nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-02053-5

Resurrecting the alternative splicing landscape of archaic hominins using machine learning

Received: 5 August 2022

Accepted: 29 March 2023

Published online: 04 May 2023

Check for updates

Colin M. Brand [1,2], Laura L. Colbran [3] & John A. Capra [1,2]

Alternative splicing contributes to adaptation and divergence in many species. However, it has not been possible to directly compare splicing between modern and archaic hominins. Here, we unmask the recent evolution of this previously unobservable regulatory mechanism by applying SpliceAI, a machine-learning algorithm that identifies splice-altering variants (SAVs), to high-coverage genomes from three Neanderthals and a Denisovan. We discover 5,950 putative archaic SAVs, of which 2,186 are archaic-specific and 3,607 also occur in modern humans via introgression (244) or shared ancestry (3,520). Archaic-specific SAVs are enriched in genes that contribute to traits potentially relevant to hominin phenotypic divergence, such as the epidermis, respiration and spinal rigidity. Compared to shared SAVs, archaic-specific SAVs occur in sites under weaker selection and are more common in genes with tissue-specific expression. Further underscoring the importance of negative selection on SAVs, Neanderthal lineages with low effective population sizes are enriched for SAVs compared to Denisovan and shared SAVs. Finally, we find that nearly all introgressed SAVs in humans were shared across the three Neanderthals, suggesting that older SAVs were more tolerated in human genomes. Our results reveal the splicing landscape of archaic hominins and identify potential contributions of splicing to phenotypic differences among hominins.

While the palaeontological and archaeological records provide evidence about some phenotypes of extinct hominins, most ancient tissues have not survived to the present. The discovery and successful sequencing of DNA genome-wide from a Denisovan[1] and multiple Neanderthal genomes[2–4] enabled direct comparisons of the genotypes of these archaic hominins to one another and to anatomically modern humans. These data also enable the potential for indirect phenotypic comparisons by predicting archaic phenotypes from their genomes[5]. Diverse molecular mechanisms collectively shape the similarities and differences between archaic hominins and modern humans. Given that the biology linking genotype to organism-level phenotype is complex and that the mapping may not generalize across human populations[6], predicting 'low-level' molecular phenotypes from genetic information is a promising alternative. Recent work has successfully explored such differences in protein-coding sequence[7] and differences relevant to gene expression, such as divergent gene regulation[8], differential methylation[9] and divergent three-dimensional genome contacts[10].

Variation in gene splicing could also underlie phenotypic differences between archaic hominins and modern humans but archaic splicing patterns have not been comprehensively quantified. Alternative splicing enables the production of multiple protein isoforms from a single gene[11–13]. The resulting proteomic diversity is essential for many processes,

[1]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA. [2]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA. [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. e-mail: tony@capralab.org

https://doi.org/10.1038/s41559-023-02053-5

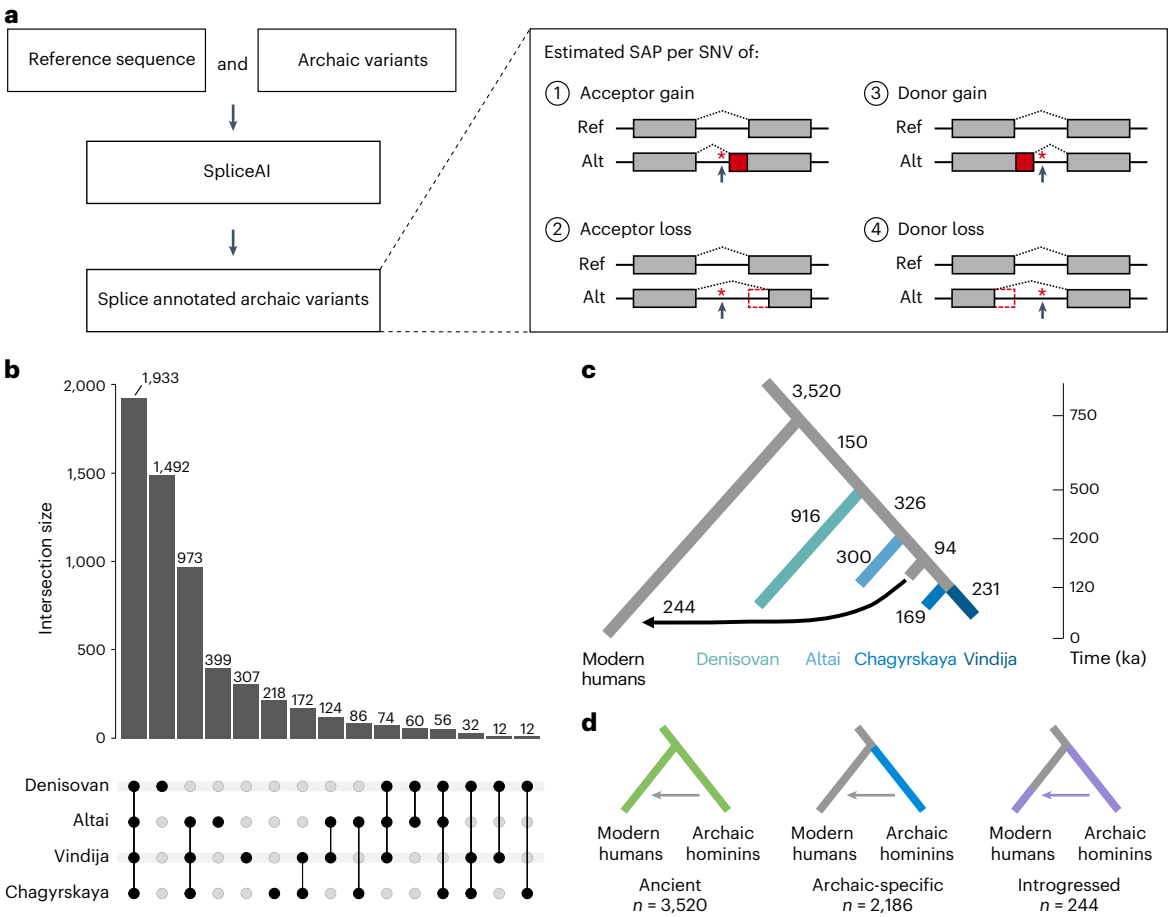


Fig. 1 | The identification, distribution and origin of archaic SAVs. a, We used SpliceAI to identify putative SAVs in archaic hominin genomes. We analysed autosomal single nucleotide variants (SNVs) from four archaic hominins compared to the reference sequence (hg19/GRCh37). SpliceAI annotates each variant with splice altering probabilities (SAPs) (Δ scores) and position changes for each class of splicing alteration: (1) acceptor gain, (2) acceptor loss, (3) donor gain or (4) donor loss. Here, we visualize one example consequence per splicing class alteration. See Extended Data Fig. 7 for all possibilities. Red asterisks and arrows indicate the variant position. Filled and dashed red boxes indicate sequence gained and lost in the predicted transcript, respectively. Ref, reference sequence; Alt, reference sequence with alternate allele. **b**, The distribution of the presence of archaic SAVs across archaic individuals. The dot matrix indicates the number of SAVs per lineage(s). **c**, The evolutionary origins of the archaic SAVs. From the distribution of SAVs across archaic and modern individuals, we inferred their origins using parsimony. We also identified introgressed archaic SAVs using two Neanderthal ancestry sets: ref. 47 (shown here) and ref. 48. The divergence times and placement of the introgression arrow reflect estimates from ref. 4 and ref. 46. The tip of the archaic hominin branches end at the estimated age of the fossil that yielded the ancient genome. We display data using $\Delta \geq 0.2$ here and these patterns are maintained when $\Delta \geq 0.5$. **d**, We consider three main categories of archaic SAVs based on their origin and presence across populations: 'ancient', archaic SAVs present in both modern and archaic hominin individuals and inferred to have origins before the last common ancestor of these groups; 'archaic-specific', archaic SAVs that are present in archaic hominins but absent or present at low frequency (allele frequency <0.0001) in modern humans; and 'introgressed', archaic SAVs that were introgressed into Eurasian populations due to archaic admixture.

including development and establishing tissue identity[14]. Defects in splicing underlie many human diseases (for example, refs. 15–22) and variation in splicing contributes to differences in traits in non-human species (see Table 1 in ref. 23). Further, alternative splicing can evolve rapidly and respond to environmental factors—suggesting that it often contributes to adaptation[23–25] and species differences[26–28].

Splicing patterns are directly influenced by the nucleotide sequences surrounding splice sites[29]. This has enabled the development of many algorithms to predict alternative splicing from RNA sequencing (RNA-seq)[30–32] or DNA sequence[33–37]. Beyond human clinical applications, methods that require only DNA sequence can be leveraged to understand alternative splicing in extant species for which acquiring RNA-seq data may be difficult to impossible, or in extinct taxa, such as archaic hominins.

Here, we resurrect the genome-wide alternative splicing landscape of archaic hominins using SpliceAI, an algorithm that predicts splicing patterns from sequence alone[35]. First, we assess the distribution of splice-altering variants (SAVs) among all four archaic individuals, identify which genes are affected and describe how the transcripts are modified. Second, we quantify which SAVs are also present in modern humans due to shared ancestry or introgression. Third, we quantify SAV enrichment among gene sets that underlie modern human phenotypes. Fourth, we estimate the effects of SAVs on the resulting transcript or protein. Fifth, we explore how selection shaped alternative splicing in archaics. Sixth, we evaluate the expression and function of archaic SAVs that are also present in modern humans. Finally, we highlight a handful of archaic SAVs with potential evolutionary significance.

Results

We examined the alternative splicing landscape in archaic hominins using all four currently available high-coverage archaic genomes, representing three Neanderthals[2–4] and a Denisovan[1]. We applied the SpliceAI classifier to sites with high-quality genotype calls where at least one archaic individual exhibited at least one allele different from the human reference (hg19/GRCh37) using the built-in GENCODE, Human Release 24, annotations to identify variants in gene bodies (Fig. 1a). SpliceAI

estimates Δ, the splice-altering probability (SAP), for each variant of: (1) an acceptor gain, (2) an acceptor loss, (3) a donor gain and (4) a donor loss (Fig. 1a). SpliceAI also indicates the positions changed for each of these four effects in base pairs (bp).

Alternative splicing occurs across nearly all eukaryotes and its molecular mechanisms are deeply conserved[38]. We therefore anticipated that the sequence patterns learned by SpliceAI in humans would generalize to archaics. To confirm this, we searched the 147 genes associated with the major spliceosome complex[39] for 'archaic-specific' variants, that is, archaic variants absent or at very low allele frequency (<0.0001) from individuals sequenced by the 1000 Genomes Project (1KG)[40] and the Genome Aggregation Database (gnomAD)[41] (Supplementary Data 1). We annotated these variants using the Ensembl Variant Effect Predictor[42]. We found only two missense variants that were scored as likely to disrupt protein function by both PolyPhen and SIFT, neither of which were fixed in all four archaics (Supplementary Data 1). We observed a similar number of predicted deleterious variants in random sets of four diverse modern humans (0–3). Thus, there is near-complete conservation of the proteins involved in splicing between archaic hominins and modern humans.

Thousands of SAVs are present in archaic hominins

We identified 1,567,894 autosomal positions with ≥1 non-reference allele among the archaic individuals (Supplementary Table 1). Many of these positions fell within a single GENCODE annotation; however, a handful were present in multiple annotated products (Supplementary Table 2). An individual variant that overlaps multiple annotations may have differential splicing effects on the different transcripts. Hereafter, we define a 'variant' as one non-reference allele for a single annotated transcript at a given genomic position.

Among these variants, 1,049 had high SAP (SpliceAI $\Delta \geq 0.5$) out of 1,607,350 archaic variants we analysed. A total of 5,950 archaic variants had moderate SAP ($\Delta \geq 0.2$). Hereafter, we refer to these variants as high-confidence SAVs and SAVs, respectively; to maximize sensitivity, we focus on the SAVs in the main text.

The number of SAVs was similar among the four archaics, ranging from 3,482 (Chagyrskaya) to 3,705 (Altai) (Supplementary Table 3). These values fell within the range of SAVs observed in individual modern humans, estimated from one randomly sampled individual per 1KG population (Supplementary Table 4). SAVs were most commonly shared among all four archaic individuals (Fig. 1b and Supplementary Fig. 1). As expected from the known phylogeny, the Denisovan exhibited the most unique SAVs, followed by all Neanderthals and each individual Neanderthal (Fig. 1b and Supplementary Fig. 1).

A total of 4,242 genes have at least one archaic SAV. A total of 3,111 genes have only one SAV; however, 1,131 had multiple SAVs (Supplementary Table 5). Among the genes with the largest number of archaic SAVs are: *WWOX* ($n = 9$), which is involved in bone growth development[43], *HLA-DPA1* ($n = 7$) and *HLA-DPB1* ($n = 10$), essential components of the immune system[44]; and *CNTNAP2* ($n = 11$), which encodes a nervous system protein associated with neurodevelopmental disorders and is also one of the longest genes in the human genome[45].

Many SAVs (47.8%) have a high SAP for only one of the four classes of splicing change (acceptor gain, acceptor loss, donor gain and donor loss) (Supplementary Fig. 2) and, as expected, the overall association between the probabilities of different change types was negative ($\rho = -0.34$ to -0.14) for variants with at least one SAP >0 (Supplementary Fig. 3). Donor gains were the most frequent result of SAVs for both thresholds (29.5% and 35.1% of variant effects, respectively) (Supplementary Fig. 2). While this may reflect the true distribution, we cannot rule out that the classifier has greater power to recognize donor gains compared to acceptor gains, acceptor losses and donor losses.

Thirty-seven per cent of archaic SAVs are archaic-specific

We inferred the origin of archaic variants on the basis of parsimony. We identified 2,186 (37%) 'archaic-specific' SAVs. These archaic SAVs are absent among modern humans in 1KG and gnomAD or occur in gnomAD at a very low (<0.0001) allele frequency (Fig. 1c). Such low-frequency variants are likely to be recurrent mutations identical by state rather than identical by descent.

The remaining 63% of archaic SAVs are present in modern humans. Archaic hominins and modern humans last shared a common ancestor ~570–752 thousand years ago (ka) (ref. 46). SAVs present in both archaic and modern humans may be the result of introgression, shared ancestry or recurrent mutation. To identify SAVs present in 1KG due to introgression, we used two datasets on archaic introgression into modern humans[47,48]. While modern human genomes retain Denisovan and Neanderthal ancestry, most 1KG samples have minimal (<1%) Denisovan ancestry[47,48]. Therefore, we focused on Neanderthal introgression and classified 244 SAVs identified by ref. 47 in 239 genes (Fig. 1d and Supplementary Fig. 4) and 386 SAVs identified by ref. 48 in 361 genes as 'introgressed' (Supplementary Fig. 4). Despite only modest overlap between the two introgression datasets (Supplementary Fig. 5), we observed qualitatively similar results in downstream analyses. Hereafter, we present results using the ref. 47 introgressed variants in the main text and include results using the ref. 48 set in the Supplementary Information.

Non-introgressed variants present in both archaic and modern humans probably evolved before our most recent common ancestor. We refer to these SAVs as 'ancient' and we consider the archaic SAVs with an allele frequency ≥0.05 in at least two 1KG superpopulations 'high-confidence ancient'. This decreases the probability of recurrent mutation or misclassification of introgressed alleles. Hereafter, 'ancient' refers to these high-confidence ancient variants unless otherwise specified. We identified 2,252 such variants on the basis of ref. 47 among 1,896 genes (Fig. 1d and Supplementary Fig. 4) and 2,195 variants on the basis of ref. 48 among 1,856 genes (Supplementary Fig. 4).

Archaic-specific SAVs are enriched in genes with diverse phenotypes

To identify the potential phenotypic consequences of archaic-specific SAVs, we tested for enrichment of functional annotations among genes with archaic-specific SAVs. Following ref. 10, we considered links between genes and phenotypes from two sources: the 2019 GWAS Catalog[49] and the Human Phenotype Ontology (HPO)[50], that capture annotations based on common and rare diseases, respectively. Structural properties of genes, such as the number of exons, influence the probability that they have SAVs (Supplementary Table 6 and Supplementary Fig. 6). To account for these different probabilities, we generated a permutation-based empirical null distribution (Methods) and used it to estimate enrichment for each phenotype and control the false-discovery rate (FDR).

Given that we cannot directly observe archaic individuals, functions associated with genes with archaic-specific SAVs are of particular interest. We found enrichment for many phenotypes among the 1,907 genes with archaic-specific SAVs (Fig. 2 and Supplementary Data 2). Only two GWAS traits were significantly enriched for these SAVs: blood metabolite levels and blood metabolite ratios (Fig. 2a). There were substantially more phenotypes from HPO enriched among genes with archaic-specific SAVs (Fig. 2b) and these included traits that are known to differentiate archaic hominins and modern humans, including skeletal traits such as lumbar hyperlordosis and several cranial features (Supplementary Data 2). At least one significantly enriched trait occurred in every system across the HPO, except for the endocrine system.

Next, we sought to characterize similarities and differences among the archaic hominin individuals. We assessed phenotype enrichment among genes that contained shared, Neanderthal- and lineage-specific SAVs (Supplementary Data 2). We found minimal enrichment among the 106 genes with shared SAVs (Extended Data Fig. 1). However, there was considerable enrichment across various systems for Neanderthal- and lineage-specific SAVs (Extended Data Figs. 2–6). For example, all

https://doi.org/10.1038/s41559-023-02053-5

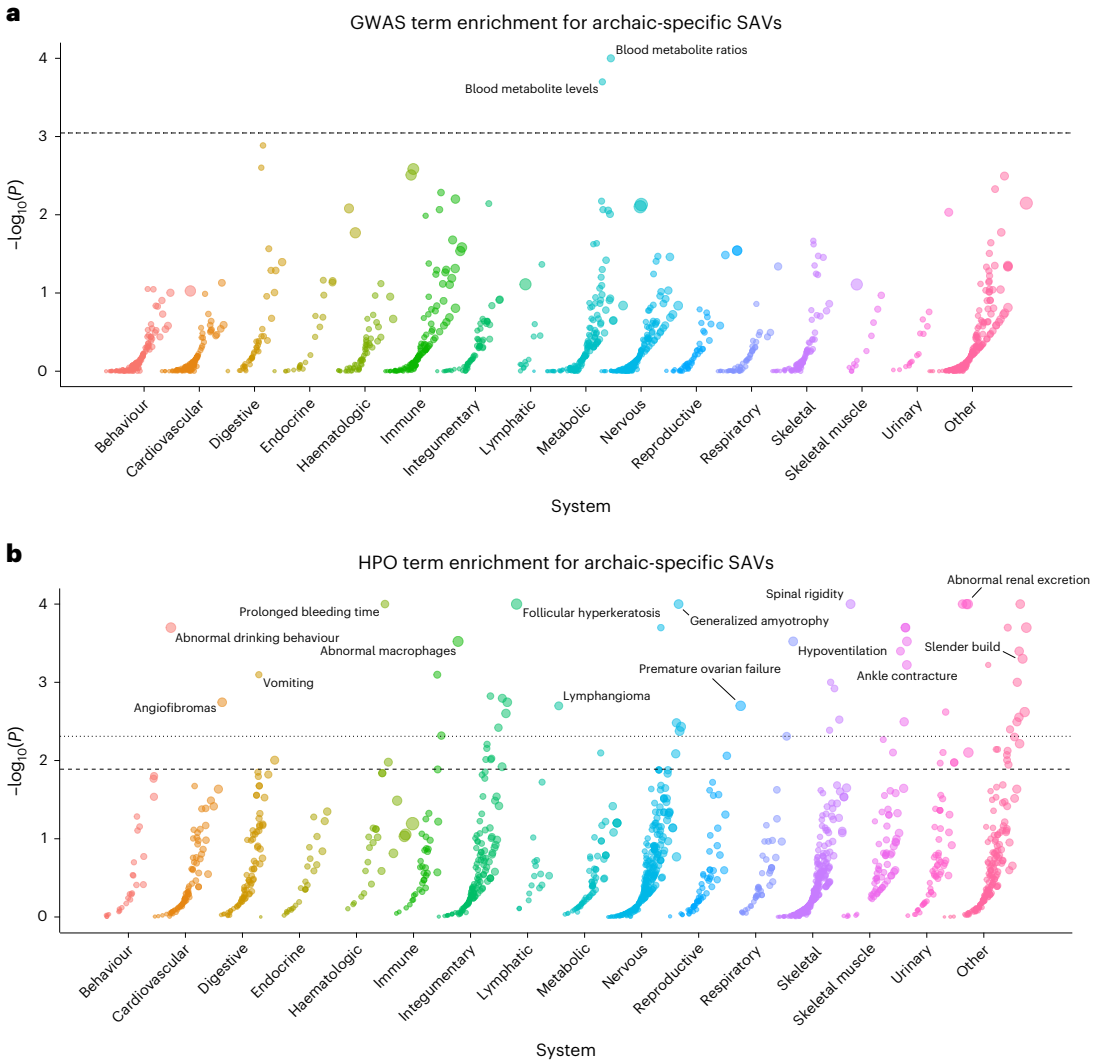


Fig. 2 | Genes with archaic-specific SAVs are enriched for roles in many phenotypes. a, Phenotype associations enriched among genes with archaic-specific SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and *P* values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent FDR-corrected *P* value thresholds at FDR = 0.05 and 0.1, respectively. One example phenotype with a *P* value less than or equal to the stricter FDR threshold (0.05) is annotated per system. **b**, Phenotypes enriched among genes with archaic-specific SAVs based on annotations from the HPO. Data were generated and visualized as in **a**. See Supplementary Data 2 for all phenotype enrichment results.

Neanderthals were enriched for SAVs in genes underlying skin conditions including abnormal skin blistering and fragile skin (Extended Data Fig. 5). The Denisovan exhibited enrichment for SAVs in genes associated with many skeletal and skeletal muscle system traits including skull defects, spinal rigidity, abnormal skeletal muscle fibre size, increased muscle fibre diameter variation and type 1 muscle fibre predominance (Extended Data Fig. 4). No traits were enriched in multiple different sets of lineage-specific SAVs at FDR-adjusted significance levels.

Most SAVs result in isoforms that trigger nonsense-mediated decay or yield altered transcripts and proteins

A SAV can result in a range of effects on the messenger RNA product, including having little to no impact. Therefore, the above analysis captures the extent of potential phenotypic consequences as inferred using gene ontologies. Next, we sought to characterize the possible functional effects of archaic SAVs on transcripts using an in silico approach.

We predicted the effect of each SAV on the resulting transcript by constructing a canonical transcript using the GENCODE exon annotations. Next, we generated a new transcript using the variant, indicated splicing alteration class (for example, acceptor gain) and Δ position for that alteration (Extended Data Fig. 7). If multiple alteration classes passed our SAP threshold, we modelled the class with the largest Δ. We compared the length and composition of the resulting transcripts and proteins for all but six SAVs with disagreements between the annotated transcript and genome sequences (Supplementary Data 3).

When considering the most likely effect per SAV, most (60%) SAVs result in a change to the transcript or protein sequence (Fig. 3a). Among these consequential SAVs, the most prevalent effect was a longer protein that included premature termination codons (PTCs) (Fig. 3a). Many such isoforms would trigger nonsense-mediated decay (NMD). The remaining SAVs resulted in altered transcripts or proteins that would not induce NMD but may yield a different or differentially stable protein.

When stratifying SAVs by allele origin, the proportion of these effects was generally similar for most classes (Fig. 3b). However, ancient SAVs more frequently resulted in no effect, whereas archaic-specific

https://doi.org/10.1038/s41559-023-02053-5

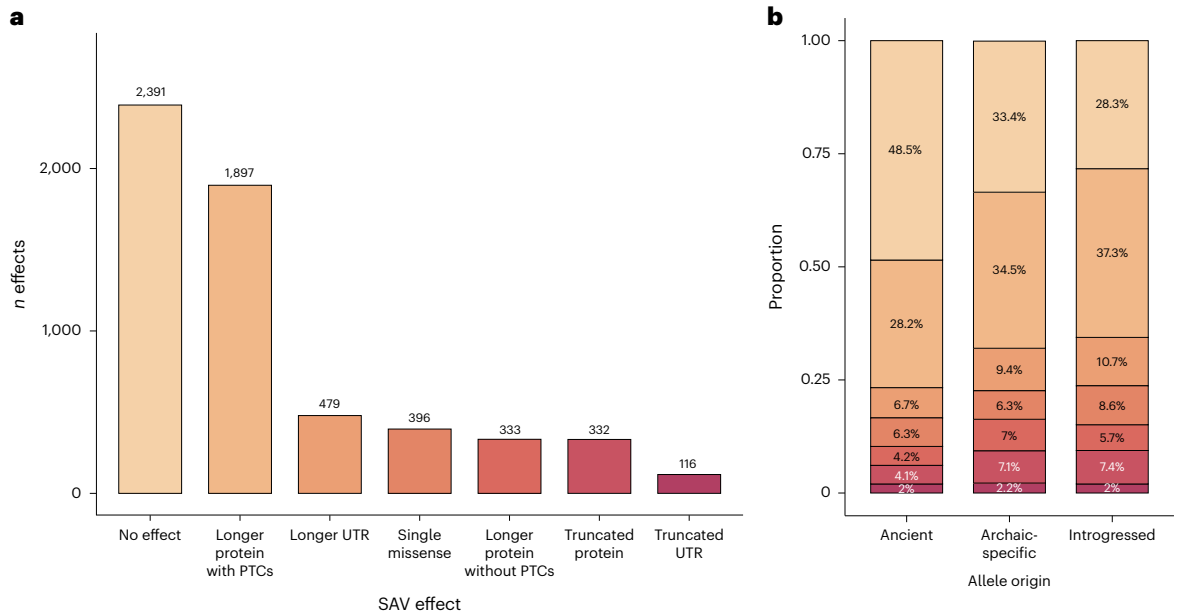


Fig. 3 | Most SAVs result in isoforms that trigger NMD or yield altered transcripts and proteins. a, The number of SAVs that result in one of seven effects based on the single, largest splicing effect per SAV. We excluded six SAVs for which the genomic and transcriptomic annotations did not match. PTCs, premature termination codons; UTRs, 5′or 3′untranslated regions. **b**, The number of protein effects stratified by allele origin[47]. Colours indicate the transcript or protein effect as in **a**.

and introgressed SAVs were more likely to alter the canonical protein or untranslated regions (UTRs) (G test of independence, $G = 138$, $P = 1.73 \times 10^{-23}$). This pattern was also observed when variant allele origin was classified as per ref. 48 ($G = 121$, $P = 3.45 \times 10^{-20}$) (Supplementary Table 7). We also note that many SAVs are predicted to result in multiple splicing alteration classes (for example, a donor gain and a donor loss) and thus, by focusing on a single class per SAV, we may miss some biologically relevant effects.

Site-level evolutionary conservation varies across SAV origin

Genes vary in their tolerance to mutation and SAVs often disrupt gene function and contribute to disease[20–22]. To evaluate if the presence of archaic SAVs is associated with evolutionary constraint on genes, we compared the per gene tolerance to missense and loss-of-function variants from gnomAD[41] among ancient, archaic-specific, introgressed and non-splice altered genes. In addition to constraint at the gene level, evolutionary constraint can be quantified at nucleotide level by methods like phyloP that quantify deviations from the expected neutral substitution rate at the site-level between species[51]. Thus, to explore the constraint on SAV sites themselves, we also compared their phyloP scores.

While we found a significant difference in the observed/expected number of missense variants per transcript among genes with different classes of SAV (Kruskal–Wallis, $H = 18.079$, $P = 0.0004$), the effect size was minimal (Supplementary Fig. 7a). Furthermore, genes with SAVs of different origins did not significantly differ in the observed/expected number of loss-of-function variants per transcript (Kruskal–Wallis, $H = 1.533$, $P = 0.675$) (Supplementary Fig. 7b). Variants classified as per ref. 48 exhibited the same pattern (Supplementary Fig. 7c,d). These results suggest that genes with alternative splicing in archaics are similar in their gene-level constraint to other genes.

In contrast, phyloP scores were significantly different between ancient SAVs, archaic-specific SAVs, introgressed SAVs and non-SAVs (Kruskal–Wallis, $H = 877.429$, $P = 6.963 \times 10^{-190}$) (Supplementary Fig. 8a). All of the variant sets exhibited a wide range of phyloP scores, indicating diverse pressures on SAVs of each type. However, ancient SAVs and non-SAVs exhibited largely negative phyloP scores, suggesting substitution rates faster than expected under neutral evolution. In contrast, archaic-specific and introgressed SAVs had higher median phyloP scores, suggesting that more of these loci experienced negative constraint. However, 84.3% occurred within the range consistent with neutral evolution ($|\text{phyloP}| \leq 1.3$). Variants classified as per ref. 48 exhibited similar patterns (Supplementary Fig. 8b); however, archaic-specific, rather than introgressed, variants had a larger mean phyloP score.

The prevalence of SAVs across lineages is consistent with purifying selection on most SAVs

Variants that disrupt splicing and/or produce new isoforms are expected to be under strong negative selection[52,53]. However, given differences in ages of archaic SAVs and the effective population sizes (N_e) of the lineages in which they arose, different SAVs were probably exposed to different strengths of selection for different periods. Thus, we proposed that the probability a given SAV would survive to the present would vary on the basis of its origin. For example, SAVs that arose in the ancestor of all archaic lineages were probably subject to purifying selection over a longer time scale than lineage-specific SAVs, especially those that arose in lineages with low N_e.

Shared archaic variants are depleted for SAVs compared to lineage-specific variants and this depletion increased with higher SAP thresholds (Fig. 4a,b and Supplementary Table 8). This result is consistent with the idea that most SAVs are deleterious and that the longer exposure to negative selection for older variants results in a smaller fraction of remaining SAVs. It is also concordant with the site-level constraint results.

Given that the population histories for each archaic lineage were probably different, we also compared within lineages. Neanderthals are thought to have lived in smaller groups and exhibited a lower N_e than Denisovans[4]. We tested this by repeating the SAV enrichment test for variants specific to each individual archaic lineage (Fig. 4a). All three Neanderthals were significantly enriched for unique SAVs compared to shared archaic variants after Bonferroni correction (odds

https://doi.org/10.1038/s41559-023-02053-5

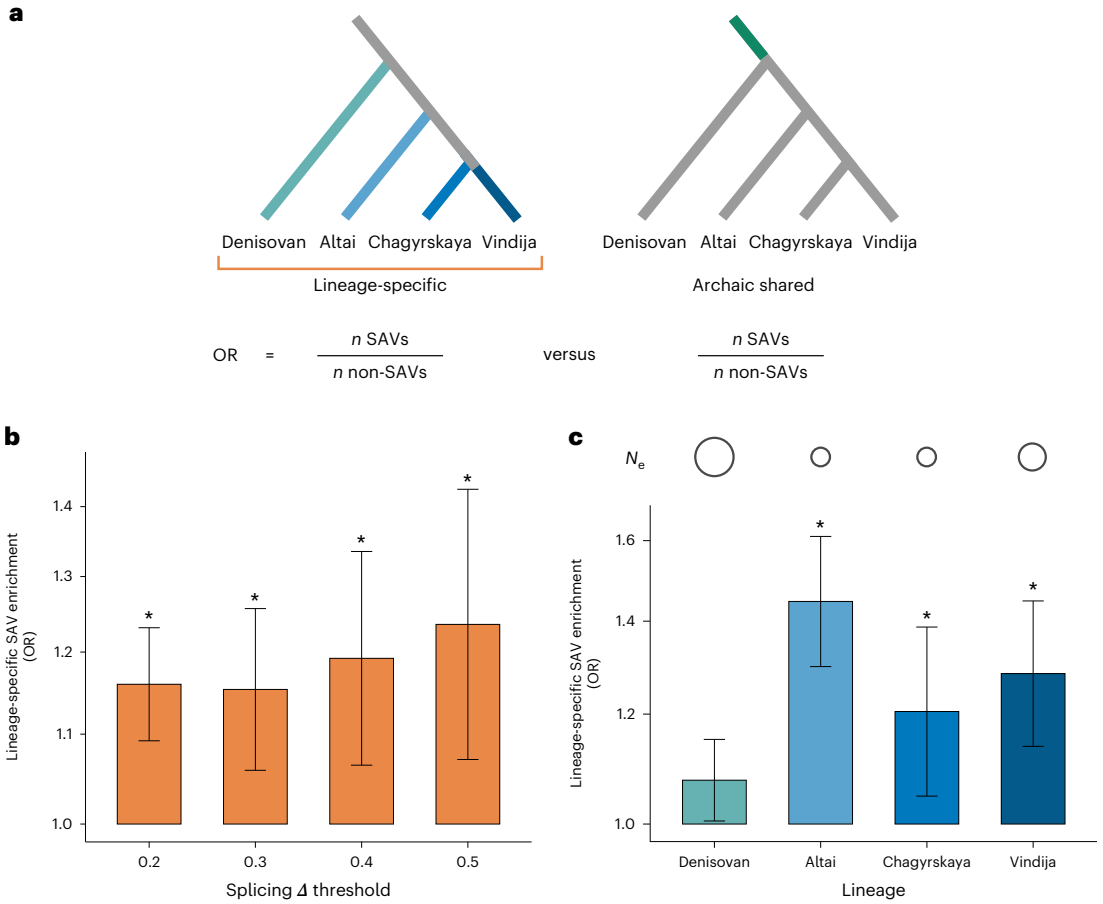


Fig. 4 | Lineage-specific archaic variants are enriched for SAVs compared to shared archaic variants. a, We proposed that lineage-specific archaic variants (left) would be enriched for SAVs compared to variants shared among the four archaic individuals (right), due to less exposure to strong negative selection. To test this, we computed the OR for being a SAV over all variants unique to each lineage (turquoise and blue edges) compared to variants shared among all four lineages (dark green edge). We also proposed that lineage-specific archaic variants would vary in their SAV enrichment compared to shared variants, based on the different effective population sizes and lengths of each branch. To test this, we computed the OR for being a SAV over variants unique to each lineage individually compared to variants shared among all four lineages. **b**, Archaic variants with origins in a specific archaic lineage (n = 618,082) are enriched for SAVs compared to variants shared among all four archaic lineages (n = 573,197). The enrichment increases at increasing SAP (Δ) thresholds. Bar height indicates the OR. Error bars denote the 95% confidence interval and are centred on the OR. Asterisks reflect significance of Fisher's exact tests using a Bonferroni-corrected α (0.0125). Note that the y axis is $\log_{10}$ transformed. The number of lineage-specific and shared SAVs/non-SAVs used in each enrichment test are listed in Supplementary Table 8. **c**, Lineage-specific archaic variants vary in their enrichment for SAVs. The Neanderthal lineage-specific variants have stronger SAV enrichment than Denisovan-specific variants. Estimated N_e per lineage is denoted by a circle above each lineage, with increasing size reflecting larger N_e. The N_e estimates are from ref. 4. Bar height indicates the OR. Error bars denote the 95% confidence interval and are centred on the OR. Asterisks reflect significance of Fisher's exact tests using a Bonferroni-corrected α (0.0125). Note that the y axis is $\log_{10}$ transformed. ORs were calculated from 81,916 Altai-specific, 53,765 Chagyrskaya-specific, 411,492 Denisovan-specific, 70,999 Vindija-specific and 573,197 shared variants. The number of lineage-specific and shared SAVs/non-SAVs used in each enrichment test are listed in Supplementary Table 9.

ratio (OR) = 1.205–1.447; Fig. 4c and Supplementary Table 9). In contrast, variants on the longer and higher N_e Denisovan lineage were not significantly enriched for SAVs (OR = 1.075). At the stricter high-confidence SAV threshold, both the Altai and Vindija Neanderthals remained significantly enriched with increased ORs (Supplementary Fig. 9). These results are consistent with experimental results that found modern humans are depleted for SAVs with strong splicing effects compared to archaics[53].

Introgressed SAVs found in modern humans were present across archaics

We proposed that the evolutionary history of SAVs might also influence their prevalence in modern human populations. For example, introgressed variants experienced strong negative selection in the generations immediately after interbreeding[54], so archaic SAVs that survived stronger and longer-term selection would be more likely to survive in modern humans. To test this, we first considered the distribution of remaining introgressed variants among the archaics.

Most introgressed SAVs were present in all Neanderthals (n = 143) or present in all archaics (n = 68; Supplementary Table 10). No SAVs private to Vindija or Chagyrskaya nor shared between both late Neanderthals were identified as introgressed, even though Neanderthal ancestry in most modern humans is most closely related to Vindija and Chagyrskaya[3,4]. This is consistent with weaker selection on lineage-specific SAVs and previous work suggesting that older introgressed archaic variants were more tolerated in humans[55–57].

To further test this, we compared the expected origin distribution for introgressed SAVs (based on the distribution of archaic-specific SAVs) to the observed distribution for introgressed SAVs. Fewer Altai-specific SAVs occur among introgressed variants whereas shared Neanderthal SAVs are more prevalent than expected (Fig. 5a). This pattern remains for SAVs from ref. 48 and high-confidence SAVs from both (Supplementary Fig. 10). These patterns suggest that older SAVs, either those that evolved before the Neanderthal common ancestor

https://doi.org/10.1038/s41559-023-02053-5

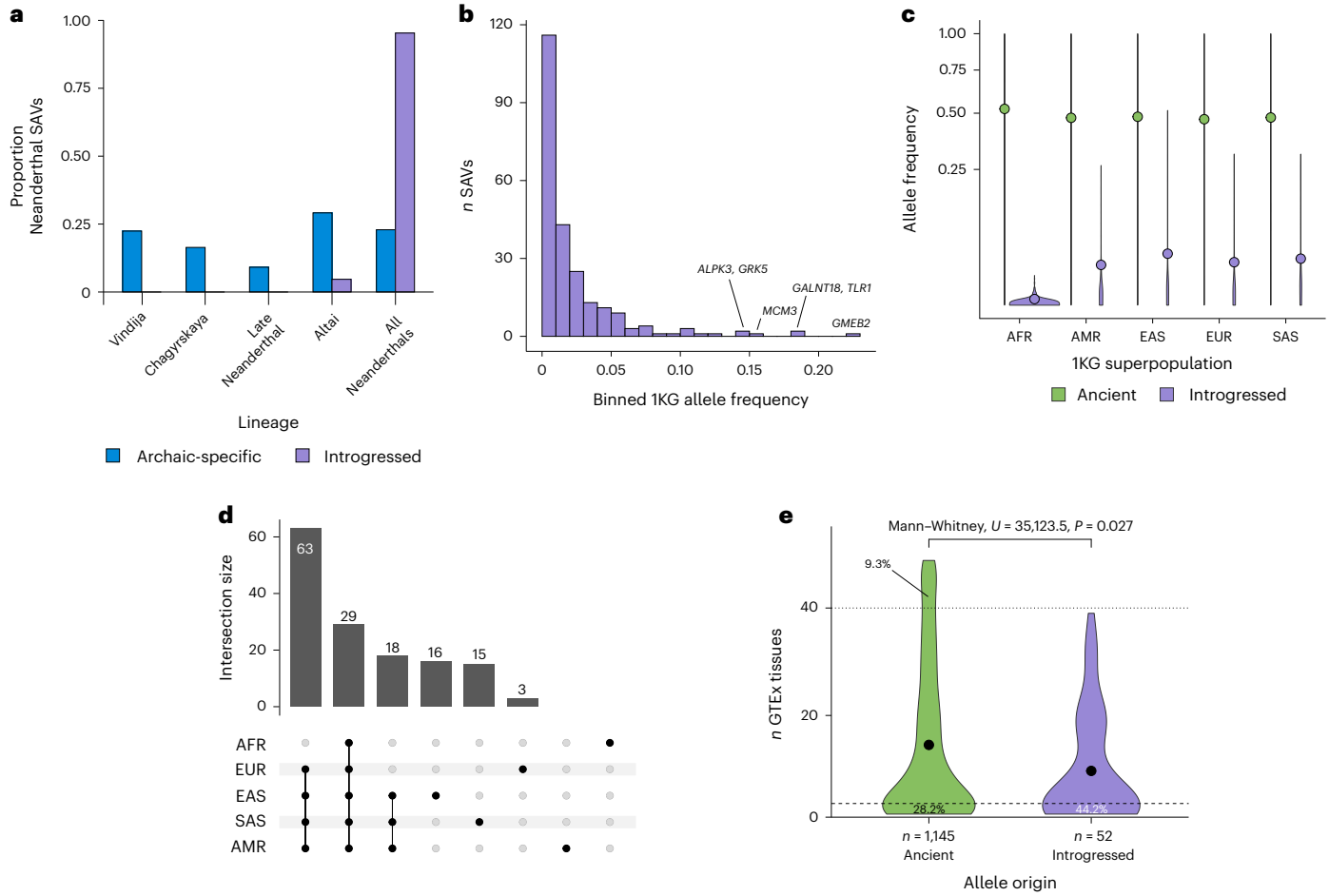


Fig. 5 | Introgressed SAVs present in modern humans were shared across archaic individuals and are associated with increased tissue specificity. **a**, Histograms comparing distributions of the presence of all Neanderthal SAVs (blue) and introgressed SAVs (purple) in different sets of Neanderthal individuals. Introgressed SAVs are older than expected from all Neanderthal SAVs. We focused on Neanderthal lineages because of low power to detect introgressed Denisovan SAVs. All data are presented in Supplementary Table 10. **b**, Allele frequency distribution for introgressed SAVs as per ref. 47. Allele frequencies represent the mean from the 1KG African, American, East Asian, European and South Asian superpopulation frequencies. **c**, Allele frequency distributions for SAVs present in both archaic and modern individuals stratified by 1KG superpopulation and origin (ancient versus introgressed) as per ref. 47. Ancient SAVs ($n = 2{,}252$) are coloured green and displayed on the left, while introgressed SAVs ($n = 237$) are coloured purple and displayed on the right per superpopulation. AFR, African; AMR, American; EAS, East Asian; EUR, European; and SAS, South Asian. The coloured dot represents the mean allele frequency for each set. The *y* axis is square-root transformed. **d**, The number of introgressed SAVs with a minimum allele frequency of at least 0.01 in each modern human population. We display all individual populations, the non-African set and Asian/American set here. See Supplementary Fig. 15 for all sets. **e**, The distribution of the number of GTEx tissues in which an ancient or introgressed SAV (as per ref. 47) was identified as an sQTL. Introgressed variants are significantly more tissue specific. We defined 'tissue-specific' variants as those occurring in one or two tissues and 'core' sQTLs as those occurring in >40 of the 49 tissues. The dashed and dotted lines represent these definitions, respectively. The proportion of SAVs below and above these thresholds are annotated for each allele origin. In all panels, introgressed SAVs and frequencies are as defined by ref. 47.

or before the Denisovan and Neanderthal common ancestor were the most tolerated after introgression.

Consistent with known introgression patterns, introgressed SAVs occurred at lower overall frequencies (Fig. 5b). However, a few introgressed SAVs occur at modest to high frequencies among genes including *GMEB2*, *GALNT18* and *TLR1* (Fig. 5b). The last occurs in an adaptively introgressed locus spanning three toll-like receptors—key components of the innate immune system[58] and this SAV have been confirmed to generate an isoform using a massively parallel splicing reporter assay[53].

In contrast, ancient SAVs occur at high frequencies in all five 1KG superpopulations (Supplementary Figs. 11 and 12) and their frequencies are significantly higher among Africans (mean (μ) = 0.522) than non-Africans ($\mu = 0.476$) (Mann–Whitney $U = 10{,}963{,}956$, $P = 2.66 \times 10^{-9}$) (Fig. 5c and Supplementary Fig. 13). Introgressed SAVs have significantly lower frequencies in all superpopulations (Fig. 5c) and are less likely to be shared among multiple populations (Fig. 5d and Supplementary Figs. 14 and 15).

It is possible these patterns reflect the general attributes of introgressed variants, rather than splicing effects of SAVs. We therefore examined the relationship between allele frequency in 1KG and SAP (Δ max) for introgressed SAVs. The 1KG populations did not generally differ in Δ max for either ancient or introgressed SAVs, although introgressed SAVs had a higher Δ max (Supplementary Fig. 16). We anticipated, however, that introgressed SAVs predicted to have stronger effects on splicing would occur at lower frequencies. Indeed, we found a significant negative association between allele frequency and Δ max for $\Delta \geq 0.2$ (Spearman, $\rho = -0.2378$, $P = 0.0002$) (Extended Data Fig. 8). This pattern is probably absent among ancient variants due to purifying selection on deleterious variants that occurred before the divergence of archaics and moderns. Further, our prediction that ancient SAVs were

more likely to have no effect on the resulting transcript and protein (Fig. 3b) is consistent with this hypothesis.

Introgressed SAVs have immune, skeletal and reproductive associations

We tested whether any functional annotations (GWAS or HPO terms) were enriched among the 361 genes with introgressed SAVs from ref. 48 and 239 genes with introgressed SAVs from ref. 47 (Supplementary Data 2). Two terms were significantly enriched among genes with introgressed SAVs[47]: adverse response to breast cancer chemotherapy (GWAS) and oligohydramnios (HPO) (Extended Data Fig. 9). Four HPO terms related to hip-girdle, pelvic and shoulder muscles were enriched among genes with ref. 48 introgressed SAVs (Extended Data Fig. 10b). However, 19 GWAS terms met our FDR-corrected significance threshold including *Helicobacter pylori* serologic status and systemic sclerosis (Extended Data Fig. 10a). Overall, these results suggest that SAVs surviving in modern human populations influence several immune, skeletal and reproductive phenotypes.

We further considered the potential functional effects of introgressed SAVs by intersecting them with Neanderthal variants exhibiting allele-specific expression (ASE) in modern humans[59]. We identified 16 SAVs out of 1,236 ASE variants, including variants in *GSDMC*, *HSPG2* and *RARS* (Supplementary Table 11). The SAV in *HSPG2*, predicted to create a donor gain, was recently validated using a massively parallel splicing reporter assay[53]. We also note that a handful of the ref. 59 ASE variants fell just under our SAV threshold. Among these is a Neanderthal variant (rs950169) in *ADAMTSL3* that results in a truncated protein[59]. SpliceAI correctly predicted the loss of the upstream acceptor (Δ = 0.19), although it did not indicate the downstream acceptor gain.

Introgressed SAVs are more tissue specific than ancient SAVs

Given their different histories of selective pressures, we proposed that introgressed SAVs would be more tissue specific than would ancient SAVs in their effects. To explore this, we identified 1,381 archaic SAVs with splicing quantitative trait loci (sQTL) data from the genotype–tissue expression (GTEx) project across 49 tissues.

Introgressed sQTL SAVs were significantly associated with tissue-specific gene expression compared to ancient sQTL SAVs (Mann–Whitney U = 35,123.5, P = 0.027) (Fig. 5e). On average, introgressed SAVs influenced splicing in 4.92 fewer tissues than did ancient SAVs. Further, all sQTL SAVs with broad effects (>40 tissues) were ancient (107 high-confidence and 5 low-confidence). A total of 74 of these were shared among all four archaics (Supplementary Table 12), suggesting that core sQTL SAVs were more likely to evolve in the deep past. These patterns were also observed among the variants from ref. 48 (Supplementary Fig. 17). Collectively, 30% of sQTL SAVs (n = 427) were associated with tissue-specific effects on splicing (one or two tissues) (Supplementary Fig. 18). Consistent with known gene expression patterns, testis had the most unique sQTL among SAVs, followed by skeletal muscle and thyroid.

Variation in gene expression among tissues may also influence the efficacy of negative selection to remove deleterious SAVs. For example, ref. 60 demonstrated that more ubiquitously expressed genes in *Paramecium tetraurelia* exhibited less alternative splicing compared to genes with more tissue-specific expression due to the differences in the efficacy of negative selection. We predicted that tissue specificity of expression would be associated with the number of SAVs per gene or maximum Δ. We quantified tissue specificity of expression using the relative entropy of the transcripts per million (TPM) count for each gene across tissues compared to the expression distribution across tissues for all genes in GTEx. This metric ranges from 0 to 1, with higher values reflecting greater tissue specificity. Most genes exhibited broad expression (Supplementary Fig. 19), so we divided genes into low, medium and high tissue specificity on the basis of the relative entropy.

Genes with the most tissue-specific expression had significantly higher median maximum SAP (Δ) than did genes with broader expression patterns (Supplementary Fig. 20a; Kruskal–Wallis H = 6.599, P = 0.037). This could indicate greater selection against SAVs likely to influence expression across many tissues; however, we note that the effect was small in magnitude. The distribution of the number of archaic SAVs per gene did not differ significantly between entropy classes (Supplementary Fig. 20b; Kruskal–Wallis H = 1.89, P = 0.388); all had a median of 1 SAV.

Archaic SAVs with potential evolutionary significance

Many archaic SAVs influence genes with known or previously suggested significance to the evolutionary divergence between archaic hominins and modern humans. For example, the 2′–5′ oligoadenylate synthetase *OAS* locus harbours an adaptively introgressed SAV at Chr. 12: 113,357,193 (G>A) that disrupts an acceptor site and results in multiple isoforms and leads to reduced activity of the antiviral enzyme[61,62]. This ancestral variant was reintroduced to modern Eurasian populations by Neanderthal introgression[63]. SpliceAI correctly predicted the acceptor loss at this site (Δ = 0.89). This locus harbours 92 additional archaic variants (n = 92). We found one additional SAV at Chr. 12: 113,355,275 in *OAS1* that potentially results in an acceptor gain (Δ = 0.26). This allele was unique to the Denisovan; it is derived and was present in only one of 2,054 1KG samples as a heterozygote. This suggests potential further splice variant evolution of this locus, with possible Denisovan-specific effects.

We also identified several variants at other well-studied loci. Variation in human populations at the *EPAS1* locus includes a Denisovan-introgressed haplotype thought to contribute to adaptation to living at high altitude among Tibetans[64]. Of 184 archaic variants occurring at this locus, we identified two as candidate SAVs. One variant (Chr. 2: 46,584,859; rs372272284) is homozygous in the Denisovan, whereas all Neanderthals have the human reference allele (Fig. 6a). The variant is introgressed and present at low frequency in East Asians in 1KG and is also the lead variant in an observed association of the introgressed haplotype with decreased haemoglobin levels in Tibetans[65]. This SAV strengthens a canonical 5′ splice site (CAA|GT to CAG|GT)[29], resulting in a donor gain (Δ = 0.37) (Fig. 6a). If used, this splice site would introduce multiple stop codons, resulting in NMD (Supplementary Data 3). This would result in the same molecular effect (decreased circulating *EPAS1* RNA) that is thought to contribute to hypoxia adaptation[66]. The other candidate SAV (Chr. 2: 46,610,904) is absent from 1KG/gnomAD and occurs as a heterozygyote in the Altai Neanderthal and is near the end of the last intron of the gene, making it much less likely to fundamentally alter the mRNA product.

We also identified three archaic SAVs within *ERAP2*, a gene subject to strong and consistent balancing selection in different human populations[67]. SpliceAI correctly identified a previously characterized human variant (Chr. 5: 96,235,896; rs2248374), which causes a donor loss (Δ = 0.51) and results in a truncated protein and subsequent NMD of the mRNA. However, we identified an additional Neanderthal SAV, which is also introgressed and occurs at low frequencies among Americans (5%), Europeans (6%) and South Asians (2%) in 1KG (Fig. 6b). This SAV, rs17486481, is a donor gain (Δ = 0.53) that introduces a canonical 5′ splice site (AT|GTAAT to AT|GTAAG) and would similarly result in NMD (Fig. 6b and Supplementary Data 3). However, this allele always occurs with the non-truncated version of rs2248374 (while being much rarer) and the need to maintain the non-truncated allele is probably why it remains uncommon. The third variant (Chr. 5: 96,248,413) was archaic-specific—occurring as a heterozygote in the Altai Neanderthal—and results in an acceptor gain (Δ = 0.24).

Discussion

Alternative splicing plays a critical role in organismal biology, particularly during development and establishing tissue

https://doi.org/10.1038/s41559-023-02053-5

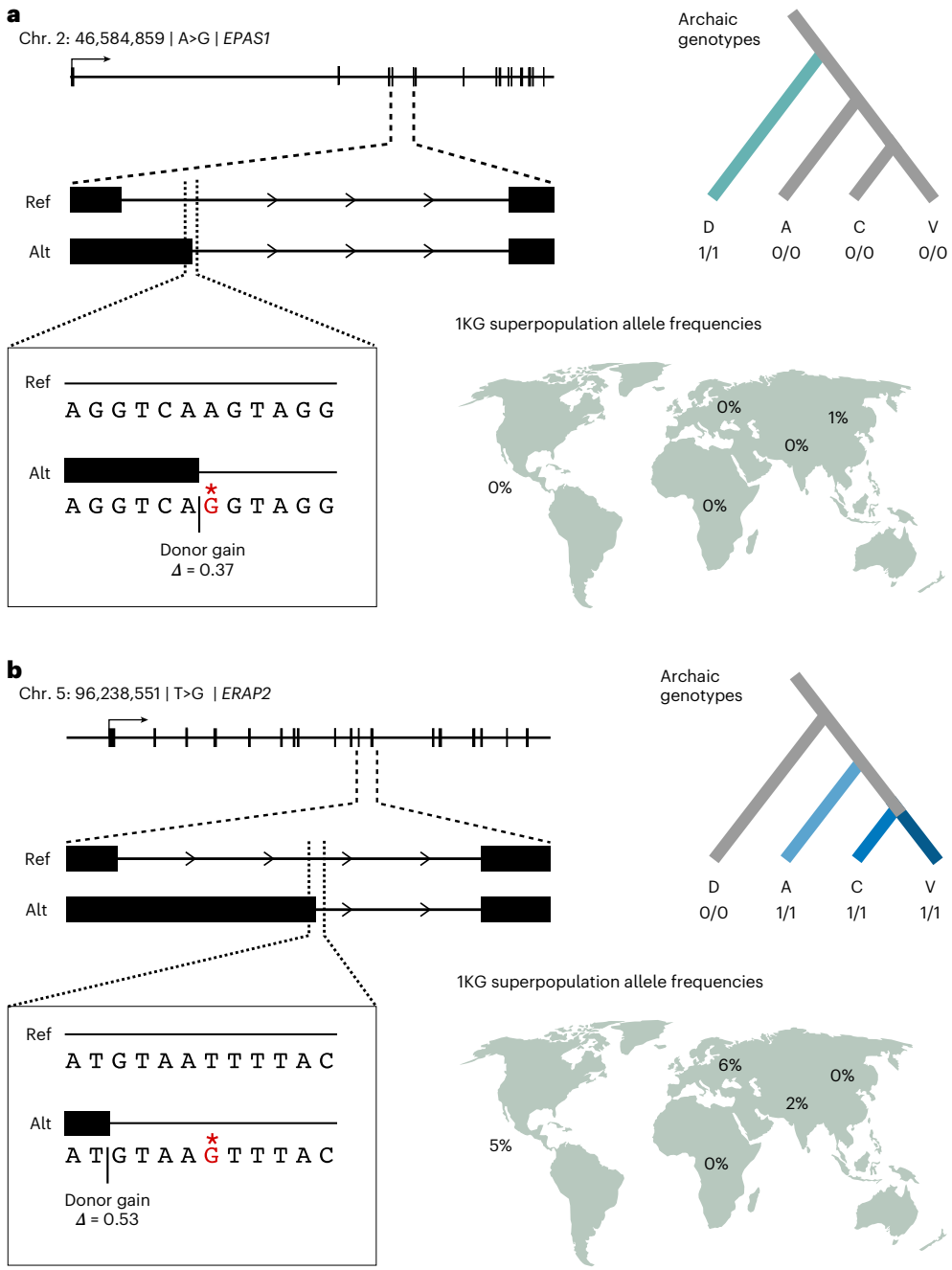


Fig. 6 | Example archaic SAVs leading to NMD in loci with evidence of recent adaptive evolution. a, A Denisovan-specific homozygous SAV results in a donor gain in *EPAS1*, hypoxia-inducible factor-2α, between the fourth and fifth exon. The transcript resulting from the SAV introduces six PTCs (Supplementary Data 3), which probably results in the elimination of the transcript via NMD. This SAV potentially contributes to the effects of the introgressed haplotype in Tibetan adaptation to living at high altitude. This variant is present as a heterozygote in 12 individuals from 1KG: 8 from the East Asian superpopulation and 4 from the South Asian superpopulation. **b**, Three archaic SAVs, including a Neanderthal-specific variant, occur in *ERAP2*, an MHC presentation gene with evidence of strong balancing selection in human populations. Consistent with this, the SNV occurs at low frequency in three of the five 1KG superpopulations. As in the *EPAS1* example, this variant results in a donor gain between the eleventh and twelfth exons, which introduces nine PTCs (Supplementary Data 3).

identity[14]. Thus, alternative splicing often contributes to adaptation and phenotypic differences between closely related species[23–28]. The development of machine-learning algorithms that can predict alternative splicing from sequence alone now enables analysis of alternative splicing in populations for which transcriptomic data are difficult or impossible to generate, including archaic hominins. Here, we use SpliceAI to uncover the previously unobservable genome-wide alternative splicing landscape of archaic hominins.

We identify thousands of putative SAVs from the high-coverage genomes of three Neanderthals and a Denisovan. We find that many of these variants do not occur in modern humans and we propose that they are implicated in specific phenotypic differences between archaic hominins and modern humans. Additionally, many SAVs are shared with modern humans and are ancient, evolving before the common ancestor of archaic hominins and modern humans. Furthermore, a few hundred SAVs are present in human populations due to introgression and these surviving introgressed SAVs are almost entirely shared across

Neanderthals. We also observe multiple lines of evidence supporting the role of negative selection in shaping SAV patterns.

Given that introgressed and ancient SAVs are present in modern humans, their splicing patterns have the potential to be directly studied to further understanding of their phenotypic effects. We found that 36.7% of non-archaic-specific SAVs were identified in modern humans in GTEx as sQTLs. There are several reasons why archaic SAVs might not have been detected as sQTL. Splicing is often tissue specific and GTEx assayed only a small fraction of tissues and contexts. Furthermore, splicing is influenced by sequence but is also influenced by other cellular dynamics, such as polymerase pausing[29]. These, along with limited statistical power in many GTEx tissues, particularly for SAVs at low frequency in Europeans, mean that many SAVs should not necessarily be detected. Indeed, we observe higher fractions of SAVs as sQTL when analysing high-frequency variants (Supplementary Fig. 21). The tissue specificity of archaic SAVs is of great interest but the degree to which SAV tissue specificity in modern humans reflects specificity in archaic hominins is unknown without further experimental study. However, such studies are challenging because the genomic and archaic cellular context cannot be perfectly replicated (that is, testing an archaic SAV in a Neanderthal genome background in a Neanderthal tissue)[5,68].

Our results offer new insight into an essential molecular mechanism and previously unstudied attributes of archaic hominins; however, we note some limitations of our approach. First, we did not include structural variants (InDels) or variants from the sex chromosomes in this analysis, both of which warrant further study. For example, the X chromosome exhibits high levels of alternative splicing[69] and splicing can occur in a sex-specific manner[26,70]. However, for now, we only have high-coverage archaic hominin sequences from females. Future development and application of models with sex-specific transcriptomic data may offer additional phenotypic insights. Second, the tag single nucleotide polymorphisms (SNPs) and modern human samples used in this analysis are best suited to identifying Neanderthal rather than Denisovan introgression[47,48]. Our conservative approach for identifying introgressed haplotypes means that the number of introgressed SAVs reported here is an underestimate and does not include Denisovan-derived SAVs. Multiple modern human populations contain considerable Denisovan ancestry, therefore, future work should consider these variants.

In summary, our approach of combining machine learning with ancient DNA and modern population genetic data identifies thousands of archaic variants that potentially alter splicing, including many that appear to be specific to archaic hominins. Genes affected by archaic SAVs are enriched for roles in a variety of phenotypes and several influence loci with known relevance to recent human evolution. For example, two archaic SAVs that we highlight probably cause NMD of the resulting *EPAS1* and *ERAP2* transcripts. Downregulation of *EPAS1* is thought to underlie high-altitude adaptation in Tibetans[66]. In *ERAP2*, another variant in human populations that induces NMD has experienced strong balancing selection[67]. These examples underscore that phenotypic effects from alternative splicing are not limited to expanded proteomic diversity but also downregulation of gene expression via NMD[71,72]. Further work is needed to understand the functional effects of these and other archaic SAVs. Others[53] recently used a massively parallel splicing reporter assay to assess exonic SAVs in archaics and modern humans, validating several predictions from the present study. However, this assay is limited to testing only a subset of exonic variants and additional assays are required to test the other types of exonic and any intronic SAVs[53]. Nonetheless, our results suggest that alternative splicing played a role in hominin divergence and offers specific molecular hypotheses for testing. The identification of archaic-specific splice variants here will enable future analysis of human-specific splice variants. We also anticipate that our sequence-based approach will enable study of alternative splicing in other extinct or difficult to sample taxa.

Methods

Archaic genomic data

We retrieved four high-coverage publicly available archaic hominin genomes representing three Neanderthals[2–4] and a Denisovan[1].

We excluded sites that were invariant among the archaic individuals (ALT=.) and variants with low site quality (QUAL < 30). Further, low-quality genotypes were set to missing on the basis of read depth (FMT/DP < 10) and genotype quality (FMT/GQ < 30). We also normalized InDels and split multi-allelic records into separate entries for positions with multiple variants (norm -m -). All filtering was completed using bcftools, v.1.13 (ref. 73).

All genomic coordinates presented in this article and Supplementary Information refer to hg19/GRCh37.

Variant annotation

We annotated variants for putative alternative splicing using SpliceAI v.1.3.1 (ref. 35). Briefly, SpliceAI uses a deep residual neural network to estimate the SAP and position change of each variant from DNA sequence alone making it ideal for studying archaic hominins, for which we cannot obtain transcript-level data. The model considers 5 kilobase pairs flanking the variant in both directions. The output includes four SAPs (Δs) for (1) acceptor gain (AG), (2) acceptor loss (AL), (3) donor gain (DG) and (4) donor loss (DL) as well as the position changes for each of the four deltas. The Δs range from 0 to 1 and represent the likelihood a variant is splice-altering for one or more of the four categories. We implemented SpliceAI in a Conda package using keras v.2.3.1 (ref. 74) and tensorflow v.1.15.0 (ref. 75). After filtering, we ran SpliceAI on sets of 5×10^3 variants using the hg19 reference genome using the GENCODE, Human Release 24, annotations[76] included with the package. We used the default parameter for maximum distance between a variant and gained/lost splice site (50 bp) and used the raw precomputed files. We concatenated all results and further split variants with multiple annotations. Among all variants, 32,105 exhibited multiple annotations with different effects on splicing (Supplementary Table 2). While we included InDels and variants on the X chromosome in this scan, we restricted all downstream analyses to autosomal single nucleotide variants (SNVs) (Discussion).

Defining SAVs

For each variant, we identified the maximum SAP (Δ) among all four classes: AG, AL, DG and DL. We then defined SAVs using two Δ thresholds: Δ max ≥ 0.2 and 0.5, 'SAVs' and 'high-confidence SAVs', respectively.

We determined whether the number of SAVs identified in each archaic individual was different than expected by randomly selecting a sample from 24 1KG populations. We annotated all variants present among these individuals using SpliceAI and the hg38 annotations included with the package. We then analysed the variants as for the archaics (that is, splitting multi-allelic sites and variants with multiple GENCODE annotations). We filtered for variants with a Δ max ≥ 0.2 and summed the number of variants per 1KG sample that had at least one alternate allele present.

Archaic variants in modern humans

We noted the distribution of each variant among the archaics using eight classes: (1) Altai, (2) Chagyrskaya, (3) Denisovan, (4) Vindija, (5) Late Neanderthal (Chagyrskaya and Vindija), (6) Neanderthal (Altai, Chagyrskaya and Vindija), (7) shared (all four archaics) and (8) other (all remaining possible subsets). The assignment was based on the presence of at least one allele with an effect.

We assessed whether any variants present among the archaics are also present in modern humans using biallelic SNVs and InDels from 1KG, Phase 3 (ref. 40) and SNVs from gnomAD v.3 (ref. 41). We used LiftOver[77] to convert archaic variants from hg19 to hg38. We then normalized variants (norm -m - -f hg38.fa) and subset variants to those

within gene bodies (view -R genes.bed). We queried these variants for allele count, allele number and allele frequencies (query -f). Further, for 1KG variants, we retrieved allele frequency per 1KG superpopulation: Africa (AFR), Americas (AMR), East Asia (EAS), Europe (EUR) and South Asia (SAS). These precomputed values had been rounded to two decimal places in the Variant Call Format files (VCFs). Normalization, filtering and querying were done using bcftools[73]. After using LiftOver to convert back to hg19 coordinates, we merged the 1KG and gnomAD variants with the archaic variants ensuring that the archaic and modern reference and alternate alleles matched. We recalculated the 1KG allele frequency for Africans as the annotated frequency included samples from an admixed African population: African ancestry in the southwestern United States (ASW). We subset samples from Esan (ESN), Mandinka (GWD), Luhya (LWK), Mende (MSL) and Yoruban (YRI) and calculated allele frequency as allele count divided by allele number per site.

We used two datasets to identify introgressed variants[47,48]. These datasets differ in their approach to recognizing introgressed sequences and partly overlap the archaic variants considered in this study. Others[47] used the S^* statistic to classify human sequences as introgressed. S^* leverages high linkage disequilibrium among variants in an admixed target population that are absent in an unadmixed reference population[78,79]. Introgressed haplotypes are then identified by maximizing the sum of scores among all SNP subsets at a particular locus[78,79]. Tag SNPs are those variants that match an archaic allele and occur with at least two other tag SNPs in a 50 kb window. Haplotypes were defined as regions encompassing ≥5 tag SNPs in LD within a given human population ($R^2 \geq 0.8$). We collated tag SNPs from all four populations: East Asian (ASN), European (EUR), Melanesian (PNG) and South Asian (SAS). We retained all metadata from ref. 47. A handful of tag SNPs encompass multiple haplotypes that reflect differences in haplotype size between modern human populations; we retained the first record per variant. Others[48] developed a modified S^* statistic, Sprime, which uses a scoring method that adjusts the score based on the local mutation and recombination rates, allows for low-frequency introgression in the unadmixed outgroup and avoids windowing to identify introgressed segments. We collated introgressed variants for 20 non-African populations and filtered for those that matched the Altai Neanderthal at high-quality loci.

A handful of sequences in the hg19 reference genome are introgressed from archaic hominins. Therefore, we maximized the number of introgressed sites we could analyse by defining sites, rather than variants, as introgressed if either the reference or alternate allele for each SAV matched any Neanderthal base at a matching position. We ensured that SpliceAI predictions were similar for these allele pairs, regardless of which was the reference and alternate, by generating a custom hg19 sequence where introgressed reference alleles (n = 7,977) from ref. 47 were replaced by the alternate allele using a custom script. We then applied SpliceAI to the introgressed reference alleles, now considered to be the alternate. We found that 24 of the 26 variants were classified as SAVs (Supplementary Table 13). One of the remaining two variants was nearly identical in splicing probability (Δ max = 0.19 and Δ max = 0.2), whereas the other variant's predictions were different (Δ max = 0.16 and Δ max = 0.31) (Supplementary Table 13). Given this overall similarity, we maintained the original predictions for introgressed[47] reference alleles in our dataframe but provide the predictions when these nucleotides are the alternate allele for all SAVs and non-SAVs in the project GitHub repository. We recalculated allele frequencies for all introgressed variants to account for sites where the reference sequence contained introgressed alleles, as the precomputed 1KG allele frequency would be incorrect. The ref. 48 metadata designate whether the reference or alternate allele is introgressed. Therefore, we used the 1KG allele frequency for sites with an introgressed alternate allele and subtracted the 1KG allele frequency from 1 for sites with an introgressed reference allele. For ref. 47 introgressed variants, we calculated an average from the metadata, which included the allele frequencies in various populations for the introgressed allele. We took the mean of the AFR, AMR, EAS, EUR and SAS frequencies for all introgressed positions.

The presence of introgressed alleles in the human reference results in previously excluded human polymorphisms due to our filtering criteria. We quantified these potential ancient or introgressed SAVs by intersecting sites that were fixed among the archaics for the human reference with introgressed alleles from refs. 47,48 using bedtools2 v.2.30 (ref. 80). We repeated the above procedure, inserting the alternate alleles from intersected sites into the hg19 reference and running SpliceAI on the reference alleles formatted as alternate alleles. This yielded 12,003 variants from 11,833 positions, of which 41 variants had a Δ max ≥ 0.2. We do not include these variants in the main text but the SpliceAI predictions for all SAVs and non-SAVs from this set are available in the project GitHub repository.

We categorized each variant's 'origin' on the basis of presence in 1KG and gnomAD as well as whether or not the variant was introgressed. Further, we classified each variant's allele origin on the basis of introgressed variants identified by ref. 47 'Vernot allele origin' and ref. 48 'Browning allele origin' due to the incomplete overlap among variants in those datasets. Variants that did not occur in 1KG or gnomAD were defined as 'archaic-specific'. Low-frequency variants in modern humans are also highly likely to be the result of recurrent mutation rather than shared ancestry. In support of this hypothesis, we found that CpGs were enriched among rare variants (allele frequency <0.0001) versus non-CpG common variants (allele frequency ≥0.01) (Fisher's exact test, OR = 1.88, $P < 0.0001$). Therefore, we also designated gnomAD variants whose allele frequency was <0.0001 as 'archaic-specific'. Sample sizes in 1KG and GTEx do not permit this level of sensitivity; therefore, allele frequency was only considered for gnomAD variants. Variants that were present in 1KG or gnomAD at an allele frequency ≥0.0001 and introgressed were defined as 'introgressed'. Variants that were present in 1KG or gnomAD at an allele frequency ≥0.0001 but not introgressed were considered 'ancient' at two confidence levels. 'High-confidence ancient' variants were present in at least two 1KG superpopulations at an allele frequency ≥0.05, while 'low-confidence ancient' variants did not meet this threshold. We report analyses on the high-confidence ancient set; this helps to remove cases of potential convergent mutation. We did not retrieve population-level allele frequency data for gnomAD variants; therefore, common variants present in gnomAD and absent from 1KG were classified as 'low-confidence ancient'. We restricted analyses including allele frequency as a variable to 1KG variants with population-level allele frequencies.

Gene characteristics, mutation tolerance and conservation

We used the SpliceAI annotation file for hg19 from GENCODE, Human Release 24 (ref. 76), to count the number of exons per gene and calculate the length in base pairs of the gene body and the coding sequence. The number of isoforms per gene were retrieved from GENCODE, Human Release 40. We retrieved missense and loss-of-function (LoF) observed/expected ratios from gnomAD[41] to quantify each gene's tolerance to mutation. We also considered conservation at the variant level. We used the primate subset of the 46-way multispecies alignment[51]. Positive phyloP scores indicate conservation or slower evolution than expected, whereas negative phyloP scores indicate acceleration or faster evolution than expected based on a null hypothesis of neutral evolution.

Phenotype enrichment

We followed the approach of ref. 10 to assess enrichment for SAVs in genes implicated in different human phenotypes. Many gene enrichment analyses suffer from low power to detect enrichment because an entire genome is used as the null distribution. Relatedly, SAVs are unevenly distributed throughout archaic genomes. We addressed this issue by generating a null distribution from the observed data. We first retrieved phenotypes and the associated genes per phenotype from Enrichr[81–83]. We used both the 2019 GWAS Catalog and the HPO.

The GWAS Catalog largely considers common disease annotations and has 1,737 terms with 19,378 genes annotated[49], whereas HPO largely considers rare disease annotations and has 1,779 terms with 3,096 genes annotated[50]. All 3,516 terms were manually curated into one of 16 systems: behavioural, cardiovascular, digestive, endocrine, haematologic, immune, integumentary, lymphatic, metabolic, nervous, other, reproductive, respiratory, skeletal, skeletal muscle and urinary.

We considered nine different gene sets, generated using SAVs with $\Delta \geq 0.2$, for our enrichment analyses: (1) genes with lineage-specific Altai SAVs ($n = 283$), (2) genes with lineage-specific Chagyrskaya SAVs ($n = 165$), (3) genes with lineage-specific Denisovan SAVs ($n = 859$), (4) genes with lineage-specific Vindija SAVs ($n = 228$), (5) genes with SAVs present in all three Neanderthals ($n = 227$), (6) genes with SAVs shared among all four archaics ($n = 106$), (7) genes with all archaic-specific SAVs ($n = 1{,}907$), (8) genes with introgressed SAVs per ref. 47 ($n = 239$) and (9) genes with introgressed SAVs per ref. 48 ($n = 361$). The shared set only included variants present in all four archaics and excluded those that were inferred from parsimony. We retained duplicated gene names to reflect genes with multiple SAVs.

We identified which genes were present in both the GWAS Catalog and HPO per set using a Boolean to calculate the observed gene counts per term per ontology. We then removed GWAS and HPO terms per set that did not include at least one gene from the set. This resulted in 631 Altai, 1,407 archaic-specific, 761 ref. 48 introgressed, 412 Chagyrskaya, 1,023 Denisovan, 515 Neanderthal, 295 shared, 627 ref. 47 introgressed and 474 Vindija terms for the 2019 GWAS Catalog and 622 Altai, 1,490 archaic-specific, 720 ref. 48 introgressed, 391 Chagyrskaya, 1,152 Denisovan, 528 Neanderthal, 306 shared, 651 Vernot et al. 2016 introgressed and 522 Vindija terms for the HPO.

The max Δ was then shuffled across all 1,607,350 variants without modifying the annotation, allele origin or distribution data. The distribution of genes for both ontologies was then recorded. We repeated this process 1×10^4 times per set and calculated enrichment as the number of observed genes divided by the mean empirical gene count per term. The P values were calculated as the proportion of empirical counts +1 ≥ the observed counts + 1. We adjusted our significance level due to multiple testing by correcting for the FDR. We used a subset ($n = 1 \times 10^3$) of the empirical null observations and selected the highest P value threshold that resulted in a $V/R < Q$ where V is the mean number of expected false discoveries and R is the observed discoveries[10]. We calculated adjusted significance levels for each set for Q at both 0.05 and 0.1.

New transcripts and proteins

We constructed a new transcript per SAV to assess downstream effects on the resulting protein. We generated a canonical transcript per gene using the exons defined from GENCODE. Next, we constructed a new transcript using the splicing alteration class (acceptor gain, acceptor loss, donor gain or donor loss) and associated position information on the maximum distance between a variant and gained/lost splice site per SAV. As we used the default SpliceAI settings for analysis, this maximum distance was set to 50 bp. For SAVs with multiple splicing class alterations, for example, a variant that results in both an acceptor gain and an acceptor loss, we modelled the alteration class per SAV with the largest Δ.

For acceptor and donor gains, we identified the relevant exon and added or removed sequence based on the SpliceAI prediction (Extended Data Fig. 7). For acceptor losses, we removed the subsequent exon from the transcript if the effect occurred at the upstream exon boundary (Extended Data Fig. 7). Similarly, we added the intronic sequence to the canonical transcript for donor losses that occurred at the exon boundary (Extended Data Fig. 7). For each of these scenarios, we included the variant when appropriate but otherwise kept the canonical sequence. We then retrieved a single start codon position from the associated general feature format file (GFF), prioritizing the longest and experimentally validated (Ensembl) versus computationally predicted (HAVANA) start sites when there were multiple start positions.

We compared the canonical and new transcripts and the resulting protein (Supplementary Data 3). We assigned each SAV as resulting in one of seven effects on the basis of transcript/protein length and composition: (1) the canonical and new transcripts/proteins are identical (no effect), (2) the new protein is longer than the canonical one and includes PTCs (longer with PTCs), (3) the new protein is longer and does not include PTCs (longer without PTCs), (4) the 5′ or 3′ UTR is longer than the canonical transcript, (5) the new protein is the same length as the canonical one but contains a single missense variant (single missense), (6) the new protein is shorter than canonical protein (truncated protein) and (7) the 5′ or 3′ UTR is more truncated than the canonical transcript (truncated UTR).

Gene expression, tissue specificity and sQTL

We used TPM counts for each gene from GTEx v.8, to analyse expression. We quantified tissue specificity as the relative entropy of each gene's expression profile across 34 tissues compared to the median across genes overall. Thus, a gene with expression only in a small number of tissues would have high relative entropy and a gene with expression across many tissues would have low relative entropy to this background distribution. The 34 tissues were selected on the basis of groupings from the Human Protein Atlas to minimize the amount of sharing between distinct tissues, for example, since brain tissues are overrepresented in GTEx. We used median expression across tissues as the null and calculated relative entropy using the entropy function from the SciPy statistics package[84]. On the basis of the observed distribution of relative entropy scores (Supplementary Fig. 19), we designated genes with scores ≤0.1 as 'low tissue specificity', genes with scores >0.1 and ≤0.1 as 'medium tissue specificity' and genes with scores >0.5 as 'high tissue specificity'. We compared both the number of SAVs per gene and the maximum Δ for SAVs among the three relative entropy categories.

We downloaded sQTL data from GTEx v.8. We collated significant variant–gene associations ($n = 24{,}445{,}206$) across all 49 tissues and intersected these with SAVs, using LiftOver[77] to convert the SAVs to hg38 and then back to hg19 after intersecting.

Major spliceosome complex

We characterized differences in the major spliceosome complex between archaics and modern humans by identifying missense variants in the 147 genes associated with the complex. We identified 1,746 variants that did not occur in 1KG or gnomAD or were present at low frequency in gnomAD (that is, 'archaic-specific'). We ran these variants through the Ensembl Variant Effect Predictor (VEP)[42] using the GRCh37.p13 assembly and all default options.

We repeated the above analysis on all four archaics and the randomly sampled 1KG individuals (defining SAVs) using all variants that occurred in the spliceosome genes and analysed them with VEP using the appropriate assembly.

Analysis

All data analyses were performed using Bash and Python scripts, some of which were implemented in Jupyter notebooks. We used samtools v.1.16, to index custom FASTAs[85]. We used non-parametric tests to analyse data including Fisher's exact test, Kruskal–Wallis tests, Mann–Whitney U tests and Spearman correlation, implemented with SciPy[84]. Partial correlations were run using the Pigouin package v.0.5.2 (ref. 86). Some additional metrics were calculated using custom functions. All reported P values are two-tailed, unless otherwise noted. The machine used to run analyses had a minimum value for representing floating numbers of $2.2250738585072014 \times 10^{-308}$. Therefore, we abbreviate values less than this as 2.23×10^{-308}.

Visualization
Results were visualized using Inkscape v.1.1 (ref. 87) and ggplot v.3.3.6 (ref. 88) implemented in R v.4.1.2 (ref. 89). Additional packages used to generate figures include: complex-upset v.1.3.3 (ref. 90), cowplot v.1.1.1, eulerr v.6.1.1 (ref. 91), reshape2 v.1.4.4 (ref. 92) and tidyverse v.1.3.2 (refs. 93).

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
The SpliceAI annotated archaic variant dataset is available on Dryad[94]. Source data are provided with this paper.

Code availability
The archived version of the code used to conduct analyses and generate figures has been deposited in Zenodo[95]. A non-archived version is available on GitHub (https://github.com/brandcm/Archaic_Splicing).

References

1. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222 (2012).
2. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
3. Prüfer, K. et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
4. Mafessoni, F. et al. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl Acad. Sci. USA* **117**, 15132 (2020).
5. Brand, C. M., Colbran, L. L. & Capra, J. A. Predicting archaic hominins phenotypes from genomic data. *Annu. Rev. Genomics Hum. Genet.* **23**, 591–612 (2022).
6. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
7. Castellano, S. et al. Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl Acad. Sci. USA* **111**, 6666 (2014).
8. Colbran, L. L. et al. Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* **3**, 1598–1606 (2019).
9. Gokhman, D. et al. Reconstructing Denisovan anatomy using DNA methylation maps. *Cell* **179**, 180–192 (2019).
10. McArthur, E. et al. Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.07.479462 (2022).
11. Lopez, A. J. Alternative slicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* **32**, 279–305 (1998).
12. Graveley, B. R. Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
13. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).
14. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
15. Cáceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and Involvement in human disease. *Trends Genet.* **18**, 186–193 (2002).
16. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
17. Nissim-Rafinia, M. & Kerem, B. Splicing regulation as a potential genetic modifier. *Trends Genet.* **18**, 123–127 (2002).
18. Krawczak, M., Reiss, J. & Cooper, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90**, 41–54 (1992).
19. Wang, G.-S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8**, 749–761 (2007).
20. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
21. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
22. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
23. Verta, J.-P. & Jacobs, A. The role of alternative splicing in adaptation and evolution. *Trends Ecol. Evol.* **37**, 299–308 (2022).
24. Singh, P. & Ahi, E. P. The importance of alternative splicing in adaptive evolution. *Mol. Ecol.* **31**, 1928–1938 (2022).
25. Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nat. Rev. Genet.* **23**, 697–710 (2022).
26. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).
27. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
28. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).
29. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17**, 407–421 (2016).
30. Jenkinson, G. et al. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics* **36**, 4609–4615 (2020).
31. Zhang, Y., Liu, X., MacLeod, J. & Liu, J. Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach. *BMC Genomics* **19**, 971 (2018).
32. Mertes, C. et al. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* **12**, 529 (2021).
33. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
34. Jagadeesh, K. A. et al. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat. Genet.* **51**, 755–763 (2019).
35. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
36. Danis, D. et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am. J. Hum. Genet.* **108**,1564–1577 (2021).
37. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using pangolin. *Genome Biol.* **23**, 103 (2022).
38. Collins, L. & Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**, 1053–1066 (2005).
39. Tweedie, S. et al. Genenames.Org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.* **49**, D939–D946 (2021).
40. Lowy-Gallego, E. et al. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res.* **4**, 50 (2019).
41. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
42. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

43. Aqeilan, R. I. et al. The WWOX tumor suppressor is essential for postnatal survival and normal bone metabolism. *J. Biol. Chem.* **283**, 21629–21639 (2008).
44. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA Genomic Loci Map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).
45. Rodenas-Cuadrado, P., Ho, J. & Vernes, S. C. Shining a light on CNTNAP2: complex functions to complex disorders. *Eur. J. Hum. Genet.* **22**, 171–178 (2014).
46. Rogers, A. R., Harris, N. S. & Achenbach, A. A. Neanderthal-Denisovan ancestors interbred with a distantly related hominin. *Sci. Adv.* **6**, eaay5483 (2020).
47. Vernot, B. et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
48. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61 (2018).
49. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
50. Köhler, S. et al. The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
51. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
52. Kriventseva, E. V. et al. Increase of functional diversity by alternative splicing. *Trends Genet.* **19**, 124–128 (2003).
53. Rong, S. et al. Large scale functional screen identifies genetic variants with splicing effects in modern and archaic humans. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.20.515225 (2022).
54. Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal introgression. *Proc. Natl Acad. Sci. USA* **116**, 1639 (2019).
55. Telis, N., Aguilar, R. & Harris, K. Selection against archaic hominin genetic variation in regulatory regions. *Nat. Ecol. Evol.* **4**, 1558–1566 (2020).
56. McArthur, E., Rinker, D. C. & Capra, J. A. Quantifying the contribution of Neanderthal introgression to the heritability of complex traits. *Nat. Commun.* **12**, 4481 (2021).
57. Aqil, A., Speidel, L., Pavlidis, P. & Gokcumen, O. Balancing selection on genomic deletion polymorphisms in humans. *eLife* https://doi.org/10.7554/eLife.79111 (2023).
58. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human toll-like receptors. *Am. J. Hum. Genet.* **98**, 22–33 (2016).
59. McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* **168**, 916–927 (2017).
60. Saudemont, B. et al. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol.* **18**, 208 (2017).
61. Mendez, F. L., Watkins, J. C. & Hammer, M. F. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. *Mol. Biol. Evol.* **29**, 1513–1520 (2012).
62. Sams, A. J. et al. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* **17**, 246 (2016).
63. Rinker, D. C. et al. Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations. *Nat. Ecol. Evol.* **4**, 1332–1341 (2020).
64. Huerta-Sánchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
65. Jeong, C. et al. Detecting past and ongoing natural selection among ethnically Tibetan women at high altitude in Nepal. *PLoS Genet.* **14**, e1007650 (2018).
66. Peng, Y. et al. Down-regulation of EPAS1 transcription and genetic adaptation of Tibetans to high-altitude hypoxia. *Mol. Biol. Evol.* **34**, 818–830 (2017).
67. Andrés, A. M. et al. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* **6**, e1001157 (2010).
68. Trujillo, C. A. et al. Reintroduction of the archaic variant of *NOVA1* in cortical organoids alters neurodevelopment. *Science* **371**, eaax2537 (2021).
69. Karlebach, G. et al. The impact of biological sex on alternative splicing. Preprint at *bioRxiv* https://doi.org/10.1101/490904 (2020).
70. Rogers, T. F., Palmer, D. H. & Wright, A. E. Sex-specific selection drives the evolution of alternative splicing in birds. *Mol. Biol. Evol.* **38**, 519–530 (2021).
71. Ge, Y. & Porse, B. T. The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. *BioEssays* **36**, 236–243 (2014).
72. Smith, J. E. & Baker, K. E. Nonsense-mediated RNA decay—a switch and dial for regulating gene expression. *BioEssays* **37**, 612–623 (2015).
73. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
74. Chollet, F. et al. Keras. Github https://github.com/fchollet/keras (2015).
75. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1603.04467 (2016).
76. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
77. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
78. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
79. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
80. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
81. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 Gene List enrichment analysis tool. *BMC Bioinf.* **14**, 128 (2013).
82. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis Web Server 2016 Update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
83. Xie, Z. et al. Gene set knowledge discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).
84. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
85. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
86. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).
87. *Inkscape Project* version 1.1.2 (Inkscape, 2020).
88. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).

89. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
90. Krassowski, M. ComplexUpset. Github https://github.com/krassowski/complex-upset (2020).
91. Larsson, J. eulerr: Area-proportional Euler and Venn diagrams with ellipses manual. R package version 6.1.1 (2021).
92. Wickham, H. Reshaping data with the RESHAPE package. *J. Stat. Softw.* **21**, 1–20 (2007).
93. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
94. Brand C. M. et al. Splice altering variant predictions in four archaic hominin genomes. *Dryad* https://doi.org/10.7272/Q6H993F9 (2023).
95. Brand C. M. et al. Code from: Resurrecting the alternative splicing landscape of archaic hominins using machine learning. *Zenodo* https://doi.org/10.5281/zenodo.7844032 (2023).

Acknowledgements

We thank M. L. Benton for kindly sharing data on tissue specificity and Z. Gao for helpful discussion on recurrent mutations. E. McArthur and D. Rinker provided comments that improved this manuscript. We also thank members of the Capra Lab for feedback on figures. This research greatly benefited from access to the Wynton high-performance compute cluster at the University of California, San Francisco. L.L.C. was funded by National Institutes of Health grant no. T32HG009495 to the University of Pennsylvania. J.A.C. and C.M.B. were funded by National Institutes of Health grant no. R35GM127087.

Author contributions

The work was conceived by C.M.B., L.L.C. and J.A.C. Formal analysis was undertaken by C.M.B. and L.L.C. The manuscript was drafted, reviewed and edited by all authors.

Competing interests

The authors declare no competing interests.

Additional information

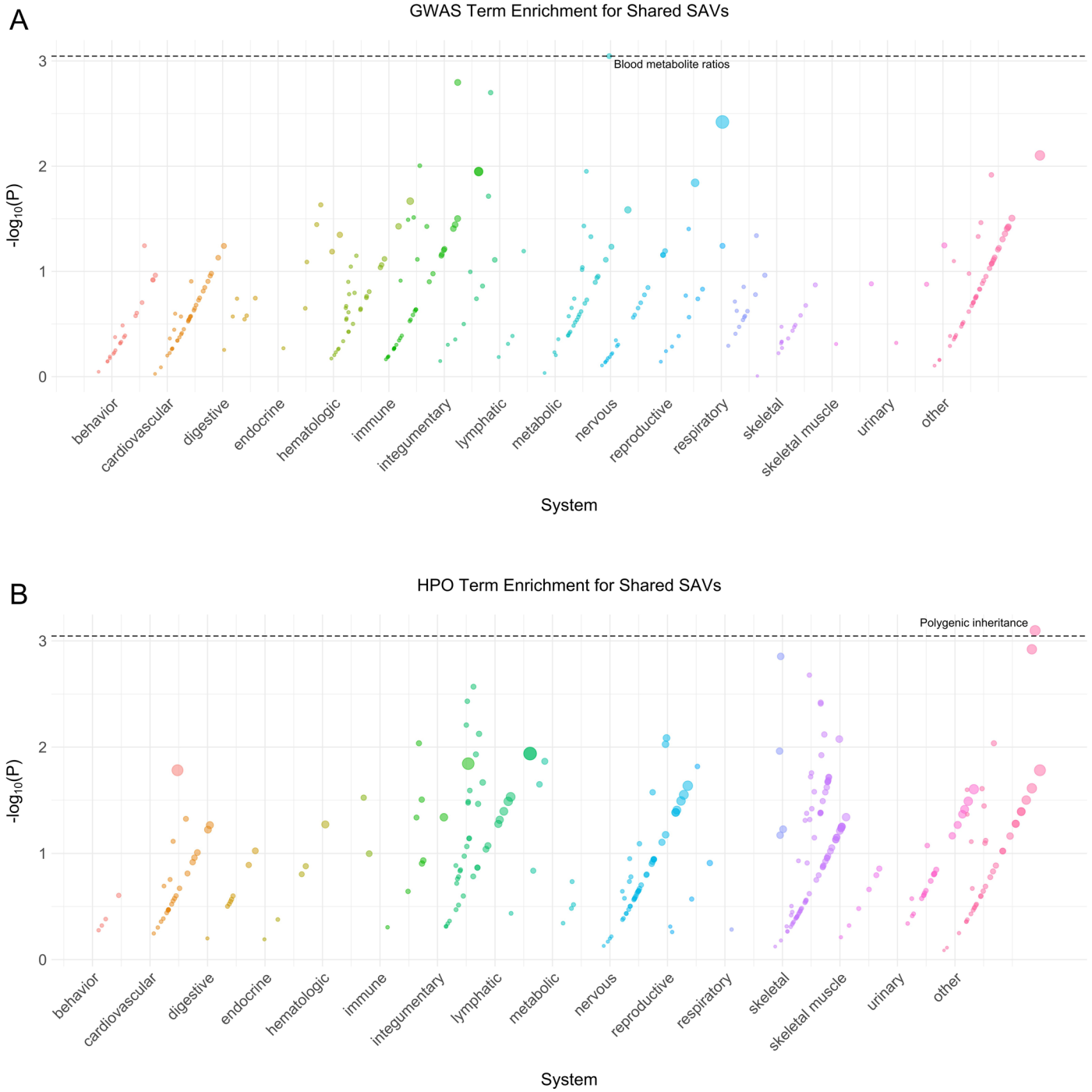
Extended data is available for this paper at https://doi.org/10.1038/s41559-023-02053-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02053-5.

Correspondence and requests for materials should be addressed to John A. Capra.

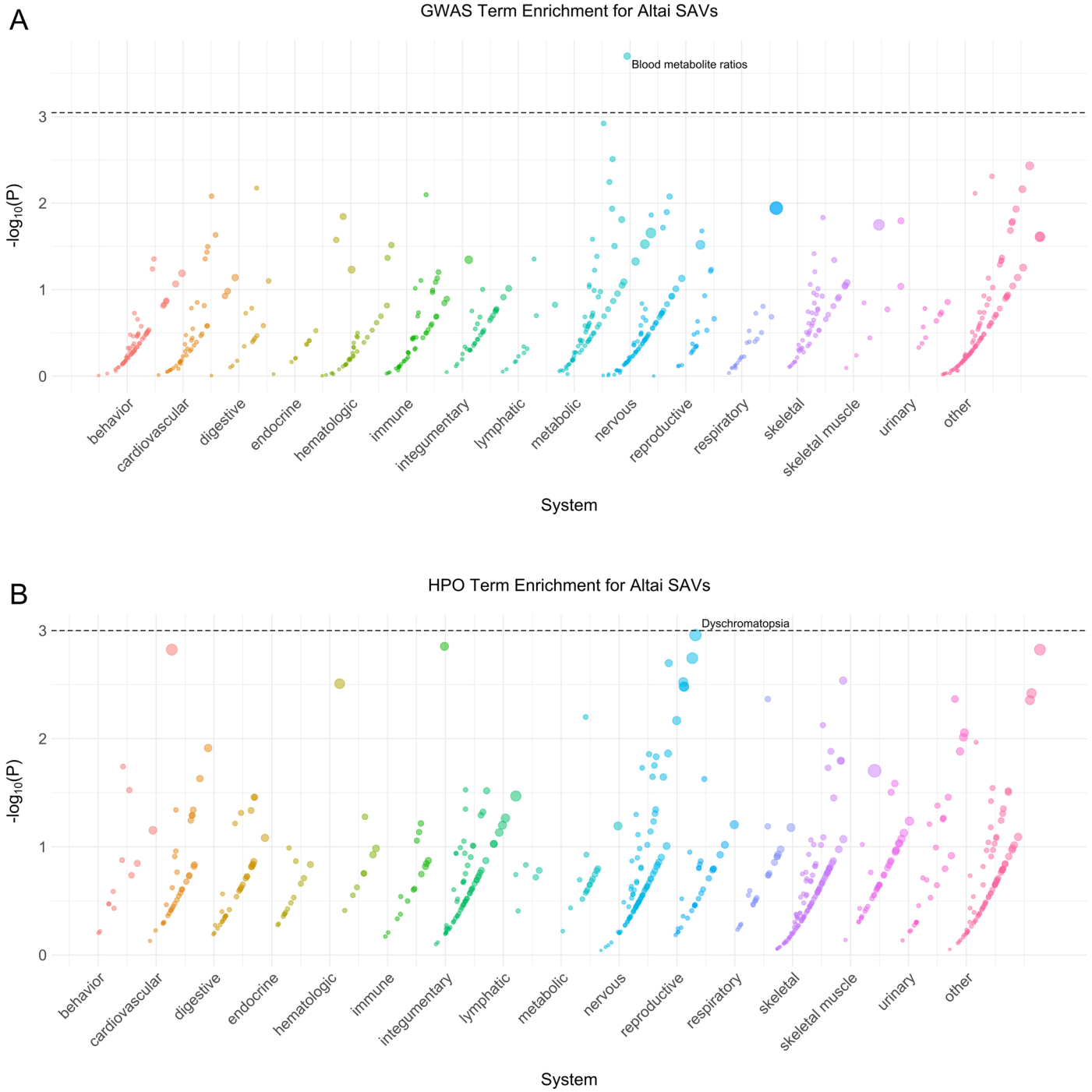
Peer review information *Nature Ecology & Evolution* thanks Maxime Rotival and Peter Robinson for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41559-023-02053-5



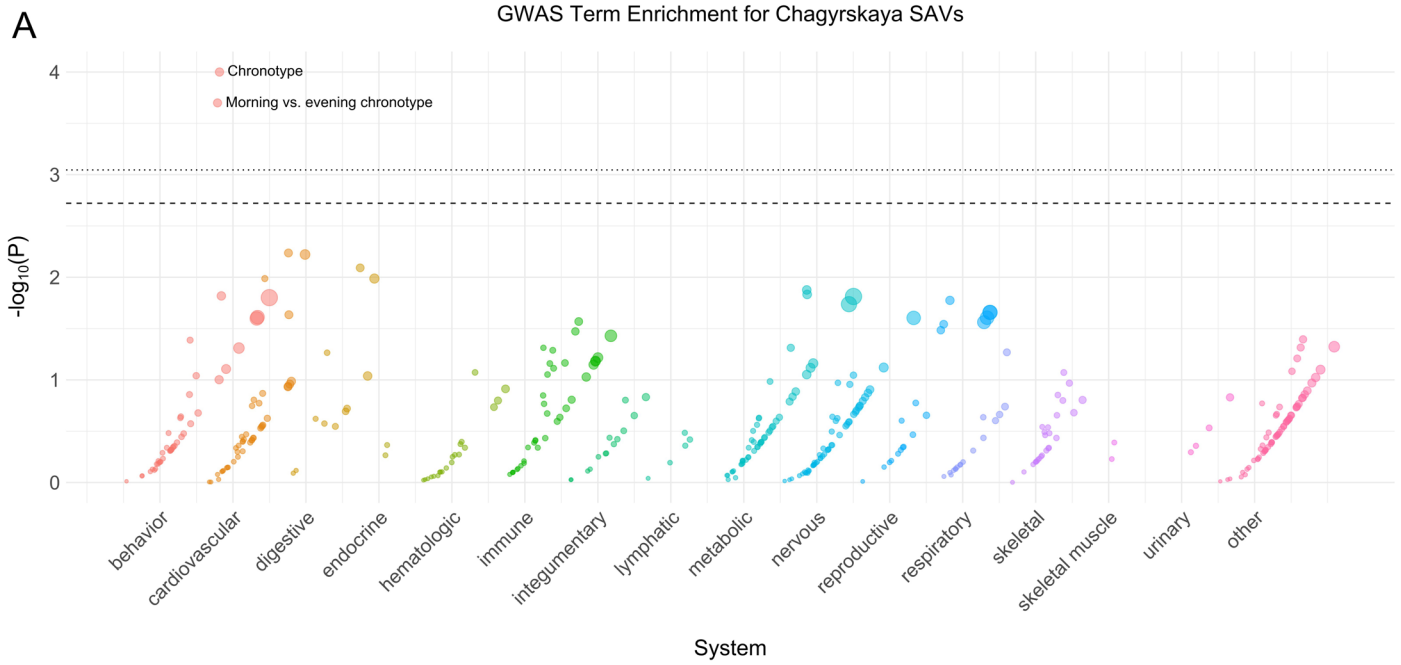
Extended Data Fig. 1 | Shared phenotype enrichment. (A) Phenotype associations enriched among genes with archaic-specific shared SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. **(B)** Phenotypes enriched among genes with archaic-specific shared SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

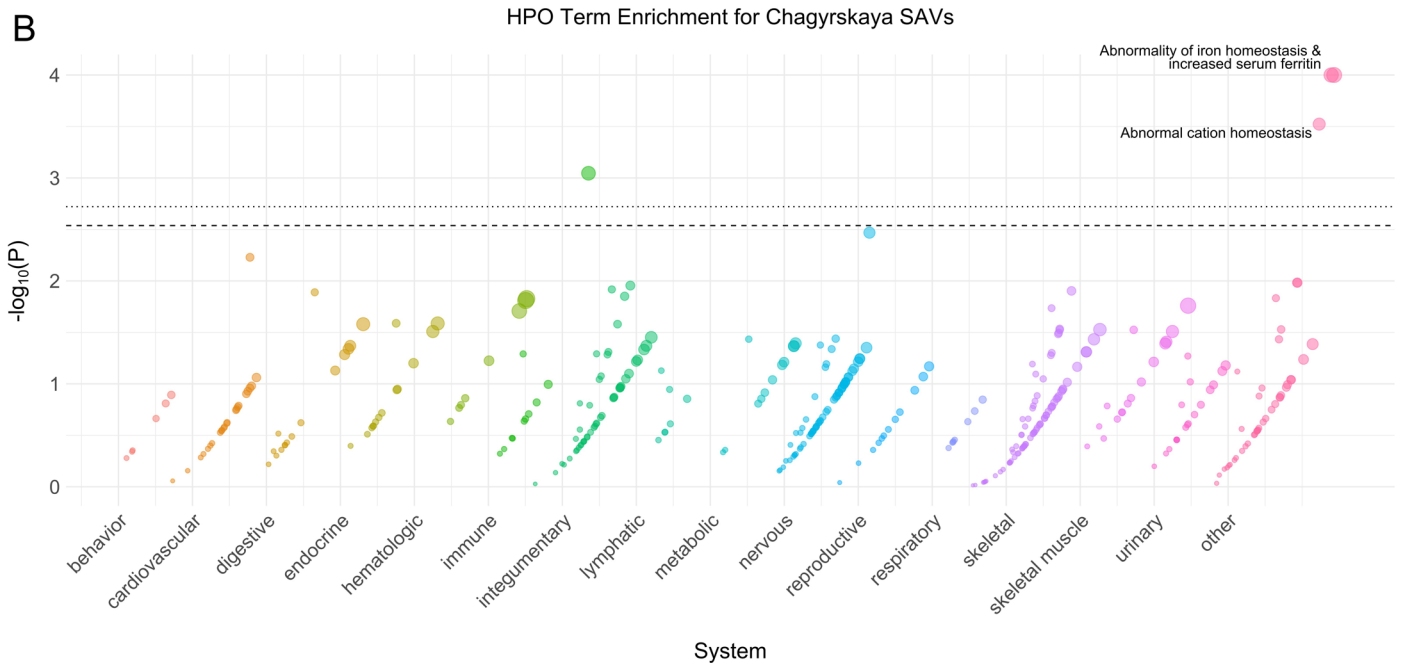
https://doi.org/10.1038/s41559-023-02053-5



Extended Data Fig. 2 | Altai phenotype enrichment. (A) Phenotype associations enriched among genes with archaic-specific Altai SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. **(B)** Phenotypes enriched among genes with archaic-specific Altai SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

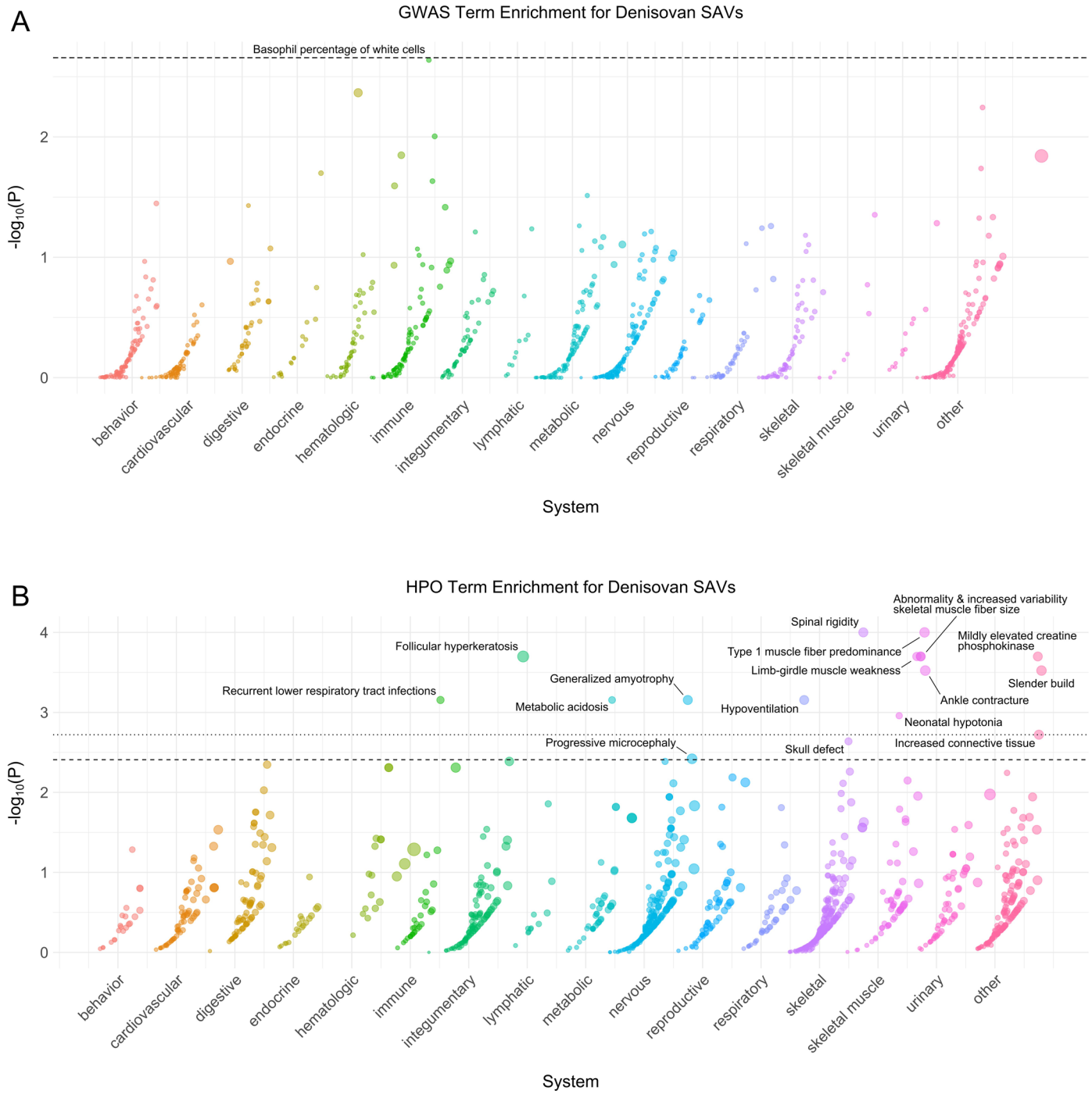
https://doi.org/10.1038/s41559-023-02053-5





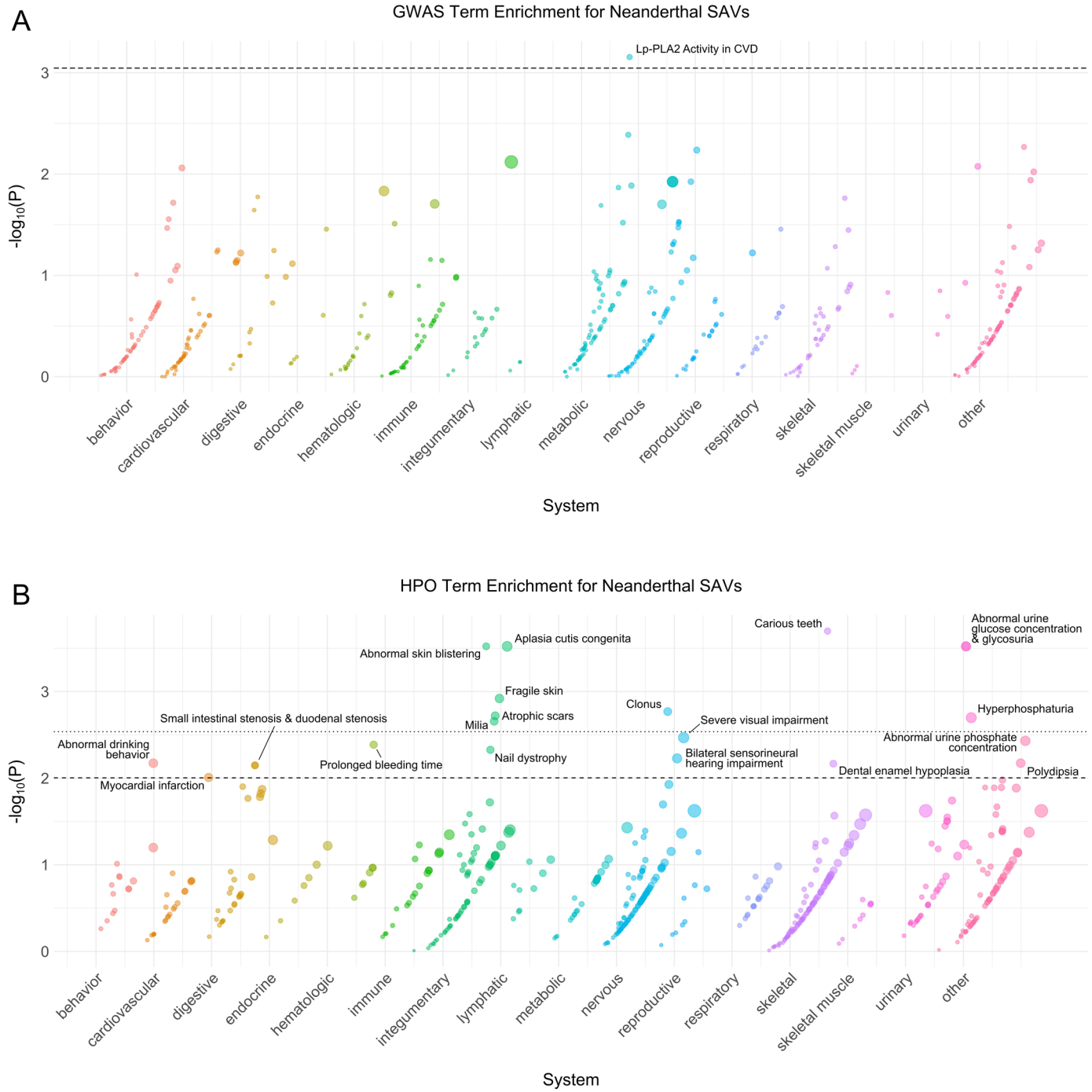
Extended Data Fig. 3 | Chagyrskaya phenotype enrichment. (A) Phenotype associations enriched among genes with archaic-specific Chagyrskaya SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. **(B)** Phenotypes enriched among genes with archaic-specific Chagyrskaya SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

https://doi.org/10.1038/s41559-023-02053-5



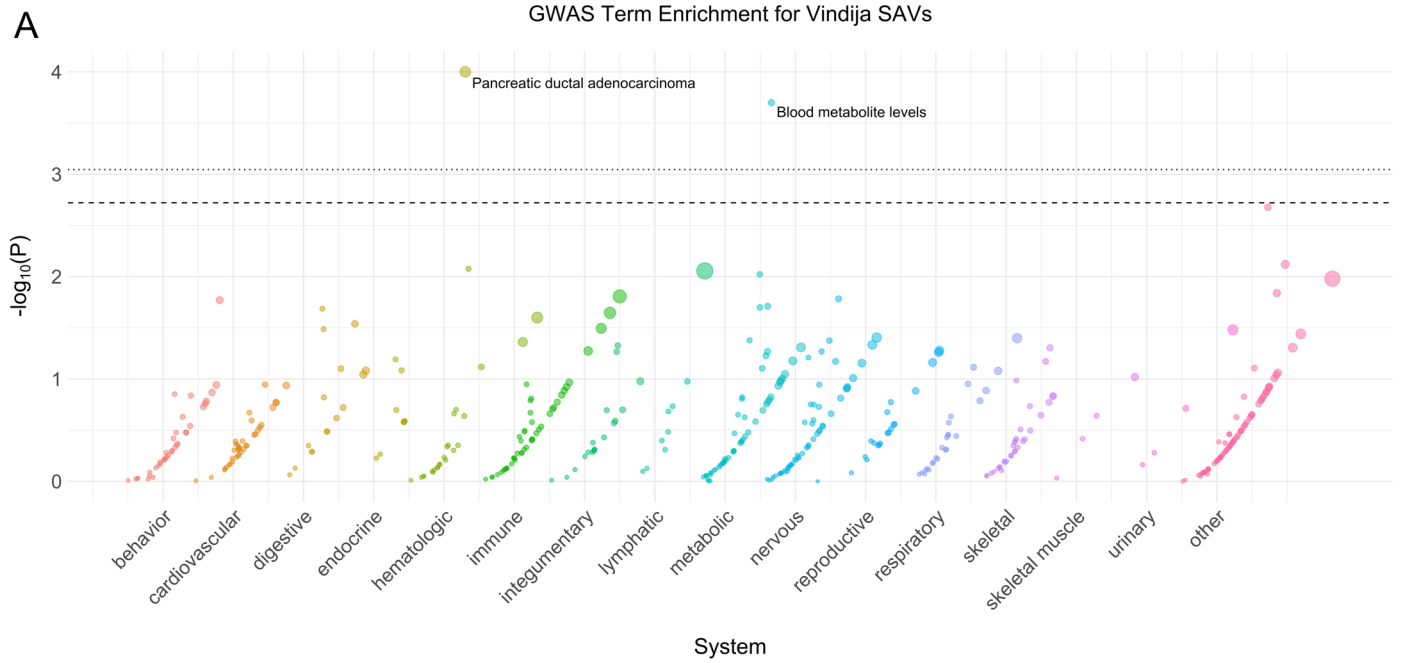
Extended Data Fig. 4 | Denisovan phenotype enrichment. (A) Phenotype associations enriched among genes with archaic-specific Denisovan SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. **(B)** Phenotypes enriched among genes with archaic-specific Denisovan SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

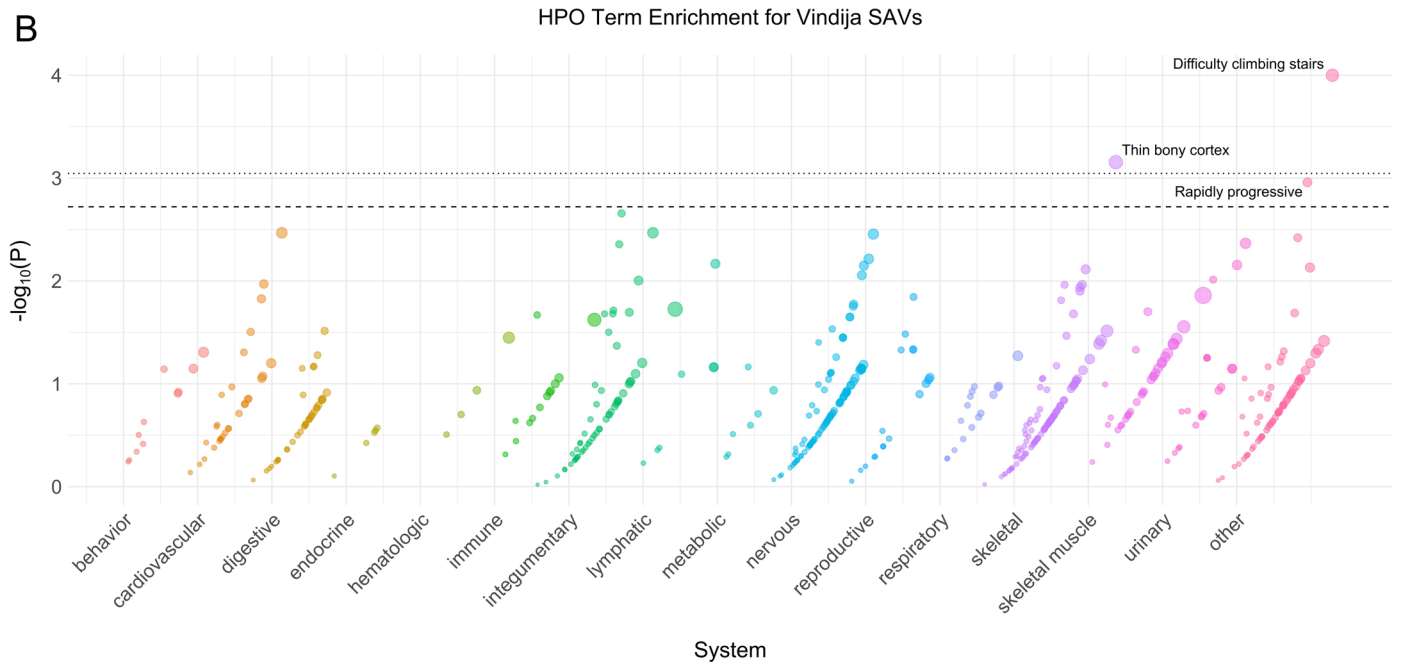
https://doi.org/10.1038/s41559-023-02053-5



Extended Data Fig. 5 | Neanderthal phenotype enrichment. (A) Phenotype associations enriched among genes with archaic-specific Neanderthal SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. CVD = cardiovascular disease, Lp-PLA2 = Lipoprotein phospholipase A2. **(B)** Phenotypes enriched among genes with archaic-specific Neanderthal SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

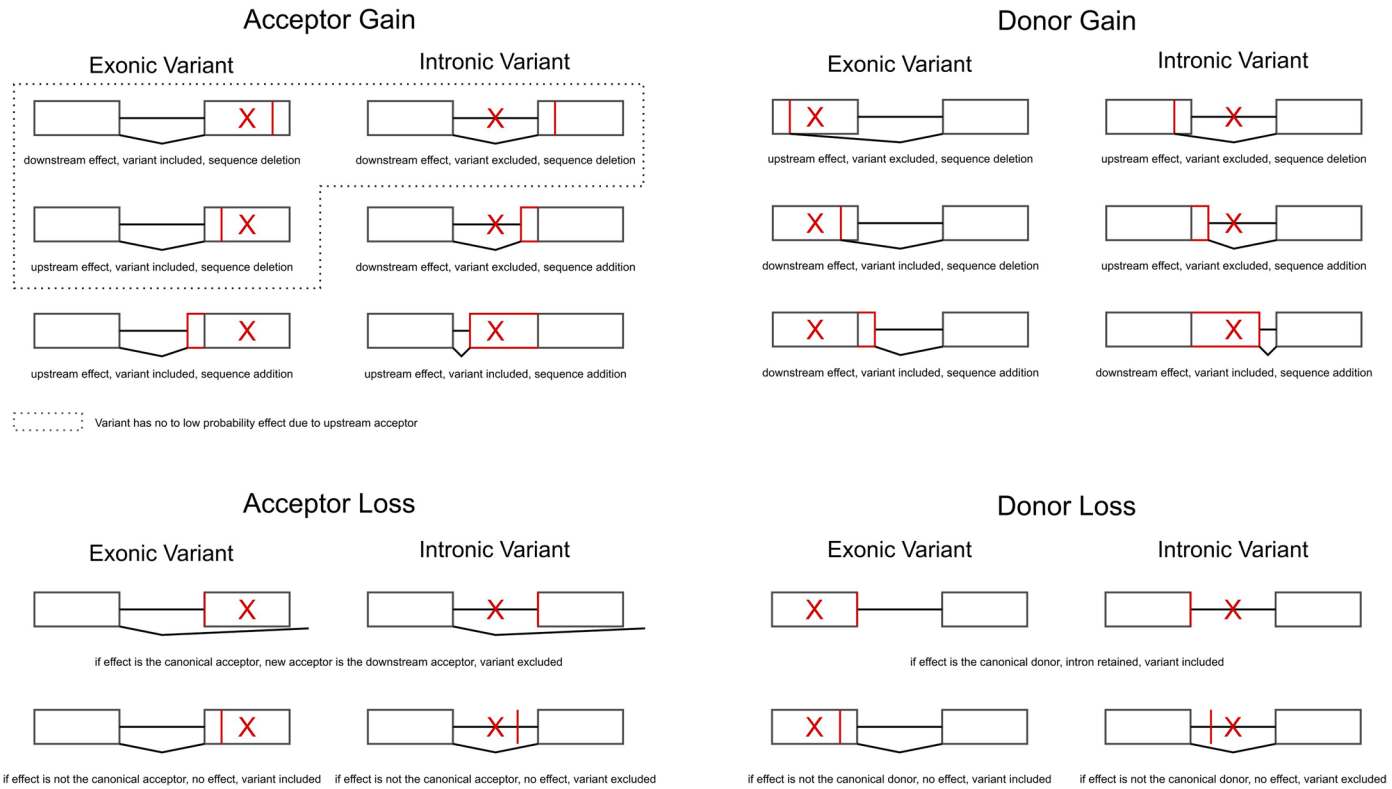
https://doi.org/10.1038/s41559-023-02053-5





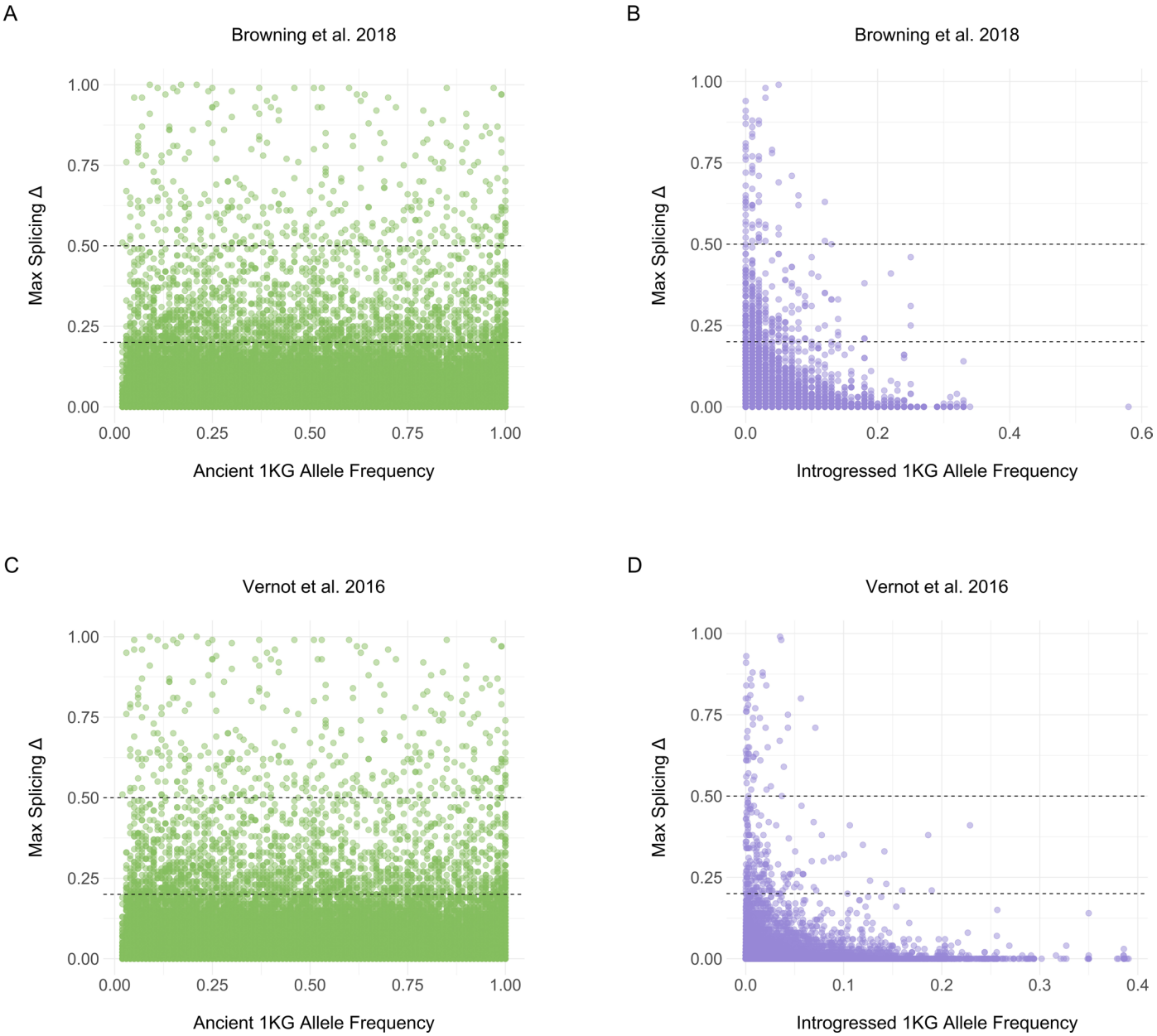
Extended Data Fig. 6 | Vindija phenotype enrichment. (A) Phenotype associations enriched among genes with archaic-specific Vindija SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. **(B)** Phenotypes enriched among genes with archaic-specific Vindija SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

https://doi.org/10.1038/s41559-023-02053-5



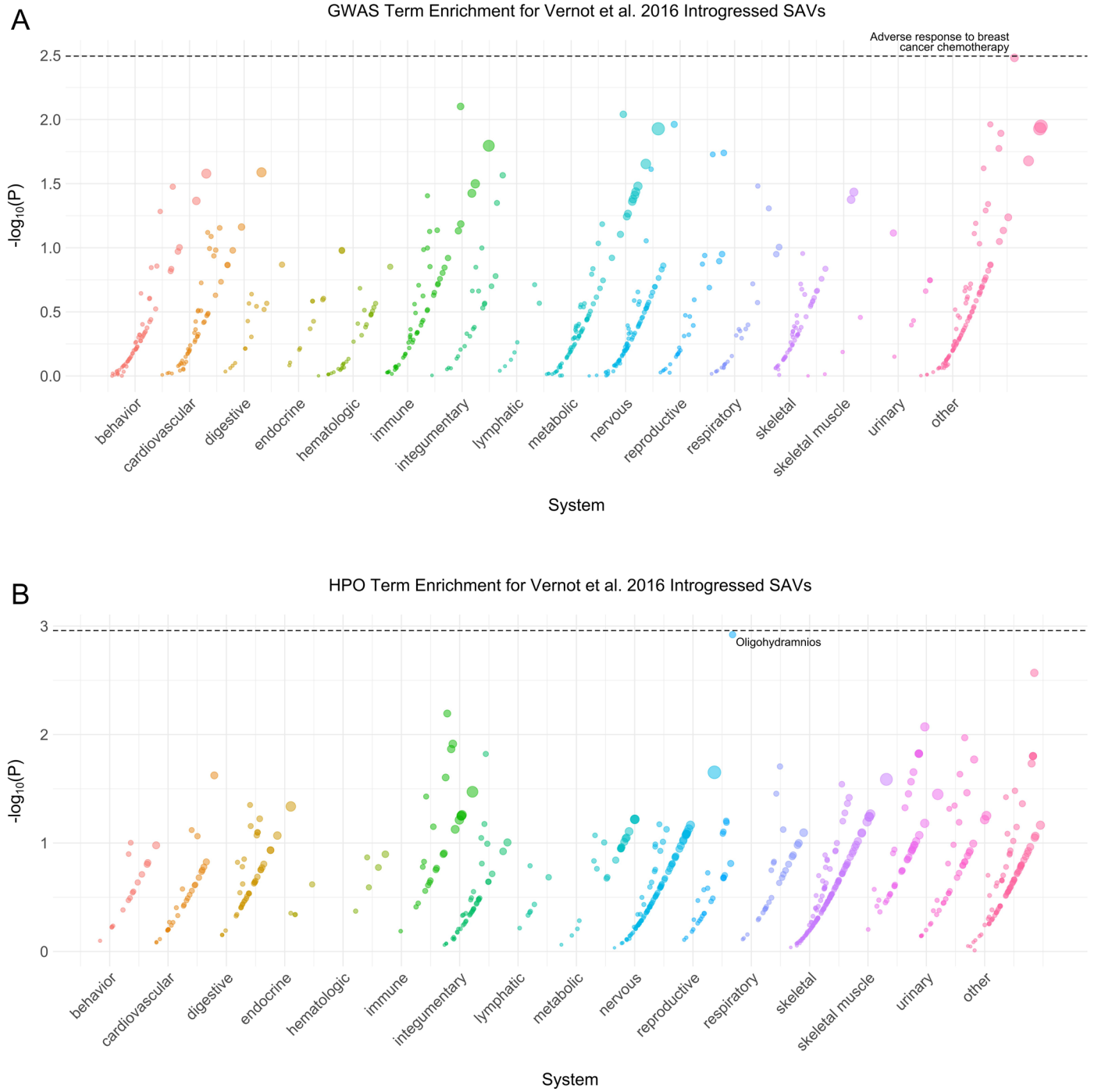
Extended Data Fig. 7 | Modelling SAV effects on the canonical transcript. We used the SpliceAI output to construct a novel transcript per SAV by modifying the canonical transcript for that gene. We considered only one effect per SAV (for example, either an acceptor gain, acceptor loss, donor gain or donor loss) based on the effect with the largest Δ. Therefore, we did not model multiple effects for a single SAV (for example, an acceptor gain and acceptor loss). Here, we illustrate all the possible consequences of a SAV for each of the four classes. We indicate the variant position with a red 'X' and the position of the effect with a red vertical line (sequence deletion) or box (sequence addition). Each scenario includes a two exon gene (boxes) with a single intron (horizontal line).

https://doi.org/10.1038/s41559-023-02053-5



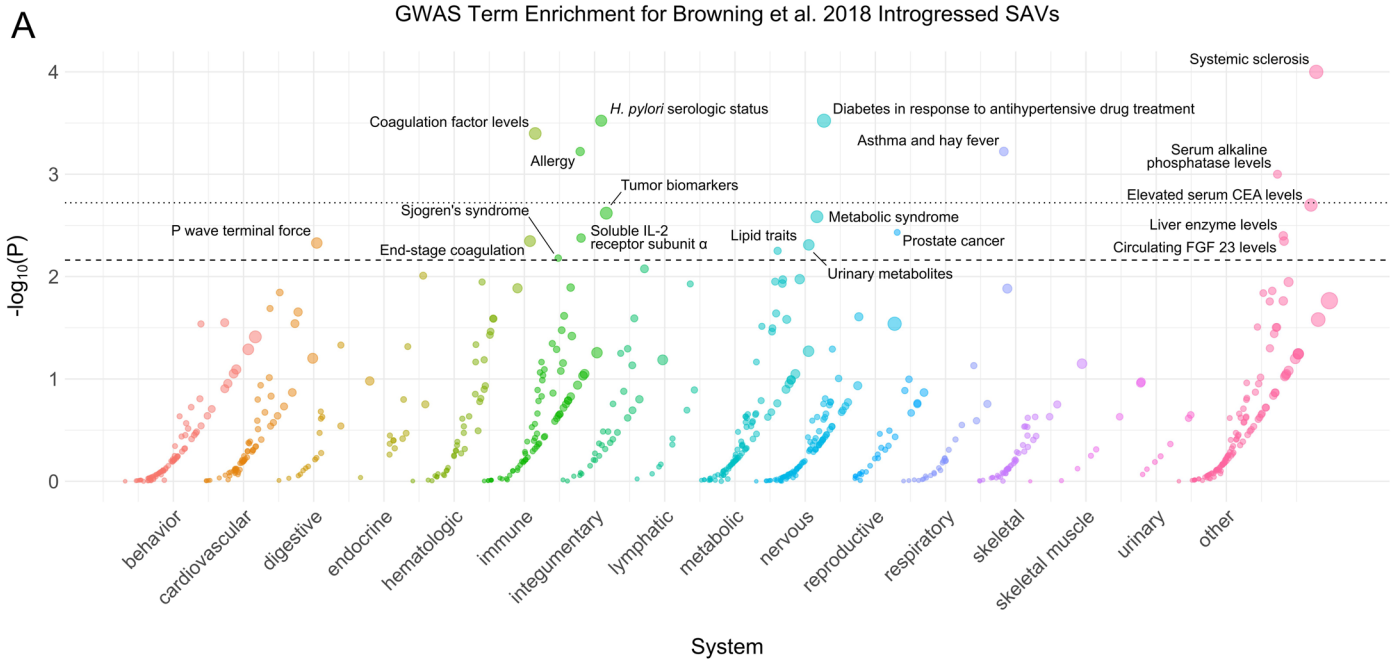
Extended Data Fig. 8 | Δ max exhibits a variable relationship to 1KG allele frequency. (A) 1KG allele frequency and Δ max for all ancient variants per[48]. Allele frequencies are from 1KG. Dashed lines reflect both Δ thresholds. **(B)** 1KG allele frequency and Δ max for all introgressed variants per[48]. Allele frequencies are from 1KG. If the introgressed allele was the reference allele, we subtracted the 1KG allele frequency from 1. **(C)** 1KG allele frequency and Δ max for all ancient variants per[47]. Allele frequencies are from 1KG. **(D)** 1KG allele frequency and Δ max for all introgressed variants per[47]. Allele frequencies represent the mean from the AFR, AMR, EAS, EUR, SAS frequencies from the[47] metadata.

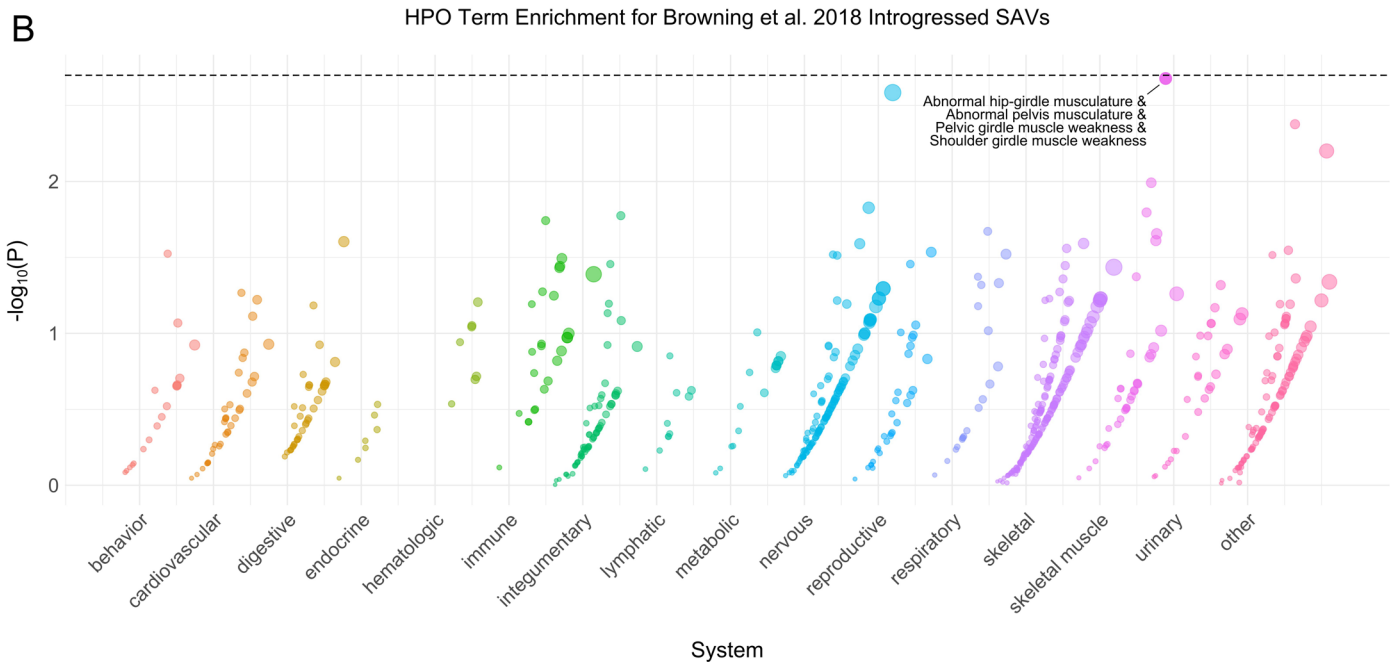
https://doi.org/10.1038/s41559-023-02053-5



Extended Data Fig. 9 | Vernot et al. 2016 introgressed phenotype enrichment. **(A)** Phenotype associations enriched among genes with[47] introgressed SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. **(B)** Phenotypes enriched among genes with[47] introgressed SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

https://doi.org/10.1038/s41559-023-02053-5





Extended Data Fig. 10 | Browning et al. 2018 introgressed phenotype enrichment. (A) Phenotype associations enriched among genes with[48] introgressed SAVs based on annotations from the 2019 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from a one-sided permutation test based on an empirical null distribution generated from 10,000 shuffles of maximum Δ across the entire dataset (Methods). Dotted and dashed lines represent false-discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. At least one example phenotype with a p-value ≤ the stricter FDR threshold (0.05) is annotated per system. CEA = carcinoembryonic antigen, FGF = fibroblast growth factor. **(B)** Phenotypes enriched among genes with[48] introgressed SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See Supplementary Data 2 for all phenotype enrichment results.

nature portfolio

Corresponding author(s): John A. Capra

Last updated by author(s): Mar 3, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used bcftools, version 1.13 to filter genomic data from archaic hominins, Thousand Genomes, and gnomAD. Splice altering variants were identified using SpliceAI, version 1.3.1.
Data analysis	All data analyses were performed using Bash and Python scripts, some of which were implemented in Jupyter notebooks.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We used publicly available data for all analyses. Archaic InDel data are from the following repository: http://ftp.eva.mpg.de/neandertal/Vindija/VCF/indels/. Archaic SNV data are from the following repositories: Altai Neanderthal (http://ftp.eva.mpg.de/neandertal/Vindija/VCF/Altai/), Chagyrskaya (http://ftp.eva.mpg.de/neandertal/Chagyrskaya/VCF/), Denisova (http://ftp.eva.mpg.de/neandertal/Vindija/VCF/Denisova/), and Vindija (http://ftp.eva.mpg.de/neandertal/Vindija/VCF/

Vindija33.19/). Modern human data are from the Thousand Genomes Project (http://hgdownload.soe.ucsc.edu/gbdb/hg38/1000Genomes/) and gnomAD (https://gnomad.broadinstitute.org/downloads#v3-variants). Introgressed tag SNPs from Vernot et al. 2016 were retrieved from: https://drive.google.com/drive/folders/0B9Pc7_zItMCVM05rUmhDc0hkWmc?resourcekey=0-zwKyJGRuooD9bWPRZ0vBzQ. Introgressed variants from Browning et al. 2018 were retrieved from: https://data.mendeley.com/datasets/y7hyt83vxr/1. gnomAD constraint data were retrieved from: https://storage.googleapis.com/gcp-public-data--gnomad/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz. phyloP data for the primate subset were retrieved from: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates.phyloP46way.bw. ASE variants were retrieved from: https://drive.google.com/file/d/10ebWfA-sboAL1SDplmIrvH-xK4x9iohV/view. TPM data were retrieved from the Human Protein Atlas (https://www.proteinatlas.org/download/rna_tissue_gtex.tsv.zip). sQTL data were retrieved from GTEx, version 8 (https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_sQTL.tar). Genes associated with the major spliceosome complex were retrieved from the HUGO Gene Nomenclature Committee (https://www.genenames.org/data/genegroup/#!/group/1518). The compiled dataset used in our analyses is available on Dryad (DOI: 10.7272/Q6H993F9).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used all available high-coverage archaic hominin genomes. We used data from the Thousand Genomes project and gnomAD to survey archaic splice altering variants in modern humans; two of the largest and geographically diverse whole genome datasets. We also used sQTLs from GTEx, which represents one of the largest and tissue diverse datasets for genetic variants linked to gene expression.
Data exclusions	We excluded some called archaic genotypes that did not pass our quality control thresholds.
Replication	We confirmed that patterns among splice altering variants broadly agreed among all four archaic individuals.
Randomization	Sample randomization was not relevant to this study because such methods were not needed to address the research questions.
Blinding	Blinding was not relevant to this study because we made no potentially subjective conclusions about specific samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging