Leveraging Biobanks and PheWAS to Uncover the Health Consequences of Recent Human Evolution

By

Corinne Nicole Simonti

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, 2017
Nashville, Tennessee

Approved:

John A. Capra, Ph.D.

Douglas P. Mortlock, Ph.D.

David C. Samuels, Ph.D.

Joshua C. Denny, M.D., M.S

William S. Bush, Ph.D.

To my family. For everything.

ACKNOWLEDGEMENTS

The work presented in this dissertation was supported in part by the ocular genomics training grant (5T32EY021453-01), originally awarded to Vanderbilt University's Center for Human Genetics Research, obtained by Jonathan Haines and now managed by Milam Brantley, as well as an Innovation Catalyst grant by the March of Dimes. This work would not have been possible without the data and time contributions of Josh Denny's group—particularly Lisa Bastarche and Robert Carroll—and Josh Akey and Ben Vernot, as well as the helpful scientific feedback from my committee members.

Thank you to all the Human Genetics students, past and present, who have given me so much support over the last five years. In particular, I would like to acknowledge all the previous graduate students who devoted so much time and energy to our practice quals, and Brittany and Carissa, who were my faithful study buddies for Phase I and helped me find all the great swag at ASHG. I also want to acknowledge all my IGP friends who I haven't seen nearly as much as I would have liked, but who have managed to find the time for coffee or brunch or drinks despite our insane schedules, and let me crash the odd happy hour for departments that weren't my own. I also must thank Rose, Megan, and all the former Suzie's baristas for keeping me caffeinated.

Another great source of support for me have been my labmates, who have helped me grow as a scientist and made coming to work fun even when science was treating me poorly. We have talked about science and TV, managed not to starve (together), and I will always appreciate the sense of camaraderie I've felt while working here. While I want to acknowledge all of you, I must particularly thank Alex Fish, who had to put up with my baby graduate student woes before we were even in the same lab. You have always been there to ask tough questions during my presentations and to cover my first drafts of emails, abstracts, and manuscripts with red ink when

needed. Finally, I have to thank my mentor, Tony Capra, for taking me into his lab even though he had just gotten to Vanderbilt, for having a managerial style that matches my working style, for thinking I was awesome when I felt mediocre, and most importantly, for all the beer.

Last but certainly not least, I would like to acknowledge my roommate, bestie, travel buddy, and sister from another mister, Lindsay, without whom I probably would have either starved to death or had a nervous breakdown. There will always be space in my guesthouse for you.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

INTRODUCTION

Human Evolutionary Genetics in the Genomic Era

One of the goals of human evolutionary genetics has always been to understand the genetic basis of human-defining traits and origins. By 1975, growing evidence had established that anatomically modern humans (AMHs) and chimpanzees showed little divergence at the protein sequence level, leading King and Wilson to hypothesize that non-coding rather than coding variation drove differences in morphology between the two species[1]. However, until the completion of the sequencing of the human genome in 2003[2–4] and the subsequent advent of high-throughput sequencing technology, most studies of human origins and dispersal were restricted to molecular studies, archeological evidence, geographic distributions of mitochondrial and Y chromosome haplogroups, and comparison of limited genomic sequences. A combination of archeological evidence and sequencing of several non-repetitive regions of the autosomal genome revealed that AMHs diverged from chimpanzees, our closest living relative, an estimated 4-6 million years ago[5,6], and mitochondrial coalescence estimates indicate that AMHs arose ~200,000 years ago in Africa[7].

The sequencing of both the human and chimpanzee genomes—the latter sequenced in 2005—confirmed the estimates of minimal coding variation between the two species made 30 years before[8]. Surprisingly, non-coding variation only affects a small proportion of the genome: 1.23% of the human genome differed from chimpanzee by single nucleotide substitutions and 1.5% by lineage-specific small insertions and deletions (indels)[8], with lineage-specific segmental duplications altering 2.7% of the genome[9]. Despite the relatively small proportion, 5% of the

3,200,000,000 base pairs (bp) that comprise the human genome is still 160,000,000 bp of sequence to investigate. Studies of these differences are further complicated by the fact that many of these lineage-specific changes are expected to differ due to genetic drift—stochastic processes that impact allele frequency and fixation—rather than selection. While there is interest in understanding "what makes us human," isolating the critical changes over the roughly 6 million years since the AMH divergence from chimpanzee is an incredibly difficult task.

One method to identify regions of the genome important to species divergence is to look for signatures of evolutionary acceleration. In general terms, this involves a statistical test comparing substitution rates with the rate expected given the phylogenetic tree. When applied to humans, this technique locates regions that have been conserved for an extended evolutionary time—typically, across mammals—but have accrued many mutations on the human lineage (reviewed in Hubisz and Pollard[10]). Thus, human accelerated regions (HARs) theoretically represent sequences whose conserved function has been lost or substantially altered solely on the human lineage. Several groups have identified HARs in genome-wide scans, the vast majority of which are non-coding and enriched near genes involved in development and the central nervous system[11–14]. Testing in mouse models has shown that many HARs act as developmental enhancers, with the tissue expression pattern of the human allele often differing from that of the conserved allele[14].

While HARs provide many interesting candidate regions for studying human-specific evolution, timing of the acceleration event is difficult to gauge without considering species more closely related to AMHs than chimpanzee. This is difficult given that our closest known relatives, Neanderthals and Denisovans, are extinct. However, advances in endogenous sequence capture from ancient samples and the discovery of fossils with remarkably well-preserved DNA

have allowed for multiple individuals from these archaic hominin groups to be sequenced[15–17], including limited autosomal DNA from several ~430,000 year old Neanderthal individuals[18]. When coupled with archeological evidence, this study revealed that the AMH lineage likely diverged from Neanderthals and Denisovans between 550,00-765,000 years ago. This dating, coupled with the high quality genome sequencing of several archaic individuals, has allowed researchers to determine that most of the periods of acceleration creating these HARs occurred before our divergence from Neanderthals and Denisovans[10]. This suggests that many identified HARs are involved in morphological traits shared between AMHs and Neanderthals, such as habitual bipedalism and increased brain-to-body size ratio. However, it also poses interesting questions about what processes are affected by human-specific accelerated regions.

The discovery of HARs is a recent advance in understanding differences between AMHs and chimpanzee, but evolutionary geneticists have been studying similar effects in the autosomal genome between human populations for years. These efforts began with the International HapMap Project, which set out to annotate common patterns of variation across individuals of European, African, and Asian ancestry[19]. This work was critical to the initial creation of genotyping platforms used in disease studies. The HapMap Project was followed by the 1000 Genomes Project, which was enabled by high-throughput sequencing technology and whose original goal was to annotate rare variation[20], and has now sequenced upwards of 2,500 individuals from Africa, Europe, East Asia, South Asia, and the Americas[21]. Through these projects and the sequencing efforts of individual labs or other consortia, the genetic basis of lactase persistence[22,23], resistance to malaria[24–30], high altitude adaptation[31–34], and many others have been uncovered (reviewed in Vitti *et al*[35]). These traits have proven somewhat easier to dissect genetically than others, due to their typically monogenic basis and that recent, strong

selection on a single variant or region—also known as a selective sweep—is relatively easy to detect between closely related populations.

Most current methods for identifying signatures of selection leverage the whole-genome sequences collected by these projects, and are optimized for detecting selection that has occurred in the last 100-1,000 generations. Because groups of AMHs began successfully migrating out of Africa into the Middle East and beyond between 50,000-60,000 years ago[36], many of these methods are suitable for studying selection on this scale. Moreover, a great deal of recent progress has been made in methods for detecting different types of selection (i.e., polygenic selection—selection occurring on many regions in the genome with only moderate effects on each—rather than selective sweeps)[37,38], as well as narrowing the window during which selection has occurred to either very recent (within the last 100 generations)[39] or very distant (since our most recent common ancestor with Neanderthal)[40] timescales.

Even with improvements in methodology, many tests for selection are susceptible to historical demography and events that would alter linkage disequilibrium (LD) or allele frequencies, as most methods rely on at least one of these features (reviewed in Vitti *et al*[35]). For example, rather than the most parsimonious scenarios of humans dispersing across Eurasia and into the Americas after leaving Africa, ancient DNA sequencing coupled with the sequencing of many individuals of varied continental ancestries has made it clear that modern populations are, perhaps unsurprisingly, the result of a complex series of migration events[41–43], admixtures between human groups[44,45], and interbreeding between humans and archaic hominins[15,17] (reviewed in Haber *et al*[46]). Beyond the complications of historical events, mutational processes with heterogeneous activity throughout the genome or between populations can also resemble positive or negative selection[47–51].

As we sequence more archaic hominin and ancient AMH samples, as well as improve methods for identifying variation with interesting evolutionary histories that take complex demography and mutational processes into account, we improve our ability to identify variation that has not been under strong selection. As selection strength should correlate with effects on fitness, connecting variation under increasingly weak selection to phenotypic expression becomes increasingly difficult. Understanding the functional mechanism through which variation acts on organismal fitness is critical to support results of studies of selection[35]. However, many evolutionary studies have been based on sequence alone, and have used tools such as gene ontology functional enrichment to support ties to phenotypes theorized to be important during the timescale considered[52–56]. Without an understanding of the functional mechanism through which variation in these regions act, it has been difficult to demonstrate precisely how regions under selection have influenced a phenotype or even what biological process is affected. This is complicated further by the fact that many tests detect regions under selection because of the presence of extended LD, obscuring identification of the causal variant (assuming the effect is due to a single causal variant and that there is not a haplotype effect). As we improve our understanding of how and when both stochastic processes and selection have shaped the genome, we must likewise improve our ability to connect variation to phenotypes.

Adapting New Tools to Answer Evolutionary Questions

One avenue for improving our understanding of the phenotypic effects of evolution is to leverage the wealth of data produced and used by functional and statistical genetics. Functional genomics and gene expression data can aid in generating hypotheses about how identified

regions or variants influence basic biology, and have been used in this manner for many genome-wide association studies (GWAS)[57–60]. One of the most extensive resources for gene expression data is the Genotype-Tissue Expression (GTEx) Project, which has sampled gene expression across many tissues in hundreds of individuals[61]. Two of the largest databases containing functional genomics data—including DNase hypersensitivity sites, histone marks, transcription factor binding sites, and others—across a wide range of cell lines and tissues are the Encyclopedia of DNA Elements (ENCODE)[62] and Roadmap Epigenomics[63] projects. These resources and many others can connect regions or variants of interest to potential effects on cellular-level phenotypes such as gene expression or alteration of enhancer activity. However, connections to an organismal-level phenotype may be more informative than functional data alone, especially in instances where the closest genes are active in many biological processes, or certain selective pressures (e.g., exposure to a pathogen) are hypothesized to have acted on a given study population. An early example of the power of this approach comes from a study that identified regions likely to be under selection in a Bangladeshi population (where cholera is endemic), and performed their own candidate association study on cholera susceptibility in independent individuals from the same population known to be exposed to cholera[64].

Rather than perform new association studies, several groups have also incorporated the results of previously conducted GWAS to identify phenotypes impacted by very recent selection[39] and archaic introgression[65]. The results of previous GWAS are manually entered by researchers into the National Human Genome Research Institute (NHGRI) GWAS catalog[66]. The catalog contains information about variants passing at least a nominal significance threshold ($P <$ 1E-06), and details the ancestry of the populations used, information about the degree of risk conferred through odds ratios (OR) or betas, allele frequencies, and much more. Many of the

studies are focused on diseases, but some have examined neutral traits such as hair color[67–70] or tendency to freckle[67,68,71]. While the catalog is extensive, it is by no means a complete record of every variant-phenotype association. As the amalgam of results from thousands of studies, it can also be difficult to determine what might cause a variant to associate with a phenotype in one study, but not a related phenotype in another study. One way to overcome these challenges would be to survey a single population for a wide range of phenotypes; however, ascertaining, phenotyping, and genotyping (if not whole-genome sequencing) thousands of individuals is an expensive, time-consuming, and difficult undertaking.

Though clinical biobanks began as tools to improve patient care and conduct clinical research[72], they provide a fertile resource for evolutionary analyses by alleviating the need to ascertain an entirely new population for many traits of interest. Many already exist across the United States and overseas, and more will be created or expanded through projects such as the Precision Medicine Initiative *All of Us* Research Program[73], which aims to collect over one million individuals. Usually linked with electronic health records (EHR), these datasets contain a dense, often longitudinal record of a patient's clinical history. Some clinical biobanks de-identify patient records when making them available to researchers. In the case of the Vanderbilt University Medical Center's Synthetic Derivative (SD), record numbers go through a one-way hash so that they cannot be traced back to a patient's identifying number, dates within a record are shifted by 1-365 days, and all Health Insurance Portability and Accountability Act (HIPAA) identifiers are removed, including names of patients and health care providers[74]. The SD has a non-human subjects designation due to the de-identification of its records, which eases the ability of researchers to access the data. However, de-identified records do not allow for contacting patients for research follow-up, hence why some biobanks used for research do not de-identify

patient records. Regardless of how the records are handled by an institution, when EHR information is paired with dense genotyping data or whole-genome DNA sequencing, researchers can perform analyses that take advantage of the wealth of phenotypic data available in EHRs[75], such as phenome-wide association studies (PheWAS).

Sometimes termed a "reverse GWAS"[76], the goal of PheWAS is to test for associations between one variant and many phenotypes. Clinical phenotypes tested in PheWAS can be derived from several sources in the EHR: International Classification of Disease version 9 (ICD-9) codes (a hierarchical way to classify diseases and symptoms), Current Procedural Terminology (CPT) codes (indicating patient procedures), lab values, and other information extracted from free text through natural language processing (NLP). Most studies have used aggregation of highly related ICD-9 codes into similarly hierarchical "phecodes" to determine case, control, and exclusion status for over 1,500 clinical phenotypes[76–79]. One such study demonstrated that 66% of SNP-phenotype associations from the GWAS catalog could be replicated when the study had sufficient power to detect the association ($P < 0.05$, consistent direction of effect) in this setting[80]. The ability to test many phenotypes at once is useful in studies of pleiotropy—the phenomenon of a single variant or gene affecting more than one phenotype—as well as untangling shared environmental and genetic risk factors between phenotypes[81–83]. It can also serve as an agnostic way to survey phenotypes affecting practically every biological system when the effect of a variant is unknown. Phecodes are grouped by primary biological system affected (i.e., immunologic, neurologic, etc.)[79,80].

Collapsing ICD-9 codes into phecodes facilitates statistical models other than PheWAS, such as genome-wide complex trait analysis (GCTA). The goal of GCTA is to use mixed linear modeling to test for the effects of many variants on a single phenotype to determine the percent

variance of a trait explained by the variants tested[84,85]. This genetic variance is often termed "chip heritability" as it only surveys variants genotyped on a "genotyping chip" rather than the whole genome. In addition to analyzing the genetic variance of a single phenotype, bivariate GCTA can also detect the genetic covariance between two phenotypes[86]. GCTA can serve as a complement to PheWAS when the effect of any individual variant may not be large enough to be detected, but all the variants tested are thought to affect the same biological system. This method's "chip heritability" is often lower than heritability estimated from twin studies[84], making it a useful tool for estimating a lower bound for heritability of a phenotype found in an EHR that has not been evaluated in twin or family studies.

Despite the many benefits of using EHR data, there are also several constraints. Because non-clinical phenotypes—such as hair or eye color—are typically of limited use in improving patient care, they are often inconsistently recorded and difficult to extract from EHRs. While not specific to EHRs, sparse genotyping coverage also poses challenges to testing all variation of interest, particularly if complex LD patterns from historic events (i.e., admixture) or mutational processes make imputation unreliable. Additionally, many clinical biobank populations are heavily biased in favor of individuals of European descent (a problem shared with most GWAS), which makes testing variation private to another population untenable. Nonetheless, as more sophisticated methods are developed for extracting phenotypes from the rich data stored in EHRs, and as EHRs are increasingly linked to dense genotyping and/or whole genome sequence data from individuals of diverse ancestry, we anticipate further insights into the phenotypic effects of many variants. Moreover, increased understanding of the functional effects of polymorphic sites coupled with their evolutionary histories may aid in understanding the effects of nearby variants that are fixed in AMHs but differ from Neanderthal or chimpanzee.

About this Dissertation

The goal of the work presented in this dissertation is to use recently developed statistical genetics methods—namely, PheWAS and GCTA—to associate evolutionarily distinct variation with modern clinical traits. While these methods cannot replace traditional molecular analyses, they provide a high-throughput approach to generate or test hypotheses about what biological systems are impacted by variants of interest. Each of the analyses performed in these chapters attempts to use PheWAS, GCTA, and other methods when appropriate to address previously generated hypotheses about how the examined events or processes affect AMH morphology or health.

In Chapter I, I examine the impact of variation that was introduced to AMHs through interbreeding between Neanderthals and the ancestors of modern Eurasians. As Neanderthals had lived outside of Africa for over 100,000 years, introgression is theorized to have been beneficial to AMHs just leaving Africa through the contribution of alleles affecting interactions with the environment, such as the immune and integumentary systems. Using GCTA and PheWAS, I show that Neanderthal introgressed variants associate with phenotypes related to these systems. Unexpectedly, these variants are enriched for associations with gene expression in the brain as well as neurologic and psychiatric phenotypes, suggesting that effects of interbreeding extended to the central nervous system. This chapter is adapted from my peer-reviewed article "The phenotypic legacy of admixture between modern humans and Neandertals" in *Science*[87].

In Chapter II, I examine the impact of alleles that have decreased sharply in frequency in humans since our most recent common ancestor with chimpanzee. This is a group of variants

where humans are nearly fixed for a derived allele and are theorized to have contributed to hominin or human-specific evolution, depending on whether the derived allele is present in Neanderthals or Denisovans. Two of the human-specific variants associate with bone fracture, which is interesting given the skeletal differences between humans and Neanderthals. For some of these variants, the ancestral allele appears to have been lost in the out-of-Africa transition, but later reintroduced into Eurasians via introgression from Neanderthals.

Finally, in Chapter III, I examine variants impacted by a mutational process, GC-biased gene conversion (gBGC), in order to examine its contribution to modern human health. This process increases the likelihood of a G or C allele being transmitted over an A or T allele in heterozygous individuals, and has been theorized to increase the frequency of weakly deleterious alleles. I compared the likelihood of variants exposed to gBGC associating with a clinical phenotype to that of variants from several matched sets. There appears to be no increased risk for diseases of modern populations for gBGC variants compared with others or permuted variants. This is likely due to how this process affects other properties of these variants and how these properties affect power to detect statistical associations.

CHAPTER I: NEANDERTHAL INTROGRESSION INFLUENCES DIVERSE BIOLOGICAL SYSTEMS IN INDIVIDUALS OF EUROPEAN DESCENT[1]

Introduction

Our understanding of human origins has improved drastically over the last several decades. As the earliest work in this field was dependent on skeletal or craniofacial morphology alone, it was unclear how the multitude of human-like fossils discovered throughout Europe and Asia fit within human history. The presence of archaic hominin fossils attributed to separate groups (*Homo erectus*, *Homo heidelbergensis*, etc.) based on morphology, geography, and age painted a complex picture of human history. Since their discovery, Neanderthals were the subject of speculation as to whether they were the ancestors of modern Europeans, or a sister species that had been stamped out by the superior *Homo sapiens*[88]. The rise of population genetics and the inception of ancient DNA sequencing began to resolve some of these complexities beginning in the 1990s. In 1997, mitochondrial DNA (mtDNA) from a Neanderthal individual discovered in Germany in 1856 was sequenced, revealing that Neanderthal mitochondrial lineages likely diverged from human lineages ~550,000-690,000 years ago[89]. These early findings were supported by later studies examining mtDNA from additional Neanderthals from the Caucasus mountains[90,91], Croatia[91,92], Belgium[92,93], Germany[91], France[92], Spain[91], Uzbekistan[94], and the Altai region of Siberia[94]. As anatomically modern human (AMH) mitochondrial lineages coalesced ~200,000 years ago[7], the totality of mtDNA evidence demonstrated that Neanderthals were a sister group to humans that had been separated for some time. This work also suggested

[1] This chapter is adapted from my peer-reviewed article titled "The phenotypic legacy of admixture between modern humans and Neandertals," published in *Science*[87].

that Neanderthals and AMHs did not interbreed[89–91,93,95], despite the two groups overlapping in space and time before Neanderthals became extinct ~28,000 years ago[96].

With advances in whole genome sequencing, the draft sequence of the Neanderthal autosomal genome was completed in 2010 using three individuals from Croatia[15]. Contrary to the mtDNA findings, the autosomal genome indicated introgression from Neanderthals into non-Africans occurred at very low levels (~1-4%). However, due to DNA degradation and human contamination, this study only managed to obtain ~1.3X sequencing depth, which limited the confidence of these claims and made identification of specific introgressed regions impossible. Improvements in ancient DNA sequencing and the discovery of archaic hominin remains with remarkably well-preserved genomic DNA in the Altai mountains allowed for the sequencing of a Neanderthal individual as well as a previously unknown hominin—designated Denisovans after the cave in which it was found—at 52X[17] and 30X[97] coverage, respectively. From these genomes, it is now clear that not only did interbreeding occur between AMHs and Neanderthals, it occurred multiple times between different combinations of these known groups, as well as other archaic hominins whose genomes have not yet been sequenced[15,17,98]. These genomes also indicated that at least one group of AMHs left Africa as early as ~100,000 years ago and interbred with Eastern but not European Neanderthals, nor did they contribute to AMH populations today[99]. Indeed, the first successful AMH migrations into Eurasia do not appear to have occurred earlier than ~60,000 years ago. These AMHs interbred with European Neanderthals roughly 50,000 years ago, which resulted in the genomes of modern Eurasians containing a small fraction (~1.5–3%) of Neanderthal DNA[42,53,54,100].

Based on this work and the sequencing of several ~430,000 year old Neanderthals from Spain[18], the divergence time of AMHs and the Neanderthal-Denisovan ancestor is currently

approximated to be between 550,000-765,000 years ago. Even with the repeated interbreeding between archaic hominin groups, such an ancient divergence time leads to questions about how compatible Neanderthal DNA would be with an AMH genome. Indeed, several genomic regions—e.g., on the X chromosome and the q-arm of chromosome 8—are strongly depleted of Neanderthal introgression, suggesting Neanderthal DNA was not well tolerated[53,54]. Two of these in particular have received a great deal of speculation: the depletion on the X chromosome, and the depletion near the forkhead box protein P2 (*FOXP2*) gene. Certain groups have suggested that the depletion of Neanderthal ancestry on the X chromosome is the result of moderate sexual incompatibility between AMHs and Neanderthal[54], though others suggest this could be the result of selection acting more efficiently on the X chromosome[101,102]. The depletion near the *FOXP2* gene has garnered attention due to the role of this gene in human language development and birdsong[103,104]. Some of the additional depleted regions may contain important human-specific changes—like those in *FOXP2*—or structural variation, as many of them are depleted of both Neanderthal and Denisovan ancestry in Melanesians, who bear ~2–3.5% Denisovan DNA in addition to Neanderthal DNA at a similar proportion to other non-Africans[105].

Despite these findings, not all introgressed Neanderthal or Denisovan DNA is necessarily expected to be detrimental to human fitness. Both of these groups moved out of Africa and into the Middle East and Europe long before AMHs, and therefore had the opportunity to adapt to the climatic and pathogenic landscapes found at these latitudes. If this were the case, introgression could have introduced variants that provided an advantage to these recently relocated AMH populations, and would be expected to occur at high frequencies. Indeed, several isolated introgressed loci have been identified with potential roles in human adaptation to pathogens and hypoxic conditions found at high altitudes[31,65,106,107]. While their phenotypic effects are currently

unknown, some Neanderthal haplotypes are found at higher than expected frequencies in non-African AMH populations, though they are not always consistent between Europeans and Asians, possibly due to different selective pressures or stochastic events[53,54,106]. Analyses of genomic regions enriched for Neanderthal ancestry have suggested potential effects on skin and hair phenotypes, lipid catabolism, neuronal function and other traits[53,54,108]. However, due to the difficulty of identifying Neanderthal-derived DNA from genotype data alone and the expense of collecting individuals to test for trait association, the impact of introgressed Neanderthal alleles on these traits in human populations has not been established.

To address these challenges, we integrated the clinical phenotypes present in electronic health records (EHRs) with variation likely to be present in human populations solely due to introgression from Neanderthals. We performed association analyses between these Neanderthal-introgressed variants and clinical phenotypes using individuals from the Electronic Medical Records and Genomics (eMERGE) Network, a consortium that unites EHRs linked to genotyping data from ten hospital sites across the United States[109]. We analyzed a set of nearly 30,000 adults of European ancestry from seven of the eMERGE sites who were genotyped on genome-wide arrays and had sufficient EHR data to define phenotypes. Based on their inclusion in the eMERGE Network Phase 1 (E1; N=13,686) or Phase 2 (E2; N=14,730) data releases[110], these individuals either fell into our discovery or replication cohorts. To understand how Neanderthal introgression influences health in modern individuals of European ancestry, we examined the association with clinical phenotypes of both: individual variants through phenome-wide association studies (PheWAS)[76,80], and these variants together through genome-wide complex trait analysis (GCTA)[85,111] (Figure 1-A). The results of these analyses can give us insight into whether Neanderthal introgression as a whole has a negative impact on modern

human health, as has been theorized[101,102], and whether certain biological systems were targeted by introgression.

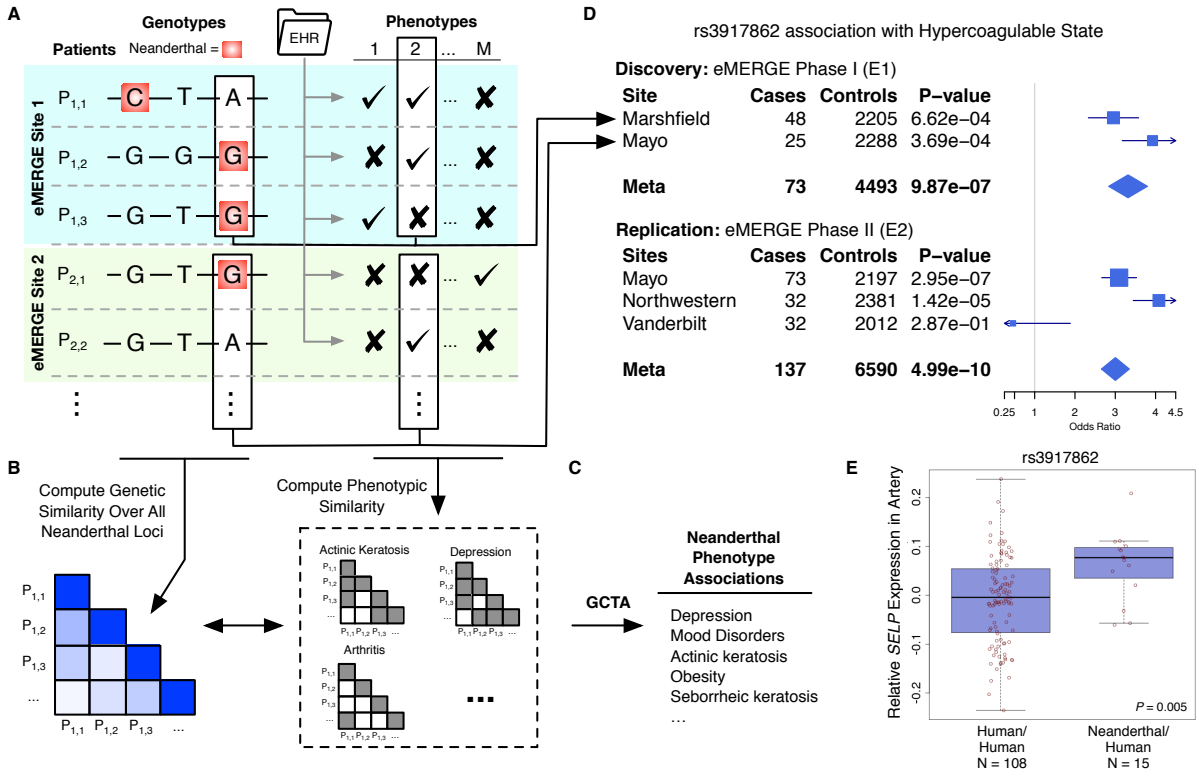


Figure 1-A*. (a) Thousands of Neanderthal alleles were identified in ~28,000 individuals of European ancestry across the eMERGE Network. We derived phecode case/control status for each individual from data in their EHRs. (b) To test Neanderthal alleles in aggregate for phenotype associations, we computed the genetic similarity of all pairs of individuals over 1,495 genotyped Neanderthal loci and their phenotypic similarity over 46 EHR-derived traits. (c) We estimated the overall variance in risk explained by Neanderthal alleles using mixed linear models in GCTA[85,111]. (d) To test individual Neanderthal alleles for trait associations, we performed a discovery meta-analysis across E1 sites with sufficient data. We then ran a replication meta-analysis over the independent E2 cohort. The example forest plot illustrates the association of Neanderthal SNP rs3917862 with hypercoagulable state in each site with >= 20 cases for the separate discovery and replication analyses. (e) rs3917862 is located in an intron of P-selectin (*SELP*), a gene that mediates leukocyte action at injuries in the early stages of inflammation. The Neanderthal allele is significantly associated (linear regression, $P = 0.005$) with increased expression of *SELP* in tibial artery data from GTEx[61].
*This figure is adapted from Figure 1 from my peer-reviewed article[87].

Methods

Variant Selection

To identify variants present in AMH populations solely due to introgression from Neanderthals, we first collected all variants falling in a set of high-confidence (FDR < 0.05) introgressed haplotypes[53] identified using the S* algorithm[98], the Altai Neanderthal sequence[17], and 1000 Genomes (1KG) data[20]. We first calculated the statistic S*, which considers both divergence and LD, and then refined the resulting set of candidate introgressed regions by directly comparing significant S* haplotypes to the Altai Neanderthal genome sequence. A "Neanderthal match *P* value" was calculated to quantify whether the observed matching is higher than would be expected by chance[53]. Only haplotypes that are significant by S* and have a significant Neanderthal match *P* value are used in the analyses here. From these, we took all biallelic variants where the putatively introgressed allele matched the Altai Neanderthal sequence and was derived with respect to chimpanzee. For each haplotype, we calculated a 90% trimmed mean allele frequency. To remove variants that were segregating in both AMHs and Neanderthals before interbreeding or were unlikely to be present on the introgressed Neanderthal haplotype at the time of introgression, we restricted our variants of interest to those within 10% of the trimmed mean allele frequency of the haplotype. After these filtering steps, we required that at least four variants remained to include a haplotype and its variants.

Study Population

The eMERGE Network is comprised of ten sites: seven with adult samples and three with pediatric samples. We used adult (18 years of age or older as of January 2015) individuals of

European ancestry from the seven adult sites: Geisinger Health System, Group Health Cooperative (Washington State), Mayo Clinic, Marshfield Clinic, Mt. Sinai, Northwestern University, and Vanderbilt University. The eMERGE Phase 1 (E1) data set comprises 13,686 individuals from five of the seven sites with adult samples: Group Health Cooperative, Mayo Clinic, Marshfield Clinic, Northwestern University, and Vanderbilt University. Phase 2 (E2) comprises an independent set of 14,730 individuals from all seven sites.

Genotyping, Quality Control, and Imputation

The eMERGE subjects were genotyped on a range of genome-wide arrays including the Affymetrix 6.0 and the Illumina 550, 610, 660, 1M, 5M, and Omni Express chips. The eMERGE Coordinating Center at Pennsylvania State University performed genotype imputations for all samples collected as part of E1 and E2. In the v3 (December 2014) data release used here, SHAPEIT2[112] and IMPUTE2[113] were used to impute all autosomes to the 1000 Genomes Project (release March 2012). Imputed data for all sites were then merged based on an intersection of successfully imputed SNPs between them. For different analyses, probabilities (dosages) or the most likely imputed genotypes (hard calls) were used as indicated in the text. For the hard call SNPs, the marker call rate threshold was set at 99% and info score threshold at 0.7.

Population structure was evaluated using Eigensoft 6.0[114] on filtered and LD pruned data. Related individuals (one from each pair of kinship >0.125) as estimated from identity-by-descent (IBD) were removed before principal components analysis (PCA). Two highly correlated regions as well as all palindromic SNPs were removed, in addition to a marker call rate filter of 99%, MAF > 10%, and LD pruning of $r^2 > 0.1$. This left 101,000 SNPs, on which 30 principal

components were calculated. See the eMERGE Network methods publications and web site for more details[115,116].

Phenotyping

Clinical phenotypes were derived using a prior EHR-based PheWAS approach, which uses established algorithms that integrate the range of diseases, signs and symptoms, causes of injury, and procedures represented by International Classification of Diseases, ninth edition (ICD-9), codes into 1,645 coherent phenotypes, such as "inflammatory bowel disease," and its child terms "Crohn's disease" and "ulcerative colitis"[79]. Some of these phenotypes and the corresponding controls (defined by lack of related codes) have seen extensive manual and computational validation across eMERGE, and proven successful in previous studies[76,78,80,117]. The ICD-9 based phenotype definition algorithms used produce phenotypes that enable replication of 66% of known associations in sufficiently powered (80%) association tests[80] and recent work has yielded even higher replication rates (communication from Joshua Denny).

ICD-9 code counts were extracted from the electronic health record (EHR) and converted to PheWAS code (phecode) counts. From the phecode count list, we used the PheWAS package (v0.9.5.1-1)[79] function "createPhewasTable" to generate case/control status for our individuals using a minimum code count of two unique dates of a diagnosis. We did not analyze phecodes with a case count less than 20.

PheWAS

We performed both a meta-analysis of PheWASes performed on each eMERGE site's data individually, and a joint PheWAS analysis over data pooled from all eMERGE sites. Both

analyses were performed separately on the independent E1 (discovery) and E2 (replication) cohorts. We analyzed 1,495 common (MAF > 1%) Neanderthal SNPs genotyped by the eMERGE Network and required that phecodes have at least 20 cases in each site analyzed in the meta-analysis or overall for the pooled analysis. For the meta-analyses, a PheWAS was performed on each eMERGE site's data using the "phewas" function in the PheWAS package[79] to run logistic regression using an additive genetic model. A meta-analysis of the site-specific scans was performed with the "phewasMeta" function. We considered age, sex, and first three principal components (PCs) as covariates. For the joint analyses, the "phewas" function in the PheWAS package was used to analyze data pooled across eMERGE sites. For the pooled analysis, we again included age, sex, and the first three PCs as covariates, and additionally used eMERGE site (dummy coded as either 4 (E1) or 6 (E2) variables). For imputed SNPs in the E2 analyses, we used dosages rather than the hard calls. We used gtool (v0.7.5) and qctool (v1.4) to select the appropriate SNPs from the IMPUTE2 files and convert to the input format for the PheWAS package. We report the *P* value and odds ratio from the fixed effect models unless otherwise stated.

Enrichment Analysis

To test whether Neanderthal SNPs were more likely to be associated with disease phenotypes than non-Neanderthal SNPs, we identified phecode associations for 1,056 common (MAF > 1%) non-redundant ($r^2 < 0.5$) Neanderthal SNPs at a relaxed significance threshold of $P < 0.001$ in the discovery set that replicated in E2 ($P < 0.05$ and same direction of effect). This yielded 60 associations after accounting for phecodes present in the same hierarchy (Appendix A). To generate a set of appropriate non-Neanderthal SNP-phecode associations for comparison,

we identified SNPs that were not within 100 kb of a Neanderthal SNP, and then we pruned these non-Neanderthal SNPs so that none were in strong LD with another variant ($r^2 > 0.5$). We then identified five independent matched control sets (for a total of 5,280 non-Neanderthal SNPs) that matched genotyping status on the Human660W-Quadv1_A genotyping platform and the allele frequency distribution (frequency difference per matched SNP < 0.005%) of the Neanderthal set. We performed a PheWAS meta-analysis of the non-Neanderthal SNPs following the same protocol as above, but using the hard calls for E2.

GCTA

GCTA uses a mixed linear model to estimate the proportion of phenotypic variance explained by SNPs of interest[85,111]. For individuals in the E1 discovery set, we used all directly genotyped Neanderthal variants (on the Human660W-Quadv1_A platform) with a MAF > 1% (1,532 variants) to compute a GRM using the "make-grm" option in the GCTA program (v1.24.4)[85,111]. For individuals in the E2 replication set, we computed a Neanderthal GRM using the same SNPs that were considered in the discovery set. Since the E2 individuals were not all genotyped on the same platform, we used imputed SNPs that passed quality control filters (info score > 0.7 and marker call rate > 99%). This resulted in 1,386 Neanderthal variants with hard call genotypes. For the non-Neanderthal GRM used in the additional two GRM replication analysis, we included all high quality non-Neanderthal variants with a MAF > 1% that were not within 100 kb of a Neanderthal variant (370,306 variants).

We tested a manually curated set of phecodes for ocular, brain, immune, lipid metabolism, digestive, or skin traits (Appendix B). These categories were selected to represent traits that Neanderthal introgression has been hypothesized to influence in previous

studies[53,54,106,108,118,119]. Phecodes tested in the discovery analyses either had a case prevalence > 20% or had an association with a nominally significant *P* value in a preliminary PheWAS analysis; 46 phecodes met these criteria. The phecodes tested in replication analyses had a *P* value < 0.1 in the discovery analyses (12 phenotypes).

In the discovery analyses, we used GCTA to estimate the variance in risk explained by Neanderthal SNPs for 46 phenotypes using a Neanderthal GRM generated as described above. In the replication analyses, we tested the 12 phecodes found to be nominally significant in the discovery analysis using the E2 Neanderthal SNP GRM. We additionally tested these 12 replication phenotypes in a GCTA analysis with a Neanderthal and non-Neanderthal GRM fitted in the same model. We included age, sex, and eMERGE site (dummy coded as for PheWAS described above) as covariates in both replication and discovery analyses. In each analysis, we used disease prevalence estimates from European descent populations when available: Depression (15.0%)[120], Actinic keratosis (38.0%)[121], Obesity (30.2%)[122], Hypercholesterolemia (26.9%)[123], and Anxiety disorder (18.0%)[124]. All other phenotypes were tested without using the prevalence GCTA function.

To ensure that the differences in percent risk estimated between E1 and E2 were not due to the variants that did not pass QC in E2, we also reran our discovery analyses without these variants. There was a negligible difference between those results and our original results (Appendix C), suggesting that these variants are not the reason for the differences in percent risk explained seen between E1 and E2.

Individual Neanderthal SNP effects were estimated using the best linear unbiased prediction (BLUP) approach in the GCTA package[85]. We calculated BLUPs for the 12 significant or nominally significant phenotypes in both E1 and E2. We analyzed the genomic

distribution of the 10% of SNPs with the highest and lowest BLUPs for actinic keratosis and depression using the Genomic Region Enrichment of Annotations Tool (GREAT) with the default basal plus extension settings[125]. These settings define the regulatory domain for each gene to be at least 5 kb upstream and 1 kb downstream of the gene boundaries. This is then extended to the nearest gene's regulatory domain, or up to 1 Mb in either direction, whichever is closer.

eQTL Data

We examined two studies that identified *cis*-eQTL in the brain. Zou *et al.*[126] quantified expression levels of 24,526 transcripts in the cerebellum and temporal cortex of autopsied patients with Alzheimer's disease (AD; 197 cerebellum, 202 temporal cortex) and patients with other brain pathologies (non-AD; 177 cerebellum, 197 temporal cortex) using Illumina's Whole Genome DASL assay. The patients were genotyped on the Illumina HumanHap300-Duo Genotyping BeadChip. They then tested SNPs within 100 kb of the quantified transcripts for association with expression level. These analyses were performed for the AD, non-AD, and combined cohorts for each tissue. To maximize power, we analyzed the association *P* values from the combined set.

We also analyzed eQTL in the cerebellum and parietal cortex from the ScanDB database. These were computed from expression and genotyping data originally collected by the Bipolar Disorder Genome Study (BiGS) Consortium[127]. ScanDB provides only the subset of the SNP–gene expression association *P* values for tests with an uncorrected $P < 0.01$. We analyzed all pairs they defined as significant by this threshold and identified whether each variant acted as an

eQTL for any tested gene. For enrichment analyses on data from both studies, we removed any variant that overlapped a gene expression probe.

Imputed Variant PheWAS

To test Neanderthal variants not directly genotyped on the Illumina 660W platform, we examined 6,359 variants with a MAF > 1% that were imputable to our quality control filters above, and were not in strong LD with each other ($r^2 < 0.8$). We used dosages for these variants for association testing in both E1 and E2 and ran a meta-PheWAS using the covariates and methods as described above.

Results

Neanderthal Variants Together Affect Risk for Depression and Actinic Keratosis

Neanderthal variants have been hypothesized to influence many phenotypes in AMHs, including lipid metabolism, immunity, depression, digestion, and hair/skin, on the basis of the enrichment of Neanderthal variants in regions of the genome relevant to these traits[17,53,54,108]. Accordingly, we first tested these hypotheses using GCTA[85] to estimate the phenotypic risk explained by 1,495 genotyped common (MAF >1%) Neanderthal SNPs for a set of 46 high-prevalence phenotypes from the hypothesized categories, using age, sex, and eMERGE site as covariates (Figure 1-A (b,c)). Neanderthal SNPs explained a significant (GCTA likelihood ratio test; FDR < 0.05 over all phenotype tests) percent of the risk in three traits in the E1 discovery cohort (Table 1-A): depression (2.03%, $P = 0.0036$), myocardial infarction (1.39%, $P = 0.0026$), and corns and callosities (1.26%, $P = 0.01$). Neanderthal SNPs also explained a nominally

significant ($P < 0.1$) percent of risk for nine additional traits, including actinic and seborrheic keratosis, coronary atherosclerosis, and obesity (Table 1-A). Of the 12 nominally significant associations, eight replicated in the independent E2 dataset, including actinic keratosis ($P = 0.0059$), mood disorders ($P = 0.018$), depression ($P = 0.020$), obesity ($P = 0.030$), and seborrheic keratosis ($P = 0.045$) at $P < 0.1$ (Table 1-A; likelihood ratio test). We also tested whether the percent of phenotypic variance explained by Neanderthal SNPs remained significant in the context of non-Neanderthal SNPs by including an additional genetic relationship matrix (GRM) computed from non-Neanderthal SNPs across the rest of the human genome in the mixed linear model. Depression ($P = 0.031$), mood disorders ($P = 0.029$), and actinic keratosis ($P = 0.036$) replicated with these stricter criteria in the independent E2 cohort.

Table 1-A. Neanderthal alleles explain risk for human clinical traits. The eight traits for which Neanderthal alleles explained a nominally significant proportion of variance in risk in both the E1 discovery and E2 replication analyses are listed (GCTA, $P < 0.1$). The depression association remained significant after controlling the false discovery rate at 5%. The Neanderthal associations with actinic keratosis, mood disorders, and depression were also maintained in a two GRM model that considered the risk explained by non-Neanderthal variants. Phenotypes are sorted by their E2 *P*-value.

Phenotype	**Discovery (E1)**		**Replication (E2)**		**Replication (E2; two GRM)**	
	Risk Explained	***P***	**Risk Explained**	***P***	**Risk Explained**	***P***
Actinic keratosis	0.64%	0.066	3.37%	0.0059	2.49%	0.036
Mood disorders	1.11%	0.0091	0.75%	0.018	0.68%	0.029
Depression	2.03%	0.0023	1.15%	0.020	1.06%	0.031
Obesity	0.59%	0.048	1.23%	0.030	0.39%	0.27
Seborrheic keratosis	0.77%	0.038	0.61%	0.045	0.41%	0.13
Overweight	0.60%	0.037	0.53%	0.052	0.23%	0.24

Acute upper respiratory infections	0.70%	0.043	0.56%	0.062	0.34%	0.18
Coronary atherosclerosis	0.68%	0.04	0.42%	0.098	0.34%	0.15

These analyses establish the influence of Neanderthal SNPs in concert on the variance in these traits. We estimated individual effects for each SNP by the best linear unbiased predictions (BLUPs); this indicated that a similar number of Neanderthal SNPs increase and decrease risk for each associated phenotype (Table 1-B). To gain insight into the loci driving these associations, we analyzed the genomic distribution of the 10% of SNPs with the highest and lowest BLUPs for actinic keratosis and depression. We found enrichment (FDR < 0.05; hypergeometric test) for many functional annotations: most notably, keratinocyte differentiation and several immune functions for actinic keratosis and regions involved in neurological diseases, cell migration, and circadian clock genes for depression[87] (Figure 1-B).

Table 1-B. Neanderthal SNP BLUP results from GCTA. Significantly replicating results in bold.

Phenotype	**E1 PRE***	**E1 % Risk SNPs**	**E2 PRE**	**E2 % Risk SNPs**
Hypercholesterolemia	0.74%	49.4%	0.20%	50.1%
Overweight	0.60%	46.6%	0.23%	48.6%
Obesity	0.59%	48.1%	0.39%	48.0%
Mood disorders	**1.11%**	**48.2%**	**0.68%**	**52.0%**
Depression	**2.03%**	**48.1%**	**1.06%**	**52.4%**
Anxiety disorder	1.70%	50.7%	0.00%	49.2%
Myocardial Infarction	1.39%	50.6%	0.13%	50.1%
Coronary atherosclerosis	0.68%	51.0%	0.34%	53.4%
Acute upper respiratory infections	0.70%	47.5%	0.34%	49.8%
Corns and callosities	1.26%	48.6%	0.21%	50.5%
Actinic keratosis	**0.64%**	**49.7%**	**2.49%**	**54.1%**
Seborrheic keratosis	0.77%	50.0%	0.41%	52.2%

* PRE = Percent Risk Explained.

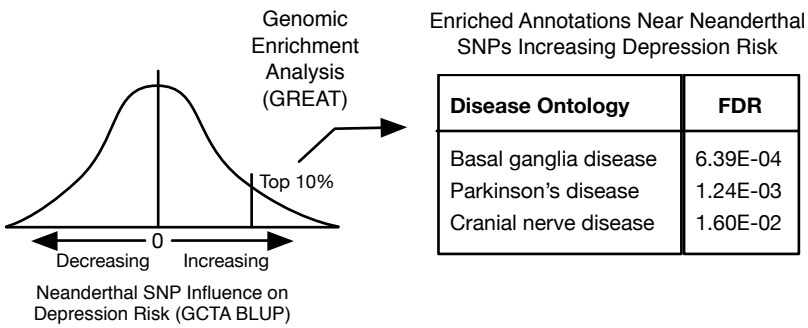


Figure 1-B. Schematic example of functional enrichment analysis on genes nearby Neanderthal SNPs with large BLUPs in the GCTA analyses. We estimated the effects of individual Neanderthal SNPs (BLUPs) and performed genomic enrichment analysis using GREAT[125] on the top 10% most protective and risk increasing SNPs for actinic keratosis and depression. We found enrichment (FDR < 0.05; hypergeometric test) for many functional annotations: most notably, keratinocyte differentiation and several immune functions for actinic keratosis and regions involved in neurological diseases, cell migration, and circadian clock genes for depression[87].

Individual Neanderthal Variants Associate with Clotting and Neurological Disorders

GCTA quantifies the overall influence of Neanderthal SNPs together on traits in AMHs. To identify individual Neanderthal loci associated with AMH phenotypes and potentially discover additional biological systems affected by Neanderthal admixture, we performed a phenome-wide association study (PheWAS) of these 1,495 Neanderthal SNPs with 1,152 EHR-derived phenotypes with at least 20 cases in at least one site (Figure 1-A (d)). PheWAS allows for large-scale characterization of the effects of variants of interest[80]. We carried out two meta-analyses across the eMERGE Network sites—one over the discovery cohort and one over the replication cohort. We focus on the meta-analyses here (Table 1-C; Appendix D), but a pooled analysis using eMERGE site as a covariate produced largely consistent results (Appendix E).

Table 1-C. Individual Neanderthal SNPs with significant replicating phenotype associations. Four locus-wise Bonferroni significant Neanderthal SNP-phenotype associations replicated (with a fixed effect $P < 0.05$ and consistent direction of effect).

					Discovery		Replication	
Phecode	**Phenotype**	**SNP**	**MAF**	**Flanking Gene(s)**	**Odds Ratio**	***P***	**Odds Ratio**	***P***
286.8	Hypercoagulable state	rs3917862	6.20%	*SELP*	3.32	9.9E−7	3.00	5.0E-10
260	Protein-calorie malnutrition	rs12049593	5.15%	*SLC35F3*	1.77	2.0E−6	1.63	5.5E-05
599.8	Symptoms involving urinary system	rs11030043	10.5%	*RHOG*, *STIM1*	1.76	7.4E−6	1.65	4.3E-02
318	Tobacco use disorder	rs901033	1.06%	*SLC6A11*	2.19	1.7E−5	1.75	7.9E-04

We found 105 SNP-phenotype associations passed a locus-wise Bonferroni corrected significance threshold ($P = 3.3E-5$) in the E1 meta-analysis (Appendix F). Four Neanderthal SNP–phenotype associations passed this discovery threshold and replicated ($P < 0.05$) with the same direction of effect in the independent E2 meta-analysis (Table 1-C). The strongest signal was a Neanderthal SNP (rs3917862, 6.5% EUR 1KG frequency) in an intron of P-selectin (*SELP*) that was significantly associated with hypercoagulable state in both E1 and E2 (Table 1-C; Figure 1-A (d)). This haplotype contains several genes directly involved in blood coagulation and inflammation, most notably *SELP*, which encodes a cell adhesion protein expressed on the surface of endothelial cells and platelets that recruits leukocytes to injuries during inflammation. Factor V (*F5*), a coagulation cofactor associated with several coagulation defects, is located ~37 kilobases (kb) downstream. The Neanderthal haplotype overlaps histone modifications suggestive of gene regulatory activity in blood cells and vein epithelial cells (Figure 1-C). Using data from the Genotype-Tissue Expression (GTEx) Project[61], we found indications that the Neanderthal allele at rs3917862 significantly increased the expression of *SELP* ($P = 0.005$) and *F5* ($P = 0.05$) in arteries (Figures 1-A (e); 1-C).

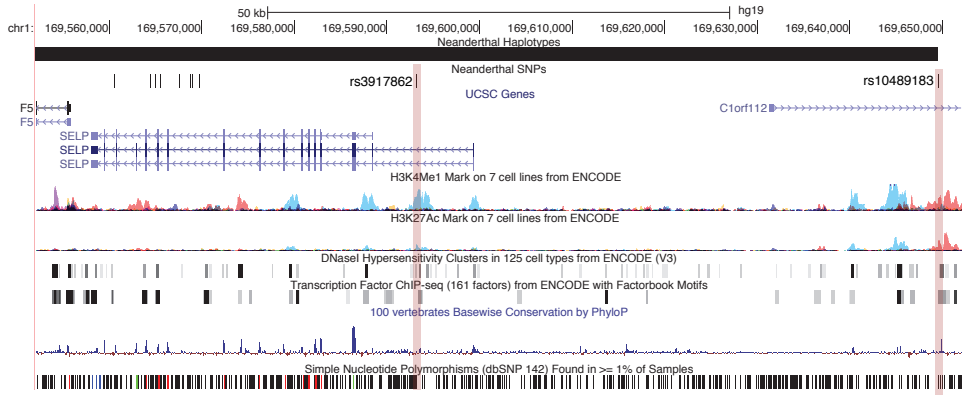


Figure 1-C. A Neanderthal allele located in *SELP* significantly associated with hypercoagulable state and has evidence of gene regulatory function. The Neanderthal SNP, rs3917862, is significantly associated with hypercoagulable state (Figure 1-A (d); Table 1-C). rs3917862 is located in an intron of P-selectin (*SELP*), a gene that mediates leukocyte action at injuries in the early stages of inflammation. This SNP is in LD with rs10489183. These SNPs have functional genomic signatures indicative of gene regulatory activity in blood cells and vein epithelial cells. The Neanderthal allele at rs3917862 is significantly associated ($P = 0.005$) with increased expression of *SELP* in tibial artery data from GTEx (Figure 1-A (e)). It also significantly associates with increased *F5* expression ($P = 0.05$; Appendix G).

The second replicating association was a SNP (rs12049593, 5.0% EUR frequency) in an intron of *SLC35F3*, a putative thiamine transporter that associates with protein-calorie malnutrition. Thiamine is crucial to carbohydrate metabolism for all cells, particularly those with increased energy requirements[128]. Variants in high LD with this SNP (r^2>0.8, D'=1) are found in regions bearing enhancer histone marks in the gastrointestinal (GI) tract, brain, and other tissues. Decreased expression of this transporter in the brain or GI could exacerbate malnutrition or its symptoms. It is possible that new dietary pressures may have caused changes in carbohydrate metabolism to be beneficial in early human migrants out of Africa; indeed, there is evidence suggesting that Neanderthal introgression likely influenced lipid catabolism in Europeans[108]. More recently, the reduction of thiamine present in foods from the grain refining process as well as increased intake of simple carbohydrates, make this a potentially harmful allele, as it could reduce thiamine availability while modern diets increase demand.

Another Neanderthal SNP (rs11030043, 9.0% EUR frequency) is upstream of stromal interaction molecule 1 (*STIM1)* and is associated with a phenotype encompassing incontinence, bladder pain, and urinary tract disorders (Figure 1-D (a)). *STIM1* is a ubiquitously expressed gene involved in intracellular calcium signaling. Variants in high LD with the Neanderthal SNP are found in regions bearing enhancer histone marks and DNase I hypersensitive sites in the brain. Because of this, we examined whether this SNP was associated with gene expression levels in brain tissues in GTEx. The Neanderthal allele is associated with significantly decreased expression of *STIM1* in the caudate basal ganglia ($P = 0.02$; Figure 1-D (b)), a region of the brain connected to bladder dysfunction, particularly in those with neurological conditions such as Parkinson's[129].

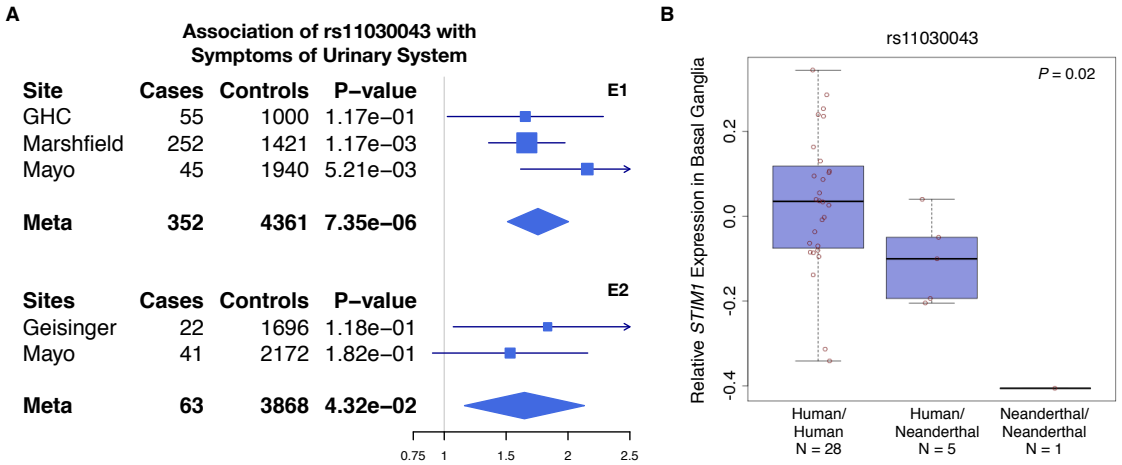


Figure 1-D. rs11030043 is significantly associated with symptoms of the urinary system, and is associated with expression levels of *STIM1*. (a) The forest plot shows odds ratios and *P* values for the association of Neanderthal SNP rs11030043 with symptoms of the urinary system in each site with >= 20 cases. This association was significant in E1 and replicated in E2. (b) rs11030043 is located ~10 kb upstream of stromal interaction molecule 1 (*STIM1*), a transmembrane protein that regulates calcium ion flux, and is significantly associated ($P = 0.02$) with its expression in the caudate basal ganglia.

The last replicated association was between rs901033 (0.5% EUR frequency) and tobacco use disorder. This SNP is in an intron of *SLC6A11*, a solute carrier family neurotransmitter transporter that is responsible for reuptake of the neurotransmitter GABA. Nicotine addiction disrupts GABAergic signaling in the brain and reduces expression of *SLC6A11*[130]. This is the second Neanderthal SNP to be associated with smoking risk[54].

rs3917862 and the Factor V Leiden Mutation

The association between rs3917862 with hypercoagulable state may be influenced by the *F5* Leiden (F5L) mutation; however, the Neanderthal allele appears to have an additional influence on risk. The Neanderthal SNP is in low linkage disequilibrium with F5L (LD, r^2=0.07, D'=0.42), and increases risk for venous thromboembolism (VTE), beyond the risk associated with F5L[131]. Furthermore, manual review of the EHRs for 16 hypercoagulable state cases revealed a diverse set of causes, and only four out of the 11 individuals tested for F5L had the mutation. Due to the direct interaction of coagulation factors with pathogens, these genes have been common targets of positive selection across vertebrate evolution, and *F5* has experienced positive selection in primates[132]. Thus, it is possible that this Neanderthal haplotype and the associated hypercoagulability provided an advantage in early AMHs outside of Africa.

However, due to the large odds ratio (~3) for the association and the proximity (~74 kb downstream) of rs3917862 to the *F5* Leiden thrombophilia mutation (F5L, rs6025), which increases risk for several conditions linked to hypercoagulability in individuals of European ancestry, we investigated whether this Neanderthal SNP could tag associations due to F5L. F5L is overlapped by a Neanderthal haplotype, but appears to postdate introgression. It was not genotyped on the arrays used by eMERGE, but we found modest linkage disequilibrium

(r^2=0.07, D'=0.42) with the imputed F5L and rs3917862. This agrees with previous studies[131] (r^2 = 0.12, D' = 0.37) and our analysis of sequencing data from the 1000 Genomes Phase 3 EUR super-population individuals (r^2 = 0.06, D' = 0.56). Furthermore, manual review of the EHRs for hypercoagulable state cases revealed that only four had a positive F5L genetic test out of 11 directly tested. Using the imputed F5L data, we tested whether we had power to detect an association with hypercoagulable state caused by F5L via rs3917862. We took the imputed frequency of the F5L mutation (2.9%). We used estimates for the genotype relative risk (Aa = 10, AA = 20) from odds ratio estimates of the association with imputed F5L with hypercoagulable state. We took the frequency of rs3917862 (6.2%), hypercoagulable state prevalence (1.6%), case numbers (92), and control numbers (9,540) from the E1 data to compute the power of rs3917862 to tag the F5L association. At our Bonferroni-corrected alpha threshold, we were significantly underpowered to detect an association driven by F5L via rs3917862 (dominant model: 36%; allelic model: 39%)[133]. We also tested a range of values that reflected the extremes of the estimates of these values from the literature. Nearly all remained significantly underpowered; however in a few situations, increasing the F5L mutation frequency to ~5% yielded power above 80%. It is worth noting that this situation is highly unlikely, even in a clinical population.

It is possible that the F5L mutation contributes to the significant observed association between hypercoagulable state and rs3917862. However, the modest LD between these SNPs, the lack of positive F5L tests in the reviewed cases, and our evidence that rs3917862 alters the expression of *F5* and *SELP* in a manner consistent with increased risk suggests an additional role for the Neanderthal allele in hypercoagulability. Furthermore, a recent well-powered study of

VTE demonstrated that rs3917862 increases the risk of VTE beyond the risk associated with F5L[131]. Thus, we conclude that this Neanderthal allele influences hypercoagulable state.

Neanderthal Variants are Enriched for Neurological and Psychiatric Associations

We tested whether Neanderthal SNPs were more likely to be associated with disease phenotypes than non-Neanderthal SNPs. We compared the Neanderthal SNP PheWAS results for an LD-pruned ($r^2 < 0.5$) set (1,056 variants) to those obtained in a PheWAS of 5,280 SNPs with low LD ($r^2 < 0.5$) and an allele frequency distribution matched to the Neanderthal SNPs. These control SNPs correspond to five separate frequency-matched sets of SNPs. The Neanderthal SNPs were 1.22 times more likely to be associated with a phenotype than non-Neanderthal SNPs; however, due to the small number of associations these differences did not reach significance at $P < 0.05$.

To consider a larger number of associations, we analyzed all Neanderthal SNP–phenotype associations at a relaxed significance threshold of $P < 0.001$ in the discovery set that replicated ($P < 0.05$ and same direction of effect). This yielded 60 associations after accounting for hierarchically related phenotypes (Appendix A). Of the 60 associations, 59 (98%) were risk increasing. We compared these results to the 260 associations obtained for the non-Neanderthal SNPs at the relaxed threshold. The Neanderthal SNPs were 1.12 times more likely to be associated with a phenotype than non-Neanderthal SNPs and were less likely to be protective (2% vs. 5%); however, these differences were not significant at $P < 0.05$ (binomial test, $P = 0.2$ and 0.13, respectively).

Next, to test whether specific classes of phenotypes were more likely to be influenced by Neanderthal SNPs, we grouped PheWAS phenotypes into 14 distinct categories used in previous

PheWAS studies[79,80] and compared the distribution of associations for Neanderthal and non-Neanderthal SNPs. Overall, the Neanderthal SNPs associated with a significantly different distribution of phenotypes (chi squared test, $P = 0.017$; Figure 1-E). They were associated with more neurological (binomial test, $P = 0.018$) and psychiatric phenotypes ($P = 0.023$), and fewer digestive phenotypes ($P = 0.004$). These analyses suggest that Neanderthal alleles influence a different set of phenotypes than expected from non-Neanderthal alleles and may be more likely to contribute to disease.

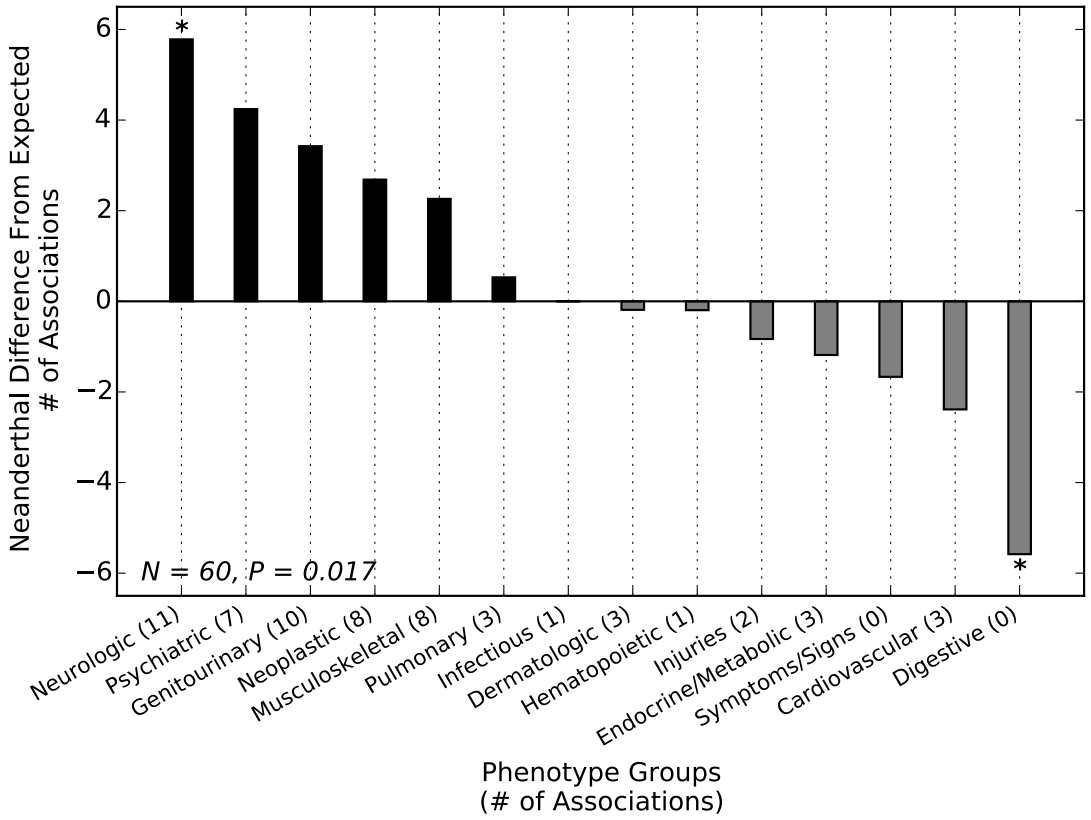


Figure 1-E*. Neanderthal SNPs associate with different phenotype categories than matched non-Neanderthal SNPs. Each bar gives the difference between the number of replicated Neanderthal SNP associations with a phenotype group (at a relaxed discovery threshold of $P < 0.001$) and the number expected from a PheWAS over five sets of non-Neanderthal sites matched to the allele frequency of tested Neanderthal SNPs. The phenotype distributions were significantly different (chi squared test, $P = 0.017$), with more Neanderthal SNPs associated with neurological and

psychiatric phenotypes than expected and fewer digestive phenotypes. The enrichment and depletion were consistent across all five matched non-Neanderthal sets (* indicates $P < 0.05$ for all five comparisons; binomial test).
*This figure is adapted from Figure 2 from my peer-reviewed article[87].

To test if these enrichments and depletions were stable, we used the fact that the 5,280 control (non-Neanderthal) alleles consisted of five independent, non-overlapping sets matched to the Neanderthal alleles tested. We compared the Neanderthal phenotype association distribution to each of these five smaller matched sets in turn, and the phenotype categories at the extremes (psychiatric, neurological, and digestive) were all consistently enriched/depleted across the five comparisons. In particular, there was enrichment for psychiatric phenotype associations in the Neanderthal set across comparisons with all five sets (binomial test, $P < 0.05$). The enrichment for neurological phenotypes was significant ($P < 0.05$) for three and consistent in direction but not significant ($P < 0.2$) for the remaining two. The depletion for digestive phenotypes was present in all five control set comparisons ($P < 0.05$). No other phenotypes were consistently enriched or depleted in more than two of the comparisons. Thus, our finding that Neanderthal alleles are associated with a significantly different set of traits than matched non-Neanderthal alleles is stable across different control sets, and the same phenotypes were consistently significantly enriched and depleted.

Neanderthal Variants are Enriched for Brain eQTL

Given the observed enrichment for psychiatric and neurological phenotype associations among Neanderthal SNPs, we tested whether Neanderthal SNPs were more likely to be expression quantitative trait loci (eQTL) in brain tissues than non-Neanderthal SNPs. We

analyzed previously generated brain eQTL datasets from cerebellum and temporal cortex from Zou *et al.*[126], and cerebellum and parietal cortex from ScanDB[134].

We identified all Neanderthal and control non-Neanderthal SNPs from the enrichment analyses that were directly genotyped and tested in the Zou *et al.*[126] study. We additionally filtered those variants that overlapped a probe on the array. This yielded 663 Neanderthal SNP–gene pairs with association *P* values and 3,295 non-Neanderthal SNP–gene pairs with *P* values. To correct for multiple testing, we calculated q-values[135] from the raw Neanderthal and non-Neanderthal SNP-gene pair *P* values. At a q-value threshold of 0.05, 22 of the 663 Neanderthal SNP–gene pairs (3.3%) and 45 of 3,295 non–Neanderthal SNP-gene pairs (1.4%) were considered significant associations in the cerebellum. This enrichment of eQTL among the Neanderthal SNPs is significant (*P* = 1.68E-04, one-tailed binomial test). These results were robust to q-value thresholds of 0.01 (14 Neanderthal vs. 29 non-Neanderthal SNP-gene pairs; one-tailed binomial test *P* = 2.72E-03) and 0.1 (32 Neanderthal vs. 65 non-Neanderthal; one-tailed binomial test *P* = 5.32E-06). We also found significant enrichment for temporal cortex eQTL among the Neanderthal SNPs: 23 of 683 Neanderthal SNP-gene pairs (3.4%) and 42 of 3,298 non-Neanderthal SNP-gene pairs (1.3%) were eQTL (*P* = 3.49E-05, one-tailed binomial test).

To ensure that our results were not biased by a few variants associating with the expression of multiple genes, we repeated the analysis using only unique SNPs from the significant SNP–gene pairs above. These comparisons also revealed significant enrichment for brain eQTL among Neanderthal SNPs: 21 of 297 Neanderthal variants (7.1%) and 44 of 1,462 non-Neanderthal variants (3.0%) were found in the cerebellum (*P* = 3.2E-04, one-tailed binomial test). In the temporal cortex, 19 of 307 Neanderthal variants (6.2%) and 42 of 1,482 non-

Neanderthal variants (2.8%) were eQTL for at least one gene (P = 1.4E-03). In all, 29 unique Neanderthal SNPs were brain eQTL for at least one transcript in the cerebellum or temporal cortex.

We then tested whether our enrichments held in an independent data set from ScanDB, which examined the cerebellum and parietal cortex. In the cerebellum, 168 of 1,056 Neanderthal variants (15.9%) and 734 of 5,280 non-Neanderthal variants (13.9%) were nominal eQTL; this represents significant enrichment (P = 0.035, one-tailed binomial test). However, in the parietal cortex, 158 Neanderthal variants (15.0%) and 742 non-Neanderthal variants (14.1%) acted as eQTL for at least one gene. This difference was not significant at the 0.05 level (P = 0.209, one-tailed binomial test).

Imputed Variant PheWAS

We tested 6,359 variants likely to be introgressed from Neanderthal, but not directly genotyped on the Illumina 660W. We considered these variants separately from those directly genotyped due to the complicated LD patterns of introgressed haplotypes[53] possibly affecting the imputation quality of the variants not directly assayed. Using a locus-wise Bonferroni corrected discovery significance threshold (7.86E-6), 16 SNP-phenotype associations passed this threshold and replicated in E2 (P < 0.05, OR in a consistent direction; Table 1-D). These represent 14 independent SNP-phenotype associations after accounting for the same variant associating with multiple phecodes in the same hierarchy. One of these associations reaches traditional GWAS genome-wide significance levels (P < 5E-8): rs73735360 with testicular hypofunction. This variant falls near the gene *ADAMTS16*, which when deleted in rats has been shown to cause cryptorchidism and infertility[136]. This result is particularly intriguing in light of the speculation

on the level of genetic, and particularly sexual, incompatibility between AMHs and Neanderthals[54,101,102].

Table 1-D. Imputed Neanderthal variants significantly associate with 16 phecodes.

Phecode	Description	SNP	E1 OR	E1 *P*	E2 OR	E2 *P*	MAF
228	Hemangioma and lymphangioma, any site	rs72668327	2.66	1.97E-06	2.23	2.26E-02	2.97%
573.8	Other specified disorders of liver	rs114638335	2.67	4.24E-06	1.59	2.35E-02	3.20%
284	Aplastic anemia	rs1726521	4.23	1.12E-06	1.77	4.32E-02	2.19%
198.3	Secondary malignant neoplasm of digestive systems	rs77893146	18.50	1.29E-07	2.99	4.11E-02	1.05%
198.3	Secondary malignant neoplasm of digestive systems	rs11098886	12.50	3.10E-06	4.24	2.49E-02	1.48%
80	Postoperative infection	rs80086934	2.67	1.35E-06	1.81	4.54E-03	1.85%
530.9	Heartburn	rs74615305	5.36	3.45E-06	2.39	4.30E-02	1.13%
257	Testicular dysfunction	rs73735360	4.29	3.96E-07	2.01	1.56E-02	3.46%
257.1	Testicular hypofunction	rs73735360	5.16	4.00E-08	2.10	1.04E-02	3.46%
574.12	Cholelithiasis with other cholecystitis	rs4869689	2.08	7.46E-06	1.68	1.89E-02	9.29%
557.1	Celiac or tropical sprue	rs115744110	5.47	6.65E-07	2.93	2.32E-03	12.28%
564	Functional digestive disorders	rs146372280	2.09	2.32E-06	1.44	4.98E-02	2.54%
536	Disorders of function of stomach	rs11982678	2.07	6.28E-06	1.80	4.78E-03	1.82%
536.8	Dyspepsia and disorders of function of stomach	rs11982678	2.20	2.68E-06	2.87	4.58E-05	1.82%
253	Disorders of the pituitary gland and its hypothalamic control	rs5743916	3.42	4.22E-06	2.38	1.24E-02	3.33%
394.7	Disease of tricuspid valve	rs62231172	4.46	7.85E-06	2.27	3.27E-02	1.03%

Discussion

Our approach establishes a new paradigm for understanding the phenotypic legacy of admixture between AMHs and archaic hominins. Using a large clinical cohort, we discovered functional associations between Neanderthal alleles and AMH traits, influencing the skin, immune system, depression, addiction, infertility, metabolism, and others.

The enrichment for nominal associations with psychiatric and neurological phenotypes, influence of Neanderthal SNPs on depression risk, and enrichment for brain eQTL suggest that Neanderthal introgression has influenced AMH brain phenotypes. The significant replicated association of Neanderthal SNPs with mood disorders, and depression in particular, is intriguing since Neanderthal alleles are enriched near genes associated with neuronal synaptic plasticity (specifically, long-term depression)[54], and human–Neanderthal DNA and methylation differences have been hypothesized to affect neurological and psychiatric phenotypes[137,138]. Depression risk in modern human populations is influenced by sunlight exposure[139], which differs between high and low latitudes, and we found enrichment of circadian clock genes near the Neanderthal alleles that contribute most to this association[87].

The replicated nominal association of Neanderthal SNPs with actinic keratosis (precancerous scaly skin lesions) further links introgressed alleles in AMHs to a phenotype directly related to sun exposure. It also suggests that the signatures of adaptive introgression and strong enrichment of Neanderthal alleles near genes associated with keratin filament formation[54] and keratinocytes[53] reflect the influence of Neanderthal alleles on a modern human phenotype. These results, as well as the association with blood coagulation, establish the impact of

Neanderthal DNA on diseases in AMHs that involve traits potentially shaped by environmental differences experienced by non-African populations.

It is possible that some Neanderthal alleles provided a benefit in early AMH populations as they moved out of Africa, but have become detrimental in modern Western environments. Available evidence suggests that Neanderthals had a smaller effective population size than AMHs, and this would have led to weakly deleterious alleles escaping selection until introduced into the larger effective population size of AMHs[102]. Our results suggest that Neanderthal variants are more likely to be deleterious in modern populations, considering the association with testicular dysfunction and the indications that introgressed SNPs may be more likely to associate with clinical phenotypes. However, it is worth noting that most of the 18 variants with replicating associations from the PheWAS are at relatively low frequency (Tables 1-C; 1-D), which does not offer much evidence for these variants being under positive selection.

Indeed, only one "adaptive" variant—defined as variants present on an introgressed haplotype with an allele frequency in excess of 40%[53]—was directly genotyped on our platform, though several more could be imputed, and no significant phenotype associations replicated with any of them. "Adaptive" variants having no clear effect on a clinical trait could be due to many scenarios: the trait it influences does not have a good phecode proxy; heterogeneity between E1 and E2 could obscure true signals; a variant that was once beneficial is now relatively neutral due to removal of the original selective pressure; this haplotype could have risen in frequency due to drift and in fact never been adaptive; or perhaps the genetic architecture of the trait it influences is highly polygenic. In studies such as these, it is important to remember that we are unlikely to discover associations that would have been responsible for any selective pressure thousands of years ago. However, they do provide insight into the biological systems potentially affected,

which may support hypotheses about the effects of introgression at the time of interbreeding or even suggest new ones.

More data are needed to resolve these questions. As more individuals are incorporated into EHR-linked genetic databases and additional whole-genome sequencing data become available for these individuals, it will be possible to more robustly test hypotheses regarding archaic introgression using our approach. As more sophisticated algorithms are developed for extracting phenotypes from EHRs, we anticipate further insights into the functional effects of archaic introgression. Ultimately, the result of these analyses will provide insight into the genetic architectures of the traits influenced by admixture and the strength of purifying selection experienced by introgressed Neanderthal alleles.

Future Directions

Functional validation of the variants identified here will be important in understanding how they impact the traits with which they associate. Given appropriate cell lines to assay, the variants found to be eQTL could be altered using CRISPR or other directed mutagenesis techniques to determine which variant on the haplotype is important for the change in gene expression. As deletion of the *ADAMTS16* gene has been previously shown to disrupt testis development in rats[136], we will test the region encompassing rs73735360 for enhancer activity in the developing gonad in mouse models. If the results of these assays prove promising, we will move on to deeper characterization of this region.

Beyond following up on the results of these analyses, there are many avenues to increase our understanding of the effect of Neanderthal ancestry in modern humans. The first of these

would be to perform similar analyses to those here, but in a clinical population of primarily Asian ancestry, or even admixed groups such as Hispanics. As individuals of Asian descent are thought to have received an additional pulse of Neanderthal introgression[53,54,100], replicating associations with the biological systems implicated in this study—if not the specific phenotypes themselves—would provide additional insight into the effects of Neanderthal introgression. This would also allow for testing of some Neanderthal variation present in Asian individuals but not European, as well as replication of the results here. There are more introgressed haplotypes at high frequency in East Asian populations, making a deeper survey of potentially "adaptive" introgression possible.

Another important area of research is into these "adaptive" haplotypes. Many of them encompass genes involved in skin functions[53,54]. Understanding their effect on human phenotypes may require study of populations who are phenotyped for non-clinical traits, such as skin pigmentation. However, the skin is also an important defense against pathogens, so a study of individuals with a predisposition to skin infections or in humanized mice may provide insight into whether the introgressed Neanderthal haplotypes at these loci were under sexual or natural selection.

CHAPTER II: HUMAN-SPECIFIC AND HOMININ-DERIVED VARIANTS IMPACT CLINICAL TRAITS RELATED TO BIPEDALISM AND IMMUNITY

Introduction

Variation arising after our divergence from chimpanzee has garnered a great deal of attention for its potential to give us insight into the traits that define us: the shift to bipedalism, extensive loss of fur, and increased brain size, to name a few. While only ~5% of the human genome differs from that of the chimpanzee genome—either through lineage-specific single nucleotide substitutions[8], indels[8], or segmental duplications[9]—this represents ~160 megabase pairs (Mbp) of sequence differences, only some of which are expected to differ due to selection rather than stochastic processes. Though the genetic changes key to large shifts in any of these human-defining traits are expected to be fixed in modern AMH populations, variation subtly influencing these traits could still be polymorphic in AMHs today. For example, brain volume and architecture is consistent across AMHs when compared with chimpanzee, but more modest variation in volume of brain structures and overall volume exists across AMHs. Identifying these variants and their mechanism of action may give us insight into which fixed differences were important in hominin and human evolution.

While there are many sites in the genome where AMHs are polymorphic for a derived allele, we expect few of these to have been or to currently be under selection, or to affect human-defining traits. The age of an allele, coupled with how quickly it has risen in frequency, can suggest that it has been under positive selection in recent history. However, an allele can take many paths to reach the same frequency. Sequence data from ancient AMH and archaic hominin

specimens allow us to date the appearance of a variant as well as the rate at which their frequencies have changed in human populations. In this analysis, we examine the effects of remaining ancestral alleles at sites in the genome where the human-derived allele is at high frequencies in modern populations. In particular, we examined polymorphic sites where the derived allele in question is not found in chimpanzee and is either completely absent—or present but not fixed—in archaic hominin individuals.

Methods

Variant Selection

We used the list of human-specific and hominin-derived variants generated by Prufer *et al*[17]. Briefly, these variants were identified as those: having the human-derived allele at >90% global allele frequency (1000 Genomes Phase 1)[20], where the ancestral allele matches the chimpanzee reference, and where the derived allele was not homozygous in both Altai Neanderthal and Denisova. We removed variants that were polymorphic in chimpanzee individuals present in the Great Ape Genome Project[140] as well as those where the chimpanzee reference allele did not match the ancestral allele determined in 1KG[20]. As in Chapter I, variants were restricted to those directly genotyped on the Illumina 660W platform with a minor allele frequency >1% and that had dosages available for E2 individuals after imputation. We calculated linkage disequilibrium with PLINK v1.9[141], and pruned variants in strong LD ($r^2 > 0.8$) with another variant that met these criteria. This resulted in 1,528 human-specific variants and 1,252 hominin-derived variants.

A human-specific variant is defined as a variant that fits the above criteria and the human (derived) allele is not present at that location in the Altai Neanderthal or Denisovan individuals, suggesting that the allele arose after the divergence of the Neanderthal-Denisovan ancestor and AMHs. Hominin-derived variants are those where at least one allele in either archaic hominin (but not all four) match the human allele, suggesting that the allele arose before the divergence of the Neanderthal-Denisovan ancestor and AMHs. It is worth noting that we do not have population samples of the Neanderthals or Denisovans, so variants we call human-specific may have in fact been shared between AMHs and these archaic hominins.

Population, Genotypes, and PheWAS

We tested for SNP-phecode associations using the same individuals and methods to generate hard call genotypes for E1 and dosages for E2 as described in Chapter I. Phecode case/control status was generated in the same way (PheWAS package v0.10.2-2). Age, sex, and the first three PCs were used as covariates in the meta-PheWAS. We considered variants with a discovery $P < 1.8E-05$ (locus-wise Bonferroni correction), with a consistent OR direction in the replication set and $P < 0.05$ to be significantly replicated.

Linkage Disequilibrium Calculations

Upon examination of the tested variants with significant associations, we noticed that several of them overlapped Neanderthal introgressed haplotypes[53]. Using European (EUR) and East Asian (EAS) individuals from 1KG Phase 3[21] and vcftools[142], we calculated D' and r^2 between Neanderthal introgressed variants and all human-specific and hominin-derived variants that met the criteria stated above. We calculated these values separately in each population. As

there is variability in the exact introgressed haplotype boundaries between 1KG individuals, for the given haplotype we report the Neanderthal introgressed variant with the strongest LD values with the human-specific or hominin-derived variant.

Results

Human-Specific Variant Associations

In a PheWAS of 1,528 human-specific variants, we found six significant associations. Two of the human-specific variants were significantly associated with skeletal phenotypes (Table 2-A). One variant falls downstream of runt-related transcription factor 2 (*RUNX2*) and chloride intracellular channel 5 (*CLIC5*), and is significantly associated with fracture of vertebral column. *RUNX2* has a well-established role in osteoblast differentiation and skeletal development[143]. Protein-coding changes in this gene are hypothesized to be one of the main drivers in skeletal morphology differences between Neanderthal and human[15], supported by indications of selection acting on this gene early in AMH history[144]. Mutations in this gene cause cleidocranial dysplasia, a disorder affecting skeletal morphology and growth[145]. Variants near this gene have never been associated with bone mineral density, though they have been associated with height[146–150] and very recently facial variation[151]. The other significant skeletal association is between a variant near macrophage scavenge receptor 1 (*MSR1*) with fracture of ankle and foot. Osteoclasts are formed from macrophages[152], and *MSR1*, also known as scavenger receptor A (*SR-A*), is involved in many macrophage functions, in particular osteoclast differentiation and function[153,154].

Table 2-A. Six human-specific variants significantly associate and replicate with a clinical phenotype.

Phecode	Description	SNP	MAF	Flanking Gene(s)	E1 OR	E1 *P*	E2 OR	E2 *P*
112.3	Candidiasis of skin and nails	rs10416005	1.80%	*UQCRFS1*	6.69	1.61E-05	3.39	1.08E-02
224.1	Benign neoplasm of eye, uveal	rs1774053	2.17%	*RFPL4B*	2.69	1.03E-05	3.68	3.84E-02
289.8	Polycythemia vera, secondary	rs13353661	1.81%	*IL21, BBS12*	5.42	1.37E-05	3.33	1.78E-02
300.8	Acute reaction to stress	rs764229*	2.90%	*WDR72, UNC13C*	2.61	1.41E-05	2.85	1.58E-02
801	Fracture of ankle and foot	rs7838403	1.09%	*MSR1*	2.70	4.63E-06	2.05	2.62E-03
805	Fracture of vertebral column without mention of spinal cord injury	rs2396558	5.58%	*RUNX2, CLIC5*	1.84	4.68E-06	1.41	4.71E-02

*Indications of ancestral allele being reintroduced on Neanderthal haplotype in 1KG EUR.

We find an association between a hominin-derived variant and a neurological phecode (vertiginous syndromes and other disorders of vestibular system; Table 2-C), and a human-specific variant with a psychiatric phecode (acute reaction to stress; Table 2-A). The variant rs764229, associated with acute reaction to stress, is nominally associated with the expression of unc-13 homolog C (*UNC13C*), a gene involved in synaptic transmission, in the frontal cortex (Figure 2-A (a), $P = 0.011$), a region of the brain that is important in stress response[155]. In whole blood, rs17032822 significantly associates with decreased expression of breast carcinoma amplified sequence 2 (*BCAS2*), which is part of the spliceosome (effect size = -0.29, $P = 0.0041$). While a direct link between *BCAS2* and vertiginous syndromes is not currently known, several neurological disorders are caused by defects in RNA splicing[156,157]. In both instances, the ancestral allele is in strong LD ($r^2 > 0.6$, D' = 1) with variants that were introduced through introgression from Neanderthals in 1KG Phase 3[21] European (EUR) and/or East Asian (EAS) individuals (Table 2-B). This suggests that the human-derived allele was fixed or close to fixed

in the ancestors of modern Eurasians at the time of Neanderthal introgression, and that this introgression reintroduced the ancestral allele at these positions.

Table 2-B. The ancestral alleles at two variants are in strong LD with Neanderthal introgressed variants.

Pop.	Tested SNP	1KG MAF	Introgressed Haplotype	Neanderthal Lead SNP	1KG MAF	r^2	D'
EUR	rs764229*	3.68%	chr1:114,911,588-115,109,434	rs138210616	3.68%	1.0	1.0
EUR	rs17032822	2.19%	chr15:54,041,869-54,164,960	rs77032474	1.49%	0.68	1.0
EAS	rs17032822	7.34%	chr15:54,041,869-54,164,960	rs78798946	7.14%	0.97	1.0

*rs764229 is monomorphic in 1KG EAS individuals.

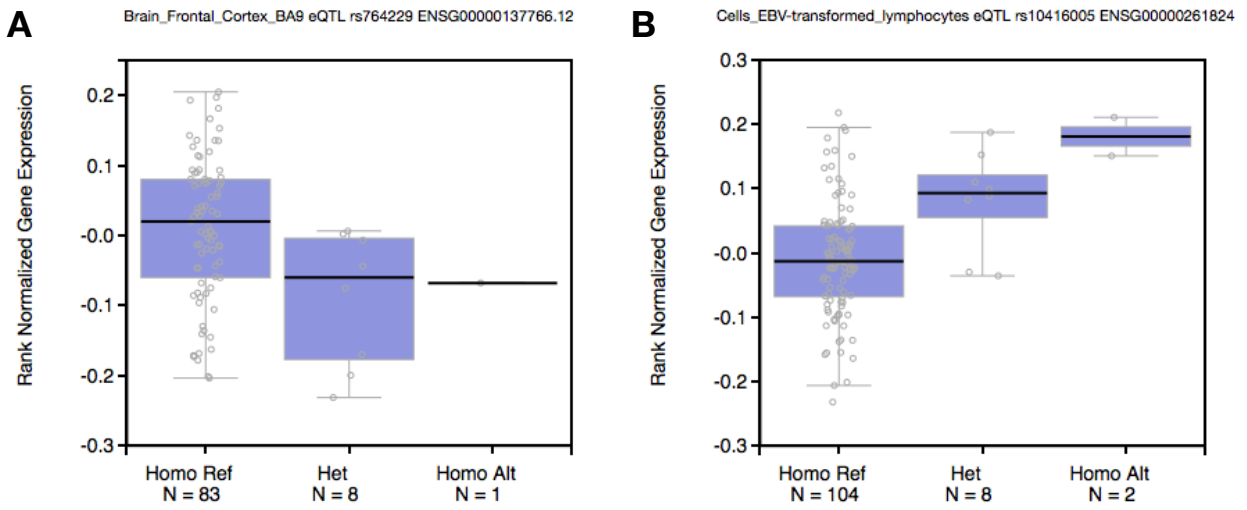


Figure 2-A. Human-specific variants significantly associate with the expression of nearby genes. (a) rs764229 nominally associates with the expression of *UNC13C* in the frontal cortex (GTEx Release V6p[61], effect size = -0.49, $P = 0.011$). (b) rs10416005 significantly associates with the expression of *LINC00662* in Epstein Barr virus-transformed lymphocytes (GTEx V6p[61], effect size = 1.1, $P = 1.1E-05$).

For the other human-specific variant associations, possible functional mechanisms are not as clear. For the association of rs10416005 with candidiasis of skin and nails, the variant is significantly associated with the expression of *LINC00662* in Epstein Barr virus-transformed

lymphocytes (Figure 2-A (b), P = 1.1E-05). However, the function of this long non-coding RNA is currently unknown. The variant rs1774053 (benign neoplasm of eye) is in the middle of a gene desert, but the closest cluster of genes involves laminin subunit alpha 4 (*LAMA4*) and WNT1 inducible signaling pathway protein 3 (*WISP3*), which have both been implicated in cancer. Secondary polycythemia vera is a disorder of the body making too many red blood cells, typically in response to chronic low oxygen levels. The variant associated with this phenotype falls between interleukin 21 (*IL21*) and fibroblast growth factor-2 (*FGF2*), both of which have a role in wound healing and hypoxia[158,159].

Hominin-Derived Variant Associations

Of the 1,252 hominin-derived variants tested, we found four significant associations (Table 2-C). As for many of the human-specific variant associations, making connections between the variant and phenotype was difficult. *LYN* is important in mast cell degranulation, and mast cells may be involved in hyperbilirubinemia[160]. Hematemesis could have many causes, and rs10981835 falls in an intron of regulator of G-protein signaling 3 (*RGS3*), which is widely expressed and involved in Wnt and ephrin-B signaling. This variant significantly decreases the expression of this gene in whole blood (Figure 2-B (a), P = 3.6E-08) and a neighboring unannotated gene *C9orf43* in multiple tissues, most significantly in testis (Figure 2-B (b), P = 3.4E-13). Due to the demographics of our population, poisoning by water, mineral, and uric acid drugs is likely predominantly driven by intolerance of uric acid drugs for the treatment of gout. This variant falls near several long non-coding RNAs of unknown function and a cluster of serine proteinase (serpin) peptidase inhibitors, class B (*SERPINB*), which are active in many biological processes.

Table 2-C. Four hominin-derived variants significantly associate and replicate with a clinical phenotype.

Code	Description	SNP	MAF	Flanking Gene(s)	E1 OR	E1 *P*	E2 OR	E2 *P*
386	Vertiginous syndromes and other disorders of vestibular system	rs17032822*	3.23%	*DENND2C*	1.86	1.67E-05	1.44	3.7E-02
573.5	Jaundice (not of newborn)	rs16922470	2.57%	*LYN*	6.29	8.53E-10	1.92	4.8E-02
578.1	Hematemesis	rs10981835	14.68%	*RGS3*	3.74	5.75E-07	1.81	7.9E-03
974	Poisoning by water, mineral, and uric acid metabolism drugs	rs12605877	2.06%	*SERPINB8*	7.94	4.02E-06	3.07	3.8E-02

*Indications of ancestral allele being reintroduced on Neanderthal haplotype in 1KG EUR & EAS.

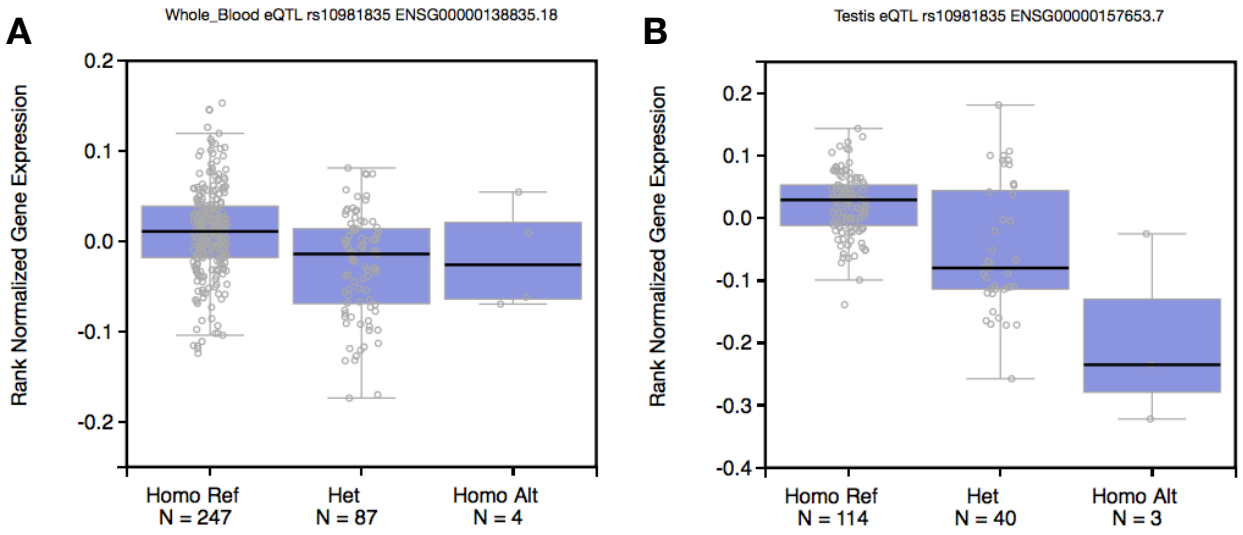


Figure 2-B. Hominin-derived variants significantly associate with expression of nearby genes. (a) rs10981835 significantly associates with expression of *RGS3* in whole blood (GTEx Release V6p[61], effect size = -0.22, *P* = 3.6E-08). (b) rs10981835 significantly associates with the expression of *C9orf43* in testis (GTEx V6p[61], effect size = -0.78, *P* = 3.4E-13).

Discussion

We found four significant associations with hominin-derived variants and six associations with human-specific variants. These associations are with clinical phenotypes impacting a range of biological systems, though four are of particular interest: the two associations with bone fracture, and the neurological and psychiatric associations. The two associations between human-specific variants and bone fracture are unlikely to reflect selection during the transition to bipedalism, as this would have been shared between humans and Neanderthals. However, there are many skeletal differences in morphology between the two—such as the attachment of the shoulder and clavicle, rib cage shape, and craniofacial morphology—in addition to Neanderthal skeletons being generally much more robust. These morphological differences make it clear that these two groups diverged in the genetics underlying development and/or maintenance of their skeletons. Alternatively, the ancestral allele at these positions may have carried a similar increased risk for fracture in Neanderthals, but the derived alleles may have not appeared on their lineage before their extinction.

The neurological and psychiatric associations are interesting in that the ancestral alleles increasing risk for both of these phecodes appear to have originated in Neanderthal. Ancestral alleles being reintroduced through interbreeding with Neanderthal raises questions about whether the causal variant in these associations is a Neanderthal derived or reintroduced ancestral allele. Recent work has suggested that at least in one instance at the 2'-5' oligoadenylate synthetase (*OAS*) locus, it is a reintroduced ancestral allele that has the functional effect on immune response rather than introduction of a Neanderthal derived allele[107]. It could be that the tested variants in this analysis are also the variants of effect; however, due to the structure of

introgressed Neanderthal haplotypes, another reintroduced ancestral allele or a derived Neanderthal allele in high LD could be the functional variants. Any of these situations is possible, though the relatively high frequency of the ancestral alleles in 1KG Phase 1[20] African super-population (AFR) individuals (rs764229: 12%, rs17032822: 13%) suggests that a variant other than the one directly tested here may be the most likely. However, it is worth noting that more complex scenarios, such as the combined effects of multiple variants on the Neanderthal haplotype, are also possible. Studies of this region in clinical populations of African descent could help shed light on which of these scenarios is most supported. Regardless of the causal variant(s), we have found two more associations that indicate that Neanderthal introgression influences clinical brain phenotypes in European descent populations.

Future Directions

As a follow-up to these studies, we will attempt to functionally validate some of the PheWAS associations. Experiments examining the role of our variants associated with skeletal phecodes are underway. Using cell lines that are skeletal-related, we will begin with determining whether these regions show signs of enhancer activity using luciferase assays in cell culture. If they do, we will delete these regions in these cell lines to determine if they alter expression of the genes discussed here.

We did not perform any analyses looking for signatures of selection on these variants. Indeed, all we know about them is that the derived allele is at high frequency in modern humans, does not appear to be present in chimpanzee populations, and does not appear to be fixed in both Neanderthals and Denisovans. In some cases, the ancestral allele appears to have been lost in the

out-of-Africa migration—though not in African populations—and reintroduced by Neanderthal introgression. This leaves a 6-million-year window where these variants could have appeared, and there may be different expectations of what phenotypes are affected based on when a variant arose and whether it has undergone positive selection or been reintroduced by Neanderthal introgression. Looking for ancient signatures of selection or admixture or estimating allele age could aid in prioritization or classification of variants for testing, or help generate hypotheses to test or refine hypotheses generated from results of PheWAS or GCTA.

CHAPTER III: GC-BIASED GENE CONVERSION INFLUENCES FACTORS THAT DETERMINE STATISTICAL POWER

Introduction

As it ensures proper segregation of chromosomes during meiosis and creates new allelic combinations, meiotic recombination is critical to the success of sexually reproducing organisms. When homologous chromosomes align during meiosis, a double-strand break (DSB) is induced in one strand and several hundred base pairs of each strand are resected in the 5' to 3' direction. One of these strands then invades the homologous DNA duplex and displaces one of the strands of the chromosome that did not experience the DSB. Where these strands from opposing parental chromosomes align is called the heteroduplex region. The invading strand is then extended by DNA synthesis so that it matches the reciprocal strand. This situation can then resolve in a crossover or non-crossover. A successful crossover event results in reciprocal exchange of genetic information, whereas non-crossover results in the invading strand being ligated back to its original strand. For a more detailed molecular discussion of the different models of meiotic recombination, see Chen *et al*[161] or Hunter[162].

Despite its importance, meiotic recombination is far from an error-free process. Major errors in this process can result in improper crossover between different chromosomes or even loss of entire chromosomes, resulting in aneuploidy. However, even successful crossovers can result in mutations. During recombination, the heteroduplex region where the DSB and DNA synthesis occurs is susceptible to a process called gene conversion in both crossover and non-crossover events[161]. If the paternal and maternal chromosomes have a mismatch in this region,

resolution of the DSB can result in one of the chromosomes being "converted" to the state of the other. If this process were random, it would have little impact on the overall population allele frequency, as either allele would be equally likely to be converted. However, a great deal of evidence suggests that the process is biased towards the promotion of G or C alleles (referred to as S for "strong") over A or T alleles (referred to as W for "weak") in eukaryotes[163–166], a phenomenon known as GC-biased gene conversion (gBGC). Here, I will refer to all biallelic variants where one allele is a W and the other an S as "WS variants." Though the strength of this bias has been estimated to be relatively weak[49,167], it can have broad ramifications for both the genome at large and evolutionary studies. Sustained gBGC near recombination hotspots increases local GC content, promotes fixation of S alleles at WS variants, and can be mistaken for signatures of positive selection[47–50].

Because gBGC promotes increased transmission of S alleles at WS variants in a way that resembles weak positive selection, this process could theoretically counteract negative selection on weakly deleterious S alleles and increase their frequency[49,50,168]. Effective population size, recombination rate, and diversity of PR domain zinc finger protein 9 (*PRDM9*) alleles (a key protein in determining recombination hotspot location[169]) influence the intensity and overall genomic impact of gBGC. African groups have the largest effective population size, as well as higher recombination average rates than European or Asian groups. Due to a lack of *PRDM9* allelic diversity, Europeans have fewer recombination hotspots dispersed across the genome than Africans[170]. This results in African groups having the strongest gBGC genome-wide of the three super-populations, but both Asians and Europeans having more concentrated regions of intense gBGC (though this is stronger in Europeans)[171]. Should gBGC appreciably affect modern human health, these differences could suggest population-specific ramifications.

Previous work has demonstrated that gBGC can shift allele frequency spectra[49] and that regions that have undergone gBGC are enriched for particular groups of pathogenic variants[50]. In this study, we evaluated the impact of gBGC on health in modern humans. Using two primarily European descent populations, we explicitly tested whether variants influenced by gBGC are more likely to associate with clinical phenotypes than variants that either cannot undergo gBGC (palindromic, or WW or SS variants), or WS variants that have not undergone detectable gBGC since divergence from chimpanzee.

Methods

We used individuals from both BioVU and eMERGE to test the effects of gBGC. As we do not have a replication set, we treated the analyses conducted in the BioVU population as exploratory analyses, and follow up on one of these comparisons in eMERGE. We additionally examine variants with associations in the GWAS catalog[66], as these should be very similar to the associations tested here.

BioVU Population, Genotypes, and PheWAS

For a description of BioVU, see Roden *et al*[74]. We used genotypes and phenotypes available for 5,357 BioVU individuals who were genotyped on the Illumina OMNI1-quad platform. Variants were removed if their marker call rate was below 95%. Individuals were removed if their sample call rate was below 98%. Phecode case/control status was generated using the PheWAS package (v0.10.2-2). Age, sex, and third party-assigned ancestry were used as covariates. We considered autosomal variants with a MAF > 1% in our tested individuals. We

calculated allele frequency and pruned variants in high LD ($r^2 > 0.8$) using PLINK v1.9[141]. To ensure that minor allele calling was accurate for the palindromic variants, we only considered those with a MAF < 45% in 1KG Phase 3[21] EUR super-population individuals and a MAF < 48% in the eMERGE individuals. We additionally removed variants whose frequency exceeded a 10% allele frequency difference from 1KG EUR individuals within at least one eMERGE site, and was not consistent with a strand flip.

BioVU Variant Selection

The first comparison using the BioVU data was to test for the effect of variants that are in phastBias[50] gBGC tracts—designed to detect gBGC occurring in AMHs since divergence from chimpanzee—and have high recombination rates (denoted "hotspot variants") compared to those that are not in gBGC tracts and have very low recombination rates (denoted "long-time coldspot variants"). We calculated the number of LD partners ($r^2 > 0.5$) for each variant in EUR super-population individuals from 1KG Phase 3[21]. The local recombination rate was calculated by taking the weighted average of the recombination rate from Kong *et al*[172] 50 kb upstream and 50 kb downstream of each variant. The number of LD partners is correlated with recombination rate, but not minor allele frequency (Figure 3-A), so we accounted for all of these in our matching. Our hotspot variants were those that were in the top 2 deciles for recombination rate among all WS variants that fell into a gBGC tract. We binned all WS variants falling into a gBGC tract into deciles for S allele frequency, local recombination rate, and number of LD partners separately. We selected up to 10 WS variants matched on decile for S allele frequency and number of LD partners for each hotspot variant that also fell into the bottom 2 deciles for recombination rate. This resulted in 480 hotspot variants and 4,017 matched variants.

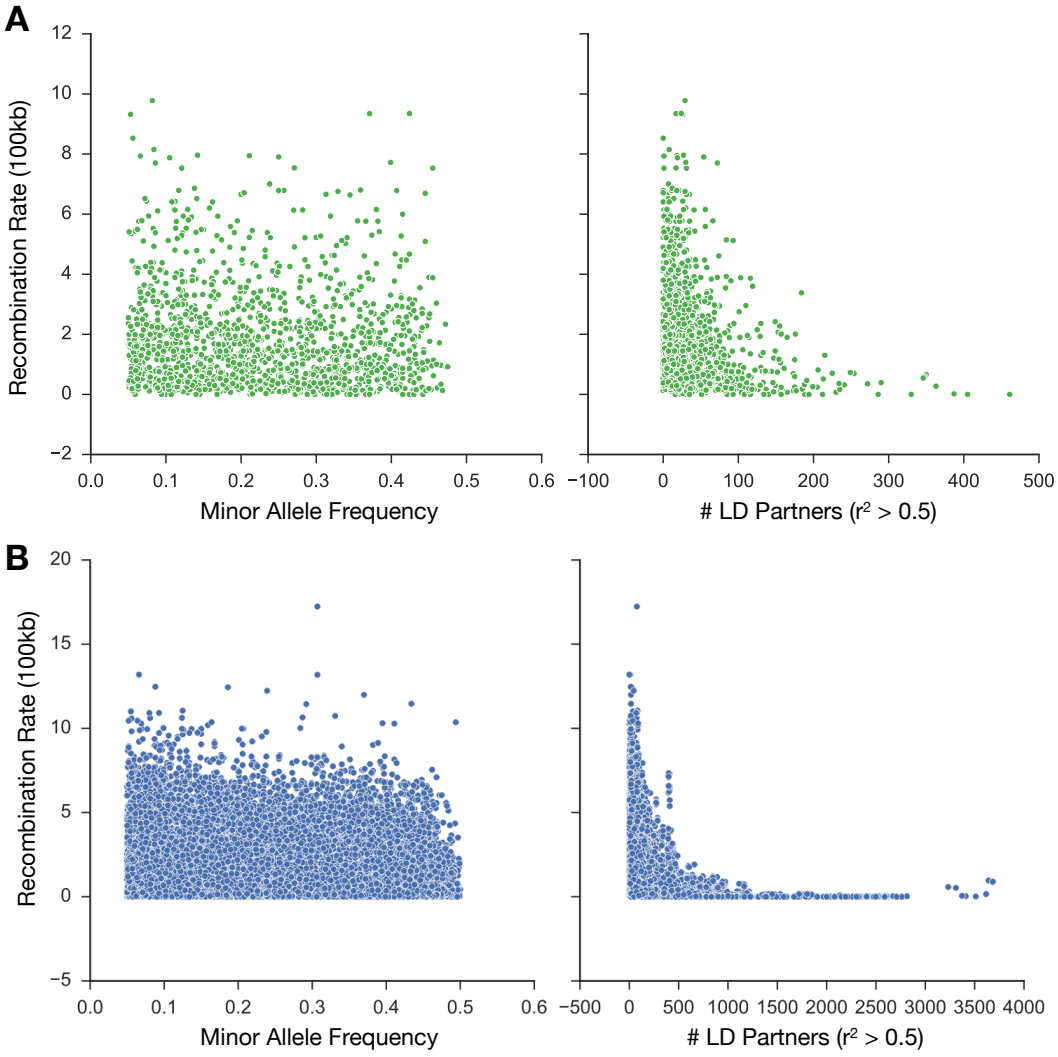


Figure 3-A. A variant's number of LD partners and recombination rate are correlated, but not minor allele frequency. (a) For each gBGC variant (N=1,469) in the eMERGE gBGC vs. palindromic comparison, the minor allele frequency, local recombination rate, and number of LD partners ($r^2 > 0.5$ in EUR individuals) are plotted. (b) For each imputable palindromic variant in the eMERGE population (N=275,537), the minor allele frequency, recombination rate, and number of LD partners are plotted. As expected, regardless of variant type, the number of LD partners is inversely correlated with the rate of recombination. Even gBGC variants with low recombination rates often have fewer LD partners than palindromic variants, possibly reflecting historic high recombination rates.

The second comparison using the BioVU data was to test for differences in phenotype associations between all variants falling in a gBGC tract (denoted "gBGC variants") when compared with palindromic variants, which cannot undergo gBGC. As above, we calculated the number of LD partners ($r^2 > 0.5$) for each variant in EUR super-population individuals from 1KG Phase 3[21]. The local recombination rate was calculated by taking the weighted average of the recombination rate[172] 50 kb upstream and 50 kb downstream of each variant. We binned all WS variants falling into a gBGC tract into deciles for minor allele frequency, local recombination rate, and number of LD partners separately. We generated one palindromic match variant per gBGC variant, matching on all three of these parameters. This resulted in the selection of 923 variants of each class.

eMERGE Population, Genotypes, and PheWAS

We tested for SNP-phecode associations using the same individuals and methods to generate dosages for both E1 and E2 as described for the E2 meta-PheWAS in Chapter I. Phecode case/control status was generated in the same way (PheWAS package v0.10.2-2). Age, sex, and the first three PCs were used as covariates in the meta-PheWAS. We used a range of discovery p-value thresholds to ensure that differences between sets were not dependent on significance threshold choice. Regardless of discovery significance threshold, we considered variant-phenotype associations with a consistent OR direction of effect between the discovery and the replication sets and E2 $P < 0.05$ to be replicated. We used the hard call genotypes to calculate allele frequency and prune variants in high LD ($r^2 > 0.8$) using PLINK v1.9[141]. As we are using imputed variants, we have a larger pool of variants to choose from. Thus, we increased our MAF threshold to > 5% across all tested individuals, which should increase statistical power.

To ensure that minor allele calling was accurate for the palindromic variants, we only considered variants with a MAF < 45% in 1KG Phase 3[21] EUR super-population individuals and a MAF < 48% in the eMERGE individuals. We additionally removed variants whose frequency exceeded a 10% allele frequency difference from 1KG EUR individuals within at least one eMERGE site, and was not consistent with a strand flip.

To determine how many associations should be expected by chance with variants of these allele frequencies, we performed a permutation analysis. We shuffled the individual-genotype relationships among the entire tested population and reran the meta-PheWAS using the same methods as discussed above. Permutation analyses enable estimations of "noise" in the data. This is useful in determining how much "signal" is detectable in the original analysis. For our purposes, this is also useful for determining how much difference between the two groups of variants is detectable.

eMERGE Variant Selection

As there were only enough variants to generate one matched set in the BioVU gBGC-palindromic comparison, this analysis serves as a tool to confirm and expand upon those results. As above, we calculated the number of LD partners ($r^2 > 0.5$) for each variant in EUR super-population individuals from 1KG Phase 3[21]. The local recombination rate was calculated by taking the weighted average of the recombination rate from Kong *et al*[172] 50 kb upstream and 50 kb downstream of each variant. We binned all WS variants falling into a phastBias[50] gBGC tract (denoted "gBGC variants") into deciles for minor allele frequency, local recombination rate, and number of LD partners separately. We selected up to 5 palindromic (WW or SS) variants

matched on decile of all three parameters for each gBGC variant. This resulted in 1,469 gBGC variants and 7,096 matched palindromic variants.

GWAS Catalog Enrichment

We downloaded all variants in the National Human Genome Research Institute (NHGRI) GWAS catalog[66] (downloaded August 10, 2016) nominally associating with any phenotype ($P <$ 1E-05), and removed any that were palindromic. We intersected previously identified GC-biased gene conversion tracts from phastBias[50] with the non-palindromic GWAS catalog variants using the intersectBed function from the bedtools suite[173] and calculated the overlap. We used the function "chi2_contingency" from the python scipy stats module with standard options to calculate the chi square test for independence of variables in a contingency table. Compared values are given in Table 3-A.

Table 3-A. Contingency table used to calculate GWAS catalog enrichment.

Category	Total	gBGC Tract	X^2	X^2 *P*
Genome Coverage (bp)	3,200,000,000*	7,389,204	-	-
WS GWAS (*P* < 1E-05)	17,794	93	63.99	1.25E-15
Significant WS GWAS (*P* < 1E-07)	6,784	30	12.16	4.88E-4
Palindromic GWAS (*P* < 1E-05)	1,890	9	3.90	0.048
Significant Palindromic (*P* < 1E-07)	713	4	2.07	0.150

*Approximation of the length of the haploid human genome used for contingency calculations.

Two comparison sets were generated. The first (not GC-matched) used shuffleBed[173] to generate 1000 sets of length-matched regions. We constrained the generated regions to the

chromosome of the corresponding observed region and did not allow shuffled regions to overlap gaps in the genome assembly or ENCODE blacklist regions[62]. The second comparison generated 100 sets of regions matched by chromosome, length, and GC content using a custom script. For both comparisons, we calculated the overlap between the generated regions and the non-palindromic GWAS catalog variants for each set. We repeated our analyses after restricting to variants that were genome-wide significant ($P < 1E-07$).

We determined the number of LD partners for GWAS catalog variants in 1KG EUR individuals as above ($r^2 > 0.5$). Additionally, we intersected these variants with the 10 kb averaged recombination rate map from Kong *et al*[172]. As averaging recombination rates over the surrounding 100 kb dramatically narrows the range of possible values, we considered only the immediate 10 kb recombination rate to better discern differences in recombination rate between variants.

Results

Few Tested Variants Have Genome-Wide Significant Associations

Between our three comparisons, we tested nearly 15,000 variants for clinical phenotype associations. In order to get a sense for the number of associations, we examined all sets for SNP-phenotype associations surpassing a genome-wide significance threshold ($P < 5E-08$). For both the hotspot-coldspot and gBGC-palindromic comparisons performed in our BioVU population, we have no replication set. If we use a genome-wide significance threshold, we find one hotspot variant and one coldspot variant each associate with a phenotype (Table 3-B). The hotspot variant rs7578066 associates with blindness and low vision. The coldspot variant

rs3129871, which falls near *HLA-DRA* and associates with Type 1 diabetes, has been previously found to associate with multiple sclerosis[174].

Table 3-B. Significant results of PheWAS.

Code	Description	SNP (Tested Allele)	Type	E1 OR	E1 *P*	E2 OR	E2 *P*
367.9	Blindness and low vision	rs7578066 (C)	Hotspot	3.53	4.9E-08	-	-
250.12	Type 1 diabetes with renal manifestations	rs3129871 (A)	Coldspot	0.44	6.7E-09	-	-
274	Gout and other crystal arthropathies	rs45499402 (C)	Palindromic	1.71	5.1E-13	1.98	4.0E-19
274.1	Gout	rs45499402 (C)	Palindromic	1.84	9.1E-15	2.06	2.9E-20

In contrast, no variant in the BioVU gBGC-palindromic comparison associates with a phenotype at this threshold. However, one of the palindromic variants tested in the eMERGE gBGC-palindromic analysis, rs45499402, has been previously found to associate with gout[175], and significantly associates (P = 9.10E-15) and replicates (P = 2.93E-20) with gout in our analysis (Table 3-B). However, none of the gBGC variants significantly associate at this level and replicate.Reassuringly, no variant associates beyond a genome-wide significance threshold and replicates in the eMERGE permutation analysis. Given the lack of genome-wide significant associations, we considered a range of discovery thresholds moving forward when comparing within these sets.

Hotspot Variants Do Not Associate with More Clinical Phenotypes than Coldspot

The phastBias[50] gBGC tracts used in these analyses were designed to detect signatures of gBGC that has occurred since divergence with chimpanzee. Thus, some of these tracts will

represent regions that used to be recombination hotspots, but are no longer. This means that selection may be able to act effectively in some of these regions once again, and may alter our expectations depending on when recombination hotspot activity was lost. To reduce our set to variants likely to still be undergoing gBGC, which we will refer to as "hotspot" variants, we limited to gBGC variants in the top 2 deciles of local recombination rate. For this comparison, we selected control WS variants that did not fall in a gBGC tract (thus are unlikely to have undergone gBGC since divergence with chimpanzee) and fell into the 2 lowest local recombination rate deciles (thus are unlikely to be in recombination hotspots too young to be detectable by phastBias). However, we matched on S allele frequency and number of LD partners. We selected between 5 and 10 matched coldspot variants per each of the 456 hotspot variants and performed a PheWAS in our BioVU population (see Methods). To compare enrichment for associations between the sets, we down-sampled to 5 matched variants per hotspot variant and calculated the number that significantly associated with at least one phenotype. We down-sampled in this way 5 times to get an idea of the variability of the proportion of matched variants that associated with a phenotype.

When we did this, we found little difference between the likelihood of association between clinical phenotypes and hotspot and coldspot variants (Figure 3-B (a)). In fact, coldspot variants appear slightly more likely to associate with a phenotype than hotspot variants. Part of the theory of gBGC leading to increased disease risk is the increased frequency of deleterious alleles despite selection, suggesting that the S allele itself may be more likely to be deleterious than the S allele of a matched variant, so we also compared this likelihood. Of all the variants with a significant association, we considered the S allele to be the risk if the S allele increased risk (e.g., OR > 1 if S is the minor allele) for its most significant association. After focusing on

the proportion of variants with significant associations where the S allele is the risk allele, we find that at our strictest threshold, the S allele is more likely to be the risk allele at hotspot variants than matched coldspot variants, but the difference is small and not consistent across p-value thresholds (Figure 3-B (b)).

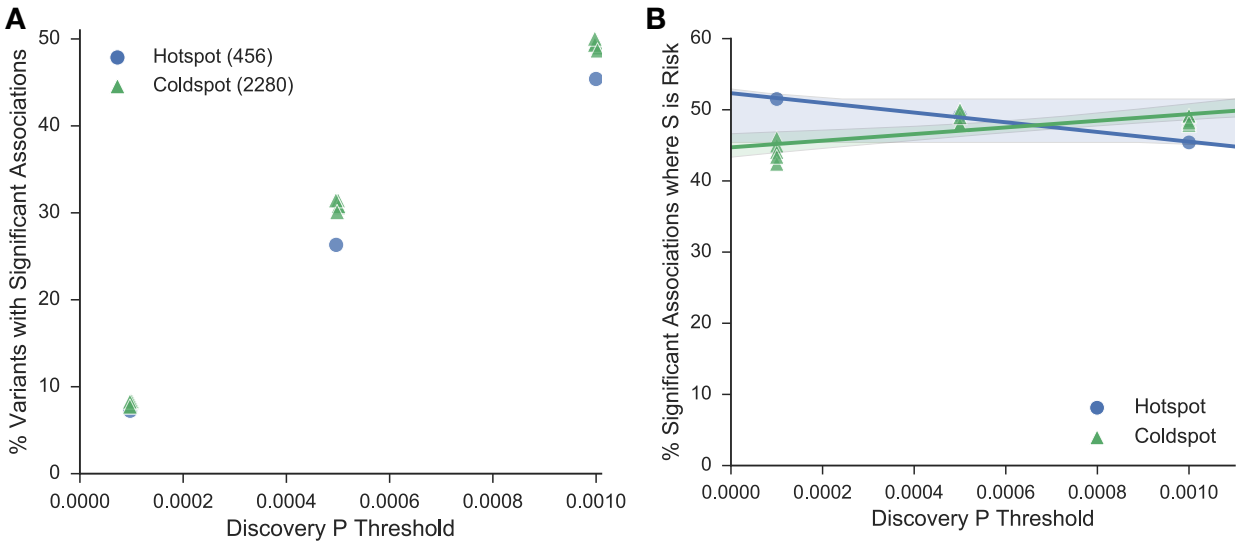


Figure 3-B. Hotspot variants are not more likely to associate with a clinical phenotype than coldspot variants matched by allele frequency and number of LD partners. (a) The percent of variants that have at least one significant association at the given discovery p-value threshold. (b) The percent of variants that have at least one significant association at the given discovery p-value threshold where the S allele is the risk allele. Each iteration of matched variants is plotted.

gBGC Variants Do Not Associate with More Clinical Phenotypes than Palindromic

To investigate whether gBGC variants overall are more likely to associate with a clinical phenotype, we directly tested this in two different populations. In our BioVU population, we identified one palindromic variant matched to a WS gBGC variant on decile for minor allele frequency, local recombination rate, and number of LD partners (see Methods). As above, we performed PheWAS in our BioVU dataset and have no replication set for this analysis. There

appears to be no difference in the likelihood of a gBGC variant associating with a clinical phenotype when compared to a palindromic variant (Figure 3-C (a)). As palindromic variants are either WW or SS by nature, we matched on minor/major allele status of the S allele for the gBGC variant to determine whether the S allele was more likely to be deleterious for gBGC variants. It appears that the gBGC S allele is slightly less likely to be the risk allele than the equivalent palindromic allele (Figure 3-C (b)), though the difference is small and may not hold if we had more comparison sets.

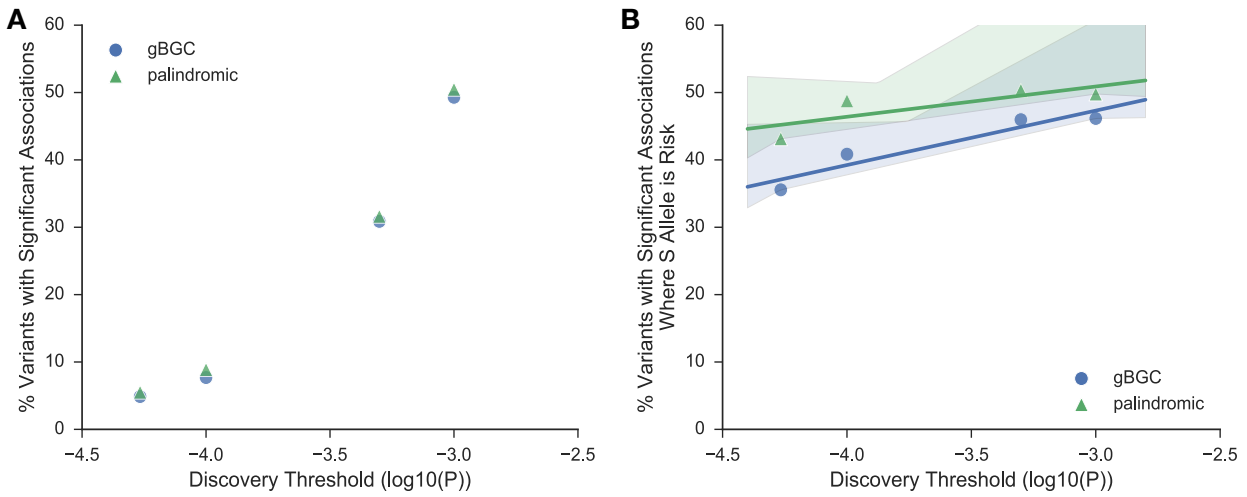


Figure 3-C. BioVU gBGC variants are not more likely to associate with a clinical phenotype than matched palindromic variants. (a) The percent of variants that have at least one significant association at the given discovery p-value threshold. (b) The percent of variants that have at least one significant association at the given discovery p-value threshold where the S allele (or equivalent) is the risk allele.

In the eMERGE population, we identified up to five palindromic variants matched to a WS gBGC variant on decile for minor allele frequency, local recombination rate, and number of LD partners (see Methods). We performed a meta-PheWAS on these variants and considered a range of discovery p-value thresholds to ensure that our results were consistent. Regardless of

discovery threshold, all associations had to have a consistent OR between E1 and E2, and an E2 $P < 0.05$ to be considered replicating. When we did this, we found that gBGC variants and their matched palindromic counterparts had the same likelihood of associating with a clinical phenotype, regardless of discovery p-value threshold (Figure 3-D (a)). Of all the variants with a significant, replicating association, we considered the S allele to be the risk if the S allele increased risk (e.g., OR > 1 if S is the minor allele) for any of its significant, replicating associations. Except for the two lowest p-value thresholds where there are very few associations (1 and 3, respectively), we see little to no difference between the gBGC and palindromic variants (Figure 3-D (b)). At these low thresholds, the variability in the percentage of associations where the S allele is the risk allele is extreme for the palindromic variants.

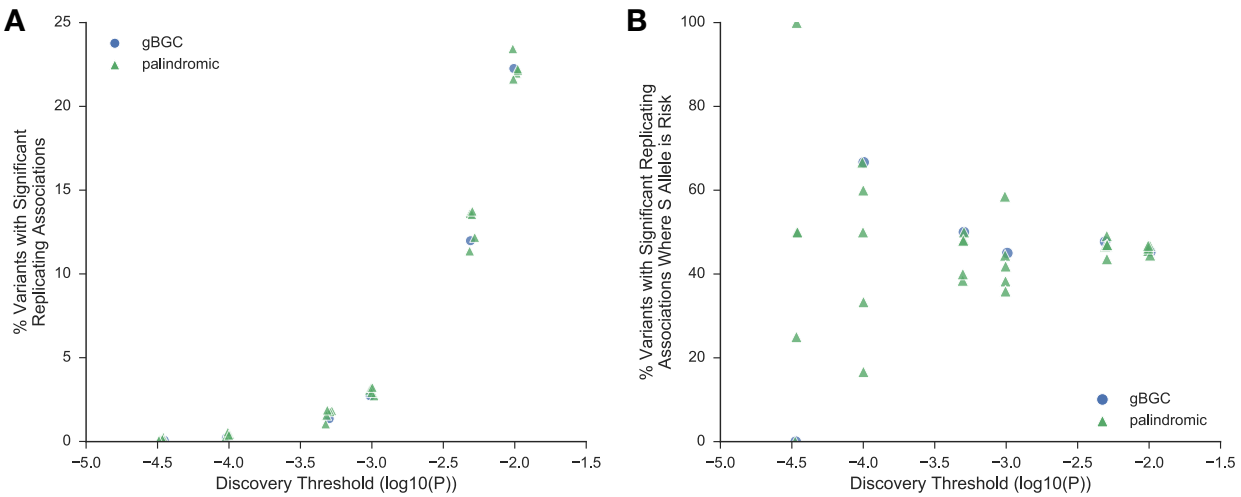


Figure 3-D. eMERGE gBGC variants are not more likely to associate with a clinical phenotype than palindromic variants matched by allele frequency, local recombination rate, and number of LD partners. (a) The percent of variants that have at least one significant, replicating association at the given discovery p-value threshold. (b) The percent of variants that have at least one significant, replicating association at the given discovery p-value threshold where the S allele (or equivalent) is the risk allele. Each set of palindromic variants is plotted (green triangles).

To investigate whether we see no difference between gBGC and palindromic variants due to an excess of noise in our data, we permuted the individual-genotype relationships and reran the meta-PheWAS. Indeed, we find that the permuted variants have roughly the same proportion of significant replicating results across discovery thresholds as the original gBGC and palindromic variants, as well as similar proportions of variants where the S allele is the risk allele when the proportion is stable enough to estimate (Figure 3-E).

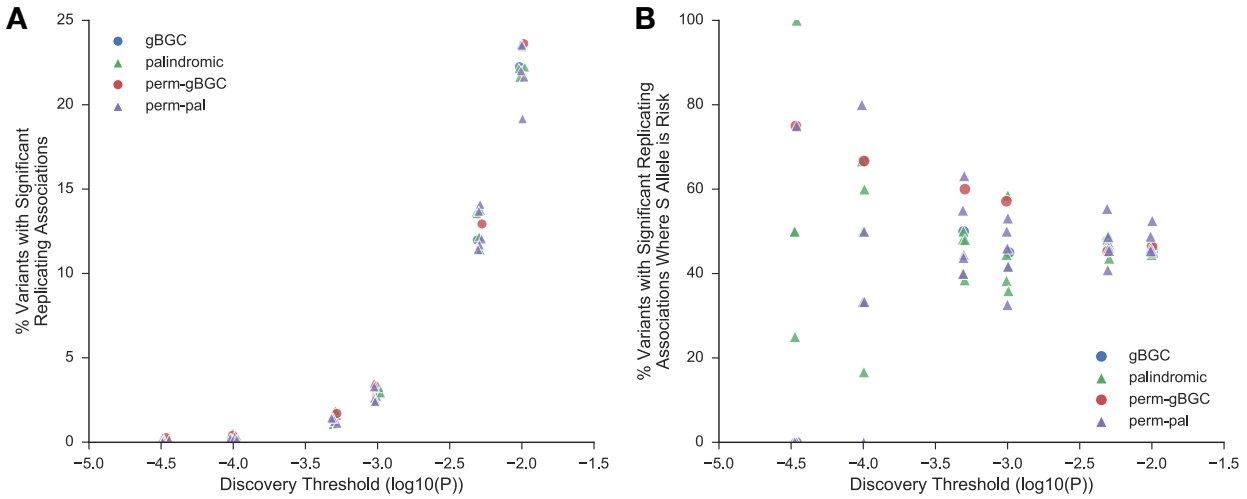


Figure 3-E. eMERGE gBGC and matched palindromic variants are not more likely to associate with a clinical phenotype than permuted variants. (a) The percent of variants that have at least one significant, replicating association at the given discovery p-value threshold. (b) The percent of variants that have at least one significant, replicating association at the given discovery p-value threshold where the S allele (or equivalent) is the risk allele. Each set of palindromic variants (green triangles) and permuted palindromic variants (purple triangles) are plotted.

gBGC Tracts are Enriched for Nominal GWAS Catalog Variants

Previous studies have found enrichment for Human Genome Mutation Database (HGMD) and dbSNP "pathogenic" variants, among others[50], in gBGC tracts, but have not examined NHGRI GWAS catalog[66] variant enrichment. As these variants are the most similar to the associations we find in PheWAS, we looked for enrichment for these variants in regions of

the genome where there has been strong, extended gBGC activity since divergence with chimpanzee (8,321 autosomal regions, covering 7,389,204 bp), as identified by phastBias[50]. As we were primarily interested in situations where gBGC has influenced disease risk, we restricted to non-palindromic (WS) variants, resulting in 17,794 variants. After intersecting these GWAS catalog variants with the gBGC tracts, 93 variants overlapped 91 gBGC tracts. When compared with the standard human genome size (3.2 Gb), this is significantly more than expected by chance (chi square contingency test, $X^2 = 63.99$, $P = 1.25E-15$; Table 3-A).

The contingency table assumes that all 3.2 Gbp of the human genome are viable for both the gBGC tracts and the GWAS variants, which is known to be untrue. To calculate enrichment, we generated two different groups of comparison sets. First, we generated 1,000 sets of regions that were matched to the gBGC tracts by chromosome and length (see Methods). These regions overlapped 51.6% (median: 48 vs 93; randomization $P < 0.001$) of the GWAS catalog variants compared to the gBGC tracts (Figure 3-F (a); Table 3-C). As GC content is by necessity altered by gBGC and high GC content regions tend to be found closer to genes, we generated an additional 100 sets of regions that were additionally matched on GC content. While these regions overlapped more GWAS catalog variants on average than the regions that were not GC-matched, they failed to exceed or even match the number of variants overlapped by the gBGC tracts (median: 55 (59.1%); randomization $P < 0.01$; Table 3-C; Figure 3-F (a)).

Table 3-C. GWAS catalog randomization enrichment results.

		Chromosome and Length Matched Regions		**GC Matched Regions**	
Variant Type	**gBGC Overlap**	Median Overlap	*P*	Median Overlap	*P*
WS (*P* < 1E-05)	93	48	<0.001	55	<0.01

Significant WS (*P* < 1E-07)	30	17	0.007	22	0.12
Palindromic (*P* < 1E-05)	9	5	0.072	6	0.12
Significant Palindromic (*P* < 1E-07)	4	2	0.114	2	0.20

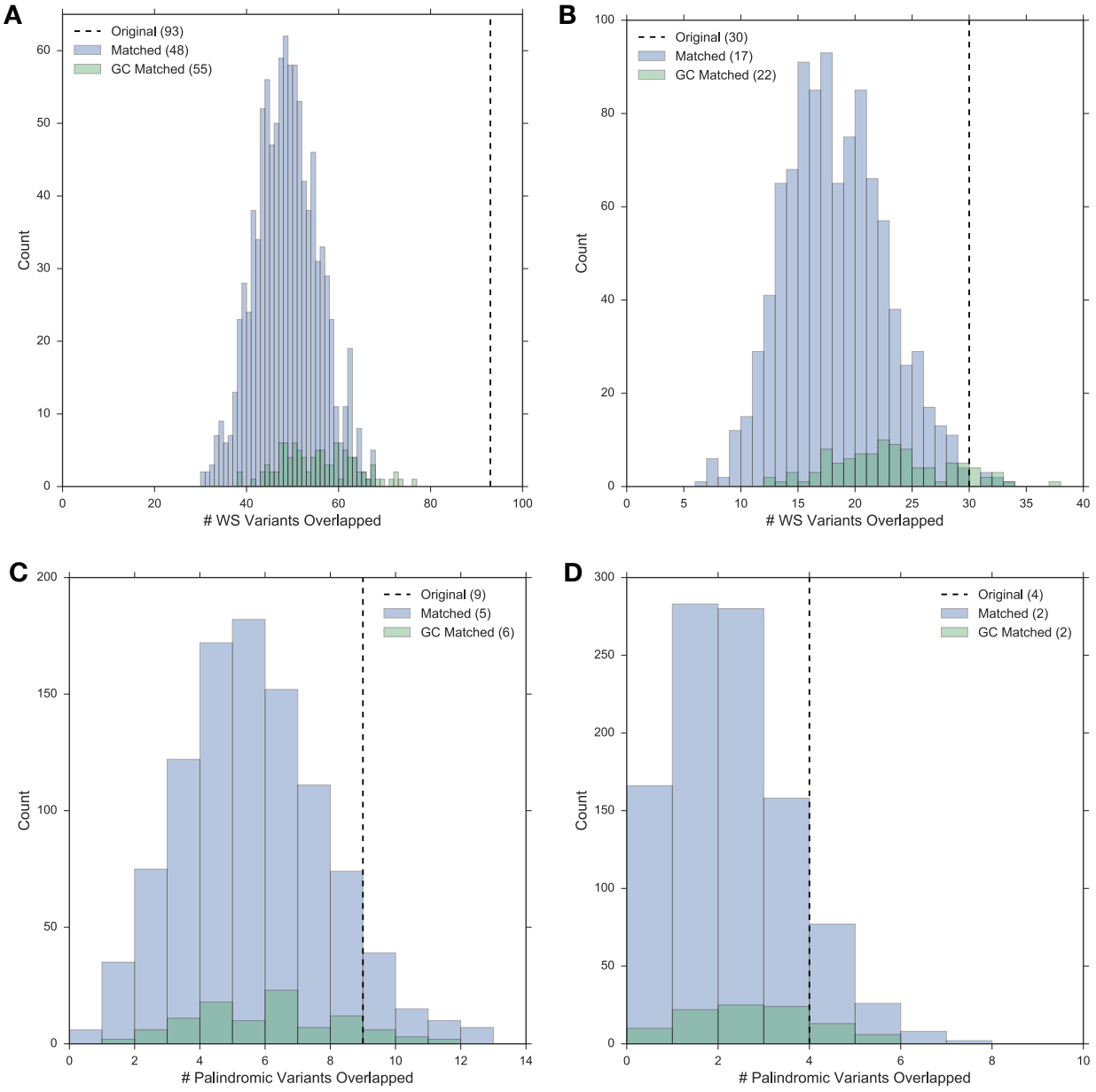


Figure 3-F. gBGC tracts are enriched for nominal, but not genome-wide significant GWAS catalog WS variants. The overlap of each set of comparison regions is plotted (only chromosome and length-matched: blue; chromosome, length, and GC-matched: green) for all GWAS catalog

WS variants (a), significant (P < 1E-07) WS variants (b), all palindromic variants (c), and significant (P < 1E-07) palindromic variants (d). The black line denotes the original number of variants overlapped by gBGC tracts. The median overlap for each group of comparison sets is found in the legend.

However, when we restrict the GWAS catalog variants to those that are genome-wide significant (P < 1E-07), we see a reduction in the enrichment. In the contingency chi square test, we still find the enrichment to be significant, but to a lesser degree (X^2 = 12.16, P = 4.88E-4; Table 3-A). The chromosome and length-matched regions overlapped 57% of the variants (median: 17 vs 30; randomization P = 0.007; Table 3-B) compared to the gBGC tracts. Matching on GC content had a larger effect for these variants, as the GC-matched regions overlapped 73% of variants (median: 22 vs 30; randomization P = 0.12; Figure 3-F (b); Table 3-B).

We considered the palindromic GWAS catalog variants as a natural control for the WS variants as they are susceptible to the reduction in LD as well as any gene proximity biases caused by falling in a recombination hotspot, but not the biased allele frequency altering of gBGC itself. While we see a reduction in enrichment between the nominal and genome-wide significant palindromic variants as we did for the WS variants, the enrichment for nominal palindromic variants is only trending towards significance (Table 3-A; Table 3-B; Figure 3-F (c,d)). This could be due to a reduction in power as there are only ~1,900 palindromic variants versus the ~17,800 WS variants.

GWAS Catalog Variants with Significant Associations Have More LD Partners

As gBGC tracts were enriched for GWAS catalog variants with nominal associations, but not significant, we investigated differences between these sets of variants. We calculated the number of LD partners (r^2 > 0.5) for each variant using 1KG EUR individuals. When we

compare number of LD partners, we find that variants with significant associations have more LD partners than nominal variants (P = 2.35E-133, Mann-Whitney U test). This is consistent regardless of variant type or whether the variant falls within a gBGC tract (Figure 3-G).

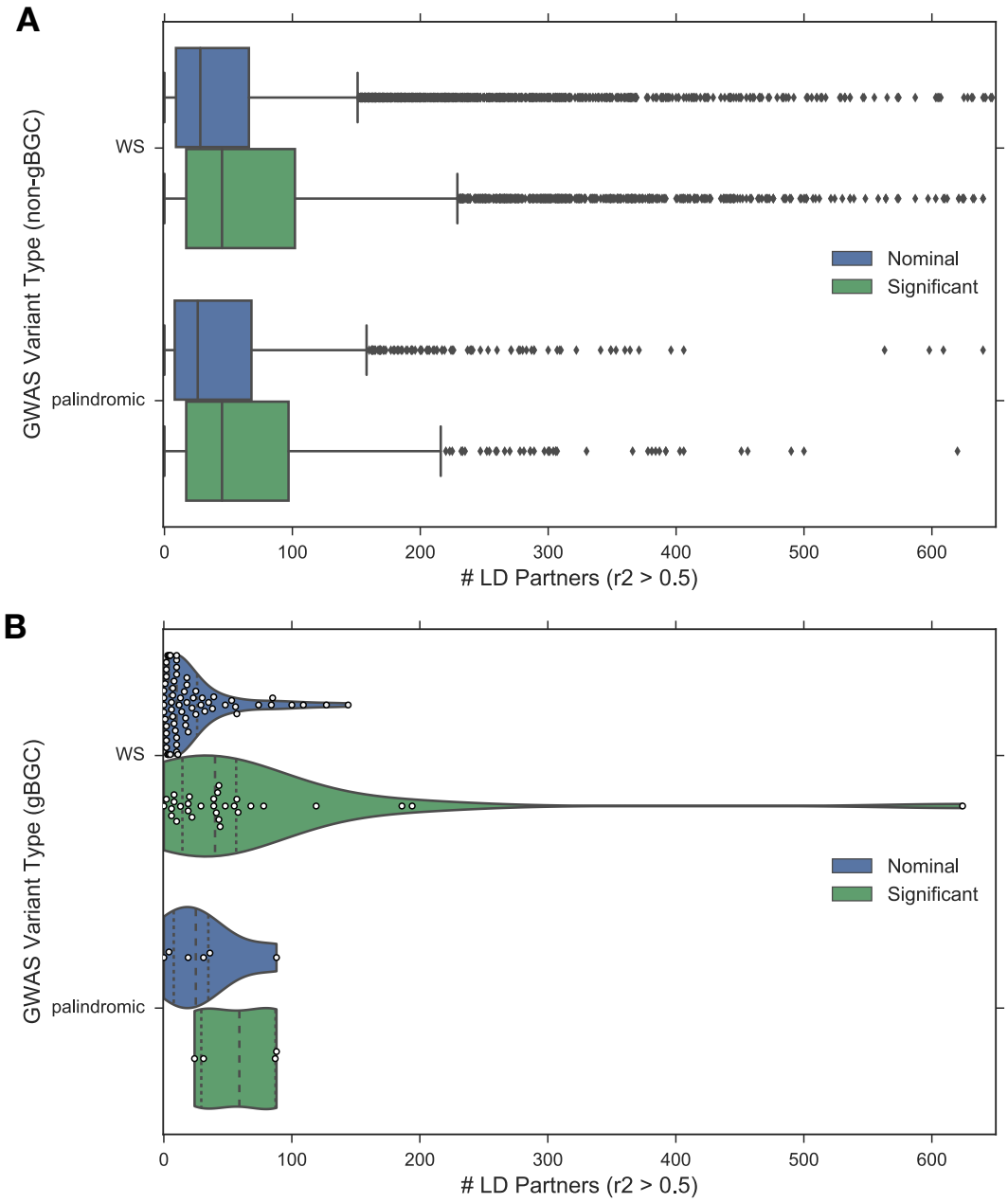


Figure 3-G. GWAS catalog variants with significant associations have more LD partners than variants with nominal associations. The distribution of number of LD partners for GWAS catalog variants that have nominal (blue) and significant (green; P < 1E-07) associations that fall outside

of gBGC tracts (a) and within (b). For comparability with the variants in (b), only variants with less than 650 LD partners are shown in (a). The full distribution of values is plotted in Appendix H. As gBGC tracts overlap few GWAS catalog variants, each variant is plotted in (b) (white dots).

We then intersected the GWAS catalog variants with the recombination rate map from Kong *et al*[172] and compared the overlapping 10 kb recombination rates. As expected from the LD results, nominal GWAS catalog variants have significantly higher recombination rates than significant variants, though this difference is smaller in magnitude (P = 8.72E-19, Mann-Whitney U test; Appendix I). This could be due to the fact that variants with high recombination rates are strongly correlated with few LD partners, but variants with low recombination rates could have a large range of LD partners (Figure 3-A). If so, this suggests that number of LD partners is more correlated with the cause of this phenomenon than recombination rate.

Discussion

In this study, we find very few genome-wide significant associations with any of our tested variants, including those affected by gBGC. Furthermore, we find that variants influenced by gBGC are not more likely to associate with a clinical phenotype than palindromic or long-time coldspot WS variants once allele frequency, recombination rate, and number of LD partners are accounted for. For the BioVU hotspot-coldspot comparison, hotspot variants may even be less likely to associate with a phenotype, depending on the discovery significance threshold considered. When we permuted the eMERGE gBGC and palindromic variants, we found that neither of these groups show any signal for phenotypic associations beyond random at any nominal discovery threshold (Figure 3-E). This suggests that gBGC does not increase the

likelihood of a variant associating with a clinical phenotype to a degree that we can detect in these cohorts.

Additionally, we find that gBGC tracts are enriched for nominal GWAS catalog WS variants, but not genome-wide significant WS variants or palindromic variants. Part of this could be due to power, as gBGC tracts overlap roughly twice as many variants as the median expectation for the chromosome and length-matched regions for all comparisons (Figure 3-A), and the nominal WS variants are by far the largest set examined (Table 3-A). For the WS variant enrichment analyses, it appears that GC content matching increases the number of GWAS variants overlapped. As GC content is associated with gene content and other genomic features, this result is not wholly unexpected. However, GC content matching does not appear to increase the palindromic variant overlap to the same degree, though this could again be due to power as there are so few palindromic variants in the GWAS catalog (Table 3-A). It is worth noting that taking LD into account in these analyses may alter our results, but this is unlikely to have a large effect considering the reduction in LD between variants within gBGC tracts and those without.

One of the unexpected findings from the GWAS catalog analyses was the association of number of LD partners and significance. The *P* values assigned to associations detected through logistic and linear regression are contingent on a host of factors influencing statistical power, including: sample size, allele frequency, and effect size. Given that a variant is directly responsible for a phenotypic change, that variant's number of LD partners should not affect its biological or statistical association with a phenotype. However, one of the foundational principles of GWAS and genotyping platform design is to leverage LD so that all variants do not have to be assayed individually, reducing multiple testing correction. Assuming that 1) LD structure itself does not confound a variant's ability to associate with a phenotype, and 2) all

variants have some likelihood of associating with a given phenotype, every additional variant in LD with the tested variant increases its likelihood of association by proxy. This would leave all variants in recombination hotspots at a disadvantage compared to those outside of recombination hotspots, but particularly WS variants as gBGC would reduce LD beyond the effects of recombination alone. As many of these variants will also be shifted to higher allele frequencies by gBGC and genic regions are biased towards high GC content, they might be simultaneously more likely to be tested in GWAS and to be nearby functional variants, but at a disadvantage to tag those variants. A balance of these factors could enrich gBGC tracts for nominal associations.

Naturally, these assumptions do not reflect biological reality as historical demography and selection alter LD patterns; and demography and selection themselves are influenced by whether a variant affects fitness or fecundity. Considering these connections, it is worth noting that most of the phenotypes in the GWAS catalog and the wider array of clinical phenotypes assayed in PheWAS are unlikely to impact fitness or fecundity, and therefore variants increasing their risk would not be under negative selection. Thus, gBGC may be acting as little more than directional genetic drift at many of these loci. Lachance and Tishkoff theorized that gBGC could have a particular effect on health by increasing homozygosity of derived variants that increase disease risk in a recessive manner[49]. The models used to identify the associations in the GWAS catalog as well as here used additive rather than recessive models, which could result in some loss of power to detect associations[176]. While we find no support for the hypothesis that gBGC increases disease risk in our data, examining the transmission of Mendelian disease-causing alleles potentially affected by gBGC using trio data, or a cohort containing sufficient cases for diseases that decrease fitness or fecundity could prove more effective.

It is important to note that while we find no support for gBGC having a large effect on the phenotypes studied here, gBGC has had a demonstrable impact on the genome. Between 5-10% of identified HARs may have been created by gBGC. Even genes have been affected by gBGC, as the adenylate cyclase activating polypeptide 1 (*ADCYAP1*) gene, which is involved in neuroendocrine stress response, has accrued 17 WS non-synonymous coding mutations since AMH divergence from chimpanzee[10,48,50]. This process may have had an additional effect on immune response, as one of the GWAS catalog WS variants that overlaps a gBGC tract (rs9271366) falls near *HLA-DRB1*, increases risk for multiple sclerosis[177,178] and ulcerative colitis[179,180], increases immunoglobulin A levels[181], and decreases risk for Crohn's disease[180,182]. There are two gBGC tracts within 500 kbp of the one encompassing this variant, both of which fall near other HLA genes. The importance of recombination hotspots in shaping the LD structure of the HLA region has been noted in the past[183], though this was not connected to disease or gBGC. Most immunological genes, but particularly HLA are some of the strongest examples of positive and balancing selection[184], so gBGC near these genes could be useful in increasing immunological diversity. Our analyses show that understanding the full scope of ramifications of gBGC on modern humans will require careful phenotype and model selection in the future.

CONCLUSION

In my dissertation, I used EHRs and PheWAS to explore the phenotypic effects of variants with three different evolutionary origins or histories. In particular, I examined: variants introduced through introgression from Neanderthals, variants whose ancestral allele frequency has declined through selection or drift since divergence with chimpanzee, and variants affected by the mutational process gBGC. Each of these projects represents the first systematic analysis of the effects of these variants across many phenotypes in a modern human population. The results of these projects have supported previously proposed hypotheses about the consequences of these events or mutational processes for modern human health in some instances, as well as failed to do so in others.

My examination of variants introgressed from Neanderthal into individuals of European ancestry supported previous hypotheses about what biological systems were influenced by introgression. As Neanderthals were presumably better adapted to life outside of Africa, systems theorized to be positively impacted by introgression included dermatological, immunological, and others. However, as Neanderthals had smaller effective population sizes throughout their history than AMHs, many deleterious variants could have accrued in their genomes and been passed to AMHs through introgression. Between this and the long divergence time between Neanderthals and AMHs, many systems could have been negatively affected by introgression, the most important of which being reproduction. My work also revealed unexpected enrichments for introgressed variants altering gene expression in the brain, as well as associating with neurological and psychiatric phenotypes. It remains unclear whether these associations suggest true differences in cognition or brain function between AMHs and Neanderthals, or whether they

represent previously neutral variation that has become deleterious in modern environments. While we can only speculate on the non-skeletal morphological differences between AMHs and Neanderthals, this work suggests that Neanderthals fell within the range of human variation in many respects.

Further inquiries into the effects of Neanderthal introgression on clinical phenotypes in other modern non-African populations and could help give the full picture of how this interbreeding event has shaped modern human health. Examining other non-African populations, such as East Asians, could also give an idea of how selection acting after introgression may have shaped what biological systems were affected. Additionally, analysis of the impact of Neanderthal introgression on non-clinical phenotypes may both inform the genetic basis of phenotypic variation in modern humans, as well as lend insight into Neanderthal soft tissue biology. Melanesians may be uniquely instructive in the effects of archaic introgression on AMHs today as they experienced both Neanderthal and Denisovan introgression during their history[105]. Beyond further statistical genetics work in diverse AMH populations, more samples of archaic hominin genomes are necessary to get the full picture of Neanderthal and Denisovan variation. Given the multiple interbreeding events between archaic hominin groups, researchers may also gain knowledge about the genomes of more anciently diverged archaic hominins that contributed to Neanderthals and Denisovans.

My study of Neanderthal introgression into humans also provides an interesting window into the hybridization process in hominins. Hybridization has been studied extensively in plants and animals with respect to both adaptive introgression as well as conservation biology (reviewed in Hamilton & Miller[185]). While hybridization is generally considered favorable in plants due to the prevalence of hybrid vigor in the offspring of different plant species[186,187], it is

often viewed negatively when used for rescue of animal populations experiencing strong inbreeding depression, though this may be anchored in historic viewpoints[185]. Indeed, adaptive introgression between species has been found to increase resistance to warfarin in mice[188], resistance to insecticides in an African mosquito[189], and diversification of MHC alleles in the alpine ibex[190]. Hybridization between humans and Neanderthals parallels a classic beneficial hybridization scenario, with one group experiencing a strong bottleneck (humans) and receiving genetic variation presumably adapted to the local environment from the introgressing group (Neanderthals). However, the historically small effective population size of Neanderthals raises questions about how beneficial this event may actually have been[101]. All told, my study as well as future analyses of the effects of Neanderthal introgression will speak not just to human evolutionary genetics, but also to issues faced in conservation biology and the ramifications of hybridization.

The human-specific and hominin-derived variant analysis revealed associations with phenotypes expected to be evolving continuously across all organisms, such as drug metabolism and immune phenotypes. Considering the many skeletal changes over hominin history and known anatomical differences between hominin groups and AMHs, the associations with bone fractures are not unexpected. Given the comparable gracility of AMH skeletons, these fracture associations could indicate that AMHs required genetic changes to protect against fracture that more robust archaic hominins did not. However, it is also possible that the human allele would have served a similarly protective role in Neanderthals and Denisovans, but simply arose after divergence. Indeed, it is difficult to determine what degree of selective pressure may have acted on variants affecting skeletal health, given the complexity of this system. The neurological and psychiatric associations with ancestral alleles reintroduced by Neanderthal further supported our

previously found enrichments for these phenotypes in the Neanderthal introgression analyses. However, it also stressed the importance of understanding the history of evolutionary variation chosen for further analysis through PheWAS or GCTA, as different histories may propose alternate hypotheses for affected biological systems.

Sequencing of additional Neanderthals and Denisovans would be useful in validating whether the variants classified as human-specific in this analysis are truly human-specific. As discussed in Chapter II, we currently lack the population samples to confidently discern whether the derived human allele was completely absent from archaic hominins. Future analyses of human-specific and hominin-derived variation would also benefit from determining when allele frequencies changed—perhaps through additional sequencing of ancient human samples—and whether these changes correlate with signatures of selection or admixture events. This would help in the delineation of variants whose alleles have changed frequency due to selection rather than genetic drift or demographic history. Analyzing variants with homogenous histories may ease interpretation of the biological systems affected and shaped by various evolutionary forces. It may also suggest phenotypes for testing through more low-throughput strategies, such as GCTA. While preliminary work in this area has been done[191], functional analysis of additional human-specific variants that are fixed in AMHs is critical to understanding changes in humans over this period. Another area of interest would be to extend the variants examined to those that are polymorphic in both chimpanzee and humans, as these could be indicative of ancient balancing selection. While such signals have been identified in previous studies, typically near immune loci such as MHC[192], no study has examined the phenotypic effects of these variants in human populations.

My analyses of variants affected by gBGC had some unexpected results. While many examples of gBGC—a mutational process that has been called the "Achilles' heel of our genome"[47]—having dramatic effects on the genome are known, I found no indication that this process led to an increased risk for common AMH morbidities compared to variants matched on variant characteristics known to be affected by gBGC, including: frequency of the S allele, recombination rate, and number of LD partners. While the noise in the data makes it impossible to claim that there is no difference between variants affected by gBGC and those that are not, it indicates that any true signal is not exceptionally strong. As discussed previously, this may not be unexpected given the phenotypes assayed, as gBGC is theorized to have the potential to counteract weak negative selection on deleterious S alleles at WS variants. Thus, if there is no negative selection to counteract, WS variants undergoing gBGC should not be expected to affect disease risk more than any other variant experiencing genetic drift. However, my analysis of variants in the GWAS catalog did demonstrate that the likelihood of detecting a significant statistical association is based on parameters affected by gBGC, such as recombination rate and number of LD partners. This suggests that any future studies of the effects of mutational processes on modern human health should consider both biological and statistical confounders of those processes when formulating their analysis plans, though this recommendation is by no means restricted to studies of mutational processes[193].

There are several important technical considerations when using PheWAS and other statistical genetics tools in clinical populations to answer evolutionary questions. In several instances, I have used enrichment of phenotype associations to test if a group of variants inordinately affects certain biological systems or simply more phenotypes than we would anticipate by chance. Choosing appropriate matched variants is critical to these analyses, as

association tests for different phecodes in a clinical population will have heterogeneous levels of statistical power. It is important to match on allele frequencies and other parameters as appropriate given the study question, and to be cognizant of the potential for overmatching. Though it does not allow for the comparison of two groups of variants, permutation analyses may prove to be the most useful in some scenarios, as they allow for a comprehensive idea of the noise level in the data in light of the complex correlation structure of phecodes.

Another technical consideration is the accuracy of the phecodes themselves. Depending on the biobank used for such analyses—thus, the practice standards for the hospital to which the biobank is attached—the ICD-9 translation to phecodes may or may not be as useful in capturing certain phenotypes. For analyses focusing on particular clinical phenotypes rather than a wide survey such as was performed here, development of an algorithm to determine case or control status may be more appropriate. Many groups are currently working on improving phenotype extraction from the EHR[194,195], and the strengthening of relationships between physicians and researchers moving forward will no doubt improve this process.

Interpretation of results can also pose challenges when using data such as these to answer evolutionary questions. For example, one important consideration is the history of the phenotype being analyzed. Many clinical phenotypes extractable from EHRs are unlikely to have been present in ancient humans, at least not as we understand them. For example, we found enrichment for associations between neurological and psychiatric phenotypes and introgressed Neanderthal variants. However, it is unlikely that AMHs or even Neanderthals suffered from any of these disorders 50,000 years ago, and certainly not in their current forms. One of the more striking examples of this phenomenon is the association of an introgressed variant with tobacco use disorder, when tobacco is absent from Eurasia and could not have had this effect before a

few hundred years ago in Europeans. Therefore, the only conclusions we can draw confidently are that many of these introgressed variants alter gene expression in the brain, and associate with neurological and psychiatric phenotypes in modern humans. These results suggest that Neanderthal neuronal function may have had some slight incompatibilities with that of AMHs in modern environments; however, speculation on the organism-level phenotypic effects of these variants in previous generations will often have to remain just that.

Psychiatric and neurologic disorders may be the most extreme examples of this, but careful consideration of the disrupted biological processes that lead to clinical phenotypes is recommended for the interpretation of all associations found in such an analysis. To give another example from the Neanderthal introgression analysis, a variant that increases pathogenic blood coagulation would seem unlikely to be maintained after introgression, given the negative implications for fertility through increased risk of miscarriage[196]. However, as coagulation factors are involved in defense from bacterial pathogens[132] and the optimal level of clotting itself is critical for balancing risk for stroke versus hemophilia, variability in this system could be indicative of a possibly beneficial trade-off in AMH or archaic hominin history. This association is also an important example of the occasional disconnect between the obvious phenotypic grouping (hematopoietic) versus the total number of affected systems (immune, hematopoietic, reproductive), as well as representative of the power of examining the full range of phecodes, rather than a limited subset.

In conclusion, though there are currently some limitations, clinical biobanks provide a promising resource for future evolutionary studies. Analysis of many densely genotyped and phenotyped individuals with statistical genetics tools such PheWAS and GCTA can be used to test or refine hypotheses generated by evolutionary models applied solely to sequence data. This

may be particularly useful when scans for various evolutionary signatures produce thousands of regions of interest containing tens of thousands of candidate variants with potentially complex LD patterns, and exhaustive functional testing of the variants involved is not feasible.

REFERENCES

1. King, M. & Wilson, A. Evolution at two levels in humans and chimpanzees. *Science (80-.).* **188,** (1975).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
3. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).
4. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–45 (2004).
5. Chen, F.-C. & Li, W.-H. Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. *Am. J. Hum. Genet.* **68,** 444–456 (2001).
6. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418,** 145–151 (2002).
7. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325,** 31–36 (1987).
8. and Analysis Consortium, T. C. S. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
9. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437,** 88–93 (2005).
10. Hubisz, M. J. & Pollard, K. S. *Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution*. *Current Opinion in Genetics and Development* **29,** 15–21 (2014).
11. Prabhakar, S., Noonan, J. P., Pääbo, S. & Rubin, E. M. Accelerated Evolution of Conserved Noncoding Sequences in Humans. *Science (80-.).* **314,** (2006).
12. Pollard, K. S. *et al.* Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genet.* **2,** e168 (2006).
13. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
14. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. London B Biol. Sci.* **368,** (2013).
15. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–22 (2010).
16. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Sci. (New York, NY)* **222,** 1–14 (2012).
17. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–9 (2014).
18. Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531,** 504–507 (2016).
19. International, T. & Consortium, H. The International HapMap Project. *Nature* **426,** 789–796 (2003).

20. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).
21. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
22. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39,** 31–40 (2007).
23. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
24. Hamblin, M. T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66,** 1669–79 (2000).
25. Kasehagen, L. J. *et al.* Reduced plasmodium vivax Erythrocyte infection in PNG duffy-negative heterozygotes. *PLoS One* **2,** e336 (2007).
26. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The Resistance Factor to Plasmodium vivax in Blacks. *N. Engl. J. Med.* **295,** 302–304 (1976).
27. Allison, A. C. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.* **1,** 290–294 (1954).
28. Hanchard, N. *et al.* Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC Genet.* **8,** 52 (2007).
29. Kwiatkowski, D. P. P. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *Am J Hum Genet* **77,** 171–190 (2005).
30. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
31. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512,** 194–197 (2014).
32. Scheinfeldt, L. B. *et al.* Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.* **13,** R1 (2012).
33. Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science (80-.).* **329,** 72–75 (2010).
34. Ge, R. L. *et al.* Metabolic insight into mechanisms of high-altitude adaptation in Tibetans. *Mol. Genet. Metab.* **106,** 244–247 (2012).
35. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet* **47,** 97–120 (2013).
36. Pagani, L. *et al.* Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96,** 986–991 (2015).
37. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet.* **10,** (2014).
38. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **advance on,** 291–295 (2015).
39. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* 1–18 (2016). doi:10.1126/science.aag0776
40. Peyrégne, S. *et al.* Detecting ancient positive selection in humans using extended lineage sorting. *bioRxiv* 1–35 (2016).

41. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505,** 87–91 (2013).
42. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–9 (2014).
43. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506,** 225–229 (2014).
44. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
45. Seguin-Orlando, A. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science (80-.).* (2014).
46. Haber, M. *et al.* Ancient DNA and the rewriting of human history: be sparing with Occam's razor. *Genome Biol.* **17,** 1 (2016).
47. Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* **23,** 273–277 (2007).
48. Ratnakumar, A. *et al.* Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. London B Biol. Sci.* **365,** (2010).
49. Lachance, J. & Tishkoff, S. A. Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles. *Am. J. Hum. Genet.* **95,** 408–420 (2014).
50. Capra, J. A. *et al.* A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genet.* **9,** e1003684 (2013).
51. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 3439–3444 (2015).
52. Simonti, C. N. *et al.* Evolution of lysine acetylation in the RNA polymerase II C-terminal domain. *BMC Evol. Biol.* **15,** 35 (2015).
53. Vernot, Benjamin, Akey, J. Resurrecting surviving Neanderthal lineages from modern human genomes. *Science (80-.).* **20,** 96–103 (2014).
54. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507,** 354–7 (2014).
55. Grossman, S. R. *et al.* A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science (80-.).* **327,** (2010).
56. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).
57. Karczewski, K. J. *et al.* Systematic functional regulatory assessment of disease-associated variants. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 9607–12 (2013).
58. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518,** 365–369 (2015).
59. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2014).
60. Zhou, X. *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat. Biotechnol.* **33,** 345 (2015).
61. Consortium, T. Gte. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348,** 648–660 (2015).

62. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
63. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
64. Karlsson, E. K. *et al.* Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. *Sci. Transl. Med.* **5,** 192ra86 (2013).
65. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am. J. Hum. Genet.* **98,** 22–33 (2016).
66. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** (2014).
67. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39,** 1443–52 (2007).
68. Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **40,** 835–7 (2008).
69. Lin, B. D. *et al.* Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. *Genes (Basel).* **6,** 559–76 (2015).
70. Han, J. *et al.* A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* **4,** e1000074 (2008).
71. Eriksson, N. *et al.* Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6,** e1000993 (2010).
72. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13,** 395–405 (2012).
73. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372,** 793–5 (2015).
74. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84,** 362–9 (2008).
75. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12,** 417–428 (2011).
76. Denny, J. C. *et al.* PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26,** 1205–1210 (2010).
77. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86,** 560–72 (2010).
78. Ritchie, M. D. *et al.* Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* **127,** 1377–1385 (2013).
79. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30,** 2375–2376 (2014).
80. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31,** 1102–10 (2013).
81. Pendergrass, S. A. *et al.* Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* **9,** (2013).

82. Hall, M. A. *et al.* Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) Study. *PLoS Genet.* **10,** (2014).
83. Pendergrass, S. A. & Ritchie, M. D. Phenome-Wide Association Studies: Leveraging Comprehensive Phenotypic and Genotypic Data for Discovery. *Curr. Genet. Med. Rep.* **3,** 92–100 (2015).
84. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Gen* **42,** 565–569 (2010).
85. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
86. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28,** 2540–2542 (2012).
87. Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science (80-.).* **351,** 737–741 (2016).
88. Hrdlicka, A. The Neanderthal Phase of Man. *J. R. Anthropol. Inst. Gt. Britain Irel.* **57,** 249 (1927).
89. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90,** 19–30 (1997).
90. Ovchinnikov, I. V. *et al.* Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404,** 490–493 (2000).
91. Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325,** 318–21 (2009).
92. Serre, D. *et al.* No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. *PLoS Biol.* **2,** e57 (2004).
93. Orlando, L. *et al.* Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Curr. Biol.* **16,** R400–R402 (2006).
94. Krause, J. *et al.* Neanderthals in central Asia and Siberia. *Nature* **449,** 902–904 (2007).
95. Currat, M. *et al.* Modern Humans Did Not Admix with Neanderthals during Their Range Expansion into Europe. *PLoS Biol.* **2,** e421 (2004).
96. Higham, T. *et al.* The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* **512,** 306–309 (2014).
97. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
98. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2,** (2006).
99. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530,** 429–433 (2016).
100. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in east Asians than in Europeans. *Genetics* **194,** 199–209 (2013).
101. Harris, K. & Nielsen, R. The genetic cost of neanderthal introgression. *Genetics* **203,** 881–891 (2016).
102. Juric, I. *et al.* The Strength of Selection against Neanderthal Introgression. *PLOS Genet.*

12, e1006340 (2016).
103. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418,** 869–872 (2002).
104. Webb, D. M. & Zhang, J. FoxP2 in Song-Learning Birds and Vocal-Learning Mammals. *J. Hered.* **96,** 212–216 (2005).
105. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science (80-.).* **352,** 235–239 (2016).
106. Abi-Rached, L. *et al.* The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science (80-.).* **334,** 89–94 (2011).
107. Sams, A. J. *et al.* Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* **17,** 246 (2016).
108. Khrameeva, E. E. *et al.* Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nat. Commun.* **5,** 3584 (2014).
109. Kho, A. N. *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* **3,** 79re1 (2011).
110. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15,** 761–71 (2013).
111. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–525 (2011).
112. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.* **10,** (2014).
113. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** (2009).
114. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
115. Crosslin, D. R. *et al.* Controlling for population structure and genotyping platform bias in the eMERGE multi-institutional biobank linked to electronic health records. *Front. Genet.* **5,** (2014).
116. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* **5,** (2014).
117. Hebbring, S. J. *et al.* A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* **14,** 187–91 (2013).
118. Pearce, E., Stringer, C. & Dunbar, R. I. M. New insights into differences in brain organization between Neanderthals and anatomically modern humans. *Proc. R. Soc. B Biol. Sci.* **280,** 20130168–20130168 (2013).
119. Pearce, E. & Dunbar, R. Latitudinal variation in light levels drives human visual system size. *Biol. Lett.* **8,** 90–93 (2012).
120. Pratt, L. A., Brody, D. J. & Gu, Q. Antidepressant use in persons aged 12 and over: United States, 2005-2008. *NCHS Data Brief* **127,** 1–8 (2011).
121. Flohil, S. C. *et al.* Prevalence of actinic keratosis and its risk factors in the general population: the Rotterdam Study. *J. Invest. Dermatol.* **133,** 1971–8 (2013).
122. Ogden, C. L., Lamb, M. M., Carroll, M. D. & Flegal, K. M. Obesity and socioeconomic status in adults: United States, 2005-2008. *NCHS Data Brief* **127,** 1–8 (2010).

123. Fryar, C. D., Hirsch, R., Eberhardt, M. S., Yoon, S. S. & Wright, J. D. Hypertension, high serum total cholesterol, and diabetes: racial and ethnic prevalence differences in U.S. adults, 1999-2006. *NCHS Data Brief* 1–8 (2010).
124. Kessler, R. C., Chiu, W. T., Demler, O. & Walters, E. E. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62,** 617 (2005).
125. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).
126. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet.* **8,** (2012).
127. Chen, C. *et al.* Two gene co-expression modules differentiate psychotics and controls. *Mol. Psychiatry* **18,** 1308–14 (2013).
128. Martin, P. R., Singleton, C. K. & Hiller-Sturmhöfel, S. The role of thiamine deficiency in alcoholic brain disease. *Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism* **27,** 134–142 (2003).
129. Pazo, J. H. & Belforte, J. E. Basal ganglia and functions of the autonomic nervous system. *Cellular and Molecular Neurobiology* **22,** 645–654 (2002).
130. Pickering, C., Bergenheim, V., Schiöth, H. B. & Ericson, M. Sensitization to nicotine significantly decreases expression of GABA transporter GAT-1 in the medial prefrontal cortex. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **32,** 1521–1526 (2008).
131. Heit, J. A. *et al.* A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *J. Thromb. Haemost.* **10,** 1521–1531 (2012).
132. Rallapalli, P. M., Orengo, C. A., Studer, R. A. & Perkins, S. J. Positive selection during the evolution of the blood coagulation factors in the context of their disease-causing mutations. *Mol. Biol. Evol.* **31,** 3040–3056 (2014).
133. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19,** 149–150 (2003).
134. Gamazon, E. R. *et al.* SCAN: SNP and copy number annotation. *Bioinformatics* **26,** 259–262 (2010).
135. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9440–9445 (2003).
136. Abdul-Majeed, S., Mell, B., Nauli, S. M. & Joe, B. Cryptorchidism and infertility in rats with targeted disruption of the Adamts16 locus. *PLoS One* **9,** (2014).
137. Gokhman, D. *et al.* Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science* **344,** 523–527 (2014).
138. Oksenberg, N., Stevison, L., Wall, J. D. & Ahituv, N. Function and Regulation of AUTS2, a Gene Implicated in Autism and Human Evolution. *PLoS Genet.* **9,** (2013).
139. Golden, R. N. *et al.* The efficacy of light therapy in the treatment of mood disorders: A review and meta-analysis of the evidence. *American Journal of Psychiatry* **162,** 656–662 (2005).
140. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499,** 471–5 (2013).
141. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).

142. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
143. Komori, T. *et al.* Targeted Disruption of Cbfa1 Results in a Complete Lack of Bone Formation owing to Maturational Arrest of Osteoblasts. *Cell* **89,** 755–764 (1997).
144. Schlebusch, C. M. *et al.* Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science (80-.).* **338,** 374–379 (2012).
145. Mundlos, S. *et al.* Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. *Cell* **89,** 773–779 (1997).
146. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
147. Berndt, S. I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45,** 501–512 (2013).
148. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–8 (2010).
149. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40,** 609–15 (2008).
150. He, M. *et al.* Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* **24,** 1791–1800 (2015).
151. Adhikari, K. *et al.* A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nat. Commun.* **7,** 1–11 (2016).
152. Teitelbaum, S. L. Bone Resorption by Osteoclasts. *Science (80-.).* **289,** 1504–1508 (2000).
153. Takemura, K. *et al.* Class A scavenger receptor promotes osteoclast differentiation via the enhanced expression of receptor activator of NF-kappaB (RANK). *Biochem. Biophys. Res. Commun.* **391,** 1675–1680 (2010).
154. Lin, Y. L. *et al.* The effect of class A scavenger receptor deficiency in bone. *J. Biol. Chem.* **282,** 4653–4660 (2007).
155. Arnsten, A. F. T. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat. Rev. Neurosci.* **10,** 410–22 (2009).
156. Bai, G. & Lipton, S. A. Aberrant RNA Splicing in Sporadic Amyotrophic Lateral Sclerosis. *Neuron* **20,** 363–366 (1998).
157. Meng, Q. *et al.* Traumatic Brain Injury Induces Genome-Wide Transcriptomic, Methylomic, and Network Perturbations in Brain and Blood Predicting Neurological Disorders. *EBioMedicine* **16,** 184–194 (2017).
158. Wang, T. *et al.* Loss of Interleukin-21 Receptor Activation in Hypoxic Endothelial Cells Impairs Perfusion Recovery After Hindlimb Ischemia. *Arterioscler. Thromb. Vasc. Biol.* **35,** 1218–1225 (2015).
159. Ortega, S., Ittmann, M., Tsang, S. H., Ehrlich, M. & Basilico, C. Neuronal defects and delayed wound healing in mice lacking fibroblast growth factor 2. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 5672–7 (1998).
160. Molderings, G. J., Brettner, S., Homann, J. & Afrin, L. B. Mast cell activation disease: a concise practical guide for diagnostic workup and therapeutic options. *J. Hematol. Oncol.* **4,** 10 (2011).
161. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion:

mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8,** 762–775 (2007).
162. Hunter, N. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harb. Perspect. Biol.* **7,** a016618 (2015).
163. Lamb, B. C. The properties of meiotic gene conversion important in its effects on evolution. *Heredity (Edinb).* **53 (Pt 1),** 113–138 (1984).
164. Jeffreys, A. J. & Neumann, R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat. Genet.* **31,** 267–271 (2002).
165. Jeffreys, A. J. & Neumann, R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.* **14,** 2277–2287 (2005).
166. Webb, A. J., Berg, I. L. & Jeffreys, A. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 10471–10476 (2008).
167. Duret, L. & Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).
168. Necşulea, A. *et al.* Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum. Mutat.* **32,** 198–206 (2011).
169. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327,** 836–40 (2010).
170. Berg, I. L. *et al.* Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12378–83 (2011).
171. Glémin, S. *et al.* Quantification of GC-biased gene conversion in the human genome. *Genome Res.* **25,** 1215–28 (2015).
172. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).
173. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).
174. Mero, I.-L. *et al.* Oligoclonal Band Status in Scandinavian Multiple Sclerosis Patients Is Associated with Specific Genetic Risk Alleles. *PLoS One* **8,** e58352 (2013).
175. Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat. Genet.* **43,** 1127–1130 (2011).
176. Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31,** 358–362 (2007).
177. Nischwitz, S. *et al.* Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis. *J. Neuroimmunol.* **227,** 162–166 (2010).
178. Bahlo, M. *et al.* Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat. Genet.* **41,** 824–828 (2009).
179. Yang, S.-K. *et al.* Genome-Wide Association Study of Ulcerative Colitis in Koreans Suggests Extensive Overlapping of Genetic Susceptibility With Caucasians. *Inflamm. Bowel Dis.* **19,** 954–966 (2013).
180. Okada, Y. *et al.* HLA-Cw*1202-B*5201-DRB1*1502 Haplotype Increases Risk for Ulcerative Colitis but Reduces Risk for Crohn's Disease. *Gastroenterology* **141,** 864–

871.e5 (2011).
181. Ferreira, R. C. *et al.* Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat. Genet.* **42,** 777–780 (2010).
182. Yang, S.-K. *et al.* Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. *Gut* **63,** 80–87 (2014).
183. Kauppi, L., Sajantila, A. & Jeffreys, A. J. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* **12,** 33–40 (2003).
184. Hughes, A. L. & Yeager, M. Natural selection and the evolutionary history of major histocompatibility complex loci. *Front. Biosci.* **3,** d509-16 (1998).
185. Hamilton, J. A. & Miller, J. M. Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conserv. Biol.* **30,** 33–41 (2016).
186. Welch, M. E. & Rieseberg, L. H. Habitat divergence between a homoploid hybrid sunflower species, Helianthus paradoxus (Asteraceae), and its progenitors. *Am. J. Bot.* **89,** 472–478 (2002).
187. Hamilton, J. A., Lexer, C. & Aitken, S. N. Genomic and phenotypic architecture of a spruce hybrid zone (Picea sitchensis × P. glauca). in *Molecular Ecology* **22,** 827–841 (2013).
188. Song, Y. *et al.* Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Curr. Biol.* **21,** 1296–1301 (2011).
189. Norris, L. C. *et al.* Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 815–20 (2015).
190. Grossen, C., Keller, L., Biebach, I., Croll, D. & Ferreira, M. Introgression from Domestic Goat Generated Variation at the Major Histocompatibility Complex of Alpine Ibex. *PLoS Genet.* **10,** e1004438 (2014).
191. Weyer, S. & Pääbo, S. Functional Analyses of Transcription Factor Binding Sites that Differ between Present-Day and Archaic Humans. *Mol. Biol. Evol.* **33,** 316–322 (2016).
192. Leffler, E. M. *et al.* Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science (80-.).* **339,** 1578–1582 (2013).
193. Fish, A. E., Capra, J. A. & Bush, W. S. Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? *Am. J. Hum. Genet.* **99,** 817–830 (2016).
194. Wei, W.-Q. & Denny, J. C. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* **7,** 41 (2015).
195. Kirby, J. C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Informatics Assoc.* **23,** 1046–1052 (2016).
196. Bick, R. L. & Hoppensteadt, D. Recurrent Miscarriage Syndrome and Infertility Due to Blood Coagulation Protein/Platelet Defects: A Review and Update. *Clin. Appl. Thromb.* **11,** 1–13 (2005).

APPENDIX

Appendix A. Neanderthal SNP-phenotype associations used in the comparison with non-Neanderthal SNP-phenotype associations. Redundant SNP-phenotype associations due to one SNP associating with multiple phecodes in the same hierarchy are highlighted in gray.

Phenotype	SNP	Category	Discovery		Replication	
			Odds Ratio	*P*	Odds Ratio	*P*
Chronic airway obstruction	rs2300659	Pulmonary	1.24	9.92E-04	1.18	2.09E-02
Chronic pain syndrome	rs2298146	Neurologic	4.02	2.80E-04	2.74	1.88E-02
Hemangioma and lymphangioma, any site	rs17114127	Neoplastic	2.61	4.47E-05	3.41	4.28E-04
Functional disorders of bladder	rs17115796	Genitourinary	2.39	5.38E-05	1.85	2.20E-02
Calculus of ureter	rs12563768	Genitourinary	2.15	8.62E-04	2.08	6.59E-03
Coagulation defects	rs3917862	Hematopoietic	1.27	8.34E-04	1.24	1.05E-03
Hypercoagulable state	rs3917862	Hematopoietic	3.32	9.87E-07	3.00	5.00E-10
Skin neoplasm of uncertain behavior	rs16848353	Neoplastic	1.50	4.73E-04	1.31	8.52E-03
Other disorders of soft tissues	rs17675600	Musculoskeletal	2.99	3.38E-04	1.71	2.13E-02
Protein-calorie malnutrition	rs12049593	Endocrine & Metabolic	1.77	1.98E-06	1.63	5.46E-05
Hepatic cancer	rs17018123	Neoplastic	4.32	5.31E-04	2.73	3.68E-03
Sleep related movement disorders	rs3771635	Neurologic	0.70	9.78E-05	0.75	3.92E-03
Personality disorders	rs2288187	Psychiatric	2.21	7.94E-04	2.62	7.44E-03
Radiotherapy	rs901033	Neoplastic	2.55	5.67E-04	3.53	9.60E-03
Tobacco use disorder	rs901033	Psychiatric	2.19	1.73E-05	1.75	7.93E-04
First degree AV block	rs901033	Cardiovascular	3.21	2.58E-04	1.92	4.74E-02
Functional disorders of bladder	rs13087234	Genitourinary	1.61	7.49E-04	1.34	3.88E-02
Infections involving bone	rs17029555	Musculoskeletal	1.62	8.49E-04	1.49	4.24E-03
Rheumatoid arthritis & related inflammatory polyarthropathies	rs12639456	Musculoskeletal	1.68	5.00E-04	1.54	1.07E-02
Rheumatoid arthritis	rs12639456	Musculoskeletal	1.76	6.58E-04	1.65	5.83E-03
Acquired foot deformities	rs1242069	Musculoskeletal	1.42	5.05E-04	1.39	1.48E-02
Gram negative septicemia	rs2050807	Infectious	4.28	2.45E-04	2.48	3.82E-02
Atherosclerosis of aorta	rs13151936	Cardiovascular	1.45	9.24E-04	1.42	4.00E-02
Leukemia	rs17527711	Neoplastic	1.78	6.90E-04	1.37	3.45E-02
Age-related osteoporosis	rs10516526	Musculoskeletal	3.56	1.08E-04	1.63	9.44E-03
Acquired toe deformities	rs10517630	Musculoskeletal	1.52	9.97E-04	1.56	1.33E-02
Psoriasis & related disorders	rs12190231	Dermatologic	1.35	9.58E-04	1.29	1.18E-02
Psoriasis	rs12190231	Dermatologic	1.40	4.56E-04	1.32	8.39E-03

Psoriasis vulgaris	rs12190231	Dermatologic	1.48	9.14E-05	1.31	1.40E-02
Dry eyes	rs12189640	Neurologic	1.37	5.62E-04	1.31	3.16E-02
Disorders of other cranial nerves	rs12662332	Neurologic	1.71	6.59E-04	2.02	1.50E-04
Subjective visual disturbances	rs1513498	Neurologic	1.53	1.37E-04	1.67	1.40E-02
Other cerebral degenerations	rs3822947	Neurologic	2.00	3.24E-04	1.47	4.87E-02
Microscopic hematuria	rs35609966	Genitourinary	2.51	6.49E-05	2.81	3.62E-02
Cancer, suspected or other	rs9366117	Neoplastic	2.38	5.15E-04	1.88	2.94E-02
Proteinuria	rs17722435	Endocrine & Metabolic	2.83	2.02E-04	1.67	4.10E-02
Other conditions of the mother complicating pregnancy	rs16868879	Genitourinary	5.89	2.34E-04	2.25	9.65E-04
Memory loss	rs16914252	Psychiatric	1.77	6.23E-04	1.70	1.52E-02
Antisocial/borderline personality disorder	rs11139709	Psychiatric	4.92	5.28E-04	4.65	2.44E-03
Chronic kidney disease, Stage IV (severe)	rs1542479	Genitourinary	1.98	7.19E-04	1.63	2.04E-02
Polyp of female genital organs	rs17742994	Genitourinary	1.63	5.08E-04	1.46	2.51E-02
Stiffness of joint	rs11817964	Musculoskeletal	1.94	8.71E-05	1.62	5.00E-02
Allergies, other	rs2394616	Injuries	2.13	3.11E-04	1.78	2.51E-02
Inflammatory diseases of female pelvic organs	rs1931553	Genitourinary	1.57	8.70E-04	1.36	3.67E-02
Conduct disorders	rs12768228	Psychiatric	2.53	8.96E-04	2.31	4.26E-02
Symptoms involving urinary system	rs11030043	Genitourinary	1.76	7.35E-06	1.65	4.32E-02
Disorders of cornea	rs16905974	Neurologic	2.69	6.12E-04	2.69	1.07E-02
Obstructive sleep apnea	rs7133666	Neurologic	1.36	7.18E-05	1.22	8.66E-03
Erythematous conditions	rs17191680	Dermatologic	1.31	7.60E-04	1.25	8.98E-03
Emphysema	rs12579609	Pulmonary	1.88	9.30E-04	1.45	4.77E-02
Other conditions of brain	rs11060784	Neurologic	2.21	9.59E-04	1.55	3.77E-02
Neoplasm of unspecified nature of digestive system	rs9316483	Neoplastic	1.84	2.58E-04	1.97	2.27E-02
Malunion fracture	rs17080490	Musculoskeletal	3.91	4.52E-04	2.22	2.18E-02
Disorders of other cranial nerves	rs12896790	Neurologic	1.72	1.55E-04	1.58	1.05E-02
Malignant neoplasm of brain and nervous system	rs3783796	Neoplastic	5.63	8.19E-05	4.04	9.57E-03
Bipolar	rs11159544	Psychiatric	1.45	5.14E-04	1.31	2.79E-02
Fracture of humerus	rs4617810	Injuries	2.77	8.77E-04	1.86	1.27E-02
Chronic obstructive asthma	rs2240903	Pulmonary	2.75	3.35E-04	2.89	3.77E-03
Alopecia	rs17765170	Dermatologic	2.21	9.55E-04	2.86	1.43E-03
Generalized anxiety disorder	rs6122806	Psychiatric	3.23	9.34E-04	2.73	1.39E-02
Stress incontinence, female	rs17766531	Genitourinary	2.95	6.94E-04	2.10	1.02E-02

Atherosclerosis of renal artery	rs5756326	Cardiovascular	1.53	5.38E-05	1.34	3.50E-02
Nontoxic multinodular goiter	rs2886122	Endocrine & Metabolic	1.45	5.14E-04	1.31	2.79E-02
Disorders of the autonomic nervous system	rs2281117	Neurologic	2.13	7.69E-04	2.05	4.28E-03

Appendix B. Phecodes tested in GCTA.

Phecode	**Phenotype Description**	**Type**
296	Mood disorders	Brain
296.2	Depression	Brain
300	Anxiety, phobic and dissociative disorders	Brain
300.1	Anxiety disorder	Brain
208	Benign neoplasm of colon	Digestive
276	Disorders of fluid, electrolyte, and acid-base balance	Digestive
276.1	Electrolyte imbalance	Digestive
276.5	Hypovolemia	Digestive
278	Overweight	Digestive
278.1	Obesity	Digestive
536	Disorders of function of stomach	Digestive
562	Diverticulosis and diverticulitis	Digestive
562.1	Diverticulosis	Digestive
564.1	Irritable Bowel Syndrome	Digestive
465	Acute upper respiratory infections	Immune
555.1	Crohn's disease	Immune
250.2	Type 2 diabetes	Lipid metabolism
272	Disorders of lipoid metabolism	Lipid metabolism
272.1	Hyperlipidemia	Lipid metabolism
272.11	Hypercholesterolemia	Lipid metabolism
401	Hypertension	Lipid metabolism
411	Ischemic Heart Disease	Lipid metabolism
411.2	Myocardial Infarction	Lipid metabolism
411.4	Coronary atherosclerosis	Lipid metabolism
443.9	Peripheral Arterial Disease	Lipid metabolism
362	Retinal disorders	Ocular
362.2	Macular degeneration	Ocular
362.29	Age-related macular degeneration	Ocular
365	Glaucoma	Ocular

366	Cataract	Ocular
367	Disorders of refraction and accommodation	Ocular
367.1	Myopia	Ocular
367.2	Astigmatism	Ocular
367.8	Hypermetropia	Ocular
368	Visual disturbances	Ocular
371	Inflammation of the eye	Ocular
371.3	Inflammation of eyelids	Ocular
379	Other disorders of eye	Ocular
379.2	Disorders of vitreous body	Ocular
681	Superficial cellulitis and abscess	Skin
687	Symptoms affecting skin	Skin
700	Corns and callosities	Skin
702	Other dermatoses	Skin
702.1	Actinic keratosis	Skin
702.2	Seborrheic keratosis	Skin
939	Atopic or contact dermatitis	Skin

Appendix C. Removing E2 untested variants from E1 does not have an appreciable effect on risk explained or *P* value.

Phenotype	**Discovery (E1; two GRM)**		**Discovery (E1; two GRM; lacking missing E2)**	
	Risk Explained	***P***	**Risk Explained**	***P***
Actinic keratosis	0.85%	0.162	0.95%	0.124
Depression	1.93%	0.0036	1.90%	0.0031

Appendix D. Nominally significant (discovery $P < 10^{-4}$) replicating results from meta-analyses. Significant results (from Table 1-C) are in bold. Results shaded in light grey were not found in pooled analysis (discovery $P < 0.001$, replication $P < 0.05$, consistent direction of effect).

			Discovery		**Replication**	
Phenotype	**SNP**	**Flanking Gene(s)**	**Odds Ratio**	***P***	**Odds Ratio**	***P***
Hypercoagulable state	**rs3917862**	**SELP**	**3.32**	**9.9E-07**	**3.00**	**5.0E-10**
Protein-calorie malnutrition	**rs12049593**	**SLC35F3**	**1.77**	**2.0E-06**	**1.63**	**5.5E-05**
Symptoms involving urinary system	**rs11030043**	**RHOG, STIM1**	**1.76**	**7.4E-06**	**1.65**	**4.3E-02**
Tobacco use disorder	**rs901033**	**SLC6A11**	**2.19**	**1.7E-05**	**1.75**	**7.9E-04**

Hemangioma and lymphangioma, any site	rs17114127	PPAP2B	2.61	4.5E-05	3.41	4.3E-04
Functional disorders of bladder	rs17115796	DAB1	2.40	5.4E-05	1.85	2.2E-02
Stress incontinence, female	rs17766531	PRDM15	1.53	5.4E-05	1.34	3.5E-02
Microscopic hematuria	rs35609966	CCR6, GPR31	2.51	6.5E-05	2.81	3.6E-02
Obstructive sleep apnea	rs7133666	PIK3C2G	1.36	7.2E-05	1.22	8.7E-03
Malignant neoplasm of brain and nervous system	rs3783796	PRKCH	5.63	8.2E-05	4.04	9.6E-03
Stiffness of joint	rs11817964	ZNF365	1.94	8.7E-05	1.62	5.0E-02
Psoriasis vulgaris	rs12190231	EEF1E1, SLC35B3	1.48	9.1E-05	1.31	1.4E-02
Sleep related movement disorders	rs3771635	PKP4	0.70	9.8E-05	0.75	3.9E-03

Appendix E. Nominally significant (discovery $P < 10^{-4}$) replicating results from pooled analyses. Results reaching locus-wise Bonferroni corrected threshold are in bold. Results shaded in light grey were not found in the meta-analysis (discovery $P < 10^{-4}$, replication $P < 0.05$, consistent direction of effect).

			Discovery		Replication	
Phenotype	**SNP**	**Flanking Gene(s)**	**Odds Ratio**	***P***	**Odds Ratio**	***P***
Protein-calorie malnutrition	**rs12049593**	**SLC35F3**	**1.72**	**3.14E-06**	**1.52**	**3.56E-04**
Hypercoagulable state	**rs3917862**	**SELP**	**2.66**	**1.46E-05**	**2.67**	**3.66E-10**
Gastroparesis	rs4963700	SOX5	2.63	4.01E-05	1.48	2.25E-02
Sleep related movement disorders	rs3771635	PKP4	0.69	5.77E-05	0.73	1.28E-03
Obstructive sleep apnea	rs7133666	PIK3C2G	1.36	7.67E-05	1.20	1.68E-02
Other conditions of the mother complicating pregnancy	rs16868879	NCALD	5.39	7.81E-05	2.11	1.44E-03
Other alveolar and parietoalveolar pneumonopathy	rs10456309	KIAA0319	3.19	7.98E-05	1.81	3.57E-02

Appendix F. All associations that pass the discovery threshold ($P < 3.3E-05$) in the Neanderthal E1 meta-PheWAS. Sorted by SNP rsID. K is number of sites with enough cases to perform PheWAS.

Phenotype	**Phecode**	**SNP**	**OR**	***P***	**Total**	**Cases**	**MAF**	**K**
Disorders of parathyroid gland	252	rs1021216	2.544	4.62E-06	12228	252	2.89%	5

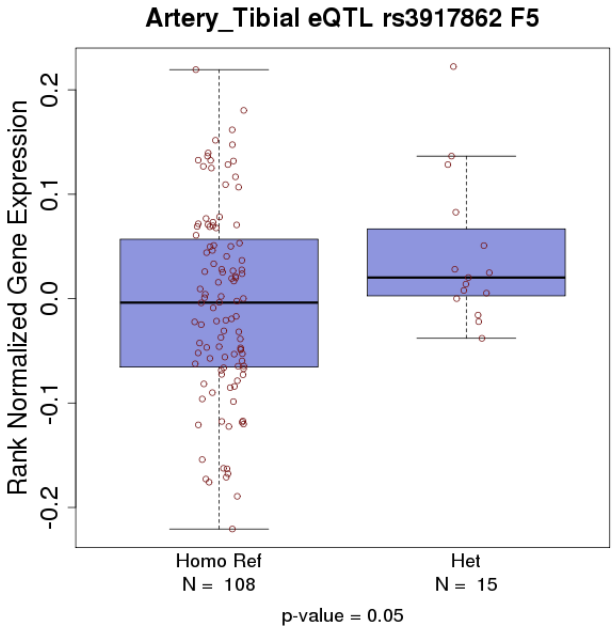
Secondary malignant neoplasm of liver	198.4	rs10487763	2.303	1.56E-06	6869	113	10.77%	3
Bariatric surgery	539	rs10492532	3.290	9.83E-07	7947	133	3.08%	3
Other disorders of intestine	569	rs10492532	1.921	1.76E-05	10705	512	3.24%	5
Intestinal infection	008	rs10498568	1.751	2.32E-05	10309	395	6.63%	4
Disorders of the globe	360	rs10509190	4.822	3.34E-05	1944	80	1.98%	2
Ill-defined descriptions and complications of heart disease	429	rs10842879	1.276	2.89E-05	10438	1905	13.20%	5
Nephritis and nephropathy without mention of glomerulonephritis	580.3	rs10962788	1.727	1.32E-05	9412	315	9.94%	5
Disorders resulting from impaired renal function	588	rs10962788	1.841	1.91E-05	7095	233	10.13%	3
Symptoms involving urinary system	599.8	rs11030043	1.757	7.35E-06	4713	352	10.49%	3
Poisoning by water, mineral, and uric acid metabolism drugs	974	rs11198978	5.928	4.88E-06	2031	39	3.84%	1
Megaloblastic anemia	281.1	rs11704728	0.438	2.08E-05	7300	171	19.88%	4
Protein-calorie malnutrition	260	rs12049593	1.766	1.98E-06	11205	648	5.15%	5
Nodular lymphoma	202.21	rs12200765	3.204	2.45E-05	6451	41	12.47%	2
Ill-defined descriptions and complications of heart disease	429	rs12230122	1.277	2.85E-05	10438	1905	13.17%	5
Other diseases of the teeth and supporting structures	525	rs12401852	2.167	3.13E-05	5007	334	3.21%	2
Complication of amputation stump	874	rs12401852	6.212	1.91E-05	2676	28	3.27%	1
Inflammatory bowel disease	555	rs1242069	2.214	2.38E-05	6773	197	4.62%	5
Cancer of other female genital organs	184	rs12431327	3.064	2.20E-05	7843	87	4.35%	3
Blister	911	rs12431327	5.620	2.10E-05	3060	26	3.92%	1
Lump or mass in breast	611.3	rs12497155	2.086	3.25E-05	10360	978	2.51%	5
Kyphosis (acquired)	737.1	rs12566055	6.074	1.61E-05	6500	91	2.02%	3
Poisoning by anticonvulsants and anti-Parkinsonism drugs	966	rs12671457	3.894	2.48E-05	2017	25	15.17%	1
Anemia in chronic	285.21	rs12770965	1.930	2.55E-05	7353	224	9.55%	4

kidney disease								
Hirsutism	704.2	rs12896790	3.899	2.02E-05	2511	40	9.84%	1
Pericarditis	420.2	rs12964811	2.163	2.49E-05	7483	123	9.71%	3
Macular degeneration, wet	362.22	rs1296793	1.755	5.27E-06	4376	267	19.65%	3
Congenital anomalies of great vessels	747.13	rs13039978	4.331	2.95E-05	2601	24	6.15%	1
Other disorders of the nervous system	349	rs13167382	4.539	8.87E-06	2840	41	2.82%	1
Neurological disorders due to brain damage	292	rs1324774	1.554	3.06E-05	9979	1315	4.59%	5
Altered mental status	292.4	rs1324774	2.339	1.16E-05	8927	263	4.44%	5
Cancer within the respiratory system	165	rs13249746	1.997	1.02E-05	13442	368	3.72%	5
Lung cancer	165.1	rs13249746	1.961	2.49E-05	13433	359	3.71%	5
Corns and callosities	700	rs13357477	0.709	2.22E-05	6144	677	23.16%	3
Hammer toe	735.21	rs1395342	1.709	2.83E-05	6763	354	8.69%	3
Corneal degenerations	364.4	rs1566479	1.806	2.67E-05	2013	149	27.50%	2
Pancreatic cancer	157	rs16918099	12.455	1.66E-05	4502	57	0.88%	2
Other unspecified back disorders	724.9	rs17116637	5.913	1.32E-05	830	34	3.37%	1
Other unspecified back disorders	724.9	rs17116658	6.227	8.72E-06	830	34	3.13%	1
Cardiac arrhythmia NOS	427.5	rs17235910	1.561	3.29E-05	6706	1271	6.33%	5
Cardiac and circulatory congenital anomalies	747	rs17235910	2.012	1.62E-05	8729	244	6.07%	4
Other diseases of the teeth and supporting structures	525	rs17244660	0.459	1.05E-05	5007	334	11.83%	2
Pain, swelling or discharge of eye	379.9	rs17304921	2.824	2.28E-06	4464	156	5.63%	3
Mild cognitive impairment	292.2	rs17324684	6.770	2.70E-06	4640	55	3.09%	2
Open wound of eye or eyelid	870.1	rs17368659	5.604	9.10E-06	2013	22	13.64%	1
Myoclonus	333.2	rs17434648	5.178	1.10E-05	2839	40	2.99%	1
Aphasia	292.11	rs17481185	2.375	1.39E-05	6438	111	7.86%	2
Other dyschromia	694.2	rs17614605	0.655	2.11E-05	6932	345	29.12%	4
Primary open angle glaucoma	365.11	rs17626479	1.669	2.72E-05	6460	361	11.81%	4
Hepatomegaly	573.3	rs17633592	5.604	1.55E-05	2620	23	5.50%	1
Lipoma	214	rs17684048	1.873	2.17E-05	11551	458	4.00%	4

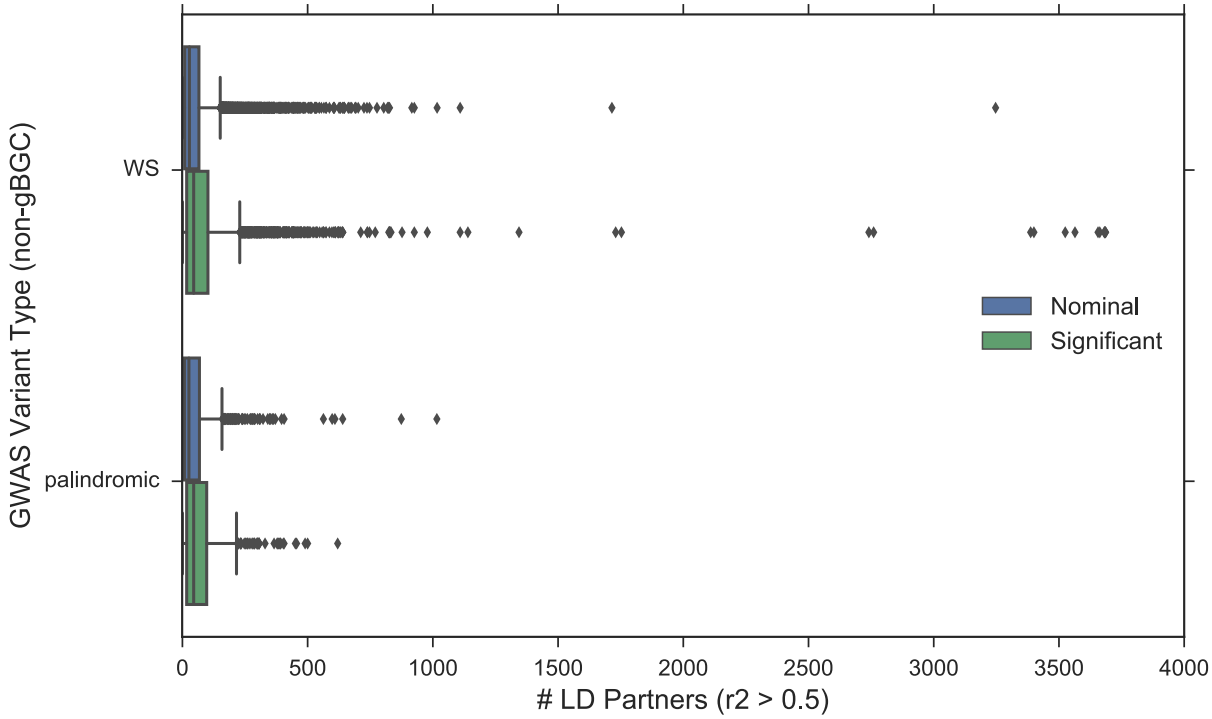
Retinal vascular changes and abnomalities	362.4	rs17695527	2.113	1.44E-06	4610	501	4.72%	3
Postoperative infection	80	rs17777982	2.069	3.78E-07	12707	399	4.52%	5
Cramp of limb	771.2	rs17787978	1.951	1.29E-05	8036	229	8.05%	4
Renal colic	594.8	rs17793551	6.248	3.35E-06	3570	32	3.63%	1
Abnormal heart sounds	396	rs1914191	1.459	6.62E-06	10706	702	11.38%	5
Cholelithiasis with other cholecystitis	574.12	rs1954003	2.188	1.76E-05	6175	173	6.19%	2
Senile osteoporosis	743.12	rs2062020	4.012	2.05E-07	2220	50	11.15%	1
Iron deficiency anemia secondary to blood loss	280.2	rs2076222	3.933	1.04E-05	5383	147	2.22%	3
Viral Enteritis	8.6	rs2289837	7.492	1.19E-05	3714	49	1.17%	1
Other specified diseases of nail	703.8	rs2298146	1.798	2.41E-05	7592	935	2.77%	3
Brain cancer	191.11	rs2319366	3.885	2.77E-05	2642	26	9.92%	1
Internal derangement of knee	835	rs2361394	1.620	3.29E-05	12115	874	4.49%	5
Lipoma	214	rs2498638	1.679	2.88E-05	11551	458	6.75%	4
Cervical radiculitis	765	rs2498638	1.667	1.95E-05	10310	507	6.48%	5
Supraventricular premature beats	427.61	rs2807345	2.129	2.37E-05	3142	246	7.65%	3
Senile osteoporosis	743.12	rs2820112	3.650	2.18E-06	2220	50	10.90%	1
Anemia of chronic disease	285.2	rs3092999	2.175	1.89E-05	8479	518	2.42%	5
Other congenital anomalies of skin	691	rs3092999	10.833	2.97E-06	1847	21	2.36%	1
Celiac or tropical sprue	557.1	rs3132630	4.562	1.10E-05	1626	21	13.01%	1
Anemia of chronic disease	285.2	rs35609966	1.883	3.02E-05	8479	518	2.68%	5
Other congenital anomalies of skin	691	rs35609966	3.998	4.41E-06	1847	21	3.38%	1
Candidiasis of skin and nails	112.3	rs3743162	2.748	1.55E-05	2398	54	23.06%	1
Visual field defects	368.4	rs3793829	2.461	9.63E-07	6835	252	4.16%	3
Abnormal sputum	516	rs3793829	2.608	1.27E-05	5332	184	4.25%	3
Hemoptysis	516.1	rs3793829	2.595	2.63E-05	5317	169	4.24%	3
Althete's foot	110.12	rs3798604	1.961	2.44E-05	7285	289	6.23%	4
Disorders of the pituitary gland and its hypothalamic control	253	rs3822947	2.003	2.71E-05	12124	148	10.35%	5
Hypercoagulable state	286.8	rs3917862	3.318	9.87E-07	4566	73	6.20%	2
Uterine/Uterovaginal	618.2	rs3917862	2.011	1.66E-05	9029	242	6.55%	3

prolapse								
Dysmetabolic syndrome X	277.7	rs41280400	3.265	2.19E-05	6293	82	4.10%	2
Abnormal reflex	350.5	rs4323776	3.744	3.32E-05	2825	24	15.59%	1
Disorders of liver	573	rs4492593	1.243	2.99E-05	11797	1702	16.67%	5
Anal and rectal polyp	565.1	rs4617810	2.622	2.69E-05	5234	327	2.07%	2
Symptoms involving nervous and musculoskeletal systems	781	rs4692446	2.458	3.00E-05	6095	273	3.15%	2
Iatrogenic hypothyroidism	244.1	rs4743645	2.927	1.06E-05	5781	155	2.71%	3
Optic atrophy	377.1	rs547136	3.017	1.50E-05	2137	98	5.08%	2
Upper gastrointestinal congenital anomalies	750.1	rs6122806	8.006	2.78E-06	2613	36	1.63%	1
Hypercoagulable state	286.8	rs668696	3.503	1.24E-05	4566	73	4.58%	2
Hirsutism	704.2	rs735225	4.875	1.90E-05	2511	40	4.80%	1
Extrapyramidal disease and abnormal movement disorders	333	rs7494783	1.497	1.95E-05	10380	576	10.80%	5
Hemorrhoids	455	rs7602743	1.355	2.21E-05	9460	2774	8.50%	5
Superficial keratitis	370.2	rs7929411	5.935	2.58E-05	2137	151	0.40%	2
Fracture of pelvis	802	rs7950298	2.568	6.65E-06	5972	171	4.00%	3
Other dyschromia	694.2	rs8007941	0.651	1.63E-05	6932	345	29.11%	4
Tobacco use disorder	318	rs901033	2.188	1.73E-05	12181	1420	1.06%	5
Renal colic	594.8	rs901033	12.497	2.90E-05	3570	32	0.95%	1
Nodular lymphoma	202.21	rs9405316	3.727	1.26E-05	6451	41	8.01%	2
Chronic liver disease and cirrhosis	571	rs9462492	1.527	3.16E-05	10479	384	12.99%	5
Chronic nonalcoholic liver disease	571.5	rs9462492	1.610	1.13E-05	8478	325	13.00%	4
Graves' disease	242.1	rs9524432	5.426	2.71E-06	5019	59	2.08%	2
Abnormal Papanicolaou smear of cervix and cervical HPV	792	rs9530050	1.858	4.05E-06	9336	437	7.75%	4
Abnormal Papanicolaou smear of cervix and cervical HPV	792	rs9530145	1.685	7.05E-07	9336	437	16.55%	4
Abnormal Papanicolaou smear of cervix and cervical HPV	792	rs9543041	1.853	4.50E-06	9336	437	7.77%	4
Hydronephrosis	595	rs9986932	1.956	2.22E-05	11266	169	10.80%	4

Appendix G. rs3917862 is significantly associated with increased expression of F5 in tibial artery.



Appendix H. The full distribution of LD partners for GWAS catalog variants falling outside of gBGC tracts.



Appendix I. Recombination rates differ significantly between GWAS catalog variants with significant associations and variants with nominal associations. (a) Recombination rates for GWAS catalog variants falling outside of gBGC tracts. (b) Recombination rates for GWAS catalog variants falling within gBGC tracts.