Predicting gene regulatory changes across human evolution using ancient DNA

By

Laura L Colbran

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

August 7, 2020

Nashville, Tennessee

Approved:

John A. Capra, PhD

Nancy J. Cox, PhD

Melinda C. Aldrich, PhD, MPH

Nicole Creanza, PhD

Emily Hodges, PhD

David C. Samuels, PhD

ACKNOWLEDGMENTS

That old saying about it taking a village to raise a child is even more true of a dissertation. My advisor, Tony Capra, is almost single-handedly responsible for the direction my career took; you could say that summer undergraduate internship was a formative experience! His support and encouragement, as well as his ability to think of the way out of boxes I've backed myself into, has been invaluable. I've also had some truly wonderful collaborators, Eric Gamazon and Dan Zhou, who really went above and beyond with the months of debugging help, among other things. I'd also like to thank everyone on my committee, who offered wisdom and suggestions while always being encouraging and positive, even when I felt like I was spinning my wheels. Seriously, not everyone can say they consistently left committee meetings feeling upbeat about their work!

Everyone in the Capra lab, both past and present, has been instrumental to my success as a student. Whether it was giving excruciatingly detailed feedback on figures and presentations, or spending hours rolling dice to navigate my made-up worlds, they've been amazing. I would especially like to thank my bay-buddy and lab twin, Mary Lauren Benton, for being a wonderful brainstorming help, occasional rubber duck stand-in and vetter of awkward emails, as well as an unfailing partner-in-mischief. All my non-Vanderbilt friends and family were excellent keepers of my sanity throughout, as well as occasional practice audiences. Thank you.

Lastly, this work would not have been possible without financial support from the Human Genetics NIH grant T32GM080178, and several Vanderbilt-based resources. Specifically, this work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University (at least, when it wasn't literally on fire!), and was based in part on data from the PredixVU system of Vanderbilt University Medical Center.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Human Evolutionary history

It has been said that humans are made by history [1], and this is particularly true in terms of evolutionary and genetic history. The human genome has been shaped by billions of years of changing environments and circumstances that today influence everything from the number of limbs humans walk on [2,3] to their susceptibility to certain diseases [4,5]. More fundamentally, exposures to certain retroviruses influenced the mechanisms by which genes are regulated [6,7], and every eukaryotic cell's source of energy was determined by a lucky encounter between a prokaryote and an archaeon [8]. Understanding ancient history and how it impacted humans at the genomic and phenotypic level is therefore fundamental to understanding how our bodies work and why diseases happen, as well as how future changes might affect us.

A large portion of what we know about human origins is based on the fossil record. While humans and chimpanzees and bonobos diverged 5 to 12 million years ago [9,10], there was a long, tangled web of history between their most recent common ancestor and the origin of anatomically modern humans ("AMH"; *Homo sapiens*). There are dozens of archaic hominins such as *Sahelanthropus* and *Australopithecus* that lived 2 to 7 million years ago, primarily in Africa [11,12]. More recent hominins such as Denisovans and Neanderthals appear around 600,000 and 200,000 years ago, respectively, and were present throughout Eurasia [13]. These different species often overlapped in timespan, and it is difficult to exactly resolve which are direct ancestors of AMHs with only the fossil record [14].

The advent of widespread genome sequencing, as well as the ability to isolate and sequence DNA from ancient bones, has increased our ability to resolve questions about more recent relationships by combining knowledge gleaned from the fossil record with analysis of genetic similarities. Some of the oldest AMH skulls were dated to 196,000 and 160,000 years ago and were both found

in East Africa[15,16]. Genetic studies using mitochondrial and Y-chromosome DNA from modern populations also suggest that the most recent common ancestor (MRCA) on the maternal and paternal lineages for modern populations was between 150,000 and 200,000 years ago in Africa[17,18,19]. More recent studies have suggested that AMH origins may have been more widely spread around Africa and as much as 100,000 years earlier[20,21].

Large-scale population movements are a recurring theme in human history, and the most significant one may have been the "Out-of-Africa" migration that was the origin of all non-African populations, which occurred 60,000 to 130,0000 years ago[22]. While the presence of AMH bones that predate that time period in the Levant and Europe indicate that that was not the first time AMHs left Africa[23], analyses of mutational rate and mitochondrial haplotypes indicate that the split-time of non-African populations was in that later time range[24,25,26,27]. This suggests that the earlier migrations contributed very little, if any, ancestry to modern populations.

Given that Neanderthals were present in Eurasia as recently as 40,000-50,000 years ago[28,29], to what extent they interacted with AMHs, if at all, has been a longstanding question[30]. While Neanderthals and AMHs are closely related, whole genome analysis done on a draft Neanderthal genome revealed that Neanderthals are more genetically similar to non-Africans than to Africans, which would not be expected if their genetic similarity was only due to common ancestry[31]. In addition, Neanderthal-like haplotypes are longer than would be expected if they had been present in AMH populations since the MRCA of Neanderthals and AMHs[32]. This suggested that it was likely that Neanderthals and the ancestors of modern Eurasians had interbred and some Neanderthal genes had introgressed into AMH genomes. Subsequent studies based on haplotype structure and sequence identity have shown that 30-48% of the Neanderthal genome is present in modern Eurasians, and that 1-2% of an average Eurasian genome is of Neanderthal ancestry[33,34,35].

The existence of the Denisovans was discovered much more recently using genetic analysis, and it has since been shown that they, too, introgressed with the ancestors of East and South Asian populations[33,36], as well as with Neanderthals[37]. Neanderthal introgression likely took place 50,000-60,000 years ago, while Denisovan introgression is more recent[22]. In both cases,

introgression likely occurred multiple times, resulting in varying amounts of archaic ancestry in different AMH populations[38,39]. Because so little physical evidence has been found of the Denisovans, they represent a particularly exciting opportunity for ancient DNA to provide new insights. While much can be learned from genome sequences about relationships between individuals, in most cases it is difficult to interpret what observed differences mean for the organism at a broader scale. Therefore, using aDNA to study physical characteristics of groups like Denisovans requires the development of methods to interpret phenotypes based only on genotype.

Much of the cultural change that occurred after the primary Out-of-Africa migration has been inferred from archaeological research on artifacts and building remains at ancient sites. While AMHs started out as hunter-gatherers, humans began domesticating plants and animals (starting with the wolf, at least 15,000 years ago)[40], and roughly 10,000 years ago these practices rapidly spread throughout the ancient world ("The Neolithic Revolution")[41]. There were two primarily lifestyles that arose as part of this. The first was nomadic pastoralism, where people managed herds of domesticated animals, moving around frequently to graze them. The second type is agriculture, which is characterized by settled communities growing domesticated plants as well as animals. Archaeologically, both can be identified by the presence of those domesticated animals, and evidence such as traces of milk in pottery, or tools one associated with domestication (for example, bit-wear on the teeth of ancient horses)[42,43]. However, agricultural sites have additional evidence of the presence of seeds and evidence of plant cultivation, as well as more permanent settlements[44,45]. However, before the availability of genetic information, it was unknown whether the spread of these different cultures were due to cultural transmission along trade routes, or to the movement of people and population turnover.

Many of the ancient DNA (aDNA) studies published in recent years have been focused on distinguishing between cultural transmission and population replacement. Because of the ready availability of ancient samples of good enough quality for genome analysis, the most detailed studies have been done on populations from Europe and Central Asia[46]. The genetic studies show that the history of the AMHs in the last tens of thousands of years is characterized by repeated

population replacement and migration[22]. For example, while some ancestry in modern Europeans can be traced to Mesolithic hunter-gatherers in the region 30,000 years ago, more of their ancestry is traceable to farmers from Anatolia who migrated into Europe around 8,000 years ago, while the largest proportion is from more recent pastoralists who came into Europe from the Asian Steppe during the Bronze Age[47,48,49]. aDNA studies have also identified populations that existed in the past without contributing substantial ancestry to later populations in the same location[50,51]. A similar story of changes in culture being accompanied by changes in genetic ancestry is true in many other places as well[52,53,54,55,56,57].

1.2 Deriving phenotypes from DNA sequence

While current aDNA studies have deepened our understanding of AMH population changes and migrations through recent history, they have not been informative about phenotypic changes that may have occurred in response to those migrations and cultural changes. Deducing phenotype based on DNA sequence is complicated; most traits, especially those of interest in recent human evolution, involve multiple genes and often are influenced by environment[58]. Model organism studies and analyses in human cells can identify the causal genetics of traits affected by a few genes, particularly if the genetic variants involve protein-coding changes in the genes. For example, missense alleles in two genes responsible for the variability in skin pigmentation in AMH populations (*SLC24A5* and *SLC45A2*) were originally identified in fish[59,60]. Using aDNA it is possible to study where and when these alleles appeared and rose in frequency[61,51], but examples like these are rare, particularly for complex traits. An attractive alternate method of studying changes in phenotypes is to use evolutionary metrics to identify interesting regions of the genome before identifying phenotypes potentially impacted by those alleles, thereby prioritizing regions of interest to target with powerful, low-throughput experimental techniques.

One area of particular focus in recent years has been the regions of the genome in modern populations with Neanderthal ancestry. Electronic health records ("EHRs") and other large collections of paired genotype and phenotype information allow Phenotype-wide Association Studies

(PheWAS) in which each introgressed haplotype is tested for higher frequency in individuals with a certain phenotype compared to controls, repeated over many phenotypes. Studies like these have identified Neanderthal introgressed haplotypes associated with a variety of neurological and skin phenotypes[62,63], as well as brain structure[64]. These have proved informative about potential differences between AMHs and Neanderthals, as well as the effects of those alleles on AMH disease. However, in isolation these methods do not provide a mechanistic explanation for observed correlations, and it can be ambiguous whether the Neanderthal-specific variants or linked ancestral alleles are causing the association[65]. An alternative is to identify introgressed alleles associated with gene expression and protein-coding changes, rather than with a wider phenotype. This has the advantage of tying hypotheses to specific genes, although interpreting findings then requires knowing the function of the gene. Introgressed alleles are more likely to have effects through gene regulation than through protein sequence changes[66], and are likely to have tissue-specific effects. For example, gene transcripts carrying Neanderthal alleles are specifically downregulated in the brain and testis[67].

These studies have several limitations in their informativeness about important phenotypic differences between Neanderthals and AMHs. The first is that they are limited to only the percentage of the Neanderthal genome that is present in modern populations, thereby missing at least half of it. This also means that everything is studied in the context of the archaic sequences' behaviour when combined with the human genome, not necessarily what the phenotypic consequences would be in the context of the Neanderthal genome. The second limitation is that these differences were not necessarily subject to strong selection. Indeed, introgressed segments of the genome were generally selected against (particularly in regulatory regions), implying that many of those that survived were unlikely to have large effects on fitness[68].

Instead, some studies start by identifying regions of the genome that appear to have been subject to strong selection, then identifying genes to explain the peaks. For example, a haplotype introgressed from Denisovans around *EPAS1* bears signs of very strong selection in Tibetans, and has been implicated in adaptation to high altitudes[69]. A genome-wide test for selection based on

identifying alleles with frequencies that are inconsistent with the alleles' frequencies in ancestral populations identifies both skin pigmentation alleles discussed previously, as well as variants near genes associated with diet and immunity[61]. A regulatory variant for *LCT* has undergone strong recent selection in Europeans, putatively in response to the advent of dairy farming[70]. However, integrating information from aDNA has suggested that this allele didn't rise in frequency until long after that time, casting doubt on the explanation for that selection[22,61]. Similarly, a selected haplotype that regulates *FADS1*, a gene involved in metabolizing nutrients from grains and associated with metabolic disorders in modern populations, remained at low frequencies for a long time after the advent of farming, suggesting that this transition was not itself the cause of the selection[71]. However, as with PheWAS studies, links between alleles and organism-level phenotypes with ready mechanistic explanations are rare, and those explanations have to be elucidated on a case-by-case basis.

Given that most recent phenotypic changes are due to changes in gene regulation[72,73], it is tempting to use our knowledge of the biology behind gene regulation to understand the mechanisms behind phenotype associations. The vast majority of genomic variants associated with disease are non-coding, and many of them overlap intergenic regulatory elements like enhancers[74]. While there are many methods to identify enhancers, identifying causal variants is complicated by our limited understanding of how those methods relate to one another[75], and by the highly combinatorial nature of enhancer function[76]. In addition, because enhancers can be as much as a megabase distant from their target genes, pairing them with the genes they regulate is complicated[77]. While it is possible to identify methylation patterns in high-coverage ancient genomes[78], ancient genomes are generally too degraded for widespread identification of enhancers in them, which limits our ability to understand gene regulation in Neanderthals and Denisovans.

In summary, existing methods of studying the recent evolution of human phenotypes are limited by our lack of ability to study combinations of genetic variants and by the complexity of many phenotypes that could of interest. Together, these challenges mean that it is often difficult to connect genetic variation to evolutionarily relevant phenotypic variation, and to identify the mechanisms

by which genetics influences the phenotype.

1.3 Overview of this dissertation

In this dissertation, I will describe analyses that address the gaps in knowledge described above. Specifically, we wanted to be able to provide mechanism-based explanations for potentially evolutionarily important regions of the genome on a large scale. Our approach centers around the idea of using combinations of genetic variants to predict patterns of gene regulation in order to identify genes and pathways with evidence of changes in gene expression during recent evolution. To do this, we used a statistical method called PrediXcan, which is trained to predict RNA-seq data based on allele counts of nearby genetic variants[79].

In Chapter 2, we applied PrediXcan models to the genomes of three archaic hominins: two Neanderthals and one Denisovan. In order to identify genes that had significant differences in regulation in these hominins, we compared the archaic predictions to predictions made on a large, diverse set of AMHs. We first focused on genes in regions of AMH genomes that contain no observed Neanderthal ancestry. While in many cases this may be due to genetic drift, some are likely to have been places where variation in gene regulation was selected against. We identified hundreds of these genes for which Neanderthals were predicted to have significantly divergent patterns of gene regulation compared to modern humans. Based on gene annotations, the affected genes influence a range of traits that could be plausibly affected by selection, including reproduction, skeletal development, language, and the immune system. We also used an EHR-linked biobank to establish that divergent regulation of these genes is associated with many clinical phenotypes in modern populations. This suggested to us that it was possible that divergent regulation of some genes might have been a barrier to archaic introgression.

In addition, in order to identify phenotypes where the archaic hominins may have differed from each other and from AMHs, we compared their predicted gene regulation. We showed Neanderthal-specific regulatory patterns in genes related to skin pigmentation and height, and regulation specific to Denisovans for morphological traits. We also identified an enrichment for

immune-related genes among those that showed divergent regulation between the two Neanderthal individuals. These analyses show that imputing ancient gene regulatory profiles has promise for studying ancient phenotypes, particularly those that cannot be studied directly from the skeletal remains, by prioritizing genes and phenotypes for more detailed analysis.

In the next chapter, we explored the possibility of applying PrediXcan to datasets with other challenging characteristics. While the archaic hominin genomes were high-coverage and fairly complete, that is not the case for the vast majority of aDNA samples. This could potentially hamper the extension of this framework into more recent AMH evolution. We therefore evaluated PrediXcan's ability to function when given low-coverage genomes, and dissected the models' behaviour to learn the conditions under which they remained accurate. We did this by simulating genomes with high amounts of missing variants, and by limiting the variants available for training. Our results suggest that PrediXcan can retain utility even when variant data are limited, as long as the models are trained for the specific application and their limitations properly taken into account. In the case of aDNA studies, this involves training on the set of variants most likely to be covered in genotyping of ancient samples, and filtering for models that, based on simulations, retain accuracy in that situation. This is encouraging for the expansion of gene regulatory studies like the ones described here into different, larger datasets, whether they involve aDNA or not.

For Chapter 4, we demonstrated the utility of PrediXcan in this context by applying our methodology to a large-scale analysis of regulatory differences between ancient hunter-gatherer, pastoralist, and farming AMH populations. We found over 5,000 genes that showed evidence for differences in regulation among the three groups. We recapitulated the pattern of regulation observed over time for *FADS1*, wherein agriculturalists and modern populations have higher expression than ancient hunter-gatherers, and also suggested new explanations for previously observed signals of selection, such as that around *LEPR*. Expanding our analyses to examine specific classes of genes, we found that housekeeping genes were enriched among those with significant differences, in part because of the increased power to model them. Overall, genes with significant differences by lifestyle were enriched for metabolic and immune processes, indicating that these

pathways are the most likely to have been affected by altered gene regulation between ancient human groups with different diets and lifestyles.

Overall, the studies described here demonstrate the power of methods that predict intermediate phenotypes, such as gene expression, to study evolution in a function-aware manner. Such genome-wide methods are well-poised to take advantage of the increasing availability of whole-genome data. PrediXcan in particular provides the opportunity to characterize gene regulation across diverse geographical and temporal ranges. These studies contribute to a larger understanding of the genome's response to large-scale environmental changes, as well as the potential impact they have on phenotypes in modern AMHs.

Chapter 2

GENE REGULATORY PATTERNS IN ARCHAIC HOMININS*

2.1 Introduction

Most aspects of archaic hominin biology cannot be directly studied due to their lack of preservation in fossils. The sequencing of DNA extracted from remains of extinct hominins has enabled the study of these groups' origins and evolutionary histories on a scale not possible from fossils alone[31,36,80,81]. However, even with whole genome sequences available, the ability to infer traits of these hominins and how they differed from one another and anatomically modern humans (AMHs) is limited[82]. Greater morphological knowledge would be especially valuable for groups like the Denisovans that lack a substantial fossil record[36,83]. A key challenge in this task is the difficulty of mapping from genetic sequence differences to function. Archaic hominins interbred with anatomically modern humans (AMHs)[31,36,84], and as a result, more than one third of the Neanderthal genome remains in introgressed sequences in AMH genomes[33,34]. However, the factors that determined the patterns of Neanderthal ancestry in AMH genomes are not fully understood. The Neanderthal DNA that remains in modern Eurasian populations influences a range of traits, with a particular influence on immune, hair and skin, and neurological phenotypes[62]. This suggests differences between Neanderthals and AMHs that could have been selected for after interbreeding. There are only a small number of protein-coding differences between archaic hominins and modern humans[85], but introgressed archaic sequences often exert their effects by modifying gene expression patterns[62,66]. One quarter of Neanderthal sequences remaining in AMHs have cis-regulatory effects, and gene transcripts carrying Neanderthal alleles are particularly downregulated in the brain and testes[67]. Thus, divergent gene regulation between archaic and AMH sequences produces physiologically relevant effects. While the functional effects of introgressed sequences have been studied in detail, much less is known about the functions of non-introgressed Nean-

*This chapter has been previously published in Colbran et al. 2019. *Nat. Eco. Evo.*

derthal sequences. Understanding the functions of these regions would provide valuable insight into barriers to introgression, the role of selection in determining the landscape of archaic DNA in modern populations, and the phenotypic differences between archaic and modern humans. We addressed this challenge by quantifying divergence in gene regulation between archaic hominin and AMH sequences and associating divergently regulated genes with AMH phenotypes using existing annotations and a large biobank linked to electronic health records (EHRs)[79]. Our results demonstrate substantial divergence in gene regulation between hominins and have the promise to highlight previously inaccessible differences in archaic hominin biology.

2.2 Quantifying gene regulatory divergence with PrediXcan

To identify archaic hominin sequences likely to have divergent gene regulatory effects compared to AMH sequences, we developed a statistic based on applying PrediXcan models to modern and archaic sequences. PrediXcan imputes the *cis* genetically regulated component of gene expression for genes in specific tissues using paired genotype and transcriptome data from human populations (Fig. 2.1A, B). Previous work has demonstrated that PrediXcan can impute the genetically regulated component of gene expression for thousands of genes, especially those whose regulatory architecture is dominated by common variants[79]. We considered accurate (FDR < 0.05) PrediXcan models of autosomal gene regulation from 44 tissues that were trained and evaluated on paired genotypes and normalized transcriptomes from the GTEx Consortium[86], which consists of 85% European ancestry and 15% African ancestry individuals (Methods).

The output of a PrediXcan model is not a direct proxy for gene expression in an individual. Instead, it is an estimate of the genetically regulated component of gene expression in reference to the distribution observed in the population used to train the model. Thus, differences in PrediXcan values between individuals reflect differences in variant genetic regulatory effects, not necessarily differences in overall gene expression (Fig. 2.1B, C). Furthermore, the regulatory effects captured only capture those variants present in modern human populations (Fig. 2.2). To emphasize this distinction, we refer to these differences as divergent regulation.

We consider the regulation of two classes of genes: 1) those that lack archaic ancestry in any variant in their PrediXcan model and 2) those with archaic ancestry in at least one modeled variant in at least one AMH (Fig. 2.3A, Methods). We will first focus on the former group and refer to them as "genes without archaic regulatory regions" (GWARRs). For simplicity, we will refer to the latter as non-GWARRs.

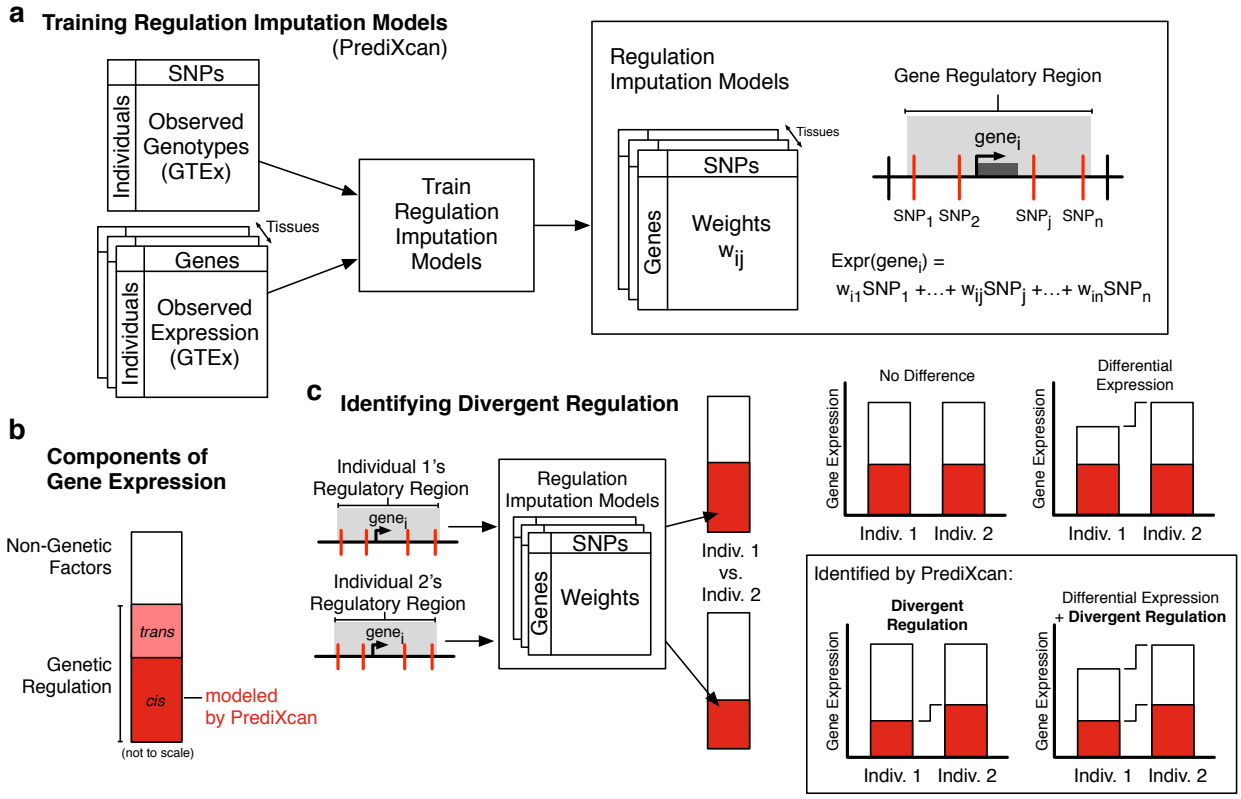

Figure 2.1: Identifying divergent gene regulation between individuals using PrediXcan. (a) Statistical models for imputing genetic regulation of gene expression (PrediXcan) were trained on genetic variants and normalized transcriptomes for 44 tissues from all individuals in the Genotype-Tissue Expression (GTEx) Project. Genetic variants within 1 Mb of each gene (Gene Regulatory Region indicated by gray box) were considered in the PrediXcan models; variants included in the models are illustrated by red vertical lines. (b) Gene expression levels are the result of genetic and non-genetic (e.g., environmental) factors. Our approach imputes the cis-genetic component of gene expression. (c) Our approach can identify divergent regulation between individuals, which reflects changes in the gene regulatory architecture, but does not necessarily imply differences in overall gene expression.

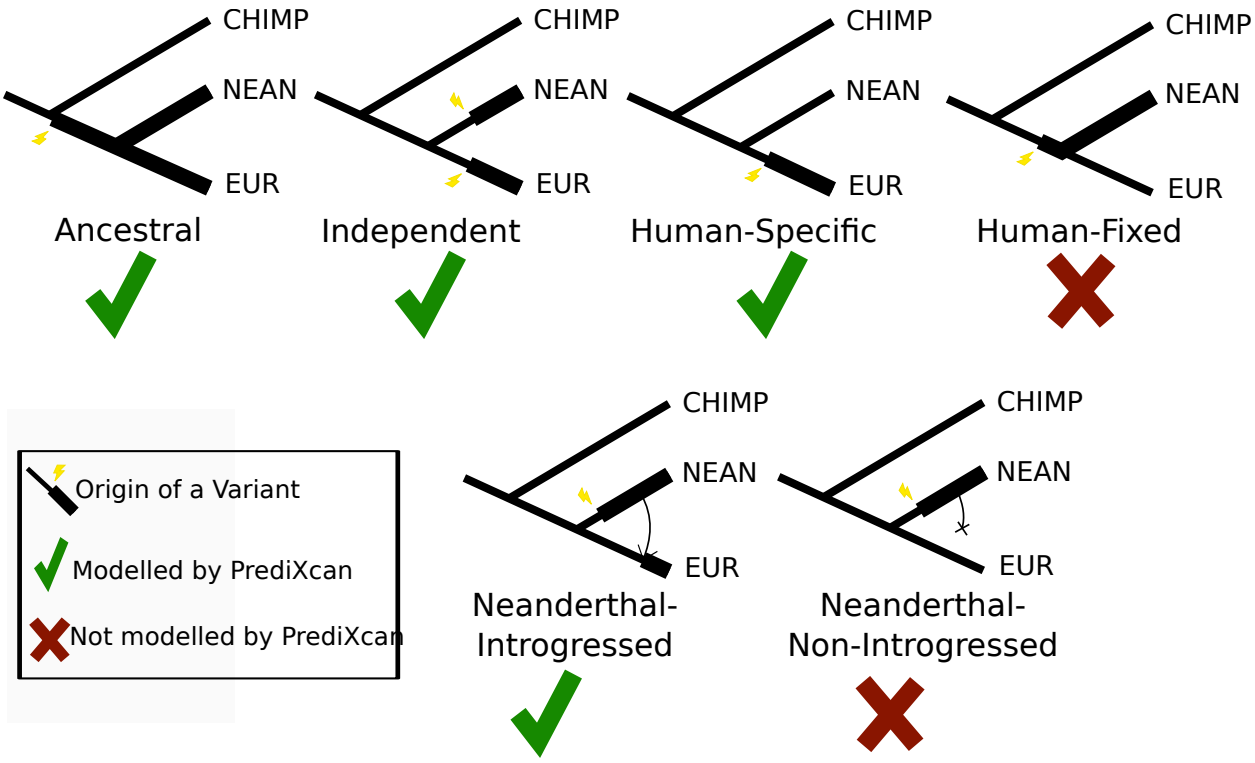

Figure 2.2: PrediXcan cannot consider all types of variant.
Specifically, it is unable to model the effects of any variant that is not present in the GTEx training population.

2.3 Identifying Neanderthal divergently regulated (DR) genes

We applied the imputation models to each gene's regulatory region from the high-quality genome sequence of the Altai Neanderthal[84]. This enabled us to estimate the effects of Neanderthal sequences on the regulation of 8587 GWARRs and 8854 non-GWARRs (Fig. 2.3A). We compared the gene regulatory effects of the Neanderthal sequence to the distributions observed when applying the same models to the corresponding regulatory regions of 2504 diverse AMH individuals from Phase 3 of the 1000 Genomes (1kG) Project[87] and computed empirical P-values for the observed differences (Fig. 2.3A). Again, since our approach estimates the genetically controlled component of gene expression in AMHs, their output should not be seen as a direct proxy for gene expression (Fig. 2.1C). Thus, we use difference in the values for a gene between AMHs and Neanderthals as a proxy for differences in the regulatory architecture between the groups. We

refer to genes for which the Neanderthal sequence's value is outside the range observed over all 1kG individuals as Neanderthal divergently regulated (DR) genes (Fig. 2.3A).

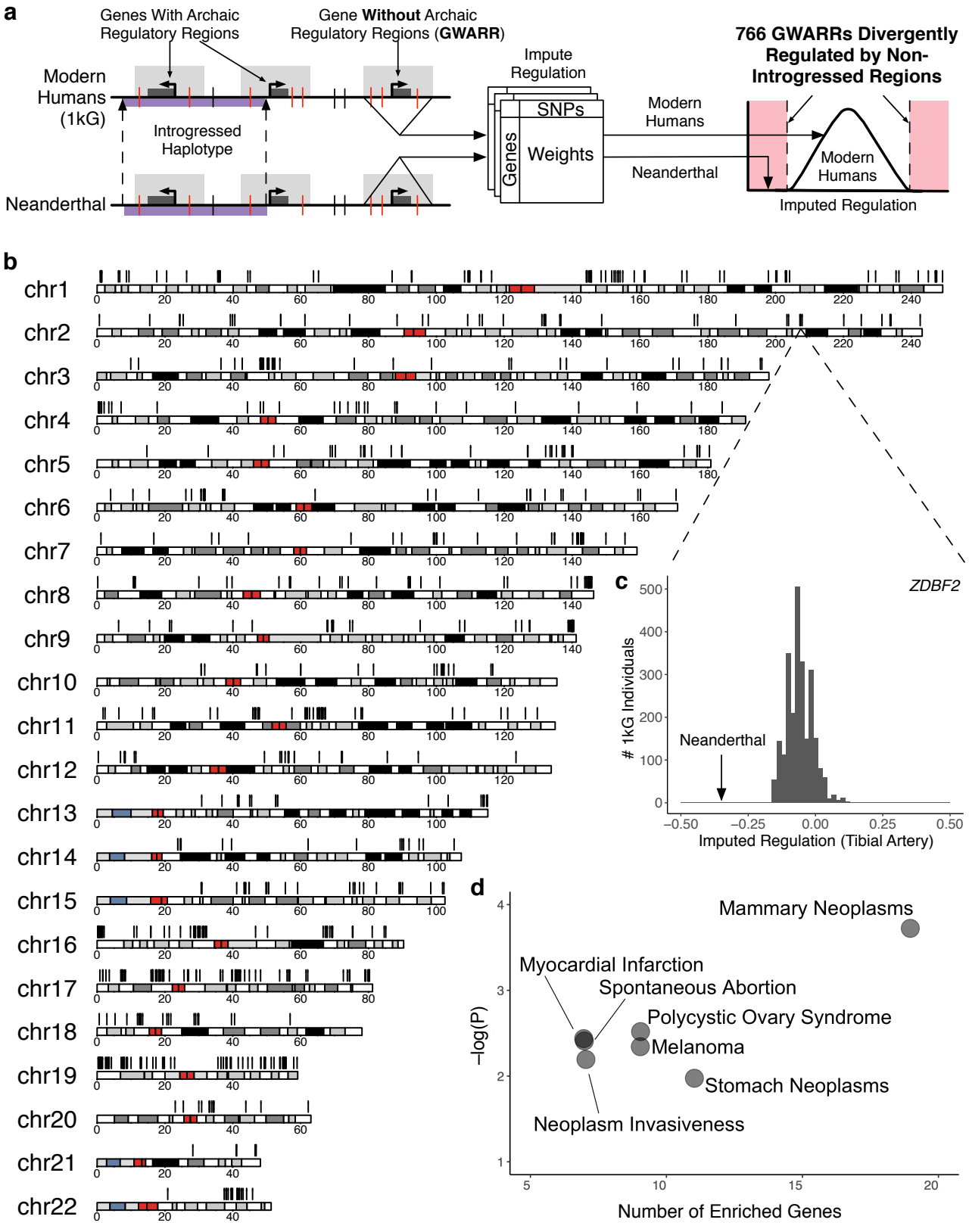
a
Genes With Archaic Regulatory Regions
Gene Without Archaic Regulatory Regions (GWARR)
Modern Humans (1kG)
Introgressed Haplotype
Neanderthal
Impute Regulation
SNPs
Genes
Weights
Modern Humans
Neanderthal
766 GWARRs Divergently Regulated by Non-Introgressed Regions
Modern Humans
Imputed Regulation
b
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
c
ZDBF2
1kG Individuals
Neanderthal
Imputed Regulation (Tibial Artery)
d
Mammary Neoplasms
Myocardial Infarction
Spontaneous Abortion
Polycystic Ovary Syndrome
Melanoma
Stomach Neoplasms
Neoplasm Invasiveness
−log(P)
Number of Enriched Genes

Figure 2.3: Neanderthal sequences drive substantial divergent regulation compared to modern humans.

(a) Pipeline for comparing the effects of modern human and archaic hominin DNA on gene regulation in modern humans. We identified genes in modern humans without archaic introgression in their regulatory regions (GWARRs). We compared the imputed gene regulatory effects of Neanderthal sequences to the regulatory effects of the corresponding human sequences in individuals from the 1000 Genomes Project (1kG). Genes for which the regulatory effect of the Neanderthal sequence was outside the range of all modern humans were labeled as divergently regulated (DR). (b) 766 GWARRs across the human genome (black lines) are divergently regulated by non-introgressed Neanderthal sequences in at least one tissue. (c) To illustrate the DR pattern, if the Altai Neanderthal sequence surrounding ZDBF2, a GWARR, were present in AMH genomes, it is predicted to drive regulation in tibial artery significantly lower than levels observed for all modern humans in 1kG (imputed regulation = –0.376, $P = 0$). (d) DR GWARRs are enriched for roles in several diseases, including spontaneous abortion, myocardial infarction, and melanoma, compared to all DR genes (FDR < 0.1, hypergeometric enrichment test on DisGeNET annotations).

2.4 Non-introgressed Neanderthal sequences divergently regulate 766 genes

Across all autosomes, non-introgressed Neanderthal sequences are predicted to divergently regulate 766 GWARRs in at least one tissue (Fig. 2.3B; table available at www.github.com/colbrall/). We refer to these genes with predicted divergent regulation as DR GWARRs. DR GWARRs are found on all autosomes, with the greatest density on gene-rich chromosome 19 (Fig. 2.3B). DR GWARRs are also observed across all tissues in GTEx (Appendix 1, Fig. 6.1), and are similarly likely to be upregulated or downregulated by the Neanderthal sequence (Appendix 1, Table 6.1).

Neanderthal sequences drive significantly more divergent regulation than observed when comparing sequences from an individual AMH to all others (12.4 times higher than maximum observed for an AMH, $P < 0.02$; Appendix 1, Fig. 6.2). Most genes exhibit similar regulatory effect distributions between human populations (Appendix 1, Fig. 6.3), and genes with large population differences are not enriched among DR GWARRs ($P = 0.821$, Fisher's exact test). This suggests that the divergent regulation is specific to Neanderthals. Additionally, DR genes have a similar number of Neanderthal-specific alleles in their regulatory regions when compared to non-DR genes, indicating that the amount of unmodeled variation is not driving the differences (Appendix 1, Fig. 6.4).

To highlight bodily systems that were not receptive to Neanderthal sequences with divergent regulatory potential, we tested for enrichment of specific disease and phenotype associations among DR GWARRs compared to all DR genes. DR GWARRs were significantly enriched (FDR < 0.1, hypergeometric test on DisGeNET annotations with Benjamini-Hochberg (BH) multiple testing correction) for genes involved in spontaneous abortion, polycystic ovary syndrome, mammary neoplasms, myocardial infarction, melanoma, and stomach neoplasms (Fig. 2.3D). Given their potential fitness effects, the DR GWARRs associated with spontaneous abortion (HSD17B1, IFI35, MUC4, IL20RA, TGFBI, TNFSF13, CD7) are of particular interest for further investigation.

We also tested for enrichment of Human Phenotype Ontology (HPO) annotations among DR GWARRs. While it did not pass multiple testing correction, the strongest enrichment was for genes involved in pectus carinatum, a deformity of the chest caused by overgrowth of the ribs and characterized by protrusion of the sternum (P = 4.3E–4; HP:0000768: GNPTG, HBA1, HBA2, MYH11, ORC4, SOS1, TNFRSF11B). The top associations also included other phenotypes that mirror physiological differences between humans and Neanderthals such as supraorbital ridge development (HP:0009891; HBA1, HBA2, PEX11A, and PEX13). Furthermore, many individual DR GWARRs function in human-specific phenotypes, including reproduction, neurotransmitter transport, circadian rhythm, and language. Overall, the large number of DR GWARRs suggests that there were substantial differences in gene regulation between modern humans and Neanderthals.

2.5 Divergent regulation of GWARRs is associated with clinical phenotypes in AMHs

To gain further insight into organism-level effects of divergent regulation of GWARRs in modern humans, we quantified the association of their imputed regulation with clinical phenotypes using BioVU, Vanderbilt University's biobank of patient DNA samples linked to de-identified EHRs. We used logistic regression to test for associations between the imputed regulatory profiles of Neanderthal DR GWARRs with phenotypes derived from the EHRs of 23,000 individuals of European descent (Fig. 2.4A).

Variation in DR GWARR regulation in BioVU is associated with many phenotypes (22 at P

$<$ 1E–7 and 284 at $P <$ 1E–5) across a broad range of phenotypic categories (Fig. 2.4B). The strongest associations include (Table 2.1): MSH5, PRSS16, VARS, and NCR3 with type 1 diabetes (T1D, Phecode: X250.1*; P = 1.3E–11, 5.2E–8, 7.1E–8, 8.0E–8, respectively), C11orf65 with transient mental disorders (Phecode: X291.1; P = 3.1E–9), SPINT1 with pulmonary embolism and infarction (Phecode: X452.1; P = 7.2E–8), and PSRC1 with hyperlipidemia (Phecode: X272.1; P = 3.1E–8). Each of the genes associated with T1D is located in or proximal to the human major histocompatibility complex (MHC) locus on chromosome 6. Certain MHC alleles may have been acquired through adaptive introgression[88]; our results suggest that variation in other regions of the MHC that were not receptive to introgression is associated with disease. Driven by the large number of associations with T1D and other autoimmune diseases, the endocrine and metabolic disorders phenotype category had the largest number of associations (Fig. 2.4B), but the raw number of associations is difficult to compare across categories due to differences in sample size, power, and between-phenotype correlations. Furthermore, the directions of effect for these associations do not always suggest that regulation by the Neanderthal haplotype increases risk. Nonetheless, divergent regulation of genes for which Neanderthal sequences likely altered regulation is associated with risk for clinical phenotypes in modern human populations. This highlights genes and bodily systems for which the lack of Neanderthal ancestry near genes may be due to divergent gene regulatory function.

Overall, the functions of DR GWARRs observed in the enrichment and biobank analyses suggest effects on a range of phenotypes, including reproductive, skeletal, cardiovascular, and immune traits. These systems are also influenced in AMHs by introgressed Neanderthal sequences[62,66]. This is consistent with a model in which these systems differed between Neanderthals and AMHs, and the genetic variants influencing these differences potentially had a range of fitness effects in the AMH context.

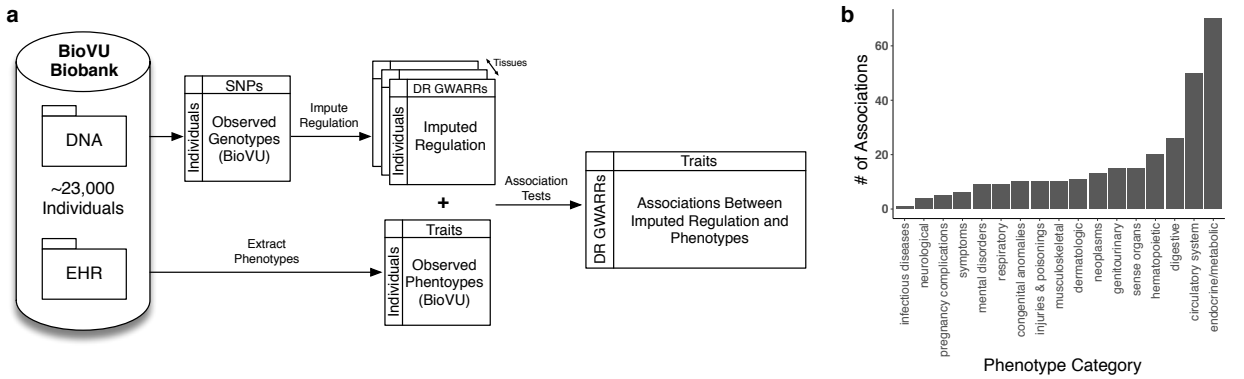

Figure 2.4: Modern human variation in the regulation of GWARRs is associated with clinical phenotypes.
(a) Pipeline for associating variation in gene regulation with diverse clinical phenotypes. Using Vanderbilt's BioVU biobank, human regulation of genes divergently regulated by non-introgressed Neanderthal sequences (DR GWARRs; Fig. 2.3) were imputed across 23,000 European ancestry individuals. Combining these genes' imputed regulation and phenotypes extracted from electronic health records (EHRs), we associated differences in imputed regulation to disease status using logistic regression controlling for standard covariates (Methods). (b) The number of associations between Neanderthal DR GWARRs and phenotypes in different phenotype categories at $P < 1E–5$. The endocrine and metabolic disorders phenotype category had the largest number of associations driven by many associations with T1D and other autoimmune diseases (Table 2.1). However, we caution against comparing across categories due to differences in sample size and power.

2.6 Genes in introgression deserts are not more likely to be divergently regulated

Given the potential importance of introgression deserts—long regions of the human genome significantly depleted of archaic ancestry—to human-specific biology, we examined the potential for divergent regulation by Neanderthal sequences among genes in six previously defined introgression deserts of greater than 8 Mb (Fig. 2.5)[33]. Each desert contained at least one DR GWARR, and the deserts contained a total of 26 DR GWARRs. 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)[89,90,91] and learning and spatial memory (CARF)[92].

Desert genes are not significantly more likely to be divergently regulated than other GWARRs (P = 0.60, permutation test). However, deserts have significantly lower recombination rates than

Table 2.1: Strongest associations between imputed regulation in BioVU and EHR-derived phenotypes for DR GWARRs.

Trait	**Gene**	**Beta**	***P*-value**
Type 1 Diabetes	*MSH5*	3.81	1.28×10^{-11}
Transient Mental Disorders	*C11orf65*	11.6	3.14×10^{-9}
Hyperlipidemia	*PSRC1*	-0.36	3.06×10^{-8}
Type 1 Diabetes with opthalmic manifestations	*PRSS16*	1.33	5.20×10^{-8}
Type 1 Diabetes	*VARS*	0.66	7.06×10^{-8}
Pulmonary Embolism	*SPINT1*	6.76	7.15×10^{-8}
Type 1 Diabetes with renal manifestations	*NCR3*	2.87	7.99×10^{-8}

**Each gene associated with T1D is located in or near the MHC locus.

other regions (Fig. 2.5B), and the deserts also have significantly lower gene densities. Controlling for these factors, there was still no significant difference in the likelihood of desert genes being DR than other GWARRs (Fig. 4C; matched recombination rate OR = 1.02; Fisher's exact test P = 0.99; matched gene density OR = 1.04, P = 0.96). Recent work suggests that recombination rate influences the retention of introgressed sequences[93], so it is possible that selection against a small number of diverged and deleterious regulatory Neanderthal haplotypes in these low recombination rate regions could have contributed to the formation of introgression deserts.

2.7 Imputing gene regulation in multiple archaic hominins

Due to the rapid degradation of most tissues and RNA, we are unlikely to ever be able to study gene expression levels directly from archaic samples. Archaic methylation status can be imputed for some regions of the genome[78], but this approach is limited to the bone cells from which archaic DNA can be extracted. In the previous analyses, we focused on the gene regulatory effects of Neanderthal DNA in the AMH genomic context. However, comparing gene regulatory profiles from archaic hominins directly may also reveal attributes of tissues in archaic hominins and their differences from one another. This approach is particularly promising for groups, like the Denisovans, that lack a substantial fossil record.

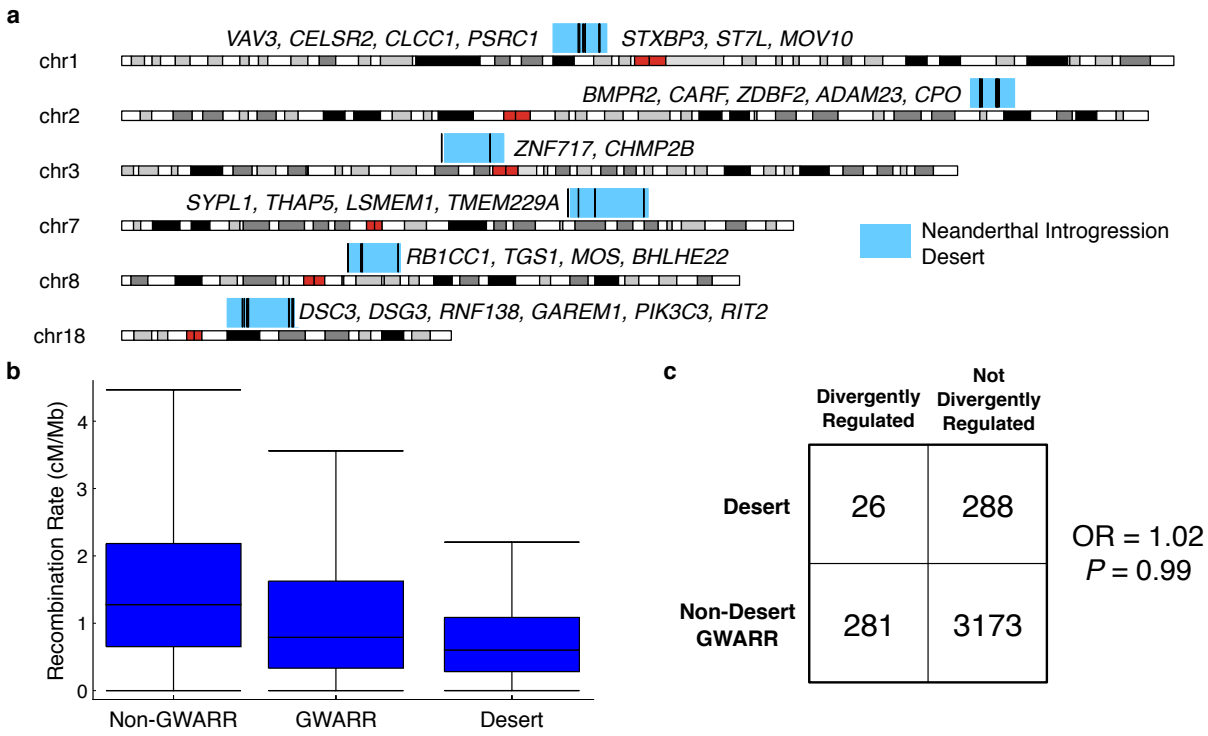

Figure 2.5: Genes in introgression deserts exhibit divergent regulation between modern humans and Neanderthals.
(a) Location of Neanderthal introgression deserts (blue boxes) and desert genes divergently regulated by Neanderthal sequences (black lines). Neanderthal DR genes are listed next to each desert. These genes have functions in a range of traits important to humanness, including neural development (CELSR2, CHMP2B) and spatial memory (CARF). (b) Recombination rate is significantly lower near genes (+/- 2 Mb) in introgression deserts than near other GWARRs or genes with archaic regulatory regions (Kruskal-Wallis test P 0, Dunn's post hoc analysis P 0.0). Box plots show the median, inner quartiles, and 95% confidence intervals. (c) Desert GWARRs are not significantly more likely to be DR compared to other GWARRs, even after controlling for recombination rate (OR = 1.02; Fisher's Exact test $P = 0.99$).

We expanded our analysis and imputed the regulation of all genes in the high-quality genomes of the Altai Neanderthal, a Neanderthal from Vindija, Croatia[80], and a Denisovan from the Altai cave[94]. To enable direct comparison, we reanalyzed the Altai Neanderthal using the smaller set of variants called in both Neanderthal genomes.

To obtain a global view of the similarity of regulatory patterns across tissues for each archaic individual compared to modern human populations from 1kG, we hierarchically clustered individuals based on the Pearson correlation of their regulatory profiles for all genes analyzed in each

tissue. This revealed that, as expected, the three archaic individuals are closer to one another than to any AMH (Fig. 2.6A). Also, as expected, despite being separated by more than 50,000 years and nearly 5,000 kilometers, the two Neanderthals' imputed regulatory profiles are more similar to one another than to the Denisovan (Fig. 2.6A, inset). Modern humans consistently group by continental population and all pairs of humans are more similar to one another than to any of the archaic individuals. Thus, the divergence of regulatory patterns in the archaic samples reflects our understanding of their evolutionary relationships with respect to one another and AMHs. These results held across all tissues analyzed and when we separated genes by the presence of archaic ancestry in their regulatory regions. We view these trees as a qualitative sanity check and caution against quantitative interpretation of the branch lengths as they are influenced by selective and demographic factors[95], as well as unmodeled archaic alleles (Appendix 1, Fig. 2.2).

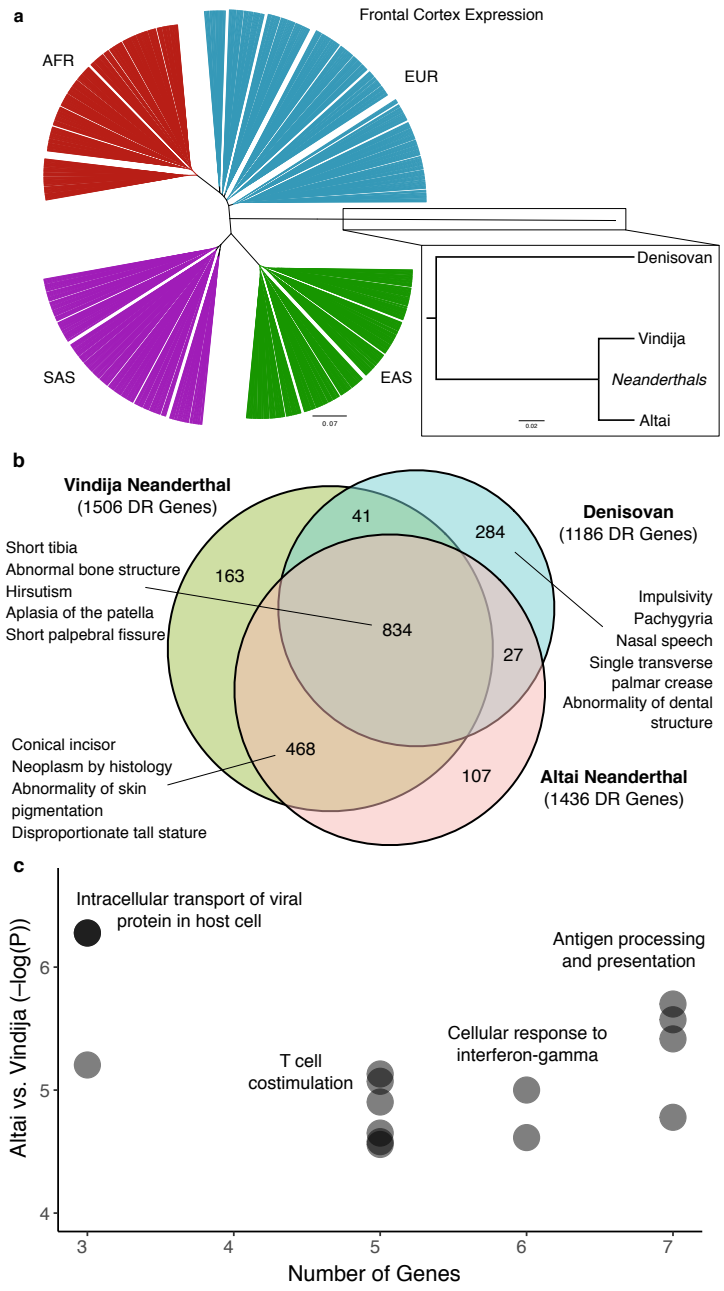

Figure 2.6: Comparison of genome-wide regulatory profiles between two Neanderthals, a Denisovan, and modern humans.
(a) Hierarchical clustering of imputed gene regulation for all genes in the frontal cortex of archaic hominins and modern human populations from 1kG. (b) Venn diagram of divergently regulated genes identified in each archaic hominin vs. all AMHs. Examples of the top 10 enriched Human Phenotype Ontology annotations among genes divergently regulated in all archaic individuals, in both Neanderthals, and in the Denisovan are shown. All terms are given in Appendix 1, Tables 6.2-6.4. (c) Enrichment for GO Biological Process annotations among the 75 genes with the largest deviation (>1 standard deviation) in imputed regulation between the Altai and Vindija Neanderthals. Immune functions are significantly enriched for differences between the two Neanderthals. Only enrichments with FDR < 0.05 are plotted.

2.8 Differences in regulation between archaic hominins reflect potential phenotypic differences

To identify specific differences in gene regulation between AMHs and the archaic groups, we determined divergently regulated genes in each archaic hominin compared to AMHs and tested for enrichment of phenotype annotations from the HPO (Fig. 2.6B). Across all tissues, 97% of DR genes in the Altai Neanderthal were also DR in Vindija with the same direction of effect. Genes divergently regulated in all three archaic individuals compared to AMHs were nominally enriched for associations with short tibia (7.15x, $P = 0.0017$, hypergeometric test), abnormal bone structure (1.62x, $P = 0.0034$), hirsutism (2.61x, $P = 0.0042$), and many other traits (Appendix 1, Table 6.2). DR genes specific to Neanderthals and the Denisovan were both nominally enriched for phenotypes involving dental morphology (Appendix 1, Tables 6.3 and 6.4). The Neanderthal-specific DR set also included genes involved in skin pigmentation (1.96x, $P = 0.0058$) and stature (7.62x, $P = 0.0063$). The repeated enrichment for genes involved in skeletal and dental morphology is striking given the known differences between modern and archaic hominins in these traits. DR genes specific to the Denisovan were uniquely enriched for several phenotypes including impulsivity, cerebral cortex development (pachygyria and lissencephaly), hand morphology, and nasal speech (Appendix 1, Table 6.4). The potential Denisovan-specific differences in speech are further supported by recent results based on imputed DNA methylation changes[96]. However, we note that due to the large number of phenotype categories these associations did not pass FDR-based multiple testing correction. Collectively, these analyses highlight genes involved in known morphological differences between archaic hominins and AMHs and suggest additional phenotypic differences that cannot be directly studied from fossils.

To identify differences in regulation between the archaic individuals without comparison to AMHs, we analyzed genes with large magnitude (>1 SD of the GTEx distribution) differences in regulation between archaic individuals. As expected, the two Neanderthals have the fewest differences (75 genes vs. 950 for each compared to the Denisovan). Immune response functions are significantly overrepresented among the genes different between the Neanderthals (Fig. 2.6C; FDR < 0.05), including transporting viral proteins (366.8x, $P = 0.0015$, hypergeometric test) and

cellular response to interferon-gamma (12.03x, $P = 0.0085$). These 75 genes include 5 MHC class II genes. This suggests that gene regulatory differences between the two Neanderthals influenced immune function, possibly reflecting adaptations in these populations. The genes that differed in the Denisovan compared to both Neanderthals are associated with many more general terms. Altogether, these results identify thousands of candidate genes for which regulation has likely diverged between archaic hominins and modern humans.

2.9 Discussion

Our application of PrediXcan to archaic genomes is a powerful approach for studying the evolution of gene regulation and the biology of archaic groups. The molecular machinery and genetic architecture of gene regulation are largely conserved across humans, and most common human regulatory variants have similar effects across populations[97,98]. Our approach enabled us to study the regulation of many genes by archaic hominin sequences. However, accurate predictions cannot be made for all genes in all populations, especially for genes with regulatory architectures dominated by rare variants[99,100] or trans effects[101]. Furthermore, since the imputation models are trained in modern humans, they do not incorporate the effects of archaic-specific alleles not present in human populations (Appendix 1, Fig. 2.2). Thus, it is likely for some genes that archaic-specific alleles could further modulate regulation. In these cases, the imputed effects are likely less accurate than in human populations, but any predicted deviations would still indicate divergence in regulatory architecture between archaic and AMH groups. As our understanding of the relationship between genotype and gene regulation improves and more tissues are characterized, our approach will enable testing of additional hypotheses about aspects of archaic hominin biology that are inaccessible to direct study.

In summary, there was substantial divergence in gene regulation between archaic hominins and modern humans. The affected genes influence a range of traits, including reproduction, skeletal development, language, and the immune system. Applying the regulation imputation models to a large, EHR-linked human biobank cohort further enabled the connection of divergent gene

regulatory patterns with clinical phenotypes in modern human populations, in particular with autoimmune and cardiovascular disease. Our results suggest that divergent regulation may have been a barrier to Neanderthal introgression in some regions of the human genome; however, more work is needed to demonstrate this. We additionally show that imputing ancient gene regulatory profiles has promise for studying ancient phenotypes. This approach is also potentially applicable to more recent ancient human genomes, where there is less sequence divergence than among Neanderthals and AMHs, and could provide an opportunity to characterize gene regulation across diverse geographical and temporal ranges.

2.10 Methods

Modern and Archaic Genetic Data

We analyzed the high-coverage genome sequences of three archaic hominins. For most comparisons to modern humans, we used the high quality archaic genome from an 122,000-year-old Neanderthal individual found in the Altai mountains ("Altai Neanderthal")[84], which was sequenced to 52x coverage and enabled PrediXcan analysis of the largest number of genes. For the comparisons that included multiple archaic individuals, we analyzed the 30x genome from a 72,000-year-old Denisovan from the Altai mountains ("Denisovan")[94], and a 30x coverage genome of a 52,000-year-old Neanderthal from Croatia ("Vindija Neanderthal")[80]. For all three genomes, we considered only autosomal SNPs from the publicly available genomes. To represent modern humans, we analyzed the genomes of 2504 individuals sequenced by the 1000 Genomes Project (1kG) and released in Phase 3[87]. These include individuals from the European (EUR), African (AFR), East Asian (EAS), South Asian (SAS), and Admixed American (AMR) continental ancestry super-populations.

PrediXcan Gene Regulation Imputation Models

We considered PrediXcan models across 44 tissues from the PredictDB Data Repository (accessed Nov. 16, 2016). The models were trained on GTEx V6p[86] using variants identified by 1kG (Phase 1)[102] within 1 Mb of the gene. We considered only those models that explained a significant

amount of variance in gene expression in each tissue (FDR < 0.05); this left us with 17,748 unique genes with an accurate model in at least one tissue (159,368 models total)[79]. We abbreviate the 44 tissues considered as follows: 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.

Imputation of Archaic Hominin and Modern Human Gene Regulation

Using the PrediXcan prediction program available from PredictDB, we applied the accurate prediction models to the relevant portions of the genome of the Altai Neanderthal to impute the effects of its sequence on gene regulation. The resulting predictions are normalized values in reference to the distribution observed in GTEx individuals used to train the original prediction models. To characterize regulatory patterns in modern human populations, we applied the same PrediXcan models to 2504 individuals from the 1kG[87]. For all cross-archaic comparisons, we applied the same models to the sequenced Vindija Neanderthal, the Altai Neanderthal, and Denisovan, which were all recently processed with the same pipeline[80]. Imputed regulation based on the previous and new versions of the Altai Neanderthal were strongly correlated (0.78–0.85 across tissues).

Identification of Genes Divergently Regulated by Archaic Sequences

To identify genes divergently regulated by archaic compared to modern human sequences, we calculated an empirical P-value for the archaic predicted regulatory profile for each gene and tissue by calculating the proportion of modern humans who had a predicted value farther from the median of the full 1kG distribution for the tissue. Genes for which the archaic sequence is predicted to drive regulation completely outside the distribution observed in 1kG in at least one tissue were considered significantly divergently regulated (DR) genes (N = 2290), 766 of these were GWARRs (see next section for more on the GWARR definition). We plotted gene locations using karyoploteR[103]. We excluded all genes which were missing genotype calls at SNPs of at least one model.

Assessment of Imputation Accuracy on Neanderthal Sequences

It is not possible to directly assess the accuracy of gene regulation imputation models trained in AMH when applied to Neanderthal sequences. To estimate how much Neanderthal-specific variation the PrediXcan models trained on AMHs could be missing, we counted the number of Neanderthal-specific alleles present in the regulatory region of each gene (1 Mb up and downstream). For this analysis, Neanderthal-specific sites include any site where the Altai Neanderthal had at least one allele not observed in 1kG. To account for different overall evolutionary rates between genes, we computed the relative amount of Neanderthal-specific variation for each gene by dividing it by the total number of variants (Neanderthal-specific alleles plus all variable sites in 1kG). We then compared relative levels of Neanderthal-specific variation between DR and non-DR genes. Further steps taken to assess model performance are described in full in Colbran *et al.*, 2019[104].

Divergent Regulation Between Humans

To aid interpretation of the number of divergently regulated genes observed with archaic sequences, we called DR genes in 50 random 1kG individuals, 10 from each continental population, using the same criteria as for archaic sequences: imputed regulation outside the range for all other 1kG individuals. For each population, we compared the distribution of the number of DR genes in each individual with the number identified in Neanderthal (Appendix 1, Fig. 6.2).

We also examined the stability of the imputed values across all 1kG populations. For all

PrediXcan models in all tissues, we computed the median imputed regulation for each 1kG population. We then found the maximum difference between populations (Appendix 1, Fig. 6.3). Only 2.7% of all gene models have a maximum difference in population median regulation greater than 1 SD.

Classification and Comparison of Non-Introgressed and Desert Genes

We used the S*-based Neanderthal introgression map from Vernot *et al.*[33] to identify the overlap between variants considered in gene regulation prediction models and introgressed sequences. After filtering out models that had no variants present in the Altai genome, we classified genes to be genes without archaic regulatory regions (GWARRs) if none of the variants considered in their prediction models were Neanderthal tag SNPs or in linkage disequilibrium ($r2 > 0.8$) with Neanderthal tag SNPs in Europeans (N = 8587). Genes with at least one introgressed Neanderthal SNP in their model were classified as "non-GWARRs."

We also analyzed the effects of genes in introgression deserts that were recently identified using coalescent simulations based on demographic models[33]. By this definition, deserts are long regions where modern humans lack introgressed sequence. Desert regions >8 Mb long are significantly more common than expected from simulations, and they also exhibit higher levels of background selection. In our analyses, "desert" genes are the subset of GWARRs for which variants in their regulatory effect prediction models overlap the bounds of an introgression desert, excluding those that also include SNPs on introgressed haplotypes (N = 311).

We calculated the enrichment of DR GWARRs within and outside deserts by shuffling GWARR locations across the genome, constrained by chromosome. For each of 1000 permutations, we counted the number overlapping a desert to compute an empirical p-value. To evaluate DR enrichment accounting for recombination rate, we first calculated the recombination rate in 250 kb non-overlapping windows across the entire genome, using recombination maps calculated in African Americans[93,105]. We then intersected those windows with the regulatory region considered by PrediXcan for each gene. For each gene, we calculated the mean recombination rate across all windows overlapping the gene region, weighted by the number of base-pairs of overlap. We then

binned genes by recombination rate (31 equal-width bins) and randomly selected 3454 GWARRs such that the overall distribution across bins was equal to the distribution of desert genes (the maximum without emptying a bin).We then performed a Fisher's Exact test on DR status in desert genes vs. the recombination rate-matched GWARRs. To match by gene density, for each gene we counted the number of genes that overlapped the region considered by PrediXcan (1 Mb flanking on either side). We then repeated the binning and Fisher's Exact test analyses as for recombination rate.

We identified gene regions overlapping human accelerated regions as those genes with at least one HAR within 1Mb[106]. We then computed an odds ratio to assess the likelihood of certain classes of genes to be nearby a HAR compared to others.

Association and Enrichment Between Divergent Regulation and Phenotypes

To investigate potential phenotypic implications of DR GWARRs, we conducted two main analyses: gene set enrichment analysis and PrediXcan on Vanderbilt's BioVU biobank. To test for enrichment of genes known to be involved in particular human phenotypes or diseases, we performed gene set overrepresentation enrichment analysis on Disgenet disease annotations and human phenotype ontology terms between DR GWARRs and other DR genes using WebGestalt[107]. We used the hypergeometric test with BH multiple testing correction, a false discovery rate (FDR) threshold of 10%, and did not consider disease categories with fewer than 10 genes.

To explore the systems potentially impacted by divergent regulation of DR GWARRs in modern human populations, we used the PredixVU system at Vanderbilt University Medical Center to discover associations between predicted regulation and clinical phenotypes. The phenotypes were extracted from de-identified electronic medical records using ICD-9 codes that were organized into PheWAS codes in 17 groups (https://phewascatalog.org/phecodes) and were linked to genotypes from the BioVU biobank. In total, this involved 23,000 subjects of European descent; the total number of cases and controls for each phenotype varied (on average 780 cases, 17176 controls). We considered only phenotypes with case counts greater than 30. The models used to impute regulation for these individuals were trained on HapMap SNPs, and there is a high correlation

between the imputed values for the models trained on HapMap and 1kG[79]. For each phenotype, we used logistic regression to regress imputed regulation onto phenotype status, and included age, sex, and genetic principal components (3 for Europeans, 10 for African Americans) as covariates. N.B. associations between divergent regulation of a gene and a phenotype in this context are not necessarily in the same direction as the divergence in the Neanderthal.

Comparing Gene Regulation Among Archaic Hominins

To visualize global similarities between different groups for each tissue, we compared the imputed regulatory profiles of non-admixed 1kG populations (excluded: MXL, CLM, PUR, ACB, ASW, PJL, PEL) and the three archaic hominins. We hierarchically clustered each individual for each tissue using Pearson correlation on imputed regulation across all genes as the distance metric. We visualized the resulting trees using FigTree. Results were similar when using Spearman correlation and when stratifying by GWARR status.

To identify specific genes of interest to differences between the archaic groups, we generated lists of genes divergently regulated between the Altai Neanderthal, the Vindija Neanderthal, and the Denisovan. First, we called DR genes versus the 1kG individuals for each archaic individual, and then intersected the DR genes. We then conducted gene set ORA over the Human Phenotype Ontology using WebGestalt[107], using only categories containing at least 10 genes.

To focus on genes with the largest differences in regulation, we computed the difference in predicted regulation between pairs of archaic individuals for each gene in each tissue for which it had an accurate model. We then picked genes that differed in imputed regulatory effect by greater than 1 (i.e., were >1 standard deviation apart with respect to the distribution of the GTEx training population). To identify general biological processes influenced by these genes that differed between the archaic hominins, we conducted gene set enrichment analyses on GO biological process terms versus the full GTEx project gene list using WebGestalt[107].

Chapter 3

MODELING GENE REGULATION WITH LOW-COVERAGE GENOMES

3.1 Introduction

The previous chapter demonstrated the potential power of PrediXcan and similar frameworks in answering evolutionary questions. However, that study represented a relatively simple case where the data we were interested in consisted of complete, high-quality genomes (despite the concerns over divergence). In many other potential use-cases available genetic data varies in coverage, depth, and quality. This creates a tradeoff between number of samples analyzed and the overall quality of the genotyping. In addition, many potential applications of PrediXcan are in populations that differ from the training populations; there is no guarantee, even if they are of similar ancestry, that all the variants the models are using to predict gene expression are actually assayed in the population of interest. Therefore it is of great interest to understand how PrediXcan behaves under these less-than-ideal conditions with varying levels of missing variant data, and develop ways to optimize its performance in such cases.

Being in the field of paleogenomics, we were particularly interested in the application of PrediXcan to ancient DNA ("aDNA") sequences. Because of time and chemistry, high-quality ancient genomes like the ones used in the previous chapter are a rarity. Given enough time in the elements, DNA gets physically degraded and contaminated in ways that must be taken into account[108,109]. Unlike studies using modern individuals, the option of resampling an individual if there was a lack of DNA the first time often is not an option due to the destructive nature of sampling. We therefore focused on the application of PrediXcan to lower-quality aDNA as a test case.

This chapter explores the impact of changing the variants available for the models during training, and quantifies the impact on predictions caused by missing variants used in training the models. As a case study, we focus on the application of PrediXcan to an aDNA dataset consisting of

thousands of individual samples compiled on a common genotyping chip, and produce recommendations for optimizing PrediXcan for this application.

3.2 Model performance decreases with missing data

First, we set out to evaluate how missing data influences the performance of models. Mathematically, if a PrediXcan model does not have genotype information for a site included in the model, it assumes the genotype was homozygous reference (i.e. dosage of alternate alleles equals zero). Therefore it multiplies the weight by zero, and because models are trained on normalized expression, a predicted value of zero (i.e. if sample was homozygous reference at all positions) results in the predicted expression being the mean of the training data. Because of this, missing variants ought to bias predictions toward that mean. The models thereby lose power to detect differences in populations and individuals, and potentially miss out on capturing the full genetic component of gene regulation. However, because PrediXcan does not explicitly filter variants that are in high linkage disequilibrium with each other, we would expect some models to be more robust to missing variants as they also include variants that are well-correlated with the missing ones. The exact dynamics of the performance of PrediXcan genotypes with missing data in practice is not well understood.

To better understand and quantify these patterns, we applied PrediXcan models trained on GTEx v8 with all available variants ("Full models") to genome sequencing information for 2504 individuals from the 1000 Genomes Project (1kG)[87]. We then selected nine thresholds for percentage of missing SNPs and downsampled 20 random European individuals per threshold such that they were missing a random set of SNPs and applied the full PrediXcan models to these downsampled genomes (Fig. 3.1A). Unsurprisingly, the agreement between the predictions on downsampled genomes and those for the full genomes was strongly correlated with the percentage of SNPs missing (Fig. 3.1B). However Spearman correlations were always above 0.75 even when genomes were missing as many as 45% of their SNPs, suggesting model accuracy can be maintained even at relatively high rates of missingness.

While this is encouraging, depending on the situation missing SNPs may not be randomly distributed throughout the genome. To check whether that nonrandomness could affect these results, we repeated the comparison described above. However, instead of randomly downsampling SNPs, we matched the patterns of missingness to a dataset (Fig. 3.1A) of particular interest to us: 3383 aDNA samples, with widely varying numbers of missing SNPs (Fig. 3.1C). Overall, the correlations were much lower (median Spearman ρ=0.39; Fig. 3.1D). This is unsurprising given that the aDNA samples were obtained on a genotyping chip, while the model training data was based on whole genome sequencing. At most, the aDNA samples had 714,959 SNPs, while the training data had over 5 million (i.e. 87% missing). This indicates that, while models can tolerate a fair amount of missing data, the missingness caused by a mismatch between genotyped SNPs and training SNPs is likely to cause problems with predictions.

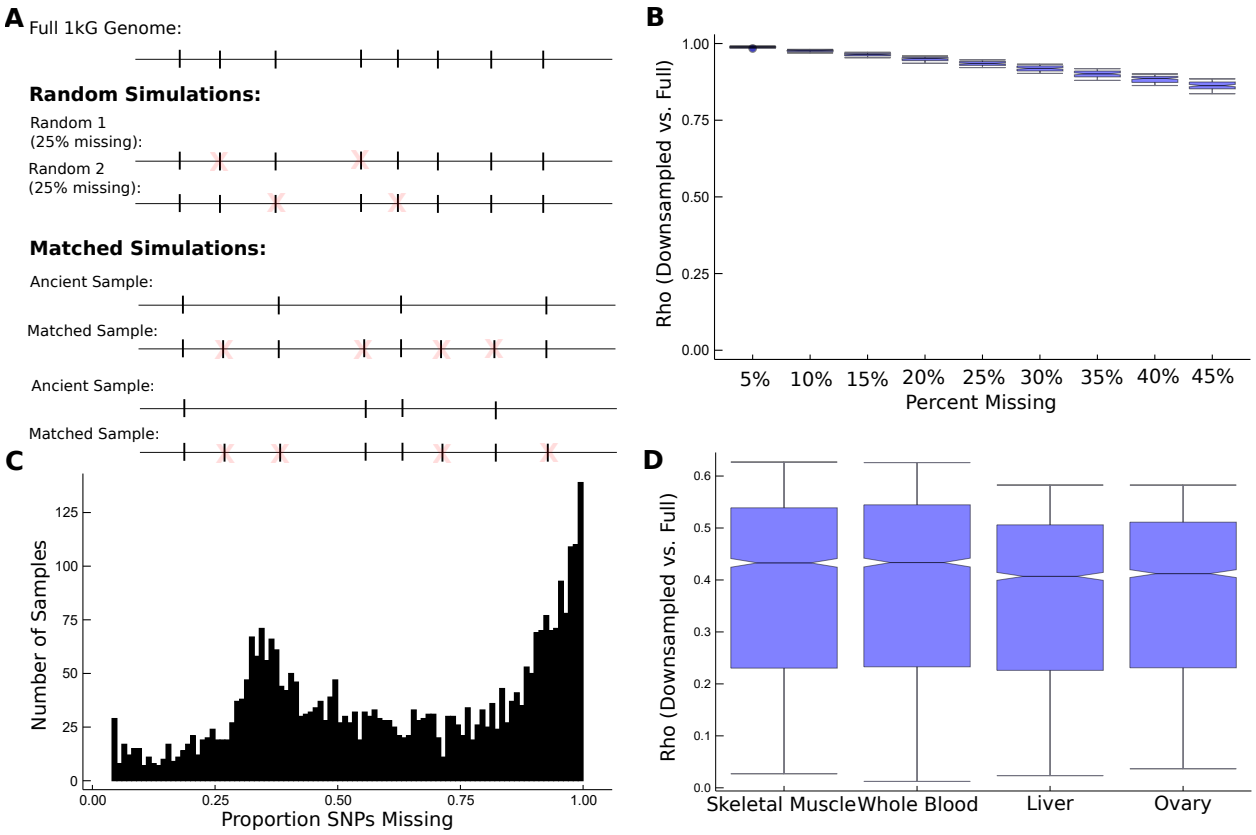

Figure 3.1: Impact of Missing SNPs on PrediXcan performance.
(a) Schematic of process for generating random and matched simulated genomes. Starting from a complete genome from 1kG, for the random simulations we mask a random number of variants corresponding to the specified missing percentage. For the matched simulations, we pair a complete modern genome and an ancient genome, and mask any variants in the modern genome that are not present in the ancient one. (b) The Spearman ρ between predictions in 4 tissues calculated for the complete genome vs. random simulations decreases as the percentage of missing variants increases. (c) Distribution of the proportion of missing variants in the aDNA data compared to all variants included in 1kG. (d) Spearman ρ between predictions in 4 tissues calculated for the complete genome vs. aDNA-matched simulated genomes.

3.3 Models can be trained with targeted variant sets

Due to the findings above, we evaluated whether models could be optimized for a given application by customizing the training data to contain only variants that will be available in the application data. This step would thereby ensure that any variants used to model gene expression were at least assayed in the application data and limit the models' reliance on SNPs they would be unable to use.

To pick a targeted set, we chose to use only the SNPs present on the 1240k genotyping chip that is commonly used in aDNA studies ("1240k set"; Fig. 3.2A). This resulted in 714,959 input variants for the models training (the number that were successfully lifted over to the hg38 genome build), as opposed to the 5,310,489 variants available for the full models. However, because most of the samples genotyped on this chip tend to be low-coverage (Fig. 3.1C), we also tested the use-case where we chose the SNPs in the dataset most likely to provide information. To do this, we ranked SNPs by the number of samples in the 3383 ancient samples used above with genotype calls, and weighted that count by the overall coverage of those samples. We then chose the 600,000 SNPs with the best ranking ("top600k set"), thereby prioritizing SNPs that were frequently present in the best-quality samples.

Unsurprisingly, we found that both sets of targeted models used fewer SNPs than when allowed access to the full data (Fig. 3.2B). While there was not a large drop in the variance in gene expression (r^2) the models explained, stricter targeting resulted in fewer significant models (Fig. 3.2C). While there was a high correlation between the variance explained in individual genes' expression, when there was disagreement, the targeted models had a lower r^2 (Fig. 3.3). This, coupled with the r^2 threshold required to be significant (Methods), likely explains the drop in the number of significant models. The models that were significant in both the Full set and targeted sets were those that had higher numbers of SNPs and explained more variance (Fig. 3.2D&E). This suggests that those models lost by selected targeted SNPs were lower-confidence ones.

While these results are encouraging for the applicability of PrediXcan to lower-coverage data, they do not speak to the performance of the targeted models on data independent of the training set. We therefore applied models trained on both the 1240k set and top600k set to all 2504 people in 1kG. We then compared the predictions obtained in those cases to those obtained by the Full models. We found that the targeted models generally predicted the same patterns of expression as the full ones did, although it varied by specific model (Fig. 3.4). This comparison could therefore be useful for filtering models in downstream analyses, allowing focus on those that are consistent when given targeted variants.

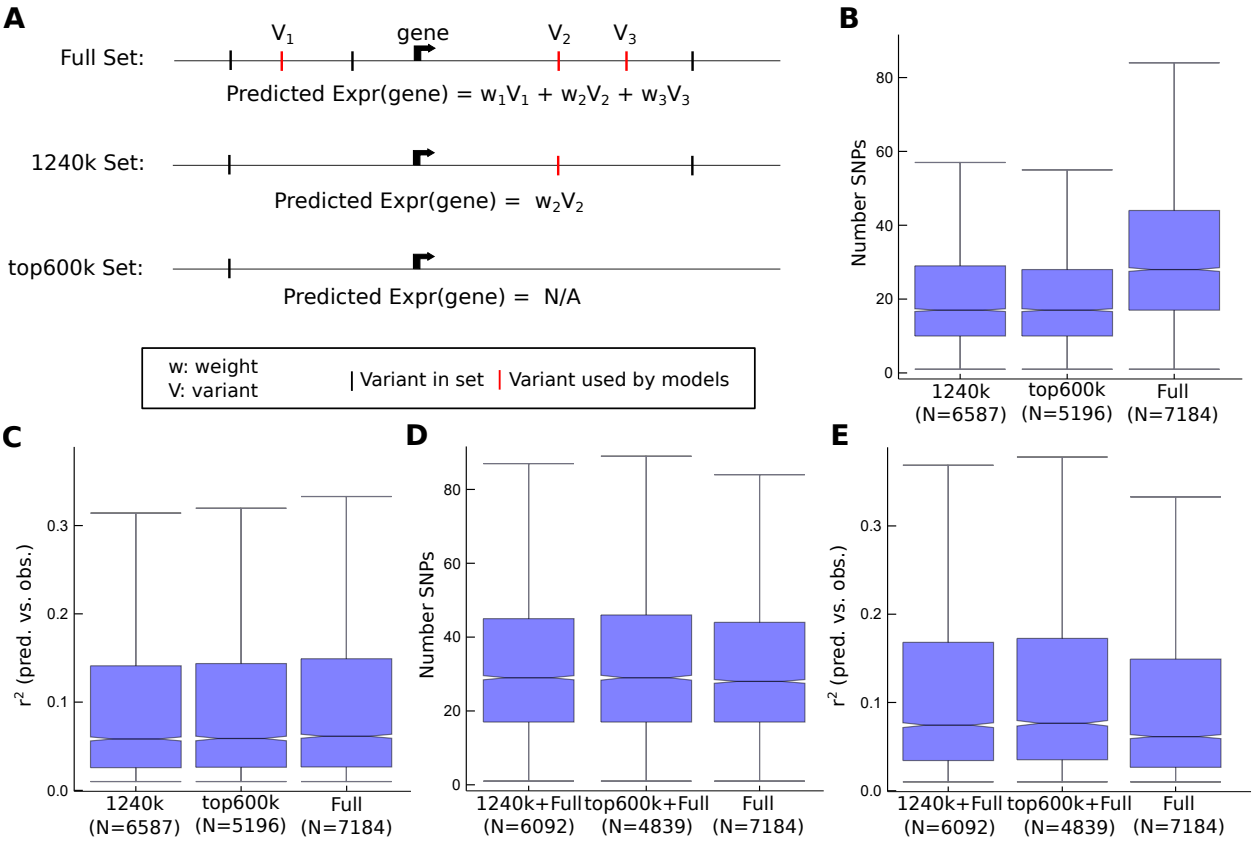

Figure 3.2: Targeted models show overall decrease in performance.
(a) Schematic of creating targeted variants sets to train PrediXcan models. (b) Number of SNPs in each set of models for Whole Blood. (c) r^2 between predicted and observed expression for each model in Whole Blood. By definition, significant models had to have $r^2 > 0.01$. (d) Number of SNPs and (f) r^2 in models identified as significant in both the Full set and either the 1240k or top600k Set (metrics plotted are those from the Full set). Full set replotted for comparison. Other tissues tested matched trends.

When comparing by individual, Both sets of targeted models were reasonably consistent with the Full models (Fig. 3.5A). Interestingly the more highly-targeted models had higher median and lower variance in agreement with the Full set (median ρ=0.67 for 1240k, ρ=0.80 for top600k). This is likely due to the enrichment for better models caused by using fewer SNPs. Overall, these results suggest that limiting the available SNPs does impact the ability of this framework to model gene expression. However, it does not hugely impact the ability to gain information about regulatory patterns of more easily-modeled genes.

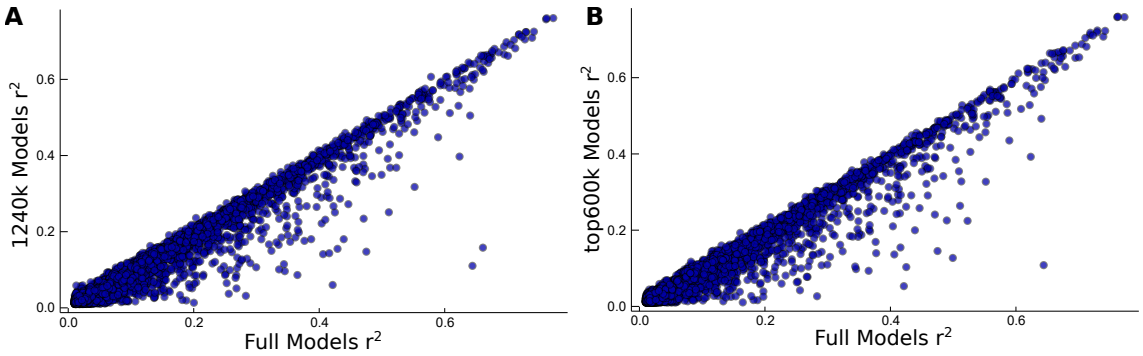

Figure 3.3: Scatterplots of training r^2
for (a) 1240k models and (b) top600k models vs. Full models in Whole Blood. Other tissues tested matched trends. r^2 is calculated over observed vs. predicted expression.

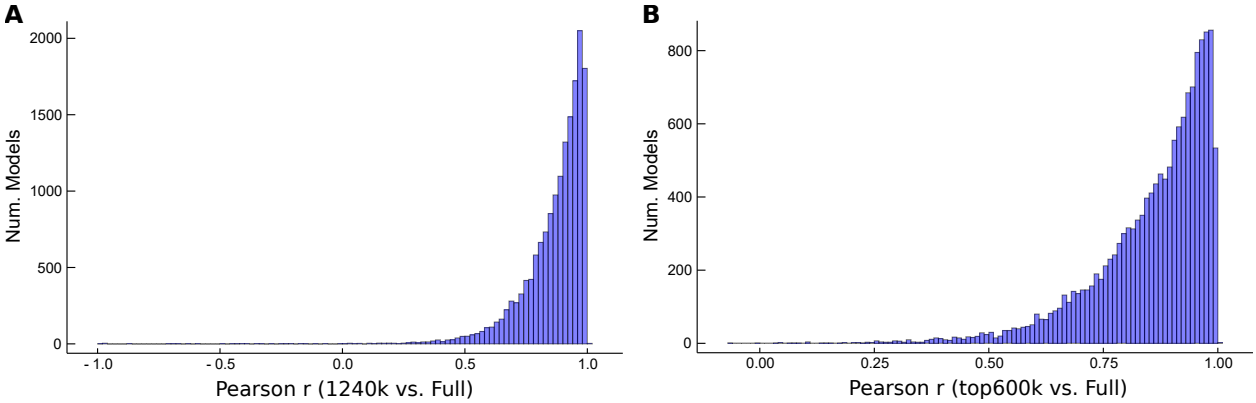

Figure 3.4: Targeted models predict consistent gene regulatory patterns.
(a) Pearson r by model between Full model 1kG predictions and (a) 1240k and (b) top600k models for all genes in 4 tissues- Liver, Ovary, Whole Blood, Skeletal Muscle. Agreement between predictions made on all 1kG individuals using Full models vs. 1240k or top600k models.

3.4 Targeting models increases susceptibility to missing data

Because models trained on variants targeted to a particular dataset use fewer variants, it is possible that they would be more susceptible to missing data than the full models. Therefore the outstanding question is whether model performance is more consistent when trained on fewer SNPs without a large missing data percentage, or if it is better to include more SNPs during training, but allow more missing data during application. To answer this, we applied our targeted models described above to 1kG genomes downsampled to match the various sets of SNPs used during model training and compared their agreement with the full models applied to the full genomes.

For the top600k models, we further downsampled the application set to 500k SNPs using the same methodology.

In line with our initial findings, we found that all models lost consistency when applied to genomes with missing SNPs (Fig. 3.5B). The 1240k models maintained the highest agreement with the Full models when applied to incomplete data (median ρ=0.61, vs 0.55 and 0.39 for Full and top600k models, respectively), though they also had a larger variance in agreement. The Full models likely did worse because there was a much larger drop in the number of SNPs available compared to training, while the top600k models' reliance on fewer SNPs may have increased their susceptibility to missing SNPs. This suggests that a balance between targeting model training for the dataset in question and allowing some missingness is the best course of action.

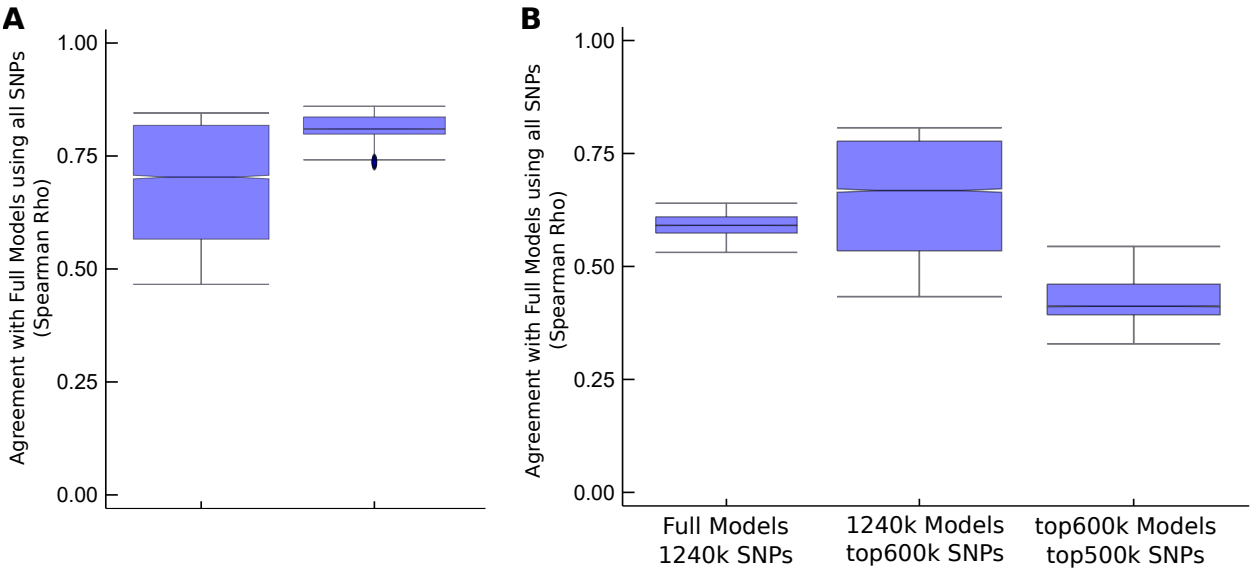

Figure 3.5: Models with fewer SNPs in training are more susceptible to missing data.
a) Spearman rho between predictions on 1kG individuals using Full vs 1240k models and top600k models with all SNPs available. b) Spearman rho between predictions of Full models applied to 1kG with all SNPs available vs downsampled 1kG genomes. In X-axis labels, the top row indicates what models were used, and the bottom row indicates what SNPs were available from the 1kG genomes.

3.5 Discussion

Overall, PrediXcan models perform best when there is a large number of variants available to train them and these variants are also assayed in the application population. If the training variants are limited in number, the significant models decrease in number (by as many as 2000 for top600k models in Whole Blood), though the performance of those models is not drastically altered. If missing variants are prevalent in the application data, model predictions become more unstable (median Spearman rho as low as 0.38 for top600k models with missing data vs. Full models and all SNPs).

However, millions of variants are not required for PrediXan models to be informative about gene expression. These simulations show that targeting the training SNPs specifically to the dataset of interest allows performance to be maintained better than if the models are trained on a larger set of SNPs that includes many that are not present in the application cohort (median Spearman ρ=0.67 for 1240k models vs. 0.55 for Full models with missing data vs. Full models and all SNPs).

Collectively, while small percentages of missing data do not cause drastic decreases in model performance, using too few SNPs in training does increase the susceptibility to incomplete data. Therefore, while it is advisable to target training data to the application, it is better to err on the side of including more SNPs initially while allowing some to be missing in some samples the models will be applied to. In our test case, for example, we would opt for models trained only with the variants on the genotyping chip commonly used in aDNA research, and then select samples with a low percentage of missing variants to run downstream analyses on.

In addition to better understanding the impact of limited data on genetic-based models, this work demonstrates the power of simulated data for better understanding how models behave under different conditions. This allows for better-informed decision making about what will work for a given dataset. Simulations also provide additional criteria for filtering both samples and genes to those for which the given framework and conditions will work. In our test case, we would filter the samples to which we'd apply PrediXcan models to just those within a certain threshold of missing SNPs, based on the simulations described here. In addition, we would filter the specific models

under consideration to include those that demonstrated robust predictions when limited data were available.

Overall, these results suggest that PrediXcan and other similar methods can retain utility even when variant data are limited, as long as the models are trained for the specific application and their limitations properly taken into account. This is encouraging for the expansion of gene regulatory studies into datasets that do not exactly match the training data, and indicates that it is possible and potentially useful to balance data quality requirements in terms of numbers of SNPs with called genotypes with obtaining a larger sample size. As newer, more accurate methods are developed, it will be important to keep this aspect of their performance in mind, and encourage their use in diverse applications.

3.6 Methods

Training and evaluating models. We trained PrediXcan models using code adapted from the PredictDB pipeline. We trained all models on RNA-seq from 4 tissues collected post-mortem, and whole genome sequencing data from GTEx version 8[110]. The tissues were Skeletal Muscle, Whole Blood, Liver, and Ovary, and were selected to represent a range of available sample sizes. We evaluated each models by calculating an r^2 between the predicted expression and observed expression. To be considered significant, a model had to pass an FDR correction in that tissue and have and $r^2 > 0.01$.

Simulating random missing data. For each percentage missing threshold, we randomly selected 20 European individuals from 1kG[87], then randomly removed that percentage of genotype calls from their genomes before applying PrediXcan models to the simulated genomes. For each downsampled genome, we calculated a Spearman correlation between the predicted regulation of each gene in 4 tissues for the downsampled vs. the full genome. Therefore, each box in Fig. 3.1B has 80 (20×4) points. We then calculated the Spearman correlation between the median correlation between downsampled and full model predictions for each threshold and the percentage of SNPs missing at that threshold.

Simulating matched missing data. For our example study population, we used 3383 ancient human samples compiled and made available by the Reich lab on March 1, 2020 (v42.4). Our set includes all that passed their QC process. We picked three random Europeans from 1kG, then for each ancient sample we created three matching downsampled genomes that were missing exactly the same complement of SNPs. For each downsampled genome, we calculated the Spearman correlation between the predicted regulation of each gene in all 4 tissues for the downsampled vs. the full genome.

Selecting target SNPs for model training. We chose three sets of variants for training PrediXcan models. The "full set" consisted of all variable sites identified in GTEx v8 (this included both SNPs and short indels, hg38 coordinates). The "1240k set" was formed by intersecting the full set with the set of variants genotyped on the 1240k chips, which totalled 714,959 SNPs after lifting them over to hg38. Lastly, we assembled the "top600k set" of SNPs, which is a subset of the 1240k set. To do this, we calculated the "support" for each SNP; for N aDNA samples, support equals $\sum_{n=1}^{N} NumSNPs_n$, where $NumSNPs$ is the number of SNPs successfully called in sample n. In other words, support for a SNP is the number of samples in which that SNP was successfully genotyped, weighted by the quality (i.e. number of genotyped SNPs) of the sample. A SNP can therefore obtain a high support either by being genotyped in many low-quality samples, or in fewer high-quality samples. We ranked the SNPs by their support, and identified the value at which 600k SNPs were above it, then used all SNPs with support equal to or greater than that value for our top600k set (N= 599,900). For the purposes of simulating the behaviour of models trained on those SNPs when applied to incomplete data, we further downsampled that SNP list to be the 500k SNPs with the highest support ("top500k"; N = 499,666)

Chapter 4

TRACING GENE REGULATORY CHANGES IN RECENT HUMAN EVOLUTION

4.1 Introduction

In previous chapters, we showed that statistical methods built to predict biological characteristics from large-scale genomic data can be instrumental in filling gaps in evolutionary studies. This is particularly true in the cases where we cannot study those characteristics by another method. While we previously focused on gene regulatory and phenotypic characteristics of Neanderthals and Denisovans, many of the same questions remain unanswered in more recent human evolution.

Within the last several years, the number of ancient DNA (aDNA) samples from anatomically modern humans (AMHs) has increased dramatically[46]. These samples span the globe, and cover time periods from several hundred to tens of thousands of years ago. While this is a potentially rich data source for understanding more recent genetic changes and adaptation that have taken place at the population level since AMHs left Africa, there are several challenges to overcome. First, while the samples are often paired with archaeological information from the site where they were found, this sort of information is limited to what can survive thousands of years in the grounds, which usually does not include soft tissues. Second, due the complexity of many phenotypes and gaps in biological understanding, there is a lack of large-scale methods to draw conclusions about phenotypes based on genetic information alone.

To date, most studies have focused on comparing aDNA from different regions in order to understand where and when people moved[54,56]. Of particular interest has been the population-level shifts from a nomadic hunter-gatherer lifestyle to that of pastoral herding and to stationary agricultural farming. This change had profound implications for multiple aspects of life. While this included changes in day-to-day activities and population density, it also involved substantial dietary shifts, including increasing reliance on domesticated grains[41,111]. This shift is likely to have resulted in changing selective pressures as populations changed circumstances.

Studies done in modern populations have identified key differences in a handful of metabolic traits[112,113]. In addition, scans for signals of recent selection have linked selected regions to changes in specific genes. This task is easier when it involves a protein-coding change, but the vast majority of recent evolutionary change is driven by more subtle effects caused by changes in gene regulation[72,114]. One of the few loci that has been identified in this category is the *FADS1* gene. Increased expression of this gene is associated with an increased ability to metabolize nutrients from grains, and alleles that increase its expression are known to have increased in frequency after populations changed to an agricultural lifestyle[71,115].

However, in many cases, the mechanistic explanation for the observed selection remains poorly understood. For example, the leptin receptor (*LEPR*) is surrounded by a haplotype that has undergone recent positive selection[116], and protein-coding changes have been implicated in increased cold tolerance[117]. However, altered expression of this gene is also associated with altered appetite regulation and metabolism[118,119], and it remains unknown which association is the source of the selective pressure, and by which mechanism[120]. In addition, examples like the previous two, where selection signals can be confidently attributed to specific genes on the basis of known function or mendelian phenotypes, are the exception. In most cases selection peaks span many genes, with little indication which might be the one that underwent changes that affected fitness.

To identify genes whose regulation potentially underwent adaptive changes in response to changes in lifestyle, we applied targeted PrediXcan models to hundreds of ancient humans representing populations from hunter-gatherer, pastoral, and agricultural lifestyles. As well as recapitulating the known pattern of the *FADS1* regulatory haplotype, we identified changes in regulation of *LEPR* among lifestyles, suggesting that its function in metabolism and appetite regulation could have been more relevant for adaptation among this group of individuals. In addition, we show that broader metabolic and immune pathways are enriched for differences between lifestyles, which is potentially reflective of both the altered metabolic requirements and immune pressures. This study is a demonstration of the potential of methods like PrediXcan to shed light on questions of recent evolution.

4.2 Defining an ancient human cohort

We collected ancient human samples that were analysed using a variety of sequencing and genotyping platforms (Methods). Based on the guidelines established in the previous chapter, we ranked individuals by the number of sites successfully genotyped, and took the top quartile of individuals, focusing on Eurasians due to sample availability (Fig. 4.1A). The samples ranged in date from 90 years before present (yBP) to 45,000 yBP, with the majority between 2,500 and 6,000 yBP (Fig. 4.1B).

We then assigned individuals to a lifestyle (hunter-gatherer, pastoralist, or agricultural) based on literature review about the associated archaeological culture. In general, hunter-gatherers were at sites that showed evidence for meat consumption, while pastoralist sites showed artifacts and structures associated with animal domestication. Agriculturalists additionally showed evidence for domesticated grains. This review resulted in 490 ancient individuals with an assigned lifestyle for study (Fig. 4.1C).

4.3 Imputing gene regulatory differences between ancient humans

Previous work has demonstrated that PrediXcan can impute the genetically regulated component of gene expression for thousands of genes, particularly those whose regulatory architecture is dominated by common variants[79]. We trained PrediXcan models in 49 tissues from GTEx v8[110], using only the roughly 700,000 variants that were genotyped in the aDNA samples and were variable in GTEx ("1240k Models" in Chapter 3). We considered accurate (FDR < 0.05m $r^2 > 0.01$) PrediXcan models of autosomal gene regulation.

We then applied these models to our 490 ancient samples, as well as to 503 modern Europeans from the 1000 Genomes Project[87]. This resulted in 210,800 models of Ancient Imputed Regulation ("AIR") for 14,873 unique genes. Because the output of a PrediXcan model is not a direct proxy for gene expression in an individual, differences in PrediXcan values between individuals reflect differences in the genetic regulatory effects of the variants involved in training. Therefore we will

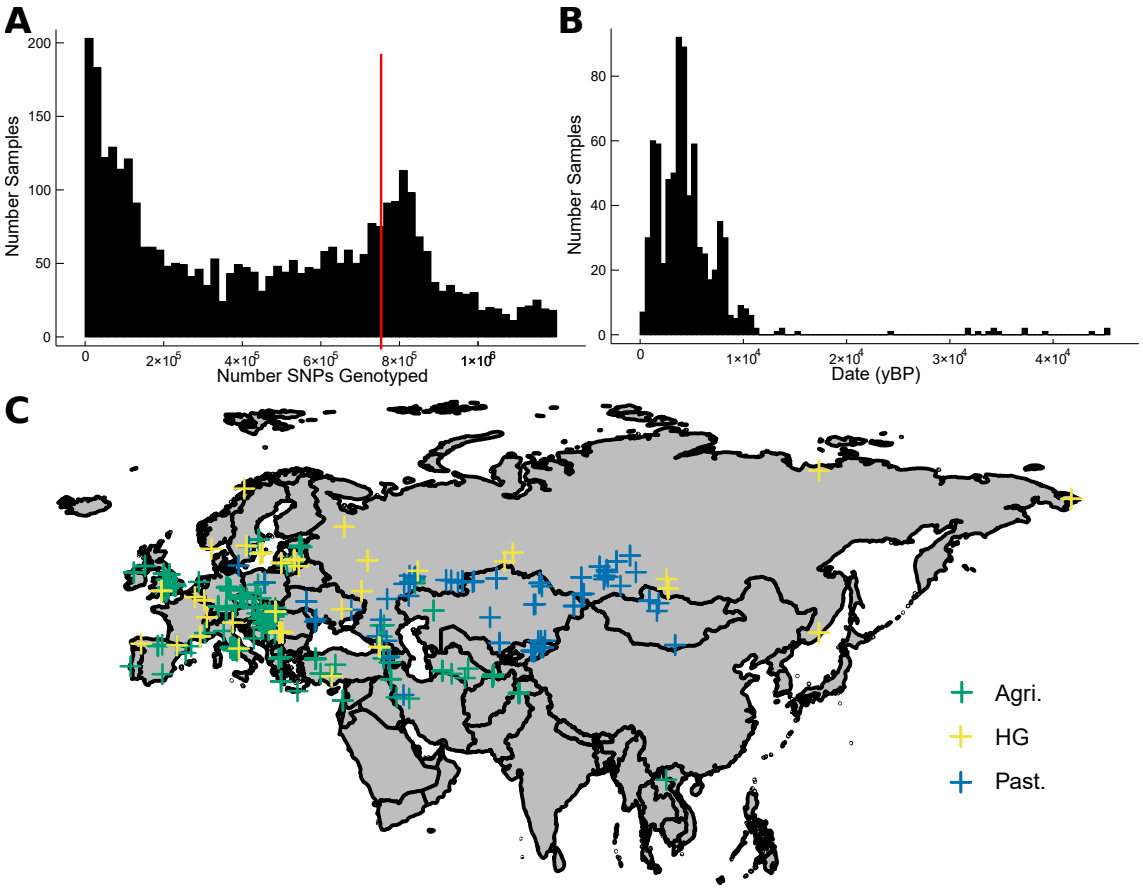

Figure 4.1: Predicting gene regulatory patterns in ancient humans.
(a) Distribution of the number of SNPs with genotype call in the aDNA samples. Maximum is 1233013, which is the number of SNPs on the 1240k genotyping chip (prior to lifting over to hg38). The red line is at the 3rd quartile (771029 SNPs), above which are the samples we focused on. (b) Distribution of the age of eurasian samples in the top quartile in years before present. (c) We analyzed 490 ancient Eurasians from three lifestyles with sufficient genomic data. Green = Agriculturalist, Blue = Pastoralist, Yellow = Hunter-gatherer.

refer to any imputed differences as differences in gene regulation.

4.4 AIR identifies regulatory changes relevant to diet changes

To demonstrate this method's potential, we applied it to genes previously suggested to be involved in shifts in ancient human diets. We first compared the predicted regulation of *FADS1* between agriculturalists, pastoralists, and hunter-gatherers (Methods). *FADS1* showed significant significant differences in 21 tissues. Furthermore, in each tissue, hunter-gatherers had significantly

lower *FADS1* levels than in agriculturalists or modern Europeans, as would be expected from their diets (Fig. 4.2A). 32 ancient Africans follow a similar trend, indicating this is not specific to the Eurasian population (Fig. 4.2B). Unsurprisingly, SNPs driving this pattern are in linkage disequilibrium (LD) with the haplotype implicated in previous evolutionary studies (Fig. 4.2C; Table 4.2)[71]. Overall, *FADS1* AIR is also negatively correlated with the date of the sample (Spearman ρ= –0.32, $P = 1.95 \times 10^{-20}$, which also agrees with known allele frequency trajectories.

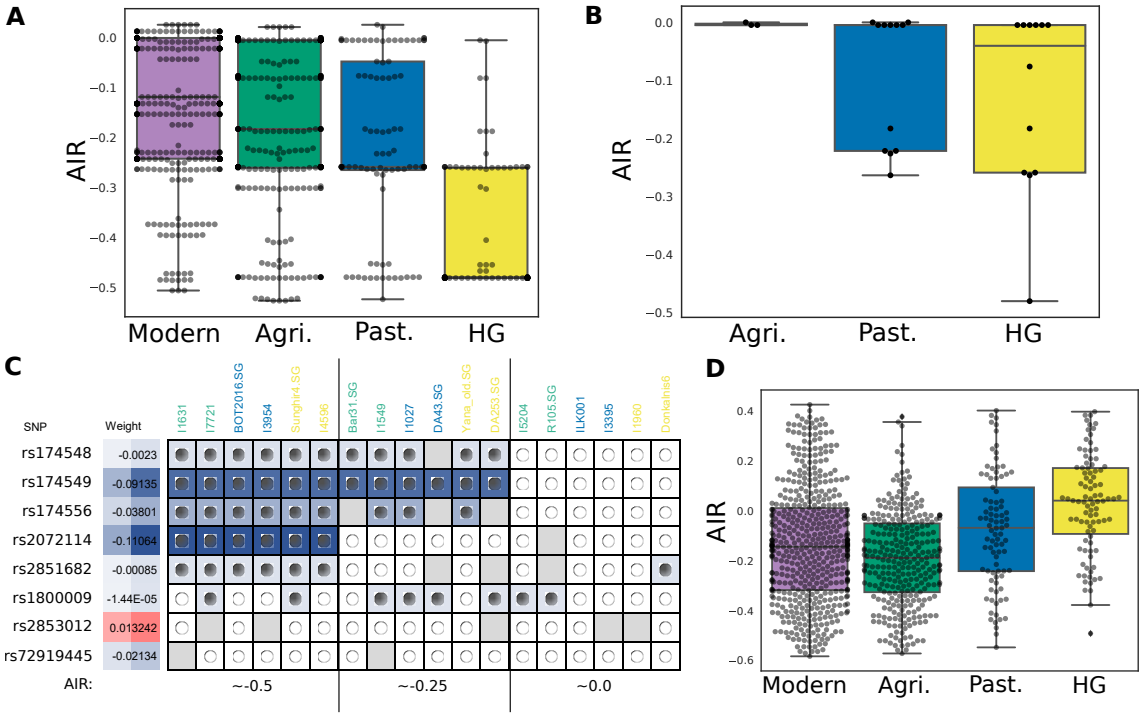

Figure 4.2: Ancient AMHs show significant differences in regulation of key diet genes
(a) *FADS1* shows significant difference in predicted normalized expression in Subcutaneous Adipose tissue by lifestyle (Kruskal-Wallis $P = 5.7 \times 10^{-24}$), as well as in 20 other tissues. (b) 27 ancient Africans follow a similar trend in expression differences to that seen in ancient Eurasian populations. (c) Breakdown of the 8 SNPs in the model of FADS1 in Adipose_Subcutaneous tissue and their presence in representative ancient Eurasians across a range of predicted normalized expression values. Cells are coloured by the weight that SNP contributed to the prediction, while the circles indicate the alleles present in that individual (filled = homozygous effect, empty = homozygous reference). A grey square indicates the SNP was ungenotyped in that individual. The vast majority of ancient samples appear homozygous due to being extremely low-coverage, such that many sites are represented by only a single read. (d) *LEPR* shows a significant difference by lifestyle in Cerebellum (Kruskal-Wallis $P = 3.6 \times 10^{-17}$). Plotted with 503 modern Europeans for comparison. Green = Agriculturalists, Blue = Pastoralists, Yellow = Hunter-Gatherers. AIR = Ancient Imputed Regulation.

Tissue	Agri. AIR	Past. AIR	HG AIR	K-W *P*
Adipose_Subcutaneous	–0.183	–0.259	–0.480	5.71×10^{-24}
Adipose_Visceral_Omentum	–0.132	–0.158	–0.158	1.79×10^{-6}
Brain_Cerebellar_Hemisphere	–0.882	–1.228	–1.422	8.07×10^{-19}
Brain_Cerebellum	–0.887	–1.184	–1.716	4.08×10^{-20}
Brain_Frontal_Cortex	–0.391	–0.602	–0.690	1.06×10^{-21}
Brain_Putamen_basal_ganglia	–0.360	–0.469	–0.530	2.51×10^{-6}
Cells_Cultured_fibroblasts	–0.712	–0.768	–1.024	7.57×10^{-9}
Colon_Sigmoid	–0.360	–0.456	–0.859	1.26×10^{-21}
Esophagus_Gastroesophageal_Junction	–0.378	–0.426	–0.579	2.27×10^{-21}
Esophagus_Mucosa	–0.620	–0.769	–0.836	4.87×10^{-10}
Esophagus_Muscularis	–0.319	–0.409	–0.506	1.47×10^{-9}
Heart_Atrial_Appendage	–0.225	–0.281	–0.382	6.03×10^{-9}
Heart_Left_Ventricle	–0.169	–0.330	–0.924	2.52×10^{-26}
Lung	–0.106	–0.154	–0.237	2.56×10^{-25}
Muscle_Skeletal	–0.410	–0.674	–0.787	3.24×10^{-28}
Nerve_Tibial	–0.560	–0.748	–0.868	2.84×10^{-9}
Pancreas	–0.452	–0.882	–1.414	4.40×10^{-18}
Stomach	–0.401	–0.641	–0.872	8.00×10^{-27}
Testis	–0.610	–0.774	–0.862	1.09×10^{-13}
Thyroid	–0.295	–0.483	–0.649	1.27×10^{-21}
Whole_Blood	0.0914	0.108	0.132	7.57×10^{-6}

Table 4.1: *FADS1* models with significant differences by lifestyle
AIR (Ancient Imputed Regulation) given as the median of that group. *FADS1* was modeled in an additional 9 tissues.

We next imputed the regulation of *LEPR*, another gene with nearby signatures of selection potentially relevant to its function in appetite or cold tolerance, across ancient individuals from the three lifestyle groups. *LEPR* was significantly divergently regulated between groups in the cerebellum (Fig. 4.2D) (the only brain tissue with a model for *LEPR*), both adipose tissues, and several other tissues. In each tissue, it was consistently predicted to be downregulated in Agricul-

Model SNP rsID	Weight	Haplotype	Tag SNP	r^2
rs2072114	–0.111	B	rs174546	0.340
rs2072114	–0.111	C	rs102274	0.329
rs2072114	–0.111	D	rs174576	0.421
rs174549	–0.0914	B	rs174546	0.918
rs174549	–0.0914	C	rs102274	0.892
rs174549	–0.0914	D	rs174576	0.673
rs174556	–0.0380	B	rs174546	0.914
rs174556	–0.0380	C	rs102274	0.889
rs174556	–0.0380	D	rs174576	0.675

Table 4.2: *FADS1* model SNPs LD with established haplotype.
The 3 highest-weight SNPs from Subcutaneous Adipose. Haplotype labels correspond to those in Mathieson & Mathieson (2018)[71].

turalists compared to the other two groups. Given that its function in appetite regulation primarily involves leptin signalling between adipose tissues and the brain, this suggests a possible mechanism behind the observed selection signal. This is particularly intriguing given the association of decreased *LEPR* function with obesity and metabolic disorders[121,122]. Collectively, these results demonstrate the potential for the imputation of gene regulation in ancient samples to both identify genes for which allele frequency shifts likely resulted in population-level changes in regulation, and to suggest mechanistic explanations for previous observations of selection.

4.5 Housekeeping genes are enriched among genes with differences

Encouraged by the identification of genes with previous support for being influenced by the shift in diets, we next explored whether PrediXcan could identify trends in the regulation of sets of genes with functions or evolutionary histories potentially relevant to the lifestyle shifts. First, we compiled a set of genes that have experienced stabilizing selection on their levels of gene expression across many species[123]. Given such long-term stability, we expect these to maintain

Tissue	Agri. AIR	Past. AIR	HG AIR	K-W *P*
Adipose_Subcutaneous	–0.504	–0.374	–0.325	4.68×10^{-9}
Adipose_Visceral_Omentum	–0.251	–0.0661	–0.0555	2.87×10^{-15}
Brain_Cerebellar_Hemisphere	0.00454	0.199	0.317	5.87×10^{-16}
Brain_Cerebellum	–0.191	0.0699	0.0395	3.62×10^{-17}
Esophagus_Gastroesophageal_Junction	–0.0109	0.0443	0.174	3.73×10^{-13}
Heart_Atrial_Appendage	–0.136	–0.0663	–0.0537	4.80×10^{-16}
Testis	–0.246	–0.0970	–0.0659	1.77×10^{-20}
Whole_Blood	0.0559	0.211	0.178	6.15×10^{-6}

Table 4.3: *LEPR* models with significant differences by lifestyle
AIRs given are medians. *LEPR* was modeled in an additional 11 tissues.

similar regulatory patterns across population, regardless of lifestyle. Indeed, these selected genes show no enrichment for significant differences between lifestyle groups across tissues (Table 4.4). Similarly, we next analyzed genes that are intolerant to loss-of-function coding variation in modern humans[124] (Methods). We predicted that regulatory variation for genes with such strong coding constraint in modern populations might also be unfavourable. As expected, these genes were significantly depleted among the genes with significant differences across the ancient groups (OR = 0.89, 95% CI [0.8131,0.9833]; Fisher's Exact $P = 0.021$).

The last set of genes we explored were housekeeping genes[125]; given their importance to fundamental cellular processes, we hypothesized that they too would be depleted among the significantly divergently regulated genes. To the contrary, they were significantly more likely to show a significant difference across the lifestyle groups in at least 1 tissue (OR = 1.14) than all genes overall. By definition, housekeeping genes tend to have robust expression in all tissues, so this pattern could partially be explained by an increased power to model changes in their regulation in multiple tissues. However, many housekeeping genes are also involved in basic cellular metabolism[126], which could require fine tuning in response to changes in nutrient sources or other environmental shifts. Together, these results demonstrate the ability of this method to highlight particular gene sets of

interest, and suggest targets for future studies.

Gene Set	N Modeled	% Significant	OR [95% CI]	Fisher's Exact *P*
Stabilizing Selection	4519	40.6	0.84 [0.6648, 1.0831]	0.19
LoF-Intolerant	2195	36.5	0.89 [0.8131, 0.9833]	0.021
Housekeeping	3127	41.2	1.14 [1.0524, 1.238]	0.0014

Table 4.4: Odds ratios for gene sets
Overall, 38.7% of genes showed a significant difference in AIR in at least one tissue (per-tissue Bonferroni correction). The odds ratio was calculated as the odds of a gene's presence in the category given it being a significant gene. Not all genes were tested in the stabilizing selection analysis, so the OR only included those that were (44.6% of tested, but unselected, genes showed a difference).

4.6 Genes divergently regulated between lifestyle groups are enriched for immune and metabolic functions

While the previous analyses targeted specific genes or gene sets of interest, our approach can also generate hypotheses through genome-wide analyses. Overall, 5759 genes showed significant divergent regulation between lifestyles in at least 1 tissue (median 2 tissues; Fig. 4.3A), and an average of 9.8% of genes in each tissue were divergent (Fig. 4.3B). Most significantly divergent genes had relatively small changes in magnitude by group (i.e. maximum 1.17 magnitude difference between hunter-gatherers and agriculturalists in Subcutaneous Adipose), and distributions generally overlapped, suggesting that, as is the case with *FADS1* (Fig. 4.2A&C), these differences are often driven by variation in allele frequencies between groups, not complete presence/absence of alleles.

We next tested for overall patterns in the function of genes that showed divergent regulation in specific tissues. In each tissue, we conducted over-representation analysis to identify the groups of genes most over-represented among the significant genes. The most-enriched term in Whole Blood was response to leptin (6.56x; Table 4.5). This category included both *LEPR*, which was the subject of a targeted analysis above, as well as two other genes in that pathway (*SIRT1* and *BBS2*). The top

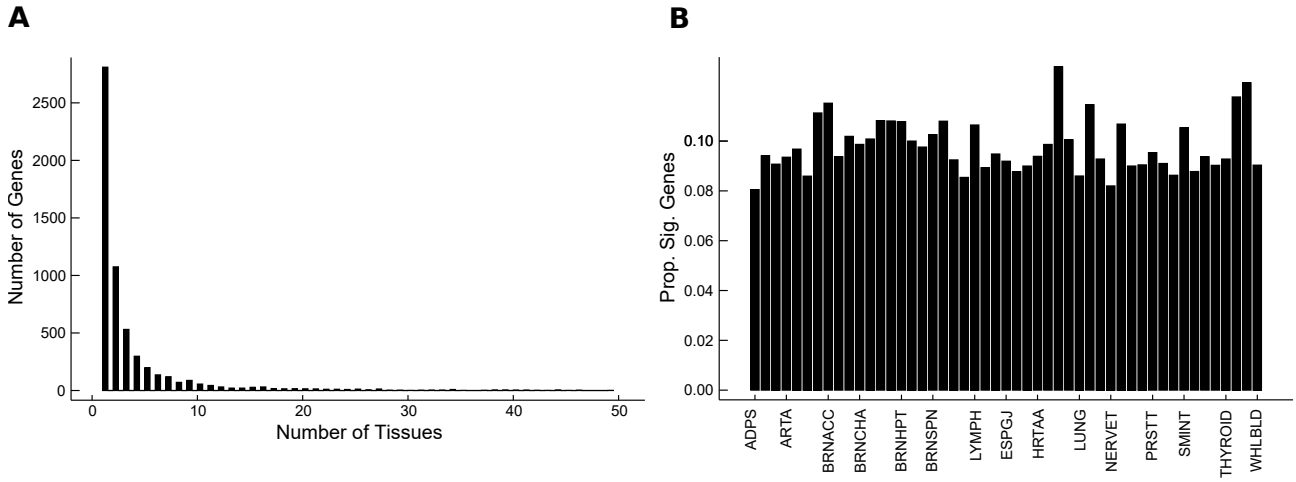

Figure 4.3: Thousands of genes are divergently regulated between lifestyle groups. (a) Distribution of the number of tissues in which each gene is significantly different between lifestyles. (b) The proportion of significant models out of all models in a tissue. Tissues are in alphabetical order: **Adipose_Subcutaneous**, Adipose_Visceral_Omentum, Adrenal_Gland, **Artery_Aorta**, Artery_Coronary, Artery_Tibial, Brain_Amygdala, **Brain_Anterior_cingulate_cortex**, Brain_Caudate_basal_ganglia, Brain_Cerebellar_Hemisphere, **Brain_Cerebellum**, Brain_Cortex, Brain_Frontal_Cortex, Brain_Hippocampus, **Brain_Hypothalamus**, Brain_Nucleus_accumbens_basal_ganglia, Brain_Putamen_basal_ganglia, **Brain_Spinal_cord_cervical_c-1**, Brain_Substantia_nigra, Breast_Mammary_Tissue, Cells_Cultured_fibroblasts, **Cells_EBV-transformed_lymphocytes**, Colon_Sigmoid, Colon_Transverse, **Esophagus_Gastroesophageal_Junction**, Esophagus_Mucosa, Esophagus_Muscularis, **Heart_Atrial_Appendage**, Heart_Left_Ventricle, Kidney_Cortex, Liver, **Lung**, Minor_Salivary_Gland, Muscle_Skeletal, **Nerve_Tibial**, Ovary, Pancreas, Pituitary, **Prostate**, Skin_Not_Sun_Exposed_Suprapubic, Skin_Sun_Exposed_Lower_leg, **Small_Intestine_Terminal_Ileum**, Spleen, Stomach, Testis, **Thyroid**, Uterus, Vagina, **Whole_Blood**.

ten terms also included glycoprotein and aminoglycan metabolic processes, as well as more basic cellular functions. The presence of more housekeeping-like functions suggests that the enrichment observed in the previous analysis is not entirely due to power differences. On the other hand, the top ten terms in Subcutaneous Adipose included several immune-related functions, including response to interferon-gamma, as well as genes involved in transporting organophosphate esters and other xenobiotic factors (Table 4.6). Differences in these genes related to interacting with the environment are more likely related to differences in living conditions between the groups than specifically related to diet.

We next tested for patterns in the function of genes that showed divergent regulation overall.

Table 4.5: Top 10 enriched GO terms in Whole Blood.

GO term	**Num. Genes**	**Enrichment**	***P***
response to leptin	3	6.56	0.0066
protein hydroxylation	5	3.91	0.0062
multivesicular body sorting pathway	4	3.65	0.019
neuromuscular process	6	2.98	0.012
cytokinesis	9	2.19	0.019
aminoglycan metabolic process	11	2.07	0.014
negative regulation of cell activation	11	2.07	0.014
glycoprotein metabolic process	22	2.02	9.3×10^{-4}
carbohydrate derivative catabolic process	12	1.90	0.021
regulation of protein stability	13	1.90	0.017

We conducted gene set enrichment analysis using biological process GO terms on significant genes ranked by the number of tissues they were significant in. The majority of the positively-associated terms were housekeeping functions (cell cycle regulation, metabolic processes) (Appendix 2, Table 6.5), recapitulating the targeted analysis in the previous section. Antigen processing and presentation, which includes 5 HLA genes that showed significant regulatory differences in as many as 44 tissues, was also enriched among genes with significant differences in many tissues.

4.7 Discussion

In this study, we applied PrediXcan models of gene regulation from 49 tissues to hundreds of ancient humans from three different lifestyles—Hunter-Gatherers, Pastoralists, and Agriculturalists. We found that over 5,000 genes showed evidence for divergent regulation among the three groups in at least one tissue. Our results recapitulated known instances of altered gene expression relevant to population changes diet and lifestyles, such as *FADS1*, and they also suggested explanations for previously observed signals of selection on *LEPR*. Altered expression of *LEPR* in these

Table 4.6: Top 10 enriched GO terms in Subcutaneous Adipose.

GO term	**Num. Genes**	**Enrichment**	***P***
DNA-templated transcription, elongation	9	4.41	9.6×10^{-5}
multi-organism cellular process	9	4.25	0.0045
secretion by tissue	3	4.25	0.028
peptide catabolic process	3	3.82	0.038
interstrand cross-link repair	6	3.48	0.0057
response to interferon-gamma	14	2.62	6.4×10^{-4}
antigen processing and presentation	12	2.51	0.0023
organophosphate ester transport	7	2.35	0.026
regulation of cell division	7	2.23	0.034
negative regulation of defense response	9	2.05	0.029

groups, as well as the regulatory divergence seen in two other genes in the leptin processing pathway (*SIRT1*, *BBS1*), suggest that appetite and adipose regulation may be relevant for adaptation in these groups in addition to the connection to cold tolerance. While housekeeping genes were enriched among significantly divergent genes, in part because of the increased power to model them, overall genes were enriched for metabolic and immune processes, indicating that these pathways are the most likely to have been affected by altered gene regulation during recent human evolution.

The frequent occurrence of metabolic and immune genes is partly to be expected. The shift from nomadic hunting and gathering to stationary farming brought many changes in lifestyle. Diet is perhaps the most striking, and population-level shifts in gene expression would likely been required to optimize metabolism of the differing types and amounts of nutrients available. Similarly, population density and interactions with other individuals and species changed as well, making it likely that human immune systems had to adapt to more and different pathogens over the thousands of years covered by this study.

Many studies have found examples of selection acting on variation that was present at low lev-

els in populations, influenced only by genetic drift, for a long time prior to selection beginning to act[71,127,128]. Our study supports this scenario in many cases by the overlapping distributions of predicted gene expression; while there are group-level shifts, the range of the predictions is fairly constant. Because this method focuses on the genetically-regulated component of gene expression, it cannot distinguish between cases where the actual expression changed and cases where the regulatory architecture turned over. Downstream analyses on highlighted genes and functions will be necessary to distinguish these scenarios and identify the functional variants.

Overall, this study demonstrates the power of methods trained to predict molecular phenotypes, such as gene expression, to study evolution by focusing on variants with a demonstrated relationship to those phenotypes. Our approach is well-positioned to take advantage of the increasing availability of modern and ancient genome data to provide both mechanistic explanations of selection signals observed in other studies and to generate hypothesis about phenotypic differences between ancient and modern groups. While this study focused on one specific question about the gene regulatory shifts in response to changes in lifestyle, similar methods could be applied in many other questions and sets of ancient samples. Given the importance of gene regulation in recent evolution, this is an important step in identifying candidate regions that have been shaped by recent human evolution. Further analyses will contribute to understanding the genome's response to large-scale environmental changes and the impact of these changes on humans today.

4.8 Methods

Human Genome Samples. We obtained ancient human genotypes from a set compiled and analyzed by the Reich lab (v42.4; accessed March 1, 2020). We filtered out any that did not pass their QC procedure, then ranked the samples by genotype count (i.e. the number of SNPs with a genotype call in that sample). We then manually assigned lifestyle by literature review based on archaeological information about the site and previous research about the associated culture. We then filtered samples by their continent of origin, and primarily focused on 490 ancient Eurasians. For a modern comparison, we used 503 European samples from the 1000 Genomes Project[87].

PrediXcan Models. Models were trained on whole genome sequencing and RNA-seq data from GTEx v8 in 49 tissues using $\tilde{1}$,240,000 SNPs that were genotyped by first enriching for those targeted SNPs ("1240k set")[48,129]. For each tissue, we filtered models to those that explained a significant amount of variance (FDR < 0.05, $r^2 > 0.01$). In addition, we filtered models such that we only included those that high correlations ($r > 0.5$) on predictions of all 2504 1kG individuals when trained on these 1240k SNPs vs. all available SNPs (see Chapter 3 for more details). We calculated LD between variants in all 1kG Populations using LDLink[130].

We abbreviate the 49 tissues considered as follows: Adipose - Subcutaneous: ADPS, Adipose - Visceral Omentum: ABPV, Adrenal Gland: ADRNLG, Artery - Aorta: ARTA, Artery - Coronary: ARTC, Artery - Tibial: ARTT, Brain - Amygdala: BRNAMY, 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, Brain- Spinal Cord Cervical C-1: BRNSPN, Brain- Substantia Nigra: BRNSN, 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, Kidney Cortex: KDNY, Liver: LIVER, Lung: LUNG, Minor Salivary Gland: MNRSG, 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.

Identifying significant differences in predicted expression. To identify genes with significant differences in predicted gene expression between the three lifestyle groups, we conducted a Kruskal-Wallis test for each gene model. To account for multiple testing, we used Bonferroni correction within each tissue. Genes that pass that correction in at least 1 tissue are said to show evidence for

a significant difference in regulation. We did not correct for the number of tissues tested, because, while gene expression can be correlated across tissues, those patterns remain poorly understood.

Gene Set Enrichment Analyses. For the targeted gene set enrichment analyses, we used three gene sets; 1) genes whose expression in particular tissues is under stabilizing selection across 17 mammalian species[123]; genes that are intolerant to loss-of-function variants in their protein products(called if the upper bound of the 95% confidence interval of the observed/expected ratio is lower than 0.35)[124]; and 3) housekeeping genes that show consistent expression across tissues[125]. We calculated an odds ratio for each, and used a Fisher's exact test to determine significance. For the genes under stabilizing selection on gene expression, we considered only those tested in that study before calculating statistics.

To conduct functional enrichment analyses on the full set of significant genes, we took two approaches. For the first, we ranked all genes by the number of tissues in which they show significant regulatory differences, then conducted gene set enrichment analysis using WebGestalt with default parameters[131]. For the second, we focused on specific tissues. For each one, we analyzed all genes with significant divergence in that tissue between the lifestyles, then conducted over-representation analysis with WebGestalt under default parameters. For both we used the annotations for the biological process GO terms.

Chapter 5

CONCLUSIONS AND FUTURE DIRECTIONS

The decades since the human genome was completed have been fruitful for understanding the impact processes of evolution have on human genetics and phenotypes. However, the genome is a big, complicated place, and much is still unknown. While we can identify regions of the genome that appear to have been subject to selection, we can rarely explain the mechanisms and reasons behind those signals. In addition, it has been difficult to confidently identify variants with effects on evolutionarily-relevant phenotypes genome-wide. Specifically, while we know that changes in gene regulation have affected recent evolution[114], we lack a genome-wide, functionally-based picture of what genes likely have been affected. In this dissertation, I described the studies we undertook to fill these gaps in knowledge. We focused on a statistical method called PrediXcan, which is trained to predict RNA-seq data based on allele counts of nearby genetic variants[79], in order to study combinations of variants and their effect on an intermediate phenotype. This allowed us to identify genes and pathways likely influenced by changes in gene expression during recent evolution.

In Chapter 2, we used this framework to study gene regulatory differences between Anatomically Modern Humans ("AMHs") and two closely-related species: Neanderthals and Denisovans ("archaic hominins"). Initially we focused on the evolutionary dynamics of Neanderthal genomic regions that had been removed from AMH genomes since introgression took place. Study of these regions is complicated to do directly, since they no longer exist in extant cells, making it difficult to distinguish between ones that were evolutionarily neutral and removed due to genetic drift, and those that might have been deleterious and selected against. We identified thousands of genes whose divergent regulation in Neanderthals, and potential phenotypic impact of altering dosage of those genes, suggested them as candidates for being in the latter category. In order to identify systems and phenotypes that may have differed in the archaic hominins more broadly,

we also compared regulatory characteristics between them in general. We identified several categories, including morphological traits and the immune system, for which many genes showed large differences in regulation. Some of these were consistent with the limited information in the archaeological record, but many were novel, which is particularly intriguing in the case of the Denisovan, given the lack of information about that group. These results established differences in gene regulatory architecture between AMHs and archaic hominins that will be interesting to follow up on in the future, and demonstrated the potential for exploring phenotypic differences between archaic groups from genomic information alone.

However, genomic information is not always of the best quality, particularly when derived from ancient samples. In Chapter 3, we examined PrediXcan's ability to be applied on low-coverage genotyping data. We simulated low-coverage genomes in a variety of ways and characterized the models' behaviour under those conditions. To study the flexibility of this framework, we additionally trained new models with several different variant sets. Overall our results demonstrated that in many cases PrediXcan models maintain reasonable performance even when applied to low-coverage genomes. However, there were problems when there was a large mismatch between the variant set available during model training and that available in the application dataset. It is therefore advisable to retrain models using specifically the variants that will be used in downstream applications. While this means it is impossible to train a set of models that would be useful in all situations, overall these results demonstrate that this framework is flexible enough to applied in many different contexts. Additionally, they demonstrate several types of analyses that are useful for focusing on gene regulation models that will work well in a given situation.

In Chapter 4, we applied what we learned in Chapter 3 to study differences in gene regulation that exist between ancient AMH groups from three different lifestyles: hunter-gatherers, pastoralists, and agriculturalists. While a few genes are known to have regulatory differences between these groups[71,132], genome-wide tests for differences had not been done. We identified thousands of genes that showed a significant difference in regulation, many of which are involved in metabolism and the immune system. These included a gene, *LEPR*, for which there was an

observed signal of recent selection without a mechanistic explanation. Follow-up studies will be informative for understanding changes in response to diet shifts, and which were in response to changes in activity or population density. Overall, this study demonstrated the utility of this framework both for suggesting explanations for known patterns of genome variation, and for suggesting phenotypes and pathways that may have differences between populations.

5.1 Contextualizing our results

Overall, our results provide additional support for known hypotheses about particular genes and pathways being altered during recent evolutionary history, and identify new candidate regions that could be under some form of selection on their regulation. Several studies in modern humans had implicated brain functions as being influenced by introgressed Neanderthal variants[62,66,63,67], and it is not terribly surprising that genes affecting the brain were included among those divergently regulated by regions that did not survive in AMH genomes. However, particularly in the case of genes in introgression deserts, our results narrow down the list of potential genes that might be interesting to study further, and suggested that, among non-introgressed regions of the genome, introgression deserts are remarkable more for their unusually low recombination rates, not due to having multiple important divergently regulated genes. Under this model, the low recombination rate at deserts causes selective pressure to be applied over a larger area of the chromosome than in regions with higher recombination rate. Similarly, the divergent regulation of genes involved in reproductive phenotypes ties in well with the hypothesis of there being barriers to introgression related to reproductive fitness[133,134], and provides additional loci to examine.

Findings related to phenotypic differences between archaic hominins are exciting and potentially novel, but are also more difficult to confirm. As the immune system is among the fastest-evolving systems, and there are many population-level differences in it in AMHs[135], identifying differences in the regulation of immune genes between different Neanderthal populations, although novel, is not surprising. However, given that population-level data for Neanderthals is increasing[81,136], it will be exciting to follow up with more individuals. In the case of Denisovans, there

is so little physical evidence about them that findings about their phenotypes are simultaneously the most exciting and the most difficult to follow up on. Skeletal differences between AMH and Neanderthals are well-established[137], and there has been a general assumption that Denisovans were similar based on the small number of fragments found[138]. A recent study done by predicting differentially methylated regions in Denisovan bone suggested specific morphological characteristics in which they may have differed from both AMHs and Neanderthals[139]. Some of those traits overlap with those affected by genes that we identified as having Denisovan-unique regulation, and in the future it will be informative to explore those more.

The proof-of-concept analysis of regulatory patterns in different AMH lifestyles provided support for pre-existing hypotheses and suggested several new ones. Studying genetic responses to diet changes is of particular interest in modern populations because of the prevalence of metabolic disorders such as obesity[140], and the understanding the evolutionary history of those traits could be informative, or help identify candidate regions to explore in more detail. In the case of *LEPR*, we demonstrated the ability of our method to provide explanations for signals of selection already observed[116]. More broadly, our analyses highlighted metabolic and immune genes, which is what we expected to find. However, it might be informative to further dissect the specific genes involved to gain a more-detailed picture of how specific pathways might have changed. Collectively, these show the utility of PrediXcan as a tool for understanding genome-wide changes in recent evolutionary history.

5.2 Limitations of PrediXcan

Despite PrediXcan's strengths, a major limitation is that it does not directly model gene expression. Rather, even in the best case scenario where a model includes all variants involved in affecting a gene's expression with the correct effect sizes, PrediXcan models the genetically-regulated component of gene expression. This quantity is capped by the heritability of that gene's expression, and does not take into account environmental influences. These include both responses to physical environment such as temperature, variables such as age or developmental stage, and also non-

sequence-based modes of regulation like DNA methylation. In addition, it focuses primarily on effects driven by common variants. In our case, particularly when applying it to archaic hominins, there is some proportion of models are missing relevant variants that were not in the training population in enough numbers to be modeled accurately. This is because the training population is primarily made up of one AMH population (85% European, 12% African-American)[110], which is not necessarily representative of other populations, and does not include variants that are specific to Neanderthals or Denisovans. Therefore, in some applications of PrediXcan, we are working with an incomplete model of genetic regulation.

However, we believe this method can be informative despite its limitations. The molecular machinery and genetic architecture of gene regulation are largely conserved across humans, and most common human regulatory variants have similar effects across populations[97,98]. While Neanderthals and Denisovans are separate species, they were related to AMHs closely enough to have interbred at least moderately successfully, so we do not expect many genes to have drastic differences in their mechanisms of regulation. Furthermore, with the aid of collaborators, we showed that PrediXcan models are capable of predicting accurate regulatory effects for genes affected by Neanderthal variants, even prevented from seeing Neanderthal variants during training[104]. Therefore we expect most of the limitations of this framework to be driven by the inherent difficulty in trying to predict a trait influenced by multiple variables based on only one.

These issues complicate the interpretation of differences identified by these models. At its most basic, a predicted difference between two individuals in a model indicates that the variants in the model are combining to do something different to the regulation of that gene. The most straightforward result of those differences would be a change in the expression of that gene between the two. Alternatively, there could be additional, unmodeled, variants in one individual that compensate for the changes occurring in the modeled variants. This latter option might be indicative of turnover in regulatory regions such as enhancers, which is a relatively common phenomenon[141]. This is impossible to disentangle without additional, detailed lines of evidence, and also means that we cannot put a direction on our hypothesis of selection acting on some of these regions. In the first

case, predicted differences could indicate the influence of directional selection changing the gene's expression, while in the second there could be stabilizing selection acting to ensure expression is maintained despite the shifting regulatory landscape. Both cases involve altered regulation (albeit with different results), and would be interesting to explore in more detail in downstream studies.

5.3 Future directions

There are several avenues that could be explored that could increase the ability of frameworks like PrediXcan to model gene expression more directly. Different types of statistical models might prove able to give a more accurate prediction of gene expression values. For example, deep learning models have proved capable of predicting complicated states of regulatory regions[142,143]. However, what these models gain in performance, they often lose in ready understanding of how they got their predictions. A more fruitful avenue of improvement might therefore be improving the training data. Datasets from more diverse populations that pair genotype and gene expression data would increase the number of variants represented in the training data, and minimize concerns about cross-population generalization. In addition, it could be worth exploring the use of additional biological information to aid predictions. This could take the form of taking greater advantage of cross-tissue correlations in gene expression to increase power[144], or by including non-sequence information such as methylation state. However the downside of including types of information in the models is that the application dataset would then need that information as well. As we learn more about the structure of regulatory landscapes and enhancers, it might be exciting to incorporate some of that information to help weight or prioritize variants, rather than including entirely separate variables.

No matter how good a model of gene expression is, however, it will gain the most power when combined with other downstream analyses. Recent developments in Massively-Parallel Reporter Assays and CRISPR-based mutagenesis mean that it should soon be (or already is) possible to test hypotheses about the behaviour of specific regulatory regions and genes at a large scale[145,77]. As developments happen in that realm this will become possible dissect much more complicated

combinations of variants to understand exactly how they impact gene expression.

As a field, paleogenomics is in an exciting phase of expansion, as the number and diversity of available aDNA samples is rapidly increasing. While, we were able to analyze a handful of ancient Africans, it will be exciting to have large enough sample sizes to extend analyses of gene regulation and selection outside Eurasia. Of particular interest would be studying which patterns of differences, for example those observed between the different lifestyles we studied in ancient Eurasians, might be universal to the transition in lifestyle as opposed to unique to specific locations and situations. Studying region-specific effects might also inform the extent to which the specific environment influences some of these regulatory shifts. These could be connected to longitudinal studies of populations who have recently made large lifestyle shifts[146,147,148], to understand some of the time scales on which environmental vs. genetic effects occur. In addition, there are increasing amounts of paleoclimatology data becoming available[149,150,151], which could be very informative when intersected with aDNA data to understand population shifts and genetic responses to specific environmental changes.

In summary, we've taken a step toward understanding genome-wide changes in gene regulation over recent human evolution. We've contributed hypotheses about specific genes that can be followed up on, and lead to greater understanding of differences between AMHs and close relatives, as well as AMH responses to large lifestyle changes. This will be useful both in understanding our history, and predicting what effect future events could have on modern populations.

Chapter 6

APPENDICES

6.1 Appendix 1

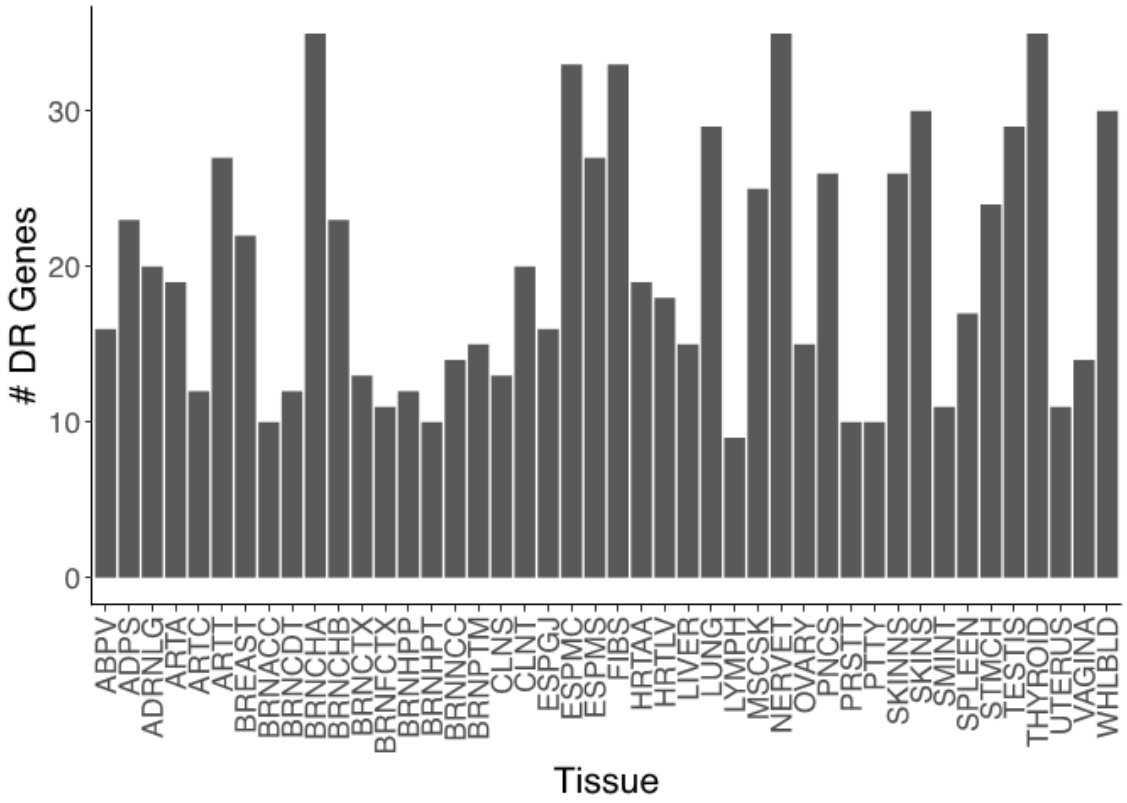

Figure 6.1: The number of DR GWARRs found in each GTEx tissue.
We caution against direct comparisons of the number of DR GWARRs in each tissue due to differences in power resulting from variation in sample size, genetic architecture, and expression levels across tissues. See Methods for tissue abbreviations.

6.2 Appendix 2

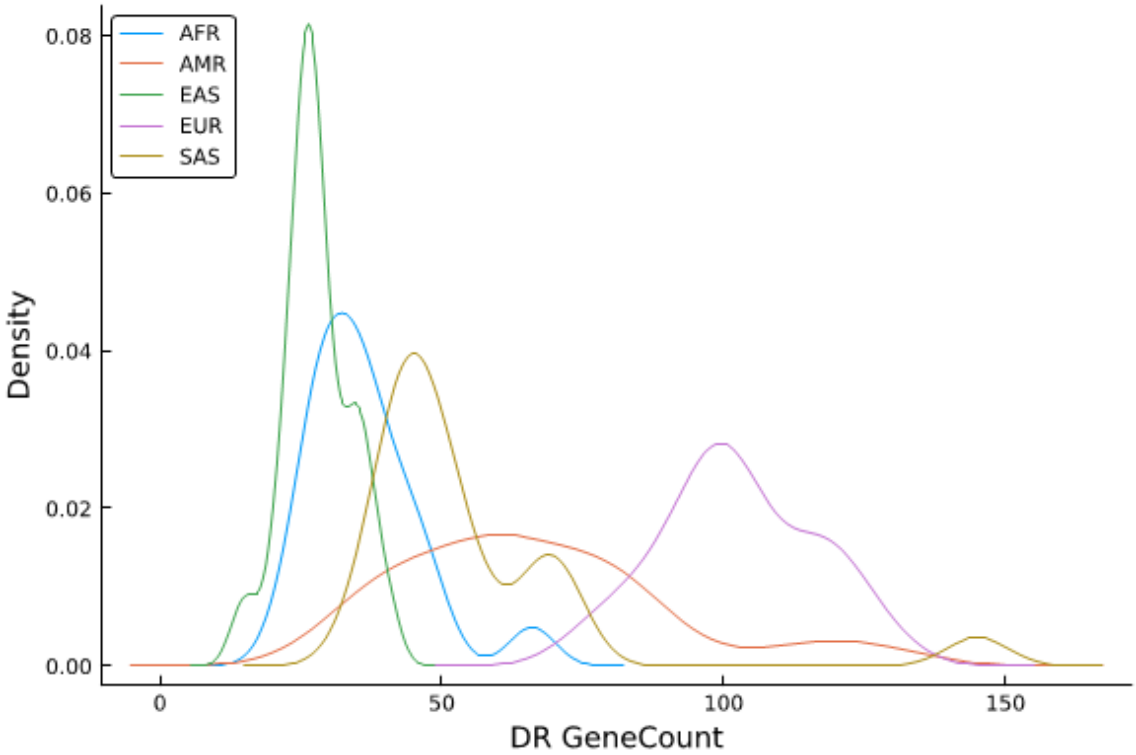

Figure 6.2: Distributions of the number of DR genes found in 50 random humans from 1kG. For 50 random individuals from the 1kG cohort (10 from each continental population), we counted the number of unique DR genes found across any of the tissues considered. Europeans have the largest number of DR genes. 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.

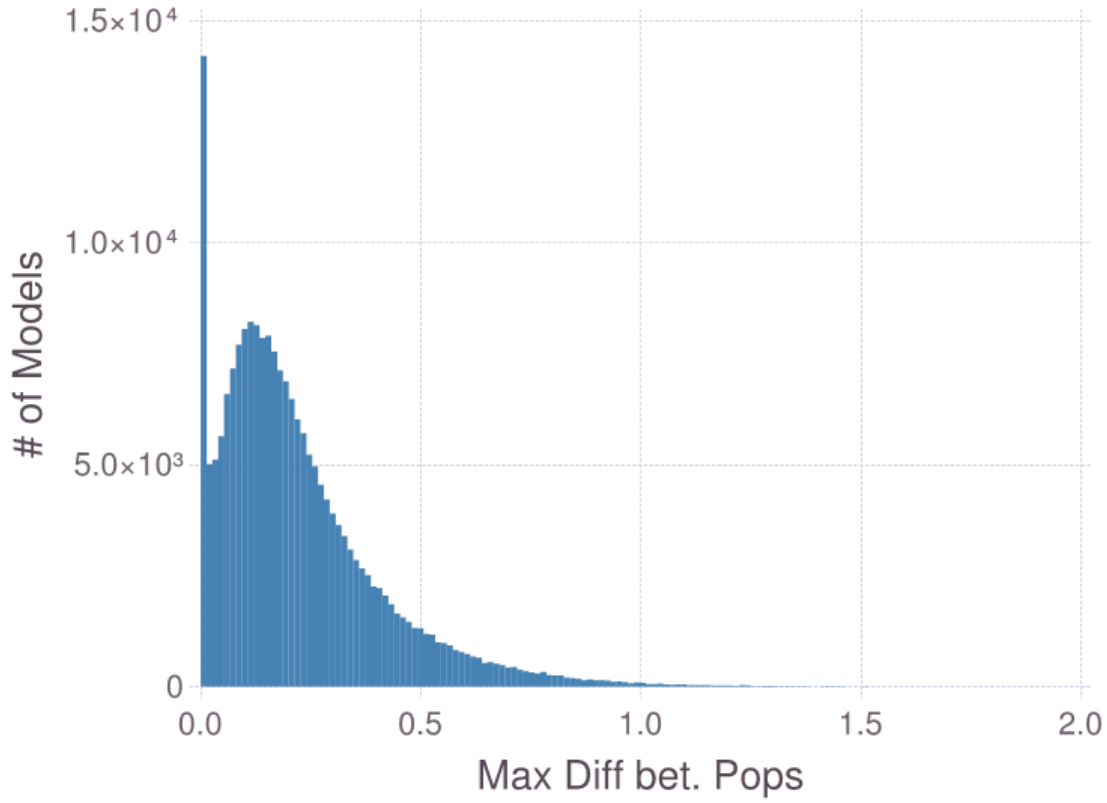

Figure 6.3: Distribution of the maximum difference in the median imputed regulation between 1000 Genomes populations for all PrediXcan models.
Very few models have large predicted regulation differences between populations.

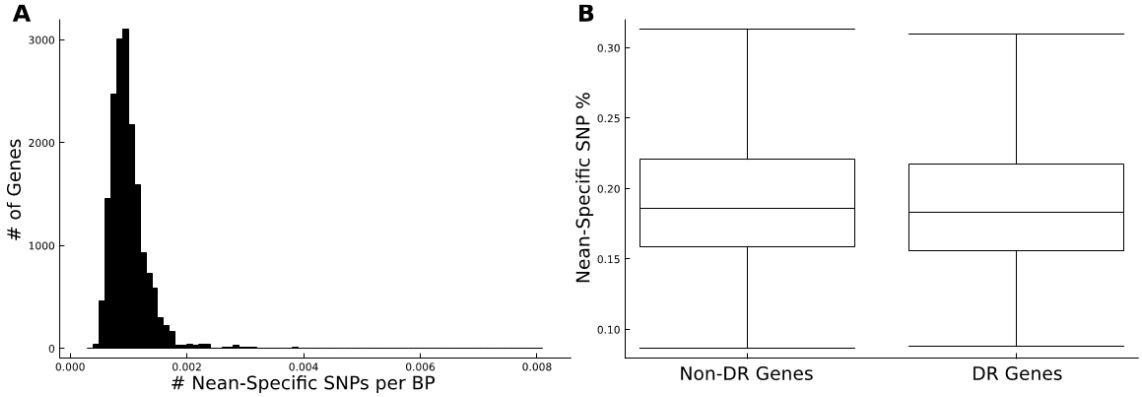

Figure 6.4: Neanderthal-specific variant density in gene regulatory regions.
(A) Density of Neanderthal-specific variants in the regulatory regions of genes. (B) The percentage of Neanderthal-specific variants out of all variable sites (observed in humans, Neanderthals, or both) in a gene region is similar 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.

Table 6.1: No tissues were significantly depleted or enriched for Neanderthal upregulation compared to the overall proportion of upregulated genes (0.43). See 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

Table 6.2: HPO phenotypes enriched in DR genes common to all archaic hominins.

gene set	description	Num. DR Genes	OR	*P*-value
HP:0005736	Short tibia	4	7.146	0.0017
HP:0004691	2-3 toe syndactyly	6	4.149	0.0026
HP:0003330	Abnormal bone structure	31	1.621	0.0034
HP:0001650	Aortic valve stenosis	6	3.783	0.0042
HP:0001007	Hirsutism	10	2.614	0.0042
HP:0006498	Aplasia/Hypoplasia of the patella	5	4.288	0.0051
HP:0002205	Recurrent respiratory infections	26	1.630	0.0071
HP:0001712	Left ventricular hypertrophy	8	2.766	0.0073
HP:0001769	Broad foot	6	3.298	0.0085
HP:0012745	Short palpebral fissure	6	3.216	0.0096

Table 6.3: HPO phenotypes enriched in DR genes among the union of all DR genes in the Altai and Vindija Neanderthals.

gene set	description	Num. DR Genes	OR	*P*-value
HP:0011065	Conical incisor	4	9.528	0.00063
HP:0006342	Peg-shaped maxillary lateral incisors	3	11.43	0.0018
HP:0011063	Abnormality of incisor morphology	4	6.929	0.0022
HP:0011792	Neoplasm by histology	12	2.499	0.0025
HP:0000698	Conical tooth	4	6.352	0.0031
HP:0000557	Buphthalmos	4	6.098	0.0037
HP:0000676	Abnormality of the incisor	5	4.537	0.0044
HP:0001019	Erythroderma	4	5.444	0.0056
HP:0001000	Abnormality of skin pigmentation	16	1.961	0.0058
HP:0001519	Disproportionate tall stature	3	7.622	0.0063

Table 6.4: HPO phenotypes enriched in DR genes unique to the Denisovan.

gene set	description	Num. DR Genes	OR	*P*-value
HP:0100710	Impulsivity	3	7.743	0.0063
HP:0001302	Pachygyria	6	3.392	0.00766
HP:0001611	Nasal speech	3	6.361	0.0110
HP:0100803	Abnormality of the periungual region	2	11.87	0.0115
HP:0000954	Single transverse palmar crease	5	3.373	0.0152
HP:0000829	Hypoparathyroidism	2	9.894	0.0165
HP:0001339	Lissencephaly	4	3.958	0.0174
HP:0001805	Thick nail	2	9.133	0.0193
HP:0200039	Pustule	2	9.133	0.0193
HP:0011061	Abnormality of dental structure	7	2.503	0.0198

GO Term	N Genes	Norm. Enrich	*P*
microtubule organizing center localization	2	1.56	0.017
CENP-A containing chromatin organization	2	1.54	0.011
multi-organism localization	3	1.53	0.015
antigen processing and presentation	12	1.50	0
DNA-templated transcription, elongation	11	1.49	0.003
deoxyribonucleotide metabolic process	3	1.49	0.023
nucleoside bisphosphate metabolic process	15	1.42	0.008
membrane docking	13	1.40	0.006
tRNA metabolic process	20	1.35	0.005
cell cycle checkpoint	20	1.33	0.009
positive regulation of cell cycle	21	1.29	0.006

Table 6.5: Overall enriched biological process GO terms
We conducted GSEA over all genes with at least one significant difference, and ranked them by the number of tissues there were significant in. None pass an FDR multiple testing correction.

REFERENCES

[1] King Jr., M. L. *Strength to Love* (1963).

[2] Skoyles, J. R. Human balance, the evolution of bipedalism and dysequilibrium syndrome. *Medical Hypotheses* **66**, 1060–1068 (2006). URL http://www.sciencedirect.com/science/article/pii/S0306987706001009.

[3] Spoor, F., Wood, B. & Zonneveld, F. Implications of early hominid labyrinthine morphology for evolution of human bipedal locomotion. *Nature* **369**, 645–648 (1994). URL https://doi.org/10.1038/369645a0.

[4] Davis, L. *et al.* Polygenic Adaptation Underlies Evolution of Brain Structures and Behavioral Traits. *European Neuropsychopharmacology* **29**, S755–S756 (2019). URL http://www.sciencedirect.com/science/article/pii/S0924977X17303796.

[5] Speakman, J. R. The evolution of body fatness: trading off disease and predation risk. *The Journal of Experimental Biology* **221**, jeb167254 (2018). URL http://jeb.biologists.org/content/221/Suppl{_}1/jeb167254.abstract.

[6] Simonti, C. N., Pavličev, M. & Capra, J. A. Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Molecular Biology and Evolution* msx219–msx219 (2017). URL http://dx.doi.org/10.1093/molbev/msx219.

[7] Huda, A., Bowen, N. J., Conley, A. B. & Jordan, I. K. Epigenetic regulation of transposable element derived human gene promoters. *Gene* **475**, 39–48 (2011). URL http://www.sciencedirect.com/science/article/pii/S0378111910004762.

[8] Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial Evolution. *Science* **283**, 1476 LP – 1481 (1999). URL http://science.sciencemag.org/content/283/5407/1476.abstract.

[9] Kumar, S., Filipski, A., Swarna, V., Walker, A. & Hedges, S. B. Placing confidence limits on the molecular age of the human–chimpanzee divergence. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18842 LP – 18847 (2005). URL http://www.pnas.org/content/102/52/18842.abstract.

[10] Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences* **113**, 10607 LP – 10612 (2016). URL http://www.pnas.org/content/113/38/10607.abstract.

[11] Maslin, M. A., Shultz, S. & Trauth, M. H. A synthesis of the theories and concepts of early human evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 1–12 (2015).

[12] Scerri, E. M. L. *et al.* Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in ecology & evolution* **33**, 582–594 (2018). URL https://pubmed.ncbi.nlm.nih.gov/30007846https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6092560/.

[13] Vattathil, S. & Akey, J. Small Amounts of Archaic Admixture Provide Big Insights into Human History. *Cell* **163**, 281–284 (2015). URL http://www.sciencedirect.com/science/article/pii/S0092867415012593.

[14] Galway-Witham, J. & Stringer, C. How did Homo sapiens evolve? *Science* **360**, 1296 LP – 1298 (2018). URL http://science.sciencemag.org/content/360/6395/1296.abstract.

[15] White, T. D. *et al.* Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423**, 742–747 (2003). URL https://doi.org/10.1038/nature01669.

[16] McDougall, I., Brown, F. H. & Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**, 733–736 (2005). URL https://doi.org/10.1038/nature03258.

[17] Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987). URL https://doi.org/10.1038/325031a0.

[18] Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000). URL https://doi.org/10.1038/35047064.

[19] Poznik, G. D. *et al.* Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. *Science* **341**, 562 LP – 565 (2013). URL http://science.sciencemag.org/content/341/6145/562.abstract.

[20] Hublin, J.-J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546**, 289–292 (2017). URL https://doi.org/10.1038/nature22336.

[21] Schlebusch, C. M. *et al.* Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652 LP – 655 (2017). URL http://science.sciencemag.org/content/358/6363/652.abstract.

[22] Skoglund, P. & Mathieson, I. Ancient Human Genomics: The First Decade. *Annu. Rev. Genom. Hum. Genet* **198**, 1–824 (2018). URL https://doi.org/10.1146/annurev-genom-083117-021749.

[23] Harvati, K. *et al.* Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia. *Nature* **571**, 500–504 (2019). URL https://doi.org/10.1038/s41586-019-1376-z.

[24] Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242 (2016). URL https://doi.org/10.1038/nature19792.

[25] Poznik, G. D. *et al.* Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature genetics* **48**, 593–599 (2016). URL https://pubmed.ncbi.nlm.nih.gov/27111036https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4884158/.

[26] Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences* **110**, 2223 LP – 2227 (2013). URL http://www.pnas.org/content/110/6/2223.abstract.

[27] Fu, Q. *et al.* A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Current Biology* **23**, 553–559 (2013). URL http://www.sciencedirect.com/science/article/pii/S0960982213002157.

[28] Higham, T. *et al.* The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* **512**, 306–309 (2014). URL https://doi.org/10.1038/nature13621.

[29] Galván, B. *et al.* New evidence of early Neanderthal disappearance in the Iberian Peninsula. *Journal of Human Evolution* **75**, 16–27 (2014). URL http://www.sciencedirect.com/science/article/pii/S0047248414001481.

[30] Huxley, T. H. The Aryan Question and Pre-Historic Man (1890).

[31] Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010). URL http://science.sciencemag.org/content/328/5979/710.abstract.

[32] Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nature reviews. Genetics* **16**, 359–371 (2015). URL https://pubmed.ncbi.nlm.nih.gov/25963373https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4478293/.

[33] Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).

[34] Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology* **26**, 1241–1247 (2016). URL http://dx.doi.org/10.1016/j.cub.2016.03.037.

[35] Skov, L. *et al.* The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature* (2020). URL https://doi.org/10.1038/s41586-020-2225-9.

[36] Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* **468**, 1053–1060 (2010).

[37] Slon, V. *et al.* The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018).

[38] Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53–61.e9 (2018). URL http://dx.doi.org/10.1016/j.cell.2018.02.031.

[39] Villanea, F. A. & Schraiber, J. G. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nature Ecology and Evolution* **3**, 39–44 (2019). URL http://dx.doi.org/10.1038/s41559-018-0735-8.

[40] McHugo, G. P., Dover, M. J. & MacHugh, D. E. Unlocking the origins and biology of domestic animals using ancient DNA and paleogenomics. *BMC biology* **17**, 98 (2019). URL https://pubmed.ncbi.nlm.nih.gov/31791340https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6889691/.

[41] Olsson, O. & Paik, C. Long-run cultural divergence: Evidence from the Neolithic Revolution. *Journal of Development Economics* **122**, 197–213 (2016). URL http://www.sciencedirect.com/science/article/pii/S0304387816300360.

[42] Robertshaw, P. T. & Collett, D. P. The Identification of Pastoral Peoples in the Archaeological Record: An Example from East Africa. *World Archaeology* **15**, 67–78 (1983). URL http://www.jstor.org/stable/124638.

[43] Outram, A. K. *et al.* The Earliest Horse Harnessing and Milking. *Science* **323**, 1332 LP – 1335 (2009). URL http://science.sciencemag.org/content/323/5919/1332.abstract.

[44] Kislev, M. E., Hartmann, A. & Bar-Yosef, O. Early Domesticated Fig in the Jordan Valley. *Science* **312**, 1372 LP – 1374 (2006). URL http://science.sciencemag.org/content/312/5778/1372.abstract.

[45] Simmons, A. H., Kohler-Rollefson, I., Rollefson, G. O., Mandel, R. & Kafafi, Z. Ain Ghazal: A Major Neolithic Settlement in Central Jordan. *Science* **240**, 35 LP – 39 (1988). URL http://science.sciencemag.org/content/240/4848/35.abstract.

[46] Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* **18**, 659–674 (2017). URL http://dx.doi.org/10.1038/nrg.2017.65.

[47] Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014). URL http://www.nature.com/doifinder/10.1038/nature13673. 1312.6639.

[48] Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015). URL http://www.nature.com/doifinder/10.1038/nature14317. 1502.02783.

[49] Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016). URL http://dx.doi.org/10.1038/nature17993http://www.nature.com/nature/journal/v534/n7606/pdf/nature17993.pdf.

[50] Bramanti, B. *et al.* Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. *Science* **326**, 137 LP – 140 (2009). URL http://science.sciencemag.org/content/326/5949/137.abstract.

[51] Brace, S. *et al.* Ancient genomes indicate population replacement in Early Neolithic Britain. *Nature Ecology and Evolution* **3**, 765–771 (2019). URL http://dx.doi.org/10.1038/s41559-019-0871-9.

[52] Schlebusch, C. M. *et al.* Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* **338**, 374 LP – 379 (2012). URL http://science.sciencemag.org/content/338/6105/374.abstract.

[53] Skoglund, P. *et al.* Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016). URL http://dx.doi.org/10.1038/nature19844http://10.0.4.14/nature19844http://www.nature.com/nature/journal/v538/n7626/abs/nature19844.html{#}supplementary-information.

[54] Posth, C. *et al.* Language continuity despite population replacement in Remote Oceania. *Nature Ecology & Evolution* (2018). URL http://www.nature.com/articles/s41559-018-0498-2.

[55] Prendergast, M. E. *et al.* Ancient dna reveals a multistep spread of the first herders into sub-saharan africa. *Science* **6275**, 1–19 (2019).

[56] Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019). URL http://science.sciencemag.org/content/365/6457/eaat7487.abstract.

[57] Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* (2019).

[58] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017). URL http://www.sciencedirect.com/science/article/pii/S0092867417306293.

[59] Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science (New York, N.Y.)* **310**, 1782–1786 (2005).

[60] Nakayama, K. *et al.* Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. *Journal of human genetics* **47**, 92–94 (2002).

[61] Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015). URL http://dx.doi.org/10.1038/nature16152.

[62] Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737–741 (2016).

[63] Dannemann, M. & Kelso, J. The contribution of Neanderthals to phenotypic variation in modern humans. *American Journal of Human Genetics* **101**, 578–589 (2017).

[64] Gunz, P. *et al.* Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Current Biology* **29**, 120–127.e5 (2019).

[65] Rinker, D. C. *et al.* Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations. *bioRxiv* 533257 (2019). URL http://biorxiv.org/content/early/2019/11/15/533257.abstract.

[66] Dannemann, M., Prüfer, K. & Kelso, J. Functional implications of Neandertal introgression in modern humans. *Genome Biology* **18**, 61 (2017). URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1181-7.

[67] McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* **168**, 916–927.e12 (2017). URL http://dx.doi.org/10.1016/j.cell.2017.01.038.

[68] Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences* **116**, 1639 LP – 1644 (2019). URL http://www.pnas.org/content/116/5/1639.abstract.

[69] Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014). URL https://doi.org/10.1038/nature13408.

[70] Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *American journal of human genetics* **74**, 1111–1120 (2004). URL https://pubmed.ncbi.nlm.nih.gov/15114531https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182075/.

[71] Mathieson, S. & Mathieson, I. FADS1 and the timing of human adaptation to agriculture. *Molecular Biology and Evolution* **35**, 2957–2970 (2018).

[72] King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

[73] Romero, I. G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet* **13**, 505–516 (2012). URL http://dx.doi.org/10.1038/nrg3229.

[74] Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Medicine* **6**, 85 (2014). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4254432/.

[75] Benton, M. L., Talipineni, S. C., Kostka, D. & Capra, J. A. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics* **20**, 511 (2019). URL https://doi.org/10.1186/s12864-019-5779-x.

[76] Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome research* **24**, 1–13 (2014). URL https://pubmed.ncbi.nlm.nih.gov/24196873https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875850/.

[77] Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* **21**, 292–310 (2020). URL https://doi.org/10.1038/s41576-019-0209-0.

[78] Gokhman, D. *et al.* Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan. *Science* **344**, 523 (2014). URL http://science.sciencemag.org/content/344/6183/523.abstract.

[79] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015). URL http://dx.doi.org/10.1038/ng.3367http://www.nature.com/ng/journal/v47/n9/pdf/ng.3367.pdf.

[80] Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).

[81] Hajdinjak, M. *et al.* Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652–656 (2018). NIHMS150003.

[82] Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLOS Genetics* **14**, e1007349 (2018). URL http://dx.plos.org/10.1371/journal.pgen.1007349.

[83] Sawyer, S. *et al.* Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proceedings of the National Academy of Sciences* **112**, 15696 LP – 15700 (2015).

[84] Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014). URL http://www.nature.com/nature/journal/v505/n7481/abs/nature12886.html. arXiv:1507.02142v2.

[85] Castellano, S. *et al.* Patterns of coding variation in the complete exomes of three Neandertals. *Proceedings of the National Academy of Sciences* **111**, 6666–6671 (2014).

[86] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). URL http://www.nature.com/doifinder/10.1038/nature24277.

[87] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL http://dx.doi.org/10.1038/nature15393http://10.0.4.14/nature15393http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html{#}supplementary-information.

[88] Abi-Rached, L. *et al.* The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science* **334**, 89–94 (2011).

[89] 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). URL http://dev.biologists.org/content/133/23/4749.abstract.

[90] Skibinski, G. *et al.* Mutations in the endosomal ESCRTIII-complex subunit CHMP2B in frontotemporal dementia. *Nature Genetics* **37**, 806 (2005). URL http://dx.doi.org/10.1038/ng1609http://10.0.4.14/ng1609https://www.nature.com/articles/ng1609{#}supplementary-information.

[91] Cox, L. E. *et al.* Mutations in CHMP2B in Lower Motor Neuron Predominant Amyotrophic Lateral Sclerosis (ALS). *PLOS ONE* **5**, e9872 (2010). URL https://doi.org/10.1371/journal.pone.0009872.

[92] McDowell, K. A. *et al.* Reduced Cortical BDNF Expression and Aberrant Memory in Carf Knock-Out Mice. *The Journal of Neuroscience* **30**, 7453 LP – 7465 (2010). URL http://www.jneurosci.org/content/30/22/7453.abstract.

[93] Schumer, M. *et al.* Natural selection interacts with the local recombination rate to shape the evolution of hybrid genomes. *Science* **3684**, 212407 (2018). 212407.

[94] Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012). URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{&}id=22936568{&}retmode=ref{&}cmd=prlinks. NIHMS150003.

[95] Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011). URL http://dx.doi.org/10.1038/nature10532http://www.nature.com/nature/journal/v478/n7369/abs/nature10532.html{#}supplementary-information.

[96] Gokhman, D., Malul, A. & Carmel, L. Inferring Past Environments from Ancient Epigenomes. *Molecular Biology and Evolution* **34**, 2429–2438 (2018).

[97] Martin, A. R. *et al.* Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS Genetics* **10**, 1004549 (2014). URL https://doi.org/10.1371/journal.pgen.1004549.

[98] Kelly, D. E., Hansen, M. E. & Tishkoff, S. A. Global variation in gene expression and the value of diverse sampling. *Current Opinion in Systems Biology* **1**, 102–108 (2017). URL http://linkinghub.elsevier.com/retrieve/pii/S2452310017300124.

[99] Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nature Genetics* **51**, 1349–1355 (2019). URL http://dx.doi.org/10.1038/s41588-019-0487-7.

[100] Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for Weak Selective Constraint on Human Gene Expression. *Genetics* **211**, 757 LP – 772 (2019). URL http://www.genetics.org/content/211/2/757.abstract.

[101] Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022–1034.e6 (2019). URL https://doi.org/10.1016/j.cell.2019.04.014.

[102] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). URL http://dx.doi.org/10.1038/nature11632http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html{#}supplementary-information.

[103] Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017). URL http://dx.doi.org/10.1093/bioinformatics/btx346.

[104] Colbran, L. L. *et al.* Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nature Ecology & Evolution* (2019). URL http://dx.doi.org/10.1038/s41559-019-0996-x.

[105] Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21775986https://www.ncbi.nlm.nih.gov/pmc/PMC3154982/.

[106] Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**, 341–354.e12 (2016). URL http://www.sciencedirect.com/science/article/pii/S0092867416311692.

[107] Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research* **45**, W130–W137 (2017). URL http://dx.doi.org/10.1093/nar/gkx356.

[108] Gilbert, M. T. P. *et al.* Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic acids research* **35**, 1–10 (2007). URL https://pubmed.ncbi.nlm.nih.gov/16920744https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1802572/.

[109] Gamba, C. *et al.* Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources* **16**, 459–469 (2016). URL https://doi.org/10.1111/1755-0998.12470.

[110] Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 787903 (2019). URL http://biorxiv.org/content/early/2019/10/03/787903.abstract.

[111] Goude, G. & Fontugne, M. Carbon and nitrogen isotopic variability in bone collagen during the Neolithic period: Influence of environmental factors and diet. *Journal of Archaeological Science* **70**, 117–131 (2016). URL http://www.sciencedirect.com/science/article/pii/S0305440316300334.

[112] Clemente, F. *et al.* A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. *The American Journal of Human Genetics* **95**, 584–589 (2014). URL http://www.sciencedirect.com/science/article/pii/S0002929714004224.

[113] Jarvis, J. P. *et al.* Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics* **8**, e1002641–e1002641 (2012). URL https://pubmed.ncbi.nlm.nih.gov/22570615https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3343053/.

[114] Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**, 206–216 (2007). URL https://doi.org/10.1038/nrg2063.

[115] Ameur, A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *American journal of human genetics* **90**, 809–820 (2012). URL https://pubmed.ncbi.nlm.nih.gov/22503634https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3376635/.

[116] Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLOS Biology* **4**, e72 (2006). URL https://doi.org/10.1371/journal.pbio.0040072.

[117] Hancock, A. M. *et al.* Adaptations to climate in candidate genes for common metabolic disorders. *PLoS genetics* **4**, e32–e32 (2008). URL https://pubmed.ncbi.nlm.nih.gov/18282109https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2242814/.

[118] Kentish, S. J., Wittert, G. A., Blackshaw, L. A. & Page, A. J. A chronic high fat diet alters the homologous and heterologous control of appetite regulating peptide receptor expression. *Peptides* **46**, 150–158 (2013). URL http://www.sciencedirect.com/science/article/pii/S0196978113002210.

[119] Loos, R. J. F. *et al.* Polymorphisms in the leptin and leptin receptor genes in relation to resting metabolic rate and respiratory quotient in the Québec Family Study. *International Journal of Obesity* **30**, 183–190 (2006). URL https://doi.org/10.1038/sj.ijo.0803127.

[120] Luca, F., Perry, G. H. & Di Rienzo, A. Evolutionary adaptations to dietary changes. *Annual review of nutrition* **30**, 291–314 (2010). URL https://pubmed.ncbi.nlm.nih.gov/20420525https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4163920/.

[121] Farooqi, I. S. *et al.* Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. *The New England journal of medicine* **356**, 237–247 (2007). URL https://pubmed.ncbi.nlm.nih.gov/17229951https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2670197/.

[122] Dehghani, M. R. *et al.* Potential role of gender specific effect of leptin receptor deficiency in an extended consanguineous family with severe early-onset obesity. *European Journal of Medical Genetics* **61**, 465–467 (2018). URL http://www.sciencedirect.com/science/article/pii/S1769721217308029.

[123] Chen, J. *et al.* A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research* **29**, 53–63 (2018).

[124] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016). URL https://doi.org/10.1038/nature19057http://10.0.4.14/nature19057https://www.nature.com/articles/nature19057{#}supplementary-information.

[125] Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics* **29**, 569–574 (2013). URL http://dx.doi.org/10.1016/j.tig.2013.05.010.

[126] Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends in Genetics* **19**, 362–365 (2003). URL https://doi.org/10.1016/S0168-9525(03)00140-9.

[127] Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature genetics* **44**, 1015–1019 (2012). URL https://pubmed.ncbi.nlm.nih.gov/22902787https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3480734/.

[128] Peter, B. M., Huerta-Sanchez, E. & Nielsen, R. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLOS Genetics* **8**, e1003011 (2012). URL https://doi.org/10.1371/journal.pgen.1003011.

[129] Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015). URL http://www.nature.com/doifinder/10.1038/nature14558.NIHMS150003.

[130] Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)* **31**, 3555–3557 (2015). URL https://pubmed.ncbi.nlm.nih.gov/26139635https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4626747/.

[131] Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research* **47**, W199–W205 (2019). URL https://pubmed.ncbi.nlm.nih.gov/31114916https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6602449/.

[132] Ségurel, L. & Bon, C. On the Evolution of Lactase Persistence in Humans. *Annual Review of Genomics and Human Genetics* **18**, 297–319 (2017). URL https://doi.org/10.1146/annurev-genom-091416-035340.

[133] Currat, M. & Excoffier, L. Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proceedings of the National Academy of Sciences* **108**, 15129 LP – 15134 (2011). URL http://www.pnas.org/content/108/37/15129.abstract.

[134] Rotival, M. & Quintana-Murci, L. Functional consequences of archaic introgression and their impact on fitness. *Genome Biology* **21**, 3 (2020). URL https://doi.org/10.1186/s13059-019-1920-z.

[135] Meyer, D., C. Aguiar, V. R., Bitarello, B. D., C. Brandt, D. Y. & Nunes, K. A genomic perspective on HLA evolution. *Immunogenetics* **70**, 5–27 (2018). URL https://doi.org/10.1007/s00251-017-1017-3.

[136] Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *bioRxiv* 2020.03.12.988956 (2020). URL http://biorxiv.org/content/early/2020/03/13/2020.03.12.988956.abstract.

[137] Sawyer, G. J. & Maley, B. Neanderthal reconstructed. *The Anatomical Record Part B: The New Anatomist* **283B**, 23–31 (2005). URL https://doi.org/10.1002/ar.b.20057.

[138] Chen, F. *et al.* A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* (2019). URL http://www.nature.com/articles/s41586-019-1139-x.

[139] Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps. *Cell* **179**, 180–192.e10 (2019). URL https://doi.org/10.1016/j.cell.2019.08.035.

[140] James, W. P. T. *et al.* Nutrition and its role in human evolution. *Journal of Internal Medicine* **285**, 533–549 (2019). URL https://doi.org/10.1111/joim.12878.

[141] Villar, D. *et al.* Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566 (2015). URL http://dx.doi.org/10.1016/j.cell.2015.01.006.

[142] Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics* **50**, 1171–1179 (2018). URL http://dx.doi.org/10.1038/s41588-018-0160-6.

[143] Chen, L., Fish, A. E. & Capra, J. A. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLOS Computational Biology* **14**, e1006484 (2018). URL https://doi.org/10.1371/journal.pcbi.1006484.

[144] Gamazon, E. R., Zwinderman, A. H., Cox, N. J., Denys, D. & Derks, E. M. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nature Genetics* **51**, 933–940 (2019). URL https://doi.org/10.1038/s41588-019-0409-8.

[145] Stricker, S. H., Koferle, A. & Beck, S. From profiles to function in epigenomics. *Nat Rev Genet* **18**, 51–66 (2017). URL http://dx.doi.org/10.1038/nrg.2016.138http://www.nature.com/nrg/journal/v18/n1/pdf/nrg.2016.138.pdf.

[146] Fagny, M. *et al.* The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nature Communications* **6**, 10047 (2015). URL https://doi.org/10.1038/ncomms10047.

[147] Mancabelli, L. *et al.* Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environmental Microbiology* **19**, 1379–1390 (2017). URL https://doi.org/10.1111/1462-2920.13692.

[148] Lea, A. J., Martins, D., Kamau, J., Gurven, M. & Ayroles, J. F. Urbanization and market-integration have strong, non-linear effects on cardio-metabolic health in the Turkana. *bioRxiv* 756866 (2019). URL http://biorxiv.org/content/early/2019/10/15/756866.abstract.

[149] Oster, J., Warken, S., Sekhon, N., Arienzo, M. & Lachniet, M. Speleothem Paleoclimatology for the Caribbean, Central America, and North America. *Quaternary* **2**, 5 (2019). URL http://www.mdpi.com/2571-550X/2/1/5.

[150] Keinan, J. *et al.* Paleoclimatology of the Levant from Zalmon Cave speleothems, the northern Jordan Valley, Israel. *Quaternary Science Reviews* **220**, 142–153 (2019). URL http://www.sciencedirect.com/science/article/pii/S0277379118307984.

[151] Hughes, M. K. Dendroclimatology in High-Resolution Paleoclimatology BT - Dendroclimatology: Progress and Prospects. 17–34 (Springer Netherlands, Dordrecht, 2011). URL https://doi.org/10.1007/978-1-4020-5725-0{_}2.