HHS Public Access

Author manuscript

Nat Ecol Evol. Author manuscript; available in PMC 2024 September 30.

Published in final edited form as:

Nat Ecol Evol. 2023 June ; 7(6): 939–953. doi:10.1038/s41559-023-02053-5.

Resurrecting the Alternative Splicing Landscape of Archaic Hominins using Machine Learning

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

[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

Abstract

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.

Introduction

While the paleontological 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 anatomically modern humans. These data also

*Correspondence to tony@capralab.org.

Author Contributions

Conceptualization, C.M.B, L.L.C, and J.A.C.; Formal Analysis, C.M.B and L.L.C; Writing – Original Draft, C.M.B, L.L.C, and J.A.C.; Writing – Review & Editing, C.M.B, L.L.C, and J.A.C.

Competing Interests

The authors declare no competing interests.

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 3D 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, including development and establishing tissue identity [14]. Defects in splicing underlie many human diseases (e.g., [15-22]), and variation in splicing contributes to differences in traits in non-human species (see [23], Table 1). Further, alternative splicing can evolve rapidly and respond to environmental factors—suggesting 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-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 basepairs.

Alternative splicing occurs across nearly all eukaryotes and its molecular mechanisms are deeply conserved [41]. 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 [42] for "archaic-specific" variants, i.e., archaic variants absent or at very low allele frequency (< 0.0001) from individuals sequenced by the 1000 Genomes Project (1KG) [43] and the Genome Aggregation Database (gnomAD) [44] (Supplementary Data 1). We annotated these variants using the Ensembl Variant Effect Predictor [45]. 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 splice altering variants (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 splice altering probability (SAP; SpliceAI Δ ≥ 0.5) out of 1,607,350 archaic variants we analyzed. 5,950 archaic variants had moderate SAP (Δ ≥ 0.2). Hereafter, we refer to these variants as high-confidence splice altering variants (SAVs) and SAVs, respectively, and 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; 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; Supplementary Fig. 1).

A total of 4,242 genes have at least one archaic SAV. 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 [46], *HLA-DPA1* (n = 7) and *HLA-DPB1* (n = 10), essential components of the immune system [47], and *CNTNAP2* (n = 11), a nervous system protein associated with

neurodevelopmental disorders that is also one of the longest genes in the human genome [48].

Many SAVs (47.8%) have a high SAP for only one of the four classes 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 ($\rho = -0.34 – -0.14$) for variants with at least one SAP greater than 0 (Supplementary Fig. 3). Donor gains were the most frequent result of splice altering variants 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.

37% of archaic SAVs are archaic-specific.

We inferred the origin of archaic variants based on 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 approximately 570–752 ka [40]. SAVs present in both archaic and modern humans may be the result of either 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 [38, 39]. While modern human genomes retain Denisovan and Neanderthal ancestry, most 1KG samples have minimal (< 1%) Denisovan ancestry [38, 39]. Therefore, we focused on Neanderthal introgression and classified 244 SAVs identified by [38] in 239 genes (Fig. 1D; Supplementary Fig. 4) and 386 SAVs identified by [39] 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 [38] introgressed variants in the main text and include results using the [39] set in the supplemental text.

Non-introgressed variants present in both archaic and modern humans likely evolved prior to 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 based on [38] among 1,896 genes (Fig. 1D; Supplementary Fig. 4) and 2,195 variants based on [39] 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

[10], we considered links between genes and phenotypes from two sources: GWAS Catalog 2019 [49] and the Human Phenotype Ontology [50], that capture annotations based on common and rare disease, respectively. Structural properties of genes, such as the number of exons, influence the probability that they have SAVs (Supplementary Table 6; 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; 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 Human Phenotype Ontology, 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- specific, 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 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 fiber size, increased muscle fiber diameter variation, and type 1 muscle fiber 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 mRNA 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 novel transcript using the variant, indicated splicing alteration class (e.g., acceptor gain), and Δ position for that alteration (Extended Data Fig. 7). If multiple alteration classes passed our SAP threshold, we modeled 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, the majority (60%) of 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 (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 and introgressed SAVs were more likely to alter the canonical protein or UTRs (G Test of Independence, G = 138, P = 1.73×10^{-23}). This pattern was also observed when variant allele origin was classified per [39] (G Test of Independence, G = 121, P = 3.45×10^{-20}) (Supplementary Table 7). We also note that many SAVs are predicted to result in multiple splicing alteration classes (e.g., 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 [44] 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 splicing altering variant 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 per [39] exhibited the same pattern (Supplementary Figs. 7C, 7D). 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 = 6.963×10^{-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 (|phyloP| <= 1.3). Variants classified per [39] 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 likely exposed to different strengths of selection for different periods. Thus, we hypothesized that the probability a given SAV would survive to the present would vary based on its origin. For example, SAVs that arose in the ancestor all archaic lineages were likely 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 splice altering probability thresholds (Fig. 4A, Fig. 4B, Supplementary Table 8). This result is consistent with the hypothesis 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 with concordant with the site-level constraint results.

Given that the population histories for each archaic lineage were likely 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 hypothesis 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 (OR = 1.205-1.447, Fig. 4C, 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 odds ratios (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 hypothesized 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 high-confidence and SAVs from [39] (Supplementary Fig. 10). These patterns suggest that older SAVs, either those that evolved prior to the Neanderthal common ancestor or prior to 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 small number of introgressed SAVs occur at modest to high frequencies among genes including *GMEB2*, *GALNT18*, and *TLR1* (Fig. 5B). The latter occurs in an adaptively introgressed locus spanning three toll-like receptors—key components of the innate immune system [58] and this SAV has 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, 12), and their frequencies are significantly higher among Africans ($\mu = 0.522$) than non-Africans ($\mu = 0.476$) (Mann-Whitney U, U = 10,963,956, P = 2.66×10^{-9}) (Fig. 5C, 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, Supplementary Figs. 14, 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 splice altering probability (Δ max) for introgressed SAVs. 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 likely absent among ancient variants due to purifying selection on deleterious variants that occurred prior to 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 [39] and 239 genes with introgressed SAVs

from [38] (Supplementary Data 2). Two terms were significantly enriched among genes with introgressed SAVs [38]: 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 [39] introgressed SAVs (Extended Data Fig. 10B). However, 19 GWAS terms met our FDR-corrected significance threshold including *H. 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 [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 hypothesized that introgressed SAVs would be more tissue-specific than ancient SAVs in their effects. To explore this, we identified 1,381 archaic SAVs with sQTL data from GTEx across 49 tissues.

Introgressed sQTL SAVs were significantly associated with tissue specific gene expression compared to ancient sQTL SAVs (Mann-Whitney U, U = 35,123.5, P = 0.027) (Fig. 5E). On average, introgressed SAVs influenced splicing in 4.92 fewer tissues than ancient SAVs. Further, all sQTL SAVs with broad effects (> 40 tissues) were ancient (107 high-confidence and 5 low-confidence). 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 [39] (Supplementary Fig. 17). Collectively, 30% of sQTL SAVs (n = 427) were associated with tissue-specific effects on splicing (1 or 2 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, [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 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 based on the relative entropy.

Genes with the most tissue-specific expression had significantly higher median maximum splice altering probability (Δ) than 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 hypothesized significance to the evolutionary divergence between archaic hominins and modern humans. For example, the 2'-5' oligoadenylate synthetase *OAS* locus harbors an adaptively introgressed splice altering variant at chr12: 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 harbors 92 additional archaic variants (n = 92). We found one additional SAV at chr12: 113,355,275 in *OAS1* that potentially results in an acceptor gain (Δ = 0.26). This allele was unique to the Denisovan, 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 (chr2: 46,584,859; rs372272284) is fixed 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 hemoglobin 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 (chr2: 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 (chr5: 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) (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 most likely why it remains uncommon. The third variant (chr5: 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 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 splice-altering variants 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 splice-altering variants should not necessarily be detected. Indeed, we observe higher fractions of SAVs as sQTL when analyzing 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 (i.e., 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 SNPs and modern human samples used in this analysis are best suited to identifying Neanderthal rather than Denisovan introgression [38, 39]. 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 likely cause in 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. [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 based on 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, version 1.13 [73].

All genomic coordinates presented in the manuscript and supplementary material refer to hg19/GRCh37.

Variant Annotation

We annotated variants for putative alternative splicing using SpliceAI, version 1.3.1 [35]. Briefly, SpliceAI uses a deep residual neural network to estimate the splice altering probability 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 kbp flanking the variant in both directions. The output includes four splice altering probabilities (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. 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, version 2.3.1 [74] and tensorflow, version 1.15.0 [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 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 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 analyzed the variants as for the archaics (i.e., 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 [43] and SNVs from gnomAD, version 3 [44]. 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 VCFs. Normalization, filtering, and querying was 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 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 SW USA (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: [38] and [39]. These datasets differ in their approach to recognizing introgressed sequences and partly overlap the archaic variants considered in this study. [38] 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 ≥ 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 [38]. 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. [39] developed a modified S* statistic, Sprime, which employs 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 analyze 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 [38] 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 [38] 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 [39] 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 [38] 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 [39] and [38] using bedtools2, version 2.30 [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" based on presence in 1KG and gnomAD as well as whether or not the variant was introgressed. Further, we classified each variant's allele origin based on introgressed variants identified by [38] "Vernot allele origin" and [39] "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 CpGs were enriched among rare variants (allele frequency < 0.0001) vs 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-confident 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 [76], to count the number of exons per gene and calculate the length in bp 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 [44] 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 multi-species 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 [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 Human Phenotype Ontology (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: behavioral, cardiovascular, digestive, endocrine, hematologic, 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 [38] (n = 239), and 9) genes with introgressed SAVs per [39] (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 Browning et al. 2018 introgressed, 412 Chagyrskaya, 1,023 Denisovan, 515 Neanderthal, 295 shared, 627 Vernot et al. 2016 introgressed, and 474 Vindija terms for the 2019 GWAS Catalog and 622 Altai, 1,490 archaic-specific, 720 Browning et al. 2018 introgressed, 391 Chagyrskaya, 1,152 Denisovan, 528 Neanderthal, 306 shared, 651 Vernot et al. 2016 introgressed, and 522 Vindija terms for the Human Phenotype Ontology.

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. p-values were calculated as the proportion of empiric counts + 1 ≥ the observed counts + 1. We adjusted our significance level due to multiple testing by correcting for the false discovery rate (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.

Novel Transcripts and Proteins

We constructed a novel 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 novel transcript using the splicing alteration class (acceptor gain, acceptor loss, donor gain, or donor loss) and associated position information per SAV. For SAVs with multiple splicing class alterations, e.g., a variant that results in both an acceptor gain and an acceptor loss, we modeled 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 GFF, prioritizing the longest and experimentally-validated (Ensembl) vs. computationally-predicted (HAVANA) start sites when there were multiple start positions.

We compared the canonical and novel transcripts and the resulting protein (Supplementary Data 3). We assigned each SAV as resulting in one of seven effects based on transcript/protein length and composition: 1) the canonical and novel transcripts/proteins are identical (“no effect”), 2) the novel protein is longer than the canonical and includes premature termination codons (“longer w/ PTCs”), 3) the novel protein is longer and does not include premature termination codons (“longer w/o PTCs”), 4) the 5’ or 3’ untranslated region (UTR) is longer than the canonical transcript, 5) the novel protein is the same length as the canonical but contains a single missense variant (“single missense”), 6) the novel protein is shorter than canonical protein (“truncated protein”), and 7) the 5’ or 3’ UTR is truncated than the canonical transcript (“truncated UTR”).

Gene Expression, Tissue Specificity, and sQTL

We used TPM counts for each gene from GTEx, version 8, to analyze 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 relatively entropy to this background distribution. The 34 tissues were selected based on groupings from the Human Protein Atlas to minimize the amount of sharing between distinct tissues, e.g., since brain tissues are over-represented 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]. Based on 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, version 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 nor gnomAD or were present at low frequency in gnomAD (i.e., "archaic-specific"). We ran these variants through the Ensembl Variant Effect Predictor (VEP) [45] 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 analyzed them with VEP using the appropriate assembly.

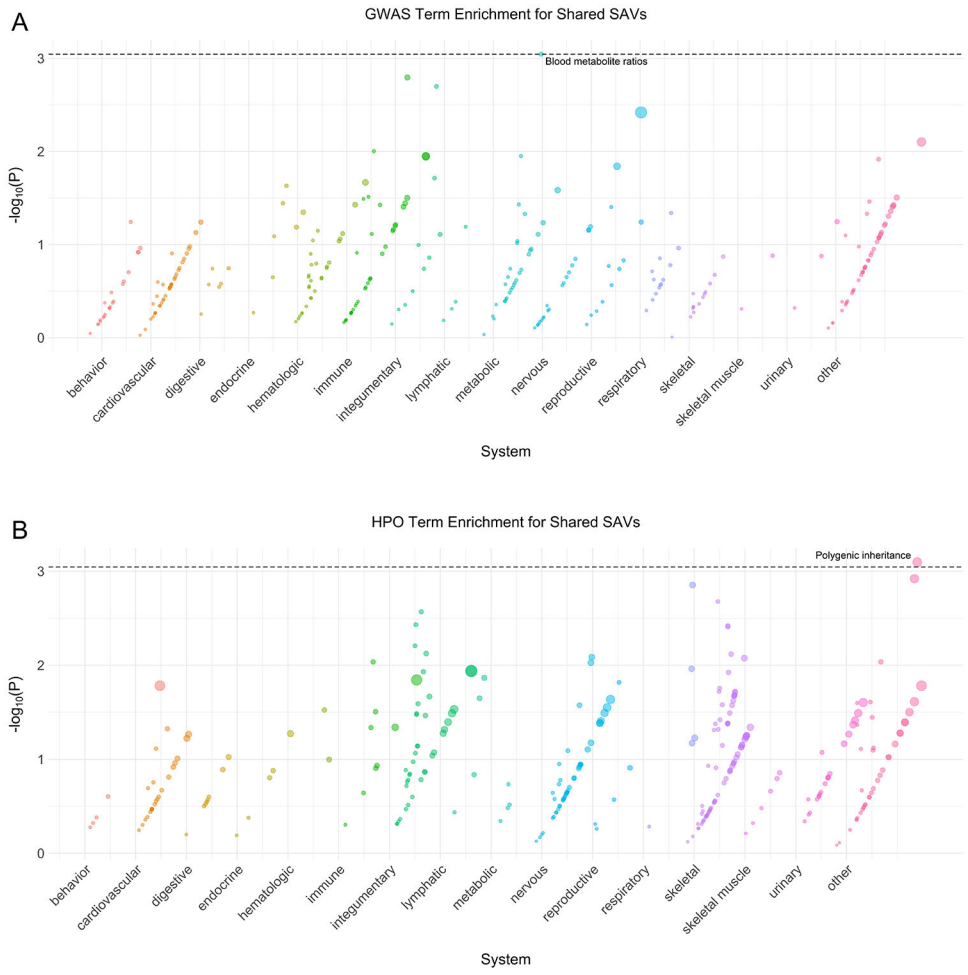
Analysis

All data analyses were performed using Bash and Python scripts, some of which were implemented in Jupyter notebooks. We used samtools, version 1.16, to index custom FASTAs [85]. We employed non-parametric tests to analyze 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, version 0.5.2 [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 2.2250738585072014e-308. Therefore, we abbreviate values less than this as 2.23e-308.

Visualization

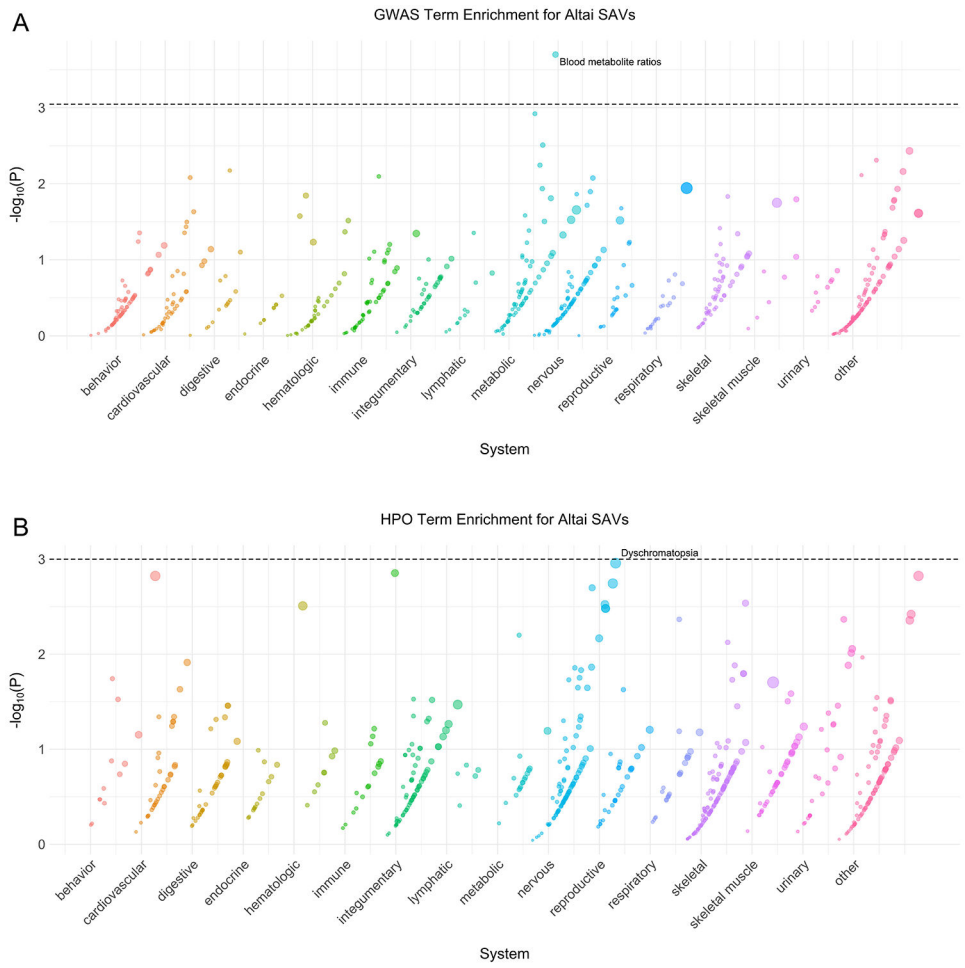
Results were visualized using Inkscape, version 1.1 [87] and ggplot, version 3.3.6 [88] implemented in R, version 4.1.2 [89]. Additional packages used to generate figures include complex-upset, version 1.3.3 [90]; cowplot, version 1.1.1; eulerr, version 6.1.1 [91]; reshape2, version 1.4.4 [92]; and tidyverse, version 1.3.2 [93].

Extended Data



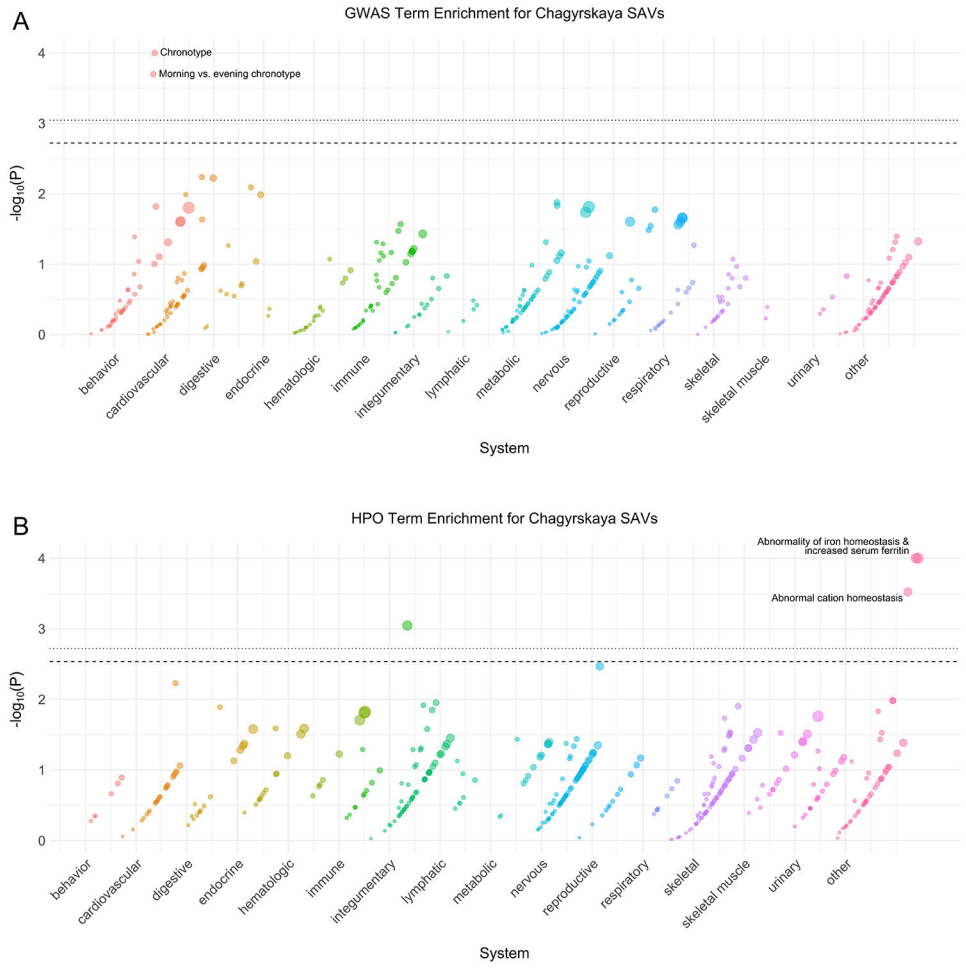
Extended Data Fig. 1: Shared phenotype enrichment.

(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.



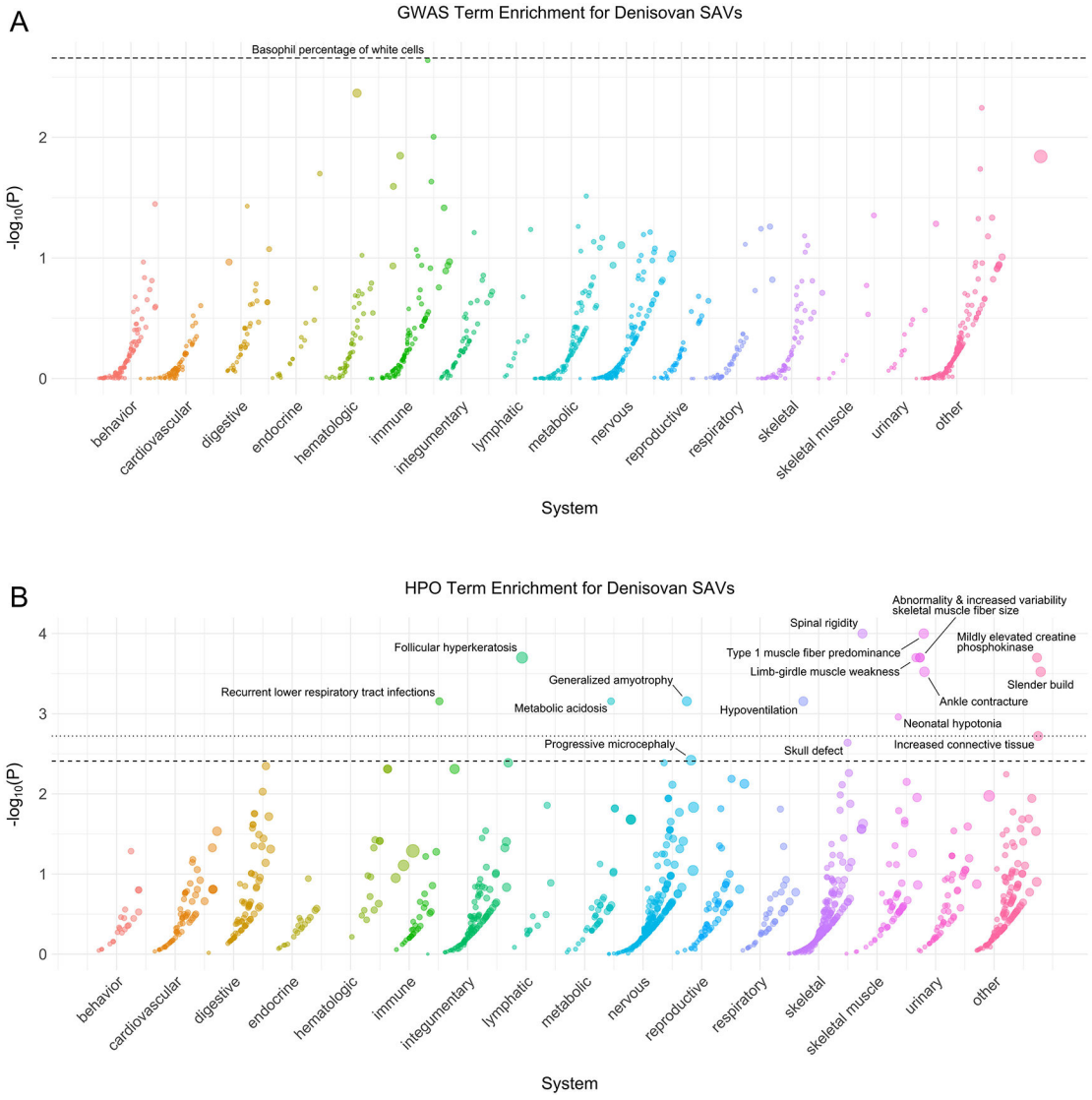
Extended Data Fig. 2: Altai phenotype enrichment.

(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.



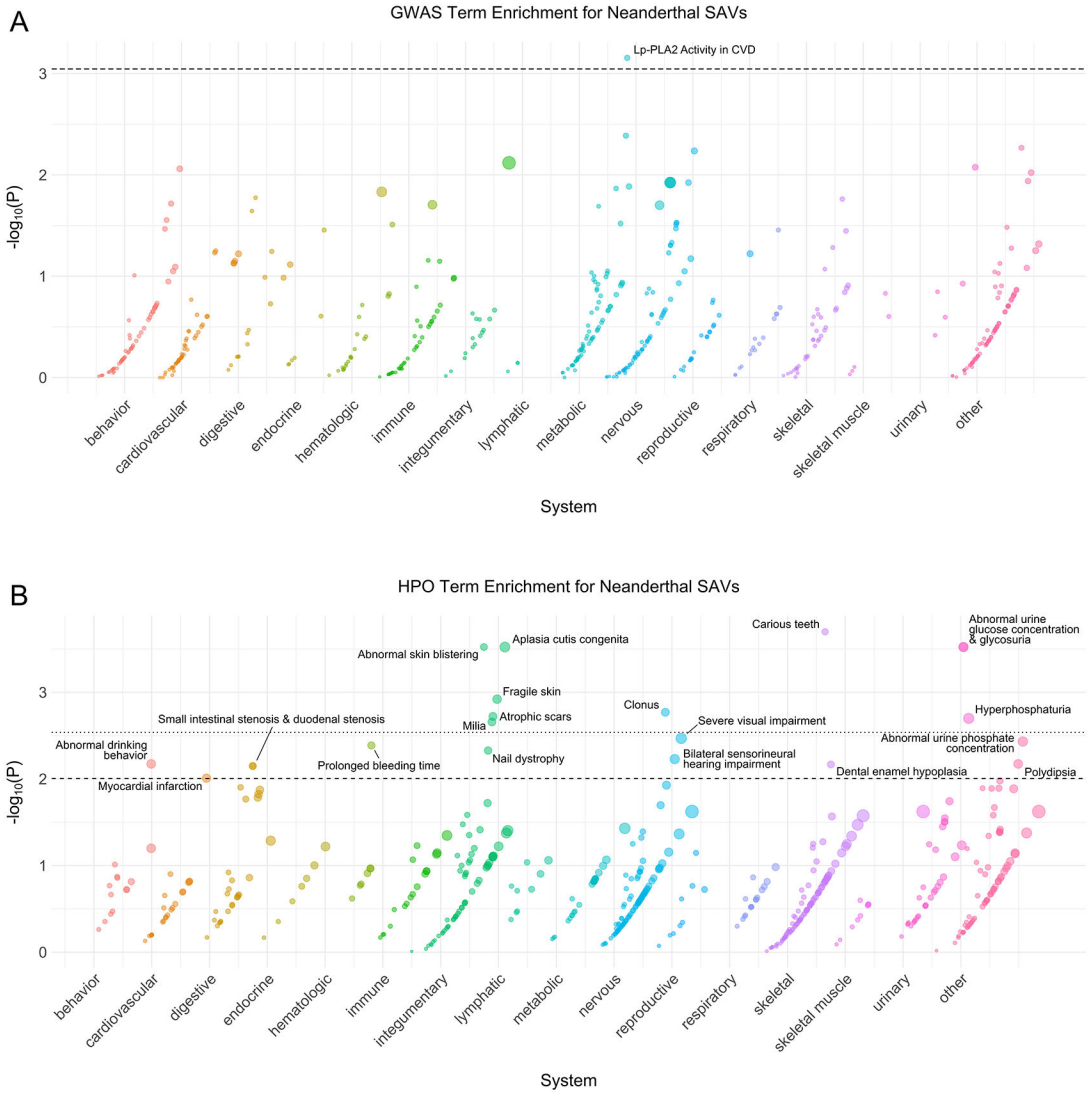
Extended Data Fig. 3: Chagyrskaya phenotype enrichment.

(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.



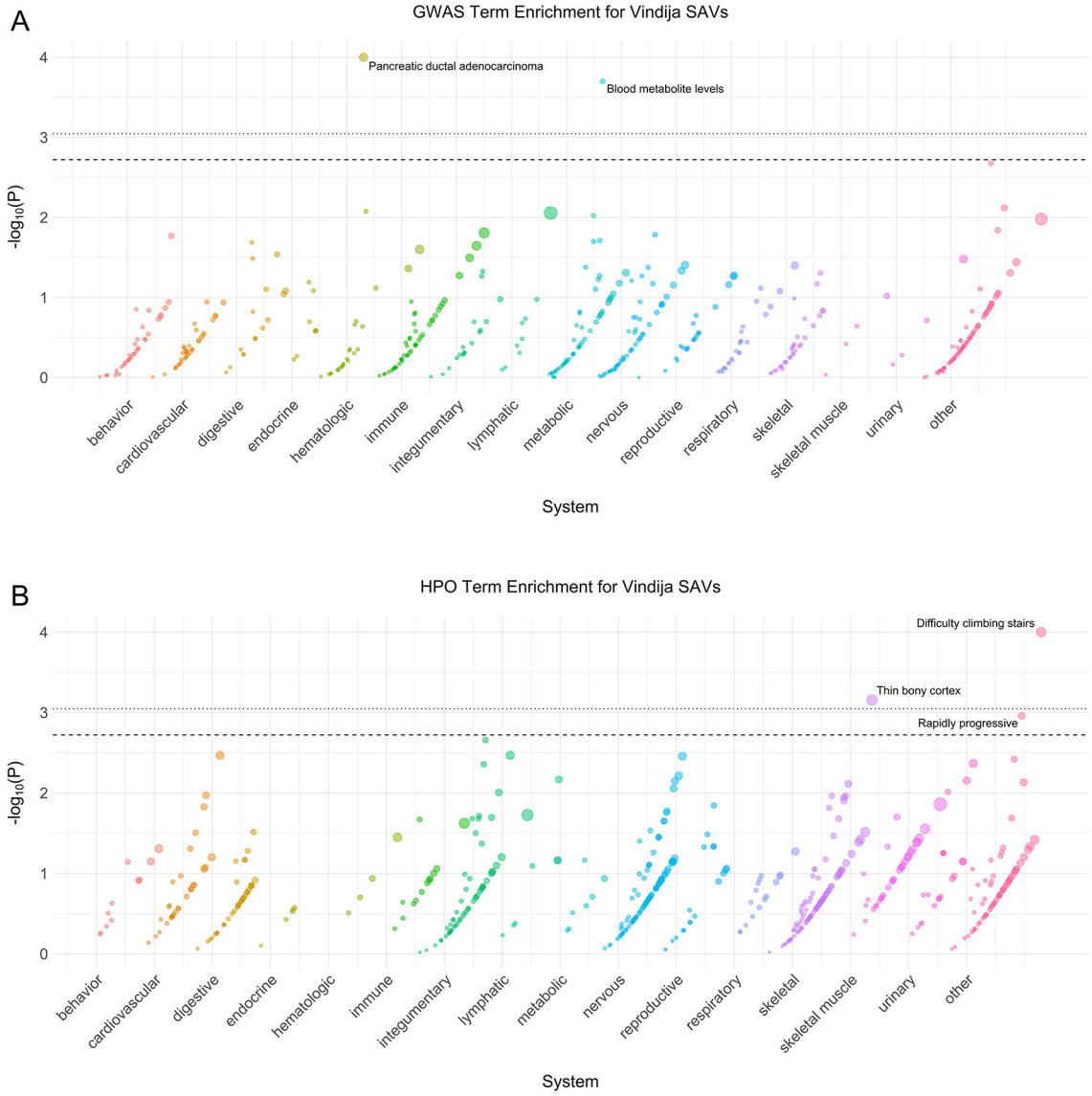
Extended Data Fig. 4: Denisovan phenotype enrichment.

(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.



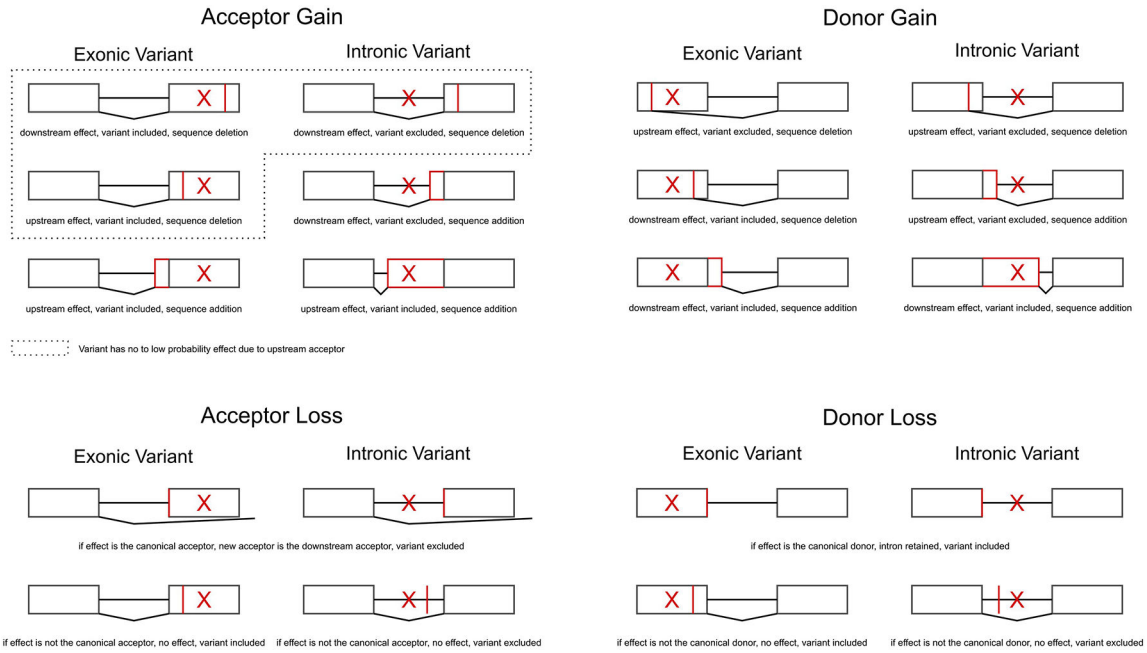
Extended Data Fig. 5: Neanderthal phenotype enrichment.

(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.



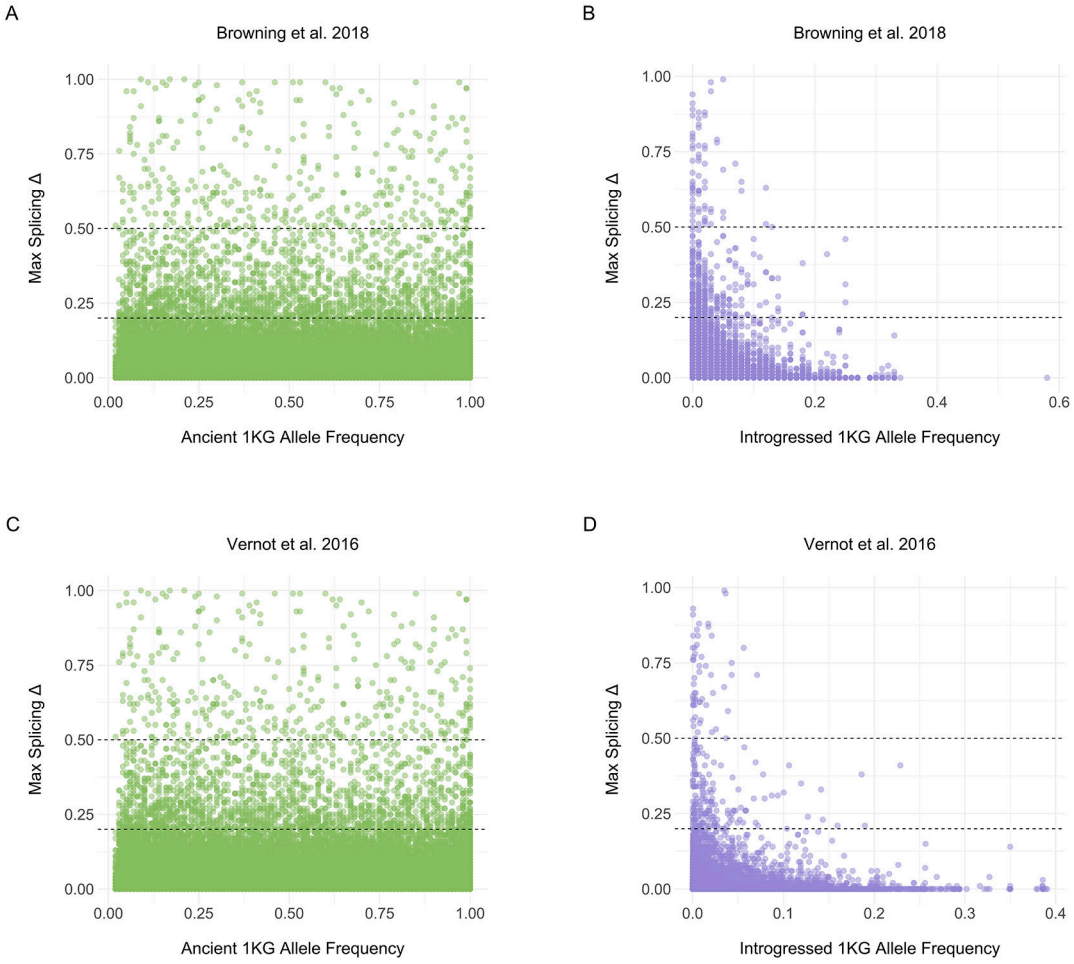
Extended Data Fig. 6: Vindija phenotype enrichment.

(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.

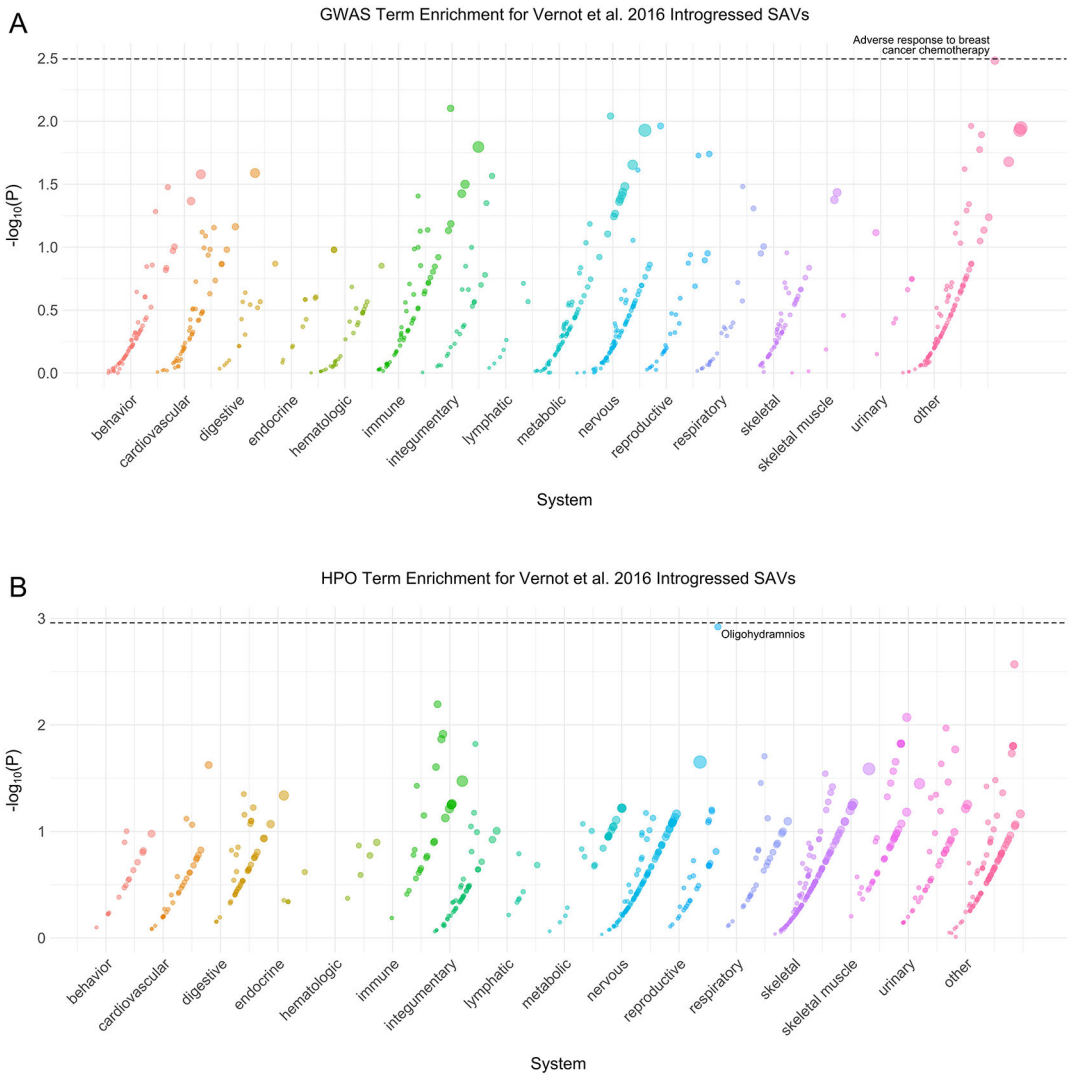


Extended Data Fig. 7: Modeling SAV effects on the canonical transcript.

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 (e.g., 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 (e.g., 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).

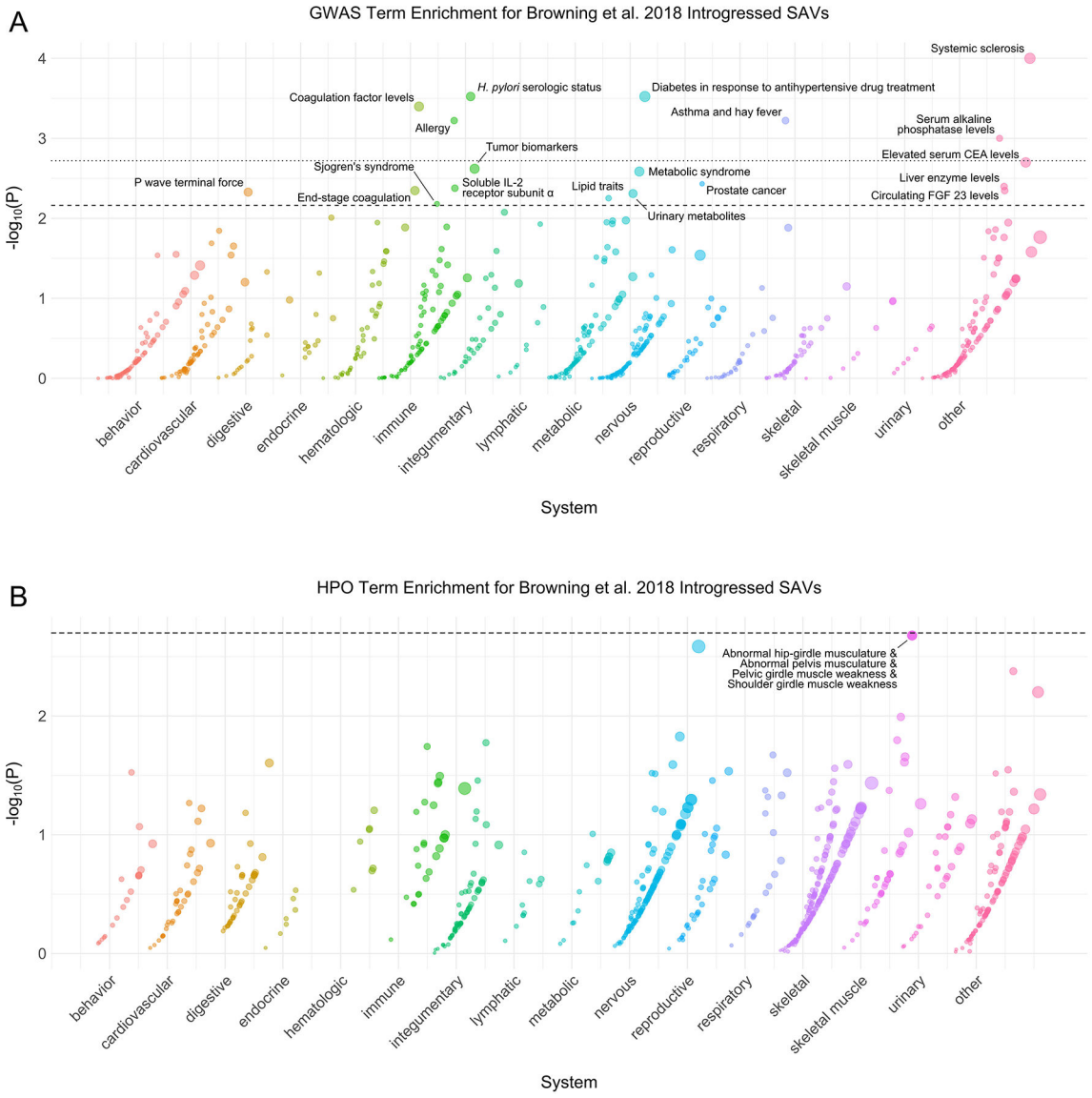


Extended Data Fig. 8: Δ max exhibits a variable relationship to 1KG allele frequency.
(A) 1KG allele frequency and Δ max for all ancient variants per [39]. Allele frequencies are from 1KG. Dashed lines reflect both Δ thresholds. (B) 1KG allele frequency and Δ max for all introgressed variants per [39]. 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 [38]. Allele frequencies are from 1KG. (D) 1KG allele frequency and Δ max for all introgressed variants per [38]. Allele frequencies represent the mean from the AFR, AMR, EAS, EUR, SAS frequencies from the [38] metadata.



Extended Data Fig. 9: Vernot et al. 2016 introgressed phenotype enrichment.

(A) Phenotype associations enriched among genes with [38] 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 [38] 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.



Extended Data Fig. 10: Browning et al. 2018 introgressed phenotype enrichment.

(A) Phenotype associations enriched among genes with [39] 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 [39] 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Mary Lauren Benton for kindly sharing data on tissue specificity and Ziyue Gao for helpful discussion on recurrent mutations. Evonne McArthur and David 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 UCSF. L.L.C. was funded by National Institutes of Health grant T32HG009495 to the University of Pennsylvania. J.A.C. and C.M.B. were funded by National Institutes of Health grant R35GM127087.

Data Availability

The SpliceAI annotated archaic variant dataset is available on Dryad (DOI: 10.7272/Q6H993F9).

Code Availability

All code used to conduct analyses and generate figures is publicly 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). [PubMed: 22936568]
2. Prüfer K. et al. The Complete Genome Sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49 (2014). [PubMed: 24352235]
3. Prüfer K. et al. A High-Coverage Neandertal Genome from Vindija Cave in Croatia. Science 358: 655–658 (2017). [PubMed: 28982794]
4. Mafessoni F. et al. A High-Coverage Neandertal Genome from Chagyrskaya Cave. Proceedings of the National Academy of Sciences 117: 15132 (2020).
5. Brand CM, Colbran LL & Capra JA Predicting Archaic Hominins Phenotypes from Genomic Data. Annual Review of Genomics and Human Genetics In press (2022).
6. Martin AR et al. Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities. Nature Genetics 51: 584–591 (2019). [PubMed: 30926966]
7. Castellano S. et al. Patterns of Coding Variation in the Complete Exomes of Three Neandertals. Proceedings of the National Academy of Sciences 111: 6666 (2014).
8. Colbran LL et al. Inferred Divergent Gene Regulation in Archaic Hominins Reveals Potential Phenotypic Differences. Nature Ecology & Evolution 3: 1598–1606 (2019). [PubMed: 31591491]
9. Gokhman D. et al. Reconstructing Denisovan Anatomy Using DNA Methylation Maps. Cell 179: 180–192.e10 (2019). [PubMed: 31539495]
10. McArthur E. et al. Reconstructing the 3D Genome Organization of Neanderthals Reveals That Chromatin Folding Shaped Phenotypic and Sequence Divergence. *bioRxiv:* 2022.02.07.479462 (2022).
11. Lopez AJ Alternative Slicing of Pre-mRNA: Developmental Consequences and Mechanisms of Regulation. Annual Review of Genetics 32: 279–305 (1998).
12. Graveley BR Alternative Splicing: Increasing Diversity in the Proteomic World. Trends in Genetics 17: 100–107 (2001). [PubMed: 11173120]
13. Black DL Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry 72: 291–336 (2003).

14. Baralle FE & Giudice J Alternative Splicing as a Regulator of Development and Tissue Identity. Nature Reviews Molecular Cell Biology 18: 437–451 (2017). [PubMed: 28488700]
15. Cáceres JF & Kornblihtt AR Alternative Splicing: Multiple Control Mechanisms and Involvement in Human Disease. Trends in Genetics 18: 186–193 (2002). [PubMed: 11932019]
16. Faustino NA & Cooper TA Pre-mRNA Splicing and Human Disease. Genes & Development 17: 419–437 (2003). [PubMed: 12600935]
17. Nissim-Rafinia M & Kerem B Splicing Regulation as a Potential Genetic Modifier. Trends in Genetics 18: 123–127 (2002). [PubMed: 11858835]
18. Krawczak M, Reiss J & Cooper DN The Mutational Spectrum of Single Base-Pair Substitutions in mRNA Splice Junctions of Human Genes: Causes and Consequences. Human Genetics 90: 41–54 (1992). [PubMed: 1427786]
19. Wang G-S & Cooper TA Splicing in Disease: Disruption of the Splicing Code and the Decoding Machinery. Nature Reviews Genetics 8: 749–761 (2007).
20. Li YI et al. RNA Splicing Is a Primary Link between Genetic Variation and Disease. Science 352: 600–604 (2016). [PubMed: 27126046]
21. Scotti MM & Swanson MS RNA Mis-Splicing in Disease. Nature Reviews Genetics 17: 19–32 (2016).
22. Li X. et al. The Impact of Rare Variation on Gene Expression across Tissues. Nature 550: 239–243 (2017). [PubMed: 29022581]
23. Verta J-P & Jacobs A The Role of Alternative Splicing in Adaptation and Evolution. Trends in Ecology & Evolution 37: 299–308 (2022). [PubMed: 34920907]
24. Singh P & Ahi EP The Importance of Alternative Splicing in Adaptive Evolution. Molecular Ecology n/a (2022).
25. Wright CJ, Smith CWJ & Jiggins CD Alternative Splicing as a Source of Phenotypic Diversity. Nature Reviews Genetics (2022).
26. Blekhman R, Marioni JC, Zumbo P, Stephens M & Gilad Y Sex-Specific and Lineage-Specific Alternative Splicing in Primates. Genome Research 20: 180–189 (2010). [PubMed: 20009012]
27. Barbosa-Morais NL et al. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science 338: 1587–1593 (2012). [PubMed: 23258890]
28. Merkin J, Russell C, Chen P & Burge CB Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. Science 338: 1593–1599 (2012). [PubMed: 23258891]
29. Sibley CR, Blazquez L & Ule J Lessons from Non-Canonical Splicing. Nature Reviews Genetics 17: 407–421 (2016).
30. Jenkinson G. et al. LeafCutterMD: An Algorithm for Outlier Splicing Detection in Rare Diseases. Bioinformatics 36: 4609–4615 (2020). [PubMed: 32315392]
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). [PubMed: 30591034]
32. Mertes C. et al. Detection of Aberrant Splicing Events in RNA-seq Data Using FRASER. Nature Communications 12: 529 (2021).
33. Cheng J. et al. MMSplice: Modular Modeling Improves the Predictions of Genetic Variant Effects on Splicing. Genome Biology 20: 48 (2019). [PubMed: 30823901]
34. Jagadeesh KA et al. S-CAP Extends Pathogenicity Prediction to Genetic Variants That Affect RNA Splicing. Nature Genetics 51: 755–763 (2019). [PubMed: 30804562]
35. Jaganathan K. et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell 176: 535–548.e24 (2019). [PubMed: 30661751]
36. Danis D. et al. Interpretable Prioritization of Splice Variants in Diagnostic Next-Generation Sequencing. *bioRxiv:* 2021.01.28.428499 (2021).
37. Zeng T & Li YI Predicting RNA Splicing from DNA Sequence Using Pangolin. Genome Biology 23: 103 (2022). [PubMed: 35449021]
38. Vernot B. et al. Excavating Neandertal and Denisovan DNA from the Genomes of Melanesian Individuals. Science 352: 235–239 (2016). [PubMed: 26989198]

39. Browning SR, Browning BL, Zhou Y, Tucci S & Akey JM Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell 173: 53–61.e9 (2018). [PubMed: 29551270]
40. Rogers AR, Harris NS & Achenbach AA Neanderthal-Denisovan Ancestors Interbred with a Distantly Related Hominin. Science Advances 6: eaay5483 (2020). [PubMed: 32128408]
41. Collins L & Penny D Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. Molecular Biology and Evolution 22: 1053–1066 (2005). [PubMed: 15659557]
42. Tweedie S. et al. Genenames.Org: The HGNC and VGNC Resources in 2021. Nucleic Acids Research 49: D939–D946 (2021). [PubMed: 33152070]
43. Lowy-Gallego E. et al. Variant Calling on the GRCh38 Assembly with the Data from Phase Three of the 1000 Genomes Project [Version 2; Peer Review: 2 Approved]. Wellcome Open Research 4 (2019).
44. Karczewski KJ et al. The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. Nature 581: 434–443 (2020). [PubMed: 32461654]
45. McLaren W. et al. The Ensembl Variant Effect Predictor. Genome Biology 17: 122 (2016). [PubMed: 27268795]
46. Aqeilan RI et al. The WWOX Tumor Suppressor Is Essential for Postnatal Survival and Normal Bone Metabolism *. Journal of Biological Chemistry 283: 21629–21639 (2008). [PubMed: 18487609]
47. Shiina T, Hosomichi K, Inoko H & Kulski JK The HLA Genomic Loci Map: Expression, Interaction, Diversity and Disease. Journal of Human Genetics 54: 15–39 (2009). [PubMed: 19158813]
48. Rodenas-Cuadrado P, Ho J & Vernes SC Shining a Light on CNTNAP2: Complex Functions to Complex Disorders. European Journal of Human Genetics 22: 171–178 (2014). [PubMed: 23714751]
49. Buniello A. et al. The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019. Nucleic Acids Research 47: D1005–D1012 (2019). [PubMed: 30445434]
50. Köhler S. et al. The Human Phenotype Ontology in 2021. Nucleic Acids Research 49: D1207–D1217 (2021). [PubMed: 33264411]
51. Pollard KS, Hubisz MJ, Rosenbloom KR & Siepel A Detection of Nonneutral Substitution Rates on Mammalian Phylogenies. Genome Research 20: 110–121 (2010). [PubMed: 19858363]
52. Kriventseva EV et al. Increase of Functional Diversity by Alternative Splicing. Trends in Genetics 19: 124–128 (2003). [PubMed: 12615003]
53. Rong S. et al. Large Scale Functional Screen Identifies Genetic Variants with Splicing Effects in Modern and Archaic Humans. *bioRxiv:* 2022.11.20.515225 (2022).
54. Petr M, Pääbo S, Kelso J & Vernot B Limits of Long-Term Selection against Neandertal Introgression. Proceedings of the National Academy of Sciences 116: 1639 (2019).
55. Telis N, Aguilar R & Harris K Selection against Archaic Hominin Genetic Variation in Regulatory Regions. Nature Ecology & Evolution 4: 1558–1566 (2020). [PubMed: 32839541]
56. McArthur E, Rinker DC & Capra JA Quantifying the Contribution of Neanderthal Introgression to the Heritability of Complex Traits. Nature Communications 12: 4481 (2021).
57. Aqil A, Speidel L, Pavlidis P & Gokcumen O Balancing Selection on Genomic Deletion Polymorphisms in Humans. *bioRxiv:* 2022.04.28.489864 (2022).
58. Dannemann M, Andrés AM & Kelso J Introgression of Neandertal- and Denisovan- like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. The Amer- ican Journal of Human Genetics 98: 22–33 (2016).
59. McCoy RC, Wakefield J & Akey JM Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell 168: 916–927.e12 (2017). [PubMed: 28235201]
60. Saudemont B. et al. The Fitness Cost of Mis-Splicing Is the Main Determinant of Alternative Splicing Patterns. Genome Biology 18: 208 (2017). [PubMed: 29084568]

61. Mendez FL, Watkins JC & Hammer MF Global Genetic Variation at OAS1 Provides Evidence of Archaic Admixture in Melanesian Populations. Molecular Biology and Evolution 29: 1513–1520 (2012). [PubMed: 22319157]
62. Sams AJ et al. Adaptively Introgressed Neandertal Haplotype at the OAS Locus Functionally Impacts Innate Immune Responses in Humans. Genome Biology 17: 246 (2016). [PubMed: 27899133]
63. Rinker DC et al. Neanderthal Introgression Reintroduced Functional Ancestral Alleles Lost in Eurasian Populations. Nature Ecology & Evolution 4: 1332–1341 (2020). [PubMed: 32719451]
64. Huerta-Sánchez E. et al. Altitude Adaptation in Tibetans Caused by Introgression of Denisovan-like DNA. Nature 512: 194–197 (2014). [PubMed: 25043035]
65. Jeong C. et al. Detecting Past and Ongoing Natural Selection among Ethnically Tibetan Women at High Altitude in Nepal. PLOS Genetics 14: e1007650 (2018). [PubMed: 30188897]
66. Peng Y. et al. Down-Regulation of EPAS1 Transcription and Genetic Adaptation of Tibetans to High-Altitude Hypoxia. Molecular Biology and Evolution 34: 818–830 (2017). [PubMed: 28096303]
67. Andrés AM et al. Balancing Selection Maintains a Form of ERAP2 That Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation. PLOS Genetics 6: e1001157 (2010). [PubMed: 20976248]
68. Trujillo CA et al. Reintroduction of the Archaic Variant of *NOVA1* in Cortical Organoids Alters Neurodevelopment. Science 371: eaax2537 (2021). [PubMed: 33574182]
69. Karlebach G. et al. The Impact of Biological Sex on Alternative Splicing. *bioRxiv:* 490904 (2020).
70. Rogers TF, Palmer DH & Wright AE Sex-Specific Selection Drives the Evolution of Alternative Splicing in Birds. Molecular Biology and Evolution 38: 519–530 (2021). [PubMed: 32977339]
71. Ge Y & Porse BT The Functional Consequences of Intron Retention: Alternative Splicing Coupled to NMD as a Regulator of Gene Expression. BioEssays 36: 236–243 (2014). [PubMed: 24352796]
72. Smith JE & Baker KE Nonsense-Mediated RNA Decay – a Switch and Dial for Regulating Gene Expression. BioEssays 37: 612–623 (2015). [PubMed: 25820233]
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). [PubMed: 21903627]
74. Chollet F. et al. Keras https://github.com/fchollet/keras. 2015.
75. Abadi M. et al. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems 2015.
76. Harrow J. et al. GENCODE: The Reference Human Genome Annotation for The ENCODE Project. Genome Research 22: 1760–1774 (2012). [PubMed: 22955987]
77. Hinrichs AS et al. The UCSC Genome Browser Database: Update 2006. Nucleic Acids Research 34: D590–D598 (2006). [PubMed: 16381938]
78. Plagnol V & Wall JD Possible Ancestral Structure in Human Populations. PLOS Genetics 2: e105 (2006). [PubMed: 16895447]
79. Vernot B & Akey JM Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science 343: 1017–1021 (2014). [PubMed: 24476670]
80. Quinlan AR & Hall IM BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. Bioinformatics 26: 841–842 (2010). [PubMed: 20110278]
81. Chen EY et al. Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. BMC Bioinformatics 14: 128 (2013). [PubMed: 23586463]
82. Kuleshov MV et al. Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. Nucleic Acids Research 44: W90–W97 (2016). [PubMed: 27141961]
83. Xie Z. et al. Gene Set Knowledge Discovery with Enrichr. Current Protocols 1: e90 (2021). [PubMed: 33780170]
84. Virtanen P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17: 261–272 (2020). [PubMed: 32015543]
85. Li H. et al. The Sequence Alignment/Map Format and SAMtools. Bioinformatics 25: 2078– 2079 (2009). [PubMed: 19505943]
86. Vallat R. Pingouin: Statistics in Python. The Journal of Open Source Software 3: 1026 (2018).

87. Inkscape Project. Inkscape 2020.
88. Wickham H. Ggplot2: Elegant Graphics for Data Analysis ISBN: 978-3-319-24277-4 (Springer-Verlag, New York, 2016).
89. R Core Team. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing. Vienna, Austria, 2020.
90. Krassowski M. ComplexUpset 2020.
91. Larsson J. eulerr: Area-proportional Euler and Venn Diagrams with Ellipses Manual (2021).
92. Wickham H. Reshaping Data with the Reshape Package. Journal of Statistical Software 21 (2007).
93. Wickham H. et al. Welcome to the Tidyverse. Journal of Open Source Software 4: 1686 (2019).
94. Chen L, Wolf AB, Fu W, Li L & Akey JM Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. Cell 180: 677–687.e16 (2020). [PubMed: 32004458]

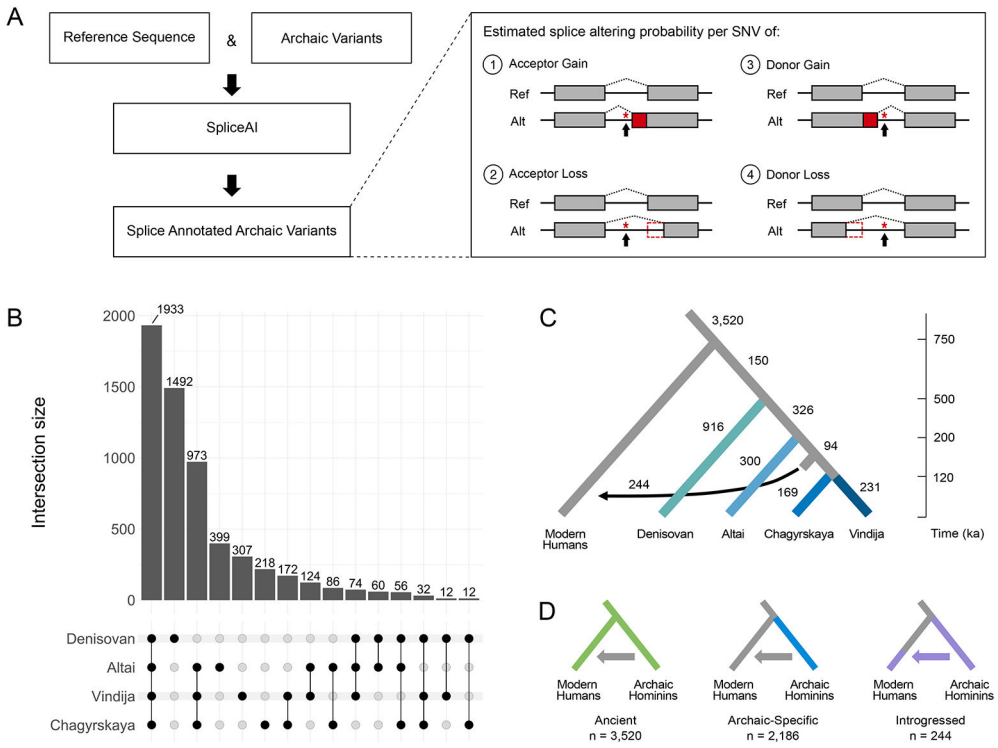


Fig. 1: The identification, distribution, and origin of archaic SAVs.

(A) We used SpliceAI to identify putative splice altering variants in archaic hominin genomes. We analyzed autosomal SNVs from four archaic hominins compared to the reference sequence (hg19/GRCh37). SpliceAI annotates each variant with splice altering probabilities (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. (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: [38] (shown here) and [39]. The divergence times and placement of the introgression arrow reflect estimates from [4] and [40]. 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” are archaic SAVs present in both modern and archaic hominin individuals and are inferred to have origins before the last common ancestor of these groups. “Archaic-Specific” are archaic SAVs that are present in archaic hominins, but absent or present at low frequency (allele frequency < 0.0001) in modern humans. “Introgressed” are archaic SAVs that were introgressed into Eurasian populations due to archaic admixture.

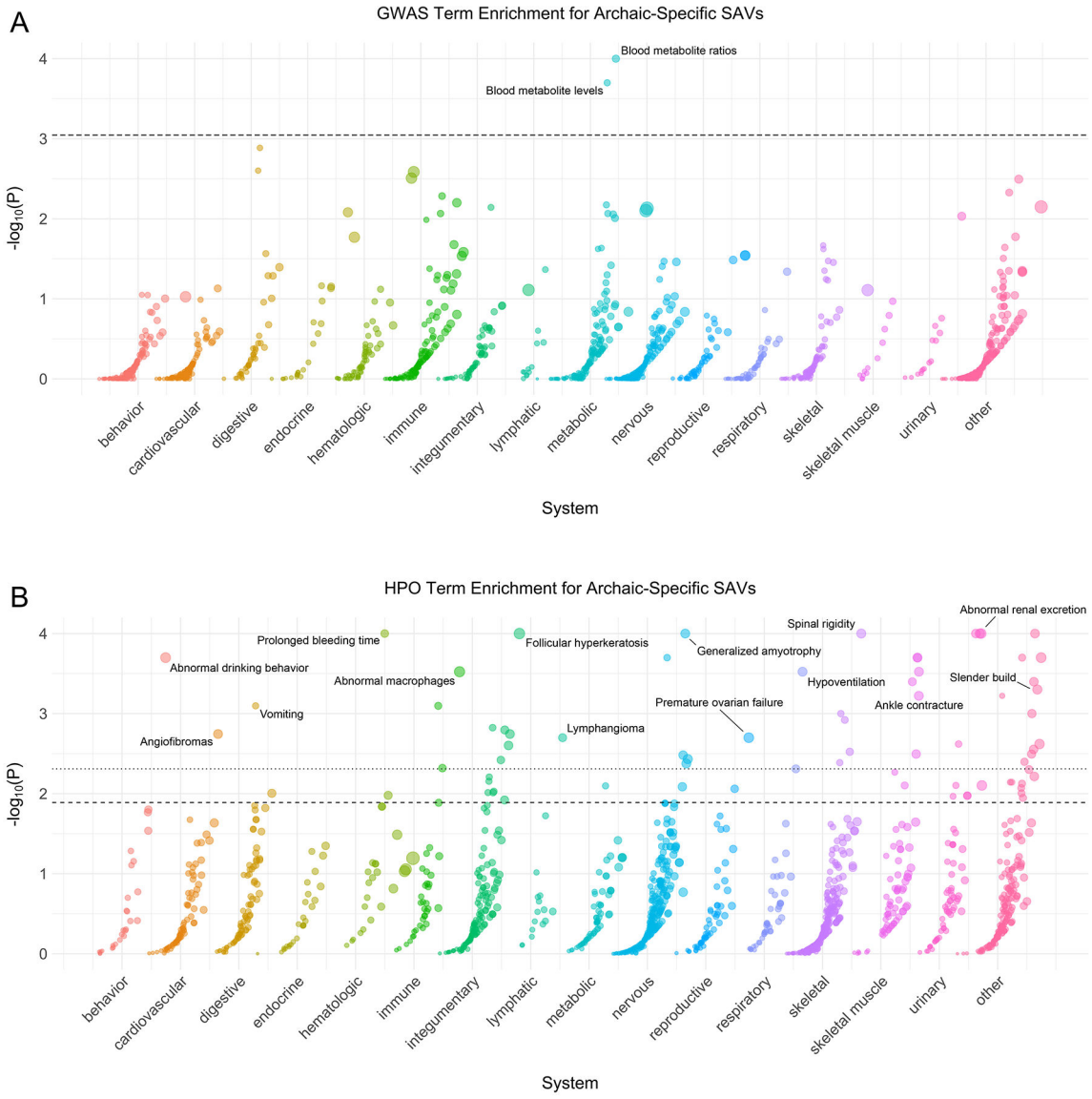


Fig. 2: Genes with archaic-specific SAVs are enriched for roles in many phenotypes.

(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 false discovery rate (FDR) corrected p-value thresholds at FDR = 0.05 and 0.1, respectively. 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 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.

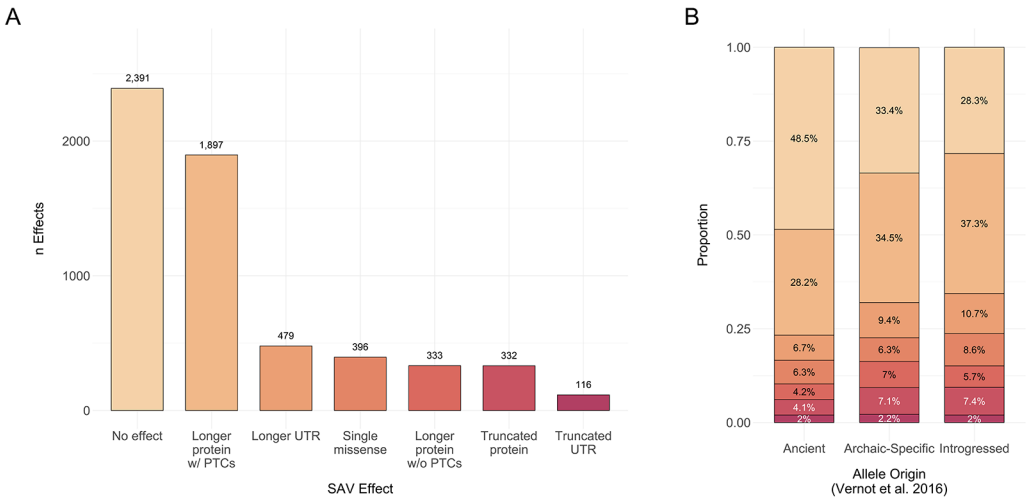


Fig. 3: Most SAVs result in isoforms that trigger nonsense mediated decay or yield altered transcripts and proteins.

(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, UTR = 5’ or 3’ untranslated region. (B) The number of protein effects stratified by allele origin [38]. Colors indicate the transcript or protein effect as in A.

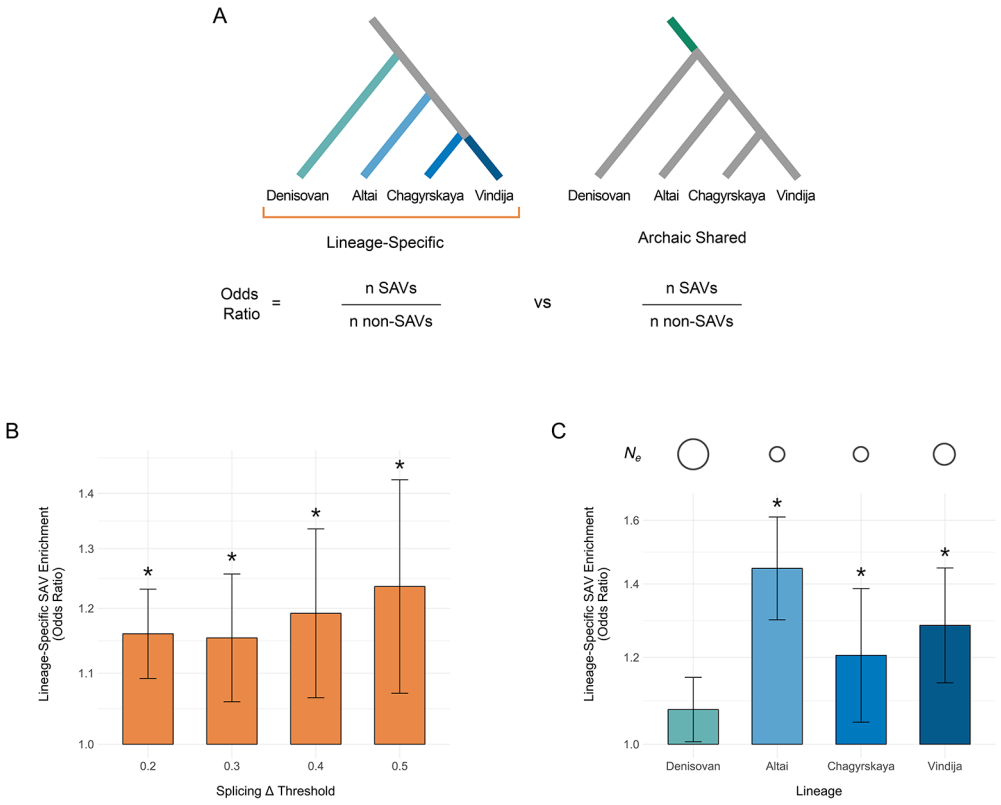


Fig. 4: Lineage-specific archaic variants are enriched for SAVs compared to shared archaic variants.

(A) We hypothesized 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 odds ratio for being an 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 hypothesized 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 odds ratio for being an 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 splice altering probability (Δ) thresholds. Bar height indicates the odds ratio. Error bars denote the 95% CI and are centered on the odds ratio. Asterisks reflect significance of Fisher's exact tests using a Bonferroni corrected α (0.0125). Note 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. N_e estimates are from [4]. Bar height indicates the odds ratio. Error bars denote the 95% CI and are centered on the odds ratio. Asterisks reflect significance of Fisher's exact tests using a Bonferroni

corrected α (0.0125). Note the y-axis is $\log_{10}$ transformed. Odds ratios 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.

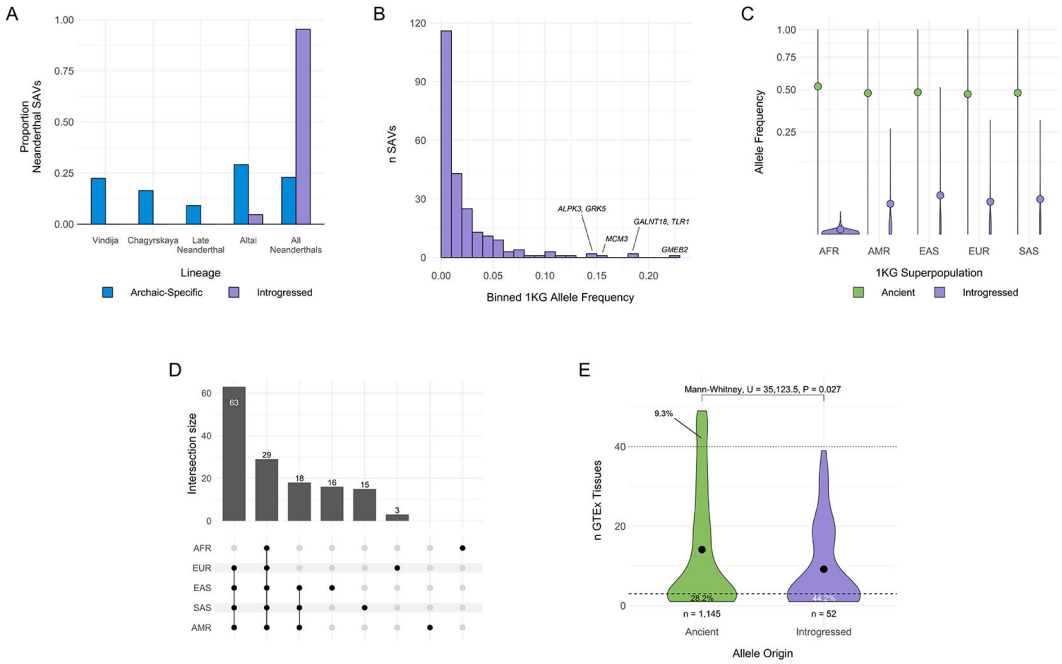


Fig. 5: Introgressed SAVs present in modern humans were shared across archaic individuals and are associated with increased tissue specificity.

(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 significantly 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 per [38]. 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 vs. introgressed) per [38]. Ancient SAVs (n = 2,252) are colored green and displayed on the left, while introgressed SAVs (n = 237) are colored purple and displayed on the right per superpopulation. AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian. The colored 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. In all panels introgressed SAVs and frequencies are as defined by [38]. (E) The distribution of the number of GTEx tissues in which an ancient or introgressed SAV (per [38]) was identified as an sQTL. Introgressed variants are significantly more tissue-specific. We defined “tissue-specific” variants as those occurring in 1 or 2 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.

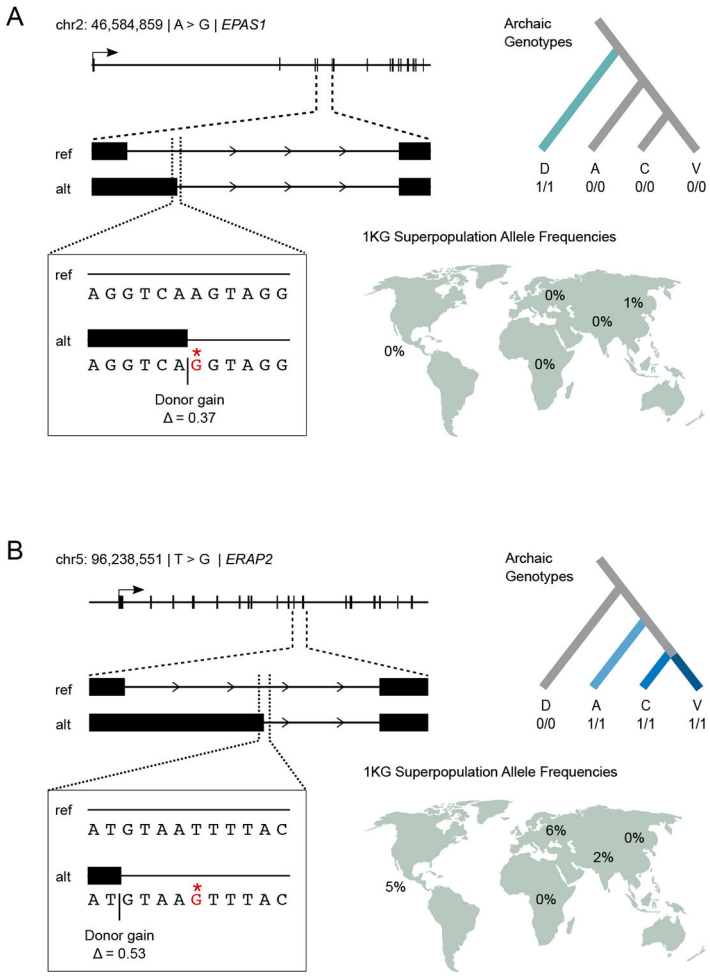


Fig. 6: Example archaic SAVs leading to nonsense mediated decay in loci with evidence of recent adaptive evolution.

(A) A Denisovan-specific homozygous SAV results in a donor gain in *EPAS1*, hypoxia-inducible factor-2alpha, between the fourth and fifth exon. The transcript resulting from the SAV introduces six premature termination codons (Supplementary Data 3), which likely results in the elimination of the transcript via nonsense mediated decay. 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*, a 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 11th and 12th exons, which introduces nine premature termination codons (Supplementary Data 3).