, Lalueza-Fox C, et al. An ancestral miR-1304 allele present in neanderthals regulates genes involved in enamel formation and could explain dental differences with modern humans. Molecular Biology and Evolution. 2012;29(7):1797–1806. doi:10.1093/molbev/mss023.

44. 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(6282):235–240.

45. Strober BJ, Elorbany R, Rhodes K, Krishnan N, Tayeb K, Battle A, et al. Dynamic genetic regulation of gene expression during cellular differentiation. Science. 2019;364(6447):1287–1290. doi:10.1126/science.aaw0040.

46. Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, et al. Latent enhancers activated by stimulation in differentiated cells. Cell. 2013;152(1-2):157–171. doi:10.1016/j.cell.2012.12.018.

47. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. PLOS Genetics. 2015;11(1):1–8. doi:10.1371/journal.pgen.1004857.

48. Dannemann M, Andrés AM, Kelso J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. American Journal of Human Genetics. 2016;98(1):22–33. doi:10.1016/j.ajhg.2015.11.015.

49. Li Y, Banerjee S, Wang Y, Goldstein SA, Dong B, Gaughan C, et al. Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(8):2241–2246. doi:10.1073/pnas.1519657113.

50. 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(7481):43–49. doi:10.1038/nature12886.

51. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 2012;338(6104):222–226. doi:10.1126/science.1224344.

52. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–1842. doi:10.1093/bioinformatics/btp328.

53. Gu Z. rGREAT: Client for GREAT Analysis; 2018.

54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–29.

55. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. Journal of Proteome Research. 2019;18(2):623–632. doi:10.1021/acs.jproteome.8b00702.

56. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13(11):2498–2504. doi:10.1101/gr.1239303.

57. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. Bioinformatics. 2017;.

Supplementary Materials

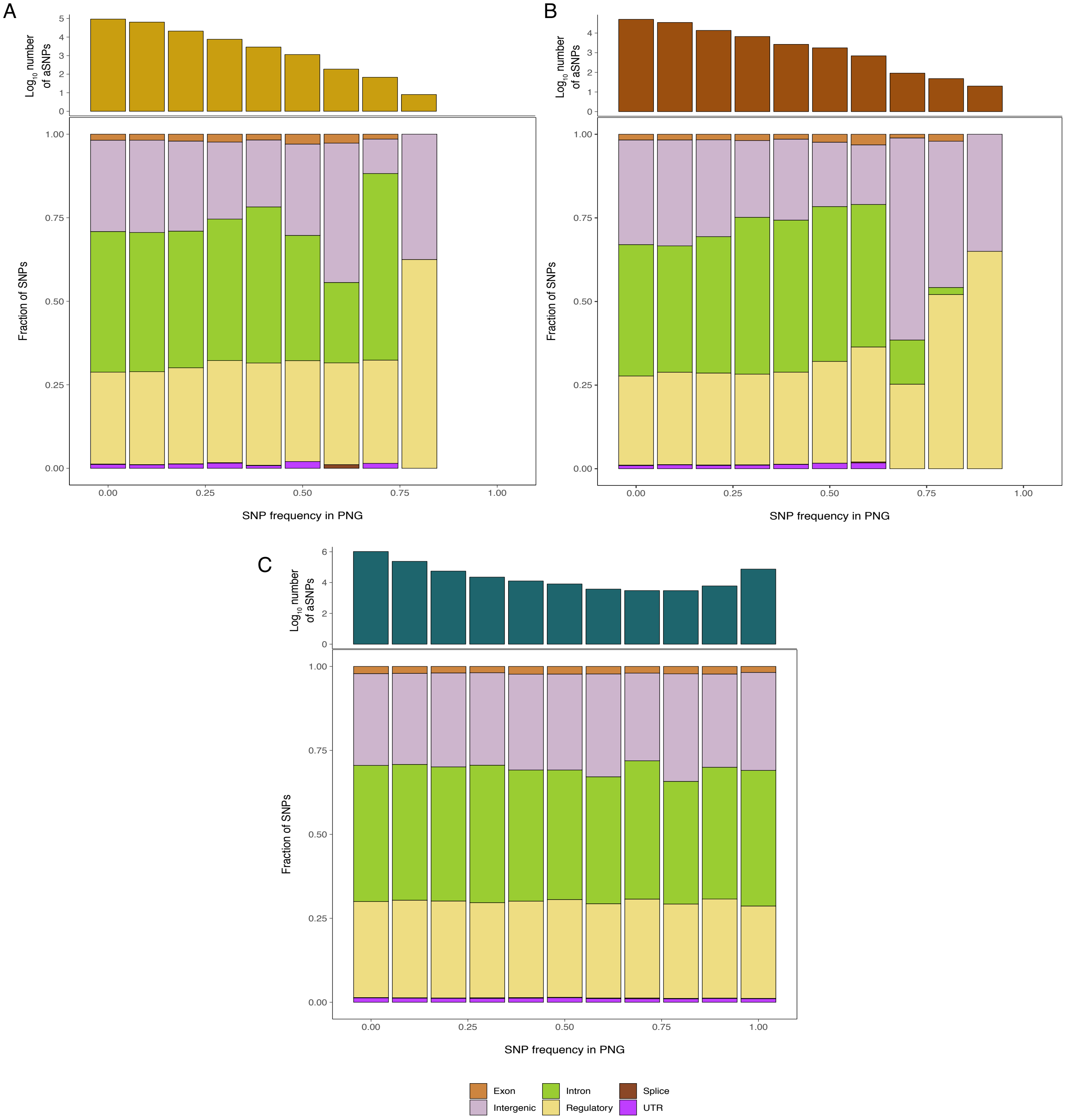

Fig S1. SNP distribution across multiple genomic elements. For A) Denisovan; B) Neanderthal and C) Non-archaic ancestries we show the proportional distribution of each SNP across multiple genomic elements, following variant annotation using the ensembl VEP tool. x-axis report the MIAF/DAF frequency bins in which aSNPs and naSNPs were respectively grouped into. Top bars report the $\log_{10}$ transformed raw number of variants within each frequency bin.

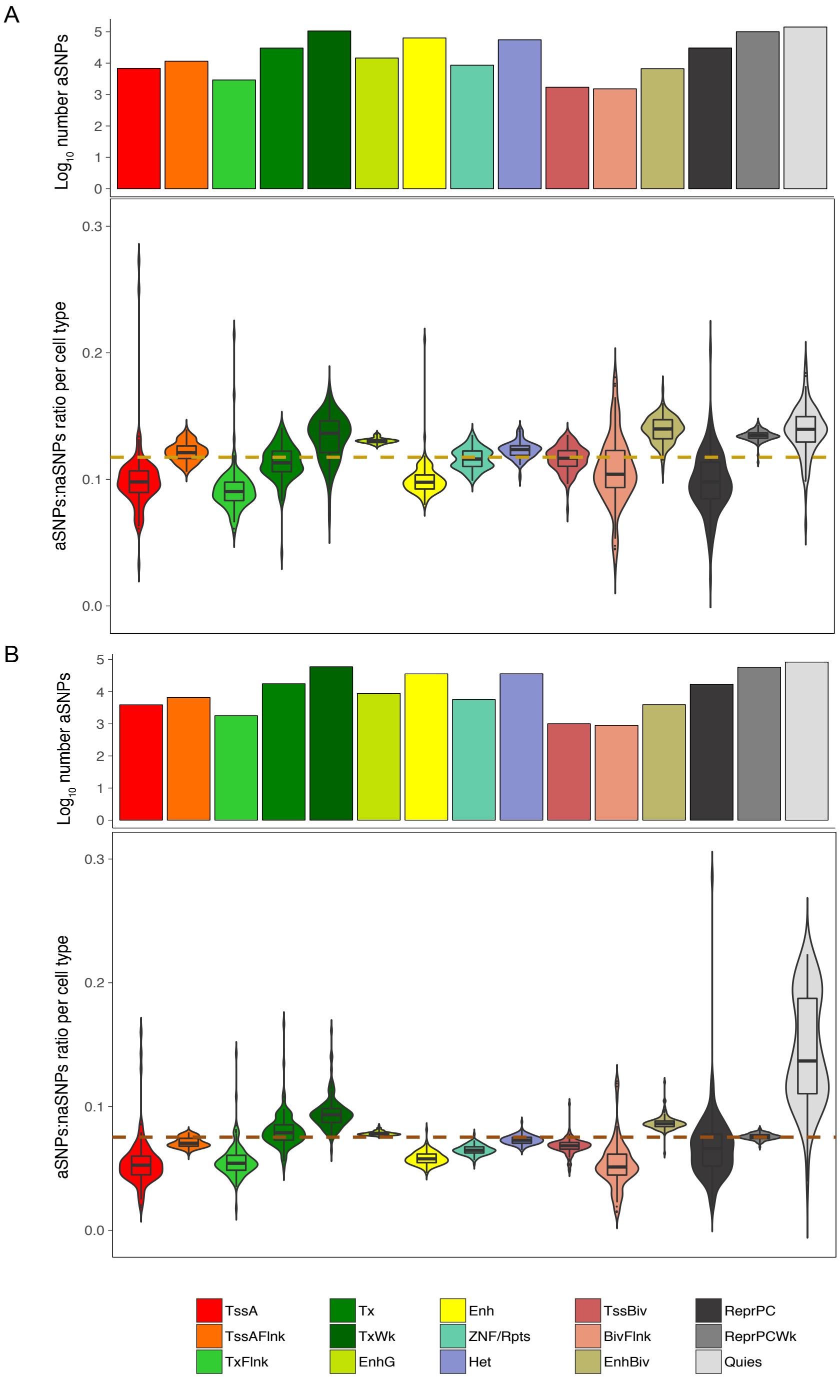

Fig S2. aSNPs distribution across Roadmap Epigenomics data. For A) Denisovan and B) Neanderthal aSNPs we show the distribution of the aSNPs:naSNPs ratio for every cell type within the 15 chromatin state. Dashed lines represent the mean of the aSNPs:naSNPs ratios calculated across all 15 chromatin states and 111 cell types for the given archaic ancestry. Top bars report the $\log_{10}$ transformed raw number of aSNPs annotated within each chromatin state across all 111 cell types.

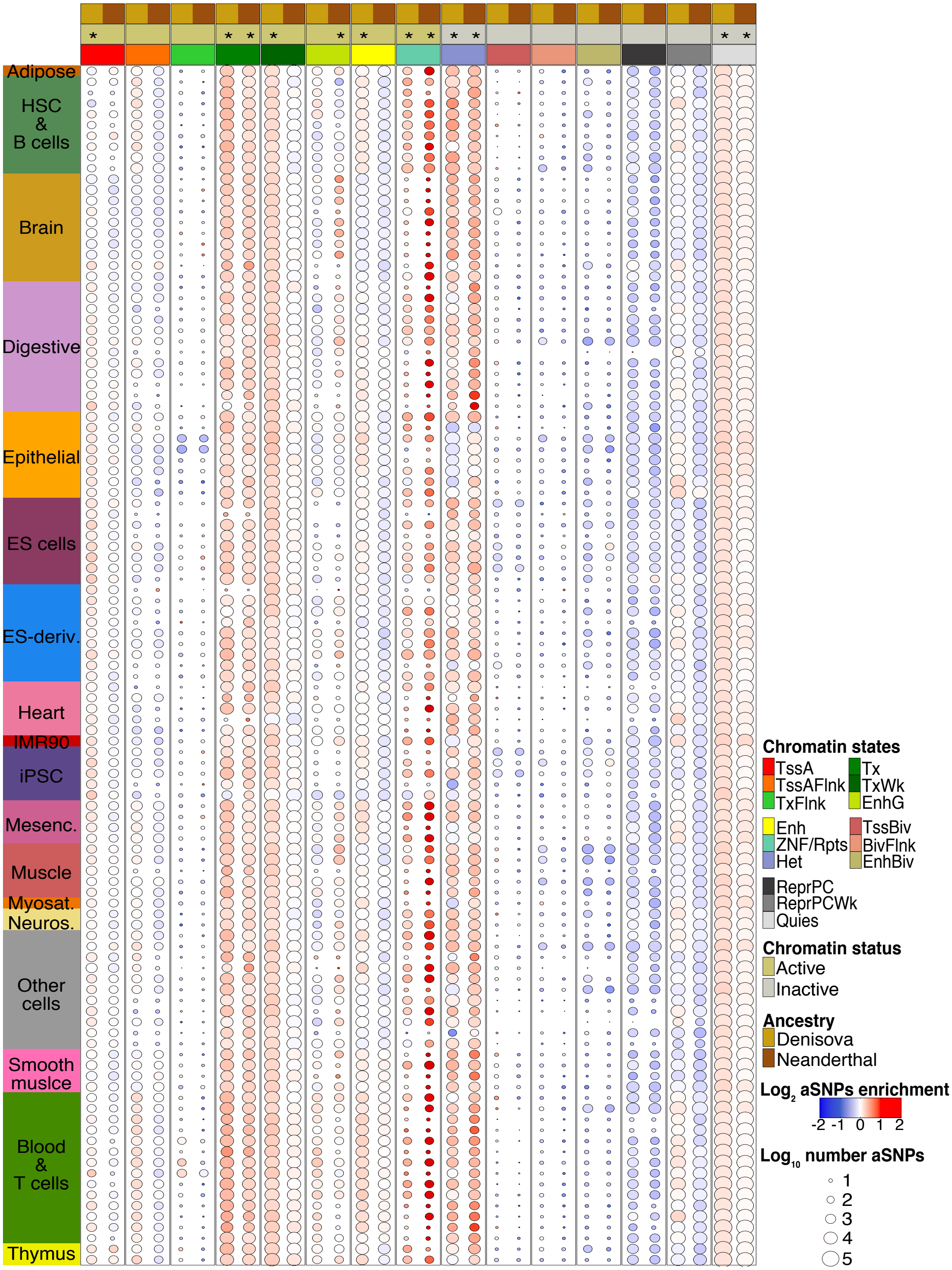

Fig S3. All-frequency aSNPs enrichment across Roadmap Epigenomics data. For all variants we show the log$_2$-transformed aSNP enrichment values for every chromatin state cell type combination computed over the mean of the aSNP:naSNP across all chromatin states and cell types (see main text). Top column annotations report (from top to bottom): ancestry information (i.e., Denisovan, Neanderthal); chromatin state functional activity (i.e., active = khaki, inactive = grey) as detailed in [26]; the 15 different chromatin states. Right column annotation shows the 18 different tissues. Colours within the heatmap cells represent the log$_2$-transformed aSNP enrichment scores. Circle diameters represent the number of Denisovan or Neanderthal aSNPs annotated within any particular cell type/chromatin state combination. Asterisks indicate whether the log$_2$-transformed aSNP enrichment was significantly higher than zero across all 111 cell types (FDR adjusted one-sample t-test p value: * = < 0.01, for detailed results, see Supp Table 2-Supp Table 3.

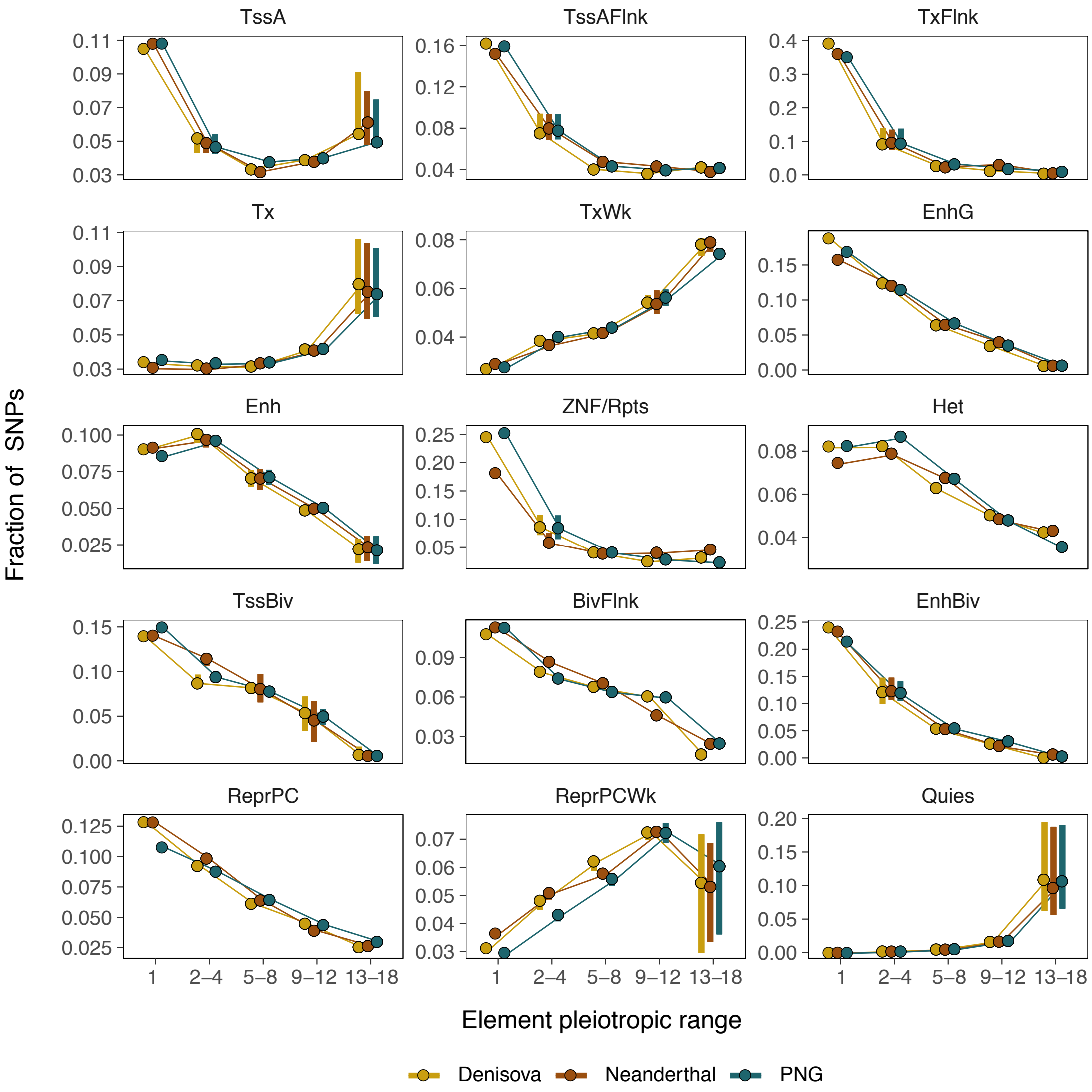

Fig S4. Pleiotropic activities of functional elements carrying all SNPs. For each chromatin state, we show the fraction of all SNPs falling within functional elements whose pleiotropic activities across the 18 tissues have been grouped into 5 different ranges. Dots and bars respectively represent the median and the IQR of the distribution of the proportion of SNPs within the given pleiotropic range. For further details see main text and Supp Table 4.

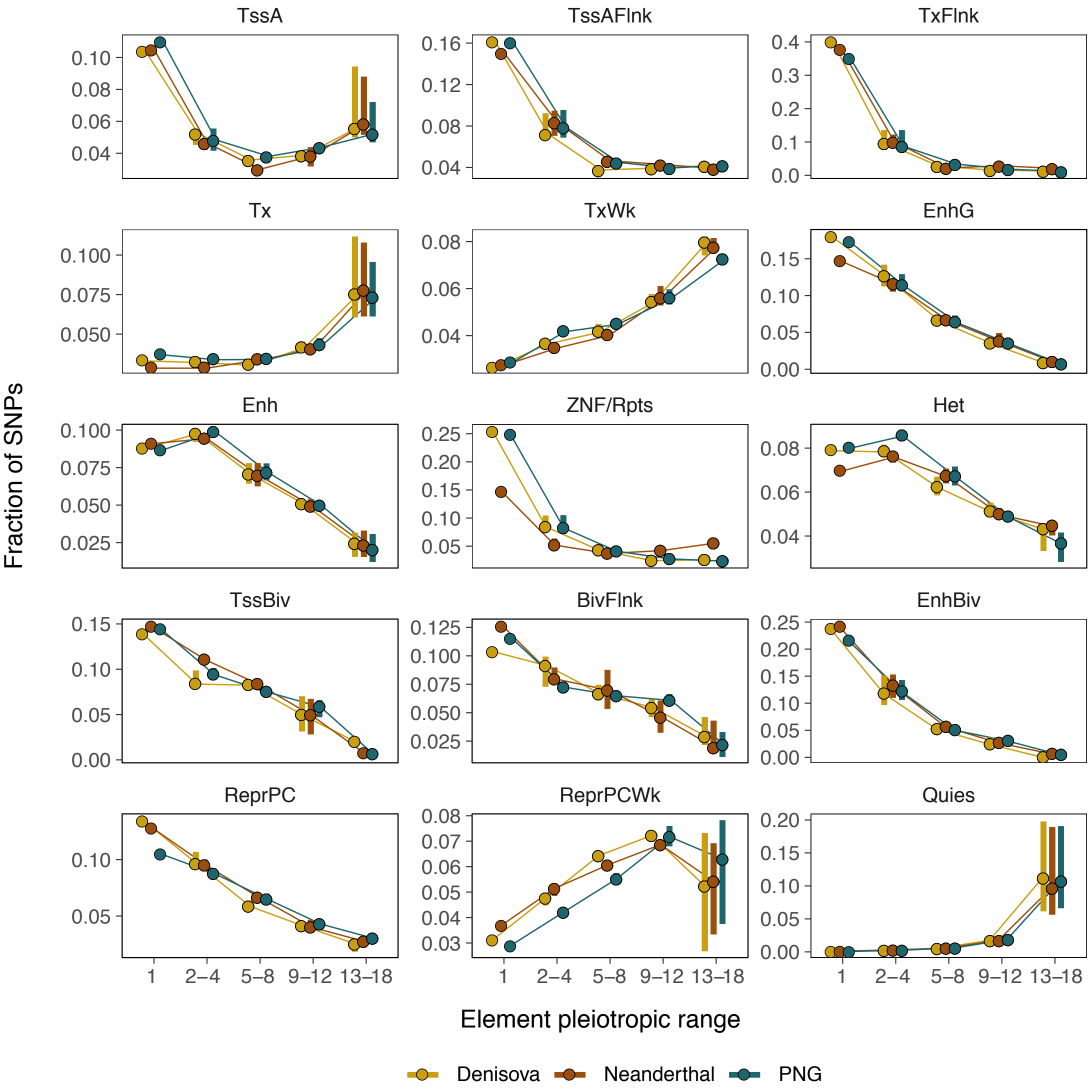

Fig S5. Pleiotropic activities of elements carrying common-to-high-frequency SNPs. For each chromatin state, we show the fraction of only common-to-high-frequency SNPs falling within functional elements whose pleiotropic activities across the 18 tissues have been grouped into 5 ranges. Dots and bars respectively represent the median and the IQR of the distribution of the proportion of SNPs within the given pleiotropic range. For further details see main text and Supp Table 4.

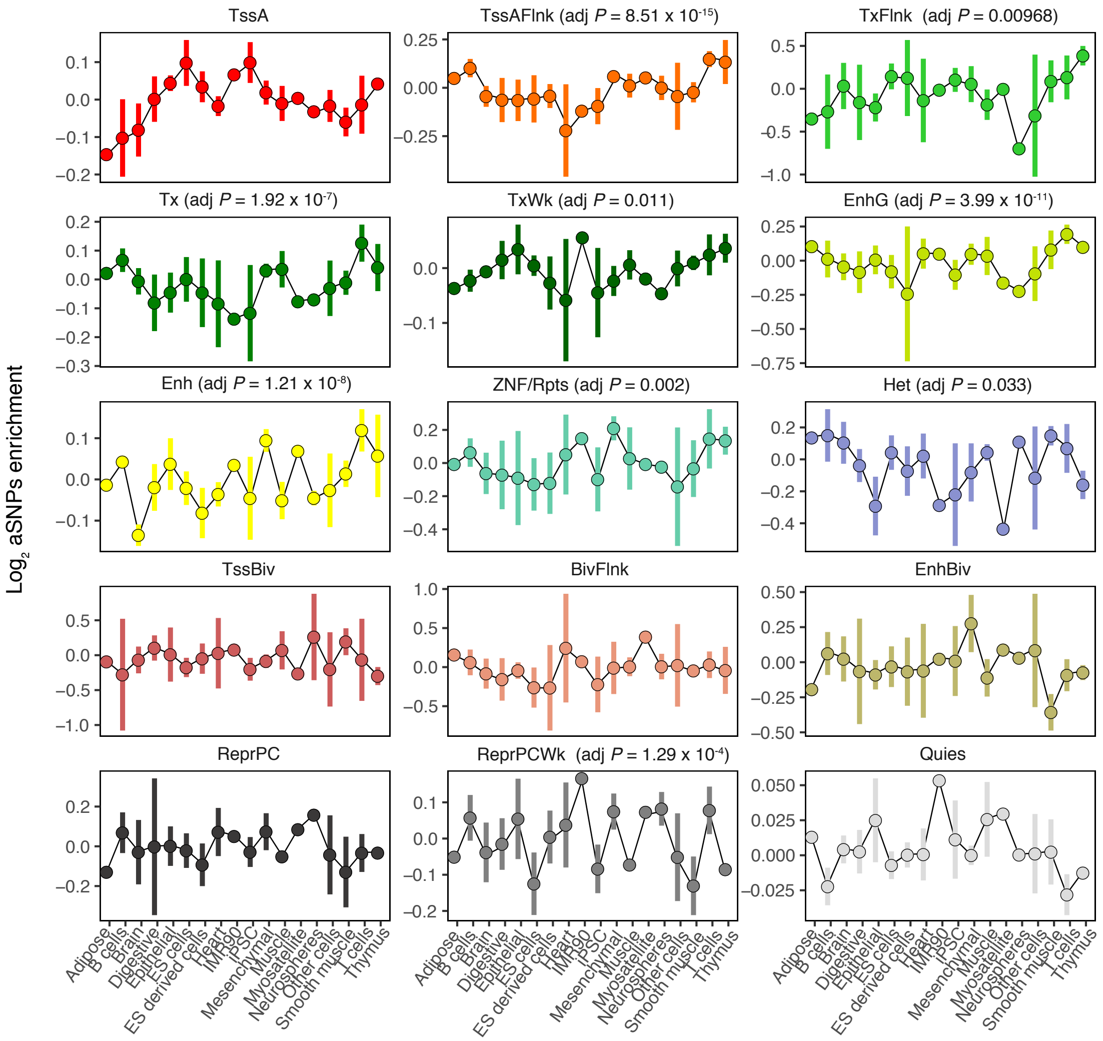

Fig S6. Denisovan aSNPs enrichment across multiple tissues. For each chromatin states we show the log$_2$-transformed aSNPs enrichment for all Denisovan variants across all cell types grouped into the 18 different tissues. Dots and cross bars respectively represent the mean and the standard deviation of the log$_2$-transformed aSNPs enrichment calculated across the sorted cell types. Numbers within parentheses indicate whether the log$_2$-transformed aSNPs enrichment was significantly higher within immune T cells compared to all the other tissues (FDR-adjusted one-tailed t-test p values). For further details, see main text and Supp Table 5.

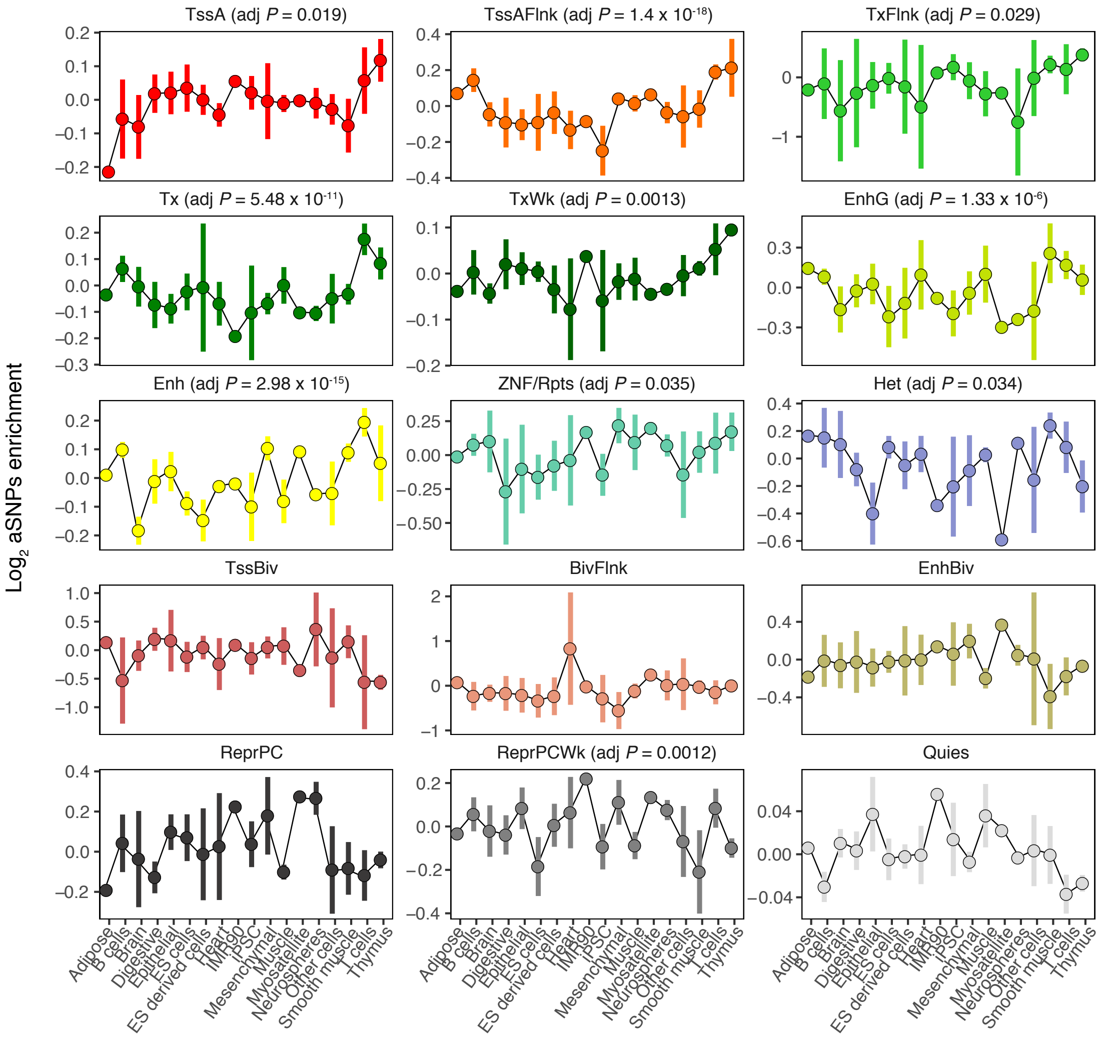

Fig S7. Denisovan common-to-high-frequency aSNPs enrichments across 18 cell lines. For each chromatin states we show the $\log_2$-transformed aSNPs enrichment only for common-to-high frequency Denisovan variants across all cell types grouped into the 18 different tissues. Dots and cross bars respectively represent the mean and the standard deviation of the $\log_2$-transformed aSNPs enrichment calculated across the sorted cell types. Numbers within parentheses indicate whether the $\log_2$-transformed aSNPs enrichment was significantly higher within immune T cells compared to all the other tissues (FDR-adjusted one-tailed t-test p values). For further details, see main text and Supp Table 5.

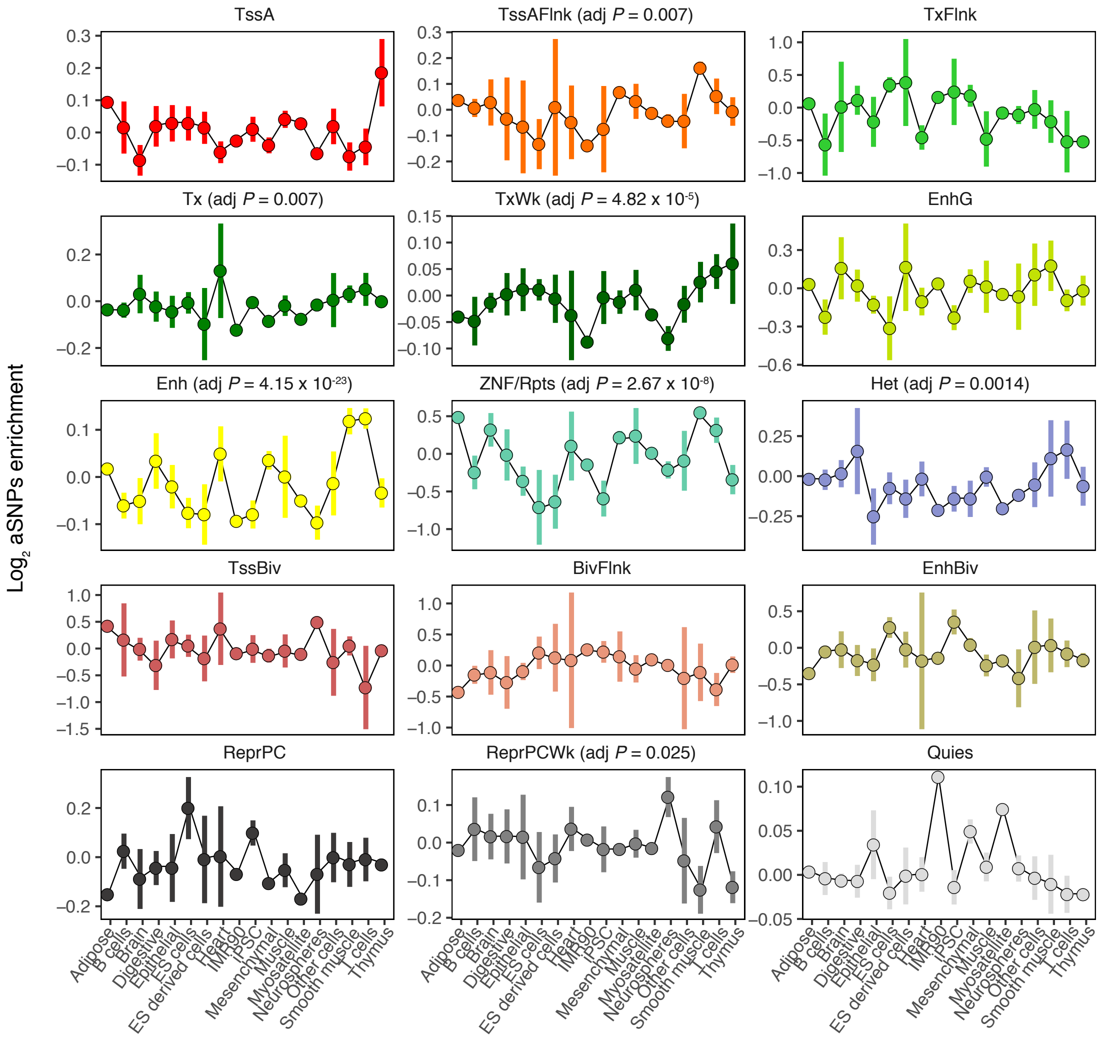

Fig S8. Neanderthal aSNPs enrichments across 18 cell lines. For each chromatin states we show the $\log_2$-transformed aSNPs enrichment for all Neanderthal variants across all cell types grouped into the 18 different tissues. Dots and cross bars respectively represent the mean and the standard deviation of the $\log_2$-transformed aSNPs enrichment calculated across the sorted cell types. Numbers within parentheses indicate whether the $\log_2$-transformed aSNPs enrichment was significantly higher within immune T cells compared to all the other tissues (FDR-adjusted one-tailed t-test p values). For further details, see main text and Supp Table 5.

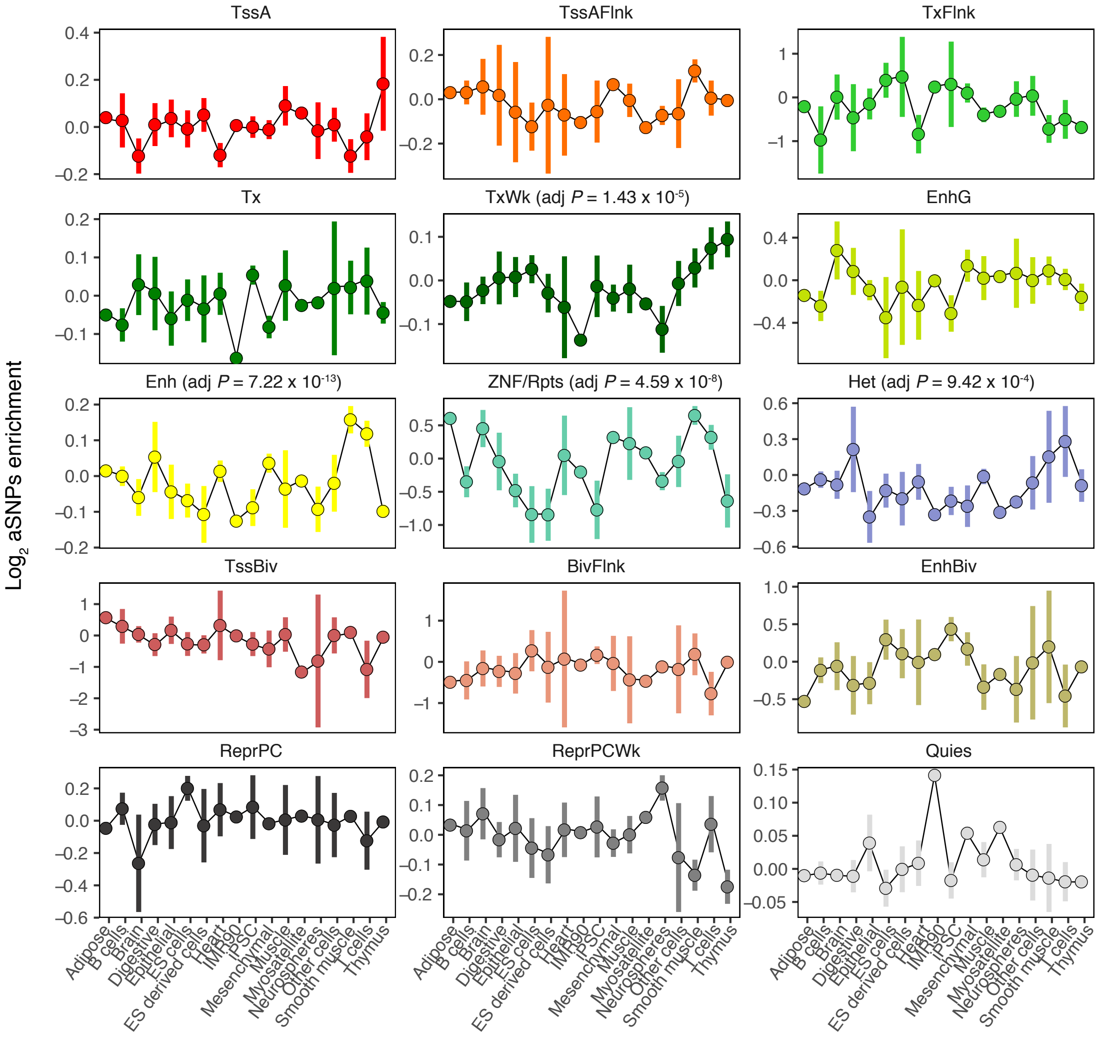

Fig S9. Neanderthal common-to-high-frequency aSNPs enrichments across 18 cell lines. For each chromatin states we show the $\log_2$-transformed aSNPs enrichment only for common-to-high frequency Neanderthal variants across all cell types grouped into the 18 different tissues. Dots and cross bars respectively represent the mean and the standard deviation of the $\log_2$-transformed aSNPs enrichment calculated across the sorted cell types. Numbers within parentheses indicate whether the $\log_2$-transformed aSNPs enrichment was significantly higher within immune T cells compared to all the other tissues (FDR-adjusted one-tailed t-test p values). For further details, see main text and Supp Table 5.

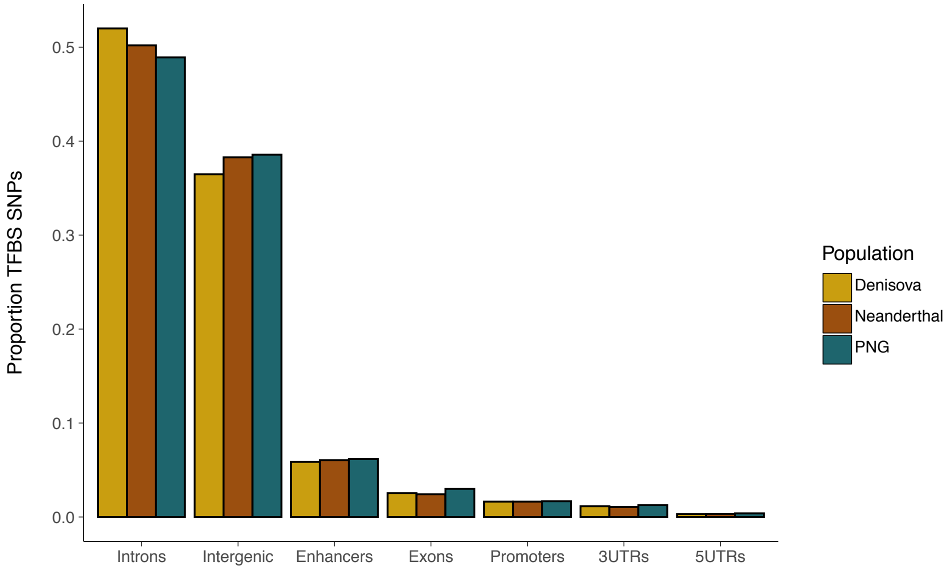

Fig S10. Genomic location of all TFBS-disrupting SNPs. For each ancestry we show the proportion of the full set of TFBS-disrupting SNPs annotated across multiple genomic elements, using the annotatr R package [57].

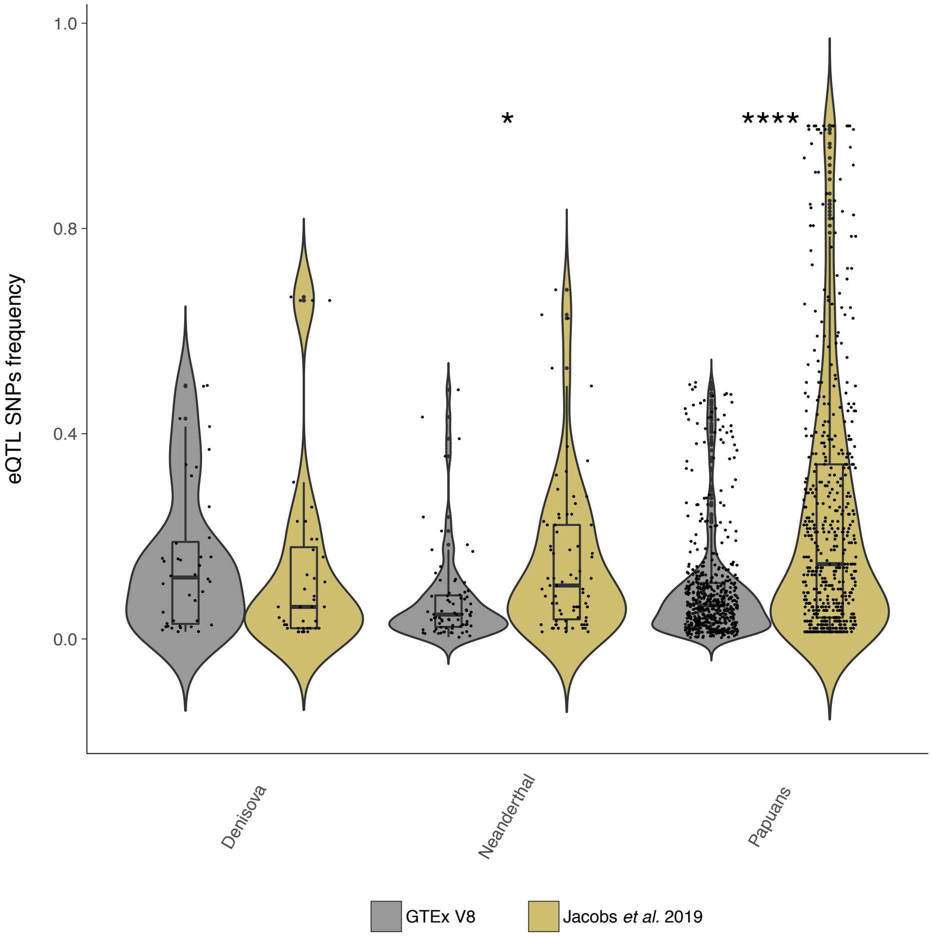

Fig S11. cis-eQTLs TFBS-disrupting SNPs. For each TFBS-disrupting allele identified as potential cis-eQTL by GTEx we show the comparison between the GTEx reported maf and the MIAF/DAF for the same alternative allele within New Guinea. Violin plots represent the full distribution of the allele frequencies. Lower and upper hinges of the boxplots correspond to the first and third quartiles of the distribution, whiskers extend to a maximum of 1.5 × IQR beyond the box. Asterisks indicate whether the allelic frequencies within New Guinea are significantly higher than that of the GTEx sample (FDR-adjusted paired Wilcoxon rank-sum test: **** = p value ≤ 0.0001; ***= p value ≤ 0.001).

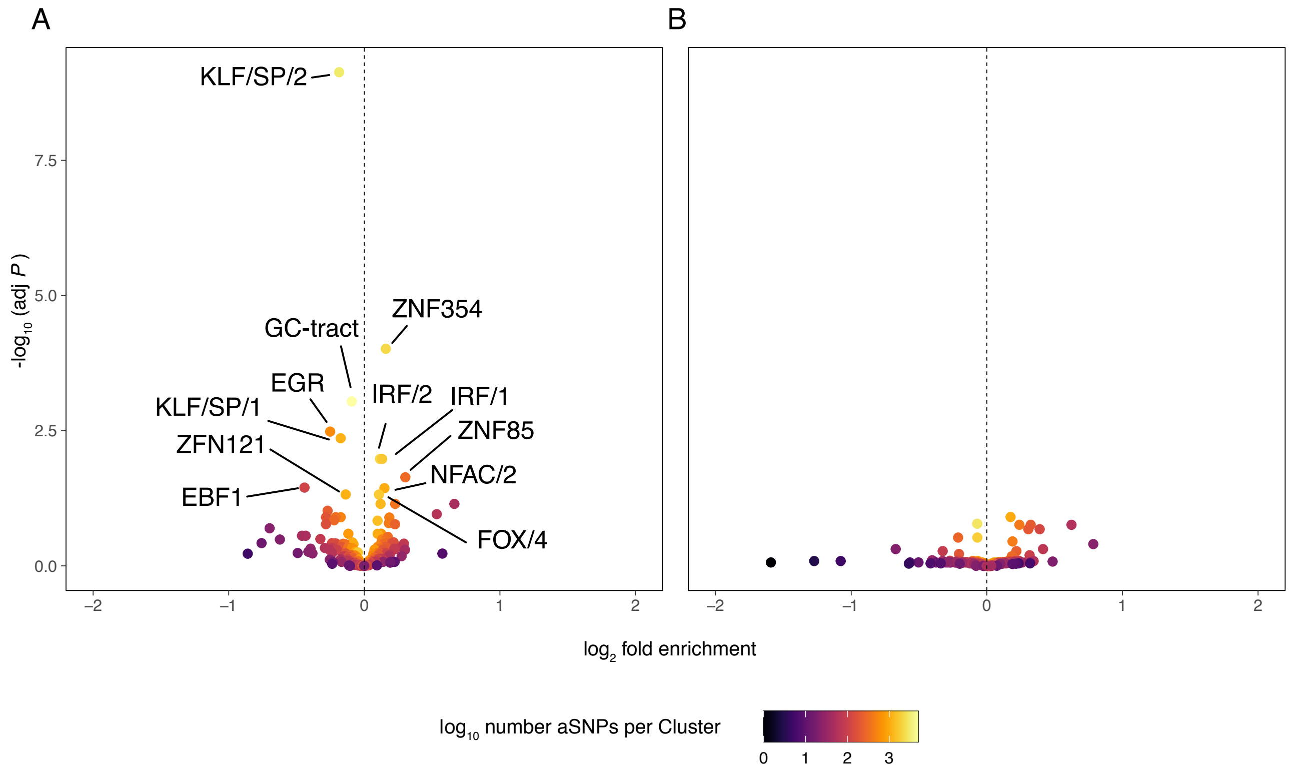

Fig S12. aSNPs enrichment for transcription factor clusters. For each TF cluster we show the $\log_2$-transformed aSNPs enrichment for all A) Denisovan and B) Neanderthal TFBS-disrupting variants relative to naSNPs against the $-\log_{10}$ FDR-corrected p values resulting from Fisher's exact test. Dots are coloured based on the number of TFBS-disrupting aSNPs annotated within the given cluster. Only clusters with a significant excess of aSNPs are labelled (FDR-adjusted $p \leq 0.05$). For further details see main text and Supp Table 6

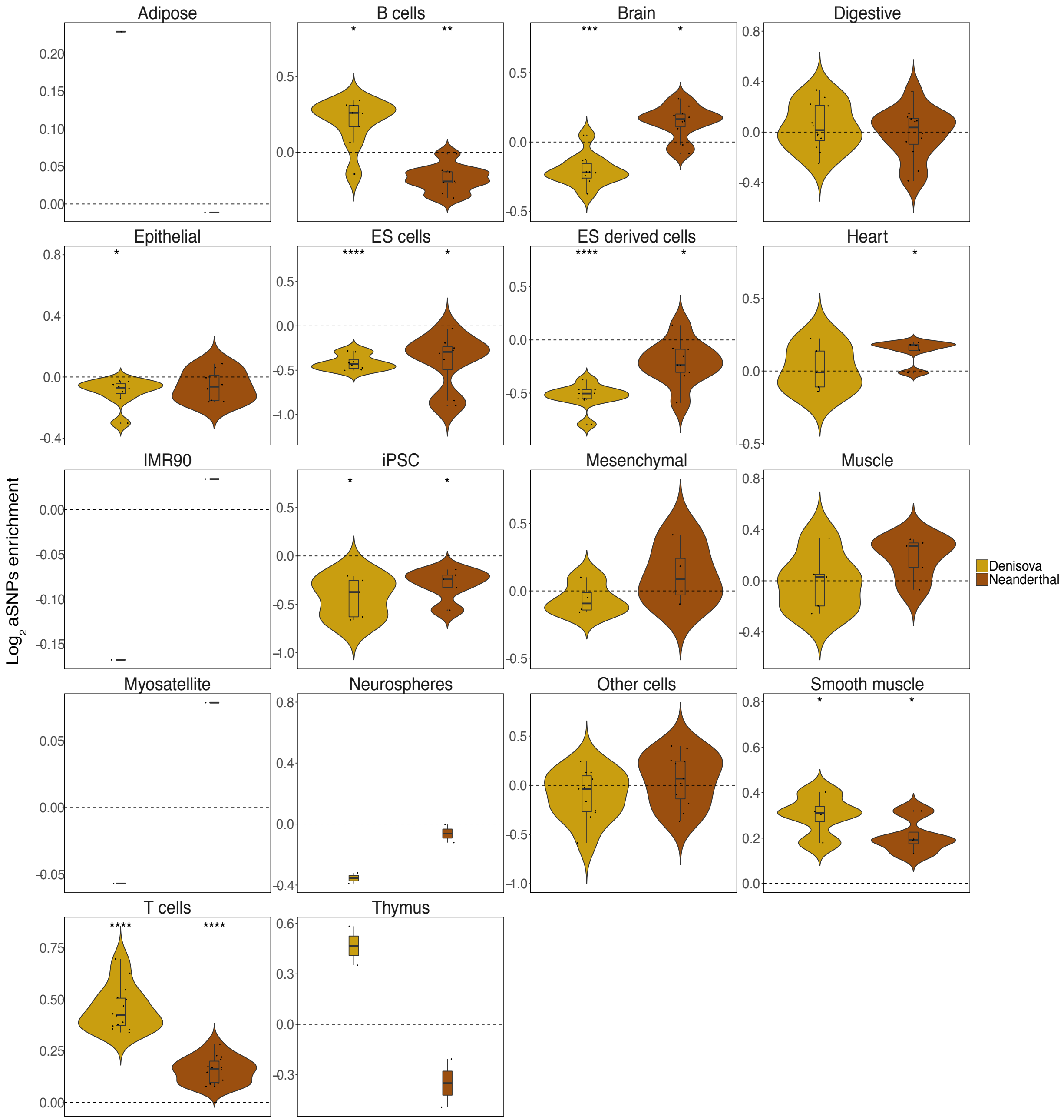

Fig S13. TFBS-disrupting aSNP enrichment across all tissues. For every cell lineage we show $\log_2$-transformed aSNPs enrichment only for TFBS-disrupting variants segregating at frequencies ≥ 0.3 within Papua New Guinea and predicted to cause a 'strong' disruption of the DNA motif as reported by motibreakR. Violin plots represent the full distribution of the aSNPs enrichment for the sorted cell types. Lower and upper hinges of the boxplots correspond to the first and third quartiles of the distribution, whiskers extend to a maximum of 1.5 × IQR beyond the box. Asterisks indicate whether, for the given cell lineage, the $\log_2$-transformed aSNPs enrichment was significantly different that the expected mean value calculated across all tissues for the given ancestry (FDR-adjusted one-sample t-test: $** = p$ value ≤ 0.01; $* = p$ value ≤ 0.05. For further details see Supp Table 7.

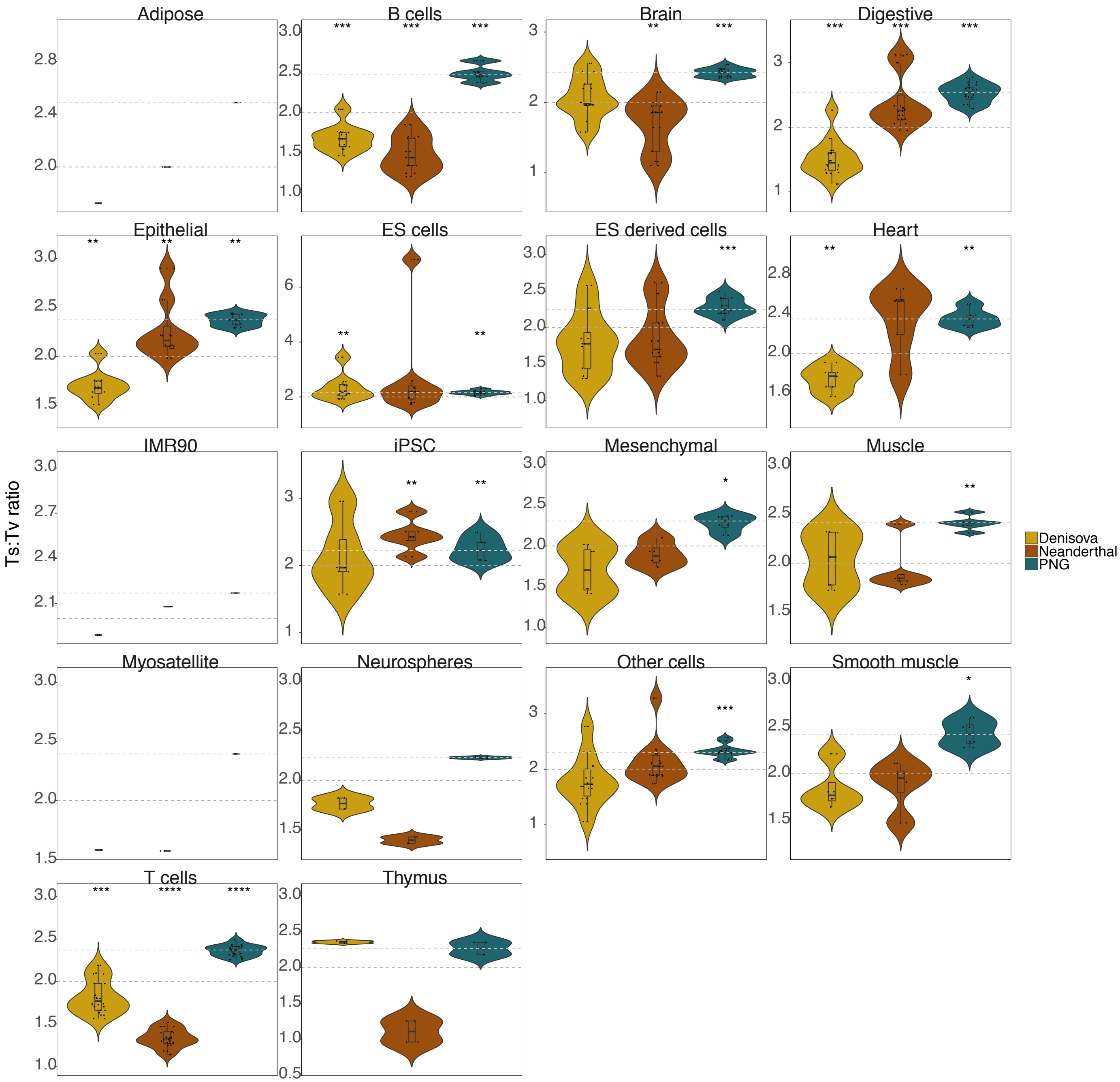

Fig S14. Ts:Tv ratio across all tissues for TFBS-dirsupting SNPs. For every cell lineage we show the Ts:Tv ratios only for TFBS-disrupting variants segregating at frequencies ≥ 0.3 within Papua New Guinea and predicted to cause a 'strong' disruption of the DNA motif as reported by motibreakR. Dashed lines represent: the expected 2:1 ratio (dark grey) and the aSNPs median ratio value calculated across all sorted cell types within the given cell lineage (light grey). Violin plots represent the full distribution of the Ts:Tv ratios for the sorted cell types. Lower and upper hinges of the boxplots correspond to the first and third quartiles of the distribution, whiskers extend to a maximum of 1.5 × IQR beyond the box. Asterisks indicate whether the distribution of ratios is significantly lower than 2 (FDR-adjusted one-tailed Wilcoxon rank-sum test: **** = p value ≤ 0.0001; *** = p value ≤ 0.001; ** = p value ≤ 0.01; ,* = $p$value ≤ 0.05. For further details see Supp Table 8.