

# 1 Resurrecting the Alternative Splicing Landscape of Archaic 2 Hominins using Machine Learning

3 Colin M. Brand<sup>1,2</sup>, Laura L. Colbran<sup>3</sup>, and John A. Capra<sup>1,2,\*</sup>

4 <sup>1</sup>Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA

5 <sup>2</sup>Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

6 <sup>2</sup>Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

7

8 \*Correspondence to [tony@capralab.org](mailto:tony@capralab.org)

9

## Abstract

10 Alternative splicing contributes to adaptation and divergence in many species. However, it  
11 has not been possible to directly compare splicing between modern and archaic hominins.  
12 Here, we unmask the recent evolution of this previously unobservable regulatory mecha-  
13 nism by applying SpliceAI, a machine-learning algorithm that identifies splice altering vari-  
14 ants (SAVs), to high-coverage genomes from three Neanderthals and a Denisovan. We  
15 discover 5,950 putative archaic SAVs, of which 2,343 are archaic-specific and 3,607 also oc-  
16 cur in modern humans via introgression (237) or shared ancestry (3,370). Archaic-specific  
17 SAVs are enriched in genes that contribute to many traits potentially relevant to hominin  
18 phenotypic divergence, such as perspiration, platelet volume, and spinal rigidity. Com-  
19 compared to shared SAVs, archaic-specific SAVs occur in sites under weaker selection and are  
20 more common in genes with tissue-specific expression. Further underscoring the impor-  
21 tance of negative selection on SAVs, Neanderthal lineages with low effective population  
22 sizes are enriched for SAVs compared to Denisovan and shared SAVs. Finally, we find that  
23 nearly all introgressed SAVs in humans were shared across Neanderthals, suggesting that  
24 older SAVs were more tolerated in modern human genomes. Our results reveal the splicing  
25 landscape of archaic hominins and identify potential contributions of splicing to phenotypic  
26 differences among hominins.

## 27 1 Introduction

28 While the paleontological and archaeological records provide evidence about some pheno-  
29 types of extinct hominins, most ancient tissues have not survived to the present. The discov-  
30 ery and successful sequencing of DNA genome-wide from a Denisovan (Meyer et al., 2012)  
31 and multiple Neanderthal genomes (Prüfer et al., 2014; Prüfer et al., 2017; Mafessoni et al.,  
32 2020) enabled direct comparisons of the genotypes of these archaic hominins to one another  
33 and anatomically modern humans. These data also enable the potential for indirect pheno-  
34 typic comparisons by predicting archaic phenotypes from their genomes (Brand et al., 2022).  
35 Diverse molecular mechanisms collectively shape the similarities and differences between ar-  
36chaic hominins and modern humans. Given that the biology linking genotype to organism-level  
37 phenotype is complex and that the mapping may not generalize across human populations  
38 (Martin et al., 2019), predicting “low-level” molecular phenotypes from genetic information is  
39 a promising alternative. Recent work has successfully explored such differences in protein-  
40 coding sequence (Castellano et al., 2014) and differences relevant to gene expression, such  
41 as divergent gene regulation (Colbran et al., 2019), differential methylation (Gokhman et al.,  
42 2019), and divergent 3D genome contacts (McArthur et al., 2022).

43 Variation in gene splicing could also underlie phenotypic differences between archaic ho-  
44 minins and modern humans, but archaic splicing patterns have not been comprehensively  
45 quantified. Alternative splicing enables the production of multiple protein isoforms from a single  
46 gene (Lopez, 1998; Graveley, 2001; Black, 2003). The resulting proteomic diversity is essential  
47 for many processes, including development and establishing tissue identity (Baralle and Giu-  
48 dice, 2017). Defects in splicing underlie many human diseases (e.g., Cáceres and Kornblihtt,  
49 2002; Faustino and Cooper, 2003; Nissim-Rafinia and Kerem, 2002; Krawczak et al., 1992;  
50 Wang and Cooper, 2007; Li et al., 2016; Scotti and Swanson, 2016; Li et al., 2017), and varia-  
51 tion in splicing contributes to differences in traits in non-human species (see Verta and Jacobs,  
52 2022, Table 1). Further, alternative splicing can evolve rapidly and respond to environmental  
53 factors—suggesting it often contributes to adaptation (Singh and Ahi, 2022; Verta and Jacobs,  
54 2022; Wright et al., 2022) and species differences (Blekhman et al., 2010; Barbosa-Morais et  
55 al., 2012; Merkin et al., 2012).

56 Splicing patterns are directly influenced by the nucleotide sequences surrounding splice  
57 sites (Sibley et al., 2016). This has enabled the development of many algorithms to predict  
58 alternative splicing from RNA-seq data (Jenkinson et al., 2020; Zhang et al., 2018; Mertes  
59 et al., 2021), RNA sequence, or DNA sequence (Cheng et al., 2019; Jagadeesh et al., 2019;  
60 Jaganathan et al., 2019; Danis et al., 2021; Zeng and Li, 2022). Beyond human clinical appli-  
61 cations, methods that require only DNA sequence can be leveraged to understand alternative  
62 splicing in extant species for which acquiring RNA-seq data may be difficult to impossible or in  
63 extinct taxa, such as archaic hominins.

64 Here, we resurrect the genome-wide alternative splicing landscape of archaic hominins  
65 using SpliceAI, a pre-trained algorithm that predicts splicing patterns from sequence alone (Ja-  
66 ganathan et al., 2019). First, we assess the distribution of splice-altering variants (SAVs) among  
67 all four archaic individuals, identify which genes are affected, and describe how the transcripts  
68 are modified. Second, we quantify which SAVs are also present in modern humans due to  
69 shared ancestry or introgression. Third, we quantify SAV enrichment among gene sets that  
70 underlie modern human phenotypes. Fourth, we explore how selection shaped alternative

71 splicing in archaics. Fifth, we evaluate the expression and function of archaic SAVs that are  
72 also present in modern humans. Finally, we highlight a handful of archaic SAVs with potential  
73 evolutionary significance.

## 74 2 Results

75 We examined the alternative splicing landscape in archaic hominins using all four currently  
76 available high-coverage archaic genomes, representing three Neanderthals (Mafessoni et al.,  
77 2020; Prüfer et al., 2014; Prüfer et al., 2017) and a Denisovan (Meyer et al., 2012). We ap-  
78 plied the SpliceAI classifier to sites with high-quality genotype calls where at least one archaic  
79 individual exhibited at least one allele different from the human reference (hg19/GRCh37) us-  
80 ing the built-in GENCODE, Human Release 24, annotations to identify variants in gene bodies  
81 (**Figure 1A**). SpliceAI estimates  $\Delta$ , the splice-altering probability (SAP), for each variant of: 1)  
82 an acceptor gain, 2) an acceptor loss, 3) a donor gain, and 4) a donor loss (**Figure 1A**). SpliceAI  
83 also indicates the positions changed for each of these four effects in basepairs.

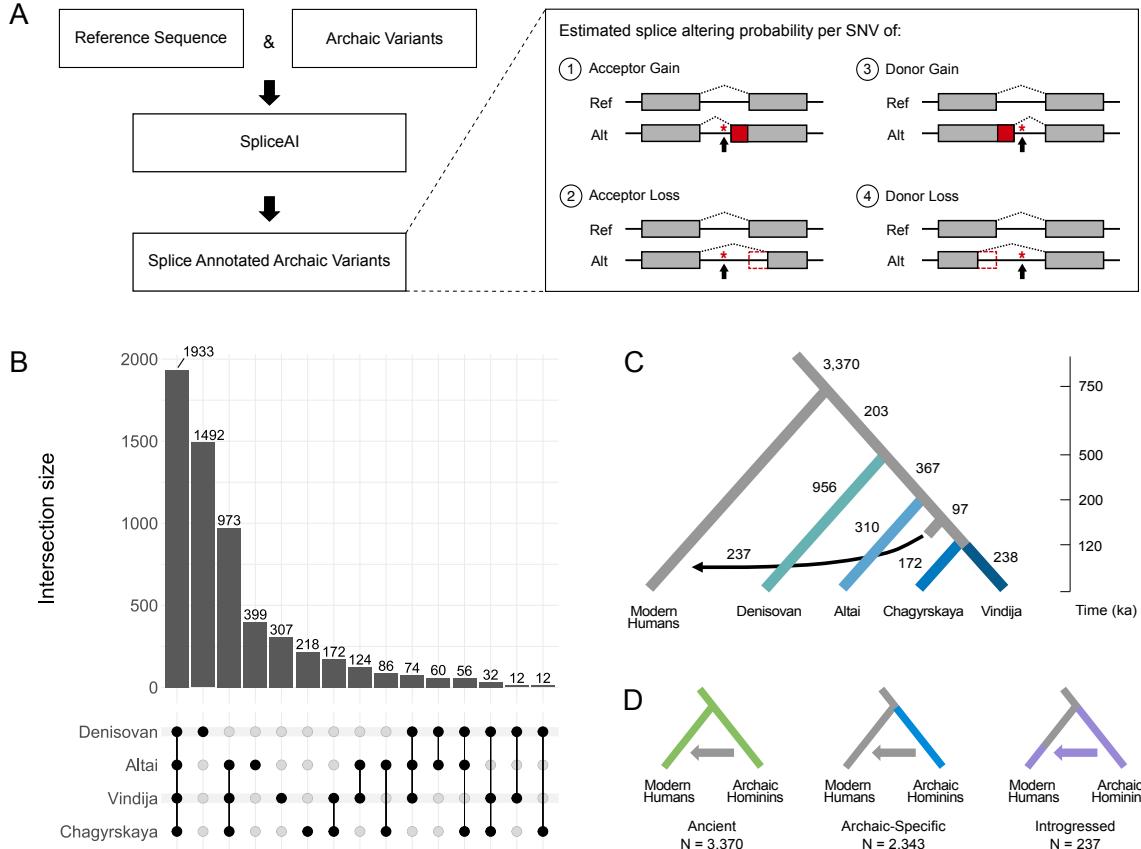
84 Alternative splicing occurs across nearly all eukaryotes and its molecular mechanisms are  
85 deeply conserved (Collins and Penny, 2005). We therefore anticipated that the sequence pat-  
86 terns learned by SpliceAI in humans would generalize to archaics. To confirm this, we searched  
87 for archaic variants in the 246 genes associated with the major spliceosome complex (Tweedie  
88 et al., 2021) that were absent from individuals sequenced by the 1000 Genomes Project (1KG)  
89 (Lowy-Gallego et al., 2019), i.e., “archaic-specific” variants (**File S1**). We annotated these vari-  
90 ants using the Ensembl Variant Effect Predictor (McLaren et al., 2016). We found only two  
91 missense variants that were scored as likely to disrupt protein function by both PolyPhen and  
92 SIFT, neither of which were fixed in all four archaics (**File S1**). Thus, there is near complete  
93 conservation of the proteins involved in splicing between archaic hominins and modern hu-  
94 mans.

### 95 2.1 Thousands of splice altering variants (SAVs) are present in archaic hominins.

96 We identified 1,567,894 autosomal positions with  $\geq 1$  non-reference allele in the archaic indi-  
97 viduals. Among these, 1,049 variants had high splice altering probability (SAP; SpliceAI  $\Delta \geq$   
98 0.5) out of 1,607,350 archaic variants we analyzed. 5,950 archaic variants had moderate SAP  
99 ( $\Delta \geq 0.2$ ). Hereafter, we refer to these variants as high-confidence splice altering variants  
100 (SAVs) and SAVs, respectively, and to maximize sensitivity, we focus on the SAVs in the main  
text.

102 The number of SAVs was similar among the four archaics, ranging from 3,482 (Chagyrskaya)  
103 to 3,705 (Altai) (**Table S1**). These values fell within the range of SAVs observed in individual  
104 modern humans, estimated from one randomly sampled individual per 1KG population (**Ta-  
105 ble S2**). SAVs were most commonly shared among all four archaic individuals (**Figure 1B**;  
106 **Figure S1**). As expected from the known phylogeny, the Denisovan exhibited the most unique  
107 SAVs, followed by all Neanderthals, and each individual Neanderthal (**Figure 1B**; **Figure S1**).

108 A total of 4,242 genes have at least one archaic SAV. 3,111 genes have only one SAV;  
109 however, 1,131 had multiple SAVs (**Table S3**). Among the genes with the largest number of  
110 archaic SAVs are: *WWOX* (N = 9), which is involved in bone growth development (Aqeilan  
111 et al., 2008), *HLA-DPA1* (N = 7) and *HLA-DPB1* (N = 10), essential components of the immune



**Figure 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 ( $\Delta$  scores) and position changes for each class of splicing alteration: 1) acceptor gain, 2) acceptor loss, 3) donor gain, or 4) donor loss. (B) The distribution of the presence of archaic SAVs at  $\Delta \geq 0.2$  across archaic individuals. The dots indicate which lineage(s) have the associated SAVs. (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: Vernot et al., 2016 (shown here) and Browning et al., 2018. The divergence times and placement of the introgression arrow reflect recent estimates from Mafessoni et al., 2020 and Rogers et al., 2020. (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 not modern humans. “Introgressed” are archaic SAVs that were introgressed into Eurasian populations due to archaic admixture.

112 system (Shiina et al., 2009), and *CNTNAP2* ( $N = 11$ ), a nervous system protein associated  
113 with neurodevelopmental disorders that is also one of the longest genes in the human genome  
114 (Rodenas-Cuadrado et al., 2014).

115 Many SAVs (47.8%) have a high SAP for only one of the four classes splicing change (ac-  
116 ceptor gain, acceptor loss, donor gain, and donor loss), and as expected, the overall association  
117 between the probabilities of different change types was negative ( $\rho = -0.34 - -0.14$ ) for variants  
118 with at least one SAP greater than 0 (**Figure S2**). Donor gains were the most frequent result  
119 of splice altering variants for both thresholds (29.5% and 35.1% of variant effects, respectively  
120 **Figure S3**). While this may reflect the true distribution, we cannot rule out that the classifier  
121 has greater power to recognize donor gains compared to acceptor gains, acceptor losses, and  
122 donor losses.

## 123 **2.2 39% of archaic SAVs are archaic-specific.**

124 We inferred the origin of archaic variants based on parsimony. We identified 2,343 SAVs that  
125 are not present in modern humans in 1KG nor present in GTEx sQTL data (**Figure 1C**). We  
126 refer to this 39% of archaic SAVs as “archaic-specific”.

127 The remaining 61% archaic SAVs are present in modern humans. Archaic hominins and  
128 modern humans last shared a common ancestor approximately 570–752 ka (Rogers et al.,  
129 2020). SAVs present in both archaic and modern humans may be the result of either intro-  
130 gression, shared ancestry, or independent occurrence. To identify SAVs present in 1KG due  
131 to introgression, we used two datasets on archaic introgression into modern humans (Vernot  
132 et al., 2016; Browning et al., 2018). While modern human genomes retain Denisovan and Ne-  
133 anderthal ancestry, most 1KG samples have minimal (< 1%) Denisovan ancestry (Browning  
134 et al., 2018; Vernot et al., 2016). Therefore, we focused on Neanderthal introgression and clas-  
135 sified 237 SAVs identified by Vernot et al., 2016 in 232 genes (**Figure 1D; Figure S4**) and 377  
136 SAVs identified by Browning et al., 2018 in 353 genes as “introgressed” (**Figure S4**). Despite  
137 only modest overlap between the two introgression datasets (**Figure S5**), we observed few  
138 qualitative differences in downstream analyses. Hereafter, we present results using the Vernot  
139 et al., 2016 introgressed variants in the main text and include results using the Browning et al.,  
140 2018 set in the supplemental text.

141 Non-introgressed variants present in both archaic and modern humans likely evolved prior  
142 to our most recent common ancestor. We refer to these SAVs as “ancient”, and we consider  
143 the archaic SAVs with an allele frequency  $\geq 0.05$  in at least two 1KG superpopulations “high-  
144 confidence ancient”. This decreases the probability of independent occurrence or misclassi-  
145 fication of introgressed alleles. Hereafter, “ancient” refers to these high-confidence ancient  
146 variants unless otherwise specified. We identified 2,252 such variants based on Vernot et al.,  
147 2016 among 1,896 genes (**Figure 1D; Figure S4**) and 2,195 variants based on Browning et al.,  
148 2018 among 1,856 genes (**Figure S4**).

## 149 **2.3 Archaic-specific SAVs are enriched in genes with diverse phenotypes.**

150 To identify the potential phenotypic consequences of archaic-specific SAVs, we tested for en-  
151 richment of functional annotations among genes with SAVs. Following McArthur et al., 2022,  
152 we considered links between genes and phenotypes from two sources: GWAS Catalog 2019

153 (Buniello et al., 2019) and the Human Phenotype Ontology (Köhler et al., 2021), that capture  
154 common and rare disease annotations, respectively. Structural properties of genes, such as  
155 the number of exons, influence the probability that they have SAVs (**Table S4; Figure S6**).  
156 To account for these different probabilities, we generated a permutation-based empirical null  
157 distribution (Methods) and used it to estimate enrichment for each phenotype and control the  
158 false-discovery rate (FDR).

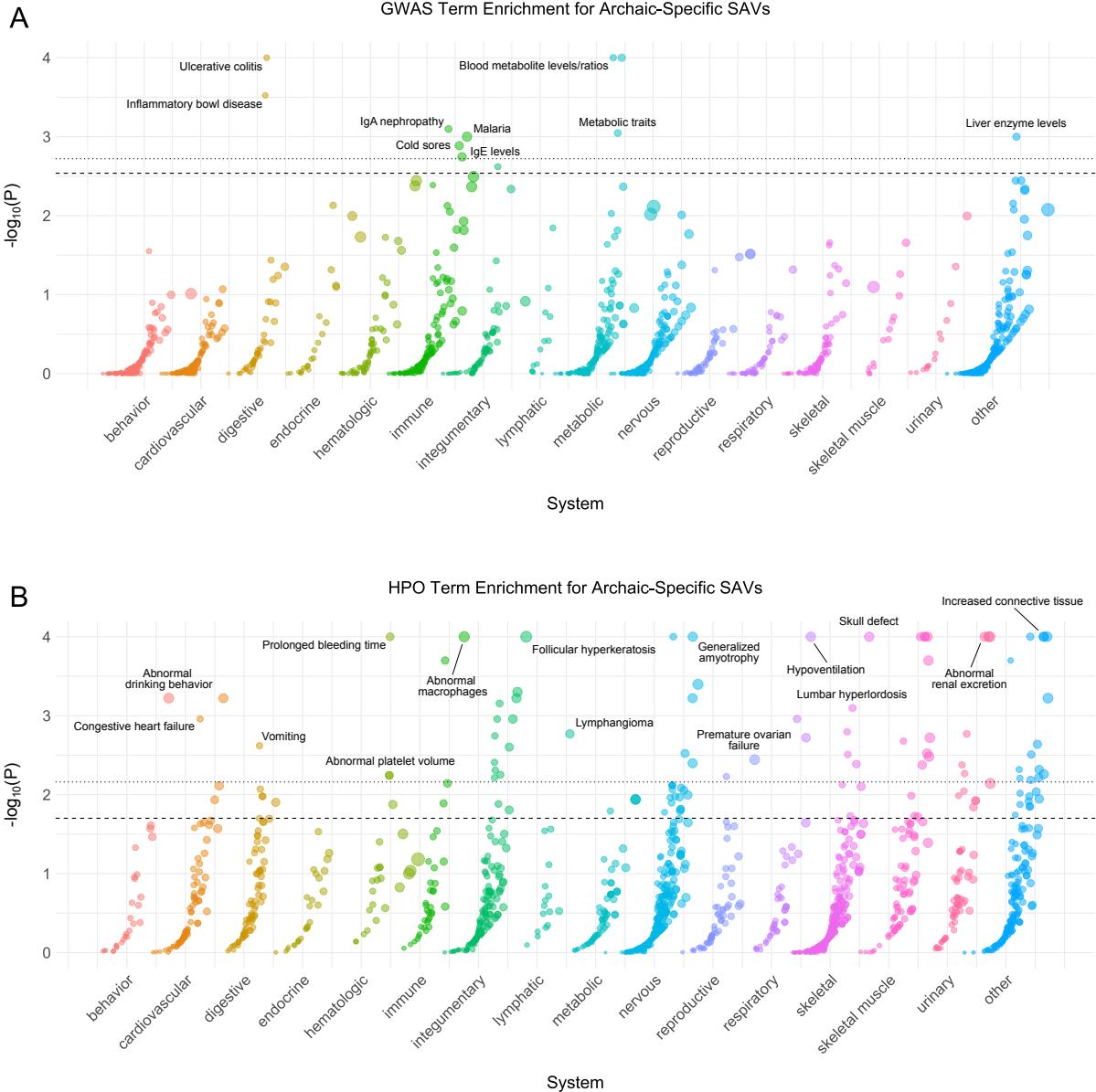
159 Given that we cannot directly observe archaic individuals, functions associated with genes  
160 with archaic-specific SAVs are of particular interest. We found enrichment for many phenotypes  
161 among the 2,012 genes with archaic-specific SAVs (**Figure 2; File S2**). The common pheno-  
162 type enrichments were dominated by immune (e.g., malaria resistance and ulcerative colitis)  
163 and metabolic (e.g., blood metabolite levels) traits. There were substantially more phenotypes  
164 enriched among genes associated with rare phenotypes (**Figure 2B**), and these included traits  
165 that are known to differentiate archaic hominins and modern humans, including skeletal traits  
166 such as lumbar hyperlordosis and several cranial features. At least one significantly enriched  
167 trait occurred in every system across the GWAS Catalog and Human Phenotype Ontology,  
168 except for the endocrine system.

169 Next, we sought to characterize similarities and differences among the archaic hominin indi-  
170 viduals. We assessed phenotype enrichment among genes that contained shared, Neanderthal-  
171 specific, and lineage-specific SAVs (**File S2**). We found minimal enrichment among the 152  
172 genes with shared SAVs (bronchitis and programulin levels) (**Figure S7**). However, there was  
173 considerable enrichment across various systems for Neanderthal- and lineage-specific SAVs  
174 (**Figure S8, Figure S9, Figure S10, Figure S11, Figure S12**). For example, all Neanderthals  
175 were enriched for SAVs in genes underlying skin conditions including abnormal skin blistering  
176 and fragile skin (**Figure S11**). The Denisovan exhibited enrichment for SAVs in genes associ-  
177 ated with many skeletal and skeletal muscle system traits including skull defects, spinal rigidity,  
178 abnormal skeletal muscle fiber size, increased muscle fiber diameter variation, and type 1 mus-  
179 cle fiber predominance (**Figure S10**). Only two traits were enriched in multiple different sets  
180 of lineage-specific SAVs at FDR-adjusted significance levels: 1) blood metabolite levels in the  
181 Altai and Vindija Neanderthal and 2) generalized amyotrophy in the Chagyrskaya Neanderthal  
182 and Denisovan.

## 183 2.4 Site-level evolutionary conservation varies across SAV origin.

184 Genes vary in their tolerance to mutation and SAVs often disrupt gene function and contribute to  
185 disease (Li et al., 2016; Scotti and Swanson, 2016; Li et al., 2017). To evaluate if the presence  
186 of archaic SAVs is associated with evolutionary constraint on genes, we compared the per  
187 gene tolerance to missense and loss-of-function variants from gnomAD (Karczewski et al.,  
188 2020) among ancient, archaic-specific, introgressed, and non-splice altered genes. In addition  
189 to constraint at the gene level, evolutionary constraint can be quantified at nucleotide level by  
190 methods like phyloP that quantify deviations from the expected neutral substitution rate at the  
191 site-level between species (Pollard et al., 2010). Thus, to explore the constraint on splicing  
192 altering variant sites themselves, we also compared their phyloP scores.

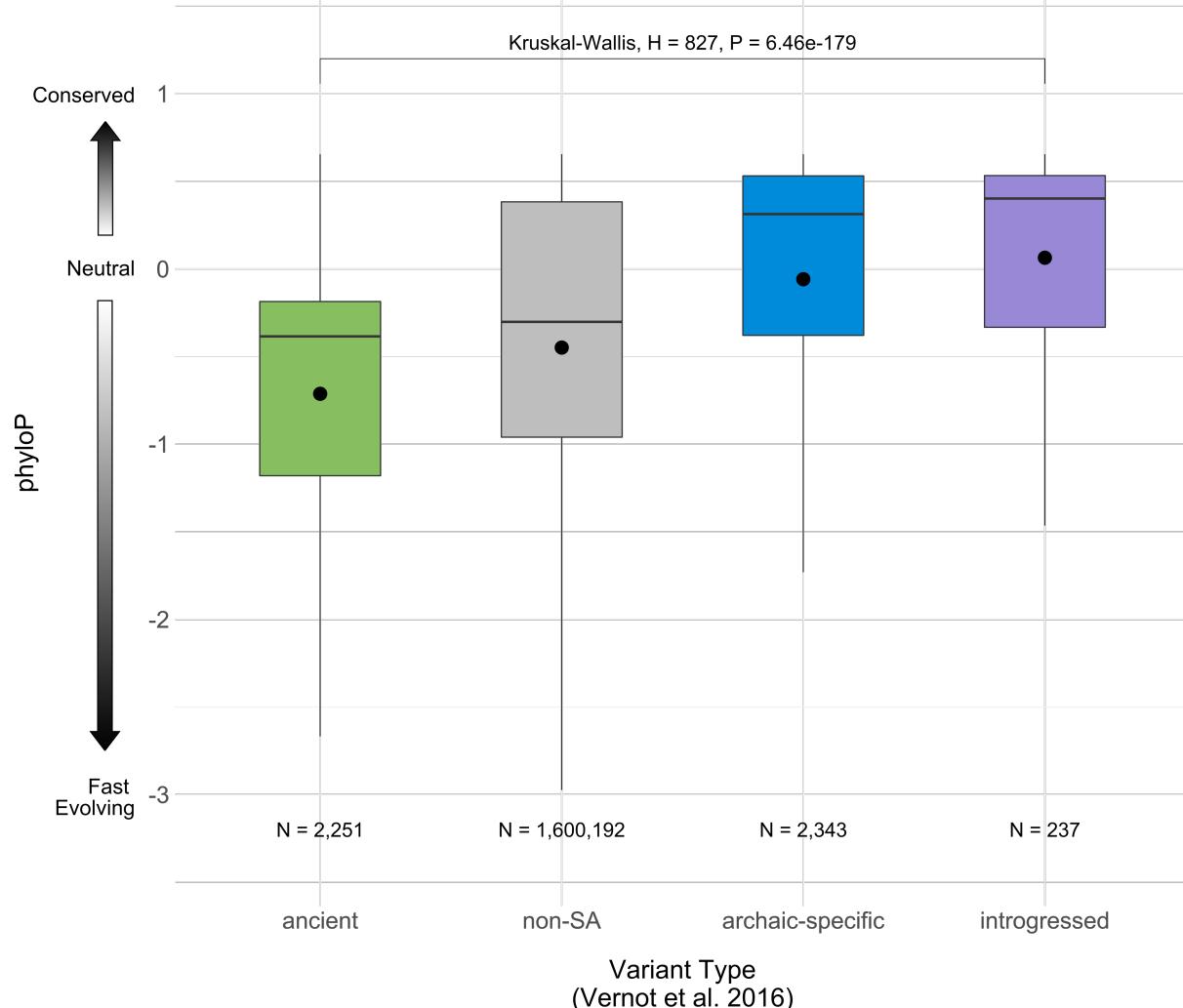
193 While we found a significant difference in the observed/expected number of missense vari-  
194 ants per transcript among genes with different classes of SAV (Kruskal-Wallis,  $H = 18.886$ ,  $P =$   
195 0.0003), the effect size was minimal (**Figure S13A**). Furthermore, genes with SAVs of different



**Figure 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 GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  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 File S2 for all phenotype enrichment results.

origins did not significantly differ in the observed/expected number of loss-of-function variants per transcript (Kruskal-Wallis,  $H = 1.697$ ,  $P = 0.638$ ) (**Figure S13B**). Variants classified per Browning et al., 2018 exhibited the same pattern (**Figure S13C**; **Figure S13D**). These results suggest that genes with alternative splicing in archaics are similar in their gene-level constraint to other genes.



**Figure 3: Site-level evolutionary conservation varies across SAVs with different origins.** phyloP evolutionary conservation score distributions for archaic SAVs of different origins and non-SAVs. Positive scores indicate substitution rates slower than expected under neutral evolution (conservation), while negative scores indicate higher substitution rates than expected (fast evolution). The boxplots give the first quantile, median, and third quantile of the distributions, and the mean is noted by the black point. Ns are the number of variants per set. SAV classifications were based on Vernot et al., 2016 introgression calls.

In contrast, phyloP scores were significantly different between ancient SAVs, archaic-specific SAVs, introgressed SAVs, and non-SAVs (Kruskal-Wallis,  $H = 826.862, P = 6.463 \times 10^{-179}$ ) (**Figure 3**). 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, they were mostly in the range consistent with neutral evolution. Variants classified per Browning et al., 2018 exhibited similar patterns (**Figure S14**); however, archaic-specific, rather than introgressed, variants had

210 a larger mean phyloP score.

## 211 2.5 The prevalence of SAVs across lineages is consistent with purifying selec- 212 tion on most SAVs.

213 Variants that disrupt splicing and/or produce new isoforms are expected to be under strong  
214 negative selection (e.g., Kriventseva et al., 2003). However, given differences in ages of archaic  
215 SAVs and the effective population sizes ( $N_e$ ) of the lineages in which they arose, different  
216 SAVs were likely exposed to different strengths of selection for different periods. Thus, we  
217 hypothesized that the probability a given SAV would survive to the present would vary based  
218 on its origin. For example, SAVs that arose in the ancestor all archaic lineages were likely  
219 subject to purifying selection over a longer time scale than lineage-specific SAVs, especially  
220 those that arose in lineages with low  $N_e$ .

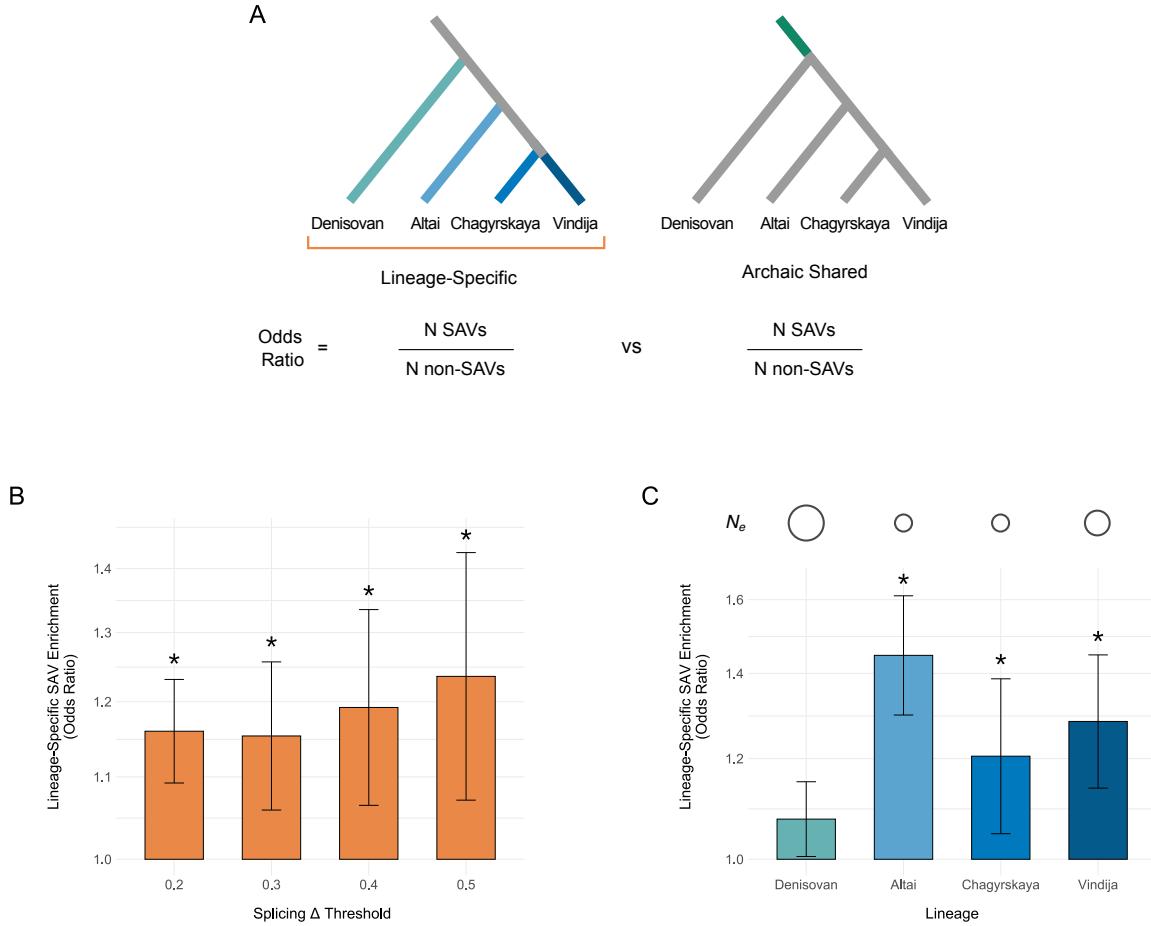
221 Shared archaic variants are depleted for SAVs compared to lineage-specific variants, and  
222 this depletion increased with higher splice altering probability thresholds (**Figure 4A**, **Figure 4B**,  
223 **Table S5**). This result is consistent with the hypothesis that most splicing variants are deleterious  
224 and that the longer exposure to negative selection for older variants results in a smaller  
225 fraction of remaining SAVs. It is also with concordant with the site-level constraint results.

226 Given that the population histories for each archaic lineage were likely different, we also  
227 compared within lineages. Neanderthals are thought to have lived in smaller groups and ex-  
228 hibited a lower  $N_e$  than Denisovans (Mafessoni et al., 2020). We tested this hypothesis by  
229 repeating the SAV enrichment test for variants specific to each individual archaic lineage (**Fig-**  
230 **ure 4A**). All three Neanderthals were significantly enriched for unique SAVs compared to shared  
231 archaic variants after Bonferroni correction (OR = 1.205-1.447, **Figure 4C**, **Table S6**). In con-  
232 trast, variants on the longer and higher  $N_e$  Denisovan lineage were not significantly enriched  
233 for SAVs (OR = 1.075). At the stricter high-confidence SAV threshold, both the Altai and Vindija  
234 Neanderthals remained significantly enriched with increased odds ratios (**Figure S15**).

## 235 2.6 Introgressed SAVs found in modern humans were present across archaics.

236 We hypothesized that the evolutionary history of SAVs might also influence their prevalence in  
237 modern human populations. For example, introgressed variants experienced strong negative  
238 selection in the generations immediately after interbreeding (Petr et al., 2019), so archaic SAVs  
239 that survived stronger and longer-term selection would be more likely to survive in modern hu-  
240 mans. To test this, we first considered the distribution of remaining introgressed variants among  
241 the archaics. Most introgressed SAVs were present in all Neanderthals (N = 141) or present in  
242 all archaics (N = 66; **Table S7**). No SAVs private to Vindija or Chagyrskaya nor shared between  
243 both late Neanderthals were identified as introgressed, even though Neanderthal ancestry in  
244 most modern humans is most closely related to Vindija and Chagyrskaya (Prüfer et al., 2017;  
245 Mafessoni et al., 2020). This is consistent with weaker selection on lineage-specific SAVs and  
246 previous work suggesting that older introgressed archaic variants were more tolerated in hu-  
247 mans (Telis et al., 2020; McArthur et al., 2021).

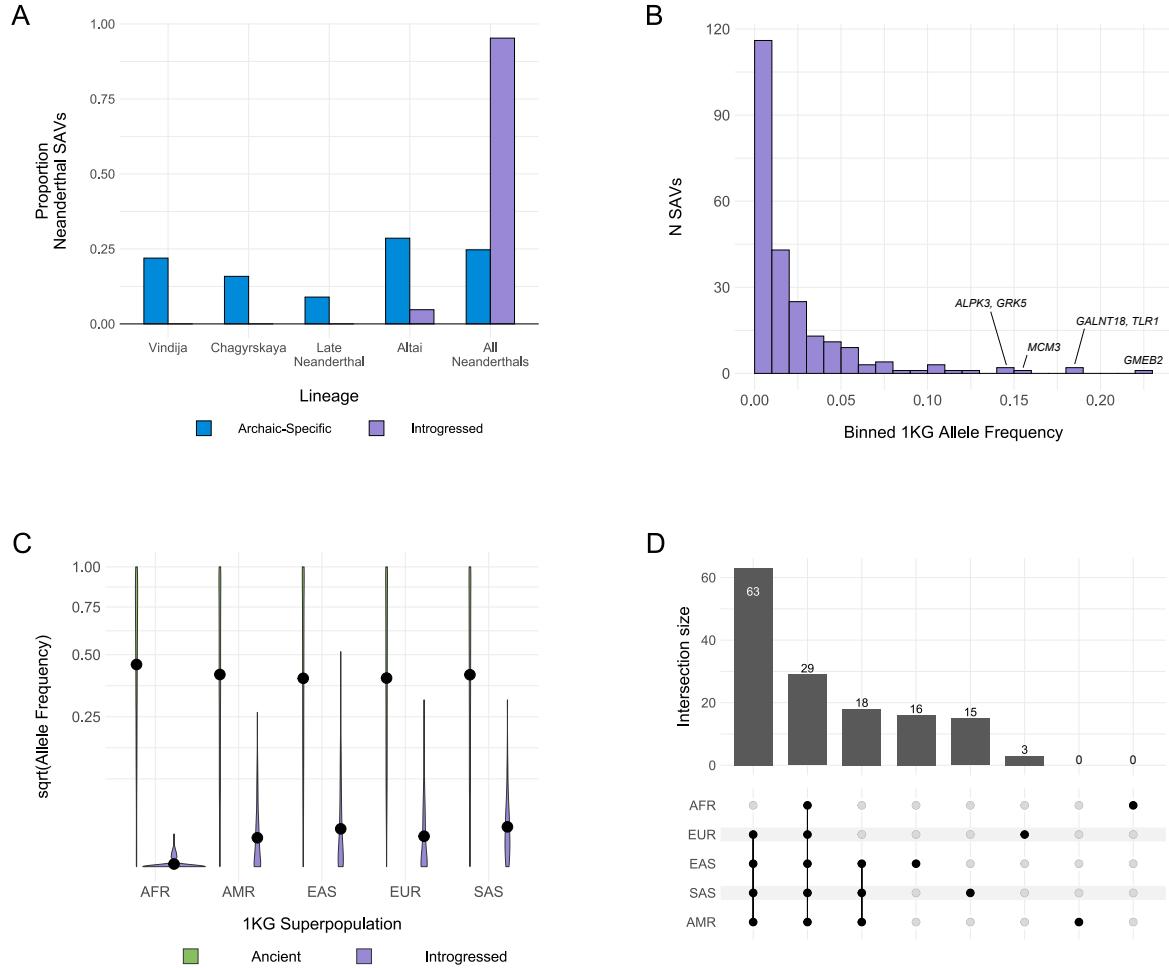
248 To further test this, we compared the expected origin distribution for introgressed SAVs  
249 (based on the distribution of archaic-specific SAVs) to the observed distribution for introgressed  
250 SAVs. Fewer Altai-specific SAVs occur among introgressed variants whereas shared Nean-



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

**(A)** We hypothesized that lineage-specific archaic variants would be enriched for SAVs compared to variants shared across the four archaic individuals, 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 specific archaic lineages are enriched for SAVs compared to variants shared among all four archaic lineages. The enrichment increases at increasing splice altering probability ( $\Delta$ ) thresholds. Note the y-axis is  $\log_{10}$  transformed. **(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 Mafessoni et al., 2020. In all panels, asterisks reflect significance of Fisher's exact tests using a Bonferroni corrected  $\alpha$  (0.0125). Error bars denote the 95% CI. The number of variants used in each enrichment test are listed in **Table S5** and **Table S6**. Note the y-axis is  $\log_{10}$  transformed.

251 Neanderthal SAVs are more prevalent than expected (**Figure 5A**). This pattern remains for high-  
 252 confidence and Browning et al., 2018 SAVs (**Figure S16**). These patterns suggest that older  
 253 SAVs, either those that evolved prior to the Neanderthal common ancestor or prior to the  
 254 Denisovan and Neanderthal common ancestor were the most tolerated after introgression.



**Figure 5: Introgressed SAVs present in modern humans were shared across archaic individuals.**

(A) Histograms comparing distributions of the presence of all Neanderthal SAVs (blue) and introgressed SAVs (purple), as defined by Vernot et al., 2016, 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 **Table S7**. (B) Allele frequency distribution for introgressed SAVs per Vernot et al., 2016 with  $\Delta \geq 0.2$ . Allele frequencies represent the mean from the AFR, AMR, EAS, EUR, SAS frequencies from the Vernot et al., 2016 metadata. (C) Allele frequency distributions for SAVs present in both archaic and modern individuals stratified by 1KG superpopulation and origin (ancient vs. introgressed). The black 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 **Figure S21** for all sets.

255 Consistent with known introgression patterns, introgressed SAVs occurred at lower overall  
 256 frequencies (**Figure 5B**). However, a small number of introgressed SAVs occur at modest to  
 257 high frequencies among genes including *GMEB2*, *GALNT18*, and *TLR1* (**Figure 5B**). The latter  
 258 occurs in an adaptively introgressed locus spanning three toll-like receptors—key components  
 259 of the innate immune system (Dannemann et al., 2016). In contrast, ancient SAVs occur at

260 high frequencies in all five 1KG superpopulations, and their frequencies are significantly higher  
261 among Africans ( $\mu = 0.522$ ) than non-Africans ( $\mu = 0.476$ ) (Mann-Whitney U,  $U = 10,963,956$ ,  
262  $P = 2.66 \times 10^{-99}$ ) (**Figure 5C, Figure S17**). Furthermore, most ancient SAVs are present at  $\geq$   
263 0.05 frequency in all five 1KG superpopulations **Figure S18, Figure S19**. Introgressed SAVs  
264 have significantly lower frequencies in all superpopulations (**Figure 5B**) and are less likely to  
265 be shared among multiple populations (**Figure 5C, Figure S20, Figure S21**).

266 It is possible these patterns reflect the general attributes of introgressed variants, rather than  
267 splicing effects of SAVs. We therefore examined the relationship between allele frequency in  
268 1KG and splice altering probability ( $\Delta_{\text{max}}$ ) for introgressed SAVs. 1KG populations did not  
269 generally differ in  $\Delta_{\text{max}}$  for either ancient or introgressed SAVs, although introgressed SAVs  
270 had a higher  $\Delta_{\text{max}}$  (**Figure S22**). We anticipated, however, SAVs that were more likely to  
271 be splice altering would occur at lower frequencies. Indeed, we found a significant negative  
272 association between allele frequency and  $\Delta_{\text{max}}$  for  $\Delta \geq 0.2$  (Spearman,  $\rho = -0.2362$ ,  $P =$   
273 0.0002) (**Figure S23**).

## 274 **2.7 Introgressed SAVs have limited functional associations.**

275 We tested whether any functional annotations were enriched among the 353 genes with Brown-  
276 ing et al., 2018 introgressed SAVs and 232 genes with Vernot et al., 2016 introgressed SAVs  
277 (**File S2**). We found little enrichment among any GWAS or HPO terms (**Figure S24, Figure S25**). Only two terms were significantly enriched among genes with Vernot et al., 2016  
278 introgressed SAVs: adverse response to breast cancer chemotherapy (GWAS) and Oligohy-  
279 dramnios (HPO) (**Figure S24**). Only one HPO term, abnormal hip-girdle musculature, was  
280 enriched among genes with Browning et al., 2018 introgressed SAVs (**Figure S25A**). How-  
281 ever, four GWAS terms met our FDR-corrected significance threshold: asthma and hay fever,  
282 diabetes in response to drug treatment, *H. pylori* serologic status, and systemic sclerosis (**Fig-**  
283 **ure S25B**). Overall, these results suggest that most SAVs surviving in modern human popula-  
284 tions are unlikely to have large phenotypic or fitness effects.  
285

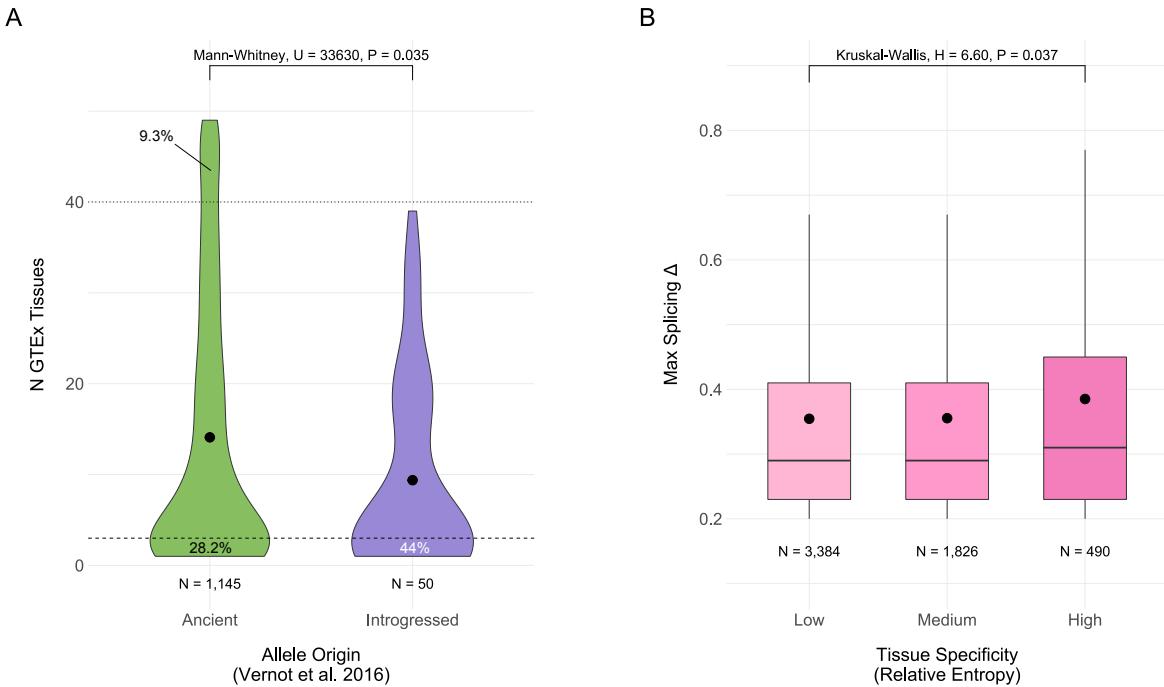
286 We further considered the potential functional effects of introgressed SAVs by intersecting  
287 the with Neanderthal variants exhibiting allele-specific expression (ASE) in modern humans  
288 (McCoy et al., 2017). We identified 16 SAVs out of 1,236 ASE variants, including variants in  
289 *GSDMC*, *HSPG2*, and *RARS* (**Table S8**). We also note that a handful of the McCoy et al.,  
290 2017 variants fell just under our SAV threshold, including a variant (rs950169) in *ADAMTSL3*  
291 highlighted in the ASE study with correctly predicted acceptor loss ( $\Delta = 0.19$ ) (McCoy et al.,  
292 2017).

## 293 **2.8 Introgressed SAVs are more tissue-specific than ancient SAVs.**

294 Given their different histories of selective pressures, we hypothesized that introgressed SAVs  
295 would be more tissue-specific than ancient SAVs in their effects. To explore this, we identified  
296 1,381 archaic SAVs with sQTL data from GTEx across 49 tissues.

297 Introgressed sQTL SAVs were significantly associated with tissue specific gene expression  
298 compared to ancient sQTL SAVs (Mann-Whitney U,  $U = 33,629.5$ ,  $P = 0.035$ ) (**Figure 6A**). On  
299 average, introgressed SAVs influenced splicing in 4.72 fewer tissues than ancient SAVs. Fur-  
300 ther, all sQTL SAVs with broad effects ( $> 40$  tissues) were ancient (107 high-confidence and

301 5 low-confidence). 74 of these were shared among all four archaics (**Table S9**), suggesting  
302 that core sQTL SAVs were more likely to evolve in the deep past. These patterns were also  
303 observed among the Browning et al., 2018 variants (**Figure S26**). Collectively, 30% of sQTL  
304 SAVs ( $N = 427$ ) were associated with tissue specific effects on splicing (1 or 2 tissues) (**Fig-**  
305 **ure S27**). Consistent with known gene expression patterns, testis had the most unique sQTL  
306 among SAVs, followed by skeletal muscle and thyroid.



**Figure 6: Increased tissue-specificity is associated with introgressed SAVs and with increased probability of alternative splicing.**

(A) The distribution of the number of GTEx tissues in which an ancient or introgressed SAV (per Vernot et al., 2016) was identified as an sQTL. Introgressed variants are significantly more tissue-specific (Mann-Whitney  $U = 33630; P = 0.035$ ). 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 SAV group. (B) The distributions of maximum splice altering probability ( $\Delta$ ) for SAVs in 4,061 genes binned by tissue-specificity of expression. We quantified the tissue-specificity of each gene as the relative entropy of expression levels across 34 tissues from GTEx compared to genes overall. Low tissue specificity reflects relative entropy  $\leq 0.1$ , medium tissue specificity reflects relative entropy  $> 0.1$  and  $\leq 0.5$ , and high tissue specificity reflects relative entropy  $> 0.5$  (**Figure S28**).

307 Variation in gene expression among tissues may also influence the efficacy of negative  
308 selection to remove deleterious SAVs. For example, Saudemont et al., 2017 demonstrated that  
309 more ubiquitously expressed genes in *Paramecium tetraurelia* exhibited less alternative splicing  
310 compared to genes with more tissue-specific expression due to the differences in the efficacy of  
311 negative selection. We predicted that tissue-specificity of expression would be associated with  
312 the number of SAVs per gene or maximum  $\Delta$ . We quantified tissue-specificity of expression  
313 using the relative entropy of the TPM count for each gene across tissues compared to the  
314 expression distribution across tissues for all genes in GTEx. This metric ranges from 0 to 1,  
315 with higher values reflecting greater tissue specificity. Most genes exhibited broad expression  
316 (**Figure S28**), so we divided genes into low, medium, and high tissue-specificity based on the  
317 relative entropy.

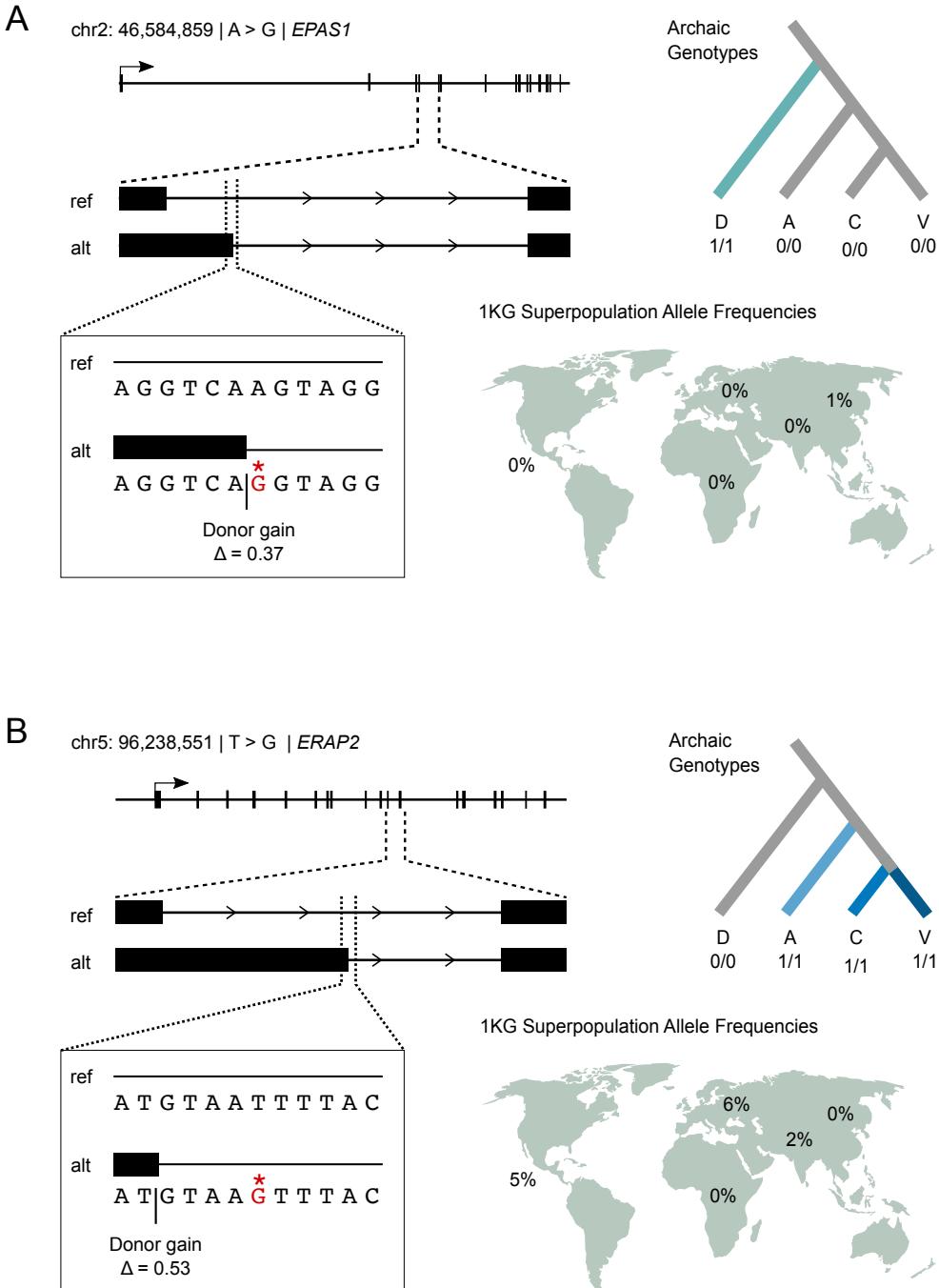
318 Genes with the most tissue-specific expression had significantly higher median maximum  
319 splice altering probability ( $\Delta$ ) than genes with broader expression patterns (**Figure 6B**; Kruskal-  
320 Wallis,  $H = 6.599$ ,  $P = 0.037$ ). This could indicate greater selection against SAVs likely to influ-  
321 ence expression across many tissues; however, we note that the effect was small in magnitude.  
322 The distribution of the number of archaic SAVs per gene did not differ significantly between en-  
323 tropy classes (Kruskal-Wallis,  $H = 1.89$ ,  $P = 0.388$ ) (**Figure S29**); all had a median of 1 SAV.

### 324 2.9 Archaic SAVs with potential evolutionary significance.

325 Many archaic SAVs influence genes with known or previously hypothesized significance to the  
326 evolutionary divergence between archaic hominins and modern humans. For example, the 2'-5'  
327 oligoadenylylate synthetase *OAS* locus harbors an adaptively introgressed splice altering variant  
328 at chr12:113,357,193 (G > A) that disrupts an acceptor site and results in multiple isoforms  
329 and leads to reduced activity of the antiviral enzyme (Mendez et al., 2012; Sams et al., 2016).  
330 This ancestral variant was reintroduced into modern Eurasian populations by Neanderthal  
331 introgression (Rinker et al., 2020). SpliceAI correctly predicted the acceptor loss at this site ( $\Delta =$   
332 0.89). This locus harbors 92 additional archaic variants ( $N = 92$ ). We found one additional SAV  
333 at chr12:113,355,275 in *OAS1* that potentially results in an acceptor gain ( $\Delta = 0.26$ ). This allele  
334 was unique to the Denisovan, is derived, and was present in only one of 2,054 1KG samples  
335 as a heterozygote. This suggests potential further splice variant evolution of this locus, with  
336 possible Denisovan-specific effects.

337 We also identified several variants at other well-studied loci. Variation in human popula-  
338 tions at the *EPAS1* locus includes a Denisovan-introgressed haplotype thought to contribute to  
339 adaptation to living at high altitude among Tibetans (Huerta-Sánchez et al., 2014). Of 184 ar-  
340chaic variants occurring at this locus, we identified two as candidate SAVs. One variant (chr2:  
341 46,584,859; rs372272284) is fixed in the Denisovan whereas all Neanderthals have the hu-  
342 man reference allele (**Figure 7A**). The variant is introgressed and present at low frequency  
343 in East Asians in 1KG, and is also the lead variant in an observed association of the intro-  
344 gressed haplotype with decreased hemoglobin levels in Tibetans (Jeong et al., 2018). This  
345 SAV, rs372272284, strengthens a canonical 5' splice site (CAA|GT to CAG|GT) (Sibley et al.,  
346 2016), resulting in a donor gain ( $\Delta = 0.37$ ) (**Figure 7A**). If used, this splice site would introduce  
347 multiple stop codons, resulting in nonsense mediated decay (NMD) of the mRNA. This would  
348 result in the same molecular effect (decreased circulating *EPAS1* RNA) that is thought to con-  
349 tribute to hypoxia adaptation (Peng et al., 2017). The other candidate SAV (chr2: 46,610,904)  
350 is absent from 1KG and occurs as a heterozygote in the Altai Neanderthal, and is near the  
351 end of the last intron of the gene, making it much less likely to fundamentally alter the mRNA  
352 product.

353 We also identified three archaic SAVs within *ERAP2*, a gene subject to strong and consis-  
354 tent balancing selection in different human populations (Andrés et al., 2010). SpliceAI correctly  
355 identified a previously characterized human variant (rs2248374), which causes a donor loss  
356 ( $\Delta = 0.51$ ) and results in a truncated protein and subsequent NMD of the mRNA. However, we  
357 identified an additional Neanderthal SAV, which is also introgressed and occurs at low frequen-  
358 cies among Americans (5%), Europeans (6%), and South Asians (2%) in 1KG (**Figure 7B**).  
359 This SAV, rs17486481, is a donor gain ( $\Delta = 0.53$ ) that introduces a canonical 5' splice site  
360 (AT|GTAAT to AT|GTAAG) and would similarly result in NMD (**Figure 7B**). However this al-



**Figure 7: 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 at least one stop codon, 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 at least one stop codon.

361     lele always occurs with the non-truncated version of rs2248374 (while being much rarer), and  
362     the need to maintain the non-truncated allele is most likely why it remains uncommon. The  
363     third variant (chr5: 96,248,413) was archaic-specific—occurring as a heterozygote in the Altai  
364     Neanderthal—and results in an acceptor gain ( $\Delta = 0.24$ ).

### 365     3 Discussion

366     Alternative splicing plays a critical role in organismal biology, particularly during development  
367     and establishing tissue identity (Baralle and Giudice, 2017). Thus, alternative splicing often  
368     contributes to adaptation and phenotypic differences between closely related species (Blekhman  
369     et al., 2010; Barbosa-Moraes et al., 2012; Merkin et al., 2012; Singh and Ahi, 2022; Verta and  
370     Jacobs, 2022; Wright et al., 2022). The development of machine learning algorithms that can  
371     predict alternative splicing from sequence alone now enables analysis of alternative splicing in  
372     populations for which transcriptomic data are difficult or impossible to generate, including ar-  
373     chaic hominins. Here, we use SpliceAI to uncover the previously unobservable genome-wide  
374     alternative splicing landscape of archaic hominins.

375     We identify thousands of putative splice altering variants from the high-coverage genomes  
376     of three Neanderthals and a Denisovan. We find that many of these variants do not occur  
377     in modern humans, and we propose that they are implicated in specific phenotypic differences  
378     between archaic hominins and modern humans. Additionally, many SAVs are shared with mod-  
379     ern humans and are ancient, evolving before the common ancestor of archaic hominins and  
380     modern humans. Furthermore, a few hundred SAVs are present in human populations due to  
381     introgression, and these surviving introgressed SAVs are almost entirely shared across Nean-  
382     derthals. We also observe multiple lines of evidence supporting the role of negative selection  
383     in shaping SAV patterns.

384     Given that introgressed and ancient SAVs are present in modern humans, their splicing  
385     patterns have the potential to be directly studied to further understanding of their phenotypic  
386     effects. We found that 35% of archaic SAVs were identified in modern humans in GTEx as  
387     sQTLs. There are several reasons why archaic SAVs might not have been detected as sQTL.  
388     Splicing is influenced by sequence, but is also a somewhat stochastic process influenced by  
389     other cellular dynamics, such as polymerase pausing (Sibley et al., 2016). This, coupled with  
390     the limited power in many GTEx tissues, particularly for SAVs at low frequency in Europeans,  
391     means that all splice-altering variants would not necessarily be detected. Indeed, we observe  
392     higher fractions of SAVs as sQTL when focusing on high frequency variants (**Figure S30**).  
393     Nonetheless, the degree to which SAV tissue-specificity in modern humans reflects specificity  
394     in archaic hominins is unknown without further experimental study. However, such studies are  
395     challenging because the genomic and archaic cellular context cannot be perfectly replicated  
396     (i.e., a Neanderthal genome in a Neanderthal tissue).

397     Our results offer new insight into an essential molecular mechanism and previously unstud-  
398     ied attributes of archaic hominins; however, we note some limitations of our approach. First, we  
399     did not include structural variants (InDels) or variants from the sex chromosomes in this anal-  
400     ysis, both of which warrant further study. For example, the X chromosome exhibits high levels  
401     of alternative splicing (Karlebach et al., 2020) and splicing can occur in a sex-specific manner  
402     (Blekhman et al., 2010; Rogers et al., 2021). Future development and application of models

403 with sex-specific transcriptomic data may offer additional phenotypic insights but, for now, we  
404 are left with high-coverage sequences from only females. Second, the modern humans used  
405 in this analysis have considerable genetic diversity, but some additional variants classified as  
406 “archaic-specific” could occur in modern humans not included here. Even so, they would be  
407 rare and a significant number of archaic-specific variants would remain. Third, the tag SNPs  
408 and modern human samples used in this analysis are best suited to identifying Neanderthal  
409 rather than Denisovan introgression (Browning et al., 2018; Vernot et al., 2016). Our conserva-  
410 tive approach means that the number of introgressed SAVs reported here is an underestimate  
411 and does not include Denisovan-derived SAVs. Multiple modern human populations contain  
412 considerable Denisovan ancestry, therefore, future work should consider these variants.

413 In summary, our approach of combining machine learning with ancient DNA and modern  
414 population genetic data identifies thousands of archaic variants that potentially alter splicing,  
415 including many that appear to be specific to archaic hominins. Genes affected by archaic SAVs  
416 are enriched for roles in a variety of phenotypes and several influence loci with known rele-  
417 vance to recent human evolution. For example, two archaic SAVs that we highlight likely cause  
418 in NMD of the resulting *EPAS1* and *ERAP2* transcripts. Downregulation of *EPAS1* is thought  
419 to underlie high-altitude adaptation in Tibetans (Peng et al., 2017). In *ERAP2*, another variant  
420 in human populations that induces NMD has experienced strong balancing selection (Andrés  
421 et al., 2010). These examples underscore that phenotypic effects from alternative splicing are  
422 not limited to expanded proteomic diversity, but also downregulation of gene expression via  
423 NMD (Ge and Porse, 2014; Smith and Baker, 2015). Further work is needed to understand the  
424 functional effects of these and other archaic SAVs. Nonetheless, our results suggest that alter-  
425 native splicing played a role in hominin divergence and offers specific molecular hypotheses for  
426 testing. We also anticipate that our sequence-based approach will enable study of alternative  
427 splicing in other extinct or difficult to sample taxa.

## 428 4 Methods

### 429 4.1 Archaic Genomic Data

430 We retrieved four high-coverage publicly available archaic hominin genomes representing three  
431 Neanderthals (Mafessoni et al., 2020; Prüfer et al., 2014; Prüfer et al., 2017) and a Denisovan  
432 (Meyer et al., 2012).

433 We excluded sites that were invariant among the archaic individuals (“ALT =.”) and variants  
434 with low site quality (“QUAL < 30”). Further, low quality genotypes were set to missing based  
435 on read depth (“FMT/DP < 10”) and genotype quality (“FMT/GQ < 30”). We also normalized  
436 InDels and split multi-allelic records into separate entries for positions with multiple variants  
437 (“norm -m -”). All filtering was completed using bcftools, version 1.13 (Li, 2011).

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

### 440 4.2 Variant Annotation

441 We annotated variants for putative alternative splicing using SpliceAI, version 1.3.1 (Jaganathan  
442 et al., 2019). Briefly, SpliceAI uses a deep residual neural network to estimate the splice alter-

443 ing probability and position change of each variant from DNA sequence alone making it ideal  
444 for studying archaic hominins, for which we cannot obtain transcript-level data. The model  
445 considers 5 kbp flanking the variant in both directions. The output includes four splice altering  
446 probabilities ( $\Delta s$ ) for 1) acceptor gain (AG), 2) acceptor loss (AL), 3) donor gain (DG), and 4)  
447 donor loss (DL) as well as the position changes for each of the four deltas.  $\Delta s$  range from 0 to  
448 1 and represent the likelihood a variant is splice altering for one or more of the four categories.  
449 We implemented SpliceAI in a Conda package using keras, version 2.3.1 (Chollet et al., 2015)  
450 and tensorflow, version 1.15.0 (Abadi et al., 2015).

451 After filtering, we ran SpliceAI on sets of  $5 \times 10^3$  variants using the hg19 reference genome  
452 using the GENCODE, Human Release 24, annotations (Harrow et al., 2012) included with  
453 the package. We concatenated all results and further split variants with multiple annotations.  
454 Among all variants, 32,105 exhibited multiple annotations with different effects on splicing (**Ta-**  
455 **ble S11**). While we included InDels and variants on the X chromosome in this scan, we re-  
456 stricted all downstream analyses to autosomal SNVs (Discussion).

### 457 4.3 Defining SAVs

458 For each variant, we identified the maximum SAP ( $\Delta$ ) among all four classes: AG, AL, DG,  
459 and DL. We then defined SAVs using two  $\Delta$  thresholds:  $\Delta_{\text{max}} \geq 0.2$  and 0.5, “SAVs” and  
460 “high-confidence SAVs”, respectively.

461 We determined whether the number of SAVs identified in each archaic individual was dif-  
462 ferent than expected by randomly selecting a sample from 24 1KG populations. We annotated  
463 all variants present among these individuals using SpliceAI and the hg38 annotations included  
464 with the package. We then analyzed the variants as for the archaics (i.e., splitting multi-allelic  
465 sites and variants with multiple GENCODE annotations). We filtered for variants with a  $\Delta_{\text{max}}$   
466  $\geq 0.2$  and summed the number of variants per 1KG sample that had at least one alternate allele  
467 present.

### 468 4.4 Archaic Variants in Modern Humans

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

474 We assessed whether any variants present among the archaics are also present in modern  
475 humans using biallelic SNVs and InDels from 1KG, Phase 3 (Lowy-Gallego et al., 2019). We  
476 used LiftOver (Hinrichs et al., 2006) to convert variants from hg19 to hg38. We then normal-  
477 ized the 1KG variants (“norm -m - -f hg38.fa”) and subset variants to those within gene bodies  
478 (“view -R genes.bed”). We queried these variants for allele count, allele number, and allele  
479 frequencies (“query -f”). Allele frequencies included the overall allele frequency and an allele  
480 frequency per 1KG superpopulation: Africa (AFR), Americas (AMR), East Asia (EAS), Europe  
481 (EUR), and South Asia (SAS). These precomputed values had been rounded to two decimal  
482 places in the VCFs. Normalization, filtering, and querying was done using bcftools (Li, 2011).  
483 After using LiftOver to convert back to hg19 coordinates, we merged the 1KG variants with the

484 archaic variants ensuring the archaic and 1KG reference and alternate alleles matched. We re-  
485 calculated the allele frequency for Africans as the annotated frequency included samples from  
486 an admixed African population: African ancestry in SW USA (ASW). We subset samples from  
487 Esan (ESN), Mandinka (GWD), Luhya (LWK), Mende (MSL), and Yoruban (YRI) and calculated  
488 allele frequency as allele count divided by allele number per site.

489 We used two datasets to identify introgressed variants: Vernot et al., 2016 and Browning  
490 et al., 2018. These datasets differ in their approach to recognizing introgressed sequences  
491 and partly overlap the archaic variants considered in this study. Vernot et al., 2016 used the S\*  
492 statistic to classify human sequences as introgressed. S\* leverages high linkage-disequilibrium  
493 among variants in an admixed target population that are absent in an unadmixed reference  
494 population (Plagnol and Wall, 2006; Vernot and Akey, 2014). Introgressed haplotypes are then  
495 identified by maximizing the sum of scores among all SNP subsets at a particular locus (Plagnol  
496 and Wall, 2006; Vernot and Akey, 2014). Tag SNPs are those variants that match an archaic  
497 allele and occur with at least two other tag SNPs in a 50 Kb window. Haplotypes were defined  
498 as regions encompassing  $\geq 5$  tag SNPs in LD within a given human population ( $R^2 \geq 0.8$ ). We  
499 collated tag SNPs from all four populations: East Asian (ASN), European (EUR), Melanesian  
500 (PNG), and South Asian (SAS). We retained all metadata from Vernot et al., 2016. A handful  
501 of tag SNPs encompass multiple haplotypes that reflect differences in haplotype size between  
502 modern human populations; we retained the first record per variant. Browning et al., 2018 de-  
503 veloped a modified S\* statistic, Sprime, which employs a scoring method that adjusts the score  
504 based on the local mutation and recombination rates, allows for low frequency introgression  
505 in the unadmixed outgroup, and avoids windowing to identify introgressed segments. We col-  
506 lated introgressed variants for 20 non-African populations and filtered for those that matched  
507 the Altai Neanderthal at high-quality loci.

508 A handful of sequences in the hg19 reference genome are introgressed from archaic ho-  
509 minins. Therefore, we maximized the number of introgressed sites we could analyze by defining  
510 sites rather than variants as introgressed if either the reference or alternate allele for each SAV  
511 matched any Neanderthal base at a matching position. We ensured that SpliceAI predictions  
512 were similar for these allele pairs, regardless of which was the reference and alternate, by  
513 generating a custom hg19 sequence where introgressed reference alleles ( $N = 7,977$ ) from  
514 Vernot et al., 2016 were replaced by the alternate allele using a custom script. We then ap-  
515 plied SpliceAI to the introgressed reference alleles, now considered to be the alternate. We  
516 found that 24 of the 26 variants were classified as SAVs (**Table S10**). One of the remain-  
517 ing two variants was nearly identical in splicing probability ( $\Delta \text{ max} = 0.19$  and  $\Delta \text{ max} = 0.2$ ),  
518 whereas the other variant's predictions were different ( $\Delta \text{ max} = 0.16$  and  $\Delta \text{ max} = 0.31$ ) (**Ta-**  
519 **ble S10**). Given this overall similarity, we maintained the original predictions for introgressed  
520 Vernot et al., 2016 reference alleles in our dataframe but provide the predictions when these  
521 nucleotides are the alternate allele for all SAVs and non-SAVs in the project GitHub repository.  
522 We recalculated allele frequencies for all introgressed variants to account for sites where the  
523 reference sequence contained introgressed alleles, as the precomputed 1KG allele frequency  
524 would be incorrect. The Browning et al., 2018 metadata designate whether the reference or  
525 alternate allele is introgressed. Therefore, we used the 1KG allele frequency for sites with an  
526 introgressed alternate allele and subtracted the 1KG allele frequency from 1 for sites with an  
527 introgressed reference allele. For Vernot et al., 2016 introgressed variants, we calculated an  
528 average from the metadata, which included the allele frequencies in various populations for the

529 introgressed allele. We took the mean of the AFR, AMR, EAS, EUR, and SAS frequencies for  
530 all introgressed positions.

531 We categorized each variant's "origin" based on presence in 1KG and whether or not the  
532 variant was introgressed. Further, we classified each variant's allele origin based on intro-  
533 gressed variants identified by Vernot et al., 2016 "Vernot allele origin" and Browning et al., 2018  
534 "Browning allele origin" due to the incomplete overlap among variants in those datasets. Vari-  
535 ants that did not occur in 1KG were defined as "archaic-specific". Variants that were present  
536 in 1KG and introgressed were defined as "introgressed". Variants that were present in 1KG  
537 but not introgressed were considered "ancient" at two confidence levels. "High-confidence an-  
538 cient" variants were present in at least two 1KG superpopulations at an allele frequency  $\geq 0.05$ ,  
539 while "low-confidence ancient" variants did not meet this threshold. We report analyses on the  
540 high-confident ancient set; this helps to remove cases of potential convergent mutation. We  
541 identified a small number of variants that were classified as introgressed by Vernot et al., 2016  
542 and Browning et al., 2018 but absent in 1KG: 22 and 12, respectively. We maintained the  
543 "archaic-specific" category for these variants. Additionally, a handful of variants ( $N = 82$ ) were  
544 absent in 1KG but present in GTEx. We reclassified these variants from "archaic-specific" to  
545 "low-confidence ancient".

## 546 4.5 Gene Characteristics, Mutation Tolerance, and Conservation

547 We used the SpliceAI annotation file for hg19 from GENCODE, Human Release 24 (Harrow et  
548 al., 2012), to count the number of exons per gene and calculate the length in bp of the gene body  
549 and the coding sequence. The number of isoforms per gene were retrieved from GENCODE,  
550 Human Release 40. We retrieved missense and loss-of-function (LoF) observed/expected ra-  
551 tios from gnomAD (Karczewski et al., 2020) to quantify each gene's tolerance to mutation. We  
552 also considered conservation at the variant level. We used the primate subset of the 46 way  
553 multi-species alignment (Pollard et al., 2010). Positive phyloP scores indicate conservation or  
554 slower evolution than expected, whereas negative phyloP scores indicate acceleration or faster  
555 evolution than expected based on a null hypothesis of neutral evolution.

## 556 4.6 Phenotype Enrichment

557 We followed the approach of McArthur et al., 2022 to assess enrichment for SAVs in genes  
558 implicated in different human phenotypes. Many gene enrichment analyses suffer from low  
559 power to detect enrichment because an entire genome is used as the null distribution. Relat-  
560 edly, SAVs are unevenly distributed throughout archaic genomes. We addressed this issue  
561 by generating a null distribution from the observed data. We first retrieved phenotypes and  
562 the associated genes per phenotype from Enrichr (Chen et al., 2013; Kuleshov et al., 2016;  
563 Xie et al., 2021). We used both the 2019 GWAS Catalog and the Human Phenotype Ontol-  
564 ogy (HPO). The GWAS Catalog largely considers common disease annotations and has 1,737  
565 terms with 19,378 genes annotated (Buniello et al., 2019), whereas HPO largely considers rare  
566 disease annotations and has 1,779 terms with 3,096 genes annotated (Köhler et al., 2021). All  
567 3,516 terms were manually curated into one of 16 systems: behavioral, cardiovascular, diges-  
568 tive, endocrine, hematologic, immune, integumentary, lymphatic, metabolic, nervous, other,  
569 reproductive, respiratory, skeletal, skeletal muscle, and urinary.

570 We considered nine different gene sets, generated using SAVs with  $\Delta \geq 0.2$ , for our en-  
571 richment analyses: 1) genes with lineage-specific Altai SAVs ( $N = 310$ ), 2) genes with lineage-  
572 specific Chagyrskaya SAVs ( $N = 172$ ), 3) genes with lineage-specific Denisovan SAVs ( $N =$   
573 956), 4) genes with lineage-specific Vindija SAVs ( $N = 238$ ), 5) genes with SAVs present in all  
574 three Neanderthals ( $N = 268$ ), 6) genes with SAVs shared among all four archaics ( $N = 160$ ), 7)  
575 genes with all archaic-specific SAVs ( $N = 2,252$ ), 8) genes with introgressed SAVs per Vernot  
576 et al., 2016 ( $N = 237$ ), and 9) genes with introgressed SAVs per Browning et al., 2018 ( $N =$   
577 373). The shared set only included variants present in all four archaics and excluded those  
578 that were inferred from parsimony. We retained duplicated gene names to reflect genes with  
579 multiple SAVs.

580 We identified which genes were present in both the GWAS Catalog and HPO per set using  
581 a boolean to calculate the observed gene counts per term per ontology. We then removed  
582 GWAS and HPO terms per set that did not include at least one gene from the set. This resulted  
583 in 713 Altai, 1,437 archaic-specific, 734 Browning et al., 2018 introgressed, 417 Chagyrskaya,  
584 1,046 Denisovan, 553 Neanderthal, 420 shared, 614 Vernot et al., 2016 introgressed, and 519  
585 Vindija terms for the 2019 GWAS Catalog and 638 Altai, 1,510 archaic-specific, 720 Browning  
586 et al., 2018 introgressed, 391 Chagyrskaya, 1,185 Denisovan, 571 Neanderthal, 389 shared,  
587 591 Vernot et al., 2016 introgressed, and 533 Vindija terms for the Human Phenotype Ontology.

588 The max  $\Delta$  was then shuffled across all 1,607,350 variants without modifying the annotation,  
589 allele origin, or distribution data. The distribution of genes for both ontologies was then  
590 recorded. We repeated this process  $1 \times 10^4$  times per set and calculated enrichment as the  
591 number of observed genes divided by the mean empirical gene count per term. p-values were  
592 calculated as the proportion of empiric counts + 1  $\geq$  the observed counts + 1. We adjusted  
593 our significance level due to multiple testing by correcting for the false discovery rate (FDR).  
594 We used a subset ( $N = 1 \times 10^3$ ) of the empirical null observations and selected the highest  
595 p-value threshold that resulted in a  $V/R < Q$  where  $V$  is the mean number of expected false  
596 discoveries and  $R$  is the observed discoveries (McArthur et al., 2022). We calculated adjusted  
597 significance levels for each set for  $Q$  at both 0.05 and 0.1.

## 598 4.7 Gene Expression, Tissue Specificity, and sQTL

599 We used TPM counts for each gene from GTEx, version 8, to analyze expression. We quantified  
600 tissue specificity as the relative entropy of each gene's expression profile across 34 tissues  
601 compared to the median across genes overall. Thus, a gene with expression only in a small  
602 number of tissues would have high relative entropy and a gene with expression across many  
603 tissues would have low relatively entropy to this background distribution. The 34 tissues were  
604 selected based on groupings from the Human Protein Atlas to minimize the amount of sharing  
605 between distinct tissues, e.g., since brain tissues are over-represented in GTEx. We used  
606 median expression across tissues as the null and calculated relative entropy using the entropy  
607 function from the SciPy statistics package (Virtanen et al., 2020). Based on the observed  
608 distribution of relative entropy scores (Figure S28), we designated genes with scores  $\leq 0.1$  as  
609 "low tissue-specificity", genes with scores  $> 0.1$  and  $\leq 0.1$  as "medium tissue-specificity", and  
610 genes with scores  $> 0.5$  as "high tissue-specificity". We compared both the number of SAVs  
611 per gene and the maximum  $\Delta$  for SAVs among the three relative entropy categories.

612 We downloaded sQTL data from GTEx, version 8. We collated significant variant-gene

613 associations (N = 24,445,206) across all 49 tissues and intersected these with SAVs, using  
614 LiftOver (Hinrichs et al., 2006) to convert the SAVs to hg38 and then back to hg19 after inter-  
615 secting.

## 616 **4.8 Major Spliceosome Complex**

617 We characterized differences in the major spliceosome complex between archaics and modern  
618 humans by identifying variants in the 246 genes associated with the complex. We identified  
619 1,866 variants that did not occur in 1KG. We ran these variants through the Ensembl Variant  
620 Effect Predictor (McLaren et al., 2016) using the GRCh37.p13 assembly and all default options.

## 621 **4.9 Analysis**

622 All data analyses were performed using Bash and Python scripts, some of which were imple-  
623 mented in Jupyter notebooks. We employed non-parametric tests to analyze data including  
624 Fisher's exact test, Kruskal-Wallis tests, Mann-Whitney U tests, and Spearman correlation, im-  
625 plemented with SciPy (Virtanen et al., 2020). Partial correlations were run using the Pigouin  
626 package, version 0.5.2 (Vallat, 2018). Some additional metrics were calculated using custom  
627 functions. All reported p-values are two-tailed.

## 628 **4.10 Visualization**

629 Results were visualized using Inkscape, version 1.1 (Inkscape Project, 2020) and ggplot, ver-  
630 sion 3.3.5 (Wickham, 2016) implemented in R, version 4.1.2 (R Core Team, 2020). Additional  
631 packages used to generate figures include complexupset, version 1.3.3 (Krassowski, 2020);  
632 cowplot, version 1.1.1; eulerr, version 6.1.1 (Larsson, 2021); reshape, version 0.8.8 (Wickham,  
633 2007); and tidyverse, version 1.3.1 (Wickham et al., 2019).

## 634 **4.11 Data availability**

635 We used publicly available data for all analyses. Archaic InDel data are from the following  
636 repository: <http://ftp.eva.mpg.de/neandertal/Vindija/VCF/indels/>. Archaic SNV  
637 data are from the following repositories: Altai Neanderthal (<http://ftp.eva.mpg.de/neandertal/Vindija/VCF/Altai/>), Chagyrskaya (<http://ftp.eva.mpg.de/neandertal/Chagyrskaya/VCF/>), Denisova (<http://ftp.eva.mpg.de/neandertal/Vindija/VCF/Denisova/>),  
640 and Vindija (<http://ftp.eva.mpg.de/neandertal/Vindija/VCF/Vindija33.19/>). Modern  
641 human data are from the Thousand Genomes Project (<http://hgdownload.soe.ucsc.edu/gbdb/hg38/1000Genomes/>). Introgressed tag SNPs from Vernot et al., 2016 were retrieved  
642 from: [https://drive.google.com/drive/folders/0B9Pc7\\_zItMCVM05rUmhDc0hkWmc?resourcekey=0-zwKyJGRuooD9bWPRZ0vBzQ](https://drive.google.com/drive/folders/0B9Pc7_zItMCVM05rUmhDc0hkWmc?resourcekey=0-zwKyJGRuooD9bWPRZ0vBzQ). Introgressed variants from Browning et al., 2018  
643 were retrieved from: <https://data.mendeley.com/datasets/y7hyt83vxr/1>. gnomAD data  
644 were retrieved from: [https://storage.googleapis.com/gcp-public-data--gnomad/release/2.1.1/constraint/gnomad.v2.1.1.lof\\_metrics\\_by\\_gene.txt.bgz](https://storage.googleapis.com/gcp-public-data--gnomad/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics_by_gene.txt.bgz). phyloP data for  
645 the primate subset were retrieved from: <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates.phyloP46way.bw>. ASE variants were retrieved from: <https://drive.google.com/file/d/10ebWfA-sboAL1SDplmIrvH-xK4x9iohV/view>. TPM data were  
646

651 retrieved from the Human Protein Atlas ([https://www.proteinatlas.org/download/rna\\_tis  
652 sue\\_gtex.tsv.zip](https://www.proteinatlas.org/download/rna_tis_sue_gtex.tsv.zip)). sQTL data were retrieved from GTEx, version 8 ([https://storage.googleapis.com/gtex\\_analysis\\_v8/single\\_tissue\\_qtl\\_data/GTEx\\_Analysis\\_v8\\_sQTL.tar](https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_sQTL.tar)).  
653 Genes associated with the major spliceosome complex were retrieved from the HUGO Gene  
654 Nomenclature Committee (<https://www.genenames.org/data/genegroup/#!/group/1518>).  
655 The compiled dataset used in our analyses is available on Dryad (DOI: 10.7272/Q6H993F9).

## 657 4.12 Code availability

658 All code used to conduct analyses and generate figures is publicly available on GitHub ([https://github.com/brandcm/Archaic\\_Splicing](https://github.com/brandcm/Archaic_Splicing)).  
659

## 660 4.13 Acknowledgements

661 We thank Mary Lauren Benton for kindly sharing data on tissue specificity. Evonne McArthur  
662 and David Rinker provided comments that improved this manuscript. We also thank members  
663 of the Capra Lab for feedback on figures. This research greatly benefited from access to the  
664 Wynton high-performance compute cluster at UCSF. L.L.C. was funded by National Institutes of  
665 Health grant T32HG009495 to the University of Pennsylvania. J.A.C. and C.M.B. were funded  
666 by National Institutes of Health grant R35GM127087.

## 667 4.14 Author Contributions

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

## 670 4.15 Competing interests

671 The authors declare no competing interests.

## 672 References

- 673 Abadi, M. et al. 2015. *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems*.
- 674 Andrés, A. M. et al. 2010. Balancing Selection Maintains a Form of ERAP2 That Undergoes  
675 Nonsense-Mediated Decay and Affects Antigen Presentation. *PLOS Genetics* 6: e1001157.  
676 DOI: [10.1371/journal.pgen.1001157](https://doi.org/10.1371/journal.pgen.1001157).
- 677 Aqeilan, R. I. et al. 2008. The WWOX Tumor Suppressor Is Essential for Postnatal Survival  
678 and Normal Bone Metabolism \*. *Journal of Biological Chemistry* 283: 21629–21639. DOI:  
679 [10.1074/jbc.M800855200](https://doi.org/10.1074/jbc.M800855200).
- 680 Baralle, F. E. and Giudice, J. 2017. Alternative Splicing as a Regulator of Development and  
681 Tissue Identity. *Nature Reviews Molecular Cell Biology* 18: 437–451. DOI: [10.1038/nrm.2017.27](https://doi.org/10.1038/nrm.2017.27).

- 683 Barbosa-Morais, N. L. et al. 2012. The Evolutionary Landscape of Alternative Splicing in Ver-  
684 tebrate Species. *Science* 338: 1587–1593. DOI: [10.1126/science.1230612](https://doi.org/10.1126/science.1230612).
- 685 Black, D. L. 2003. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of  
686 Biochemistry* 72: 291–336. DOI: [10.1146/annurev.biochem.72.121801.161720](https://doi.org/10.1146/annurev.biochem.72.121801.161720).
- 687 Blekhman, R. et al. 2010. Sex-Specific and Lineage-Specific Alternative Splicing in Primates.  
688 *Genome Research* 20: 180–189. DOI: [10.1101/gr.099226.109](https://doi.org/10.1101/gr.099226.109).
- 689 Brand, C. M., Colbran, L. L., and Capra, J. A. 2022. Predicting Archaic Hominins Phenotypes  
690 from Genomic Data. *Annual Review of Genomics and Human Genetics* In press.
- 691 Browning, S. R. et al. 2018. Analysis of Human Sequence Data Reveals Two Pulses of Archaic  
692 Denisovan Admixture. *Cell* 173: 53–61.e9. DOI: [10.1016/j.cell.2018.02.031](https://doi.org/10.1016/j.cell.2018.02.031).
- 693 Buniello, A. et al. 2019. The NHGRI-EBI GWAS Catalog of Published Genome-Wide Asso-  
694 ciation Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Research*  
695 47: D1005–D1012. DOI: [10.1093/nar/gky1120](https://doi.org/10.1093/nar/gky1120).
- 696 Cáceres, J. F. and Kornblihtt, A. R. 2002. Alternative Splicing: Multiple Control Mechanisms and  
697 Involvement in Human Disease. *Trends in Genetics* 18: 186–193. DOI: [10.1016/S0168-9525\(01\)02626-9](https://doi.org/10.1016/S0168-9525(01)02626-9).
- 699 Castellano, S. et al. 2014. Patterns of Coding Variation in the Complete Exomes of Three Ne-  
700 andertals. *Proceedings of the National Academy of Sciences* 111: 6666. DOI: [10.1073/pnas.1405138111](https://doi.org/10.1073/pnas.1405138111).
- 702 Chen, E. Y. et al. 2013. Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment  
703 Analysis Tool. *BMC Bioinformatics* 14: 128. DOI: [10.1186/1471-2105-14-128](https://doi.org/10.1186/1471-2105-14-128).
- 704 Chen, L. et al. 2020. Identifying and Interpreting Apparent Neanderthal Ancestry in African  
705 Individuals. *Cell* 180: 677–687.e16. DOI: [10.1016/j.cell.2020.01.012](https://doi.org/10.1016/j.cell.2020.01.012).
- 706 Cheng, J. et al. 2019. MMSplice: Modular Modeling Improves the Predictions of Genetic Variant  
707 Effects on Splicing. *Genome Biology* 20: 48. DOI: [10.1186/s13059-019-1653-z](https://doi.org/10.1186/s13059-019-1653-z).
- 708 Chollet, F. et al. 2015. *Keras*. <https://github.com/fchollet/keras>.
- 709 Colbran, L. L. et al. 2019. Inferred Divergent Gene Regulation in Archaic Hominins Reveals  
710 Potential Phenotypic Differences. *Nature Ecology & Evolution* 3: 1598–1606. DOI: [10.1038/s41559-019-0996-x](https://doi.org/10.1038/s41559-019-0996-x).
- 712 Collins, L. and Penny, D. 2005. Complex Spliceosomal Organization Ancestral to Extant Eu-  
713 karyotes. *Molecular Biology and Evolution* 22: 1053–1066. DOI: [10.1093/molbev/msi091](https://doi.org/10.1093/molbev/msi091).
- 714 Danis, D. et al. 2021. Interpretable Prioritization of Splice Variants in Diagnostic Next-Generation  
715 Sequencing. *bioRxiv*: 2021.01.28.428499. DOI: [10.1101/2021.01.28.428499](https://doi.org/10.1101/2021.01.28.428499).
- 716 Dannemann, M., Andrés, A. M., and Kelso, J. 2016. Introgression of Neandertal- and Denisovan-  
717 like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *The Amer-  
718 ikan Journal of Human Genetics* 98: 22–33. DOI: [10.1016/j.ajhg.2015.11.015](https://doi.org/10.1016/j.ajhg.2015.11.015).

- 719 Faustino, N. A. and Cooper, T. A. 2003. Pre-mRNA Splicing and Human Disease. *Genes &*  
720 *Development* 17: 419–437. DOI: [10.1101/gad.1048803](https://doi.org/10.1101/gad.1048803).
- 721 Ge, Y. and Porse, B. T. 2014. The Functional Consequences of Intron Retention: Alternative  
722 Splicing Coupled to NMD as a Regulator of Gene Expression. *BioEssays* 36: 236–243. DOI:  
723 [10.1002/bies.201300156](https://doi.org/10.1002/bies.201300156).
- 724 Gokhman, D. et al. 2019. Reconstructing Denisovan Anatomy Using DNA Methylation Maps.  
725 *Cell* 179: 180–192.e10. DOI: [10.1016/j.cell.2019.08.035](https://doi.org/10.1016/j.cell.2019.08.035).
- 726 Graveley, B. R. 2001. Alternative Splicing: Increasing Diversity in the Proteomic World. *Trends*  
727 *in Genetics* 17: 100–107. DOI: [10.1016/S0168-9525\(00\)02176-4](https://doi.org/10.1016/S0168-9525(00)02176-4).
- 728 Harrow, J. et al. 2012. GENCODE: The Reference Human Genome Annotation for The EN-  
729 CODE Project. *Genome Research* 22: 1760–1774. DOI: [10.1101/gr.135350.111](https://doi.org/10.1101/gr.135350.111).
- 730 Hinrichs, A. S. et al. 2006. The UCSC Genome Browser Database: Update 2006. *Nucleic Acids*  
731 *Research* 34: D590–D598. DOI: [10.1093/nar/gkj144](https://doi.org/10.1093/nar/gkj144).
- 732 Huerta-Sánchez, E. et al. 2014. Altitude Adaptation in Tibetans Caused by Introgression of  
733 Denisovan-like DNA. *Nature* 512: 194–197. DOI: [10.1038/nature13408](https://doi.org/10.1038/nature13408).
- 734 Inkscape Project. 2020. *Inkscape*.
- 735 Jagadeesh, K. A. et al. 2019. S-CAP Extends Pathogenicity Prediction to Genetic Variants That  
736 Affect RNA Splicing. *Nature Genetics* 51: 755–763. DOI: [10.1038/s41588-019-0348-4](https://doi.org/10.1038/s41588-019-0348-4).
- 737 Jaganathan, K. et al. 2019. Predicting Splicing from Primary Sequence with Deep Learning.  
738 *Cell* 176: 535–548.e24. DOI: [10.1016/j.cell.2018.12.015](https://doi.org/10.1016/j.cell.2018.12.015).
- 739 Jenkinson, G. et al. 2020. LeafCutterMD: An Algorithm for Outlier Splicing Detection in Rare  
740 Diseases. *Bioinformatics* 36: 4609–4615. DOI: [10.1093/bioinformatics/btaa259](https://doi.org/10.1093/bioinformatics/btaa259).
- 741 Jeong, C. et al. 2018. Detecting Past and Ongoing Natural Selection among Ethnically Tibetan  
742 Women at High Altitude in Nepal. *PLOS Genetics* 14: e1007650. DOI: [10.1371/journal.pgen.1007650](https://doi.org/10.1371/journal.pgen.1007650).
- 744 Karczewski, K. J. et al. 2020. The Mutational Constraint Spectrum Quantified from Variation in  
745 141,456 Humans. *Nature* 581: 434–443. DOI: [10.1038/s41586-020-2308-7](https://doi.org/10.1038/s41586-020-2308-7).
- 746 Karlebach, G. et al. 2020. The Impact of Biological Sex on Alternative Splicing. *bioRxiv*: 490904.  
747 DOI: [10.1101/490904](https://doi.org/10.1101/490904).
- 748 Köhler, S. et al. 2021. The Human Phenotype Ontology in 2021. *Nucleic Acids Research* 49: D1207–  
749 D1217. DOI: [10.1093/nar/gkaa1043](https://doi.org/10.1093/nar/gkaa1043).
- 750 Krassowski, M. 2020. *ComplexUpset*. DOI: [10.5281/zenodo.3700590](https://doi.org/10.5281/zenodo.3700590).
- 751 Krawczak, M., Reiss, J., and Cooper, D. N. 1992. The Mutational Spectrum of Single Base-  
752 Pair Substitutions in mRNA Splice Junctions of Human Genes: Causes and Consequences.  
753 *Human Genetics* 90: 41–54. DOI: [10.1007/BF00210743](https://doi.org/10.1007/BF00210743).

- 754 Kriventseva, E. V. et al. 2003. Increase of Functional Diversity by Alternative Splicing. *Trends  
755 in Genetics* 19: 124–128. DOI: [10.1016/S0168-9525\(03\)00023-4](https://doi.org/10.1016/S0168-9525(03)00023-4).
- 756 Kuleshov, M. V. et al. 2016. Enrichr: A Comprehensive Gene Set Enrichment Analysis Web  
757 Server 2016 Update. *Nucleic Acids Research* 44: W90–W97. DOI: [10.1093/nar/gkw377](https://doi.org/10.1093/nar/gkw377).
- 758 Larsson, J. 2021. *eulerr: Area-proportional Euler and Venn Diagrams with Ellipses*. Manual.
- 759 Li, H. 2011. A Statistical Framework for SNP Calling, Mutation Discovery, Association Map-  
760 ping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics*  
761 27: 2987–2993. DOI: [10.1093/bioinformatics/btr509](https://doi.org/10.1093/bioinformatics/btr509).
- 762 Li, X. et al. 2017. The Impact of Rare Variation on Gene Expression across Tissues. *Nature*  
763 550: 239–243. DOI: [10.1038/nature24267](https://doi.org/10.1038/nature24267).
- 764 Li, Y. I. et al. 2016. RNA Splicing Is a Primary Link between Genetic Variation and Disease.  
765 *Science* 352: 600–604. DOI: [10.1126/science.aad9417](https://doi.org/10.1126/science.aad9417).
- 766 Lopez, A. J. 1998. Alternative Slicing of Pre-mRNA: Developmental Consequences and Mech-  
767 anisms of Regulation. *Annual Review of Genetics* 32: 279–305. DOI: [10.1146/annurev.genet.32.1.279](https://doi.org/10.1146/annurev.genet.32.1.279).
- 769 Lowy-Gallego, E. et al. 2019. Variant Calling on the GRCh38 Assembly with the Data from  
770 Phase Three of the 1000 Genomes Project [Version 2; Peer Review: 2 Approved]. *Wellcome  
771 Open Research* 4. DOI: [10.12688/wellcomeopenres.15126.2](https://doi.org/10.12688/wellcomeopenres.15126.2).
- 772 Mafessoni, F. et al. 2020. A High-Coverage Neandertal Genome from Chagyrskaya Cave.  
773 *Proceedings of the National Academy of Sciences* 117: 15132. DOI: [10.1073/pnas.2004944117](https://doi.org/10.1073/pnas.2004944117).
- 775 Martin, A. R. et al. 2019. Clinical Use of Current Polygenic Risk Scores May Exacerbate Health  
776 Disparities. *Nature Genetics* 51: 584–591. DOI: [10.1038/s41588-019-0379-x](https://doi.org/10.1038/s41588-019-0379-x).
- 777 McArthur, E., Rinker, D. C., and Capra, J. A. 2021. Quantifying the Contribution of Neanderthal  
778 Introgression to the Heritability of Complex Traits. *Nature Communications* 12: 4481. DOI:  
779 [10.1038/s41467-021-24582-y](https://doi.org/10.1038/s41467-021-24582-y).
- 780 McArthur, E. et al. 2022. Reconstructing the 3D Genome Organization of Neanderthals Reveals  
781 That Chromatin Folding Shaped Phenotypic and Sequence Divergence. *bioRxiv*: 2022.02.07.479462.  
782 DOI: [10.1101/2022.02.07.479462](https://doi.org/10.1101/2022.02.07.479462).
- 783 McCoy, R. C., Wakefield, J., and Akey, J. M. 2017. Impacts of Neanderthal-Introgressed Se-  
784 quences on the Landscape of Human Gene Expression. *Cell* 168: 916–927.e12. DOI: [10.1016/j.cell.2017.01.038](https://doi.org/10.1016/j.cell.2017.01.038).
- 786 McLaren, W. et al. 2016. The Ensembl Variant Effect Predictor. *Genome Biology* 17: 122. DOI:  
787 [10.1186/s13059-016-0974-4](https://doi.org/10.1186/s13059-016-0974-4).
- 788 Mendez, F. L., Watkins, J. C., and Hammer, M. F. 2012. Global Genetic Variation at OAS1  
789 Provides Evidence of Archaic Admixture in Melanesian Populations. *Molecular Biology and  
790 Evolution* 29: 1513–1520. DOI: [10.1093/molbev/msr301](https://doi.org/10.1093/molbev/msr301).

- 791 Merkin, J. et al. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian  
792 Tissues. *Science* 338: 1593–1599. DOI: [10.1126/science.1228186](https://doi.org/10.1126/science.1228186).
- 793 Mertes, C. et al. 2021. Detection of Aberrant Splicing Events in RNA-seq Data Using FRASER.  
794 *Nature Communications* 12: 529. DOI: [10.1038/s41467-020-20573-7](https://doi.org/10.1038/s41467-020-20573-7).
- 795 Meyer, M. et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Indi-  
796 vidual. *Science* 338: 222. DOI: [10.1126/science.1224344](https://doi.org/10.1126/science.1224344).
- 797 Nissim-Rafinia, M. and Kerem, B. 2002. Splicing Regulation as a Potential Genetic Modifier.  
798 *Trends in Genetics* 18: 123–127. DOI: [10.1016/S0168-9525\(01\)02619-1](https://doi.org/10.1016/S0168-9525(01)02619-1).
- 799 Peng, Y. et al. 2017. Down-Regulation of EPAS1 Transcription and Genetic Adaptation of Ti-  
800 betans to High-Altitude Hypoxia. *Molecular Biology and Evolution* 34: 818–830. DOI: [10.1093/molbev/msw280](https://doi.org/10.1093/molbev/msw280).
- 802 Petr, M. et al. 2019. Limits of Long-Term Selection against Neandertal Introgression. *Proceed-  
803 ings of the National Academy of Sciences* 116: 1639. DOI: [10.1073/pnas.1814338116](https://doi.org/10.1073/pnas.1814338116).
- 804 Plagnol, V. and Wall, J. D. 2006. Possible Ancestral Structure in Human Populations. *PLOS  
805 Genetics* 2: e105. DOI: [10.1371/journal.pgen.0020105](https://doi.org/10.1371/journal.pgen.0020105).
- 806 Pollard, K. S. et al. 2010. Detection of Nonneutral Substitution Rates on Mammalian Phyloge-  
807 nies. *Genome Research* 20: 110–121. DOI: [10.1101/gr.097857.109](https://doi.org/10.1101/gr.097857.109).
- 808 Prüfer, K. et al. 2014. The Complete Genome Sequence of a Neanderthal from the Altai Moun-  
809 tains. *Nature* 505: 43–49. DOI: [10.1038/nature12886](https://doi.org/10.1038/nature12886).
- 810 Prüfer, K. et al. 2017. A High-Coverage Neandertal Genome from Vindija Cave in Croatia.  
811 *Science* 358: 655–658. DOI: [10.1126/science.aao1887](https://doi.org/10.1126/science.aao1887).
- 812 R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation  
813 for Statistical Computing. Vienna, Austria.
- 814 Rinker, D. C. et al. 2020. Neanderthal Introgression Reintroduced Functional Ancestral Alleles  
815 Lost in Eurasian Populations. *Nature Ecology & Evolution* 4: 1332–1341. DOI: [10.1038/s41559-020-1261-z](https://doi.org/10.1038/s41559-020-1261-z).
- 817 Rodenas-Cuadrado, P., Ho, J., and Vernes, S. C. 2014. Shining a Light on CNTNAP2: Complex  
818 Functions to Complex Disorders. *European Journal of Human Genetics* 22: 171–178. DOI:  
819 [10.1038/ejhg.2013.100](https://doi.org/10.1038/ejhg.2013.100).
- 820 Rogers, A. R., Harris, N. S., and Achenbach, A. A. 2020. Neanderthal-Denisovan Ancestors  
821 Interbred with a Distantly Related Hominin. *Science Advances* 6: eaay5483. DOI: [10.1126/sciadv.aay5483](https://doi.org/10.1126/sciadv.aay5483).
- 823 Rogers, T. F., Palmer, D. H., and Wright, A. E. 2021. Sex-Specific Selection Drives the Evolution  
824 of Alternative Splicing in Birds. *Molecular Biology and Evolution* 38: 519–530. DOI: [10.1093/molbev/msaa242](https://doi.org/10.1093/molbev/msaa242).

- 826 Sams, A. J. et al. 2016. Adaptively Introgressed Neandertal Haplotype at the OAS Locus Func-  
827 tionally Impacts Innate Immune Responses in Humans. *Genome Biology* 17: 246. DOI:  
828 [10.1186/s13059-016-1098-6](https://doi.org/10.1186/s13059-016-1098-6).
- 829 Saudemont, B. et al. 2017. The Fitness Cost of Mis-Splicing Is the Main Determinant of Alter-  
830 native Splicing Patterns. *Genome Biology* 18: 208. DOI: [10.1186/s13059-017-1344-6](https://doi.org/10.1186/s13059-017-1344-6).
- 831 Scotti, M. M. and Swanson, M. S. 2016. RNA Mis-Splicing in Disease. *Nature Reviews Genetics*  
832 17: 19–32. DOI: [10.1038/nrg.2015.3](https://doi.org/10.1038/nrg.2015.3).
- 833 Shiina, T. et al. 2009. The HLA Genomic Loci Map: Expression, Interaction, Diversity and Dis-  
834 ease. *Journal of Human Genetics* 54: 15–39. DOI: [10.1038/jhg.2008.5](https://doi.org/10.1038/jhg.2008.5).
- 835 Sibley, C. R., Blazquez, L., and Ule, J. 2016. Lessons from Non-Canonical Splicing. *Nature*  
836 *Reviews Genetics* 17: 407–421. DOI: [10.1038/nrg.2016.46](https://doi.org/10.1038/nrg.2016.46).
- 837 Singh, P. and Ahi, E. P. 2022. The Importance of Alternative Splicing in Adaptive Evolution.  
838 *Molecular Ecology* n/a. DOI: [10.1111/mec.16377](https://doi.org/10.1111/mec.16377).
- 839 Smith, J. E. and Baker, K. E. 2015. Nonsense-Mediated RNA Decay – a Switch and Dial for  
840 Regulating Gene Expression. *BioEssays* 37: 612–623. DOI: [10.1002/bies.201500007](https://doi.org/10.1002/bies.201500007).
- 841 Telis, N., Aguilar, R., and Harris, K. 2020. Selection against Archaic Hominin Genetic Variation  
842 in Regulatory Regions. *Nature Ecology & Evolution* 4: 1558–1566. DOI: [10.1038/s41559-020-01284-0](https://doi.org/10.1038/s41559-020-01284-0).
- 844 Tweedie, S. et al. 2021. Genenames.Org: The HGNC and VGNC Resources in 2021. *Nucleic*  
845 *Acids Research* 49: D939–D946. DOI: [10.1093/nar/gkaa980](https://doi.org/10.1093/nar/gkaa980).
- 846 Vallat, R. 2018. Pingouin: Statistics in Python. *The Journal of Open Source Software* 3: 1026.
- 847 Vernot, B. and Akey, J. M. 2014. Resurrecting Surviving Neandertal Lineages from Modern  
848 Human Genomes. *Science* 343: 1017–1021. DOI: [10.1126/science.1245938](https://doi.org/10.1126/science.1245938).
- 849 Vernot, B. et al. 2016. Excavating Neandertal and Denisovan DNA from the Genomes of Melane-  
850 sian Individuals. *Science* 352: 235–239. DOI: [10.1126/science.aad9416](https://doi.org/10.1126/science.aad9416).
- 851 Verta, J.-P. and Jacobs, A. 2022. The Role of Alternative Splicing in Adaptation and Evolution.  
852 *Trends in Ecology & Evolution* 37: 299–308. DOI: [10.1016/j.tree.2021.11.010](https://doi.org/10.1016/j.tree.2021.11.010).
- 853 Virtanen, P. et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.  
854 *Nature Methods* 17: 261–272. DOI: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
- 855 Wang, G.-S. and Cooper, T. A. 2007. Splicing in Disease: Disruption of the Splicing Code and  
856 the Decoding Machinery. *Nature Reviews Genetics* 8: 749–761. DOI: [10.1038/nrg2164](https://doi.org/10.1038/nrg2164).
- 857 Wickham, H. 2007. Reshaping Data with the Reshape Package. *Journal of Statistical Software*  
858 21.
- 859 —— 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. ISBN: 978-  
860 3-319-24277-4.

- 861 Wickham, H. et al. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4: 1686.  
862 DOI: [10.21105/joss.01686](https://doi.org/10.21105/joss.01686).
- 863 Wright, C. J., Smith, C. W. J., and Jiggins, C. D. 2022. Alternative Splicing as a Source of  
864 Phenotypic Diversity. *Nature Reviews Genetics*. DOI: [10.1038/s41576-022-00514-4](https://doi.org/10.1038/s41576-022-00514-4).
- 865 Xie, Z. et al. 2021. Gene Set Knowledge Discovery with Enrichr. *Current Protocols* 1: e90. DOI:  
866 [10.1002/cpz1.90](https://doi.org/10.1002/cpz1.90).
- 867 Zeng, T. and Li, Y. I. 2022. Predicting RNA Splicing from DNA Sequence Using Pangolin.  
868 *Genome Biology* 23: 103. DOI: [10.1186/s13059-022-02664-4](https://doi.org/10.1186/s13059-022-02664-4).
- 869 Zhang, Y. et al. 2018. Discerning Novel Splice Junctions Derived from RNA-seq Alignment: A  
870 Deep Learning Approach. *BMC Genomics* 19: 971. DOI: [10.1186/s12864-018-5350-1](https://doi.org/10.1186/s12864-018-5350-1).

## 871 5 Supplementary Information

### 872 5.1 Supplementary Text

#### 873 5.1.1 The major spliceosome complex is strongly conserved between archaic hominins 874 and modern humans.

875 Alternative splicing occurs across nearly all eukaryotes and its molecular mechanisms are  
876 deeply conserved (Collins and Penny, 2005). Thus, we anticipated that the sequence patterns  
877 learned by SpliceAI in humans would generalize to archaic hominins. To test for any signif-  
878 icant differences, we first compared the sequences of 246 genes associated with the major  
879 spliceosome complex (Tweedie et al., 2021) between archaic hominins and modern humans.  
880 Analyzing all archaic variants absent in the 1000 Genomes Project (1KG) (Lowy-Gallego et al.,  
881 2019), we identified 19 non-synonymous archaic-specific variants, only 8 of which were pre-  
882 dicted to be deleterious by SIFT and 5 possibly or probably damaging by PolyPhen 2 (McLaren  
883 et al., 2016; **File S1**). Additionally, only two variants were predicted to disrupt protein function  
884 by both PolyPhen and SIFT (**File S1**). Furthermore, among these archaic-specific variants,  
885 none were shared among all four individuals and many were Denisovan-specific. Therefore,  
886 the major spliceosome complex is nearly identical between archaic hominins and modern hu-  
887 mans, and there is no evidence that the sequence determinants of splicing have diverged.

#### 888 5.1.2 Physical characteristics of a gene are associated with the number of SAVs.

889 Alternative splicing requires genes have at least one intron and the extent of splice altering  
890 variants may further be related to gene length. We considered the relationship between the  
891 the number of splicing variants ( $N = 0 - 11$ ) in each gene and three gene traits: 1) number of  
892 exons, 2) length of gene body, and 3) length of coding sequence. These characteristics were  
893 positively associated at both  $\Delta$  thresholds (**Table S4**) (**Figure S6**). We also considered the re-  
894 lationship between the number of known isoforms per gene and the number of SAVs. We found  
895 a significant, positive association at both thresholds (**Table S4**). However, this relationship is  
896 likely driven by the number of exons. When we conducted a partial correlation, controlling for  
897 the number of exons, both associations were non-significant with minimal effect size ( $\Delta \geq 0.2$ :  
898 partial Spearman,  $\rho = -0.003$ ,  $P = 0.702$ ,  $\Delta \geq 0.5$ : partial Spearman,  $\rho = 0.0008$ ,  $P = 0.9174$ ).

#### 899 5.1.3 Sprime identifies more introgressed SAVs present in all five 1KG superpopula- 900 tions.

901 Similarly to Vernot et al., 2016, most ancient SAVs occurred at  $\geq 0.05$  frequency in all five 1KG  
902 superpopulations (**Figure S18**). The next most frequent were SAVs that were present in all but  
903 east Asians. SAVs shared among non-Africans and those present in all but Europeans were  
904 also common.

905 There were some differences between the datasets for introgressed SAVs. We found that  
906 among those introgressed SAVs classified per Vernot et al., 2016, many were shared among  
907 non-Africans (**Figure 5D**). However, introgressed SAVs classified per Browning et al., 2018  
908 were most commonly present in both Africans and non-Africans, followed by a smaller set  
909 of non-African SAVs (**Figure S20**). This difference likely reflects low frequency introgressed

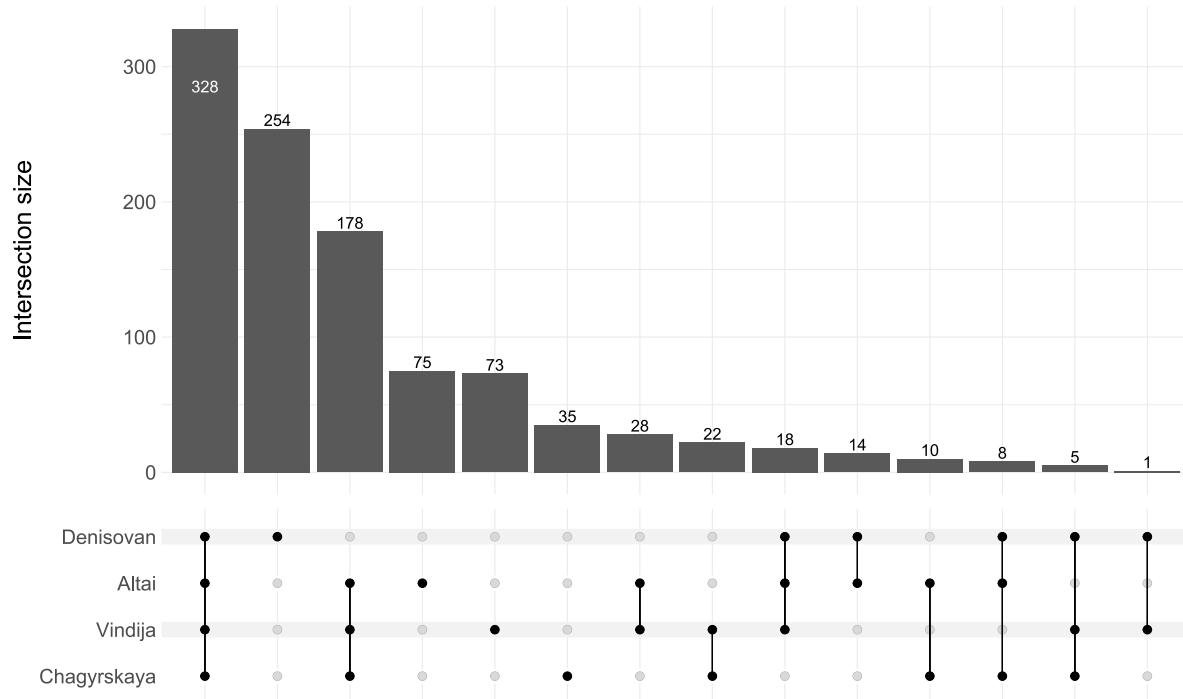
910 variants that are allowed to occur in the reference population in Sprime. While introgression  
911 between modern humans and archaics occurred outside of Africa, many Africans have Ne-  
912 anderthal ancestry due to backflow from Eurasians into Africans (Chen et al., 2020). Among  
913 non-African subsets, both datasets were largely similar in their distribution (**Figure S20, Fig-**  
914 **ure S21**).

915 **5.2 Supplementary Files**

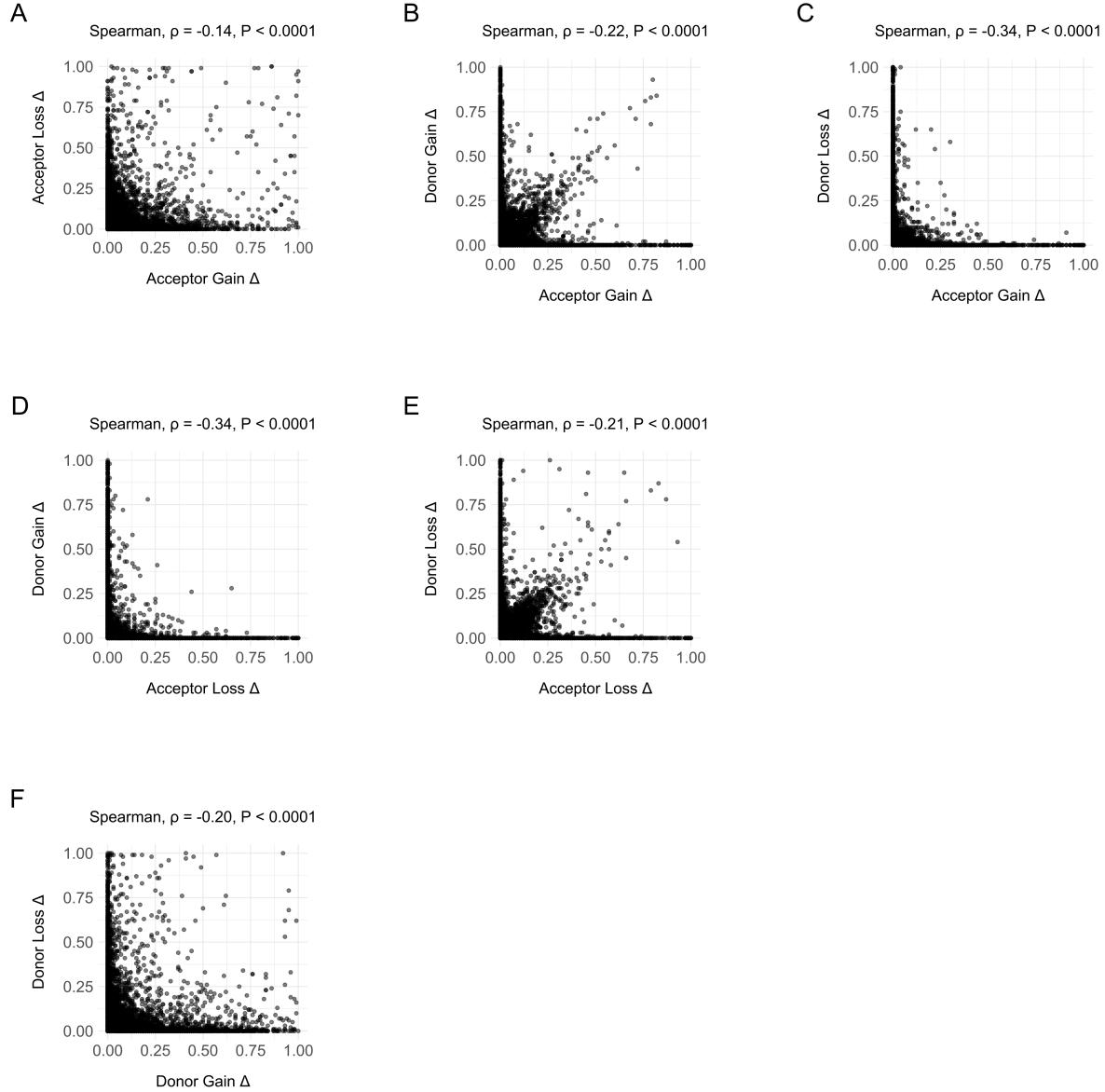
916 **File S1.** This file is the variant effect predictor (VEP) output from Ensembl for 1,866 archaic-  
917 specific variants in genes associated with the major spliceosome complex.

918 **File S2.** This file contains the outputs from the gene ontology analysis for GWAS Catalog 2019  
919 and Human Phenotype Ontology terms.

920 **5.3 Supplemental Figures**

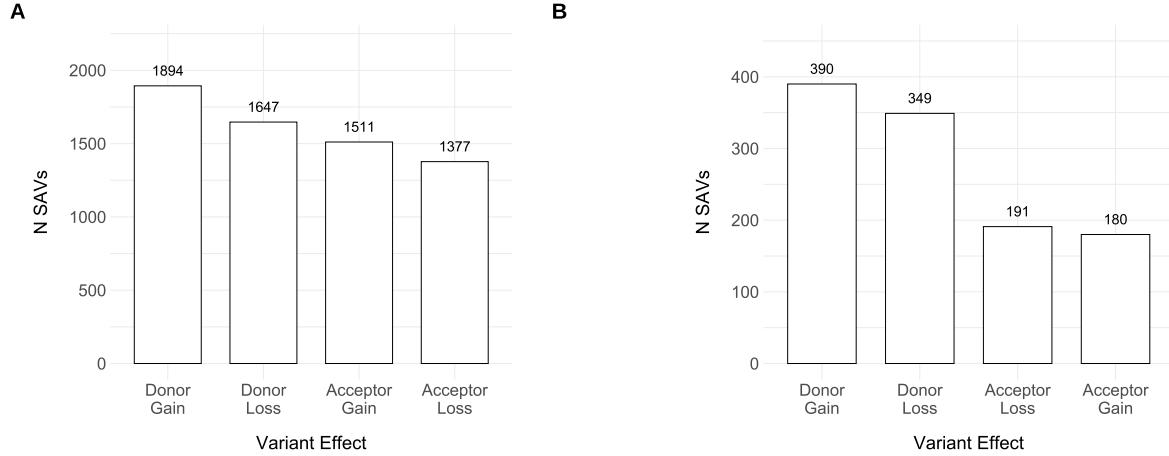


**Figure S1: The distribution of SAVs is consistent with the archaic hominin phylogeny.**  
Unique and shared SAVs at  $\Delta \geq 0.5$ .



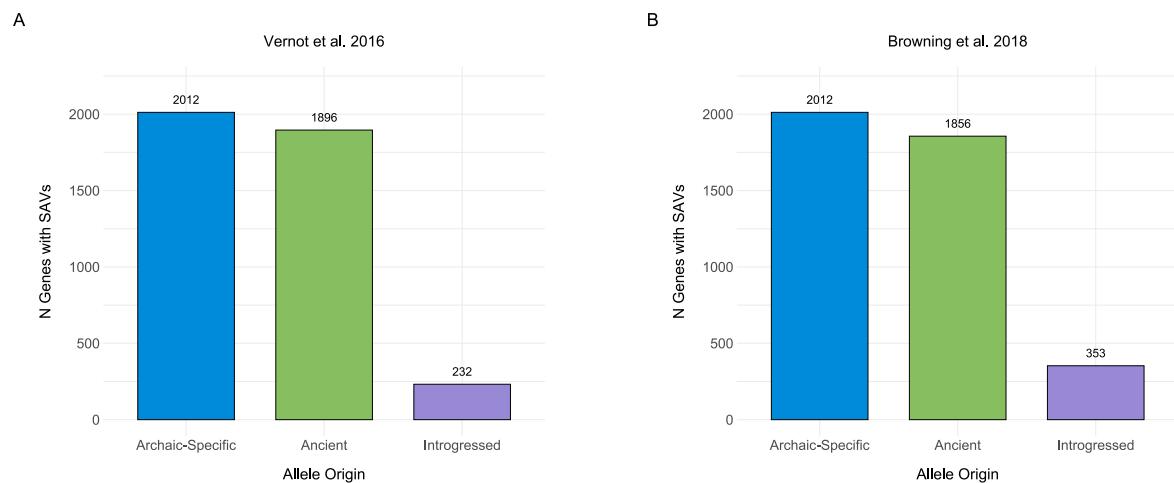
**Figure S2: Delta scores between classes are negatively associated.**

(**A**) Association between acceptor gain  $\Delta$ s and acceptor loss  $\Delta$ s for all variants with at least one  $\Delta > 0$ . (**B**) Association between acceptor gain  $\Delta$ s and donor gain  $\Delta$ s for all variants with at least one  $\Delta > 0$ . (**C**) Association between acceptor gain  $\Delta$ s and donor loss  $\Delta$ s for all variants with at least one  $\Delta > 0$ . (**D**) Association between acceptor loss  $\Delta$ s and donor gain  $\Delta$ s for all variants with at least one  $\Delta > 0$ . (**E**) Association between acceptor loss  $\Delta$ s and donor loss  $\Delta$ s for all variants with at least one  $\Delta > 0$ . (**F**) Association between donor gain  $\Delta$ s and donor loss  $\Delta$ s for all variants with at least one  $\Delta > 0$ .



**Figure S3: SAVs most commonly result in a donor gain.**

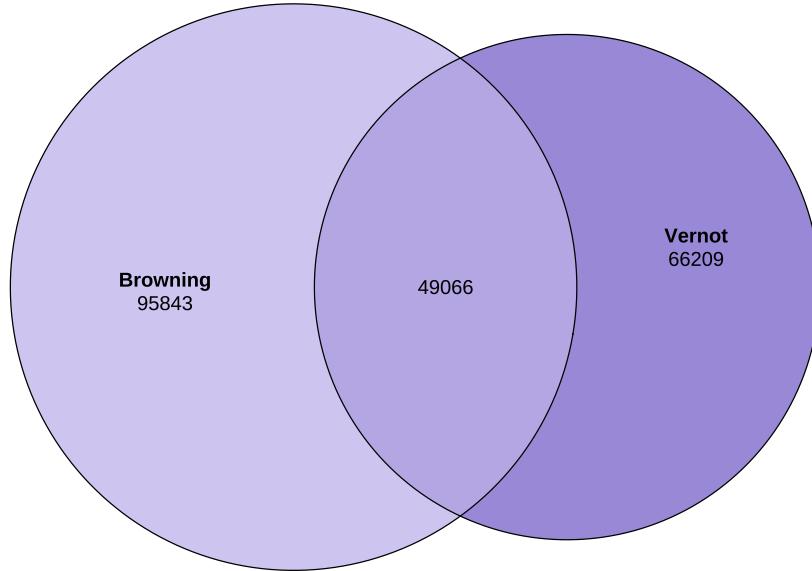
(A) The distribution of variant effects for all variants at the  $\Delta \geq 0.2$  threshold. Some variants result in multiple effects; therefore, the sum of these classes  $\neq 5,950$ . (B) The distribution of variant effects for all variants at the  $\Delta \geq 0.5$  threshold. Some variants result in multiple effects; therefore, the sum of these classes  $\neq 1,049$ .



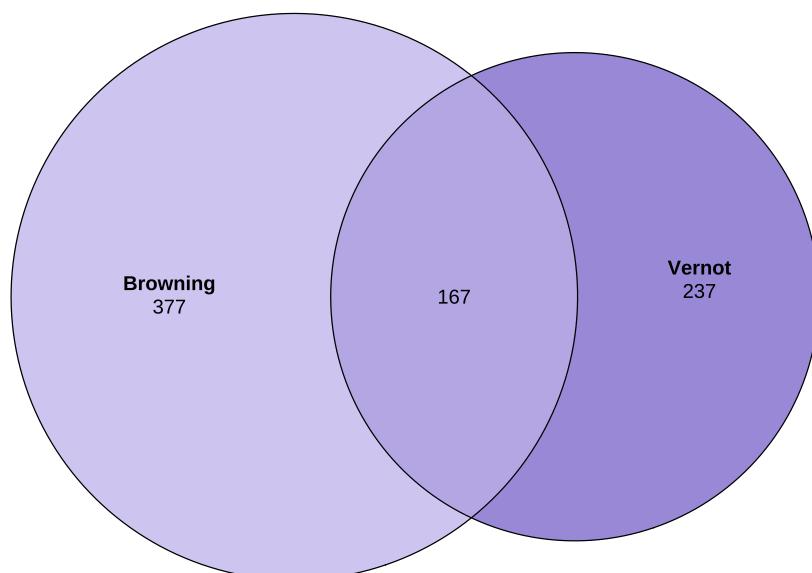
**Figure S4: N genes with SAVs by allele origin.**

(A) The number of genes encompassed by each SAV allele origin per Browning et al., 2018 at  $\Delta \geq 0.2$ . Some genes may occur in multiple categories (B) The number of genes encompassed by each SAV allele origin per Vernot et al., 2016 at  $\Delta \geq 0.2$ . Some genes may occur in multiple categories.

A

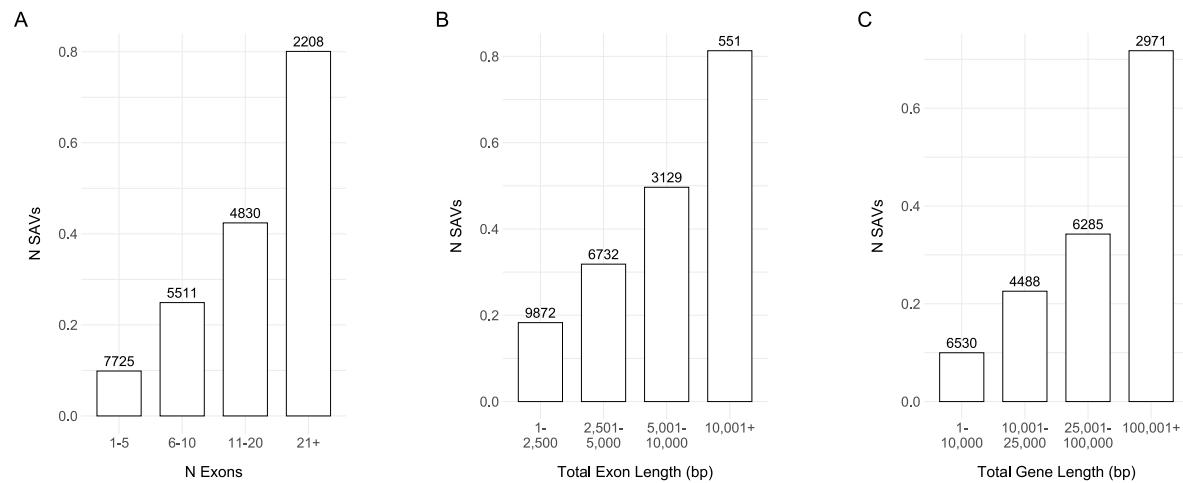


B



**Figure S5: Two different sets of introgressed variants modestly overlap.**

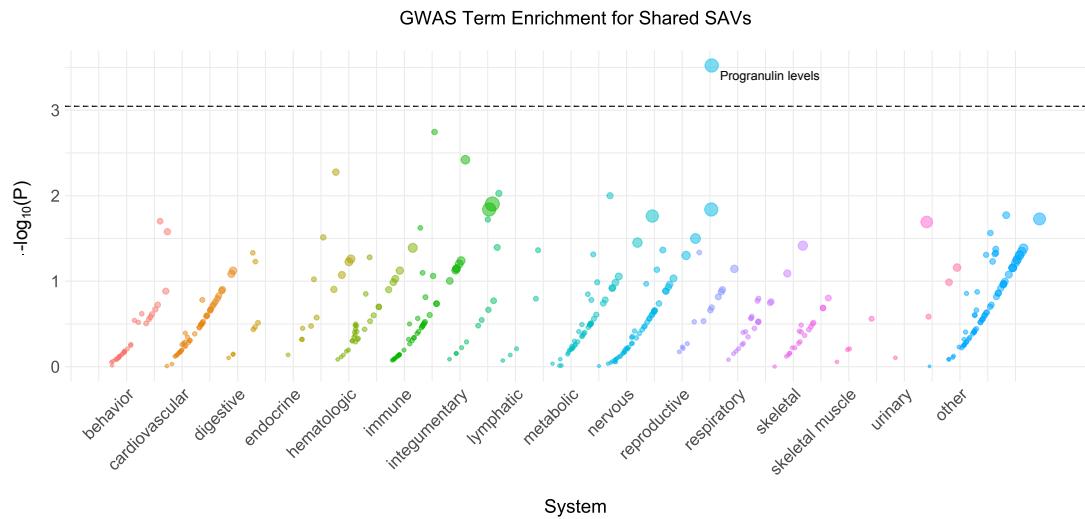
(A) The overlap between introgressed variants from Browning et al., 2018 and Vernet et al., 2016 that match a quality filtered locus in this study and are present in Thousand Genomes. (B) A subset of the overlap for variants with  $\Delta \geq 0.2$ .



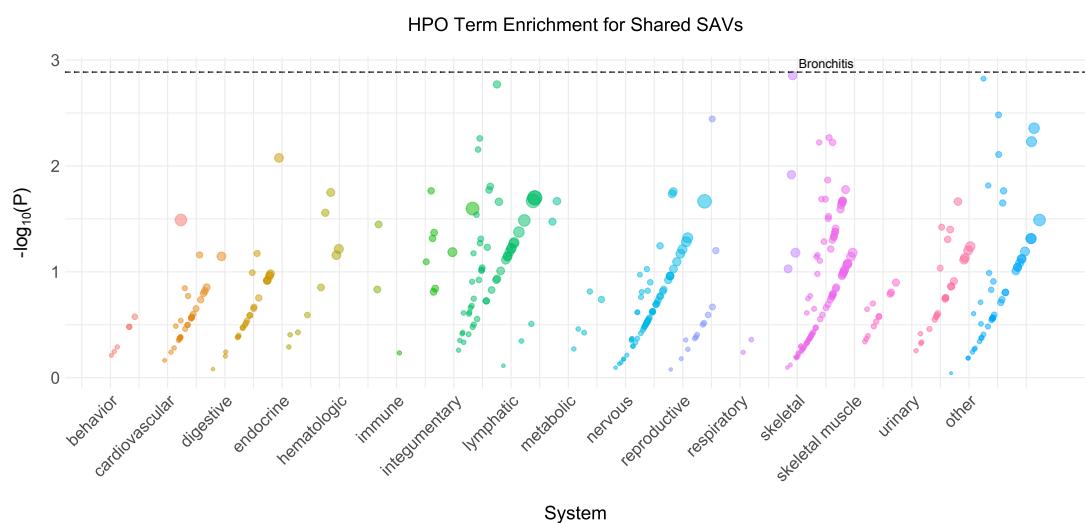
**Figure S6: The number of exons and gene length are associated with more SAVs.**

**(A)** Binned number of exons and the mean number of SAVs at  $\Delta \geq 0.2$  per bin. N reflects the number of genes per bin. **(B)** Binned exon length in bp and the mean number of SAVs at  $\Delta \geq 0.2$  per bin. N reflects the number of genes per bin. **(C)** Binned gene length in bp and the mean number of SAVs at  $\Delta \geq 0.2$  per bin. N reflects the number of genes per bin.

A



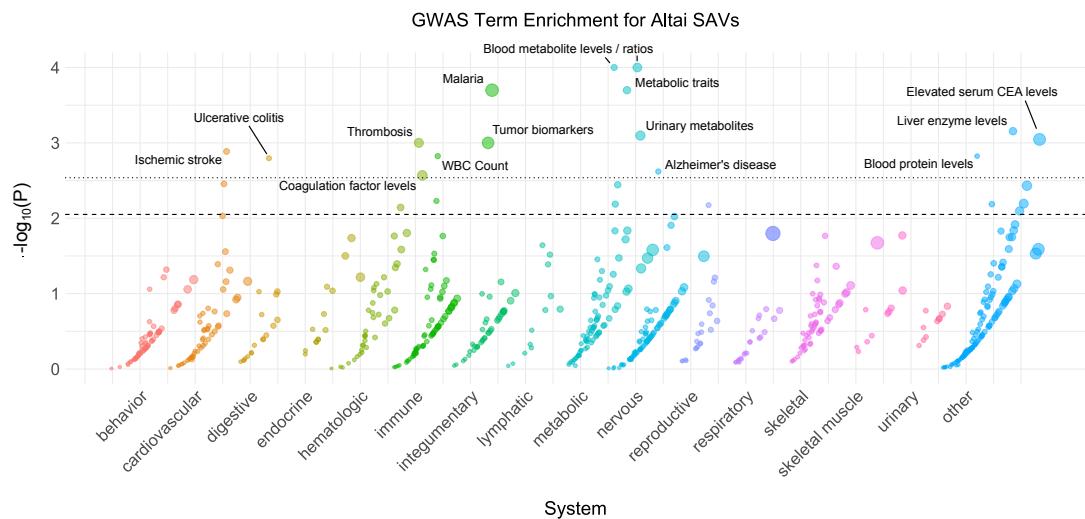
B



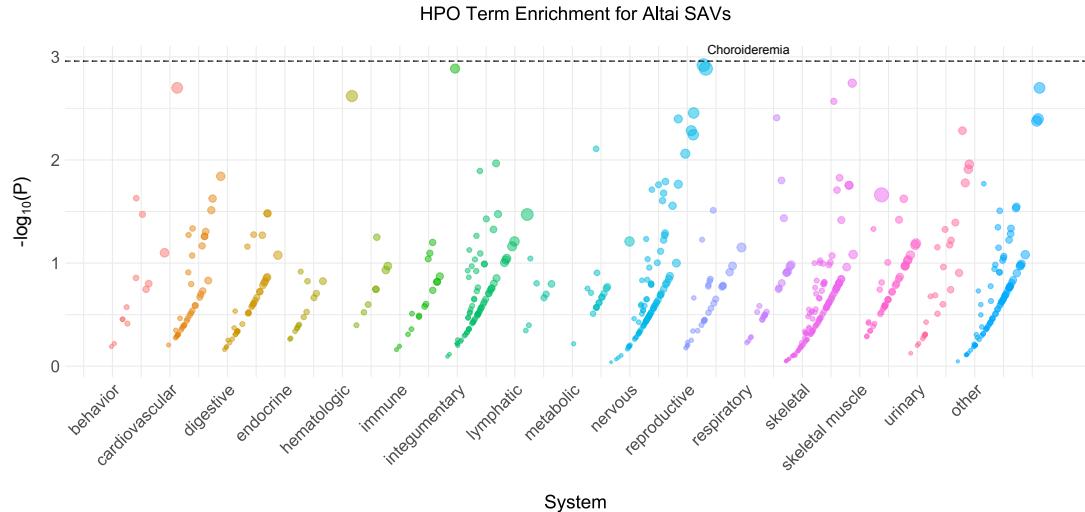
### Figure S7: Shared phenotype enrichment.

**(A)** Phenotype associations enriched among genes with archaic-specific shared SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  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 **File S2** for all phenotype enrichment results.

A



B

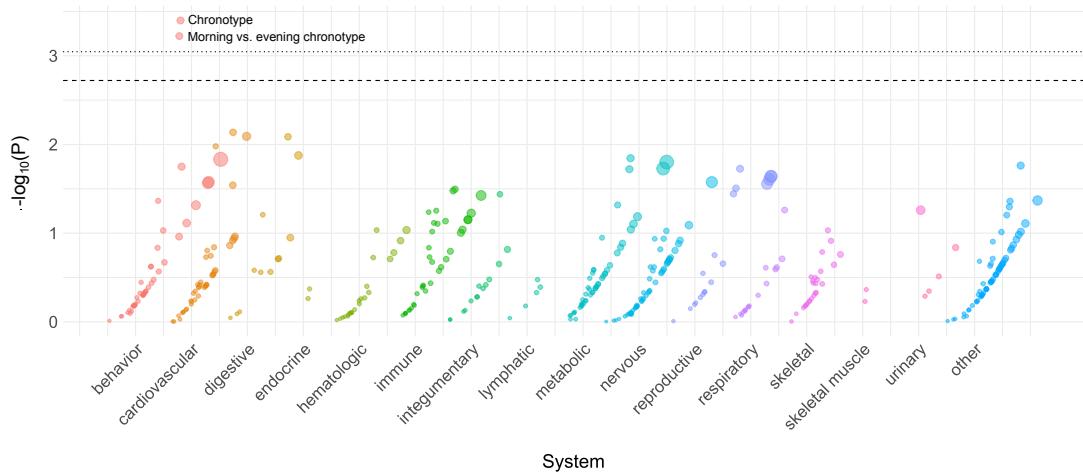


### Figure S8: Altai phenotype enrichment.

**(A)** Phenotype associations enriched among genes with archaic-specific Altai SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  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 **File S2** for all phenotype enrichment results.

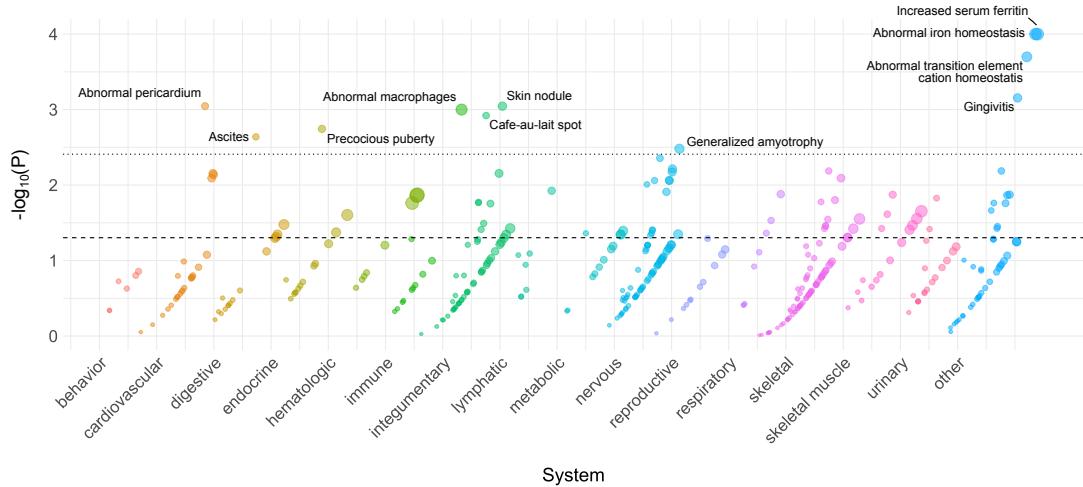
A

GWAS Term Enrichment for Chagyrskaya SAVs



B

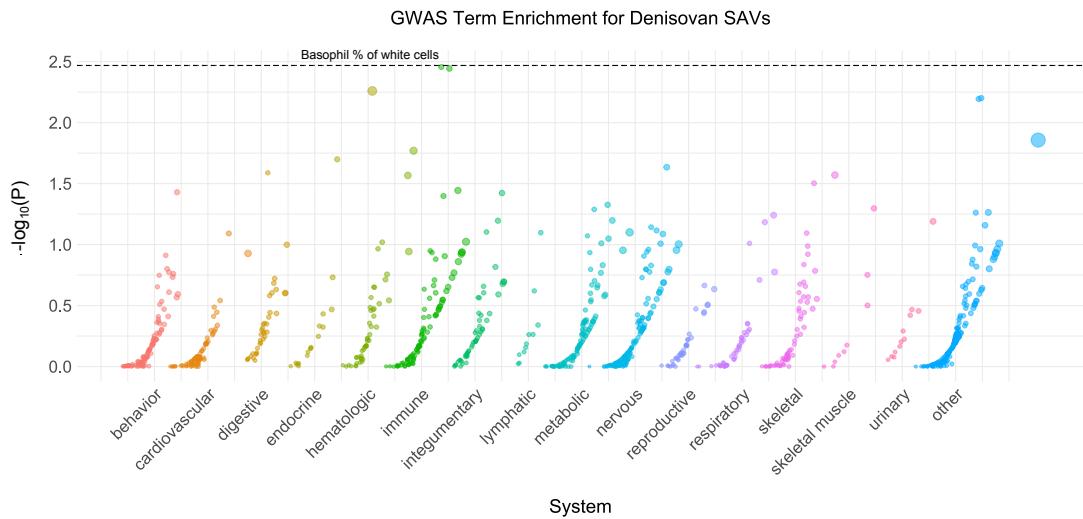
HPO Term Enrichment for Chagyrskaya SAVs



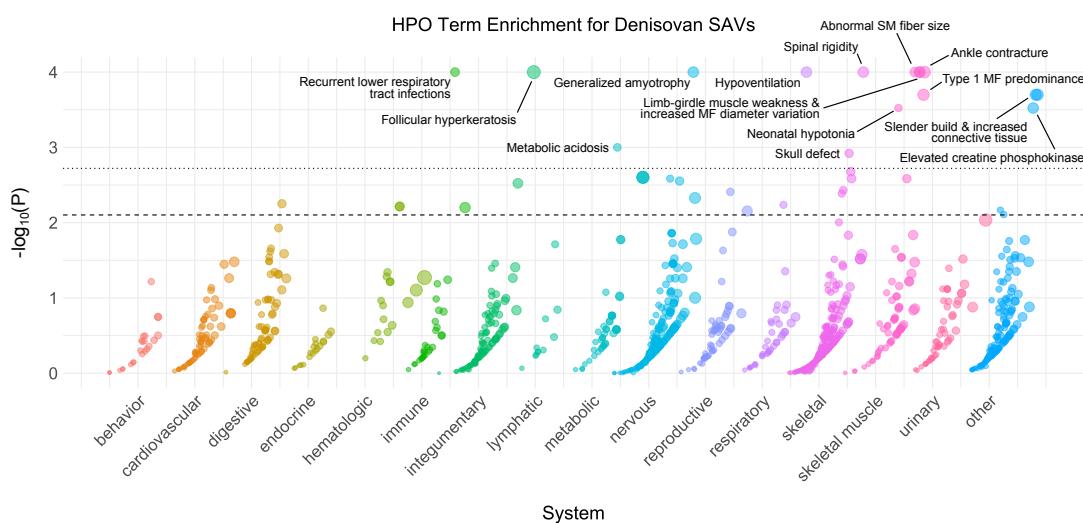
### Figure S9: Chagyrskaya phenotype enrichment.

(A) Phenotype associations enriched among genes with archaic-specific Chagyrskaya SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  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 File S2 for all phenotype enrichment results.

A



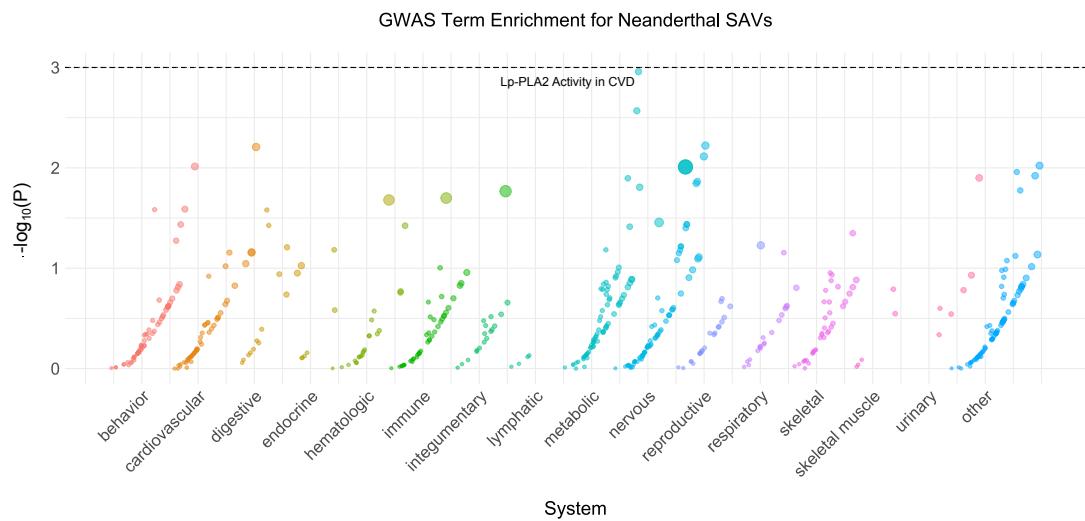
B



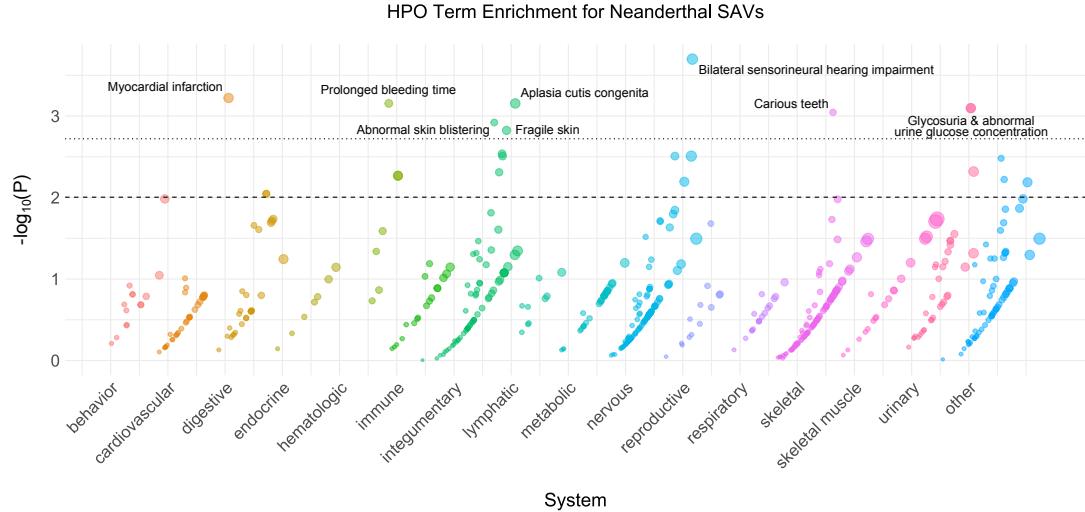
### Figure S10: Denisovan phenotype enrichment.

**(A)** Phenotype associations enriched among genes with archaic-specific Denisovan SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  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 **File S2** for all phenotype enrichment results.

A



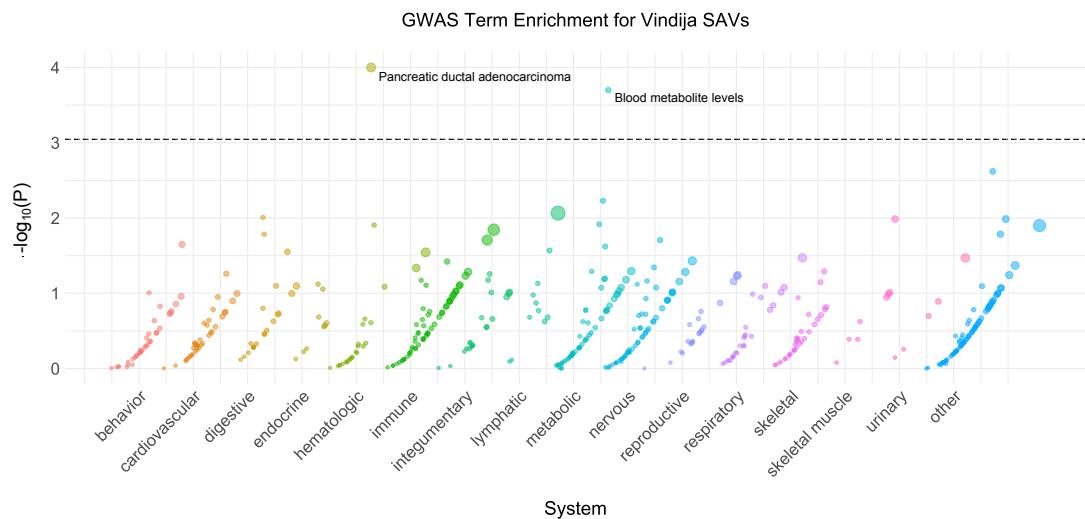
B



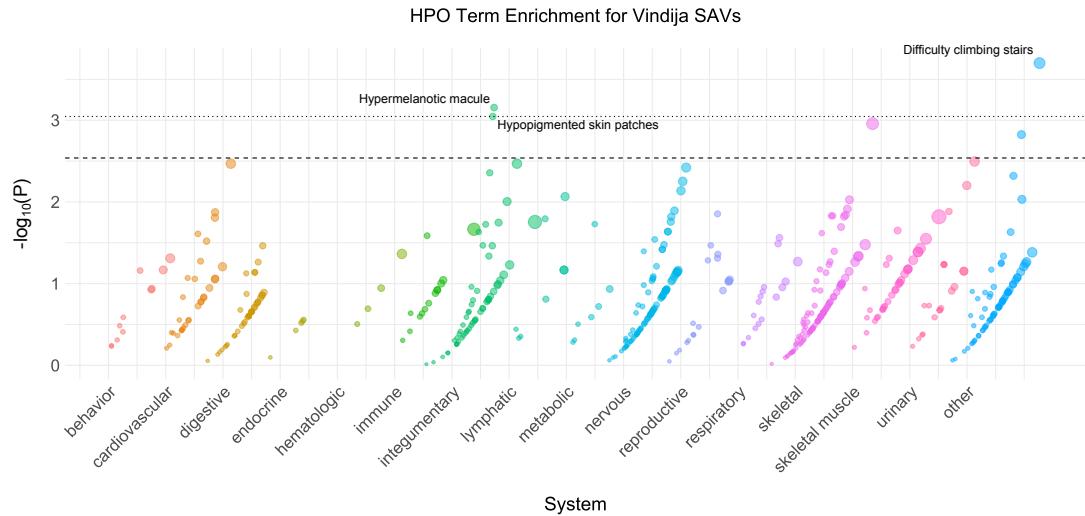
### Figure S11: Neanderthal phenotype enrichment.

**(A)** Phenotype associations enriched among genes with archaic-specific Neanderthal SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  the stricter FDR threshold (0.05) is annotated per system. **(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 **File S2** for all phenotype enrichment results.

A

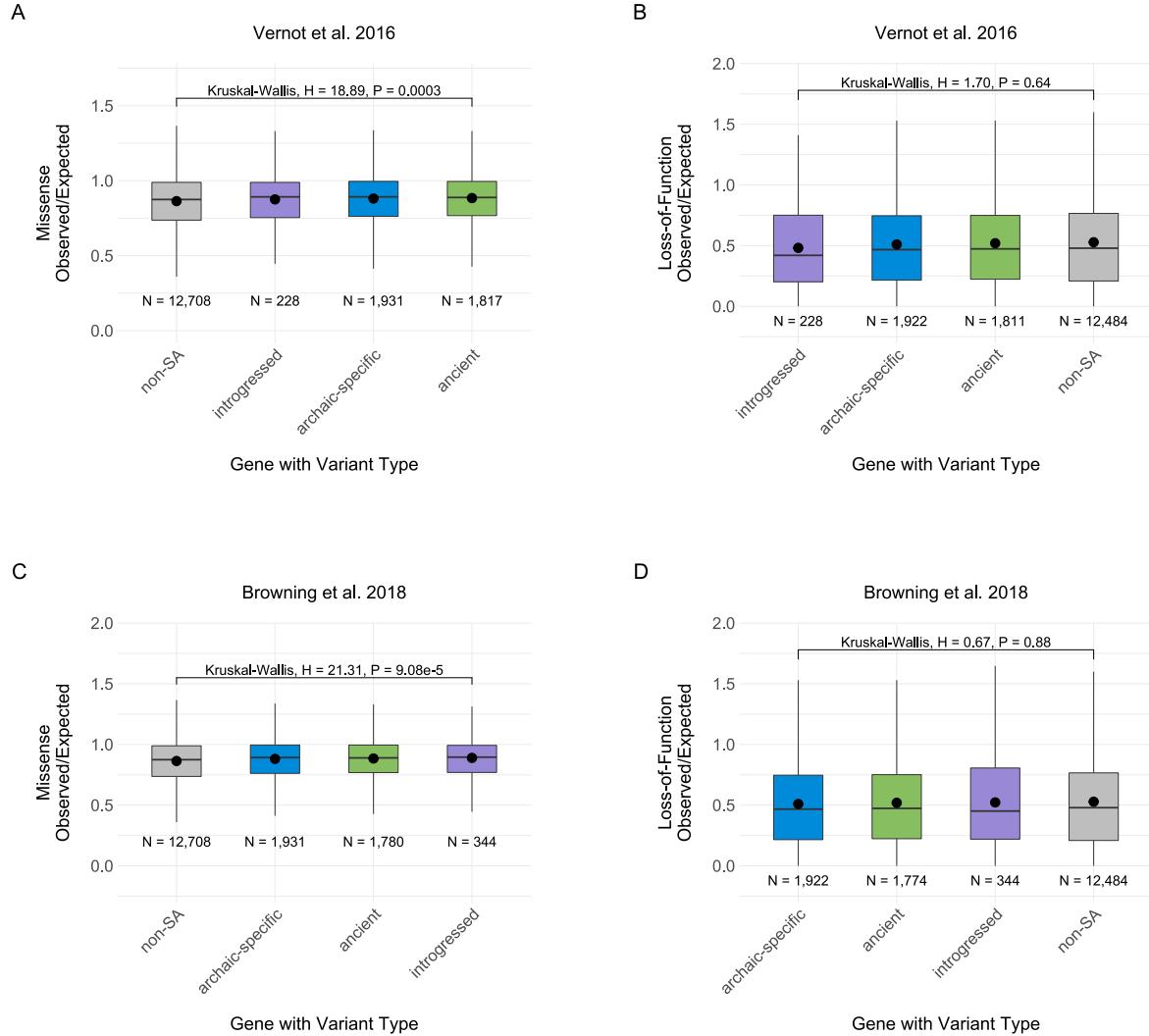


B



### Figure S12: Vindija phenotype enrichment.

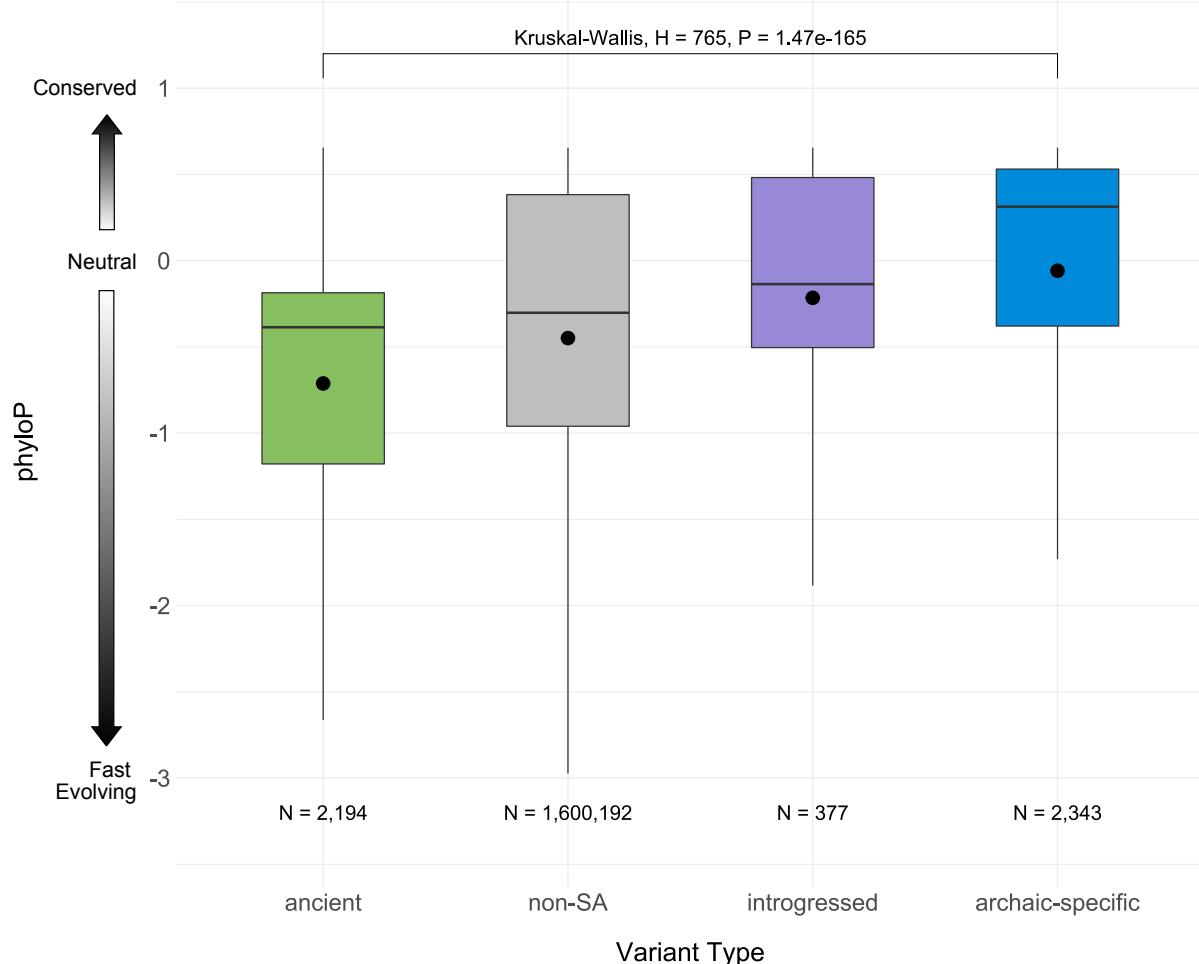
**(A)** Phenotype associations enriched among genes with archaic-specific Vindija SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  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 **File S2** for all phenotype enrichment results.



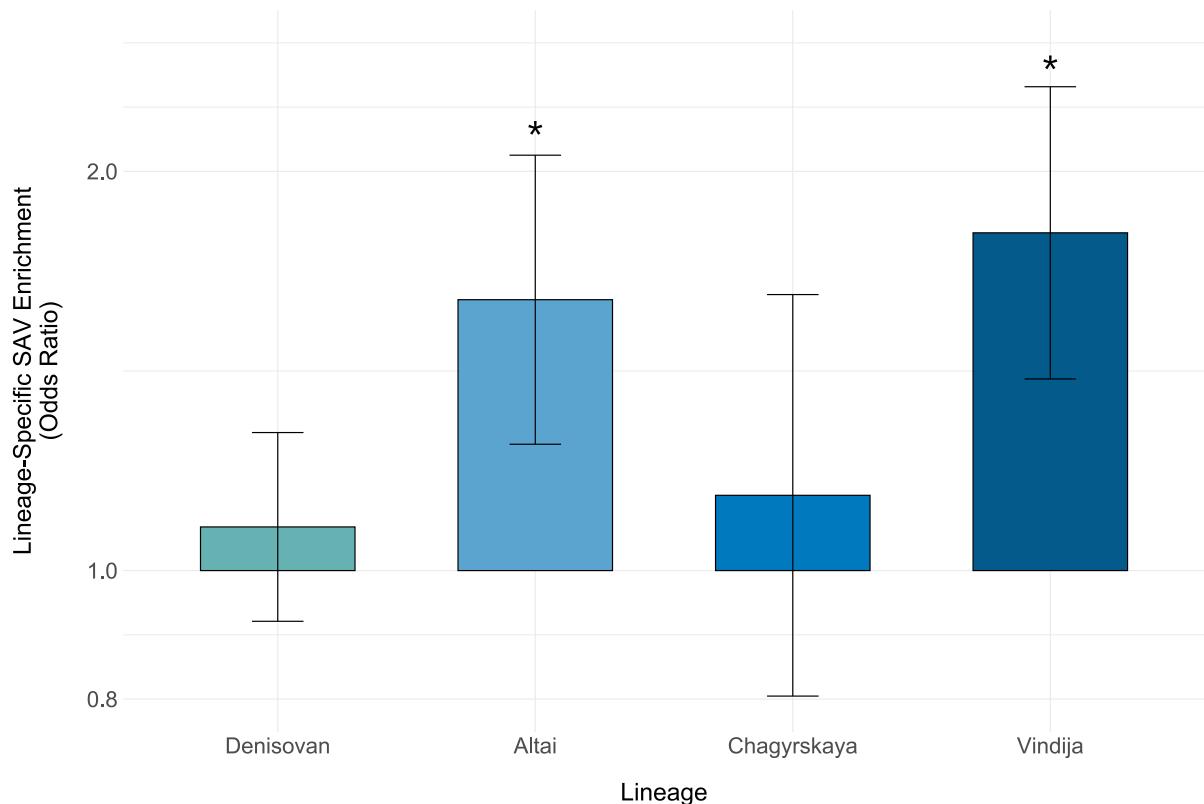
**Figure S13: Gene mutational tolerance does not constrain the evolutionary history of SAVs.**

**(A)** Observed over expected ratio for missense variants per gene from gnomAD among four different sets. The ancient, archaic-specific, and introgressed sets from Browning et al., 2018 include any gene that had  $\geq 1$  SAV at the  $\Delta \geq 0.2$  threshold. Non-SA genes are all genes that do not occur in any of the other three sets. Boxplots indicate the first quartile, median, and third quartile and the mean is noted by the black point. Ns reflect the number of variants per set. **(B)** Observed over expected ratio for loss-of-function variants per gene from gnomAD among four different sets from Browning et al., 2018. The ancient, archaic-specific, and introgressed sets include any gene that had  $\geq 1$  SAV at the  $\Delta \geq 0.2$  threshold. Non-SA genes are all genes that do not occur in any of the other three sets. Boxplots indicate the first quartile, median, and third quartile and the mean is noted by the black point. Ns reflect the number of variants per set. **(C)** Observed over expected ratio for missense variants per gene from gnomAD among four different sets from Vernot et al., 2016. **(D)** Observed over expected ratio for loss-of-function variants per gene from gnomAD among four different sets from Vernot et al., 2016.

Browning et al. 2018

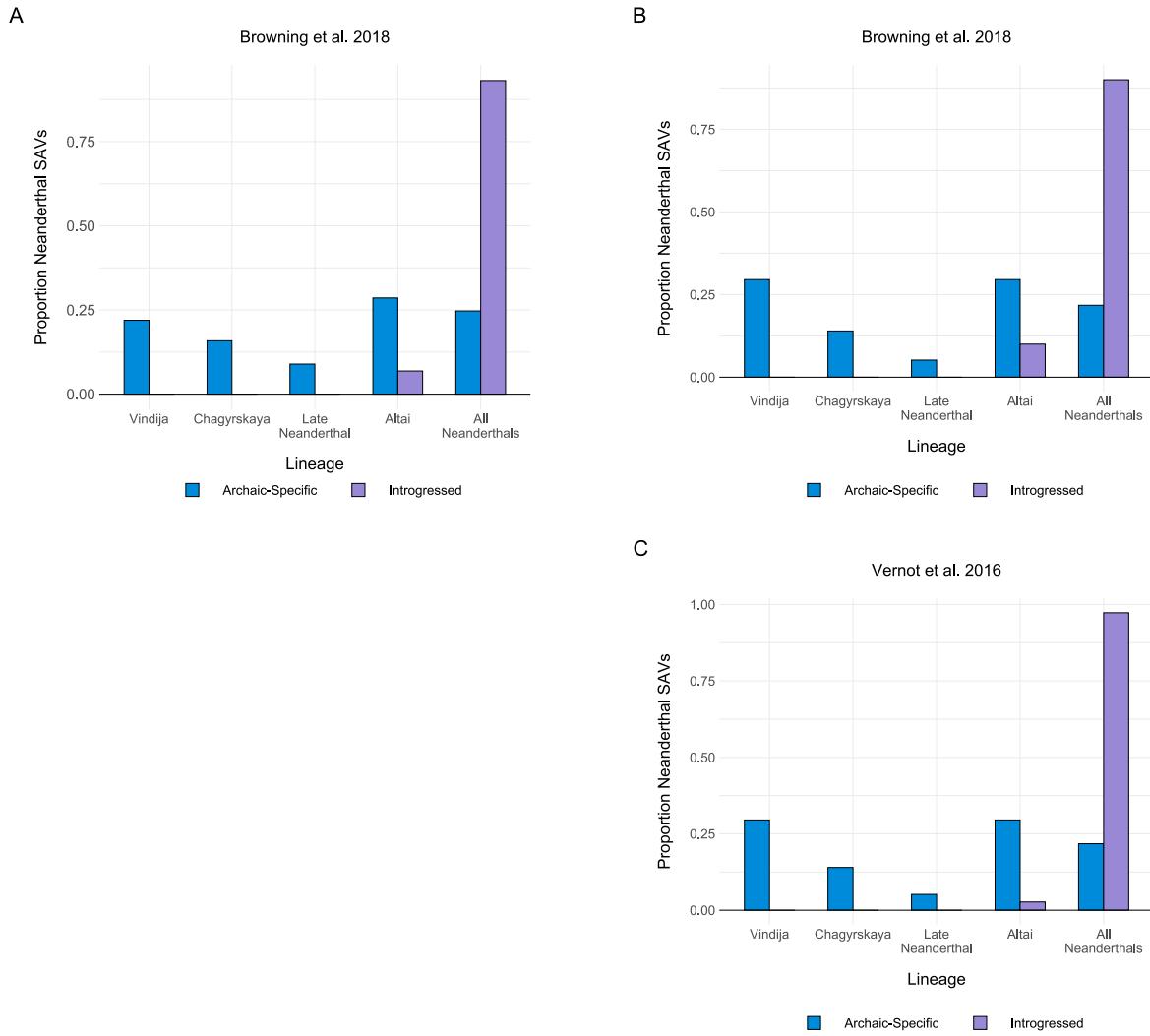


**Figure S14: Site-level evolutionary conservation varies across SAVs with different origins.** phyloP evolutionary conservation score distributions for archaic SAVs of different origins and non-SAVs. Positive scores indicate substitution rates slower than expected under neutral evolution (conservation), while negative scores indicate higher substitution rates than expected (fast evolution). The boxplots give the first quantile, median, and third quantile of the distributions, and the mean is noted by the black point. Ns are the number of variants per set. SAV classifications were based on Browning et al., 2018 introgression calls.



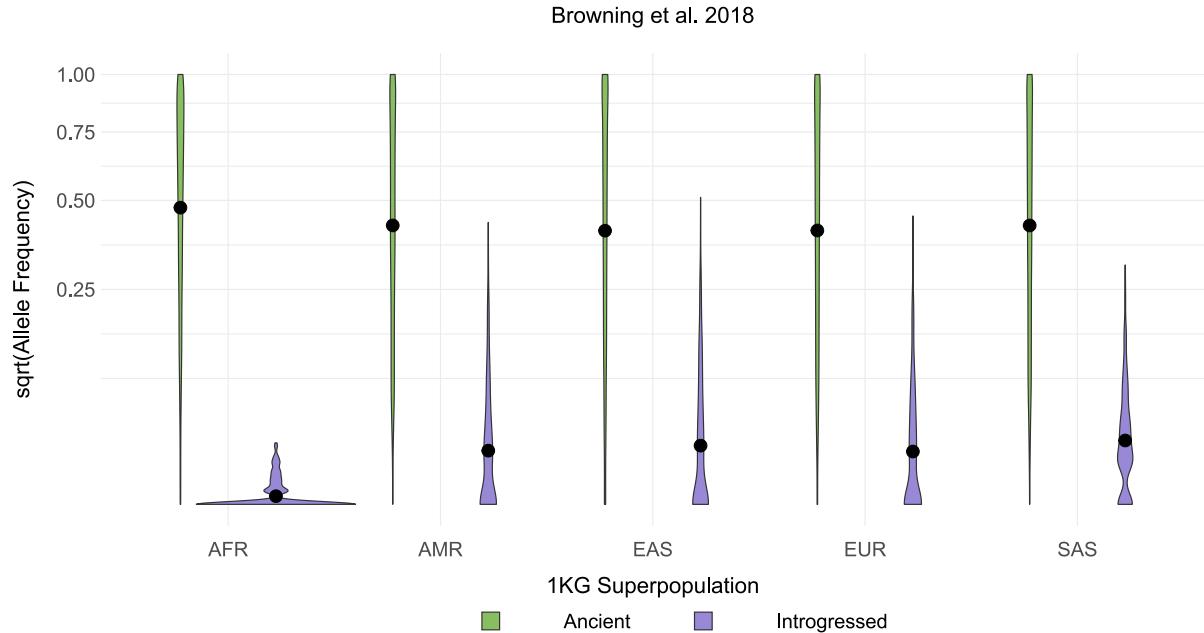
**Figure S15: Lineage-specific high-confidence SAVs are enriched in two Neanderthals.**

Odds ratios from Fisher's exact test performed for each lineage's unique high confidence SAVs/non-SAVs compared to those shared among all four individuals. The number of variants used in each enrichment test are listed in **Table S6**. Asterisks reflect significance using a Bonferroni corrected  $\alpha$  (0.0125). Error bars denote the 95% CI. Note the y-axis is  $\log_{10}$  transformed.

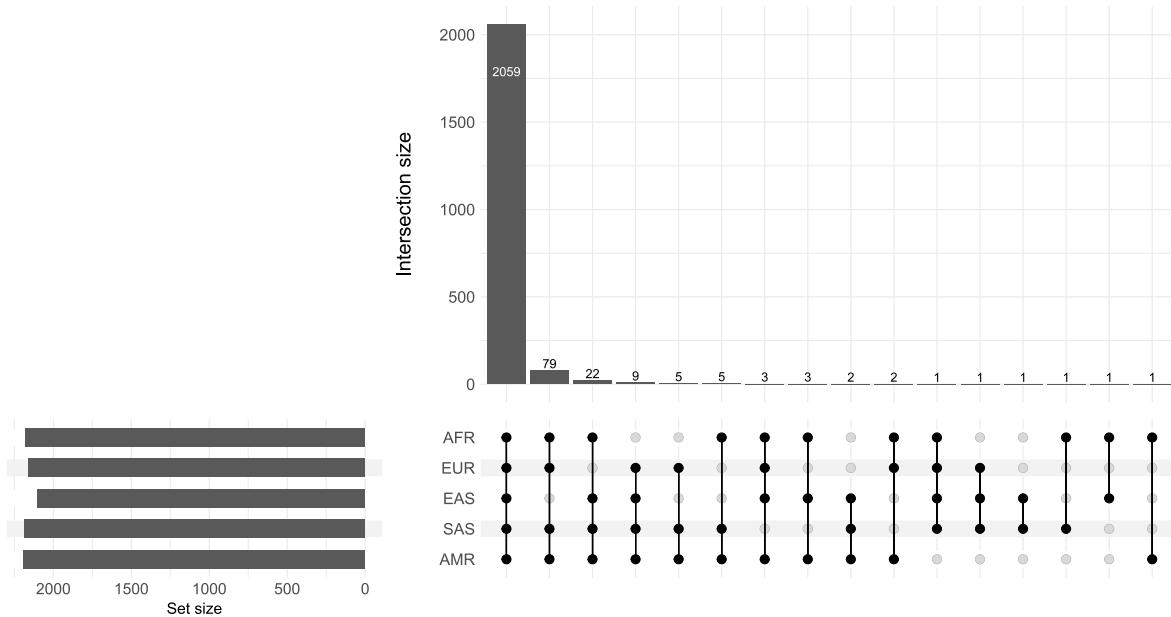


**Figure S16: Introgressed SAVs are largely older variants.**

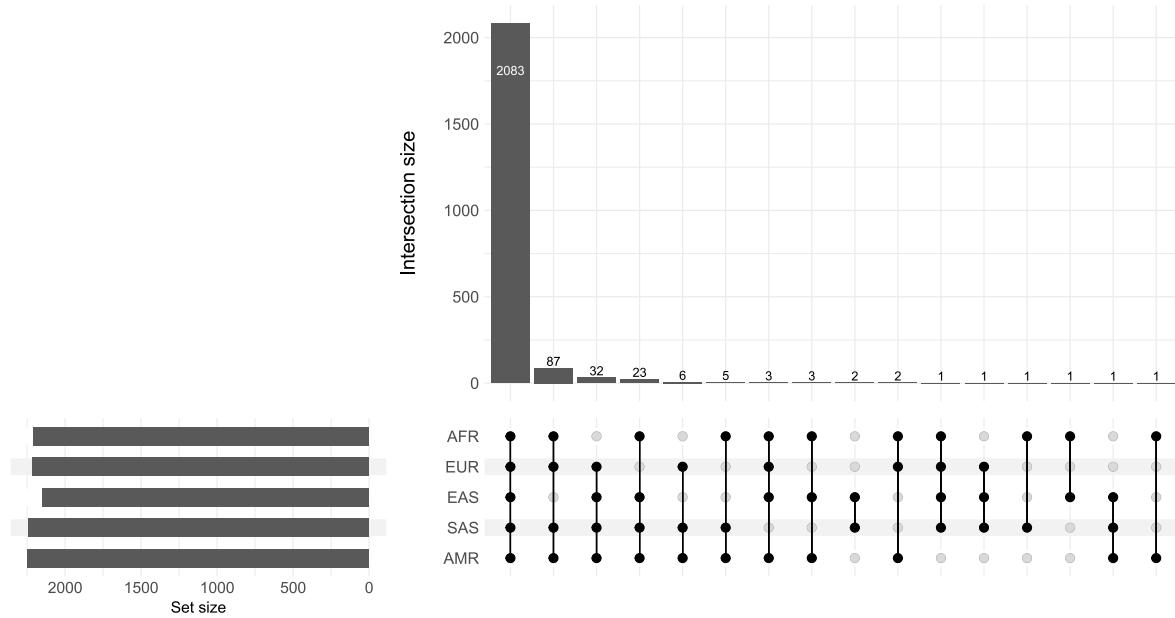
(A) Proportion of SAVs at  $\Delta \geq 0.2$  per Neanderthal lineage among archaic-specific SAVs (expected) and introgressed SAVs (observed) from Browning et al., 2018. Proportions were calculated from the sum of all Neanderthal lineages because power to detect introgressed Denisovan SAVs is low. All data are presented in **Table S7**. (B) Proportion of SAVs at  $\Delta \geq 0.5$  using introgressed SAVs from Browning et al., 2018. (C) Proportion of SAVs at  $\Delta \geq 0.5$  using introgressed SAVs from Vernot et al., 2016.



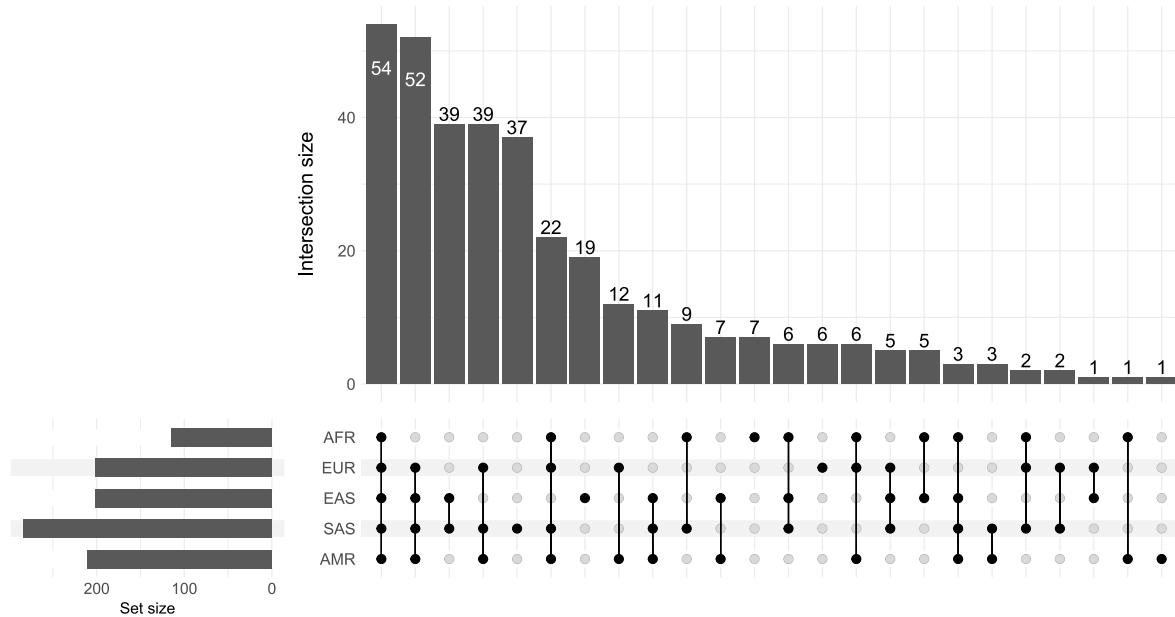
**Figure S17: Ancient SAVs occur a moderate to high frequencies, whereas introgressed SAVs occur at lower frequencies.** Allele frequency distributions for SAVs at  $\Delta \geq 0.2$  per Browning et al., 2018 by 1KG superpopulation. Allele frequencies are from 1KG. If the introgressed allele was the reference allele, we subtracted the 1KG allele frequency from 1. The black dot represents the mean allele frequency. Note the y-axis is square root transformed.



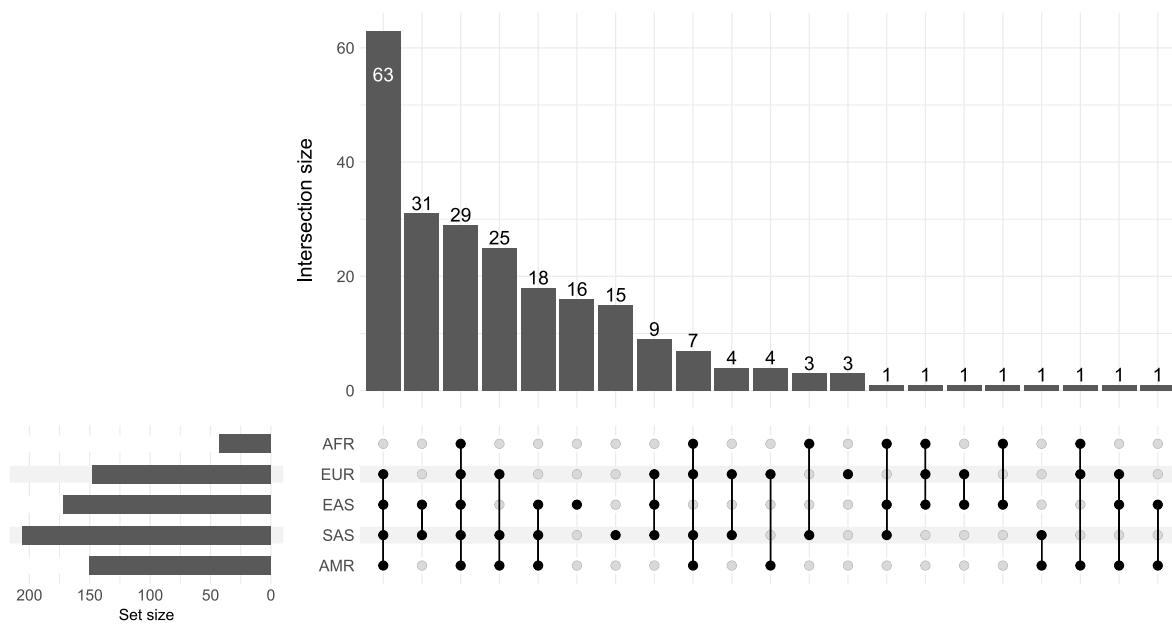
**Figure S18: Most ancient SAVs occur in all 1KG superpopulations.** Unique and shared ancient SAVs per Browning et al., 2018 at  $\Delta \geq 0.2$ . Allele frequencies are from 1KG. By definition, each ancient SAV was considered present in a population if the allele frequency was  $\geq 0.05$  for at least two superpopulations. AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian.



**Figure S19: Most ancient SAVs occur in all 1KG superpopulations.** Unique and shared ancient SAVs per Vernot et al., 2016 at  $\Delta \geq 0.2$ . Allele frequencies are from 1KG. By definition, each ancient SAV was considered present in a population if the allele frequency was  $\geq 0.05$  for at least two superpopulations. AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian.



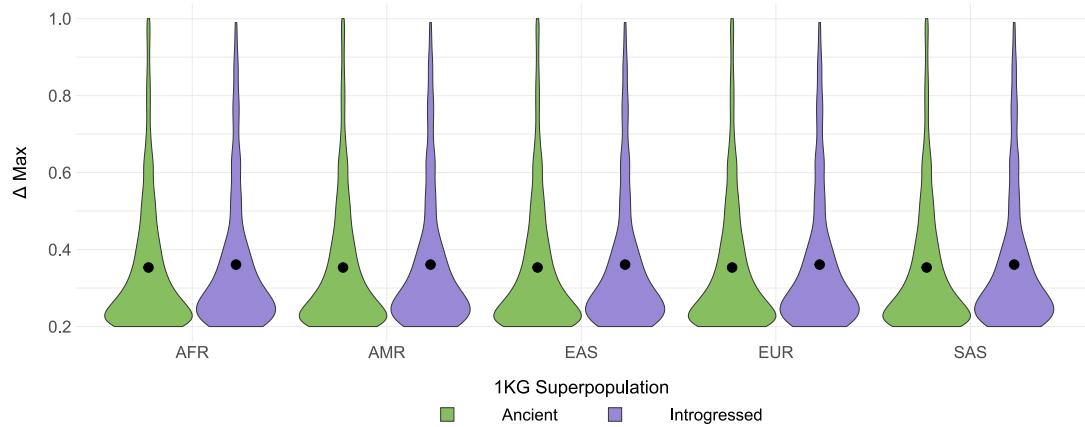
**Figure S20: Many introgressed SAVs occur in multiple 1KG superpopulations.** Unique and shared introgressed SAVs per Browning et al., 2018 at  $\Delta \geq 0.2$ . Allele frequencies are from 1KG. If the introgressed allele was the reference allele, we subtracted the 1KG allele frequency from 1. Each SAV was considered present in a population if the allele frequency was  $> 0.01$  ( $N = 349$ ). The remaining 28 variants occurred at  $\leq 0.01$  allele frequency. AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian.



**Figure S21: Many introgressed SAVs occur in multiple 1KG superpopulations.** Unique and shared introgressed SAVs per Vernot et al., 2016 at  $\Delta \geq 0.2$ . Allele frequencies are from the Vernot et al., 2016 metadata. Each SAV was considered present in a population if the allele frequency was  $> 0.01$  ( $N = 203$ ). The remaining 34 variants occurred at  $\leq 0.01$  allele frequency. AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian.

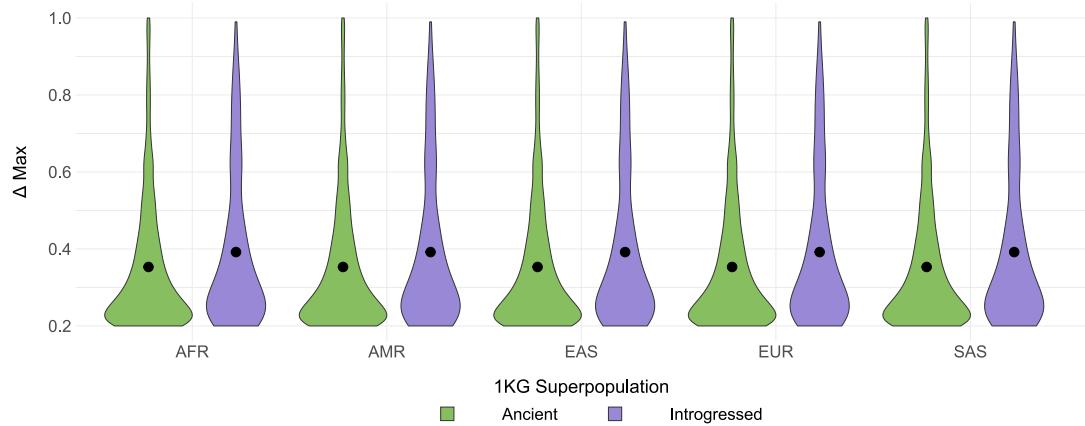
A

Browning et al. 2018



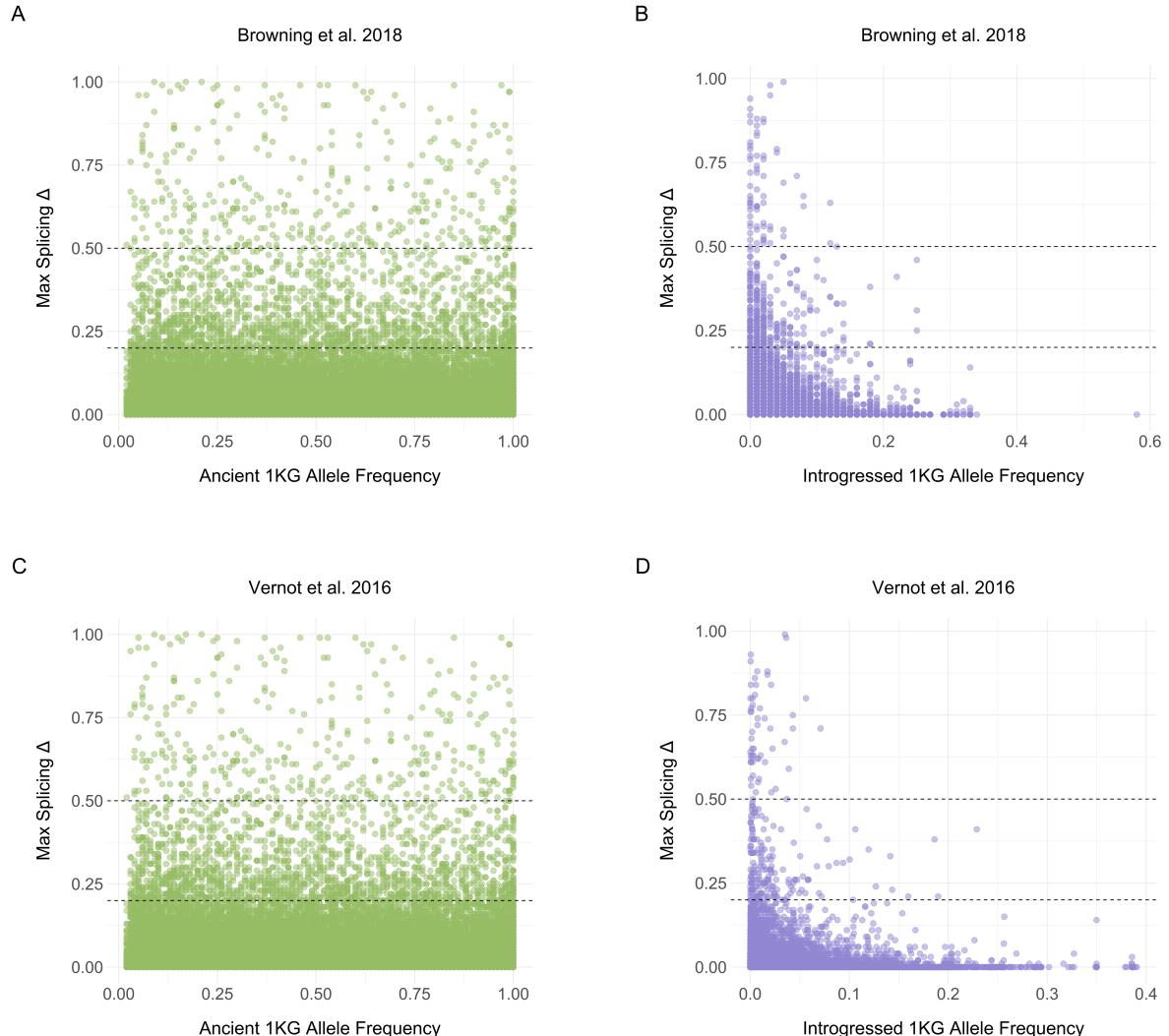
B

Vernot et al. 2016



**Figure S22:  $\Delta_{\text{max}}$  does not differ between 1KG superpopulations.**

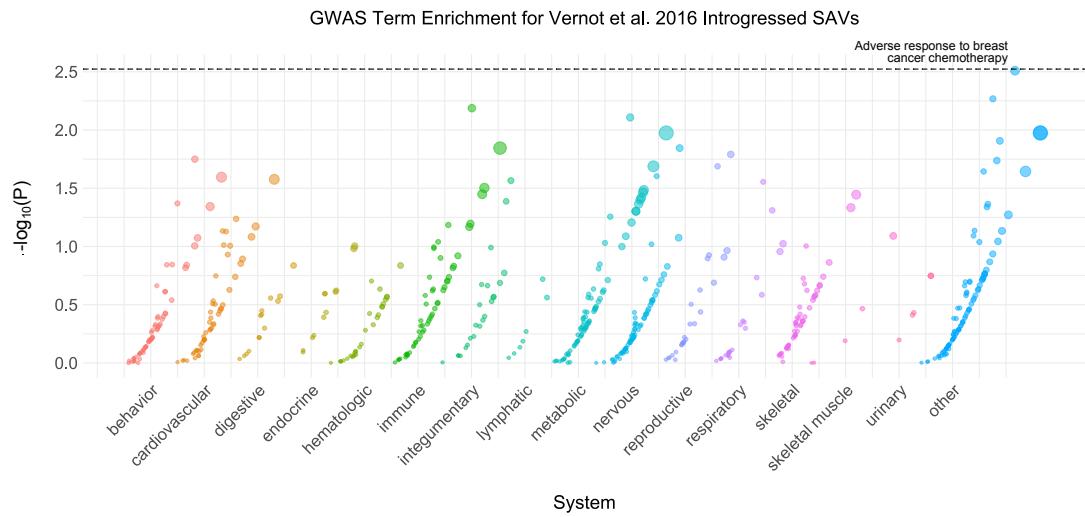
(A)  $\Delta_{\text{max}}$  for SAVs by 1KG superpopulation and allele origin per Browning et al., 2018. AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian. (B)  $\Delta_{\text{max}}$  for SAVs by 1KG superpopulation and allele origin per Vernot et al., 2016.



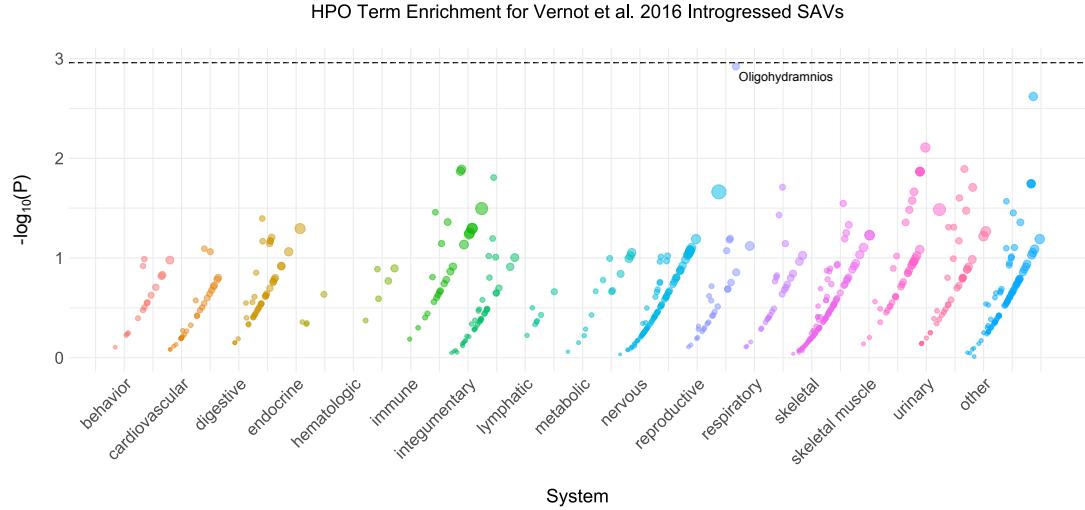
**Figure S23:  $\Delta$  max exhibits a variable relationship to 1KG allele frequency.**

(A) 1KG allele frequency and  $\Delta$  max for all ancient variants per Browning et al., 2018. Allele frequencies are from 1KG. Dashed lines reflect both  $\Delta$  thresholds. (B) 1KG allele frequency and  $\Delta$  max for all introgressed variants per Browning et al., 2018. 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  $\Delta$  max for all ancient variants per Vernot et al., 2016. Allele frequencies are from 1KG. (D) 1KG allele frequency and  $\Delta$  max for all introgressed variants per Vernot et al., 2016. Allele frequencies represent the mean from the AFR, AMR, EAS, EUR, SAS frequencies from the Vernot et al., 2016 metadata.

A



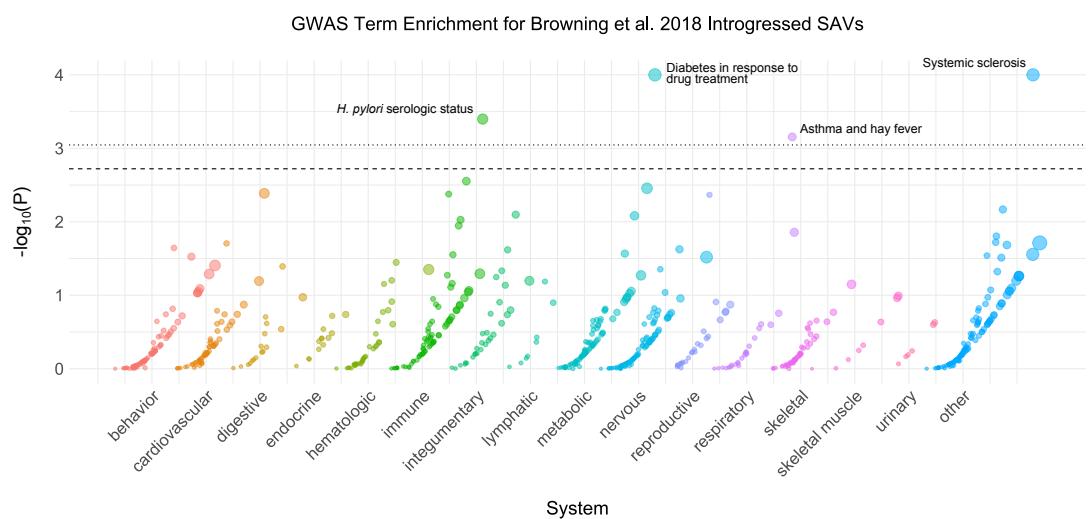
B



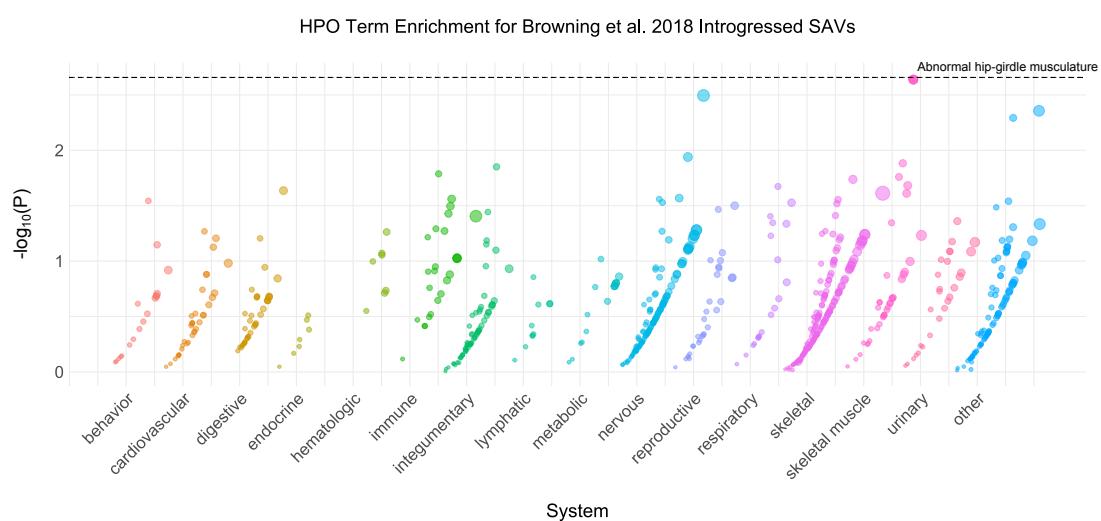
### Figure S24: Vernot et al. 2016 introgressed phenotype enrichment.

(A) Phenotype associations enriched among genes with Vernot et al., 2016 introgressed SAVs based on annotations from GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  the stricter FDR threshold (0.05) is annotated per system. (B) Phenotypes enriched among genes with Vernot et al., 2016 introgressed SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See File S2 for all phenotype enrichment results.

A



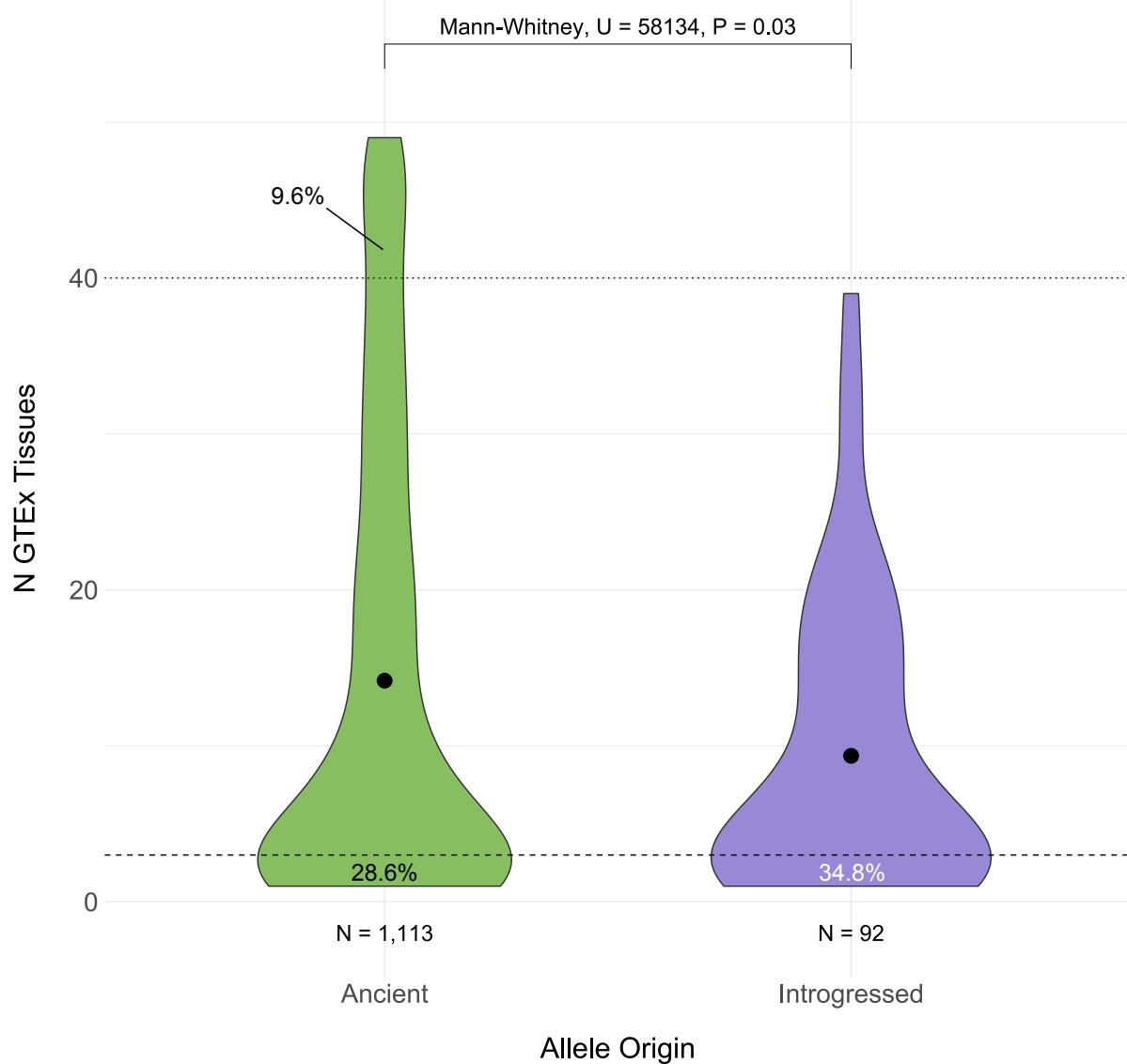
B



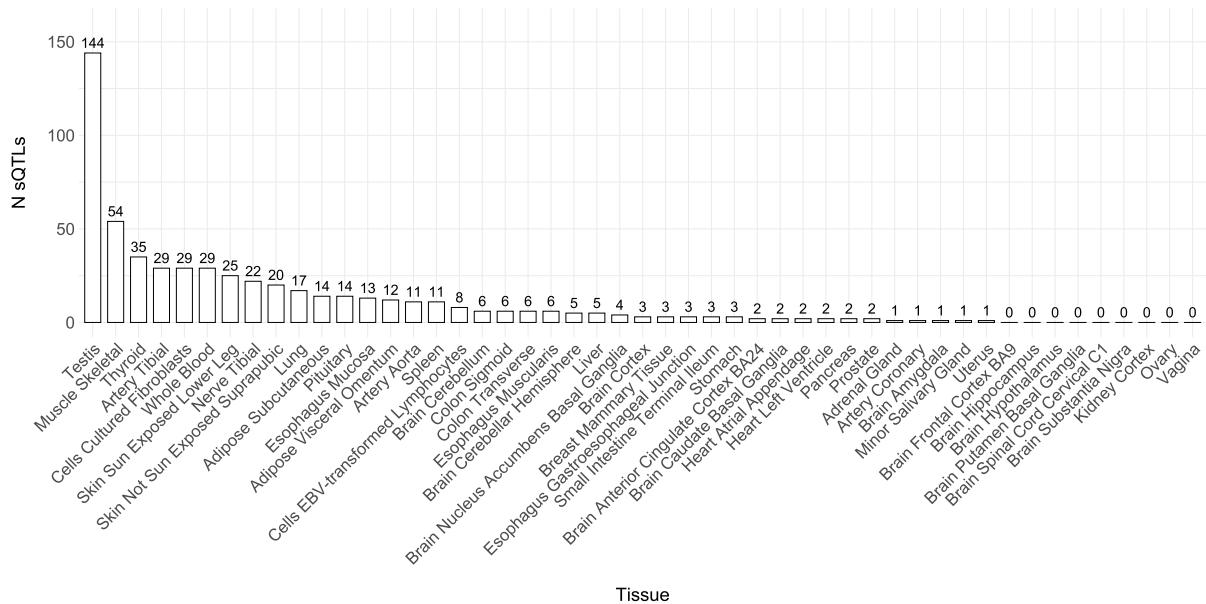
**Figure S25: Browning et al. 2018 introgressed phenotype enrichment.**

(A) Phenotype associations enriched among genes with Browning et al., 2018 introgressed SAVs based on annotations from the GWAS Catalog. Phenotypes are ordered by increasing enrichment within manually curated systems. Circle size indicates enrichment magnitude. Enrichment and p-values were calculated from an empirical null distribution generated from 10,000 shuffles of maximum  $\Delta$  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  $\leq$  the stricter FDR threshold (0.05) is annotated per system. (B) Phenotypes enriched among genes with Browning et al., 2018 introgressed SAVs based on annotations from the Human Phenotype Ontology (HPO). Data were generated and visualized as in A. See File S2 for all phenotype enrichment results.

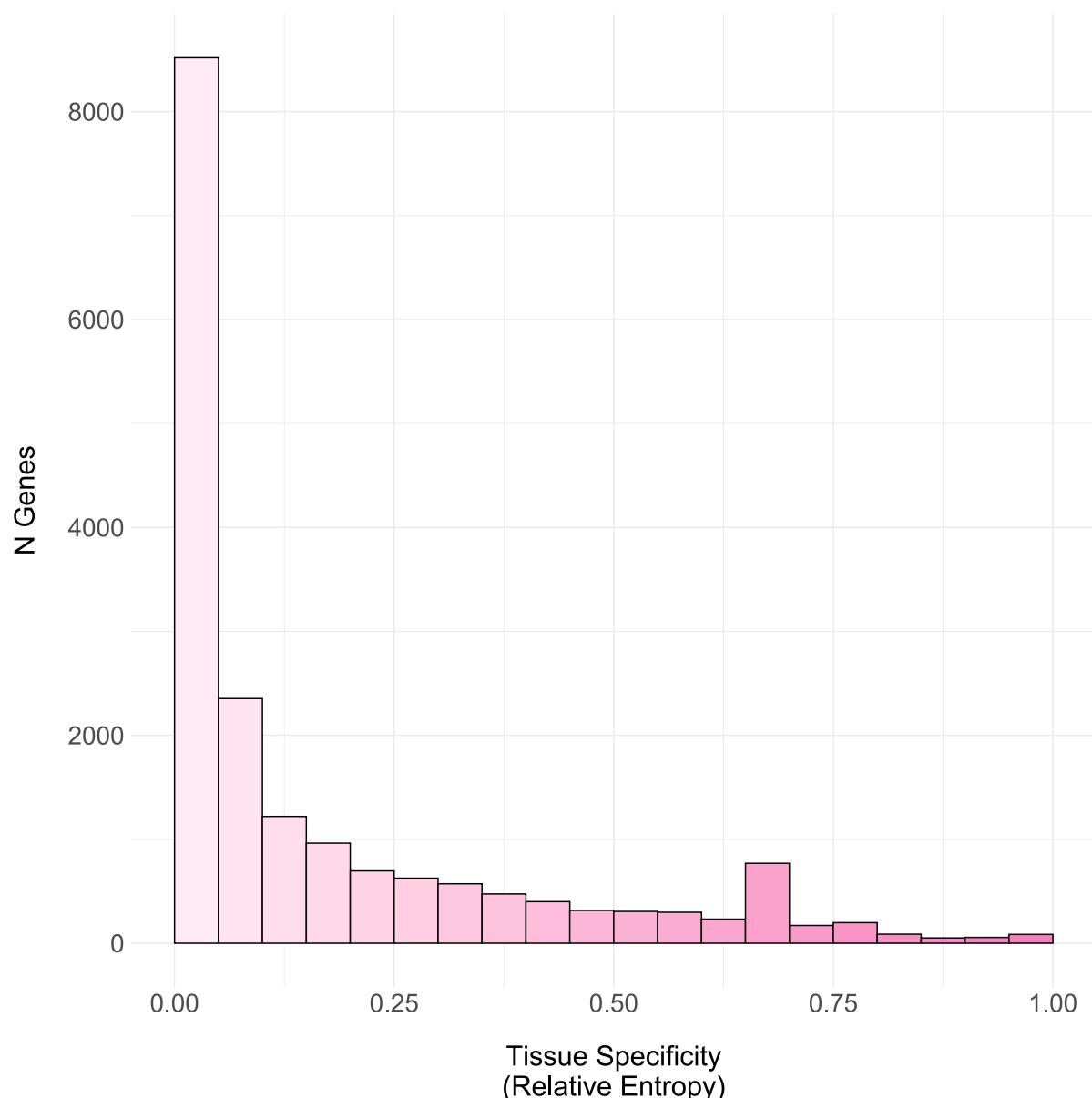
Browning et al. 2018



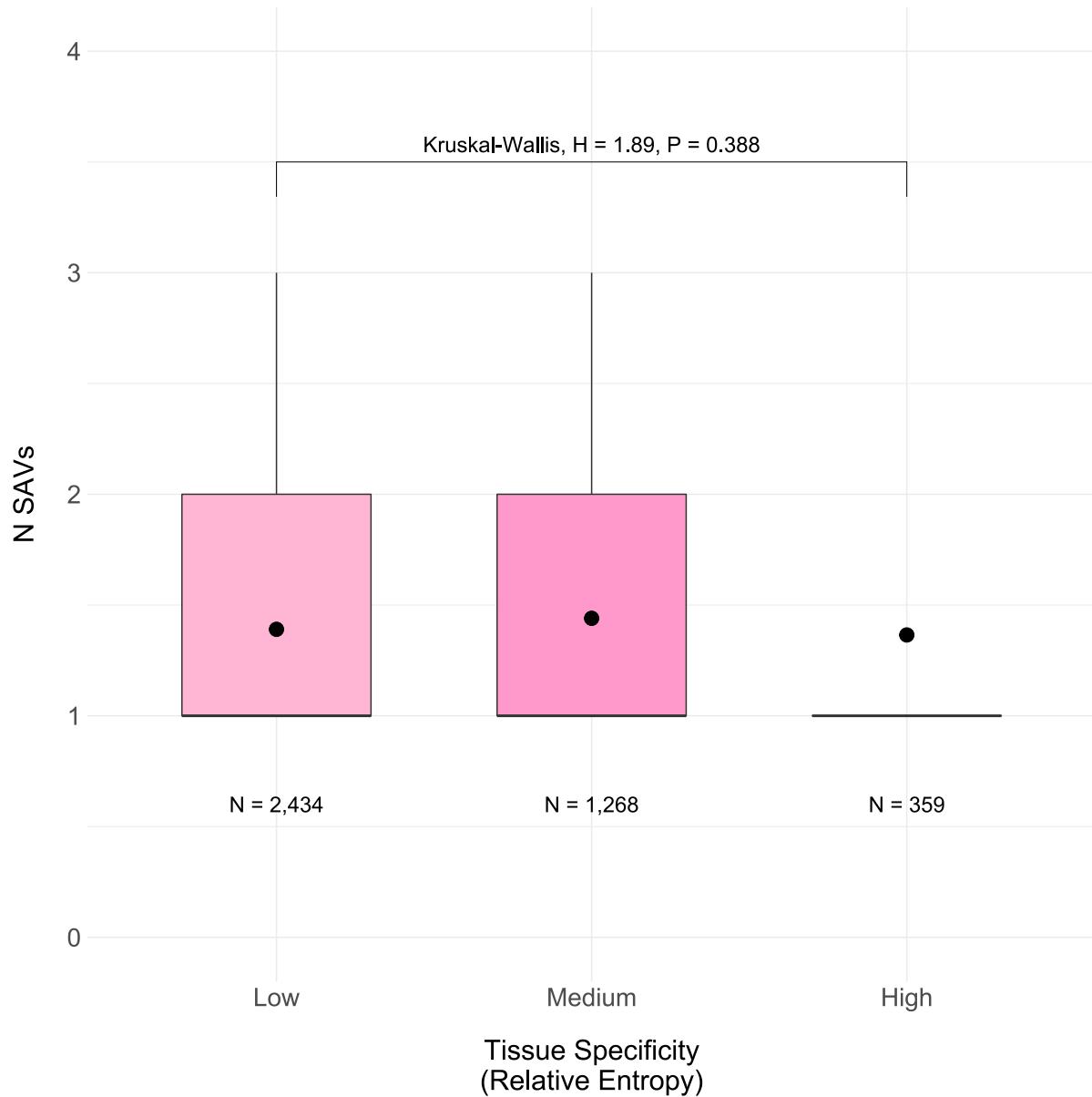
**Figure S26: Introgressed SAV sQTLs are more tissue-specific than ancient variants.** The distribution of GTEx tissues in which an ancient or introgressed SAV per Browning et al., 2018 was identified as an sQTL. We defined “tissue-specific” variants as those occurring in 1 or 2 tissues and “core” sQTLs 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 for both origins are annotated.



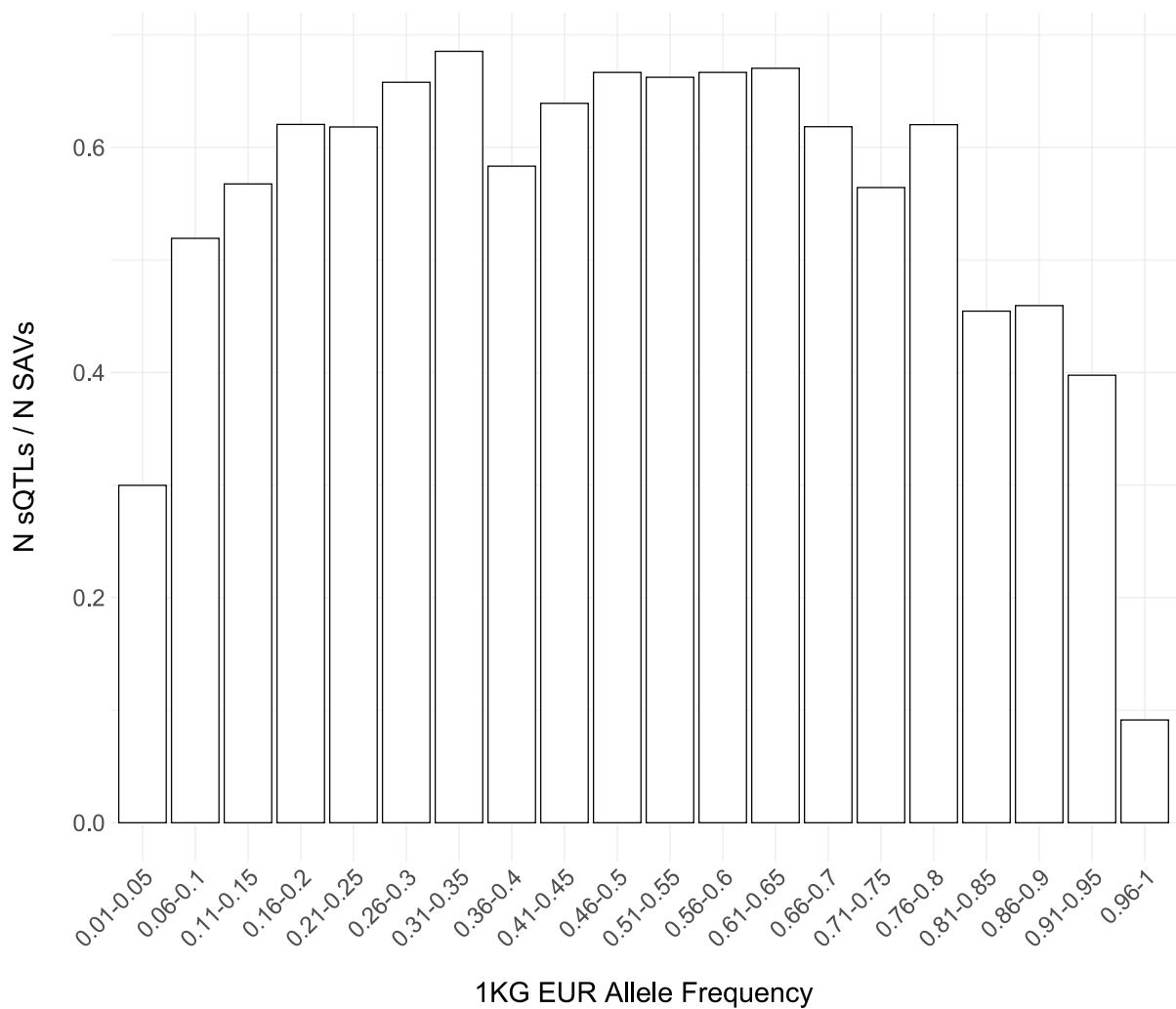
**Figure S27: sQTL SAVs predominantly occur in testis, muscle skeletal, thyroid, tibial artery, fibroblasts, skin, and tibial nerve tissues.** The number of tissue-specific (N GTEx tissues = 1 or 2) sQTL SAVs per GTEx tissue.



**Figure S28: Most gene expression is not tissue-specific.** The distribution of relative entropy in 0.05 bins calculated from GTEx TPM counts across 34 tissues for 18,392 genes.



**Figure S29: The number of SAVs is unrelated to tissue-specific gene expression.** The number of SAVs by binned gene expression calculated from GTEx TPM counts across 34 tissues. We analyzed a total of 4,061 genes with SAVs. Low tissue specificity reflects relative entropy  $\leq 0.1$ , medium tissue specificity reflects relative entropy  $> 0.1$  and  $\leq 0.5$ , and high tissue specificity reflects relative entropy  $> 0.5$  (Figure S28).



**Figure S30: > 50% of SAVs are sQTLs.** The proportion of sQTLs also identified SAVs  $\geq$  binned allele frequency thresholds among 1KG European individuals. We used the precalculated frequencies provided in the 1KG VCFs for this analysis.

921 **5.4 Supplemental Tables**

All Variants				
	./.	0/0	1/0	1/1
Altai	6,226	591,371	126,440	883,313
Chagyrskaya	37,156	583,787	98,076	888,331
Denisovan	21,574	546,792	139,774	899,210
Vindija	29,672	570,175	119,918	887,585

Variants with $\Delta \text{Max} \geq 0.2$				
	./.	0/0	1/0	1/1
Altai	18	2,227	538	3,167
Chagyrskaya	266	2,202	397	3,085
Denisovan	157	2,122	587	3,084
Vindija	163	2,160	496	3,131

Variants with $\Delta \text{Max} \geq 0.5$				
	./.	0/0	1/0	1/1
Altai	4	386	101	558
Chagyrskaya	69	394	61	525
Denisovan	45	376	107	521
Vindija	35	361	108	545

**Table S1: Archaic genotype distribution.** The number of genotypes per individual for all variants and both delta thresholds. “./.” reflects a missing genotype. “0/0” reflects a homozygous site for the reference allele. “1/0” reflects a heterozygote. “1/1” reflects a homozygote for the alternate allele.

1KG Superpopulation	1KG Population	Sample Name	N SAVs
AFR	ESN	HG00137	3,266
	GWD	HG00338	3,344
	LWK	HG00619	3,320
	MSL	HG01198	3,514
	YRI	HG01281	3,493
AMR	CLM	HG01524	3,363
	MXL	HG01802	3,363
	PEL	HG02142	3,349
	PUR	HG02345	3,385
EAS	CDX	HG02629	4,120
	CHB	HG03060	4,160
	CHS	HG03190	4,142
	JPT	HG03708	3,368
	KHV	HG03711	3,379
EUR	CEU	HG03800	3,435
	FIN	HG04014	3,421
	GBR	NA11830	3,200
	GIH	NA18552	3,382
	IBS	NA18868	4,102
	TSI	NA19011	3,321
SAS	BEB	NA19452	4,078
	ITU	NA19741	3,284
	PJL	NA20537	3,168
	STU	NA21141	3,487

**Table S2: The average number of SAVs per randomly sampled individual in 1KG.** We randomly sampled one individual per 1KG population and ran SpliceAI on their combined variants. There were 4,582,422 total SpliceAI annotated variants among autosomal SNVs and 14,006 variants with a  $\Delta_{\text{max}} \geq 0.2$ . These N SAVs reported here include any site where an individual had at least one alternate allele present (i.e., heterozygotes and homozygotes for the alternate allele).

N SAVs	N Genes	Genes
1	3,111	
2	769	
3	246	
4	71	
5	19	
6	14	
7	6	<i>CDH13, CDH23, CSMD1, DNAH17, HLA-DPA1, LAMA5</i>
8	1	<i>ADARB2</i>
9	2	<i>SDK1, WWOX</i>
10	1	<i>HLA-DPB1</i>
11	2	<i>CNTNAP2, SSPO</i>

**Table S3: Distribution of SAVs per gene.** The distribution of SAVs per gene with the specific genes listed for those with  $\geq 7$  SAVs.

$\Delta$	Gene Characteristic	$\rho$	P Value
0.2	N Exons	0.3161	0
	CDS/Exon Length	0.1851	1.07E-135
	Gene Length	0.2920	0
	N Isoforms	0.1767	1.87E-113
0.5	N Exons	0.1330	2.20E-70
	CDS/Exon Length	0.0625	1.01E-16
	Gene Length	0.1017	8.99E-42
	N Isoforms	0.0769	1.24E-22

**Table S4: Physical gene characteristic associations with N SAVs per gene.** Spearman correlation between four variables: 1) the number of exons, 2) length of coding sequence in bp, 3) gene body length in bp, 4) the number of isoforms and the N SAVs per gene for both  $\Delta$  thresholds.

$\Delta$	N Unique SAVs	N Unique Non SAVs	N Shared SAVs	N Shared Non SAVs	OR	95% CI Lower Bound	95% CI Upper Bound	P Value	Bonferroni
0.2	2,416	615,666	1,933	571,264	1.160	1.092	1.231	0.000	Y
0.3	1,175	616,907	945	572,252	1.153	1.059	1.257	0.001	Y
0.4	685	617,397	533	572,664	1.192	1.064	1.335	0.002	Y
0.5	437	617,645	328	572,869	1.236	1.071	1.426	0.004	Y

**Table S5: Enrichment tests for the sum of lineage-specific SAVs at different  $\Delta$ s.** Input data for Fisher's exact tests. OR = odds ratio; 95% CI Lower Bound and 95% CI Upper Bound = the lower and upper bounds of the 95% confidence interval, respectively; P Value = unadjusted p-value.

$\Delta$	Distribution	N Unique SAVs	N Unique Non SAVs	N Shared SAVs	N Shared Non SAVs	OR	95% CI Lower Bound	95% CI Upper Bound	P Value	Bonferroni
0.2	Altai	399	81,517	1,933	571,264	1.447	1.298	1.612	0.000	Y
	Chagyrskaya	218	53,457	1,933	571,264	1.205	1.047	1.387	0.011	Y
	Denisovan	1492	410,000	1,933	571,264	1.075	1.004	1.151	0.036	N
	Vindija	307	70,692	1,933	571,264	1.283	1.138	1.448	0.000	Y
	Total	2,416	615,666	1,933	571,264	1.160	1.092	1.231	0.000	NA
0.5	Altai	75	81,841	328	572,869	1.601	1.245	2.057	0.000	Y
	Chagyrskaya	35	53,640	328	572,869	1.140	0.804	1.615	0.452	N
	Denisovan	254	411,238	328	572,869	1.079	0.916	1.271	0.377	N
	Vindija	73	70,926	328	572,869	1.798	1.395	2.317	0.000	Y
	Total	437	617,645	328	572,869	1.236	1.071	1.426	0.004	NA

**Table S6: Enrichment tests for lineage-specific SAVs.** Input data for Fisher's exact tests. OR = odds ratio; 95% CI Lower Bound and 95% CI Upper Bound = the lower and upper bounds of the 95% confidence interval, respectively; P Value = unadjusted p-value.

$\Delta$	Distribution	$\geq 0.2$				$\geq 0.5$			
		Archaic-Specific N SAVs	Browning Introgressed N SAVs	Vernot Introgressed N SAVs	Archaic-Specific N SAVs	Browning Introgressed N SAVs	Vernot Introgressed N SAVs	Archaic-Specific N SAVs	Browning Introgressed N SAVs
Altai	312	15	7	57	4	4	1	0	0
Chagyrskaya	172	0	0	27	0	0	0	0	0
Denisovan	968	2	3	176	2	2	1	1	1
Late Neanderthal	97	0	0	10	0	0	0	0	0
Neanderthal	281	203	0	46	36	36	36	36	36
Other	154	37	141	31	4	4	4	4	4
Shared	203	120	20	43	23	23	23	23	23
Vindija	238	0	66	57	0	0	0	0	0

**Table S7: Distribution of archaic-specific and introgressed SAVs.** N SAVs for both  $\Delta$  thresholds among archaic-specific and introgressed SAVs per Browning et al., 2018 and Vernot et al., 2016. Other denotes any combination of archaics not already listed.

Chromosome	Position	Reference Allele	Alternate Allele	Delta Max	Browning Allele Origin	Vernot Allele Origin	Annotation
chr1	22174518	G	T	0.98	introgressed	introgressed	<i>HSPG2</i>
chr1	55,537,474	C	G	0.33	low-confidence ancient	introgressed	<i>USP24</i>
chr1	161,681,848	C	T	0.20	introgressed	introgressed	<i>FCRLA</i>
chr1	212,985,592	G	A	0.52	introgressed	introgressed	<i>TATDN3</i>
chr11	86,159,859	G	A	0.26	ancient	introgressed	<i>ME3</i>
chr12	133,272,470	G	T	0.26	introgressed	introgressed	<i>PXMP2</i>
chr12	133,272,470	G	T	0.26	introgressed	introgressed	<i>RP13-672B3.2</i>
chr15	85,403,496	G	A	0.33	introgressed	introgressed	<i>ALPK3</i>
chr16	88,924,425	C	G	0.41	introgressed	introgressed	<i>TRAPPC2L</i>
chr19	40,913,595	G	A	0.23	introgressed	low-confidence ancient	<i>PRX</i>
chr22	50,684,852	G	C	0.25	introgressed	introgressed	<i>HDAC10</i>
chr4	170,990,750	G	A	0.22	introgressed	introgressed	<i>AADAT</i>
chr5	6,753,013	C	T	0.24	introgressed	introgressed	<i>PAPD7</i>
chr5	167,919,825	G	A	0.38	introgressed	introgressed	<i>RARS</i>
chr6	33,264,115	A	C	0.34	low-confidence ancient	introgressed	<i>RGL2</i>
chr8	130,763,783	C	T	0.87	introgressed	introgressed	<i>GSDMC</i>

**Table S8: SAVs that exhibit allele-specific expression in modern humans.** 16 SAVs from our dataset that matched variants from McCoy et al., 2017.

Distribution	N sQTLs
Altai	2
Chagyrskaya	3
Denisovan	8
Late Neanderthal	2
Neanderthal	9
Other	14
Shared	74

**Table S9: Core sQTL SAV distribution among the archaics.** The distribution of core sQTL SAVs (N = 1,145) among the archaics. Core sQTLs were defined as those variants detected in > 40 tissues. All the variants represented here are either low or high-confidence ancient.

Chromosome	Position	Reference Allele	Alternate Allele	Delta Max	Alternate Allele	$\Delta$ Max Reference as Alternate
chr1	52,820,680	C	T	0.23		0.23
chr1	152,944,501	A	T	0.23		0.23
chr1	183,750,131	C	A	0.26		0.26
<b>chr1</b>	<b>216,945,756</b>	<b>C</b>	<b>A</b>	<b>0.20</b>		<b>0.19</b>
chr10	17,373,518	T	C	0.22		0.22
chr11	84,191,111	T	C	0.40		0.40
chr11	99,345,359	C	G	0.67		0.67
chr17	14,109,360	T	G	0.20		0.21
chr18	60,003,587	G	A	0.31		0.30
chr19	33,410,289	T	C	0.71		0.71
chr20	62,224,595	G	A	0.41		0.41
chr3	52,486,376	T	C	0.34		0.35
chr4	38,805,942	G	C	0.38		0.37
chr4	100,460,531	A	G	0.35		0.35
chr4	135,122,662	G	A	0.80		0.80
chr5	118,669,326	G	C	0.22		0.22
chr5	126,989,007	A	G	0.91		0.91
chr5	148,430,560	A	C	0.32		0.32
chr6	51,889,358	T	C	0.42		0.42
chr6	52,138,226	C	G	0.21		0.21
chr7	119,524,266	C	T	0.30		0.30
chr7	126,797,408	T	A	0.43		0.43
chr7	140,801,296	T	C	0.25		0.25
chr7	140,958,656	A	G	0.26		0.26
<b>chr7</b>	<b>157,177,273</b>	<b>C</b>	<b>T</b>	<b>0.31</b>		<b>0.16</b>
chr7	158,543,832	C	A	0.21		0.22

**Table S10: Maximum  $\Delta$  for SAVs with an introgressed reference allele.**  $\Delta$  maximum for 26 SAVs whose reference allele rather than the alternate allele matched an introgressed tag SNP from Vernot et al., 2016. We switched the reference and alternate alleles for these (and non-SAV) loci and reran SpliceAI. “Delta Max Alternate Allele” indicates the originally predicted  $\Delta$  where the reference allele is introgressed and “Delta Max Reference as Alternate” indicates the  $\Delta$  when the reference and alternate alleles are switched. Bolded variants are those whose values did not pass the SAP threshold after switching alleles.

N Annotations	N
1	153,7451
2	30,448
3	1,154
4	75
5	36
6	15
7	25
8	22
9	36
10	20
11	20
12	15
13	58
14	70
15	34
16	6
17	6
18	32
19	6
20	3
21	24

**Table S11: Distribution of multiple annotations.** The number of variants with the given number of annotations from GENCODE, Human Release 24.