Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences

Supplementary Information

Laura L. Colbran, Eric R. Gamazon, Dan Zhou, Patrick Evans, Nancy J. Cox, and John A. Capra

Correspondence to: tony.capra@vanderbilt.edu

Supplementary Text

Cross-Population Analyses of Gene Regulation

Recent work has raised concerns about the accuracy of predictions based on genetic models when applied both across and within human populations due to demographic, environmental, and other confounding factors[1–4]. Given the divergence of archaic hominins and modern humans, addressing these issues is particularly relevant in this study[5]. Several lines of evidence support our approach.

First, our analyses are based on the identification of divergent regulation, not differential expression. We do not interpret the PrediXcan output as a direct proxy for gene expression, but rather a proxy for divergence in gene regulatory architecture. Furthermore, in many of our analyses, we evaluate the effects of Neanderthal sequences in the modern human genomic context, where the models were trained.

Second, concerns about genetics-based prediction models are largely based on polygenic risk scores (PRSs), which analyze genetic variants genome-wide to predict risk for organism-level phenotypes. In contrast, our approach considers only genetic variants in the *cis*-regulatory region of a gene and a much narrower phenotype, gene expression. Comparisons of the gene regulatory architecture across humans have shown that effects of common variants are largely shared across diverse populations and that the majority of expression quantitative trait loci (eQTL) are conserved[6,7]. In other words, the molecular machinery and genetic architecture of gene regulation are largely conserved across humans; most common human alleles have similar regulatory effects across populations[7]. Despite this there are still differences in gene expression that are attributable to population-level genetic differences[6]. To test how well the PrediXcan models, which were trained on GTEx (a primarily European-ancestry population), generalize across human populations, we applied models trained on GTEx LCLs to expression and genotype data from LCLs from 1000 Genomes European ancestry populations (CEU, GBR, FIN, TSI) and a sub-Saharan African population (YRI)[8]. We then compared the accuracy of predictions between the European and African populations. We find that our approach maintains significant accuracy when models are trained in a European-ancestry population and applied to African individuals, though there is a decrease in the amount of variance in gene expression for many genes (Supplementary Fig. 1).

Third, while there is a reduction in performance across populations[9], cross-population application of PrediXcan has previously enabled identification of relevant cardiometabolic gene-trait associations, supporting the biological relevance of differences observed[10]. In all of our analyses, we compare Neanderthal sequences to individuals from all 1kG human populations, thereby taking cross-population variation into account when interpreting our results. We examined the stability of the imputed values across all 1000 Genomes populations and found that most genes analyzed here have similar imputed distributions between populations (Supplementary Fig. 7). In more detail, for all PrediXcan models in all tissues, we computed the median imputed regulation for each 1000 Genomes population, then found the maximum difference between populations (Supplementary Fig. 7). Only 2.7% of all gene models have a

maximum difference in population median regulation greater than 1 standard deviation. To determine whether genes with population-specific regulation differences were more likely to be divergently regulated in the Altai Neanderthal, we computed the odds ratio for population-specific patterns and divergent regulation. At a threshold of max difference > 1 (matching the threshold set for comparisons across archaic hominins), population-specific models were not more likely to be divergently regulated (OR = 1.06; $P = 0.689$, Fisher's exact test). However, if we used a less stringent threshold for defining population differences (max difference in imputed regulation > 0.5), population-specific genes are more likely to be divergently regulated (OR = 1.35, $P = 3.5E{-}11$). The pattern is similar when the analysis only considers the DR GWARRs. Overall, this suggests that, while some DR genes show moderate differences among human populations, this is not true of the most extreme differences. Furthermore, when excluding genes with evidence of population-specificity from our analyses, hundreds of DR GWARRs remain.

Finally, and most importantly, as described in the Main Text, we demonstrate that models trained on ancestral human sequences have significant accuracy when applied to Neanderthal sequences remaining in modern human genomes (Supplementary Fig. 3).

GWARR Functional Annotations

Beyond the phenotypes significantly enriched for DR GWARRs, several additional DR GWARRs have functions relevant to potentially human-specific phenotypes, like language. For example, *GNPTG*, the gamma subunit of GlcNAc-1-phosphotransferase, is associated with stuttering[11,12], and *CHMP2B*, a member of the chromatin-modifying protein/charged multivesicular body protein (CHMP) family, is associated with aphasia, frontotemporal dementia, and amyotrophic lateral sclerosis (ALS). *CHMP2B* is also in a Neanderthal introgression desert and has seven human accelerated regions (HARs) within its regulatory region.

Several DR GWARRs are involved in processing of melatonin and serotonin (*CYP2U1*, *SULT1A1*, *SULT1A2*) and in regulating photoperiod (*CRY2*, *ARNTL*)[13]. This suggests potential gene regulatory differences in circadian biology between Neanderthal and AMH populations[14,15].

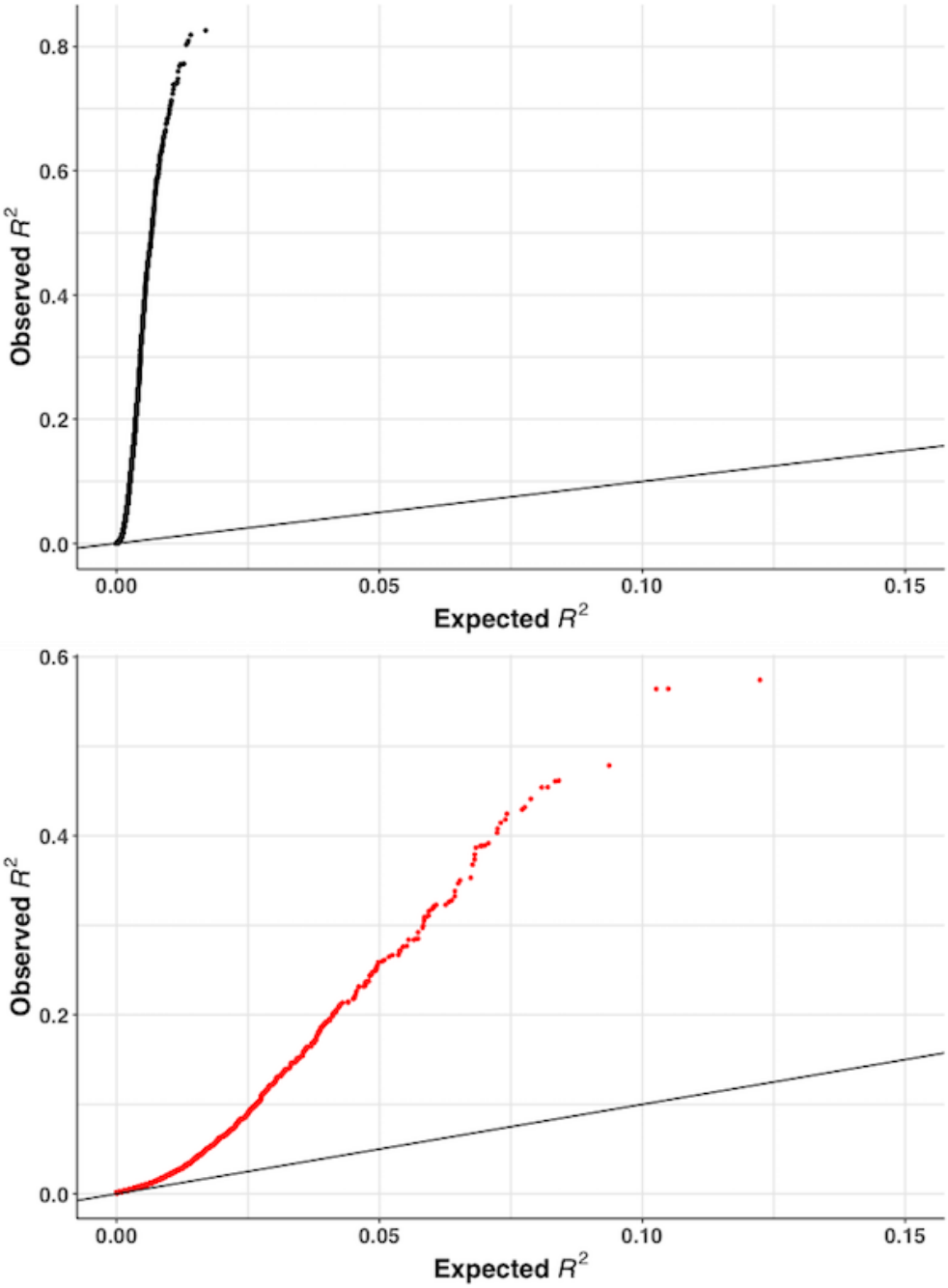
DR genes in introgression deserts are of particular interest, due to their potential relevance to human-specific biology. Several additional DR desert genes have been implicated—either in previous work or our biobank association tests—with a variety of traits important to humanness, including neural development (*CELSR2, CHMP2B*)[16–18] and learning and spatial memory (*CARF*)[19]. *CELSR2* and *CARF* both also have associations with heart disease. In addition, five DR desert genes are loss-of-function intolerant according to gnomAD (<0.35 for the upper bound of the o/e 95% confidence interval): *CELSR2*, *RB1CC1*, *BMPR2*, *ADAM23*, and *MOV10*. Desert genes are also significantly more likely to have a human accelerated region (HAR) within 1 Mb than other GWARRs (OR = 1.36, $P = 0.01$); this trend becomes even stronger when comparing DR desert genes to other DR GWARRs (OR = 2.36, $P = 0.07$). This suggests that regions of the deserts have experienced significant sequence changes since the last

common ancestor of humans and chimpanzees (but prior to the effects captured by our models), and that these regions might therefore be important to human-specific phenotypes.

GWARR Phenotype Associations in BioVU

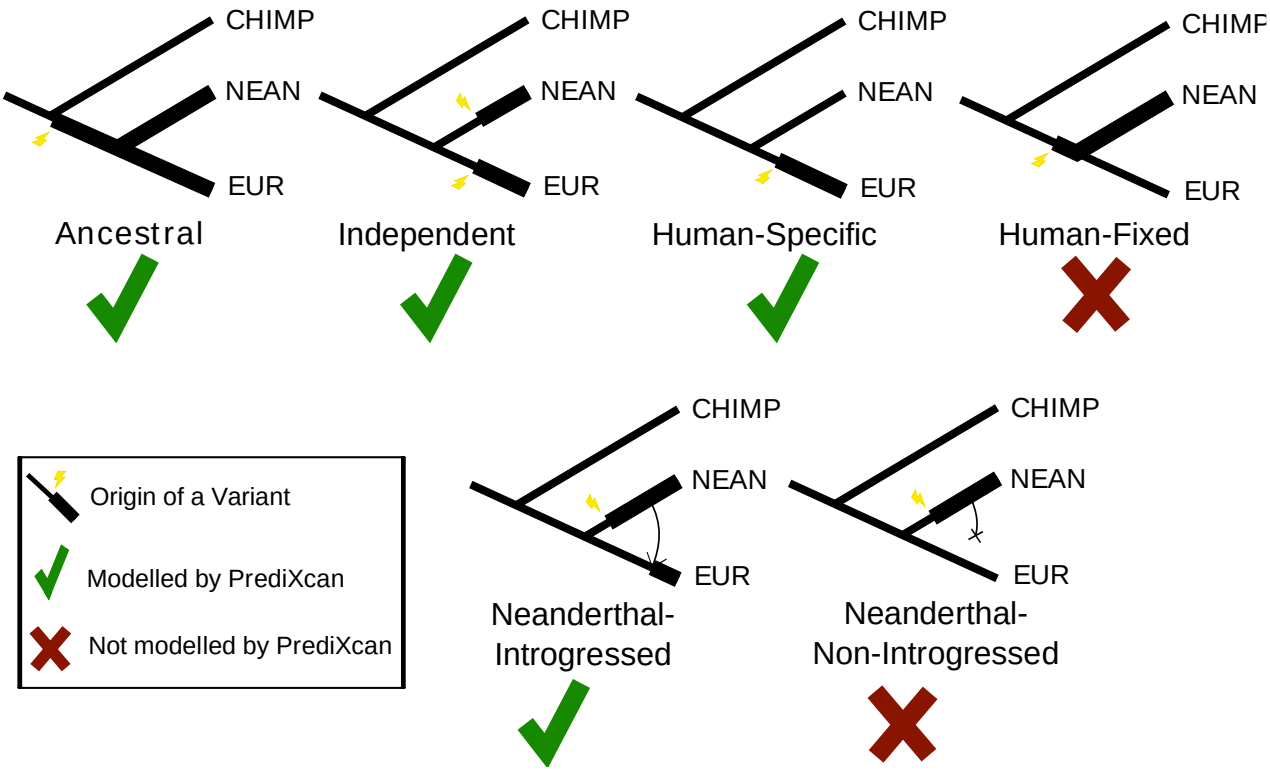
Among the strongest associations of DR GWARRs with phenotypes in BioVU were *MSH5*, *PRSS16*, *VARS*, and *NCR3* with type 1 diabetes (T1D; P = 1.3E–11, 5.2E–8, 7.1E–8, 8.0E–8, respectively, from PrediXcan association tests), *C11orf65* with transient mental disorders (P = 3.1E–9), *SPINT1* with pulmonary embolism and infarction (P = 7.2E–8), and *PSRC1* with hyperlipidemia (P = 3E–8). With the exception of *C11orf65*, each of these genes has evidence of function in pathways relevant to the associated phenotypes. *MSH5* encodes a mutS family protein involved in DNA damage repair and meiotic recombination. DNA damage repair has been linked with various forms of diabetes[20,21] and differential splicing of MSH5 was observed between lean individuals with normoglycemia and overweight individuals with type 2 diabetes[22]. *PRSS16* is a serine protease expressed in the thymus that is involved in T cell maturation, and polymorphism in *PRSS16* has been associated with T1D[23,24]. *VARS* encodes a Valyl-tRNA synthetase and is located in the class III region of the major histocompatibility complex. *NCR3* encodes a cell membrane receptor that activates natural killer (NK) cells in response to extracellular ligands. Genetic variation in *NCR3* has been linked to chronic autoimmune diseases, like Sjögren's syndrome, and *NCR3* has significantly lower expression in NK cells of T1D patients compared to controls[25–27]. *SPINT1* encode a serine protease inhibitor that regulates the activity of hepatocyte growth factor in injured tissue; SPINT1 inhibits hepatocyte growth factor activator (HGFA), which is activated during blood coagulation[28]. *PSRC1* encodes a proline-rich protein that regulates mitotic spindle dynamics; variants in the *CELSR2–PSRC1–SORT1* locus have been associated with lipid traits and coronary artery disease in multiple GWAS studies[29–32]. There is limited knowledge about the *C11orf65* locus. It is possible that it is functional or that regulatory signals influencing other molecules are captured by its prediction models.

Supplementary Figures



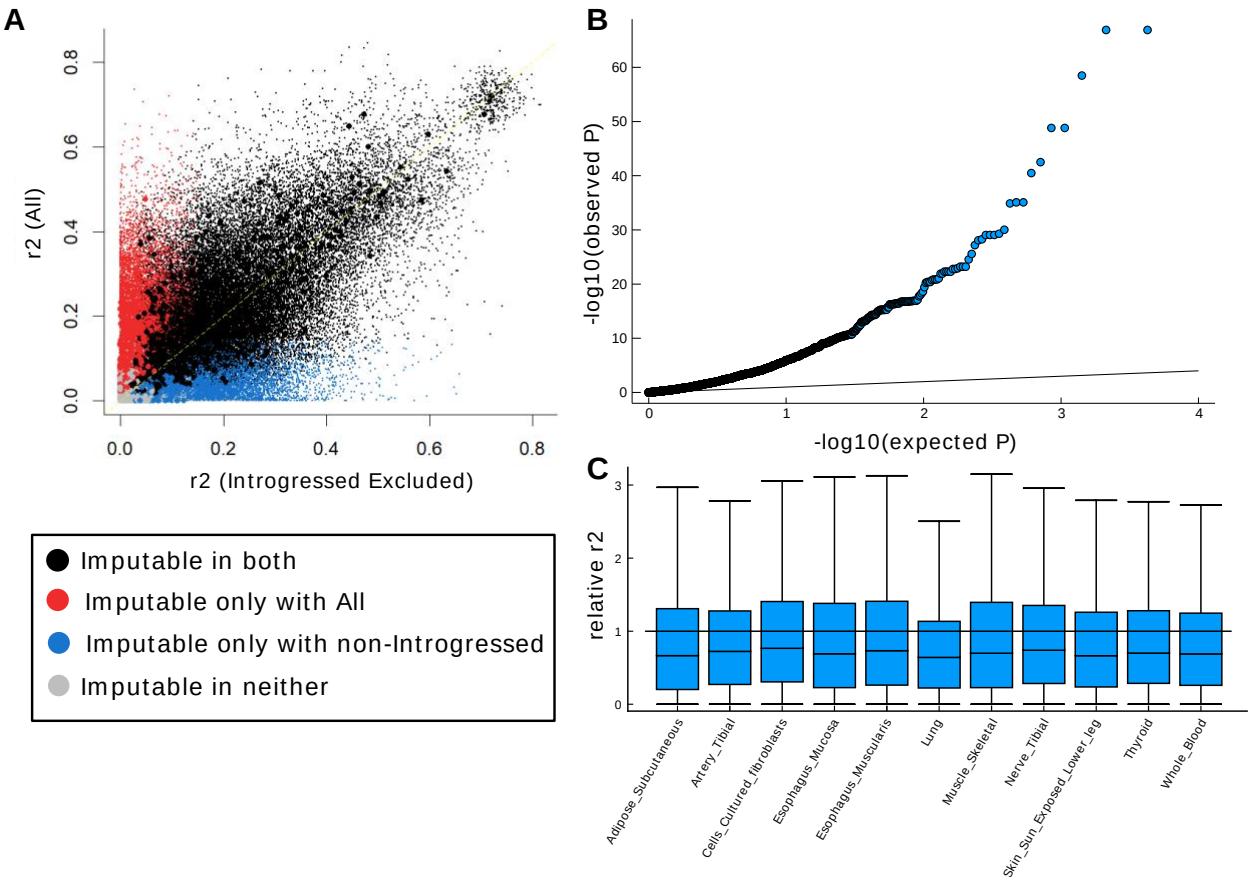
Supplementary Figure 1. PrediXcan models retain predictive power across populations.

QQ plots for applying GTEx LCL PrediXcan models to (A) the 1000 Genomes Europeans (CEU, GBR, FIN, TSI) and (B) YRI. The plots show the observed and expected square of the Spearman correlation between observed and imputed regulation.

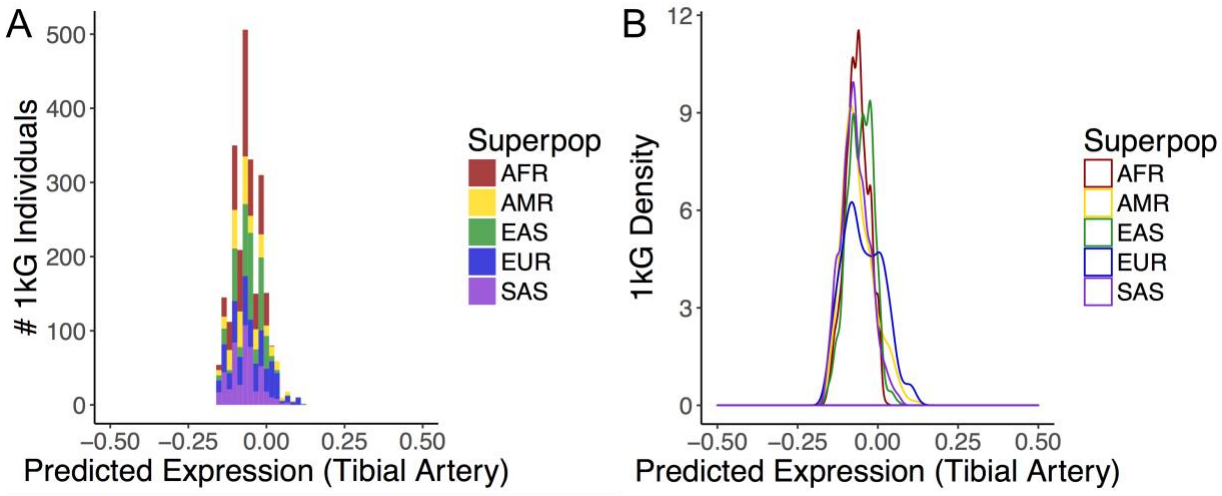


Supplementary Figure 2. PrediXcan considers genomic variants with different evolutionary histories, but does not model Neanderthal-specific variants.

This schematic illustrates the evolutionary histories of variants whose effects on gene regulation are modelled by PrediXcan. These include: variants ancestral to Neanderthals and AMH, variants that occurred on each lineage independently, variants specific to AMH (where Neanderthal retains the ancestral or a different variant), and variants that appeared in the Neanderthal lineage and were introgressed into AMH populations. The PrediXcan approach cannot directly model the effects of Neanderthal-specific variants.

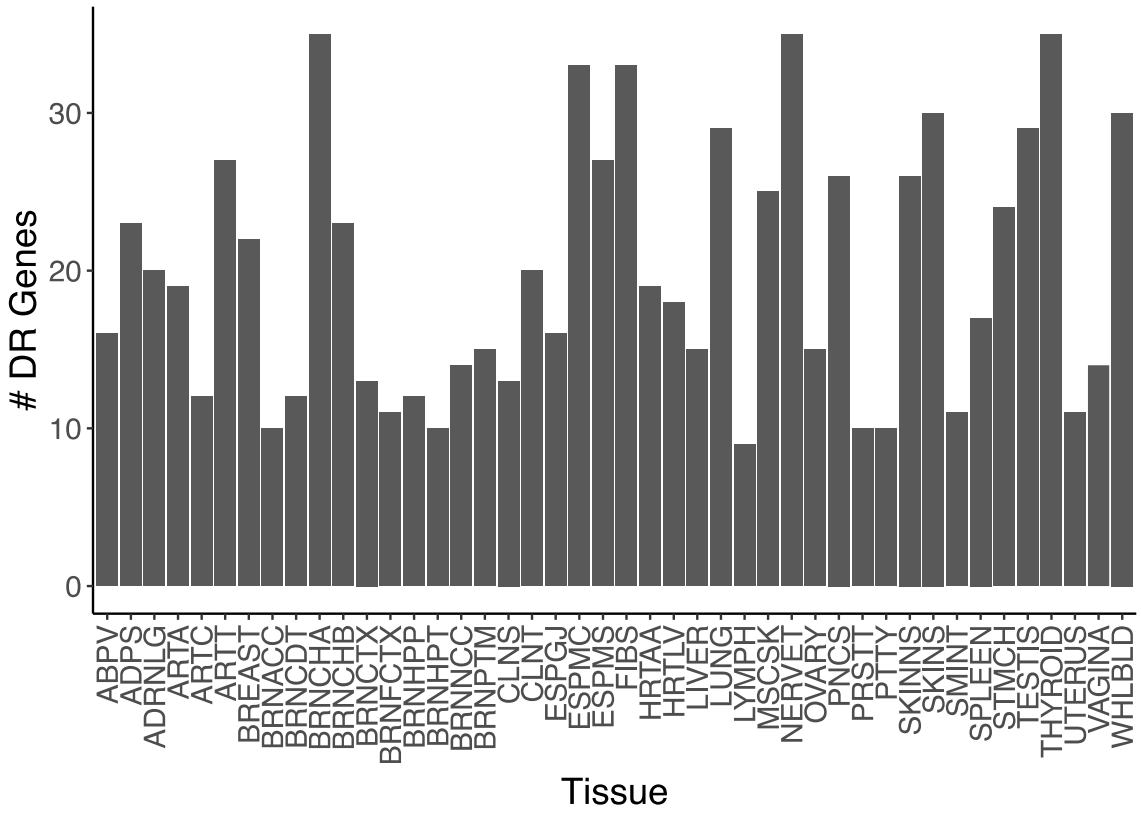


Supplementary Figure 3. PrediXcan models trained without introgressed Neanderthal haplotypes retain significant accuracy on Neanderthal-ancestry regulatory regions. (A) Scatter plot comparing the accuracy of PrediXcan models for non-GWARRs in skeletal muscle when trained and evaluated on sequences of human and Neanderthal ancestry (All) vs. trained on human-ancestry sequences and evaluated on sequences of Neanderthal ancestry (Introgressed Excluded). Accuracy was quantified as the r^2 between observed expression and PrediXcan prediction. Models with $r > 0.1$ and $P < 0.05$ are defined as imputable. Patterns are similar for other tissues. The large dots represent r^2 computed over all individuals in the testing set. To illustrate the variation in the performance, the small dots represent results over 99 resampled sets of 50% of the testing set (Methods). (B) QQ Plot of observed vs. expected –log10(*P*) for the Introgression Excluded skeletal muscle models when applied to individuals with Neanderthal ancestry regulatory regions. (One representative sampling per gene is plotted for simplicity.) (C) Relative r^2 ($r^2_{\text{Introgression-Excluded}}/r^2_{\text{All}}$) for models with $r_{\text{All}} > 0.1$ across a representative set of tissues. Outliers were not plotted. As expected, there is a decrease in accuracy: medians between 0.647 and 0.769. Nonetheless, thousands of models remain significant predictive ability when applied to Neanderthals.

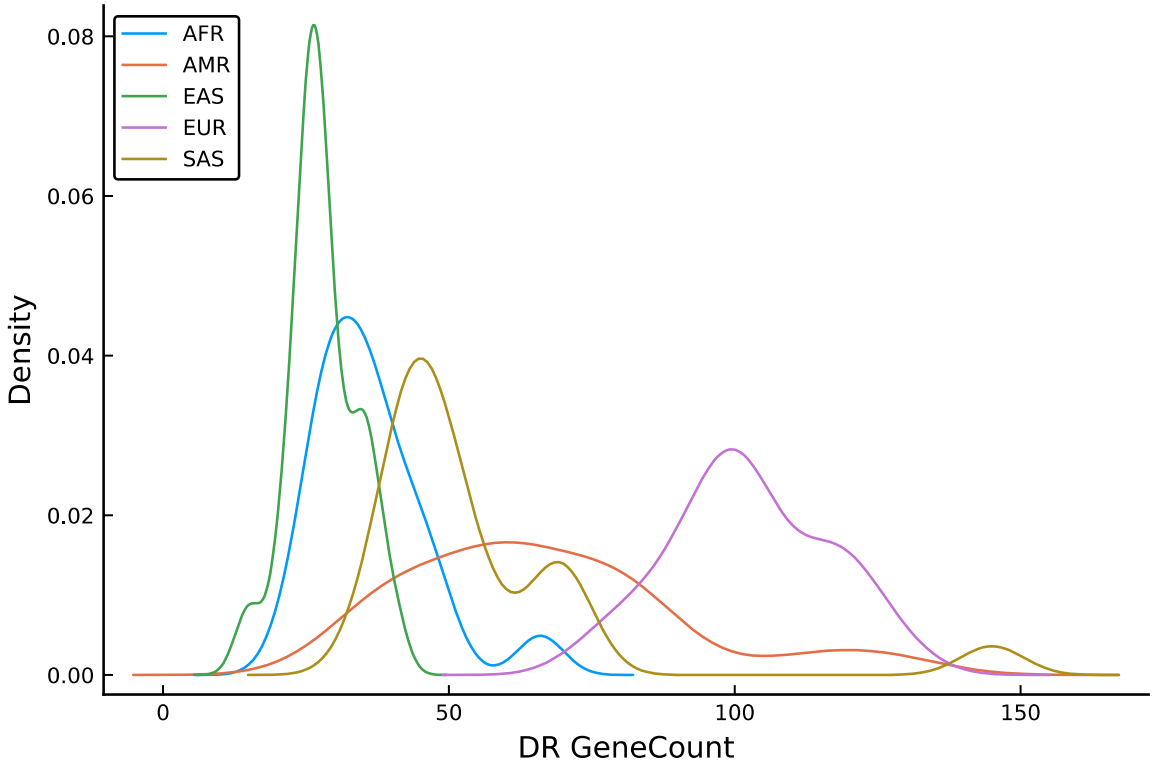


Supplementary Figure 4. *ZDBF2* has similar predicted regulation distributions across AMH populations.

(A) Stacked bar plot showing predicted regulation of *ZDBF2* in the tibial artery, colored by 1kG super-population. (B) Density plot of *ZDBF2* predicted regulation in the tibial artery across 1kG super-populations.

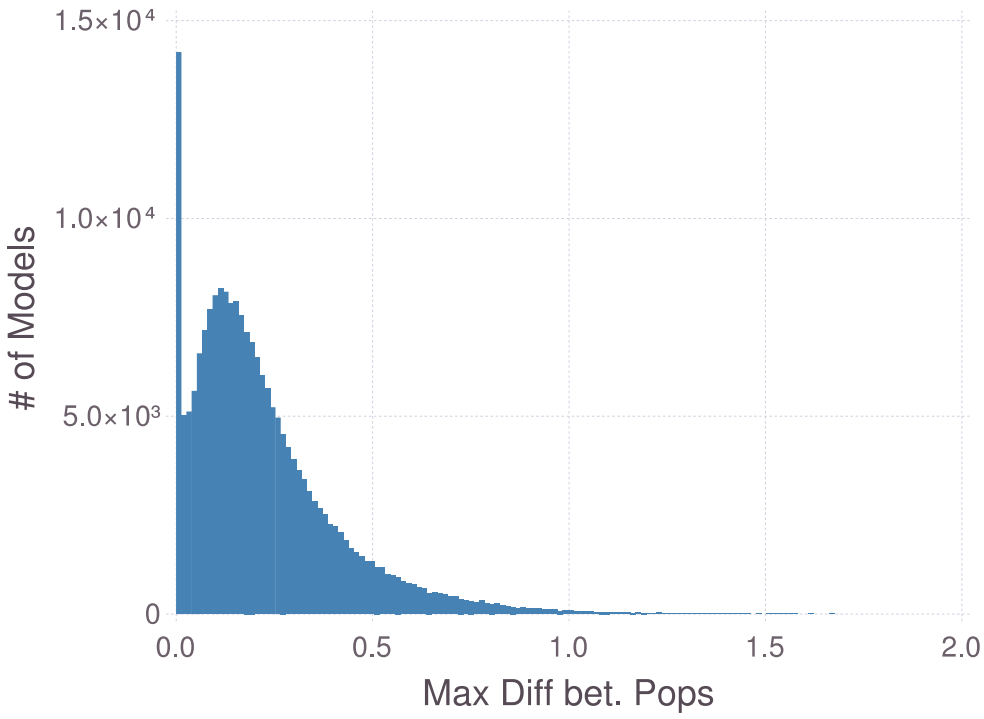


Supplementary Figure 5. The number of DR GWARRs found in each GTEx tissue.
DR GWARRs are found across diverse tissues. See Supplementary Methods for abbreviations. We caution against direct comparisons of the number of DR GWARRs in each tissue due to differences in power resulting from variation in sample size, genetic architecture, and expression levels across tissues. For example, the number of DR GWARRs per tissue is correlated with the GTEx sample size for the tissue (Supplementary Figure 13). We abbreviate the 44 tissues considered as follows throughout the supplement: Adipose - Subcutaneous: ADPS, Adipose - Visceral Omentum: ABPV, Adrenal Gland: ADRNLG, Artery - Aorta: ARTA, Artery - Coronary: ARTC, Artery - Tibial: ARTT, Brain - Anterior Cingulate Cortex: BRNACC, Brain - Caudate: BRNCDT, Brain - Cerebellar Hemisphere: BRNCHB, Brain - Cerebellum: BRNCHA, Brain - Cortex: BRNCTX, Brain - Frontal Cortex: BRNFCTX, Brain - Hippocampus: BRNHPP, Brain - Hypothalamus: BRNHPT, Brain - Nucleus Accumbens basal ganglia: BRNNCC, Brain - putamen basal ganglia: BRNPTM, Breast: BREAST, Cells - Transformed Fibroblasts: FIBS, Colon - Sigmoid: CLNS, Colon - Transverse: CLNT, Esophagus - Gastroesophageal Junction: ESPGJ, Esophagus - Mucosa: ESPMC, Esophagus - Muscularis: ESPMS, Heart - Atrial Appendage: HRTAA, Heart - Left Ventricle: HRTLV, Liver: LIVER, Lung: LUNG, Cells- EBV-transformed Lymphocytes: LYMPH, Ovary: OVARY, Pancreas: PNCS, Pituitary: PTTY, Prostate: PRSTT, Skeletal Muscle: MSCSK, Skin - Not sun-exposed: SKINNS, Skin - Sun-exposed: SKINS, Small Intestine: SMINT, Spleen: SPLEEN, Stomach: STMCH, Testis: TESTIS, Thyroid: THYROID, Tibial Nerve: NERVET, Uterus: UTERUS, Vagina: VAGINA, Whole Blood: WHLBLD.



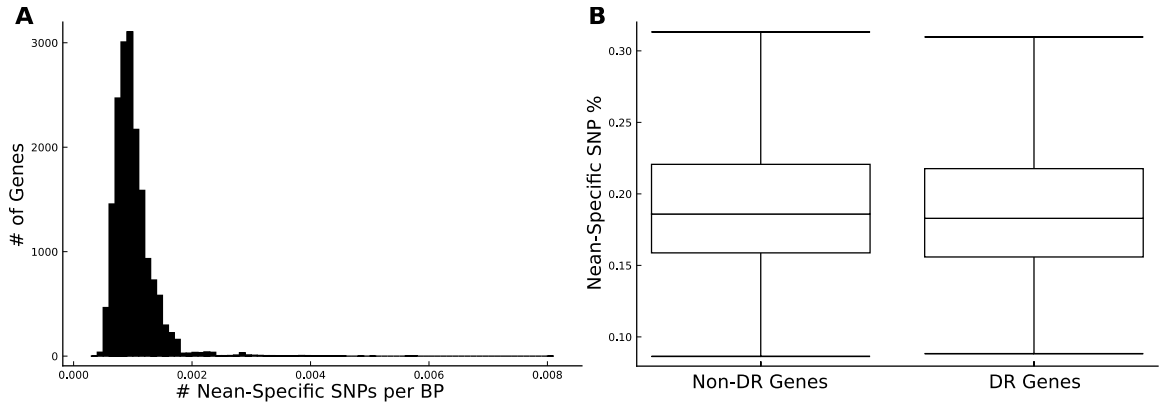
Supplementary Figure 6. Distributions of the number of DR genes found in 50 random humans from 1kG.

For 50 random individuals from the 1kG cohort (10 from each continental population), we counted the number of unique DR genes found across any of the tissues considered. Europeans have the largest number of DR genes. The other individuals with high DR gene counts are from populations with significant amounts of admixture with Europeans (AMR; PJL and GIH from SAS (N=6); ASW and ACB from AFR (N=2)). This suggests that power to detect DR is greatest in the training population, and that divergence from the training population is unlikely to cause a large number of false positives. The Altai Neanderthal has significantly more DR genes (2325 total; $P < 0.02$) than any modern human, despite its greater evolutionary distance from the training population.

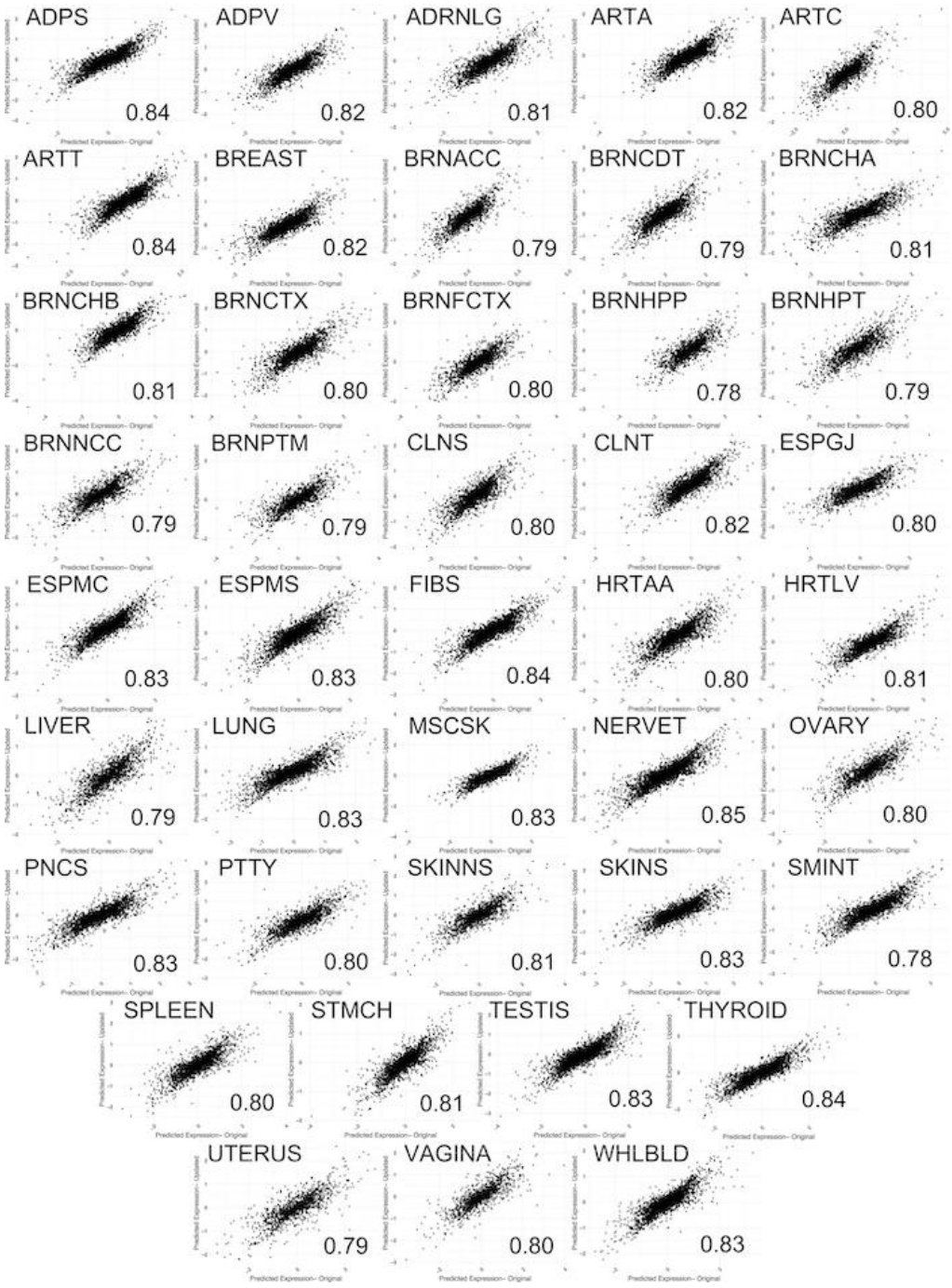


Supplementary Figure 7. Distribution of the maximum difference in the median imputed regulation between 1000 Genomes populations for all PrediXcan models.

Very few models have large predicted regulation differences between populations.

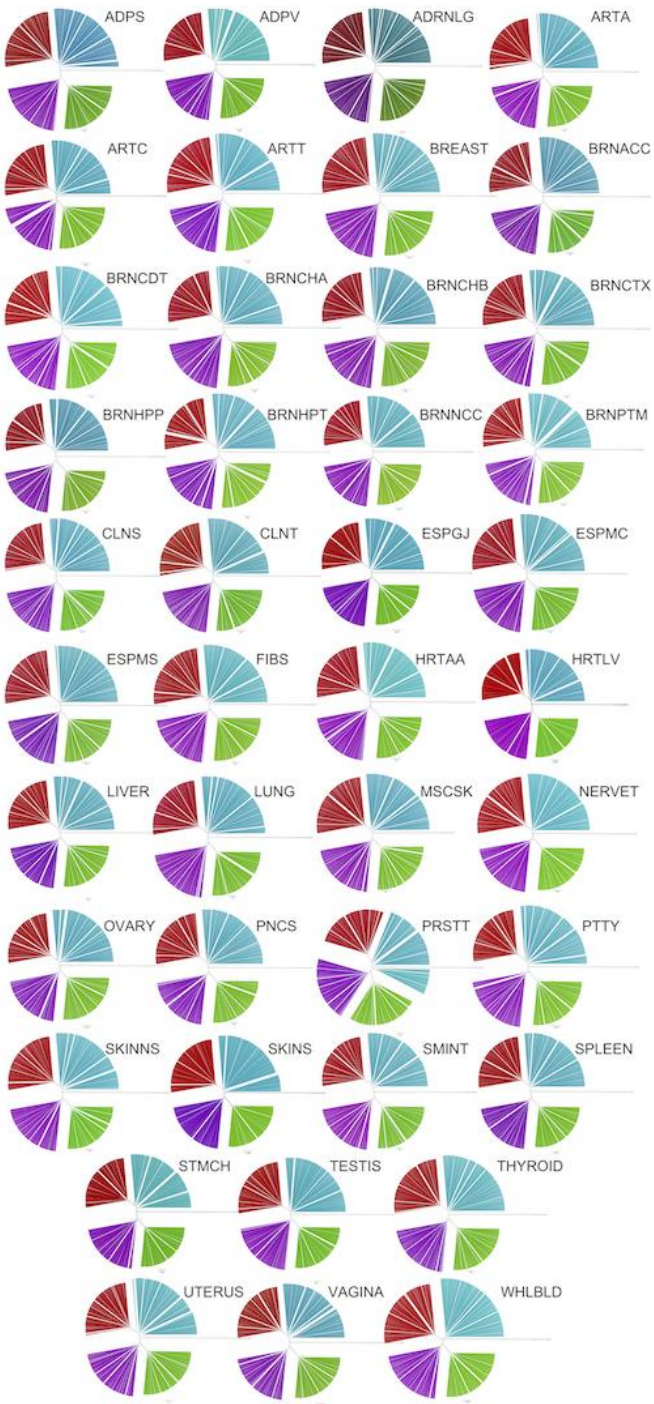


Supplementary Figure 8. Neanderthal-specific variant density in gene regulatory regions. (A) Density of Neanderthal-specific variants in the regulatory regions of genes. (B) The percentage of Neanderthal-specific variants out of all variable sites (observed in humans, Neanderthals, or both) in a gene region is similar for both DR and Non-DR genes: median 0.182 for DR genes, 0.186 for non-DR genes. The difference is significant due to the large number of genes compared (P = 0.0095, MWU Test), but is very small in magnitude. The regulatory region is defined as the gene plus 1 Mb flanking on either side, corresponding to the region considered by PrediXcan.



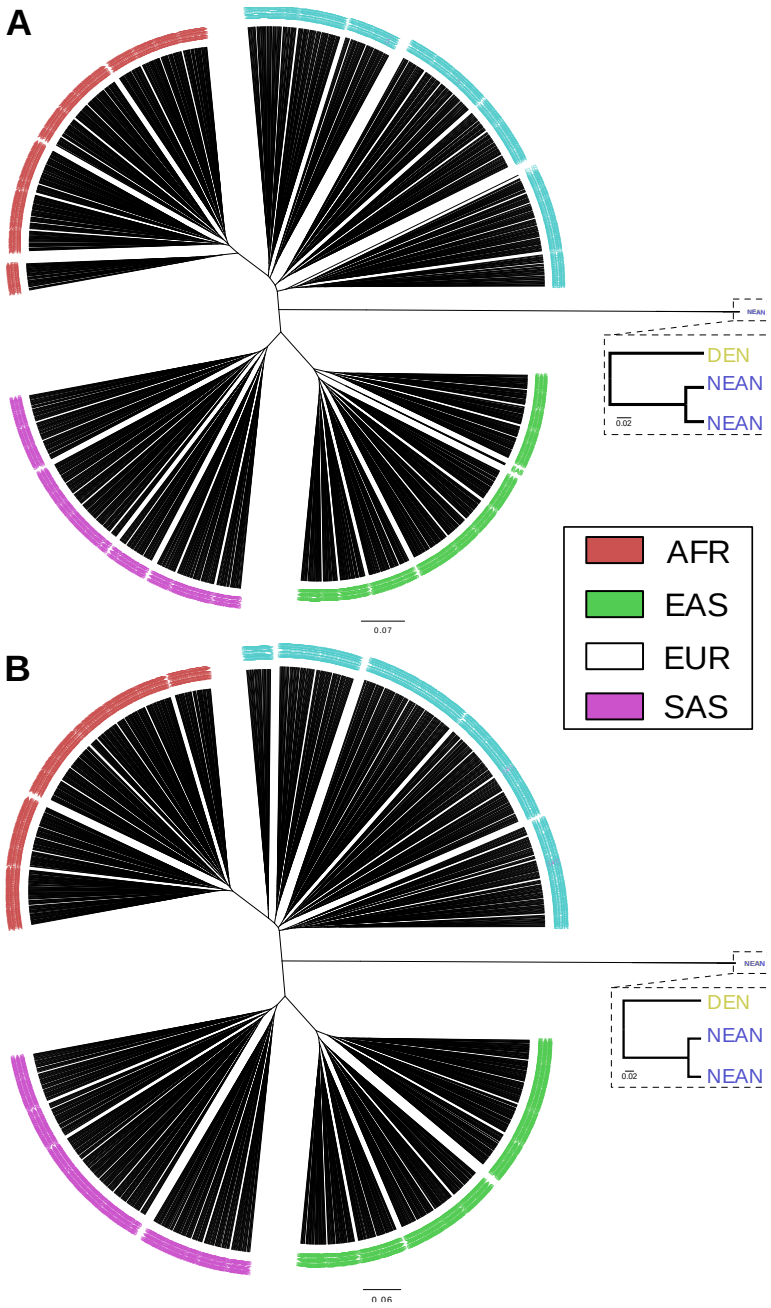
Supplementary Figure 9. Regulation predictions are consistent on the Altai genome.

Scatterplots of predicted regulation of all genes in all tissues based on the recently re-processed Altai genome (y-axis) vs. the original release of the Altai genome (x-axis). We excluded gene models that were missing all SNPs in either version. The mean Spearman correlation across all tissues is 0.81. Individual tissue correlations are in the bottom right of the corresponding plot. Overall, 97% of DR genes called in Altai are also DR in Vindija, with the same predicted level of effect (92% vice versa).

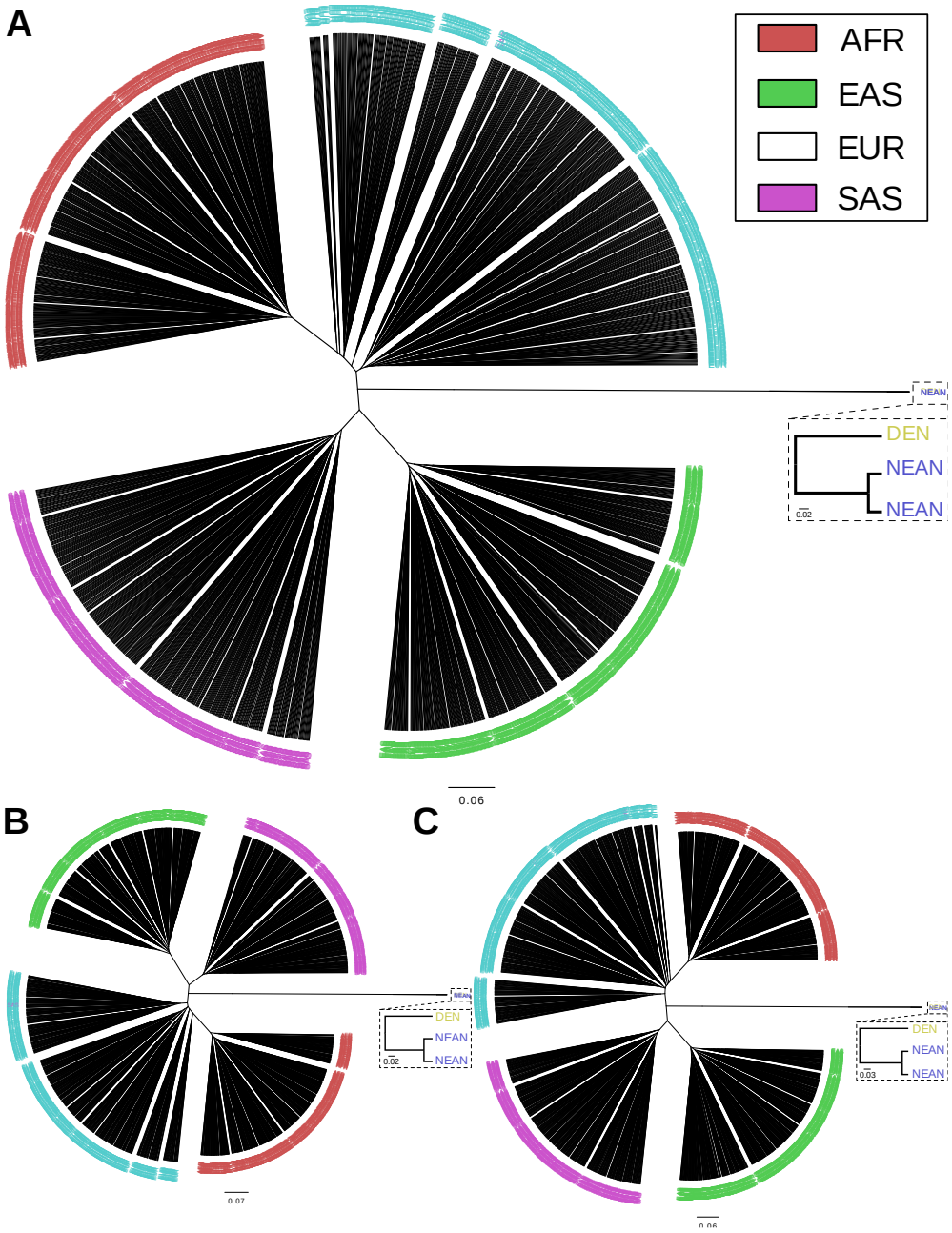


Supplementary Figure 10. Archaic hominins have distinct imputed gene regulatory profiles compared to AMH in all tissues analyzed.

Hierarchical clustering dendrogram by the Pearson correlation of predicted gene regulation for all analyzed genes in all tissues of archaic hominins and non-admixed AMH populations from 1kG. Red=AFR, Blue=EUR, Green=EAS, Purple=SAS.

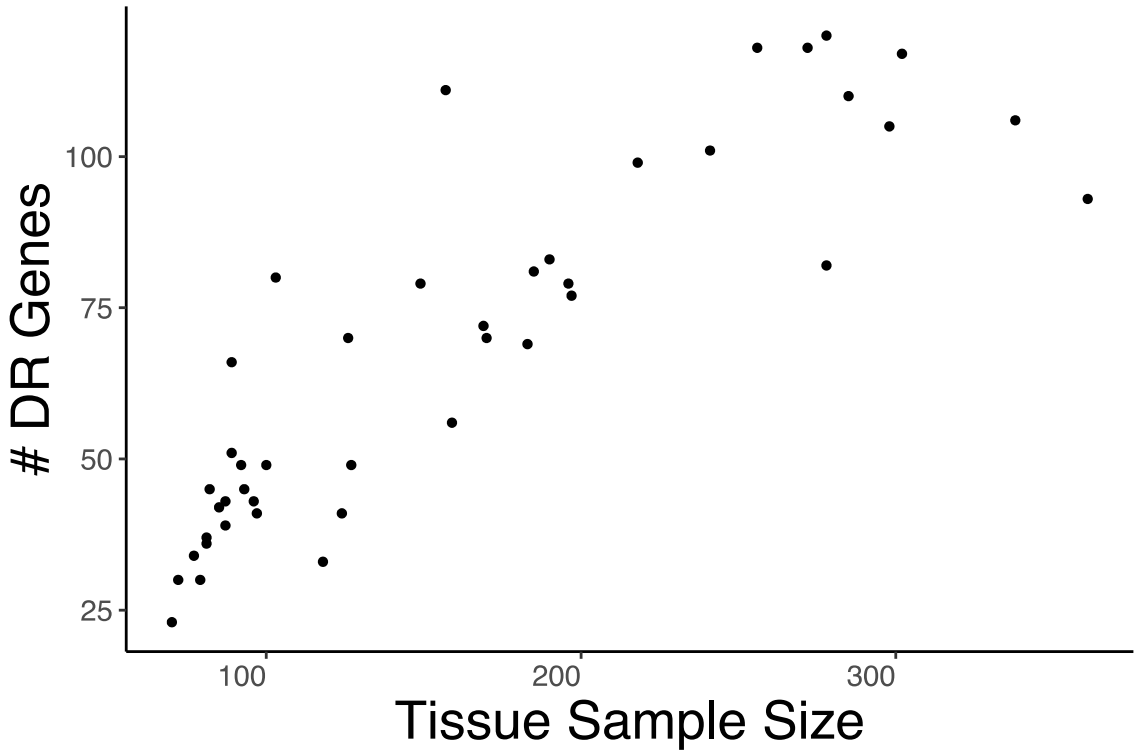


Supplementary Figure 11. Archaic hominins have distinct imputed regulatory profiles for both GWARRs (A) and non-GWARRs (B).
Trees were constructed as in Fig. 5 using PrediXcan values from the Brain-Frontal Cortex models. Other tissues show similar patterns (e.g., as in Supplementary Figure 11).



Supplementary Figure 12. Clustering in imputed regulatory profiles is consistent when using Spearman correlation as the similarity metric.

Trees were generated with hierarchical clustering on imputed regulatory profiles from Brain Frontal Cortex models using Spearman correlation. (A) Tree based on all genes. (B) Tree based on GWARRs. (C) Tree based on non-GWARRs. Patterns were similar for all other tissues.



Supplementary Figure 13. The number of DR genes found in a tissue is correlated with the sample size used in training models for that tissue.

The correlation between sample size and # of DR genes is significant (Spearman's ρ = 0.89; *P* = 1.3E-15). The raw number of DR genes or phenotype associations should not be directly compared across tissues.

Supplementary Table 1. No bias in the direction of divergent regulation among DR GWARRs and tissues.

No tissues were significantly depleted or enriched for Neanderthal upregulation compared to the overall proportion of upregulated genes (0.43). *P*-values were calculated using the binomial test. The Bonferroni-corrected significance threshold is 0.001. See Supplementary Methods for tissue abbreviations.

Tissue	Prop. Up	*P*-value	Tissue	Prop. Up	*P*-value
ADPS	0.43	1.00	ESPMS	0.67	0.02
ABPV	0.56	0.32	HRTAA	0.47	0.82
ADRNLG	0.35	0.51	HRTLV	0.50	0.64
ARTA	0.29	0.33	LIVER	0.60	0.21
ARTC	0.42	1.00	LUNG	0.48	0.71
ARTT	0.37	0.56	LYMPH	0.44	1.00
BRNACC	0.60	0.35	OVARY	0.60	0.21
BRNCDT	0.50	0.77	PNCS	0.50	0.56
BRNCHB	0.52	0.41	PTTY	0.50	0.76
BRNCHA	0.57	0.12	PRSTT	0.44	1.0
BRNCTX	0.38	0.79	MSCSK	0.32	0.31
BRNFCTX	0.36	0.77	SKINNS	0.60	0.11
BRNHPP	0.18	0.13	SKINS	0.47	0.72
BRNHPT	0.50	0.76	SMINT	0.45	1.00
BRNNCC	0.36	0.60	SPLEEN	0.47	0.81
BRNPTM	0.40	1.00	STMCH	0.46	0.84
BREAST	0.59	0.20	TESTIS	0.41	0.85
FIBS	0.52	0.38	THYROID	0.50	0.49
CLNS	0.62	0.26	NERVET	0.63	0.03
CLNT	0.45	1.00	UTERUS	0.45	1.00
ESPGJ	0.38	0.80	VAGINA	0.43	1.00
ESPMC	0.39	0.73	WHLBLD	0.47	0.72

Supplementary Table 2. HPO phenotypes enriched in DR genes common to all archaic hominins.

Enrichments are the ratio of the number of genes in the category observed among DR genes found in all archaic hominins compared to the number expected under no association. *P*-values were calculated using gene set over-representation analysis for phenotypes containing >10 genes in the Human Phenotype Ontology via the hypergeometric test. See Supplementary File S18 for all associations.

HPO ID	Description	# DR Genes	Enrichment	*P*-value
HP:0005736	Short tibia	4	7.15	0.0017
HP:0004691	2-3 toe syndactyly	6	4.15	0.0026
HP:0003330	Abnormal bone structure	31	1.62	0.0034
HP:0001650	Aortic valve stenosis	6	3.78	0.0042
HP:0001007	Hirsutism	10	2.61	0.0042
HP:0006498	Aplasia/Hypoplasia of the patella	5	4.29	0.0051
HP:0002205	Recurrent respiratory infections	26	1.63	0.0071
HP:0001712	Left ventricular hypertrophy	8	2.77	0.0073
HP:0001769	Broad foot	6	3.30	0.0085
HP:0012745	Short palpebral fissure	6	3.22	0.0096

Supplementary Table 3. HPO phenotypes enriched in DR genes specific to Neanderthals.

Enrichments are the ratio of the number of genes in the category observed among the union of DR genes in the Altai and Vindija Neanderthals that are not DR in the Denisovan compared to the number expected under no association. *P*-values were calculated using gene set over-representation analysis for phenotypes containing >10 genes in the Human Phenotype Ontology via the hypergeometric test. See Supplementary File S18 for all associations.

HPO ID	Description	# DR Genes	Enrichment	*P*-value
HP:0011065	Conical incisor	4	9.53	0.0006
HP:0006342	Peg-shaped maxillary lateral incisors	3	11.43	0.0018
HP:0011063	Abnormality of incisor morphology	4	6.93	0.0022
HP:0011792	Neoplasm by histology	12	2.50	0.0025
HP:0000698	Conical tooth	4	6.35	0.0031
HP:0000557	Buphthalmos	4	6.10	0.0037
HP:0000676	Abnormality of the incisor	5	4.54	0.0044
HP:0001019	Erythroderma	4	5.44	0.0056
HP:0001000	Abnormality of skin pigmentation	16	1.96	0.0058
HP:0001519	Disproportionate tall stature	3	7.62	0.0063

Supplementary Table 4. HPO phenotypes enriched in DR genes specific to the Denisovan. Enrichments are the ratio of the number of genes in the category observed among Denisovan-specific DR genes compared to the number expected under no association. *P*-values were calculated using gene set over-representation analysis for phenotypes containing >10 genes in the Human Phenotype Ontology via the hypergeometric test. See Supplementary File S18 for all associations.

HPO ID	Description	# DR Genes	Enrichment	*P*-value
HP:0100710	Implusivity	3	7.74	0.0063
HP:0001302	Pachygyria	6	3.39	0.0077
HP:0001611	Nasal Speech	3	6.36	0.0110
HP:0100803	Abnormality of the periungual region	2	11.87	0.0115
HP:0000954	Single transverse palmar crease	5	3.37	0.0152
HP:0000829	Hypoparathyroidism	2	9.89	0.0165
HP:0001339	Lissencephaly	4	3.96	0.0174
HP:0001805	Thick nail	2	9.13	0.0193
HP:0200039	Pustule	2	9.13	0.0193
HP:0011061	Abnormality of dental structure	7	2.50	0.0198

References

1. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
2. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
3. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol.* **19**, 179 (2018).
4. Mostafavi, H., Harpak, A., Conley, D., Pritchard, J. K. & Przeworski, M. Variable prediction accuracy of polygenic scores within an ancestry group. *bioRxiv* (2019). doi:10.1101/629949
5. Berens, A. J., Cooper, T. L. & Lachance, J. The genomic health of ancient hominins. *Hum. Biol.* **89**, 7–19 (2017).
6. Martin, A. R. *et al.* Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. *PLoS Genet.* **10**, 1004549 (2014).
7. Kelly, D. E., Hansen, M. E. B. & Tishkoff, S. A. Global variation in gene expression and the value of diverse sampling. *Curr. Opin. Syst. Biol.* **1**, 102–108 (2017).
8. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
9. Mikhaylova, A. V & Thornton, T. A. Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. *Front. Genet.* **10**, 261 (2019).
10. Petty, L. E. *et al.* Functionally oriented analysis of cardiometabolic traits in a trans-ethnic sample. *Hum. Mol. Genet.* **28**, 1212–1224 (2019).
11. Alm, P. A. Stuttering and the basal ganglia circuits: a critical review of possible relations. *J. Commun. Disord.* **37**, 325–369 (2004).
12. Carlos, F. & Dennis, D. Genetic contributions to stuttering: the current evidence. *Mol. Genet. Genomic Med.* **5**, 95–102 (2017).
13. Ko, C. H. & Takahashi, J. S. Molecular components of the mammalian circadian clock. *Hum. Mol. Genet.* **15**, R271–R277 (2006).
14. Dannemann, M. & Kelso, J. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *Am. J. Hum. Genet.* **101**, 578–589 (2017).
15. Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737–741 (2016).
16. Wada, H., Tanaka, H., Nakayama, S., Iwasaki, M. & Okamoto, H. Frizzled3a and Celsr2 function in the neuroepithelium to regulate migration of facial motor neurons in the developing zebrafish hindbrain. *Development* **133**, 4749 LP – 4759 (2006).
17. Skibinski, G. *et al.* Mutations in the endosomal ESCRTIII-complex subunit CHMP2B in frontotemporal dementia. *Nat. Genet.* **37**, 806 (2005).
18. Cox, L. E. *et al.* Mutations in CHMP2B in Lower Motor Neuron Predominant Amyotrophic Lateral Sclerosis (ALS). *PLoS One* **5**, e9872 (2010).
19. McDowell, K. A. *et al.* Reduced cortical BDNF expression and aberrant memory in Carf knock-out mice. *J. Neurosci.* **30**, 7453–7465 (2010).
20. Blasiak, J. *et al.* DNA damage and repair in type 2 diabetes mellitus. *Mutat. Res. Mol. Mech. Mutagen.* **554**, 297–304 (2004).
21. Ye, C. *et al.* Diabetes causes multiple genetic alterations and downregulates expression of DNA repair genes in the prostate. *Lab. Investig.* **91**, 1363 (2011).

22. Kaminska, D. *et al.* Regulation of alternative splicing in human obesity loci. *Obesity* **24**, 2033–2037 (2016).
23. Viken, M. K. *et al.* Reproducible association with type 1 diabetes in the extended class I region of the major histocompatibility complex. *Genes Immun.* **10**, 323 (2009).
24. Viret, C. *et al.* Thymus-specific serine protease controls autoreactive CD4 T cell development and autoimmune diabetes in mice. *J. Clin. Invest.* **121**, 1810–1821 (2011).
25. Rodacki, M. *et al.* Altered natural killer cells in type 1 diabetic patients. *Diabetes* **56**, 177–185 (2007).
26. Rusakiewicz, S. *et al.* NCR3/NKp30 contributes to pathogenesis in primary Sjögren's syndrome. *Sci. Transl. Med.* **5**, 195ra96--195ra96 (2013).
27. Pende, D. *et al.* Identification and Molecular Characterization of Nkp30, a Novel Triggering Receptor Involved in Natural Cytotoxicity Mediated by Human Natural Killer Cells. *J. Exp. Med.* **190**, 1505–1516 (1999).
28. Eigenbrot, C., Ganesan, R. & Kirchhofer, D. Hepatocyte growth factor activator (HGFA): molecular structure and interactions with HGFA inhibitor-1 (HAI-1). *FEBS J.* **277**, 2215–2222 (2010).
29. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).
30. Davis, J. P. *et al.* Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in Finnish men from the METSIM study. *PLoS Genet.* **13**, e1007079 (2017).
31. Arvind, P., Nair, J., Jambunathan, S., Kakkar, V. V & Shanker, J. CELSR2–PSRC1–SORT1 gene expression and association with coronary artery disease and plasma lipid levels in an Asian Indian cohort. *J. Cardiol.* **64**, 339–346 (2014).
32. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714 (2010).