The contribution of the 3D genome to gene regulation, human evolution, and disease

By

Evonne McArthur

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

May 13th, 2022

Nashville, Tennessee

Approved:

John A. Capra, Ph.D.

Nancy J. Cox, Ph.D.

Lea K. Davis, Ph.D.

Douglas M. Ruderfer, Ph.D.

Emily Hodges, Ph.D.

Alexander G. Bick, M.D, Ph.D.

ACKNOWLEDGMENTS

More than any discovery, my journey was truly about the community of people who made this pursuit fulfilling and fun. On paper, many things went “wrong” during my PhD: my lab moved across the country and a pandemic forced more than half of my experience to be physically isolated. But, I never felt alone. In fact, these last few years have been a grand adventure and, despite the challenges, I would not want it any other way. For this, I am eternally grateful and indebted to the scientific community, friends, and family that supported me along the way.

I was once told that a PhD is like a four year-long conversation: you *really* better like the people you are talking to. Luckily, I could not have asked for a better group of conversation partners. Being at the edge of knowledge can be intimidating and frustrating, but my peers and my mentors made it exciting and enjoyable. I want to thank the Vanderbilt MSTP, VGI, and ACCRE administrators, staff, teachers, and faculty that facilitate these conversations. By coordinating seminars, providing resources, teaching challenging topics, and directing these programs, you all shape the future of science. I am grateful for my peers in both the Vanderbilt MSTP, human genetics graduate program, and collaborators at UCSF for their continually interesting ideas and help in making my ideas a reality. I want to thank each member of my committee—Lea Davis, Doug Ruderfer, Emily Hodges, Alex Bick, and Nancy Cox—for inspiring me at every meeting and giving me the encouragement to keep going. To each member of the Capra Lab, both past and present, I am forever appreciative of all your advice, camaraderie. . . and especially all of our happy hours. Not only have you left a unique imprint on this work, but you have also made me a better scientist and person. I especially want to thank Sarah Fong and David Rinker for your willingness to be my recurring sounding-board both inside and outside of the lab. Most importantly, I want to express my deepest gratitude to Tony Capra for his unfailingly thoughtful mentorship. You made me feel at home, both scientifically and personally, even in trying times.

Finally, I want to acknowledge my family for making me whole during such a focused undertaking. Thank you for always reminding me of the important things and for being my biggest cheerleaders. To Nikki, thank you for the most empathy. To my Mom, thank you for the most encouragement. To my Dad, thank you for the most advice. All of your support means more than you can ever know. And, to Ankith, who was my partner on this crazy adventure—travelling across the country 4 times, living in 6 apartments in 3 states, working 9 different jobs, planning 3 weddings, and co-parenting Hero the dog—I could not have done it without you.

TABLE OF CONTENTS

[*] This chapter has been previously published in McArthur, E., Rinker, D.C., & Capra, J.A. 2021. *Nat. Comms.*

[†] This chapter has been previously published in McArthur, E. & Capra, J.A. 2021. *AJHG.*

[‡] This chapter is available as a preprint in McArthur, E. *et al.* 2022. *BioRxiv.*

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION*

1.1 The evolution of humans

> *"And today we inherit the genetic scars of an ancient struggle. What Mendel discovered was not a law so much as a battleground."*
>
> — Carl Zimmer. "She Has Her Mother's Laugh: The Powers, Perversions, and Potential of Heredity." (2018), pp. 162.[1]

1.1.1 Motivation

Although anatomically modern humans first appeared in Africa 200,000–300,000 years ago, the human genome has been shaped by billions of years of diverse physical environments, exposures, and hybridization[2]. This process of evolution has enabled our ancestors to biologically adapt to diverse challenges. Through mutations to the genome, evolved features become encoded and inherited through generations. However, the processes that permitted our species to become diploid, multi-cellular, sentient, and social organisms are the same processes that cause genetic disease. These evolutionarily ancient systems, like DNA replication, methylation, recombination, and repair, lead to genome diversity which provides the substrate for both adaptation and dysfunction[3,4]. Consequently, all of human genetic predisposition—risk for cardiovascular disease, protection against autoimmune disease, or distaste for cilantro—is a consequence of evolution[3]. To explain human biology, it is imperative to understand human evolution.

Variation in traits and diseases represent a network of intertwined trade-offs[3]: benefits and costs exist for both short and tall stature, high and low blood sugar, the risk for infection and over-active autoimmunity. And none of these traits exist in a silo, rather they are evaluated by selection as a product of interactions with both the external and internal environment. For example, lighter versus darker skin pigmentation modulates exposure to UV radiation, which is necessary for Vitamin D_3 synthesis, but an excess increases risk for skin neoplasms. Therefore, selection on skin color can be adaptive to populations at different latitudes that have different sun exposure. Vitamin D levels subsequently affect the physiology of the skeleton, parathyroid, metabolism, heart, and immune system—each of which also interacts with the environment and other body systems[5]. Furthermore, genomic loci responsible for skin color are nearby and physically "linked" to loci that contribute to risk of prostate cancer, which illustrates another entwined, or "hitchhiking", trade-off[6]. This example illustrates how environmental pressures select for traits that benefit multiple human biological processes, but maintain genetic variants that increase the risk for others. This complexity highlights the need for an evolutionary perspective to move from isolated trait associations towards mechanistic biological insights that can enable prognostic, preventative, and therapeutic strategies.

*Parts of this chapter have been adapted from McArthur, E., Rinker, D.C., & Capra, J.A. 2021. *Nat. Comms.*; McArthur, E. & Capra, J.A. 2021. *AJHG.*; and, McArthur, E. *et al.* 2022. *BioRxiv.* with permission of the publisher and co-authors.

One such mechanism that drives the variation in human traits is the regulation of how genes are expressed to produce the diversity of cell types needed across the human lifespan[7–10]. Tracing the "genetic scars" left in gene regulatory regions of the genome can help us to understand how humans have evolved and why individuals have risk for certain traits. Just as an individual's family tree holds clues about both their past ancestry and future risk for disease[3], improving our understanding of the human evolutionary tree can teach us both about the challenges our ancestors faced and help us to understand, and maybe even manipulate, our genome for tomorrow.

1.1.2 Archaic hominins and admixture

Today, humans (*H. sapiens*) are the only remaining hominin species. However, the archeological and paleontological record suggests that the closest members of our family tree—archaic hominins—survived until 30–40 kya[2,11,12]. Evidence of Neanderthals (*H. neanderthalensis*) have been found in the Middle East, Europe, and Asia, while Denisovans are thought to have lived in East and Southeast Asia[2]. The sequencing of these archaic hominin genomes and their comparison with diverse modern human genomes has transformed our understanding of human history, evolution, and biology[13–17]. To date, high-quality genome sequencing is available for three Neanderthals and one Denisovan. Each are named for where they were discovered: *Vindija 33.13* Neanderthal is from Vindija cave in Croatia[14]; *Chagyrskaya 8* Neanderthal is from Chagyrskaya cave[15]; and, both the *Altai* Neanderthal[13] (also named *Denisova 5*) and the *Denisova 3* Denisovan[16] are from the Denisova cave. Both the Denisova and Chagyrskaya caves are in the Altai mountains in Siberia, Russia near the border of Kazakhstan. It is estimated that the archaic and modern human lineage diverged between 0.85–1.2 mya and the Neanderthal-Denisovan lineage split around 600–750 kya[18–20]. These individuals estimated relationships, ages, and geographic locations are depicted in Fig. 1.1.

Ever since the first Neanderthal fossil was discovered in 1856 in the German Neander Valley[21], there has been considerable curiosity as to how their lifestyles, anatomy, and cognition were similar or different from modern humans. Archaeological artifacts, paleontological evidence, and their geographic dispersal suggest they were approximately 5 feet tall, with a prominent supra-orbital ridge, large thorax and skull, had a hunter-gatherer lifestyle, and made a variety of complex tools[22–25]. For more than a century, scientists hypothesized whether modern and archaic humans hybridized (i.e., interbred), especially once it was established that humans and archaic hominins were in the same place at the same time[2].

Genome sequencing finally revealed that, over the past 50,000 years, modern humans interbred with multiple archaic hominin groups—including both Neanderthals and Denisovans—on multiple occasions and in several locations after migrating out of Africa. Archaic hominin groups also hybridized with other archaic hominins creating an intertangled web of relatedness[26]; notably, *Denisovan 11*, "Denny," was a first-generation hybrid with a Neanderthal mother and Denisovan father[27]. As a result of these introgression events, nearly all Eurasians have approximately 2% Neanderthal ancestry[14,28,29]. The amount of Denisovan ancestry is more population-specific, with some Oceanian individuals having up to 5% Denisovan ancestry[28]. Collectively, at least 38% of archaic hominin genomes still remain in fragments across Eurasian genomes[30,31]. Therefore, archaic hominins are not only a closely-related foil that can help highlight uniquely human parts of our genome, but they actually contribute to human genome diversity today.

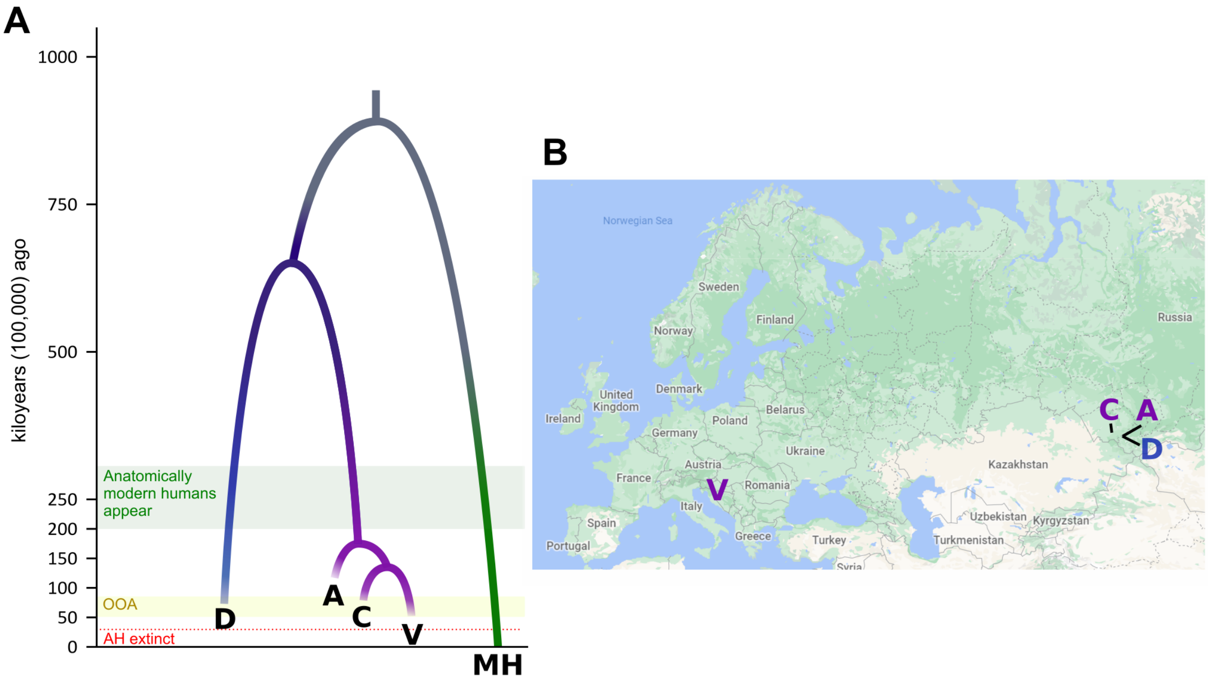


Figure 1.1: Evolutionary and geographic relationships between archaic hominins with high-quality genome sequences.
(A) The estimated evolutionary relationships between modern humans (**MH** [green]), three Neanderthals (**V**: Vindija[14], **A**: Altai[13], **C**: Chagyrksyaka[15] [purple]), and one Denisovan[16] (**D** [blue]) are depicted. The phylogeny is approximately to scale; however, many of the dates are imprecise estimates and depend on the simulation parameters, like mutation rate. The end of each archaic branch represents the date estimate for each individual (not the extinction of that population). Other events like the appearance of anatomically modern humans (green bar), the most significant recent out-of-Africa (OOA) migration (yellow bar), and an estimate of when archaic hominins (AH) disappeared are highlighted (red line). **(B)** The map depicts where each archaic sample was identified within Eurasia. These figures were adapted from Mafessoni et al.[15] (Fig. 1) with additional data from Rogers et al.[18], Rogers et al.[19], and Gómez-Robles[20]. The map was reproduced from Google Maps.

1.1.3 Phenotypic legacy of archaic ancestry

The archaic ancestry remaining in modern human genomes is far from silent. Admixture added multiple pulses of genetic diversity to human genomes[26,32], providing both potential fitness advantages and disadvantages to those who inherited them. For example, archaic DNA may have facilitated the ability of modern humans to inhabit diverse environments as they spread around the globe[33]. Some archaic alleles have functions and evolutionary signatures suggestive of positive selection due to beneficial effects[33–35]. Many of these alleles influence systems that directly interact with the environment[26], such as the immune system[36–43], hair and skin[44–46], response to oxygen[47], and metabolism[34,48–50].

Despite these potential adaptive benefits of admixture, simulations and empirical analyses of the distribution of introgressed alleles across the genome suggest that they were largely deleterious in modern humans[51,52]. Several lines of evidence support selection against introgressed Neanderthal DNA in most functional regions of human genomes shortly after hybridization[51–54]. First, Neanderthal ancestry is depleted in regions of the genome with strong background selection and evolutionary conservation[44,45,53]. Second, Neanderthal ancestry is depleted in regions of the genome with annotated molecular functions (e.g., genes and gene regulatory elements), and this depletion is strongest in annotated brain and testis regulatory regions[53,55,56]. Furthermore, remaining alleles of Neanderthal ancestry—–i.e., introgressed alleles that were maintained by either selection or drift since admixture—are predicted to be less likely to modify protein and regulatory function than matched sets of alleles that arose on the human lineage, suggesting that functional introgressed variants were less tolerated[57,58]. Finally, the majority of archaic alleles that are strongly associated with disease in single-locus tests are risk-increasing in the context of modern human populations[59].

Several non-exclusive scenarios may explain the apparent genetic cost of Neanderthal introgression. The introgressing Neanderthals had a smaller effective population size than modern human populations. The resulting lower efficacy of selection allowed the accumulation of weakly deleterious alleles in Neanderthal populations[13]. After introgression, these variants were subject to more effective selection in larger modern human populations[51,52]. It is also possible that hybrid incompatibilities and deleterious epistatic interactions between archaic and human alleles reduced the fitness of early hybrids[44,45,55,60,61].

These benefits and costs of introgression in modern humans have been primarily characterized based on overlap with molecular annotations[44,45,53,55] or existing genome-wide association study (GWAS) hits[14,59,62]. Phenotypes associated with Neanderthal ancestry are diverse and range from immune system response[36–43,59,63], hair and skin coloration[44–46,59,63], metabolism[34,48–50], cardiopulmonary function[47,63], skeletal morphology[63,64], and behavior[26,59,63]. However, most medically and evolutionarily relevant traits are complex, with hundreds or thousands of loci across the genome contributing to them[65,66]. Thus, studies of individual loci are not sufficient to address the overall influence of Neanderthal admixture on human traits. Furthermore, despite these numerous phenotypic associations, understanding the specific mechanisms by which archaic variants contribute to disease risk and protection remains difficult.

1.2 Contribution of the non-coding genome to complex trait architecture and evolution

> *"But one researcher's trash is another researcher's treasure, and a growing number of scientists believe that hidden in the junk DNA are intellectual riches that will lead to a better understanding of diseases [. . .], normal genome repair and regulation, and perhaps even the evolution of multicellular organisms."*
>
> — Rachel Nowak. "Mining Treasures from 'Junk DNA'."
> *Science*. 263.5147, (1994), pp. 608–610.[67]

1.2.1 Non-coding genome revolution

Just as the advent of genome sequencing facilitated our ability to accurately trace human evolutionary origins, it also spurred a transformation in human genetics and genomics. Although initial drafts of the human genome were published in 2001 by the International Human Genome Sequencing Consortium and Celera Genomics, it was only very recently (2021) that the full 3.055 billion base pair genome was truly completed after filling the missing 8%[68]. In addition to the improved genome sequence quality, the past 20 years have also been witness to an explosion in genome sequence quantity. A single pass of sequencing for an entire genome can take as little as 2 minutes[69] and, today, over 200,000 whole genome sequences are available[70].

One of the most striking surprises unveiled by the human genome sequence is the modest number of protein-coding genes. At surface-level, the central dogma of molecular biology implies that the primary purpose of DNA is to transcribe RNA that is translated into proteins which are the "functional" units of the cell. In the 1990s, it was hypothesized that humans had over 100,000 protein-coding genes and that the rest is non-functional or redundant "junk DNA"[71,72]. Remarkably, there are only approximately 20,000 protein-coding genes that account for just 2% of the human genome. Interpretation of the remaining 98%—the "non-coding" genome—has proved an ongoing puzzle.

Over the past two decades, the Encyclopedia of DNA Elements (ENCODE) project, among others, has demonstrated the importance of the non-coding genome in processes such as transcriptional and translational regulation, DNA replication, chromatin structure, and histone modification, among others. The power and complexity of the non-coding genome is illustrated by the ability for diverse cell types to arise from the same genome: transcriptional control of the right gene, at the right time, and in the right cell type is necessary for complex multi-cellular life. While this serves as the foundation for phenotypic diversity across cell types, individuals, and species, it also creates opportunity for dysfunction leading to disease. While most initial attempts to isolate variants implicated in disease focused on the exome, the genome-wide association study (GWAS) revolution has further highlighted the necessity and challenges of interpreting the phenotypic effects of the non-coding genome.

The increased availability of genotype data for large cohorts of individuals paved the way for GWAS to become a standard experimental design for population-scale studies. Simply, GWAS test for associations between a genotype and a phenotype. Over the past 15 years, GWAS have led to the discovery of thousands of loci, genes, and pathways associated with complex traits[65]. Notably, more than 88% of variants associated with common disease in GWAS are in non-coding regions[73]. These associations are enriched in functionally-annotated regulatory elements (e.g. enhancers and promoters), often in a cell-type-specific manner, suggesting that gene expression modulation in the non-coding genome largely mediates common disease risk[73–76]. There are many tools that incorporate these functional annotations with machine learning and probabilistic models to aid in variant interpretation[77]. However, prioritizing non-coding variants for functional investigation remains a challenge given the genome's complex regulatory landscape and context specificity[78,79]. Ultimately, moving from association to function is necessary for developing prognostic, preventative, and therapeutic strategies[80–84].

1.2.2 Partitioned heritability

Results of a GWAS also allow for a variety of other types of analysis such as developing polygenic risk scores, conducting Mendelian randomization, estimating genetic correlations and SNP-based heritability (SNP: single nucleotide polymorphism)[85]. Heritability of a trait describes the proportion of the phenotypic variance in a population that is attributable to genetic factors ($\hat{h}^2$), which was historically estimated with twin or family studies[86]. SNP-based heritability (h^2_{SNP}) uses data from a GWAS to estimate the proportion of phenotypic variance explained by any set of SNPs (e.g., in a genotyping array, whole genome sequencing)[87]. While heritability describes how the entire set of SNPs across the whole genome contributes to the trait's phenotypic variation, a related concept—partitioned heritability—quantifies how a subset of SNPs across a partition of the genome contributes to phenotypic variation[75,88]. For example, early work partitioned the heritability of certain traits, like height, by chromosome[88,89]. They found that the variance explained by each chromosome was directly proportional to the length of the chromosome. This suggests that height is highly polygenic; if you normalize the phenotypic variance explained by the number of SNPs in the partition, each chromosome contributes relatively equally to the heritability of height. In contrast, some traits like von Willebrand factor (vWF) levels were dominated by contribution from chromosome 9[88]. Genetic variation on chromosome 9 explains 14% of the variance in vWF levels despite chromosome 9 only accounting for 4.5% of the bases in the genome. Thus, in this example, chromosome 9 is enriched for partitioned heritability contribution to vWF levels. This preliminarily suggests that part of chromosome 9 contributes an outsized effect to vWF levels. Indeed, the *ABO* locus, which is important for vWF levels, is on chromosome 9[88]. In summary, enrichment from partitioned heritability analysis is calculated as $Enrichment_c = \frac{\%h^2_{(c)}}{\%SNP_{(c)}}$ where $h^2_{(c)}$ is the heritability explained by SNPs in partition c and $\%SNP_{(c)}$ is the proportion of SNPs in that partition of the genome[75]. This is visually depicted with a toy example in Fig. 1.2.

Although partitioned heritability can be conducted with a variety of statistical frameworks, for this work I use stratified LD score regression (S-LDSC). S-LDSC quantifies the heritability of a trait explained by common (minor allele frequency [MAF] $> 5\%$) variants in a set of regions of interest, explicitly conditioned on the association statistics and the underlying linkage disequilibrium (LD) structure[75,90].

Practically, partitioned heritability describes the genetic architecture landscape of a trait with relevance to functional, evolutionary, or population genetics-based annotations. For example, contributions of SNPs at different allele frequencies can be used to highlight the differences in genetic architecture between traits, such as obsessive-compulsive disorder and Tourette's syndrome[91]. Partitions can also be made on functional elements (i.e., enhancers, promoters, or eQTLs) across cell-types to link diseases to relevant cellular contexts. For example, heritability for body mass index (BMI) is more concentrated in central nervous system (CNS) enhancers than in other metabolic tissues (e.g., adrenal or pancreas)[75,76]. Heritability is also enriched in enhancers and promoters that have older sequence age (i.e., sequence synteny with more distant species) when compared to those with younger sequence age[92]. Together, these illustrate the power of leveraging the influx of GWAS data with partitioned heritability analyses to reveal differences in biology and evolutionary history underlying diverse traits.

1.2.3 Non-coding variation in hominin evolution and divergence

Understanding the non-coding genome is not only essential for interpreting trait-associated variations in humans, it has also proved to be a fundamental mechanism in the evolution of species. Over 50 years ago, early results from comparisons between chimpanzee and human blood proteins pointed to a "paradox": despite obvious phenotypic differences between the species, their blood protein amino acid sequences are almost identical[7,72]. Likewise, the protein-coding genomes of archaic hominins and modern humans are highly similar. Previous work has identified some non-synonymous changes leading to amino acid substitutions between archaic and modern humans that impact genes involved in metabolism, hair distribution, body morphology, cognition, and behavior[17,93]. However, these represent only a small fraction of the genome

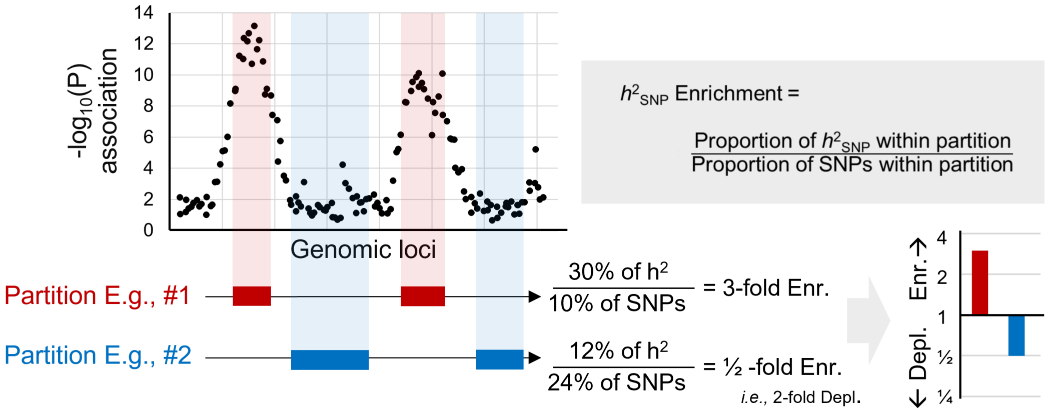

14
12
10
8
6
4
2
0
-log$_{10}$(P) association
Genomic loci
Partition E.g., #1
Partition E.g., #2
30% of h^2 / 10% of SNPs = 3-fold Enr.
12% of h^2 / 24% of SNPs = ½ -fold Enr.
i.e., 2-fold Depl.
h^2_{SNP} Enrichment =
Proportion of h^2_{SNP} within partition / Proportion of SNPs within partition
Enr. ↑
↓ Depl.
4
2
1
½
¼

Figure 1.2: Partitioned heritability analysis quantifies a genomic region's contribution to the heritability of a trait.

Pictured is a toy example of how partitioned heritability enrichment is calculated across two example genome partitions. Using an LD reference panel and regression conditioned on a set of baseline annotations (not pictured), GWAS association statistics for a given trait (Manhattan plot) are partitioned by given genomic partition(s). For the SNPs in each partition, their collective contribution to the trait heritability is calculated. For example, in partition #1, 30% of the trait's SNP-based heritability is contributed by the variants (black dots) in the partition (red bar). In partition #2, only 12% of the SNP-based heritability is contributed (blue bar), owing to the lack of SNP association signal in this part of the genome. The proportion of SNP-based heritability in a window is normalized by the total proportion of SNPs in the window (over all SNPs considered in the genome). For example, partition #1 accounts for 10% of the SNPs in the "genome" while partition #2 is larger and accounts for 24% of the SNPS in the "genome". The equation for partitioned heritability used by S-LDSC[75] is depicted and the calculation for each example is shown. Generally, heritability enrichment indicates that genetic variants in the regions are more associated with phenotypic variation in the trait than expected given a null hypothesis of polygenicity (Example #1). Heritability depletion means that the variants associated with less phenotypic variation than expected (Example #2). Throughout the dissertation, heritability enrichment is indicated with red color or elevation above the baseline of one (no enrichment). Heritability depletion is indicated with blue or depression from the baseline.

differences between archaic and modern humans. In modern and archaic hominins respectively, there are only 42 and 167 fixed non-synonymous single nucleotide changes[17]. Furthermore, they tend to have less predicted functional effect than mutations that arose on the modern human lineage[57]. Thus, similar to the paradox from early comparisons between chimpanzees and humans, these protein-coding changes alone cannot explain the phenotypic divergence between modern and archaic humans.

Instead, the phenotypic evolution of hominins is largely driven by changes in the non-coding genome that affect the regulation of conserved proteins[7–10]. However, because of ancient sample degradation, gene expression in archaic hominins cannot be directly assayed[94]. Previous studies have used diverse approaches to understand the gene regulatory differences between modern and archaic humans. Many studies that have considered the effect of archaic variants on gene expression leverage genomes of modern Eurasians with Neanderthal ancestry. These investigations have found widespread expression differences between Neanderthal and human alleles. One-quarter of Neanderthal haplotypes inherited by humans show *cis*-

regulatory effects[55] and these introgressed alleles contribute more to expression variation than expected[57].

Yet, these studies have major limitations. They only investigate regions where modern humans have remaining Neanderthal ancestry. Some regions of modern human genomes have little or no evidence of introgression largely due to negative selection on deleterious or incompatible haplotypes within a few generations after hybridization[44,45,53–57]. Thus, considering only introgressed variation provides a very limited view into hominin biology and the differences between archaic and modern humans. It also limits our ability to understand why certain regions of MH genomes tolerated Neanderthal DNA while others did not. Colbran et al.[94] addressed this challenge by imputing AH gene regulation genome-wide through models trained on gene expression data in MHs[95]. They estimated that over 1900 genes had different patterns of regulation between AHs and MHs. However, the molecular mechanisms through which archaic variants alter gene expression remain unclear.

Gokhman et al.[96] and Batyrev et al.[97] aimed to elucidate these mechanisms by computationally reconstructing maps of archaic DNA methylation. They found 2,000 differentially methylated regions that associate with genes predominantly related to facial and limb anatomy. Silvert et al.[98] evaluated the overlap of archaic variants with enhancers, promoters, and miRNAs and found links to adipogenesis and cancer susceptibility. Together, these illustrate the potential to mechanistically link archaic genotypes with regulatory functions via the prediction of molecular phenotypes.

1.3 Form and function of the 3D genome

> *"Whether it be the sweeping eagle in his flight, or the open apple-blossom, the toiling workhorse, the blithe swan, the branching oak, the winding stream at its base, the drifting clouds, over all the coursing sun, form ever follows function, and this is the law."*
>
> — Louis Henry Sullivan. "The Tall Office Building
> Artistically Considered." (1896).[99]

1.3.1 3D chromatin organization and consequences

Given the importance of gene regulation in interpreting both inter- and intra-species diversity, it is critical to consider the physical form of how transcription control occurs within the cell nucleus. Despite the discovery of chromosome territories in 1909, the physical folding of the genome has only more recently been explored as an integral part of genome function[100,101]. During interphase, the human genome is organized in three-dimensional (3D) nuclear space to allow for proper genome replication and transcription[102]. Chromatin contact maps are experimentally measured using chromosome-conformation-capture technologies (3C, 4C, 5C, Hi-C, MicroC)[103–107], which allow for quantification of genome folding at various resolutions. These range from large-scale chromosome territories to small-scale local structure, like loops and "architectural stripes", which can reflect gene regulatory function by enhancers[105,108–110].

At an intermediate sub-chromosomal scale, chromatin physically compartmentalizes into topologically associating domains (TADs)[105,111–113]. TADs are megabase-long genomic regions that self-interact, but rarely contact regions outside the domain[105,111–114] (Fig. 1.3A). They are likely formed and maintained through interactions between protein complexes including CTCF zinc-finger transcription factors and cohesin ring-shaped complexes[102,105]. TADs have been characterized as basic units of 3D genome structure given their stability across cell divisions[114,115], cell-types[111,112,116–118], and syntenic sequences between species[105,111,119–121].

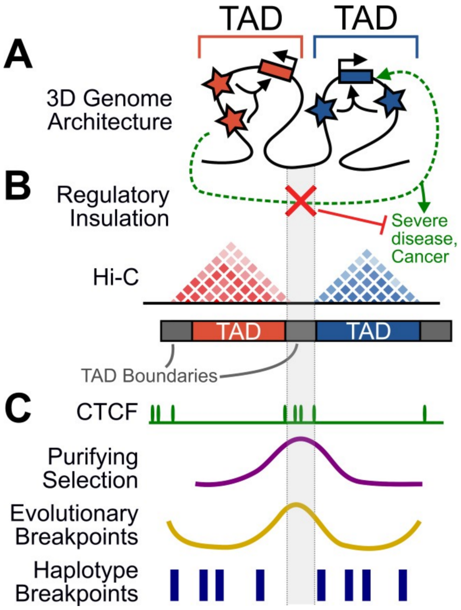


Figure 1.3: Topologically associating domains (TADs) are fundamental units of 3D genome nuclear organization.
(A) TADs are megabase-long regions of chromatin organization that self-interact, but rarely contact regions outside the domain. TADs facilitate communication between genes (rectangle) and enhancers (stars) via physical looping. **(B)** The regions insulating TADs (gray vertical bar) are TAD boundaries. They contribute to the regulation of gene expression by restricting interactions (red "x") of cis-regulatory sequences to their appropriate target genes. TAD structure disruption via large structural variants (SV) has been implicated in rare disease and cancer parthenogenesis[122]. **(C)** We summarize the findings of previous work which has begun to characterize the functional and constraint landscape across the 3D genome. CTCF binding and clustering are enriched at boundaries[102,105]. TAD boundaries have more evidence of purifying selection on SV compared to TADs. Boundaries are enriched for syntenic breaks[121,123]. Finally, human haplotype breakpoints do not align with chromatin boundaries (Fig. 1.3C)[124].

Although 3D genome organization is essential to many aspects of normal cell physiology including cell-type identity, differentiation, and replication timing, one of its primary roles is in facilitating enhancer-promoter interactions to regulate gene expression[125–131]. TADs modulate gene regulation by restricting interactions of cis-regulatory sequences, like enhancers, to their target genes[105]. Specifically, regions between TADs—termed TAD boundaries—have insulatory properties. TAD boundaries prevent "enhancer hijacking": a process where enhancers act on an inappropriate target gene. Removing insulatory TAD boundaries leads to ectopic gene expression both *in vitro* and *in vivo*. For example, TAD structure disruption at the *EPHA4* locus leads to inappropriate rewiring of developmental genes implicated in limb formation defects[105,122,132]. In cancer, large structural alterations that disrupt TAD boundaries cause pathogenic gene expression in acute myeloid leukemia (AML) and medulloblastoma[133,134] (Fig. 1.3B). TADs and boundaries are formed through the binding and interaction of several factors including CTCF and cohesin; consequently, altering CTCF binding promotes oncogenic gene expression in gliomas[135]. Structural variation (SV) that disrupts TAD structure causes gain-of-function, loss-of-function, and misexpression in many forms of rare neurodevelopmental disease[122].

Although we are beginning to understand the role of 3D genome disruption in rare disease and cancer, the relationship between common disease and 3D genome architecture has not been investigated. Even common small-scale variation (e.g., SNPs) influences 3D genome structure[136], for example, through modifying CTCF-binding site motifs necessary for TAD formation. However, the contribution of common variation in different 3D contexts to complex traits is unknown.

1.3.2 Relationship between evolution and the 3D genome

Previous studies disagree about the functional importance and evolutionary pressures across the landscape of the 3D genome. TAD organization has two basic features: the “self-associating” TAD and the “insulatory” TAD boundary[105]. TADs are functional units; their disruption is often deleterious. For example, across species, enrichment of syntenic breaks at TAD boundaries suggests an evolutionary preference for mutations that “shuffle” intact TADs, rather than “break” them[121,123] (Fig. 1.3C). Additionally, TADs often contain clusters of co-regulated genes—e.g., cytochrome genes and olfactory receptors[105]. Together, these suggest that TADs are more functionally important and evolutionarily constrained than boundaries.

In contrast, other studies have highlighted the greater importance of TAD boundaries. SVs that disrupt TAD boundaries are implicated in rare disease and cancer[105,122,133–135]. Accordingly, TAD boundaries have evidence of purifying selection on SV (Fig. 1.3C). Moreover, boundaries shared between two cell-types experience stronger purifying selection than unique boundaries, suggesting that shared boundaries are more intolerant of disruption[137]. TAD boundaries are also enriched for housekeeping genes and transcription start sites (TSS)[105,111]. Finally, human haplotype breakpoints do not align with chromatin boundaries, which indicates that recombination may be deleterious at TAD boundaries[124](Fig. 1.3C). Collectively, these suggest that TAD boundaries are more functionally important, especially at the scale of human evolution.

1.3.3 Predictions of 3D genome folding

Understanding the mechanisms, function, and evolution of 3D genome folding has parallels in another decades-long “folding” challenge. Namely, the “protein folding problem” addresses how an amino acid sequence (rather than DNA sequence) encodes a protein’s 3D atomic structure (rather than chromatin’s 3D nuclear structure)[138]. The protein folding problem encompasses many related sub-puzzles[138]. First, what is the “code” that determines folding? Second, how does the folding physically occur *in vivo*? Third, how can the structure be predicted from sequence computationally? Finally, how does the form relate to function? These questions first emerged around 1960. Since then, both experimental data and methods have seen big advances. With the development of AlphaFold2—a deep learning approach that combines physio-chemical knowledge with multiple-sequence alignments—some have deemed the protein folding problem solved[139]. Despite the continued potential for improvement, comprehensive knowledge of the protein-folding code provides necessary context to understand the relationship between form and function for interpretation of disease variants and evolutionary constraint across proteins.

The "chromatin folding problem" may follow a parallel trajectory towards a similar ultimate goal of improving variant interpretation and functional understanding of non-coding genome biology. Yet, work on this complex problem is in its early stages. Chromatin folding is likely facilitated by a complex interplay of transcription factors. While DNA only encodes for 20 amino acids, it encodes for many more transcription factor motifs that do not respect the same tri-nucleotide codon template. Transcription factors and chromatin folding are also cell type specific. Furthermore, protein folding can be compartmentalized to the scale of a single protein or multimer. In contrast, chromatin interactions can occur at every pair of sites across the entire genome, even across different chromosomes. This is immense in scope with 4.7×10^{18} possible pair-wise interactions across our 3.055 billion base pair genome. Finally, although protein folding is dynamic, there is evidence that chromatin folding is transient and exists in multiple configurations. Current experiments are limited in both their resolution and their ability to measure dynamics.

A variety of both polymer-based and statistical approaches have been developed to address these challenges[140]. Some incorporate other experimental data (e.g., epigenetic data, transcription factor binding) to inform their predictions. Challenges include the uncertainty of specifying biophysical parameters needed for modelling, high computational demands, non-linear dependencies between data used to predict structure, and lack of availability of high-quality experimental data in the proper cellular context[140]. Recently, deep learning methods have been developed to learn the sequence "grammar" underlying 3D genome folding patterns[140–143]. Because they predict 3D organization from sequence alone, they avoid the need for additional experimental data or pre-specified biophysical parameters. Using a convolutional neural network (CNN) allows the model to "learn" non-linear combinations of motifs, without any *a priori* knowledge, to predict chromatin contact maps. Because the molecular mechanisms governing genome organization, like CTCF binding and co-localization with cohesin, are largely evolutionarily conserved[111,121], models trained using human data perform well even when applied to DNA sequences from mice[141]. Thus, unlike genome-wide methods for predicting organism-level phenotype (e.g., polygenic risk scores), these models can likely be applied across diverse hominins to provide insight into the role of 3D genome folding in disease and evolution.

1.4 Overview

Understanding the consequence of any of the millions of variants in an individual genome[78]—modern or archaic—requires context of human evolutionary history coupled with a mechanistic knowledge of genome function. Given the established importance of gene regulation combined with the emerging role of chromatin folding in gene regulation, there is a critical need to integrate 3D genome architecture into genome interpretation both for the understanding of evolution and disease risk.

In summary, this dissertation will leverage techniques in human genetics, functional genomics, evolutionary biology, and machine learning to interrogate the relationship between recent human evolution, 3D genome organization, gene regulation, and complex human disease with the following aims:

- Comprehensively quantify the contribution of Neanderthal ancestry to diverse human traits. (Chapter 2)

- Map the relationship between 3D genome architecture and the genetic architecture of complex traits (Chapter 3)
- Reconstruct the 3D genome organization of Neanderthals to evaluate how chromatin folding shaped human evolution (Chapter 4)

In order to ultimately assess the 3D genome contribution to the phenotypic diversity of modern and archaic humans, we must first comprehensively catalog the influence of archaic ancestry on traits in modern humans today. Given the pervasiveness of archaic ancestry in modern human genomes and the complexity of most evolutionary-relevant traits, studies of individual loci one-by-one have not sufficiently addressed the overall influence of Neanderthal admixture on human traits. In Chapter 2, I use partitioned heritability to comprehensively quantify the contribution of Neanderthal ancestry to over 400 traits. Integrating the results, I propose a model for using trait heritability and direction of effect to understand how selection acted on different traits and how introgression may have facilitated adaptation to non-African environments.

Despite evidence that archaic ancestry influences certain traits in modern humans, elucidating mechanisms through which variation contributes to traits remains difficult. One such putative mechanism is via the 3D genome. Although 3D genome disruption has been implicated in rare disease, I evaluate the role of the 3D genome in complex traits in Chapter 3. Additionally, I synthesize 3D genome maps across cell types to quantify their functional and evolutionary conservation to help resolve conflicting evidence about evolutionary pressures across the 3D genome landscape.

Given the established importance of the 3D genome organization to both rare disease and complex traits, in Chapter 4, I computationally resurrect the 3D genome organization of Neanderthals and Denisovans using deep learning models that predict 3D genome folding patterns. Because these models can be applied to novel sequences, it overcomes limitations of previous work that only investigated genomic regions where present-day humans have remaining Neanderthal ancestry. Using the resulting chromatin contact maps, I demonstrate how differences in 3D genome folding between archaic and modern humans provide a putative molecular mechanism for the phenotypic differences between the species. Finally, comparisons with contact maps across modern humans highlight that 3D genome organization constrained sequence divergence and patterns of introgression in hominin evolution.

Together, mapping these relationships quantifies the functional contribution of variants in different 3D contexts. This will provide a framework to ask previously unanswerable questions about the role of the 3D genome in human evolution and disease. Ultimately, this work highlights the 3D genome as a mechanism linking genotype and phenotype, both within present-day humans and across hominins.

CHAPTER 2

QUANTIFYING THE CONTRIBUTION OF NEANDERTHAL INTROGRESSION TO THE HERITABILITY OF COMPLEX TRAITS[*]

2.1 Introduction

Anatomically modern humans (AMH) interbred with archaic hominin groups on multiple occasions and in several locations over the past 50,000 years. As a result, nearly all Eurasians have approximately 2% Neanderthal ancestry resulting from interbreeding events that occurred shortly after their ancestors left Africa[14,28]. Analyses of available genome-wide association studies and large-scale biobank data revealed that alleles of Neanderthal ancestry are associated with diverse traits in modern Eurasians[14,58,59,62]. However, due to limited phenotype data and technical challenges quantifying associations between archaic alleles and traits[31,58], previous studies have not comprehensively characterized the genome-wide influence of Neanderthal introgression on modern human diseases and traits.

Archaic admixture may have facilitated the ability of AMH to inhabit diverse environments as they spread around the globe[33]. Despite the potential benefits of admixture, simulations and empirical analyses of the distribution of introgressed alleles across the genome suggest that they were largely deleterious in AMH[51,52]. Given the broad evidence for negative selection against alleles of Neanderthal ancestry in functional regions coupled with evidence of positive selection on specific introgressed Neanderthal alleles, there is a need to more comprehensively characterize and reconcile the functional effects of introgressed alleles on variation in diverse AMH traits. Previously, the legacy of introgression in AMHs has been primarily characterized based on overlap with molecular annotations[44,45,53,55] or existing genome-wide association study (GWAS) hits[14,59,62]. However, most medically and evolutionarily relevant traits are complex, with hundreds or thousands of loci across the genome contributing to them[65,66]. Thus, studies of individual loci are not sufficient to address the overall influence of Neanderthal admixture on human traits.

Here, we leverage recent maps of Neanderthal ancestry[32] with new techniques to characterize the contribution of Neanderthal introgression to the heritability of common complex traits[75,90] and identify trends in introgressed variants' direction of effect on these traits. Using well-powered GWASs for 405 diverse traits from existing studies and the UK Biobank[144], we estimate trait heritability in genetic variation in regions of the human genome in which detectable Neanderthal ancestry remains and in introgressed Neanderthal variants themselves. This broad view of the influence of Neanderthal ancestry genome-wide supports selection against Neanderthal ancestry in regions of the genome that influence nearly all complex traits. However, it reveals that common introgressed Neanderthal alleles, especially those shared across Neanderthal populations, have a greater than expected effect on several traits with potential relevance for AMH adaptation into non-African environments. Integrating our results, we propose a framework (see 2.3:Discussion) for using trait heritability and direction of effect in introgressed regions to understand how selection acted on different

[†] This chapter is adapted from McArthur, E., Rinker, D.C., & Capra, J.A. 2021. *Nat. Comms.* and has been reproduced with permission of the publisher and co-authors D.C. Rinker and J.A. Capra.

traits and how introgression may have facilitated adaptation to non-African environments.

2.2 Results

2.2.1 Genomic regions with Neanderthal ancestry are depleted of complex trait heritability

To quantify the relationship between the heritability of complex traits and Neanderthal introgression, we first investigated genomic regions where detectable Neanderthal ancestry remains in some AMHs. Hereafter, we will refer to these as "regions with Neanderthal ancestry" (Fig. 2.1A). We consider introgressed regions in Europeans identified by the Sprime algorithm. This algorithm identifies regions in individuals' genomes that contain a high density of single nucleotide variants absent in unadmixed African populations and that frequently match Neanderthal alleles[32]. Filtering for introgressed regions matching the Altai Neanderthal genome, we identified 1345 segments of the human genome with remaining Neanderthal ancestry that have a median length of 299 kb (IQR: 174 – 574 kb), covering 19% of the genome (2.4:Methods, Fig. 6.1). This high confidence set reflects the state-of-the-art, but likely does not include all regions with Neanderthal ancestry; some archaic fragments are too short or too similar to non-archaic fragments to detect. As more modern and archaic individuals are sequenced, additional regions in AMHs with Neanderthal ancestry may be detected. We also separately considered introgressed segments defined based on comparison to the Vindija Neanderthal and using the S* algorithm (Figs. 6.2,6.3).

To estimate the contribution of variation in regions with Neanderthal ancestry to trait heritability, we conducted partitioned heritability analysis using stratified LD score regression (S-LDSC). S-LDSC quantifies the heritability of a trait explained by common (minor allele frequency [MAF] $> 5\%$) variants in a set of regions of interest, explicitly conditioned on the association statistics and the underlying linkage disequilibrium (LD) structure[75,90]. To start, we considered summary statistics from a curated representative set of 41 diseases and complex traits with high-quality GWAS used in previous S-LDSC analyses (average number of individuals [N] = 329,378; SNPs in GWAS [M] = 1,155,239; $h^2_{SNP} = 0.19$; Table 6.1)[92,144–154].

In this context, heritability depletion indicates that genetic variants in regions in which some individuals have Neanderthal ancestry are less associated with phenotypic variation in the trait than expected given a null hypothesis of complete polygenicity. Heritability enrichment means that the variants associate with more phenotypic variation than expected. Heritability enrichment (or depletion) in a set of variants provides evidence of functional relevance for the region to the trait and suggests the action of selection[155] (see model in 2.3:Discussion). Regions with Neanderthal ancestry are broadly depleted of variation that contributes to complex trait heritability (Fig. 2.1B). These regions are 1.10-fold (i.e. 10%) depleted for contribution to trait heritability compared to the heritability expected from the background genome (two-tailed one-sample t-test $P = 8 \times 10^{-7}$, 95% confidence interval [CI]:1.07–1.14). Most variants segregating in Eurasian populations in regions of the genome with Neanderthal ancestry are not of Neanderthal origin (Fig. 2.1A); yet, even after removing introgressed variants (LD expanded to $r^2 > 0.5$ [2.4:Methods]), these regions are still 1.06-fold depleted for trait heritability ($P = 0.003$, CI:1.02–1.10). The heritability depletion observed after removing introgressed variants (and those in LD with them) suggests that introgressed variants account for some, but not all, of the heritability depletion in these regions. The depletion across traits also holds for introgressed haplotypes identified by the earlier S* method (Fig. 6.2A) and based on matching the Vindija

Neanderthal (rather than Altai) genome (Fig. 6.3A)[156]. Previous studies have shown that regions with Neanderthal ancestry have less evidence for evolutionary constraint and function at the molecular level[53,55,56]. Our results demonstrate that regions of the genome that retain Neanderthal ancestry are also depleted for variation influencing a diverse array of complex traits.

We find three exceptions to the complex trait heritability depletion: sunburn, skin color, and tanning (Fig. 2.1B). In contrast to all other traits, regions with Neanderthal ancestry are not depleted for heritability of these traits ($P = 0.3$–0.4). These three traits are genetically correlated with magnitudes between $r = 0.55$ and 0.86. Several previous hypotheses suggest that the introgression of Neanderthal alleles related to hair and skin pigmentation could have provided non-African AMHs with adaptive benefits as they moved to higher latitudes[44,45,59,62]. Our results suggest that introgressed Neanderthal haplotypes were not selected against in regions of the genome involved in skin pigmentation, in contrast to regions associated with other traits.

2.2.2 Neanderthal introgressed variants are depleted for heritability of most complex traits

In the previous section, we demonstrated that non-introgressed variants in regions with remaining Neanderthal ancestry are depleted for heritability of most complex traits. We now focus on the heritability contributed by introgressed variants specifically.

We quantified the relationship between heritability of the representative 41 complex traits and several sets of common Neanderthal-introgressed variants with different evolutionary histories. The largest set included all variants with evidence of introgression in any Eurasian population according to Sprime[32] ($N = 900,902$, 2.4:Methods); this set will be referred to as “introgressed variants” throughout the manuscript. This set includes not only high-confidence Neanderthal-origin introgressed variants, but also ancestral alleles lost in Africans that were reintroduced to Eurasians through archaic introgression[58], variants with origins in other archaic hominins such as Denisovans, and possibly variants tightly linked to introgressed haplotypes that arose in Eurasians shortly after introgression. The most stringent and high-confidence sets include Neanderthal-introgressed alleles that are observed in Europeans and explicitly match the either the Altai genome ($N = 138,774$) or the Vindija genome ($N = 167,927$, 2.4:Methods); these sets will be referred to as “Altai-matching” and “Vindija-matching” introgressed variants, respectively. We calculated partitioned heritability on these sets and two other intermediate-stringency sets (2.4:Methods); results from all sets are in Fig. 6.4.

Consistent with our observations on non-introgressed variants in regions with Neanderthal ancestry (Fig. 2.1B), the set of all introgressed variants is 1.28-fold depleted for contribution to trait heritability (two-tailed one-sample t-test $P = 0.0004$, CI: 1.13-1.45). (Fig. 2.1C, Table 6.2). We observed the strongest depletion for heritability for cholesterol level (4.7-fold depleted, CI:1.82–12.1, $q = 0.02$ after Benjamini-Hochberg FDR-correction at the 0.05 level), platelet count (1.7-fold depleted, CI:1.18–2.42, $q = 0.04$), systolic blood pressure (1.6-fold depleted, CI:1.22–2.01, $q = 0.01$), years of education (1.5-fold depleted, CI:1.14–1.96, $q = 0.04$), and body mass index (BMI, 1.5-fold depleted, CI:1.18–1.89, $q = 0.02$). Due their distinct evolutionary histories, introgressed variants have a different allele frequency distribution than other sets of common variants; however, this difference is not responsible for the number of significantly depleted

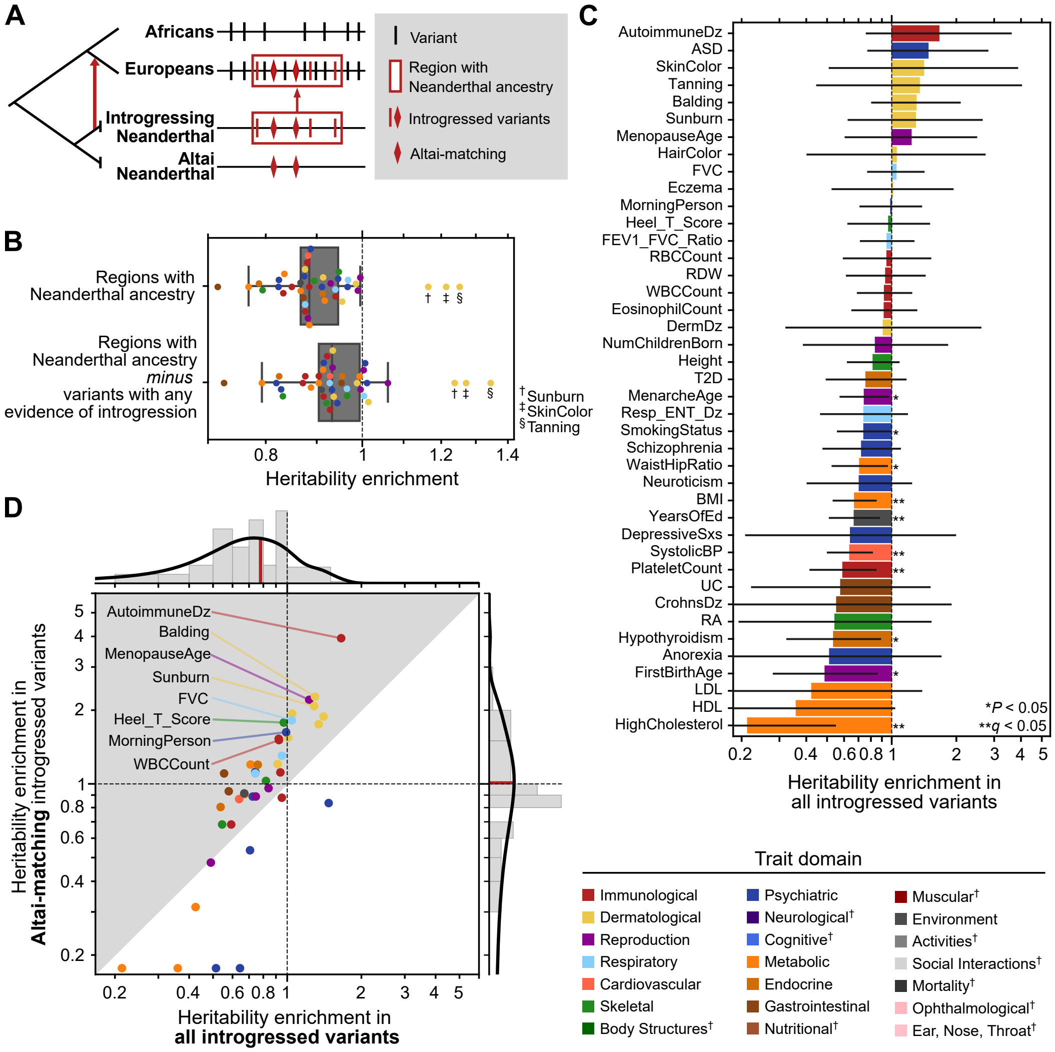


Figure 2.1: Complex trait heritability is broadly depleted in regions with Neanderthal ancestry and in introgressed variants.

(A) We focus on variants in regions of the human genome with remaining Neanderthal ancestry (red box). These variants (vertical lines and diamonds) have multiple evolutionary histories: most are segregating on non-introgressed haplotypes (black), many are present in Eurasians due to introgression (red), and some of these introgressed alleles were shared among multiple Neanderthal populations including both the Altai Neanderthal and the introgressing Neanderthal population (diamonds). (B) Regions of the genome where Neanderthal ancestry remains (all variants in the red box in A) are depleted for heritability of 41 diverse complex traits (mean: 1.10 fold-depleted, two-tailed one-sample t-test $P = 8 \times 10^{-7}$) except for sunburn, skin color, and tanning. Each dot represents the heritability enrichment or depletion for a single trait estimated by stratified LD score regression (S-LDSC). Removing introgressed variants (red lines and diamonds in A, LD expanded to $r^2 > 0.5$), these regions are still broadly depleted for trait heritability (mean: 1.06 fold-depleted, two-tailed one-sample t-test $P = 0.003$). The boxplot centers represent medians, the boxes are bounded by the first and third quartiles, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. (C) Introgressed variants (red lines and diamonds in A) contribute varying levels of heritability to different 41 traits. Most (76%) traits trend towards heritability depletion in introgressed variants (one-tailed Binomial test $P = 0.007$). Bars for individual traits represent heritability enrichment estimates with 95% confidence intervals which are calculated by S-LDSC standard errors using a block jackknife ($n = 200$). Traits are colored by their domain (legend); marked domains appear in later figures. These colors will be used in all figures. (D) Altai-matching introgressed variants (y-axis, diamonds in A) are more enriched for heritability than all introgressed variants (x-axis) for 78% of traits (one-tailed Binomial test $P = 0.0002$, 1.02x vs. 0.78x). Traits with depletion below 0.125 are plotted on the y-axis baseline. These patterns are consistent when considering the Vindija Neanderthal (Fig. 6.3).

traits we observe (Fig. 6.5).

2.2.3 Older introgressed variants contribute more trait heritability

The Altai-matching set contains alleles that originated in the Neanderthal lineage and were likely common among diverse Neanderthal groups given the substantial genetic, geographical, and temporal divergence of the Altai Neanderthal from the introgressing population[14,15]. However, it excludes many true introgressed Neanderthal alleles, such as those that were not present in the Altai Neanderthal. The Vindija Neanderthal was closer to the introgressing population, so the Vindija-matching set additionally includes many younger Neanderthal alleles, as does the set of all introgressed variants (Fig. 6.6).

Despite the overall depletion for complex trait heritability in regions of the genome with introgression (Fig. 2.1B; 1.10-fold depleted, $P = 8 \times 10^{-7}$) and in all introgressed variants (Fig. 2.1D; 1.28-fold depleted, $P = 0.0004$), the trait heritability in Altai-matching variants is not depleted (Fig. 2.1D, 1.02-fold more heritability contribution, $P = 0.9$). The Altai results are very similar to partitioned heritability estimates when introgressed variants are identified using the S* approach ($r^2 = 0.79$), suggesting their robustness to technical variation (Fig. 6.2B-C)[156]. The heritability enrichments for Vindija-matching variants across traits are highly correlated with those for the Altai-matching variants ($r^2 = 0.93$, Fig. 6.3B-C). However, Altai-matching variants contribute more heritability than Vindija-matching variants to 66% of traits (one-tailed Binomial test $P = 0.03$, Fig. 6.3C).

The greater contribution of Altai-matching variants to trait heritability compared to all introgressed variants and Vindija-matching variants supports our hypothesis that older variants that were shared among multiple Neanderthal populations were more tolerated after introgression. On average across the 41 traits, 79.2% (CI: 73.6-84.8%) of nominally trait-associated introgressed variants are observed in the Altai Neanderthal ($P = 1 \times 10^{-4}$, pruned associations with $r^2 = 0.5$). However, we note one exception: only 50% of the Crohn's disease risk-associated variant clusters (two of four) are present in Altai; the remainder are likely younger as they are observed only in the Vindija Neanderthal ($P = 4 \times 10^{-13}$, Fig. 6.7); these contribute to the increased heritability enrichment for Crohn's disease in Vindija-matching variants compared to other introgressed variants (Fig. 6.3C, Supplemental Text 6.1 in Appendix 1).

Finally, we hypothesize that selection contributed to the heritability enrichment observed for certain traits. Supporting this, we find that high-frequency introgressed variants (MAF $> 21\%$) contribute more to heritability enrichment than rarer variants (Fig. 6.8, Supplemental Text 6.1 in Appendix 1) and that many genomic windows contributing to the heritability enrichment of sunburn risk and white blood cell count overlap introgressed haplotypes predicted to be adaptive (Fig. 6.9, Supplemental Text 6.1 in Appendix 1). Together, these findings suggest that selection acted differently on Neanderthal variation with specific histories (older vs. younger) and differently across traits.

2.2.4 Neanderthal introgressed variants are most enriched for heritability of dermatologic traits and most depleted for cognitive traits

To evaluate heritability trends across more traits and bodily systems, we analyzed GWAS summary statistics for 405 traits from the UK BioBank and FinnGen divided into domains, chapters, and subchapters from the

GWAS Atlas (2.4:Methods)[144,157–159]. We performed partitioned heritability analysis on these traits using the sets of Neanderthal-introgressed variants described above (Fig. 2.2, Figs. 6.10-6.12).

In this diverse set of traits, Altai-matching introgressed variants are most enriched for heritability of dermatologic (hair-related) traits (2.7-fold enriched, CI: 2.4–3.1, $q = 0.04$, two-tailed one-sample t-test) and most depleted for cognitive (higher-level cognitive and memory functions) traits (2.0-fold depleted, CI:1.4–2.7, $q = 0.04$) (Fig. 2.2A, Table 6.3). We also observed heritability enrichment in traits related to body structure (e.g. fractures, dental diseases, 1.9-fold enriched, CI:1.3–2.8, $q = 0.06$), endocrine (1.7-fold enriched, CI:1.4–2.2, $q = 0.11$), respiratory (1.3-fold enriched, CI:1.1–1.5, $q = 0.05$), and the skeletal system (1.1-fold enriched, CI:1.0–1.3, $q = 0.1$). Traits related to eye structure (1.8-fold depleted, CI:1.2–3.1, q = 0.04), environment (1.5-fold depleted, CI:1.1–2.1, $q = 0.05$), and daily activities (1.3-fold depleted, CI:1.0–1.7, $q = 0.07$) are depleted in addition to cognitive traits. The depletion in cognitive traits suggests that the previously observed strong depletion for Neanderthal alleles in regulatory regions active in the brain may be due to effects on brain-related complex traits[55,56,58].

Other trait domains exhibit substantial intra-domain diversity in the heritability patterns with some traits showing strong enrichment and others showing depletion in Altai-matching introgressed variants. Thus, we also quantified enrichment and depletion for traits at the more granular chapter and subchapter levels. Dividing immunologic traits into subchapters, Altai-matching variants contribute more heritability to WBC-related traits (1.3-fold enriched, CI:1.0–1.6) than to RBC-related traits (1.5-fold depleted, CI:1.0–2.4) ($P = 0.02$, two-tailed two-sample t-test, Fig. 2.2B). For skeletal traits, bone mineral density-related traits show the most enrichment for heritability in introgressed variants (1.2-fold enriched, CI:1.1–1.4, $q = 0.01$, Fig. 2.2C). For reproductive traits, puberty- and menstruation-related traits are enriched for heritability (1.5-fold enriched, CI:1.0–2.2, $q = 0.10$), whereas sexual and procreation functions are depleted (1.5-fold depleted, CI:1.2–2.0, $q = 0.05$, Fig. 2.2D), possibly reflecting reproductive barriers to introgression. For psychiatric traits, tobacco use disorders trend towards enrichment (1.2-fold enriched, CI:1.0-1.5, $q = 0.23$), consistent with previous observations, while introgressed variants are depleted for contribution to personality-related functions (1.4-fold depleted, CI:1.1–1.8, $q = 0.07$, Fig. 2.2E)[59,62]. Domain, chapter, and subchapter-level results across all traits for all the sets of introgressed variants are in Figs. 6.10-6.12 and Tables 6.3-6.5.

2.2.5 Neanderthal alleles confer directional effects for some traits

Partitioned heritability analyses quantify the overall contribution of introgressed loci to variation in traits across humans; however, they do not test for consistent directional effects on a trait across introgressed loci. We now test whether introgressed alleles consistently have effects in the same direction (e.g., mostly risk increasing) for eight traits spanning phenotypic domains for which Altai-matching introgressed variants contributed more heritability than expected (AutoimmuneDz, Balding, Sunburn, FVC, Heel_T_Score, MorningPerson, MenopauseAge, WBCCount, Fig. 2.1C). We quantify Neanderthal introgressed allele direction of effect in two ways.

First, focusing on the trait-associated variants with the strongest effects, we intersected Altai-matching introgressed alleles with associated variants from the eight GWAS. We then quantified if there is an overrepresentation of introgressed alleles in the risk-increasing or risk-decreasing direction. We considered GWAS

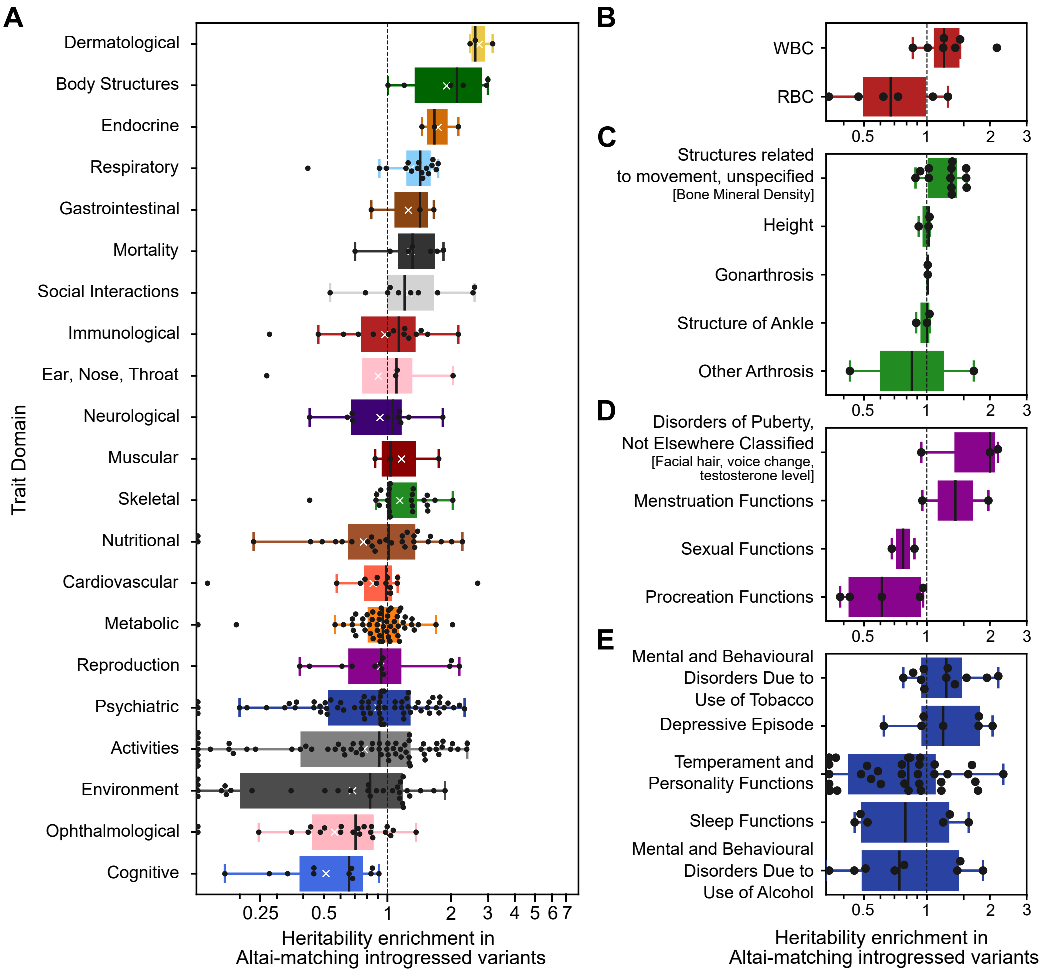


Figure 2.2: Heritability enrichment and depletion in introgressed variants across 405 traits clustered by domain.

(A) Altai-matching introgressed variants are most enriched for dermatological traits (hair-related) and most depleted for cognitive traits (higher-level cognitive and memory functions). Each point represents heritability enrichment or depletion of one trait among Altai-matching introgressed variants. Traits with depletion less than 0.125 are plotted on the baseline for visualization. Within some domains, introgressed variants also show variable heritability enrichment. (B) Dividing immunologic traits into subchapters, Altai-matching introgressed variants contribute more to heritability of WBC-related traits (1.3-fold enriched, $n = 7$) and less to RBC-related traits (1.5-fold depleted, $n = 6$) ($P = 0.02$, two-tailed two-sample t-test). (C) For skeletal traits, bone mineral density-related traits show the most enrichment for heritability in introgressed variants (1.2-fold enriched, $q = 0.01$, $n = 12$). (D) For reproductive traits, puberty- and menstruation-related traits are enriched for heritability (1.5-fold enriched, $q = 0.1$, $n = 5$), whereas sexual and procreation functions are depleted (1.5-fold depleted, $q = 0.05$, $n = 7$). (E) For psychiatric traits, tobacco use disorders trend towards enrichment (1.2-fold enriched, $q = 0.23$, $n = 11$), consistent with previous observations, while introgressed variants are depleted for contribution to personality-related functions (1.4-fold depleted, $q = 0.07$, $n = 35$) Unless otherwise specified, all q values are from two-tailed one-sample t-tests with Benjamini-Hochberg FDR-correction at the 0.05 level. The domain, chapter, and subchapter-level results across all traits are similar when considering other sets of introgressed variants (Figs. 6.10-6.12, Tables 6.3-6.5). The boxplot centers represent medians, the white Xs denote means, the boxes are bounded by the first and third quartile, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box.

variants with $P < 1 \times 10^{-8}$ and pruned variants in perfect LD ($r^2 = 1$) to reduce redundant counts due to linked variants. Results from using less strict thresholds ($P < 5 \times 10^{-8}$, $P < 1 \times 10^{-6}$ and $r^2 > 0.8$, $r^2 > 0.5$) show consistent directions of effect with some modest differences in the strength of directionality (Fig. 6.13).

Four traits show a difference ($q < 0.05$, one-tailed χ^2 goodness of fit test) in the direction of effect of introgressed variants: balding, menopause age, forced vital capacity, and morning person (Fig. 2.3A). Respectively, Altai-matching introgressed alleles were more associated with hair loss ($q = 0.01$, less Type 1 Balding), younger age at menopause ($q = 0.04$), larger lung volumes (FVC, $q = 0.03$) and increased likelihood of being a morning person ($q = 0.03$). Additionally, introgressed alleles may be more likely to be associated with increased bone density ($q = 0.19$) and with increased sunburn risk ($q = 0.21$), which would support previous findings, but requires further validation.

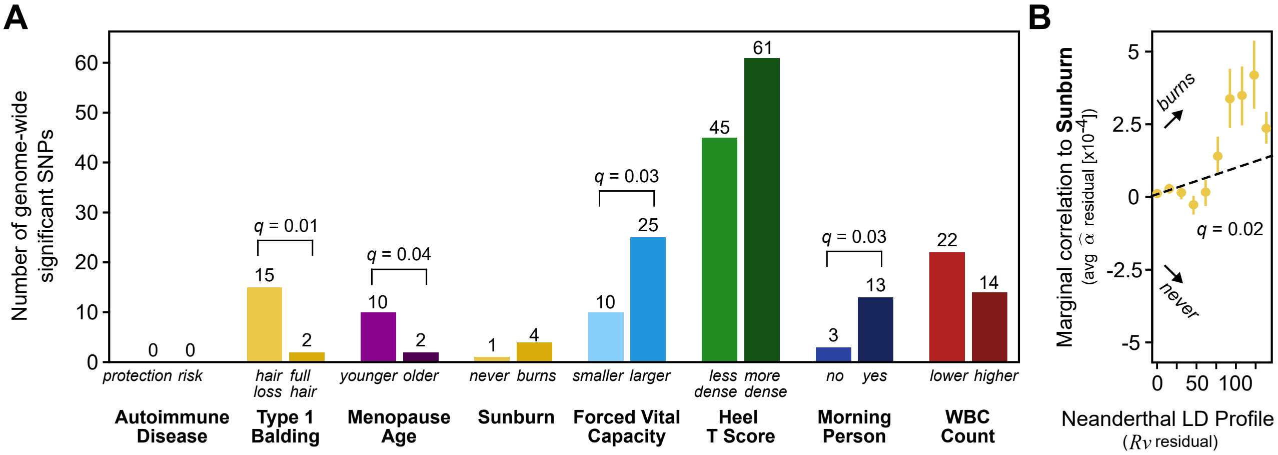


Figure 2.3: Neanderthal alleles confer directional effects for some traits.
For eight traits with heritability enrichment in Altai-matching introgressed variants (Fig. 2.1D), we assessed the direction of effect of the Neanderthal alleles with two approaches. The first intersects introgressed Altai-matching Neanderthal alleles (LD-expanded to $r^2 = 1$) with strongly associated ($P < 1 \times 10^{-8}$) variants from each GWAS. (A) For each trait, we plot the number of variants by the direction of effect of the Neanderthal allele. Variants in perfect LD ($r^2 = 1$) are pruned. Four traits show a significant difference ($q < 0.05$, one-tailed χ^2 goodness of fit test) in direction of effect: increased balding, younger menopause age, increased forced vital capacity, and morning person. For example, of the 17 Neanderthal alleles associated with balding, 15 are associated with hair loss and only two with full hair. Sunburn, Heel T score, and WBC count also show modest biases. (B) This second approach, signed LD profile regression, considers the direction of effect over all variants ($n = 1,187,349$), not just those with the largest effects. For each variant, we compute the marginal correlation ($\hat{\alpha}$) of the variant to the trait versus the Neanderthal LD profile (Rv). For the sunburn trait, we observe a positive correlation indicating a significant uni-directional relationship genome-wide between Neanderthal introgressed alleles and risk for sunburn (empirical null distribution $P = 0.001$, $q = 0.02$). For visualization, we bin Rv into 10 equally spaced intervals and plot the average $\hat{\alpha}$ with 95% bootstrapped confidence intervals.

By considering only variants passing a strict "genome-wide significant" P-value threshold, this directionality analysis tests for a relationship at the extremes of effect size and P-value. To assess if there is a uni-directional bias among Neanderthal-introgressed alleles on these traits across the effect distribution, we used signed LD profile regression (SLDP)[160]. Signed LD profile regression assesses whether variant effects on a trait ($\hat{\alpha}$ from GWAS summary statistics) are correlated genome-wide with a signed genomic annotation via the signed LD profile (Rv). Our genomic annotation quantifies how tightly linked each variant is to

Altai-matching Neanderthal introgressed alleles (Neanderthal LD profile [$R\nu$], 2.4:Methods).

Using signed LD profile regression on eight traits with heritability enrichment, we find a strong genome-wide correlation between higher LD to introgressed alleles (Neanderthal LD profile, $R\nu$) and increased risk for Sunburn $\hat{\alpha}$ ($r_f = 0.18\%$, $q = 0.02$, Fig. 2.3B, Table 6.6). Other traits, including menopause age and morningness, show directionality trends similar to the P-value threshold analysis (Fig. 6.14, Table 6.6). Some traits with heritability enrichment, like WBC count and autoimmunity, do not show consistent directionality genome-wide; instead, these traits have both genomic windows where Neanderthal alleles associate with risk-increase and other windows that associate with risk-reduction (i.e., bi-directional). Expanding to the 41 representative traits, introgressed alleles also have strong genome-wide uni-directional effects of protection from anorexia ($r_f = -0.93\%$, $q = 4 \times 10^{-5}$) and schizophrenia ($r_f = -0.27\%$, $q = 0.01$; Table 6.6). Even though Neanderthal variants contribute less to the heritability of these traits than expected, the associated introgressed alleles that remain are disproportionately risk-decreasing.

2.2.6 LD-aware identification of introgressed alleles with directional effects on human traits

In this section, we present examples of specific associations based on consistent directions of effect of introgressed alleles on traits identified by signed LD profile regression[160]. In contrast to previous approaches that simply intersected introgressed alleles with estimates of trait effects from association studies, we locate regions of interest based on strong correlations between LD to Altai-matching introgressed alleles (Neanderthal LD profile, $R\nu$) and trait-associated risk or protection ($\hat{\alpha}$) in sliding windows across the genome (2.4:Methods). This provides additional evidence of biologically relevant effects for Neanderthal variants, and has the benefit over simple GWAS intersections because directional effects are less confounded by genomic co-localization of Neanderthal ancestry with other functional elements and can have more power when applied to rare variation or diverse populations[160]. With this method we can identify candidate trait-associated regions that are not tagged by a single genome-wide significant association, yet still have a significant directional relationship between Neanderthal LD profile and a trait.

Applying this method to the eight traits from Fig. 2.3, we found many previously reported introgressed loci with trait associations. For example, we identify a window with a strong positive correlation between the Neanderthal LD profile and sunburn risk (chr9:16641651–16787775, $r = +0.83$). This window includes the gene *BNC2* and a high-frequency introgressed haplotype that influences skin pigmentation levels in Europeans (Fig. 6.15)[161]. We also recover previous links between Neanderthal introgression and chronotype surrounding ASB1 (overall $r = -0.92$, Fig. 6.15)[62]. Recapitulating these established findings supports the utility of this method for identifying regions where Neanderthal introgression influences phenotypes in modern Europeans.

We also identify several hundred additional windows with strong associations between LD to introgressed alleles and directional effects on traits. For example, we discovered two windows near *NMUR2* (within chr5:151745423-151931514) that show a positive relationship between increased LD to Neanderthal alleles and increased propensity to be a morning person (overall $r = +0.91$, Fig. 6.16). In the Supplemental Text 6.1 in Appendix 1, we describe eQTL, PheWAS, and model organism evidence supporting the hypothesis that introgressed alleles downregulate *NMUR2* in the brain leading to increased morningness.

This introgressed haplotype also has a genome-wide significant association with being a morning person (rs4958561: $P = 8.5 \times 10^{-12}$).

In contrast, no introgressed alleles individually had associations with autoimmune disease in the UK Biobank ($n = 459,324$) that pass genome-wide significance thresholds. Yet, illustrating the potential of the signed LD profile regression approach to discover candidate associations, we identify a window in which variants show a strong negative correlation (i.e., a protective relationship) between LD to Neanderthal introgression and autoimmune disease risk (chr7:50649920-50739129, $r = -0.84$, Fig. 2.4A-C). In this approximately 90 kb window, there are six introgressed GWAS tag variants; rs17544225 has the strongest single-locus association with autoimmune disease ($P = 9.8 \times 10^{-5}$; Fig. 2.4B). Within 1 Mb there are only two other variants with a similar association to autoimmune disease (rs2886554, 361 kb upstream, $P = 4.0 \times 10^{-5}$; rs6583440, 326 kb upstream, $P = 6.8 \times 10^{-5}$). These variants are not introgressed or in LD with rs17544225 ($r^2 = 0.0001$ and 0.0026, respectively).

Considering the introgressed variants and others in LD together provides power to test if the association signal in this region is likely related to the Neanderthal alleles or other nearby variation (Fig. 2.4C). This region contains *GRB10*, which encodes a growth factor receptor-bound protein known to interact with several tyrosine kinase receptors and signaling molecules[162]. *GRB10* has been associated with a subtype of systemic sclerosis (lcSSc); patients with systemic sclerosis have higher expression of *GRB10* in monocytes[163,164]. Studies of *Grb10* deficient mice demonstrated *Grb10*'s role in hematopoietic regeneration in vivo[165]. Additionally, in a transcriptome study of CD4+ Effector Memory T cells (CD4+ TEM), *GRB10* was the most downregulated gene after T-cell receptor stimulation[166]. Notably, in both human and mouse, *GRB10* mRNA is highly alternatively spliced, resulting in four to seven unique isoforms[167]. Of the 20 introgressed variants overlapping this window, 17 are splicing quantitative trait loci (sQTL, increasing intron excision, tag variant rs17544225: $P = 3 \times 10^{-9}$ [Bonferroni critical value $P = 1 \times 10^{-3}$], Fig. 2.4B) in the spleen[95]. In a PheWAS, traits associated with the introgressed haplotype (tagged by rs17544225) include monocyte count ($P = 4 \times 10^{-8}$) and monocyte percentage ($P = 3 \times 10^{-6}$; both pass the Bonferroni correction threshold $P = 1 \times 10^{-5}$)[159]. Therefore, we hypothesize that Neanderthal introgressed alleles regulate the expression or splicing of *GRB10* contributing to changes in monocytes that may lead to protection from autoimmunity.

2.3 Discussion

Here we estimate heritability patterns across more than 400 diverse traits in genomic regions influenced by Neanderthal introgression. Regions with remaining Neanderthal ancestry in modern populations are depleted of heritability for all traits considered, except those related to skin and hair. Introgressed alleles are also depleted for heritability of most traits; however, there is modest enrichment for heritability of several traits among alleles with older Neanderthal origins, including autoimmune disorders, hair and skin traits, chronotype, bone density, lung capacity, and age at menopause (Fig. 2.1). Summarizing these heritability patterns over trait domains, we find that dermatological, endocrine, and respiratory traits are consistently enriched for heritability among Altai-matching Neanderthal introgressed variants, whereas cognitive and ophthalmological domains are the most depleted (Fig. 2.2A). Additionally, several trait domains show di-

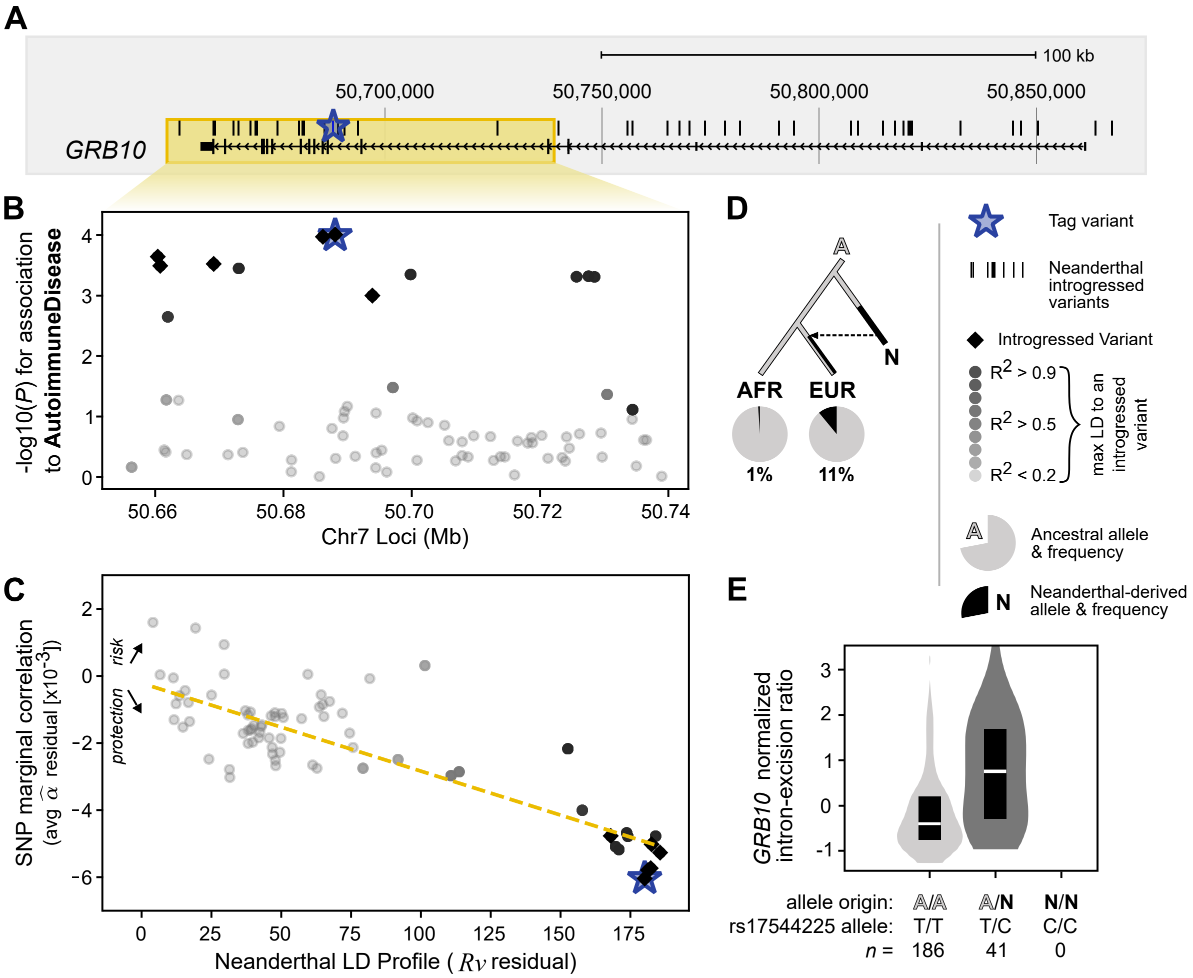


Figure 2.4: Signed LD profile regression identifies a candidate functional association between an introgressed haplotype in *GRB10* and autoimmune disease.
(A) A genomic region overlapping *GRB10* (chr7:50,649,920-50,739,129, yellow box) contains an introgressed haplotype. (B) GWAS Manhattan plot for this region showing associations with autoimmune disease from the UK Biobank ($n = 459,324$). The strongest single-variant association is at an introgressed variant, but it does not reach genome-wide significance (blue star, $P = 9.8 \times 10^{-5}$ at rs17544225). (C) Using signed LD Profile regression, we discover a strong negative relationship between LD to introgressed alleles and autoimmune disease in this region. The negative correlation between Neanderthal LD profile and autoimmune disease risk suggests a protective relationship between Neanderthal introgression at this locus and autoimmune disease ($r = -0.84$). Thus, while the single-variant association alone is not sufficient to implicate this introgressed haplotype in autoimmune disease risk, considering LD to Neanderthal alleles and the direction of effect across variants identifies it as a candidate. (D) The haplotype (tagged by starred variant rs17544225) is derived in Neanderthals (N) and at 11% frequency in modern Europeans (EUR, n = 503) with 1% frequency in Africans (AFR, $n = 661$, only observed in admixed African Americans and Caribbeans) (1000G super-populations). (E) The introgressed allele is also is an sQTL in which Neanderthal alleles associate with increased *GRB10* intron excision in spleen (two-tailed t-test $P = 3 \times 10^{-9}$). The boxplot centers represent medians and the boxes are bounded by the first and third quartiles.

vergent heritability patterns, e.g. among psychiatric and reproductive traits (Fig. 2.2D-E). Using two methods for evaluating the direction of effect of variants on traits, we find uni-directional biases for introgressed alleles with balding risk, younger menopause age, sunburn risk, forced vital capacity increase, and morning preference (Figs. 2.3,6.14). Finally, we show how our approaches can highlight novel candidate introgressed variants that influence risk for disease (Figs. 2.4,6.16, Supplemental Text 6.1 in Appendix 1).

To contextualize the implications of our results and to provide a framework for future studies, we propose a model that links observed patterns of heritability and direction of effect to hypotheses about the history of selective pressures on introgressed haplotypes (Fig. 2.5). Along the dimensions of heritability enrichment vs. depletion and uni-directional vs. bi-directional associations, traits fall into four general quadrants (Fig. 2.5B). First, most traits show heritability depletion among introgressed variants and no bias in the direction of effect. This suggests selection against introgressed variants that influenced these traits (Fig. 2.5B, bottom left). Second, the opposite pattern—enrichment for heritability in introgressed variants and a directional bias in their direction of effect—suggests that introgression introduced functional alleles that were positively selected in AMHs (Fig. 2.5B, top right). For example, the enrichment for heritability of sunburn and tanning in Altai-matching introgressed alleles and the bias in direction of effect in AMH suggests that these introgressed alleles decreased hair and skin protection against sun exposure in ways that may have been beneficial, perhaps in response to decreased UV at higher latitudes. Third, traits, like autoimmune disease risk and WBC count, have heritability enrichment among introgressed variants, but no directional bias. In this case, introgression likely contributed increased diversity—both trait increasing and decreasing—into AMHs that was beneficial as they adapted to non-African environments (Fig. 2.5B, bottom right). We found support for the action of positive selection on two traits with heritability enrichment; high-frequency putatively adaptive introgressed haplotypes are enriched for overlap with windows associated with both sunburn and white blood cell count (Fig. 6.9). Fourth, traits like anorexia and schizophrenia, show depletion for heritability among introgressed variants, but in contrast to most depleted traits, the remaining introgressed variants have a bias towards trait-protective effects (Fig. 2.5B, top left). We hypothesize that this pattern could be produced by negative selection purging most introgressed alleles that influence the trait paired with selection for a small number of introgressed protective alleles. Supporting this interpretation, remaining Altai-matching variation has the strongest correlation with protective benefit against serious fitness-reducing diseases (anorexia, schizophrenia)[168]. In summary, our results reveal signatures of contrasting patterns of selection since admixture on introgressed variation associated with different traits. Further work is needed to determine how these introgressed variants influence traits and resolve the dynamics of selection.

Our results expand the current understanding of the functional effects of introgressed variants in several dimensions. First, previous studies of regions with Neanderthal ancestry found depletion for evidence of background selection and functional annotations, such as genes and gene regulatory elements active in specific tissues[44,45,53,56]. We extend beyond these proxies for function and show depletion for effects on diverse complex traits in a human population. This further supports selection against Neanderthal introgression in trait-associated genomic regions. However, we also find an exception to this pattern for variation associated with skin color and tanning. This is consistent with previous hypotheses that genomic regions associated with skin traits tolerated introgression and with previous tests for genome-wide effects of Neanderthal ancestry on complex traits that found enrichment for traits related to skin and hair[45,59].

Second, our analyses increase the scope and accuracy of estimates of the genome-wide influence of Neanderthal introgression on human phenotypes. S-LDSC requires only GWAS summary statistics, rather than individual-level data as in the GCTA analysis of 46 specific traits in Simonti et al. 2016[59]. This enabled

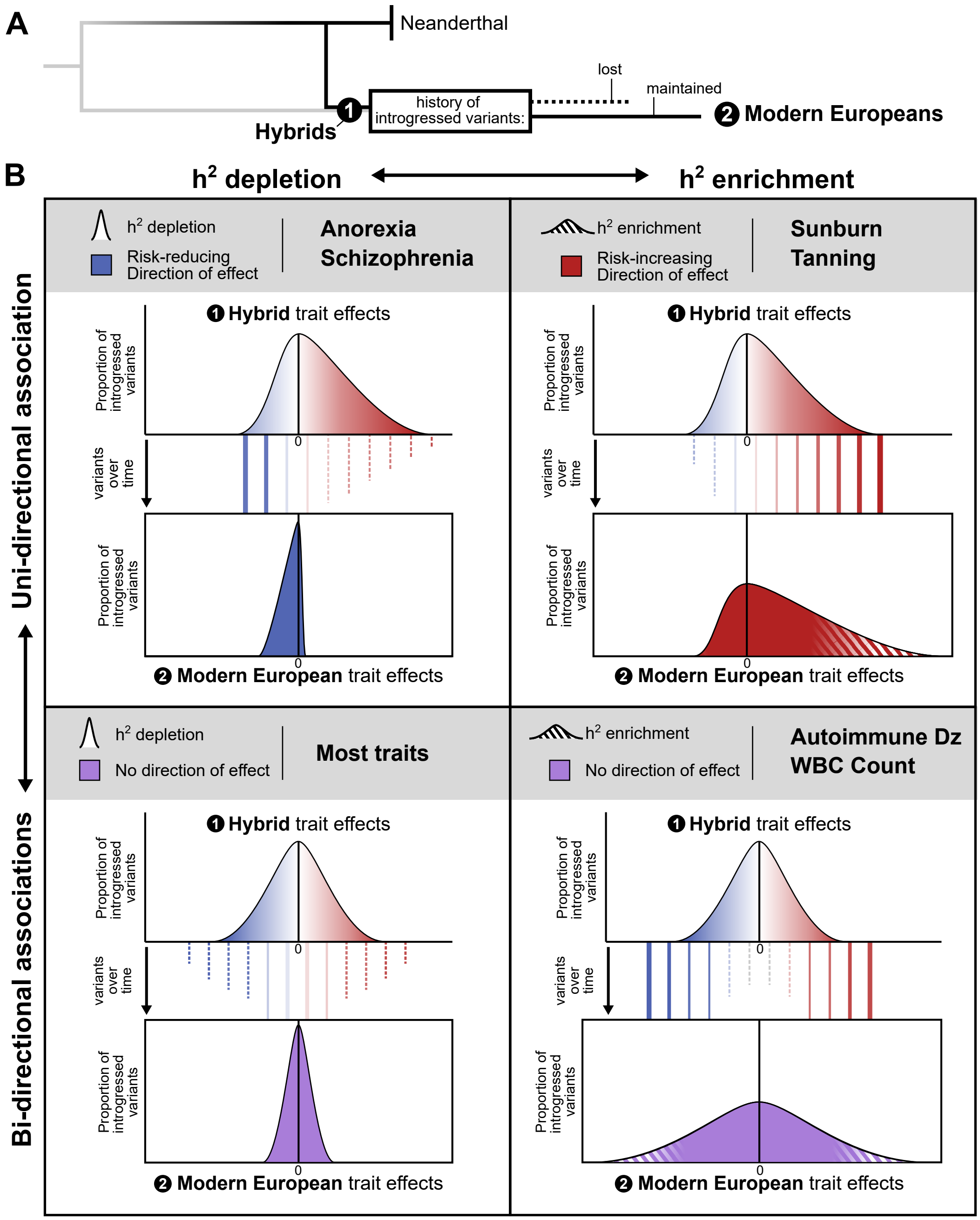


Figure 2.5: Patterns of heritability and direction of effect suggest contrasting selective pressures on introgressed variation associated with different traits.
See next page for full caption.

Figure 2.5: (Previous page.) Patterns of heritability and direction of effect suggest contrasting selective pressures on introgressed variation associated with different traits.
(A) After admixture, many Neanderthal variants segregated in hybrid populations. As these populations evolved into modern Eurasians, some introgressed variants were lost due to drift or negative selection (dashed line) and some were maintained due to drift or positive selection (solid line). (B) Among the remaining introgressed variants in modern Europeans, traits fall into four general quadrants on the axes of heritability enrichment vs. depletion (x-axis) and uni-directional vs. bi-directional trait effects (y-axis). For each quadrant, we depict potential variant histories and selective pressures leading to the observed distribution of introgressed variants' trait effects (solid and dashed lines). (Bottom Left) Heritability for most traits is depleted among introgressed variants (narrow effect distribution with most variants conferring no effect) with no bias in the direction of effect (centered at zero). This suggests selection against introgressed variants that influenced these traits. (Top Right) The opposite pattern is observed in traits such as sunburn and tanning. These traits are enriched for heritability among introgressed variants (thick tail with more variants conferring trait effects than expected), and they have a bias in their direction of effect (skewed). This pattern suggests that introgression introduced some functional alleles that were positively selected in AMHs. (Bottom Right) Traits, like autoimmune disease risk and white blood cell (WBC) count, have heritability enrichment among introgressed variants (thick tails with many variants conferring trait effect), but no directional bias (centered). In this case, introgression likely contributed increased diversity relevant to the trait—both trait-increasing and decreasing—into AMHs that was beneficial as they adapted to non-African environments. (Top Left) Finally, traits like anorexia and schizophrenia, show depletion for heritability among introgressed variants (narrow distribution), but they have a significant directionality bias in the few introgressed variants with effects (skewed). This pattern could be produced by negative selection purging most introgressed alleles that influence the trait paired with selection for a small number of introgressed beneficial alleles.

us to test effects on over 400 traits across many domains in a larger cohort. Furthermore, the partitioned heritability method for identifying enrichment considers LD and the full distribution of variant effect sizes from a GWAS rather than selecting an ad hoc significance threshold and attempting to generate appropriate comparison sets of non-introgressed alleles as in the analysis of 136 traits in an earlier release of the UK Biobank by Dannemann et al.[62] Highlighting the importance of accounting for LD, a recent analysis of introgression in whole-genome sequences from 27,566 Icelanders by Skov et al.[31] suggested based on locus-by-locus trait association tests that many previous associations between traits and introgressed variants were better explained by non-introgressed variants in LD. Our approach addresses this important concern without the need for arbitrary filters and assumptions about the causal variant that complicate locus-level analyses. Furthermore, in contrast to simply associating the absolute number of archaic alleles in each individual with traits[31], our approach assesses the genome-wide influence of archaic introgression on phenotypes by considering the specific archaic alleles present across individuals and the effects of each allele on traits.

Third, we analyze trait heritability patterns for different sets of variants in regions with Neanderthal ancestry (Fig. 2.1A). Considering non-introgressed variants and remaining introgressed variants with different histories separately enables us to identify differences in the effects of introgressed variants based on their origins and genomic context. For example, we find modest enrichment for heritability of several traits among introgressed alleles, even though they are in regions of the genome with overall depletion for heritability of these traits. Our analyses also suggest differences in heritability among different subsets of introgressed variants. The introgressed variants that remain in AMH genomes are the result of complex selective and demographic pressures following admixture[35,51,53]. Introgressed haplotypes carry alleles of different origins, including ancestral alleles lost in some modern Eurasian populations[58].Our analysis of different sets of alleles on introgressed haplotypes revealed that introgressing alleles matching the Altai Neanderthal are less depleted for heritability than those matching introgressed alleles overall (Fig. 2.1D). The introgressing Neanderthal population diverged from the Altai Neanderthal population more than 100 kya, while the Vindija was much closer genetically and geographically[14,15]. Thus, we hypothesized that the Altai-matching introgressed alleles were likely at higher frequency in different Neanderthal populations and were thus less likely to have strong deleterious effects than younger introgressed Neanderthal alleles. The lower levels of depletion (and modest enrichment for some traits) of heritability in Altai-matching variants support this hypothesis.

Fourth, we introduce a new approach for testing for consistent direction of effects for introgressed alleles on traits. Using this approach, we show that Neanderthal introgression generally increased propensity for sunburn, balding, larger lung capacity, and younger menopause, while it had both increasing and decreasing effects on most other traits. With this directionality metric, we also highlight hundreds of candidate functional introgressed variants including many that would not have been identified by simply intersecting introgressed alleles with GWAS results.

Several limitations must be considered when interpreting our results. First, we quantify the contribution to heritability of common introgressed variants (MAF > 0.05); genome-wide investigation of rarer introgressed variant effects will be possible in the future as more dense sequencing cohorts and new statistical methods become available[169]. Second, because some of the partitions of the genome considered are small (e.g. common Altai-matching introgressed variants), some of the enrichment, depletion, and directionality tests we performed are underpowered. Third, many introgressed alleles likely had pleiotropic effects and different fitness effects in modern versus archaic environments, complicating the inference of the history of selection. Fourth, recent analyses have demonstrated that estimates of heritability enrichment are sensitive to the assumed heritability model and that variation in heritability estimates from different statistical methods are influenced by demographic factors[170,171]. Nonetheless, our results are consistent in direction across many traits and are correlated across variant sets. Given this consistency, that the overall differences in heritability estimates in previous evaluations are small, and that none of our interpretations rely on magnitude of effect, we anticipate that other estimation methods would identify similar overall depletion for trait-associated variation in genomic regions with Neanderthal ancestry. Fifth, we only analyze the effects of introgressed variation in the context of Europeans. Further work in new cohorts[172] and continued expansion of GWAS across diverse traits are needed to comprehensively understand the role of introgressed variation in other (e.g. East and South Asian) populations, especially given that Asians have evidence of pulses of introgression from different Neanderthal populations than Europeans[173]. Sixth, in the direction of effect analyses, we were conservative in considering only Altai-matching alleles and expanding for LD in mapping introgressed variants to GWAS hits. Thus, some introgressed alleles with effects on traits considered may have been missed (2.4:Methods); however, our genome-wide signed LD profile regression approach considers all variants and effects. Finally, while we identify associations between many introgressed haplotypes and traits, molecular validation is needed to determine the specific causal allele(s) behind the association.

With the growth of large cohorts including linked genotype and phenotype data, it will be valuable to extend these heritability analyses to large-scale biobank data sets from diverse populations. This will enable further quantification of the functional effects and selective pressures on introgressed variants, including introgression from Denisovans, and other alleles with unique evolutionary histories (e.g., reintroduced ancestral alleles, high frequency derived alleles). We also anticipate that simulation studies can inform our understanding of the types of selective pressures required after introgression to produce the heritability patterns observed. Ultimately, knowledge of how remaining introgressed Neanderthal alleles influence AMH populations provides a window into understanding the phenotypic variation of Neanderthal populations over 50,000 years ago and how this variation contributed to AMH adaptation to diverse environments.

2.4 Methods

Defining Neanderthal-introgressed regions and variants

Genomic regions with Neanderthal ancestry

To define genomic regions with Neanderthal ancestry we used "segments" identified by Browning et al.[32] using Sprime, a heuristic scoring strategy that compares high LD regions in a target admixed populations (i.e. Europeans) with an unadmixed outgroup (i.e. Africans) to identify putatively introgressed regions. We considered the Sprime-identified segments identified using five European subpopulations (CEU, TSI, FIN, GBR, IBS). To isolate regions with Neanderthal ancestry, as recommended by Browning et al.[32], we (1) considered segments identified in these five populations that have at least 30 putatively introgressed variants that could be compared to the Altai Neanderthal genome and (2) had a match rate of at least 30% to the Altai Neanderthal allele. We provide data on these sets in Fig. 6.1. After applying these two filters to the segments identified independently in the five European subpopulations, we merged these sets. This ultimately defines a set of segments with strong evidence of Neanderthal ancestry in Europeans used for the top panel of Fig. 2.1B. To define the non-introgressed variants in segments of Neanderthal ancestry (bottom panel of Fig. 2.1B), we identified 1000G variants in these segments and subtracted out introgressed variants (LD expanded to $r^2 > 0.5$, see set four below). Finally, in Fig. 6.3A, we repeat this analysis with regions that have at least a 30% match rate to the Vindija Neanderthal genome (instead of Altai).

Neanderthal introgressed variants (All introgressed variants, Altai-matching, and Vindija-matching)

We consider several sets of Neanderthal introgressed alleles based on Sprime analyses. From most stringent to least stringent, these sets are: (1) putatively introgressed variants identified in European subpopulations matching the Altai Neanderthal allele (used predominately in analyses in Fig. 2.1D - Fig. 2.4, $N = 138,774$), (2) putatively introgressed variants identified in any modern subpopulation matching the Altai Neanderthal allele ($N = 276,902$), (3) putatively introgressed variants identified in European subpopulations regardless of evidence of matching the Neanderthal allele ($N = 350,577$), and (4) putatively introgressed variants identified in any subpopulation regardless of evidence of matching the Neanderthal allele (used in Fig. 2.1C-D, $N = 900,902$). In sets three and four, the variants might not match the Altai Neanderthal allele at the site or a comparison might not have been possible due to lack of coverage or high confidence allele call. We present results from set one ("Altai-matching introgressed variants") and set four ("introgressed variants") in the main text. Fig. 6.8 reports heritability enrichments by trait for the set one Altai-matching variants but further stratified by minor allele frequency.

Of all Altai-matching variants (set one) and introgressed variants (set four), respectively, 44,537/138,774 (32.1%) and 139,118/900,902 (15.4%) are at MAF > 0.05 and are used to calculate heritability enrichment by S-LDSC. However, all variants at MAF $>= 0.52\%$ (Allele Count ¿= 5) are used to compute LD scores. This includes 82.9% (115,081/138,774) of Altai-matching variants and 41.5% (374,172/900,902) of all introgressed variants.

Finally, we created a "Vindija-matching introgressed variants" set to investigate evolutionarily younger variants shared among the Neanderthals closer to the introgressing population. This set includes putatively introgressed variants identified in European subpopulations that match the Vindija Neanderthal allele ($N = 167,927$, used in Fig. 6.3).

Vernot 2016 S-identified haplotypes and variants*

For completeness, we also considered the introgressed Neanderthal haplotypes previously identified by Vernot et al.[156]. These introgressed regions were identified using the S* statistic which, like Sprime, infers introgressed regions in the absence of any archaic reference genome. Like Sprime, S* uses a heuristic scoring strategy between introgressed target populations and a non-introgressed outgroup. Sprime differs from S* in that it simultaneously considers multiple members of the target population, and Sprime allows for limited gene flow between the target population and the outgroup.

For introgressed haplotypes identified by S* in Europeans (5851), 3243 (55%) are more than 50% covered by at least one EUR segment identified by Sprime, and 2370 S* haplotypes (40%) have 0% coverage. Conversely, for introgressed segments identified by Sprime in Europeans (1733), 1128 (65%) are more than 50% covered by at least one EUR haplotype identified by S*, while 282 (16%) have 0% coverage.

GWAS summary statistics

41 representative traits

We considered GWAS summary statistics from a previously-described representative set of 41 diseases and complex traits[92,144–154]. Previous studies using these traits had GWAS replicates (genetic correlation > 0.9) for six of these traits (BMI, Height, High Cholesterol, Type 2 Diabetes, Smoking status, Years of Education). For these six traits, we considered only the GWAS with the largest sample size so our combined analysis did not overrepresent these six. All GWAS are European-ancestry only. Many are from UK Biobank, but we note that their coding may be different than coding used in other UK Biobank heritability analyses[144]. For example, morning person is converted into a binary variable (morning person vs. evening person) rather than the categorical ordinal scale of the underlying data ("definitely a morning person", "more a morning person", "more an evening person", "definitely an evening person"). Information on these traits is in Table 6.1.

405 UK Biobank Traits

For a more diverse set of traits, we considered GWAS from the UK Biobank and 15 from FinnGen formatted for LDSC by the Neale Lab[144,157]. For reliability of S-LDSC heritability estimates, we apply two thresholds to select GWAS based on recommendations from Finucane et al.[75] and the Neale lab[157,174]. We only consider traits that meet the following criteria:

1. High confidence estimates of SNP heritability: traits with an effective sample size of greater than 40,000, a standard error of less than 6 times expected based on the GWAS sample size, sex bias less than 3:1, no nonlinear ordinal coding of numeric values
2. Significantly heritable traits: phenotypes that have heritability estimates with $P < 1.28 \times 10^{-12}$ ($z > 7$)

Together, these two criteria define a set of 405 traits (average $n = 288,130$, $h^2_{SNP} = 0.108$). Some traits are genetically independent of the other traits considered, but many of these traits are also correlated with each other (e.g., the shared genetic architecture of depression and anxiety). Traits from the previous set of 41 are only included if they meet the criteria for this high-confidence set from UK Biobank/FinnGen.

Defining phenotypic domains

To explore heritability on a trait domain level, we categorize traits by their phenotypic "domains," "chapters," and "subchapters". We derive these designations from the GWAS Atlas, a database of publicly available GWAS summary statistics[159]. The GWAS Atlas has categorized many of the 405 UK Biobank traits; however, because the GWAS Atlas uses different criteria for inclusion into their database, some of the traits analyzed here were uncategorized. We manually assigned the uncategorized UK Biobank traits and the 41 representative traits into the domain, chapter, and subchapter hierarchy based on similar categorized traits. The only change we made to the existing designations was among subchapter labels of the immunologic domain. All its subchapter instances ($N = 14$) were labeled "Immunological System Functions." We manually changed this generic label to either red blood cell (RBC) or white blood cell (WBC). For example, reticulocyte count and mean corpuscular hemoglobin fall under RBC, while eosinophil count and neutrophil fall under WBC. The 405 GWAS cross 21 domains, 31 chapters, and 62 subchapters. However, we note that this organization is not purely hierarchical (e.g. some traits in the same subchapter belong in two different domains).

Quantifying partitioned heritability with S-LDSC

We quantified partitioned heritability using Stratified-LD Score Regression v1.0.1 (S-LDSC) to test whether an annotation of interest (e.g., introgressed regions or introgressed variants) is enriched for heritability of a trait[75,90]. We use 1000 Genomes for the LD reference panel (variants with MAF ¿ 0.05 in European samples)[175]and HapMap Project Phase 3 (HapMap 3)[176] excluding the MHC region for our regression variants to estimate heritability enrichment and standardized effect size metrics following previous recommendations for S-LDSC[75].

S-LDSC estimates the heritability enrichment, defined as the proportion of heritability explained by common variants (MAF > 0.05) in the annotation divided by the proportion of all variants considered that are in the annotation. The enrichment of annotation c is estimated as

$$Enrichment_c = \frac{\%h^2_{(c)}}{\%SNP_{(c)}} = \frac{h^2_{(c)}/h^2}{|c|/M}$$

where $h^2_{(c)}$ is the heritability explained by common variants in annotation c, h^2 is the heritability explained by the common variants over the whole genome, $|c|$ is the number of common variants that lie in the annotation, and M is the number of common variants considered over the genome[75,92]. We use the baseline v2.1 model which includes 86 diverse annotations including coding, UTR, promoter and intronic regions, histone marks (H3K4me1, H3K4me3, H3K9ac, H3K27ac), DNAse I hypersensitivity sites (DHSs), chromHMM and Segway predictions, super-enhancers, FANTOM5 enhancers, GERP annotations, MAF bins, LD-related, and conservation annotations[75,155,169].

Direction of effect: Intersection with genome-wide significant variants

To intersect introgressed variants with genome-wide significant variants, we first used PLINK to LD expand the Altai-matching introgressed Neanderthal variants (set one, described in "Neanderthal introgressed variants" methods section) to perfect LD ($r^2 > 0.999$)[177]. LD was calculated for variants within 1 Mb of each introgressed variant using the 1000G European reference population while preserving the "phase" of the allele in LD with the Neanderthal allele[175]. We eliminated any duplicates (i.e., if two introgressed variants in perfect LD were both tagging another variant). We intersected this LD-expanded set of introgressed variants with the GWAS summary statistics using rsIDs. We oriented the sign of the summary statistic (the z-score) relative to the archaic allele (or the allele in perfect LD to the archaic allele). For example, if a variant is positively associated with a trait (z-score is +6 with GWAS effect allele "A" and alternative allele "C"), but the archaic allele is "C", we flip the z-score to be -6 because the archaic allele "C" is negatively associated with the trait.

For eight traits (AutoimmuneDz, Balding, Sunburn, FVC, Heel_T_Score, MorningPerson, Menopause-Age, WBCCount), we filtered the introgressed variant-summary statistic intersection at different thresholds of genome-wide significance ($P < 1 \times 10^{-8}$, $P < 5 \times 10^{-8}$, $P < 1 \times 10^{-6}$). We then pruned variants at various levels of LD ($r^2 = 1$, $r^2 = 0.8$, $r^2 = 0.5$) to reduce redundant counts due to linked loci. We used the LDmatrix tool in the LDlink API to calculate the pairwise LD to prune linked variants (with the 1000G EUR as a reference)[178]. We then counted the number of introgressed alleles associated with the positive and the negative directions of the trait. With quantified significance with a chi-squared goodness of fit test.

Limitations of genome-wide significant variant intersection

We caution overinterpretation of these results and highlight some of the limitations of this method. First, despite LD expansion, only 29% of the introgressed alleles could be intersected with variation interrogated by the GWAS (LD expanded to r2 = 1). Therefore, this analysis does not investigate directionality of introgressed variants in regions not perfectly tagged by the genotyping array used for the GWAS. However, 61% of the Sprime segments (larger windows with Neanderthal introgression) have at least one introgressed variant interrogated by the GWAS; therefore, we feel confident that this analysis samples broadly across introgressed regions. Second, by considering only genome-wide significant variants, this directionality analysis is limited to loci in the extremes of the GWAS distribution. It does not consider the global genome-wide relationship between introgressed alleles and directionality of trait-associated variation at varying levels of effect size and significance. However, we show these results are consistent at a less stringent level of genome-wide significance ($P < 5 \times 10^{-8}$, Fig. 6.13).

Direction of effect: Signed LD profile regression (SLDP) analysis

SLDP quantifies the genome-wide directional effect of a signed functional annotation on polygenic disease risk. SLDP calculates the correlation between a vector of variant effects on a trait (from GWAS summary statistics, $\hat{\alpha}$) and a vector of those variants' aggregate tagging of an annotation (Rv)[160]. Our annotation is each variant's maximum LD to a Neanderthal introgressed allele (which we term "Neanderthal LD profile"). This allows us to quantify if there is a genome-wide relationship between a variant's LD to a Neanderthal allele and the direction of that variant's trait association. This is distinct from previous stratified-LD score regression (S-LDSC) analyses because S-LDSC quantifies heritability enrichment in an annotation of interest independent of directionality.

More specifically, SLDP regresses $\hat{\alpha}$ (the vector of marginal correlations between variant alleles and a trait) on vector v (the signed functional annotation) to estimate r_f, the functional correlation between the annotation and trait, using

$$E(\hat{\alpha}|v) = \sqrt{h_g^2}Rv$$

where R is the LD matrix from the 1000G Phase 3 European reference, h_g^2 is the trait's SNP-heritability. Together, Rv is a vector quantifying each variant's aggregate tagging of the annotation, termed the "signed LD profile". SLDP uses generalized least-squares regression across HapMap 3 variants excluding the MHC region ($M = 1,187,349$). It also conditions the regression on a "signed background model" that quantifies the directional effects of minor alleles to reduce confounding due to genome-wide negative selection or population stratification (using five equally-sized MAF binds). False discovery rates and P-values are obtained by empirically generating a null distribution by randomly flipping the signs of v in large blocks. For a detailed description of the SLDP method, derivation, estimands, and validation see Reshef et al.[160].

We conducted SLDP analysis on the 41 representative GWAS summary statistics (Fig. 2.3C and Table 6.6). To generate our functional annotation, we used PLINK to calculate pair-wise LD between the Altai-matching introgressed variants (set 1, described in "Neanderthal introgressed variants" methods section) and the 1000 Genomes Phase 3 European reference panel (approx. 10M variants)[175]. We considered LD limited to variant pairs within 1 Mb and $r^2 > 0.2$. For each variant in the reference panel, the annotation (v) is the maximum r^2 value to the Neanderthal variants. The input annotation (v) is generated with reference to the allele that is on the Neanderthal haplotype. However, for the SLDP regression, the signs (for both $\hat{\alpha}$ and Rv) are oriented with reference to the European minor allele.

For interpretability of the visualizations, all plots show $\hat{\alpha}$ and Rv with reference to the Neanderthal allele. For example, if a Neanderthal variant, "X", is in LD (and in-phase) with SLDP regression variant, "Y" at $r^2 = 0.5$, variant Y's functional annotation (v) is 0.5. We plot the sign of $\hat{\alpha}$ (from the GWAS) with reference to Y as the effect allele (A1). All plots describing SLDP results display the residuals of $\hat{\alpha}$ (y-axis) and Rv (x-axis) for each variant. This residual reflects that all analyses are conditioned on the "signed background model" described above.

Identifying genomic windows with an association between Neanderthal LD profile and trait effect

To locate regions in which Neanderthal introgression likely influences a trait of interest, we identify genomic windows with a strong correlation between LD to introgressed alleles and trait-associated risk or protection. From the per-variant output from SLDP regression ($M = 1,187,349$), we calculated Pearson correlation coefficients (r) between the residuals of $\hat{\alpha}$ and $R\nu$ for 30 kb sliding windows centered around each SLDP regression variant. We select windows that have at least 15 SLDP regression variants and an $r^2 > 0.5$ (correlation in either direction), and we join overlapping windows. Therefore, the final windows are often bigger than 30 kb and can have a correlation coefficient less than 0.5. We only consider windows that have at least one variant marginally associated with the trait ($P < 1 \times 10^{-4}$) and windows that overlap at least one Altai-matching Neanderthal introgressed allele (set one; see above).

Figures depicting the windows of interest identified were generated using the UCSC Genome Browser[179]. eQTL and sQTL analysis and plots were generated using the Genotype-Tissue Expression (GTEx) Project (V8 release) Portal on 4/29/2020[95]. GTEx V8 results are in GRCh38 and were lifted over to GRCh37 (hg19) for comparison with the windows of interest. PheWAS results are from the GWAS Atlas and consider 4756 traits[159].

Overlap between genomic windows and high-frequency haplotypes

To test if the windows with Neanderthal trait-associated heritability enrichment and directionality have evidence of recent positive selection, we compared them with high-frequency haplotypes defined by Gittelman et al.[35] (European only) and Chen et al.[180] (excluding haplotypes identified in Africans only). We calculated an empirical null distribution by shuffling identified trait-associated windows within the universe of genomic regions that could have been identified through the above method (30 kb sliding windows centered around each SLDP regression variant with at least 15 regression variants that, when merged into non-overlapping windows, had to overlap at least one Altai-matching allele). For the observed trait-associated windows and 10,000 shuffled sets of the windows, we quantified the proportion that overlapped the high-frequency haplotypes and compared the observed to the shuffled (Fig. 6.9).

Data analysis and figure generation

All genomic coordinates and analysis refer to Homo sapiens (human) genome assembly GRCh37 (hg19) unless otherwise specified. All P-values are two-tailed and q-values are Benjamini-Hochberg FDR-corrected at $\alpha = 0.05$, unless otherwise specified. All measures of central tendencies are means, unless otherwise specified. Data and statistical analyses were conducted using Python 3.5.4 (Anaconda distribution), R 3.6.1, Jupyter Notebook, BedTools v2.26, and PLINK 1.9[177,181]. Figure generation was significantly aided by Matplotlib, Seaborn, and Inkscape[182–184].

Data Availability

The publicly available data used for analysis are available in the following repositories: introgressed variants and segments from Sprime Version 1 [https://data.mendeley.com/datasets/y7hyt83vxr] (Browning et al.[32]), introgressed variants and segments from S* [http://akeylab.gs.washington.edu/vernot_et_al_2016_release_ data/introgressed_tag_snp_frequencies/] (Vernot et al.[156]), GWAS traits formatted for LDSC v1.0.1 from the Alkes Price lab [https://data.broadinstitute.org/alkesgroup/LDSCORE/independent_sumstats/], UK Biobank traits formatted for LDSC from the Neale lab [http://www.nealelab.is/uk-biobank][157], GWAS Atlas [https://atlas.ctglab.nl/] (Watanabe et al.[159]), the GTEx Project Portal [https://gtexportal.org/home/] (Lonsdale et al.[95]), 1000 Genomes for the LD reference (Auton et al.[175]), and HapMap Project Phase 3 (HapMap 3) (Altshuler et al.[176]). The datasets we generated are available in the trait-h2-neanderthals GitHub repository [https://github.com/emcarthur/trait-h2-neanderthals][185]. They include bed files of all genomic partitions considered (regions with Neanderthal ancestry, sets of introgressed variants), all results of partitioned heritability analysis output (for the 41 traits formatted from the Price Lab and the 405 traits from the UKBioBank formatted by the Neale Lab) and signed LD profile regression results.

Code Availability

The publicly available software are available in the following repositories: LDSC v1.0.1 [https://github.com/bulik/ldsc](Finucane et al.[75] and Bulik-Sullivan et al.[90]) and Signed LD Profile Regression [https://github.com/yakirr/sldp](Reshef et al.[160]). The trait-h2-neanderthals GitHub repository [https://github.com/emcarthur/trait-h2-neanderthals][185] contains a Jupyter notebook with custom code used for data analysis and all figure generation.

CHAPTER 3

TOPOLOGICALLY ASSOCIATING DOMAIN BOUNDARIES THAT ARE STABLE ACROSS DIVERSE CELL TYPES ARE EVOLUTIONARILY CONSTRAINED AND ENRICHED FOR HERITABILITY[*]

3.1 Introduction

The three-dimensional (3D) conformation of the genome facilitates the regulation of gene expression[186–189]. Using chromosome-conformation-capture technologies (3C, 4C, 5C, Hi-C),[103–105] recent studies have demonstrated that modulation of gene expression via 3D chromatin structure is important for many physiologic and pathologic cellular functions, including cell-type identity, cellular differentiation, and risk for multiple rare diseases and cancer[125–130,190]. Nonetheless, many fundamental questions about the functions of and evolutionary constraints on 3D genome architecture remain. For example, how does genetic variation in different 3D contexts contribute to the risk of common complex disease? Furthermore, disease-causing regulatory variation is known to be tissue-specific; however, only recently has there been characterization of 3D-structure variation across multiple cell types and individuals[190–192]. Understanding how different attributes of 3D genome architecture influence disease risk in a cell-type-specific manner is crucial for interpreting human variation and, ultimately, moving from disease associations to an understanding of disease mechanisms[81].

3D genome organization can be characterized at different scales. Globally, chromosomes exist in discrete territories in the cell nucleus[105]. On a sub-chromosomal scale, chromatin physically compartmentalizes into topologically associating domains (TADs). TADs are megabase-scale genomic regions that self-interact but rarely contact regions outside the domain (Fig.3.1A)[105,111–113]. They are formed and maintained through interactions between CTCF zinc-finger transcription factors and cohesin ring-shaped complexes, among other proteins both known and unknown[102,105]. TADs are identified based on regions of enriched contact density in Hi-C maps (Fig.3.1). TADs modulate gene regulation by limiting interactions of cis-regulatory sequences to target genes[105]. The extent to which chromatin 3D topology affects gene expression is still debated[193]. In extensively rearranged Drosophila balancer chromosomes, few genes had expression changes[194]. In contrast, subtle chromatin interaction changes in induced pluripotent stem cells (iPSCs) from seven related individuals were associated with proportionally large differential gene expression[195]. Thus, further cell-type-specific investigation into properties of TAD organization and disruption will need to clarify which parts of the genome are sensitive to changes in 3D structure and how these changes influence gene regulation and traits.

At the highest level, TAD organization can be divided into two basic features: the TAD and the TAD boundary. TADs are the self-associating, loop-like domains that contain interacting cis-regulatory elements and target genes. TAD boundaries—regions in between TADs—are insulatory elements that restrict in-

[*] This chapter is from McArthur, E. & Capra, J.A. 2021. *AJHG.* and has been reproduced with permission of the publisher and co-author J.A. Capra.

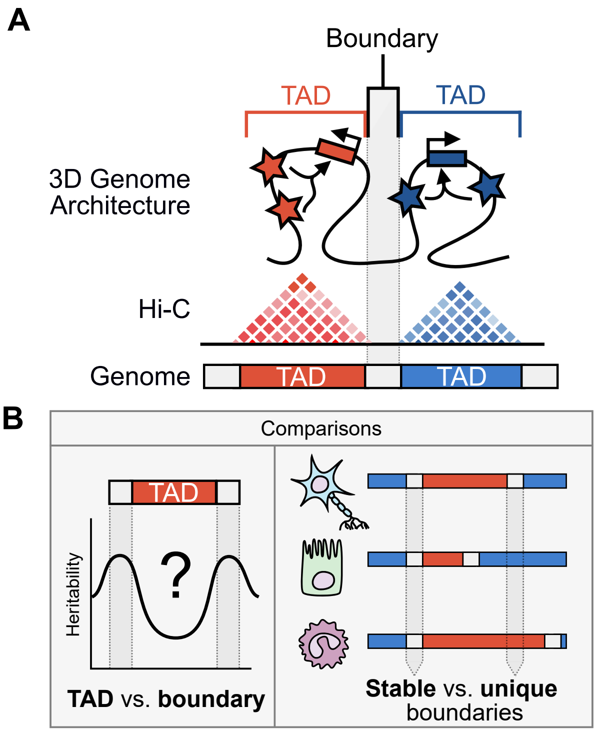


Figure 3.1: Schematic depiction of our analyses of 3D chromatin TAD-boundary stability and function. **(A)** Chromatin is organized in 3D space into topologically associating domains (TADs), which are identified by Hi-C experiments. Regions within a TAD are much more likely to interact with one another than are regions outside of the TAD. Regions bordering TADs are TAD boundaries. Boxes with right-angled arrows represent genes, and stars represent gene regulatory elements, such as enhancers. **(B)** This work addresses two main questions: (1) How are complex-trait heritability and evolutionary sequence conservation partitioned between TADs and TAD boundaries? (2) Do stable TAD boundaries (i.e., those observed across multiple tissues) have different contributions to trait heritability or sequence conservation than TAD boundaries unique to specific tissues?

teractions of cis-regulatory sequences, such as enhancers, to target genes[105]. Previous work suggests the functional importance of maintaining both the self-associating TADs and the insulatory boundaries. For example, in cross-species multiple sequence alignments, syntenic break enrichment at TAD boundaries suggests a long-term evolutionary preference for rearrangements that "shuffle" intact TADs, rather than "break" them[121,123]. Additionally, 3D genome structure correlates with similar functional features, such as histone modifications and replication timing, across species[196]. TADs also often contain clusters of co-regulated genes—e.g., cytochrome genes and olfactory receptors[105,112,135]. Intra-TAD structural variation that deletes or duplicates enhancers has been implicated in polydactyly, B cell lymphoma, and aniridia[122]. Together, these data suggest that the genome is under pressure to preserve TADs as functional units.

Other evidence suggests the greater importance of maintaining TAD boundaries. TAD boundaries are enriched for housekeeping genes and transcription start sites[105,111]. Removing insulatory TAD boundaries leads to ectopic gene expression in cultured cells and in vivo. For example, TAD structure disruption at the EPHA4 locus leads to inappropriate rewiring of developmental genes implicated in limb-formation defects[105,122,132]. In cancer, large structural alterations that disrupt TAD boundaries cause pathogenic gene expression in acute myeloid leukemia (AML) and medulloblastoma[133,134]. Structural variation (SV) that disrupts TAD boundaries causes gain-of-function, loss-of-function, and misexpression in many forms of rare neurodevelopmental disease[132]. Accordingly, TAD boundaries and CTCF sites have evidence of purifying selection on SVs[137,197]. Finally, human haplotype breakpoints do not align with chromatin boundaries, which indicates that recombination might be deleterious at TAD boundaries[124]. Collectively, these findings suggest that TAD boundaries are functionally important and constrained, especially on the scale of human evolution.

In addition to the need for further characterization of the constraint on and functions of TADs versus TAD boundaries, there is also a gap in our understanding of the variability in TAD organization across cell types. TADs and TAD boundaries have been characterized as largely invariant across cell types[111,112,116–118] and species[105,111,119–121]. However, previous pairwise comparisons of five 3D maps suggest that 30%–50% of TADs differ across cell types[78,117]. More comprehensive recent investigations have observed large differences in the percent of boundaries not shared across cell lines (20%–80%), which contrasts with previous claims of extensive TAD conservation[198,199]. Boundaries shared across two cell types have evidence of stronger SV purifying selection than boundaries unique to a cell type, suggesting that shared boundaries are more intolerant of disruption[137]. Additionally, stratifying boundaries by their strength (in a single cell type) facilitated discovery that greater CTCF binding confers stronger insulation and that super-enhancers are preferentially insulated by the strongest boundaries[200]. Stratifying by hierarchical properties of TADs—TADs often have sub-TADs—demonstrated that boundaries flanking higher-level structures are enriched for CTCF, active epigenetic states, and higher gene expression[201].

Despite these preliminary indications that the stability of components of the 3D architecture might influence functional constraint, there has been no comprehensive analysis comparing genomic features and disease associations between 3D structural elements stable across multiple cell types and those that are unique to single cell types. Quantifying stability across cell types is important for interpreting new variation within the context of the 3D genome given our knowledge that disease-associated regulatory variation is often tissue-specific[190–192].

To investigate differences in TAD boundaries across cell types, we quantify boundary "stability" as the number of tissues that share a TAD boundary. If a TAD boundary is found in many tissues, it is "stable," whereas if it is found in few tissues, it is "unique" (Fig. 3.1B). Using this characterization, we address two main questions that aim to expand our framework for cell-type-aware interpretation of genetic variation and disease associations in the context of the 3D genome (Fig. 3.1B):

1. How do TADs and TAD boundaries differ in their contribution to complex-trait heritability and their evolutionary constraint?

2. Are there functional and evolutionary differences in TAD boundaries that are stable across multiple cell types versus TAD boundaries that are unique to specific tissues?

Synthesizing 3D genome maps across 37 diverse cell types with multiple functional annotations and genome-wide association studies (GWASs), we show that TAD boundaries are more enriched for heritability of common complex traits and more evolutionarily conserved than TADs. Furthermore, genetic variation in TAD boundaries stable across multiple cell types contributes more to the heritability of immunologic, hematologic, and metabolic traits than variation in TAD boundaries unique to a single cell type. Finally, these cell-type-stable TAD boundaries are also more evolutionarily constrained and enriched for functional elements. Together, our work suggests that TAD boundary stability across cell types provides valuable context for understanding the genome's functional landscape and enabling variant interpretation that accounts for genome 3D structure

3.2 Results

3.2.1 Estimating complex-trait heritability across the 3D genome landscape

Disruption of 3D genome architecture plays a role in rare disease and cancer; however, the contribution of common variation in different 3D contexts to common phenotypes is unknown. To investigate complex-trait heritability patterns across the 3D genome landscape, we use 37 TAD maps from the 3D Genome Browser (Table 6.7)[202]. The cellular contexts include primary tissues, stem cells, and cancer cell lines[116–118,203–206]; for simplicity, we will refer to these as "cell types." All TAD maps were systematically predicted from Hi-C data with the HMM pipeline from Dixon et al.[111] at either 40 kb or 25 kb resolution (Supplemental Text 6.2 in Appendix 2)[202]. We estimated common-trait heritability enrichment among common variants within these 3D genome annotations by using stratified-LD score regression (S-LDSC)[75,90]. S-LDSC is a method of partitioning heritability across the genome by using GWAS summary statistics and LD patterns to test whether variants in an annotation of interest (e.g., TADs or TAD boundaries) are enriched for heritability of a trait in comparison to the rest of the genome. We considered GWAS summary statistics from a previously described representative set of 41 diseases and complex traits (Table 6.8)[92,144–153].

To investigate patterns of heritability across the 3D genome landscape, we used two strategies for defining genomic partitions. In the first, we analyzed TADs plus 50% of their length on each side. Motivated by the approach to partitioning TADs from Krefting et al.[123], we subdivided these regions into 20 equally sized partitions. Bins 1–5 and 16–20 "bookend" the TAD, whereas the center bins 6–15 are inside the TAD (see 3.4:Methods). In addition to characterizing heritability patterns in bins across the TAD landscape, we also explicitly defined TAD boundary windows as fixed-size (40 kb, 100 kb, or 200 kb) regions bookending TADs. We conducted S-LDSC across the 37 cell types for the 41 traits to estimate the enrichment (or depletion) of heritability for each trait across the 20 partitions over the TAD landscape and the 100 kb TAD boundaries.

3.2.2 TAD boundaries are enriched for complex-trait heritability and evolutionary sequence conservation

Regions flanking TADs are enriched for complex-trait heritability; whereas partitions in TADs are marginally depleted for heritability overall (1.07× enrichment in flanking regions versus 0.99× enrichment in TADs, $P = 1 \times 10^{-193}$) (Fig. 3.2A). We also observed enrichment in regions flanking TADs when when we used the 100 kb TAD boundary definition (1.07× background, $P = 0.001$, Fig. 6.19). The results are consistent whether averaged across traits or meta-analyzed with a random-effects model[75,92,207] ($r^2 = 0.85$, $P = 7 \times 10^{-9}$ Fig. 6.17); therefore, further analyses of heritability across traits will use averaging for simplicity and interpretability. There is also a spike of heritability enrichment in the center of TADs; we explore this further in a subsequent section.

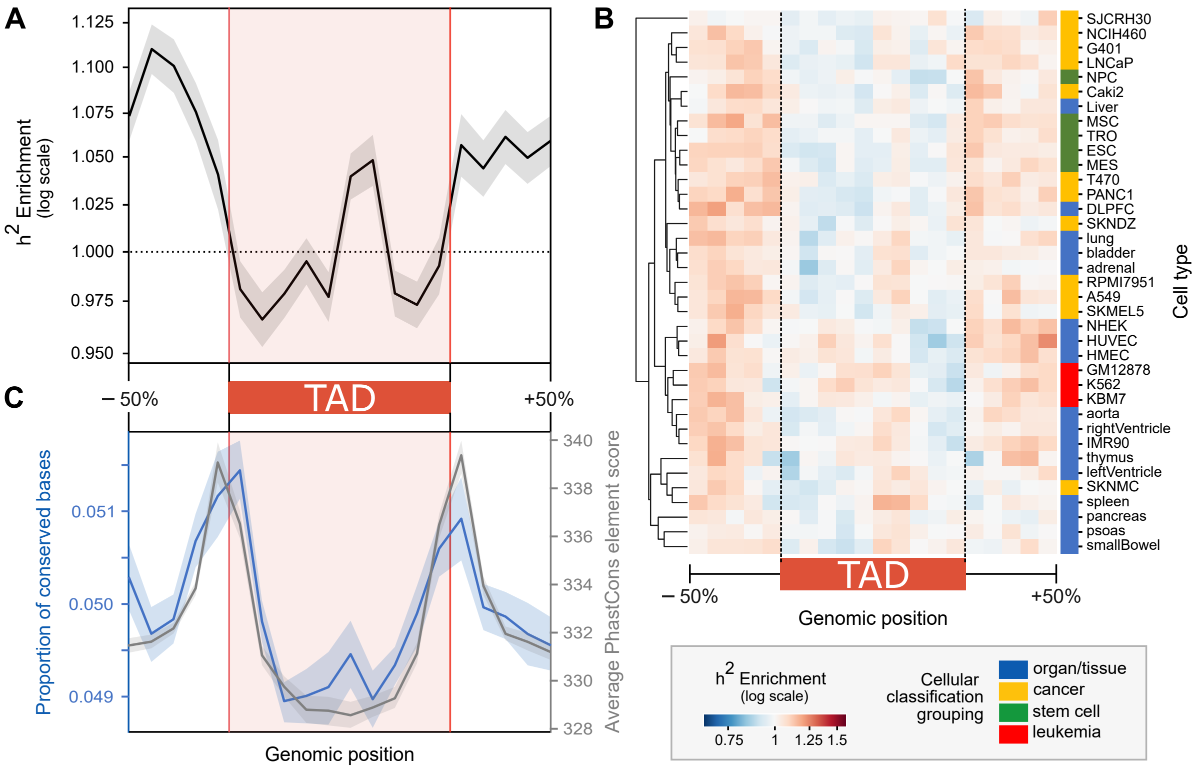


Figure 3.2: Regions flanking TADs are enriched for heritability of diverse common complex traits and evolutionary sequence conservation.
(A) Contribution to trait heritability (h^2) is enriched across variation in TAD-flanking regions and in the center of TADs when averaged across 41 common complex phenotypes and TAD maps from 37 cell types ($P = 1 \times 10^{-193}$). Enrichment was computed within 20 equally sized bins centered on each TAD ±50% of its length. **(B)** Heritability patterns are consistent across the 3D genome landscape for 37 cell types. **(C)** Regions flanking TADs have increased sequence-level constraint. They have a higher proportion of conserved bases (overlap with PhastCons elements; $P = 5 \times 10^{-11}$) (left blue axis) and a higher average conservation score across those overlapping PhastCons elements (right gray axis; $P = 3 \times 10^{-29}$). Error bands signify 99% confidence intervals. Trends are similar for fixed-size 100 kb TAD boundaries bookending TADs; TAD boundaries are enriched for heritability ($P = 0.001$, Fig. 6.19) and conservation ($P = 3 \times 10^{-29}$, Fig. 6.20A).

The complex-trait heritability enrichment flanking TADs is also consistent across cell types (Fig. 3.2B). The heritability enrichment values are significant but relatively small in magnitude. This is expected in light of the large genomic regions considered by this analysis—only a small fraction of the base pairs in a boundary are likely to be functionally relevant.

To assess functionality via a complementary approach, we compared between-species sequence-level conservation for TADs and boundaries. Regions flanking TADs are more evolutionarily conserved than sequences in TADs (Fig. 3.2C). We quantified evolutionary conservation in terms of the proportion of base pairs in a region in a conserved element identified by PhastCons elements and by the average PhastCons element score across the region. On average, 5.02% of regions flanking TADs are overlapped by PhastCons elements, versus 4.97% of TADs ($P = 5 \times 10^{-11}$), Fig. 3.2C). Furthermore, across these PhastCons elements, regions flanking TADs have average higher conservation scores than TADs (334 versus 331, $P = 3 \times 10^{-29}$, Fig. 3.2C). The 100 kb TAD boundary set corroborates these results; 5.21% of bases in TAD boundaries are conserved versus 4.91% in intra-TAD 100 kb windows ($P = 3 \times 10^{-29}$, Fig. 6.20A). This supports previous findings underscoring the importance of maintaining TAD boundaries.

The heritability enrichment and conservation at TAD boundaries are most likely due to their known overlap with functional elements such as CTCF binding sites and genes. Many such elements are enriched for heritability and conservation themselves[75]. To assess whether the heritability enrichment flanking TADs is greater than expected given the known functional elements overlapping TAD boundaries, we calculated standardized enrichment effect sizes (τ_c^*)[92,155]. This statistic quantifies heritability unique to the focal annotation by conditioning on a broad set of 86 gene regulatory, evolutionary, gene, allele frequency, and LD-based annotations (baseline v2.1)[75,155,169,175]. TAD boundaries did not show more heritability than expected on the basis of their enrichment for the 86 other annotations (Fig. 6.21). Similarly, to assess whether the greater evolutionary conservation flanking TADs is the result of the known enrichment in functional elements, we evaluated the conservation of bases in 100 kb boundaries and matched intra-TAD windows that do not overlap CTCF ChIP-seq peaks or exons. Filtering the base pairs that overlap CTCF peaks, we found that TAD boundaries still overlap more PhastCons elements and have a higher average PhastCons element score than windows in TADs (Fig. 6.20). When removing all exonic base pairs, we found that TAD boundaries have less overlap with PhastCons elements than do windows in TADs. However, the conserved non-exonic regions of TAD boundaries have higher conservation scores than conserved non-exonic regions in TADs (Fig. 6.20). Thus, existing annotations probably capture most of the relevant functional elements (e.g., CTCF, genes, and other regulatory element-binding sites) that determine and maintain boundary function.

3.2.3 TAD boundaries vary in stability across cellular contexts

The heritability enrichment patterns we observed are similar across cell types, and TADs have been characterized as largely invariant across cell types[111,112,116–118]. However, previous work suggests distinct functional properties among TAD boundaries with different insulatory strengths, hierarchical structures, and cell types[137,200,201]. Thus, we hypothesized that the stability of TAD boundaries across cell types would be informative about their functional roles and conservation. To characterize the stability of TAD boundaries across diverse cellular contexts, we focused on the 100 kb bookended TAD boundaries (described above), since these can be directly compared across the 37 cell types. The maps for each cell type are defined with respect to the same 100 kb windows across the genome, so we identify shared, or "stable," boundaries on the basis of these 100 kb windows (Fig. 3.3A). Our results are robust to different definitions of TAD boundaries, including 40 kb windows surrounding (±20 kb) TAD start and stop sites ("40 kb boundaries") and 200 kb windows flanking the TAD start and stop sites ("200 kb bookend boundaries") (see Fig. 6.22 and 3.4:Methods).

Using the cross-cell-type TAD boundary intersection, we found that boundaries vary substantially across cell types. Less than 10% of TAD boundaries are shared in 25+ of the 37 cell types, and 22.6% of TAD boundaries are unique to a single cell type (Fig. 3.3B). With the more granular 40 kb boundaries, 33.9% of

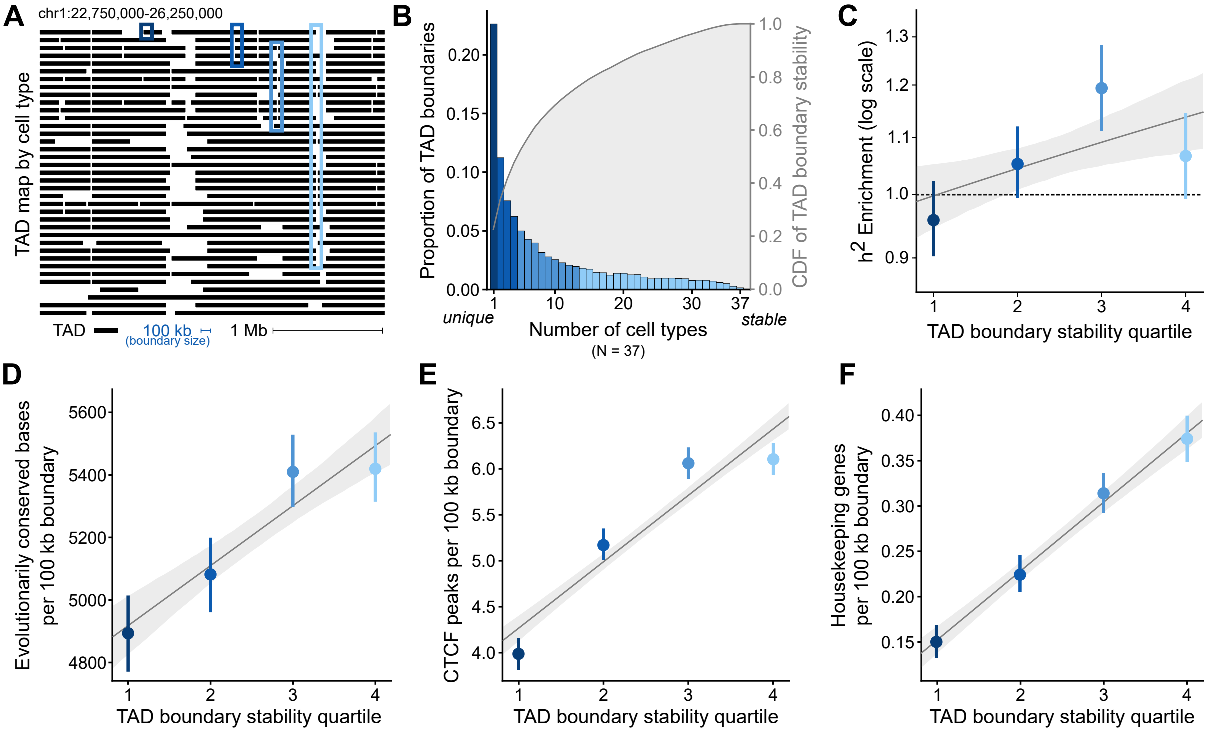


Figure 3.3: Stable TAD boundaries are enriched for complex-trait heritability, evolutionary conservation, and functional elements.

(A) Example TAD maps from 37 cell types (rows) for a 3.5 Mb window from human chromosome 1 (hg19). Each black line represents the genomic extent of a TAD. Example boundaries of different stability quartiles are outlined in blue (quartile 1 [most cell-type unique] in the darkest blue and quartile 4 [most cell-type stable] in light blue). **(B)** Histogram of TAD boundaries by the number of cell types they are observed in (this quantifies their "stability," colored by quartiles). The right axis and gray distribution represent the empirical cumulative distribution function (CDF) of boundary stability shown in the histogram. **(C–F)** Across TAD-boundary stability quartiles, there is a correlation between increased cell-type stability and increased **(C)** complex-trait heritability enrichment ($P = 0.006$), **(D)** conserved bases (overlap with PhastCons elements, $P = 6 \times 10^{-13}$), (E) CTCF binding (overlap with ChIP-seq peaks, $P = 1 \times 10^{-83}$), and (F) housekeeping genes ($P = 8 \times 10^{-58}$). All error bars signify 95% confidence intervals. These trends hold at different boundary definitions (40 kb and 200 kb), for germ-layer informed measures of cell type stability, and for other measurements of conservation, CTCF binding, and gene overlap (Figs. 6.25–6.28).

boundaries are unique to one tissue (Fig. 6.22A). Even with the permissive 200 kb resolution boundaries, 18.3% of boundaries are unique to a single tissue (Fig. 6.22B). To quantify boundary stability for further analyses, we bin boundaries into their cell-type stability quartile: boundaries present in only one context of 37 (cell-type unique) are in the first quartile of stability, boundaries in 2–4 cell types are in the second quartile, boundaries in 5–13 cell types are in the third quartile, and boundaries in 14 or more of the 37 contexts are the fourth quartile of cell type stability (Fig. 3.3B, examples in Fig. 3.3A).

Although there is high variability in the landscape of TAD boundaries across different cell types, we found that biologically similar cell types have more similar TAD boundary maps. For example, cell type classes (e.g., organ or tissue, stem cell, and cancer) generally cluster together. The two neuroblastoma cell lines cluster together, as do left ventricle, right ventricle, aorta, and skeletal muscle (Fig. 6.23B). This trend of biologically similar clusters also held at the 40 kb and 200 kb boundary resolution (Figs. 6.23A and

6.23C). Previous studies have found contrasting results about the level and patterns of similarity across cell types (Supplemental Text 6.2 in Appendix 2), but our similarity quantifications between cell types agree with some previous estimates[112,117,190,198].

In summary, although TADs and TAD boundaries have been characterized as largely invariant across cell types, we demonstrate that there is substantial variability between cell types[111,112,116–118]. We also find that biologically related cell types have more similar TAD maps, providing preliminary evidence for the cell-type specificity of the 3D genome and providing further rationale for investigating differences in TAD maps between cell types.

3.2.4 Stable TAD boundaries are enriched for complex-trait heritability, evolutionary constraint, and functional elements

When stratifying the 100 kb boundaries by their cell-type stability we found a positive relationship between cell-type-stability and trait-heritability enrichment ($r^2 = 0.045$, $P = 0.006$, Fig. 3.3C). The most stable boundaries (fourth quartile, darkest blue) have 1.07× enrichment of trait heritability, as opposed to 0.96× enrichment in unique boundaries (first quartile). This positive relationship between heritability and boundary stability holds at both the 40 kb and 200 kb resolution (Figs. 6.24A and 6.24D).

We also explored the relationship between TAD boundary stability and other evolutionary and functional attributes. Although TAD boundaries, when compared to TADs, are enriched for CTCF binding[111,200], evidence of evolutionary constraint[124,137] (Fig. 3.2C) and housekeeping genes are enriched at TAD boundaries[105,111] (compared to TADs), it is unknown how these features relate to boundary stability across cell types.

We found that TAD boundary stability is positively correlated with increased evolutionary sequence constraint (Fig. 3.3D, $P = 3 \times 10^{-13}$); compared to cell-type-unique TAD boundaries, boundaries in the highest quartile of stability have an additional 527 base pairs of overlap with PhastCons elements (5,420 versus 4,893 per 100 kb boundary). This extends previous observations that investigated two cell types to show that shared boundaries have evidence of stronger purifying selection on structural variants than boundaries present in only one of the cell types[137]. On the basis of on our result, we conclude that stable boundaries are more intolerant of disruption, not only on the scale of structural variants, but also at the base-pair level.

TAD boundary stability is also correlated with increased CTCF binding (Fig. 3.3E, $P = 1 \times 10^{-83}$). Boundaries in the highest quartile of stability have 1.5× more CTCF sites on average than TAD boundaries unique to one cell type (6.1 versus 4.0). This aligns with previous findings that boundary insulatory strength (in a single cell type) is positively associated with CTCF binding[111,200]; however, it expands this finding to stability across cell types.

Finally, we found that TAD boundary stability is correlated with increased overlap with genes (1.56×, Figs. 6.25A–6.25C, $P = 1 \times 10^{-74}$), protein-coding genes (1.65×, Figs. 6.25D–F, $P = 7 \times 10^{-90}$), and housekeeping genes (2.50××, Figs. 3.3F, 6.25G–I, $P = 8 \times 10^{-58}$). Boundaries in the highest quartile of stability overlap 2.5× more housekeeping genes than do cell-type-unique TAD boundaries (0.37 versus 0.15 per 100 kb boundary). The relationship between stable TAD boundaries and housekeeping-gene enrichment might result from many factors, including strong enhancer-promoter interactions, specific transcription-factor binding, or chromatin insulation caused by highly active sites of transcription[129].

Motivated by the observation that closely related cell types have more similar boundary maps (Fig. 6.23) and given the non-uniform sampling of cell types considered here, we defined an additional measure of boundary stability based on cellular development. We determined the germ layer of origin (endoderm, mesoderm, ectoderm) for each of the 37 cell types and stratified boundaries on the basis of their presence across cells of different origins. Consistent with our results based on the raw count of cell types, boundaries observed in cell types from all three germ layers are enriched for trait heritability, conserved bases, CTCF binding, and housekeeping genes in comparison to boundaries unique to one germ layer (Fig. 6.28). This shows that the greater contribution to complex trait heritability for more stable boundaries is probably robust to the sample of cell types considered.

Although our measure of TAD boundary stability correlates highly with these functional annotations, we note a slight drop-off in enrichment at the fourth quartile (compared to the third quartile), especially for trait heritability, conservation, and CTCF binding (Figs. 3.3C–E). We identify two factors—one technical and one biological—contributing to this. First, TADs must necessarily start and stop at the edges of chromosomes, centromeres, and gap regions; these regions will be identified as highly stable TAD boundaries independent of their functional importance and constraint. When boundaries within 5 Mb of genomic gaps[179,208] or blacklist regions are removed[209], the enrichment drop-off is diminished (Fig. 6.29). Second, the 37 cellular contexts considered are not uniformly sampled; some are more closely related than others. Thus, a boundary present in a well-sampled set of cell types might appear more stable than a boundary present in less densely sampled cell types. The germ-layer-based definition of stability has lower resolution but is less subject to sampling biases. We do not observe a decrease in the enrichment for heritability or other functional annotations among the most stable set when we use the germ-layer stability scores (Fig. 6.28). Thus, it will be important in future work to incorporate more detailed understanding of the developmental relationships of the considered cell types into comparisons of TAD maps.

In summary, TAD boundaries stable across multiple cell types are enriched for complex-trait heritability, evolutionary constraint, CTCF binding, and housekeeping genes. These trends hold at different boundary definitions (40 kb and 200 kb), for germ-layer-informed measures of cell type stability, and for other measurements of conservation, CTCF binding, and gene overlap (Figs. 6.25–6.28).

3.2.5 The heritability landscape across the 3D genome varies across phenotypes

The previous analyses have shown that trait heritability is generally enriched at TAD boundaries and further enriched in boundaries stable across cell types. Given preliminary evidence that different traits have unique enrichment profiles among different functional annotations[75], we hypothesized that variation in TAD boundaries might influence certain traits more than others. To investigate trait-specific heritability across the TAD landscape, we computed heritability enrichment profiles across the 3D genome partitions by trait and hierarchically clustered them (Fig. 3.4A). We observed two distinct trait clusters (Fig. 3.4A, Table 6.8).

One cluster of traits ("boundary-enriched" cluster) is strongly enriched for complex-trait heritability at regions flanking TADs (Fig. 3.4B) and in the 100 kb TAD boundaries (Fig. 6.19). Across TAD maps

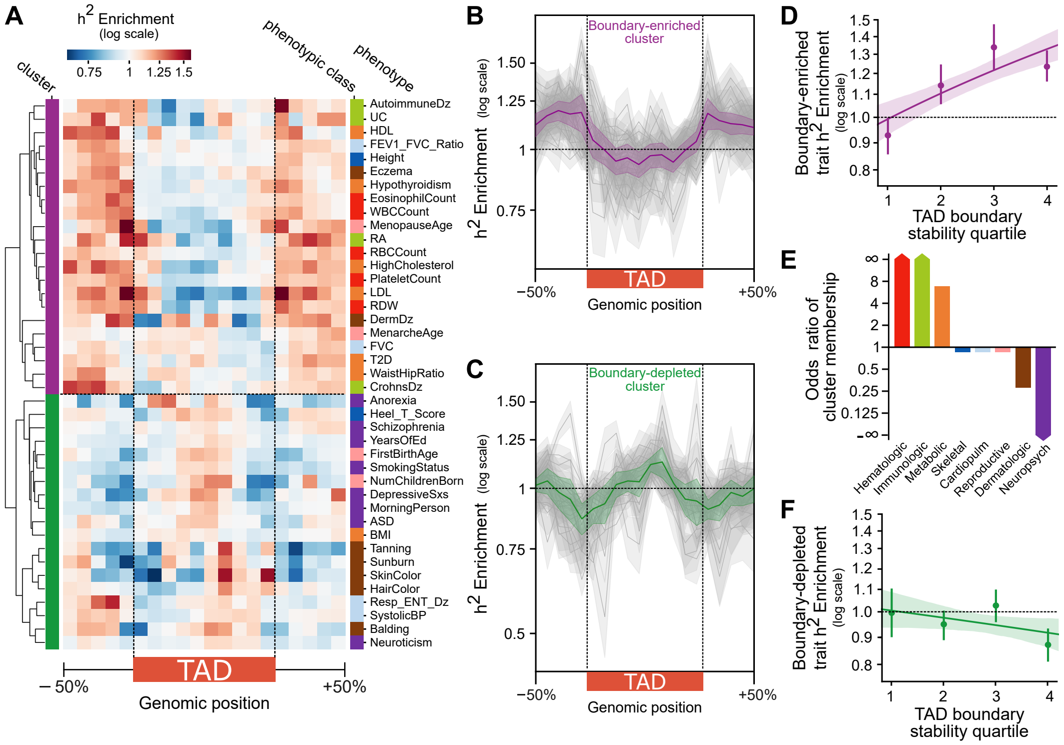


Figure 3.4: The heritability landscape across the 3D genome varies across phenotypes.
(A) Trait heritability patterns across the 3D genome organize into two clusters. Some traits are strongly enriched for complex-trait heritability at TAD boundaries ("boundary-enriched" cluster, purple), whereas others are weakly depleted at TAD boundaries and enriched centrally within the TAD ("boundary-depleted" cluster, green). **(B)** Heritability enrichment landscape over TADs for traits in the boundary-enriched cluster ($n = 22$). The gray lines represent the heritability pattern for each trait in the cluster; the purple line is the average over all the traits. **(C)** Heritability enrichment landscape over TADs for traits in the boundary-depleted cluster ($n = 19$). The green line is the average over all the traits. **(D)** The positive correlation between boundary stability and trait heritability (Fig. 3.3C) is driven by the subset of traits in the boundary-enriched cluster ($r^2 = 0.23$, $P = 2 \times 10^{-6}$). **(E)** Odds of cluster membership across phenotype categories. The boundary-enriched cluster is predominantly hematologic, immunologic, and metabolic traits. The boundary-depleted cluster is predominantly neuropsychiatric traits. **(F)** There is a weak negative correlation between boundary stability and trait heritability for traits in the boundary-depleted cluster ($r^2 = 0.04$, $P = 0.09$). Error bars signify 99% confidence intervals in (B) and (C) and 95% confidence intervals in (D) and (F).

in 37 cell types, these traits have on average 1.16× heritability enrichment at 100 kb TAD boundaries in comparison to genomic background ($P = 1 \times 10^{-7}$, Fig. 6.19). The other cluster of traits ("boundary-depleted" cluster) shows a weak inverted pattern in comparison to the boundary-enriched cluster; there is marginal heritability depletion at TAD boundaries (0.97× enrichment, $P = 0.06$, Fig. 6.19) and a spike of heritability enrichment within the TAD center (Fig. 3.4C).

The traits in the boundary-enriched cluster are predominantly hematologic (e.g., counts of white and red blood cells), immunologic (e.g., rheumatoid arthritis, Crohn disease), and metabolic traits (e.g., type 2 diabetes, lipid counts) (Fig. 3.4E). The traits in the boundary-depleted cluster are mostly neuropsychiatric (e.g., schizophrenia, years of education, Autism spectrum disorder) and dermatologic (e.g., skin color, balding) (Fig. 3.4E). This stratification of complex diseases into phenotypic classes does not perfectly reflect the traits' pathophysiology. For example, some dermatologic traits fall into the boundary-enriched cluster. However, these dermatologic traits, such as eczema, also have a substantial immunologic and hematologic basis, which is a hallmark of other traits in the boundary-enriched cluster. Additionally, body mass index (BMI) clustered with the psychiatric-predominant boundary-depleted cluster instead of with other metabolic traits in the boundary-enriched cluster. This is interesting in light of previous findings that BMI heritability is enriched in central nervous system (CNS)-specific annotations rather than metabolic-tissue (liver, adrenal, pancreas) annotations[75]. Skeletal, cardiopulmonary, and reproductive traits do not consistently segregate into one of the clusters (Fig. 3.4E). This is most likely because of the small sample size and heterogeneity of traits in these phenotypic classes.

The relationship between heritability enrichment in TAD boundaries and the trait clusters is not confounded by GWAS trait sample size (n), number of SNPs (M), or the traits' SNP-based heritability (h^2_{SNP}) (Fig. 6.30). Despite using a diverse set of cell types, we recognize that the heritability pattern differences between traits could be affected by the representation of investigated cell types. However, given that the pattern of heritability enrichment is consistent across all cell types (Fig. 3.2B), we are confident that no single cluster of cell types is driving the differences in heritability patterns between traits. Furthermore, these patterns are maintained even when we call TADs by a variety of computational methods (Armatus, Arrowhead, DomainCaller, HiCseg, TADbit, TADtree, TopDom), suggesting that the finding of immunologic and hematologic heritability enrichment at TAD boundaries is robust to technical variation (Fig. 6.31).

Although analysis across all traits revealed a positive relationship between boundary cell-type-stability and heritability enrichment (Fig. 3.3C), we found that this trend is driven by traits in the boundary-enriched cluster: they have further heritability enrichment in cell-type-stable boundaries ($r^2 = 0.23$, $P = 2 \times 10^{-6}$, Fig. 3.4D). The most stable boundaries (fourth quartile) have $1.23\times$ enrichment of trait heritability as compared to $0.93\times$ enrichment in unique boundaries (first quartile). In contrast, traits in the boundary-depleted cluster have a non-significant negative relationship between stability and heritability ($r^2 = 0.04$, $P = 0.09$, Fig. 3.4F). These trends also hold when the germ-layer-informed measurement of boundary stability is used (Figs. 6.28C and 6.28D). Thus, boundary stability might be more relevant when interpreting variation associated with hematologic, immunologic, and metabolic traits.

3.3 Discussion

Although we are beginning to understand the role of 3D genome disruption in rare disease and cancer, we have a limited framework for integrating maps of 3D genome structure into the study of genome evolution and the interpretation of common disease-associated variation. Here, we show that TAD boundaries, in comparison to TADs, are enriched for common complex-trait heritability. Additionally, in exploring TAD boundaries stable across cell types, we find they are further enriched for heritability of hematologic, immunologic, and metabolic traits, as well as evolutionary constraint, CTCF binding, and housekeeping genes. These findings demonstrate a relationship between 3D genome structure and the genetic architecture of common complex disease and reveal differences in the evolutionary pressures acting on different components of the 3D genome.

Previous work has predominantly characterized the importance and evolutionary constraint of different components of the 3D genome from the perspective of SV and rearrangement events. We address the relationship between genome 3D structure across cell types at the level of common single nucleotide variation. We consider evolutionary constraint within humans (approx. 100,000 ya) and constraint across diverse vertebrate species (approx. 13-450 mya).

At the scale of common human variation, we show that TAD boundaries are enriched for common variants that account for the heritability of common complex traits. This relationship between 3D genome structure and common disease-associated variation aligns with the finding of Whalen et al.[124] that human haplotype breakpoints—which are associated with increased variation as a result of the mutagenic properties of recombination—are depleted at chromatin boundaries. Together, these findings suggest that TADs and TAD boundaries differ in their tolerance to genetic variation.

Over vertebrate evolution, we show that TAD boundaries have more sequence-level constraint than TADs. This provides a complementary perspective to that of Krefting et al.,[123] who found that human TAD boundaries are enriched for syntenic breaks when they compared humans to 12 other vertebrate species, and they thus concluded that intact TADs are shuffled over evolutionary time. While shuffling a TAD may "move" its genomic location, preserving the TAD unit also requires maintaining at least part of its boundary. Our work suggests that even though TADs are shuffled, the boundary-defining sequences are under more constraint than the sequences within the TAD. This is further supported by the high concordance of TAD boundaries within syntenic blocks across different species and by depletion of SVs at TAD boundaries in humans and primates[105,111,119–121,137].

Slight variation in 3D structure can cause large changes in gene expression[193,195]. For example, CTCF helps maintain and form TAD boundaries; consequently, altering CTCF binding often leads to functional gene expression changes, e.g., oncogenic gene expression in gliomas[135]. We hypothesize that altering gene regulation though common-variant disruption of transcription-factor motifs, such as CTCF, that are important in 3D structure organization contributes to the enrichment for complex-disease heritability. However, variation at TAD boundaries most likely also modifies genes or regulatory elements, such as enhancers, that are known to be enriched at boundaries without disrupting the TAD architecture. A deeper mechanistic understanding of TAD formation will be critical to further understanding how TAD-boundary disruption contributes to both rare and common disease at potentially nucleotide-level and cell-type resolution.

Our finding of divergent patterns of TAD boundary heritability enrichment for different traits (enrichment for hematologic, immunologic, and metabolic traits versus depletion for psychiatric and dermatologic traits) suggests that the 3D genome architecture might play differing roles in the genetic architecture of different traits. As a preliminary test of this hypothesis, we evaluated the relationship between boundary stability and intra-TAD heritability enrichment. We find that, for traits with heritability depletion at boundaries (psychiatric, dermatologic traits), TADs with stable boundaries have greater intra-TAD heritability enrichment (Fig. 6.32). Thus, for these traits, we speculate that stable boundaries might function to insulate important intra-TAD functional elements (e.g., enhancers or genes). This idea is consistent with previous work showing that super-enhancers are insulated by the strongest boundaries (in a single cell type)[200]. However, for the boundary-enriched traits (hematologic, immunologic, metabolic), we hypothesize that essential functional elements are enriched at the stable boundaries (rather than inside the TAD). This is supported by previous work that detected a positive association between genome-wide binding of CTCF, a transcription factor intimately involved in TAD boundary formation, and eczema, an immunologic trait that we identified as part of the boundary-enriched trait cluster[160]. Thus, it will be important to further explore how TAD boundaries (or other functional elements at TAD boundaries) might play different regulatory roles in different traits and diseases. This will be especially interesting to consider from an evolutionary perspective in light of evidence that certain subtypes of TADs, depending on the regulatory role of genes they contain, are under different selective pressures[210].

Finally, we identify substantial variation among 3D maps across cell types. Whereas TAD stability across cell types is greater than expected by chance, our findings expand the number and diversity of compared cell types and identify a large proportion of boundaries unique to single cell types (see Supplemental Text 6.2 in Appendix 2). Furthermore, using our measurement of cell-type stability to stratify TAD boundaries identifies meaningful biological differences: stable boundaries are enriched for common-trait heritability, evolutionary constraint, and functional elements. Although we identify this enrichment for stable boundaries, we anticipate that cell-type-specific TAD boundaries often have functional significance relevant to their context; however, we are underpowered to detect trait-heritability enrichment in cell-type-specific TAD boundaries.

Several limitations should be considered when interpreting our results. First, they are based on available Hi-C data and existing methods for calling TADs. The Hi-C data were generated by different groups, so there could be batch- or protocol-specific effects. However, previous work suggests that biological differences dominate lab-of-origin effects in comparisons of structural similarity[198]. Furthermore, we showed that the conclusions are robust to the computational method used (Fig. 6.31) and that our stability results are not contingent on the specific set of cell types considered (Fig. 6.28). Nonetheless, higher-resolution Hi-C across diverse cell types in multiple replicates is needed. Second, there is no standard for defining TAD boundaries. We use two complementary approaches and show our conclusions are robust. The first approach considers heritability across the 3D structural landscape by partitioning TADs and their flanking regions into 20 equal-size bins and enable comparison with previous work[123]. The second defines fixed-size boundaries at multiple resolutions: 40, 100, and 200 kb. Continued efforts to integrate data from multiple TAD-calling algorithms to more precisely define TAD boundaries, especially given their hierarchical nature, will further refine our observations[201,210]. Despite the complexities inherent in identifying TAD boundaries, our findings replicate with all our boundary definitions and with different TAD calling pipelines.

Here, we introduce a method for quantifying the stability of a TAD boundary across cell types and demonstrate enrichment of complex-trait heritability, sequence-level constraint, and CTCF binding among stable TAD boundaries. Our work suggests the utility of incorporating 3D structural data across multiple cell types to aid context-specific non-coding variant interpretation. Starting from this foundation, much further work is needed to elucidate the molecular mechanisms, evolutionary history, and cell-type-specificity of TAD-structure disruption. Furthermore, although we have focused on properties of TAD boundaries stable across cell types, it will also be valuable to identify differences in TAD boundary stability across species and find human-specific structures across diverse cell types[196]. Finally, as high-resolution Hi-C becomes more prevalent from diverse tissues and individuals, we anticipate that computational prediction of personalized cell-type-specific TAD structure[141,142] will facilitate understanding of how specific genetic variants are likely to affect 3D genome structure, gene regulation, and disease risk.

3.4 Methods

We examine heritability and functional annotation enrichment across the 3D genome landscape in two ways: (1) across the genome in windows centered and scaled around each TAD and (2) in fixed-size TAD boundaries defined with varying resolution (40–200 kb) at the ends of each TAD. We then characterize the stability of TAD boundaries across diverse cellular contexts. By splitting boundaries into quartiles of stability—from those unique to a single tissue to those shared across many tissues—we test whether there is a relationship between boundary stability and annotation enrichment. The annotations considered include contribution to complex trait heritability enrichment, base-pair-level evolutionary constraint, CTCF binding, and genic content. We demonstrate the robustness of our results by using multiple definitions of TAD boundaries, TADs called by a variety of methods, and different measurements of the annotations investigated to replicate our experiments.

Defining TADs

TAD maps for 37 different cell types were obtained from the 3D genome browser (Table 6.7)[202]. All TAD maps were systematically predicted from Hi-C data with the hidden Markov model (HMM) pipeline from Dixon et al.[111,116,202]. The maps were defined with respect to the same 40 kb windows, except in the case of seven cell types (GM12878, HMEC, HUVEC, IMR90, K562, KBM7, and NHEK) that were defined with respect to 25 kb windows. For details about the length and number of TADs per map, see Supplemental Text 6.2 in Appendix 2.

Quantifying partitioned heritability with S-LDSC

We conducted partitioned heritability by using stratified-LD Score Regression v1.0.1 (S-LDSC) to test whether an annotation of interest (e.g., TADs or TAD boundaries) is enriched for heritability of a trait[75,90]. We considered GWAS summary statistics from a previously described representative set of 41 diseases and complex traits (average $n = 329,378$, $M = 1,155,239$, $h^2_{SNP} = 0.19$, Table 6.8)[92,144–153]. Previous studies using these traits had GWAS replicates (genetic correlation > 0.9) for six traits (BMI, height, high cholesterol, type 2 diabetes, smoking status, years of education). For these, we considered only the GWAS with the largest sample size. All GWASs involved subjects of European ancestry only. We used 1000 Genomes for the LD reference panel[175] and HapMap Project Phase 3 (HapMap 3)[176] excluding the MHC region to estimate heritability enrichment and standardized effect size. Heritability was estimated from common variants with minor-allele frequency (MAF) > 0.05, and standard errors were computed by LDSC via a block-jackknife.

Heritability enrichment

S-LDSC estimates the heritability enrichment, defined as the proportion of heritability explained by single-nucleotide polymorphisms (SNPs) in the annotation divided by the proportion of SNPs in the annotation.

The enrichment of annotation c is estimated as

$$Enrichment_c = \frac{\%h^2_{(c)}}{\%SNP_{(c)}} = \frac{h^2_{(c)}/h^2}{|c|/M}$$

where $h^2_{(c)}$ is the heritability explained by common SNPs in annotation c, h^2 is the heritability explained by the common SNPs over the whole genome, $|c|$ is the number of common SNPs that lie in the annotation, and M is the number of common SNPs considered over the genome[75,92]. To investigate trends across all traits, we computed the average heritability enrichment and a confidence interval. When compared to meta-analysis using a random-effects model conducted with Rmeta[75,92,207,211], the trends are consistent (Fig. 6.17); therefore, we report results based on averaging to simplify interpretation and reduce over-representation of higher-powered GWAS traits.

Standardized effect size

In contrast to heritability enrichment, the standardized effect size (τ_c^*) quantifies effects that are unique to the focal annotation compared to a set of other annotations[92,155]. The estimate of τ_c^* is conditioned on 86 diverse annotations from the baseline v.2.1 model; these include coding, UTR, promoter and intronic regions, histone marks (H3K4me1, H3K4me3, H3K9ac, and H3K27ac), DNase I hypersensitivity sites (DHSs), chromHMM and Segway predictions, super-enhancers, FANTOM5 enhancers, GERP annotations, MAF bins, LD relation, and conservation annotations[75,155,169].

Heritability enrichment across the TAD landscape

We partitioned the genome with respect to TAD annotations by using two different strategies. In the first, motivated by Krefting et al.[123], we considered TADs plus 50% of their total length flanking each side and subdivided these into 20 equal-sized partitions. Hence, the center 10 bins (6–15) are inside the TAD. Bins 1–5 are upstream of the TAD, and 16–20 are downstream of the TAD. In cases where a TAD is adjacent to another TAD, the ± 50% region flanking the TAD (bins 1–5 and 16–20) often partially extends into a neighboring TAD (Fig. 6.18A). However, the ± 50% flanking region extends into the center of a neighboring TAD less than 20% of the time (Fig. 6.18B). We ran S-LDSR on these 20 bins across TAD maps from 37 cell types to calculate heritability enrichment over 41 traits. We investigated the heritability enrichment (or depletion) trends averaged across all traits and cell types, by cell type, and by trait. Second, we analyzed heritability in fixed-size TAD boundary windows of 40, 100, and 200 kb (see subsection on TAD stability below).

For the analyses by cell type and by trait, we clustered the heritability landscapes to determine whether related cell types or related traits had similar patterns of heritability across the 3D genome. To do so, correlation distance was used as the distance metric with average linkage clustering. When clustering traits by their heritability landscape across the 3D genome, we identified two agglomerative clusters and termed these “boundary enriched” and “boundary depleted.”

Evaluating robustness on other TAD callers

To assess the influence of technical variation of TAD calling on our findings, we assessed the heritability patterns in human embryonic stem cells across TADs called by seven diverse methods (Armatus, Arrowhead, DomainCaller, HiCseg, TADbit, TADtree, and TopDom). The TADs were called and published by Dali et al.[212] with Hi-C from Dixon et al.[116]

Sequence-level conservation across the TAD landscape

We considered PhastCons element overlap and score to quantify evolutionary constraint across the TAD landscape. Other researchers previously determined PhastCons elements by fitting a phylo-HMM across a group of 46 vertebrate genomes to predict conserved elements[213]. We downloaded these conserved element loci from the UCSC table browser[179,208]. Each element has a score describing its level of conservation (a transformed log-odds score between 0 and 1000). We intersected the PhastCons elements with regions of interest (e.g., TAD boundaries) across the TAD landscape. Across each region, we quantified the number of PhastCons base pairs (regardless of score) and the average PhastCons element score.

Evolutionary constraint in TADs versus boundary windows

To specifically measure the constraint in TAD boundaries versus TADs, we investigated base-pair-level conservation at 100 kb TAD boundaries (below) and matched randomly shuffled equally sized windows in TADs. For the windows in TADs, we shuffled the 100 kb boundaries for each of the 37 cell types three times and required them to fall inside TADs ($n = 111$). For both the TAD boundaries and TAD set, we calculated overlap with conserved (PhastCons) elements. To investigate whether conserved element overlap is influenced by the density of CTCF binding and exons, we repeated this analysis after subtracting bases (from both the boundaries and TAD windows) overlapping CTCF ChIP-seq peaks or exons.

Quantifying boundary overlap and stability

For each cell type, we defined a set of boundaries with regard to the same windows across the genome.

100 kb boundaries

We defined 100 kb boundaries (results shown in main text) as regions 100 kb upstream of the TAD start and 100 kb downstream of the TAD end. For example, if a TAD was at chr1: 2,000,000–3,000,000, we would define its TAD boundaries to be at chr1:1,900,000–2,000,000 (boundary around the start) and chr1: 3,000,000–3,100,000 (boundary around the end). To quantify stability, we examined each 100 kb window across the genome. We removed boundaries that had any overlap with genomic gaps (centromeric/telomeric repeats from UCSC table browser)[179,208]. If there was a TAD boundary in the window for any of the cell types, we counted how many cell types (out of 37) shared the boundary. If only one cell type had a boundary at that location, it was considered a “unique” boundary, whereas if it was observed in many cell types, it was considered “stable.” These boundaries were divided into quartiles of cell-type-stability.

40 kb and 200 kb bookend boundaries

To test whether our results were robust to different resolutions of boundary definitions, we defined 40 kb and 200 kb bookend boundaries (see results in Supplemental Text 6.2 in Appendix 2). 40 kb boundaries are 40 kb windows surrounding (±20 kb) TAD start and stop sites. For example, if a TAD was located at chr1: 2,000,000–3,000,000, we would define its TAD boundaries to be at chr1: 1,980,000–2,020,000 and chr1: 2,980,000-3,020,000. 200 kb bookend boundaries are 200 kb upstream of the TAD start and 200 kb downstream of the TAD end. For example, if a TAD was at chr1: 2,000,000–3,000,000, we would define its TAD boundaries to be at chr1: 1,800,000–2,000,000 and chr1: 3,000,000–3,200,000. We removed boundaries that had any overlap with genomic gaps[179,208]. Both sets of boundaries were divided into quartiles of cell-type-stability.

Boundaries distant from genomic gap or blacklist regions

To investigate whether boundaries near genome assembly gaps or repetitive sequences affect the relationship between annotation enrichment and stability quartile, we defined a very conservative set of 100 kb TAD boundaries by excluding those within 5 Mb of a genomic gap (UCSC table browser[179,208]) or blacklist region ([209]).

Germ-layer-informed boundary-stability measure

Of the 37 cell types considered, some are more closely related than others, therefore we grouped 34 of them by germ-layer origin (endoderm [$n = 12$], mesoderm [$n = 13$], ectoderm [$n = 9$]; Table 6.7). Germ layers for each of the cell types were defined via ENCODE documentation of common cell types[203,204]. Embryonic stem cell, mesendoderm, and trophoblast were omitted because they have no single germ-layer classification. We defined a measurement of stability on the basis of whether each 100 kb boundary (above) was found in cells from one, two, or all three germ layers.

Quantifying TAD boundary similarity across cell types

To quantify TAD boundary similarity between two cell types, we calculate the Jaccard similarity coefficient by counting the number of shared boundaries (intersection) and dividing by the total boundaries over both tissues (union). For the TAD boundary similarity heatmaps, we clustered the cell types by using complete linkage (i.e., farthest neighbor) with the Jaccard distance (1-stability).

Heritability and annotation enrichment by TAD boundary stability

Complex-trait heritability

S-LDSC was conducted on each quartile of stability for all 41 traits. Partitions for each quartile include TAD boundaries of that stability (see above). We computed a linear regression on log-scaled enrichment values by regressing log10(heritability enrichment) on quartile of stability. by regressing log10(heritability enrichment) on quartile of stability.

Evolutionary constraint

Evolutionary constraint was quantified by PhastCons[213] as described above. The PhastCons elements were intersected with the TAD boundaries, partitioned by stability. The two overlap quantifications are the number of PhastCons base pairs per boundary regardless of score (base pairs per boundary) and the average PhastCons element score per boundary (average score of elements in the boundary).

CTCF enrichment

CTCF binding sites were determined through ChIP-seq analyses from ENCODE[203,204]. We downloaded all CTCF ChIP-seq data with the following criteria: experiment, released, ChIP-seq, human (hg19), all tissues, adult, BED NarrowPeak file format. We excluded any experiments with biosample treatments. Across all files, the CTCF peaks were concatenated, sorted, and merged into a single file; thus, overlapping peaks were merged into a single larger peak. We quantified the number of CTCF ChIP-seq peaks per TAD boundary (peaks per boundary) and the number of CTCF peak base pairs overlapping each boundary (base pairs per boundary).

Genes and protein-coding genes

RefSeq genes were downloaded from the UCSC table browser[179,208,214] and filtered to include coordinates of only one transcript per gene (the longest) and only autosomal and sex chromosome genes. From the simplified list of RefSeq genes, a subset of protein-coding genes was also created (these were identified on the basis of RefSeq accession numbers starting with NM). The simplified RefSeq gene list contains 27,090 genes. The simplified protein-coding RefSeq gene list contains 19,225 genes. We quantified the number of genes or protein-coding genes per TAD boundary stratified by boundary stability.

Housekeeping genes

Housekeeping genes ($N = 3804$) are from Eisenberg & Levanon[215]. We retrieved the coordinates by intersecting with the RefSeq genes (above), resulting in coordinates for 3681 genes (coordinates for a small number of genes were not found in the RefSeq list)[179,208,214]. We quantified the number of housekeeping genes or protein-coding genes per TAD boundary stratified by boundary stability.

Defining GWAS phenotypic classes

To determine whether similar traits had similar heritability patterns across the 3D genome, we defined eight different phenotypic classes (Table 6.8): cardiopulmonary ($n = 4$), dermatologic ($n = 7$), hematologic ($n = 5$), immunologic ($n = 4$), metabolic ($n = 7$), neuropsychiatric ($n = 8$), reproductive ($n = 4$), and skeletal ($n = 2$). Our clusters originated from domains in the GWAS Atlas[159]; however, the categories were modified to place more emphasis on disease pathophysiology instead of organ system (e.g., Crohn disease and Rheumatoid Arthritis were moved from the gastrointestinal and connective-tissue categories, respectively, to an immunologic category). Similar categories were also combined (e.g., metabolic and endocrine, cardiovascular and respiratory).

Data analysis and figure generation

All analyses were conducted with the hg19 genome build. Intersections of genomic regions were computed with the pybedtools wrapper for BedTools[181,216]. Data and statistical analyses were conducted in Python 3.5.4 (Anaconda distribution) and R 3.6.1. Figure generation was aided by Matplotlib, Seaborn, and Inkscape[182–184]. This work was conducted in part with the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN.

Data and Code Availability

The datasets we generated are available in the TAD-stability-heritability GitHub repository [https://github.com/emcarthur/TAD-stability-heritability][217] and at Zenodo: https://doi.org/10.5281/zenodo.3601559 and include all results of our boundary calling (40 kb, 100 kb bookend, and 200 kb bookend) and all partitioned heritability analysis output (by cell type and trait). The repository also contains a Jupyter notebook with code for analysis, statistics, and figure generation.

CHAPTER 4

RECONSTRUCTING THE 3D GENOME ORGANIZATION OF NEANDERTHALS REVEALS THAT CHROMATIN FOLDING SHAPED PHENOTYPIC AND SEQUENCE DIVERGENCE[*]

4.1 Introduction

The sequencing of archaic hominin (AH) and modern human (MH) genomes has transformed our understanding of human history, evolution, and biology[13–17]. However, even with these whole-genome sequences available, our understanding of how and why AHs differed from MHs is limited[2]. A major challenge in understanding the phenotypic and sequence differences between AHs and MHs is bridging the gap between genetic variation and function. The evolution of hominins is largely driven by changes in the regulation of conserved proteins[7–10,55,57,94], but the mechanisms through which archaic variants influence gene expression, and ultimately phenotype, are incompletely understood[2,94,98].

Previous works have theorized the mechanistic effect of archaic variants on enhancers, promoters, miRNA, cis-eQTLs, and DNA methylationColbran et al.[94], Gokhman et al.[96], Batyrev et al.[97], and Silvert et al.[98]. Yet, these studies have been unable to address a fundamental aspect of gene regulation and genome function—the physical three-dimensional (3D) organization of the genome. Regulation of gene expression is facilitated by the 3D looping and folding of chromatin in the cell nucleus, which is central to enhancer-promoter (E-P) communication and insulation[186–189,218–220]. Thus, to fully understand the consequences of genetic variation between AHs and MHs, we must consider the 3D genome folding. However, the role of 3D genome organization in the divergence between AHs and MHs has never been explored because chromatin contacts cannot be assayed in ancient DNA.

Recent deep learning methods have been developed that learn the sequence "grammar" underlying 3d genome folding patterns[140–143]. We hypothesized that these deep learning methods would allow us to infer genome-wide 3D chromatin contact maps of Neanderthals and Denisovans. Because the molecular mechanisms that determine genome organization, like CTCF binding and co-localization with cohesin, are largely evolutionarily conserved[111,121], models trained using human data perform well even when applied to DNA sequences from distantly related species, such as mouse[141].

To elucidate the contribution of 3D genome folding to recent hominin evolution, we apply novel deep learning methods for inferring 3D genome organization from DNA sequence patterns to Neanderthal, Denisovan, and diverse MH genomes. Using the resulting genome-wide 3D genome folding maps, we identify 167 loci that are divergent in 3D organization between AHs and MHs. We show that these 3D-diverged loci are enriched for physical links to genes related to the function and morphology of the eye, supra-orbital ridge, hair, lung function, immune response, and cognition. We also find that 3D genome organization constrained recent human evolution and patterns of introgression. Finally, we evaluate the legacy of introgression on the 3D organization of humans and identify examples where introgression imparted divergent 3D genome folding to Eurasians. In summary, our application of deep learning to predict archaic 3D genome fold-

[*] This chapter is adapted from the preprint McArthur, E. *et al.* 2022. *BioRxiv.*

ing provides a window into previously unobservable molecular mechanisms linking genetic differences to phenotypic consequences in hominin evolution.

4.2 Results

4.2.1 Reconstructing the 3D genome organization of archaic hominins

To evaluate the role of 3D genome organization changes in recent human evolution, we apply deep learning to infer 3D genome organization from DNA sequences of archaic hominins (AHs) and modern humans (MHs) (Fig. 4.1). We consider the genomes of four AHs—one Denisovan and three Neanderthals, each named for where they were discovered (*Altai* mountains, *Vindija* and *Chagyrskaya* caves) [13–16]. We compare these to 20 diverse MHs from the 1000 Genomes Project (Table 6.9)[175].

For each individual, we predict chromatin contact maps across the genome. Each contact map gives a 2D representation of the predicted 3D chromatin physical contacts, which will refer to as "3D genome organization". We predict these maps using approximately 1 Mb (1,048,576 bp) tiled sliding windows overlapping by half with Akita, a convolutional neural network (CNN) trained on high-quality experimental chromatin contact maps (Hi-C and Micro-C)[141]. Each resulting contact map represents pairwise physical 3D contact frequencies at approximately 2 kb (2,048 bp) resolution for a single individual. Previous work demonstrated that Akita accurately infers 3D contact organization at this resolution[141]. We only consider windows with full (100%) sequence coverage in the MH reference, and we conservatively mask missing archaic sequence with the human reference sequence (Figs. 6.34,6.35,6.36 and Methods).

We compare contact maps from two genomes using a "3D divergence" score, namely, one minus the Spearman's rank correlation coefficient ($1-\rho$) for all pixels in the maps. Genomic windows with more different 3D genome maps have higher 3D divergence and, conversely, a window with lower 3D divergence will reflect more 3D similarity (Fig. 4.1). Other divergence metrics (e.g., based on Pearson's correlation coefficient and mean squared difference) are strongly correlated (Fig. 6.37). Akita is trained simultaneously on Hi-C and Micro-C across five cell types in a multi-task framework. In the main text we focus on predictions from the highest resolution cell type, human foreskin fibroblast (HFF). Results are similar when considering other cell types (e.g. embryonic stem cells) (Fig. 6.38), likely because of limited cell-type-specific differences in both available experimental data and model predictions[141].

4.2.2 Archaic hominin and modern human genomes exhibit a range of 3D divergence

Reconstructing the genome-wide 3D genome organization of AHs and MHs revealed genomic windows with a range of 3D divergence (Fig. 4.2A). Most of the genome has very similar 3D genome organization between AHs and MHs (circle example in Fig. 4.2A-B). However, we also found regions of AH-MH 3D genome divergence. Some of these differences are changes in predicted chromatin contact intensity but similar overall organization (diamond example in Fig. 4.2A-B). Others reveal reorganization with evidence of new sub-organization (neo-TADs or -loops) or lost structures (fused TADs or loops) (indicated with an "x" example in Fig. 4.2A-B). At the 95th percentile of observed divergence, differences in the contact maps are substantial. However, because the 3D divergence measure considers the entire window, strong focal changes may not rank as highly as structural differences that influence a large segment of the window (diamond vs.

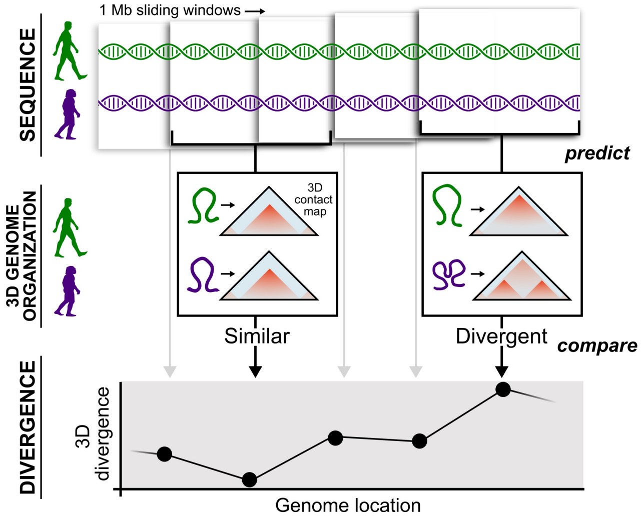


Figure 4.1: Reconstructing the 3D genome organization of archaic hominins.
We infer 3D genome organization from sequence across the genomes of modern humans (MHs, green) and archaic hominins (AHs, purple). Using approximately 1 Mb (1,048,576 bp) sliding windows (overlapping by half), we input the genome sequences into Akita, a convolutional neural network, to predict 3D genome contact maps[141]. The resulting contact maps are compared between MHs and AHs to identify regions that have similar 3D genome organization (left, low divergence) and regions that have different 3D organization (right, high divergence).

"x" examples in Fig. 4.2B).

To illustrate genome-wide patterns of divergence in 3D organization, we plotted the average divergence of each of the AHs to five modern African individuals from different subpopulations (Fig. 4.2C). We show the landscape of 3D divergence across the entire genome for all four AHs in Fig. 6.39. Some AH-MH divergences are shared across all four archaics, while others are specific to a single lineage like the Denisovan individual (Fig. 4.2C). We only considered sub-Saharan Africans in these comparisons, because they have low levels of AH introgression. We consider how introgressed variation in Eurasians influences 3D divergence in a subsequent section.

4.2.3 3D genome organization diverges between AH and MH at 167 genomic loci

To consistently identify regions with divergent 3D genome organization between AH and MH, we compared the 3D contact maps at each locus for each AH to 20 MH (African) individuals. We applied a conservative procedure that required all 20 AH-MH comparisons to be more 3D divergent than all MH-MH comparisons (Fig. 4.3A). In other words, the differences between the 3D genome organization of an AH to all MHs must be more extreme than the differences between each MHs to all other MHs. Furthermore, we required the average AH-MH 3D divergence to be in the 80th percentile of the most diverged. This identified regions with consistent 3D differences between AHs and MHs (Fig. 4.3A, left) while excluding regions with a large 3D diversity in modern humans (Fig. 4.3A, right) (Methods).

We find 167 total AH-MH consistently 3D diverged loci: 67, 70, 71, and 73 for Altai, Vindija, Chagyrskaya, and Denisova compared to MHs, respectively (Fig. 4.3B). 3D diverged loci are found throughout the genome on every chromosome (Fig. 4.3B). As suggested by Fig. 4.2C, some 3D divergences are shared by all four

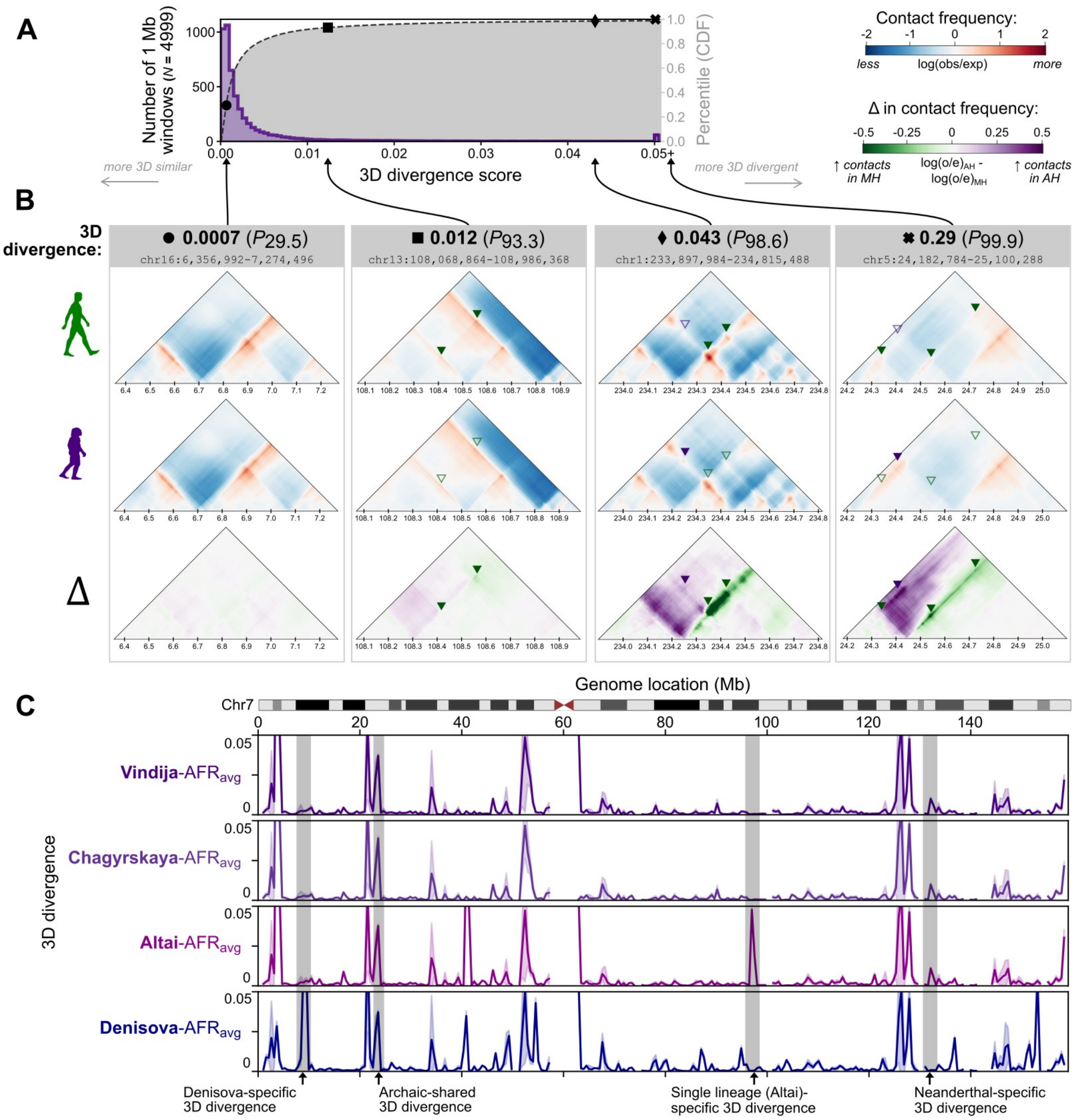


Figure 4.2: 3D genome divergence between archaic hominins (AHs) and modern humans (MHs) varies across the genome.

(A) Distribution of 3D genome divergence between AHs and modern humans MHs for 1 Mb windows across the genome. Most windows have similar 3D genome organization between MHs and AHs (low 3D divergence). The cumulative density function (CDF) of this distribution is overlaid in gray with percentiles on the right vertical axis. **(B)** We highlight four examples (shapes) along the 3D divergence distribution illustrating low 3D divergence (left) to high divergence (right). Each example compares a representative African MH (top, HG03105) to a Neanderthal (bottom, Vindija) in terms of both raw score and relative percentile of 3D divergence. Examples with scores near the 95th percentile have visible contact map differences, but the type of differences vary from re-organization (neo-TADs or TAD-fusions) to altered contact intensity (stronger vs. weaker TAD/loop). Green and purple triangles indicate regions with increased contact frequency in MH versus AH, respectively. **(C)** Average 3D divergence along chromosome 7 between AHs and five representative African MHs. The error band indicates the 95% confidence interval (CI). Comparing the 3D genomes of Neanderthals (purple) or Denisova (blue) with MHs reveals windows of both similarity and divergence (peaks). Featured examples (gray overlays) highlight regions of 3D divergence that are shared (e.g., shared across all archaics) or lineage-specific (e.g., specific to the Denisovan individual).

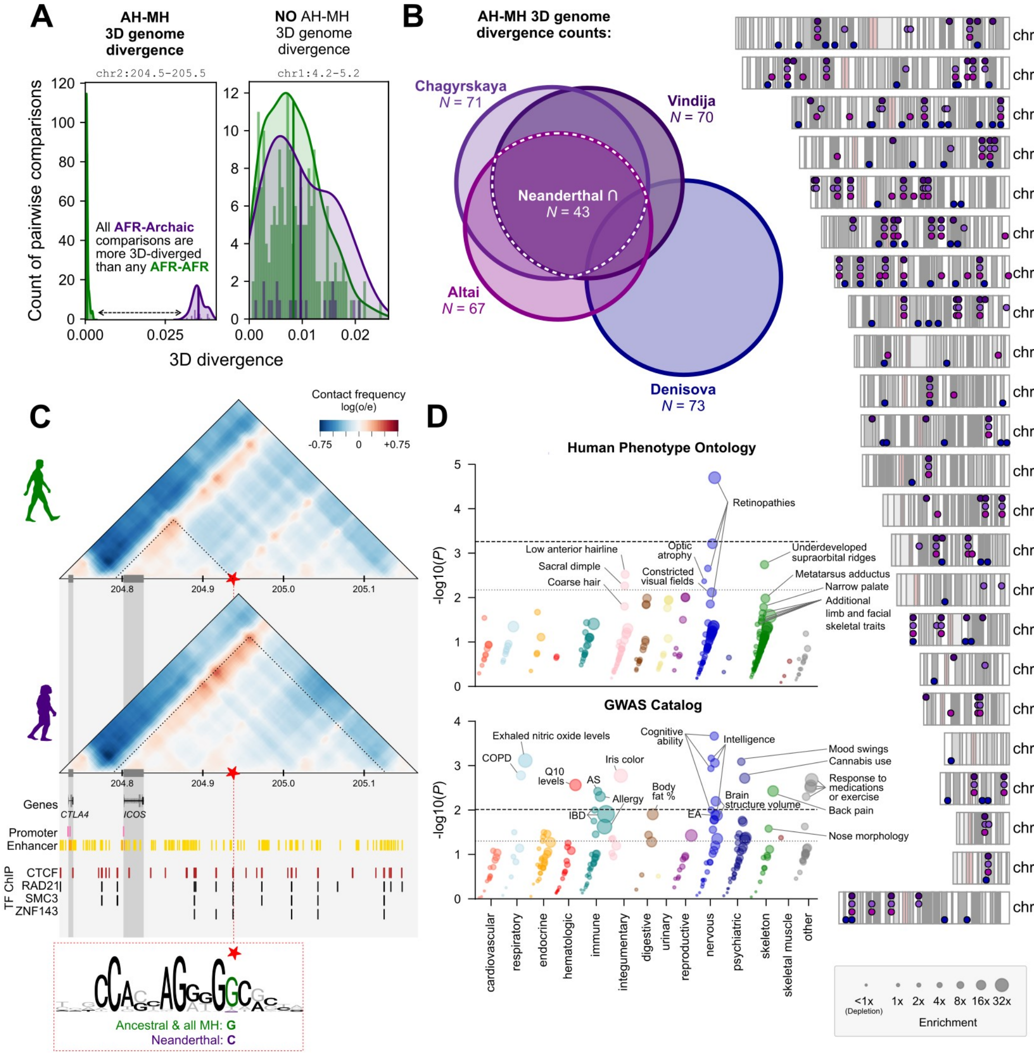


Figure 4.3: Regions with 3D divergence between MHs and AHs highlight loci linked to phenotypic differences. **(A)** We identified genomic windows with 3D divergence between AH and MH by comparing distributions of pairwise divergence in 3D contact maps. We used a conservative procedure that required all 20 comparisons of each AH to 20 MH (African) individuals (purple, $n = 20$) to be more 3D-diverged than all MH-MH comparisons (green, $n = \binom{20}{2}$ = 190) and the mean of the AH-MH divergences (purple) to be in the 95th percentile of most diverged. The left plot shows an example that meets these criteria (chr2:204,472,320-205,520,896). The right shows an example where there is diversity in 3D genome organization, but not an AH-MH divergence (chr1:4,194,304-5,242,880). **(B)** We identified 167 AH-MH 3D divergent windows across the genome. Many are shared (Euler-diagram), but some are unique to a single lineage, with the most unique divergence in the Denisovan. **(C)** Contact maps for the example Neanderthal-MH 3D divergent window shown in **A** (zoomed to chr2:204,722,176-205,166,592). All MHs have a smaller domain insulated by a CTCF site (red star). In Neanderthals (Vindija and Altai), the CTCF motif is disrupted with a C instead of a G (red dashed box, chr2:204,937,347). We predict that this leads to ectopic connections with the promoter of *ICOS* (T-cell costimulator). **(D)** Phenotype enrichment for the 43 Neanderthal 3D diverged loci identified in **B** (white dashed line). We computed functional annotation enrichment for genes physically linked to 3D-modifying variants at these 3D divergent loci using HPO (top, $n = 271$) and GWAS catalog (bottom, $n = 208$) annotations (Methods). Within each phenotypic domain, traits are organized along the vertical axis by significance and along the horizontal axis by enrichment (also indicated by size). Genes nearby AH-MH 3D divergence are enriched for functions related to the retina and visual field, skeletal morphology (notably, supra-orbital ridge), hair, lung function, immune and medication response, and cognitive traits. Significance lines represent the P-value thresholds that controls the FDR with $q = 0.05$ (dotted) and $q = 0.1$ (dashed). (COPD: chronic obstructive pulmonary disease, AS: ankylosing spondylitis, IBD: inflammatory bowel disease, EA: educational attainment)

AHs ($N = 7$), and many are shared by all three Neanderthals ($N = 43$) (Fig. 4.3B). We summarize the AH-MH 3D divergent windows in Tables 6.10.

To illustrate the properties of a AH-MH 3D divergent window, we highlight a divergent locus on chromosome 2 that is nearby several immune genes (Fig. 4.3C). MHs have an approximately 140 kb loop linking the promoter of *ICOS* at 204.80 Mb to a CTCF motif at 204.94 Mb. This CTCF motif is overlapped by many ChIP-seq peaks for transcription factors (TFs) involved in determining chromatin folding (CTCF, RAD21, SMC3, and ZNF143). The contact maps for both Vindija and Altai Neanderthal show a more prominent "architectural stripe"—an asymmetric loop-like contact often reflecting enhancer activity[108–110]—starting near the promoter of *ICOS*. However, in contrast to MHs, the loop does not end at the same CTCF site and instead has greater contact frequency with a CTCF site at 205.2 Mb. Thus, the resulting loop in Neanderthals is predicted to be over 400 kb—three times as large as the MH loop.

To determine which AH-MH nucleotide differences cause the largest change in the contact maps, we used *in silico* mutagenesis (Methods). Using an African MH (HG03105) background, we inserted every allele unique to the AH genome one-by-one and measured the resulting 3D genome divergence. This identifies the archaic variant resulting in the largest 3D organization changes between the AH and MH genomes, a G to C change at chr2:204,937,347 (Methods). This change disrupts a high information-content site in the CTCF binding site described above. All MHs carry an ancestral C allele, but Vindija and Altai have a derived G allele. In summary, we predict that the Neanderthal-derived allele weakens CTCF binding leading to reduced insulation between *ICOS*, a T-cell costimulator, with downstream contacts.

4.2.4 Regions with 3D divergence highlight AH-MH phenotypic differences

To explore the functional effects of AH-MH 3D genome divergence, we tested for phenotypic annotation enrichment. We considered the 43 loci with shared divergence between MHs and all three Neanderthals (Fig. 4.3B). Although the loci were identified at approximately 1 Mb resolution, most 3D modifications disrupt a smaller sub-window. Thus, as described in the example above (Fig. 4.3C), we used *in silico* mutagenesis to identify the AH-MH sequence change(s) that produced the largest disruption in the contact maps. We will refer to these as "3D-modifying variants" (Methods). We then intersected the predicted 3D-modifying variants with experimentally defined TADs to determine the genes to which they are physically linked. Ultimately, we found 88 physical links to protein-coding genes (85 unique genes) for the 45 3D-modifying variants in the 43 Neanderthal-MH 3D divergent loci (Tables 6.10).

We tested if these genes are enriched for phenotypic annotations using both gene-phenotype links from rare disease (OMIM Human Phenotype Ontology [HPO] terms) and common disease databases (GWAS Catalog 2019)[221–225]. 3D genome organization perturbation has been linked to both types of disease: large-scale disruption leading to severe disease and subtle changes in regulatory insulation contributing to complex traits disease[122,132,136,154,195]. We find links to 271 and 208 candidate traits from the rare and common disease ontologies, respectively. For each trait, we test if the observed overlap with 3D divergent loci is more than expected by chance using an empirically-generated null distribution (Methods). In summary, this sequential process links 3D divergent windows to variants to TADs to genes and, ultimately, phenotypes (Fig. 6.40).

With the HPO annotations, we found enrichment for effects of these genes related to the eye (retinopathies, optic atrophy, constricted visual field [most significant association: 27× enriched, $P = 2 \times 10^{-5}$]), skeletal system (notably, supraorbital ridge morphology [12×, $P = 0.002$]), and hair (e.g. low anterior hairline [12×, $P = 0.003$]) (Fig. 4.3D, top). In the GWAS Catalog annotations, we find enrichment related to intelligence and cognition (13×, $P = 0.0002$), lung function (NO levels, COPD [35×, $P = 0.0008$]), response to certain medications (30×, $P = 0.002$), immunologic response (ankylosing spondylitis, allergy, inflammatory bowel disease [12×, $P = 0.004$]), and brain region volumes (putamen, subcortex [17×, $P = 0.006$]) (Fig. 4.3D, bottom). Trait enrichments for 3D-modifying variants found in Denisova are highlighted in Fig. 6.41. Because Denisova and Neanderthal share many alleles, some similar traits are enriched (retinopathy, intelligence, lung function, etc.); however, overall, we find fewer enriched traits.

In summary, genomic loci with 3D divergence between Neanderthals and MHs are enriched for physical proximity to genes associated with a diversity of traits related to the skeleton, eye, hair, lung, immune response, brain region volume, and cognitive ability. These findings align with and expand what we know from both the fossil-record and previous work based on variants in MHs[55,59,62,63,98,156,226,227]. Importantly, our approach permitted the interrogation of variants unobserved in MHs (76% of predicted 3D-modifying variants), and it provides a putative molecular mechanism for the phenotypic differences.

4.2.5 Relationship between sequence divergence and 3D divergence

Given that we observe 3D differences between AH and MH genomes, we quantified the relationship between 3D and sequence divergence on both genome-wide and more local scales. First, we computed the genome-wide 3D genome divergence for all pairs of AH and MH individuals. We find the mean 3D genome divergence largely follows sequence divergence (Figs. 4.4A,6.42). Neanderthals are the most similar in 3D genome organization to other Neanderthals, then to the Denisova, and then to MHs (mean 3D divergences: 9.8×10^{-4}, 3.4×10^{-3}, and 4.3×10^{-3}, respectively). Genome-wide 3D divergence also tracks with sequence divergence within the Neanderthal: Vindija and Chagyrskaya are more similar than they are to the outgroup Altai (Vindija-Chagyrskaya mean 3D divergence of 8.4×10^{-4} vs. Vindija-Altai of 1.0×10^{-3})[15].

Next, we evaluated if sequence divergence and 3D divergence are correlated on the local scale. We find a very weak positive relationship between 3D and sequence divergence at the 1 Mb window level (Fig. 4.4B, $r^2 = 0.01$, $P = 2.3 \times 10^{-13}$). As suggested by the weak correlation, many windows with low sequence divergence have high 3D divergence, and many windows with high sequence divergence have low 3D divergence.

Given the weak relationship between sequence and 3D divergence, we sought to identify some properties of sequence differences that result in large 3D divergence. Based on the importance of CTCF-binding in maintaining 3D genome organization[102,105,219,228], we quantified the effects of AH-MH nucleotide differences overlapping CTCF binding motifs. Disruption of CTCF binding sites is important, but not all disruptions are likely to influence 3D divergence. Leveraging additional functional genomics data on CTCF binding and TAD boundaries, we find that the quantity, quality, and context (e.g., strength of a motif and proximity to a TAD boundary) influence whether AH-MH sequence divergence will result in a 3D organization divergence (Fig. 6.43). For example, if a window has at least one AH-MH nucleotide difference

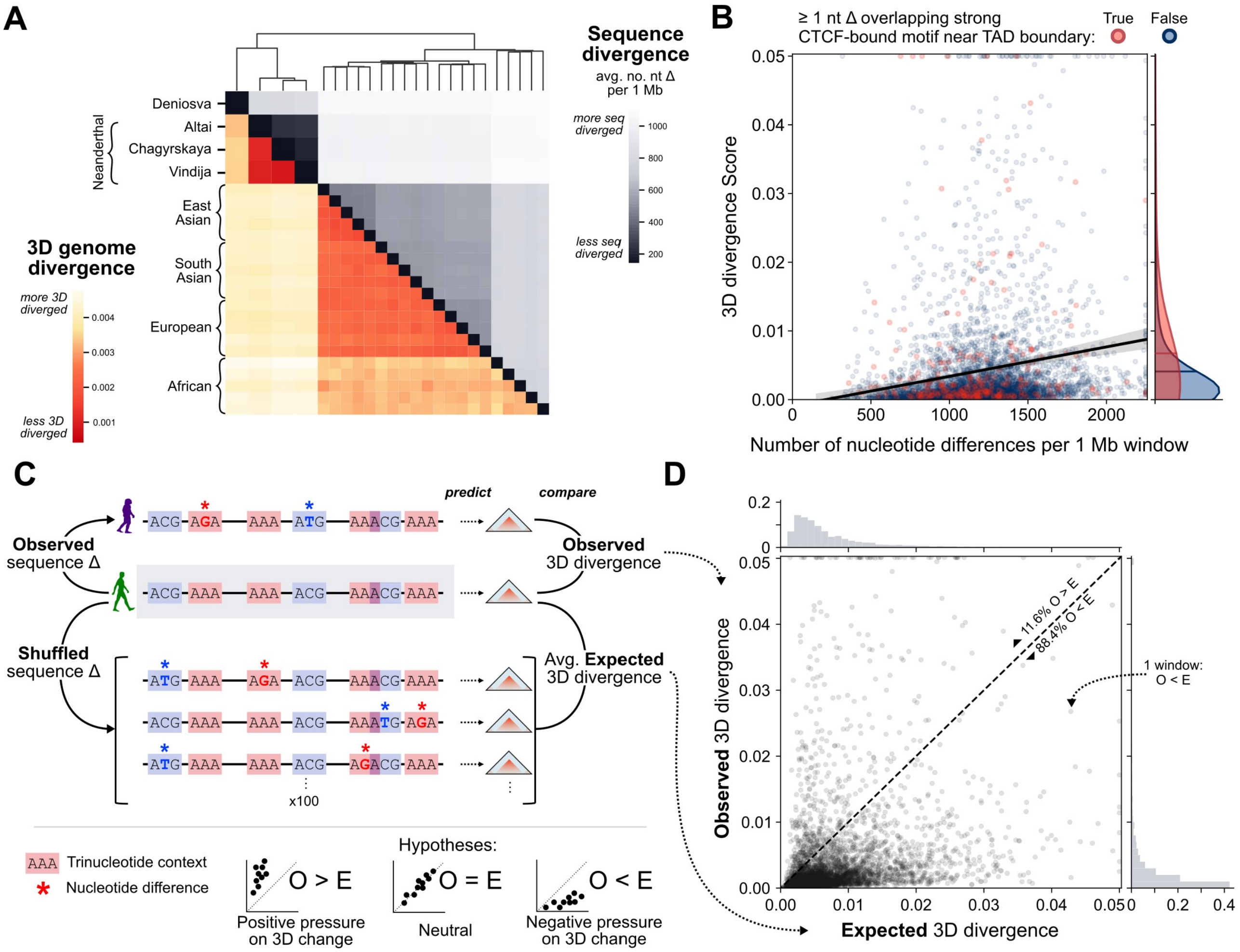


Figure 4.4: 3D genome organization constrained human sequence divergence.
(A) 3D genome divergence (lower triangle) follows patterns of sequence divergence (upper triangle). AHs have more similar 3D genome organization to each other than to 15 MHs from different 1000G super-populations. Clustering is based on sequence divergence; see Fig. 6.42 for clustering by 3D genome divergence and data for each sub-population. **(B)** Sequence divergence is only very modestly correlated with 3D genome divergence ($r^2 = 0.011$, $P = 2.3 \times 10^{-13}$, $N = 4999$). Each point represents a 1 Mb window from a genome-wide comparison between the 3D genome organization of a Neanderthal (Vindija) and African MH (HG03105) individual and the black line with band represents a linear regression with 95% CI. Windows with large 3D divergence are enriched for MH-AH nucleotide (nt) differences overlapping a strong CTCF-bound motif within 15 kb of a TAD boundary (red) (two-tailed Mann–Whitney U $P = 0.00077$). **(C)** To evaluate whether 3D genome organization constrained sequence divergence, we estimate the null distribution of expected 3D divergence based on sequence differences between the Neanderthal (Vindija) and African MH (HG03105) genomes. We shuffle observed nucleotide differences (stars) while preserving tri-nucleotide context (colored rectangles) and predict 3D genome organization for 100 shuffled sequences for each window. Under a model of no sequence constraint due to 3D organization, observed 3D divergence would equal the expected 3D divergence ($O = E$). Alternatively, observing more 3D divergence than expected would suggest positive selection on sequence changes that cause 3D divergence ($O > E$). Finally, observing less 3D divergence than expected would suggest negative pressure on sequence changes that cause 3D divergence ($O < E$). **(D)** Observed 3D divergence is significantly less than the mean expected 3D divergence based on sequence ($O < E$: 88.4% of $N = 4{,}999$ windows below the diagonal, binomial-test $P < 5 \times 10^{-324}$). The mean expected 3D divergence is on average 1.78-times higher than the observed 3D divergence (t-test $P = 1.8 \times 10^{-48}$). 3D divergence scores greater than 0.05 and nucleotide differences greater than 2250 are clipped to the baseline for visualization purpose

overlapping a strong CTCF-bound motif near a TAD boundary (within 15 kb), the AH-MH 3D divergence is 1.64-times greater ($P = 0.00077$, $N = 260/4999$ windows, Fig. 4.4B). Thus, we are observing complex sequence patterns underlying 3D genome folding that could not be determined by simply considering sequence divergence or intersecting AH variants with all CTCF sites. This is concordant with previous results which suggest that 3D genome folding is governed by a complex CTCF binding grammar[141,142,219,228].

4.2.6 Maintenance of 3D genome organization constrained sequence divergence in recent hominin evolution

Next, we evaluated if the pressure to maintain 3D genome organization constrained recent human sequence evolution. We estimated whether the amount of 3D divergence between AHs and MHs is more or less than expected given the observed sequence divergence. To compute the expected 3D divergence distribution for each 1 Mb window, we shuffled observed nucleotide differences between an African MH (HG03105) and AH (Vindija Neanderthal) 100 times and applied Akita to predict the resulting 3D genome divergence (Fig. 4.4C). We controlled for the non-uniform probability of mutation across sites using a model that preserved the tri-nucleotide context of all variants in each window with each shuffle. For each 1 Mb window, we compared the observed 3D divergence with the expected 3D divergence from the 100 shuffled sequences with the same nucleotide divergence.

If the 3D genome does not influence sequence divergence, the observed 3D divergence would be similar to the expected 3D divergence (Fig. 4.4C, bottom-middle). Alternatively, if the observed 3D divergence is greater than expected based on sequence divergence (Fig. 4.4C, bottom-left), this suggests positive selection on variation contributing to 3D differences. Finally, if the observed 3D divergence is less than expected based on sequence divergence (Fig. 4.4C, bottom-right), this suggests negative pressure on variation contributing to 3D differences.

We find that observed 3D divergence is significantly less than expected based on sequence divergence (Fig. 4.4D). 88.4% of 1 Mb windows have less 3D divergence that expected based on their observed sequence differences (binomial-test $P < 5 \times 10^{-324}$). Genome-wide, the mean expected 3D divergence is 78% higher than the observed 3D divergence (t-test $P = 1.8 \times 10^{-48}$). This suggests that, in recent hominin evolution, pressure to maintain 3D genome organization constrained sequence divergence. This aligns with previous studies that demonstrated depletion of variation at 3D genome-defining elements (e.g., TAD boundaries, CTCF sites)[123,137,154,196,229], but it specifically implicates 3D genome folding.

4.2.7 3D genome organization constrained introgression in MHs

Eurasian individuals have on average 2% AH ancestry due to introgression; however, AH ancestry is not evenly distributed throughout the genome[14,45,156]. Our previous analyses demonstrate that AH and MH exhibit a range of 3D genome organization divergence across the genome (Fig. 4.2C) and that pressure to maintain 3D genome organization constrained sequence divergence (Fig. 4.4D). Thus, we hypothesized that for a given genomic window, its tolerance to 3D genome organization variation in MHs would influence the probability that introgressed AH DNA is maintained in MH.

To test this, we first quantified the levels of 3D genome diversity for 20 modern Africans in 1 Mb sliding windows across the genome. We then computed the average African-African 3D genome divergence and term this "3D genome variability". Genomic windows with low 3D genome variability have similar 3D genome organization among all Africans, suggesting these loci are less tolerant of 3D folding changes. In contrast, regions with high 3D genome variability suggest a diversity of 3D genome organization present. Finally, we computed the amount of introgressed sequence in Eurasian populations for each window (Methods,[32]).

Genomic windows with high levels of introgression across Eurasians are enriched for windows with higher 3D genome variability (Fig. 4.5A, Mann-Whitney U $P = 0.0007$). On average, windows with evidence of introgression have 72% higher 3D genome variability than windows without introgression. Moreover, the magnitude of 3D genome variability is predictive of the average amount (proportion of bp) of introgressed sequence remaining in a 1 Mb window ($P = 5.7 \times 10^{-9}$, Fig. 4.5B, vertical axis). Even when conditioning on sequence variability, 3D genome variability provides additional information about the amount of AH ancestry in a window (Fig. 4.5B, conditional $P = 5.7 \times 10^{-4}$). In other words, even if two windows have the same level of sequence variability in MHs, windows that are more 3D variable are more likely to retain introgressed sequence. We also find that 3D genome variability is more strongly predictive of introgression shared among all three super-populations than an introgressed sequence unique to a single super-population (Supplemental Text, Tables 6.11,6.12). Using earlier introgressed Neanderthal haplotype predictions from Vernot et al.[156] and other thresholds yield similar results (Figs. 6.44,6.45). Because we compute variability in Africans with very low levels of AH ancestry, the increased 3D genome variability in MHs is not a result of introgression.

These results suggest that 3D genome organization shaped the landscape of AH introgression in modern Eurasian genomes. Previous findings demonstrated Neanderthal ancestry is depleted in regions of the genome with strong background selection, evolutionary conservation, and annotated molecular function (e.g. genes and regulatory elements)[44,45,53,55,56]. Our results expand this to implicate the 3D genome as a contributor to the landscape of AH ancestry in MHs today.

4.2.8 Introgression shaped the 3D genome organization of present-day Eurasians

Given the differences between AH and MH 3D genome organization at many loci, we hypothesized that introgressed AH sequences could have introduced novel 3D contact patterns to Eurasian MHs. To test this, we integrated Eurasians into our previous comparisons of AHs and African MHs.

For example, we found an AH-MH 3D divergent window on chromosome 7 with a striking pattern of 3D genome diversity across modern Eurasians (Fig. 4.6A). As required to be an AH-MH divergent locus, the 3D genome divergence between all Africans and AH (Vindija Neanderthal) was consistently high. And, out of 15 Eurasians, 11 had similar divergent organization compared to the Neanderthal 3D contact map. However, four Eurasians had very low 3D divergence from the Neanderthal.

When examining the contact maps of this window, all Africans have a large approximately 450 kb loop domain starting near the promoter of *IGFBP3*, a gene encoding insulin-like growth factor binding protein 3 (Fig. 4.6B). In contrast, Neanderthals (Vindija, Chagyrskaya, and Altai) have two smaller sub-domains insulated by a CTCF site. Using *in silico* mutagenesis, we identify that the variant with the largest effect

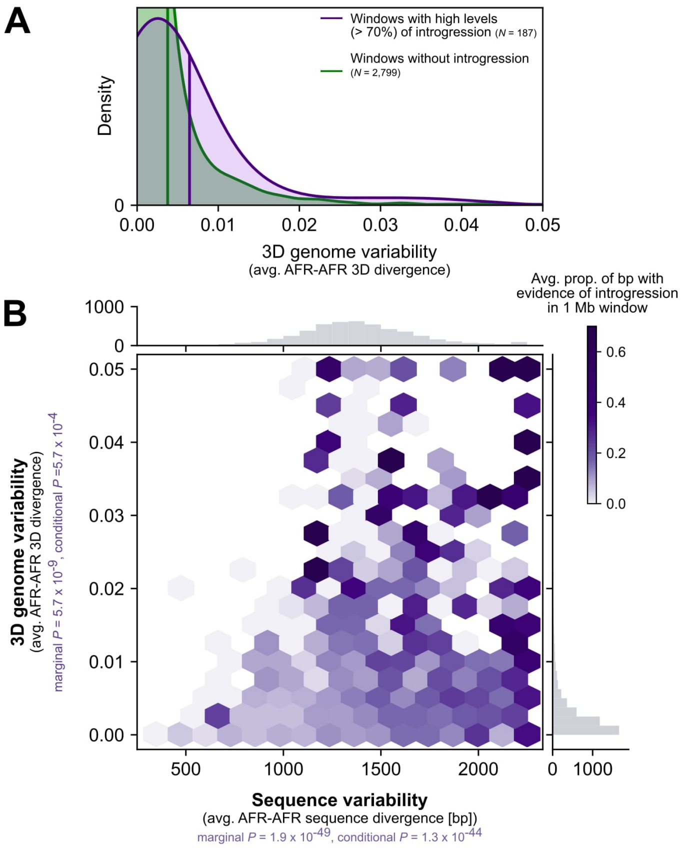


Figure 4.5: 3D variable windows in MH have more evidence of AH introgression.

(A) Windows with high levels of introgression across present-day non-African populations (purple, $N = 187$) are more 3D-variable in modern Africans (horizontal axis) than windows without evidence of introgression (green, $N = 2,799$; two-tailed Mann–Whitney U $P = 0.0007$). Vertical lines represent the distribution means. Introgression is called based on Sprime[32]. To focus on regions consistently tolerant of AH ancestry, we considered introgression shared across 1000 Genomes super-populations and covering at least 70% of bases in a 1 Mb window (Methods). Results from other introgression sets and thresholds are similar (Figs. 6.44–6.45 and Tables 6.11–6.12). **(B)** The relationship between sequence variability (horizontal axis) and 3D genome variability (vertical axis) with amount of AH ancestry in a window. Darker purple indicates a higher proportion of introgression in a 1 Mb genomic window. Sequence variability ($P = 1.9 \times 10^{-49}$) and 3D genome variability ($P = 5.7 \times 10^{-9}$) both independently predict amount of introgression. Additionally, even when controlling for sequence variability in a window, 3D genome variability is informative about the amount of introgression ($P = 5.7 \times 10^{-4}$).

on 3D organization is a G to A change at chr7:46,169,621 (rs12536129). The derived A allele, which strengthens the CTCF motif, appeared along the Neanderthal lineage. The four Eurasians (two Europeans (EUR), two South Asians (SAS)) with 3D genome organization very similar to Neanderthals all have an introgressed haplotype carrying the Neanderthal-derived A allele overlapping this CTCF site[180]. None of the other 11 Eurasians have introgression at this site (although some have introgression in the larger 1 Mb window). Across human populations, this introgressed allele remains at high-frequency today, especially in Peru (28% AMR, 2% EAS, 16% EUR, 11% SAS, 0% AFR, Fig. 6.46A).

In addition to influencing the strength of a CTCF site, this introgressed allele is also an eQTL in GTEx for the physically linked gene *IGFBP3*, Insulin-like growth factor-binding protein 3 (Fig. 6.46B, $P = 0.00014$ in artery tissue)[95]. In MHs, this variant is associated with traits including standing height ($P = 9.9 \times 10^{-7}$), fat distribution (trunk fat ratio, impedance measures, $P = 1.3 \times 10^{-5}$), and diastolic blood pressure ($P = 2.1 \times 10^{-5}$) (Fig. 6.46C).

Of the 191 3D-modifying variants identified in 167 AH-MH 3D diverged windows, 45 are observed in MHs (Table 6.10). Of note, 18 are common ($>$ 5% MAF) and 6 are at high frequency ($>$ 10%) in at least one MH 1000 Genomes Project (1KGP) super-population which motivates the hypothesis that some introgressed 3D changes were adaptive. We find very modest non-significant enrichment for these loci in previously proposed adaptive haplotypes[180] (2.3-fold enrichment, $P = 0.24$).

Given these examples of Neanderthal introgression contributing novel 3D folding to present-day Eurasians, we searched for similar patterns genome-wide. We considered 4,749 autosomal 1 Mb windows for 15 Eurasians (total $n = 71,235$) to quantify the relationship between the amount of introgression and 3D similarity to Neanderthals. We find that the amount of introgression (bp per window) is significantly correlated with 3D divergence to the Vindija Neanderthal ($P = 0.00011$, Fig. 4.6C). Results from comparisons to the other Neanderthals are consistent (Fig. 6.47). On average, in a 1 Mb window, if an individual has 80% Neanderthal ancestry, their 3D genome is 2.4 times more similar to the Neanderthal 3D genome than if they have no (0%) Neanderthal ancestry.

In summary, we find that Eurasians with more Neanderthal ancestry in a window have more Neanderthal-like 3D genome folding patterns. Furthermore, at an example locus, we demonstrate how the influence of Neanderthal introgression on 3D genome organization highlights a putative molecular mechanism for the effect of Neanderthal ancestry on human traits.

4.3 Discussion

The role of 3D genome organization in human biology is increasingly recognized[105,123,137,154,196,229]; however, current techniques for measuring 3D folding cannot be applied to the study of ancient DNA. Furthermore, despite methodological improvements in assays of the 3D genome, high-resolution experiments across many diverse individuals, species, and cell types remain prohibitive. To address these gaps, we provide a framework for inferring 3D genome organization at population-scale that facilitates evaluation of previously untestable hypotheses.

First, we apply this framework to resurrect archaic 3D genome organization. We find that 3D genome organization constrained sequence divergence and patterns of introgression in hominin evolution. We cat-

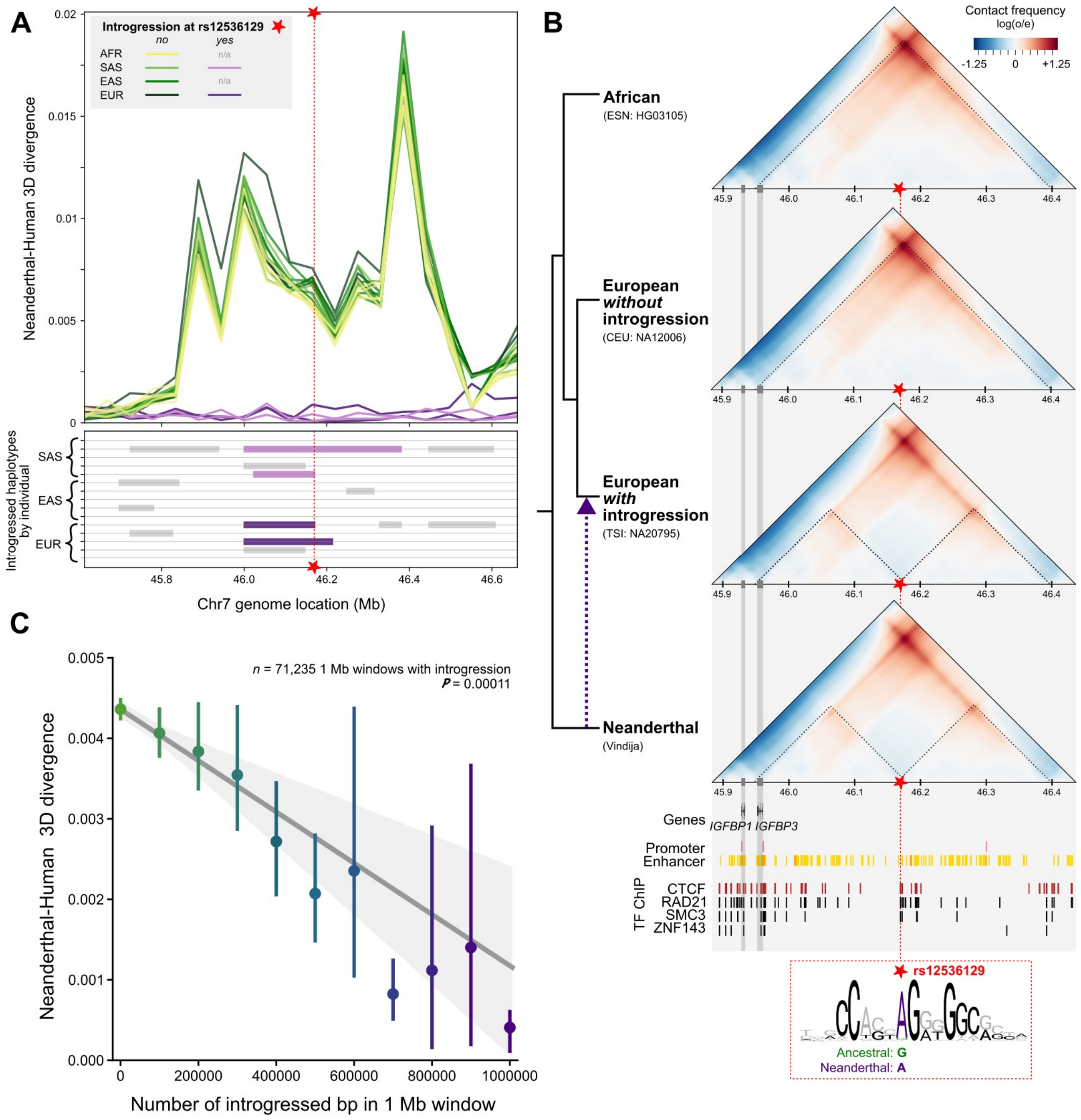


Figure 4.6: Introgression introduced novel 3D genome organization patterns to modern Eurasians. **(A)** Comparison of the 3D contact maps between Neanderthal (Vindija) and 20 MHs for a window on chromosome 7 reveals that most MHs (yellow, green) have different 3D organization compared to Neanderthals. In contrast, four MHs with introgression (purple boxes) overlapping chr7:46,169,621 (red star) have similar 3D organization compared to Neanderthals across this part of the genome (purple). (AFR: African, SAS: Southeast Asian, EAS: East Asian, EUR: European) This example 3D-divergent locus **(B)** was introgressed into MH and remains at high frequency (28% AMR, 2% EAS, 16% EUR, 11% SAS, 0% AFR, Fig. 6.46). At this locus (zoomed to chr7:45,883,392-46,436,352), Neanderthals and individuals with introgression have two domains insulated by a CTCF site (red box). In MHs without introgression, this motif is disrupted with a G instead of an A (star, chr7:46,169,621, rs12536129) leading to a larger fused domain and differential contacts with the promoter of *IGFBP3*. **(C)** The amount of introgression in a 1 Mb window (number of bp, horizontal axis) is significantly correlated with the similarity of an individual's 3D genome organization to a Neanderthal's (Vindija) genome organization (vertical axis) ($P = 0.00011$, $n = 71,235$ 1 Mb windows across 15 Eurasians). The error bars signify 95% bootstrapped CIs and the error band signifies the 95% bootstrapped CI for the linear regression estimate.

alog genomic regions where AH and MH 3D genome organization diverged and illustrate how this novel mechanism links sequence differences to phenotypic differences. Importantly, our approach permitted the evaluation of variants unobserved in MHs, and it provides a putative molecular mechanism for AH-MH phenotypic differences including those that may have been selected against after hybridization (e.g. cognitive and brain morphology traits)[44,45,53–56,63]. Finally, we identify regions in which introgression introduced AH 3D genome folding that are novel to MHs in Eurasians with Neanderthal ancestry. Together, these results illustrate the power of imputing unobservable molecular phenotypes to resolve evolutionary questions about functional divergence.

Second, we anticipate that our framework for comparing and interpreting hundreds of genome-wide 3D genome contact maps will be helpful for testing hypotheses beyond archaic DNA. In the interpretation of genetic variants of unknown significance, it will be key to consider the effect of inter-individual and inter-species variation on 3D genome architecture, especially given recent evidence that even common DNA sequence variants can influence 3D organization and human phenotypic variation[136]. Our work establishes the groundwork to answer many diverse questions. For example, we illustrate how *in silico* mutagenesis can highlight the role of a variant in 3D genome organization and how to integrate this with other functional annotations. This allows us to examine the 3D effects of variants never before observed in MHs, which is essential to non-coding variant interpretation from the lens of both evolution and disease. Our new measure of "3D genome variability" provides genome-wide quantification of how different regions tolerate variation in 3D genome folding. We also demonstrate a simulation approach for testing how 3D genome folding constrains sequence evolution across the genome. Finally, we develop a method to robustly identify 3D divergent windows between populations. With the recent growth of 3D genome *in silico* predictors[140–143], we anticipate that our work can provide a foundation for both hypothesis generation and prioritization of experimental resources.

Although our approach provides many novel benefits, it also has limitations that we hope future work will address. First, our comparisons likely underestimate 3D diversity. We only investigate windows of the genome with complete sequence coverage. Because of ancient sample degradation, we do not have full coverage of AH genomes. We use a conservative approach to effectively mask regions of the genome lacking coverage in AHs (Fig. 6.34 and Methods). Furthermore, we only consider the effects single nucleotide variants. We do not consider structural variation (SV) due to the challenges of calling SV accurately in ancient samples. We anticipate new methods in ancient DNA sequencing will allow us to model the 3D genome organization of AHs more completely. Second, our 3D genome organization comparisons are based on a correlation-based metric. We demonstrate concordance with comparisons using other more biologically informed methods (Fig. 6.37); however, more sophisticated methods to quantify the type and resolution of change (e.g. neo-TAD vs TAD-fusion event, scale of TAD vs. loop) would benefit the 3D genome community[140]. Third, although Akita is trained simultaneously across five cell types, 3D genome organization is largely conserved across cell types and predictors only identify limited cell-type-specific differences. Therefore, we focused on the highest resolution predictions in a single context (HFF). As more high-resolution Hi-C and Micro-C becomes available across diverse cell types, our framework can be applied to identify cell-type-specific AH-MH differences.

Several practical caveats must be considered when interpreting some of our results. For example, to conduct *in silico* mutagenesis we manipulate every single nucleotide separately against the same background rather than considering the prohibitively large number of possible combinatorial variant sets. Additionally, while our null model of genome divergence accounts for context-dependent mutation probabilities, we suggest that future study of the influence of 3D folding on genome evolution would benefit from the use of forward-time genomic simulations. The annotations that link 3D-modifying variants to genes and functions are also based on studies in MHs (HPO and GWAS). It is possible, though unlikely, that a gene disrupted in MHs would not lead to the same traits in AHs. Finally, given the scope of our study and the nature of archaic DNA, direct experimental validation is not possible with current technology. To date, Gorkin et al.[136] provides the largest set of Hi-C across 19 MH individuals in the same cell type (LCL GM12878). However, the resolution is too low to call chromatin loops (40 kb vs. 2 kb in our analyses), and 13 of the 19 individuals are African and have almost no Neanderthal ancestry. Thus, we use complementary experimental data, like CTCF ChIP-seq and experimentally-derived TAD maps, to provide independent support for the influence of variants on 3D genome organization and to link variants with genes in true physical proximity. Moreover, even if high-resolution Hi-C were available across many Eurasians, an experimental approach would still not capture all AH variation, highlighting the necessity of our computational approach.

In conclusion, our framework for inferring archaic 3D genome organization provides a window into previously unobservable molecular mechanisms which shaped the sequence and phenotypic evolution of hominins.

4.4 Methods

Modern human and archaic genomes

Obtaining genomes

All genomic analysis was conducted using the GRCh37 (hg19) genome assembly and coordinates (www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/). Genomic variation within modern humans (MH) came from 1000 Genomes Project (1KGP), Phase 3 from Auton et al.[175]. All MH genomes were selected randomly from each subpopulation with a filter for females only to facilitate comparisons of the X chromosome. The 1KGP individuals used are listed in Table 6.9. Archaic genomes are from Prüfer et al.[13] (Altai), Prüfer et al.[14] (Vindija), Mafessoni et al.[15] (Chagyrskaya), and Meyer et al.[16] (Denisova).

Building individual genomes

We constructed full-length genomes for each MH or AH based upon the genotyping information in their respective vcf file. Given the difficulty of distinguishing heterozygous genotypes in the ancient DNA samples, we treated all individuals as if they were homozygous (pseudo-haploid). We built each individual genome using GATK's FastaAlternateReferenceMaker tool[230]. If an individual had an alternate allele (homozygous or heterozygous), we inserted it into the reference genome to create a pseudo-haploid, or "flattened" genome for each individual. This procedure is illustrated in step 1 of Fig. 6.34.

Accounting for missing data in the archaic genomes

Ancient DNA is both fragmented and degraded. These characteristics present challenges to both sequencing and alignment, resulting in gaps in coverage, particularly in genomic regions of low complexity. To account for this missing data, we "masked" all genomic regions lacking archaic genotyping information by reverting nucleotide states to the hg19 reference. For analyses that compared 3D genome organization between MHs and AHs, we masked both MH and AH genomes. This procedure is illustrated in steps 2-4 of Fig. 6.34. Archaic genome coverage is shown in Fig. 6.35. For analyses that only considered MHs (e.g. quantifying 3D genome variability across the genome in MHs), this masking procedure was not applied.

3D genome organization predictions with Akita

After the genomes were prepared, we input them into Akita for predictions using a 1 Mb sliding window (1,048,576 bp) overlapping by half (e.g. 524,288-1,572,864, 1,048,576-2,097,152, 1,572,864-2,621,440). Although Akita is trained simultaneously on Hi-C and Micro-C across five cell types in a multi-task framework to achieve greater accuracy, we focus on predictions in the highest resolution maps, human foreskin fibroblast (HFF). We note that the results are similar when considering other cell types (e.g. embryonic stem cells), likely because of limited cell-type-specific differences (Fig. 6.38). Akita considers the full window to generate predictions, but the resulting predictions are generated for only the middle 917,504 bp. Each contact map is a prediction for a single individual, and each cell represents physical 3D contacts at approximately 2 kb (2,048 bp) resolution. The value in each cell is $\log_2(obs/exp)$-scaled to account for the distance-dependent nature of chromatin contacts. Darker red pixels indicate more physical contacts and darker blue pixels denote fewer physical contacts. For all analyses, we only considered windows with full (100%) coverage in the hg19 reference genome for a total of 4749 autosomal and 250 chromosome X windows. Fudenberg et al.[141] provides further details on the CNN architecture and training data used.

3D genome comparisons

After predictions were made on all 1 Mb windows for all individuals, we compared the resulting predictions using a variety of measures. All measures are scaled to indicate divergence: higher indicates more difference while lower indicates more similarity. In the maintext we transform the Spearman's rank correlation coefficient $(1-\rho)$ to describe 3D divergence. We consider measures based on the Pearson correlation coefficient $(1-r)$ and mean squared difference $(\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2)$ in Fig. 6.37. Percentiles of 3D divergence shown in Fig. 4.2A-B are calculated with reference to a universe of 4 AHs × 5 African MHs × 4999 genomic windows for a total of 99,980 comparisons. Figs. 4.4A,6.42 averages the 3D divergence $(1-\rho)$ across all 4999 1 Mb windows (lower triangle) to compare to the average number of bp differences (after the masking procedure described above) in the same pair of individuals (upper triangle). Clustering is done with the "complete" (Farthest Point) method.

Sequence comparisons

Some analyses compare 3D genome divergence with sequence divergence. To calculate the sequence divergence between two individuals, we counted the proportion of bases at which the two individuals differ in the 1 Mb window. For comparisons of divergence when including AHs, we applied the same masking procedure as used to facilitate 3D genome comparisons (i.e. windows with missingness in AHs are filled with hg19 reference).

CTCF motif overlap

We consider how nucleotide differences in a window (between Neanderthal [Vindija] and an African MH [HG03105]) impacts 3D genome divergence in Figs. 4.4B,6.43. We stratified variants by if they overlap a bound CTCF motif and their distance to TAD boundaries. CTCF motifs are from Vierstra et al.[231]. CTCF-bound open chromatin candidate cis-regulatory elements (cCREs) in the HFF cell type are from Abascal et al.[232]. TAD boundaries in the HFF cell type are from processed MicroC data from Akgol Oksuz et al.[233]. These annotations were all lifted over to hg19[234]. A window was considered to have a CTCF-overlapping variant if an AH-MH nucleotide difference intersected a CTCF-bound HFF cCRE and a CTCF motif. Results were further stratified by varying levels of motif strength ("match_score" in the top 10th,25th, 50th, or any percentile), distance to TAD boundary (within 15 kb, 30 kb, or anywhere), and whether the CTCF motif overlap occurs in the middle 50% of the 1 Mb window or not.

Empirical distribution of expected 3D genome divergence

To compute the expected 3D divergence in a window given the observed sequence divergence, we generate genomes with shuffled nucleotide differences. We match these shuffled differences to the same number and tri-nucleotide context of the observed sequence differences between the Neanderthal (Vindija) and an African MH (HG03105) genome (Fig. 4.4C). Variants are not shuffled into masked regions of the genome. For each 1 Mb window of the genome ($N = 4999$) we generate 100 shuffled sequences. We calculate an empirical distribution of expected 3D divergence from comparing the contact maps of the shuffled sequences with the MH sequence. Finally, we compare the average expected 3D divergence from this distribution to the observed AH-MH 3D divergence.

AH-MH 3D divergent loci

Identifying loci

To identify loci with AH-MH 3D genome organization divergence, we compared the 3D contact map at each 1 Mb loci between each AH and 20 African MHs. To call a region as divergent, we required all 20 AH-MH comparisons to be more 3D divergent than all MH-MH comparisons (Fig. 4.3A). This identifies regions with consistent 3D differences between AHs and MHs while excluding regions with a large 3D diversity in modern humans. We also required the minimum AH-MH 3D divergence to be in the 80th percentile or greater of most 3D diverged (Fig, 4.2A, 3D divergence > 0.0042). Because 20 MHs do not capture the full MH genome diversity, it is possible that these criteria would still capture 3D patterns segregating in modern Africans that are not truly AH-MH diverged. Thus, we removed any windows where the 3D-modifying variant determined by *in silico* mutagenesis (below) was observed in 1KGP MHs if it was not introgressed (LD of $r^2 = 1$ with introgressed variants called by Browning et al.[32] or Vernot et al.[156]). For the counts of

AH-MH divergent windows (Fig. 4.3B), we considered overlapping 1 Mb windows as a single observation. We summarize and report the AH-MH 3D divergent windows in Tables 6.10.

In silico mutagenesis
To identify the variant(s) contributing to the most prominent 3D differences in each identified AH-MH divergent window, *3D-modifying variants*, we use *in silico* mutagenesis. For example, for an Altai Neanderthal divergent window, we identify every bp difference that is unique to the Altai genome when compared to 20 African MH genomes. In the background of the MH (HG03105) genome, we insert each different Altai allele one-at-a-time. We then compare the resulting contact map between the original MH genome and the MH genome with each Altai allele. We then identify both the allele resulting in the largest 3D divergence and any other variants that contribute to a 3D divergence $>=$ 0.0042 and term these “3D-modifying variants” (Table 6.10).

Phenotype ontology enrichment
To test if AH-MH 3D-modifying variants are enriched near genes related to particular phenotypes we follow a procedure visually described in Fig. 6.40. 3D-modifying variants (above) are linked to genes in their TAD because this provides evidence of physical proximity. TADs are defined as regions between TAD boundaries as defined with MicroC data in HFF from Akgol Oksuz et al.[233] (lifted over to hg19). Genes are defined as the longest transcript from protein-coding genes (NM prefix) from NCBI RefSeq downloaded from the UCSC Table Browser[208]. Genes are linked to phenotypes from the Human Phenotype Ontology (HPO) and GWAS Catalog 2019 downloaded from Enrichr[223–225]. Annotations are further grouped into phenotypic systems via system-level annotations from Gene ORGANizer[235] and manual curation. HPO largely considers rare disease annotations and has 1779 terms with 3,096 genes annotated[221]. The GWAS Catalog largely considers common disease annotations and has 737 terms with 19,378 genes annotated[222]. Through this procedure, we counted the number of ontology terms linked to the set of 3D-modifying variants. We considered 3 different sets, those shared (intersect) by all Neanderthals (Fig. 4.3), those in any Neanderthal (union), and those in Denisova (Fig. 6.41, Table 6.10).

We test enrichment for ontology terms linked to at least one 3D-modifying variant. While the annotations are downloaded from Enrichr, we did enrichment analyses with a more appropriate null. For each set, we shuffle the observed 3D-modifying variants into the background genome. We defined the background genome as any place where a 3D-modifying variant could have been identified (i.e. regions with full coverage in modern humans used for Akita predictions). We then use the same procedure (Fig. 6.40 to link the shuffled variants to genes and then ontology terms. We repeat this shuffle 500,000 times to create an empirical distribution for how many times we would observe each annotation under the null. We used these distributions to calculate an enrichment and P-value for each ontology term. The FDR-corrected significance level was determined empirically using these null observations (a subset of $n = 50,000$). We select the highest p-value threshold that led to a $V/R < Q$ where V is the mean number of expected false discoveries and R is the observed discoveries (which includes both true and false positives).

Relationship between 3D genome organization and introgression

3D genome variability

To consider how 3D organization may have constrained where we observe introgression in the genome, we calculated 3D genome variability across the genome in MHs. Because we are not comparing these predictions with AH 3D genome organization, we did not mask the genomes before 3D genome predictions (above). In the same 1 Mb sliding windows across the genome, we predicted the contact maps for 20 modern Africans (because they have no or very little introgression). For each window, we calculate the 3D genome divergence between all 190 $\binom{20}{2}$ pairs of contact maps. We then computed the "3D genome variability" by taking the mean of these 190 divergences for each 1 Mb window across the genome. High 3D genome variability indicates a high average pairwise 3D divergence (i.e. diversity of 3D organization), while low 3D genome variability indicates low pairwise 3D divergence (i.e. similar 3D organization across all individuals).

Genomic windows with evidence of introgression

To define genomic regions with Neanderthal ancestry we used "segments" identified by Browning et al.[32] using Sprime, a heuristic scoring strategy that compares high-LD regions in a target admixed population (i.e., Europeans) with an unadmixed outgroup (i.e., Africans) to identify putatively introgressed regions. We considered a set of Sprime-identified segments shared (intersection) among East Asians (EAS), EUR, and SAS. We repeat the analysis using a more stringent subset of Sprime segments that (1) have at least 30 putatively introgressed variants that could be compared to the Altai Neanderthal genome and (2) had a match rate of at least 30% to the Altai Neanderthal allele (Neanderthal filter). We also considered the introgressed Neanderthal haplotypes previously identified by Vernot et al.[156] identified using the S* statistic. Finally, we consider introgressed segments unique to a single population (EAS, EUR, or SAS). Because these introgression calls only consider autosomes, we do not use the X chromosome for these analyses. Results from these sets of Neanderthal ancestry are in Figs. 4.5,6.44,6.45 and Tables 6.11,6.12.

In the main text (Fig. 4.5), we compare the 3D genome variability between 1 Mb windows with no introgression (0%) versus windows where at least 70% of the bp have evidence of introgression. Other thresholds are shown in Fig. 6.44.

Predicting the amount of introgression

To test if 3D genome variability can be uniquely informative to predict tolerance of introgression, we conducted a simple linear regression. We predict the amount of introgression in a 1 Mb window while conditioning on the amount of sequence variability in a window. $Y = B_0 + B_1 X_{\text{3D variability}} + B_2 X_{\text{Sequence Variability}}$, where Y is the proportion of the 1 Mb window with evidence of introgression defined using the previously described sets of Neanderthal ancestry. For comparison, we also conducted some regressions where Y was modeled from only 3D variability or sequence variability alone. Results from these models are in Figs. 4.5B,6.45, Tables 6.11,6.12.

Individual-level introgression calls

We used introgression calls in 1KGP individuals from Chen et al.[180], which applied IBDmix with the Altai Neanderthal genome to identify introgressed segments in MHs. We identified windows with AH-MH divergence with evidence of introgression by intersecting with the introgression calls.

We also test the relationship between the amount of introgression an individual has and their 3D divergence from AHs. For each window, we compare the amount of introgression (% of bp) for an individual in a 1 Mb window with that individual's 3D divergence from Neanderthals. We do this analysis for 15 Eurasians across 4,749 1 Mb autosomal windows (total $n = 71,235$). In Fig. 4.6C we compare Eurasians to the Vindija Neanderthal 3D genome and in Fig. 6.47 we compare to Altai and Chagyrskaya. We also repeat the analysis removing windows with no evidence (0% bp) of introgression.

eQTL and PheWAS analysis

eQTL analysis and plots were generated using the Genotype-Tissue Expression (GTEx) Project (V8 release) Portal (lifted over to hg19)[95]. PheWAS results are from the GWAS Atlas and consider 4756 traits[159]. Allele frequencies come from 1KGP Phase 3[175].

Examples

The examples visualized in Figs. 4.3,4.6 are annotated using the UCSC genome browser[234]. They were each manually zoomed to highlight the regions of interest. We use ENCODE open chromatin candidate cis-regulatory elements (cCREs)[232] to highlight promoters (promoter-like signature, pink) and enhancers (proximal [orange] and distal [yellow] enhancer-like signature) combined from all cell types downloaded from the UCSC table browser (lifted over to hg19)[208]. We use Transcription Factor (TF) ChIP-seq Clusters (130 cell types) from ENCODE 3[236,237] downloaded from UCSC table browser[208]. We show the motif sequence logo with reference to the positive strand of hg19.

Data analysis and figure generation

The datasets we generated are available in the GitHub repository "neanderthal-3d-genome" available here https://github.com/emcarthur/neanderthal-3D-genome/ which will be formally cited and versioned upon publication.

All genomic coordinates and analysis refer to Homo sapiens (human) genome assembly GRCh37 (hg19), unless otherwise specified. All *P* values are two-tailed, unless otherwise specified. All measures of central tendencies are means, unless otherwise specified. Data and statistical analyses were conducted using Python 3.6.10 (Anaconda distribution), Jupyter Notebook, BedTools v2.26, and PLINK 1.9[177,181]. Figure generation was significantly aided by Matplotlib, Seaborn, and Inkscape[182–184].

Data availability

The publicly available data used for analysis are available in the following repositories. MH genome vcfs are from 1000 Genomes Project (1KGP) (ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_ genomes_project/release/20190312_biallelic_SNV_and_INDEL/[175]. Archaic genotypes are from the following repositories: Altai Neanderthal[13] (ftp.eva.mpg.de/neandertal/Vindija/VCF/Altai/), Denisova (ftp.eva.mpg.de/neandertal/Vindija/VCF/Denisova/)[16], Vindija Neanderthal[14] (ftp.eva.mpg.de/neandertal/Vindija/VCF/Vindija33.19/), and Chagyrskaya Neanderthal[15] (ftp.eva.mpg.de/neandertal/Chagyrskaya/VCF/). Introgressed variants and segments are from Sprime Version 1 (https://data.mendeley.com/datasets/y7hyt83vxr)[32]. An alternative set of introgressed variants and segments are from S*[156]]. Individual level 1KGP introgression calls are from the Akey Lab[180].

CTCF motifs are from genome-wide motif scans v1.0[231], CTCF-bound open chromatin candidate cis-regulatory elements (cCREs) in the HFF cell type (https://screen.encodeproject.org/ > Downloads > by cell type > HFF-Myc male newborn originated from foreskin fibroblast, lifted-over to hg19)[232], TAD boundaries in the HFF cell type are from processed MicroC data available at the 4D nucleome data portal (https://data.4dnucleome.org/experiment-set-replicates/4DNES9X112GZ/, lifted-over to hg19)[233]. RefSeq genes, TF ChIP-seq Clusters, enhancer and promoter cCREs are downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables)[208]. Gene ontology annotations are downloaded from Enrichr (https://maayanlab.cloud/Enrichr/#libraries)[223–225]. System-level groupings of disease ontology terms were aided by Gene ORGANizer annotations(http://geneorganizer.huji.ac.il/downloads/)[235]. eQTL data is from the GTEx Portal (https://www.gtexportal.org/, lifted-over to hg19)[95]. PheWAS results are from the GWAS Atlas (https://atlas.ctglab.nl/)[159].

Code availability

Akita is in the “basenji” GitHub repository available here https://github.com/calico/basenji/tree/master/manuscripts/akita[141]. The “neanderthal-3d-genome” GitHub repository (above) contains a Jupyter notebook with custom code used for data analysis and all figure generation.

CHAPTER 5

CONCLUSIONS

5.1 Summary

The goal of this dissertation was to synthesize interdisciplinary methods and data to quantify the relationships between the recent human evolution, 3D genome organization, gene regulation, and complex human disease. The aims were to:

- Comprehensively quantify the contribution of Neanderthal ancestry to diverse human traits. (Chapter 2)

- Map the relationship between 3D genome architecture and the genetic architecture of complex traits (Chapter 3)

- Reconstruct the 3D genome organization of Neanderthals to evaluate how chromatin folding shaped human evolution (Chapter 4)

Chapter 2 addressed a gap in our understanding of how Neanderthal ancestry influences trait variation in modern humans. These results expand our understanding of the consequences of introgression in several ways. First, we use partitioned heritability to consider the genome-wide effects of introgression, overcoming limitations of previous work that only studied individual loci. Additionally, previous work found evidence for depletion of functional elements in regions of the genome with Neanderthal ancestry. Our work goes beyond these proxies for function to demonstrate depletion for contribution to diverse complex traits. Despite this depletion, we found that introgressed variants shared across multiple Neanderthal populations are enriched for heritability contribution to several traits with potential relevance to human adaptation to non-African environments, including hair and skin traits, autoimmunity, chronotype, bone density, lung capacity, and menopause age. Integrating our findings, we proposed a model in which selection against introgressed functional variation was the dominant trend (especially for cognitive traits); however, for a few traits, introgressed variants provided beneficial variation via uni-directional (e.g., lightening skin color) or bi-directional (e.g., modulating immune response) effects.

Chapter 3 transitioned to investigate the 3D genome architecture across diverse cell types to demonstrate its relevance to evolutionary conservation and trait-associated variation. Although the 3D genome has an established role in rare disease, this work demonstrated that genetic variation in TAD boundaries contributes more to complex-trait heritability, especially for immunologic, hematologic, and metabolic traits. We also unify seemingly contradictory findings about evolutionary pressures across the 3D genome landscape: we suggest that even though TADs are shuffled, the boundary-defining sequences are under more constraint than the sequences within the TAD. We also demonstrate that TAD boundaries shared across cell types are further enriched for complex-trait heritability, evolutionary constraint, CTCF binding, and housekeeping genes. Ultimately, we highlight how considering the 3D genome across cell types provides valuable context for understanding the genome's functional landscape and enabling variant interpretation that takes 3D structure into account.

Chapter 4 synthesized knowledge about 3D genome folding with outstanding questions about the mechanism of archaic variation contribution to phenotype. Using novel deep learning tools we reconstruct the 3D genome organization of Neanderthals and Denisovans. Using the resulting 3D contact maps, we identify 167 regions with 3D genome divergence between AHs and MHs and find enrichment for phenotypes related to the eye, supra-orbital ridges, hair, lungs, immune response, and cognition. We demonstrate that the 3D genome organization constrained sequence divergence and patterns of introgression in hominin evolution. Finally, we highlight loci where modern Eurasians inherited novel 3D genome folding from AH ancestors. Together, our findings illustrate the power of inferring molecular phenotypes to reveal previously unobservable biological differences.

5.2 Future directions

Given the complexity of the human genome, many of the models and ideas presented in this dissertation simplify certain aspects of both evolution and genome folding that are presently unknown. Yet, these models build the foundation for future hypotheses and discovery. Detailed limitations specific to each aim are included at the end of each chapter. Here, I will synthesize limitations with common threads across all the presented work to highlight big-picture avenues for future exploration.

5.2.1 Replication across diverse populations

Despite advances in genome sequencing, there are still significant gaps in equitable representation across populations in human genomics research. For example, non-European participants represent only 22% of individuals in GWAS[222,238], despite the vast majority of the world's population living in Africa and Asia[239]. This not only hinders our knowledge about biology, but it also prohibits equitable benefits of genomics to all. However, there are reasons to be hopeful as the number of the non-European individuals respresented in GWAS has increased five-fold from 4% in 2009 thanks to community engagement, increased diversity of scientists, better funding, and work by many consortia[239,240]. Increasing diversity and sample sizes will permit further population-specific investigations into the architecture of complex traits.

This will be especially important in the context of understanding the contribution of archaic hominin variation to traits across the globe. Many of the associations we and others have found are related to interaction with the environment (e.g., immune response, skin color). Because humans in different populations have faced unique environments and challenges, I hypothesize that the genome-wide consequences of introgression are different across populations. Furthermore, studies across diverse humans will facilitate a better understanding of Denisovan ancestry because it is population-specific. Accordingly, recent work has built upon our findings to begin addressing these gaps. Dannemann[227] used summary statistics from GWAS cohorts of 212,000 individuals in Biobank Japan to identify population-specific associations with Neanderthal DNA. Koller et al.[241] conducted phenome-wide association studies (PheWAS) of Denisovan and Neanderthal alleles in participants of six different ancestries in the UK Biobank to highlight the specific contribution of Denisovan introgression to East Asian populations across diverse phenotypes. Both studies were limited by sample size and power differences that impeded the ability to compare results with European-ancestry cohorts. New statistical methods can also address this concern. For example, S-LDXR[242] aims to highlight population-specific associations. Ultimately, as sample-sizes increase and consistent phenotype data in biobanks become more available, I envision that the application of our framework will provide important insights into the diverse selective pressures humans faced and how they lead to disease today.

A more diverse catalog of modern human genetic variation would also benefit our mechanistic understanding of the non-coding genome. Although all human populations share underlying biology, most variants are population-specific[239]. Understanding the consequences of a given variant provide insight into the function of that part of the genome; thus, sampling diverse individuals will make conclusions more robust and our models more accurate. Accordingly, our results mapping the relationship between 3D genome architecture and the architecture of complex traits should be replicated across diverse populations. Improvements in statistical methods to calculate partitioned heritability will also enable more robust conclusions. For example, cov-LDSC can be applied in admixed populations[243], and I anticipate methods will be developed to estimate partitioned heritability by meta-analyzing across populations, even when the sample size for sub-populations are smaller[244].

Finally, models to predict 3D genome organization can be applied to diverse populations to identify variation in 3D folding. While larger-scale 3D genome organization is conserved across species[105,111,119–121], I hypothesize that finer-scale loops are more population specific[136]. Creating a genome-wide catalog of 3D genome organization across diverse populations may highlight differences between populations that lead to trait variation.

5.2.2 Consideration of rare variation

In addition to most genetic variation being population-specific, most genetic variation is also rare and observed in less than 5% of the population ($\text{MAF} < 0.05$)[239,245]. The work in this dissertation that leverages GWAS data with heritability analyses only considers the effect of common variation on traits ($\text{MAF} \geq 0.05$). Future work should consider the effects of rare and ultra-rare variation. New methods, larger LD reference panels, and greater sample size now permit preliminary investigations into the genome-wide effect of rare variation ($0.5\% \leq \text{MAF} < 5\%$) to complex traits[169]. Progress in sampling rare variation should be united with the efforts to increase diversity: an increased sample size should not come at the convenience of excluding diverse populations.

Extension of these results to consider rare variation should provide insight towards interpreting the effect of both Neanderthal variation and 3D genome context. As more humans have been sequenced, an increasing number of archaic ancestry segments have been identified. Early (2014) estimates suggested that approximately 20%[44] of archaic genomes exist in modern humans. This was revised to 40%[31] (2020), and more recent methods to detect introgression have proposed that closer to 93%[30] (2021) of the human genome has evidence of admixture or incomplete lineage sorting. Identifying this archaic ancestry is exciting, but most of these archaic variants segregate at very low frequencies in present-day populations, which makes phenotypic associations more difficult. I hypothesize that considering rare variants will highlight more deleterious trait-associated variation under negative selection, even though the heritability they explain is small[169].

In the context of the 3D genome, our work was motivated by studies of ultra-rare structural variation

(SV) disrupting 3D genome structure. Many of the SVs that disrupt 3D organization (e.g., via TAD boundary rearrangement) cause neurological and developmental traits[105,122,132]. Thus, I originally hypothesized that common variants in TAD boundaries would associate with related traits. Instead, we found that TAD boundaries enriched for the heritability of metabolic, hematologic, and immunologic traits. However, this only considers common variation. In light of these results, I now hypothesize that negative selection purged common variation in TAD boundaries with consequences for the brain and development traits[154]. To resolve these differences, we must quantify the 3D genome contribution to traits across the allele-frequency spectrum[169,245]. This may reveal which parts of the genome are more intolerant to 3D genome variation which is critical for understanding the genetic architecture of rare and common disease.

5.2.3 Applications to other mechanisms of gene regulation

This work specifically considers the role of the 3D genome in evolution and disease; however, this is only one piece of the gene regulatory puzzle. Our work, along with others that considered methylation[96,97] and gene regulation prediction models[94], have demonstrated the potential to impute molecular phenotypes to mechanistically link genotype to phenotype to ask diverse questions. Together, these studies provide a foundation that can be used to consider other mechanisms of gene regulation. Owing to more data and improved deep learning architectures, novel methods that predict molecular phenotypes from sequence have emerged. For example, SpliceAI models mRNA splicing and predicts cryptic splice mutations[246]. Some methods also directly predict gene expression. For example, Basenji2[247], ExPecto[248], and Enformer[249] predict genomic tracks including open chromatin (e.g., DNase), histone modifications (e.g., H3K27ac), and CAGE gene expression. Applying these to both archaic and modern human genomes will paint a picture of gene regulatory landscapes to facilitate comparisons.

This dissertation demonstrates the utility of imputing molecular phenotypes of extinct species for which we would never be able to assay 3D genome structure or RNA levels. However, these methods can also be applied across present-day humans to answer outstanding questions. For example, although we have the technical capability to measure RNA expression across multiple individuals in multiple cell types at population-scale, it would be impractical. Consequently, imputing gene regulation has demonstrated fruitful in highlighting genes and pathways implicated in disease via transcriptome-wide association studies (TWAS)[250,251]. I hypothesize that integrating 3D genome organization predictions (or other molecular predictions) with phenotype data may provide a complementary benefit. This hypothesis is preliminarily supported by Gorkin et al.[136] who identified common sequence variants that influence experimentally-determined 3D folding and demonstrated that these differences correlate with epigenomic and transcriptomic annotations. However, these Hi-C data were only for 19 individuals, in one cell type, and were low resolution. I propose a 3DWAS (3D genome-wide association study) as an intriguing idea to directly quantify the contribution of differential 3D genome folding across cell types on traits or gene expression at population-scale.

5.2.4 Evaluation of differences across cell types and species

Finally, as alluded to in the Introduction (Chapter 1), our understanding of the "how" and "why" of chromatin folding is still in its infancy: genome-wide assays of chromatin configuration are just a decade old[104]. Early studies largely characterized 3D genome organization, especially TADs, as conserved across cell types[111,112,116–118] and species[105,111,119–121]. Consequently, models that infer folding from sequence can only predict limited cell type differences[141]. However, with increasingly high-resolution maps, single-cell Hi-C, and samples from diverse tissues and species, we are now appreciating the diversity in 3D genome organization and its implications for cell-type differences[154,198] and evolution[199,252].

These differences across species and cell types make the "chromatin folding problem" even more complex; however, it also poses exciting questions about how the genome encodes such a diversity of 3D organization patterns. The aim of current 3D genome prediction models is simply to "learn" how to best predict the experimental map. However, the architecture and goal of these deep learning models could be adjusted to incentivize predictions of cell-type or species-specific regions. In addition to better models, increased availability of data across cell types will allow for better training that may enable novel applications. For example, high-quality cell-type-specific predictions will be necessary to evaluate the consequences of variation in rare disease cohorts, especially as many rare diseases have tissue-specific consequences (e.g., neurologic, cardiac). Increased data availability will also facilitate comparisons that could identify species-specific or cell-type-specific patterns of 3D genome folding. In Chapter 3, we were under-powered to test for a relationship between cell-type-specific TAD boundaries and heritability contribution to specific traits. More robust detection of cell-type-specific chromatin features may provide the necessary resolution to consider the relationship between cell types and traits in the context of the 3D genome.

In addition to comparisons between experimental maps, I anticipate that comparisons between experimental and predicted maps may provide opportunities to test hypotheses about the "code" underlying genome folding across contexts. For example, when models trained in humans were applied to mice, they predicted mouse 3D genome structure poorly in certain regions[141]. Notably, these poorly predicted regions contain B2 SINE elements, which harbor CTCF motifs and have expanded in Muridae lineages[141]. Thus, identifying regions where real maps and predictions differ could provide clues to the evolution and cell-type specificity of 3D genome folding mechanisms.

We are in the midst of a genomics revolution. The influx of genome sequences and complementary functional data will enable exciting investigations into some of the future directions outlined here, among others we cannot even conceptualize yet. However, as genome sequencing becomes more inexpensive and routine, this acceleration will pose new challenges. Human genomes differ at up to 5 million sites[78], and the majority of this variation is in the enigmatic non-coding genome. To interpret this vast amount of variation, this dissertation highlights that we must consider the role of the 3D genome given its importance to both our risk for disease and our evolutionary history.

CHAPTER 6

APPENDICES

6.1 Appendix 1: Supporting information for Chapter 2

Supplemental Text

Crohn's disease risk in Vindija-matching variants

The heritability enrichments across traits for Vindija-matching variants are highly correlated with those for the Altai-matching variants ($r^2 = 0.93$, Fig. 6.3B-C). However, heritability enrichment for Crohn's disease is higher in Vindija-matching variants than in other introgressed sets (2.1-fold vs. 1.1-fold enriched, Fig. 6.3C). We note that, given the large overlap between the variant sets that match Altai and Vindija (Jaccard similarity = 77%), the increased enrichment is not significant genome-wide ($P = 0.4$). Nonetheless, to explore the specific loci underlying this difference, we identified variants that contribute more to the heritability enrichment in Vindija-matching set compared to Altai-matching variants. We found that these introgressed Crohn's disease associated variants have diverse evolutionary histories. For example, as expected, we identify several variants that appeared on the Neanderthal lineage after the split of the ancestors of the Vindija and Altai Neanderthals among the introgressed alleles most associated with Crohn's disease (Fig. 6.7B). In addition, we observe ancestral variants from before the divergence of AMH and Neanderthals that were lost in AMHs and the Altai Neanderthal, but that remained in Vindija. These ancestral Crohn's disease risk variants were reintroduced to AMH by introgression (Fig. 6.7A)[58].

Contribution of selection to observed heritability enrichment

We hypothesized that selection contributed to the heritability enrichment observed among introgressed variants for certain traits. Many tests for selection are confounded by introgression, but high frequencies in modern populations suggest selection for introgressed alleles[33,35]. On a variant-level, introgressed variants with high frequency in modern Europeans ($> 21\%$ MAF) contribute more to the heritability enrichment than rarer variation (Fig. 6.8), suggesting that after introgression these trait-associated variants increased in frequency in European populations potentially due to selection. Irrespective of their origin, common variants contribute more to complex trait heritability than rarer variants. However, this MAF-dependent architecture pattern is consistent with the action of selection on variants affecting complex traits[155,253]. To further investigate the type of selection, we find that, on a haplotype level, genomic windows that contribute most to the heritability of sunburn and white blood cell count overlap more putatively adaptive introgressed haplotypes than expected by chance[35,180] (Fig. 6.9, Sunburn: $P = 0.02$ [$q = 0.09$]; WBC: $P = 0.02$ [$q = 0.09$]). Together, these findings support our hypothesis that selection acted differently on Neanderthal variation associated with different traits.

Relationship between introgression and morningness at *NMUR2*

We identify multiple windows near *NMUR2* with a positive relationship between Neanderthal LD profile and morning person status. We focus on two windows (at chr5:151745423-151931514) nearest to *NMUR2* with the strongest positive relationship to morningness (overall $r = +0.91$, Fig. 6.16A-C). This positive correlation suggests that increased LD to Neanderthal alleles is associated with an increased propensity to be a morning person. The variants most associated with morningness at this locus are all introgressed (Fig. 6.16B-C), but they have different histories: some are Neanderthal-derived (e.g. rs4958561: $P = 9 \times 10^{-12}$, Fig. 6.16D and some were lost ancestral alleles reintroduced through introgression (e.g. rs10045463: $P = 3 \times 10^{-12}$). In addition, 168 of 169 introgressed variants in high LD in this window negatively associate with expression of *NMUR2* in frontal cortex cells from GTEx[95] (Fig. 6.16E, rs4958561:

$P = 1 \times 10^{-9}$ [Bonferroni critical value $P = 1 \times 10^{-3}$]). In a PheWAS across the UK Biobank, traits most associated with this introgressed haplotype (tagged by rs4958561) include ease of getting up in the morning ($P = 1 \times 10^{-14}$), chronotype ($P = 2 \times 10^{-12}$), morningness ($P = 4 \times 10^{-12}$), sedentary behavior ($P = 4 \times 10^{-9}$), and tea intake ($P = 1 \times 10^{-8}$) (all pass Bonferroni correction $P = 1 \times 10^{-5}$)[159]. *NMUR2* encodes the Neuromedin-U receptor 2, a receptor for neuromedins U (NMU) and S (NMS) that phase shifts circadian rhythm activity[254–256]. In vivo, *NMU* shows circadian expression in rat brains in response to melatonin and a genetic overexpression screen in zebrafish larvae identified *Nmu* to promote hyperactivity through Nmu receptor 2[257,258]. Integrating these data, we hypothesize that Neanderthal introgressed alleles downregulate *NMUR2* in the brain leading to an association with increased morning person propensity.

There are four clusters of morningness-associated variants at this broader locus (within 1 Mb of rs4958561), further suggesting the putative biological importance of this region to chronotype. The first cluster, pictured in Fig. 6.16 and discussed above, is tagged by rs4958561 and rs10045463, which are both introgressed variants. The second cluster's lead SNP is rs17489682 ($P = 7.4 \times 10^{-13}$) which is also introgressed. The third cluster's lead SNP is rs2910032 ($P = 1.9 \times 10^{-13}$) and this cluster does not contain introgressed variants; however, rs4958561 (introgressed tag SNP) and rs2910032 are not in high LD ($r^2 = 0.046$). The fourth cluster's lead SNP is rs4533947 and is not introgressed ($P = 4 \times 10^{-12}$). rs4958561 (introgressed SNP) and rs4533947 are in moderate LD ($r^2 = 0.39$). Therefore, we believe cluster 1's association with morningness (shown in the figure) is driven by the Neanderthal introgressed variants. While we cannot fully exclude that some of the signal observed at rs4958561 is not shared by the cluster of variants 330 kb downstream (rs4533947), we find this to be unlikely due to the degree of LD ($r^2 = 0.39$) and the similarly strong association at both loci ($P = 9 \times 10^{-12}$ and $P = 4 \times 10^{-12}$).

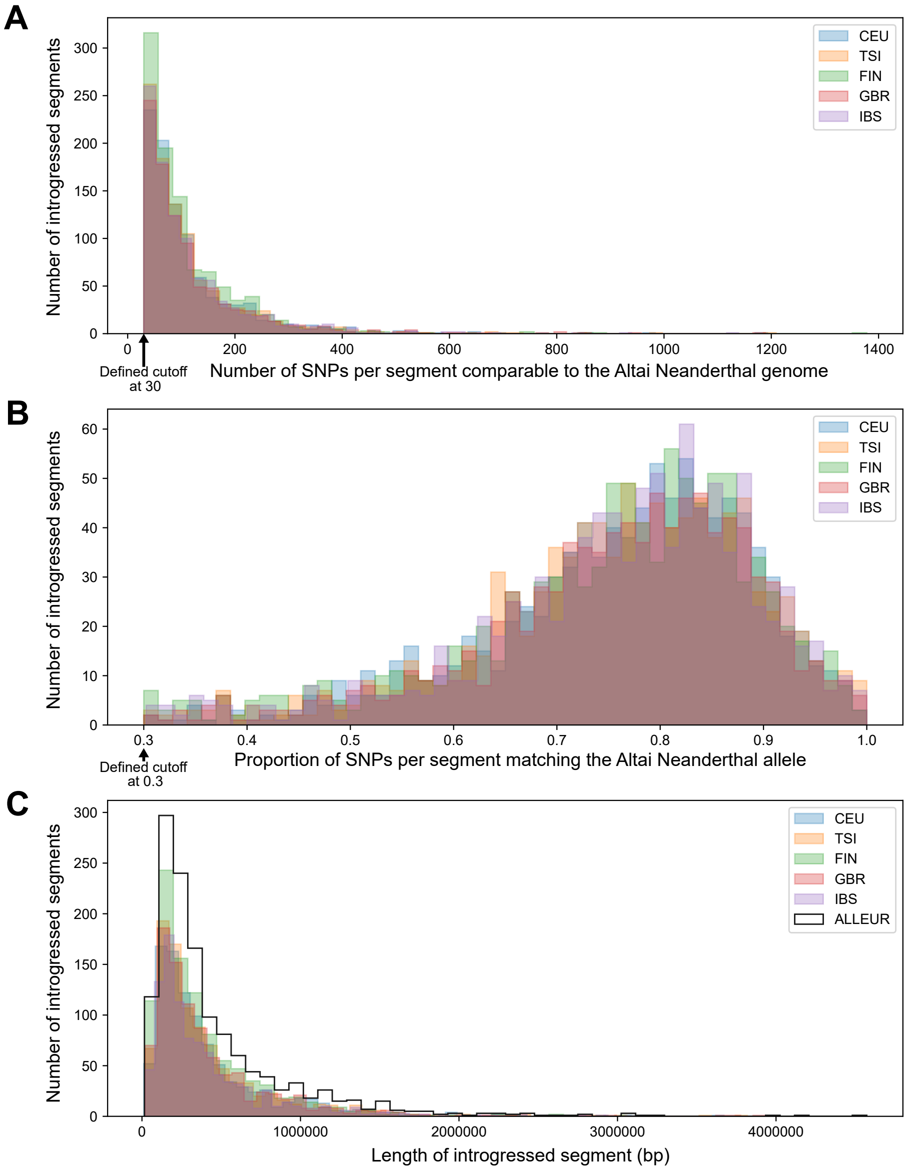


Figure 6.1: Defining genomic regions with Neanderthal ancestry.

Of the introgressed segments defined by Browning et al. 2018 we consider those observed in any of the European subpopulations (CEU, TSI, FIN, GBR, IBS), (A) that have at least 30 putative introgressed variants that are comparable to the Altai Neanderthal genome (after filtering, these segments have an average of 116 comparable variants) and (B) that these putative introgressed variants have at least 30% match to the Altai Neanderthal allele. After filtering, these segments have a 76% match on average. (C) The size distribution of these independently identified segments after applying these two filters. We also consider the union of these sets (black). Ultimately, we define 1345 segments that have a median length of 299 kb (IQR: 174 – 574 kb). This set is used in Fig. 2.1B. (CEU: Utah Residents with Northern and Western European Ancestry, TSI: Toscani in Italia, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian Population in Spain).

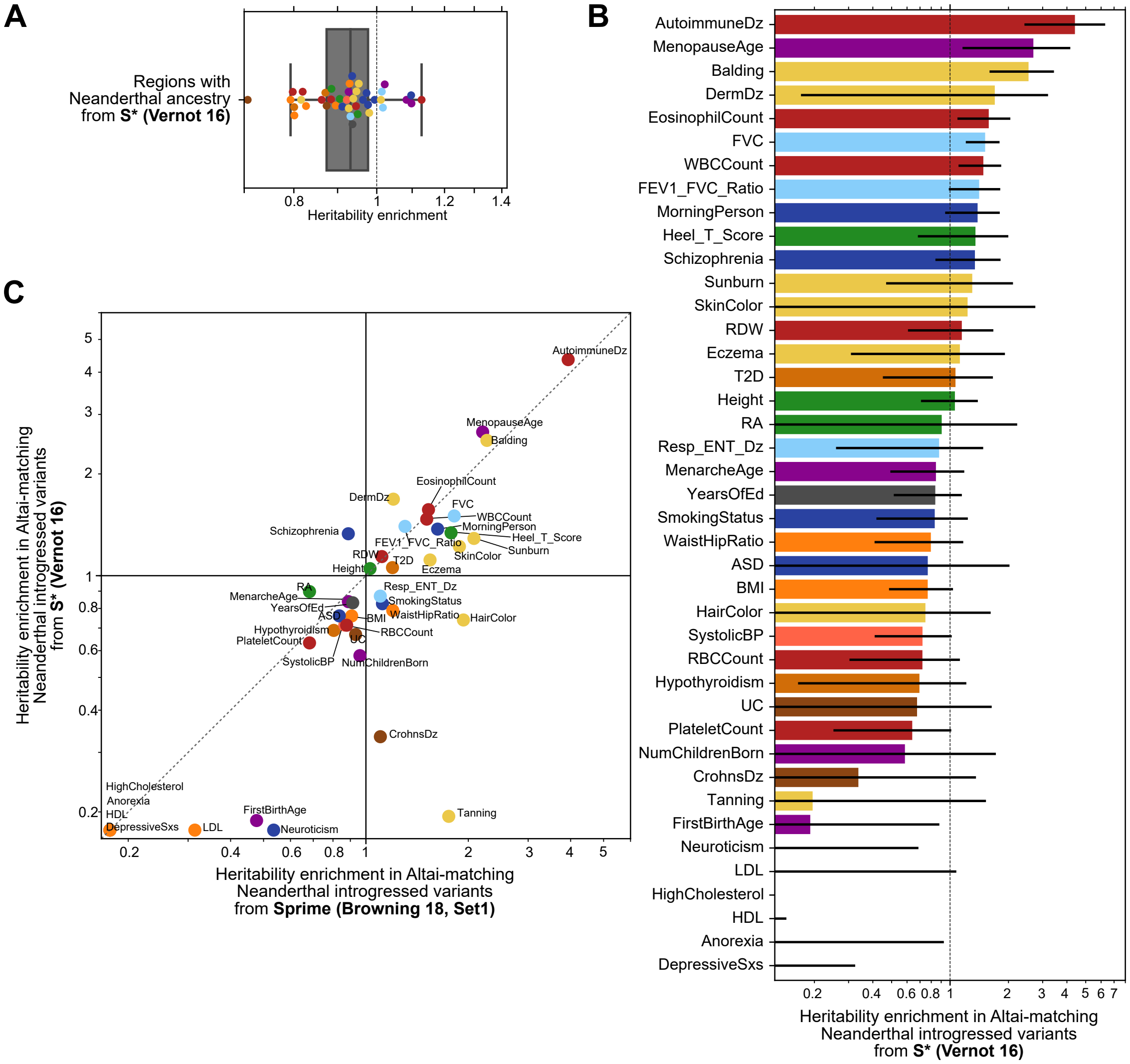


Figure 6.2: Trait heritability patterns in regions with Neanderthal ancestry and introgressed variants are consistent when defined based on variants identified by S* from Vernot et al. 2016.
(A) Similar to Fig. 2.1B, we show that traits ($n = 41$) are broadly depleted of heritability in regions with Neanderthal ancestry defined using haplotypes from S* (0.93x background expectation, two-tailed one-sample t-test $P = 1 \times 10^{-5}$). The boxplot centers represent medians, the boxes are bounded by the first and third quartile, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. Traits (Crohn's disease) with depletion less than 0.7 are plotted on the baseline. (B) For a set of Altai-matching Neanderthal introgressed variants identified by S*, we show the trait-by-trait partitioned heritability analysis. Bars for individual traits represent heritability enrichment estimates and error bars are standard errors calculated by LDSC using a block jackknife ($n = 200$). Trait heritability depletion less than 0.125 are truncated. This set includes 132,296 variants and is comparable to the Altai-matching "set 1" variants identified by Sprime ($N = 138,774$) which is shown in Fig. 6.4A. (C) Trait heritability compared between Sprime-identified Altai-matching "set 1" introgressed variants (x-axis) and S*-identified high-confidence variants (y-axis), are highly correlated ($r^2 = 0.79$).

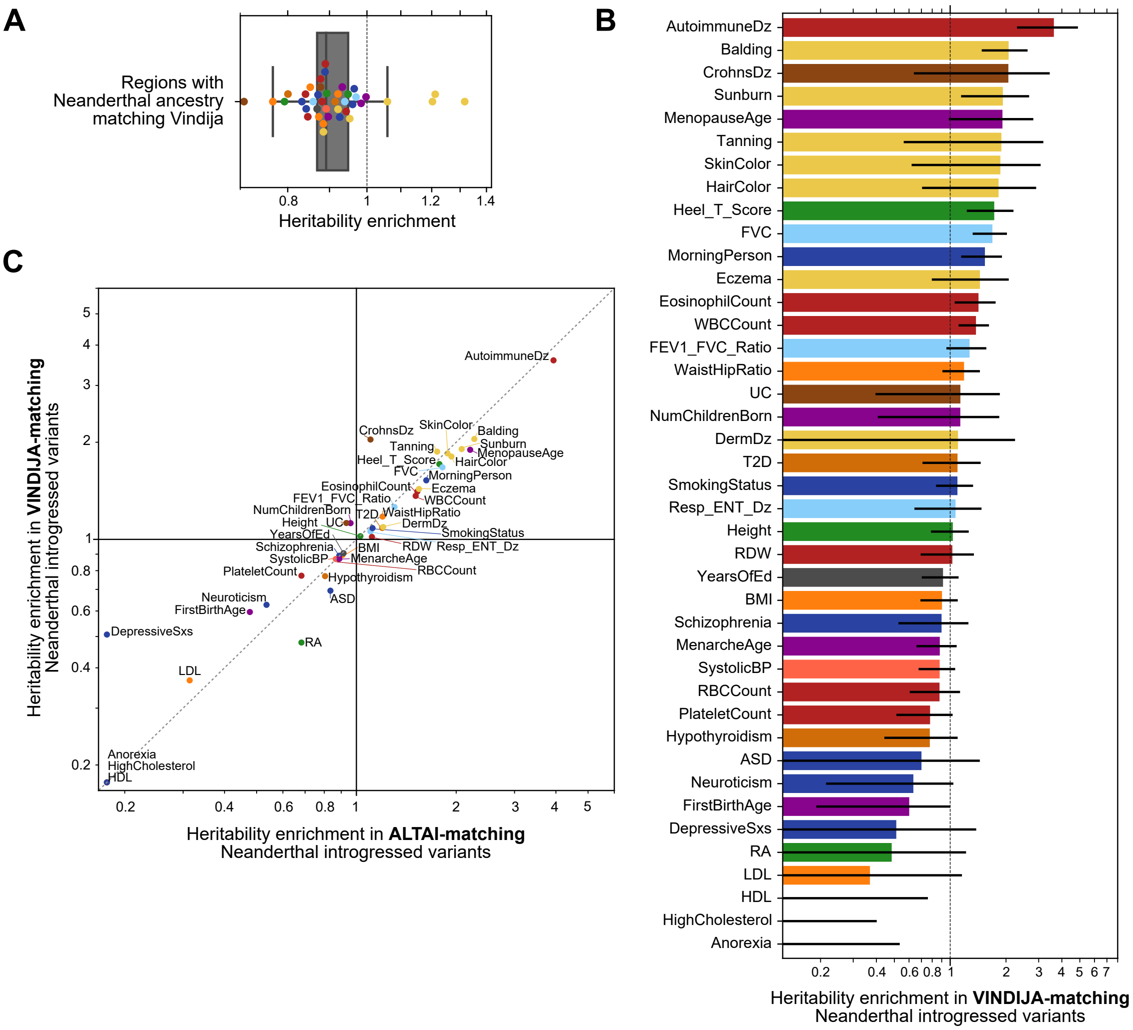


Figure 6.3: Trait heritability patterns in regions with Neanderthal ancestry and introgressed variants are consistent when defined based on match to the Vindija Neanderthal genome.

(A) Similar to Fig. 2.1B, we show that traits ($n = 41$) are broadly depleted of heritability in regions with Neanderthal ancestry defined using Vindija-matching haplotypes from Sprime (0.92x expectation, two-tailed one-sample t-test $P = 1 \times 10^{-5}$). The boxplot centers represent medians, the boxes are bounded by the first and third quartile, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. Traits (Crohn's disease) with depletion less than 0.7 are plotted on the baseline. (B) For a set of Vindija-matching Neanderthal introgressed variants, we show the trait-by-trait partitioned heritability analysis. Bars for individual traits represent heritability enrichment estimates and error bars are standard errors calculated by LDSC using a block jackknife ($n = 200$). Trait heritability depletion less than 0.125 are truncated. This set includes 167,927 variants and is comparable to the Altai-matching "set 1" variants ($N = 138,774$) which is shown in Fig. 6.2A. (C) Trait heritability compared between Altai-matching "set 1" introgressed variants (x-axis) and Vindija-matching variants (y-axis), are highly correlated ($r^2 = 0.93$). 66% of traits are more enriched for heritability in Altai-matching variants compared to Vindija-matching variants (bottom right triangle, one-tailed Binomial test $P = 0.03$). Heritability of Crohn's disease is the one trait that is notably enriched in Vindija-variants (2.1-fold) compared to Altai-variants (1.1-fold).

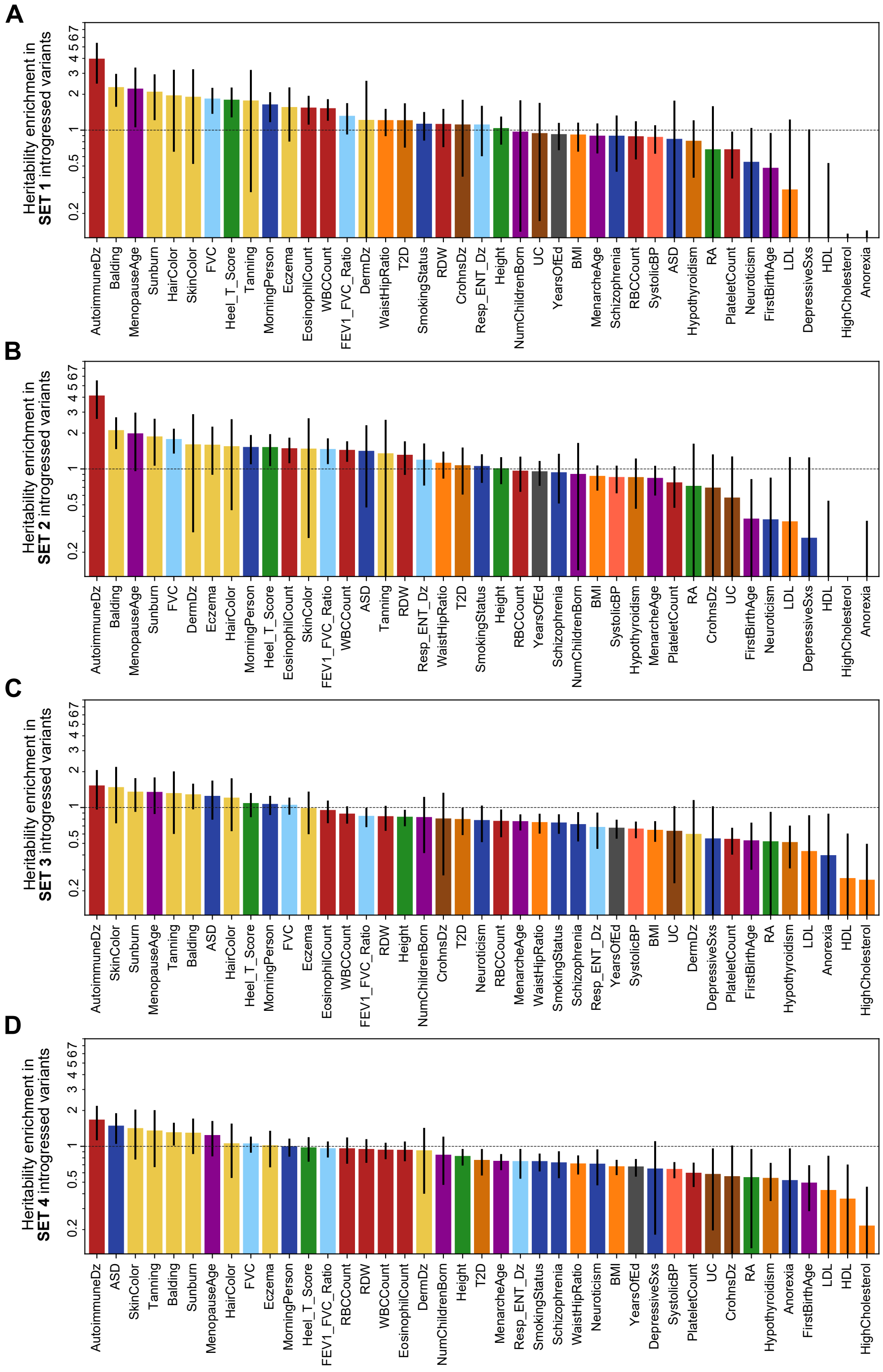


Figure 6.4: Patterns of complex trait heritability are similar across four different sets of Neanderthal introgressed variants.

From the most stringent set of Altai-matching variants observed in Europeans (set 1, **A**) to the most inclusive set of introgressed variants observed in any subpopulation (set 4, **D**), we show the heritability enrichment (or depletion) ordered by magnitude. Bars for individual traits represent heritability enrichment estimates and error bars are standard errors calculated by LDSC using a block jackknife ($n = 200$). Traits with depletion less than 0.125 are truncated. Set 4 (D) is the same as Fig. 2.1C. The relationship between Set 4 (D) and Set 1 (A) is shown in Fig. 2.1D. Details of each set are in the methods.

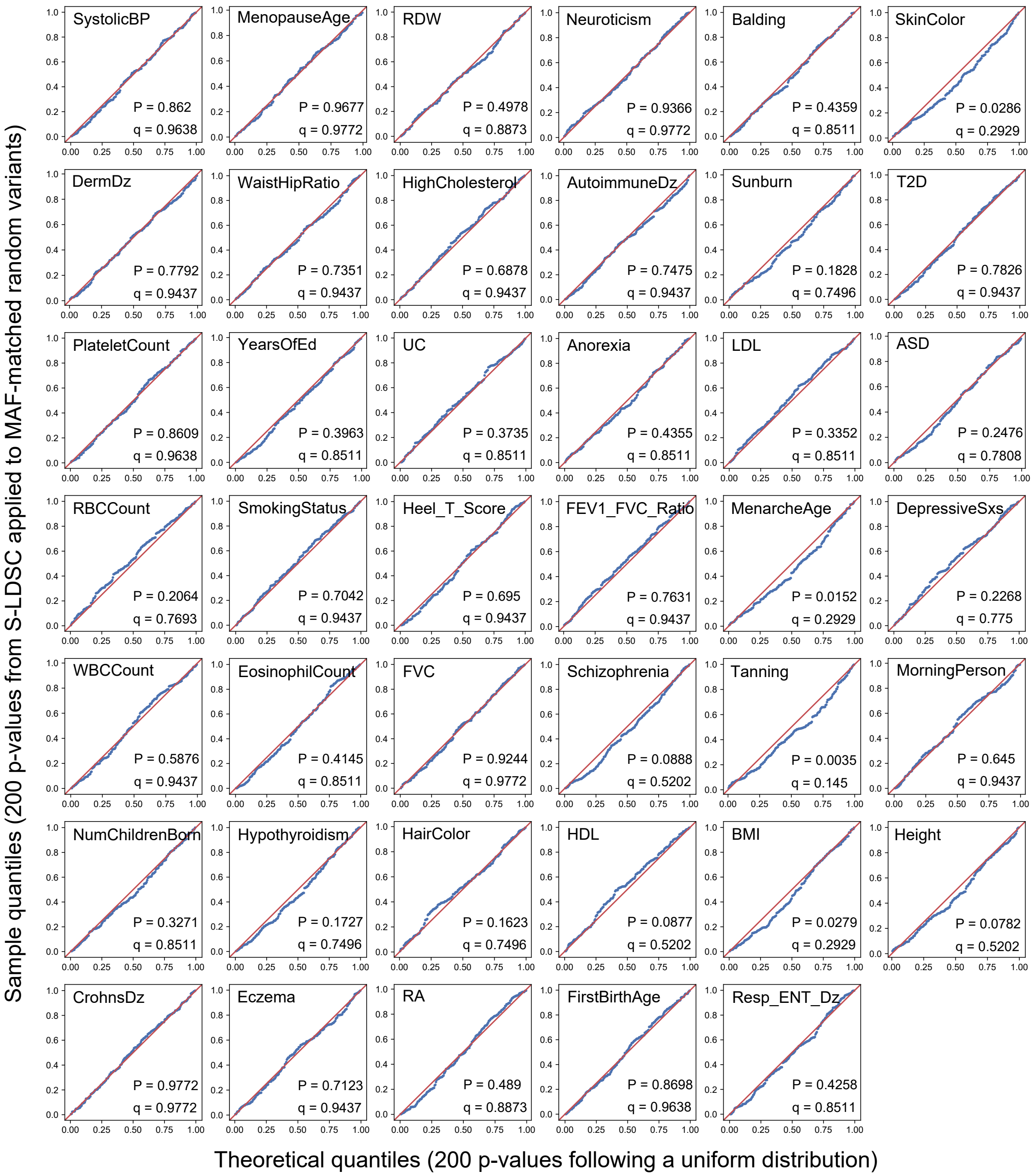


Figure 6.5: Partitioned heritability enrichment *P*-values are not biased by the allele frequency distribution. non-introgressed common variants. We tested whether the minor allele frequency (MAF) distribution for introgressed variants could alone be responsible for the number of traits with significant enrichment or depletion observed. We generated 200 sets of random variants matching the MAF distribution of the Altai-matching variant set in 10 bins (5-7%, 7-10%, 10-13%, 13-17%, 17-21%, 21-26%, 26-32%, 32-38%, 38-44%, 44-50%). We chose the Altai-matching variant set because its distribution is the most skewed compared to the set of all 1000G variants (the most variants in the 5-7% bin and fewest in the 44-50% bin). For the 200 MAF-matched random sets, we calculated *P*-values for h2 enrichment (or depletion) for each trait with S-LDSC (which uses a block jackknife approach). For each trait, we plot a *P*-value qq-plot and calculate the two-tailed Kolmogorov–Smirnov test to assess if the *P*-values from the MAF-matched random variant h^2 enrichment tests follow the uniform distribution. We find no *P*-value inflation for any of the 41 traits (K-S FDR-controlled $q = 0.145 - 0.977$). Thus, the test is well calibrated and the results observed for the introgressed variants are not due to their allele frequency distribution alone.

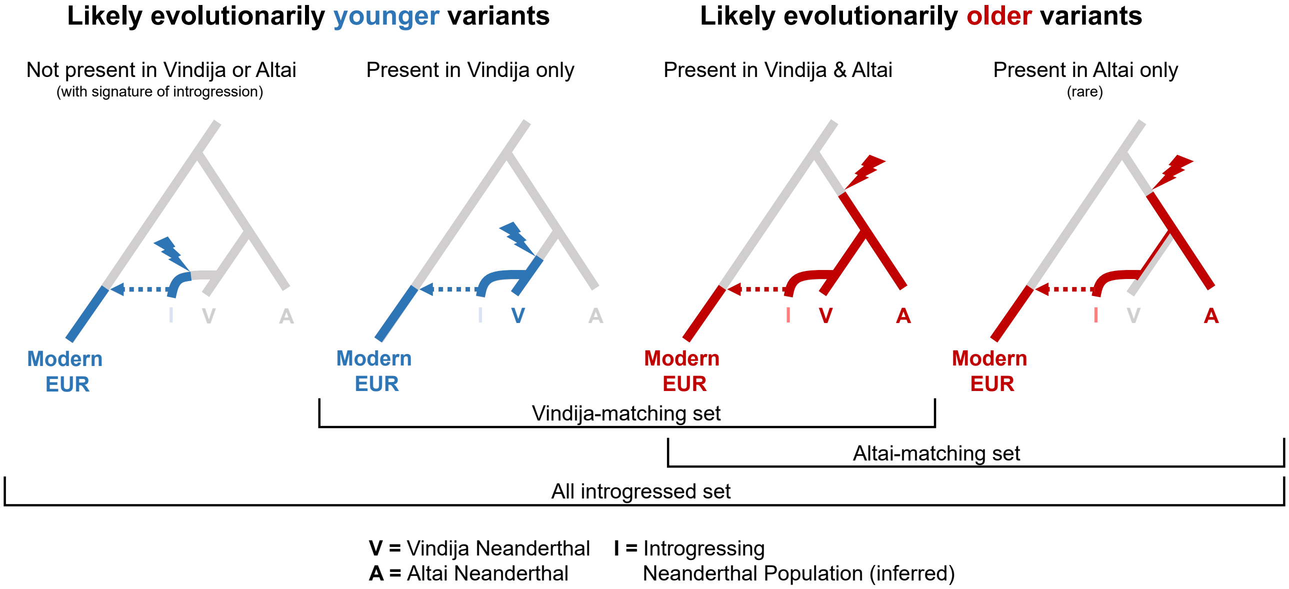

Likely evolutionarily younger variants
Likely evolutionarily older variants
Not present in Vindija or Altai
(with signature of introgression)
Present in Vindija only
Present in Vindija & Altai
Present in Altai only
(rare)
I
V
A
Modern
EUR
Vindija-matching set
Altai-matching set
All introgressed set
V = Vindija Neanderthal
A = Altai Neanderthal
I = Introgressing
Neanderthal Population (inferred)

Figure 6.6: Schematic of likely evolutionary trajectories and ages of introgressed alleles in the different sets considered.
We represent possible evolutionary histories for different sets of putatively introgressed variants considered in our study (not to scale). For each set of possible observed genotypes (observed indicated by colored label) in modern Europeans (EUR), Vindija Neanderthal (V) and Altai Neanderthal (A), the most likely evolutionary trajectories under parsimony assumptions are depicted. The introgressing Neanderthal (I) genotype is inferred. The branch on which the introgressed allele appeared in each scenario is depicted by a bolt. The two major scenarios that we interpret as "younger" are introgressed variants that (1) appeared after the split of the introgressing population from Vindija and related populations and (2) appeared after the split of Altai and Vindija. In the second set, the modern European genotype will only match the Vindija Neanderthal but not the Altai Neanderthal. Alternatively, "older" variants arose prior to the split between Vindija and Altai populations; throughout the paper these are referred to as the "Altai-matching set" and we demonstrate that these "likely older" variants are enriched for trait heritability. Altai-matching variants that are not observed in Vindija Neanderthal are rare ($N = 5,685$ out of 900,902 total introgressed variants), but are also classified as "likely older" because two independent mutation events (on the Altai lineage and the introgressing population lineage) is less likely than scenarios where the genotype is unobserved in Vindija for other reasons (e.g., not sequenced, allele was not fixed). We note that these evolutionary histories are a simplification and not all sites matching each pattern followed the trajectories shown.

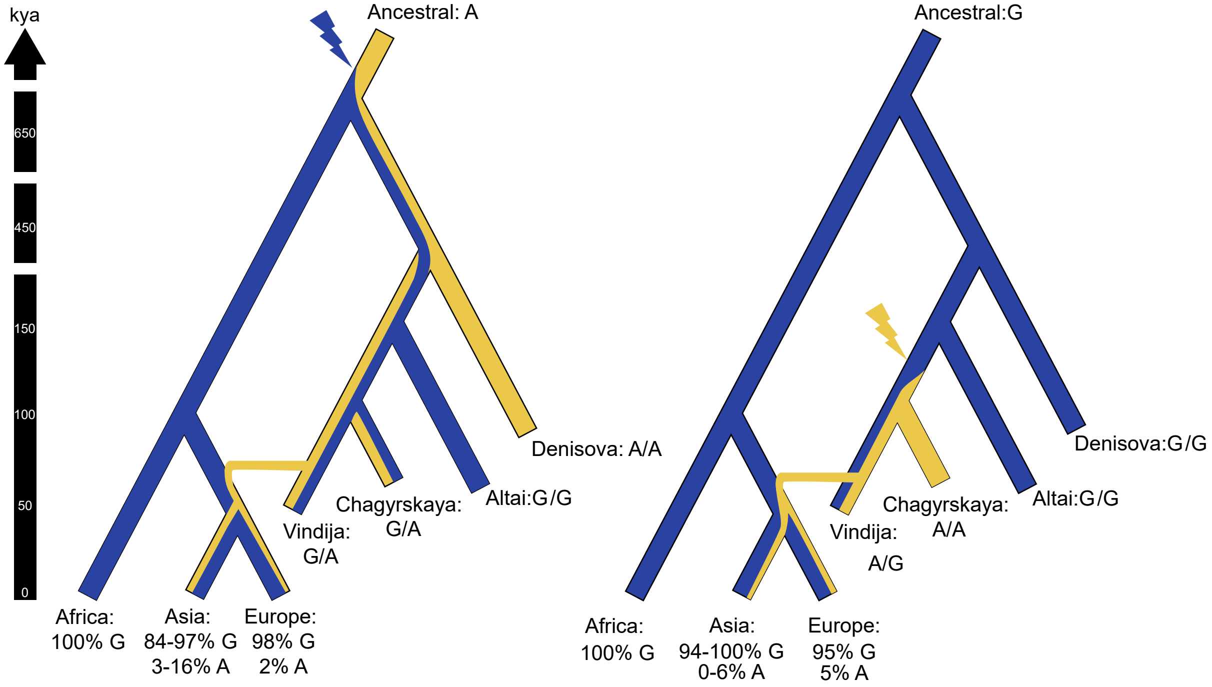


Figure 6.7: Variants that contribute to Crohn's Disease risk observed in Vindija that are absent in Altai have diverse evolutionary origins.
(A) rs17467144 is a missense variant in MUC19 tagging a region associated with Crohn's disease on chr12 ($P = 1 \times 10^{-17}$). The A allele is the ancestral allele that was lost in AMHs but was maintained in Vindija and Chagyrskaya Neanderthals and reintroduced to Eurasian populations. (B) rs17768654 is a missense variant in TTC6 tagging a region nominally associated with Crohn's disease on chr14 ($P = 1.7 \times 10^{-3}$). The G allele is the ancestral allele. The A allele is Neanderthal-derived, likely after the split of younger Neanderthal populations (Vindija/Chagyrskaya) from Altai. The A allele was then introgressed into Eurasians and is associated with Crohn's disease risk. These associations contribute to the Crohn's disease heritability enrichment seen in Vindija-matching variants when compared to Altai-matching variants (Fig. 6.3C).

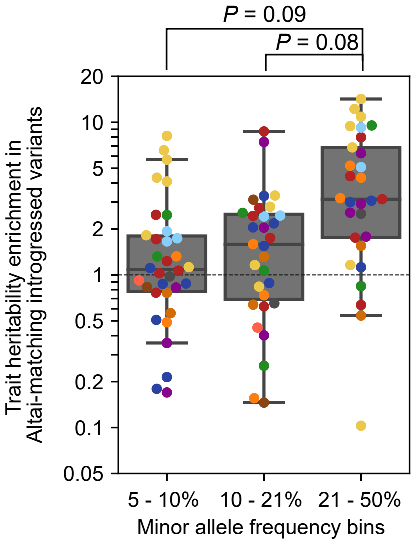


Figure 6.8: Introgressed variants at higher allele frequency in modern European populations contribute more to trait heritability than rarer variants.

Partitioned heritability was calculated on the Altai-matching introgressed variants (Fig. 2.1C, S4A) stratified by minor allele frequency (MAF). The number of Altai-matching introgressed variants that fall in each MAF bin are 24,598, 16,016, and 3,923 respectively for 5-10%, 10-21%, and 21-50% (70,544 variants with frequency $< 5\%$ are not included). Each dot represents heritability enrichment or depletion of one of the traits ($n = 41$) (legend in Fig. 2.1). Traits with heritability depletion less than 0.05 are truncated. *P*-values are from two-tailed two-sample t-tests. The boxplot centers represent medians, the boxes are bounded by the first and third quartiles, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box.

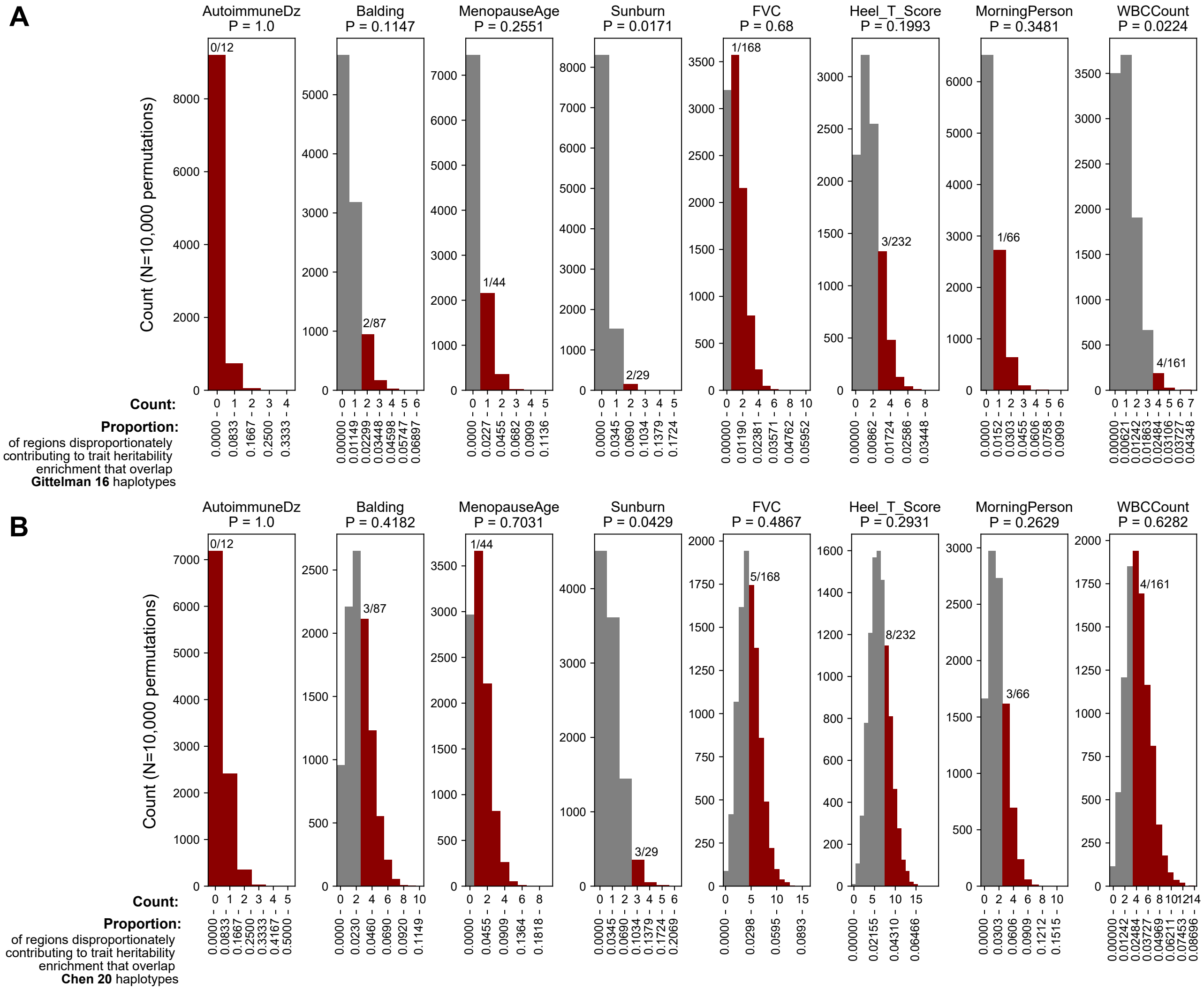


Figure 6.9: Neanderthal introgressed regions that disproportionately contribute to heritability of sunburn and WBC count are enriched in haplotypes with evidence of adaptive selection.
For the eight traits investigated in Fig. 2.3, we identified regions that contribute to directional trait heritability (see 2.4:Methods, Fig. 2.4). We intersected these regions of interest with high-frequency haplotypes with evidence of adaptive selection identified by (A) Gittelman et al. 2016 and (B) Chen et al. 2020. This observed overlap is reported as a fraction in each sub-figure (e.g., we identify 29 Sunburn-related regions of interest; two of these overlap Gittelman haplotypes [2/29]). The histogram is an empirical distribution ($n = 10,000$ permutations) for the expected overlap of the heritability-enriched regions of interest with the adaptive haplotypes. Empirical observations equal to or more extreme than the observed overlap are in red and are used to calculate the empirical one-tailed P-value. For example, we observe that 2/29 Sunburn-related regions of interest overlap Gittelman haplotypes; under the null, you would expect this (or more overlap) 1.7% of the time ($P = 0.017$, FDR $q = 0.09$).

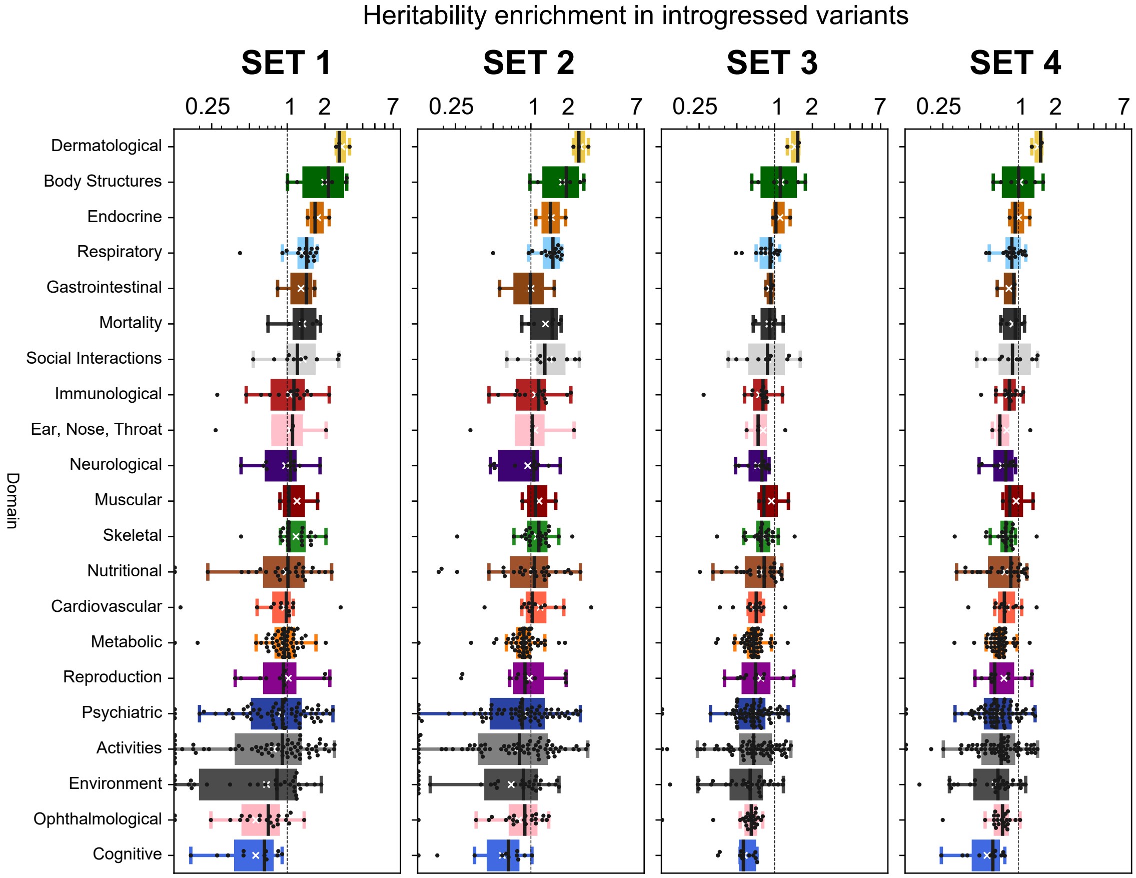


Figure 6.10: Patterns of complex trait heritability across 405 traits organized by DOMAIN across four different sets of Neanderthal introgressed variation.
Across four sets of Neanderthal introgressed variation (from most stringent to least stringent [2.4:Methods]), we show the trait heritability enrichment (or depletion) across 21 phenotypic domains (across $n = 405$ traits). Domains are ordered by the magnitude of the median enrichment in Set 1 variants for comparison across sets. Results from Set 1 are the same as those depicted in Fig. 2.2A. Each point represents heritability enrichment or depletion of one trait in Altai-matching introgressed variants. The boxplot centers represent medians, the white Xs denote means, the boxes are bounded by the first and third quartile, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. Traits with depletion less than 0.125 are plotted on the baseline for visualization.

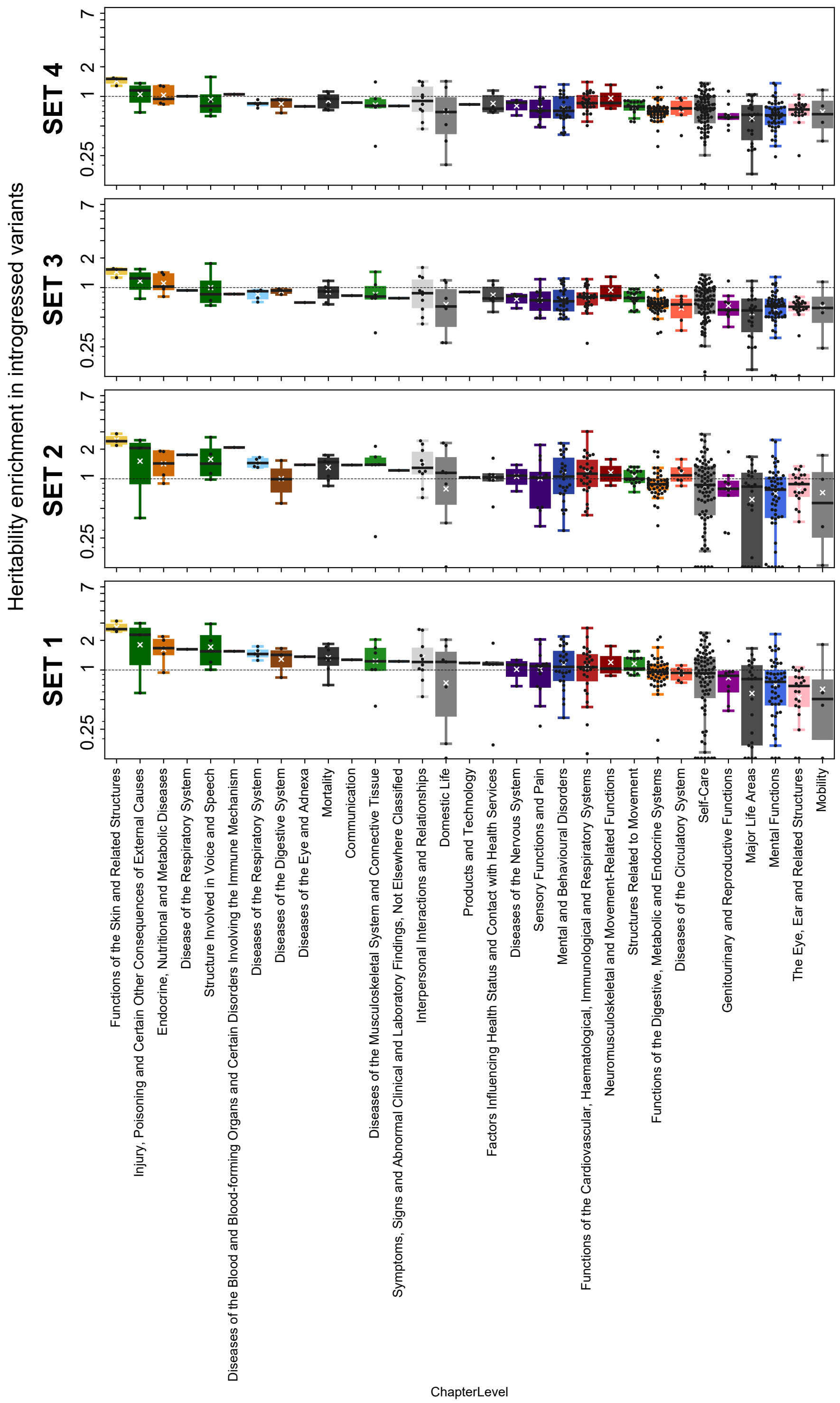


Figure 6.11: Patterns of complex trait heritability across 405 traits organized by CHAPTER across four different sets of Neanderthal introgressed variation.

Across four sets of Neanderthal introgressed variation (from most stringent to least stringent [2.4:Methods]), we show the trait heritability enrichment (or depletion) across 31 phenotypic chapters (across $n = 405$ traits). Chapters are ordered by the magnitude of the median enrichment in Set 1 variants for comparison across sets. Each point represents heritability enrichment or depletion of one trait in Altai-matching introgressed variants. The boxplot centers represent medians, the white Xs denote means, the boxes are bounded by the first and third quartile, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. Traits with depletion less than 0.125 are plotted on the baseline for visualization.

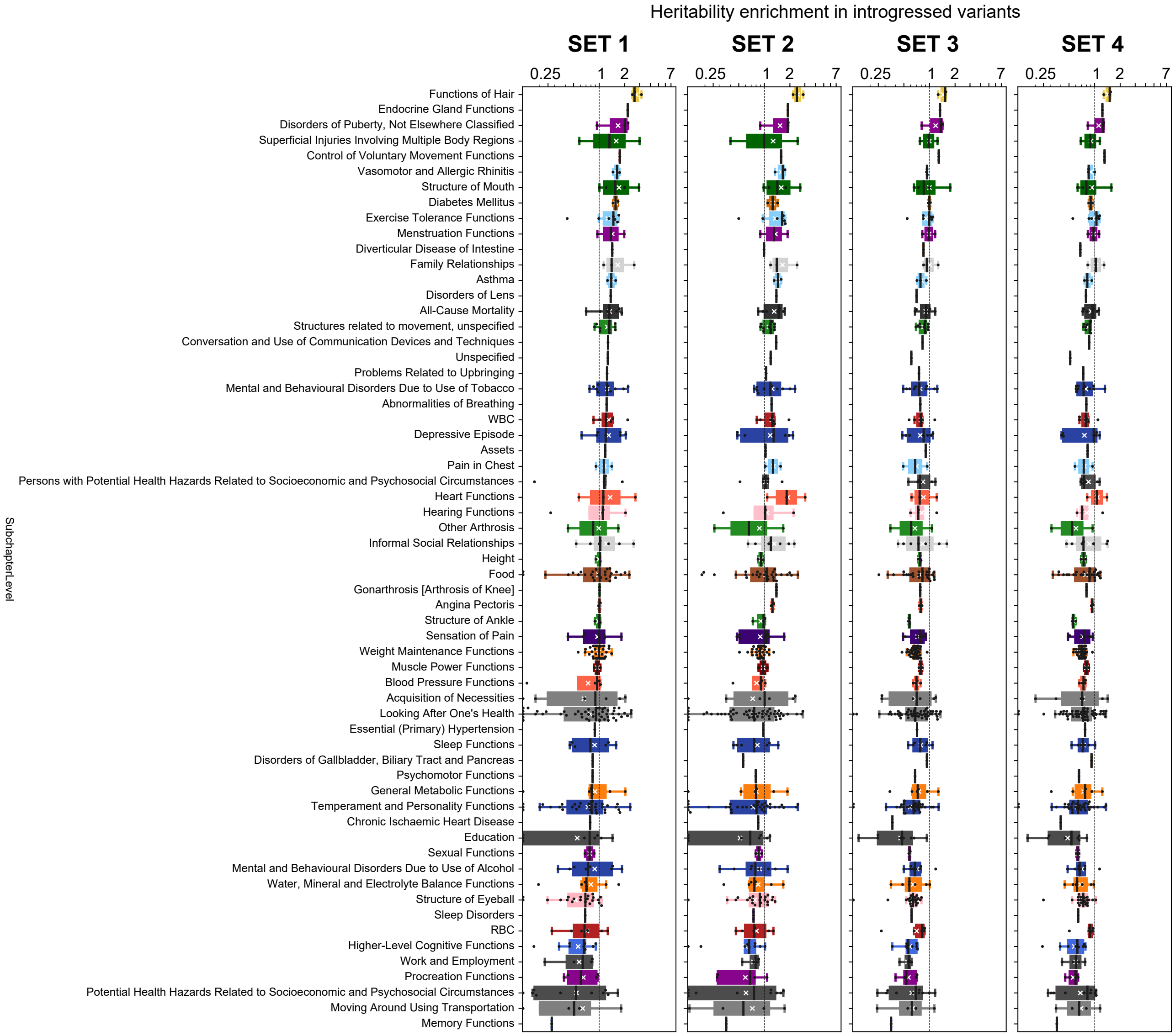
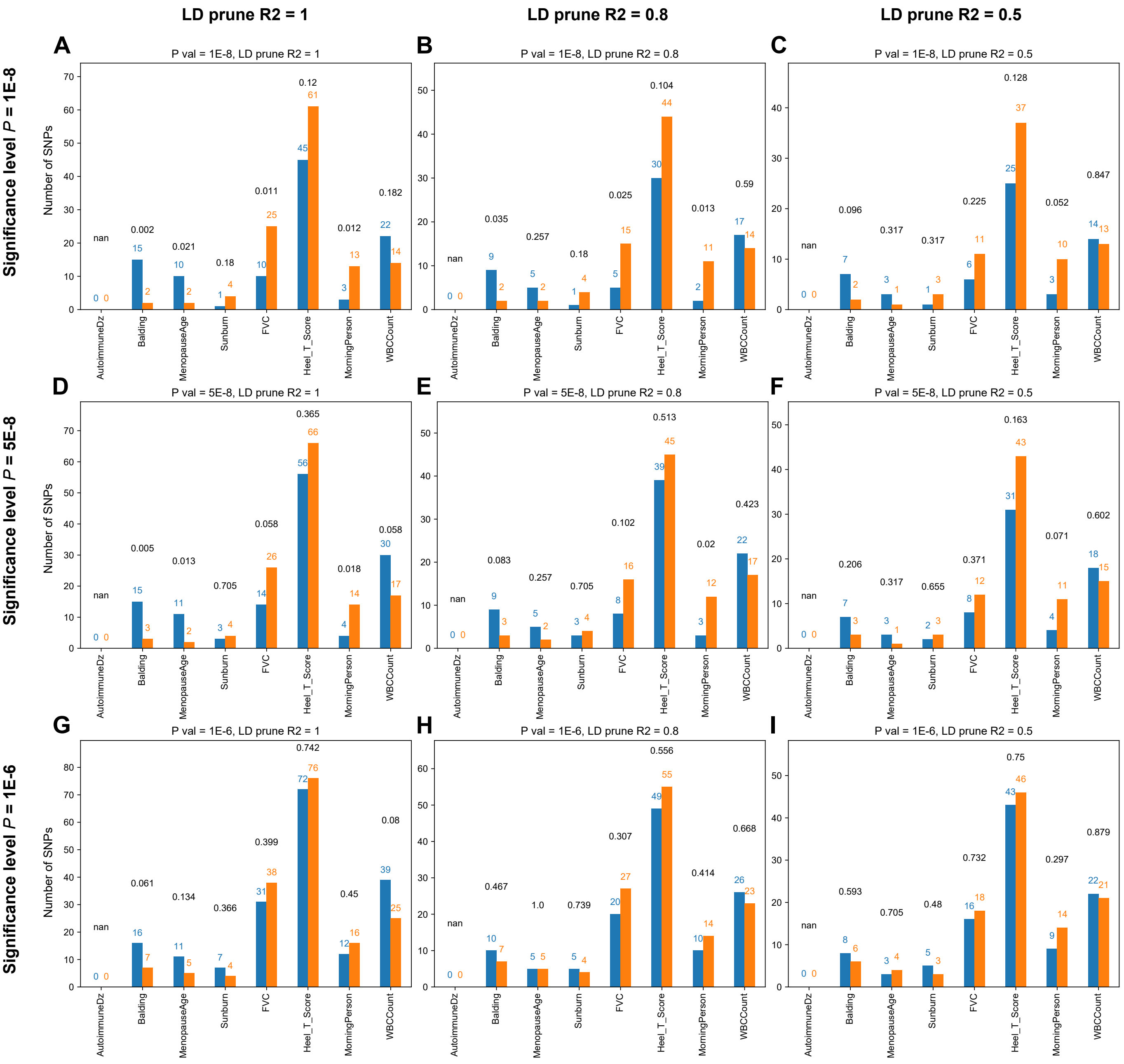


Figure 6.12: Patterns of complex trait heritability across 405 traits organized by SUBCHAPTER across four different sets of Neanderthal introgressed variation.

Across four sets of Neanderthal introgressed variation (from most stringent to least stringent [2.4:Methods]), we show the trait heritability enrichment (or depletion) across 62 phenotypic subchapters (across $n = 405$ traits). Subchapters are ordered by the magnitude of the median enrichment in Set 1 variants for comparison across sets. A subset of the results from Set 1 is the same as those depicted in Fig. 2.2B-E. Each point represents heritability enrichment or depletion of one trait in Altai-matching introgressed variants. The boxplot centers represent medians, the white Xs denote means, the boxes are bounded by the first and third quartile, and the Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. Traits with depletion less than 0.125 are plotted on the baseline for visualization.



	AutoimmuneDz	Balding	MenopauseAge	Sunburn	FVC	Heel_T_Score	MorningPerson	WBCCount
Direction of effect	protection	hair loss	younger	never burnt	smaller	less dense	evening	lower
	risk	full hair	older	sunburns	larger	more dense	morning	higher

Figure 6.13: Directionality of effects for introgressed variants with the strongest trait associations is stable at different significance levels and pruning thresholds.

For eight traits, we intersected introgressed Altai-matching Neanderthal alleles (LD-expanded to $r^2 = 1$) with the genome-wide significant variants from each GWAS. After pruning for linked variants, we plot the number of significantly associated introgressed variants by their direction of effect (risk-increasing or risk-decreasing [legend]). Fig. 2.3A shows this result for the threshold $P < 1 \times 10^{-8}$ and pruning threshold of $r^2 = 1$. Here, we show these results are consistent at different thresholds ([**A, B, C**] $P < 1 \times 10^{-8}$; [**D, E, F**]; $P < 5 \times 10^{-8}$; [**G, H, I**] $P < 1 \times 10^{-6}$) and different LD pruning thresholds ([**A, D, E**] $r^2 = 1$; [**B, E, H**] $r^2 > 0.8$; [**C,F,I**] $r^2 > 0.5$). Black numbers above the bars represent P values (one-tailed χ^2 goodness of fit test).

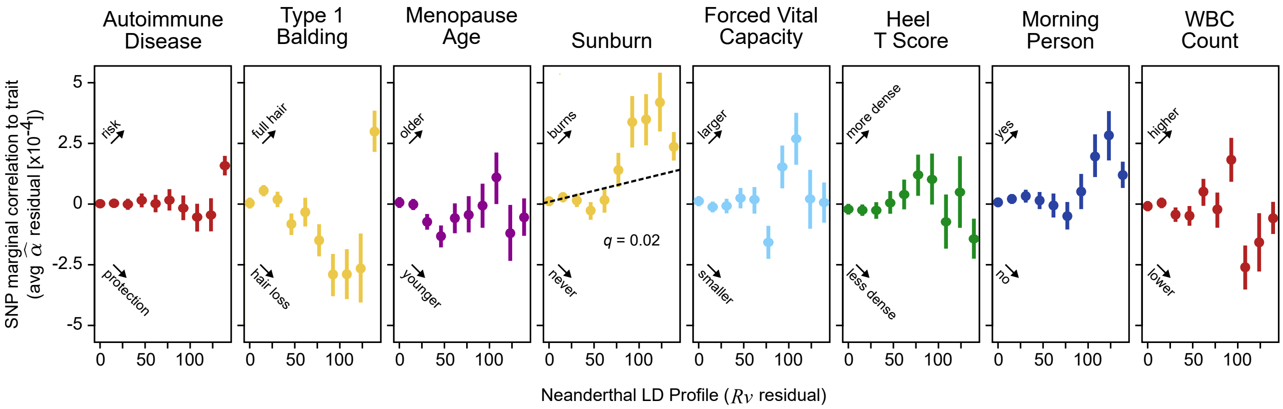


Figure 6.14: Neanderthal alleles confer genome-wide uni-directional effects for some traits. We use signed LD profile regression to consider the direction of effect over all introgressed variants, not just those with the largest effects. Here, we only consider eight traits with evidence of heritability enrichment in introgressed variants. For each variant (genome-wide, $n = 1,187,349$), we plot the marginal correlation ($\hat{\alpha}$) of the variant to the trait versus the Neanderthal LD profile ($R\nu$). Increased signed LD Profile reflects increased conditional LD to a Neanderthal introgressed allele. For visualization, we bin $R\nu$ into 10 equally spaced intervals and plot the average $\hat{\alpha}$ with 95% bootstrapped confidence intervals. The correlation between $R\nu$ and $\hat{\alpha}$ indicates the genome-wide direction of effect. For example, the positive correlation for sunburn indicates a significant uni-directional relationship genome-wide between Neanderthal introgressed alleles and risk for sunburn (empirical null distribution $P = 0.001$, $q = 0.02$, same as Fig. 2.3B). Other traits show directionality similar to the P-value threshold analysis (Fig. 2.3A). For example, the Neanderthal LD profile correlates with risk for younger menopause age ($r_f = -0.091\%$) and increased propensity to be a morning person ($r_f = 0.033\%$). The remaining traits (like autoimmune disease and WBC Count) do not show consistent directionality genome-wide; instead, these traits have genomic windows where Neanderthal alleles contribute in risk-increasing directions and other windows with risk-reducing directions (i.e., bi-directional). Results for all 41 representative traits are in Table 6.6.

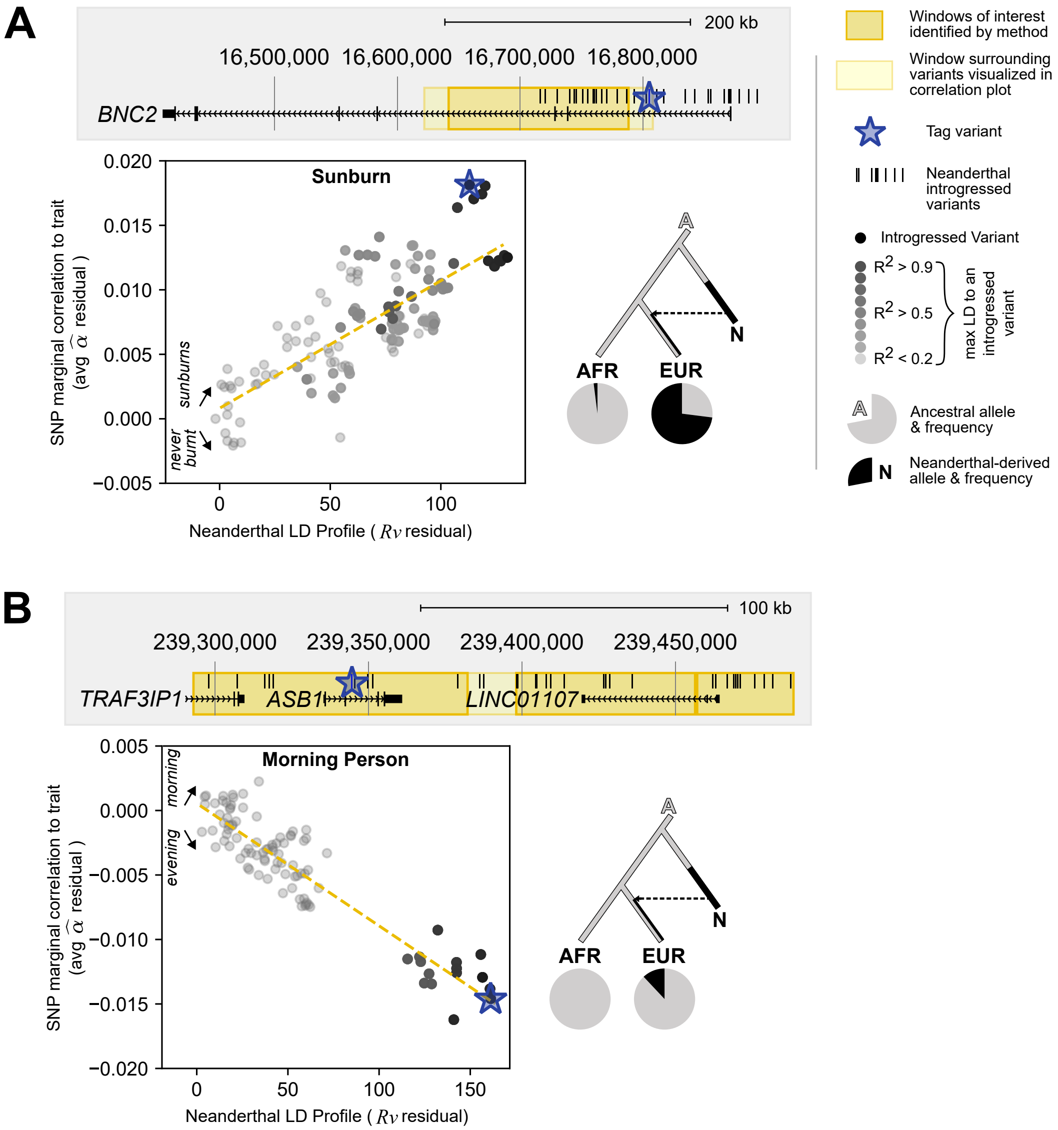


Figure 6.15: Windows with strong correlations between Neanderthal LD profile and trait-association highlights genes implicated in introgression's effect on sunburn risk and chronotype.
A) The genomic window (chr9:16,641,651-16,787,775) overlaps BNC2 and has a positive relationship ($r = +0.82$) between Neanderthal LD profile and sunburn risk. Supporting this association, the starred variant (rs10962612; EUR AF: 0.73; AFR AF: 0.02) was previously shown to tag an introgressed haplotype and associate with childhood sunburn risk and poor tanning[62,161]. Among these regions, we identify many promising candidates, including one nearby *SPATA33* which has been implicated in tanning response, facial pigmentation, and skin cancer[259] and one nearby *MC1R* which is a key genetic determinant of pigmentation and hair color[62]. (B) The genomic region shown highlights three windows (chr2:239,292,973-239,382,296, chr2:239,398,170-239,456,308, chr2:239,457,097-239,488,435) around *ASB1* with a negative relationship between Neanderthal LD profile and morning person status. The scatter plot shows this negative correlation ($r = -0.92$) for chr2:239,292,973-239,488,435; hence, increased Neanderthal LD profile is associated with increased eveningness. Supporting this association, the starred variant (rs3191996; EUR AF:0.12; AFR AF; 0.00) was previously identified as an archaic allele associated with preference for being an evening person[62]. For each example, we display the genomic region overlapped by the identified window(s) of interest (dark yellow box), genes, all Altai-matching Neanderthal-introgressed variants (black marks). For the region in light yellow, we display a scatter plot between the variant's Neanderthal LD Profile (Rv) and trait marginal correlation ($\hat{\alpha}$). Each variant is colored by its maximum LD to an introgressed variant. We display the evolutionary history (dendrogram) of each discussed tag variant (blue star) with its allele frequency in Africans (AFR) and Europeans (EUR) (pie charts).

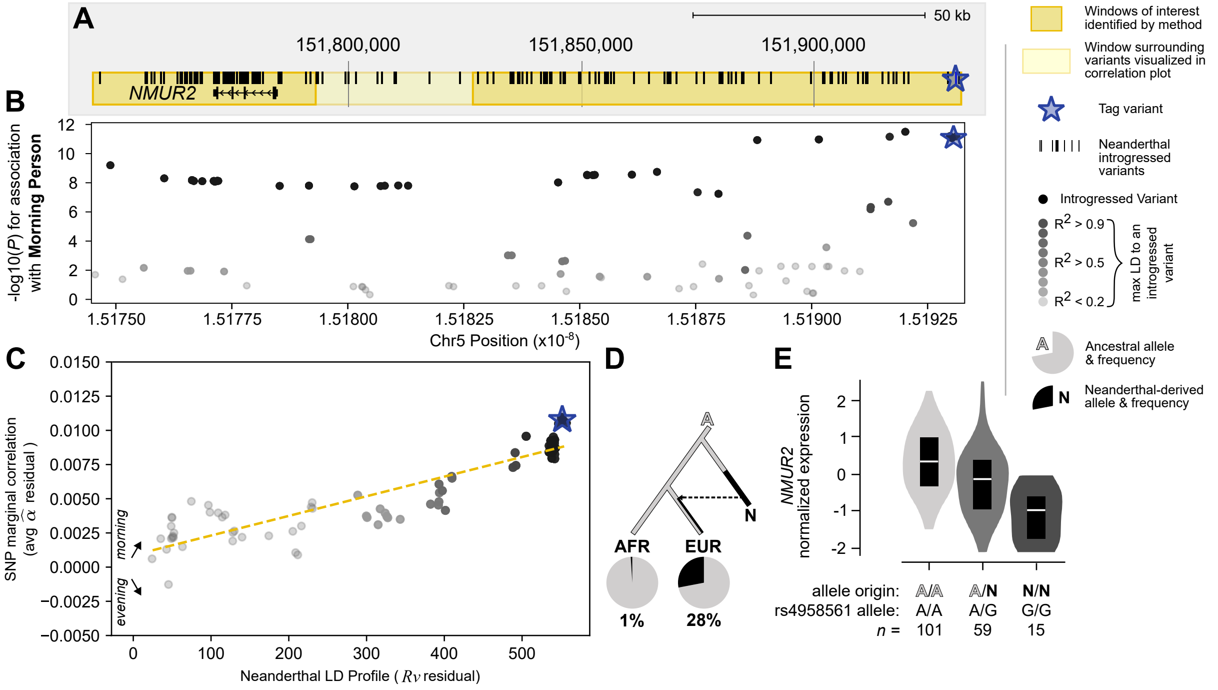


Figure 6.16: Correlations between Neanderthal LD profile and trait-association at *NMUR2* highlight putative mechanisms for the effect of introgression on morningness.
(A) We discover a positive relationship between Neanderthal LD profile and morning person status in two regions (dark yellow boxes, chr5:151,745,423-151,793,214, chr5:151,826,774-151,931,514) near *NMUR2*. (B) The Manhattan plot for this region highlights that introgressed variants have the strongest association to the morning person GWAS (blue star, $P = 9 \times 10^{-12}$ at rs4958561). We note that the variant most strongly associated with morningness (rs10045463: $P = 3 \times 10^{-12}$) is introgressed; however, it is a reintroduced ancestral allele (lost in AMHs but reintroduced through introgression, not pictured). There are additional clusters of both introgressed and non-introgressed variants significantly associated with morningness downstream of the highlighted region (Supplemental Text 6.1); however, they are only in moderate LD ($r^2 = 0.046 - 0.39$) with the other non-introgressed clusters, suggesting that this particular signal is likely driven by variants on the introgressed haplotype. (C) The scatter plot shows the positive relationship ($r = +0.91$) for chr5:151,745,423-151,931,514 (entire light yellow box) which indicates Neanderthal introgression at this locus is associated with increased morningness. (D) The starred variant (rs4958561) is derived in Neanderthals (N) and at 28% frequency in modern Europeans (EUR) with 1% frequency in Africans (AFR)(1000G super-populations). (E) This haplotype (tagged by rs4958561) is an eQTL in which Neanderthal alleles associate with increased *NMUR2* expression in frontal cortex (two-tailed t-test $P = 1 \times 10^{-9}$) and cortex (not shown; $P = 9 \times 10^{-5}$). The boxplot centers represent medians and the boxes are bounded by the first and third quartiles.

Nickname	Trait	M	h2	h2_SE	N	Source
Anorexia	Anorexia	931184	0.2153	0.0169	32143	Boraska et al. 2014 Mol Psych(Boraska et al. 2014)
ASD	Autism_Spectrum	1173307	0.4607	0.0517	10263	PGC Cross-Disorder Group, 2013 Lancet(Smoller et al. 2013)
AutoimmuneDz	Auto_Immune_Traits_(Sure)	1187056	0.0068	0.0013	459324	UKBiobank(Sudlow et al. 2015)
Balding	Balding_Type_I	1187056	0.2154	0.019	208336	UKBiobank(Sudlow et al. 2015)
BMI	BMI	1187056	0.252	0.0071	457824	UKBiobank(Sudlow et al. 2015)
CrohnsDz	Crohn's_Disease	1051514	0.4723	0.0575	20883	Jostins et al., 2012 Nature(Jostins et al. 2012)
DepressiveSxs	Depressive_symptoms	1115393	0.0473	0.0037	161460	Okbay et al., 2016 Nat Genet(Okbay et al. 2016)
DermDz	Dermatologic_Diseases	1187056	0.0094	0.0014	459324	UKBiobank(Sudlow et al. 2015)
Eczema	Eczema	1187056	0.0675	0.0038	458699	UKBiobank(Sudlow et al. 2015)
EosinophilCount	Eosinophil_Count	1187056	0.1977	0.0143	439938	UKBiobank(Sudlow et al. 2015)
FEV1_FVC_Ratio	FEV1-FVC_Ratio	1187056	0.2336	0.0113	371949	UKBiobank(Sudlow et al. 2015)
FirstBirthAge	Age_first_birth	1079424	0.0617	0.0033	222037	Barban et al., 2016 Nat Genet(Barban et al. 2016)
FVC	Forced_Vital_Capacity_(FVC)	1187056	0.2068	0.0065	371949	UKBiobank(Sudlow et al. 2015)
HairColor	Hair_Color	1187056	0.4523	0.1497	452720	UKBiobank(Sudlow et al. 2015)
HDL	HDL	1019272	0.1362	0.0166	99900	Teslovich et al., 2010 Nature(Teslovich et al. 2010)
Heel_T_Score	Heel_T_Score	1187056	0.3628	0.0307	445921	UKBiobank(Sudlow et al. 2015)
Height	Height	1187056	0.6034	0.027	458303	UKBiobank(Sudlow et al. 2015)
HighCholesterol	High_Cholesterol	1187056	0.0468	0.0039	459324	UKBiobank(Sudlow et al. 2015)
Hypothyroidism	Hypothyroidism	1187056	0.0459	0.0037	459324	UKBiobank(Sudlow et al. 2015)
LDL	LDL	1017973	0.121	0.0166	95454	Teslovich et al., 2010 Nature(Teslovich et al. 2010)
MenarcheAge	Age_at_Menarche	1187056	0.2457	0.0102	242278	UKBiobank(Sudlow et al. 2015)
MenopauseAge	Age_at_Menopause	1187056	0.1215	0.0086	143025	UKBiobank(Sudlow et al. 2015)
MorningPerson	Morning_Person	1187056	0.1002	0.0035	410520	UKBiobank(Sudlow et al. 2015)
Neuroticism	Neuroticism	1187056	0.1113	0.0037	372066	UKBiobank(Sudlow et al. 2015)
NumChildrenBorn	Number_children_ever_born	1080059	0.0256	0.0018	318863	Barban et al., 2016 Nat Genet(Barban et al. 2016)
PlateletCount	Platelet_Count	1187056	0.349	0.0294	444382	UKBiobank(Sudlow et al. 2015)
RA	Rheumatoid_Arthritis	1125155	0.1694	0.023	38242	Okada et al., 2014 Nature(Okada et al. 2014)
RBCCount	Red_Blood_Cell_Count	1187056	0.2434	0.0191	445174	UKBiobank(Sudlow et al. 2015)
RDW	Red_Blood_Cell_Distribution_Width	1187056	0.2234	0.0198	442700	UKBiobank(Sudlow et al. 2015)
Resp_ENT_Dz	Respiratory_and_Ear-nose-throat_Diseases	1187056	0.0483	0.0034	459324	UKBiobank(Sudlow et al. 2015)
Schizophrenia	Schizophrenia	1083014	0.4512	0.0189	70100	SCZ Working Group of the PGC, 2014 Nature(Ripke et al. 2014)
SkinColor	Skin_Color	1187056	0.1896	0.0539	453609	UKBiobank(Sudlow et al. 2015)
SmokingStatus	Smoking_Status	1187056	0.0972	0.0032	457683	UKBiobank(Sudlow et al. 2015)
Sunburn	Sunburn_Occasion	1187056	0.0915	0.0162	344229	UKBiobank(Sudlow et al. 2015)
SystolicBP	Systolic_Blood_Pressure	1187056	0.1966	0.007	422771	UKBiobank(Sudlow et al. 2015)
T2D	Type_2_Diabetes	1187056	0.043	0.0025	459324	UKBiobank(Sudlow et al. 2015)
Tanning	Tanning	1187056	0.172	0.0609	449984	UKBiobank(Sudlow et al. 2015)
UC	Ulcerative_Colitis	1076834	0.2424	0.032	27432	Jostins et al., 2012 Nature(Jostins et al. 2012)
WaistHipRatio	Waist-hip_Ratio	1187056	0.1423	0.0067	458417	UKBiobank(Sudlow et al. 2015)
WBCCount	White_Blood_Cell_Count	1187056	0.1873	0.0105	444502	UKBiobank(Sudlow et al. 2015)
YearsOfEd	College_Education	1187056	0.1299	0.0037	454813	UKBiobank(Sudlow et al. 2015)

Table 6.1: Traits used for partitioned heritability analyses with S-LDSC.

trait	enrichment	depletion= 1/enr	stderr	pval	qval (FDR-BH α=0.05)	lower_CI	upper_CI
HighCholesterol	0.213421	4.685574	0.24457	0.001	0.018754	0.08288	0.549572
HDL	0.359528	2.781422	0.34365	0.059	0.218119	0.124579	1.037576
LDL	0.424695	2.35463	0.4084	0.156	0.491944	0.130093	1.386438
FirstBirthAge	0.489845	2.041464	0.20429	0.013	0.08828	0.279058	0.859848
Anorexia	0.51414	1.944994	0.44698	0.277	0.51604	0.154979	1.705648
Hypothyroidism	0.536614	1.863537	0.18933	0.016	0.096183	0.322724	0.892262
RA	0.54439	1.836917	0.40492	0.25	0.511929	0.193291	1.533237
CrohnsDz	0.554532	1.803323	0.46237	0.348	0.594849	0.161768	1.900905
UC	0.579111	1.726785	0.38215	0.266	0.51604	0.221084	1.516934
PlateletCount	0.592574	1.687554	0.13719	0.004	0.035661	0.413566	0.849063
SystolicBP	0.63814	1.567053	0.09903	4E-04	0.014974	0.498461	0.81696
DepressiveSxs	0.643197	1.554734	0.46195	0.445	0.729273	0.207427	1.99445
YearsOfEd	0.669496	1.493661	0.1133	0.004	0.035661	0.508953	0.880679
BMI	0.671583	1.489019	0.09843	1E-03	0.018754	0.530272	0.850552
Neuroticism	0.706502	1.415424	0.23609	0.229	0.511929	0.401172	1.244217
WaistHipRatio	0.709536	1.409372	0.12957	0.026	0.133387	0.524501	0.959849
Schizophrenia	0.72441	1.380435	0.18433	0.133	0.454533	0.475667	1.103227
SmokingStatus	0.742216	1.347317	0.12711	0.044	0.181154	0.555173	0.992276
Resp_ENT_Dz	0.742449	1.346894	0.20765	0.216	0.511929	0.463422	1.189479
MenarcheAge	0.745924	1.34062	0.11537	0.031	0.142824	0.571187	0.974116
T2D	0.759389	1.316848	0.1904	0.213	0.511929	0.492229	1.171552
Height	0.819823	1.219775	0.1302	0.169	0.493549	0.617883	1.087763
NumChildrenBorn	0.840035	1.190426	0.36535	0.661	0.847165	0.385243	1.831727
DermDz	0.914293	1.093742	0.51444	0.867	0.95684	0.31959	2.615633
EosinophilCount	0.923507	1.082829	0.17424	0.66	0.847165	0.648118	1.315909
WBCCount	0.925591	1.080391	0.14581	0.61	0.833849	0.687596	1.245962
RDW	0.937912	1.066199	0.2109	0.769	0.913303	0.611518	1.438516
RBCCount	0.950773	1.051776	0.23659	0.835	0.95076	0.59159	1.528034
FEV1_FVC_Ratio	0.952129	1.050278	0.14646	0.743	0.913303	0.710316	1.276264
Heel_T_Score	0.968281	1.032758	0.22507	0.887	0.95684	0.621173	1.509349
MorningPerson	0.989816	1.010288	0.17079	0.952	0.976188	0.707391	1.385001
Eczema	1.008652	0.991422	0.34224	0.979	0.979414	0.524268	1.940571
FVC	1.045004	0.956934	0.16076	0.78	0.913303	0.767689	1.422494
HairColor	1.047467	0.954684	0.50478	0.925	0.972034	0.40081	2.737422
MenopauseAge	1.228115	0.814256	0.40422	0.571	0.833849	0.603101	2.50085
Sunburn	1.285756	0.777753	0.42415	0.497	0.777689	0.622973	2.653673
Balding	1.296168	0.771505	0.28119	0.29	0.516738	0.801729	2.095537
Tanning	1.341133	0.745638	0.67219	0.602	0.833849	0.444541	4.046056
SkinColor	1.40469	0.711901	0.63006	0.512	0.777689	0.508524	3.880157
ASD	1.472969	0.678901	0.42449	0.243	0.511929	0.769216	2.820585
AutoimmuneDz	1.655501	0.604047	0.53184	0.207	0.511929	0.756705	3.621864

Table 6.2: Heritability enrichment for introgressed variants.

Heritability enrichment and depletion results with confidence intervals and statistics for Neanderthal introgressed variants shown in Fig. 2.1C. *P*-values are calculated empirically by LDSC using a block jackknife ($n = 200$). *q*-values are corrected for multiple comparisons using the Benjamini-Hochberg FDR-correction at the 0.05 level. Confidence intervals are at the 95% level.

Domain	Enr (median)	Depletion (median)	Enr (mean)	Depletion (mean)	Lower CI	Upper CI	P	q	N (traits)
Dermatological	2.602	NA	2.719	NA	2.351	3.145	0.006	0.039	3
Body structures	2.135	NA	1.907	NA	1.323	2.750	0.018	0.063	6
Endocrine	1.668	NA	1.738	NA	1.384	2.183	0.042	0.109	3
Respiratory	1.429	NA	1.291	NA	1.087	1.532	0.011	0.046	16
Gastrointestinal	1.425	NA	1.256	NA	0.838	1.882	0.385	0.524	3
Mortality	1.314	NA	1.293	NA	1.009	1.658	0.089	0.187	7
Social interactions	1.208	NA	1.259	NA	0.928	1.707	0.172	0.302	10
Immunological	1.133	NA	0.971	1.030	0.734	1.284	0.837	0.837	14
Ear, nose, throat	1.103	NA	0.904	1.106	0.389	2.099	0.829	0.837	4
Neurological	1.064	NA	0.922	1.085	0.710	1.197	0.557	0.650	10
Muscular	1.034	NA	1.165	NA	0.774	1.753	0.540	0.650	3
Skeletal	1.030	NA	1.142	NA	1.009	1.293	0.047	0.110	24
Nutritional	1.014	NA	0.770	1.299	0.516	1.150	0.212	0.343	28
Cardiovascular	0.984	1.017	0.855	1.169	0.602	1.214	0.399	0.524	13
Metabolic	0.951	1.052	0.948	1.055	0.862	1.037	0.253	0.379	52
Reproduction	0.934	1.070	0.928	1.077	0.673	1.280	0.659	0.729	12
Psychiatric	0.918	1.089	0.901	1.110	0.766	1.042	0.171	0.302	67
Activities	0.915	1.093	0.776	1.288	0.600	0.965	0.024	0.073	66
Environment	0.828	1.207	0.684	1.461	0.478	0.909	0.011	0.046	31
Ophthalmological	0.707	1.415	0.562	1.779	0.325	0.824	0.005	0.039	22
Cognitive	0.659	1.518	0.512	1.954	0.373	0.702	0.002	0.039	11

Table 6.3: Domain enrichment for 405 traits.

For each of the phenotypic domains, we list the median and mean heritability enrichment. These results are also plotted in Figs. 2.2A, 6.10. For those domains which are depleted (enrichment below 1), we also report the fold-depletion (1/Enrichment). Domains are ordered by their median enrichment. *P*-values are from two-tailed one-sample t-tests. *q*-values are corrected for multiple comparisons using the Benjamini-Hochberg FDR-correction at the 0.05 level. Confidence intervals are at the 95% level. The mean enrichment, confidence intervals, and *P*-values were calculated on the log-transformed enrichment values.

trait	Enrichment (median)	Fold-depletion (median)	Enrichment (mean)	Fold-depletion (mean)	lower_CI	upper_CI	stderr	P	N (traits)
Functions of the Skin and Related Structures	2.6021	NA	2.7191	NA	2.3506	3.1453	0.0323	0.0055	3
Injury, Poisoning and Certain Other Consequences of External Causes	2.28	NA	1.584	NA	0.5888	4.2612	0.2193	0.4585	3
Endocrine, Nutritional and Metabolic Diseases	1.6678	NA	1.5855	NA	1.1852	2.1208	0.0645	0.036	5
Disease of the Respiratory System	1.6239	NA	1.6239	NA	NA	NA	NA	NA	1
Structure Involved in Voice and Speech	1.5952	NA	1.6308	NA	1.0108	2.6309	0.106	0.1388	4
Diseases of the Blood and Blood-forming Organs and Certain Disorders Involving the Immune Mechanism	1.5488	NA	1.5488	NA	NA	NA	NA	NA	1
Diseases of the Respiratory System	1.4535	NA	1.4699	NA	1.3193	1.6378	0.024	0.0022	5
Diseases of the Digestive System	1.4246	NA	1.2555	NA	0.8378	1.8817	0.0897	0.3852	3
Diseases of the Eye and Adnexa	1.3657	NA	1.3657	NA	NA	NA	NA	NA	1
Mortality	1.3138	NA	1.2931	NA	1.0086	1.6579	0.0551	0.089	7
Communication	1.2704	NA	1.2704	NA	NA	NA	NA	NA	1
Diseases of the Musculoskeletal System and Connective Tissue	1.2496	NA	1.1427	NA	0.7326	1.7825	0.0985	0.582	6
Symptoms, Signs and Abnormal Clinical and Laboratory Findings, Not Elsewhere Classified	1.223	NA	1.223	NA	NA	NA	NA	NA	1
Interpersonal Interactions and Relationships	1.208	NA	1.259	NA	0.9285	1.7073	0.0675	0.1724	10
Domestic Life	1.2077	NA	0.7389	1.3533	0.0387	1.6863	0.067	0.5693	7
Products and Technology	1.1789	NA	1.1789	NA	NA	NA	NA	NA	1
Factors Influencing Health Status and Contact with Health Services	1.1468	NA	0.9122	1.0962	0.4669	1.7823	0.1484	0.7988	6
Diseases of the Nervous System	1.1297	NA	0.9918	1.0083	0.6864	1.433	0.0815	0.969	3
Sensory Functions and Pain	1.0942	NA	0.9302	1.075	0.6859	1.2615	0.0675	0.6499	13
Mental and Behavioural Disorders	1.0822	NA	1.0651	NA	0.882	1.2863	0.0418	0.518	26
Functions of the Cardiovascular, Haematological, Immunological and Respiratory Systems	1.0688	NA	0.9589	1.0428	0.7564	1.2156	0.0526	0.7317	27
Neuromusculoskeletal and Movement-Related Functions	1.0342	NA	1.1651	NA	0.7743	1.753	0.0905	0.5398	3
Structures Related to Movement	1.0304	NA	1.1415	NA	1.0417	1.2509	0.0203	0.0114	18
Functions of the Digestive, Metabolic and Endocrine Systems	0.9558	1.0463	0.9674	1.0337	0.875	1.0628	0.0071	0.5015	53
Diseases of the Circulatory System	0.9351	1.0694	0.9142	1.0939	0.8053	1.0378	0.0281	0.2243	6
Self-Care	0.927	1.0788	0.8493	1.1774	0.7058	1.0007	0.0117	0.0544	82
Genitourinary and Reproductive Functions	0.8726	1.1461	0.7737	1.2924	0.5617	1.0658	0.071	0.1551	9
Major Life Areas	0.8093	1.2357	0.5782	1.7294	0.3432	0.8374	0.0217	0.0052	23
Mental Functions	0.7572	1.3206	0.7149	1.3988	0.5795	0.8576	0.0116	0.0004	52
The Eye, Ear and Related Structures	0.6878	1.4539	0.529	1.8904	0.2925	0.7905	0.0223	0.0028	21
Mobility	0.5122	1.9525	0.3075	3.2521	0.0456	2.0752	0.4231	0.3127	4

Table 6.4: Chapter enrichment for 405 traits.

For each of the phenotypic chapters, we list the median and mean heritability enrichment. These results are also plotted in Fig. 6.11. For those subchapters which are depleted (enrichment below 1), we also report the fold-depletion (1/Enrichment). Subchapters are ordered by their median enrichment. *P*-values are from two-tailed one-sample t-tests. *q*-values are corrected for multiple comparisons using the Benjamini-Hochberg FDR-correction at the 0.05 level. Confidence intervals are at the 95% level. The mean enrichment, confidence intervals, and p-values were calculated on the log-transformed enrichment values.

trait	Enrichment (median)	Fold-depletion (median)	Enrichment (mean)	Fold-depletion (mean)	lower_CI	upper_CI	stderr	P	N (traits)
Functions of Hair	2.60212	NA	2.719	NA	2.351	3.145	0.032	0.0055	3
Endocrine Gland Functions	2.1638	NA	2.164	NA	NA	NA	NA	NA	1
Disorders of Puberty, Not Elsewhere Classified	2.00493	NA	1.604	NA	0.949	2.711	0.116	0.2197	3
Superficial Injuries Involving Multiple Body Regions	1.78428	NA	1.32	NA	0.267	6.53	0.354	0.791	2
Control of Voluntary Movement Functions	1.74776	NA	1.748	NA	NA	NA	NA	NA	1
Vasomotor and Allergic Rhinitis	1.62395	NA	1.6	NA	1.446	1.772	0.023	0.012	3
Structure of Mouth	1.59515	NA	1.631	NA	1.011	2.631	0.106	0.1388	4
Diabetes Mellitus	1.56164	NA	1.558	NA	1.363	1.78	0.03	0.097	2
Exercise Tolerance Functions	1.46753	NA	1.19	NA	0.826	1.716	0.081	0.3868	7
Menstruation Functions	1.46099	NA	1.369	NA	0.671	2.793	0.158	0.5462	2
Diverticular Disease of Intestine	1.42457	NA	1.425	NA	NA	NA	NA	NA	1
Family Relationships	1.39983	NA	1.597	NA	0.983	2.597	0.108	0.1995	3
Asthma	1.38791	NA	1.396	NA	1.23	1.584	0.028	0.0354	3
Disorders of Lens	1.36575	NA	1.366	NA	NA	NA	NA	NA	1
All-Cause Mortality	1.31381	NA	1.293	NA	1.009	1.658	0.055	0.089	7
Structures related to movement, unspecified	1.31122	NA	1.232	NA	1.101	1.379	0.025	0.0039	12
Conversation and Use of Communication Devices and Techniques	1.2704	NA	1.27	NA	NA	NA	NA	NA	1
Unspecified	1.26565	NA	1.266	NA	NA	NA	NA	NA	1
Problems Related to Upbringing	1.23877	NA	1.239	NA	NA	NA	NA	NA	1
Mental and Behavioural Disorders Due to Use of Tobacco	1.23545	NA	1.208	NA	0.989	1.474	0.044	0.0931	11
Abnormalities of Breathing	1.22295	NA	1.223	NA	NA	NA	NA	NA	1
WBC	1.20753	NA	1.271	NA	1.022	1.58	0.048	0.0745	7
Depressive Episode	1.19377	NA	1.232	NA	0.893	1.698	0.071	0.2504	7
Assets	1.17888	NA	1.179	NA	NA	NA	NA	NA	1
Pain in Chest	1.16059	NA	1.135	NA	0.746	1.726	0.093	0.6608	2
Persons with Potential Health Hazards Related to Socioeconomic and Psychosocial Circumstances	1.14678	NA	0.912	1.0962	0.467	1.782	0.148	0.7988	6
Heart Functions	1.11758	NA	1.197	NA	0.501	2.86	0.193	0.7242	3
Hearing Functions	1.10312	NA	0.904	1.1063	0.389	2.099	0.187	0.8294	4
Other Arthrosis	1.05344	NA	0.848	1.179	0.223	3.231	0.296	0.8493	2
Informal Social Relationships	1.02973	NA	1.137	NA	0.779	1.659	0.084	0.5305	7
Height	1.01943	NA	0.987	1.0135	0.916	1.063	0.017	0.758	3
Food	1.01416	NA	0.77	1.2986	0.516	1.15	0.089	0.2124	28
Gonarthrosis [Arthrosis of Knee]	1.01014	NA	1.01	NA	1.01	1.01	0	0	2
Angina Pectoris	1.00776	NA	1.007	NA	0.961	1.056	0.01	0.8089	2
Structure of Ankle	1.00197	NA	0.972	1.0289	0.89	1.062	0.02	0.5932	3
Sensation of Pain	0.999	1.001005	0.893	1.1193	0.627	1.273	0.078	0.5553	7
Weight Maintenance Functions	0.99321	1.006834	0.974	1.0263	0.919	1.033	0.013	0.39	39
Muscle Power Functions	0.95457	1.047593	0.951	1.0513	0.807	1.121	0.036	0.6569	2
Blood Pressure Functions	0.94752	1.055391	0.601	1.6636	0.233	1.553	0.21	0.3705	4
Acquisition of Necessities	0.94042	1.063349	0.66	1.5152	-0.111	1.755	0.078	0.5237	6
Looking After One's Health	0.91496	1.092941	0.783	1.2764	0.596	0.984	0.016	0.04	54
Essential (Primary) Hypertension	0.8866	1.127904	0.887	1.1279	NA	NA	NA	NA	1
Sleep Functions	0.85658	1.167432	0.805	1.2423	0.51	1.271	0.101	0.3944	6
Disorders of Gallbladder, Biliary Tract and Pancreas	0.83863	1.192415	0.839	1.1924	NA	NA	NA	NA	1
Psychomotor Functions	0.83594	1.196256	0.836	1.1963	NA	NA	NA	NA	1
General Metabolic Functions	0.81556	1.226145	0.891	1.1218	0.397	1.49	0.042	0.7165	6
Temperament and Personality Functions	0.81248	1.230793	0.737	1.3571	0.552	0.935	0.016	0.0146	35
Chronic Ischaemic Heart Disease	0.78304	1.277077	0.783	1.2771	NA	NA	NA	NA	1
Education	0.77912	1.283499	0.551	1.8141	0.174	0.995	0.036	0.0884	8
Sexual Functions	0.77625	1.288244	0.77	1.2983	0.603	0.984	0.054	0.2837	2
Mental and Behavioural Disorders Due to Use of Alcohol	0.73803	1.354964	0.789	1.2669	0.513	1.216	0.096	0.3186	8
Water, Mineral and Electrolyte Balance Functions	0.70757	1.413285	0.702	1.4253	0.425	1.159	0.111	0.2159	7
Structure of Eyeball	0.68781	1.453888	0.529	1.8904	0.293	0.79	0.022	0.0028	21
Sleep Disorders	0.68495	1.459958	0.685	1.46	NA	NA	NA	NA	1
RBC	0.67416	1.483319	0.656	1.5252	0.421	1.021	0.098	0.1209	6
Higher-Level Cognitive Functions	0.65861	1.518338	0.519	1.9268	0.368	0.733	0.076	0.0058	9
Work and Employment	0.65774	1.520355	0.528	1.894	0.293	0.952	0.131	0.1237	4
Procreation Functions	0.60936	1.641079	0.617	1.621	0.426	0.894	0.082	0.0633	5
Potential Health Hazards Related to Socioeconomic and Psychosocial Circumstances	0.58137	1.720078	0.54	1.8503	0.128	1.035	0.04	0.1	10
Moving Around Using Transportation	0.51217	1.952477	0.307	3.2521	0.046	2.075	0.423	0.3127	4
Memory Functions	0.2764	3.61797	0.276	3.618	NA	NA	NA	NA	1

Table 6.5: Subchapter enrichment for 405 traits.

For each of the phenotypic subchapters, we list the median and mean heritability enrichment. These results are also plotted in Figs. 2.2B-E and Fig. 6.12. For those subchapters which are depleted (enrichment below 1), we also report the fold-depletion (1/Enrichment). Subchapters are ordered by their median enrichment. P-values are from two-tailed one-sample t-tests. q-values are corrected for multiple comparisons using the Benjamini-Hochberg FDR-correction at the 0.05 level. Confidence intervals are at the 95% level. The mean enrichment, confidence intervals, and P-values were calculated on the log-transformed enrichment values.

	S-LDSC partitioned h^2			SLDP direction of effect					
Phenotype	**h^2 Enr**	**SE**	**P**	r_f	**Z**	**P**	**q**	**mu**	**SE(mu)**
AutoimmuneDz	3.934	1.475	0.028	6.97E-04	1.064	0.287	0.592	2.11E-07	2.03E-07
Balding	2.269	0.699	0.068	1.52E-03	0.189	0.850	0.914	1.33E-06	2.00E-06
MenopauseAge	2.205	1.148	0.293	-9.06E-04	-1.020	0.308	0.592	-5.58E-07	5.38E-07
Sunburn	2.078	0.865	0.208	1.82E-03	3.184	0.001	0.020	9.94E-07	2.98E-07
HairColor	1.935	1.277	0.465	-4.79E-04	-0.628	0.530	0.714	-5.33E-07	6.59E-07
SkinColor	1.883	1.363	0.508	4.54E-04	0.873	0.383	0.604	3.45E-07	4.23E-07
FVC	1.819	0.448	0.069	-5.37E-04	-0.154	0.878	0.914	-4.41E-07	1.08E-06
Heel_T_Score	1.780	0.499	0.126	-7.20E-04	-1.180	0.238	0.592	-7.75E-07	6.43E-07
Tanning	1.752	1.451	0.592	-8.12E-04	-1.759	0.079	0.592	-6.06E-07	3.38E-07
MorningPerson	1.625	0.458	0.174	3.28E-04	0.589	0.556	0.714	1.80E-07	3.02E-07
Eczema	1.544	0.742	0.459	-8.00E-04	-1.456	0.145	0.592	-3.80E-07	2.61E-07
EosinophilCount	1.528	0.414	0.202	-9.39E-04	-1.247	0.212	0.592	-7.97E-07	6.55E-07
WBCCount	1.510	0.310	0.100	-1.42E-04	-0.222	0.824	0.913	-1.15E-07	5.45E-07
FEV1_FVC_Ratio	1.303	0.383	0.430	-6.82E-04	-1.351	0.177	0.592	-6.02E-07	4.36E-07
DermDz	1.204	1.390	0.884	-2.30E-05	-0.033	0.973	0.973	-6.97E-09	2.19E-07
WaistHipRatio	1.199	0.309	0.517	6.75E-04	0.587	0.557	0.714	4.60E-07	6.42E-07
T2D	1.197	0.484	0.681	9.20E-05	0.137	0.891	0.914	3.35E-08	2.48E-07
SmokingStatus	1.119	0.301	0.694	-8.24E-04	-1.235	0.217	0.592	-4.48E-07	3.57E-07
RDW	1.114	0.394	0.772	1.09E-03	0.763	0.446	0.677	9.21E-07	1.00E-06
CrohnsDz	1.103	0.696	0.882	-8.98E-04	-0.664	0.507	0.714	-1.13E-06	1.65E-06
Resp_ENT_Dz	1.102	0.498	0.836	-5.65E-04	-1.017	0.309	0.592	-2.25E-07	2.21E-07
Height	1.028	0.272	0.918	-2.00E-04	-0.372	0.710	0.832	-2.92E-07	7.65E-07
NumChildrenBorn	0.960	0.820	0.961	1.36E-03	0.723	0.470	0.688	4.14E-07	5.23E-07
UC	0.933	0.760	0.929	-2.96E-04	-0.291	0.771	0.878	-2.91E-07	1.03E-06
YearsOfEd	0.915	0.236	0.718	4.35E-04	0.551	0.581	0.722	2.78E-07	4.99E-07
BMI	0.908	0.248	0.711	-6.41E-04	-1.439	0.150	0.592	-5.74E-07	4.05E-07
MenarcheAge	0.888	0.251	0.657	7.30E-04	0.913	0.361	0.592	6.33E-07	6.54E-07
Schizophrenia	0.888	0.440	0.798	-2.72E-03	-3.438	0.001	0.012	-3.22E-06	8.58E-07
RBCCount	0.878	0.310	0.692	1.48E-03	1.363	0.173	0.592	1.33E-06	1.13E-06
SystolicBP	0.867	0.233	0.568	-9.65E-04	-1.770	0.077	0.592	-7.74E-07	4.38E-07
ASD	0.835	0.934	0.860	3.31E-03	1.597	0.110	0.592	2.92E-06	1.86E-06
Hypothyroidism	0.804	0.406	0.631	1.09E-03	1.464	0.143	0.592	4.29E-07	2.85E-07
RA	0.683	0.906	0.725	1.35E-03	1.427	0.154	0.592	1.04E-06	7.07E-07
PlateletCount	0.683	0.290	0.281	-5.65E-04	-1.209	0.226	0.592	-5.96E-07	5.57E-07
Neuroticism	0.535	0.508	0.373	1.25E-03	0.926	0.354	0.592	7.25E-07	7.27E-07
FirstBirthAge	0.477	0.470	0.270	-9.79E-04	-0.914	0.361	0.592	-3.77E-07	4.02E-07
LDL	0.314	0.914	0.456	-6.33E-04	-0.476	0.634	0.764	-2.93E-07	6.72E-07
DepressiveSxs	-0.027	1.042	0.336	-1.08E-03	-1.014	0.311	0.592	-4.00E-07	4.36E-07
HDL	-0.177	0.706	0.102	1.18E-03	1.039	0.299	0.592	5.58E-07	5.65E-07
HighCholesterol	-0.322	0.457	0.006	-5.09E-04	-0.919	0.358	0.592	-1.96E-07	2.19E-07
Anorexia	-0.956	1.100	0.085	-9.30E-03	-4.892	1.00E-06	4.1E-05	-5.21E-06	1.44E-06

Table 6.6: Partitioned heritability and direction of effect results for 41 representative traits. For 41 traits, we calculated partitioned heritability and direction of effect for the Altai-matching introgressed variants (Set 1, 2.4:Methods). The first set of columns describes the partitioned heritability results calculated with S-LDSC (enrichment, standard error [SE], and *P*-value). Enrichments above 1 indicate depletion. The second set of columns describes the direction of effect results calculated with SLDP (functional correlation [r_f], Z-score, corresponding *P*-value, mu, and mu standard error [see 2.4:Methods]). *P*-values are calculated from an empirical null distribution described in Reshef et al.[160]. *q*-values are corrected for multiple comparisons using the Benjamini-Hochberg FDR-correction at the 0.05 level. Positive Functional correlations and Z-scores indicate a positive relationship with the trait in introgressed variants, whereas negative values indicate a negative relationship with the trait (all with reference to the coding of the GWAS).

6.2 Appendix 2: Supporting information for Chapter 3

Supplemental Text

TAD maps and length

TAD maps for 37 different cell types were obtained from the 3D genome browser (Table S1). All cell types were available in hg19 format, except the liver data, which we downloaded in hg38 and used the UCSC liftOver tool to convert to hg19[179,260].The median TAD length across all cell types is 1.15 Mb (IQR: 0.71 - 1.82 Mb) and the median number of TADs per cell type is 1844 (IQR: 1625 - 2277). We observed an inverse relationship between TAD length and number of TADs in a cell type: cells with longer TADs have fewer TADs (Fig. S17). Primary tissues have longer TADs, whereas naïve cell types like stem cells and de-differentiated leukemia cell-lines have shorter TADs (Fig. S17). This is consistent with previous examination of neuronal development which found that, during differentiation, TAD number decreases with a corresponding increase in size[129].

Similarity between TAD maps

Our finding of TAD map similarity among functionally similar cell types contrasts with previous work by Sauerwald et al. (2018)[190] that found that most similar TAD map pairs have no biological connection; however, they investigate a different set of cells (predominantly cancer cell lines) . Comparisons with highly mutated cancer cell lines that may not reflect natural boundary patterns. Both our results and the Sauerwald et al. (2018)[190] comparisons could be influenced by batch effects because the Hi-C data considered were generated by different groups. However, an important follow-up by Sauerwald et al. (2020)[198] finds that lab specific differences have little impact on TAD map similarity comparisons and that cell type is the greater driver of biological variation in TAD structures.

Our similarity quantifications agree with some previous estimates. We find that the median pairwise Jaccard similarity for all 37 x 37 cell type comparisons is 0.18 (IQR: 0.15 - 0.23), 0.32 (IQR: 0.26 - 0.37), 0.41 (IQR: 0.35 - 0.47) at 40 kb, 100 kb, and 200 kb resolution, respectively. Our pairwise Jaccard similarity between 200 kb boundaries (0.41) aligns with previous analyses that examined cell type TAD map similarity among larger windows have reported similarity coefficients between 0.4 - 0.5[190,198]. At a finer resolution, Rao et al.[117] reported Jaccard indices from 0.21 - 0.30 for comparisons of GM12878 to each of IMR90, HMEC, HUVEC, K562, KBM7 and NHEK[117,190]. The Jaccard similarity for our comparisons of these cell types is 0.24 - 0.37 (40 kb resolution).

Overall, this variability in TAD similarity across different cell types highlights the sensitivity of stability comparisons to the definition of TAD boundaries used. For example, the median pairwise Jaccard similarity between 40 kb boundaries across 21 tissues defined by Schmitt et al.[118] is 0.106 (IQR: 0.086 - 0.123). However, they collapsed boundaries to 200 kb "boundary regions" to conclude that TAD boundaries are highly stable (stating that over 35% of TAD boundaries are present in 21 of 21 tissues). These previous studies often investigated more homogenous groups of cell types which could lead to higher estimates of stability. Ultimately, we stress than when interpreting claims of similarity between TAD maps of different cell types, the method of defining TADs (versus loop domains or boundary "regions"), the genomic resolution, and the breadth of cell types considered should be considered for context.

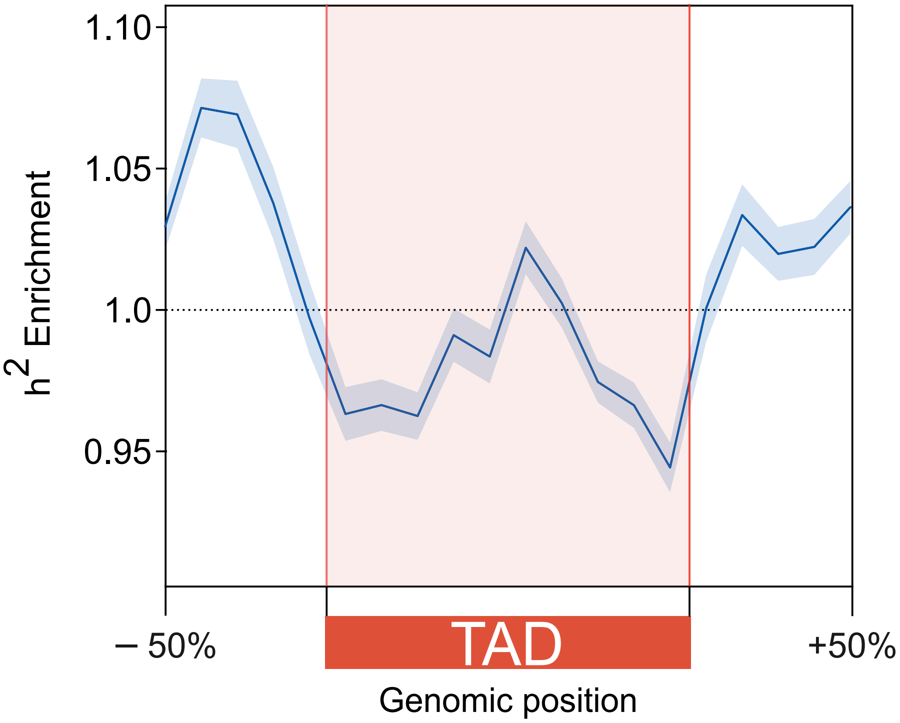


Figure 6.17: Meta-analysis of heritability patterns across cell types yields similar results to averaging. TADs across 37 cell types, heritability is enriched near regions flanking TADs when meta-analyzed across 41 common complex phenotypes. When combining data across traits, the heritability enrichment results are consistent using random-effects meta-analysis model (here) versus averaging ($r^2 = 0.85$, $P = 7 \times 10^{-9}$), Fig. 3.2A). The error band signifies a 99% confidence interval.

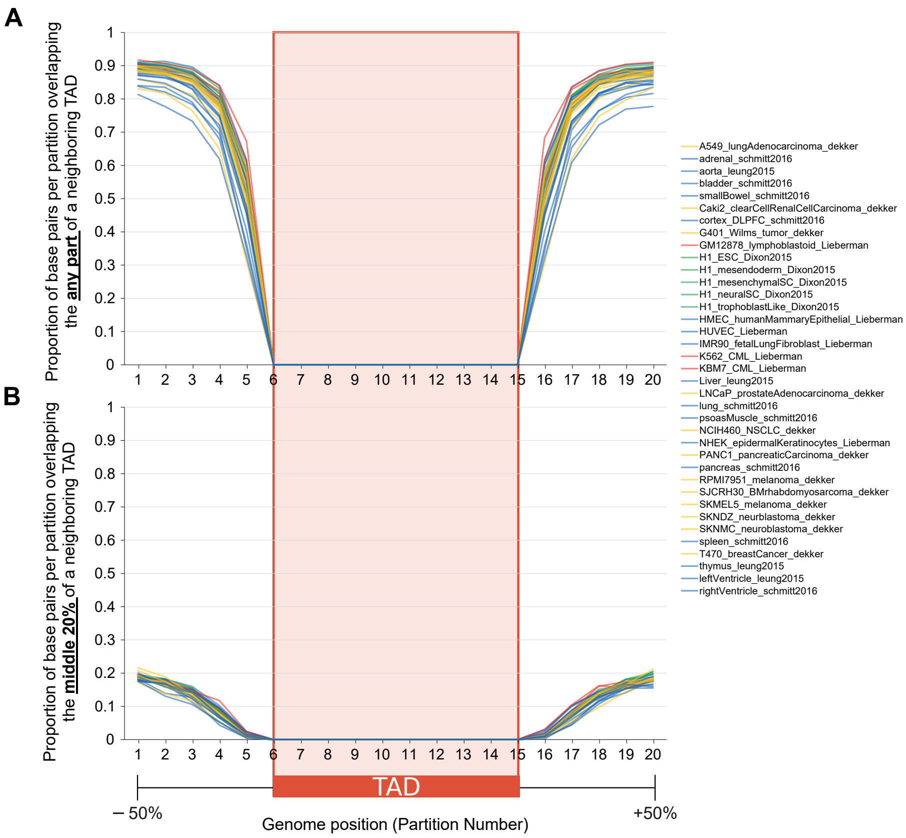


Figure 6.18: Overlap between region flanking TADs and neighboring TADs.
In Fig. 3.2 and 3.4A-C we analyzed TADs plus 50% of their total length on each side and subdivided this region into 20 equal-sized partitions. Bins 1-5 and 16-20 "bookend" the TAD, while the center bins 6-15 are inside the TAD. Because TADs are often adjacent, we quantify how often the ±50% region flanking the TAD (bins 1-5,16-20) overlaps a neighboring TAD. Per partition across the TAD landscape (x-axis) we calculate the proportion of bases that overlap **(A)** any part of a neighboring TAD and **(B)** the middle 20% of a neighboring TAD. A higher proportion of the partitions further from the edge of the TAD overlap a neighboring TAD, as expected. At the bin farthest from the TAD (bins 1 and 20), 80-90% extend into a neighboring TAD. However, less than 20% extend into the center of a neighboring TAD.

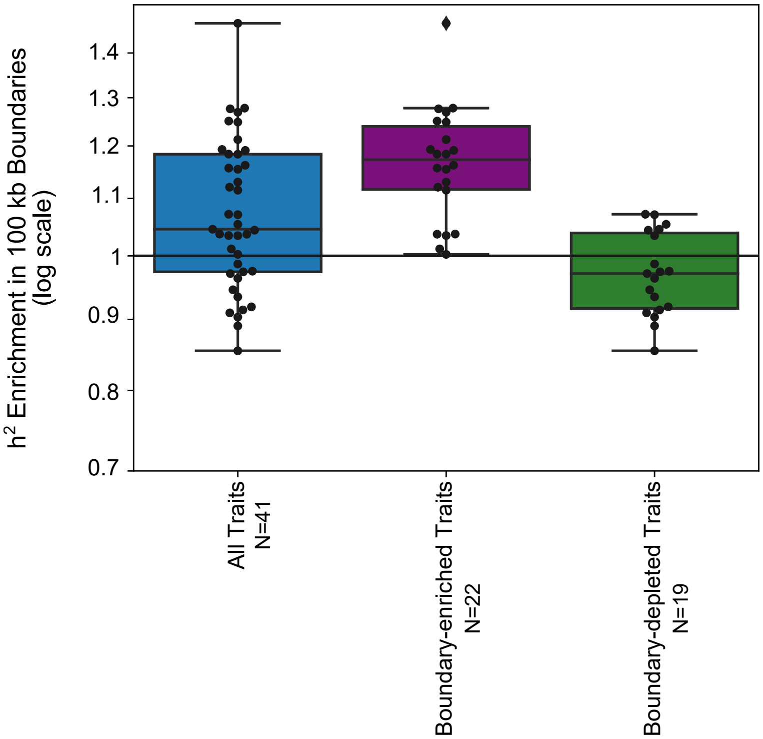


Figure 6.19: TAD boundaries are enriched for heritability.

When defining TAD boundaries as the 100 kb region flanking TADs, boundaries are generally enriched for heritability across 41 common complex traits (blue box, 1.07x, $P = 0.001$). These are the same data shown in Fig. 3.3C; however, the boundaries are not stratified by their stability across cell types. When we split the traits into the clusters defined in Fig. 3.4, Boundary-enriched traits are further enriched for trait heritability (purple box, 1.16x, $P = 1 \times 10^{-7}$) while Boundary-depleted traits show no significant enrichment (green box, 0.97x, $P = 0.06$). These are the same data shown in Fig. 3.4 and 3.4F, respectively, without stratification by stability across cell types. These findings are consistent with the heritability patterns across the TAD landscapes shown in Fig. 3.2A, 3.4B-C, but with fixed-window 100 kb boundary definitions.

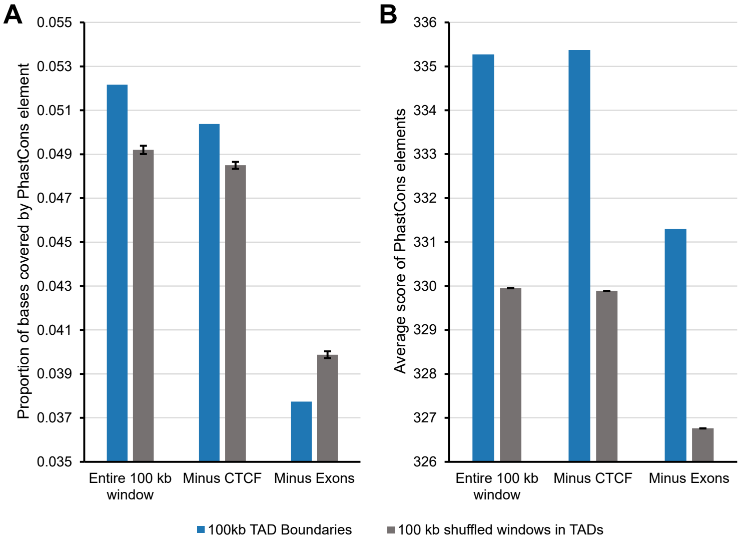


Figure 6.20: TAD boundaries are more conserved than windows inside TADs.
We quantified evolutionary sequence conservation in terms of **(A)** the proportion of base pairs in a region overlapping a conserved element identified by PhastCons and **(B)** by the element-wise average PhastCons conservation score across the region. Using these two measures we compared base pair level conservation in 100 kb TAD boundaries (blue) and matched 100 kb windows shuffled inside TADs ($n = 111$, gray). When considering the entire 100 kb window, TAD boundaries have more overlap with PhastCons elements and a higher average PhastCons element score than windows in TADs (left bars). When considering the 100 kb windows with CTCF ChIP-seq peaks removed, TAD boundaries still have more overlap and higher score than windows in TADs (middle bars). When considering the 100 kb windows with all exons removed, TAD boundaries have less overlap with PhastCons elements, but the remaining PhastCons elements still have a higher conservation score (right bars).

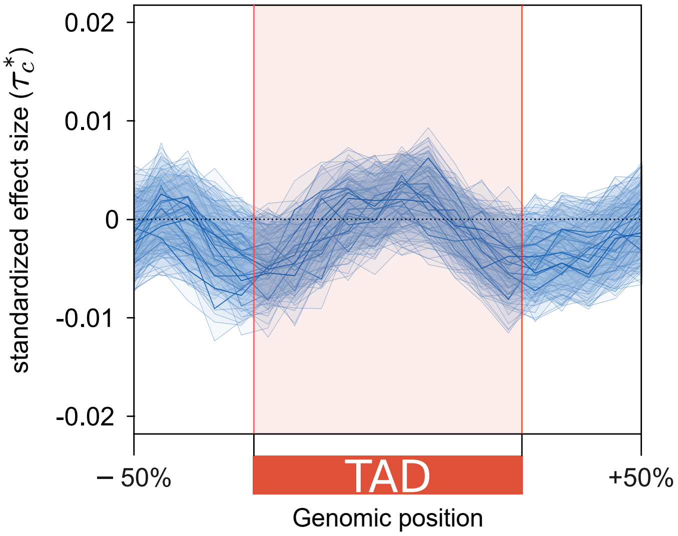


Figure 6.21: Trait heritability conditioned on 86 annotations.
In contrast to heritability enrichment, the standardized effect size (τ_c^*) quantifies effects that are unique to the focal annotation compared to a set of other 86 annotations (e.g. regulatory annotations, evolutionary conservation, coding regions, LD, minor allele frequency). When meta-analyzed across all traits, the standardized effect sizes for partitions across the 3D genome are non-significant compared to the unconditioned enrichment analyses (Fig. 3.2). This indicates that enrichment for these known annotations (e.g., CTCF binding sites and genes) across partitions explains much of the observed heritability enrichment for regions flanking TADs. Each line represents the standardized effect size meta-analyzed across all traits for that cell type ($n = 37$). The error bands signify 99% confidence intervals.

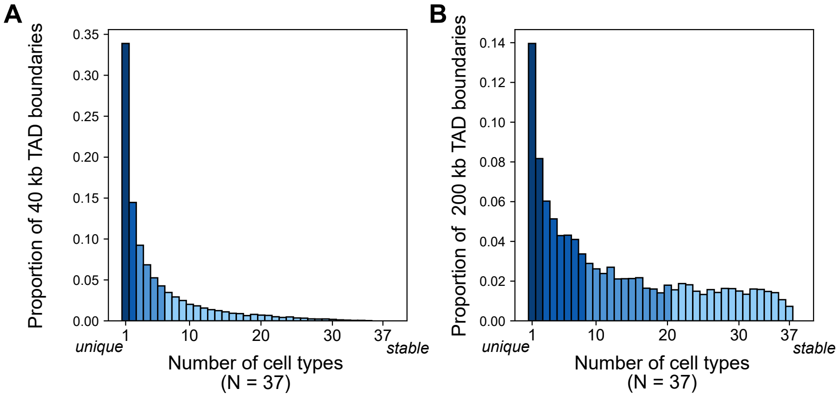


Figure 6.22: Histograms of boundary stability based on alternate definitions of TAD boundaries.
Histograms of TAD boundaries by the number of cell types they are observed in (their "stability") colored by quartiles. In addition to the 100 kb bookend boundary definitions (Fig. 3.3B), our supplemental analysis investigates **(A)** 40 kb centered boundaries and **(B)** 200 kb bookend boundaries. Using the 40 kb definition, 33.9% of boundaries are unique to a single context and 2.0% of boundaries are observed in 25+ of 37 cell types. Using the 200 kb definition, 14.0% of boundaries are unique to a single context and 18.3% of boundaries are observed in 25+ of 37 cell types.

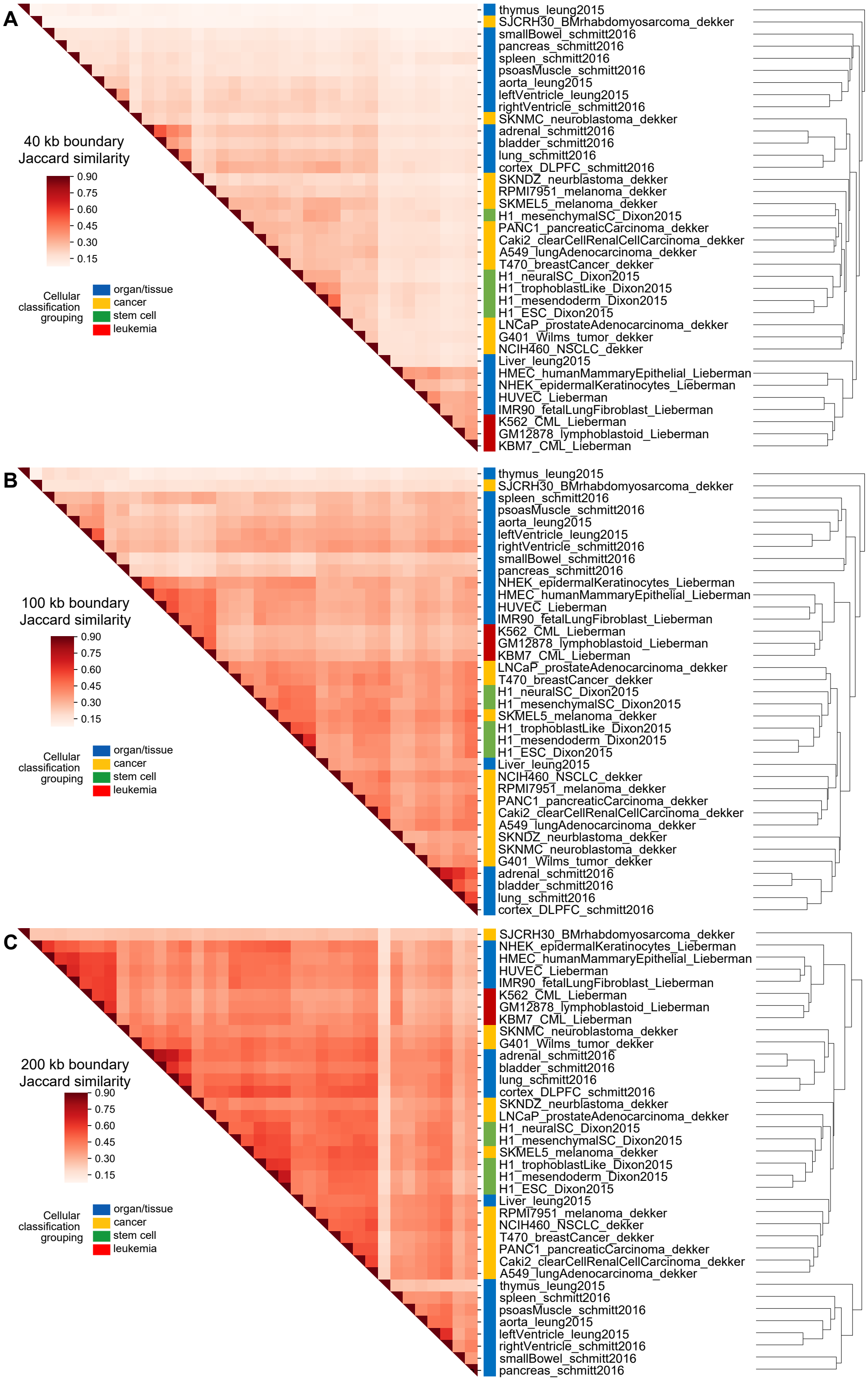


Figure 6.23: Biologically similar cell types cluster by TAD map similarity.
Clustering for 37 cell types using the pairwise Jaccard similarity metric with colors labelling cellular groups for **(A)** 40 kb boundaries, **(B)** 100 kb boundaries, and **(C)** 200 kb boundaries.

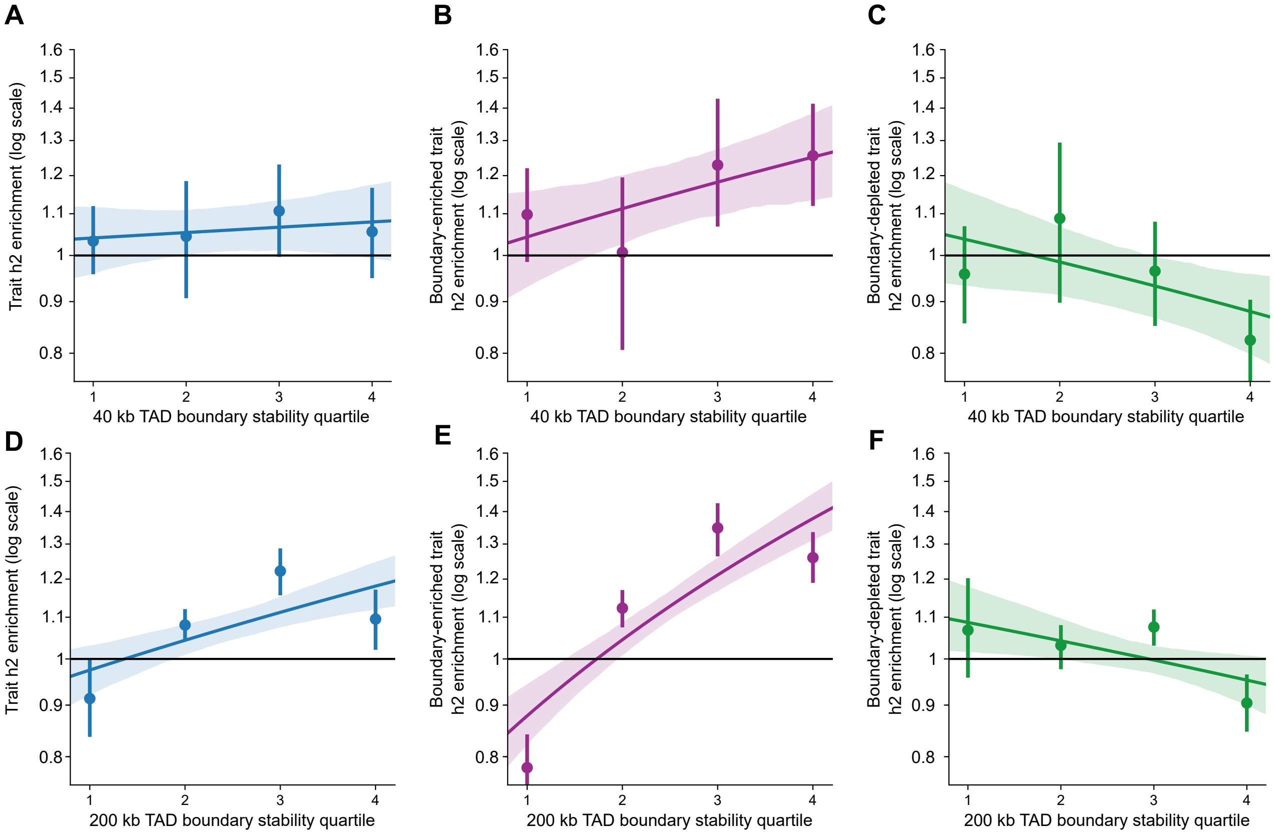


Figure 6.24: Relationship between heritability enrichment and boundary stability is robust to different boundary definitions.

Over all traits, there is a positive relationship between boundary stability and heritability enrichment using 40 kb boundaries (**A**, $P = 0.61$), 100 kb boundaries (Fig. 3.3C, P = 0.006), and 200 kb boundaries (**D**, $P = 2 \times 10^{-5}$). For traits in the boundary-enriched cluster (Fig. 3.4B), there is a stronger positive relationship between boundary stability and heritability in 40 kb boundaries (**B**, $P = 0.06$), 100 kb boundaries (Fig. 3.4D, $P = 2 \times 10^{-6}$), and 200 kb boundaries (**E**, $P = 3 \times 10^{-14}$). For traits in the boundary-depleted cluster (Fig. 3.4C), there is a weak negative relationship between boundary stability and heritability using 40 kb boundaries (**C**, $P = 0.09$), 100 kb boundaries (Fig. 3.4F, $P = 0.09$), and 200 kb boundaries (**F**, $P = 0.01$). Error bars/bands signify 95% confidence intervals.

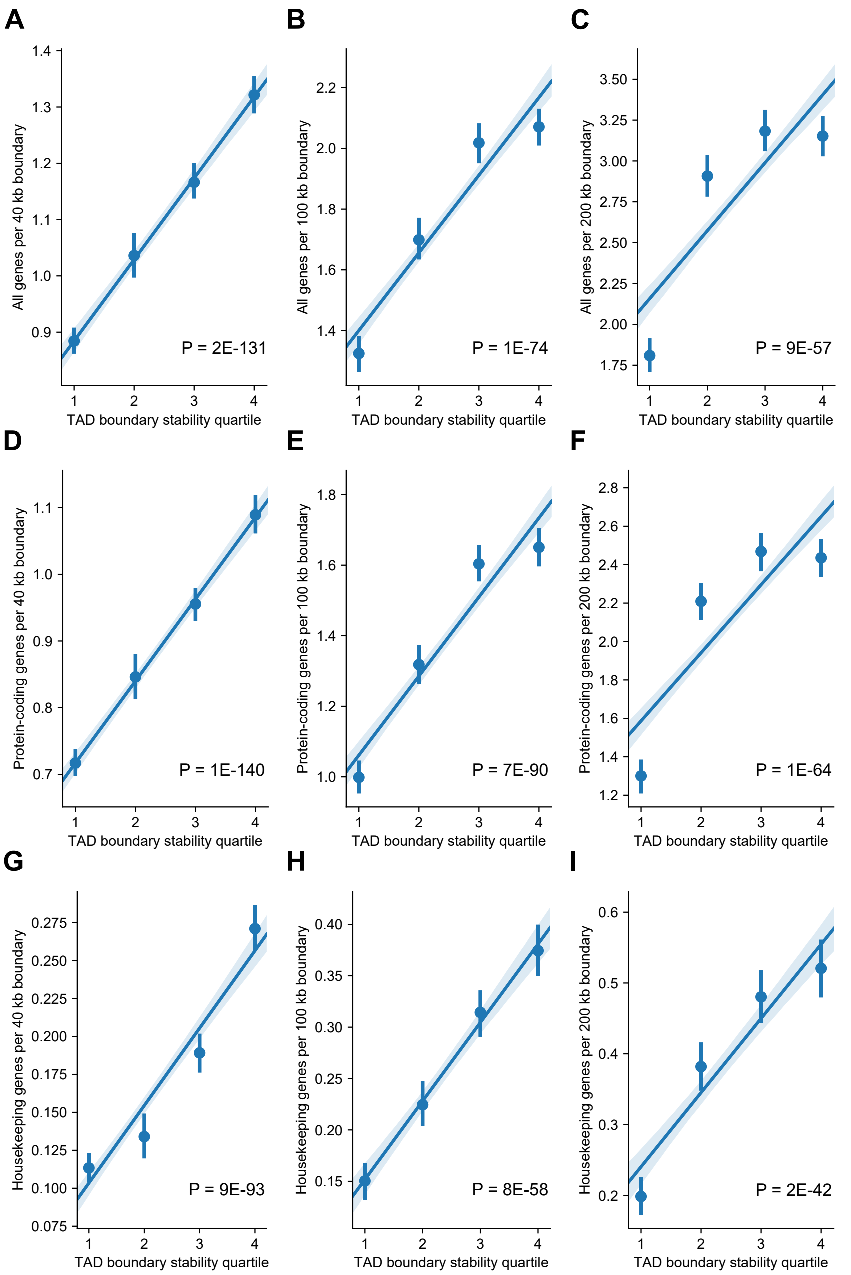


Figure 6.25: The enrichment of stable TAD boundaries for genes is robust to gene set and boundary definitions. The relationship between increased TAD boundary stability and gene overlap using 40 kb boundaries **(A,D,G)**, 100 kb boundaries **(B,E,H)**, and 200 kb boundaries **(C,F,I)**. We also demonstrate this trend using three types of genes: all RefSeq genes **(A-C)**, protein-coding genes **(D-F)**, and housekeeping genes **(G-I)**. Panel H is shown in the main text (Fig. 3.3F). TAD boundary stability quartiles are defined by the empirical distributions shown in Fig. 6.22A (40 kb), Fig. 3.3B (100 kb), and Fig. 6.22B (200 kb). Boundaries in the first quartile are unique to a single cell type, while boundaries in higher quartiles are stable across multiple cell types. Error bars/bands signify 95% confidence intervals.

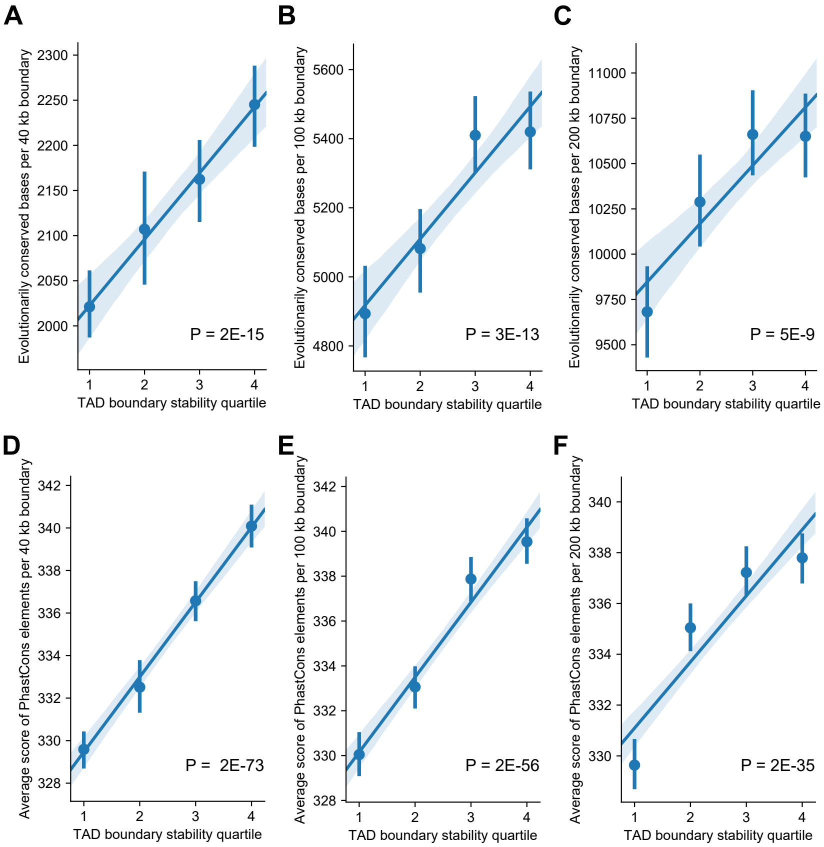


Figure 6.26: The enrichment of stable TAD boundaries for sequence-level conservation is robust to boundary definitions.

The relationship between increased TAD boundary stability and sequence-level conservation quantified (via PhastCons element overlap) considering 40 kb boundaries (**A & D**), 100 kb boundaries **(B & E)**, and 200 kb boundaries **(C & F)**. We also demonstrate this trend holds with two different measures of evolutionary conservation: number of bases overlapping PhastCons elements **(A-C)** and average PhastCons element score per boundary **(D-F)**. Panel B is shown in the main text (Fig. 3.3D). TAD boundary stability quartiles are defined by the empirical distributions shown in Fig. 6.22A (40 kb), Fig. 3.3B (100 kb), and Fig. 6.22B (200 kb). Boundaries in the first quartile are unique to a single cell type, while boundaries in higher quartiles are stable across multiple cell types. Error bars/bands signify 95% confidence intervals.

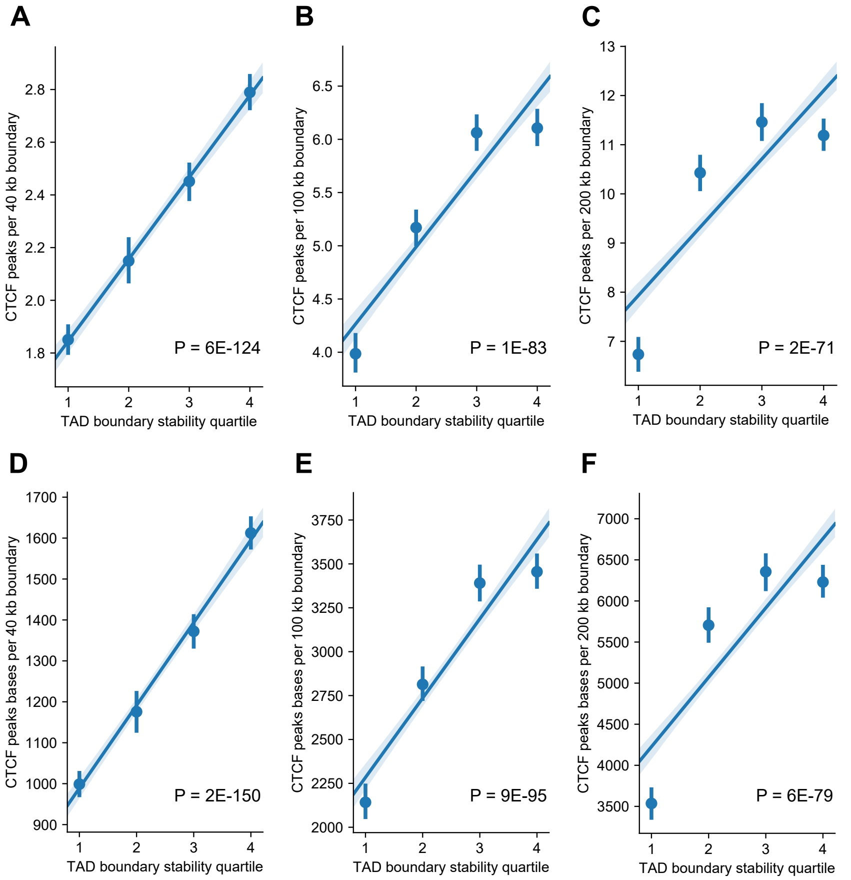


Figure 6.27: The enrichment of stable TAD boundaries for CTCF binding is robust to boundary definitions. The relationship between increased TAD boundary stability and CTCF binding considering 40 kb boundaries **(A & D)**, 100 kb boundaries **(B & E)**, and 200 kb boundaries **(C & F)**. We also demonstrate this trend holds with two different quantifications of CTCF overlap: count of CTCF ChIP-seq peaks per boundary **(A-C)** and number of CTCF ChIP-seq peak bases overlapping each boundary **(D-F)**. Panel B is shown in the main text (Fig. 3.3E). TAD boundary stability quartiles are defined by the empirical distributions shown in Fig. 6.22A (40 kb), Fig. 3.3B (100 kb), and Fig. 6.22B (200 kb). Boundaries in the first quartile are unique to a single cell type, while boundaries in higher quartiles are stable across multiple cell types. Error bars/bands signify 95% confidence intervals.

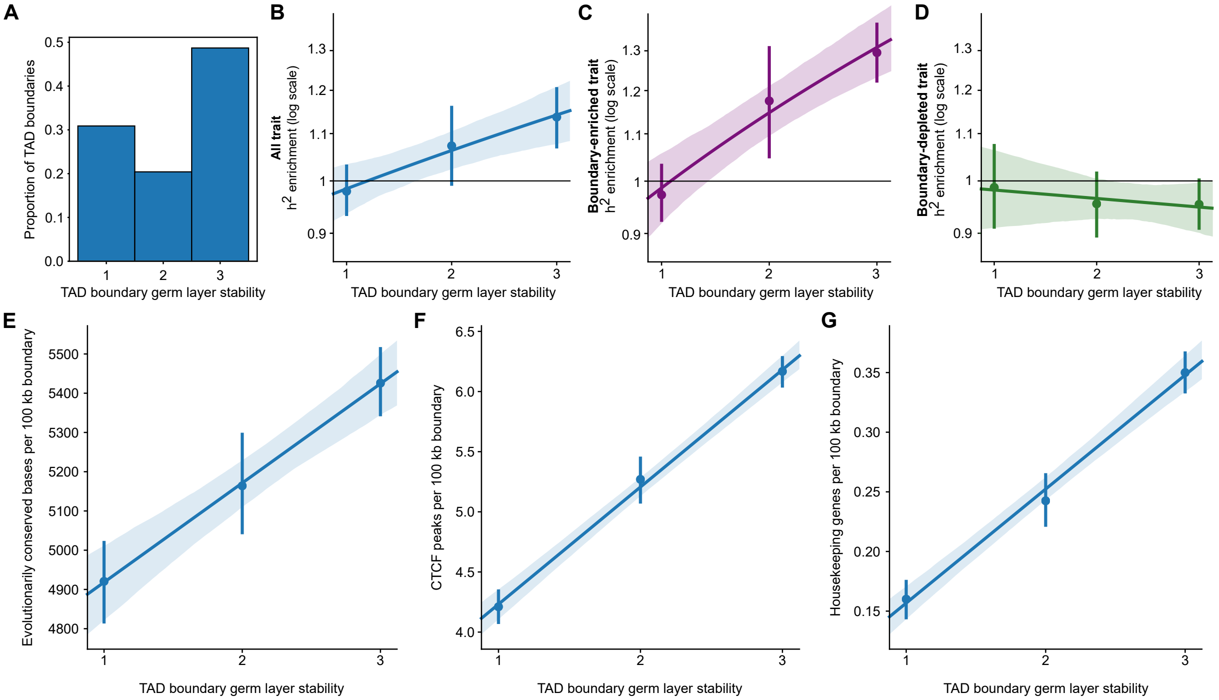


Figure 6.28: Heritability enrichment and conservation at TAD boundaries stable across cell types replicates using a germ-layer-informed measure of stability.
Of the 37 cell types considered, some are more closely related than others, therefore we grouped 34 of them by germ layer (endoderm [$N = 12$], mesoderm [$N = 13$], ectoderm [$N = 9$]; Table S1). We then quantified stability based on whether the boundary was found in one, two, or all three germ layers. **(A)** The proportion of 100 kb boundaries that fall into each stability measurement. For example, if a boundary was found in muscle, spleen, and mesenchymal stem cells, but no other tissues, it is a "mesoderm-only" boundary and in the "1" category for germ layer stability. If a boundary was found in muscle, cortex, and lung, it is a boundary found across all three germ layers and in the "3" category for germ layer stability. These examples were assigned the same level of stability in the raw cell type count measure because they are both present in 3/37 cell types (Fig. 3.3, 3.4D, and 3.4F). Increased stability using this germ layer informed measure is correlated with increased: **(B)** complex trait heritability enrichment ($P = 0.002$), **(E)** conserved bases (overlap with PhastCons elements, $P = 2 \times 10^{-14}$), **(F)** CTCF binding (overlap with ChIP-seq peaks, $P = 3 \times 10^{-97}$), and **(G)** housekeeping genes ($P = 3 \times 10^{-58}$). When we split the traits into the clusters defined in Fig. 3.4, **(C)** the positive correlation between boundary stability and trait heritability is even stronger for the subset of traits in the boundary-enriched cluster ($P = 2 \times 10^{-5}$), while **(D)** the boundary-depleted traits show no significant trend between boundary stability and trait heritability ($P = 0.49$). Respectively, these replicate the results in Figs.. 3.3C-F, 3.4D, and 3.4F with the germ-layer stability measurement. All error bars/bands signify 95% confidence intervals.

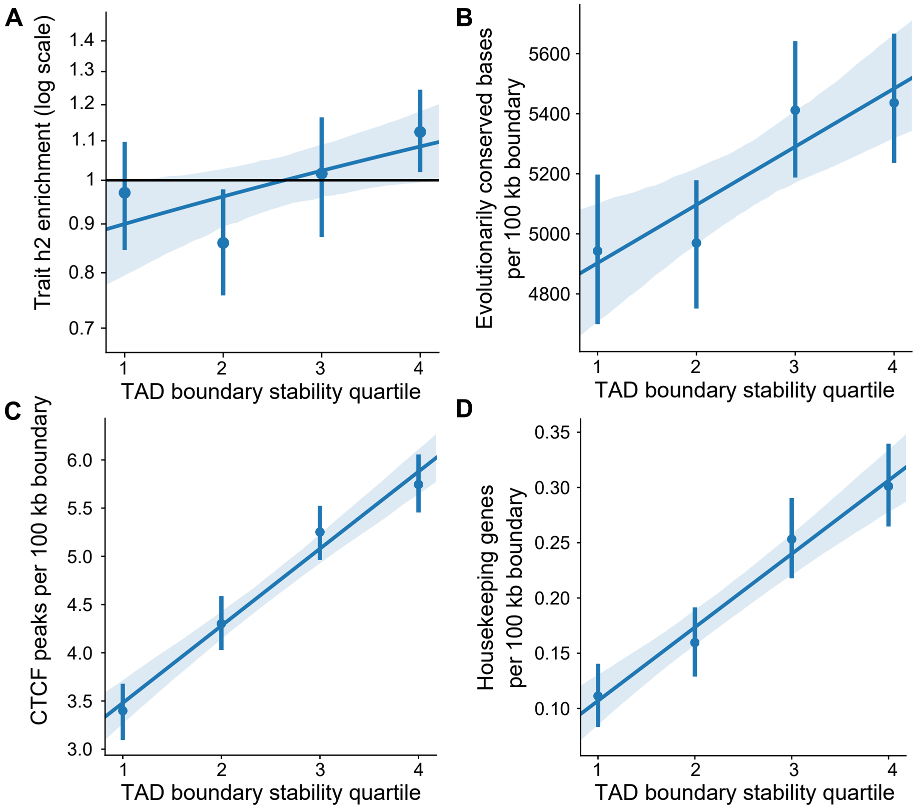


Figure 6.29: Removing boundaries near genomic gaps or blacklist regions increases the correlations between stability and functional attributes.
In Figs. 3.3C-F we note that there is a positive trend between TAD boundary stability quartile and functional annotation; however, we find that the fourth quartile “drops-off” and has equal or slightly lower enrichment compared to the third quartile. We hypothesize that this trend is, in part, due to technical factors. For example, TADs must be called at the starts and ends of chromosomes, centromeres, and assembly gaps in all tissues. This may create highly stable TAD boundaries independent of their functional significance. To test this, we apply a conservative filter and remove all boundaries within 5 MB of a genomic gap or blacklist region. Across TAD boundary stability quartiles, we replicate the correlation between increased cell type stability and increased **(A)** complex trait heritability enrichment ($P = 0.03$), **(B)** conserved bases (overlap with PhastCons elements, $P = 0.0002$), **(C)** CTCF binding (overlap with ChIP-seq peaks, $P = 1 \times 10^{-37}$), and **(D)** housekeeping genes ($P = 1 \times 10^{-18}$). The enrichment “drop-off” is reduced or absent in the relationship with heritability, CTCF, and genes suggesting that technical bias partially contributes to a drop-off of enrichment in the fourth quartile. All error bars/bands signify 95% confidence intervals.

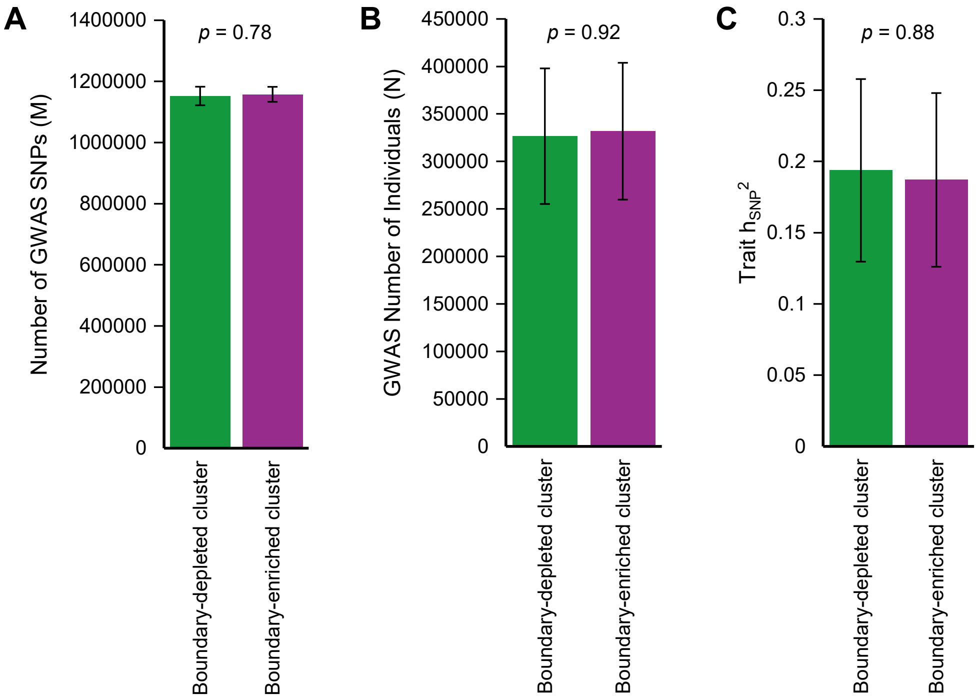


Figure 6.30: Traits in the boundary-depleted cluster and boundary-enriched cluster do not differ in GWAS parameters.

(A) Number of GWAS SNPs ($P = 0.78$, t-test with equal variances), **(B)** Number of individuals in the GWAS ($P = 0.92$), or **(C)** SNP-based heritability ($P = 0.88$). Error bars signify 95% confidence intervals.

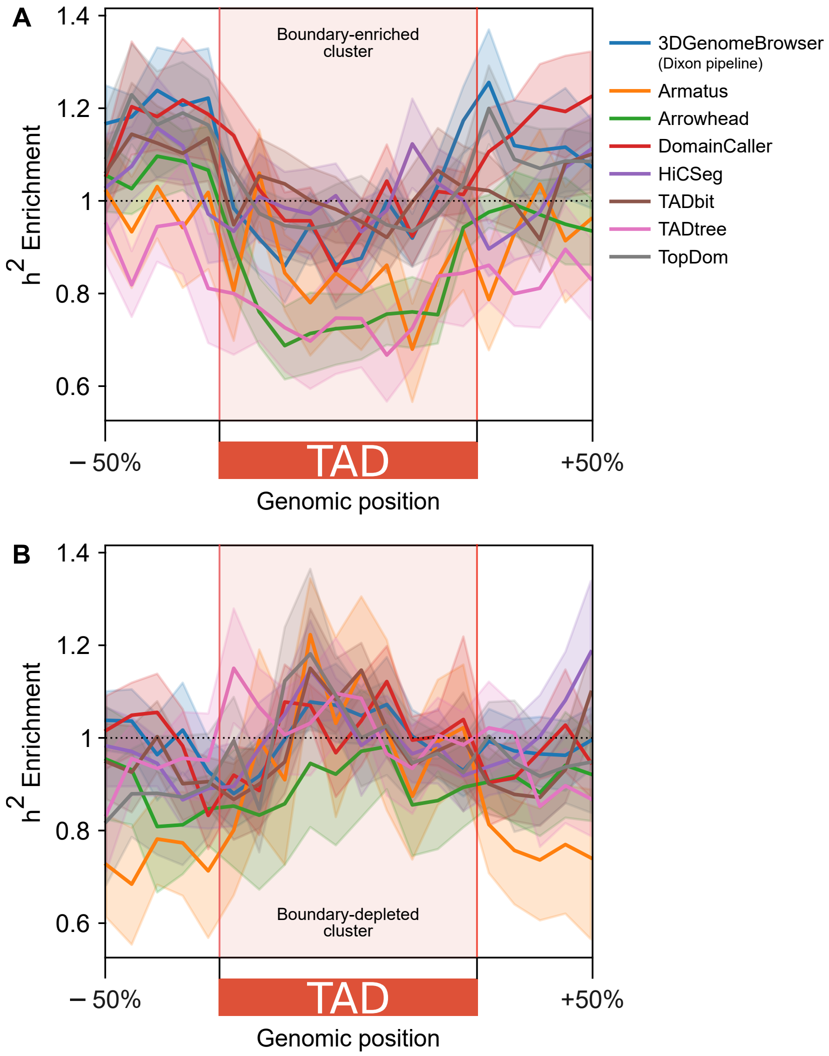


Figure 6.31: Patterns of heritability enrichment across the 3D genome in human embryonic stem cells (ESC) are robust to the TAD calling algorithm used.

(A) Heritability enrichment landscape over TADs in ESCs called by eight different algorithms for traits in the boundary-enriched cluster. Similar to the results shown in Fig. 3.4B (which use TADs from the Dixon pipeline), regions flanking TADs are enriched for heritability compared to TADs. **(B)** Heritability enrichment landscape over TADs in ESCs for traits in the boundary-depleted cluster. Similar to the results shown in Fig. 3.4C (which use TADs from the Dixon pipeline), TADs are centrally enriched for heritability. Error bands signify 95% confidence intervals.

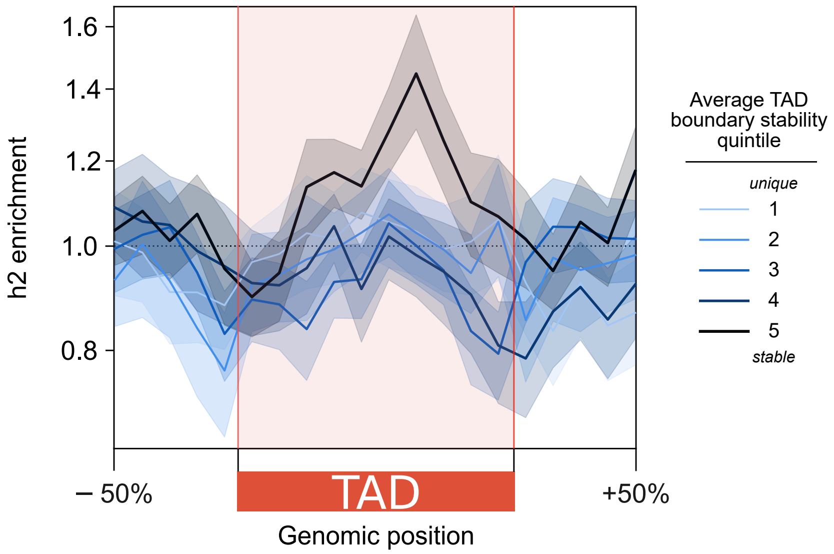


Figure 6.32: Among boundary-depleted traits, stable boundaries associate with stronger heritability enrichment in TAD centers.
For the boundary-depleted cluster traits, TADs flanked by the most stable boundaries (measured by taking the average stability of its two boundaries and binning into quintiles) have increased heritability in the TAD center. This analysis was performed in a random subset of 7 cell types (aorta, H1_ESC, leftVentricle, Liver, psoasMuscle, SKNDZ, T470). Error bands signify 95% confidence intervals.

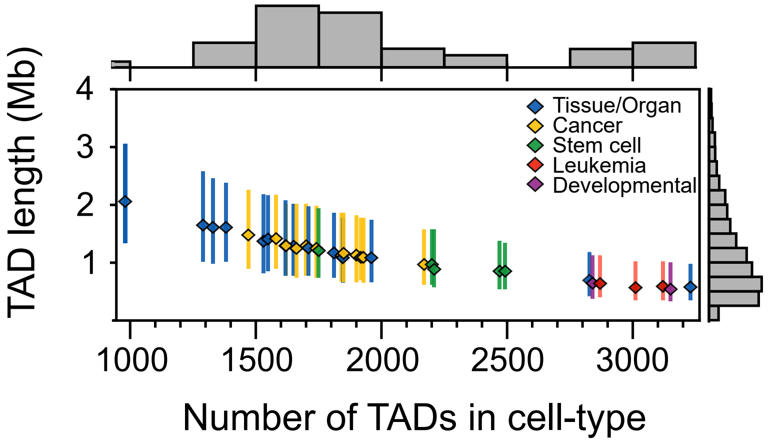


Figure 6.33: Average TAD length in a cell type negatively correlates with number of TADs.
Across 37 cell types, there is an inverse relationship between TAD length and number of TADs. Organ/tissue cell types generally have the longest (and fewest) TADs. Leukemia and stem cells have the shortest (and most) TADs. Error bands signify the IQR.

FileNameFrom3DGenomeBrowser	CellTypeDescription	Abbreviation	BiologicalCluster	GermLayer	Citation
A549_raw-rep1_TADs.txt	A549_lungAdenocarcinoma_dekker	A549	cancer	endoderm	Lajoie, Dekker et al. (2015), ENCODE
AdrenalGland_Donor-AD2-raw_TADs.txt	adrenal_schmitt2016	adrenal	organ/tissue	endoderm	Schmitt et al. (2016)
Aorta_STL002_Leung2015-raw_TADs.txt	aorta_leung2015	aorta	organ/tissue	mesoderm	Leung et al. (2015)
Bladder_Donor-BL1-raw_TADs.txt	bladder_schmitt2016	bladder	organ/tissue	endoderm	Schmitt et al. (2016)
Bowel_Small_Donor-SB2-raw_TADs.txt	smallBowel_schmitt2016	smallBowel	organ/tissue	endoderm	Schmitt et al. (2016)
Caki2_raw-rep1_TADs.txt	Caki2_clearCellRenalCellCarcinoma_dekker	Caki2	cancer	mesoderm	Lajoie, Dekker et al. (2015), ENCODE
Cortex_DLPFC_Donor-CO-raw_TADs.txt	cortex_DLPFC_schmitt2016	DLPFC	organ/tissue	ectoderm	Schmitt et al. (2016)
G401_raw-rep1_TADs.txt	G401_Wilms_tumor_dekker	G401	cancer	mesoderm	Lajoie, Dekker et al. (2015), ENCODE
GM12878_Lieberman-raw_TADs.txt	GM12878_lymphoblastoid_Lieberman	GM12878	leukemia	mesoderm	Rao et al. (2014)
H1-ESC_Dixon2015-raw_TADs.txt	H1_ESC_Dixon2015	ESC	stem cell	NA	Dixon et al. (2015)
H1-MES_Dixon2015-raw_TADs.txt	H1_mesendoderm_Dixon2015	MES	stem cell	NA	Dixon et al. (2015)
H1-MSC_Dixon2015-raw_TADs.txt	H1_mesenchymalSC_Dixon2015	MSC	stem cell	mesoderm	Dixon et al. (2015)
H1-NPC_Dixon2015-raw_TADs.txt	H1_neuralSC_Dixon2015	NPC	stem cell	ectoderm	Dixon et al. (2015)
H1-TRO_Dixon2015-raw_TADs.txt	H1_trophoblastLike_Dixon2015	TRO	stem cell	NA	Dixon et al. (2015)
HMEC_Lieberman-raw_TADs.txt	HMEC_humanMammaryEpithelial_Lieberman	HMEC	organ/tissue	ectoderm	Rao et al. (2014)
HUVEC_Lieberman-raw_TADs.txt	HUVEC_Lieberman	HUVEC	organ/tissue	mesoderm	Rao et al. (2014)
IMR90_Lieberman-raw_TADs.txt	IMR90_fetalLungFibroblast_Lieberman	IMR90	organ/tissue	endoderm	Rao et al. (2014)
K562_Lieberman-raw_TADs.txt	K562_CML_Lieberman	K562	leukemia	mesoderm	Rao et al. (2014)
KBM7_Lieberman-raw_TADs.txt	KBM7_CML_Lieberman	KBM7	leukemia	mesoderm	Rao et al. (2014)
Liver_STL011_Leung_2015-raw_TADs_hg19From38.txt	Liver_leung2015	Liver	organ/tissue	endoderm	Leung et al. (2015)
LNCaP_raw-rep1_TADs.txt	LNCaP_prostateAdenocarcinoma_dekker	LNCaP	cancer	endoderm	Lajoie, Dekker et al. (2015), ENCODE
Lung_Donor-LG1-raw_TADs.txt	lung_schmitt2016	lung	organ/tissue	endoderm	Schmitt et al. (2016)
Muscle_Psoas_Donor-PO1-raw_TADs.txt	psoasMuscle_schmitt2016	psoas	organ/tissue	mesoderm	Schmitt et al. (2016)
NCIH460_raw-rep1_TADs.txt	NCIH460_NSCLC_dekker	NCIH460	cancer	endoderm	Lajoie, Dekker et al. (2015), ENCODE
NHEK_Lieberman-raw_TADs.txt	NHEK_epidermalKeratinocytes_Lieberman	NHEK	organ/tissue	ectoderm	Rao et al. (2014)
PANC1_raw-rep1_TADs.txt	PANC1_pancreaticCarcinoma_dekker	PANC1	cancer	endoderm	Lajoie, Dekker et al. (2015), ENCODE
Pancreas_Donor-PA2-raw_TADs.txt	pancreas_schmitt2016	pancreas	organ/tissue	endoderm	Schmitt et al. (2016)
RPMI7951_raw-rep1_TADs.txt	RPMI7951_melanoma_dekker	RPMI7951	cancer	ectoderm	Lajoie, Dekker et al. (2015), ENCODE
SJCRH30_raw-rep1_TADs.txt	SJCRH30_BMrhabdomyosarcoma_dekker	SJCRH30	cancer	mesoderm	Lajoie, Dekker et al. (2015), ENCODE
SKMEL5_raw-rep1_TADs.txt	SKMEL5_melanoma_dekker	SKMEL5	cancer	ectoderm	Lajoie, Dekker et al. (2015), ENCODE
SKNDZ_raw-rep1_TADs.txt	SKNDZ_neurblastoma_dekker	SKNDZ	cancer	ectoderm	Lajoie, Dekker et al. (2015), ENCODE
SKNMC_raw-rep1_TADs.txt	SKNMC_neuroblastoma_dekker	SKNMC	cancer	ectoderm	Lajoie, Dekker et al. (2015), ENCODE
Spleen_Donor-PX1-raw_TADs.txt	spleen_schmitt2016	spleen	organ/tissue	mesoderm	Schmitt et al. (2016)
T470_raw-rep1_TADs.txt	T470_breastCancer_dekker	T470	cancer	ectoderm	Lajoie, Dekker et al. (2015), ENCODE
Thymus_STL001_Leung2015-raw_TADs.txt	thymus_leung2015	thymus	organ/tissue	endoderm	Leung et al. (2015)
VentricleLeft_STL003_Leung2015-raw_TADs.txt	leftVentricle_leung2015	leftVentricle	organ/tissue	mesoderm	Leung et al. (2015)
VentricleRight_Donor-RV3-raw_TADs.txt	rightVentricle_schmitt2016	rightVentricle	organ/tissue	mesoderm	Schmitt et al. (2016)

Table 6.7: Cell types used for all analyses from the 3DGenomeBrowser.

Nickname	Trait	M	h2	h2_SE	N	Phenotypic class	actual cluster	Source
Anorexia	Anorexia	931184	0.2153	0.0169	32143	Neuropsych	Boundary-depleted	Boraska et al. 2014 Mol Psych
ASD	Autism_Spectrum	1173307	0.4607	0.0517	10263	Neuropsych	Boundary-depleted	PGC Cross-Disorder Group, 2013 Lancet
AutoimmuneDz	Auto_Immune_Traits_(Sure)	1187056	0.0068	0.0013	459324	Immunologic	Boundary-enriched	UKBiobank
Balding	Balding_Type_I	1187056	0.2154	0.019	208336	Dermatologic	Boundary-depleted	UKBiobank
BMI	BMI	1187056	0.252	0.0071	457824	Metabolic	Boundary-depleted	UKBiobank
CrohnsDz	Crohn's_Disease	1051514	0.4723	0.0575	20883	Immunologic	Boundary-enriched	Jostins et al., 2012 Nature
DepressiveSxs	Depressive_symptoms	1115393	0.0473	0.0037	161460	Neuropsych	Boundary-depleted	Okbay et al., 2016 Nat Genet
DermDz	Dermatologic_Diseases	1187056	0.0094	0.0014	459324	Dermatologic	Boundary-enriched	UKBiobank
Eczema	Eczema	1187056	0.0675	0.0038	458699	Dermatologic	Boundary-enriched	UKBiobank
EosinophilCount	Eosinophil_Count	1187056	0.1977	0.0143	439938	Hematologic	Boundary-enriched	UKBiobank
FEV1_FVC_Ratio	FEV1-FVC_Ratio	1187056	0.2336	0.0113	371949	Cardiopulmonary	Boundary-enriched	UKBiobank
FirstBirthAge	Age_first_birth	1079424	0.0617	0.0033	222037	Reproductive	Boundary-depleted	Barban et al., 2016 Nat Genet
FVC	Forced_Vital_Capacity_(FVC)	1187056	0.2068	0.0065	371949	Cardiopulmonary	Boundary-enriched	UKBiobank
HairColor	Hair_Color	1187056	0.4523	0.1497	452720	Dermatologic	Boundary-depleted	UKBiobank
HDL	HDL	1019272	0.1362	0.0166	99900	Metabolic	Boundary-enriched	Teslovich et al., 2010 Nature
Heel_T_Score	Heel_T_Score	1187056	0.3628	0.0307	445921	Skeletal	Boundary-depleted	UKBiobank
Height	Height	1187056	0.6034	0.027	458303	Skeletal	Boundary-enriched	UKBiobank
HighCholesterol	High_Cholesterol	1187056	0.0468	0.0039	459324	Metabolic	Boundary-enriched	UKBiobank
Hypothyroidism	Hypothyroidism	1187056	0.0459	0.0037	459324	Metabolic	Boundary-enriched	UKBiobank
LDL	LDL	1017973	0.121	0.0166	95454	Metabolic	Boundary-enriched	Teslovich et al., 2010 Nature
MenarcheAge	Age_at_Menarche	1187056	0.2457	0.0102	242278	Reproductive	Boundary-enriched	UKBiobank
MenopauseAge	Age_at_Menopause	1187056	0.1215	0.0086	143025	Reproductive	Boundary-enriched	UKBiobank
MorningPerson	Morning_Person	1187056	0.1002	0.0035	410520	Neuropsych	Boundary-depleted	UKBiobank
Neuroticism	Neuroticism	1187056	0.1113	0.0037	372066	Neuropsych	Boundary-depleted	UKBiobank
NumChildrenBorn	Number_children_ever_born	1080059	0.0256	0.0018	318863	Reproductive	Boundary-depleted	Barban et al., 2016 Nat Genet
PlateletCount	Platelet_Count	1187056	0.349	0.0294	444382	Hematologic	Boundary-enriched	UKBiobank
RA	Rheumatoid_Arthritis	1125155	0.1694	0.023	38242	Immunologic	Boundary-enriched	Okada et al., 2014 Nature
RBCCount	Red_Blood_Cell_Count	1187056	0.2434	0.0191	445174	Hematologic	Boundary-enriched	UKBiobank
RDW	Red_Blood_Cell_Distribution_Width	1187056	0.2234	0.0198	442700	Hematologic	Boundary-enriched	UKBiobank
Resp_ENT_Dz	Respiratory_and_Ear-nose-throat_Diseases	1187056	0.0483	0.0034	459324	Cardiopulmonary	Boundary-depleted	UKBiobank
Schizophrenia	Schizophrenia	1083014	0.4512	0.0189	70100	Neuropsych	Boundary-depleted	SCZ Working Group of the PGC, 2014 Nature
SkinColor	Skin_Color	1187056	0.1896	0.0539	453609	Dermatologic	Boundary-depleted	UKBiobank
SmokingStatus	Smoking_Status	1187056	0.0972	0.0032	457683	Neuropsych	Boundary-depleted	UKBiobank
Sunburn	Sunburn_Occasion	1187056	0.0915	0.0162	344229	Dermatologic	Boundary-depleted	UKBiobank
SystolicBP	Systolic_Blood_Pressure	1187056	0.1966	0.007	422771	Cardiopulmonary	Boundary-depleted	UKBiobank
T2D	Type_2_Diabetes	1187056	0.043	0.0025	459324	Metabolic	Boundary-enriched	UKBiobank
Tanning	Tanning	1187056	0.172	0.0609	449984	Dermatologic	Boundary-depleted	UKBiobank
UC	Ulcerative_Colitis	1076834	0.2424	0.032	27432	Immunologic	Boundary-enriched	Jostins et al., 2012 Nature
WaistHipRatio	Waist-hip_Ratio	1187056	0.1423	0.0067	458417	Metabolic	Boundary-enriched	UKBiobank
WBCCount	White_Blood_Cell_Count	1187056	0.1873	0.0105	444502	Hematologic	Boundary-enriched	UKBiobank
YearsOfEd	College_Education	1187056	0.1299	0.0037	454813	Neuropsych	Boundary-depleted	UKBiobank

Table 6.8: GWAS traits used for heritability analyses and phenotypic cluster membership.

6.3 Appendix 3: Supporting information for Chapter 4

Supplementary Text

When evaluating the relationship between 3D genome variability and introgression (Results section 4.2.7:"3D genome organization constrained introgression in MHs"), we considered a variety of subsets of genomic windows to fully explore these results. We show that the maintext results (Fig. 4.5) replicate when using earlier introgressed Neanderthal haplotype predictions from Vernot et al.[156] and other thresholds (Figs. 6.44,6.45). We also find that 3D genome variability is more strongly predictive of introgression shared among all three super-populations than an introgressed sequence unique to a single super-population (Table 6.11). We hypothesize this is because the maintenance of a haplotype across diverse populations indicates stronger tolerance of the AH 3D organization pattern in diverse human genomic contexts. Additionally, 3D variability is relatively more informative about the amount of introgression when only considering windows of the genome with any introgressed sequence present (Table 6.12). Thus, we hypothesize that in 1 Mb windows with strong purifying selection against a large-effect introgressed variant (e.g., a deleterious protein-coding variant), 3D genome variability is less relevant. Ultimately, the pressures shaping the landscape of introgression across the genome were multi-factorial, but we demonstrate that 3D genome organization likely played a role.

Supplementary Figures

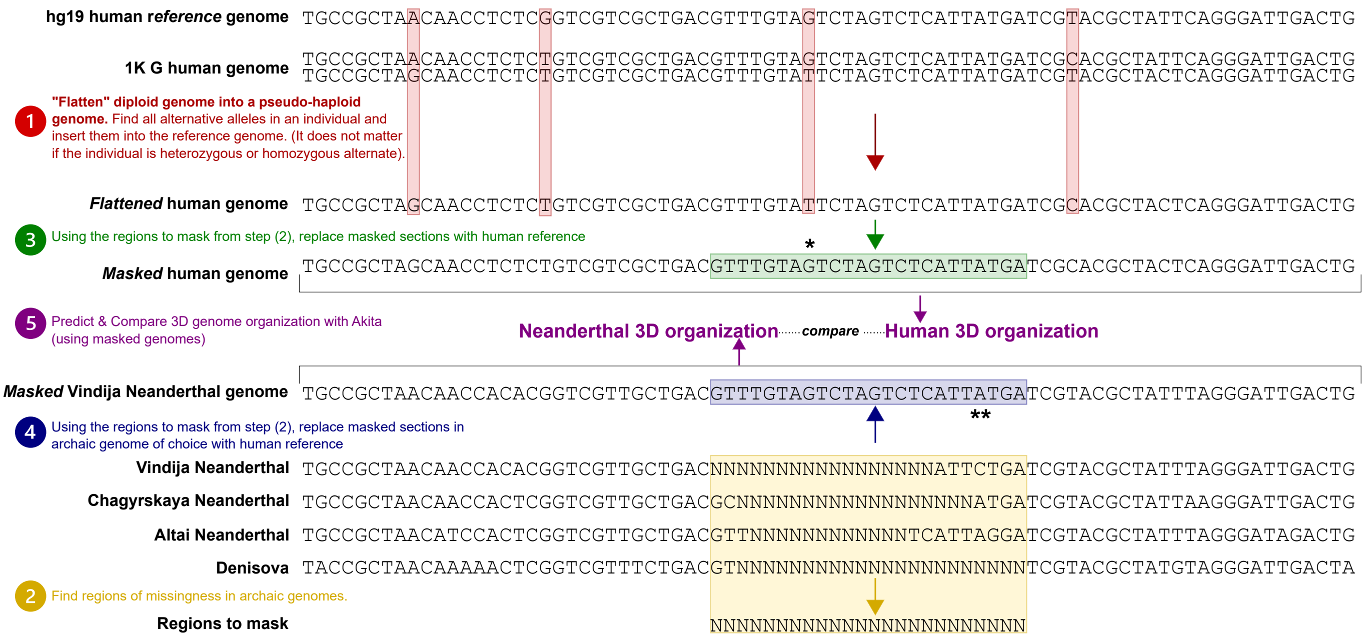

hg19 human *reference* genome
TGCCGCTAACAACCTCTCGGTCGTCGCTGACGTTTGTAGTCTAGTCTCATTATGATCGTACGCTATTCAGGGATTGACTG
1K G human genome
TGCCGCTAACAACCTCTCTGTCGTCGCTGACGTTTGTAGTCTAGTCTCATTATGATCGCACGCTATTCAGGGATTGACTG
TGCCGCTAGCAACCTCTCTGTCGTCGCTGACGTTTGTATTCTAGTCTCATTATGATCGTACGCTACTCAGGGATTGACTG
1
"Flatten" diploid genome into a pseudo-haploid genome. Find all alternative alleles in an individual and insert them into the reference genome. (It does not matter if the individual is heterozygous or homozygous alternate).
Flattened human genome
TGCCGCTAGCAACCTCTCTGTCGTCGCTGACGTTTGTATTCTAGTCTCATTATGATCGCACGCTACTCAGGGATTGACTG
3
Using the regions to mask from step (2), replace masked sections with human reference
*
Masked human genome
TGCCGCTAGCAACCTCTCTGTCGTCGCTGACGTTTGTAGTCTAGTCTCATTATGATCGCACGCTACTCAGGGATTGACTG
5
Predict & Compare 3D genome organization with Akita (using masked genomes)
Neanderthal 3D organization
compare
Human 3D organization
Masked Vindija Neanderthal genome
TGCCGCTAACAACCACACGGTCGTTGCTGACGTTTGTAGTCTAGTCTCATTATGATCGTACGCTATTTAGGGATTGACTG
**
4
Using the regions to mask from step (2), replace masked sections in archaic genome of choice with human reference
Vindija Neanderthal
TGCCGCTAACAACCACACGGTCGTTGCTGACNNNNNNNNNNNNNNNNNNATTCTGATCGTACGCTATTTAGGGATTGACTG
Chagyrskaya Neanderthal
TGCCGCTAACAACCACTCGGTCGTTGCTGACGCNNNNNNNNNNNNNNNNNNNNATGATCGTACGCTATTAAGGGATTGACTG
Altai Neanderthal
TGCCGCTAACATCCACTCGGTCGTTGCTGACGTTNNNNNNNNNNNNNTCATTAGGATCGTACGCTATTTAGGGATAGACTG
Denisova
TACCGCTAACAAAAACTCGGTCGTTTCTGACGTNNNNNNNNNNNNNNNNNNNNNNNTCGTACGCTATGTAGGGATTGACTA
2
Find regions of missingness in archaic genomes.
Regions to mask
NNNNNNNNNNNNNNNNNNNNNNNN

Figure 6.34: Handling missingness in the archaic hominin genomes.
We constructed full-length genomes for each MH or AH based upon their genotyping information. Here, we illustrate a schematic of the procedure used to account for the challenges of archaic DNA. **(1)** Given the difficulty of distinguishing heterozygous genotypes in the ancient DNA samples, we treated all individuals as if they were homozygous (pseudo-haploid). If an individual had an alternate allele (homozygous or heterozygous), we inserted it into the reference genome to create a pseudo-haploid, or "flattened" genome for each individual (hightlighted in red boxes). **(2)** Because of gaps in coverage resulting from the challenges of ancient DNA, particularly in genomic regions of low complexity, we "masked" all genomic regions lacking archaic genotyping information by reverting nucleotide states to the hg19 reference (yellow box). For analyses that compared 3D genome organization between MHs and AHs, and MHs we do this masking procedure for both **[3]** MHs (green box) and **[4]** AHs (blue box) to facilitate appropriate comparisons. **[5]** We run Akita on each processed genome separately and then compare the resulting contact maps. By filling both genomes with the same sequence, there will be no differences between the AH-MH predictions or resulting comparisons. Although AHs and MHs certainly did not have the same genome sequences in these regions of missingness, we preferred this as a conservative approach to minimize identifying regions of interest if there were missing data. For example, we illustrate that at the nucleotide *****, although we observe an MH alternative allele (`T`), it gets masked and replaced with the hg19 reference (`G`) because that locus is not comparable to AH genomes. Many of the regions of missingness are shared by all or most of the AHs because those regions are just inherently difficult to sequence (Fig. 6.35). However, at the nucleotide ******, we illustrate another example where an allele observed in the Vindija genome (`C`) is masked with hg19 reference (`A`) so that it facilitates comparisons between the AHs (some of which have missingness at that locus).

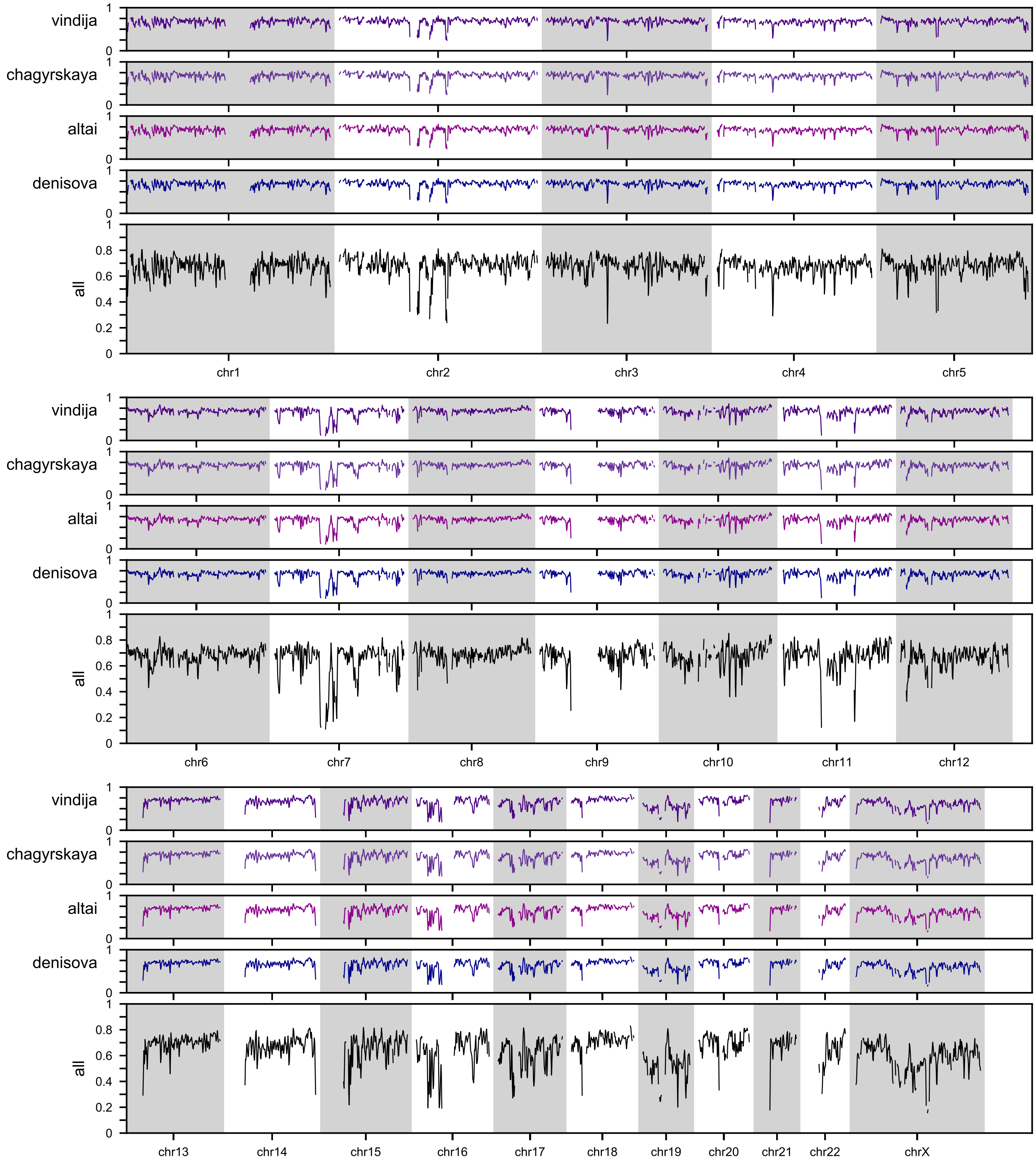


Figure 6.35: Archaic hominin sequence coverage across the genome.
Ancient DNA fragmentation and degradation present challenges to both sequencing and alignment resulting in gaps in coverage, particularly in genomic regions of low complexity. Here, we show coverage across the genome for the 4 AHs. The horizontal axis represents genomic loci at the same sliding approximately 1 Mb window resolution ($N = 4,999$) used to do all analyses (Methods). The vertical axis unit is the proportion of bp with coverage (for the 1 Mb window). Bins without full coverage in modern humans (often near centromeres or telomeres) are excluded from all analyses and this figure. The bottom trace (black, labeled "all") represents the union of the missing segments for all 4 AHs. These regions are masked (Methods, Fig. 6.34) to facilitate 3D genome and sequence variation comparisons.

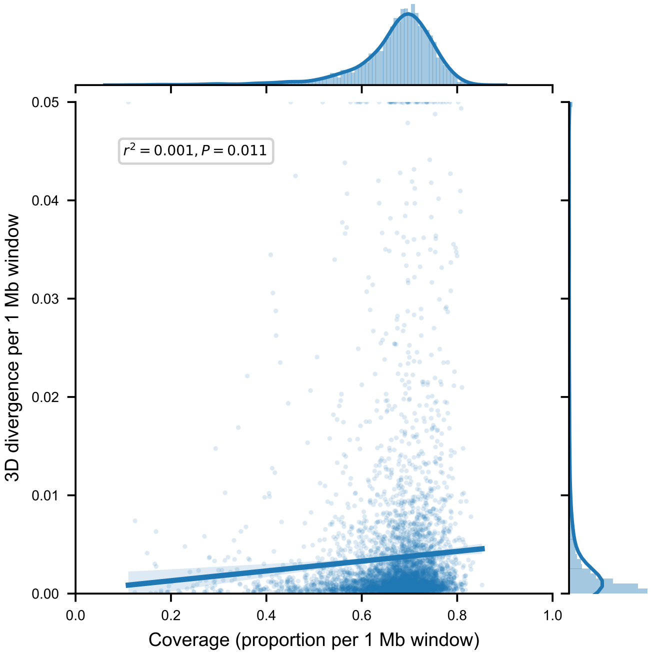


Figure 6.36: 3D divergence in 1 Mb genomic window is weakly correlated with coverage. Because we mask archaic missingness (Methods, Fig. 6.34,6.35), regions with less coverage have more masking and the resulting processed sequences may have less AH-MH sequence variation. For 1 Mb windows across the genome ($N = 4999$), we compare AH (Vindija Neanderthal) and African MH (HG03105) 3D divergence (vertical axis) with the amount of coverage in that window (horizontal axis). The amount masked is equal to $1 - \text{coverage}$. 3D divergence is positively correlated with coverage ($r^2 = 0.001$, $P = 0.01$). This is likely because there is more opportunity to find variation that results in contact map changes when less of the region is masked; however, this correlation is very weak suggesting that more coverage of the archaic genomes may not uncover many additional examples of divergent organization.

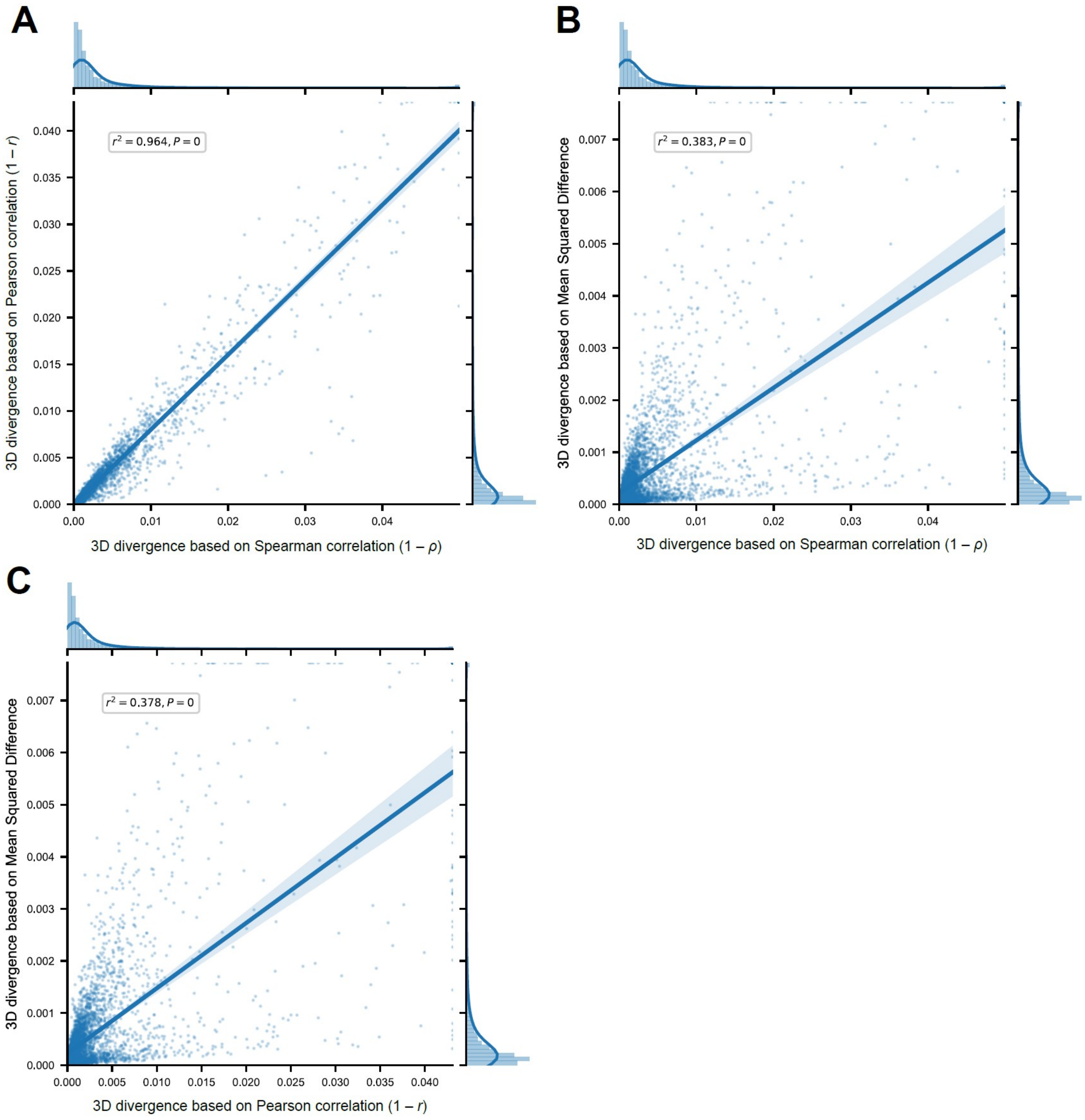


Figure 6.37: Alternative measures of contact map comparison correlate with the 3D divergence derived from the Spearman's rank correlation coefficient.

In the main text, we compare chromatin contact maps using a 3D divergence score based on Spearman's rank correlation coefficient ($1-\rho$). Here, for the same windows across the genome ($N = 4999$), we compare AH (Vindija Neanderthal) and African MH (HG03105) predictions using this Spearman-derived 3D divergence to others based on **(A)** Pearson's correlation coefficient ($1-r$) ($r^2 = 0.964$) and **(B)** mean squared difference ($\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$) ($r^2 = 0.383$). We also compare **(C)** these alternative measures (mean squared difference vs. Pearson's correlation) to each other ($r^2 = 0.378$). The correlations between all measures are highly significant (all $P < 5 \times 10^{-324}$).

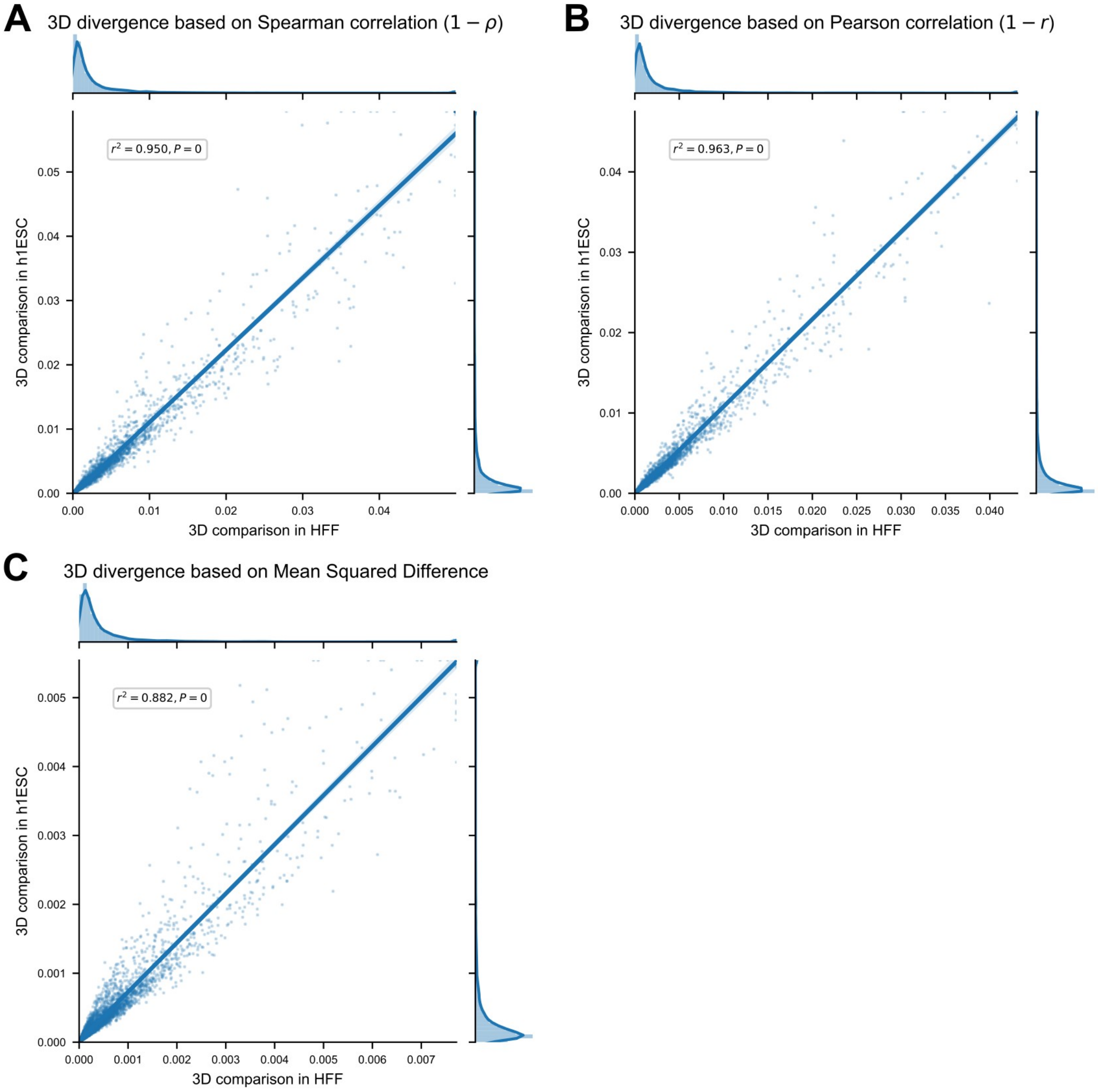


Figure 6.38: 3D genome organization comparisons with chromatin contact maps from embryonic stem cell (ESC) are similar to those from human foreskin fibroblast (HFF).
For the same windows across the genome ($N = 4999$), we compare AH (Vindija Neanderthal) and African MH (HG03105) predictions in embryonic stem cell (ESC) (vertical axis) versus human foreskin fibroblast (HFF) (horizontal axis) cell types. The comparisons across cell types are highly correlated regardless of the measure used to quantify their divergence. We consider comparison measures defined using the **(A)** Spearman correlation ($r^2 = 0.95$), **(B)** Pearson correlation ($r^2 = 0.96$), and **(C)** mean squared difference ($r^2 = 0.88$) (all $P < 5 \times 10^{-324}$).

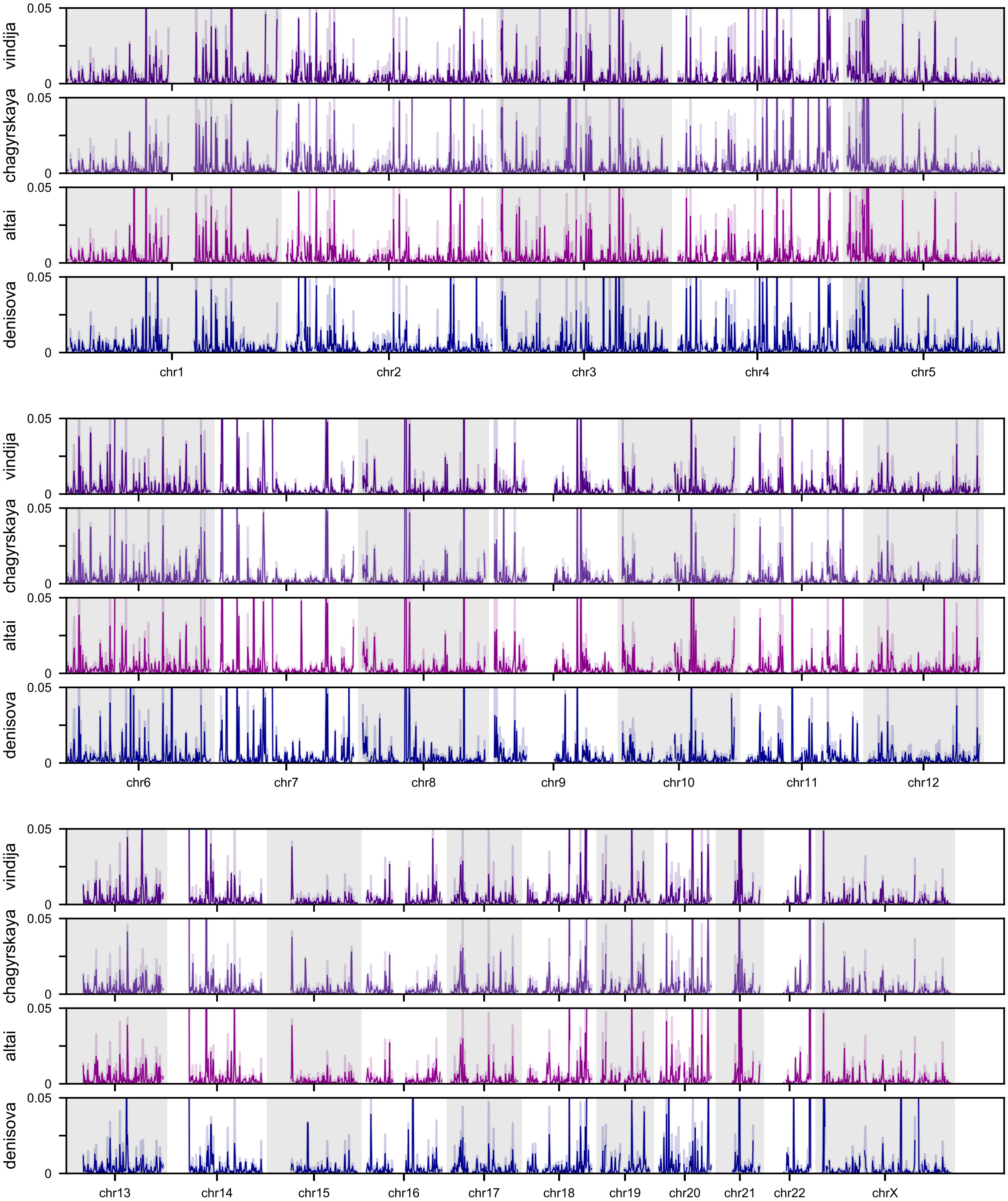


Figure 6.39: AH-MH 3D divergence across the whole genome.
Across the genome, we plotted the average divergence of each of the AHs to five modern African individuals from different subpopulations. The horizontal axis represents genomic loci at the same sliding 1 Mb window resolution ($N = 4,999$) used to do all analyses (Methods). This expands Fig. 4.2C from chr7 to the whole genome. The error band indicates the 95% CI. Comparing the 3D genomes of Neanderthals (purple) or Denisova (blue) with MHs reveals windows of both similarity and divergence (peaks).

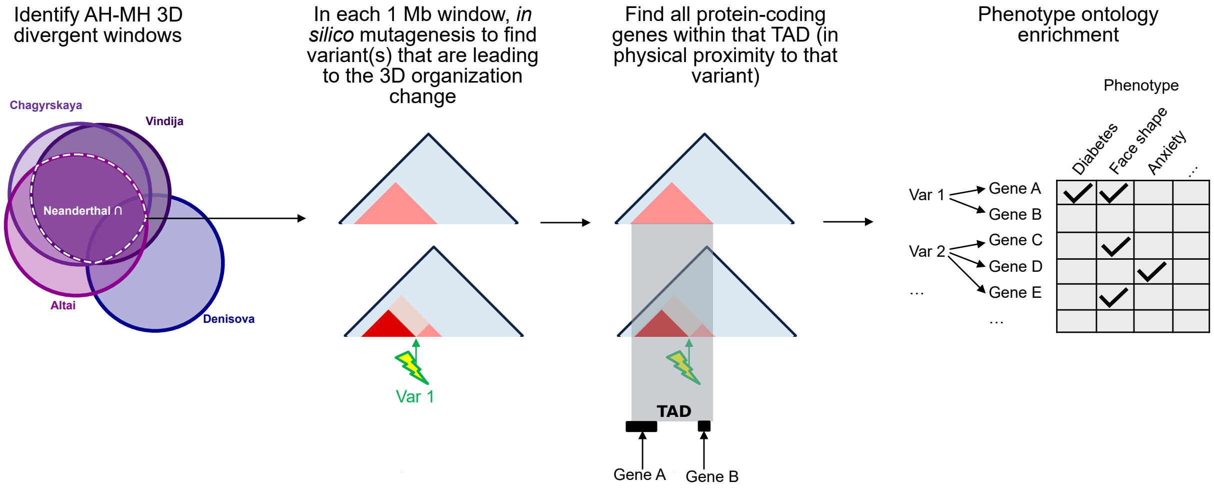

Identify AH-MH 3D divergent windows
Chagyrskaya
Vindija
Neanderthal ∩
Altai
Denisova
In each 1 Mb window, *in silico* mutagenesis to find variant(s) that are leading to the 3D organization change
Var 1
Find all protein-coding genes within that TAD (in physical proximity to that variant)
TAD
Gene A
Gene B
Phenotype ontology enrichment
Phenotype
Diabetes
Face shape
Anxiety
...
Var 1
Var 2
...
Gene A
Gene B
Gene C
Gene D
Gene E
...

Figure 6.40: Method for linking 3D divergent windows to test phenotype ontology term enrichment. To test if differences in AH-MH 3D organization are enriched near genes related to particular phenotypes we follow a procedure that sequentially links 3D divergent windows to variants to TADs to genes and, ultimately, to phenotypes. We identify AH-MH 3D divergent windows in Fig. 4.3A–B. We consider three different sets of AH-MH divergent windows, those shared (intersect) by all Neanderthals, those in any Neanderthal (union), and those in Denisova. Results from the set shared by all Neanderthals ($N = 43$ windows) are shown in the main text (Fig. 4.3D). In each 1 Mb 3D divergent window, we identify the variant(s) contributing to the most prominent 3D differences using *in silico* mutagenesis (lightning bolt) (Methods). 3D-modifying variants are then linked to protein-coding genes (black bars) in their TAD (gray rectangle) because this provides evidence of physical proximity. Genes are linked to phenotypes from the Human Phenotype Ontology (HPO) and genome-wide association studies (GWAS) Catalog 2019. Through this procedure, we counted the number of ontology terms linked to the set of 3D-modifying variants. We test enrichment for ontology terms linked to at least one 3D-modifying variant using a shuffling approach to create an empirical distribution for how many times we would observe each annotation under the null. We used these distributions to calculate an enrichment and P-value for each ontology term. The specific data sets used in this procedure are detailed in the Methods. Counts of the number of windows, 3D-modifying variants, genes, and phenotypes for each set are in Table 6.10. Results for enrichment are in Figs. 4.3D,6.41.

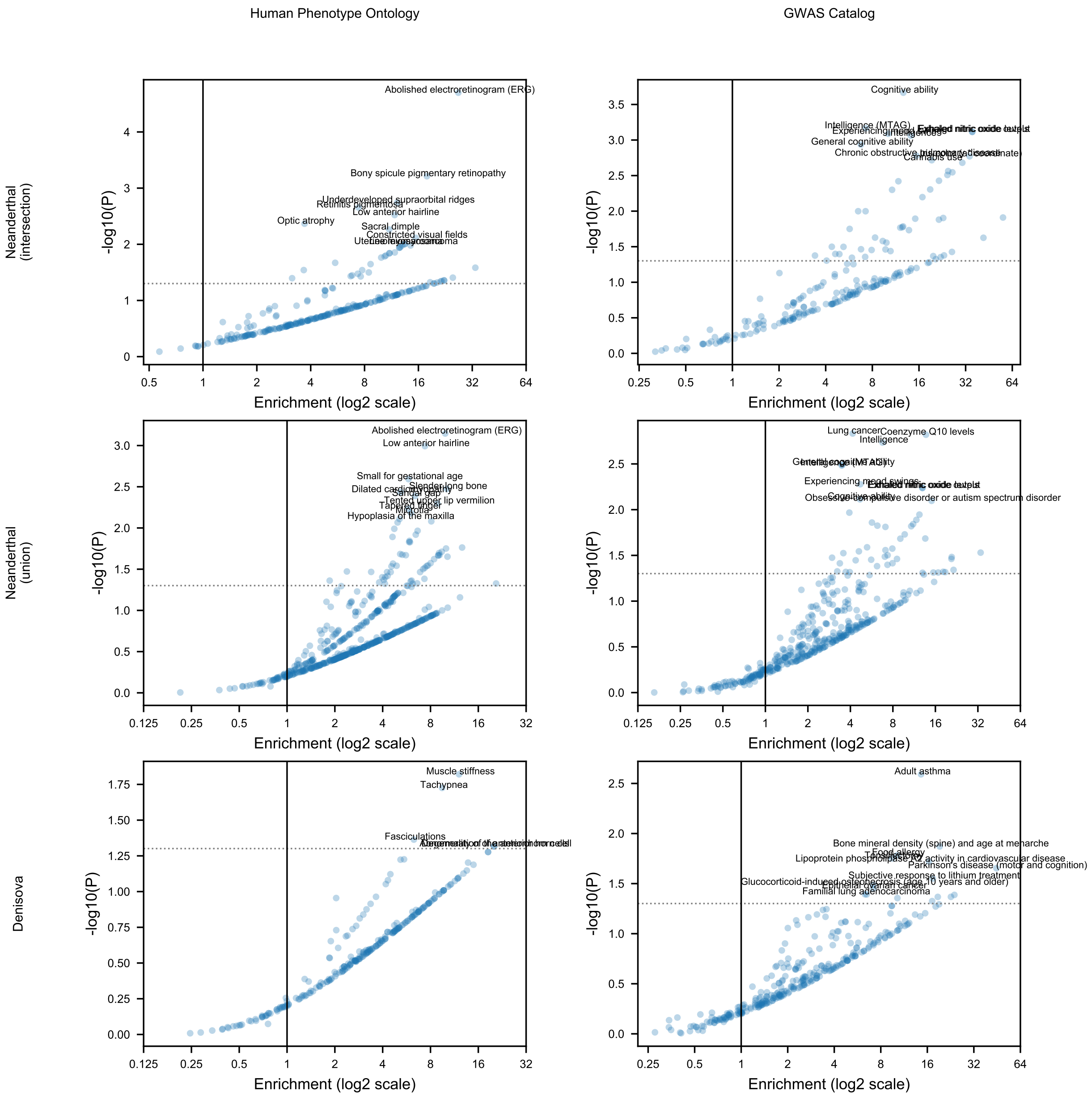


Figure 6.41: Phenotype ontology enrichment across other sets of AH-MH 3D divergent windows implicate similar phenotypes.

When testing if differences in AH-MH 3D organization are enriched near genes related to particular phenotypes, we used three different sets of AH-MH 3D divergent windows (rows) and two different sets of gene-phenotype links (columns). The top set is from 43 3D-divergent windows shared by Neanderthals (intersect) (also shown in the main text, Fig. 4.3D). The middle is from 110 divergent windows in any Neanderthal (union). The bottom is from 73 divergent windows in Denisova. Each volcano plot has enrichment on the horizontal axis and significance on the vertical axis which were calculated with reference to a shuffled null distribution ($n = 500,000$, Methods). Each point represents one ontology term. Only terms linked to the 3D divergent windows in each set were tested for enrichment or depletion. The most significant 10 terms are labeled if $P < 0.05$ (dotted line). Similar to the Neanderthal (intersection) set, phenotypes related to the retina, hair, immune response, skeleton, cognition, and lung capacity are highlighted. Additional phenotypes at nominal significance include traits related to the heart, muscle, cancer, and bone density. Details about the process to link the 3D divergent windows to genes and phenotypes are in the Methods and Fig. 6.40. Details about the number of windows, variants, and phenotypes considered for each set are in Table 6.10.

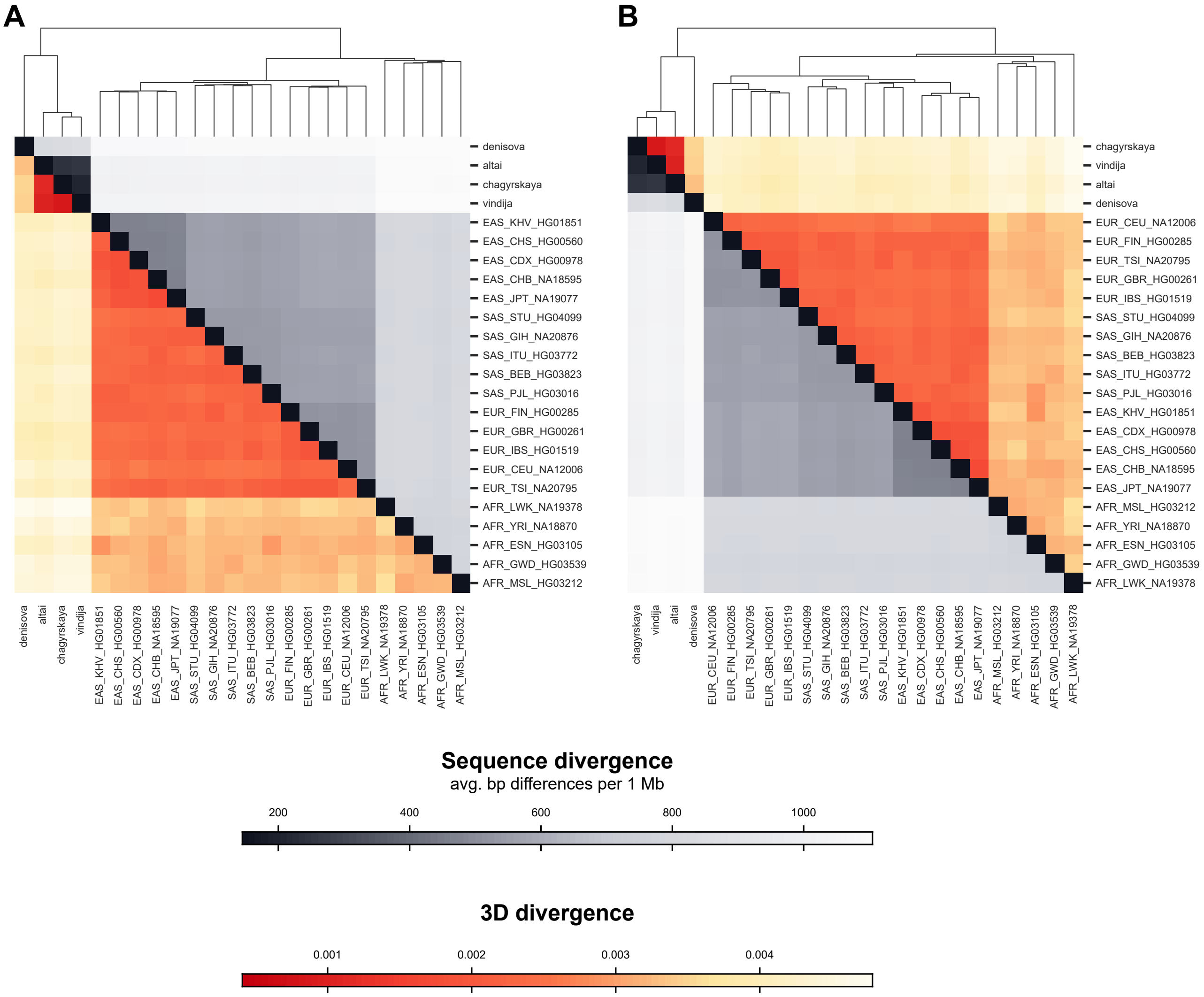


Figure 6.42: Full pairwise heatmaps clustered by both sequence 3D divergence and sequence divergence. We calculated the mean genome-wide 3D divergence for all pairs of AH and MH individuals (oranges) to compare with the genome-wide mean sequence divergence (grays). Fig. 4.4A displays these heatmaps when clustered by sequence divergence. Fig. 4.4A is reproduced in **(A)** with the full labels of all 1KGP individuals and their sub- and super-population information. **(B)** We also show the heatmap clustered by 3D genome divergence. Overall, global patterns of 3D genome divergence follow global patterns of sequence divergence. Lists of 1KGP individuals used and their abbreviation codes are defined in Table 6.9.

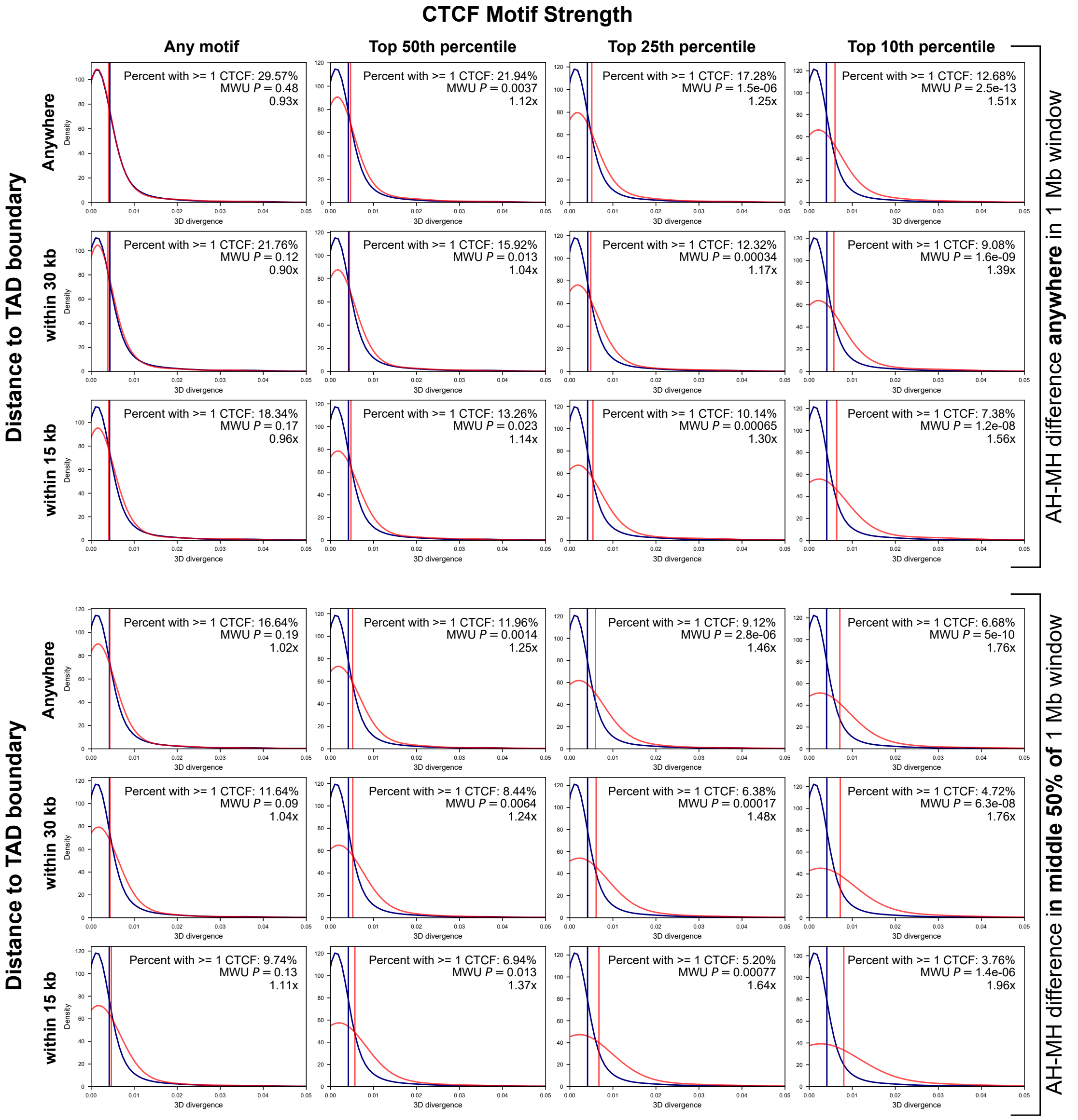


Figure 6.43: 3D genome divergence depends on both the strength and context of the CTCF motif disrupted. Based on the importance of CTCF-binding in maintaining 3D genome organization, we quantified the effects of AH-MH nucleotide differences overlapping CTCF binding motifs on 3D divergence. Given the complexity in the "grammar" of encoding 3D genome organization, we hypothesized that not all CTCF disruptions are equally likely to influence 3D divergence. Fig. 4.4B demonstrates this. But, here we replicate this with other thresholds and filters. We considered if each 1 Mb window ($N = 4,999$) had a sequence difference between a Neanderthal (Vindija) and a MH (HG03105) genome that overlapped a CTCF site. We plotted the distribution of 3D divergence in a window by whether there was a "CTCF overlapping variant" (red) or not (blue). We further filtered windows by multiple annotations describing the context and strength of the CTCF site overlapped. First, we stratified windows by if the "CTCF overlapping variant" occurs within the middle half of the 1 Mb window (right vertical axis). Second, we stratified windows by the proximity of the "CTCF overlapping variant" to a TAD boundary (anywhere, within 30 kb, or within 15 kb) (left vertical axis). Finally, we stratified windows by the strength of the overlapped CTCF motif in percentiles (any, top 50%, 25%, or 10%) (horizontal axis). All three features describing context and strength are informative about the likelihood of 3D divergence. For example, when filtering for the strongest CTCF motifs overlapped by a variant, 3D divergence increases 1.96-fold compared to 1.11-fold if strength is ignored (bottom left vs. bottom right). When considering by proximity to TAD boundaries, 3D divergence always increases when a "CTCF overlapping variant" is closer to a TAD boundary (4th row vs. 6th row). This illustrates that our approach has learned the complex sequence patterns underlying 3D genome folding that could not be determined by simply intersecting AH variants with all CTCF sites.

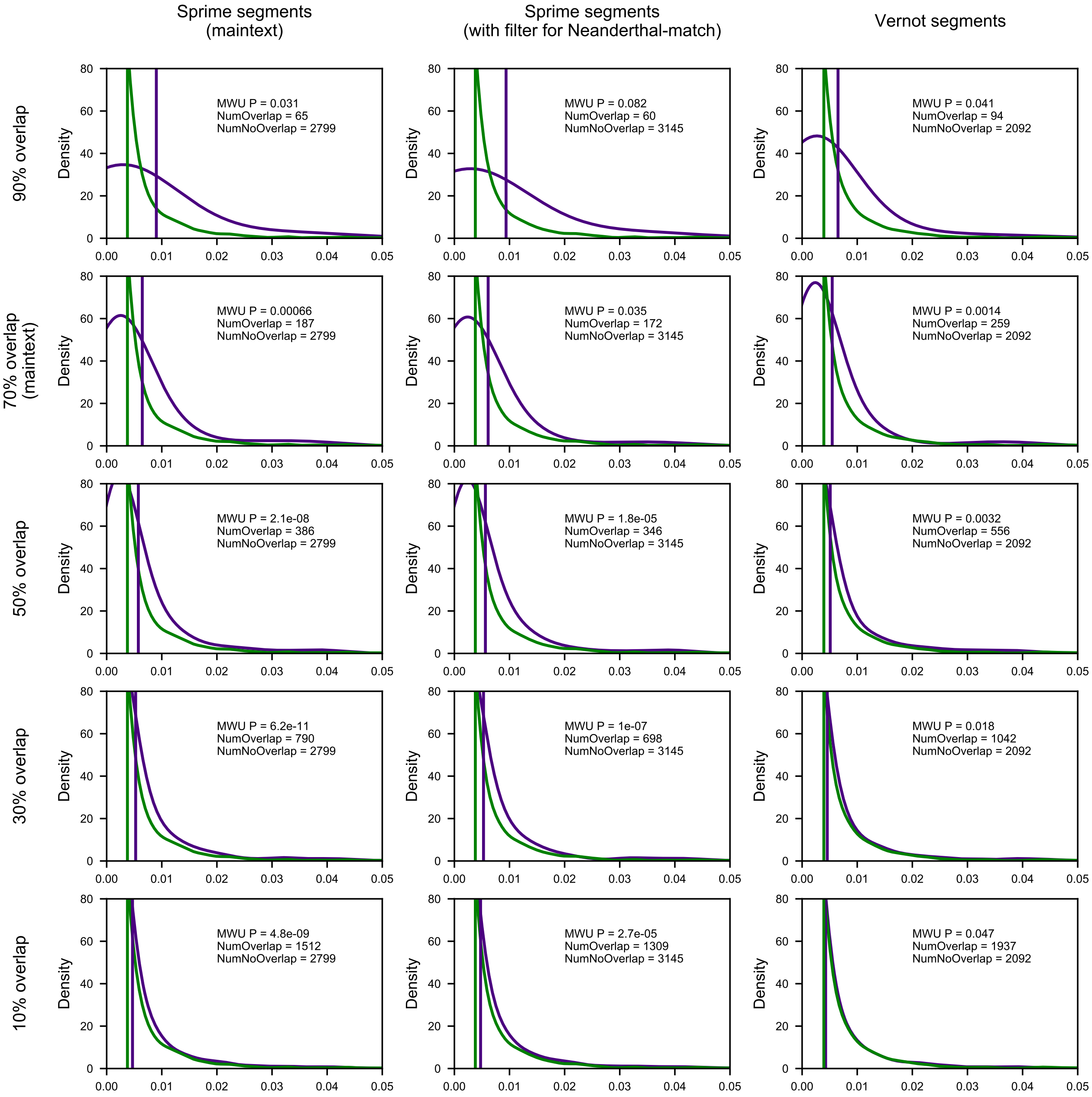


Figure 6.44: Windows with evidence of AH introgression are more 3D variable in MHs even when using different definitions of introgression.

Genomic windows with high levels of introgression across present-day non-African populations (purple distribution) are more 3D-variable in modern Africans (horizontal axis) than windows without evidence of introgression (green distribution). In the main text, we considered introgression defined by segments from Browning et al.[32] (first column) covering at least 70% of bases in a 1 Mb window (second row). This identifies 187 autosomal 1 Mb windows with introgression and 2,799 without (same figure as Fig. 4.5A). Here, we show that this trend is consistent even when using different sets of introgressed haplotypes (columns) and thresholds for overlap (rows). Sprime segments are from Browning et al.[32]. Sprime segments with Neanderthal-matching filter are a subset of the Browning et al.[32] introgressed segments that have 30 putatively introgressed variants that could be compared to the Altai Neanderthal genome and had a match rate of at least 30% to the Altai Neanderthal allele. S* Vernot segments are from Vernot et al.[156]. Vertical lines represent the distribution means. *P*-values are from a two-tailed Mann–Whitney U test.

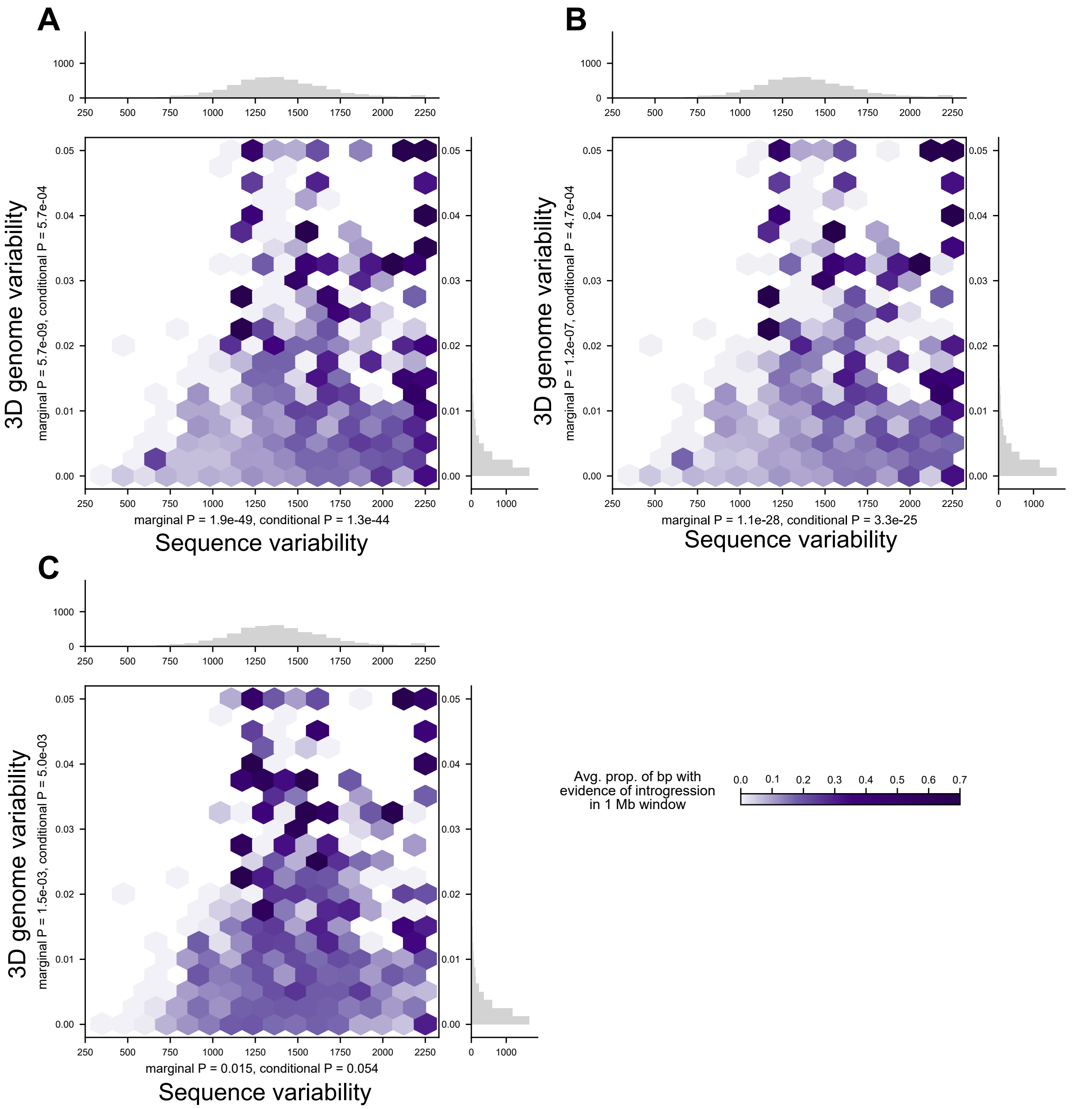


Figure 6.45: 3D variable windows in MH have more evidence of AH introgression even when using different definitions of introgression.

For three different sets of introgressed haplotypes (**A**-**C**), we plot the relationship between sequence variability (horizontal axis) and 3D genome variability (vertical axis) with amount of AH ancestry in a window (purples). Darker purple indicates a higher proportion of introgression in a 1 Mb genomic window. 3D genome variability is defined as the average modern-African pairwise 3D genome diversity. Sequence variability is defined as the average pairwise nucleotide differences per modern-African in a 1 Mb window. *P*-values correspond to the significance of sequence variability or 3D genome variability to predict amount of introgression in a 1 Mb window. 3D genome variability is predictive of the amount of introgression both independently and when conditioned on sequence variability for all three sets of introgression. For, **A**,**B**, and **C**, respectively, introgressed haplotypes are from Sprime segments, Sprime segments with a Neanderthal-sequence match filter, and S* segments. **A** is shown in the maintext in Fig. 4.5B. Sprime segments are from Browning et al.[32]. Sprime segments with Neanderthal-matching filter are a subset of the Browning et al.[32] introgressed segments that have 30 putatively introgressed variants that could be compared to the Altai Neanderthal genome and had a match rate of at least 30% to the Altai Neanderthal allele. Vernot segments are from Vernot et al.[156].

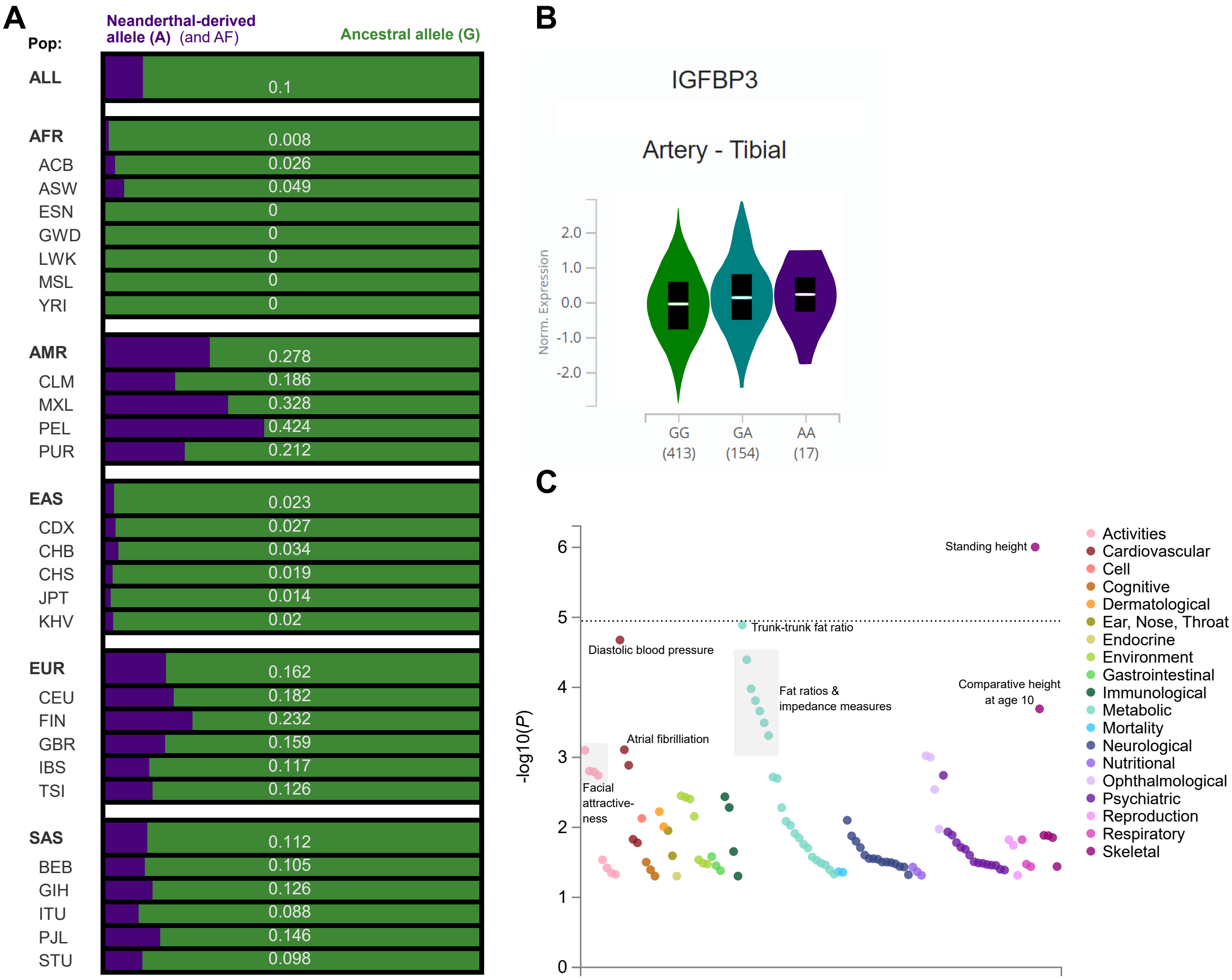


Figure 6.46: rs12536129 is a high-frequency introgressed allele with regulatory and phenotypic associations. In Fig. 4.6A–B, we describe an AH-MH 3D divergent window that was introgressed into some modern Eurasians. *In silico* mutagenesis of this window revealed a G to A change at chr7:46,169,621 (rs12536129) associated with the largest change in 3D genome organization. **(A)** Across human populations, this introgressed allele remains at high-frequency today, especially in Peru (28% AMR, 2% EAS, 16% EUR, 11% SAS, 0% non-admixed sub-Saharan AFR). Purple bars represent the frequency of the introgressed Neanderthal-derived allele. **(B)** This introgressed allele is also an eQTL in GTEx for the physically linked gene *IGFBP3*, Insulin-like growth factor-binding protein 3 ($P = 0.00014$ in artery tissue)[95]. **(C)** In MHs, this variant is associated with traits including standing height ($P = 9.9 \times 10^{-7}$), fat distribution (trunk fat ratio, impedance measures, $P = 1.3 \times 10^{-5}$), and diastolic blood pressure ($P = 2.1 \times 10^{-5}$). This figure was generated with the GWASAtlas from Watanabe et al.[159] and is sorted by domain and *P*-value. The dotted line represents a highly conservative Bonferroni corrected *P*-value (1.05×10^{-5}) for testing 4756 traits (including many correlated traits and GWASs in which the SNP was not tested).

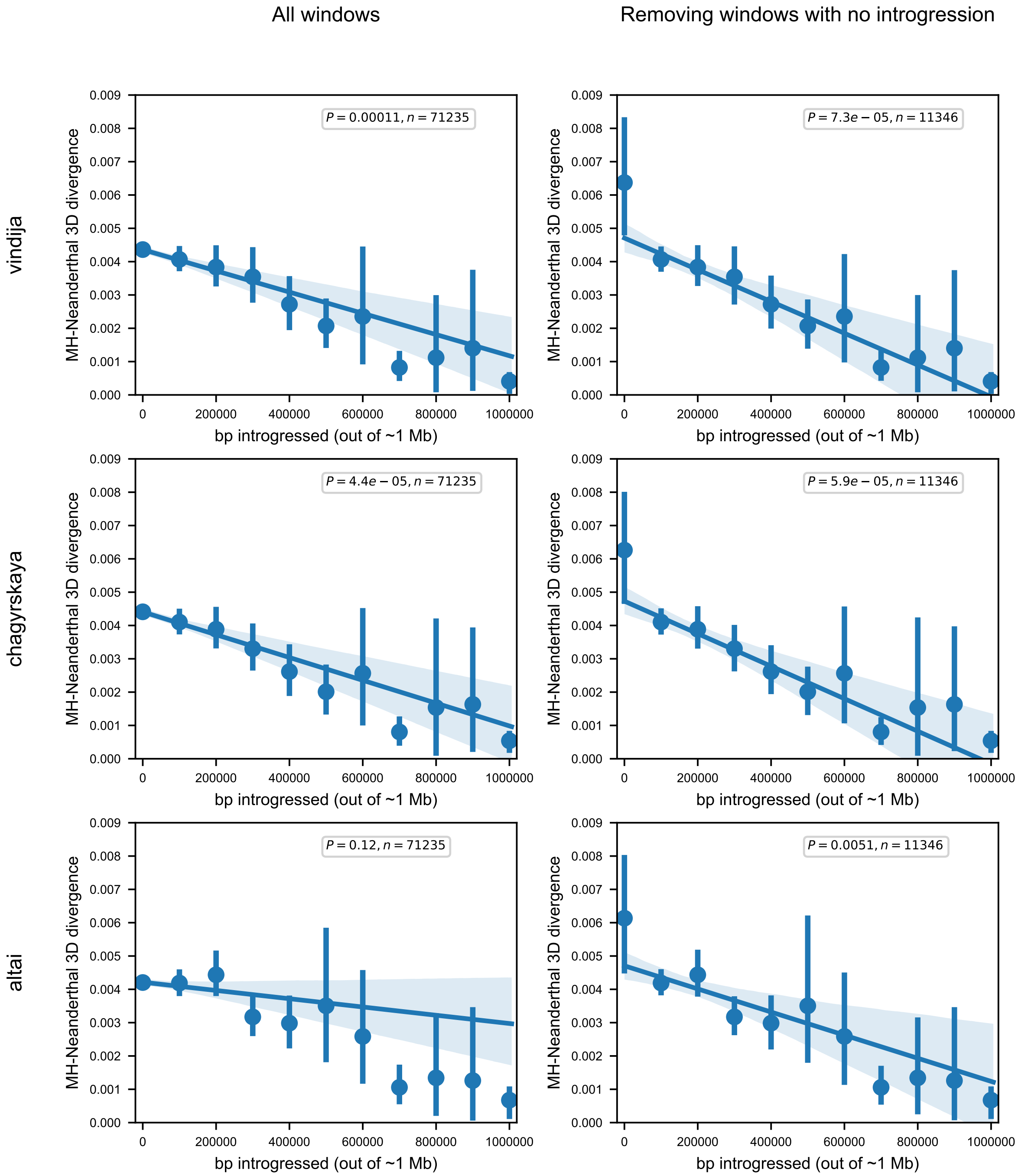


Figure 6.47: Amount of introgression is negatively correlated with 3D divergence to all Neanderthal individuals. The amount of introgression in a 1 Mb window (number of bp, horizontal axis) is significantly correlated with the similarity of an individual's 3D genome organization to a Neanderthal's genome organization (vertical axis). This is demonstrated across all three Neanderthal individuals: Vindija in the top panel (also shown in Fig. 4.6C), Chagyrskaya in the middle, and Altai at the bottom. We hypothesize the trend is weakest in Altai because it is less related to the introgressing Neanderthal population compared to the Vindija Neanderthal[14]. The left column considers all 4,749 autosomal 1 Mb windows for 15 Eurasians (total $n = 71,235$, 1KGP individuals in Table 6.9). In the right column, this trend also holds when you remove 1 Mb windows with no (0 bp) introgression in the 15 considered Eurasian individuals $n = 11,346$. The P-values are the significance of the correlation. The error bars signify 95% bootstrapped confidence intervals and the error band signifies the 95% bootstrapped confidence interval for the linear regression estimate.

Supplementary Tables

	Superpopulation	Subpopulation	ID	Subpopulation Description
Individuals in initial and Eurasian introgression analyses	EAS	CDX	HG00978	Chinese Dai in Xishuangbanna, China
	EAS	CHB	NA18595	Han Chinese in Beijing, China
	EAS	CHS	HG00560	Han Chinese South
	EAS	JPT	NA19077	Japanese in Tokyo, Japan
	EAS	KHV	HG01851	Kinh in Ho Chi Minh City, Vietnam
	EUR	CEU	NA12006	Utah residents (CEPH) with Northern and Western European ancestry
	EUR	FIN	HG00285	Finnish in Finland
	EUR	GBR	HG00261	British in England and Scotland
	EUR	IBS	HG01519	Iberian populations in Spain
	EUR	TSI	NA20795	Toscani in Italia
	SAS	BEB	HG03823	Bengali in Bangladesh
	SAS	GIH	NA20876	Gujarati Indian in Houston, TX
	SAS	ITU	HG03772	Indian Telugu in the UK
	SAS	PJL	HG03016	Punjabi in Lahore, Pakistan
	SAS	STU	HG04099	Sri Lankan Tamil in the UK
	AFR	GWD	HG03539	Gambian in Western Division, The Gambia
	AFR	LWK	NA19378	Luhya in Webuye, Kenya
	AFR	MSL	HG03212	Mende in Sierra Leone
	AFR	YRI	NA18870	Yoruba in Ibadan, Nigeria
	AFR	ESN	HG03105*	Esan in Nigeria
Africans in AH-MH divergence and 3D genome variability analyses	AFR	ESN	HG03105	Esan in Nigeria
	AFR	ESN	HG03499	Esan in Nigeria
	AFR	ESN	HG03511	Esan in Nigeria
	AFR	ESN	HG03514	Esan in Nigeria
	AFR	ESN	HG02922	Esan in Nigeria
	AFR	GWD	HG03539	Gambian in Western Division, The Gambia
	AFR	GWD	HG03025	Gambian in Western Division, The Gambia
	AFR	GWD	HG03028	Gambian in Western Division, The Gambia
	AFR	GWD	HG03040	Gambian in Western Division, The Gambia
	AFR	GWD	HG03046	Gambian in Western Division, The Gambia
	AFR	LWK	NA19378	Luhya in Webuye, Kenya
	AFR	LWK	NA19017	Luhya in Webuye, Kenya
	AFR	LWK	NA19434	Luhya in Webuye, Kenya
	AFR	LWK	NA19445	Luhya in Webuye, Kenya
	AFR	LWK	NA19019	Luhya in Webuye, Kenya
	AFR	MSL	HG03212	Mende in Sierra Leone
	AFR	MSL	HG03086	Mende in Sierra Leone
	AFR	MSL	HG03085	Mende in Sierra Leone
	AFR	MSL	HG03437	Mende in Sierra Leone
	AFR	MSL	HG03378	Mende in Sierra Leone

Table 6.9: 1000 Genomes Project (1KGP) individual genomes used for 3D genome predictions. The top set of individuals were used in the initial 3D genome survey (Figs. 4.2, 4.4A) and introgression analyses (Fig. 4.6). The bottom set of African individuals was used to more robustly call AH-MH 3D genome divergence windows (Fig. 4.3) and to calculate MH 3D genome variability (Fig. 4.5). For consistency, the genome of HG03105 was used for all examples.

‘	Number of 1 Mb 3D divergent windows (includes partially overlapped windows)	Number of unique 3D divergent windows (merging overlapping windows)	Number of “3D-modifying variants” observed in the 3D divergent windows	Number of “3D-modifying variants” that have evidence of introgression	Proportion of “3D-modifying variants” that have evidence of introgression	Number of gene-”3D-modifying variants” links*	Number of unique genes linked to 3D-modifying variants”*	Number of unique HPO terms linked to “3D-modifying variants” (used in enrichment test)*	Number of unique GWAS terms linked to “3D-modifying variants” (used in enrichment test)*
Vindija	93	70	76	38	0.500				
Chagyrskaya	95	71	78	32	0.410				
Altai	82	67	73	33	0.452				
Denisova	105	73	83	9	0.108	130	129	248	318
All Neanderthals (intersection)	54	43	45	28	0.622	88	85	271	208
All Neanderthals (union)	144	110	121	43	0.355	224	206	535	435
All archaic hominins (intersection)	10	7	8	6	0.750				
All archaic hominins (union)	234	167	191	45	0.236				

Table 6.10: Counts of 3D divergent windows and 3D-modifying variants.
The number of 3D divergent windows per different AH individuals (rows) are in the first two columns. The first column is the raw 1 Mb windows, while the second column counts overlapping windows as one merged window (these values are depicted in Fig. 4.3B). The number of 3D-modifying variants in each window is in column three. The number and fraction of these 3D-modifying variants that are introgressed are in columns four and five. To conduct the phenotype ontology enrichment analyses shown in Figs. 4.3D,6.41, we linked the 3D-modifying variants to genes (column six and seven, see Methods). These genes were then linked to terms using Human Phenotype Ontology (HPO) (column eight) and the GWAS Catalog. These two sets of terms were then tested for enrichment. The phenotype ontology enrichment analysis parts (columns denoted with *) were only calculated on certain sets of 3D-diverged windows.

		Sequence variability		3D genome variability	
		marginal P	conditional P	marginal P	conditional P
Browning introgressed haplotypes	introgression SHARED across populations	1.9E-49	1.3E-44	5.7E-09	0.00057
	introgression UNIQUE to one population	0.039	0.019	0.14	0.066
Browning introgressed haplotypes with Neanderthal filter	introgression SHARED across populations	1.1E-28	3.3E-25	1.2E-07	0.00047
	introgression UNIQUE to one population	0.067	0.014	0.00054	0.00013
Vernot introgressed haplotype	introgression SHARED across populations	0.015	0.054	0.0015	0.005
	introgression UNIQUE to one population	0.48	0.79	0.0094	0.012

Table 6.11: Both 3D genome and sequence variability are more important in predicting introgression shared across super-populations than introgression unique to a single super-population.

When considering the relationships between 3D genome variability, sequence variability, and amount of introgression (Supplemental Text, Figs. 4.5, 6.45), we consider introgression that was shared across 1KGP super-populations (EAS, EUR, SAS) (white rows) compared to introgression unique to only one super-population (gray rows). We find that 3D genome variability (last two columns) is more strongly predictive of introgression shared among all three super-populations. The analysis was replicated on three sets of introgressed haplotypes. Browning introgressed haplotypes are Sprime segments Browning haplotypes with Neanderthal-matching filter are a subset of the Browning et al.[32] introgressed segments that have 30 putatively introgressed variants that could be compared to the Altai Neanderthal genome and had a match rate of at least 30% to the Altai Neanderthal allele. Vernot haplotypes are S* segments from Vernot et al.[156].

		Sequence variability		3D genome variability	
		marginal P	conditional P	marginal P	conditional P
Browning introgressed haplotypes	ALL windows (N = 4749)	1.90E-49	1.30E-44	5.70E-09	0.00057
	ONLY windows with any evidence of introgression (N = 1950)	0.0004	0.0072	1.90E-06	3.00E-05
Browning introgressed haplotypes with Neanderthal filter	ALL windows (N = 4749)	1.10E-28	3.30E-25	1.20E-07	0.00047
	ONLY windows with any evidence of introgression (N = 1604)	0.042	0.19	0.0001	0.00038
Vernot introgressed haplotype	ALL windows (N = 4749)	1.50E-02	5.40E-02	0.0015	0.005
	ONLY windows with any evidence of introgression (N = 2657)	3.40E-05	8.40E-07	0.00068	1.60E-05

Table 6.12: Compared to sequence variability, 3D variability is a relatively more informative predictor of amount of introgression when considering windows of the genome with any introgression.
When considering the relationships between 3D genome variability, sequence variability, and amount of introgression (Supplemental Text, Figs. 4.5, 6.45), we consider a subset of windows with any evidence of introgression (gray rows) compared to all windows (white rows). 3D variability is relatively more informative about the amount of introgression when only considering windows of the genome with any introgressed sequence present (last column). The analysis was replicated on three sets of introgressed haplotypes. Browning introgressed haplotypes are Sprime segments Browning haplotypes with Neanderthal-matching filter are a subset of the Browning et al.[32] introgressed segments that have 30 putatively introgressed variants that could be compared to the Altai Neanderthal genome and had a match rate of at least 30% to the Altai Neanderthal allele. Vernot haplotypes are S* segments from Vernot et al.[156].

Bibliography

[1] Zimmer, C. *She has her mother's laugh: The powers, perversions and potential of heredity*. New York: Dutton, 2019.

[2] Wolf, A. B. and Akey, J. M. "Outstanding questions in the study of archaic hominin admixture". In: *PLoS Genetics* 14.5 (May 2018), e1007349. ISSN: 15537404. DOI: 10.1371/journal.pgen.1007349.

[3] Benton, M. L., Abraham, A., LaBella, A. L., Abbot, P., Rokas, A., and Capra, J. A. "The influence of evolutionary history on human health and disease". In: *Nature Reviews Genetics* 22.5 (May 2021), pp. 269–283. ISSN: 14710064. DOI: 10.1038/s41576-020-00305-9.

[4] Loewe, L. and Hill, W. G. "The population genetics of mutations: Good, bad and indifferent". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1544 (Apr. 2010), pp. 1153–1167. ISSN: 14712970. DOI: 10.1098/rstb.2009.0317.

[5] Bikle, D. D. "Vitamin D metabolism, mechanism of action, and clinical applications". In: *Chemistry and Biology* 21.3 (Mar. 2014), pp. 319–329. ISSN: 10745521. DOI: 10.1016/j.chembiol.2013.12.016.

[6] Lachance, J., Berens, A. J., Hansen, M. E., Teng, A. K., Tishkoff, S. A., and Rebbeck, T. R. "Genetic hitchhiking and population bottlenecks contribute to prostate cancer disparities in men of african descent". In: *Cancer Research* 78.9 (May 2018), pp. 2432–2443. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-17-1550.

[7] King, M. C. and Wilson, A. C. "Evolution at two levels in humans and chimpanzees". In: *Science* 188.4184 (1975), pp. 107–116. ISSN: 00368075. DOI: 10.1126/science.1090005.

[8] Wray, G. A. "The evolutionary significance of cis-regulatory mutations". In: *Nature Reviews Genetics* 8.3 (Mar. 2007), pp. 206–216. ISSN: 14710056. DOI: 10.1038/nrg2063.

[9] Carroll, S. B. *Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution*. 2008. DOI: 10.1016/j.cell.2008.06.030.

[10] Wittkopp, P. J. and Kalay, G. *Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence*. 2012. DOI: 10.1038/nrg3095.

[11] Finlayson, C., Fa, D. A., Jiménez Espejo, F., Carrión, J. S., Finlayson, G., Giles Pacheco, F., Rodríguez Vidal, J., Stringer, C., and Martínez Ruiz, F. "Gorham's Cave, Gibraltar-The persistence of a Neanderthal population". In: *Quaternary International* 181.1 (Apr. 2008), pp. 64–71. ISSN: 10406182. DOI: 10.1016/j.quaint.2007.11.016.

[12] Higham, T., Douka, K., Wood, R., et al. "The timing and spatiotemporal patterning of Neanderthal disappearance". In: *Nature* 512.7514 (Aug. 2014), pp. 306–309. ISSN: 14764687. DOI: 10.1038/nature13621.

[13] Prüfer, K., Racimo, F., Patterson, N., et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), pp. 43–49. ISSN: 00280836. DOI: 10.1038/nature12886.

[14] Prüfer, K., De Filippo, C., Grote, S., et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia". In: *Science* 358.6363 (2017), pp. 655–658. ISSN: 10959203. DOI: 10.1126/science.aao1887.

[15] Mafessoni, F., Grote, S., Filippo, C. D., et al. "A high-coverage neandertal genome from chagyrskaya cave". In: *Proceedings of the National Academy of Sciences of the United States of America* 117.26 (June 2020), pp. 15132–15136. ISSN: 10916490. DOI: 10.1073/pnas.2004944117.

[16] Meyer, M., Kircher, M., Gansauge, M. T., et al. "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104 (Oct. 2012), pp. 222–226. ISSN: 10959203. DOI: 10.1126/science.1224344.

[17] Kuhlwilm, M. and Boeckx, C. “A catalog of single nucleotide changes distinguishing modern humans from archaic hominins”. In: *Scientific Reports* 9.1 (June 2019), pp. 1–14. ISSN: 20452322. DOI: 10.1038/s41598-019-44877-x.

[18] Rogers, A. R., Bohlender, R. J., and Huff, C. D. “Early history of Neanderthals and Denisovans”. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.37 (Sept. 2017), pp. 9859–9863. ISSN: 10916490. DOI: 10.1073/pnas.1706426114.

[19] Rogers, A. R., Harris, N. S., and Achenbach, A. A. “Neanderthal-Denisovan ancestors interbred with a distantly related hominin”. In: *Science Advances* 6.8 (2020). ISSN: 23752548. DOI: 10.1126/sciadv.aay5483.

[20] Gómez-Robles, A. “Dental evolutionary rates and its implications for the Neanderthal–modern human divergence”. In: *Science Advances* 5.5 (May 2019). ISSN: 23752548. DOI: 10.1126/sciadv.aaw1268.

[21] Drell, J. R. “Neanderthals: A history of interpretation”. In: *Oxford Journal of Archaeology* 19.1 (Feb. 2000), pp. 1–24. ISSN: 02625253. DOI: 10.1111/1468-0092.00096.

[22] Helmuth, H. “Body height, body mass and surface area of the Neanderthals.” In: *Zeitschrift für Morphologie und Anthropologie* 82.1 (1998), pp. 1–12. ISSN: 0044314X. DOI: 10.1127/zma/82/1998/1.

[23] Gómez-Olivencia, A., Barash, A., García-Martínez, D., Arlegi, M., Kramer, P., Bastir, M., and Been, E. “3D virtual reconstruction of the Kebara 2 Neandertal thorax”. In: *Nature Communications* 9.1 (Oct. 2018), pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-018-06803-z.

[24] Ahern, J. C. M. “Neanderthals Revisited: New Approaches and Perspectives”. In: *Neanderthals Revisited. New Approaches and Perspectives* (2006), p. 333.

[25] Kolobova, K. A., Roberts, R. G., Chabai, V. P., et al. “Archaeological evidence for two separate dispersals of Neanderthals into southern Siberia”. In: *Proceedings of the National Academy of Sciences of the United States of America* 117.6 (Feb. 2020), pp. 2879–2885. ISSN: 10916490. DOI: 10.1073/pnas.1918047117.

[26] Ahlquist, K. D., Bañuelos, M. M., Funk, A., Lai, J., Rong, S., Villanea, F. A., and Witt, K. E. “Our Tangled Family Tree: New Genomic Methods Offer Insight into the Legacy of Archaic Admixture”. In: *Genome biology and evolution* 13.7 (July 2021). ISSN: 17596653. DOI: 10.1093/gbe/evab115.

[27] Slon, V., Mafessoni, F., Vernot, B., et al. “The genome of the offspring of a Neanderthal mother and a Denisovan father”. In: *Nature* 561.7721 (Aug. 2018), pp. 113–116. ISSN: 14764687. DOI: 10.1038/s41586-018-0455-x.

[28] Reich, D., Green, R. E., Kircher, M., et al. “Genetic history of an archaic hominin group from Denisova cave in Siberia”. In: *Nature* 468.7327 (2010), pp. 1053–1060. ISSN: 00280836. DOI: 10.1038/nature09710.

[29] Green, R. E., Krause, J., Briggs, A. W., et al. “A draft sequence of the neandertal genome”. In: *Science* 328.5979 (May 2010), pp. 710–722. ISSN: 00368075. DOI: 10.1126/science.1188021.

[30] Schaefer, N. K., Shapiro, B., and Green, R. E. “An ancestral recombination graph of human, Neanderthal, and Denisovan genomes”. In: *Science Advances* 7.29 (July 2021), pp. 776–792. ISSN: 23752548. DOI: 10.1126/sciadv.abc0776.

[31] Skov, L., Coll Macià, M., Sveinbjörnsson, G., et al. “The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes”. In: *Nature* 582.7810 (June 2020), pp. 78–83. ISSN: 14764687. DOI: 10.1038/s41586-020-2225-9.

[32] Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., and Akey, J. M. “Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture”. In: *Cell* 173.1 (2018), 53–61.e9. ISSN: 10974172. DOI: 10.1016/j.cell.2018.02.031.

[33] Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. *Evidence for archaic adaptive introgression in humans*. 2015. DOI: 10.1038/nrg3936.

[34] Racimo, F., Marnetto, D., and Huerta-Sánchez, E. “Signatures of archaic adaptive introgression in present-day human populations”. In: *Molecular Biology and Evolution* 34.2 (2017), pp. 296–317. ISSN: 15371719. DOI: 10.1093/molbev/msw216.

[35] Gittelman, R. M., Schraiber, J. G., Vernot, B., Mikacenic, C., Wurfel, M. M., and Akey, J. M. “Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments”. In: *Current Biology* 26.24 (2016), pp. 3375–3382. ISSN: 09609822. DOI: 10.1016/j.cub.2016.10.041.

[36] Abi-Rached, L., Jobin, M. J., Kulkarni, S., et al. “The shaping of modern human immune systems by multiregional admixture with archaic humans”. In: *Science* 334.6052 (2011), pp. 89–94. ISSN: 10959203. DOI: 10.1126/science.1209202.

[37] Mendez, F. L., Watkins, J. C., and Hammer, M. F. “A haplotype at STAT2 introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea”. In: *American Journal of Human Genetics* 91.2 (2012), pp. 265–274. ISSN: 00029297. DOI: 10.1016/j.ajhg.2012.06.015.

[38] Sams, A. J., Dumaine, A., Nédélec, Y., Yotova, V., Alfieri, C., Tanner, J. E., Messer, P. W., and Barreiro, L. B. *Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans*. 2016. DOI: 10.1186/s13059-016-1098-6.

[39] Dannemann, M., Andrés, A. M., and Kelso, J. “Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors”. In: *American Journal of Human Genetics* 98.1 (2016), pp. 22–33. ISSN: 15376605. DOI: 10.1016/j.ajhg.2015.11.015.

[40] Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J. L., Patin, E., and Quintana-Murci, L. “Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes”. In: *American Journal of Human Genetics* 98.1 (2016), pp. 5–21. ISSN: 15376605. DOI: 10.1016/j.ajhg.2015.11.014.

[41] Quach, H., Rotival, M., Pothlichet, J., et al. “Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations”. In: *Cell* 167.3 (2016), 643–656.e17. ISSN: 10974172. DOI: 10.1016/j.cell.2016.09.024.

[42] Nédélec, Y., Sanz, J., Baharian, G., et al. “Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens”. In: *Cell* 167.3 (2016), 657–669.e21. ISSN: 10974172. DOI: 10.1016/j.cell.2016.09.025.

[43] Enard, D. and Petrov, D. A. “Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans”. In: *Cell* 175.2 (2018), 360–371.e13. ISSN: 10974172. DOI: 10.1016/j.cell.2018.08.034.

[44] Vernot, B. and Akey, J. M. “Resurrecting surviving Neandertal lineages from modern human genomes”. In: *Science* 343.6174 (2014), pp. 1017–1021. ISSN: 10959203. DOI: 10.1126/science.1245938.

[45] Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. “The genomic landscape of Neanderthal ancestry in present-day humans”. In: *Nature* 507.7492 (2014), pp. 354–357. ISSN: 14764687. DOI: 10.1038/nature12961.

[46] Ding, Q., Hu, Y., Xu, S., Wang, J., and Jin, L. “Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in east asians”. In: *Molecular Biology and Evolution* 31.3 (2014), pp. 683–695. ISSN: 15371719. DOI: 10.1093/molbev/mst260.

[47] Huerta-Sánchez, E., Jin, X., Asan, et al. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA". In: *Nature* 512.7513 (2014), pp. 194–197. ISSN: 14764687. DOI: 10.1038/nature13408.

[48] Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sanchez, E., and Nielsen, R. "Archaic adaptive introgression in TBX15/WARS2". In: *Molecular Biology and Evolution* 34.3 (2017), pp. 509–524. ISSN: 15371719. DOI: 10.1093/molbev/msw283.

[49] Khrameeva, E. E., Bozek, K., He, L., et al. "Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans". In: *Nature Communications* 5 (2014). ISSN: 20411723. DOI: 10.1038/ncomms4584.

[50] Gouy, A., Excoffier, L., and Nielsen, R. "Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens". In: *Molecular Biology and Evolution* 37.5 (May 2020), pp. 1420–1433. ISSN: 15371719. DOI: 10.1093/molbev/msz306.

[51] Harris, K. and Nielsen, R. "The genetic cost of neanderthal introgression". In: *Genetics* 203.2 (2016), pp. 881–891. ISSN: 19432631. DOI: 10.1534/genetics.116.186890.

[52] Juric, I., Aeschbacher, S., and Coop, G. "The Strength of Selection against Neanderthal Introgression". In: *PLoS Genetics* 12.11 (2016). ISSN: 15537404. DOI: 10.1371/journal.pgen.1006340.

[53] Petr, M., Pääbo, S., Kelso, J., and Vernot, B. "Limits of long-term selection against Neandertal introgression". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.5 (Jan. 2019), pp. 1639–1644. ISSN: 10916490. DOI: 10.1073/pnas.1814338116.

[54] Hajdinjak, M., Mafessoni, F., Skov, L., et al. "Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry". In: *Nature* 592.7853 (Apr. 2021), pp. 253–257. ISSN: 14764687. DOI: 10.1038/s41586-021-03335-3.

[55] McCoy, R. C., Wakefield, J., and Akey, J. M. "Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression". In: *Cell* 168.5 (2017), 916–927.e12. ISSN: 10974172. DOI: 10.1016/j.cell.2017.01.038.

[56] Telis, N., Aguilar, R., and Harris, K. "Selection against archaic hominin genetic variation in regulatory regions". In: *Nature Ecology and Evolution* 4.11 (Aug. 2020), pp. 1558–1566. ISSN: 2397334X. DOI: 10.1038/s41559-020-01284-0.

[57] Dannemann, M., Prüfer, K., and Kelso, J. "Functional implications of Neandertal introgression in modern humans". In: *Genome Biology* 18.1 (Apr. 2017), pp. 1–11. ISSN: 1474760X. DOI: 10.1186/s13059-017-1181-7.

[58] Rinker, D. C., Simonti, C. N., McArthur, E., Shaw, D., Hodges, E., and Capra, J. A. "Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations". In: *Nature Ecology and Evolution* 4.10 (July 2020), pp. 1332–1341. ISSN: 2397334X. DOI: 10.1038/s41559-020-1261-z.

[59] Simonti, C. N., Vernot, B., Bastarache, L., et al. "The phenotypic legacy of admixture between modern humans and Neandertals". In: *Science* 351.6274 (Feb. 2016), pp. 737–741. ISSN: 10959203. DOI: 10.1126/science.aad2149.

[60] Jégou, B., Sankararaman, S., Rolland, A. D., Reich, D., and Chalmel, F. "Meiotic genes are enriched in regions of reduced archaic ancestry". In: *Molecular Biology and Evolution* 34.8 (2017), pp. 1974–1980. ISSN: 15371719. DOI: 10.1093/molbev/msx141.

[61] Schumer, M., Xu, C., Powell, D. L., et al. “Natural selection interacts with recombination to shape the evolution of hybrid genomes”. In: *Science* 360.6389 (2018), pp. 656–660. ISSN: 10959203. DOI: 10.1126/science.aar3684.

[62] Dannemann, M. and Kelso, J. “The Contribution of Neanderthals to Phenotypic Variation in Modern Humans”. In: *American Journal of Human Genetics* 101.4 (2017), pp. 578–589. ISSN: 15376605. DOI: 10.1016/j.ajhg.2017.09.010.

[63] McArthur, E., Rinker, D. C., and Capra, J. A. “Quantifying the contribution of Neanderthal introgression to the heritability of complex traits”. In: *Nature Communications* 12.1 (July 2021), p. 2020.06.08.140087. ISSN: 20411723. DOI: 10.1038/s41467-021-24582-y.

[64] Gunz, P., Tilot, A. K., Wittfeld, K., et al. “Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity”. In: *Current Biology* 29.1 (Jan. 2019), 120–127.e5. ISSN: 09609822. DOI: 10.1016/j.cub.2018.10.065.

[65] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. *10 Years of GWAS Discovery: Biology, Function, and Translation*. 2017. DOI: 10.1016/j.ajhg.2017.06.005.

[66] Boyle, E. A., Li, Y. I., and Pritchard, J. K. *An Expanded View of Complex Traits: From Polygenic to Omnigenic*. June 2017. DOI: 10.1016/j.cell.2017.05.038.

[67] Nowak, R. “Mining treasures from ’junk DNA’”. In: *Science* 263.5147 (1994), pp. 608–610. ISSN: 00368075. DOI: 10.1126/science.7508142.

[68] Nurk, S., Koren, S., Rhie, A., et al. “The complete sequence of a human genome”. In: *bioRxiv* (May 2021), p. 2021.05.26.445798. DOI: 10.1101/2021.05.26.445798.

[69] Gorzynski, J. E., Goenka, S. D., Shafin, K., et al. “Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting”. In: *New England Journal of Medicine* (Jan. 2022). ISSN: 0028-4793. DOI: 10.1056/nejmc2112090.

[70] Karczewski, K. J., Francioli, L. C., Tiao, G., et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. In: *Nature* 581.7809 (May 2020), pp. 434–443. ISSN: 14764687. DOI: 10.1038/s41586-020-2308-7.

[71] Salzberg, S. L. “Open questions: How many genes do we have?” In: *BMC Biology* 16.1 (Aug. 2018). ISSN: 17417007. DOI: 10.1186/s12915-018-0564-x.

[72] Franchini, L. F. and Pollard, K. S. “Human evolution: The non-coding revolution”. In: *BMC Biology* 15.1 (Oct. 2017). ISSN: 17417007. DOI: 10.1186/s12915-017-0428-9.

[73] Rojano, E., Seoane, P., Ranea, J. A., and Perkins, J. R. “Regulatory variants: From detection to predicting impact”. In: *Briefings in Bioinformatics* 20.5 (June 2019), pp. 1639–1654. ISSN: 14774054. DOI: 10.1093/bib/bby039.

[74] Lee, T. I. and Young, R. A. “Transcriptional regulation and its misregulation in disease”. In: *Cell* 152.6 (Mar. 2013), pp. 1237–1251. ISSN: 10974172. DOI: 10.1016/j.cell.2013.02.014.

[75] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1228–1235. ISSN: 15461718. DOI: 10.1038/ng.3404.

[76] Finucane, H. K., Reshef, Y. A., Anttila, V., et al. “Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types”. In: *Nature Genetics* 50.4 (Apr. 2018), pp. 621–629. ISSN: 15461718. DOI: 10.1038/s41588-018-0081-4.

[77] Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. "A general framework for estimating the relative pathogenicity of human genetic variants". In: *Nature Genetics* 46.3 (2014), pp. 310–315. ISSN: 15461718. DOI: 10.1038/ng.2892.

[78] Krijger, P. H. L. and De Laat, W. "Regulation of disease-associated gene expression in the 3D genome". In: *Nature Reviews Molecular Cell Biology* 17.12 (Dec. 2016), pp. 771–782. ISSN: 14710080. DOI: 10.1038/nrm.2016.138.

[79] Zhu, Y., Tazearslan, C., and Suh, Y. "Challenges and progress in interpretation of non-coding genetic variants associated with human disease". In: *Experimental Biology and Medicine* 242.13 (July 2017), pp. 1325–1334. ISSN: 15353699. DOI: 10.1177/1535370217713750.

[80] Boyle, A. P., Hong, E. L., Hariharan, M., et al. "Annotation of functional variation in personal genomes using RegulomeDB". In: *Genome Research* 22.9 (Sept. 2012), pp. 1790–1797. ISSN: 10889051. DOI: 10.1101/gr.137323.112.

[81] Edwards, S. L., Beesley, J., French, J. D., and Dunning, M. "Beyond GWASs: Illuminating the dark road from association to function". In: *American Journal of Human Genetics* 93.5 (Nov. 2013), pp. 779–797. ISSN: 15376605. DOI: 10.1016/j.ajhg.2013.10.012.

[82] He, Z., Liu, L., Wang, K., and Ionita-Laza, I. "A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs". In: *Nature Communications* 9.1 (Dec. 2018), p. 5199. ISSN: 20411723. DOI: 10.1038/s41467-018-07349-w.

[83] Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., and Beer, M. A. "A method to predict the impact of regulatory variants from DNA sequence". In: *Nature Genetics* 47.8 (2015), pp. 955–961. ISSN: 15461718. DOI: 10.1038/ng.3331.

[84] Zhou, J. and Troyanskaya, O. G. "Predicting effects of noncoding variants with deep learning-based sequence model". In: *Nature Methods* 12.10 (2015), pp. 931–934. ISSN: 15487105. DOI: 10.1038/nmeth.3547.

[85] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *American Journal of Human Genetics* 101.1 (July 2017), pp. 5–22. ISSN: 15376605. DOI: 10.1016/j.ajhg.2017.06.005.

[86] Manolio, T. A., Collins, F. S., Cox, N. J., et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (Oct. 2009), pp. 747–753. ISSN: 00280836. DOI: 10.1038/nature08494.

[87] Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. "Concepts, estimation and interpretation of SNP-based heritability". In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1304–1310. ISSN: 15461718. DOI: 10.1038/ng.3941.

[88] Yang, J., Manolio, T. A., Pasquale, L. R., et al. "Genome partitioning of genetic variation for complex traits using common SNPs". In: *Nature Genetics* 43.6 (May 2011), pp. 519–525. ISSN: 15461718. DOI: 10.1038/ng.823.

[89] Lee, S. H., Decandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., and Wray, N. R. "Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs". In: *Nature Genetics* 44.3 (Feb. 2012), pp. 247–250. ISSN: 10614036. DOI: 10.1038/ng.1108.

[90] Bulik-Sullivan, B., Loh, P. R., Finucane, H. K., et al. "LD score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (Feb. 2015), pp. 291–295. ISSN: 15461718. DOI: 10.1038/ng.3211.

[91] Davis, L. K., Yu, D., Keenan, C. L., et al. "Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture". In: *PLoS Genetics* 9.10 (2013), e1003864. ISSN: 15537404. DOI: 10.1371/journal.pgen.1003864.

[92] Hujoel, M. L., Gazal, S., Hormozdiari, F., Geijn, B. van de, and Price, A. L. "Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species". In: *American Journal of Human Genetics* 104.4 (Apr. 2019), pp. 611–624. ISSN: 15376605. DOI: 10.1016/j.ajhg.2019.02.008.

[93] Castellano, S., Parra, G., Sánchez-Quinto, F. A., et al. "Patterns of coding variation in the complete exomes of three Neandertals". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.18 (May 2014), pp. 6666–6671. ISSN: 10916490. DOI: 10.1073/pnas.1405138111.

[94] Colbran, L. L., Gamazon, E. R., Zhou, D., Evans, P., Cox, N. J., and Capra, J. A. "Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences". In: *Nature Ecology and Evolution* 3.11 (2019), pp. 1598–1606. ISSN: 2397334X. DOI: 10.1038/s41559-019-0996-x.

[95] Lonsdale, J., Thomas, J., Salvatore, M., et al. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (June 2013), pp. 580–585. ISSN: 10614036. DOI: 10.1038/ng.2653.

[96] Gokhman, D., Lavi, E., Prüfer, K., Fraga, M. F., Riancho, J. A., Kelso, J., Pääbo, S., Meshorer, E., and Carmel, L. "Reconstructing the DNA methylation maps of the neandertal and the Denisovan". In: *Science* 344.6183 (May 2014), pp. 523–527. ISSN: 10959203. DOI: 10.1126/science.1250368.

[97] Batyrev, D., Lapid, E., Carmel, L., and Meshorer, E. "Predicted Archaic 3D Genome Organization Reveals Genes Related to Head and Spinal Cord Separating Modern from Archaic Humans". In: *Cells* 9.1 (Dec. 2019). ISSN: 20734409. DOI: 10.3390/cells9010048.

[98] Silvert, M., Quintana-Murci, L., and Rotival, M. "Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation". In: *American Journal of Human Genetics* 104.6 (June 2019), pp. 1241–1250. ISSN: 15376605. DOI: 10.1016/j.ajhg.2019.04.016.

[99] Sullivan, L. H. *The tall office building artistically considered*. Philadelphia: J.B. Lippincott Co., 1896.

[100] Boveri, T. "Die Blastomerenkerne von Ascaris megalocephala und die Theorie der Chromosomenindividualität". In: *Archiv für Zellforschung* 3 (1909), pp. 181–268.

[101] Rabl, C. "U¨ ber Zellteilung". In: ().

[102] Rowley, M. J. and Corces, V. G. "Organizational principles of 3D genome architecture". In: *Nature Reviews Genetics* 19.12 (Dec. 2018), pp. 789–800. ISSN: 14710064. DOI: 10.1038/s41576-018-0060-8.

[103] Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. "Capturing chromosome conformation". In: *Science* 295.5558 (2002), pp. 1306–1311. ISSN: 00368075. DOI: 10.1126/science.1067799.

[104] Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *Science* 326.5950 (Oct. 2009), pp. 289–293. ISSN: 00368075. DOI: 10.1126/science.1181369.

[105] Dixon, J. R., Gorkin, D. U., and Ren, B. "Chromatin Domains: The Unit of Chromosome Organization". In: *Molecular Cell* 62.5 (June 2016), pp. 668–680. ISSN: 10974164. DOI: 10.1016/j.molcel.2016.05.018.

[106] Krietenstein, N., Abraham, S., Venev, S. V., et al. "Ultrastructural Details of Mammalian Chromosome Architecture". In: *Molecular Cell* 78.3 (May 2020), 554–565.e7. ISSN: 10974164. DOI: 10.1016/j.molcel.2020.03.003.

[107] Hsieh, T. H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Rando, O. J., Tjian, R., and Darzacq, X. "Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding". In: *Molecular Cell* 78.3 (May 2020), 539–553.e8. ISSN: 10974164. DOI: 10.1016/j.molcel.2020.03.002.

[108] Vian, L., Pękowska, A., Rao, S. S., et al. "The Energetics and Physiological Impact of Cohesin Extrusion". In: *Cell* 173.5 (May 2018), 1165–1178.e20. ISSN: 10974172. DOI: 10.1016/j.cell.2018.03.072.

[109] Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. "Formation of Chromosomal Domains by Loop Extrusion". In: *Cell Reports* 15.9 (May 2016), pp. 2038–2049. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.04.085.

[110] Kraft, K., Magg, A., Heinrich, V., et al. "Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations". In: *Nature Cell Biology* 21.3 (Feb. 2019), pp. 305–310. ISSN: 14764679. DOI: 10.1038/s41556-019-0273-x.

[111] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398 (Apr. 2012), pp. 376–380. ISSN: 00280836. DOI: 10.1038/nature11082.

[112] Nora, E. P., Lajoie, B. R., Schulz, E. G., et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485.7398 (2012), pp. 381–385. ISSN: 00280836. DOI: 10.1038/nature11049.

[113] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. "Three-dimensional folding and functional organization principles of the Drosophila genome". In: *Cell* 148.3 (2012), pp. 458–472. ISSN: 10974172. DOI: 10.1016/j.cell.2012.01.010.

[114] Jackson, D. A. and Pombo, A. "Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells". In: *Journal of Cell Biology* 140.6 (1998), pp. 1285–1295. ISSN: 00219525. DOI: 10.1083/jcb.140.6.1285.

[115] Ma, H., Samarabandu, J., Devdhar, R. S., Acharya, R., Cheng, P. C., Meng, C., and Berezney, R. "Spatial and temporal dynamics of DNA replication sites in mammalian cells". In: *Journal of Cell Biology* 143.6 (1998), pp. 1415–1425. ISSN: 00219525. DOI: 10.1083/jcb.143.6.1415.

[116] Dixon, J. R., Jung, I., Selvaraj, S., et al. "Chromatin architecture reorganization during stem cell differentiation". In: *Nature* 518.7539 (Feb. 2015), pp. 331–336. ISSN: 14764687. DOI: 10.1038/nature14222.

[117] Rao, S. S., Huntley, M. H., Durand, N. C., et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7 (Dec. 2014), pp. 1665–1680. ISSN: 10974172. DOI: 10.1016/j.cell.2014.11.021.

[118] Schmitt, A. D., Hu, M., Jung, I., et al. "A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome". In: *Cell Reports* 17.8 (Nov. 2016), pp. 2042–2059. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.10.061.

[119] Dekker, J. *Two ways to fold the genome during the cell cycle: Insights obtained with chromosome conformation capture*. 2014. DOI: 10.1186/1756-8935-7-25.

[120] Dekker, J. and Heard, E. *Structural and functional diversity of Topologically Associating Domains.* 2015. DOI: 10.1016/j.febslet.2015.08.044.

[121] Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. "Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture". In: *Cell Reports* 10.8 (Mar. 2015), pp. 1297–1309. ISSN: 22111247. DOI: 10.1016/j.celrep.2015.02.004.

[122] Spielmann, M., Lupiáñez, D. G., and Mundlos, S. "Structural variation in the 3D genome". In: *Nature Reviews Genetics* 19.7 (July 2018), pp. 453–467. ISSN: 14710064. DOI: 10.1038/s41576-018-0007-0.

[123] Krefting, J., Andrade-Navarro, M. A., and Ibn-Salem, J. "Evolutionary stability of topologically associating domains is associated with conserved gene regulation". In: *BMC Biology* 16.1 (Dec. 2018), p. 87. ISSN: 17417007. DOI: 10.1186/s12915-018-0556-x.

[124] Whalen, S. and Pollard, K. S. "Most chromatin interactions are not in linkage disequilibrium". In: *Genome Research* 29.3 (Jan. 2019), pp. 334–343. ISSN: 15495469. DOI: 10.1101/gr.238022.118.

[125] Fudenberg, G., Getz, G., Meyerson, M., and Mirny, L. A. "High order chromatin architecture shapes the landscape of chromosomal alterations in cancer". In: *Nature Biotechnology* 29.12 (2011), pp. 1109–1113. ISSN: 10870156. DOI: 10.1038/nbt.2049.

[126] Hnisz, D., Weintrau, A. S., Day, D. S., et al. "Activation of proto-oncogenes by disruption of chromosome neighborhoods". In: *Science* 351.6280 (2016), pp. 1454–1458. ISSN: 10959203. DOI: 10.1126/science.aad9024.

[127] Meaburn, K. J., Gudla, P. R., Khan, S., Lockett, S. J., and Misteli, T. "Disease-specific gene repositioning in breast cancer". In: *Journal of Cell Biology* 187.6 (2009), pp. 801–812. ISSN: 00219525. DOI: 10.1083/jcb.200909127.

[128] Misteli, T. *Higher-order genome organization in human disease.* 2010. DOI: 10.1101/cshperspect.a000794.

[129] Bonev, B., Mendelson Cohen, N., Szabo, Q., et al. "Multiscale 3D Genome Rewiring during Mouse Neural Development". In: *Cell* 171.3 (Oct. 2017), 557–572.e24. ISSN: 10974172. DOI: 10.1016/j.cell.2017.09.043.

[130] Bruijn, S. E. de, Fiorentino, A., Ottaviani, D., et al. "Structural Variants Create New Topological-Associated Domains and Ectopic Retinal Enhancer-Gene Contact in Dominant Retinitis Pigmentosa". In: *American Journal of Human Genetics* 107.5 (Oct. 2020), pp. 802–814. ISSN: 15376605. DOI: 10.1016/j.ajhg.2020.09.002.

[131] Pope, B. D., Ryba, T., Dileep, V., et al. "Topologically associating domains are stable units of replication-timing regulation". In: *Nature* 515.7527 (Nov. 2014), pp. 402–405. ISSN: 14764687. DOI: 10.1038/nature13986.

[132] Lupiáñez, D. G., Kraft, K., Heinrich, V., et al. "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". In: *Cell* 161.5 (May 2015), pp. 1012–1025. ISSN: 10974172. DOI: 10.1016/j.cell.2015.04.004.

[133] Gröschel, S., Sanders, M. A., Hoogenboezem, R., et al. "A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia". In: *Cell* 157.2 (2014), pp. 369–381. ISSN: 10974172. DOI: 10.1016/j.cell.2014.02.019.

[134] Northcott, P. A., Lee, C., Zichner, T., et al. "Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma". In: *Nature* 511.7510 (2014), pp. 428–434. ISSN: 14764687. DOI: 10.1038/nature13379.

[135] Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., and Bernstein, B. E. “Insulator dysfunction and oncogene activation in IDH mutant gliomas”. In: *Nature* 529.7584 (2016), pp. 110–114. ISSN: 14764687. DOI: 10.1038/nature16490.

[136] Gorkin, D. U., Qiu, Y., Hu, M., et al. “Common DNA sequence variation influences 3-dimensional conformation of the human genome”. In: *Genome Biology* 20.1 (Nov. 2019), pp. 1–25. ISSN: 1474760X. DOI: 10.1186/s13059-019-1855-4.

[137] Fudenberg, G. and Pollard, K. S. “Chromatin features constrain structural variation across evolutionary timescales”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.6 (Feb. 2019), pp. 2175–2180. ISSN: 10916490. DOI: 10.1073/pnas.1808631116.

[138] Dill, K. A., Ozkan, S. B., Shell, M. S., and Weikl, T. R. “The protein folding problem”. In: *Annual Review of Biophysics* 37 (2008), pp. 289–316. ISSN: 1936122X. DOI: 10.1146/annurev.biophys.37.092707.153558.

[139] Jumper, J., Evans, R., Pritzel, A., et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (July 2021), pp. 583–589. ISSN: 14764687. DOI: 10.1038/s41586-021-03819-2.

[140] Belokopytova, P. and Fishman, V. “Predicting Genome Architecture: Challenges and Solutions”. In: *Frontiers in Genetics* 11 (Jan. 2021), p. 1776. ISSN: 16648021. DOI: 10.3389/fgene.2020.617202.

[141] Fudenberg, G., Kelley, D. R., and Pollard, K. S. “Predicting 3D genome folding from DNA sequence with Akita”. In: *Nature Methods* 17.11 (Oct. 2020), pp. 1111–1117. ISSN: 15487105. DOI: 10.1038/s41592-020-0958-x.

[142] Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., Teh, Y. W., Lunter, G., and Hughes, J. R. “DeepC: predicting 3D genome folding using megabase-scale transfer learning”. In: *Nature Methods* 17.11 (Oct. 2020), pp. 1118–1124. ISSN: 15487105. DOI: 10.1038/s41592-020-0960-3.

[143] Zhou, J. “Sequence-based modeling of genome 3D architecture from kilobase to chromosome-scale”. In: *bioRxiv* (May 2021), p. 2021.05.19.444847. DOI: 10.1101/2021.05.19.444847.

[144] Sudlow, C., Gallacher, J., Allen, N., et al. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLoS Medicine* 12.3 (Mar. 2015), e1001779. ISSN: 15491676. DOI: 10.1371/journal.pmed.1001779.

[145] Hormozdiari, F., Gazal, S., Van De Geijn, B., et al. “Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits”. In: *Nature Genetics* 50.7 (2018), pp. 1041–1047. ISSN: 15461718. DOI: 10.1038/s41588-018-0148-2.

[146] Boraska, V., Franklin, C. S., Floyd, J. A., et al. “A genome-wide association study of anorexia nervosa”. In: *Molecular Psychiatry* 19.10 (2014), pp. 1085–1094. ISSN: 14765578. DOI: 10.1038/mp.2013.187.

[147] Smoller, J. W., Kendler, K., Craddock, N., et al. “Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis”. In: *The Lancet* 381.9875 (2013), pp. 1371–1379. ISSN: 1474547X. DOI: 10.1016/S0140-6736(12)62129-1.

[148] Jostins, L., Ripke, S., Weersma, R. K., et al. “Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease”. In: *Nature* 491.7422 (2012), pp. 119–124. ISSN: 14764687. DOI: 10.1038/nature11582.

[149] Okbay, A., Baselmans, B. M., De Neve, J. E., et al. “Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses”. In: *Nature Genetics* 48.6 (2016), pp. 624–633. ISSN: 15461718. DOI: 10.1038/ng.3552.

[150] Barban, N., Jansen, R., De Vlaming, R., et al. “Genome-wide analysis identifies 12 loci influencing human reproductive behavior”. In: *Nature Genetics* 48.12 (2016), pp. 1462–1472. ISSN: 15461718. DOI: 10.1038/ng.3698.

[151] Teslovich, T. M., Musunuru, K., Smith, A. V., et al. “Biological, clinical and population relevance of 95 loci for blood lipids”. In: *Nature* 466.7307 (2010), pp. 707–713. ISSN: 14764687. DOI: 10.1038/nature09270.

[152] Okada, Y., Wu, D., Trynka, G., et al. “Genetics of rheumatoid arthritis contributes to biology and drug discovery”. In: *Nature* 506.7488 (2014), pp. 376–381. ISSN: 00280836. DOI: 10.1038/nature12873.

[153] Ripke, S., Neale, B. M., Corvin, A., et al. “Biological insights from 108 schizophrenia-associated genetic loci”. In: *Nature* 511.7510 (2014), pp. 421–427. ISSN: 14764687. DOI: 10.1038/nature13595.

[154] McArthur, E. and Capra, J. A. “Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability”. In: *American Journal of Human Genetics* 108.2 (Feb. 2021), pp. 269–283. ISSN: 15376605. DOI: 10.1016/j.ajhg.2021.01.001.

[155] Gazal, S., Finucane, H. K., Furlotte, N. A., et al. “Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection”. In: *Nature Genetics* 49.10 (2017), pp. 1421–1427. ISSN: 15461718. DOI: 10.1038/ng.3954.

[156] Vernot, B., Tucci, S., Kelso, J., et al. “Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals”. In: *Science* 352.6282 (2016), pp. 235–239. ISSN: 10959203. DOI: 10.1126/science.aad9416.

[157] *Neale Lab: Heritability of ¿4,000 traits & disorders in UK Biobank*. 2018.

[158] FinnGen. *FinnGen research project*. 2018.

[159] Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., Leeuw, C. de, Polderman, T. J., Sluis, S. van der, Andreassen, O. A., Neale, B. M., and Posthuma, D. “A global overview of pleiotropy and genetic architecture in complex traits”. In: *Nature Genetics* 51.9 (2019), pp. 1339–1348. ISSN: 15461718. DOI: 10.1038/s41588-019-0481-0.

[160] Reshef, Y. A., Finucane, H. K., Kelley, D. R., et al. “Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk”. In: *Nature Genetics* 50.10 (Oct. 2018), pp. 1483–1493. ISSN: 15461718. DOI: 10.1038/s41588-018-0196-7.

[161] Jacobs, L. C., Wollstein, A., Lao, O., Hofman, A., Klaver, C. C., Uitterlinden, A. G., Nijsten, T., Kayser, M., and Liu, F. “Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans”. In: *Human Genetics* 132.2 (2013), pp. 147–158. ISSN: 03406717. DOI: 10.1007/s00439-012-1232-9.

[162] Nantel, A., Mohammad-Ali, K., Sherk, J., Posner, B. I., and Thomas, D. Y. “Interaction of the Grb10 adapter protein with the Raf1 and MEK1 kinases”. In: *Journal of Biological Chemistry* 273.17 (1998), pp. 10475–10484. ISSN: 00219258. DOI: 10.1074/jbc.273.17.10475.

[163] Gorlova, O., Martin, J. E., Rueda, B., et al. “Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy”. In: *PLoS Genetics* 7.7 (2011). ISSN: 15537390. DOI: 10.1371/journal.pgen.1002178.

[164] Moreno-Moral, A., Bagnati, M., Koturan, S., et al. “Changes in macrophage transcriptome associate with systemic sclerosis and mediate GSDMA contribution to disease risk”. In: *Annals of the Rheumatic Diseases* 77.4 (2018), pp. 596–601. ISSN: 14682060. DOI: 10.1136/annrheumdis-2017-212454.

[165] Yan, X., Himburg, H. A., Pohl, K., et al. "Deletion of the Imprinted Gene Grb10 Promotes Hematopoietic Stem Cell Self-Renewal and Regeneration". In: *Cell Reports* 17.6 (2016), pp. 1584–1594. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.10.025.

[166] Hu, X., Kim, H., Raj, T., et al. "Regulation of Gene Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4+ Effector Memory T Cells". In: *PLoS Genetics* 10.6 (2014). ISSN: 15537404. DOI: 10.1371/journal.pgen.1004404.

[167] Plasschaert, R. N. and Bartolomei, M. S. "Tissue-specific regulation and function of Grb10 during growth and neuronal commitment". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.22 (2015), pp. 6841–6847. ISSN: 10916490. DOI: 10.1073/pnas.1411254111.

[168] Power, R. A., Kyaga, S., Uher, R., MacCabe, J. H., Långström, N., Landen, M., McGuffin, P., Lewis, C. M., Lichtenstein, P., and Svensson, A. C. "Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings". In: *Archives of General Psychiatry* 70.1 (2013), pp. 22–30. ISSN: 0003990X. DOI: 10.1001/jamapsychiatry.2013.268.

[169] Gazal, S., Loh, P. R., Finucane, H. K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A. L. "Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations". In: *Nature Genetics* 50.11 (2018), pp. 1600–1607. ISSN: 15461718. DOI: 10.1038/s41588-018-0231-8.

[170] Speed, D. and Balding, D. J. "SumHer better estimates the SNP heritability of complex traits from summary statistics". In: *Nature Genetics* 51.2 (Feb. 2019), pp. 277–284. ISSN: 15461718. DOI: 10.1038/s41588-018-0279-5.

[171] Hou, K., Burch, K. S., Majumdar, A., Shi, H., Mancuso, N., Wu, Y., Sankararaman, S., and Pasaniuc, B. "Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture". In: *Nature Genetics* 51.8 (Aug. 2019), pp. 1244–1251. ISSN: 15461718. DOI: 10.1038/s41588-019-0465-0.

[172] Ishigaki, K., Akiyama, M., Kanai, M., et al. "Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases". In: *Nature Genetics* 52.7 (July 2020), pp. 669–679. ISSN: 15461718. DOI: 10.1038/s41588-020-0640-3.

[173] Taskent, O., Lin, Y. L., Patramanis, I., Pavlidis, P., and Gokcumen, O. "Analysis of haplotypic variation and deletion polymorphisms point to multiple archaic introgression events, including from altai neanderthal lineage". In: *Genetics* 215.2 (June 2020), pp. 497–509. ISSN: 19432631. DOI: 10.1534/genetics.120.303167.

[174] *Neale Lab: Defining Confidence Levels for UKB Round 2 LDSR Analyses*. 2019.

[175] Auton, A., Abecasis, G. R., Altshuler, D. M., et al. *A global reference for human genetic variation*. 2015. DOI: 10.1038/nature15393.

[176] Altshuler, D. M., Gibbs, R. A., Peltonen, L., et al. "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311 (Sept. 2010), pp. 52–58. ISSN: 14764687. DOI: 10.1038/nature09298.

[177] Purcell, S., Neale, B., Todd-Brown, K., et al. "PLINK: A tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795.

[178] Machiela, M. J. and Chanock, S. J. "LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants". In: *Bioinformatics* 31.21 (2015), pp. 3555–3557. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv402.

[179] Haeussler, M., Zweig, A. S., Tyner, C., et al. "The UCSC Genome Browser database: 2019 update". In: *Nucleic Acids Research* 47.D1 (2019), pp. D853–D858. ISSN: 13624962. DOI: 10.1093/nar/gky1095.

[180] Chen, L., Wolf, A. B., Fu, W., Li, L., and Akey, J. M. "Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals". In: *Cell* 180.4 (Feb. 2020), 677–687.e16. ISSN: 10974172. DOI: 10.1016/j.cell.2020.01.012.

[181] Quinlan, A. R. and Hall, I. M. "BEDTools: A flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq033.

[182] Waskom, M., Botvinnik, O., O'Kane, D., et al. "mwaskom/seaborn: v0.9.0 (July 2018)". In: (July 2018). DOI: 10.5281/ZENODO.1313201.

[183] Hunter, J. D. "Matplotlib: A 2D graphics environment". In: *Computing in Science and Engineering* 9.3 (2007), pp. 90–95. ISSN: 15219615. DOI: 10.1109/MCSE.2007.55.

[184] InkscapeProject. *Inkscape*. 2018.

[185] McArthur, E. *emcarthur/trait-h2-neanderthals*. 2021. DOI: http://doi.org/10.5281/zenodo.4900031.

[186] Cavalli, G. and Misteli, T. *Functional implications of genome topology*. 2013. DOI: 10.1038/nsmb.2474.

[187] Cremer, T. and Cremer, C. *Chromosome territories, nuclear architecture and gene regulation in mammalian cells*. 2001. DOI: 10.1038/35066075.

[188] Duggal, G., Wang, H., and Kingsford, C. "Higher-order chromatin domains link eQTLs with the expression of far-away genes". In: *Nucleic Acids Research* 42.1 (2014), pp. 87–96. ISSN: 03051048. DOI: 10.1093/nar/gkt857.

[189] Le Dily, F. L., Baù, D., Pohl, A., et al. "Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation". In: *Genes and Development* 28.19 (2014), pp. 2151–2162. ISSN: 15495477. DOI: 10.1101/gad.241422.114.

[190] Sauerwald, N. and Kingsford, C. "Quantifying the similarity of topological domains across normal and cancer human cell types". In: *Bioinformatics* 34.13 (July 2018), pp. i475–i483. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty265.

[191] Yu, J., Hu, M., and Li, C. "Integrative analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes". In: *bioRxiv* (2018). ISSN: 2692-8205. DOI: 10.1101/392266.

[192] Aguet, F., Brown, A. A., Castel, S. E., et al. "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675 (2017), pp. 204–213. ISSN: 14764687. DOI: 10.1038/nature24277.

[193] Xiao, J., Hafner, A., and Boettiger, A. N. "How subtle changes in 3d structure can create large changes in transcription". In: *eLife* 10 (Oct. 2021), p. 2020.10.22.351395. ISSN: 2050084X. DOI: 10.7554/eLife.64320.

[194] Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O., and Furlong, E. E. "Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression". In: *Nature Genetics* 51.8 (July 2019), pp. 1272–1282. ISSN: 15461718. DOI: 10.1038/s41588-019-0462-3.

[195] Greenwald, W. W., Li, H., Benaglio, P., et al. "Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression". In: *Nature Communications* 10.1 (Dec. 2019), p. 1054. ISSN: 20411723. DOI: 10.1038/s41467-019-08940-5.

[196] Yang, Y., Zhang, Y., Ren, B., Dixon, J. R., and Ma, J. "Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF". In: *Cell Systems* 8.6 (June 2019), 494–505.e14. ISSN: 24054720. DOI: 10.1016/j.cels.2019.05.011.

[197] Han, L., Zhao, X., Benton, M. L., et al. "Functional annotation of rare structural variation in the human brain". In: *Nature Communications* 11.1 (June 2020), p. 2990. ISSN: 20411723. DOI: 10.1038/s41467-020-16736-1.

[198] Sauerwald, N., Singhal, A., and Kingsford, C. "Analysis of the structural variability of topologically associated domains as revealed by Hi-C". In: *NAR Genomics and Bioinformatics* 2.1 (Mar. 2020). ISSN: 2631-9268. DOI: 10.1093/nargab/lqz008.

[199] Eres, I. E. and Gilad, Y. *A TAD Skeptic: Is 3D Genome Topology Conserved?* 2021. DOI: 10.1016/j.tig.2020.10.009.

[200] Gong, Y., Lazaris, C., Sakellaropoulos, T., Lozano, A., Kambadur, P., Ntziachristos, P., Aifantis, I., and Tsirigos, A. "Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries". In: *Nature Communications* 9.1 (Dec. 2018), p. 542. ISSN: 20411723. DOI: 10.1038/s41467-018-03017-1.

[201] An, L., Yang, T., Yang, J., Nuebler, J., Xiang, G., Hardison, R. C., Li, Q., and Zhang, Y. "OnTAD: Hierarchical domain structure reveals the divergence of activity among TADs and boundaries". In: *Genome Biology* 20.1 (Dec. 2019), p. 282. ISSN: 1474760X. DOI: 10.1186/s13059-019-1893-y.

[202] Wang, Y., Song, F., Zhang, B., et al. "The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions". In: *Genome Biology* 19.1 (Dec. 2018), p. 151. ISSN: 1474760X. DOI: 10.1186/s13059-018-1519-9.

[203] Dunham, I., Kundaje, A., Aldred, S. F., et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 14764687. DOI: 10.1038/nature11247.

[204] Davis, C. A., Hitz, B. C., Sloan, C. A., et al. "The Encyclopedia of DNA elements (ENCODE): Data portal update". In: *Nucleic Acids Research* 46.D1 (2018), pp. D794–D801. ISSN: 13624962. DOI: 10.1093/nar/gkx1081.

[205] Leung, D., Jung, I., Rajagopal, N., et al. "Integrative analysis of haplotype-resolved epigenomes across human tissues". In: *Nature* 518.7539 (Feb. 2015), pp. 350–354. ISSN: 14764687. DOI: 10.1038/nature14217.

[206] Lajoie, B. R., Dekker, J., and Kaplan, N. "The Hitchhiker's guide to Hi-C analysis: Practical guidelines". In: *Methods* 72.C (2015), pp. 65–75. ISSN: 10959130. DOI: 10.1016/j.ymeth.2014.10.031.

[207] Hormozdiari, F., Geijn, B. van de, Nasser, J., et al. "Functional disease architectures reveal unique biological role of transposable elements". In: *bioRxiv* (2018). DOI: 10.1101/482281.

[208] Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. "The UCSC table browser data retrieval tool". In: *Nucleic Acids Research* 32.DATABASE ISS. (Jan. 2004). ISSN: 03051048. DOI: 10.1093/nar/gkh103.

[209] Amemiya, H. M., Kundaje, A., and Boyle, A. P. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: *Scientific Reports* 9.1 (2019). ISSN: 20452322. DOI: 10.1038/s41598-019-45839-z.

[210] Torosin, N. S., Anand, A., Golla, T. R., Cao, W., and Ellison, C. E. “3D genome evolution and reorganization in the Drosophila melanogaster species group”. In: *PLoS Genetics* 16.12 (Dec. 2020). Ed. by B. Payseur, e1009229. ISSN: 15537404. DOI: 10.1371/journal.pgen.1009229.

[211] Lumley, T. *rmeta: Meta-Analysis. R package version 3.0.* 2018.

[212] Dali, R. and Blanchette, M. “A critical assessment of topologically associating domain prediction tools”. In: *Nucleic Acids Research* 45.6 (2017), pp. 2994–3005. ISSN: 13624962. DOI: 10.1093/nar/gkx145.

[213] Siepel, A., Bejerano, G., Pedersen, J. S., et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome Research* 15.8 (Aug. 2005), pp. 1034–1050. ISSN: 10889051. DOI: 10.1101/gr.3715005.

[214] O’Leary, N. A., Wright, M. W., Brister, J. R., et al. “Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D733–D745. ISSN: 13624962. DOI: 10.1093/nar/gkv1189.

[215] Eisenberg, E. and Levanon, E. Y. “Human housekeeping genes, revisited”. In: *Trends in Genetics* 29.10 (Oct. 2013), pp. 569–574. ISSN: 01689525. DOI: 10.1016/j.tig.2013.05.010.

[216] Dale, R. K., Pedersen, B. S., and Quinlan, A. R. “Pybedtools: A flexible Python library for manipulating genomic datasets and annotations”. In: *Bioinformatics* 27.24 (2011), pp. 3423–3424. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr539.

[217] McArthur, E. and Capra, J. A. *emcarthur/TAD-stability-heritability*. 2020.

[218] Beagan, J. A. and Phillips-Cremins, J. E. “On the existence and functionality of topologically associating domains”. In: *Nature Genetics* 52.1 (Jan. 2020), pp. 8–16. ISSN: 15461718. DOI: 10.1038/s41588-019-0561-1.

[219] Hsieh, T.-H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Darzacq, X., and Tjian, R. “Enhancer-promoter interactions and transcription are maintained upon acute loss of CTCF, cohesin, WAPL, and YY1”. In: *bioRxiv* (July 2021), p. 2021.07.14.452365. DOI: 10.1101/2021.07.14.452365.

[220] Baur, B., Schreiber, J., Shin, J., Zhang, S., Zhang, Y., Manjunath, M., Song, J. S., Noble, W. S., and Roy, S. “Leveraging epigenomes and three-dimensional genome organization for interpreting regulatory variation”. In: *bioRxiv* (Aug. 2021), p. 2021.08.29.458098. DOI: 10.1101/2021.08.29.458098.

[221] Köhler, S., Gargano, M., Matentzoglu, N., et al. “The human phenotype ontology in 2021”. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D1207–D1217. ISSN: 13624962. DOI: 10.1093/nar/gkaa1043.

[222] Buniello, A., Macarthur, J. A., Cerezo, M., et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1005–D1012. ISSN: 13624962. DOI: 10.1093/nar/gky1120.

[223] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma’ayan, A. “Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool”. In: *BMC Bioinformatics* 14.1 (Apr. 2013), pp. 1–14. ISSN: 14712105. DOI: 10.1186/1471-2105-14-128.

[224] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., et al. “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update”. In: *Nucleic acids research* 44.W1 (July 2016), W90–W97. ISSN: 13624962. DOI: 10.1093/nar/gkw377.

[225] Xie, Z., Bailey, A., Kuleshov, M. V., et al. “Gene Set Knowledge Discovery with Enrichr”. In: *Current Protocols* 1.3 (Mar. 2021), e90. ISSN: 26911299. DOI: 10.1002/cpz1.90.

[226] Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. "The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans". In: *Current Biology* 26.9 (May 2016), pp. 1241–1247. ISSN: 09609822. DOI: 10.1016/j.cub.2016.03.037.

[227] Dannemann, M. "The Population-Specific Impact of Neandertal Introgression on Human Disease". In: *Genome biology and evolution* 13.1 (Jan. 2021). ISSN: 17596653. DOI: 10.1093/gbe/evaa250.

[228] Chang, L.-H., Ghosh, S., Papale, A., et al. "A complex CTCF binding code defines TAD boundary structure and function". In: *bioRxiv* (Apr. 2021), p. 2021.04.15.440007. DOI: 10.1101/2021.04.15.440007.

[229] Liao, Y., Zhang, X., Chakraborty, M., and Emerson, J. J. "Topologically associating domains and their role in the evolution of genome structure and function in Drosophila". In: *Genome Research* 31.3 (Mar. 2021), pp. 397–410. ISSN: 15495469. DOI: 10.1101/GR.266130.120.

[230] Van der Auwera, G. A. and O'Connor, B. *Genomics in the cloud : using Docker, GATK, and WDL in Terra*. 2020. ISBN: 1-4919-7518-0.

[231] Vierstra, J., Lazar, J., Sandstrom, R., et al. "Global reference mapping of human transcription factor footprints". In: *Nature* 583.7818 (July 2020), pp. 729–736. ISSN: 14764687. DOI: 10.1038/s41586-020-2528-x.

[232] Abascal, F., Acosta, R., Addleman, N. J., et al. "Expanded encyclopaedias of DNA elements in the human and mouse genomes". In: *Nature* 583.7818 (July 2020), pp. 699–710. ISSN: 14764687. DOI: 10.1038/s41586-020-2493-4.

[233] Akgol Oksuz, B., Yang, L., Abraham, S., et al. "Systematic evaluation of chromosome conformation capture assays". In: *Nature Methods* 18.9 (Sept. 2021), pp. 1046–1055. ISSN: 15487105. DOI: 10.1038/s41592-021-01248-7.

[234] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, and David. "The Human Genome Browser at UCSC". In: *Genome Research* 12.6 (June 2002), pp. 996–1006. ISSN: 1088-9051. DOI: 10.1101/gr.229102.

[235] Gokhman, D., Kelman, G., Amartely, A., Gershon, G., Tsur, S., and Carmel, L. "Gene ORGANizer: Linking genes to the organs they affect". In: *Nucleic Acids Research* 45.W1 (July 2017), W138–W145. ISSN: 13624962. DOI: 10.1093/nar/gkx302.

[236] Wang, J., Zhuang, J., Iyer, S., et al. "Factorbook.org: A Wiki-based database for transcription factor-binding data generated by the ENCODE consortium". In: *Nucleic Acids Research* 41.D1 (Jan. 2013), pp. D171–D176. ISSN: 03051048. DOI: 10.1093/nar/gks1221.

[237] Wang, J., Zhuang, J., Iyer, S., et al. "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors". In: *Genome Research* 22.9 (Sept. 2012), pp. 1798–1812. ISSN: 10889051. DOI: 10.1101/gr.139105.112.

[238] Sirugo, G., Williams, S. M., and Tishkoff, S. A. "The Missing Diversity in Human Genetic Studies". In: *Cell* 177.1 (Mar. 2019), pp. 26–31. ISSN: 10974172. DOI: 10.1016/j.cell.2019.02.048.

[239] Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E., Hutter, C. M., Manolio, T. A., and Green, E. D. "Prioritizing diversity in human genomics research". In: *Nature Reviews Genetics* 19.3 (Nov. 2018), pp. 175–185. ISSN: 14710064. DOI: 10.1038/nrg.2017.89.

[240] Popejoy, A. B. and Fullerton, S. M. "Genomics is failing on diversity". In: *Nature* 538.7624 (Oct. 2016), pp. 161–164. ISSN: 14764687. DOI: 10.1038/538161a.

[241] Koller, D., Wendt, F. R., Pathak, G. A., Lillo, A. D., and De, F. "The impact of evolutionary processes in shaping the genetics of complex traits in East Asia and Europe : a specific contribution from Denisovan and Neanderthal introgression". In: *bioRxiv* 1.203 (Aug. 2021), p. 2021.08.12.456138. DOI: 10.1101/2021.08.12.456138.

[242] Shi, H., Gazal, S., Kanai, M., et al. "Population-specific causal disease effect sizes in functionally important regions impacted by selection". In: *Nature Communications* 12.1 (Feb. 2021), pp. 1–15. ISSN: 20411723. DOI: 10.1038/s41467-021-21286-1.

[243] Luo, Y., Li, X., Wang, X., et al. "Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations". In: *Human molecular genetics* 30.16 (July 2021), pp. 1521–1534. ISSN: 14602083. DOI: 10.1093/hmg/ddab130.

[244] Brown, B. C., Ye, C. J., Price, A. L., and Zaitlen, N. "Transethnic Genetic-Correlation Estimates from Summary Statistics". In: *American Journal of Human Genetics* 99.1 (July 2016), pp. 76–88. ISSN: 15376605. DOI: 10.1016/j.ajhg.2016.05.001.

[245] Quintana-Murci, L. "Understanding rare and common diseases in the context of human evolution". In: *Genome Biology* 17.1 (Nov. 2016), pp. 1–14. ISSN: 1474760X. DOI: 10.1186/s13059-016-1093-y.

[246] Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., et al. "Predicting Splicing from Primary Sequence with Deep Learning". In: *Cell* 176.3 (Jan. 2019), 535–548.e24. ISSN: 10974172. DOI: 10.1016/j.cell.2018.12.015.

[247] Kelley, D. R. "Cross-species regulatory sequence activity prediction". In: *PLoS Computational Biology* 16.7 (July 2020), e1008050. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1008050.

[248] Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". In: *Nature Genetics* 50.8 (Aug. 2018), pp. 1171–1179. ISSN: 15461718. DOI: 10.1038/s41588-018-0160-6.

[249] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature Methods* 18.10 (Oct. 2021), pp. 1196–1203. ISSN: 15487105. DOI: 10.1038/s41592-021-01252-x.

[250] Li, B. and Ritchie, M. D. "From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries". In: *Frontiers in Genetics* 12 (Sept. 2021), p. 1502. ISSN: 16648021. DOI: 10.3389/fgene.2021.713230.

[251] Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., et al. "Opportunities and challenges for transcriptome-wide association studies". In: *Nature Genetics* 51.4 (Mar. 2019), pp. 592–599. ISSN: 15461718. DOI: 10.1038/s41588-019-0385-z.

[252] Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E., and Gilad, Y. "Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates". In: *PLoS Genetics* 15.7 (July 2019). Ed. by H. S. Malik, e1008278. ISSN: 15537404. DOI: 10.1371/journal.pgen.1008278.

[253] Schoech, A. P., Jordan, D. M., Loh, P. R., Gazal, S., O'Connor, L. J., Balick, D. J., Palamara, P. F., Finucane, H. K., Sunyaev, S. R., and Price, A. L. "Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection". In: *Nature Communications* 10.1 (Dec. 2019), pp. 1–10. ISSN: 20411723. DOI: 10.1038/s41467-019-08424-6.

[254] Mori, K., Miyazato, M., Ida, T., Murakami, N., Serino, R., Ueta, Y., Kojima, M., and Kangawa, K. "Identification of neuromedin S and its possible role in the mammalian circadian oscillator system". In: *EMBO Journal* 24.2 (2005), pp. 325–335. ISSN: 02614189. DOI: 10.1038/sj.emboj.7600526.

[255] Nakahara, K., Hanada, R., Murakami, N., Teranishi, H., Ohgusu, H., Fukushima, N., Moriyama, M., Ida, T., Kangawa, K., and Kojima, M. "The gut-brain peptide neuromedin U is involved in the mammalian circadian oscillator system". In: *Biochemical and Biophysical Research Communications* 318.1 (2004), pp. 156–161. ISSN: 0006291X. DOI: 10.1016/j.bbrc.2004.04.014.

[256] Novak, C. M. *Neuromedin S and U*. 2009. DOI: 10.1210/en.2009-0448.

[257] Chiu, C. N., Rihel, J., Lee, D. A., et al. "A Zebrafish Genetic Screen Identifies Neuromedin U as a Regulator of Sleep/Wake States". In: *Neuron* 89.4 (2016), pp. 842–856. ISSN: 10974199. DOI: 10.1016/j.neuron.2016.01.007.

[258] Aizawa, S., Sakata, I., Nagasaka, M., Higaki, Y., and Sakai, T. "Negative Regulation of Neuromedin U mRNA Expression in the Rat Pars Tuberalis by Melatonin". In: *PLoS ONE* 8.7 (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0067118.

[259] Visconti, A., Duffy, D. L., Liu, F., et al. "Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure". In: *Nature Communications* 9.1 (2018). ISSN: 20411723. DOI: 10.1038/s41467-018-04086-y.

[260] Hinrichs, A. S., Karolchik, D., Baertsch, R., et al. "The UCSC Genome Browser Database: update 2006." In: *Nucleic acids research* 34.Database issue (Jan. 2006), pp. D590–D598. ISSN: 13624962. DOI: 10.1093/nar/gkj144.