The role of ancient genetic variation on human adaptation: Insights from trans-species and introgressed variation

By

Keila S. Velazquez-Arcelay

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

Biological Sciences

December 16, 2023

Nashville, Tennessee

Approved:

Antonis Rokas, Ph.D.

Douglas McMahon, Ph.D.

Nicole Creanza, Ph.D.

Jada Benn Torres, Ph.D.

John A. Capra, Ph.D.

In memory of my grandmothers, Maria de los Santos Ortiz and Paula Martinez, who passed away during the preparation of this work.

ACKNOWLEDGMENTS

During the writing of this dissertation, I jumped down so many compelling rabbit holes—but rabbit holes nonetheless—that you might think my name is Alice. This work would not have been possible without my advisor, Tony Capra, whose expert guidance and occasional rescue missions to pull me out of these trips ensured I stayed on course. Thanks to the members of the Capra Lab for their brilliant and critical suggestions and feedback on my work. I would also like to thank my committee for their wisdom and suggestions, which were always communicated in an encouraging manner. Thank you especially to my parents for their unconditional love and support and consistent willingness to guide me through the world. To my aunts and uncles, who played a crucial role in supporting me unconditionally during my development years, and to my brothers and cousins for being my friends when making friends was extremely hard. To my friends for their willingness to geek out with me. This dissertation process would have been much more difficult without the support of my wife. My dog Amy consistently gave me a lot of emotional support for over a decade, but she will probably never know this. This work would not have been possible without financial support from the National Institutes of Health [Grant R35GM127087; Grant T32LM012412], and the Burroughs-Wellcome Fund and the resources from the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

1.1 Evolutionary history of the genus *Homo*

Today, humans stand as a distinctive species within the animal kingdom, demonstrating a capacity for creating complexity in the form of tools, religion, architecture, transport systems, computers, the internet, and even weapons of mass destruction. But how did humans come to possess such extraordinary capabilities? One of the elements that made this possible is a unique skill: the cultural transmission of knowledge, or what Richard Dawkins referred to as "memes" in 1976 (Dawkins, 1976). These memes, like viral ideas spreading through the collective consciousness, have driven human culture in a constant development of complexity. However, the current human state is the result of millions of years of evolution and adaptation.

Humans are primates, belonging to the lineage that originated in Africa during the Paleocene epoch approximately 60 million years ago (Ma) (Gagneux and Varki, 2001; Page and Goodman, 2001). Primates are divided in two groups: the strepsirrhines and the haplorrhines. The strepsirrhines include the lemuriforms and lorises, which are characterized by their wet nose and grooming claws, and nocturnal lifestyle. The haplorrhines, in turn, are divided into the tarsiers and anthropoids: the platyrrhines (New-World monkeys) and the catarrhines (Old-World monkeys and apes) (Bonner, Heinemann and Todaro, 1980; Koop *et al.*, 1989; Bailey *et al.*, 1991, 1992; Porter *et al.*, 1995; Porter, Czelusniak, *et al.*, 1997; Porter, Page, *et al.*, 1997; Goodman *et al.*, 1998; Zietkiewicz, Richer and Labuda, 1999). The main difference between the groups is in their continental origin and nose shape. The New-World monkeys evolved in South America with flat noses and wide nostrils. The Old-World monkeys and apes evolved in Africa with narrow noses and nostrils. This latter group gave rise to several species of apes, including gorillas, orangutans, chimpanzees, and bonobos. The apes also include humans, and our aim is to delve deeper into evolutionary aspects of this group (Shoshani *et al.*, 1996; Goodman *et al.*, 1998).

The chimpanzees are the closest living relatives of humans, sharing a common ancestor in Africa around 7 Ma (Prado-Martinez *et al.*, 2013). Genetic analysis revealed that humans and chimpanzees share around 98% of genomic identity (Chimpanzee Sequencing and Analysis Consortium, 2005; Patterson *et al.*, 2006; Prado-Martinez *et al.*, 2013; Kronenberg *et al.*, 2018).

Although *Homo sapiens* is the only extant *Homo*, up until 35 Ma they coexisted with many other species of *Homo* (Finlayson *et al.*, 2006; Stepanchuk *et al.*, 2017; Bokelmann *et al.*, 2019). Archaeological findings across Africa and Eurasia during the 20th century revealed a complex evolutionary history of the genus *Homo* in Africa. A few examples of precursor species have been found across Africa. The oldest known group, *Sahelanthropus tchadensis*, was found in the central-northern part of Africa (Djurab Desert, Chad) and dates back to approximately the Late Miocene 7 Ma (Brunet *et al.*, 2002; Vignaud *et al.*, 2002; Michel Brunet *et al.*, 2005). For instance, the archaeological evidence has not provided clear evidence of bipedalism (Meyer *et al.*, 2023), and its cranial structure was similar to both apes and hominins. Additionally, it lived close to the divergence between humans and chimpanzees. Another early hominin, *Orrorin tugenensis*, was found in the year 2000 in eastern Africa (Togen Hills, Kenya) and dates back to the Late Miocene 6 million years ago (Senut *et al.*, 2001). Based on the upper limb anatomy, this species could have been bipedal while also still adapted to climbing and tree-dwelling. However, there is not enough evidence to support full bipedalism. Approximately 1,200 km to the north, another site in Ethiopia yielded remains from *Ardipithecus ramidus* specimens that lived between the Late Miocene and Early Pliocene 4.4 million years ago (White *et al.*, 1993, 2009; White, Suwa and Asfaw Berhane, 1994). Anatomical features found in *A. ramidus* remains also suggest that it was adapted to both tree-dwelling and upright walking, and its pelvis and foot bones were similar to those found in later bipedal hominin species. Still, whether it was fully bipedal is still up for debate.

Australopithecines lived 4-3 Ma in eastern and southern Africa. The genus is divided into robust (e.g. *robustus*, *boisei*) and gracile (e.g. *anamensis*, *afarensis*, *africanus*, *sediba*, *ghari*) species. This is the first genus to show clear evidence of bipedalism (Strait, 2010). A representative skull of *Au. africanus*, the Taung Child discovered in 1924 in South Africa. Its foramen magnum is positioned at the base of the skull, indicating that it walked upright (Dart, 1925). Three decades later in 1974, archaeologists found one of the most important specimens in the study of human evolution. A nearly complete representative fossil of the *Australopithecus afarensis* was discovered at Hadar in the Awash Valley region of Ethiopia (Kimbel, Johanson and Rak, 1994), and dated to the Pliocene 3.2 million years ago (Johanson and Taieb, 1976; Ferguson, 1989). This fossil was affectionately named "Lucy" after the Beatles' song "Lucy in the Sky with Diamonds."

The factors leading to the evolution of the genus *Homo* are still debated. Evolution is a continuous interaction between genetic variation in a population and the environment, where the environment shapes the frequencies of traits prevailing in populations. Thus, when asking questions about evolutionary changes, it is essential to consider the evolutionary factors acting on the phenotypic traits of a population. A series of environmental hypotheses have been proposed to explain the development of *Homo*. The first idea, called the turnover-pulse hypothesis, proposes that the cooling period 2.5 million years ago and retreat of land water due to a Northern Hemisphere glaciation was a potential major force behind the emergence of *Homo*. It is proposed that this environmental shift overlapped some transitions in hominin evolution, potentially influencing adaptations to grasslands and the development of an increased brain size (Prentice and Denton, 1988; Vrba, 1988, 1994). Periods of global cooling were also suggested to have caused aridification in Africa Plio-Pleistocene (Vrba, 1995a, 1995b). This idea is supported by flora (Cerling and Hay, 1986; Cerling, 1992), fauna (Bobe and Eck, 2001; Bobe and Behrensmeyer, 2004) and soil studies (DeMenocal and Bloemendal, 1995; deMenocal, 2004). These environmental factors overlapped with the emergence of various hominin groups (DeMenocal and Bloemendal, 1995; deMenocal, 2004; DeMenocal, 2011). Another hypothesis suggests that over 4 million years of warming periods in the Turkana basin of Ethiopia and Kenya played a role in hominin evolution (Passey *et al.*, 2010). It was proposed that this adaptive challenge could have led to the evolution of thermoregulation mechanisms such as upright bodies, exposed skin, and sweating. This period overlaps the evolution of protective dark skin pigmentation around 1.2 million years ago (Rogers, Iltis and Wooding, 2004). Opposite to the aridity hypothesis, it was proposed that periods of humidity played an important role in driving hominin evolution (Trauth *et al.*, 2005, 2007, 2010; Trauth, Larrasoaña and Mudelsee, 2009). The suggested periods of humidity overlap the warming periods, and it has thus been suggested that there was a constant shift between the arid and the humid phases during this period. Following the idea of variation in environmental factors, a variability selection hypothesis was proposed. This hypothesis proposes that constantly fluctuating environmental factors favored the emergence of a biological plasticity in the capacity of organisms to adapt to a variety of environments (Potts, 1996, 1998b, 1998a, 2012b, 2012a).

One key trait defining the genus *Homo* is the production and utilization of tools (Semaw *et al.*, 1997; Potts, 1998a; McPherron *et al.*, 2010; Stout *et al.*, 2015). *Homo habilis* is the first Homo group unmistakably evidenced to possess the ability to produce stone tools extracted from flakes of stone (Leakey, 1971; Semaw *et al.*, 2003). The first specimen was discovered at Olduvai Gorge in Tanzania by Louis Leakey in the early 1960s and is dated to 2.8 million years ago (Leakey, Tobias and Napier, 1964). Thus, the technology of producing pounders, choppers, and scrapers is called the Oldowan industrial complex (Susman, 2017). Subsequently, archaeological excavations across Africa have yielded thousands of additional stone tools

(Semaw, 2000). These early tools are believed to have been employed in the processing of food. Another key trait in humans is the capacity for complex language. Behaviors do not fossilize, but this trait can be detected by proxy through the presence of Broca's area in the brain. In humans, Broca's area is located in the posterior part of the frontal cortex, above the lateral sulcus, and is involved in language production and comprehension. Although brains do not fossilize either, this pattern can be observed in the inner surface of the skull. The human brain contains convolutions and folds (gyri and sulci). During development, the brain presses against the skull, leaving endocranial imprints (endocasts) in its inner surface (Holloway *et al.*, 2009). The endocasts of early *Homo* can be analyzed to compare their sulcal patterns to that of humans and great apes. For instance, the sulcal pattern in *Australopithecines* has an ape-like shape (Falk, 1980). The oldest finding containing a sulcal pattern more similar to humans comes from an early *Homo* specimen (KNM-ER 1470) found near Lake Turkana in Kenya and dated to 2 million years ago (Holloway, 1983, 2015). Its classification is still debated, but it was classified as a *Homo rudolfensis*. *Homo erectus* fossils also show a similar configuration (Zeitoun *et al.*, 2010; Hillert, 2021). It is important to note that even if these early *Homo* had the capacity for language, it is not enough evidence to support the physical capacity for speech based on the anatomy of the hyoid bone (Capasso, Michetti and D'Anastasio, 2008; Steele, Clegg and Martelli, 2013).

One of the striking traits that stand out about humans is their ability to survive in many environments, partially due to their tool making ability. The first demographic expansion *Homo* is evidenced in *Homo erectus*. *Homo erectus* likely originated in Africa around 1.8 million years ago (Asfaw et al., 2002; Brown et al., 1985; Simpson et al., 2008; Suwa et al., 2007). Fossil remains have been found in Ethiopia (Asfaw *et al.*, 2002)[Asfaw 2002], Kenya [Leakey 1976] (Leakey, 1976; Brown *et al.*, 1985), and Tanzania (Rightmire, 1979). Not long after their appearance in the fossil record, individuals in this group dispersed to Eurasia and Maritime Southeast Asia. Many other fossils have been excavated in Georgia (Lordkipanidze *et al.*, 2013), China (Weidenreich, 1938; Yue *et al.*, 2004; Shen *et al.*, 2009), Indonesia (Zaim *et al.*, 2011), Turkey (Lebatard *et al.*, 2014). *Homo erectus* was capable of making more sophisticated tools, representing a transition from Oldowan to Acheulean culture (Joordens *et al.*, 2015).

1.2 The common ancestor between Humans and Neanderthals

During the Middle Pleistocene, a number of *Homo* were already living across Eurasia. The closest known relatives of humans are the Neanderthals and the Denisovans. Neanderthals and

Denisovans are closely related groups that originated in Eurasia at least a hundred thousand years before humans originated in Africa. These groups are believed to have diverged around 650 thousand years ago (ka) (Prüfer *et al.*, 2014; Meyer *et al.*, 2016; Steinrücken *et al.*, 2018). The common ancestor between these groups is still unknown. In Europe, some of the archaic human remains were found in the Atapuerca Mountains, in the province of Burgos in Spain. The caves in Atapuerca are one of the places with the largest abundance of ancient *Homo* remains. In the early 1900s (late 1800s) a trench was dug up across the Atapuerca Mountains to build a railway. The trench exposed a series of sediment-filled caves, containing human and animal remains (Cervera *et al.*, 1998; Santos Ganges, 2003). The trench, called Trinchera del Ferrocarril, or Railway Trench, contains three caves that yielded important human findings. The first cave, Sima del Elefante TE9, yielded bones from an unidentified *Homo* group, *Homo sp*., dating to 1.2 million years ago (Arsuaga *et al.*, 2000; López-Valverde, López-Cristiá and Gómez De Diego, 2012). In 2022, a partial face was excavated from TE7 and dated to 1.4 million years, which means that this is one of the oldest fossils found in the region. This specimen is still under study and no additional information is yet available, but this fossil will help better understand the identity of the *Homo* group that first colonized this area (Equipo Investigador de Atapuerca, 2022). The Galería cave system is composed of three sections: Galería (TG), Covacha de los Zarpazos (TZ), and Tres Simas Norte (TN) (Bermúdez de Castro and Rosas, 1992; Arsuaga *et al.*, 2000; Demuro *et al.*, 2014; Núñez-Lahuerta *et al.*, 2022). The Galería lay below a sinkhole that functioned as a trap to animals and thus served as a scavenging site for hominins. Several human-made items were found in this cave that likely served as butchering tools. Animal remains show evidence of burning and cut marks. Two human remains were found in TZ: a jaw fragment found in the 1970s and a skull fragment in 1995, dated to 0.6-0.4 Ma and attributed to *Homo heidelbergensis* (Carbonell *et al.*, 1999; Arsuaga *et al.*, 2000; Falguères *et al.*, 2013; Ollé *et al.*, 2013). The Gran Dolina cave TD6 yielded 6 *Homo antecessor* individuals dating to around 1.2-0.8 Ma (Arsuaga *et al.*, 2000; Bermúdez de Castro *et al.*, 2004; Ollé *et al.*, 2013). Some of the human bones show cut marks and were broken likely to access the marrow, suggesting cannibalism (Fernández-Jalvo *et al.*, 1973). Another complex of caves is found south east of the Railway Trench, called Cueva Mayor. The most intriguing cave in Cueva Mayor is Sima de los Huesos (Pit of Bones). The cave contains a subterranean chamber with a complex network of passages and galleries. The bottom is accessed through a 13m deep vertical shaft (Arsuaga, Martínez, Gracia, Carretero, *et al.*, 1997). The Sima de los Huesos cave yielded over 1,600 bone fragments from at least 28 individuals (Arsuaga, Martínez, Gracia and Lorenzo, 1997; Bischoff *et al.*, 2003; Sala *et al.*, 2016; Bermúdez de Castro *et al.*, 2021). The remains have an approximate age of 0.4 Ma and belong to a group closely related to Neanderthals, or pre-Neanderthals, sometimes classified as *Homo heidelbergensis* (Meyer *et al.*, 2016). This includes one the most preserved craniums in the world, cranium 5 or Miguelón, who has become a celebrity in social media (Arsuaga *et al.*, 1993). Other *Homo heidelbergensis* specimens have been found in Italy (Manzi, Mallegni and Ascenzi, 2001), Germany (Schoetensack, 1908) (Wagner et al., 2010), England (Marston, 1937; Lockey *et al.*, 2022), Greece (Stringer, Howell and Melentis, 1979; Grün, 1996), Ethiopia (Rightmire, 1996), Zambia (Woodward, 1921; Grün *et al.*, 2020). Although *Homo heidelbergensis* was previously hypothesized to represent the last common ancestor between modern humans and Neanderthals, their emergence in the fossil record is too close to the divergence of humans and Neanderthals, and thus the validity of this claim still being debated. Another candidate ancestor has been *Homo antecessor*, although only remains have been found in the Iberian Peninsula. Due to the limited fossil record and incomplete specimens, it is challenging to reconstruct the past.

Neanderthals appear in the Eurasian fossil record around 0.4 ka, which is earlier than the estimated origin of humans. They are named after the first fossil of their kind found in the mid-19th century in the Neander Valley in Germany (King, 1864). Many other findings have been excavated all across Eurasia. This human lineage went extinct around 35 ka after they lived in refugia across Europe (Finlayson *et al.*, 2006; Higham *et al.*, 2014). Following the divergence between the ancestor of humans and Neanderthals, another split happened approximately 400 ka (Kelso and Prüfer, 2014), leading to the emergence of Neanderthal and Denisovans. Denisovan specimens have been identified in Asia only (F. Chen et al., 2019; M. Meyer et al., 2012; Reich et al., 2010; Slon et al., 2017).

1.3 The origin of *Homo sapiens* and out-of-Africa expansion

The most accepted model of human origins points to the emergence of *Homo sapiens* 300 ka (kilo annum) in Africa (Chan et al., 2019; J. J. Hublin et al., 2017; Jouganous et al., 2017; Steinrücken et al., 2018; C. Stringer, 2016). Several early specimens were discovered across the continent, most importantly in Morocco, South Africa, Kenya/Ethiopia, and the Levant. Two crania from Jebel Irhoud in Morocco were dated to approximately 300 ka (J. J. Hublin et al., 2017; Richter et al., 2017). While the findings represent the oldest human remains containing facial characteristics similar to modern humans, they contain more pronounced brow ridges similar to Neanderthals. Located in the northwestern side of Africa, this early individual is an isolated finding in that region. In eastern Africa, three crania found in the Omo Kibish Formation in Kenya are dated to 233 ka and they show a more modern facial structure (Vidal *et al.*, 2022a). To the north in Herto, Ethiopia, two more crania were found and dated to 160,000 BP (White *et al.*, 2003), but their facial traits appear more archaic than Djebel Irhoud. The age of this sample is likely underestimated, as determined by stratigraphy studies (Vidal *et al.*, 2022b). Several early human fossils have been found in the Levant region. In Misliya Cave, Israel, a maxillary fragment dated to 180,000 ka shows some *Homo sapiens* features (Hershkovitz *et al.*, 2018). Djebel Qafzeh yielded remains from 15 individuals dated to about 100 ka (Schwarcz *et al.*, 1988). Skhul Caves yielded 10 individuals to approximately 120 ka (Grün *et al.*, 2005). Many of the features found in these specimens are modern, although a few have pronounced brow ridges and occipital bun. These individuals may represent an early wave of migration out of Africa (Bons *et al.*, 2019). However, there is no evidence that these first expansion attempts produced any offspring into the present.

The out-of-Africa migration that resulted in all present-day non-Africans took place between 70 and 65 ka (Groucutt *et al.*, 2015; Bae, Douka and Petraglia, 2017; Rito *et al.*, 2019). This event coincides with the onset of the Marine Isotope Stage (MIS) 4 glacial period spanning 71 and 59 ka and peaking at 65 ka (Tierney, deMenocal and Zander, 2017; De Deckker *et al.*, 2019), which is part of a series of environmental changes called the Last Glacial Period (115-11.7 ka). The resulting deterioration of weather conditions across Africa leading to arid conditions might have influenced the different waves of migrations of humans across east Africa and into the Arabian Peninsula. These changes likely posed challenges to early humans and led to migrations along Arabia into Eurasia. The first human groups resulting from this out-of-Africa expansion migrated across southern Eurasian coastal lines, which were at lower sea levels due to the formation of glaciers, arriving in Australia as early as 65 ka (Clarkson *et al.*, 2017). There is no evidence of human colonization in Europe until later. During the MIS3 57 ka, climate patterns across Eurasia continually fluctuated between cooler and warmer periods, leading to waves of human colonization attempts (Benazzi *et al.*, 2011; Higham *et al.*, 2014; Fu *et al.*, 2016; Hublin, 2020; Prüfer *et al.*, 2021). None of the recovered human specimens in Europe from this period contributed to the ancestry of present-day Europeans. A migration wave dated to around 37 ka contributed entirely to the human ancestry in humans living in Europe up to around 14 ka. (Fu *et al.*, 2016). This date coincides with the end of the Last Glacial Maximum (LGM) and the beginning of the Late Glacial Interstadial warm period of the Late Glacial Period. At this time, groups of western hunter-gatherers from the Near East moved to Europe (Fu *et al.*, 2016; Lazaridis *et al.*, 2016), followed by groups of farmers from Anatolia 8 ka (Mathieson *et al.*, 2015; Hofmanová *et al.*, 2016).

The initial dispersal of small human groups to higher latitudes exposed them to new environments that they had never experienced, triggering new selective pressures associated with new factors such as pathogens, colder temperatures, reduced intake of UV radiation, extreme seasonal and photoperiodic patterns. Humans and their ancestors evolved over millions of years in near-equatorial regions, leading to a high degree of adaptation to this particular environment. Near the Equator, located at latitude 0 degrees, sunlight is more direct year-round leading to consistently warm temperatures and little fluctuation in the length of daylight hours and seasonal patterns. Most organisms evolved to synchronize their biological functions to the Earth's near 24-hour day cycles. These biological clocks are synchronized, or entrained, mainly by a light input (Roenneberg, Daan and Merrow, 2003; Foster and Roenneberg, 2008). Because of genetic variation, chronotypes (daytime or nighttime preference) exist in a population following a normal distribution, with most people falling under an intermediate chronotype (Toh *et al.*, 2001; Archer *et al.*, 2003; Roenneberg *et al.*, 2004). Over 351 single nucleotide polymorphisms have been identified to affect chronotype in a cohort of British individuals (Jones *et al.*, 2019; Wang *et al.*, 2019).

1.4 Sequencing of present-day human populations

The development of next-generation sequencing (NGS) technologies at the beginning of the 21st century opened the door for a new field of study of human variation and evolution (1000 Genomes Project Consortium, 2010a, 2015; Nielsen *et al.*, 2017). The ability to genotype DNA rapidly led to the development of several large-scale human sequencing projects. The International HapMap project aimed to sequence genomes from populations around the world and to create a catalog of human genetic variation. The project's first version was released in 2005 (Belmont *et al.*, 2005), and it contained a haplotype map (HapMap) of human variation. Initially, it included more than 1 million SNPs from 269 individuals in four populations from different continental ancestries: Africa, Europe, East Asia. The phase 3 of the project introduced 7 new populations, including South Asian and admixed American groups. The 1000 Genomes Project is a follow-up project published in 2010 (1000 Genomes Project Consortium, 2010a), which expanded the set of populations to 26. Today, it includes 88 million variants from 2,504 individuals from Africa, Asia, Europe, and admixed Americans (1000 Genomes Project Consortium, 2010a, 2015; Sudmant *et al.*, 2015). In this dataset, ~64 million variants are rare (frequency <0.5%), ~12 million have a frequency between 0.5% and 5%, and ~8 million have a frequency of >5%. Even though in the set most variants are rare, each individual genome contains mostly common variants. The Simons Genome Diversity Project (SGDP) was launched in 2016 and aimed to sequence the genomes of 300 individuals from 142 different populations, with a focus on historically underrepresented populations, including the San from South Africa, Papuans from New Guinea, Mayans from Mexico, Basque from Spain, Sardinians from Italy (Mallick *et al.*, 2016). The ExAC and Genome Aggregation Database (gnomAD) combined exome sequence data from 125,748 individuals from diverse populations around the world, and 15,708 whole-genomes (Lek et al., 2016) (Karczewski et al., 2020). The Cancer Genome Atlas (TCGA) project was developed in 2006 with the specific purpose of understanding the genetic basis of cancer (Weinstein *et al.*, 2013). It involved sequencing the genomes of over 11,000 cancer samples from 33 different cancer types. The project identifies genomic alterations that contribute to the development of cancer.

Other projects have aimed to sequencing individuals from specific populations. The UK Biobank project launched in 2006 and has sequenced 500,000 individuals in the United Kingdom, providing health and lifestyle information, genetic data, and biological samples (Sudlow *et al.*, 2015). Because of the variety of information collected from each individual, it has been widely used in biomedical research to understand the genetic basis of diseases. Other population-specific projects include the FinGen project that includes 500,000 individuals from Finland (University of Helsinki, 2017), the Irish DNA Atlas contains genome data for 194 Irish individuals (Gilbert *et al.*, 2017), Biobank Japan contains data for 200,000 Japanese individuals (Nagai *et al.*, 2017), and the EstonianBiobank sequenced nearly 52,000 Estonian individuals (Leitsalu *et al.*, 2015). Projects aimed specifically at gathering genomic data from patients include Vanderbilt University's BioVU in Nashville Tennessee, and University of California San Francisco (UCSF) and Stanford University electronic health record databases in California.

1.5 Ancient DNA sequencing technology and sequencing ancient humans and archaic hominins

Following the invention of polymerase chain reaction (PCR) in 1983, efforts were made to apply this technology could next be applied to the recovery of ancient genomes (Saiki *et al.*, 1985). However, sequencing ancient DNA poses additional challenges that first needed to be overcome to accurately sequence ancient DNA. Over time, the chemical bonds that hold DNA molecules together break down due to chemical and physical degradation, resulting in fragmentation and damage to the DNA. Moisture, heat, and UV radiation accelerate the degradation of DNA. The target genetic material also becomes contaminated by microbes, making it challenging to avoid sequencing the contaminants. In addition, microbial contamination can lead to the degradation of DNA, as microorganisms can break down DNA as a source of nutrients. Because of these factors, the type of environment in which the DNA is preserved is relevant to the speed of degradation. DNA is most likely to be preserved in cold and dry environments, such as permafrost or caves, where the DNA is protected from exposure to moisture and other environmental factors. In contrast, DNA is less likely to be preserved in warm and humid environments, such as tropical regions or coastal areas, where exposure to moisture and microbial activity can cause rapid degradation (Orlando *et al.*, 2021).

Before PCR technology was widely available, Svante Pääbo demonstrated that ancient DNA (aDNA) could survive and be isolated from ancient samples (Pääbo, 1984, 1985), initially from a 2,400-year-old Egyptian mummy. He then proceeded to clone genetic fragments in plasmid vectors. He used Alu elements, which are unique to monkeys and primates, as a probe to avoid cloning bacteria and fungi DNA. He successfully cloned a 3.4k long piece of DNA. However, this work received criticism due to the potential of contamination. For years he focused on improving the technology to successfully sequence aDNA using PCR while reducing contamination in the sampling process (Pääbo, 1989; Höss and Pääbo, 1993; Poinar *et al.*, 1996). In 1997 he published mitochondrial DNA extracted from a Neanderthal fossil found in Neander valley in 1856, and dated to 40,000 years old (King, 1864; Krings *et al.*, 1997).

Sequencing of ancient DNA is different from modern DNA due to natural processes influencing the degradation of DNA. Endonucleases are enzymes that cleave DNA at the phosphodiester bonds where nucleotides are linked. After death, the cease of biological activities and these enzymes are no longer in check. This leads to shorter fragments of DNA due to the degradation of the backbone of the DNA, modification or even absence of nucleotides (abasic sites), and hydantoin formation where cytosine or thymine bases are converted to uracil or hypoxanthine and react with carbonyl compounds to form hydantoin derivatives. As a result, these modifications, amplification of DNA through PCR becomes challenging. Another factor hindering this process is contamination of the ancient samples with present-day human DNA. Colder and drier conditions can slow down this degradation process (Briggs *et al.*, 2010; Dabney, Meyer and Pääbo, 2013).

The many years of effort to improve the ancient DNA sequencing technology resulted in the sequencing of thousands of genomes from long-deceased individuals. Starting 2010, the number of sequenced ancient human genomes has been growing exponentially. A compendium of all the available ancient and present-day human genotype data was created by the David

Reich lab reported that as of March 2023, 9,990 unique ancient genomes are available (Mallick and Reich, 2023). Most of these sequenced individuals lived in Europe, but many more are available from Africa, Asia, Oceania, and the Americas. Over 30 archaic hominins have also been sequenced, although due to their older age, most of them resulted in low-coverage genomes (Reich *et al.*, 2010; Hajdinjak *et al.*, 2018; Skov *et al.*, 2022). Only four of those archaic genomes are high- coverage: 3 Neanderthal and 1 Denisovan (Green et al., 2010; Mafessoni et al., 2020; M. Meyer et al., 2012; Prüfer et al., 2014, 2017). Additionally, the offspring of a Neanderthal and a Denisovan from the Denisova Cave has also been sequenced (Slon *et al.*, 2018).

1.6 Admixture between humans and archaic hominins

The question of whether humans and Neanderthals interbred sparked the curiosity of researchers even years before the genomes of Neanderthals became available (Currat and Excoffier, 2004; Hodgson and Disotell, 2008). The first genetic material from this archaic group came from mitochondrial DNA a decade before the sequencing of the first archaic (i.e. Neanderthal and Denisovan). Mitochondrial DNA from around 15 archaic individuals were sequenced from sites in Europe, Central Asia, and Siberia (Krings *et al.*, 1997, 1999, 2000; Ovchinnikov *et al.*, 2000; Schmitz *et al.*, 2002; Serre *et al.*, 2004; Beauval *et al.*, 2005; Lalueza-Fox *et al.*, 2005, 2006; Orlando *et al.*, 2005; Caramelli *et al.*, 2006; Krause *et al.*, 2007; Krause, Fu, *et al.*, 2010). Early analyses of Neanderthal mitochondrial DNA did not provide any evidence of admixture between the archaic hominins and humans (Krings *et al.*, 1999; Currat and Excoffier, 2004; Serre *et al.*, 2004; Hodgson and Disotell, 2008). However, the nuclear genome sequencing of the Altai Neanderthal, published in 2010, for the first time indicated allele sharing between humans and archaic hominins (Green *et al.*, 2010). This is evidenced by the presence of shared alleles between Eurasians and archaics, excluding sub-Sahara Africans. The time of the introgression events were estimated at around 2,000 generations ago, or ~55 kya (Sankararaman *et al.*, 2012; Fu *et al.*, 2014; Prüfer *et al.*, 2014). The total fraction of the Neanderthal genetic variation remaining in the human genome is estimated at ~40% (Vernot and Akey, 2014; Skov *et al.*, 2020), although the contribution to individual genomes is approximately 2% of the total variation in Eurasians (Green et al., 2010; M. Meyer et al., 2012; Prüfer et al., 2014; Sankararaman et al., 2016; Vernot et al., 2016; Wall et al., 2013). Additionally, Denisovan-human admixture contributed 2-4% of the variation in Southeast Asians individuals (Reich *et al.*, 2010; Mallick *et al.*, 2016; Sankararaman *et al.*, 2016; Vernot *et al.*, 2016; Browning *et al.*, 2018).

Several different methods were developed to detect introgression in the human genome. The D-statistic was developed following the sequencing of the first Neanderthal draft genome to detect the proportion of introgressed regions in the human genome (Green *et al.*, 2010). This method integrates the concept of ABBA BABA allele configurations. ABBA BABA is used to describe patterns of allele sharing between populations using data from two closely related human populations, an archaic population, and an outgroup, e.g. D(H1, H2, N, Chimp), with the null hypothesis that A did not share alleles with either H1 or H2. Under the ABBA scenario, H1 carries the outgroup (Chimp) allele and H2 carries the derived allele together with the archaic group (N). Under the BABA scenario, H1 and N carry the derived allele and H2 carries the same allele found in the outgroup. The test compares the total number of ABBA sites and BABA sites in a specific genomic region and calculates the difference between these two counts and normalized by the number of observations. The D-statistic was tested for robustness to different demographic models, as long as the individuals are mating randomly (Durand *et al.*, 2011). This method was applied in other studies to compare the sharing of Denisovan derived alleles between Eurasian and African populations and thus infer introgression (Reich *et al.*, 2010).

Other methods are based on Hidden Markov models (HMMs), where the observed state of the data is influenced by hidden states that can be modeled as a Markov process (Baum and Petrie, 1966; Baum *et al.*, 1970). These HMM-based methods use an archaic reference genome, and a low-admixture human reference genome to infer introgressed segments in Eurasian human genomes. The hidden state represents the ancestry of a haplotype from two possible states: archaic or non-archaic. One such approach was developed to infer Neanderthal introgression in Eurasians and Denisovan introgression in Asians and Oceanians (Prüfer *et al.*, 2014). This study was able to identify that the Neanderthal specimen (among the specimens available at that time) that is closest to the introgressing Neanderthal is the Mezmaiskaya Neanderthal from the Caucasus Mountains. It also revealed that approximately 2% of the genome in Eurasians is inherited from Neanderthals. diCal-admix (Steinrücken *et al.*, 2018) is an extension of diCal2 (Demographic Inference using Composite Approximate Likelihoods; (Steinrücken *et al.*, 2019), which infers ancient demographic histories from genome data using a Hidden Markov Model (HMM) framework. diCal-admix uses a hidden state to represent whether a target non-African genome coalesces with an African or archaic haplotype and at what point in time this coalescence event occurred. The probability of the hidden state is estimated using forward-backward simulations. Unlike the previous HMM-based methods, the HMM-based method developed by Skov (Skov *et al.*, 2018) is reference-free. Instead, it considers the excess of genomic variant density found in introgressed regions.

Statistical modeling methods have also been applied to identify introgression in the human genome. Conditional random fields (CRFs) are a type of classification technique used to predict labels in data based on observed patterns in the datasets that they are trained on (Lafferty, McCallum and Pereira, 2001). This method has been applied to infer the ancestral state of each allele in a set of haplotypes in Eurasians using data from an archaic and a low-admixture human reference genome (Sankararaman *et al.*, 2014). The CRF was trained on a demographic simulation. The method identified archaic haplotypes using data from Eurasian populations, Neanderthal genomes, and a reference African population (west-African Yoruba). This method searches for regions of the human genome where a derived allele in Neanderthals is present in non-Africans but absent in the African population, where the target haplotype in non-Africans is more diverged from the African haplotypes than the Neanderthal haplotype, and the length of the haplotype left by recombination agrees with the years after interbreeding (Sankararaman *et al.*, 2012). This method estimated with high confidence a 1.15% in Europeans and 1.38% in east-Asian Neanderthal ancestry. This method also identified an enrichment of Neanderthal ancestry in some regions of the human region, which can contain up to 64% Neanderthal ancestry in Europeans and 62% in east-Asians, suggesting that there was a process of natural selection on Neanderthal introgressed haplotypes. A modified version of this method was also applied to detect Denisovan ancestry in Southeast Asians (Sankararaman *et al.*, 2016).

The S* statistic is a linkage disequilibrium (LD) based method developed in 2006 and later modified in 2014 (Plagnol and Wall, 2006; Vernot and Akey, 2014; Simonti *et al.*, 2016). This approach does not require an archaic genome to detect introgression in the human genome. Instead, archaic introgression is identified by looking for the signatures that ancient admixture leaves in the genome. Recombination happening in Neanderthal introgressed segments since the time of introgression leave LD blocks approximately 50 kb in length (Sankararaman *et al.*, 2012). This results in high LD between archaic alleles that are not found in African human genomes. The linkage is analyzed in a pairwise manner, and assigning scores that rate the linkage level of the loci. Candidate introgressed haplotypes are detected by calculating S* in a sliding window-based method, and then the candidate windows are filtered by comparing the identified sequences to the reference Neanderthal genome and testing if they match significantly more than expected. The 2014 version of S* analyzes batches of 20 individuals, and the 2016 version analyzes only 1 individual simultaneously to avoid the effects of population structure. Building up on this method, the Sprime statistic doesn't use windows (increasing the power and performance of the method compared to windowed methods), and also accounts for the back-migration to Africa. The data used for this analysis are the genomes of a target population and an outgroup (Yoruba) expected to have received little admixture from archaics, and a linkage-disequilibrium genetic map (Browning *et al.*, 2018). Recently, Chen 2020 introduced a method, IBDmix (Chen *et al.*, 2020) that can detect introgressed haplotypes without the use of an unadmixed outgroup.

1.7 Adaptive Introgression

The time of introgression of Neanderthal DNA in the human genome is estimated to 60–50 ka (Plagnol and Wall, 2006; Green *et al.*, 2010; Fu *et al.*, 2014; Prüfer *et al.*, 2014, 2017; Sankararaman *et al.*, 2014; Vernot and Akey, 2014; Vernot *et al.*, 2016) and 44–54 kya for Denisovans (Sankararaman *et al.*, 2016). Although in present-day non-African humans, only approximately 2% of each genome originated in the Neanderthal line, this proportion has been shown to have been higher in earlier humans and decreasing over time, mostly nearby genic regions (Seguin-Orlando *et al.*, 2014; Fu *et al.*, 2016). For instance, the Oase 1 individual from Peştera cu Oase, Romania and dated to approximately 40 ka contains 6–9% Neanderthal ancestry (Trinkaus *et al.*, 2003; Fu *et al.*, 2015). Although this proportion is atypical to the proportion observed in other Late Pleistocene humans in Eurasia, due to the lack of samples of earliest modern humans in Eurasia, an inverse correlation has been observed between age and Neanderthal ancestry (Krause, Briggs, *et al.*, 2010; Fu *et al.*, 2013, 2014; Benazzi *et al.*, 2015; Yang *et al.*, 2017; Hublin, 2020; Prüfer *et al.*, 2021). However, this correlation was later shown to result from a bias in the application of the F statistics, and further modeling revealed a rapid reduction in archaic ancestry in the generations immediately following introgression (Petr *et al.*, 2019). This reduction in archaic ancestry has been attributed to a process of purifying selection where introgressed alleles that were deleterious in humans were purged from the gene pool (Harris and Nielsen, 2016; Juric, Aeschbacher and Coop, 2016). This process is also evidenced by the presence of archaic deserts in the genome depleted of archaic ancestry (Sankararaman *et al.*, 2014, 2016; Vernot and Akey, 2014; Vernot *et al.*, 2016) suggesting that some genomic regions had incompatibilities with archaic variants. Archaic alleles have also been shown to be downregulated in brain tissues (McCoy, Wakefield and Akey, 2017). However, some disease associated alleles have been identified, including systemic lupus erymatosus, addiction, Crohn’s disease, type-2 diabetes, actinic keratosis, mood disorders, and depression (Sankararaman *et* *al.*, 2014; Simonti *et al.*, 2016). Simulations show that Neanderthal introgression in individual genomes can decrease to 3% just after 20 generations from the admixture event, and the purging of deleterious alleles slows down when the fraction of admixture is balanced in the population, and the load created by these alleles are more likely to affect polygenic traits (Harris and Nielsen, 2016).

However, among the remaining archaic alleles in the human genome, there is strong evidence of a beneficial contribution to human adaptation (Huerta-Sánchez *et al.*, 2014; Racimo *et al.*, 2015, 2017; Dannemann, Andrés and Kelso, 2016; Racimo, Marnetto and Huerta-Sánchez, 2017). Adaptively introgressed alleles associated with immune response may have helped humans adapt to new pathogens (Abi-Rached *et al.*, 2011; Dannemann, Andrés and Kelso, 2016). Other introgressed alleles associated with skin phenotypes may have been adaptively involved in decreased intake of sunlight and UV radiation (Vernot and Akey, 2014). An introgressed allele associated with lower expression of hemoglobin levels as an adaptation to living in high altitude environments (Huerta-Sánchez *et al.*, 2014).

Despite all the previous work aimed at understanding the adaptive contribution of archaic introgression in the human genome, the field still requires more comprehensive descriptions of this contribution. Specifically, there is a need for further investigation into the molecular mechanisms through which archaic alleles contributed to the adaptive ability of humans to migrate and inhabit nearly every region of the planet, even with the new environmental challenges posed by these environments.

1.8 Long-term balancing selection

Introgressed variants are not the only source of ancient potentially adaptive variants in the human genome. One of the first descriptions of adaptation in humans was reported in the mid 20th century, when Haldane hypothesized that the sickle cell anemia red blood cells in heterozygous individuals protected them from malaria infections in countries where the infection was prevalent (Haldane, 1949). A few years later Allison published results showing that African populations where malaria was endemic had an increased incidence by 10% of the trait (Allison, 1954). Malaria is a highly deadly disease transmitted by various species of protozoans in the *Plasmodium* genus through the bite of *Anopheles* mosquito. The disease has affected human populations over millennia (World Health Organization (WHO), 2022). Today we understand that the sickle cell mutation in the hemoglobin beta gene (HBB) follows allele frequencies that are higher than in other populations (Hedrick, 2011; Piel *et al.*, 2013; Serjeant, 2013). Sickle cell anemia is a disease where the red blood cells are deformed because of the abnormal hemoglobin that they carry. The organism is thus unable to carry oxygen as efficiently. The disease is deadly when an individual carries two mutated alleles, but mild when the locus is heterozygous. Alterations in the biochemistry of the sickle red blood cell causes it to become a suboptimal environment for the *Plasmodium* (Elguero et al., 2015; Gong et al., 2013; Jinam et al., 2017; Luzzatto, 2012; Taylor & Fairhurst, 2014), providing some protection against the infection. This process is an example of balancing selection.

Balancing selection maintains intermediate levels of genetic variation at a given locus in the individuals of a population. In this case, maintaining diversity at a functional locus is beneficial for the population, thus an allele doesn't become fixed. The idea was introduced by Theodosius Dobzhansky in a study of chromosomal inversion polymorphism in *Drosophila* (Dobzhansky, 1950). Under this type of selection, maintaining variation at a given locus confers higher fitness, compared to directional selection. The example of the sickle allele protection against malaria represents a type of balancing selection called heterozygote advantage, also called overdominance (Wright, 1931; Lewontin, 1964). In this case, the fitness obtained from the combination of two different alleles is greater than the fitness of either of those alleles in a homozygous form. In addition to heterozygote advantage, this natural selection process can be the result of several other mechanisms, including frequency-dependent selection (Kojima and Yarbrough, 1967; Kojima and Tobari, 1968; Huang, Singh and Kojima, 1971; Hedrick, 1972; Gromko, 1977) and spatial/temporal heterogeneity or periodical environmental shifts (Howard Levene, 1953; Gillespie and Langley, 1974; Bergland *et al.*, 2014). These natural pressures on the genome generate signatures that can be used to detect balancing selection. These signatures become more evident when this process has been maintained at a locus over long periods of time, namely long-term balancing selection (Bitarello *et al.*, 2018).

Under balancing selection, linkage disequilibrium (LD) can create an accumulation of higher-than-expected variation around a specific locus (Hudson and Kaplan, 1988; Charlesworth, Nordborg and Charlesworth, 1997; Takahata and Satta, 1998). This process generates a genetic load that would not be possible under a neutrally evolving scenario (Haldane, 1957; Charlesworth, Nordborg and Charlesworth, 1997; Uyenoyama, 1997, 2005; Le Veve *et al.*, 2023), Leach 1986, Lenz 2016, Tezenas 2023). For example, the human leukocyte antigen (HLA; MHC, the major histocompatibility complex in mammals), which is a classic example region under balancing selection and associated with the immune system in humans (Hedrick and Thomson, 1983; Hughes and Nei, 1988; Takahata and Nei, 1990; Black and Hedrick, 1997), contains many linked deleterious variants that are associated with many diseases (Lenz *et al.*, 2016; Matzaraki *et al.*, 2017). When this process has been maintained over long time it can have an effect on the coalescence time, making it deeper in the balanced locus than the species coalescence time (Gillespie and Langley, 1974; Wiuf *et al.*, 2004).

Although examples of this type of natural selection have been described for decades, it has been understudied. During the late 20th century, statistical methods to detect signatures of natural selection in the genome were applied to detect balancing selection. Many of these methods detect deviations from the neutral theory of molecular evolution in DNA sequence data. The neutral theory states that most genetic variation has no effect on fitness, so the prevalence in the population changes over time according to genetic drift (Kimura, 1968; Hellmann *et al.*, 2003). The HKA test compares the density of polymorphism and divergence at a locus to the neutral expectation using a chi-square test (Hudson, Kreitman and Aguadé, 1987; Wright and Charlesworth, 2004). Tajima's D evaluates the site frequency spectrum (SFS) at a locus and looks for deviations from the neutral SFS pattern (Tajima, 1989). When applying Tajima's D, values of -2 or lower suggest an excess of rare alleles (positive selection or selective sweep) and values higher than 2 suggest an excess of common alleles (balancing selection) (Tajima, 1989). Fu & Li's D and F can also detect over-representation of common alleles and looks for the amount of derived singleton mutations and the amount of derived nucleotide variants (Y.-X. Fu and Li 1993). Another statistic, Watterson's θw estimator (Watterson, 1975) looks for loci with many variants independently of their allele frequency. It estimates the population mutation rate or genetic diversity based on the number of segregating sites in a region. When using any of these statistics, it is important to consider demographic history (e.g. bottlenecks), which are confounders of balancing selection.

In the past two decades, new methods were developed specifically to identify balancing selection in the genome. These methods aim to identify signatures of balancing selection, which include an excess of heterozygosity, an excess of intermediate frequency variation (e.g. analysis of the SFS), and an excess of nucleotide diversity at a specific locus. The first balancing selection specific method, built upon the HKA test to detect specific patterns of genetic variation, is composed of two statistical tests: HKAlow and MWUhigh (Andrés *et al.*, 2009). HKAlow detects increased coalescence times and excess of polymorphism in LD loci. MWUhigh detects long-term balancing selection at intermediate frequencies in cases of overdominance and frequency-dependent selection. Another study identified regions under long-term balancing selection in humans by using trans-species polymorphisms as a proxy (Leffler *et al.*, 2013). To do this, they identified haplotypes shared between humans and chimpanzees spanning a maximum length of 4 kb. To determine the maximum distance at which balancing selection maintains two shared SNPs in high LD, the team used the Kaplan-Darden-Hudson coalescent model to identify a segment that coalesces before the speciation event between the two species (Hudson and Kaplan, 1988; Kaplan, Hudson and Iizuka, 1988; Leffler *et al.*, 2013). The T1 and T2 statistics also build upon the Kaplan-Darden-Hudson coalescent model (Hudson and Kaplan, 1988; Kaplan, Hudson and Iizuka, 1988; DeGiorgio, Lohmueller and Nielsen, 2014). The method calculates the composite likelihood that a locus is under balancing selection by: 1) estimating the neutral expectation in coalescence time between the target species and an outgroup, 2) calculating the expected tree length and height at a genomic locus, 3) creating a spatial distribution of polymorphisms flanking a target site, 4) calculating the probability of observing either a polymorphism or a substitution at each site flanking a target site as a function of the distance of the neutral site to the balanced polymorphism, adjusted based on the distance to the target site, and 5) multiplying all the probabilities in 4. Similar to the HKA test (Hudson, Kreitman and Aguadé, 1987; Wright and Charlesworth, 2004), T1 and T2 identify regions containing an excess of polymorphisms relative to fixed sites. T2 additionally measures allele frequencies at that locus and has more power than T1.

Neutral polymorphisms in balancing selection loci tend to exist in frequencies similar to the balanced polymorphism due to LD. After many generations and recombination events, the frequencies of the neutral sites will slowly drift apart (Hey, 1991; Charlesworth, 2006). The β summary statistic uses this signature to detect regions under balancing selection in humans: it looks at the variants in each region to identify loci containing an excess of variants with similar allele frequency to the target variant (Siewert and Voight, 2017). This method uses a weighted sum to score genomic regions composed of a putative polymorphism under balancing selection and flanking polymorphisms in a window of 1kb. Regions containing an overrepresentation of flanking SNPs at similar allele frequency to the target SNP are scored higher.

The non-central deviation (NCD) statistic looks for signatures of long-term balancing selection in the form of increased polymorphic sites and an excess of alleles at intermediate frequencies in the LD region flanking a genomic locus (Bitarello *et al.*, 2018). Linkage disequilibrium shapes the flanking region around a SNP under balancing selection, leaving an increased ratio of polymorphic to divergent sites and an overrepresentation of intermediate frequency alleles. NCD detects deviations from neutral expectations in the SFS using polymorphisms (NCD1) or substitutions compared to an outgroup (NCD2). NCD2 improves power.

BetaScan2 (Siewert and Voight, 2020) is another summary statistics method that detects long-term balancing selection from substitutions and polymorphisms using the β2 statistic. β2 builds up on the β statistic, which uses 1000 Genomes Project data to detect genomic regions containing an excess of variants at similar frequencies to a target of balancing selection (Siewert and Voight, 2017). β2 has higher power than β1 because it uses an estimator of the number of substitutions compared to an outgroup. β2 also adds a normalization of β by deriving the variance of the statistics, allowing the scores to be compared across a range of parameters which can affect its distribution.

Most of the long-term balancing selection regions identified and studied are associated with immune-related function. Other targets of long-term balancing selection have been detected in genes with functions such as sleep-related phenotypes (MYRIP), corneal astigmatism (BICC1), diabetes (WFS1), and brain-related biology (CADM2) (Siewert and Voight, 2017). These associations have not been explored in detailed, and the driver of long-term balancing selection could be related to other traits different from the main described function of the gene.

1.9 Tools used for the annotation of functional effects of genomic loci

In the previous sections, I described the process of acquiring genomes to extract genotype data and how this variation interacts with the process of adaptation. However, the tools that have been described so far are not intended for understanding the phenotypic or functional role of these genomic sites. Many of these methods work at the single nucleotide polymorphism (SNP) level. Over the past few decades, many methods and tools have been developed with the aim to annotate genomic loci and understand their function at the molecular and phenotypic level. Some of these tools take advantage of the new capacity to sequence large cohorts of individuals and perform statistical analyses on the genotypes of many loci to identify phenotypic functions. Other methods take advantage of functional assays to annotate genomic loci and to interpret their functional impact.

Genome-wide association studies (GWAS) involve scanning the genomes of large numbers of individuals to identify genetic variations associated with particular traits or diseases, leveraging SNPs as markers for genetic differences. A caveat to this approach is that the contribution to phenotype of many of the identified genetic variants is polygenic, they only have a small effect on the trait or disease. One of the largest GWAS available is the GWAS Catalog (Sollis *et al.*, 2023). Phenome-wide association studies (PheWAS) were later developed with a focus on the clinical phenotype level and integrate electronic health records (EHR) and other medical databases to find a link between genetic variants and a range of clinical phenotypes. One of the largest PheWAS database is the UKBiobank (Ollier, Sprosen and Peakman, 2005).

Expression quantitative trait loci (eQTLs) are genomic loci where the alleles associate with the level of gene expression. The first genome-wide eQTL map was published in 2003 (Cheung *et al.*, 2003). The Genotype-Tissue Expression (GTEx) Project, initiated in 2010 by the National Institutes of Health (NIH), is a large-scale project that identified eQTLs across ~50 human tissues from hundreds of individuals (Carithers *et al.*, 2015).

Tools like the Variant effect predictor (VEP) and ANNOVAR aim to understand the association of genomic loci at the functional molecular level (Wang, Li and Hakonarson, 2010; McLaren *et al.*, 2016). These aim to identify molecular descriptions for each site such as their impact on protein structure and function, their potential effect on regulatory regions of genes, whether it affects the protein-coding sequence, the type of regulatory feature, and their contribution to disease.

Using a combination of these tools it is now possible to describe the function of genomic regions. However, when trying to understand the potential drivers for natural selection on specific loci, it is important to understand that genomic loci are often associated with many different functions.

1.10 Overview of this dissertation

Thus far, many studies have focused on better understanding the role of genetic variation on human evolution in response to various environmental conditions and other adaptive pressures, as described above. In this dissertation I aim to contribute to the gaps in knowledge found in two different natural selection scenarios: 1) long-term balancing selection on old variation present in the closest living relative to humans, the chimpanzees, and 2) variation that originated in Neanderthals, another closely related group, and was introduced into humans through admixture, potentially enabling adaptation to new environments.

In Chapter 2, we focus on exploring the functions and potential adaptive roles of LTBS in humans. Specifically, we analyzed non-coding regions of the genome containing multiple shared polymorphisms (SPs) between humans and chimpanzees. By integrating diverse genomic annotations, such as functional genomics assays and association studies, this project aims to elucidate the functions and potential adaptive roles of these shared polymorphisms beyond the immune system. These findings shed light on the genetic basis of diversity in humans. This investigation discovered a wide range of traits potentially influenced by LTBS, including immune system phenotypes, risk-taking behavior, cognitive performance, body size, and more. By shedding light on the functions of non-coding regions under LTBS, this research contributes to our understanding of genetic diversity and behavioral diversity in humans.

In Chapter 3, we aimed to evaluate the contribution of archaic introgressed alleles to the circadian biology of Eurasian humans. We leveraged the available genome sequence data from different human populations and from archaic hominins to explore the contribution of archaic introgression to circadian biology and chronotype in Eurasians. To achieve this objective, we investigated differences in circadian gene sequences, splicing, and regulation between archaic hominins and modern humans. We also studied the direction of effect of these introgressed alleles on human chronotype. We also analyzed results from newly developed machine learning methods used to identify adaptive introgression in human genomes. These findings contribute to the growing knowledge of specific phenotypes that were influenced by archaic alleles and became adaptive in environments that were new to migrating humans after the Middle-Paleolithic.

Chapter 2

DIVERSE FUNCTIONS ASSOCIATE WITH NON-CODING POLYMORPHISMS SHARED BETWEEN HUMANS AND CHIMPANZEES*

2.1 Introduction

The interaction between populations and environments is dynamic. Over time, allele frequencies in a population shift due to drift and adaptive responses to specific environmental pressures. Most genetic variants are short-lived compared to the timescale of species. But on rare occasions variants persistently segregate at intermediate frequencies for millions of years, sometimes pre-dating the most recent common ancestor (MRCA) between two sister species (Leffler *et al.*, 2013; DeGiorgio, Lohmueller and Nielsen, 2014; Teixeira *et al.*, 2015; Siewert and Voight, 2017; Bitarello *et al.*, 2018; Cheng and DeGiorgio, 2019). These trans-species polymorphisms are often a sign of genomic regions under long-term balancing selection (LTBS). Over time, instances of LTBS leave signatures in the genome that differentiate them from those under other forms of selection (Leffler *et al.*, 2013; Key *et al.*, 2014; Siewert and Voight, 2017; Bitarello *et al.*, 2018), such as maintenance of alleles at intermediate frequency alleles than expected by chance, increased level of neutral variation near the target site, and deep coalescence times.

Several instances of LTBS regions have been observed in humans and other primates, mostly within the major histocompatibility complex (MHC) or the ABO blood group locus. For example, the MHC, or human leukocyte antigen (HLA) system in humans, is a family of varied proteins expressed on the cell surface with essential functions in adaptive immune response and regulation. Balancing selection on different components of the HLA region dates to the common ancestor between chimpanzees and humans (Lawlor *et al.*, 1988; Mayer *et al.*, 1988) (Azevedo *et al.*, 2015). Similarly, the ABO gene has three alleles, and its variants lead to different blood cell antigens, or lack of thereof, on the surface of the cell. Variation in this group could have a benefit in the immune response to pathogens, and balanced polymorphisms at this locus are present in gorillas, orangutans, and humans, and thus likely date back to their last common ancestor (Ségurel *et al.*, 2012). Several other immune-related genes show LTBS between humans and other primates, e.g.: *TRIM5*, a RING finger protein 88 (Cagliani *et al.*, 2010; Battivelli *et al.*, 2011; Ganser-Pornillos and Pornillos, 2019), and *ZC3HAV1*, a zinc finger CCCH-type antiviral protein 1 (Cagliani *et al.*, 2012; Mao *et al.*, 2013; Todorova, Bock and Chang, 2015; De Filippo *et al.*, 2016). These genes have important roles in host/pathogen response through inhibition of virus replication.

The high allelic variation maintained by balancing selection at a locus can also enable adaptation to new environments. For example, some variants found under balancing selection in African and ancestral human populations have experienced directional selection in non-African populations (European and Asian), with one allele becoming predominant in the population (De Filippo *et al.*, 2016). This suggests the adaptive potential of the variation maintained under balancing selection; however, in some cases the adaptive variants themselves may have hitchhiked with those under LTBS.

Recent studies have developed statistical methods to identify instances of balancing selection in genome-wide data (DeGiorgio, Lohmueller and Nielsen, 2014; Siewert and Voight, 2017, 2020; Bitarello *et al.*, 2018; Cheng and DeGiorgio, 2019). Some have focused on detecting LTBS using trans-species data, while others have considered balancing selection over shorter timescales based on single-species data. For example, DeGiorgio (DeGiorgio, Lohmueller and Nielsen, 2014) developed likelihood-ratio tests (T_1 and T_2) based on computing probabilities of polymorphism and substitution under LTBS based on inter-species coalescent modeling to test the spatial distribution of polymorphisms and mutations around genomic sites. With this method they identified balancing selection on HLA regions, but also in a gene that had no previous associations with balancing selection, *FANK1*, which is involved in the suppression of apoptosis during/after the process of meiosis. They also found enrichment for signals in genes with other functions: cell adhesion, membrane protein activity, and components of membranes. A more recent study (Cheng and DeGiorgio, 2019) expanded the T_2 method to seek trans-species balancing selection without direct consideration of trans-species polymorphism and identified a handful of additional LTBS candidates. Bitarello et al. (Bitarello *et al.*, 2018) developed Non-central Deviation (NCD) statistics that quantify the deviation of the local site frequency spectrum (SFS) under balancing selection from neutral expectations. The statistic identifies genomic windows with variants at intermediate frequencies and higher than

*This chapter has been previously published in Velazquez-Arcelay et al. 2022. BMC Ecol Evo.

expected levels of variation as a signature of balancing selection (Andrés *et al.*, 2009). Applying the statistics to African and European 1000 Genomes populations, they found thousands of candidates for balancing selection in humans. They also showed varying directional selection in different populations, providing evidence for the adaptive potential of regions under balancing selection. Siewert & Voight (Siewert and Voight, 2017) developed ß, a summary statistic for detecting genomic windows with clusters of intermediate frequency alleles suggestive of balancing selection. They also recently updated the ß statistic to consider both polymorphism and substitution data (Siewert and Voight, 2020). Among the highest scoring windows in these two analyses, they highlighted three genes (*CADM2*, *WFS1*, and *ACSBG2*) with functions outside the immune system.

Shared polymorphisms (SPs) between species, especially when more than one falls on a haplotype, suggest the action of LTBS. For example, Leffler et al. (Leffler *et al.*, 2013) compared polymorphisms across the genome in Yoruba individuals from the 1000 Genomes Project to those found in Western chimpanzees sequenced by the PanMap Project. They identified more than 100 non-coding haplotypes with multiple SPs within 4 kilobases (kb) and in high LD as candidates for LTBS. However, sequencing errors and regions with high mutation rates can create patterns that can be mistaken for LTBS. Further modeling has shown that it is unlikely to observe haplotypes with more than two TSPs in close proximity by chance without balancing selection (Gao, Przeworski and Sella, 2015; Cheng and DeGiorgio, 2019).

Despite the importance and prevalence of balancing selection, most of the non-coding haplotypes bearing potential signatures of LTBS (e.g., multiple SPs), have not been functionally characterized. Here, we focus on a high confidence subset of the non-coding SPs identified by Leffler et al. (Leffler *et al.*, 2013). Determining the candidate functional roles of these SPs in human adaptation and health would deepen our understanding of the dynamics of balancing and positive selection and their roles in adaptation to new environments.

We identify potential functions associated with SP regions in humans by applying several genome-wide functional annotations and association tests. Our results identify diverse functions, including effects unrelated to the immune system, that may have been targets of balancing selection on the human and chimpanzee lineages.

2.2 Human-chimpanzee shared SNPs

We consider 125 human genomic regions containing multiple variants segregating in both humans and chimpanzees in close proximity and in high LD (Leffler *et al.*, 2013). The set was defined based on identifying groups of human-chimp shared-polymorphisms (SPs) within 4 kb of each other outside the major histocompatibility (MHC) locus. Based on coalescent theory, this pattern is unlikely to result from neutral processes, (Ségurel *et al.*, 2012; Leffler *et al.*, 2013) and are thus candidates for LTBS (Appendix 5.1.1). However, these criteria alone are insufficient to guarantee that the SPs are the result of identity-by-descent and driven by LTBS (Gao, Przeworski and Sella, 2015).

To identify regions with stronger evidence of balancing selection, we consider two additional recent genome-wide balancing selection scans (Bitarello *et al.*, 2018; Siewert and Voight, 2020) and additional evidence of identity-by-descent (Figure 1). The first scan is based on NCD, a balancing selection detection statistic that uses the allele frequency spectrum to find regions enriched for intermediate frequency alleles (Andrés *et al.*, 2009). The second is based on BetaScan2, which detects balancing selection by identifying deviation from neutrality in the vicinity of a haplotype from variance in substitutions and mutation rate. We apply a filter based on regions containing evidence in NCD from at least one population or regions containing at least one SP with a BetaScan2 score of 2.0 or higher. Of the initial set of 125 candidate haplotypes, 60 were highlighted in these recent balancing selection scans. We refer to the 133 variants on these haplotypes as candidate balanced shared polymorphisms (cbSPs). Next, to identify variants with the strongest evidence of LTBS, we further filtered these regions based on additional evidence of human-chimp identity-by-descent to create set of candidate trans-species polymorphisms (ctSPs). For this set, we required the candidate haplotypes additionally to have either extremely ancient times to most recent common ancestor (TMRCA) as estimated by ARGweaver(Rasmussen *et al.*, 2014) (>140,000 generations ago) or more than 3 SPs per candidate haplotype. This resulted in 19 haplotypes with 51 ctSPs. In summary, 60 out of the original 125 candidate regions show evidence of balancing selection from at least one of BetaScan2 or NCD (Methods), and 19 of these show additional evidence of identity by descent (Figure 1).

In the following, we analyze functional annotations and associations for both cbSPs and ctSPs. In some analyses, to capture associations tagged by variants in high linkage disequilibrium (LD) with cbSPs, we also considered potential tag SNPs in high LD ($R^2 \geq 0.8$) in African, European, or East Asian populations from the 1000 Genomes Project. This LD-expanded set for cbSPs includes 6,171 variants across the 60 regions (Appendix 5.1.2). By expanding to include variants in high LD, we capture additional associations, but may also identify functions unrelated to balancing selection; thus, we report results on both sets.

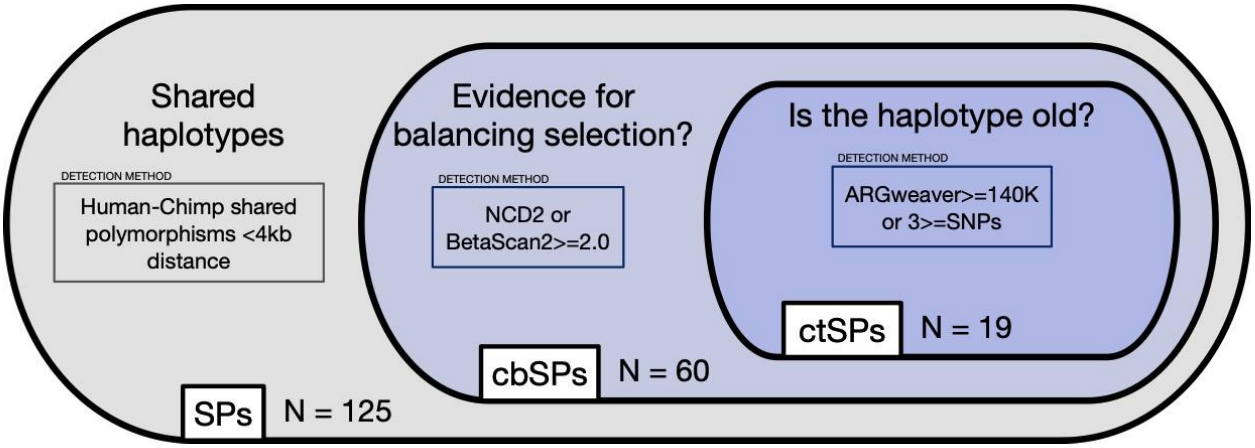


Figure 2.1. Schematic of the criteria for identifying the SP sets used in this study. A previous study(Leffler *et al.*, 2013) reported a set of 125 candidate regions with two or more non-coding human-chimp shared polymorphisms (SP) within 4 kb. We refined this set based on several additional lines of evidence. First, we considered scores from two balancing selection statistics (NCD and BetaScan2) to create a set of 60 haplotypes with 133 candidate balanced SPs (cbSP). We consider regions with evidence for balancing selection in at least one population from NCD, or regions containing variants with BetaScan2 scores equal or higher than 2.0. We further filtered this set to the 19 haplotypes additionally predicted to be at least 140,000 generations old by ARGweaver or contain at least 3 SPs within 4 kb. These haplotypes include 51 candidate trans-species SP (ctSP) with the highest likelihood of LTBSs.

2.3 Shared polymorphisms overlap diverse functional annotations

We intersected the cbSPs with diverse lines of functional evidence from large-scale genomic studies, including genome-wide functional genomics assays, eQTL, GWAS, and PheWAS. We found at least one functional annotation for 98% (59 of 60) of the cbSP regions and all of the ctSP regions, covering 77 SPs and 772 LD SNPs (Figure 2). Limiting only to the SPs themselves, we found annotations for 68% (41 of 60) of cbSP regions and 84% (16 of 19) of ctSP regions. Here, we provide an overview of the overlap with these annotations. In future sections, we provide details about each of these annotations. Variants in 93% (56 out of 60) of regions overlap annotated gene regulatory regions. This includes 23 cbSPs and 599 LD variants. We also found 64 cbSPs across 34 regions with evidence of being expression quantitative trait loci (eQTL) across 48 tissues. We found genome-wide significant associations with phenotypes in available genome- or phenome-wide association studies in 32% of the LD expanded regions (19 out of 60; 14 GWAS Catalog and 11 UK Biobank from geneAtlas and NealeLab).

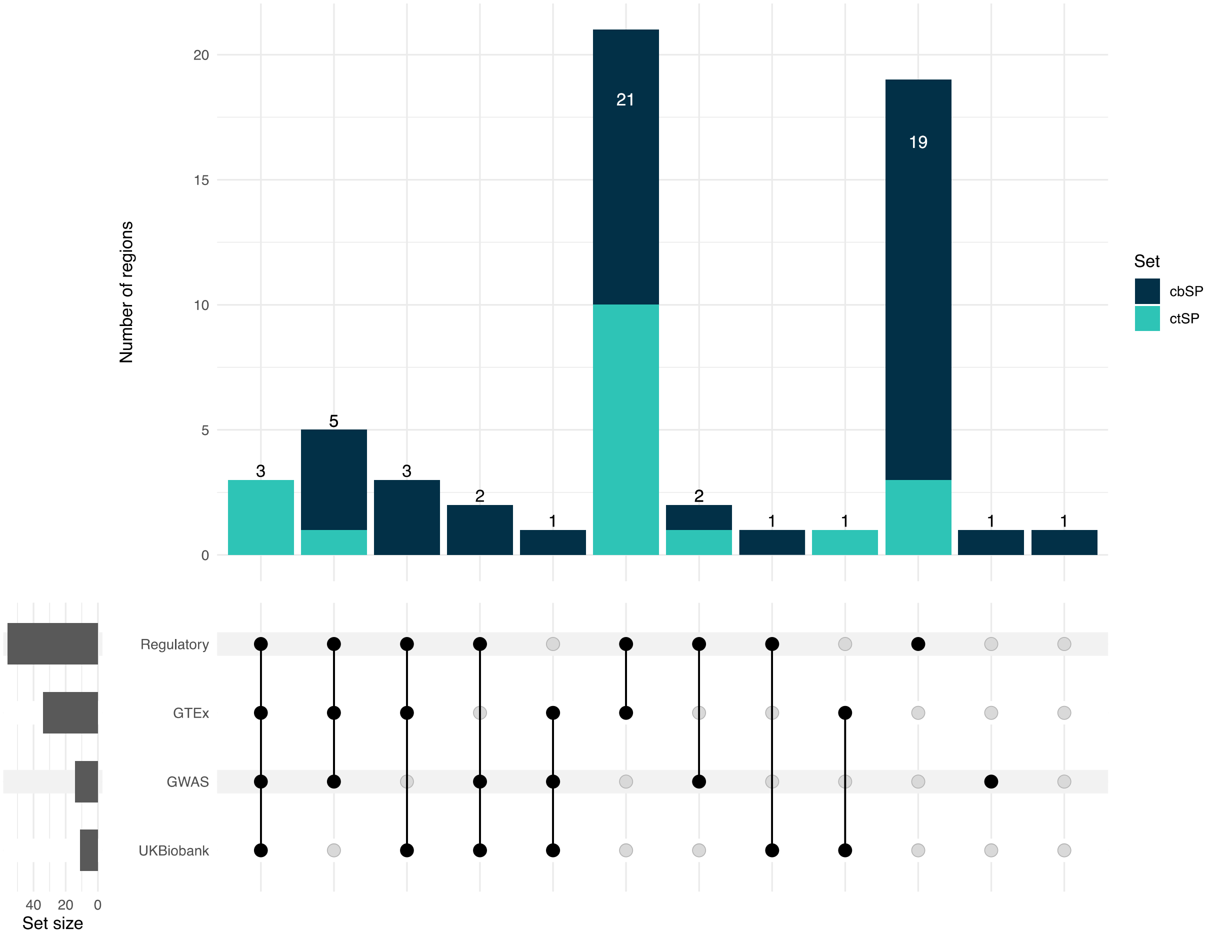


Figure 2.2. Functional annotations available for the expanded cbSP regions. Summary of the annotations of each type available for cbSP regions, including tagging variants in high LD with cbSPs. A total of 59 out of 60 cbSP regions contain at least one line of functional evidence. The analysis from GWAS, UK Biobank, and regulatory function include annotations for SNPs that are in high LD (0.80 R^2) with the cbSP set. The UKBiobank set included analysis from geneAtlas and the Neale Lab set. Associations for the highest confidence subset (ctSP regions) are shown in aqua.

2.4 Evidence of gene regulatory function for SPs

We hypothesized that many of the non-coding SPs in our set perform gene regulatory functions. To evaluate this possibility, we intersected the cbSPs and variants in high LD with maps of functional regulatory regions from the Ensembl regulatory build (Zerbino *et al.*, 2015). We found 23 cbSPs with regulatory annotations and additionally 599 LD variants in 56 cbSP regions. These include variants in CTCF binding sites, open chromatin regions, promoter flanking regions, enhancers, promoters, and known TF binding sites. We also tested cbSP regions for enrichment in any specific types of regulatory regions. We compared the observed overlap between cbSP regions and each type of regulatory annotation to the distribution of overlaps expected if cbSP regions were randomly distributed across the genome. We shuffled the cbSP regions 1,000 times maintaining their length and chromosome distributions and avoiding genome assembly gaps, ENCODE blacklist regions, and the MHC locus. We compared the number of overlaps observed with regulatory elements with the number from each random permutation (Appendix 5.1.3). cbSPs showed more overlap with enhancer and promoter elements than expected, but this was not significant, perhaps due to the small sample size.

Overlap of a variant with a regulatory annotation does not necessarily imply a regulatory function. To consider additional evidence of regulatory function, we examined eQTL in GTEx from 50 tissues for overlap with cbSPs. At least one eQTL was found for 34 of the regions (57%). Among these 34 regions, 64 cbSPs are themselves eQTL in 48 tissues. We tested for enrichment of eQTL in cbSPs compared to the background across all genomic regions and found enrichment for eQTL activity in a diversity of GTEx tissues, including liver, whole blood, skin, and pancreas (Figure 3).

We found diverse gene ontology (GO) terms among the genes influenced by cbSP eQTL, but no individual terms remained significant after multiple testing correction. These results suggest that the targets of balancing selection in these regions may have functions in gene regulation across diverse tissues beyond the immune system.

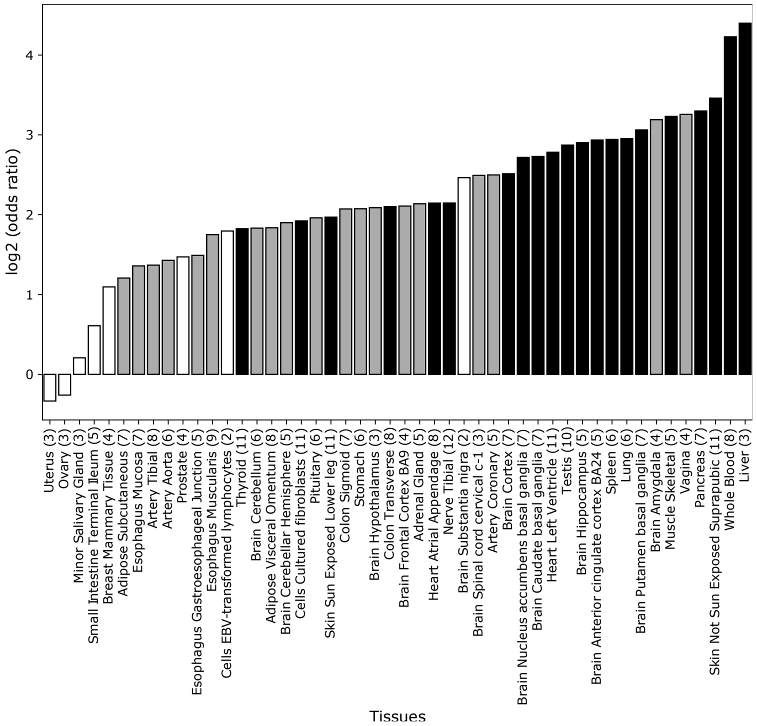


Figure 2.3. cbSPs are eQTLs in diverse tissues. cbSP regions are enriched for eQTL activity in many tissues compared to genomic background levels. Statistical significance is represented by black, gray, and white bars, where black indicates significance at the Bonferroni correction threshold, gray significance at $p < 0.05$, and white is not significant. The number of eQTL in the cbSP set for each tissue are given in parentheses following the tissue names. Tissues that are not present had a count of 0 eQTL (i.e. Kidney Cortex).

2.5 Genome-wide association studies link cbSPs to traits

Genome-wide association studies have identified thousands of associations between genetic variants and human traits. We intersected the cbSP regions with associations reported in the GWAS Catalog (downloaded 2021/12), which is composed of over 170,000 associations in 4,070 terms. Since cbSPs themselves were not always directly tested in GWAS studies, we also include genome-wide significant (p <= 5E-8) associations with the tag variants in high LD with SPs. We found significant associations for 52 different variants (Figure 4A). Among the functional associations we found immunological functions, hematological/blood measurements, and anthropometric traits. The associations with immune traits were expected given the results of previous balancing selection studies and the few well-characterized instances of LTBS. We identified many variants in LD with cbSPs that are associated with blood measurement phenotypes and diseases related to immune response. These traits include ulcerative colitis and other chronic inflammatory diseases (chr2 near cbSPs rs13426764/rs11694806).

We also found many neurological and behavior-related associations among cbSP region variants. These traits include cognitive performance (rs13426764 and rs11694806 on chromosome 2 and rs9869178/rs2118072 on chromosome 3), alcohol and smoking status (alcohol use: chromosome 16 near rs9933768 and rs57790054; smoking: chromosome 2 near rs13426764 and rs11694806), risky behavior (automobile speeding propensity: chromosome 3, rs9869178/rs2118072), experiencing mood swings (chromosome 2 near rs13426764 and rs11694806), insomnia, neuroticism, sun-seeking behavior, and age at first sexual intercourse. In addition to the immune response and neurological categories, we observed associations in reproductive traits (polycystic ovary syndrome, testosterone levels), urate levels, pancreatic cancer, and gut microbiota. An enrichment analysis found significant results for GWAS categories including blood and immune related traits, uric acid levels (including urate and gout), cognitive performance measurements (intelligence, educational attainment, math ability), smoking status, and gut microbiome measurement (Appendix 5.1.4). We discuss several of these associations in more detail in following sections.

2.6 Phenome-wide association studies link cbSPs to additional diverse traits

The growth of biobanks with linked genetic and phenotypic data has enabled the testing of the association of genetic variants with diverse traits within a single cohort. This PheWAS approach enables exploration of the functional and potentially pleiotropic effects of variants of interest (Bush, Oetjens and Crawford, 2016). Using published associations from the UK Biobank (geneAtlas and NealeLab), we analyzed the association of cbSPs with over a thousand traits; all 60 of the regions were tested. Overall, we found that 150 different variants in 11 regions had at least one genome-wide significant association ($P < 1E-8$, Figure 4B). Though testing different phenotypes than the GWAS, these associations were qualitatively similar to the GWAS results, in that blood and immune system phenotypes had many associations with cbSPs, but the cbSPs were also associated with a more diverse set of phenotypes. We found associations in categories of blood assays, body measurement, and lifestyle and environment. Among the observed associations we found, for example: hair color, standing height, number of days/week walked 10+ minutes, and 28 variants associated with alcohol intake frequency.

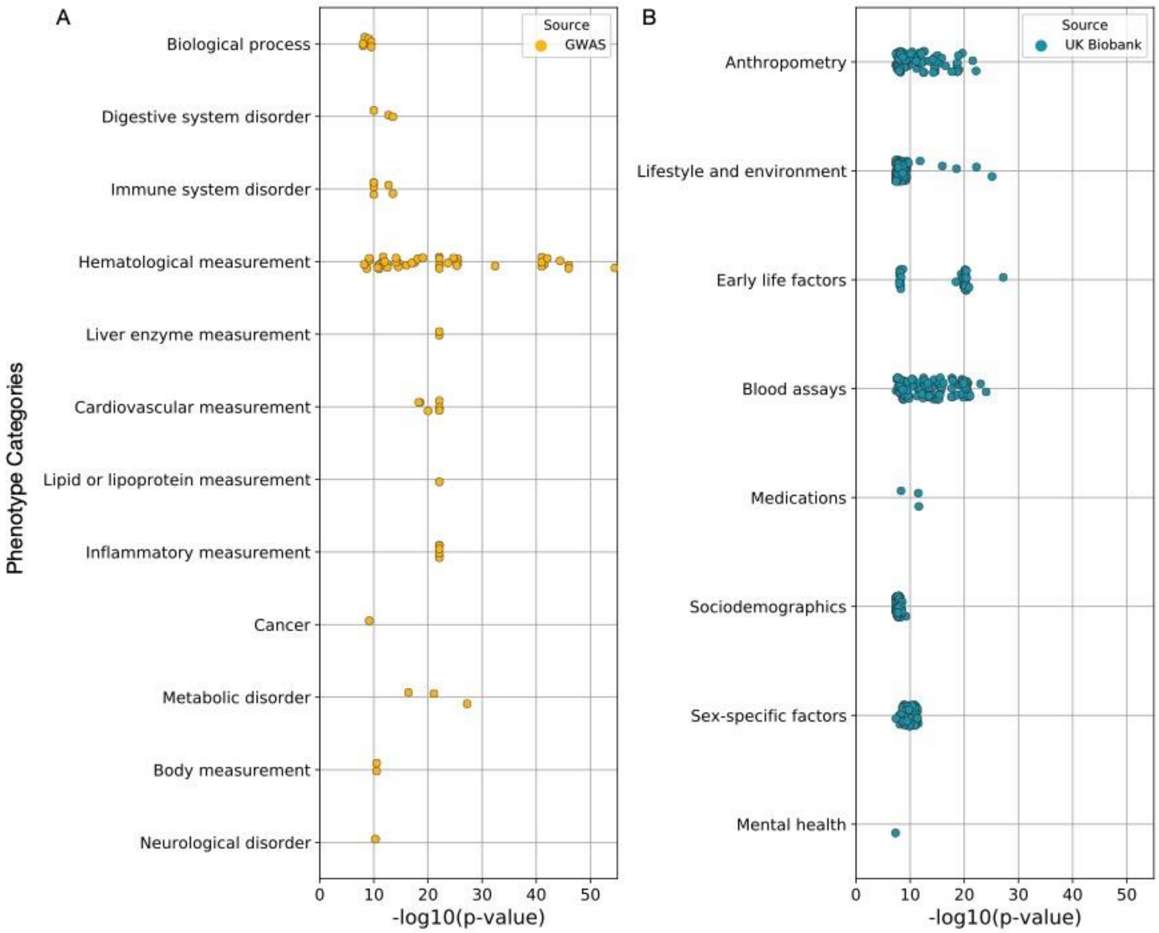


Figure 2.4. Genome- and phenome-wide association studies link cbSPs to diverse traits. **A)** Genome-wide significant (P < 1E-8) associations from the GWAS Catalog and **B)** PheWAS over the UK Biobank (from the geneAtlas and NealeLab(Watanabe *et al.*, 2019). Each dot represents an association between a cbSP region and a trait. Many immune-related traits (under immune system disorder, blood assays, and other measurements) are associated with cbSPs, but there are also associations with a wider variety of phenotypes including lifestyle and environment, neurological traits, and cognitive performance. Since few cbSPs themselves were directly tested in GWAS, we include GWAS Catalog associations with tag variants in high LD (r^2 > 0.8) with the cbSPs. We also observed associations in the "other measurements" and "other disease" parent categories, which include miscellaneous measurements and traits that did not fit in the listed categories. For the most enriched GWAS categories, see Appendix 5.1.4.

2.7 Illustrative examples of diverse functions associated with cbSP regions

Integrating the above data, we found 38 cbSP regions with two or more lines of functional evidence (Figure 2). This includes 13 regions with annotations from at least three evidence sources. To illustrate the diverse functions associated with cbSPs, we highlight three of these regions (Bitarello *et al.*, 2018; Siewert and Voight, 2020). In these detailed analyses, we also considered additional manually identified annotations and associations from the literature and sources like the gwasAtlas (Watanabe *et al.*, 2019).

Risky behavior and cognitive performance. A ctSP region on chromosome 3q24 is more than 235,000 generations old, and thus has strong evidence of identity by descent between humans and chimpanzees. Both ctSPs in this region (rs9869178, rs2118072) are associated with a risky behavior, automobile speeding propensity. The ctSPs are also modestly associated with variation in brain white matter microstructure (Anterior corona radiata mean diusivities, P = 1.96E-6) (Zhao *et al.*, 2019), as reported in the gwasAtlas database. Variants in the expanded ctSPs region in 3q24 (hg19.chr3:143636420-143740729) are associated with risky behavior and cognitive performance traits in multiple individual GWAS studies (Figure 5A). For example, they are associated with automobile speeding propensity (P = 1E-8) (Linnér, 2019), cognitive performance (P = 5E-9), educational attainment (P = 1E-10) (Lee *et al.*, 2018), and self-reported math ability and highest math class taken (both P = 3E-10). Many of the variants in high LD with the ctSPs in this region overlap annotated regulatory regions: open chromatin region, promoter, promoter flanking region, CTCF binding sites, and enhancer. Furthermore, the ctSPs are significant eQTLs (P ≤ 1E-5) for the gene *DIPK2A* (*C3orf58*) across four GTEx tissues (small intestine terminal ileum, transformed fibroblasts, skin from the lower leg, and suprapubic skin). The DIPK2A protein has not been comprehensively functionally characterized, but it contains a protein kinase domain and is broadly expressed, including in the developing and adult brain. Deletion of this gene has been linked to autism, and its expression is responsive to neuronal activity(Morrow *et al.*, 2008).

Urate levels. Two cbSPs (rs1839333, rs1913638) on chromosome 8q21.11 are both significantly associated (P < 2.0e-18) with uric acid levels in multiple GWAS in European and Asian ancestry populations (Figure 5B) (Köttgen *et al.*, 2013; Kanai *et al.*, 2018; Tin *et al.*, 2019). These variants are also associated with a range of body mass traits in the UK Biobank. Another variant in this locus (rs2941471, R^2=0.97 and R^2=0.82 in East Asians and Europeans respectively) is associated with pancreatic cancer (p=7E-10). Though elevated uric acid in the blood is associated with many conditions, it is a marker for pancreatic cancer (Stotz *et al.*, 2014). This locus also contains LD SNPs (rs1805098 and rs2943549) in East Asians that are expression and splicing QTL for the gene *HNF4G* in testis, pancreas, and brain (P ≤ 5E-5). Variants in *HNF4G* are associated with several traits, including the development of hyperuricemia (Chen *et al.*, 2017). One of the cbSPs (rs1839333, p=2.65E-05) is also associated with gout, although the p-value did not meet our strict threshold.

Body mass and alcohol intake. A cbSP (rs57790054) on 16p12.3 (hg19.chr16: 20006097-20006986) is strongly associated with several growth and body mass phenotypes as well as alcohol intake frequency (Figure 5C; P < 5E-8 for all). Another variant in high LD in Europeans (rs72771074, R^2=0.89) with a cbSP (rs57790054) in this locus was associated with alcohol use disorder in a previous GWAS in a European cohort (P = 5E-8) (Sanchez-Roige *et al.*, 2019). The nearest gene, *GPR139*, encodes for a G-protein coupled receptor expressed in the brain that is involved in alcohol drinking behavior and withdrawal symptoms in rats (Kononoff *et al.*, 2018). This region contains several variants in LD with cbSPs in regulatory regions, such as CTCF binding sites (rs117293173, rs13338055, rs74011247, and rs79521770). One cbSP (rs57790054, p=1.89E-5) is an eQTL for the gene *KNOP1* (aka *C16orf88*). This gene has been associated with obsessive compulsive disorder, among other diseases(Mattheisen *et al.*, 2015). These results suggest that effects on growth and BMI or on addictive behaviors could be under LTBS. We note that there is some evidence of ethanol consumption in chimpanzees, but it is unclear how widespread its availability was over the past several million years (Hockings *et al.*, 2015).

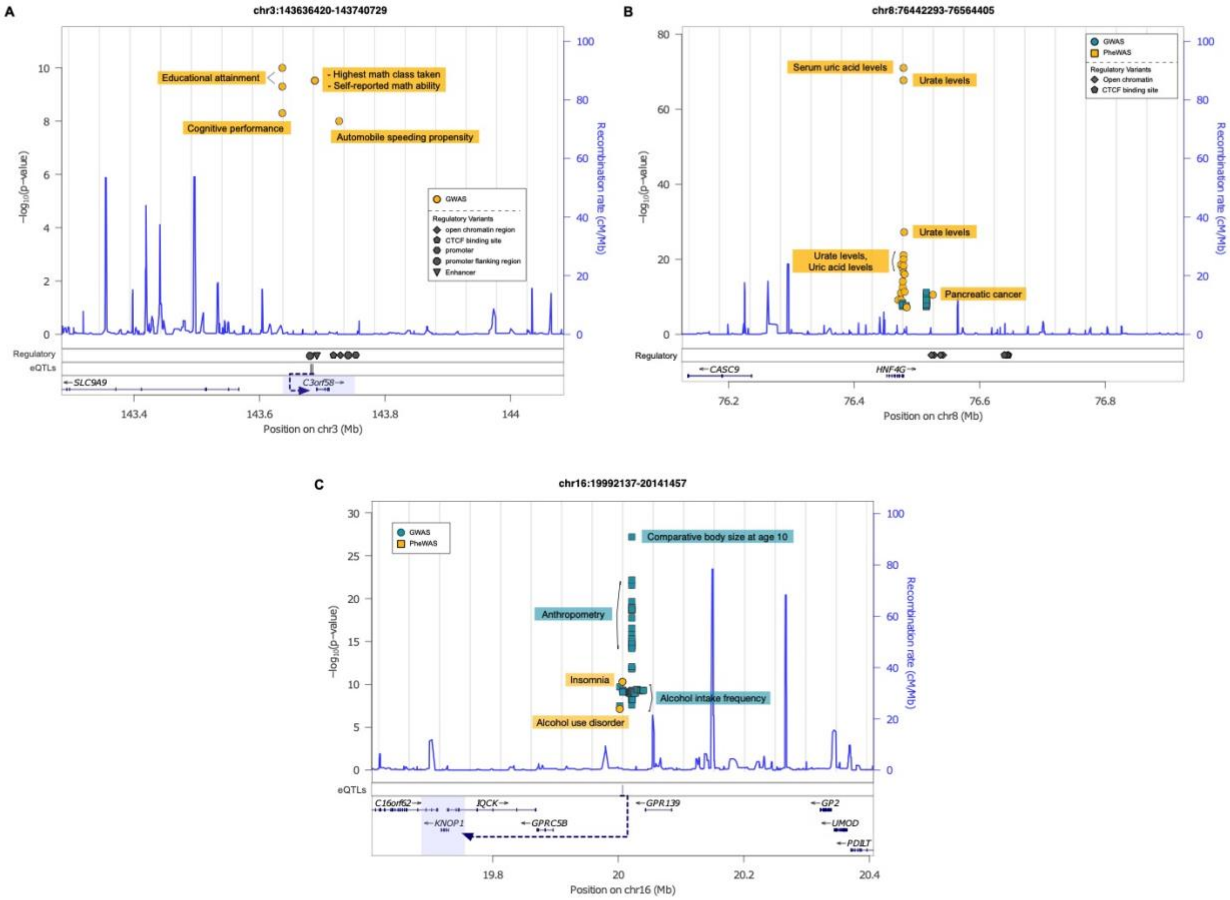


Figure 2.5. Illustrative examples of non-immune functions associated with cbSPs. A) LD SNPs in ctSP locus on 3q24 is associated with cognitive performance and risky behavior. Regional association plot showing statistically significant genome- and phenome-wide associations (threshold p ≤ 1E-08), regulatory and eQTLs. This locus is characterized by neurological traits involved in educational attainment, cognitive performance, and risky behavior (automobile speeding propensity). Both ctSPs in this region (rs9869178, rs2118072) are eQTL in the gene *DIPK2A* (*C3orf58*). LD SNPs are found in enhancer and promoter flanking regions. **B)** SNPs in high LD with cbSPs in 8q21.11 are associated with uric acid and urate levels. Regional association plot showing statistically significant genome- and phenome-wide associations (P ≤ 1E-08), eQTL, and regulatory (open chromatin, CTCF binding site) SNPs. LD SNPs in this region are associated with urate (rs2941484, rs2943539) and uric acid (rs2977944, rs2941484) levels, and pancreatic cancer (rs2941471, p=7E-10). **C)** A cbSP in 16p12.3 is associated with alcohol intake frequency and comparative body size at age 10. The regional association plot shows statistically significant genome- and phenome-wide associations (threshold p ≤ 1E-08), and eQTLs from GTEx. One of the cbSPs (rs57790054, yellow) is associated with alcohol intake in the UK Biobank. A variant in high LD (rs72771074, green) has been associated with alcohol use disorder in a previous GWAS. The cbSP is also strongly associated with insomnia (5e-11). The cbSPs are nearby *GPR139*, a gene encoding a G-protein coupled receptor expressed in the brain, whose expression levels influence alcohol drinking behavior in rats. Figures created with LocusZoom (Pruim *et al.*, 2011)

2.8 Discussion

In this study we aimed to characterize the function of genomic regions with multiple lines of evidence of LTBS on the human lineage. We started with candidate regions containing two or more human-chimp SPs in LD and close proximity. We then considered additional evidence from genome-wide scans for balancing selection with BetaScan2 and NCD, and allele age estimates from ARGweaver. Variants in the resulting candidate sets likely have deep ancestry in the common ancestor between humans and chimpanzees and have persisted in the genomes of both species for millions of years. However, the majority of the non-coding candidate LTBS regions previously identified do not have known functions.

We addressed this challenge with the help of newly developed genomic annotation tools and identified at least one functional annotation for 59 out of 60 cbSP regions and all the ctSP regions. These annotations suggest that non-coding SPs likely maintained by LTBS have diverse functions beyond enabling a flexible immune response to pathogens. This expands on several recent studies of balancing selection over shorter timescales that have also identified regions with functions outside the immune system (Siewert and Voight, 2017; Bitarello *et al.*, 2018; Sato and Kawata, 2018; Viscardi *et al.*, 2018).

To explore the gene regulatory potential of cbSPs, we analyzed eQTL data from 48 tissues from the GTEx Atlas. We found that cbSPs are often eQTL for genes in tissues beyond the immune system, and we observed significant enrichment for eQTL activity in diverse tissues, including many brain and reproductive tissues. A recent study of genes potentially evolving under LTBS identified by the NCD2 statistic found enrichment for genes expressed in the lung, adipose tissue, adrenal tissue, kidney, and prostate (Bitarello *et al.*, 2018). Among our non-coding candidate regions, there is significant enrichment in lung, nominally significant enrichment for adipose and adrenal tissues, and none for prostate or kidney (Figure 3). These differences suggest that the functions of coding vs. non-coding regions subject to LTBS may differ. However, we note that the number of regions considered in each analysis is relatively small.

The phenotype associations we observe for candidate variants in GWAS and PheWAS studies suggest possible behavioral, neurological, and morphological traits that may be targets of LTBS. In particular, our results provide support and candidate loci for previous hypotheses about the need for neurological and behavioral diversity in populations. For example, we found evidence for association with risky behavior and cognitive performance in one ctSP region. Selection has recently been shown to act on risk-taking behavior in anole lizards (Lapiedra *et* *al.*, 2018). Thus, our identification of associations between ctSPs and human risk-taking behavior (Figure 4A) suggests that LTBS may have maintained genetic variants that contribute to variation in risk taking behavior in humans and chimpanzees. The ctSPs are eQTL for *DIPK2A* (*C3orf58*), which encodes for a protein kinase and has been associated with autism and other neurological disorders (Dudkiewicz, Lenart and Pawłowski, 2013). Associations with behavioral and cognitive traits must be interpreted with caution as these traits are very challenging to quantify and strongly influenced by social factors that may vary with other characteristics. Nonetheless, these associations point to an influence of the ctSPs on behaviors relevant to risk tolerance. Thus, it is possible that maintaining a diversity of risk tolerance in human and chimpanzee populations has been beneficial.

Our results also raise the intriguing possibility that variants that modulate urate levels have been under LTBS. Uricase, the enzyme that metabolizes uric acid into an easily excreted water-soluble form in most mammals, has been lost in great apes. This gene was disabled by a series of mutations that slowly decreased activity over primate evolution, increasing the levels of uric acid in blood (Kratzer *et al.*, 2014; Li *et al.*, 2022). It has been hypothesized that this loss of uricase activity was driven by increase fructose in primate diets due to fruit eating (Johnson *et al.*, 2009; Kratzer *et al.*, 2014). It has also been proposed that high levels of uric acid, a potent antioxidant, played an important role in the evolution of intelligence, acting as antioxidant in the brain (Álvarez-Lario and Macarrón-Vicente, 2010). However, as reflected in the associations with this locus, elevated uric acid levels contribute to many common diseases in modern humans, including chronic hypertension, cardiovascular disease, kidney and liver diseases, metabolic syndrome, diabetes, and obesity (Gustafsson and Unwin, 2013). This suggests potential functional tradeoffs at this locus; however, proving the environmental drivers of past selection is challenging.

Some of the phenotype associations we discovered may reflect manifestations of variation on traits in modern environments that could not be long-term drivers of balancing selection. As an extreme example, influence on smoking behavior could not have been the cause of LTBS given the relatively recent wide availability of nicotine. Though we note that there is some evidence of ethanol consumption in chimpanzees (Hockings *et al.*, 2015). Even if they reflect modern environments, these associations provide hints about possible behavioral, neurological, or other traits that may have driven LTBS. For instance, plant chemicals can hijack reward systems in the brain that motivate repetition and learning (U.S. Department of Health & Human Services, 2016). The same systems that influence these actions and consequently reproductive fitness could be a byproduct of excessive seeking of dopamine or other reward chemicals.

There are several caveats to our work. First, factors other than LTBS, such as high mutation rates and sequencing errors, can produce signals similar to those of LTBS. However, our use of additional evidence from balancing selection detection methods, and filters by evidence of ancient origins or the presence of multiple cbSPs in the regions we considered strongly suggest LTBS. Nonetheless, candidate regions of interest for future study should be further analyzed for possible confounders. Moreover, additional approaches for identifying signatures of LTBS have recently been developed. For example, the $T_{2,trans}$ statistic has been shown to have higher power than single species metrics in many scenarios (Cheng and DeGiorgio, 2019). Considering this metric in the definition of cbSPs only identified one additional locus (defined by rs16872492, rs114975228), and it did not have clear functional annotations. Future work will likely identify additional candidate regions that could be characterized using our approaches.

Even with recent growth of genetic and phenotypic databases, our knowledge of the functions of most regions of the genome is sparse. Thus, failure to observe a functional association does not imply that a region does not have an important function. The genome- and phenome-wide association tools we used are limited to the samples that have been analyzed; available data do not represent the full scope of human variation. Most of the individuals analyzed in available genetic association studies are of European ancestry (Sirugo, Williams and Tishkoff, 2019). Variant functions and the ability to detect associations vary across human populations; however, we anticipate that SPs should have functional effects across populations, unless modern environments have masked the pressure driving LTBS. Nonetheless, even in PheWAS, a limited number of phenotypes have been quantified across individuals, and these studies are focused on a subset of clinically relevant rather than evolutionarily relevant traits. To expand the potential to identify candidate functions, in some analyses we considered annotations based on trait associations with variants in high LD ($r^2 > 0.8$) with cbSPs. This could potentially introduce false positives if the variant also tags a different causal variant that is not subject to LTBS. Nonetheless, these associations would still implicate the regions with signatures of LTBS in the associated functions., but functional studies are needed to confirm the role of the candidate variants in these associations. Finally, our analyses have focused on the human context. Due to lack of functional data, it is not possible to explore the function of cbSPs in chimpanzees. Nonetheless, we feel that our integration of genome-scale annotations and biobank data highlight the diversity of functions associated with LTBS.

2.9 Conclusions

In conclusion, we assign putative functions to many non-coding haplotypes carrying human-chimpanzee SPs that likely persisted due to balancing selection dating back to at least their common ancestor. These annotations expand beyond immune functions to traits relevant to behavior, cognition, and body shape. Notably, we also find that most regions with multiple cbSPs overlap gene regulatory annotations suggesting balancing selection on gene expression levels. As methods improve for quantifying the effects of variants on gene regulation in different tissues and how these relate to organism-level phenotypes, we anticipate deeper mechanistic understanding of the functions and potential evolutionary pressures on these regions.

2.10 Methods

Human-chimpanzee shared polymorphisms and balancing selection scans

The initial set of 125 regions containing 263 human-chimp shared polymorphisms analyzed in this study was published by Leffler et al. (Leffler *et al.*, 2013). The set is composed of regions that: 1) contain at least two trans-species polymorphisms—i.e., variants that are segregating in both 51 Yoruba individuals in the 1000 Genomes Pilot 1 and 10 chimpanzees from the PanMap project—within 4 kb of each other in both species, and 2) are in high LD in humans and chimpanzees.

We overlapped the shared polymorphism (SP) regions with balancing selection candidate regions from two different methods developed to detect balancing selection. BetaScan2 (Siewert and Voight, 2020) is a statistic for detecting balancing selection based enrichment for variants in a region with low variation in allele frequency and a deficit of substitutions. We identified overlaps between the SP regions and genomic regions detected by BetaScan2. Among the regions with Beta scores, 48% (60/125) had a SP with value greater than the 2.0 standardized beta score threshold used by the authors. We also computed overlap with regions identified by the NCD statistic (Bitarello *et al.*, 2018). The overlap with the regions detected by NCD containing evidence from at least one population is 14% (18/125 regions). In total, 48% (60/125) of the SP regions were supported by either the BetaScan2 or NCD. We refer to the resulting set of 60 regions as candidate balanced shared polymorphism (cbSP) regions.

Candidate trans-species polymorphisms

We further filtered the cbSP set to find high-confidence candidate trans-species balanced shared polymorphisms (ctSPs). To achieve this, we first selected all cbSP regions that contain three or more SPs, since this is estimated to substantially reduce the false positive rate(Gao, Przeworski and Sella, 2015). We additionally considered time to more recent common ancestor (TMRCA) predictions for the cbSPs from an ancestral recombination graph method, ARGweaver (Rasmussen *et al.*, 2014). ARGweaver reconstructs the recombination history of a genomic site and estimates its age. Following the threshold used in the original ARGweaver analysis of LTBS candidate regions, we filtered cbSP regions to those that are estimated to be 140,000 generations or older, and thus approach the human-chimpanzee divergence. The ctSP subset contains 19 cbSPs.

To increase our ability to identify trait annotations in each locus, we also created an expanded set that includes variants in high LD (threshold R^2=0.8) with each of the SPs as is common in association studies. We computed linkage disequilibrium for the SP variants from 1000 Genomes Project Phase 3 data using the SNiPA Proxy Search web tool developed by the German Research Center for Environmental Health (https://snipa.helmholtz-muenchen.de/snipa3/). We considered LD in African, East Asia, and European populations. Variants with no reported RSID name were excluded from the analysis. The dataset was thus expanded by 6,038 SNPs in high LD with the cbSPs for a total of 6,171 SNPs.

Genome- and Phenome-wide associations

The GWAS Catalog (https://www.ebi.ac.uk/gwas/) collects variant-trait associations from published genome-wide association studies. The database is currently composed of more than 200,000 associations. We used the GWAS Catalog (download date: December 2021) to find functional associations for the LTBS variants. The search was done using the BEDTools intersect function between the GWAS catalog and the LD-expanded SP dataset (Quinlan, 2014).

We performed an enrichment analysis for Experimental Factor Ontology (EFO) trait categories associated with cbSPs in the GWAS catalog using a binomial test based on the background probability of each category across the full catalog. We apply a Bonferroni correction for the number of EFO terms tested (0.05/394 categories tested). However, given the small number of associations with any specific trait, relative enrichment is challenging to quantify.

PheWAS is an analysis strategy built on top of medical records with information about patient phenotypes and associated variants. The geneAtlas (http://geneatlas.roslin.ed.ac.uk/) and the NealeLab (http://www.nealelab.is/uk-biobank) catalogs take advantage of the data provided by the UK Biobank cohort, which contains medically relevant data from nearly 500,000 British individuals of European ancestry. The geneAtlas database contains 3 million variants in 778 traits and the NealeLab database contains more 50,000 variants in more than 4,000 phenotypes. We matched our set of variants against these databases to search for traits associated with balancing selection.

GTEx eQTL data

To evaluate potential gene regulatory effects of SPs in non-coding regions, we analyzed data from GTEx, a project developed to quantify the consequence of genetic variation on expression at the tissue level (https://www.gtexportal.org/). The GTEx project v8 data have identified eQTL across 50 tissues based on analyses of nearly 1,000 individuals to identify differential expression through SNP variation. The intersection between the SPs and LD SNPs and the GTEx eQTL returned a large collection of SPs with evidence of eQTL. To explore the patterns of the cbSPs on regulatory function, we performed an enrichment analysis on these results by calculating the odds ratio on the number of eQTLs for each tissue in the GTEx catalog.

Enrichment for overlap with regulatory regions

We used a permutation framework to calculate whether SPs were more enriched for overlap with regulatory regions than expected by chance (Benton *et al.*, 2019). We quantified the number of overlapping SPs for each type of regulatory region (open chromatin, promoter, enhancer, promoter-flanking, CTCF binding site, TF binding site). We then compared the observed SP overlap to a null distribution of expected overlap generated by randomly shuffling the regulatory regions 1000 times across the genome. We maintain the original length and chromosome distributions for shuffled regions and exclude all ENCODE blacklist and gap regions (Kundaje, 2013), as well as the human MHC locus, since SPs in this region were excluded from the Leffler et al. set. We then computed an empirical p-value for the observed SP overlap based on the distribution of overlaps for the set of matched shuffled regions.

Chapter 3

ARCHAIC INTROGRESSION SHAPED CIRCADIAN TRAITS*

3.1 Introduction

All anatomically modern humans (AMH) trace their origin to the African continent around 300 thousand years ago (ka) (Stringer, 2016; Hublin *et al.*, 2017), where environmental factors shaped many of their biological features. Approximately seventy-thousand years ago (Bae, Douka, and Petraglia 2017), the ancestors of modern Eurasian AMH began to migrate out of Africa, where they were exposed to diverse new environments. In Eurasia, the novel environmental factors included greater seasonal variation in temperature and photoperiod.

Changes in the pattern and level of light exposure have biological and behavioral consequences in organisms. For example, *D. melanogaster* that are native to Europe harbor a polymorphism in *timeless*, a key gene in the light response of the circadian system, that follows a latitudinal cline in allele frequency (Sandrelli et al. 2007; Tauber et al. 2007). The ancestral haplotype produces a short TIM (S-TIM) protein that is sensitive to degradation by light because of its strong affinity to cryptochromes (CRY), photoreceptor proteins involved in the entrainment of the circadian clock. An insertion of a G nucleotide in the 5' coding region of the gene originated approximately 10 kya in Europe and created a start codon that produces a new long TIM isoform (L-TIM). The L-TIM variant has a lower affinity to CRY, creating a change in photosensitivity and altering the length of the period. L-TIM flies are at a higher frequency in southern Europe, while S-TIM flies are more prevalent in northern Europe. Another example is found in pacific salmon. Chinook salmon (*Oncorhynchus tshawytscha*) populations show a latitudinal cline in the frequency and length of repeat motifs in the gene *OtsClock1b*, strongly suggesting that this locus is under selection associated with latitude and photoperiod (O'Malley, Ford, and Hard 2010; O'Malley and Banks 2008). The evolution of circadian adaptation to diverse environments has also been widely studied in insects, plants (Michael *et al.*, 2003; Zhang *et al.*, 2008), and fishes, but it is understudied in humans. Adaptive processes could have helped to align human biology and chronotype to new natural conditions.

Previous studies in humans found a correlation between latitude and chronotype (morningness vs. eveningness) variation (Leocadio-Miguel et al. 2017; Lowden et al. 2018; Randler and Rahafar 2017) and a latitudinal cline in some circadian allele frequencies

*This chapter is under revision for publication: Velazquez-Arcelay et al. 2023. bioRxiv.

(Dorokhov *et al.*, 2018; Putilov, Dorokhov and Poluektov, 2018; Putilov *et al.*, 2019), highlighting the contribution of the environment to behavior and circadian biology. Many human health effects are linked to the misalignment of chronotype (Knutson and von Schantz 2018), including cancer, obesity (Gyarmati *et al.*, 2016; Papantoniou *et al.*, 2016, 2017; Gan *et al.*, 2018; Shi *et al.*, 2020; Yousef *et al.*, 2020), and diabetes (Gan *et al.*, 2015; Larcher *et al.*, 2015, 2016). There is also evidence of a correlation between evening chronotype and mood disorders, most notably seasonal affective disorder (SAD), depression, and worsening of bipolar disorder episodes (Srinivasan *et al.*, 2006; Kivelä, Papadopoulos and Antypa, 2018; Taylor and Hasler, 2018). Thus, we hypothesize that the differences in geography and environment encountered by early AMH populations moving into higher latitudes created potential for circadian misalignment and health risk.

Although AMHs arrived in Eurasia ~70 ka, other hominins (e.g., Neanderthals and Denisovans) lived there for more than 400 ka (Arnold *et al.*, 2014; Meyer *et al.*, 2014, 2016). These archaic hominins diverged from AMHs around 700 ka (Meyer *et al.*, 2012b; Prüfer *et al.*, 2014, 2017; Nielsen *et al.*, 2017; Gómez-Robles, 2019; Mafessoni *et al.*, 2020), and as a result, the ancestors of AMHs and archaic hominins evolved under different environmental conditions. While there was substantial variation in the latitudinal ranges of each group, the Eurasian hominins largely lived at consistently higher latitudes and, thus, were exposed to higher amplitude seasonal variation in photoperiods. Given the influence of environmental cues on circadian biology, we hypothesized that these separate evolutionary histories produced differences in circadian traits adapted to the distinct environments.

When AMH migrated into Eurasia, they interbred with the archaic hominins that were native to the continent, initially with Neanderthals (Green et al. 2010; Villanea and Schraiber 2019) around 60 ka (Sankararaman et al. 2012; Skoglund and Mathieson 2018) and later with Denisovans (Jacobs et al. 2019). Due to this, a substantial fraction (>40%) of the archaic variation remains in present-day Eurasians (Skov et al. 2020; Vernot and Akey 2014), although each human individual carries only ~2% DNA of archaic ancestry (Vernot *et al.*, 2016; Prüfer *et al.*, 2017). Most of the archaic ancestry in AMH was subject to strong negative selection, but some of these introgressed alleles remaining in AMH populations show evidence of adaptation (Racimo *et al.*, 2015; Gower *et al.*, 2021). For example, archaic alleles have been associated with differences in hemoglobin levels at higher altitude in Tibetans, immune resistance to new pathogens, levels of skin pigmentation, and fat composition (Huerta-Sánchez *et al.*, 2014; Racimo *et al.*, 2015, 2017; Dannemann and Kelso, 2017; Racimo, Marnetto and Huerta-Sánchez, 2017; McArthur, Rinker and Capra, 2021). Previous work also suggests that introgressed alleles could have adaptively influenced human chronotype. First, a phenome-wide association study (PheWAS) in the UK Biobank found loci near *ASB1* and *EXOC6* with introgressed variants that significantly associated with self-reported sleeping patterns (Dannemann and Kelso, 2017). One of these alleles showed a significant association between frequency and latitude. Second, summarizing effects genome-wide, introgressed alleles are also moderately enriched for heritability of chronotype compared to non-introgressed alleles (McArthur, Rinker and Capra, 2021). These results suggest a potential role for introgressed alleles in adaptation to pressures stemming from migration to higher latitudes.

Motivated by the potential for a role of archaic introgression in AMH circadian variation, we explore two related questions: 1) Can comparative genomic analysis identify differences in AMH and archaic hominin circadian biology?, and 2) Do introgressed archaic alleles influence human circadian biology? Understanding the ancient history and evolution of chronotypes in humans will shed light on human adaptation to high latitudes and provide context for the genetic basis for the modern misalignment caused by the development of technology and night shiftwork.

3.2 Did archaic hominins and modern humans diverge in circadian biology?

Following divergence ~700,000 years ago (ka) (Nielsen *et al.*, 2017; Gómez-Robles, 2019), archaic hominins and AMH were geographically isolated, resulting in the accumulation of lineage-specific genetic variation and phenotypes (Figure 1). In the next several sections, we evaluate the genomic evidence for divergence in circadian biology between archaic hominin and modern human genomes.

3.2.1 Identifying archaic-hominin-specific circadian gene variation

With the sequencing of several genomes of archaic hominins, we now have a growing, but incomplete, catalog of genetic differences specific to modern and archaic lineages. Following recent work (Kuhlwilm and Boeckx, 2019), we defined archaic-specific variants as genomic positions where archaic hominins (Altai Neanderthal, Vindija Neanderthal, and Denisovan) all have the derived allele while in humans the derived allele is absent or present at such an extremely low frequency in the 1000 Genome Project (<0.00001) that it is likely an independent occurrence. We defined human-specific variants as positions where all individuals in the 1000 Genomes Project carry the derived allele and all the archaics carry the ancestral allele.

We evaluated archaic-specific variants for their ability to influence proteins, splicing, and regulation of 246 circadian genes (Methods). The circadian genes were identified by a combination of literature search, expert knowledge, and existing annotations (Appendix 5.2.1; Methods).

The core circadian clock machinery is composed of a dimer between the CLOCK and ARNTL (BMAL1) transcription factors, which binds to E-box enhancer elements and activates the expression of the Period (*PER1/2/3*) and Cryptochrome (*CRY1/2*) genes (Figure 1C). PERs and CRYs form heterodimers that inhibit the positive drive of the CLOCK-BMAL1 dimer on E-boxes, inhibiting their own transcription in a negative feedback loop. CLOCK-BMAL1 also drives the expression of many other clock-controlled genes (CCG), including *NR1D1/2* (Nuclear Receptor Subfamily 1 Group D Member 1 and 2), *RORA/B* (RAR Related Orphan Receptor A and B), and *DBP* (D-Box Binding PAR BZIP Transcription Factor). ROR and REV-ERB are transcriptional regulators of BMAL1. CK1 binds to the PER/CRY heterodimer, phosphorylating PER and regulating its degradation. Similarly, FBXL3 marks CRY for degradation. Beyond the core clock genes, we included other upstream and downstream genes that are involved in maintenance and response of the clock.

We identified 1,136 archaic-specific variants in circadian genes, promoters, and candidate distal cis-regulatory elements (cCREs). The circadian genes with the most archaic-specific variants are *CLDN4*, *NAMPT*, *LRPPRC*, *ATF4*, and *AHCY* (125, 112, 110, 104, 102 respectively).

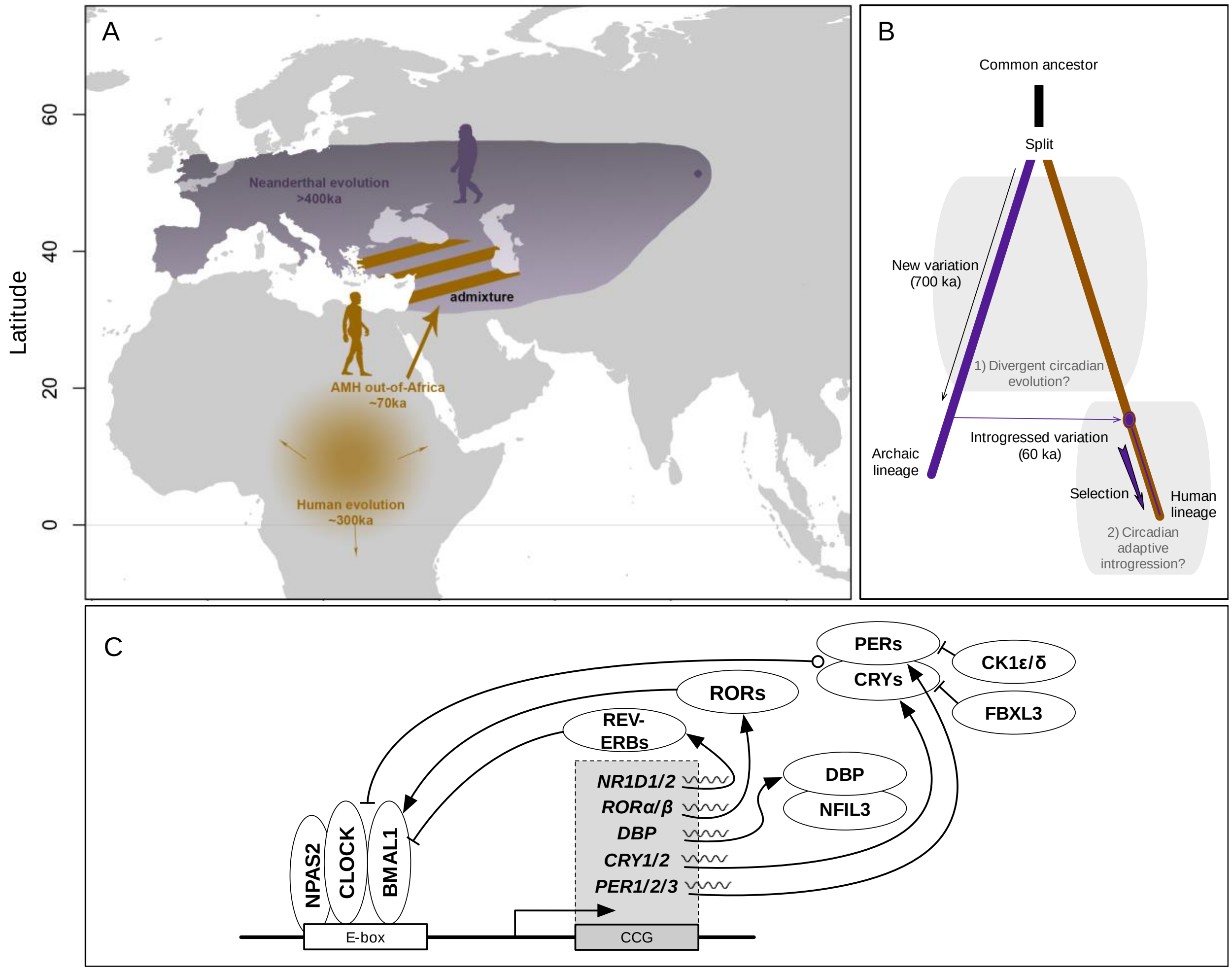


Figure 3.1. Did the sharing of functionally diverged alleles from archaic hominins influence human circadian biology? A) Anatomically modern humans and archaic hominins evolved separately at different latitudes for hundreds of thousands of years. The ancestors of modern Eurasian humans left Africa approximately 70 thousand years ago (ka) and admixed with archaics, likely in southwestern Asia. The shaded purple range represents the approximate Neanderthal range. The purple dot represents the location of the sequenced Denisovan individual in the Altai Mountains; the full range of Denisovans is currently unknown. Silhouettes from phylopic.org. **B)** After the split between the human and archaic lineages, each group accumulated variation and evolved in their respective environments for approximately 700 ka. We first test for evidence for divergent circadian evolution during this time. Humans acquired introgressed alleles from Neanderthals and from Denisovans around 60 and 45 ka, respectively. These alleles experienced strong selective pressures; however, ~40% of the genome retains archaic ancestry in some modern populations. The second question we explore is whether introgression made contributions to human circadian biology. **C)** The core circadian clock machinery is composed of several transcription factors (ovals) that dimerize and interact with E-box enhancer elements and each other to create a negative feedback loop. We defined a set of 246 circadian genes through a combination of literature search, expert knowledge, and existing annotations (Appendix 5.2.1; Methods). Lines with arrows represent activation, and lines with bars represent suppression.

3.2.2 Fixed human- and archaic-specific variants are enriched in circadian genes and associated regulatory elements

After the archaic and AMH lineages diverged, each group accumulated genetic variation specific to each group. Variants fixed in each lineage are likely to be enriched in genomic regions that influence traits that experienced positive selection. We tested whether human- and archaic-specific fixed variants are enriched compared to other variants in circadian genes, their promoters, and in annotated candidate cis-regulatory elements within 1 megabase (Mb) (Figure 2). We found that human- and archaic-specific fixed variants are enriched in circadian genes (Fisher's exact test; human: OR=1.84, P=7.06e-12; archaic: OR=1.13, P=0.023) and distal regulatory elements (Fisher's exact test; human: OR=1.25, P=8.39e-4; archaic: OR=1.16, P=6.15e-5) compared to variants derived on each lineage, but not fixed. Promoter regions have a similar enrichment pattern as that in gene and regulatory regions, but the p-values are high (Fisher's exact test; human: OR=1.21, P=0.65; archaic: OR=1.09, P=0.63). This is likely due to the small number of such variants in promoters. These results suggest that both groups had a greater divergence in genomic regions related to circadian biology than expected.

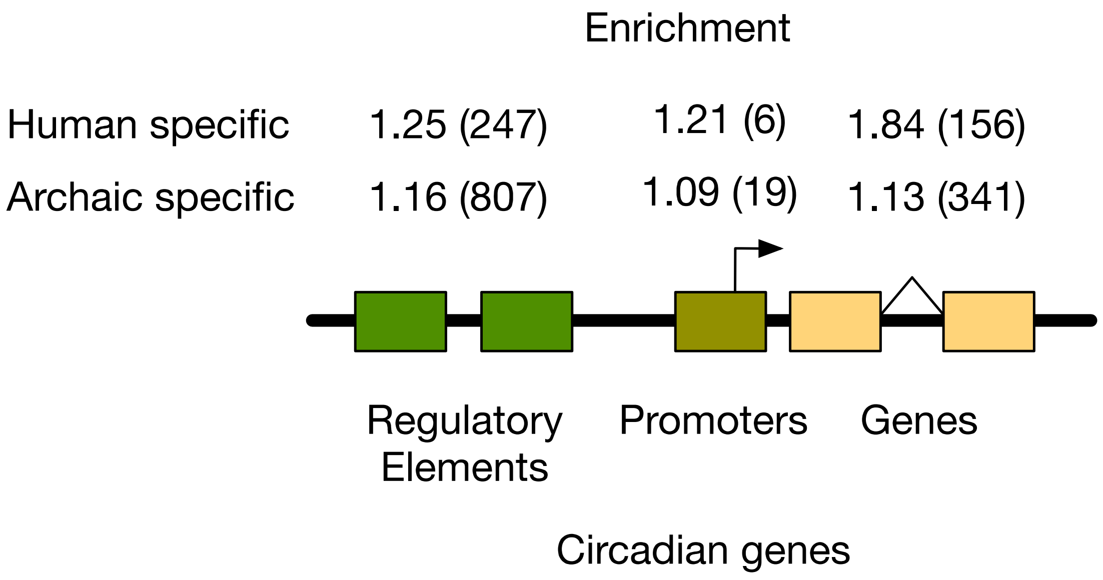


Figure 3.2. Human- and archaic-specific fixed variants are enriched in circadian regulatory, promoter, and gene regions. Human-specific fixed variants are significantly enriched compared to variants that are not fixed in circadian regulatory elements (Fisher's exact: OR=1.25, P=8.39e-4) and gene regions (Fisher's exact: OR=1.84, P=7.06e-12). Promoters show a similar enrichment, but the higher p-value is the result of the small number of variants (Fisher's exact test: OR=1.21, P=0.65). Likewise, archaic-specific variants are enriched in circadian regulatory regions (Fisher's exact: OR=1.16, P=6.15e-5) and gene regions (Fisher's exact: OR=1.13, P=0.023), with the promoters showing a similar trend (Fisher's exact test:

OR=1.09, P=0.63). The numbers in parentheses give the counts of fixed variants observed in each type of element. Regulatory elements were defined based on the ENCODE candidate cis-regulatory elements.

3.2.3 Several core circadian genes have evidence of alternative splicing between humans and archaic hominins

We find only two archaic-specific coding variants in circadian genes: one missense and one synonymous. The missense variant (hg19: chr17_46923411_A_G) is in the gene *CALCOCO2*, calcium-binding and coiled-coil domain-containing protein 2. SIFT, PolyPhen, and CADD all predict that the variant does not have damaging effects. The second variant (hg19: chr7_119914770_G_T) is in the gene *KCND2*, which encodes a component of a voltage-gated potassium channel that contributes to the regulation of the circadian rhythm of action potential firing, but it is synonymous and the variant effect predictors suggest it is tolerated.

To explore potential splicing differences in circadian genes between humans and archaics, we applied SpliceAI to predict whether any sequence differences between modern humans and archaics are likely to modify splicing patterns. Four archaic individuals were included in this analysis (the Altai, the Vindija, the Chagyrskaya Neanderthals, and the Altai Denisovan) (Meyer *et al.*, 2012b; Prüfer *et al.*, 2014, 2017; Mafessoni *et al.*, 2020). We found that 28 genes contained at least one archaic-specific variant predicted to result in alternative splicing in archaics. These included several of the core clock genes *CLOCK*, *PER2*, *RORB*, *RORC*, and *FBXL13* (Figure 3A,C; Appendix 5.2.7). For example, the variant chr2:239187088-239187089 in the 1st intron of *PER2* is predicted to result in a longer 5' UTR. The splice-altering variants were largely specific to the two different archaic linages (Figure 3A), with 13 specific to the Denisovan, 8 shared among the three Neanderthals, and only one shared among all four archaic individuals.

3.2.4 Circadian gene regulatory divergence between humans and archaic hominins

Given the enrichment of variants in regulatory regions of circadian genes, we sought to explore the potential for differences in circadian gene regulation between humans and archaics with causes beyond single lineage-specific variants. We leveraged an approach we recently developed for predicting gene regulatory differences between modern and archaic individuals from combinations of genetic variants (Colbran *et al.*, 2019). The approach uses PrediXcan, an elastic net regression method, to impute gene transcript levels in specific tissues from genetic variation. Previous work demonstrated that this approach has a modest decrease in performance when applied to Neanderthals, but that it can accurately applied between humans and Neanderthals for thousands of genes. Here, we quantify differences in predicted regulation of the 246 circadian genes between 2,504 humans in the 1000 Genomes Project (1000 Genomes Project Consortium, 2010b) and the archaic hominins. The predicted regulation values are normalized to the distribution in the training set from the Genotype Tissue Expression Atlas (GTEx).

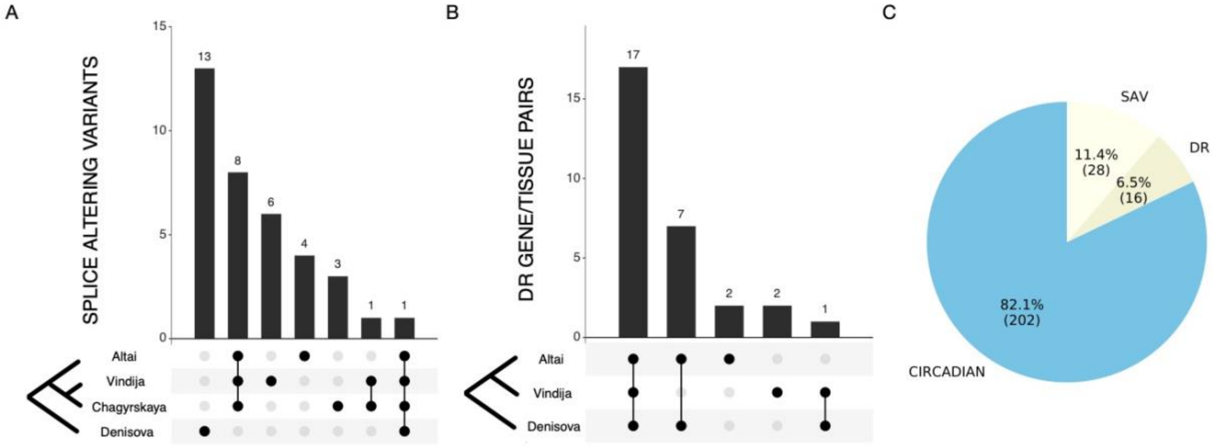


Figure 3.3. Many circadian genes have evidence of alternative splicing and divergent regulation between modern and archaic hominins. A) The distribution of the 28 predicted archaic-specific splice-altering variants (SAV) in circadian genes across archaic individuals. Most are specific to either the Denisovan or Neanderthal lineage (Appendix 5.2.7). **B** The sharing of predicted divergently regulated (DR) gene/tissue pairs across three archaic individuals. (Predictions were not available for the Chagyrskaya Neanderthal.) Seventeen divergently regulated gene/tissue pairs were present in all three archaics (representing 16 unique genes). Additionally, 7 gene/tissue DR pairs are shared between the Altai Neanderthal and the Denisovan individual. One pair is shared between the Vindija Neanderthal and the Denisovan (Appendix 5.2.8). **C)** The proportion of circadian genes containing archaic splice-altering variants predicted by SpliceAI (SAV; 11.4%) or divergently regulated circadian genes predicted by PrediXcan (DR; 6.5%). Thus, 17.9% of the circadian genes are predicted to contain differences to AMH via these mechanisms.

We first analyzed gene regulation predictions in the core circadian clock genes. Archaic gene regulation was at the extremes of the human distribution for many core clock genes including *PER2*, *CRY1*, *NPAS2*, *RORA, NR1D1* (Figure 4; Appendix 5.2.2). For example, the regulation of *PER2* in the two Neanderthals is lower than 2,491 of the 2,504 (99.48%) modern humans considered. The Denisovan has a predicted *PER2* regulation that is lower than 2,410 (96.25%).

Expanding to all circadian genes and requiring archaic regulation to be more extreme than all humans (Methods), we identified 24 circadian genes across 23 tissues with strong divergent regulation between humans and at least one archaic hominin (Figure 3B; Appendix 5.2.8). For example, all three archaic individuals' regulation values for *RORA*, a core clock gene, are lower than for any of the 2,504 modern humans. We found that 16 of these genes (Appendix 5.2.3; Appendix 5.2.8), including *RORA*, *MYBBP1A*, and *TIMELESS*, were divergently regulated in all archaic individuals. This represents 6.5% of all the circadian genes (Figure 3C). Surprisingly, the two Neanderthals only shared one DR gene not found in the Denisovan, while the Altai Neanderthal and Denisovan shared seven not found in Vindija (Figure 3B). The Altai and Vindija Neanderthals represent deeply diverging lineages, and this result suggests that they may have experienced different patterns of divergence in the regulation of their circadian genes.

Given these differences in circadian gene regulation between humans and archaics, we tested whether circadian genes are more likely to be divergently regulated than other gene sets. Each archaic individual shows nominal enrichment for divergent regulation of circadian genes, and the enrichment was stronger (~1.2x) in the Altai Neanderthal and Denisovan individual. However, given the small sample size, the P-values are moderate (Permutation test; Altai: OR=1.21, P=0.19, Vindija: OR=1.05, P=0.43, Denisovan: OR=1.20, P=0.24).

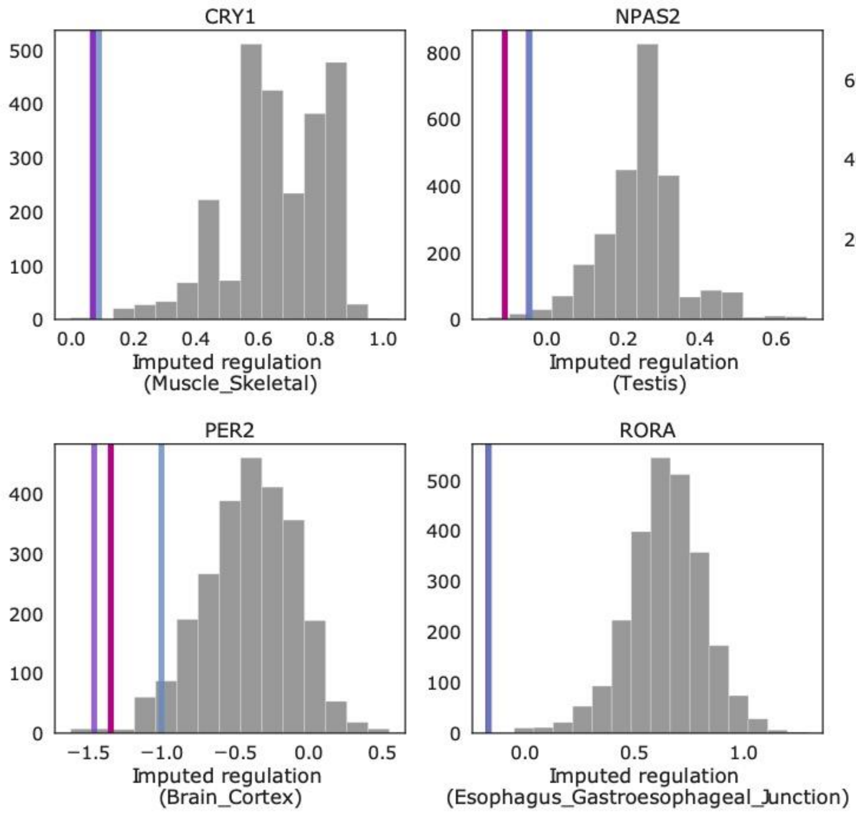


Figure 3.4. Many circadian genes are divergently regulated between modern humans and archaic hominins. Comparison of the imputed regulation of core circadian genes between 2504 humans in 1000 Genomes Phase 3 (gray bars) and three archaic individuals (vertical lines). For each core circadian gene, the tissue with the lowest average P-value for archaic difference from humans is plotted. Archaic gene regulation is at the extremes of the human distribution for several core genes: *CRY1*, *PER2*, *NPAS2, NR1D1 RORA*. See Appendix 5.2.2 for all core clock genes and Appendix 5.2.3 for all divergently regulated circadian genes.

3.3 Did introgressed archaic variants influence modern human circadian biology?

The previous sections demonstrate lineage-specific genetic variation in many genes and regulatory elements essential to the function of the core circadian clock and related pathways. Given this evidence of functional differences between archaic hominins and AMH in these systems, we next evaluated the influence of archaic introgression on AMH circadian biology.

3.3.1 Introgressed variants are enriched in circadian gene eQTL

Given the differences between archaic and modern sequences of circadian genes and their regulatory elements, we investigated whether Neanderthal introgression contributed functional circadian variants to modern Eurasian populations. We considered a set of 863,539 variants with evidence of being introgressed from archaic hominins to AMH (Browning *et al.*, 2018). These variants were identified using the Sprime algorithm, which searches for regions containing a high density of alleles in common with Neanderthals and not present or at very low frequency in Africans. Since many approaches have been developed to identify introgressed variants, we also considered two stricter sets: 47,055 variants that were supported by all of six different introgression maps (Sankararaman *et al.*, 2014; Vernot *et al.*, 2016; Browning *et al.*, 2018; Steinrücken *et al.*, 2018; Skov *et al.*, 2020; Schaefer, Shapiro and Green, 2021) and 755,653 variants that were supported by Sprime and at least one other introgression map. As described below, our main results replicated on both of these stricter sets.

We first tested whether the presence of introgressed variants across modern individuals associated with the expression levels of any circadian genes, i.e., whether the introgressed variants are expression quantitative trait loci (eQTL). We identified 3,857 introgressed variants associated with the regulation of circadian genes in modern non-Africans. The genes *PTPRJ*, *HTR1B*, *NR1D2*, *CLOCK*, and *ATOH7* had the most eQTL (304, 273, 262, 256, and 252 respectively). We found introgressed circadian eQTL for genes expressed in all tissues in GTEx, except kidney cortex. Notably, several of these circadian genes (e.g., *NR1D2* and *CLOCK*) with introgressed eQTL were also found to be divergently regulated in our comparison of modern and archaic gene regulation. This indicates that some of the archaic-derived variants that contributed to divergent regulation were retained after introgression and continue to influence circadian regulation in modern humans.

Introgressed variants are significantly more likely to be eQTL for circadian genes than expected by chance from comparison to all eQTL (Figure 5A; Fisher's exact test: OR=1.45, P=9.71e-101). The stricter set of introgressed variants identified by Browning et al. plus at least one other introgression map had similar levels of eQTL enrichment for circadian genes (OR=1.47; P=2.4e-103). The highest confidence set of introgressed variants that were identified by all six maps considered had even stronger enrichment (OR=1.68; P=6.5e-23).

Most core circadian genes are expressed broadly across tissues; the fraction expressed in each GTEx tissue ranges from 57% (whole blood) to 83% in testis, and an average of 72% (Appendix 5.2.9). As a result, we anticipated that the enrichment of introgressed variants among eQTL for circadian genes would hold across tissues. Examining the associations in each tissue, we found that introgressed eQTL showed significant enrichment for circadian genes in most tissues (34 of 49; Figure 5B; Appendix 5.2.10) and trended this way in all but five. Given that tissues in GTEx have substantial differences in sample size and cellular heterogeneity, statistical power to detect enrichment differs. We anticipate that this is the main driver of differences in enrichment across tissues.

These results suggest that circadian pressures were widespread across tissues. Given the previously observed depletion for introgressed variants in regulatory elements and eQTL (Petr *et al.*, 2019; Rinker *et al.*, 2020; Telis, Aguilar and Harris, 2020), this enrichment for circadian genes among introgressed eQTL is surprising and suggests that the archaic circadian alleles could have been beneficial after introgression.

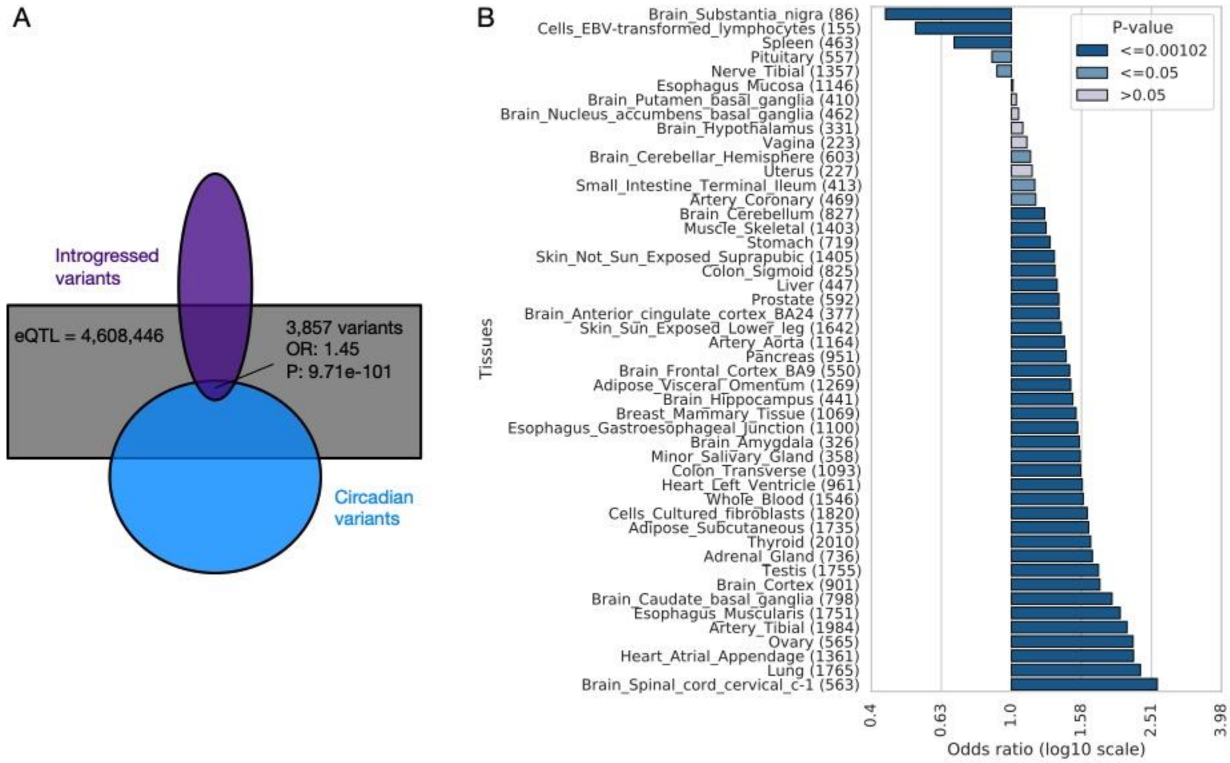


Figure 3.5. Circadian genes are enriched for introgressed eQTL. A) Archaic introgressed variants are more likely to be eQTL for circadian genes in GTEx than for non-circadian genes (Fisher's exact test: OR=1.45, P=9.71e-101). Purple represents the set of introgressed variants, and blue represents the set of circadian variants. 3,857 are introgressed eQTL in circadian genes. Gray represents the universe of all GTEx eQTLs lifted over to hg19. The overlaps are not to scale. **B)** The enrichment for circadian genes among the targets of introgressed eQTLs in each GTEx tissue. Introgressed eQTL in most tissues show significant enrichment for circadian genes (Fisher's exact test; Appendix 5.2.10). Kidney cortex did not have any circadian introgressed eQTLs and thus is not shown. Numbers inside the parenthesis indicate the count of variants in each tissue. Gray bars indicate lack of statistical significance; light blue bars indicate nominal significance (p <= 0.05); and dark blue bars indicate significance at the 0.05 level after Bonferroni multiple testing correction (p <= 0.00102).

3.3.2 Introgressed variants predominantly increase propensity for morningness

After observing that circadian gene expression is influenced by archaic variants, we evaluated whether these effects are likely to result in a change in organism-level phenotype. To do this, we evaluated evidence that introgressed variants influence chronotype. The heritability of chronotype has been estimated in a range from 12 to 38% (Jones *et al.*, 2016, 2019; Lane *et al.*, 2016), and previous studies have identified two introgressed loci associated with sleep patterns (Dannemann and Kelso, 2017; Putilov *et al.*, 2019). We recently found modest enrichment for heritability of chronotype (morning/evening person phenotype in a GWAS of the UK Biobank) among introgressed variants genome-wide using stratified LD score regression (heritability enrichment: 1.58, P=0.25) (McArthur, Rinker and Capra, 2021). This analysis also suggested that introgressed variants were more likely to increase morningness.

To test for this proposed directional effect, we calculated the cumulative fraction of introgressed loci associated with chronotype in the UK Biobank that increase morningness (after collapsing based on LD at R^2>0.5 in EUR). The introgressed loci most strongly associated with chronotype increase propensity for morningness (Figure 6). As the strength of the association with morningness decreases, the bias begins to decrease, but the effect is maintained well past the genome-wide significance threshold (P<5e-8). When focusing the analysis on introgressed variants in proximity (<1 Mb) to circadian genes, the pattern becomes even stronger. The bias toward morningness remains above 80% at the genome-wide significance threshold. This result also held when limiting to introgressed variants found in Browning plus one or all other introgression maps considered (Appendix 5.2.4). This suggests that introgressed variants act in a consistent direction on chronotype, especially when they influence circadian genes..

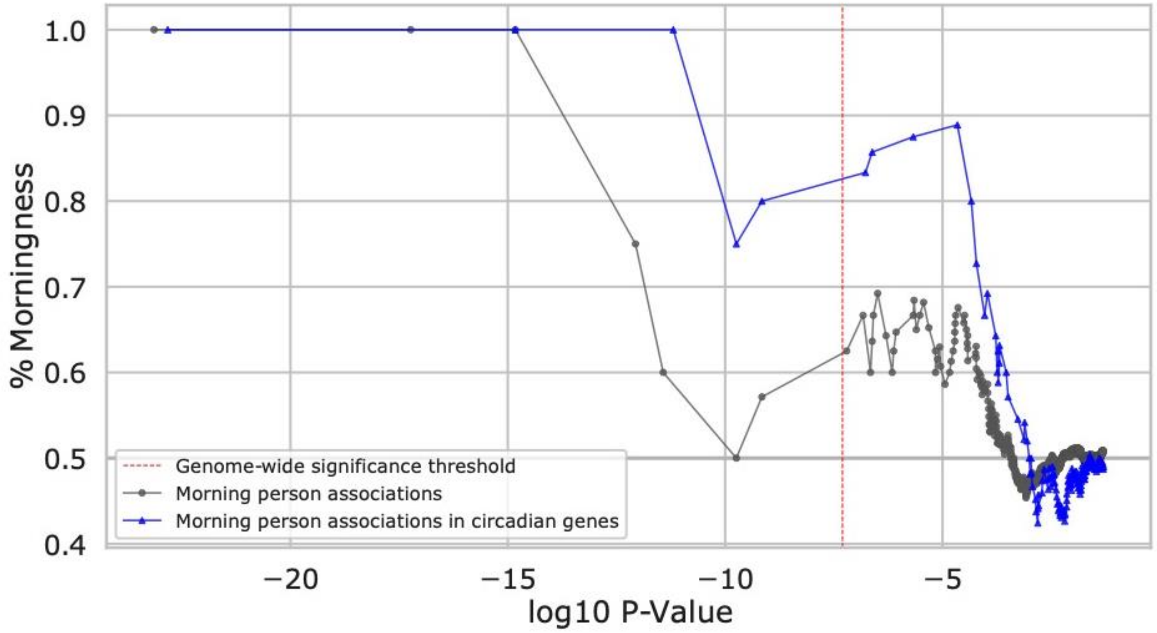


Figure 3.6. Introgressed variants associate with increased morningness. The cumulative fraction of introgressed loci significantly associated with the morning vs. evening person trait in the UK Biobank that increase morningness (y-axis) at a given p-value threshold (x-axis). Introgressed loci associated with chronotype are biased towards increasing morningness, and this effect is greatest at the most strongly associated loci. Introgressed variants nearby (<1 Mb) circadian genes (blue) are even more strongly biased towards increasing morningness than introgressed variants overall (gray). Each dot (triangle) represents an associated locus; variants were clumped by LD for each set (R^2>0.5 in EUR).

Circadian rhythms are involved in a wide variety of biological systems. To explore other phenotypes potentially influenced by the introgressed circadian variants, we evaluated evidence for pleiotropic associations. First, we retrieved all the genome-wide associations reported for introgressed variants in the Open Targets Genetics (https://genetics.opentargets.org) database, which combines GWAS data from the GWAS Catalog, UK Biobank, and several other sources. Introgressed circadian variants are associated with traits from a diverse range of categories. Associations with blood related traits are by far the most common; however, this is likely because they have more power in the UK Biobank. Overall, circadian introgressed variants are significantly more likely to have at least one trait association than introgressed variants not in the circadian set (Fisher's exact test: OR=1.25, P=7.03e-25) (Appendix 5.2.5A). The circadian variants also associate with significantly more traits per variant than the non-circadian set (Mann-Whitney U: P=9.93e-14) (Appendix 5.2.5B). These results suggest effects for introgressed circadian variants beyond chronotype.

3.3.3 Evidence for adaptive introgression at circadian loci

The gene flow from Eurasian archaic hominins into AMH contributed to adaptations to some of the new environmental conditions encountered outside of Africa (Racimo *et al.*, 2015). The above analyses demonstrate the effects of introgressed variants on circadian gene regulation and chronotype. To explore whether these circadian regions show evidence of adaptive introgression, we considered three sets of introgressed regions predicted to have contributed to AMH adaptation: one from an outlier approach based on allele frequency statistics (Racimo, Marnetto and Huerta-Sánchez, 2017) and two from recent machine learning algorithms: *genomatnn* (Gower *et al.*, 2021) and *MaLAdapt* (Zhang *et al.*, 2023). We intersected the circadian introgressed variants with the adaptive introgression regions from each method.

We identified 47 circadian genes with evidence of adaptive introgression at a nearby variant from at least one of the methods. No region was supported by all three methods; however, six were shared between Racimo and *MaLAdapt* and three were shared by Racimo and *genomatnn*. The relatively small overlap between these sets underscores the challenges of identifying adaptive introgression. Nonetheless, these represent promising candidate regions for further exploration of the effects of introgressed variants on specific aspects of circadian biology. For example, an introgressed haplotype on chr10 tagged by rs76647913 was identified by both *MaLAdapt* and Racimo. This introgressed haploype is an eQTL for the nearby *ATOH7* gene in many GTEx tissues. *ATOH7* is a circadian gene that is involved in retinal ganglion cell development, and mice with this gene knocked out are unable to entrain their circadian clock based on light stimuli (Brzezinski *et al.*, 2005).

3.3.4 Latitudinal clines for introgressed circadian loci

Motivated by the previous discovery of an introgressed haplotype on chr2 that is associated with chronotype and increases in frequency with latitude (Dannemann and Kelso, 2017; Putilov *et al.*, 2019), we also tested each introgressed circadian variant for a correlation between allele frequency and latitude in modern non-African populations from the 1000 Genomes Project.

The strongest association between latitude and frequency was a large chromosome 2 haplotype that contains the previously discovered introgressed SNP (rs75804782, R=0.85) associated with chronotype. This haplotype is present in all non-African populations, and rs61332075 showed the strongest latitudinal cline (R=0.87). The second strongest consisted of a smaller haplotype of introgressed variants a few kb upstream of the previous haplotype (tagged by rs35333999 and rs960783) that overlaps the core circadian gene *PER2*. These variants have a correlation between latitude and frequency of ~0.68 They are also in moderate LD (R^2 of ~0.35 in EUR) with an additional introgressed variant (rs62194932) that has a similar latitudinal cline of 0.70 (Appendix 5.2.6). These variants are each in very low LD with the previously discovered haplotype (R^2 of ~0.01) and are each supported by multiple introgression maps. Moreover, these introgressed variants are absent in all EAS populations, absent or at very low frequency in SAS (<3%), and at higher frequency in EUR populations (~13%).

The EUR-specific introgressed variant rs35333999 causes a missense change in the PER2 protein (V903I) that overlaps a predicted interaction interface with PPARG. PER2 controls lipid metabolism by directly repressing PPARG’s proadipogenic activity (Grimaldi *et al.*, 2010). The rs62194932 variant is an eQTL of *HES6* in the blood in the eQTLGen cohort (Võsa *et al.*, 2021). *HES6* encodes a protein that contributes to circadian regulation of LDLR and cholesterol homeostasis (Lee *et al.*, 2012).

Thus, this genomic region, that includes circadian genes and introgressed variants associated with chronotype, has population-specific structure and at least two distinct sets of introgressed variants with latitudinal clines and functional links to lipid metabolism. *PER2* is also predicted to have lower gene regulation in archaic hominins than most humans (Figure 4). and the Vindija Neanderthal carries a lineage-specific variant in this gene that has splice-altering effects. These results together suggests that *PER2* may have experienced multiple functional changes in different modern and archaic lineages, with potential adaptive effects mediated by introgression.

We did not discover any other significant associations between latitude and frequency for other introgressed circadian loci. The rapid migration and geographic turnover of populations in recent human history is likely to obscure many latitude-dependent evolutionary signatures, so we did not anticipate many circadian loci would have a strong signal.

3.4 Discussion

The Eurasian environments where Neanderthals and Denisovans lived for several hundred thousand years are located at higher latitudes with more variable photoperiods than the landscape where AMH evolved before leaving Africa. Evaluating genetic variation that arose separately in each of the archaic and AMH lineages after their split ~700 MYA, we identified lineage-specific genetic variation in circadian genes, their promoters, and flanking distal regulatory elements. We found that both archaic- and human-specific variants are observed more often than expected in each class of functional region. This result suggests that, while each group evolved separately during hundreds of thousands of years in divergent environments, both experienced pressure on circadian related variation. Leveraging sequence-based machine learning methods, we identified many archaic-specific variants likely to influence circadian gene splicing and regulation. For example, core clock genes (*CLOCK*, *PER2*, *RORB*, *RORC*, and *FBXL13*) have archaic variants predicted to cause alternative splicing compared to AMH. Several core genes were also predicted in archaics to be at the extremes of human gene regulation, including *PER2*, *CRY1*, *NPAS2*, *RORA, NR1D1*. Surprisingly, the Altai Neanderthal shared more divergent regulation in the circadian genes with the Denisovan individual than the Vindija Neanderthals. The two Neanderthals represent populations that were quite distantly diverged with substantially different histories and geographical ranges. The Denisovan and Altai Neanderthal also come from the same region in Siberia, while the Vindija Neanderthal came from a region in Croatia with slightly lower latitude.

Introgression introduced variation that first appeared in the archaic hominin lineage into Eurasian AMH. While most of this genetic variation experienced strong negative selection in AMH, a smaller portion is thought to have provided adaptive benefits in the new environments (Racimo *et al.*, 2015). Given the divergence in many circadian genes' regulation, we explored the landscape of introgression on circadian genes. We first looked at introgressed circadian variants that are likely to influence gene regulation in AMH. Variants in this set are observed more often than expected, suggesting the importance of maintaining circadian variation in the population. We also verified that these results held over variants identified by different methods for calling archaic introgression.

We then evaluated the association of these introgressed variants with variation in circadian phenotypes of Eurasians. We previously reported a modest enrichment among introgressed variants for heritability of the morning/evening person phenotype (McArthur, Rinker and Capra, 2021). Here, we further discovered a consistent directional effect of the introgressed circadian variants on chronotype. The strongest associated variants increase the probability of being a morning person in Eurasians.

While it is not immediately clear why increased morningness would be beneficial at higher latitudes, considering this directional effect in the context of clock gene regulation and the challenge of adaptation to higher latitudes suggests an answer. In present day humans, behavioral morningness is correlated with shortened period of the circadian molecular clockworks in individuals. This earlier alignment of sleep/wake with external timing cues is a consequence of a quickened pace of the circadian gene network (Brown *et al.*, 2008). Therefore, the morningness directionality of introgressed circadian variants may indicate selection toward shortened circadian period in the archaic populations living at high latitudes. Supporting this interpretation, shortened circadian periods are required for synchronization to the extended summer photoperiods of high latitudes in *Drosophila*, and selection for shorter periods has resulted in latitudinal clines of decreasing period with increasing latitude, as well as earlier alignment of behavioral rhythms (Hut *et al.*, 2013). In addition, *Drosophila* populations exhibit decreased amplitude of behavioral rhythms at higher latitudes which is also thought to aid in synchronization to long photoperiods (Hut *et al.*, 2013).

Our finding that introgressed circadian variants generally decrease gene regulation of circadian genes suggests that they could lead to lower amplitude clock gene oscillations. However, when assayed in present day humans there is not a strong correlation between the overall expression level of *NR1D1* and the transcriptional amplitudes of other clock genes within individuals (Brown *et al.*, 2008), and quantitative modeling of the mammalian circadian clockworks suggests that stable clock gene rhythms can result across a wide range of absolute levels of gene expression as long as the stoichiometric ratios of key positive and negative clock genes are reasonably conserved (Kim and Forger, 2012). Interestingly, lower transcriptional amplitude of *NR1D1* does confer greater sensitivity of the present-day human clockworks to resetting stimuli, a potentially adaptive characteristic for high latitudes (Brown *et al.*, 2008).

Thus, given the studies of latitudinal clines and adaptation from *Drosophila* and the nascent understanding of clock gene contributions to behavioral phenotypes in present day humans, the directional effects of introgressed circadian gene variants toward early chronotype and decreased gene regulation we observed can be viewed as potentially adaptive. More complex chronotype phenotyping and mechanistic studies of the variants of interest are needed to fully understand these observations.

Finally, to explore evidence for positive selection on introgressed variants in AMH, we analyzed results from three recent methods for detecting adaptive introgression. All methods identified circadian loci as candidates for adaptive introgression. However, we note that the predictions of these methods have only modest overlap with one another, underscoring the difficulty of identifying adaptive introgression. Nonetheless, many of these loci, especially those supported by both Racimo and *MaLAdapt*, are good candidates for adaptive introgression given their functional associations with circadian genes

Several limitations must be considered when interpreting our results. First, it is challenging to quantify the complexity of traits with a large behavioral component (like chronotype) and infer their variation from genomic information alone. Nevertheless, we believe our approach of focusing on molecular aspects (splicing, gene regulation) of genomic loci with relevance to circadian biology, in parallel to GWAS-based associations, lends additional support to the divergence in chronotype between archaic hominins and modern humans. Second, we also note that circadian rhythms contribute to many biological systems, so the variants in these genes tend to be associated with a variety of phenotypes. Thus, there is also the potential that selection acted on other phenotypes influenced by circadian variation than those related directly to chronotype. Third, given the complexity of circadian biology, there is no gold standard set of circadian genes. We focus on the core clock genes and a broader set of expert-curated genes relevant to circadian systems, but it is certainly possible that other genes with circadian effects are not considered. Fourth, recent adaptive evolution is challenging to identify, and this is especially challenging for introgressed loci. Nonetheless, we find several circadian loci with evidence of adaptive introgression from more than one method. Finally, given the many environmental factors that differed between African and non-African environments, it is difficult to definitively determine whether selection on a particular locus was the result of variation in light levels vs. other related factors, such as temperature. Nonetheless, given the observed modern associations with chronotype for many of these variants, we believe it is a plausible target.

In conclusion, studying how humans evolved in the face of changing environmental pressures is necessary to understanding variation in present-day phenotypes and the potential tradeoffs that influence propensity to different diseases in modern environments (Benton *et al.*, 2021). Here, we show that genomic regions involved in circadian biology exhibited substantial functional divergence between separate hominin populations. Furthermore, we show that introgressed variants contribute to variation in AMH circadian phenotypes today in ways that are consistent with an adaptive benefit.

3.5 Methods

Circadian gene selection

Circadian biology is a complex system due to its high importance in the functioning of biological timing in diverse biological systems. For that reason, determining which genes are crucial for selection to environment response related to light exposure is not a straight forward process. To address this issue, we look at different sources of genome annotation databases and searched for genes and variants associated with circadian related phenotypes. We considered all human protein-coding genes in the Gene Ontology database annotated with the GO:0007623 ("circadian rhythm") term or terms annotated with relationship "is_a", "part_of", "occurs_in", or "regulates" circadian rhythm. We also considered genes containing experimental or orthologous evidence of circadian function in the Circadian Gene Database (CGDB), the GWAS Catalog genes containing "chronotype" or "circadian rhythm" associated variants, and a curated set of genes available in WikiPathways (Martens *et al.*, 2021). The final set of circadian genes was curated by Dr. Douglas McMahon.

To select the candidate circadian genes with the highest confidence, we defined a hierarchy system where genes annotated by McMahon or annotated in 3 out of 4 other sources receive a "High" level of confidence. Genes with evidence from 2 out of 4 of the sources are assigned a "Medium" level of confidence. Genes annotated as circadian only in 1 out of 4 sources are assigned to Low confidence and not considered in our circadian gene set. We then defined our set of circadian variants from the 1000 Genomes Project using the official list of circadian genes. The variants are included in analysis of coding, non-coding, regulatory, eQTL, human-specific, archaic-specific, and introgressed variants.

Definition of lineage-specific variants

To identify candidate variants that are specific to the human and the archaic lineages, we used a set of variants published by Kuhlwilm and Boeckx (Kuhlwilm and Boeckx, 2019). The variants were extracted from the high-coverage genomes of three archaics: a 122,000-year-old Neanderthal from the Altai Mountains (52x coverage), a 52,000-year-old Neanderthal from Vindija in Croatia (30x coverage), and a 72,000-year-old Denisovan from the Altai Mountains (30x coverage). The variants were called in the context of the human genome hg19/GRCh37 reference. The total number of variant sites after applying filters for high coverage sites and genotype quality is 4,437,803. A human-specific variant is defined as a position where all the humans in the 1000 Genomes Project carry the derived allele and all the archaics carry the ancestral allele. An archaic-specific is defined as a position where all the archaics carry the derived allele and the derived allele is absent or extremely rare (<= 0.00001) across all human populations. Note that introgressed archaic alleles are not included in the "archaic-specific" set. These criteria resulted in 9,424 human specific and 33,184 archaic-specific variants.

Enrichment of lineage-specific variants among functional regions of the genome

We intersected the sets of lineage-specific variants with several sets of annotated functional genomic regions. Inside circadian gene regions (Gencode v29), we found 156 human-specific variants and 341 archaic-specific variants. In circadian promoter regions, we found 6 human-specific variants and 19 archaic-specific variants. Promoters were defined as regions 5 kb up- to 1 kb downstream from a transcription start site. In distal regulatory elements, we found 247 human-specific variants and 807 archaic-specific variants. For this last set, we considered candidate cis-regulatory elements (cCREs) published by ENCODE (Moore *et al.*, 2020) within 1 Mb of the circadian genes.

To compute whether lineage-specific variants are more abundant than expected in circadian genes, we applied a Fisher's exact test to the sets of human- and archaic-specific variants in regulatory, promoter, and gene regions. Human and archaic-specific variants are significantly enriched in both regulatory (Human: OR=1.25, P=8.39e-4; Archaic: OR=1.16, P=6.15e-5) and gene (Human: OR=1.84, P=7.06e-12; Archaic: OR=1.13, P=0.023) regions. The enrichment observed in the promoters of both lineages is not supported by a significant p-value (Human: OR=1.21, P=0.65; Archaic: OR=1.09, P=0.63).

Genes containing archaic variants with evidence of alternative splicing

We used a set of archaic variants annotated with the splice altering probabilities to identify circadian genes that may be differentially spliced between archaic hominins and AMH (Brand, Colbran and Capra, 2023). We considered variants from four archaic individuals: the Altai, Chagyrskaya, and Vindija Neanderthals and the Altai Denisovan. These archaic variants were annotated using SpliceAI (Jaganathan *et al.*, 2019) and we considered any variant with a maximum delta, or splice altering probability, > 0.2. We identified 36 archaic-specific splice altering variants, defined as those variants absent from 1000 Genomes Project, among 28 circadian genes. Next, we tested for enrichment among this gene set using an empirical null approach (McArthur *et al.*, 2022; Brand, Colbran and Capra, 2023). We shuffled the maximum deltas among 1,607,350 variants 10,000 times and counted the number of circadian genes with a splice altering variant each iteration. Enrichment was calculated as the number of observed genes (N = 28) divided by the mean gene count among 10,000 shuffles. In addition to all genes with archaic-specific variants, we considered six other subsets among these variants: 1) genes with variants private to the Altai Neanderthal, 2) genes with variants private to the Chagyrskaya Neanderthal, 3) genes with variants private to the Altai Denisovan, 4) genes with variants private to all Neanderthals, 5) genes with variants shared among all archaic individuals, and 6) genes with variants private to the Vindija Neanderthal. Finally, we considered a subset of splice altering variants that were identified as tag SNPs by Vernot et al. (Vernot *et al.*, 2016).

PrediXcan

To understand the difference in circadian biology between present-day humans and archaic hominins, we analyzed predictions on gene regulation. We considered the results from PrediXcan gene regulation predictions across 44 tissues from the PredictDB Data Repository (http://predictdb.org/). The models were trained on GTEx V6 using variants identified in 2,504 present-day humans in the 1000 Genomes Project phase 3 within 1 Mb of each circadian gene. The original analysis includes predictions for 17,748 genes for which the models explained a significant amount of variance in gene expression in each tissue (FDR < 0.05). The prediction models were also applied to the Altai and Vindija Neanderthals and the Denisovan. The resulting predictions are normalized values of the distribution observed in GTEx individuals used to train the original prediction models. Each prediction contains an empirical P-value which was calculated for each gene and tissue pair to define genes that are divergently regulated between archaic hominins and humans. The P-value is obtained by calculating the proportion of humans from the 1000 Genomes Project that have predictions more extreme compared to the human median than the archaic individual. Significantly DR genes are defined as those where the archaic prediction falls outside the distribution of humans in the 1000 Genomes Project predictions.

We tested whether the circadian genes in our set are more likely to be DR compared to an empirical null distribution from random gene sets of the same size. We account for the fact that some genes are modeled in more tissues than others by matching the distribution of tissues in which each gene could be modeled in the random sets to our set. Among 1,467 DR genes in the Altai Neanderthal we find 23 DR circadian genes out of the total 236 genes in the circadian set. We iterate through the permutation analysis 1,000,000 times and find an enrichment of 1.21 (P=0.19). A similar analysis is done in the Vindija Neanderthal (1,536 total DR, 21 circadian DR, enrichment of 1.05, P=0.43) and the Denisovan individual (1,214 total DR, 19 circadian DR, enrichment of 1.20, P=0.24). In this study, we define a set of DR genes as the intersection between DR genes in all three archaics, resulting in a set of 16 genes.

Enrichment of introgressed variants in eQTL

We performed an enrichment analysis using Pearson's chi-squared test to evaluate if there is overrepresentation of introgressed alleles in our set of circadian variants using the GTEx dataset. We did a liftOver of the GTEx v8 dataset from hg38 to hg19. The original hg38 set contains 4,631,659 eQTLs across 49 tissues. After the LiftOver, 4,608,446 eQTLs remained, with the rest not mapping. We used the archaic introgressed variants dataset from Browning 2018. The set contains 863,539 variants that are introgressed in humans originating in archaic hominins. We performed an intersection between the set of genes containing evidence for eQTLs and our set of 246 circadian genes to retrieve a subset of variant sites with evidence of being eQTL in circadian genes. The resulting subset contained 97,441 circadian eQTLs in 49 tissues and 239 genes. We further intersected the introgressed variants and the set of eQTL, resulting in 128,138 introgressed eQTLs. The final set of eQTLs that are circadian and also introgressed is 3,857.

Direction of effect of chronotype associations

To explore the effect of archaic introgression in circadian dreams on human chronotype, we quantified the direction of effect of variants associated to a Morning/Evening person trait in a GWAS analysis of the UK Biobank (http://www.nealelab.is/uk-biobank/). The variants were LD clumped using PLINK v1.9 (R^2>0.5). We generated cumulative proportion values on the beta values assigned to each associated variant on an ascending order of P-values.

Detection of pleiotropy in the set of introgressed circadian variants

To understand the extent of different phenotypes associated with the introgressed circadian variants, we first extracted genome-wide associations from Open Targets Genetics (https://genetics.opentargets.org/) for each of the variants with evidence of introgression (Browning et al. 2018). Only the variants with significant p-values were analyzed. The p-value threshold was set at the genome-wide significance level (P=5e-8). We split the variants in two sets: introgressed circadian and introgressed non-circadian. Many of these variants are not associated with any phenotype. We performed a Fisher's exact test to analyze which of the two sets contains a higher ratio of SNPs with at least one association versus SNPs with no association. The result showed that the circadian set had a significantly higher ratio (OR=1.36, P=5e-29). Then we calculated the total of unique traits associated with each of the variants, given that the SNP has at least one association. We used a Mann-Whitney U test to understand which set is represented by a higher level of traits per SNP. The circadian set was slightly more pleiotropic, and the result is supported by a significant p-value (P=5.4e-3).

Detection of latitudinal clines in chronotype associations

To evaluate latitudinal clines in chronotype-associated variants, we assigned a latitude to each of the Eurasian 1000 Genomes Project populations. The latitude of diaspora populations was set to their ancestral country (GIH Gandhinagar in Gujarat: 23.223, STU Sri Jayawardenepura Kotte: 6.916667, ITU Amaravati in Andhra Pradesh: 16.5131, CEU: 52.372778). CEU was assigned a latitude in Amsterdam, following an analysis that shows that this group is more closely related to Dutch individuals (Lao *et al.*, 2008). We then used the LDlink API to retrieve allele frequencies for each introgressed morningness variant in Eurasian individuals (Machiela and Chanock, 2015). Variants that follow a latitudinal cline were identified using linear regression statistics requiring correlation coefficient (R >= 0.65) and P-value (P <= 0.5).

Identifying introgressed circadian variants with evidence of adaptive introgression

We sought to identify circadian variants that contain evidence of adaptive introgression (AI). To achieve this, we collected AI predictions from a method that applied various summary statistics on 1000 Genomes Project data (Racimo, Marnetto and Huerta-Sánchez, 2017) and two sets of genomic regions that were measured for their likelihood to be under AI by two machine learning methods: *genomatnn* and *MaLAdapt*. *genomatnn* is a convolutional neural network trained to identify adaptive introgression based on simulations (Gower *et al.*, 2021). *MaLAdapt* is a machine learning algorithm trained to find adaptive introgression based on simulations using an extra-trees classifier (ETC) (Zhang *et al.*, 2023). Following the thresholds used in each paper, a region is considered to be under AI if the prediction value assigned to it meets a threshold of 0.5 or 0.9, respectively. To find the variants of interest that fall into AI regions, we intersected the set of introgressed circadian SNPs with the Racimo et al. 2015, *genomatnn* and the *MaLAdapt* regions individually. The set of introgressed circadian variants contains variants inside circadian genes, in circadian promoter regions (5 kb up- and 1 kb downstream of the TSS), and variants with regulatory function (cCREs) flanking circadian genes by 1 Mb.

Chapter 4

DISCUSSION

4.1 Summary of Contributions

The sequencing of thousands of human genomes, from present-day and ancient individuals, during the last two decades has deepened our understanding of human adaptation. These data shed light on the genetic diversity among humans and how natural selection has acted upon this variation to help humans adapt to different selective pressures, shaping their evolutionary history.

My work has specifically contributed to this understanding by exploring two mechanisms of selection on "old" genetic variation from different origins. Trans-species polymorphisms are old variants that appeared before the last common ancestor between two species at specific sites of the human genome and persist in both lineages. These variants are often found in regions under long-term balancing selection. In Chapter 2, we described non-immune human phenotypes linked to trans-species polymorphisms and likely targeted by long-term balancing selection on non-coding regions of the human genome. To do this, we analyzed a set of balanced trans-species polymorphisms persisting in humans and their closest living relatives, the chimpanzees. The development of genomic functional annotations and association studies have been essential in understanding of the role of specific variants on the expression of different phenotypes. Using genomic annotations and association studies, we identified potential drivers of balancing selection associated with behavioral or neuropsychiatric traits and metabolism. Among the behavioral and neuropsychiatric traits, we found one region associated with risk taking behavior and cognitive performance. The nearest gene, *C3orf58*, is associated with autism. We also discovered associations with addiction, specifically alcohol intake. It is possible that variation in tolerance for risky or repetitive behaviors might have been useful to maintain during human evolutionary history. As for metabolic traits, our analysis revealed that a region near the gene *HNF4G* is associated with uric acid levels in the blood. Interestingly, a region under balancing selection in the *SLC2A9* gene is also associated with urate levels (DeGiorgio, Lohmueller and Nielsen, 2014; Siewert and Voight, 2017, 2020; Bitarello *et al.*, 2018; Cheng and DeGiorgio, 2019). These findings are consistent with previous knowledge about balancing selection maintaining alleles with potential for different fitness consequences in different environments.

While balancing selection can induce an increase in variation in the genome and maintain this variation long after speciation events, other processes decrease variation. In Chapter 3, we explored a set of variants that did not originate in humans or their ancestor, but instead were introduced in the human genome through a process of gene flow, i.e. introgression. Recently, starting around 55 ka, humans acquired new variation through the admixture with Eurasian archaic groups, i.e. Neanderthals and Denisovans (Fu et al., 2014; Prüfer et al., 2014; Sankararaman et al., 2012). Following the time of introgression, many of these alleles experienced strong negative selection; however, some have been reported to have experienced positive selection (Racimo *et al.*, 2015). New technologies allowing the sequencing of hominin ancient DNA that has undergone thousands of years of deterioration has been essential to better understand the process of adaptation to new selective pressures (new environment, pathogens, etc.) through the introgression of alleles that have already gone through the process of natural selection in a specific environment. In Chapter 3, we explored the role of archaic introgression in shaping circadian biology and sleep/wake patterns in humans migrating to higher latitudes from Africa to Eurasia. We first found that human specific and archaic specific variants are significantly enriched in circadian genes and in regulatory regions. This suggests that selective pressures were acting in different directions in different environments. When we explored further the differences in circadian genes between these groups, we found that 16 circadian genes are divergently regulated between humans and archaics and 28 circadian genes contain variants predicted to alter splicing in archaics. Given that we observe divergent evolution in and near circadian associated genes, we explored the contribution of archaic introgression to human circadian biology and potential for adaptation to the higher latitudes where archaics evolved. We found that introgressed variants are enriched among circadian variants that have regulatory function. We also found that most of the variants most strongly associated with chronotype consistently increase morning preference. Two different signals on chromosome 2 follow a latitudinal cline in Eurasian populations. This direction of effect supports the idea that higher latitudes triggered a selective pressure on circadian periods in geographical regions where photoperiods are extreme. The results of these studies provide examples of how variants of ancient origin function and have experienced different selective pressure in humans. They thus advance our understanding of how the diverse origin of human genetic variation facilitate adaptation.

4.2 Limitations

Despite our contributions, the projects presented in this dissertation have some limitations.

Genetic data are biased towards individuals of European ancestry

The studies presented here were made possible by the availability of present-day human sequencing data. Most human sequencing projects have focused on sequencing groups of individuals from specific populations, which can mean at the continent, country, or culture level. Genetic variation, allele frequencies, and LD may vary from one population to another. For that reason, the population being studied is relevant to the results that will be obtained. However, most of the association studies aiming to identify the connection between genotype and phenotype (e.g. genome- and phenome-wide studies) are predominantly biased to groups of people of European descent. This limitation affects the ability of these methods to detect many of the functions of genomic sites on phenotype, because there is less power to detect phenotypic associations with smaller samples sizes and many variants are not present across all populations. Most notably, African populations carry more genetic variation and diversity than Eurasians (Yu *et al.*, 2002; 1000 Genomes Project Consortium, 2015).

Functional data are biased in their coverage of phenotypes and populations

Many of the available annotations for regions that we described here are also based on mostly individuals from European ancestry. Moreover, the phenotypes and cell types that have been investigated in genome-wide studies do not reflect the full range of phenotypes potentially relevant to recent human evolution. This caveat potentially limited our ability to detect phenotypes associated with the regions under study.

Pleiotropy limits the ability to infer which traits drove selection

Many genes and genetic variants have functional effects on many different traits and biological systems. In Chapter 2 we ensured that the regions we reported as non-immune related long-term balancing selection regions did not have any annotation related to the immune system. However, many of these regions still associated with multiple different traits, each of which could be a candidate for driving selection. These analyses reveal the diversity of functions potentially influenced by long-term balancing selection, but for most specific variants or genomic regions, multiple different associated phenotypes were found.

Limited high-quality sequence data from archaic hominins

The sequencing of archaic hominins made the analyses of Chapter 3 possible. Although many attempts have been made at sequencing archaic genomes (Hajdinjak *et al.*, 2018; Skov *et al.*, 2022), today only 4 high-coverage genomes from 3 Neanderthals and 1 Denisovan are available. The power of population genetics lies in quantifying patterns of variation across an abundance of sequenced individuals. This low number of individuals limits our ability to gather an understanding of the extent of variation available in this group of hominins and the frequencies of the alleles in these populations. Although there is evidence that these archaic groups had very low genetic diversity (Skov *et al.*, 2022), acquiring more samples is essential to provide a more representative understanding of the genetic structure of these populations.

Finally, we relied on various statistics and algorithms that have been developed to detect balancing selection and introgression. These genomic processes are complex and challenging to identify. Many methods have been developed to identify balancing selection in the human genome. However, these methods differ substantially in the genomic regions they identify. In some cases, this is due to differences in power over different timescales, but we also anticipate that some methods have substantial false positive rates. Similarly, many methods that aim to identify introgressed regions in the human genome do not always agree with each other. This means that without supporting evidence from multiple sources, interpretations at specific sites could be based on false positives. Thus, we in our analyses focused on regions with support from multiple methods, but acknowledge that this may result in lower sensitivity.

4.3 Future directions

Improved methods for detecting signatures of ancient and long-term selection based on machine learning.

Although several different of methods to detect regions of the genome containing ancient haplotypes (e.g., resulting from long-term balancing selection or archaic introgression) have been developed, often the results from these methods do not agree with each other. Thus, because these evolutionary processes leave complex signatures in the genome, these methods have limitations that affect the capacity to accurately identify these processes. There is a need to develop comprehensive methods capable of more accurately detecting these regions. Not having accurate methods leads to false positives or false negatives in studies that rely on them to analyze the process of natural selection. The rapid development of machine learning methods is currently helping make predictions in diverse areas of medical and biological research. Similar methods can be used to build machine learning models that better predict or identify signatures of different selective processes using simulated data based on human population demographic models (Schrider and Kern, 2017; Kern and Schrider, 2018). These approaches could be used to refine the set of genomic regions considered in Chapter 2 and Chapter 3.

Incorporation of ancient human DNA into studies of selection and introgression

During the past decade thousands of genomes from ancient modern humans were sequenced, ranging in age from 45 ka to 100 years ago (Fu et al., 2015; Hajdinjak et al., 2021; Mallick & Reich, 2023; Prüfer et al., 2021). These data allow direct study of the process of natural selection through time/generations in the past instead of just relying on the snapshot of genetic variation in the present. In this dissertation we analyzed archaic variants remaining in the genomes of present-day humans. However, we know that the amount and frequencies of archaic variants in human populations changed during the generations after the introgression event. Also, admixture events between human populations across Eurasia have masked the signatures of selection on these variants. Thus, we are missing historical information about the adaptive process by limiting our studies to modern genomes. Exploring the historical direction of selection on introgressed variants associated with circadian genes using ancient DNA could support and refine our observations. Previous studies have used this approach to detect old hard sweeps that were obscured by the Bronze Age admixtures that took place across Europe (Souilmi *et al.*, 2022). Similar methods could be applied in Chapter 3 to detect how natural selection shaped human circadian biology near circadian genes during the generations after the introgression events.

The impact of past evolutionary adaptations in rapidly changing modern society

Throughout the human evolutionary history, environments directed adaptation, shaping the frequencies of variants in human populations in response to various environmental challenges. As society and technology changed and advanced, these factors influenced the ongoing development of humans, and are continuing to shape human biology in modern times.

In early human evolution, humans lived in diverse environments and faced a wide range of challenges. Maintaining a variety of traits in populations may have allowed for adaptive flexibility and contributing to the survival and success of populations in diverse environments. For example, a mix of risk-averse and risk-taking behaviors could increase the chance of survival of a population. A few individuals prone to taking risks could lead to potential rewards for the entire group at the expense of higher dangers for the individual taking risks. Risk-taking behavior can translate into individuals who ventured into new territories to find more abundant food, water, and shelter sources, allowing the group to adapt to changing environmental conditions. In modern society, some people are willing to take financial, personal, or social risks, while others prefer caution. This diversity contributes to advancement of human rights, the development of new companies that develop innovative technologies and health therapies, and in the exploration of new territories.

The exploration of new territories is not a new trait in humans. The ancestors of humans evolved near equatorial regions, and after the emergence of humans as a distinctive species in Africa, they continued to evolve there for 200 thousand years before moving into other regions of the planet with a variety of climates and photoperiods. These new environments posed a pressure on many bodily systems including the capacity of circadian clocks to respond to variable light-dark patterns and entrain their internal clocks.

In modern society, the invention of the incandescent light bulb in 1879 gave some populations the capacity to transition to a longer-lasting source of light. Electric light allowed humans to better simulate the daytime level of brightness at night. However, it also disrupts the entrainment of the body's clock with the external light-dark cycle (Blume, Garbazza and Spitschan, 2019). Artificial light at night can disrupt the rhythms of various physiological processes, including sleep-wake cycles, hormone secretion, cognitive function, energy metabolism, glucose regulation, adipose tissue function, appetite regulation (Tordjman *et al.*, 2017), by suppressing the production of melatonin (Blume, Garbazza and Spitschan, 2019). This occurs through the wide range of the visible spectrum emitted by light bulbs, including blue light, which is particularly disruptive as it mimics daylight (Blume, Garbazza and Spitschan, 2019; Moyano, Sola and González-Lezcano, 2020; Wong and Bahmani, 2022). Electronic devices such as smartphones, tablets, and computers have a stronger emission of blue light. This new lifestyle occurred during the span of a few decades, and it differs from the environment where humans evolved for hundreds of thousands of years. Similar to how humans migrating from near equatorial regions to higher latitudes faced adaptive pressures associated to their circadian biology, we need to evaluate the effect of light emitted by electronics on modern human biology.

With the increasing specialization of labor in today’s society, many industries (e.g., healthcare, transportation, hospitality) require personnel working during late shifts. Late shift work creates a misalignment between work hours and individuals' biological clocks and can result in an increased risk of chronic diseases, including obesity, diabetes, cardiovascular diseases, cancer, and mental health issues (Baron and Reid, 2014; Khan *et al.*, 2018).

Another important aspect that society needs to address is the diversity of chronotypes. Throughout this work we have focused on describing the adverse effects on the health of humans that fall close to the average chronotype pattern. However, it is essential to consider those who fall on the extreme ends of the chronotype spectrum. These individuals have to adapt to society's standard work patterns, living constantly in asynchrony with their internal clocks.

In summary, in this dissertation contributes to knowledge about how ancient genetic variation has served as the raw material for natural selection. This work expanded current knowledge about the non-immune targets of long-term balancing selection, specifically in non-coding regions. It also discovered and quantified the influence of ancient introgression from Neanderthals on the evolution of circadian biology in environments that were new to humans.

Chapter 5

APPENDICES

Appendix 5.1

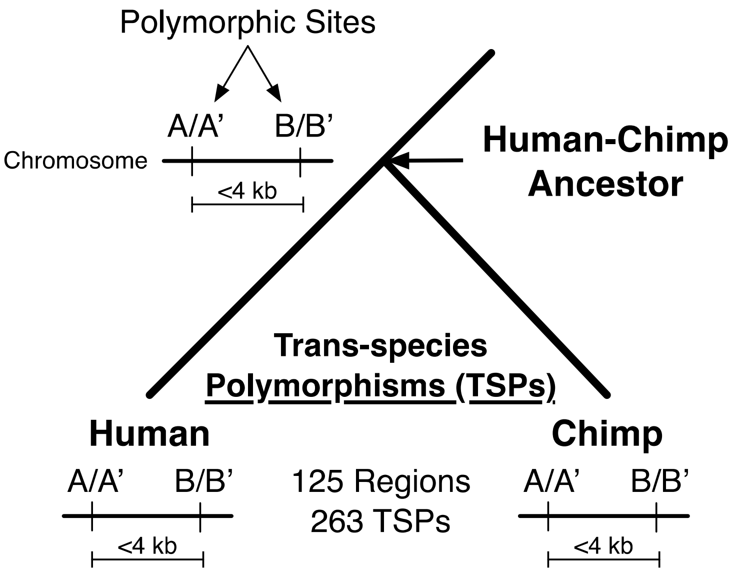


Figure 5.1.1. Human-chimpanzee shared polymorphisms (SPs) previously reported as candidate targets of long-term balancing selection (LTBS). Schematic showing the criteria used by Leffler et al. (2013) to identify SPs likely maintained by LTBS. Each line represents a chromosome with polymorphisms segregating in a species. A/A' are two alleles segregating in both humans and chimpanzees at one site (i.e., an SP), and B/B' are two alleles segregating in both species at a nearby SP site. SPs are very unlikely to appear nearby (within 4 kb) without the action of balancing selection. Within these regions, multiple functional scenarios are possible. For example, one SP may be under LTBS while the other is neutral, but maintained due to tight linkage. Alternatively, the SPs may have epistatic functions and both be under selection.

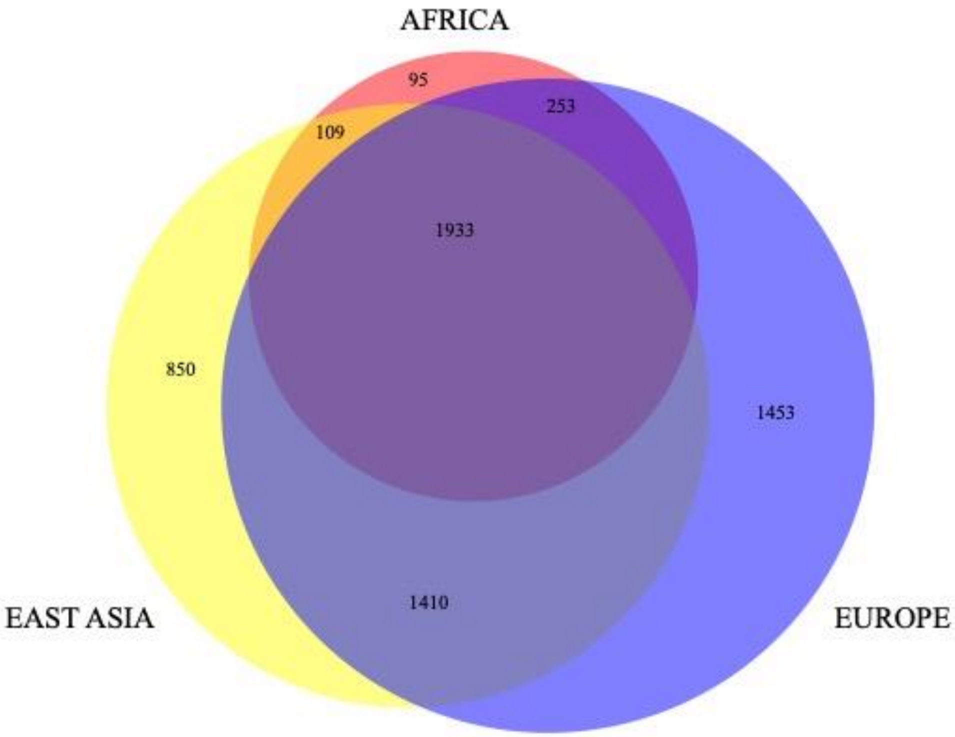


Figure 5.1.2. SNPs in LD with candidate balanced shared polymorphisms (cbSPs). We consider 60 regions containing 133 cbSPs. For each of these SNPs we find variants in high LD (R^2 >= 0.8). As a result, we obtain an additional 6,038 LD variants from the 1000 Genomes Project. Counts include LD SNPs and cbSPs. Figure created with www.biovenn.nl

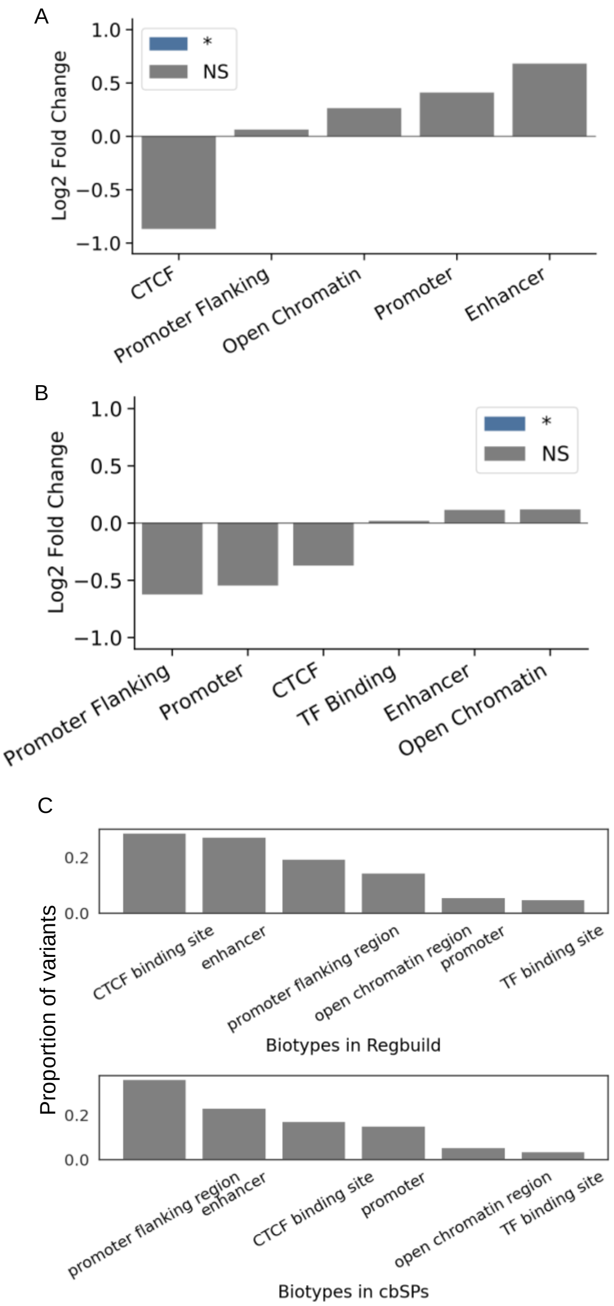
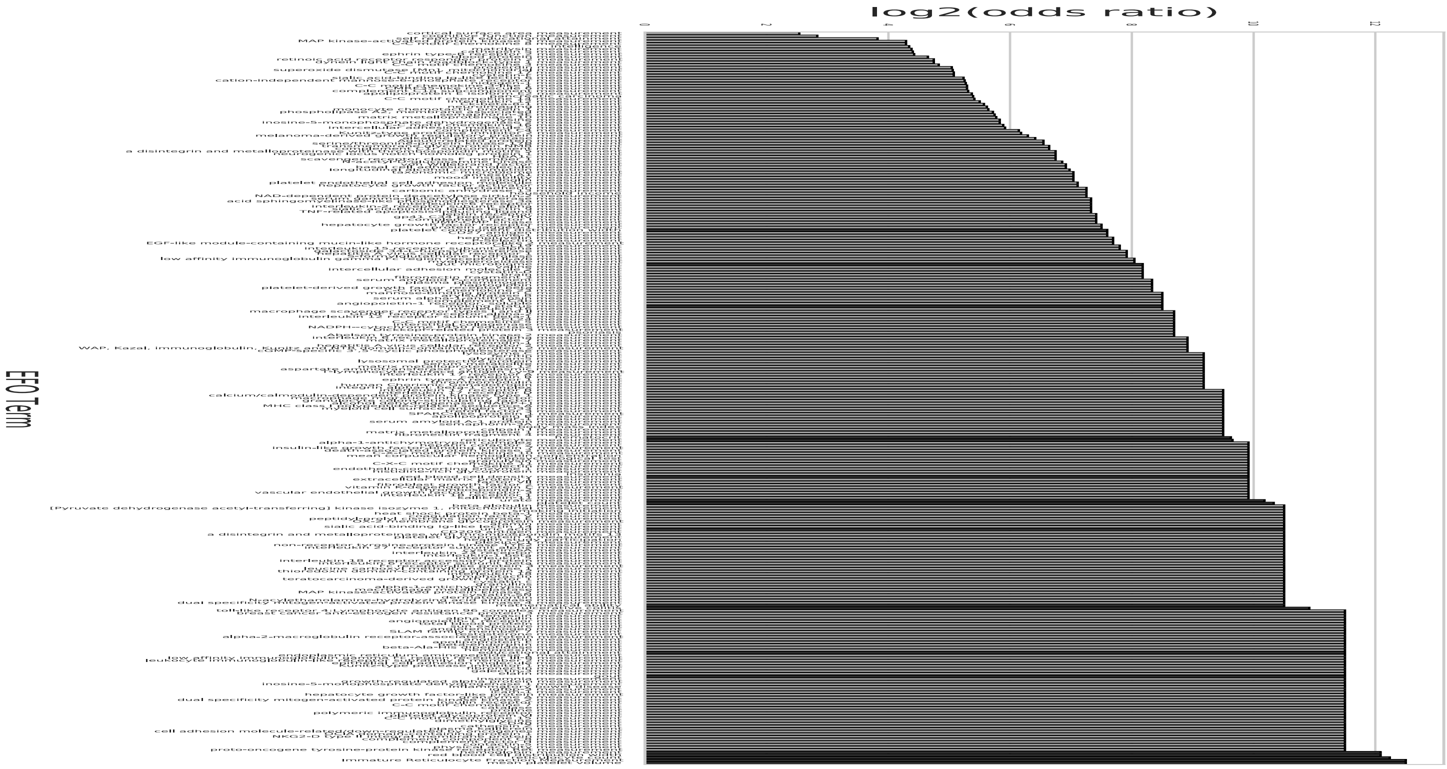


Figure 5.1.3: Enrichment analysis of cbSP in annotated regulatory regions. cbSPs overlap more enhancers promoters, and open chromatin regions and fewer CTCF binding sites than expected compared to length- and chromosome-matched non-coding regions from the genomic background. However, these signals were not statistically significant. Enrichment was tested in the cbSP haplotype region (A) and in the LD region (B). Since variants in CTCF regions are likely to influence regulation of many genes in many tissues (e.g., compared to enhancers which are often context-specific), this suggests that individual cbSPs may be less pleiotropic than expected by chance. C) The proportion of LD variants observed in each regulatory feature type (bottom) and genome-wide (top).


log2(odds ratio)
EFO Term
0
2
4
6
8
10
12

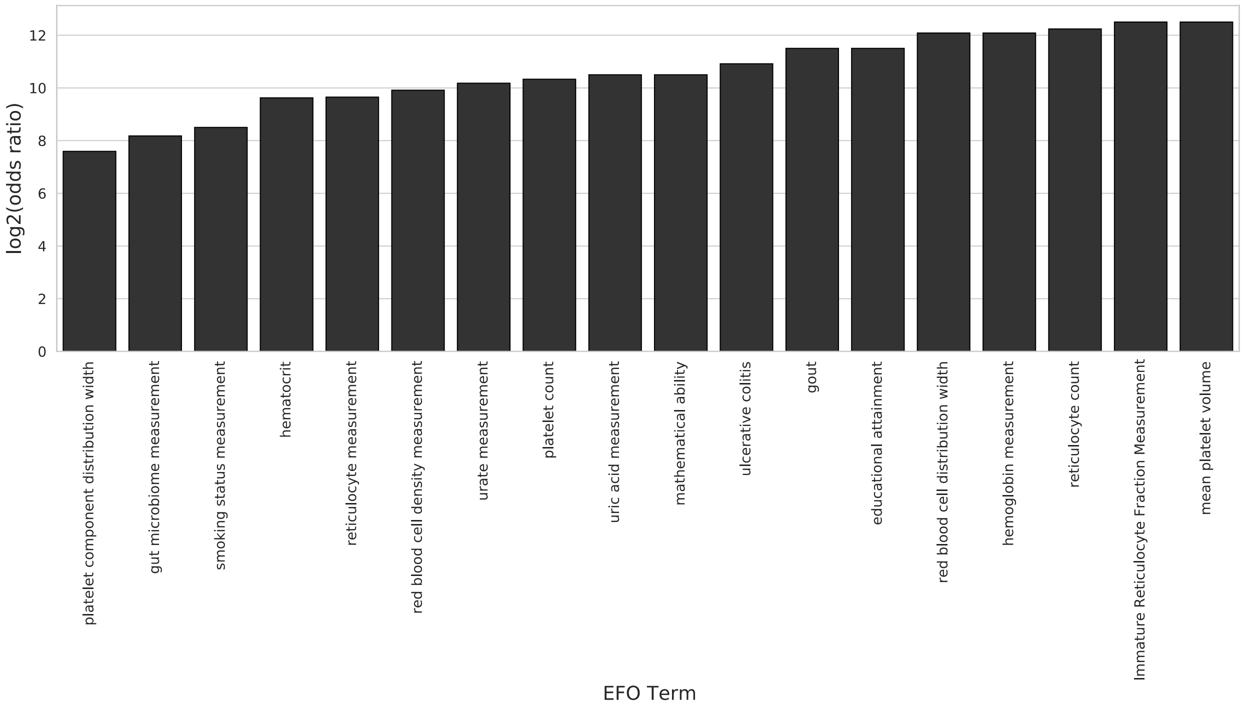


Figure 5.1.4: Enrichment analysis of GWAS phenotype categories. (Top) We performed an enrichment analysis on the GWAS phenotype categories (EFOs) and found significant enrichment in many of the categories. Bars colored in gray meet a significant threshold of 0.05 P-value (binomial test), and bars colored in black pass a Bonferroni correction. (Bottom) The most enriched GWAS EFO categories include blood and immune related traits, and also cognitive, smoking status, and uric acid related traits, including urate levels and gout. All the categories represent a significant enrichment under a Bonferroni correction (binomial test). However, we note that the absolute number of associations driving these enrichments are very small.

Appendix 5.2

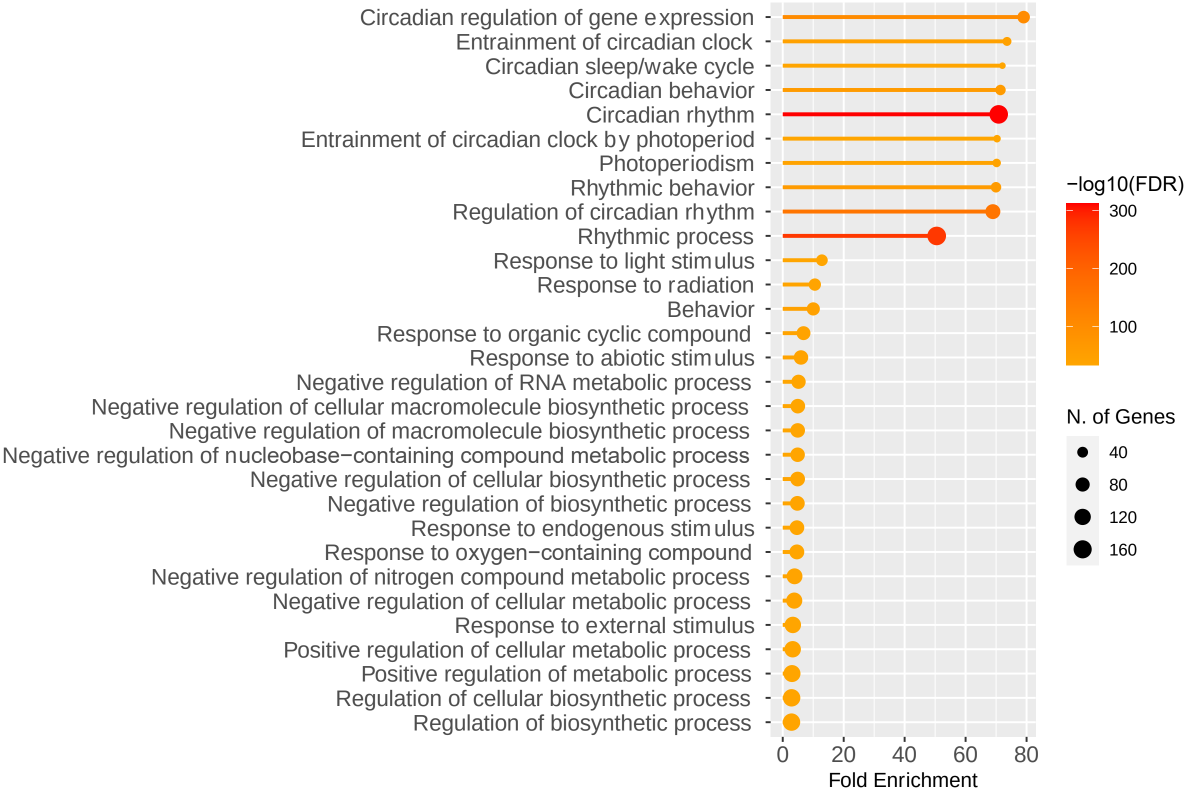


Figure 5.2.1. **GO terms associated with the 246 circadian genes.** Generated by ShinyGO, http://bioinformatics.sdstate.edu/go75/ . The strong enrichment for circadian terms supports their relevance, but we note that GO annotations were used to select some of these genes, so this should not be viewed as independent.

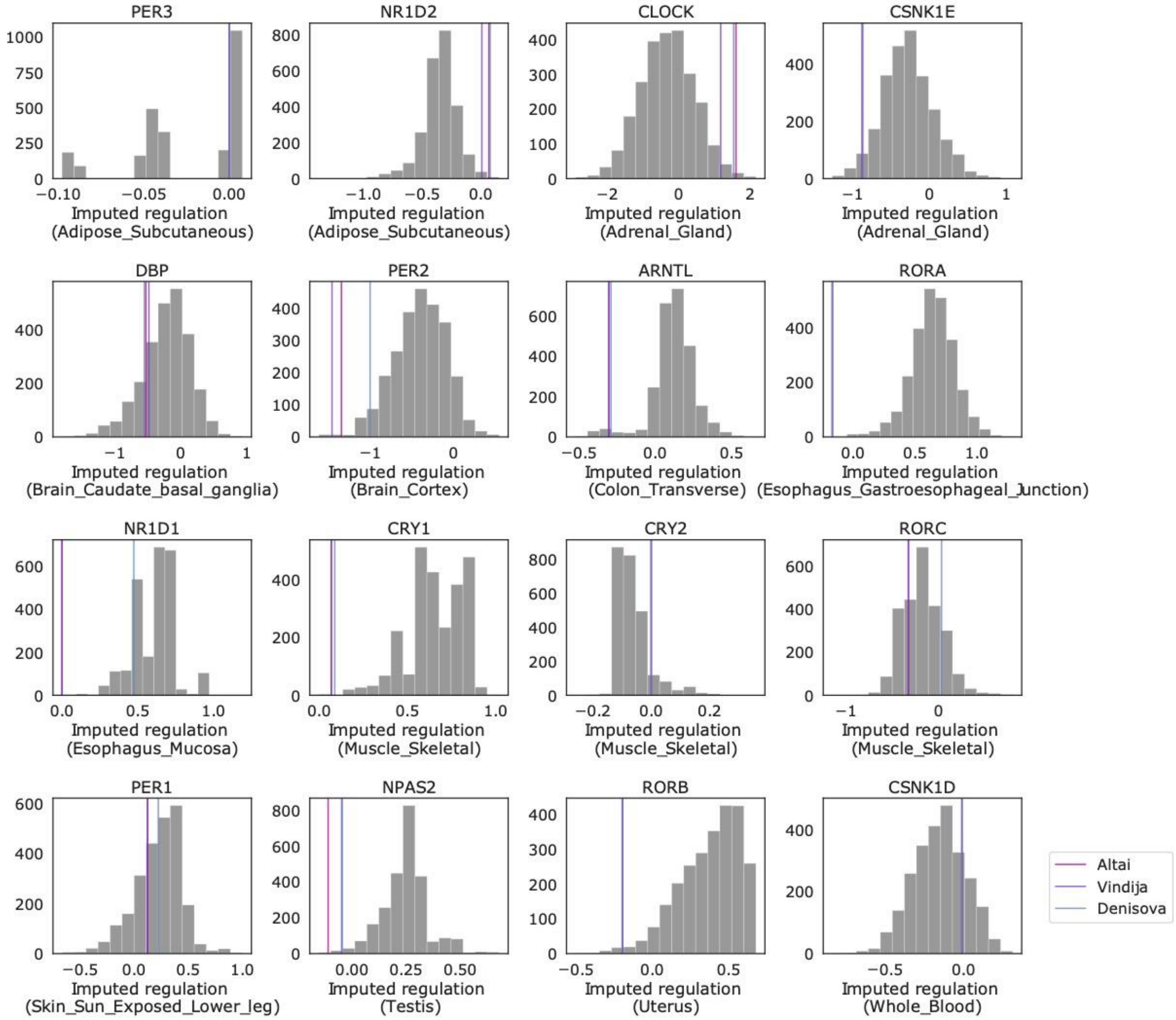


Figure 5.2.2. Comparison of imputed regulation for core circadian genes between modern humans and archaic hominins. Comparison of the imputed regulation of core circadian genes between 2504 humans in 1000 Genomes Phase 3 (gray bars) and three archaic individuals (vertical lines). For each core circadian gene, the tissue with the lowest average P-value for archaic difference from humans is plotted. Archaic gene regulation is at the extremes of the human distribution for several core genes: *CRY1*, *PER2*, *NPAS2, NR1D2*, *RORA*.

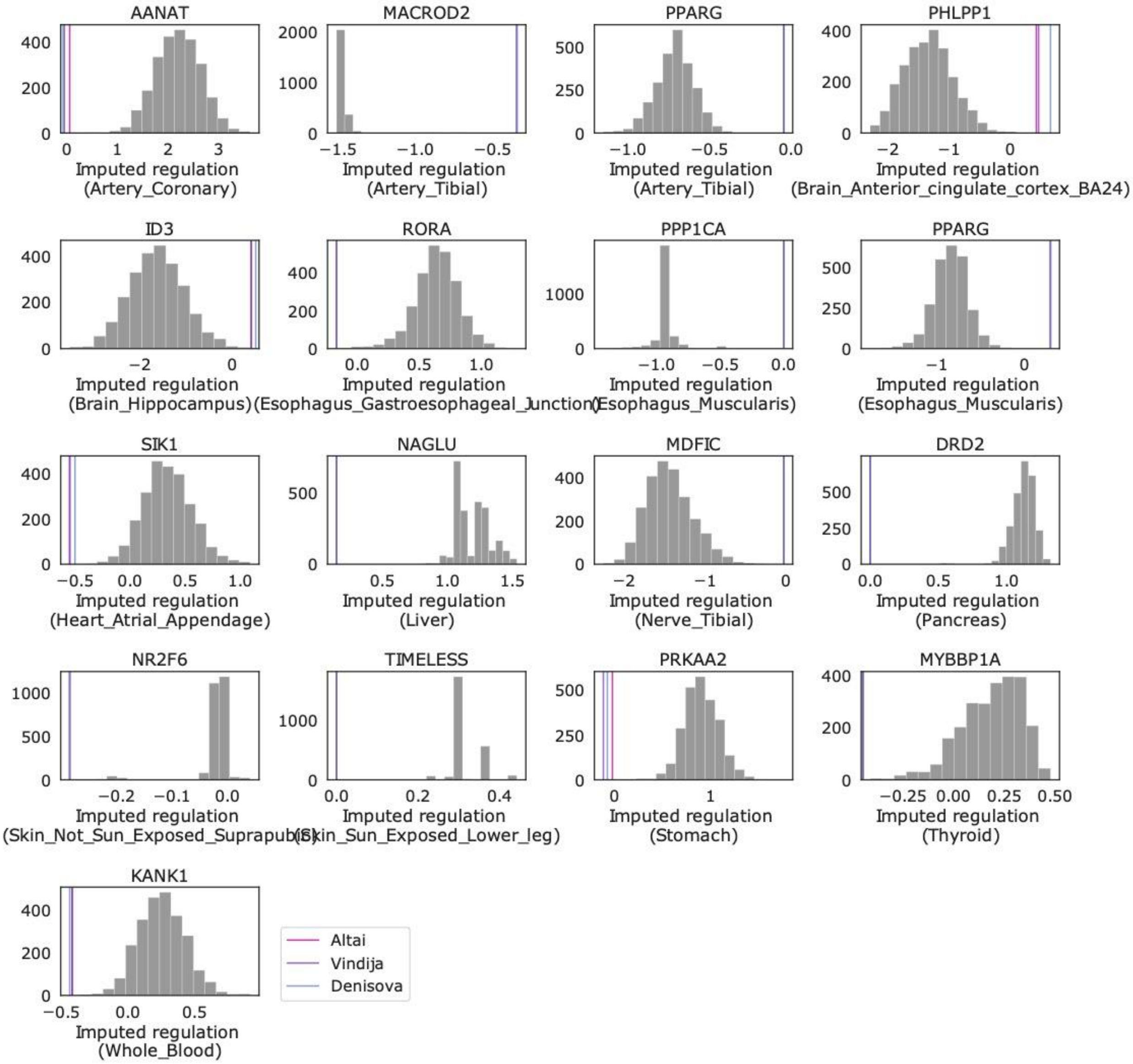


Figure 5.2.3. **Distributions of gene regulation predictions in all divergently regulated circadian genes.** Divergent regulation indicates that the archaic individuals (colored lines) each had imputed regulation more extreme than all 2504 modern humans from the 1000 Genomes Project (gray bars). For each gene, the tissue with largest average archaic difference from humans is plotted.

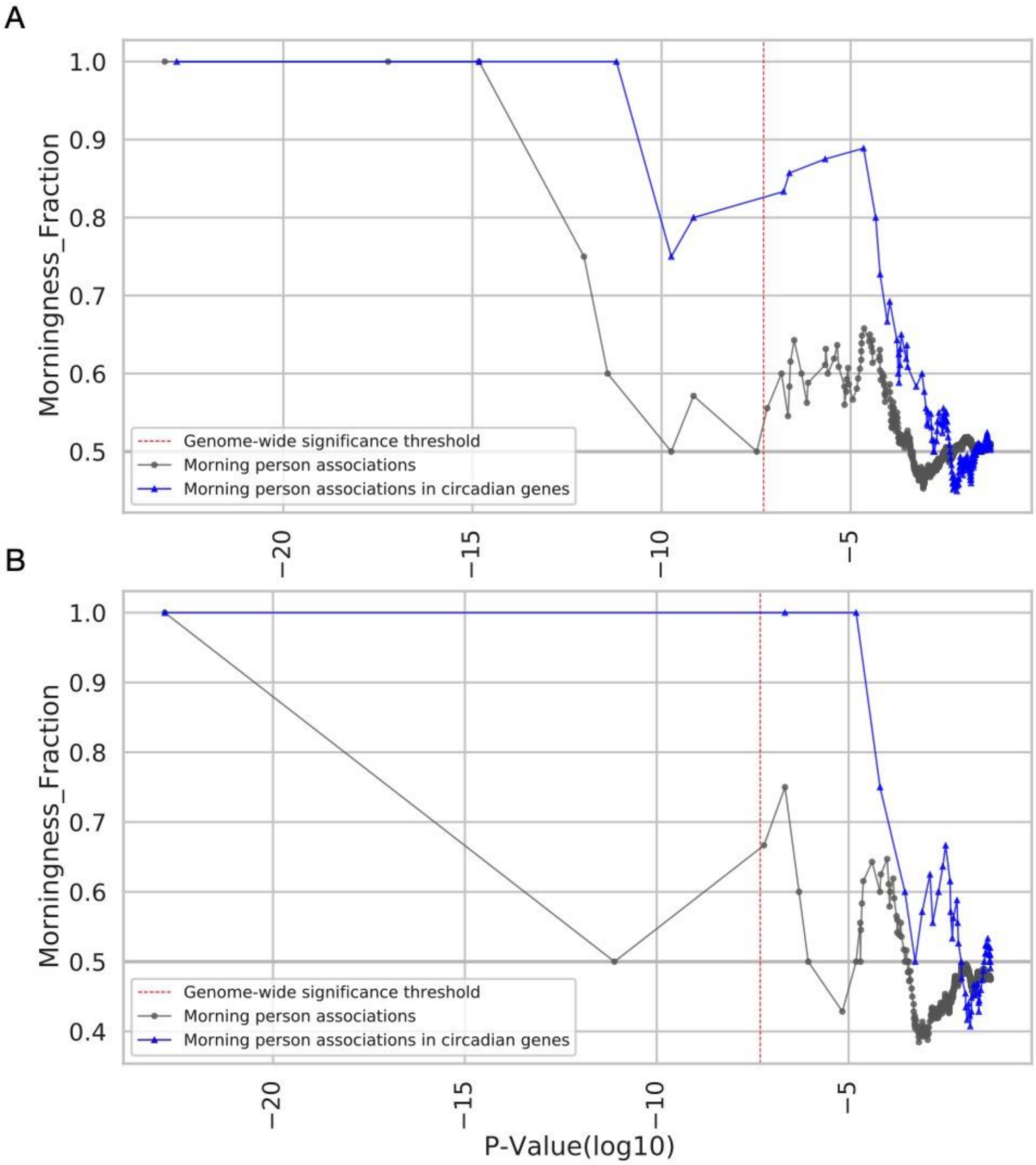
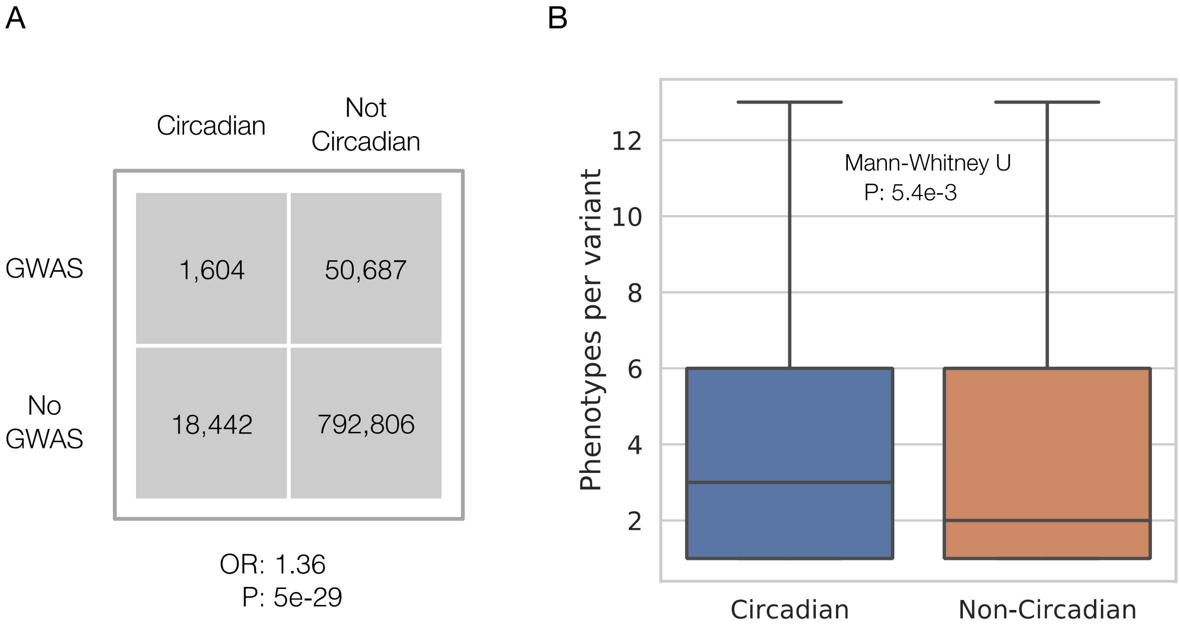
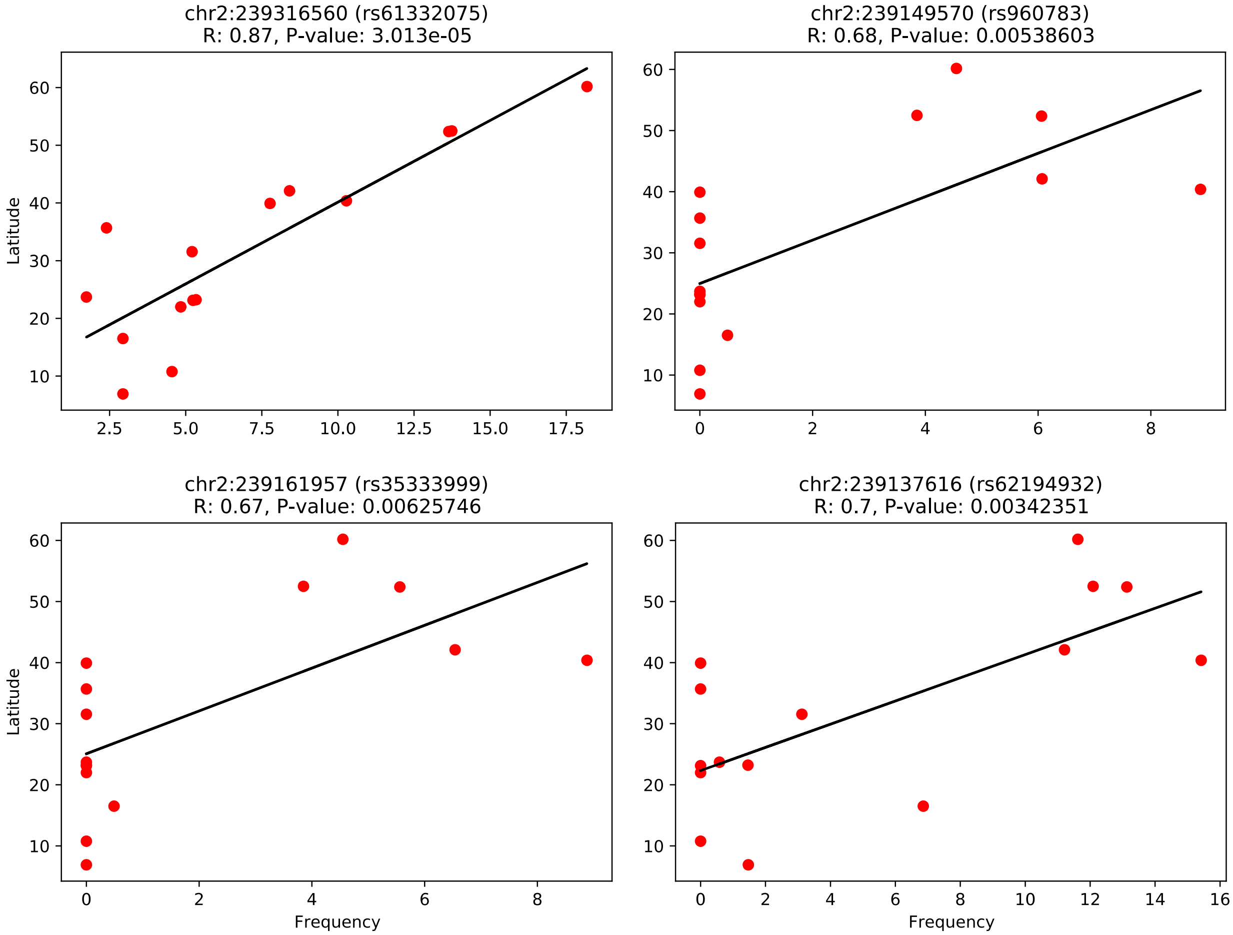


Figure 5.2.4. Introgressed variants associate with increased morningness. A) Cumulative fraction of morningness increasing variants reported as introgressed by Browning et al. 2018, and at least one other introgression detection method. B) Cumulative fraction of morningness increasing variants reported as introgressed by Browning et al. 2018 and five other introgression detection methods.



	Circadian	Not Circadian
GWAS	1,604	50,687
No GWAS	18,442	792,806

Figure 5.2.5. Pleiotropy in introgressed circadian variants. A) The relationship between GWAS associations and circadian introgressed variants versus non-circadian introgressed variants. The circadian set contains significantly more sites with at least one association than the non-circadian set (Fisher's exact: OR=1.36, P=5e-29). The GWAS associations were retrieved from Open Targets Genetics and were filtered by the genome-wise significance (P=5e-8). **B)** The distribution of phenotype associations per variant in the set of circadian and non-circadian introgressed variants. The circadian set shows significantly higher pleiotropy (Mann-Whitney U test: P=5.4e-3). Outliers (beyond 1.5 times the interquartile range) were removed for visualization.


chr2:239316560 (rs61332075)
R: 0.87, P-value: 3.013e-05

chr2:239149570 (rs960783)
R: 0.68, P-value: 0.00538603

chr2:239161957 (rs35333999)
R: 0.67, P-value: 0.00625746

chr2:239137616 (rs62194932)
R: 0.7, P-value: 0.00342351

Figure 5.2.6. Introgressed haplotypes associated with morningness follow a latitudinal cline. rs61332075 is the leading SNP in a large chromosome 2 haplotype that contains a SNP (rs75804782) previously reported to show a latitudinal cline in Eurasia. rs960783 and rs35333999 are tag SNPs in a nearby haplotype that also shows a latitudinal cline in Eurasia. rs62194932 is in moderate LD (R^2 of ~0.35 in EUR) with this haplotype and shows a similar latitudinal cline. Red dots represent 1000 Genomes Project populations of Eurasian ancestry.

Table 5.2.7. Genes containing variants predicted to be splice-altering by the SpliceAI method in the Altai Neanderthal (A), Vindija Neanderthal (V), Chagyrskaya Neanderthal (C), and Denisova (D).

GeneID	GeneName	Description	Locus	A	C	D	V
ENSG00000153064	BANK1	B cell scaffold protein with ankyrin repeats 1	chr4_102911885	0	0	1	0
ENSG00000158941	CCAR2	cell cycle and apoptosis regulator 2	chr8_22472265	0	0	0	1
ENSG00000107736	CDH23	cadherin related 23	chr10_73298326	1	1	1	1
			chr10_73405409	1	1	0	1
			chr10_73466969	0	1	0	0
			chr10_73487652	0	0	1	0
ENSG00000134852	CLOCK	clock circadian regulator	chr4_56296172	0	0	1	0
ENSG00000105662	CRTC1	CREB regulated transcription coactivator 1	chr19_18867740	1	1	0	1
ENSG00000057593	F7	coagulation factor VII	chr13_113769975	0	0	1	0
ENSG00000161040	FBXL13	F-box and leucine rich repeat protein 13	chr7_102667940	1	1	0	1
ENSG00000119771	KLHL29	kelch like family member 29	chr2_23805847	1	1	0	1
ENSG00000205213	LGR4	leucine rich repeat containing G protein-coupled receptor 4	chr11_27443527	0	0	1	0
ENSG00000005810	MYCBP2	MYC binding protein 2	chr13_77853722	0	0	1	0
ENSG00000140396	NCOA2	nuclear receptor coactivator 2	chr8_71180317	0	0	1	0
ENSG00000141027	NCOR1	nuclear receptor corepressor 1	chr17_16040743	0	0	1	0
ENSG00000064300	NGFR	nerve growth factor receptor	chr17_47581270	0	0	1	0
ENSG00000204640	NMS	neuromedin S	chr2_101097107	1	0	0	0
ENSG00000132326	PER2	period circadian regulator 2	chr2_239187089	0	0	0	1
ENSG00000149177	PTPRJ	protein tyrosine phosphatase receptor type J	chr11_48009909	0	0	0	1
			chr11_48041749	0	0	1	0
ENSG00000173482	PTPRM	protein tyrosine phosphatase receptor type M	chr18_7666317	0	0	0	1
			chr18_7666418	0	1	0	0
ENSG00000152061	RABGAP1L	RAB GTPase activating protein 1 like	chr1_174606762	0	1	0	0
			chr1_174876469	1	0	0	0
ENSG00000173933, ENSG00000173914	RBM4, RBM4B	RNA binding motif protein 4, RNA binding motif protein 4B	chr11_66433175	1	1	0	1
ENSG00000141576	RNF157	ring finger protein 157	chr17_74152940	1	1	0	1
ENSG00000134318	ROCK2	Rho associated coiled-coil containing protein kinase 2	chr2_11370990	0	0	1	0
			chr2_11484478	0	0	0	1
ENSG00000198963	RORB	RAR related orphan receptor B	chr9_77113671	0	0	0	1
ENSG00000143365	RORC	RAR related orphan receptor C	chr1_151796494	1	1	0	1
ENSG00000103546	SLC6A2	solute carrier family 6 member 2	chr16_55707020	0	0	1	0
ENSG00000108576	SLC6A4	solute carrier family 6 member 4	chr17_28547727	1	1	0	1
ENSG00000072310	SREBF1	sterol regulatory element binding transcription factor 1	chr17_17719515	1	0	0	0
			chr17_17719517	1	0	0	0
ENSG00000141510	TP53	tumor protein p53	chr17_7578263	0	0	1	0
ENSG00000140836	ZFHX3	zinc finger homeobox 3	chr16_72966622	0	1	0	1

Table 5.2.8. Genes predicted to be divergently regulated (DR) in the archaics from Altai (A), Vindija (V), and Denisova (D) by the PrediXcan method.

GeneID	GeneName	Description	GTEx Tissue	A	V	D
ENSG00000129673	AANAT	aralkylamine N-acetyltransferase	Artery_Coronary	1	1	1
ENSG00000174080	CTSF	cathepsin F	Adipose_Visceral_Omentum	1	0	1
			Brain_Cerebellar_Hemisphere	1	0	1
ENSG00000149295	DRD2	dopamine receptor D2	Pancreas	1	1	1
ENSG00000107485	GATA3	GATA binding protein 3	Skin_Sun_Exposed_Lower_leg	0	1	0
ENSG00000115738	ID2	inhibitor of DNA binding 2	Skin_Not_Sun_Exposed_Suprapubic	1	0	0
ENSG00000117318	ID3	inhibitor of DNA binding 3, HLH protein	Brain_Hippocampus	1	1	1
ENSG00000143772	ITPKB	inositol-trisphosphate 3-kinase B	Esophagus_Mucosa	1	0	0
ENSG00000107104	KANK1	KN motif and ankyrin repeat domains 1	Whole_Blood	1	1	1
ENSG00000172264	MACROD2	mono-ADP ribosylhydrolase 2	Artery_Tibial	1	1	1
ENSG00000135272	MDFIC	MyoD family inhibitor domain containing	Nerve_Tibial	1	1	1
ENSG00000132382	MYBBP1A	MYB binding protein 1a	Muscle_Skeletal	1	0	1
			Thyroid	1	1	1
ENSG00000108784	NAGLU	N-acetyl-alpha-glucosaminidase	Liver	1	1	1
ENSG00000126368	NR1D1	nuclear receptor subfamily 1 group D member 1	Adipose_Subcutaneous	0	1	0
			Esophagus_Mucosa	1	0	1
ENSG00000160113	NR2F6	nuclear receptor subfamily 2 group F member 6	Skin_Not_Sun_Exposed_Suprapubic	1	1	1
ENSG00000140538	NTRK3	neurotrophic receptor tyrosine kinase 3	Adipose_Visceral_Omentum	1	0	1
ENSG00000081913	PHLPP1	PH domain and leucine rich repeat protein phosphatase 1	Brain_Anterior_cingulate_cortex_BA24	1	1	1
ENSG00000132170	PPARG	peroxisome proliferator activated receptor gamma	Esophagus_Muscularis	1	1	1
			Artery_Tibial	1	1	1
ENSG00000172531	PPP1CA	protein phosphatase 1 catalytic subunit alpha	Esophagus_Muscularis	1	1	1
ENSG00000162409	PRKAA2	protein kinase AMP-activated catalytic subunit alpha 2	Stomach	1	1	1
ENSG00000069667	RORA	RAR related orphan receptor A	Esophagus_Gastroesophageal_Junction	1	1	1
ENSG00000142178	SIK1	salt inducible kinase 1	Heart_Atrial_Appendage	1	1	1
ENSG00000111602	TIMELESS	timeless circadian regulator	Skin_Sun_Exposed_Lower_leg	1	1	1
ENSG00000141510	TP53	tumor protein p53	Brain_Nucleus_accumbens_basal_ganglia	1	0	1
ENSG00000140836	ZFHX3	zinc finger homeobox 3	Cells_Cultured_fibroblasts	1	0	1
			Small_Intestine_Terminal_Ileum	0	1	1

Table 5.2.9. Counts of circadian genes expressed in each GTEx tissue (TPM >= 1).

Tissue	Gene_count	Percentage
Testis	206	83.74
Pituitary	197	80.08
Brain_Nucleus_accumbens_basal_ganglia	195	79.27
Brain_Hypothalamus	195	79.27
Brain_Frontal_Cortex_BA9	195	79.27
Brain_Cortex	195	79.27
Brain_Caudate_basal_ganglia	193	78.46
Brain_Anterior_cingulate_cortex_BA24	191	77.64
Lung	190	77.24
Brain_Amygdala	189	76.83
Brain_Putamen_basal_ganglia	188	76.42
Fallopian_Tube	187	76.02
Brain_Hippocampus	187	76.02
Prostate	187	76.02
Cervix_Endocervix	186	75.61
Vagina	186	75.61
Breast_Mammary_Tissue	185	75.2
Brain_Cerebellum	185	75.2
Cervix_Ectocervix	184	74.8
Small_Intestine_Terminal_Ileum	184	74.8
Brain_Substantia_nigra	184	74.8
Nerve_Tibial	183	74.39
Thyroid	183	74.39
Adipose_Subcutaneous	181	73.58
Artery_Coronary	180	73.17
Brain_Cerebellar_Hemisphere	179	72.76
Colon_Transverse	179	72.76
Bladder	179	72.76
Skin_Sun_Exposed_Lower_leg	178	72.36
Adipose_Visceral_Omentum	177	71.95
Ovary	176	71.54
Spleen	175	71.14
Artery_Tibial	175	71.14
Skin_Not_Sun_Exposed_Suprapubic	175	71.14
Uterus	175	71.14
Brain_Spinal_cord_cervical_c-1	175	71.14
Stomach	173	70.33
Esophagus_Gastroesophageal_Junction	172	69.92
Kidney_Medulla	172	69.92
Colon_Sigmoid	172	69.92
Adrenal_Gland	172	69.92
Minor_Salivary_Gland	171	69.51
Esophagus_Muscularis	171	69.51
Kidney_Cortex	170	69.11
Heart_Atrial_Appendage	170	69.11
Artery_Aorta	169	68.7
Esophagus_Mucosa	165	67.07
Pancreas	164	66.67
Liver	162	65.85
Heart_Left_Ventricle	158	64.23
Muscle_Skeletal	156	63.41
Cells_Cultured_fibroblasts	156	63.41
Cells_EBV-transformed_lymphocytes	150	60.98
Whole_Blood	141	57.32

Table 5.2.10. Enrichment analysis on the introgressed circadian variants with evidence of being eQTL in each GTEx tissue (Fisher's exact test).

Tissue	OR	P-value	Variants	log10(OR)	Bonferroni
Adipose_Subcutaneous	1.66152116	2.71E-52	1735	0.220505877	0.00102
Adipose_Visceral_Omentum	1.476860866	9.13E-28	1269	0.169339583	0.00102
Adrenal_Gland	1.702549549	1.63E-33	736	0.23109976	0.00102
Artery_Aorta	1.416931434	1.48E-21	1164	0.151348835	0.00102
Artery_Coronary	1.173064989	1.72E-03	469	0.069322073	<=0.05
Artery_Tibial	2.137431738	4.40E-117	1984	0.329892254	0.00102
Brain_Amygdala	1.563173117	9.38E-13	326	0.194007078	0.00102
Brain_Anterior_cingulate_cortex_BA24	1.367795929	4.51E-08	377	0.136021307	0.00102
Brain_Caudate_basal_ganglia	1.936875177	1.14E-52	798	0.287101633	0.00102
Brain_Cerebellar_Hemisphere	1.13396327	5.97E-03	603	0.054598988	<=0.05
Brain_Cerebellum	1.24382714	8.53E-08	827	0.094760029	0.00102
Brain_Cortex	1.787971179	8.41E-46	901	0.252360514	0.00102
Brain_Frontal_Cortex_BA9	1.465715751	6.62E-15	550	0.166049755	0.00102
Brain_Hippocampus	1.497305427	9.35E-14	441	0.175310399	0.00102
Brain_Hypothalamus	1.077370207	2.06E-01	331	0.032364961	>0.05
Brain_Nucleus_accumbens_basal_ganglia	1.048180951	3.56E-01	462	0.020436263	>0.05
Brain_Putamen_basal_ganglia	1.032429183	5.37E-01	410	0.013860272	>0.05
Brain_Spinal_cord_cervical_c-1	2.608684479	1.44E-75	563	0.416421554	0.00102
Brain_Substantia_nigra	0.436786642	1.14E-17	86	-0.359730652	0.00102
Breast_Mammary_Tissue	1.528807665	3.16E-29	1069	0.184352852	0.00102
Cells_Cultured_fibroblasts	1.647755942	1.91E-51	1820	0.216892886	0.00102
Cells_EBV-transformed_lymphocytes	0.531777911	6.51E-17	155	-0.274269707	0.00102
Colon_Sigmoid	1.330811081	3.46E-12	825	0.124116408	0.00102
Colon_Transverse	1.577020134	1.01E-33	1093	0.197837238	0.00102
Esophagus_Gastroesophageal_Junction	1.54633457	3.40E-31	1100	0.189303465	0.00102
Esophagus_Mucosa	1.010777799	7.61E-01	1146	0.004655695	>0.05
Esophagus_Muscularis	2.043355917	1.92E-100	1751	0.31034402	0.00102
Heart_Atrial_Appendage	2.231494768	5.88E-110	1361	0.348595873	0.00102
Heart_Left_Ventricle	1.593604706	3.49E-32	961	0.202380604	0.00102
Liver	1.350488287	1.54E-08	447	0.130490822	0.00102
Lung	2.335072493	2.13E-139	1765	0.368300368	0.00102
Minor_Salivary_Gland	1.571257565	4.34E-14	358	0.196247382	0.00102
Muscle_Skeletal	1.255375135	3.87E-11	1403	0.098773522	0.00102
Nerve_Tibial	0.90900835	5.29E-03	1357	-0.041432127	<=0.05
Ovary	2.225014847	2.48E-55	565	0.347332913	0.00102
Pancreas	1.433186205	4.62E-20	951	0.156302619	0.00102
Pituitary	0.879911135	5.73E-03	557	-0.055561187	<=0.05
Prostate	1.365834247	3.60E-11	592	0.135397998	0.00102
Skin_Not_Sun_Exposed_Suprapubic	1.326473329	2.59E-16	1405	0.122698522	0.00102
Skin_Sun_Exposed_Lower_leg	1.387255205	1.29E-22	1642	0.142156363	0.00102
Small_Intestine_Terminal_Ileum	1.165873368	4.21E-03	413	0.066651382	<=0.05
Spleen	0.685638014	4.39E-15	463	-0.163905111	0.00102
Stomach	1.289923197	3.51E-09	719	0.110563853	0.00102
Testis	1.772493272	5.46E-66	1755	0.248584595	0.00102
Thyroid	1.683584451	9.27E-57	2010	0.226234906	0.00102
Uterus	1.145211042	5.54E-02	227	0.058885527	>0.05
Vagina	1.106729697	1.53E-01	223	0.044041564	>0.05
Whole_Blood	1.603439669	1.03E-43	1546	0.205052624	0.00102

REFERENCES

1000 Genomes Project Consortium (2010a) ‘A map of human genome variation from population-scale sequencing’, *Nature*, 467(7319), p. 1061. Available at: https://doi.org/10.1038/nature09534.

1000 Genomes Project Consortium (2010b) ‘A map of human genome variation from population-scale sequencing’, *Nature*, 467(7319), p. 1061. Available at: https://doi.org/10.1038/nature09534.

1000 Genomes Project Consortium (2015) ‘A global reference for human genetic variation’, *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

Abi-Rached, L. *et al.* (2011) ‘The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans’, *Science*, 334(6052), pp. 89–94. Available at: https://doi.org/10.1126/science.1209202.

Allison, A.C. (1954) ‘The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria’, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 48(4), pp. 312–318. Available at: https://doi.org/10.1016/0035-9203(54)90101-7.

Álvarez-Lario, B. and Macarrón-Vicente, J. (2010) ‘Uric acid and evolution’, *Rheumatology*, 49(11), pp. 2010–2015. Available at: https://doi.org/10.1093/rheumatology/keq204.

Andrés, A.M. *et al.* (2009) ‘Targets of balancing selection in the human genome’, *Molecular biology and evolution*, 26(12), pp. 2755–2764. Available at: https://doi.org/10.1093/molbev/msp190.

Archer, S.N. *et al.* (2003) ‘A length polymorphism in the circadian clock gene Per3 is linked to delayed sleep phase syndrome and extreme diurnal preference’, *Sleep*, 26(4), pp. 413–415. Available at: https://doi.org/10.1093/sleep/26.4.413.

Arnold, L.J. *et al.* (2014) ‘Luminescence dating and palaeomagnetic age constraint on hominins from Sima de los Huesos, Atapuerca, Spain’, *Journal of Human Evolution*, 67(1), pp. 85–107. Available at: https://doi.org/10.1016/j.jhevol.2013.12.001.

Arsuaga, J.L. *et al.* (1993) ‘Three new human skulls from the Sima de los Huesos Middle Pleistocene site in Sierra de Atapuerca, Spain’, *Nature*, 362, pp. 534–537. Available at: https://doi.org/10.1038/362534a0.

Arsuaga, J.L., Martínez, I., Gracia, A., Carretero, J.M., *et al.* (1997) ‘Sima de los Huesos (Sierra de Atapuerca, Spain). The site’, *Journal of Human Evolution*, 33(2–3), pp. 109–127. Available at: https://doi.org/10.1006/jhev.1997.0132.

Arsuaga, J.L., Martínez, I., Gracia, A. and Lorenzo, C. (1997) ‘The Sima de los Huesos crania (Sierra de Atapuerca, Spain). A comparative study’, *Journal of Human Evolution*, 33(2–3), pp. 219–281. Available at: https://doi.org/10.1006/jhev.1997.0133.

Arsuaga, J.L. *et al.* (2000) ‘The Atapuerca human fossils’, *Human Evolution*, 15, pp. 75–82. Available at: https://doi.org/10.1007/BF02436236.

Asfaw, B. *et al.* (2002) ‘Remains of Homo erectus from Bouri, Middle Awash, Ethiopia’, *Nature*, 416(6878), pp. 317–320. Available at: https://doi.org/10.1038/416317a.

Azevedo, L. *et al.* (2015) ‘Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response’, *Human Genomics*, 9(1). Available at: https://doi.org/10.1186/s40246-015-0043-1.

Bae, C.J., Douka, K. and Petraglia, M.D. (2017) ‘On the origin of modern humans: Asian perspectives’, *Science*, 358(6368), p. eaai9067. Available at: https://doi.org/10.1126/science.aai9067.

Bailey, W.J. *et al.* (1991) ‘Molecular evolution of the Ψη-globin gene locus: Gibbon phylogeny and the hominoid slowdown’, *Molecular Biology and Evolution*, 8(2), pp. 155–184. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a040641.

Bailey, W.J. *et al.* (1992) ‘Reexamination of the African hominoid trichotomy with additional sequences from the primate β-globin gene cluster’, *Molecular Phylogenetics and Evolution*, 1(2), pp. 97–135. Available at: https://doi.org/10.1016/1055-7903(92)90024-B.

Baron, K.G. and Reid, K.J. (2014) ‘Circadian misalignment and health’, *International Review of Psychiatry*, 26(2), pp. 139–154. Available at: https://doi.org/10.3109/09540261.2014.911149.

Battivelli, E. *et al.* (2011) ‘Gag Cytotoxic T Lymphocyte Escape Mutations Can Increase Sensitivity of HIV-1 to Human TRIM5 , Linking Intrinsic and Acquired Immunity’, *Journal of Virology*, 85(22), pp. 11846–11854. Available at: https://doi.org/10.1128/jvi.05201-11.

Baum, L.E. *et al.* (1970) ‘A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains’, *The Annals of Mathematical Statistics*, 41(1), pp. 164–171. Available at: https://doi.org/10.1214/aoms/1177697196.

Baum, L.E. and Petrie, T. (1966) ‘Statistical Inference for Probabilistic Functions of Finite State Markov Chains’, *The Annals of Mathematical Statistics*, 37(6), pp. 1554–1563. Available at: https://doi.org/10.1214/aoms/1177699147.

Beauval, C. *et al.* (2005) ‘A late Neandertal femur from Les Rochers-de-Villeneuve, France’, *Proceedings of the National Academy of Sciences*, 102(20), pp. 7085–7090. Available at: https://doi.org/10.1073/pnas.0502656102.

Belmont, J.W. *et al.* (2005) ‘A haplotype map of the human genome’, *Nature*, 437(7063), pp. 1299–1320. Available at: https://doi.org/10.1038/nature04226.

Benazzi, S. *et al.* (2011) ‘Early dispersal of modern humans in Europe and implications for Neanderthal behaviour’, *Nature*, 479(7374), pp. 525–528. Available at: https://doi.org/10.1038/nature10617.

Benazzi, S. *et al.* (2015) ‘The makers of the Protoaurignacian and implications for Neandertal extinction’, *Science*, 348(6236), pp. 793–796. Available at: https://doi.org/10.1126/science.aaa2773.

Benton, M.L. *et al.* (2019) ‘Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function’, *BMC Genomics*, 20(1), pp. 1–22. Available at: https://doi.org/10.1186/s12864-019-5779-x.

Benton, M.L. *et al.* (2021) ‘The influence of evolutionary history on human health and disease’, *Nature Reviews Genetics*. Nature Research, pp. 269–283. Available at: https://doi.org/10.1038/s41576-020-00305-9.

Bergland, A.O. *et al.* (2014) ‘Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila’, *PLoS Genetics*, 10(11). Available at: https://doi.org/10.1371/journal.pgen.1004775.

Bermúdez de Castro, J.M. *et al.* (2004) ‘The Atapuerca Sites and Their Contribution to the Knowledge of Human Evolution in Europe’, *Evolutionary Anthropology*, 13(1), pp. 25–41. Available at: https://doi.org/10.1002/evan.10130.

Bermúdez de Castro, J.M. *et al.* (2021) ‘The Sima de los Huesos Middle Pleistocene hominin site (Burgos, Spain). Estimation of the number of individuals’, *Anatomical Record*, 304(7), pp. 1463–1477. Available at: https://doi.org/10.1002/ar.24551.

Bermúdez de Castro, J.M. and Rosas, A. (1992) ‘A human mandibular fragment from the Atapuerca Trench (Burgos, Spain)’, *Journal of Human Evolution*, 22(1), pp. 41–46. Available at: https://doi.org/10.1016/0047-2484(92)90028-8.

Bischoff, J.L. *et al.* (2003) ‘The Sima de los Huesos hominids date to beyond U/Th equilibrium (>350 kyr) and perhaps to 400-500 kyr: New radiometric dates’, *Journal of Archaeological Science*, 30(3), pp. 275–280. Available at: https://doi.org/10.1006/jasc.2002.0834.

Bitarello, B.D. *et al.* (2018) ‘Signatures of long-term balancing selection in human genomes’, *Genome Biology and Evolution*, 10(3), pp. 939–955. Available at: https://doi.org/10.1093/gbe/evy054.

Black, F.L. and Hedrick, P.W. (1997) ‘Strong balancing selection at HLA loci: Evidence from segregation in South Amerindian families’, *Proceedings of the National Academy of Sciences of the United States of America*, 94(23), pp. 12452–12456. Available at: https://doi.org/10.1073/pnas.94.23.12452.

Blume, C., Garbazza, C. and Spitschan, M. (2019) ‘Effects of light on human circadian rhythms, sleep and mood’, *Somnologie*. Dr. Dietrich Steinkopff Verlag GmbH and Co. KG, pp. 147–156. Available at: https://doi.org/10.1007/s11818-019-00215-x.

Bobe, R. and Behrensmeyer, A.K. (2004) ‘The expansion of grassland ecosystems in Africa in relation to mammalian evolution and the origin of the genus Homo’, *Palaeogeography, Palaeoclimatology, Palaeoecology*, 207(3–4), pp. 399–420. Available at: https://doi.org/10.1016/j.palaeo.2003.09.033.

Bobe, R. and Eck, G.G. (2001) ‘Responses of African bovids to Pliocene climatic change’, *Paleobiology*, 27, pp. 1–47. Available at: https://doi.org/10.1666/0094-8373(2001)027<0001:roabtp>2.0.co;2.

Bokelmann, L. *et al.* (2019) ‘A genetic analysis of the Gibraltar Neanderthals’, *Proceedings of the National Academy of Sciences*, 116(31), pp. 15610–15615. Available at: https://doi.org/10.1073/pnas.1903984116.

Bonner, T.I., Heinemann, R. and Todaro, G.J. (1980) ‘Evolution of DNA sequences has been retarded in malagasy primates’, *Nature*, 286(5771), pp. 420–423. Available at: https://doi.org/10.1038/286420a0.

Bons, P.D. *et al.* (2019) ‘Out of Africa by spontaneous migration waves’, *PLOS ONE*. Edited by Q. Ayub, 14(4), p. e0201998. Available at: https://doi.org/10.1371/journal.pone.0201998.

Brand, C.M., Colbran, L.L. and Capra, J.A. (2023) ‘Resurrecting the alternative splicing landscape of archaic hominins using machine learning’, *Nature Ecology and Evolution*, 7(6), pp. 939–953. Available at: https://doi.org/10.1038/s41559-023-02053-5.

Briggs, A.W. *et al.* (2010) ‘Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA’, *Nucleic Acids Research*, 38(6), p. 87. Available at: https://doi.org/10.1093/nar/gkp1163.

Brown, F. *et al.* (1985) ‘Early Homo erectus skeleton from west Lake Turkana, Kenya’, *Nature*, 316(6031), pp. 788–792. Available at: https://doi.org/10.1038/316788a0.

Brown, S.A. *et al.* (2008) ‘Molecular insights into human daily behavior’, *Proceedings of the National Academy of Sciences of the United States of America*, 105(5), pp. 1602–1607. Available at: https://doi.org/10.1073/pnas.0707772105.

Browning, S.R. *et al.* (2018) ‘Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture’, *Cell*, 173(1), pp. 53–61. Available at: https://doi.org/https://doi.org/10.1016/j.cell.2018.02.031.

Brunet, M. *et al.* (2002) ‘A new hominid from the Upper Miocene of Chad, Central Africa’, *Nature*, 418(August), pp. 145–151.

Brzezinski, J.A. *et al.* (2005) ‘Loss of circadian photoentrainment and abnormal retinal electrophysiology in Math5 mutant mice’, *Investigative Ophthalmology and Visual Science*, 46(7), pp. 2540–2551. Available at: https://doi.org/10.1167/iovs.04-1123.

Bush, W.S., Oetjens, M.T. and Crawford, D.C. (2016) ‘Unravelling the human genome-phenome relationship using phenome-wide association studies’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 129–145. Available at: https://doi.org/10.1038/nrg.2015.36.

Cagliani, R. *et al.* (2010) ‘Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene’, *Human Genetics*, 128(6), pp. 577–588. Available at: https://doi.org/10.1007/s00439-010-0884-6.

Cagliani, R. *et al.* (2012) ‘A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis’, *Molecular Biology and Evolution*, 29(6), pp. 1599–1613. Available at: https://doi.org/10.1093/molbev/mss002.

Capasso, L., Michetti, E. and D’Anastasio, R. (2008) ‘A homo erectus hyoid bone: Possible implications for the origin of the human capability for speech’, *Collegium Antropologicum*, 32(4), pp. 1004–1011.

Caramelli, D. *et al.* (2006) ‘A highly divergent mtDNA sequence in a Neandertal individual from Italy’, *Current Biology*, 16(16), pp. 630–632. Available at: https://doi.org/10.1016/j.cub.2006.07.043.

Carbonell, E. *et al.* (1999) *Atapuerca: Ocupaciones Humanas y Paleoecología del Yacimiento de Galería*. Junta de Castilla y Leon Consejeria de Educacion y Cultura.

Carithers, L.J. *et al.* (2015) ‘A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project’, *Biopreservation and Biobanking*, 13(5), pp. 311–317. Available at: https://doi.org/10.1089/bio.2015.0032.

Cerling, T.E. (1992) ‘Development of grasslands and savannas in East Africa during the Neogene’, *Palaeogeography, Palaeoclimatology, Palaeoecology*, 97(3), pp. 241–247. Available at: https://doi.org/10.1016/0921-8181(92)90013-Z.

Cerling, T.E. and Hay, R.L. (1986) ‘An isotopic study of paleosol carbonates from Olduvai Gorge’, *Quaternary Research*, 25(1), pp. 63–78. Available at: https://doi.org/10.1016/0033-5894(86)90044-X.

Cervera, J. *et al.* (1998) *Atapuerca. Un millón de años de historia*. Madrid: Complutense S A.

Chan, E.K.F. *et al.* (2019) ‘Human origins in a southern African palaeo-wetland and first migrations’, *Nature*, 575(7781), pp. 185–189. Available at: https://doi.org/10.1038/s41586-019-1714-1.

Charlesworth, B., Nordborg, M. and Charlesworth, D. (1997) ‘The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations’, *Genetics*, 70(2), pp. 155–174. Available at: https://doi.org/10.1534/genetics.115.178558.

Charlesworth, D. (2006) ‘Balancing selection and its effects on sequences in nearby genome regions’, *PLoS Genetics*, 2(4), pp. 379–384. Available at: https://doi.org/10.1371/journal.pgen.0020064.

Chen, B.D. *et al.* (2017) ‘TT genotype of rs2941484 in the human HNF4G gene is associated with hyperuricemia in Chinese Han men’, *Oncotarget*, 8(16), pp. 26918–26926. Available at: https://doi.org/10.18632/oncotarget.15851.

Chen, F. *et al.* (2019) ‘A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau’, *Nature*, 569(7756), pp. 409–412. Available at: https://doi.org/10.1038/s41586-019-1139-x.

Chen, L. *et al.* (2020) ‘Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals’, *Cell*, 180(4), pp. 677–687. Available at: https://doi.org/10.1016/j.cell.2020.01.012.

Cheng, X. and DeGiorgio, M. (2019) ‘Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism’, *Molecular Biology and Evolution*, 36(1), pp. 177–199. Available at: https://doi.org/10.1093/molbev/msy202.

Cheung, V.G. *et al.* (2003) ‘Natural variation in human gene expression assessed in lymphoblastoid cells’, *Nature Genetics*, 33(3), pp. 422–425. Available at: https://doi.org/10.1038/ng1094.

Chimpanzee Sequencing and Analysis Consortium (2005) ‘Initial sequence of the chimpanzee genome and comparison with the human genome’, *Nature*, 437(7055), pp. 69–87. Available at: https://doi.org/10.1038/nature04072.

Clarkson, C. *et al.* (2017) ‘Human occupation of northern Australia by 65,000 years ago’, *Nature*, 547(7663), pp. 306–310. Available at: https://doi.org/10.1038/nature22968.

Colbran, L.L. *et al.* (2019) ‘Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences’, *Nature Ecology & Evolution*, 3(November), pp. 1598–1606. Available at: https://doi.org/10.1038/s41559-019-0996-x.

Currat, M. and Excoffier, L. (2004) ‘Modern humans did not admix with Neanderthals during their range expansion into Europe’, *PLoS Biology*, 2(12), p. e421. Available at: https://doi.org/10.1371/journal.pbio.0020421.

Dabney, J., Meyer, M. and Pääbo, S. (2013) ‘Ancient DNA damage’, *Cold Spring Harbor Perspectives in Biology*, 5(7), p. a012567. Available at: https://doi.org/10.1101/cshperspect.a012567.

Dannemann, M., Andrés, A.M. and Kelso, J. (2016) ‘Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors’, *American Journal of Human Genetics*, 98(1), pp. 22–23. Available at: https://doi.org/10.1016/j.ajhg.2015.11.015.

Dannemann, M. and Kelso, J. (2017) ‘The Contribution of Neanderthals to Phenotypic Variation in Modern Humans’, *American Journal of Human Genetics*, 101(4), pp. 578–589. Available at: https://doi.org/10.1016/j.ajhg.2017.09.010.

Dart, R.A. (1925) ‘Australopithecus africanus: The Man-Ape of South Africa’, *Nature*, 115, pp. 195–199. Available at: https://doi.org/10.1038/115195a0.

Dawkins, R. (1976) ‘The Selfish Gene’, in. Oxford University Press, pp. 192–195.

De Deckker, P. *et al.* (2019) ‘Marine Isotope Stage 4 in Australasia: A full glacial culminating 65,000 years ago – Global connections and implications for human dispersal’, *Quaternary Science Reviews*, 204, pp. 187–207. Available at: https://doi.org/10.1016/j.quascirev.2018.11.017.

DeGiorgio, M., Lohmueller, K.E. and Nielsen, R. (2014) ‘A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data’, *PLoS Genetics*, 10(8), p. e1004561. Available at: https://doi.org/10.1371/journal.pgen.1004561.

deMenocal, P.B. (2004) ‘African climate change and faunal evolution during the Pliocene-Pleistocene’, *Earth and Planetary Science Letters*, 220(1–2), pp. 3–24. Available at: https://doi.org/10.1016/S0012-821X(04)00003-2.

DeMenocal, P.B. (2011) ‘Climate and human evolution’, *Science*, 331(6017), pp. 540–541. Available at: https://doi.org/10.1126/science.1190683.

DeMenocal, P.B. and Bloemendal, J. (1995) ‘Plio-Pleistocene climatic variability in subtropical Africa and the paleoenvironment of hominid evolution’, in *Paleoclimate and Evolution, with Emphasis on Human Origins*. Yale University Press, pp. 262–288.

Demuro, M. *et al.* (2014) ‘New luminescence ages for the Galería Complex archaeological site: Resolving chronological uncertainties on the Acheulean record of the Sierra de Atapuerca, Northern Spain’, *PLoS ONE*, 9(10), p. e110169. Available at: https://doi.org/10.1371/journal.pone.0110169.

Dobzhansky, T. (1950) ‘Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudoobscura.’, *Genetics*, 35(3), pp. 288–302. Available at: https://doi.org/10.1093/genetics/35.3.288.

Dorokhov, V.B. *et al.* (2018) ‘An hour in the morning is worth two in the evening: association of morning component of morningness – eveningness with single nucleotide polymorphisms in circadian clock genes’, *Biological Rhythm Research*, 49(4), pp. 622–642. Available at: https://doi.org/10.1080/09291016.2017.1390823.

Dudkiewicz, M., Lenart, A. and Pawłowski, K. (2013) ‘A Novel Predicted Calcium-Regulated Kinase Family Implicated in Neurological Disorders’, *PLoS ONE*, 8(6). Available at: https://doi.org/10.1371/journal.pone.0066427.

Durand, E.Y. *et al.* (2011) ‘Testing for ancient admixture between closely related populations’, *Molecular Biology and Evolution*, 28(8), pp. 2239–2252. Available at: https://doi.org/10.1093/molbev/msr048.

Elguero, E. *et al.* (2015) ‘Malaria continues to select for sickle cell trait in Central Africa’, *Proceedings of the National Academy of Sciences of the United States of America*, 112(22), pp. 7051–7054. Available at: https://doi.org/10.1073/pnas.1505665112.

Equipo Investigador de Atapuerca (2022) *Encuentran en Atapuerca la cara del Primer Europeo*, *Fundación Atapuerca*. Available at: https://www.atapuerca.org/es/ficha/ZB0E4E1F1-AF49-B47F-68B4E90B14BC706A/encuentran-en-atapuerca-la-cara-del-primer-europeo (Accessed: 31 May 2023).

Falguères, C. *et al.* (2013) ‘Combined ESR/U-series chronology of Acheulian hominid-bearing layers at Trinchera Galería site, Atapuerca, Spain’, *Journal of Human Evolution*, 65(2), pp. 168–184. Available at: https://doi.org/10.1016/j.jhevol.2013.05.005.

Falk, D. (1980) ‘A reanalysis of the South African australopithecine natural endocasts’, *American Journal of Physical Anthropology*, 53(4), pp. 525–539. Available at: https://doi.org/10.1002/ajpa.1330530409.

Ferguson, W.W. (1989) ‘A new species of the genus Australopithecus (primates: hominidae) from Plio/Pleistocene deposits west of Lake Turkana in Kenya’, *Primates*, 30(2), pp. 223–232. Available at: https://doi.org/10.1007/BF02381307.

Fernández-Jalvo, Y. *et al.* (1973) ‘Evidence of early cannibalism’, (51).

De Filippo, C. *et al.* (2016) ‘Recent Selection Changes in Human Genes under Long-Term Balancing Selection’, *Molecular Biology and Evolution*, 33(6), pp. 1435–1447. Available at: https://doi.org/10.1093/molbev/msw023.

Finlayson, C. *et al.* (2006) ‘Late survival of Neanderthals at the southernmost extreme of Europe’, *Nature*, 443(7113), pp. 850–853. Available at: https://doi.org/10.1038/nature05195.

Foster, R.G. and Roenneberg, T. (2008) ‘Human Responses to the Geophysical Daily, Annual and Lunar Cycles’, *Current Biology*, 18(17), pp. 784–794. Available at: https://doi.org/10.1016/j.cub.2008.07.003.

Fu, Q. *et al.* (2013) ‘DNA analysis of an early modern human from Tianyuan Cave, China’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), pp. 2223–2227. Available at: https://doi.org/10.1073/pnas.1221359110.

Fu, Q. *et al.* (2014) ‘Genome sequence of a 45,000-year-old modern human from western Siberia’, *Nature*, 514(7523), pp. 445–449. Available at: https://doi.org/10.1038/nature13810.

Fu, Q. *et al.* (2015) ‘An early modern human from Romania with a recent Neanderthal ancestor’, *Nature*, 524, p. 216. Available at: https://doi.org/10.1038/nature14558.

Fu, Q. *et al.* (2016) ‘The genetic history of Ice Age Europe’, *Nature*, 534(7606), pp. 200–205. Available at: https://doi.org/10.1038/nature17993.

Fu, Y.-X. and Li, W.-H. (1993) ‘Statistical tests of neutrality of mutations’, *Genetics*, 133(3), pp. 693–709. Available at: https://doi.org/10.1093/genetics/133.3.693.

Gagneux, P. and Varki, A. (2001) ‘Genetic differences between humans and great apes’, *Molecular Phylogenetics and Evolution*, 18(1), pp. 2–13. Available at: https://doi.org/10.1006/mpev.2000.0799.

Gan, Y. *et al.* (2015) ‘Shift work and diabetes mellitus: A meta-analysis of observational studies’, *Occupational and Environmental Medicine*, pp. 72–78. Available at: https://doi.org/10.1136/oemed-2014-102150.

Gan, Y. *et al.* (2018) ‘Association between shift work and risk of prostate cancer: A systematic review and meta-analysis of observational studies’, *Carcinogenesis*, pp. 87–97. Available at: https://doi.org/10.1093/carcin/bgx129.

Ganser-Pornillos, B.K. and Pornillos, O. (2019) ‘Restriction of HIV-1 and other retroviruses by TRIM5’, *Nature Reviews Microbiology*. Nature Publishing Group, pp. 546–556. Available at: https://doi.org/10.1038/s41579-019-0225-2.

Gao, Z., Przeworski, M. and Sella, G. (2015) ‘Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species’, *Evolution*, 69(2). Available at: https://doi.org/10.1111/evo.12567.

Gilbert, E. *et al.* (2017) ‘The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland’, *Scientific Reports* [Preprint]. Available at: https://doi.org/10.1038/s41598-017-17124-4.

Gillespie, J.H. and Langley, C.H. (1974) ‘A general model to account for enzyme variation in natural populations’, *Genetics*, 76(4), pp. 837–848. Available at: https://doi.org/10.1093/genetics/76.4.837.

Gómez-Robles, A. (2019) ‘Dental evolutionary rates and its implications for the Neanderthal–modern human divergence’, *Science Advances*, 5(5), p. eaaw1268. Available at: https://doi.org/10.1126/sciadv.aaw1268.

Gong, L. *et al.* (2013) ‘Biochemical and immunological mechanisms by which sickle cell trait protects against malaria’, *Malaria Journal*, 12(1), pp. 1–9. Available at: https://doi.org/10.1186/1475-2875-12-317.

Goodman, M. *et al.* (1998) ‘Toward a Phylogenetic Classification of Primates Based on DNA Evidence Complemented by Fossil Evidence’, *Molecular Phylogenetics and Evolution*, 9(3), pp. 585–598. Available at: https://doi.org/10.1006/mpev.1998.0495.

Gower, G. *et al.* (2021) ‘Detecting adaptive introgression in human evolution using convolutional neural networks’, *eLife*, 10, p. e64669. Available at: https://doi.org/10.7554/eLife.64669.

Green, R.E. *et al.* (2010) ‘A Draft Sequence of the Neandertal Genome’, *Science*, 328(5979), pp. 710–722. Available at: https://doi.org/10.1126/science.1188021.

Grimaldi, B. *et al.* (2010) ‘PER2 controls lipid metabolism by direct regulation of PPARγ’, *Cell Metabolism*, 12(5), pp. 509–520. Available at: https://doi.org/10.1016/j.cmet.2010.10.005.

Gromko, M.H. (1977) ‘What is Frequency-Dependent Selection?’, *Evolution*, 31(2), pp. 438–442. Available at: https://doi.org/10.2307/2407763.

Groucutt, H.S. *et al.* (2015) ‘Rethinking the dispersal of Homo sapiens out of Africa’, *Evolutionary Anthropology: Issues, News, and Reviews*, 24(4), pp. 149–164. Available at: https://doi.org/10.1002/evan.21455.

Grün, R. (1996) ‘A re-analysis of electron spin resonance dating results associated with the Petralona hominid’, *Journal of Human Evolution*, 30(3), pp. 227–241.

Grün, R. *et al.* (2005) ‘U-series and ESR analyses of bones and teeth relating to the human burials from Skhul’, *Journal of Human Evolution*, 49(3), pp. 316–334. Available at: https://doi.org/10.1016/j.jhevol.2005.04.006.

Grün, R. *et al.* (2020) ‘Dating the skull from Broken Hill, Zambia, and its position in human evolution’, *Nature*, 580, pp. 372–375. Available at: https://doi.org/10.1038/s41586-020-2165-4.

Gustafsson, D. and Unwin, R. (2013) ‘The pathophysiology of hyperuricaemia and its possible relationship to cardiovascular disease, morbidity and mortality’, *BMC Nephrology*. Available at: https://doi.org/10.1186/1471-2369-14-164.

Gyarmati, G. *et al.* (2016) ‘Night shift work and stomach cancer risk in the MCC-Spain study’, *Occupational and Environmental Medicine*, 73(8), pp. 520–527. Available at: https://doi.org/10.1136/oemed-2016-103597.

Hajdinjak, M. *et al.* (2018) ‘Reconstructing the genetic history of late Neanderthals’, *Nature*, 555, pp. 652–659. Available at: https://doi.org/10.1038/nature26151.

Hajdinjak, M. *et al.* (2021) ‘Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry’, *Nature*, 592(7853), pp. 253–257. Available at: https://doi.org/10.1038/s41586-021-03335-3.

Haldane, J.B.S. (1949) ‘The Rate of Mutation of Human Genes’, *Hereditas*, 35(S1), pp. 267–273. Available at: https://doi.org/10.1111/j.1601-5223.1949.tb03339.x.

Haldane, J.B.S. (1957) ‘The cost of natural selection’, *Journal of Genetics*, 55, pp. 511–524. Available at: https://doi.org/10.1007/BF02984069.

Harris, K. and Nielsen, R. (2016) ‘The genetic cost of neanderthal introgression’, *Genetics*, 203(2), pp. 881–891. Available at: https://doi.org/10.1534/genetics.116.186890.

Hedrick, P.W. (1972) ‘Maintenance of genetic variation with a frequency-dependent selection model as compared to the overdominant model’, *Genetics*, 72(4), pp. 771–775. Available at: https://doi.org/10.1093/genetics/72.4.771.

Hedrick, P.W. (2011) ‘Population genetics of malaria resistance in humans’, *Heredity*, 107(4), pp. 283–304. Available at: https://doi.org/10.1038/hdy.2011.16.

Hedrick, P.W. and Thomson, G. (1983) ‘Evidence for balancing selection at HLA’, *Genetics*, 104(3), pp. 449–456. Available at: https://doi.org/10.1093/genetics/104.3.449.

Hellmann, I. *et al.* (2003) ‘A neutral explanation for the correlation of diversity with recombination rates in humans’, *American Journal of Human Genetics*, 72(6), pp. 1527–1535. Available at: https://doi.org/10.1086/375657.

Hershkovitz, I. *et al.* (2018) ‘The earliest modern humans outside Africa’, *Paleoanthropology*, 459, pp. 456–459. Available at: https://doi.org/10.1126/science.aap8369.

Hey, J. (1991) ‘A multi-dimensional coalescent process applied to multi-allelic selection models and migration models’, *Theoretical Population Biology*, 39, pp. 30–48. Available at: https://doi.org/10.1016/0040-5809(91)90039-I.

Higham, T. *et al.* (2014) ‘The timing and spatiotemporal patterning of Neanderthal disappearance’, *Nature*, 512(7514), pp. 306–309. Available at: https://doi.org/10.1038/nature13621.

Hillert, D. (2021) ‘How did language evolve in the lineage of higher primates?’, *Lingua*, 264, p. 103158. Available at: https://doi.org/10.1016/j.lingua.2021.103158.

Hockings, K.J. *et al.* (2015) ‘Tools to tipple: Ethanol ingestion by wild chimpanzees using leaf-sponges’, *Royal Society Open Science*, 2(6). Available at: https://doi.org/10.1098/rsos.150150.

Hodgson, J.A. and Disotell, T.R. (2008) ‘No evidence of a Neanderthal contribution to modern human diversity’, *Genome Biology*, 9(2), p. 206. Available at: https://doi.org/10.1186/gb-2008-9-2-206.

Hofmanová, Z. *et al.* (2016) ‘Early farmers from across Europe directly descended from Neolithic Aegeans’, *Proceedings of the National Academy of Sciences*, 113(25), pp. 6886–6891. Available at: https://doi.org/10.1073/pnas.1523951113.

Holloway, R.L. (1983) ‘Human paleontological evidence relevant to language behavior’, *Human neurobiology*, 2(3), pp. 105–114.

Holloway, R.L. *et al.* (2009) ‘Evolution of the Brain in Humans – Paleoneurology’, *Encyclopedia of Neuroscience*. Available at: https://doi.org/10.1007/978-3-540-29678-2_3152.

Holloway, R.L. (2015) ‘The Evolution of the Hominid Brain’, *Handbook of Paleoanthropology,* 3. Available at: https://doi.org/10.1007/978-3-642-39979-4.

Höss, M. and Pääbo, S. (1993) ‘DNA extraction from Pleistocene bones by a silica-based purification method’, *Nucleic Acids Research*, 21(16), pp. 3913–3914. Available at: https://academic.oup.com/nar/article/21/16/3913/2386436.

Howard Levene (1953) ‘Genetic Equilibrium When More Than One Ecological Niche is Available’, *American Society of Naturalists*, 87(836), pp. 331–333.

Huang, S.L., Singh, M. and Kojima, K.-I. (1971) ‘A study of frequency-dependent selection observed in the esterase-6 locus of Drosophila melanogaster using a conditioned media method’, *Genetics*, 68(1), pp. 97–104. Available at: https://doi.org/10.1093/genetics/68.1.97.

Hublin, J.J. *et al.* (2017) ‘New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens’, *Nature*, 546(7657), pp. 289–292. Available at: https://doi.org/10.1038/nature22336.

Hublin, J.-J. (2020) ‘Initial Upper Palaeolithic Homo sapiens from Bacho Kiro Cave, Bulgaria’, *Nature*, 581(7808), pp. 299–302. Available at: https://doi.org/10.1038/s41586-020-2259-z.

Hudson, R.R. and Kaplan, N.L. (1988) ‘The coalescent process in models with selection and recombination’, *Genetics*, 120(3), pp. 831–840. Available at: https://doi.org/10.1017/S0016672300029074.

Hudson, R.R., Kreitman, M. and Aguadé, M. (1987) ‘A Test of Neutral Molecular Evolution Based on Nucleotide Data’, *Genetics*, 116(1), pp. 153–159. Available at: https://doi.org/10.1093/genetics/116.1.153.

Huerta-Sánchez, E. *et al.* (2014) ‘Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA’, *Nature*, 512(7513), pp. 194–197. Available at: https://doi.org/10.1038/nature13408.

Hughes, A. and Nei, M. (1988) ‘Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection’, *Nature*, 335(1), pp. 167–170. Available at: https://doi.org/10.1038/335167a0.

Hut, R.A. *et al.* (2013) ‘Latitudinal clines: An evolutionary view on biological rhythms’, *Proceedings of the Royal Society B: Biological Sciences*, 280(1765). Available at: https://doi.org/10.1098/rspb.2013.0433.

Jacobs, G.S. *et al.* (2019) ‘Multiple Deeply Divergent Denisovan Ancestries in Papuans’, *Cell*, 177. Available at: https://doi.org/10.1016/j.cell.2019.02.035.

Jaganathan, K. *et al.* (2019) ‘Predicting Splicing from Primary Sequence with Deep Learning’, *Cell*, 176(3), pp. 535-548.e24. Available at: https://doi.org/10.1016/j.cell.2018.12.015.

Jinam, T.A. *et al.* (2017) ‘Discerning the origins of the negritos, first sundal and people: Deep divergence and archaic admixture’, *Genome Biology and Evolution*, 9(8), pp. 2013–2022. Available at: https://doi.org/10.1093/gbe/evx118.

Johanson, D.C. and Taieb, M. (1976) ‘Plio-Pleistocene hominid discoveries in Hadar, Ethiopia’, *Nature*, 260(5549), pp. 293–297. Available at: https://doi.org/10.1038/260293a0.

Johnson, R.J. *et al.* (2009) ‘Lessons from comparative physiology: Could uric acid represent a physiologic alarm signal gone awry in western society?’, *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology*. Springer Verlag, pp. 67–76. Available at: https://doi.org/10.1007/s00360-008-0291-7.

Jones, S.E. *et al.* (2016) ‘Genome-Wide Association Analyses in 128,266 Individuals Identifies New Morningness and Sleep Duration Loci’, *PLoS Genetics*, 12(6), p. e1006125. Available at: https://doi.org/10.1371/journal.pgen.1006125.

Jones, S.E. *et al.* (2019) ‘Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms’, *Nature Communications*, 10(1), p. 343. Available at: https://doi.org/10.1038/s41467-018-08259-7.

Joordens, J.C.A. *et al.* (2015) ‘Homo erectus at Trinil on Java used shells for tool production and engraving’, *Nature*, 518(7538), pp. 228–231. Available at: https://doi.org/10.1038/nature13962.

Jouganous, J. *et al.* (2017) ‘Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation’, *Genetics*, 206(3), pp. 1549–1567. Available at: https://doi.org/10.1534/genetics.117.200493.

Juric, I., Aeschbacher, S. and Coop, G. (2016) ‘The Strength of Selection against Neanderthal Introgression’, *PLoS Genetics*, 12(11), p. e1006340. Available at: https://doi.org/10.1371/journal.pgen.1006340.

Kanai, M. *et al.* (2018) ‘Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases’, *Nature Genetics*, 50(3), pp. 390–400. Available at: https://doi.org/10.1038/s41588-018-0047-6.

Kaplan, N., Hudson, R.R. and Iizuka, M. (1988) ‘The coalescent process in models with selection’, *Genetical Research*, 57(1), pp. 83–91. Available at: https://doi.org/10.1017/S0016672300029074.

Karczewski, K.J. *et al.* (2020) ‘The mutational constraint spectrum quantified from variation in 141,456 humans’, *Nature*, 581(7809), pp. 434–443. Available at: https://doi.org/10.1038/s41586-020-2308-7.

Kelso, J. and Prüfer, K. (2014) ‘Ancient humans and the origin of modern humans’, *Current Opinion in Genetics and Development* [Preprint]. Available at: https://doi.org/10.1016/j.gde.2014.09.004.

Kern, A.D. and Schrider, D.R. (2018) ‘DiploS/HIC: An updated approach to classifying selective sweeps’, *G3: Genes, Genomes, Genetics*, 8(6), pp. 1959–1970. Available at: https://doi.org/10.1534/g3.118.200262.

Key, F.M. *et al.* (2014) ‘Advantageous diversity maintained by balancing selection in humans’, *Current Opinion in Genetics and Development*, 29, pp. 45–51. Available at: https://doi.org/10.1016/j.gde.2014.08.001.

Khan, S. *et al.* (2018) ‘Health risks associated with genetic alterations in internal clock system by external factors’, *International Journal of Biological Sciences*. Ivyspring International Publisher, pp. 791–798. Available at: https://doi.org/10.7150/ijbs.23744.

Kim, J.K. and Forger, D.B. (2012) ‘A mechanism for robust circadian timekeeping via stoichiometric balance’, *Molecular Systems Biology*, 8(1). Available at: https://doi.org/10.1038/msb.2012.62.

Kimbel, W.H., Johanson, D.C. and Rak, Y. (1994) ‘The first skull and other new discoveries of Australopithecus afarensis at Hadar, Ethiopia’, *Nature*, 368(6470), pp. 449–451. Available at: https://doi.org/10.1038/368449a0.

Kimura, M. (1968) ‘Evolutionary Rate at the Molecular Level’, *Nature*, 217, pp. 624–626. Available at: https://doi.org/10.1006/rwgn.2001.0432.

King, W. (1864) ‘The Reputed Fossil Man of the Neanderthal’, in James Samuelson and William Crookes (eds) *Quarterly Journal of Science*, pp. 88--97.

Kivelä, L., Papadopoulos, M.R. and Antypa, N. (2018) ‘Chronotype and Psychiatric Disorders’, *Current Sleep Medicine Reports*, 4(2), pp. 94–103. Available at: https://doi.org/10.1007/s40675-018-0113-8.

Knutson, K.L. and von Schantz, M. (2018) ‘Associations between chronotype, morbidity and mortality in the UK Biobank cohort’, *Chronobiology International*, 35(8), pp. 1045–1053. Available at: https://doi.org/10.1080/07420528.2018.1454458.

Kojima, K. and Yarbrough, K.M. (1967) ‘Frequency-dependent selection at the esterase 6 locus in Drosophila melanogaster’, *Proceedings of the National Academy of Sciences*, 57(3), pp. 645–649. Available at: https://doi.org/10.1073/pnas.57.3.645.

Kojima, K.-I. and Tobari, Y.N. (1968) ‘The pattern of viability changes associated with genotype frequency at the alcohol dehydrogenase locus in a population of Drosophila melanogaster’, *Genetics*, 61(1), pp. 201–209. Available at: https://doi.org/10.1093/genetics/61.1.201.

Kononoff, J. *et al.* (2018) ‘Systemic and intra-habenular activation of the orphan G protein-coupled receptor GPR139 decreases compulsive-like alcohol drinking and hyperalgesia in alcohol-dependent rats’, *eNeuro*, 5(3). Available at: https://doi.org/10.1523/ENEURO.0153-18.2018.

Koop, B.F. *et al.* (1989) ‘A molecular view of primate phylogeny and important systematic and evolutionary questions.’, *Molecular biology and evolution*, 6(6), pp. 580–612. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a040574.

Köttgen, A. *et al.* (2013) ‘Genome-wide association analyses identify 18 new loci associated with serum urate concentrations’, *Nature Genetics*, 45(2), pp. 145–154. Available at: https://doi.org/10.1038/ng.2500.

Kratzer, J.T. *et al.* (2014) ‘Evolutionary history and metabolic insights of ancient mammalian uricases’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(10), pp. 3763–3768. Available at: https://doi.org/10.1073/pnas.1320393111.

Krause, J. *et al.* (2007) ‘Neanderthals in central Asia and Siberia’, *Nature*, 449(7164), pp. 902–904. Available at: https://doi.org/10.1038/nature06193.

Krause, J., Briggs, A.W., *et al.* (2010) ‘A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia’, *Current Biology*, 20(3), pp. 231–236. Available at: https://doi.org/10.1016/j.cub.2009.11.068.

Krause, J., Fu, Q., *et al.* (2010) ‘The complete mitochondrial DNA genome of an unknown hominin from southern Siberia’, *Nature*, 464(7290), pp. 894–897. Available at: https://doi.org/10.1038/nature08976.

Krings, M. *et al.* (1997) ‘Neandertal DNA sequences and the origin of modern humans’, *Cell*, 90(1), pp. 19–30. Available at: https://doi.org/10.1016/S0092-8674(00)80310-4.

Krings, M. *et al.* (1999) ‘DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen’, *Proceedings of the National Academy of Sciences*, 96(10), pp. 5581–5585. Available at: https://doi.org/10.1073/pnas.96.10.5581.

Krings, M. *et al.* (2000) ‘A view of Neandertal genetic diversity’, *Nature Genetics*, 26, pp. 144–146. Available at: https://doi.org/10.1038/79855.

Kronenberg, Z.N. *et al.* (2018) ‘High-resolution comparative analysis of great ape genomes’, *Science*, 360(6393), p. eaar6343. Available at: https://doi.org/10.1126/science.aar6343.

Kuhlwilm, M. and Boeckx, C. (2019) ‘A catalog of single nucleotide changes distinguishing modern humans from archaic hominins’, *Nature Scientific Reports*, 9(8463). Available at: https://doi.org/10.1038/s41598-019-44877-x.

Kundaje, A. (2013) ‘A comprehensive collection of signal artifact blacklist regions in the human genome’.

Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001) ‘Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data’, *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289. Available at: https://doi.org/10.29122/mipi.v11i1.2792.

Lalueza-Fox, C. *et al.* (2005) ‘Neandertal evolutionary genetics: Mitochondrial DNA data from the Iberian Peninsula’, *Molecular Biology and Evolution*, 22(4), pp. 1077–1081. Available at: https://doi.org/10.1093/molbev/msi094.

Lalueza-Fox, C. *et al.* (2006) 'Mitochondrial DNA of an Iberian Neandertal suggests a population affinity with other European Neandertals', *Current Biology*, 16(16), pp. 629–630. Available at: https://doi.org/10.1016/j.cub.2006.07.044.

Lane, J.M. *et al.* (2016) 'Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank', *Nature Communications*, 7(1), pp. 1–10. Available at: https://doi.org/10.1038/ncomms10889.

Lao, O. *et al.* (2008) 'Correlation between Genetic and Geographic Structure in Europe', *Current Biology*, 18(16), pp. 1241–1248. Available at: https://doi.org/10.1016/j.cub.2008.07.049.

Lapiedra, O. *et al.* (2018) 'Predator-driven natural selection on risk-taking behavior in anole lizards', *Science*, 360(6392), pp. 1017–1020. Available at: https://doi.org/10.1126/science.aap9289.

Larcher, S. *et al.* (2015) 'Sleep habits and diabetes', *Diabetes and Metabolism*, pp. 263–271. Available at: https://doi.org/10.1016/j.diabet.2014.12.004.

Larcher, S. *et al.* (2016) 'Impact of sleep behavior on glycemic control in type 1 diabetes: The role of social jetlag', *European Journal of Endocrinology*, 175(5), pp. 411–419. Available at: https://doi.org/10.1530/EJE-16-0188.

Lawlor, D.A. *et al.* (1988) 'HLA-A and B polymorphisms predate the divergence of humans and chimpanzees', *Nature*, 335(6187), pp. 268–271. Available at: https://doi.org/10.1038/335268a0.

Lazaridis, I. *et al.* (2016) 'Genomic insights into the origin of farming in the ancient Near East', *Nature Publishing Group*, 536, pp. 419–426. Available at: https://doi.org/10.1038/nature19310.

Leakey, L.S.B., Tobias, P. V. and Napier, J.R. (1964) 'A New Species of The Genus Homo From Olduvai Gorge', *Nature*, 202(4927), pp. 7–9. Available at: https://doi.org/10.1038/202007a0.

Leakey, M.D. (1971) 'Olduvai Gorge: Excavations in Beds I and II; 1960–1963', *Cambridge University Press* [Preprint].

Leakey, R.E.F. (1976) 'New hominid fossils from the Koobi Fora formation in Northern Kenya', *Nature*, 261(5561), pp. 574–576. Available at: https://doi.org/10.1038/261574a0.

Lebatard, A.E. *et al.* (2014) 'Dating the Homo erectus bearing travertine from Kocabaş (Denizli, Turkey) at at least 1.1 Ma', *Earth and Planetary Science Letters*, 390, pp. 8–18. Available at: https://doi.org/10.1016/j.epsl.2013.12.031.

Lee, J.J. *et al.* (2018) 'Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals', *Nature Genetics*, 50(8), pp. 1112–1121. Available at: https://doi.org/10.1038/s41588-018-0147-3.

Lee, Y.J. *et al.* (2012) 'Circadian regulation of low density lipoprotein receptor promoter activity by CLOCK/BMAL1, Hes1 and Hes6', *Experimental and Molecular Medicine*, 44(11), pp. 642–652. Available at: https://doi.org/10.3858/emm.2012.44.11.073.

Leffler, E.M. *et al.* (2013) 'Multiple instances of ancient balancing selection shared between humans and chimpanzees', *Science*, 340(6127), pp. 1578–1582. Available at: https://doi.org/10.1126/science.1234070.

Leitsalu, L. *et al.* (2015) 'Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu', *International Journal of Epidemiology*, 44(4), pp. 1137–1147. Available at: https://doi.org/10.1093/ije/dyt268.

Lenz, T.L. *et al.* (2016) 'Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection', *Molecular Biology and Evolution*, 33(10), pp. 2555–2564. Available at: https://doi.org/10.1093/molbev/msw127.

Leocadio-Miguel, M.A. *et al.* (2017) 'Latitudinal cline of chronotype', *Scientific Reports*, 7(5437), pp. 2–7. Available at: https://doi.org/10.1038/s41598-017-05797-w.

Lewontin, R.C. (1964) 'The interaction of selection and linkage. I. General considerations; heterotic models', *Genetics*, 49(1), pp. 49–67. Available at: https://doi.org/10.1093/genetics/49.1.49.

Li, Z. *et al.* (2022) 'Phylogenetic Articulation of Uric Acid Evolution in Mammals and How It Informs a Therapeutic Uricase', *Molecular Biology and Evolution*, 39(1). Available at: https://doi.org/10.1093/molbev/msab312.

Linnér, R.K. (2019) 'Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences', *Nature Genetics*, 51(2), pp. 245–257. Available at: https://doi.org/10.1038/s41588-018-0309-3.

Lockey, A.L. *et al.* (2022) 'Comparing the Boxgrove and Atapuerca (Sima de los Huesos) human fossils: Do they represent distinct paleodemes?', *Journal of Human Evolution*, 172, p. 103253. Available at: https://doi.org/10.1016/j.jhevol.2022.103253.

López-Valverde, A., López-Cristiá, M. and Gómez De Diego, R. (2012) 'Europe's oldest jaw: Evidence of oral pathology', *British Dental Journal*, 212(5), pp. 243–245. Available at: https://doi.org/10.1038/sj.bdj.2012.176.

Lordkipanidze, D. *et al.* (2013) 'A Complete Skull from Dmanisi, Georgia, and the Evolutionary Biology of Early Homo', *Science*, 342(6156), pp. 326–332.

Lowden, A. *et al.* (2018) 'Delayed Sleep in Winter Related to Natural Daylight Exposure among Arctic Day Workers', *Clocks & Sleep*, 1(1), pp. 105–116. Available at: https://doi.org/10.3390/clockssleep1010010.

Luzzatto, L. (2012) 'Sickle cell anaemia and malaria', *Mediterranean Journal of Hematology and Infectious Diseases*, 4(1). Available at: https://doi.org/10.4084/MJHID.2012.065.

Machiela, M.J. and Chanock, S.J. (2015) 'LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants', *Bioinformatics*, 31(21), pp. 3555–3557. Available at: https://doi.org/10.1093/bioinformatics/btv402.

Mafessoni, F. *et al.* (2020) 'A high-coverage Neandertal genome from Chagyrskaya Cave', *Proceedings of the National Academy of Sciences of the United States of America*, 117(26), pp. 15132–15136. Available at: https://doi.org/10.1073/pnas.2004944117.

Mallick, S. *et al.* (2016) 'The Simons Genome Diversity Project: 300 genomes from 142 diverse populations', *Nature*, 538(7624), pp. 201–206. Available at: https://doi.org/10.1038/nature18964.

Mallick, S. and Reich, D. (2023) 'The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes'. Harvard Dataverse, V8.

Manzi, G., Mallegni, F. and Ascenzi, A. (2001) 'A cranium for the earliest Europeans: Phylogenetic position of the hominid from Ceprano, Italy', *Proceedings of the National Academy of Sciences*, 98(17), pp. 10011–10016. Available at: https://doi.org/10.1073/pnas.151259998.

Mao, R. *et al.* (2013) 'Inhibition of Hepatitis B Virus Replication by the Host Zinc Finger Antiviral Protein', *PLoS Pathogens*, 9(7). Available at: https://doi.org/10.1371/journal.ppat.1003494.

Marston, A.T. (1937) 'The Swanscombe skull', *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 67, pp. 339–406.

Martens, M. *et al.* (2021) 'WikiPathways: Connecting communities', *Nucleic Acids Research*, 49(D1), pp. D613–D621. Available at: https://doi.org/10.1093/nar/gkaa1024.

Mathieson, I. *et al.* (2015) 'Genome-wide patterns of selection in 230 ancient Eurasians', *Nature*, 528(7583), pp. 499–503. Available at: https://doi.org/10.1038/nature16152.

Mattheisen, M. *et al.* (2015) 'Genome-wide association study in obsessive-compulsive disorder: Results from the OCGAS', *Molecular Psychiatry*, 20(3). Available at: https://doi.org/10.1038/mp.2014.43.

Matzaraki, V. *et al.* (2017) 'The MHC locus and genetic susceptibility to autoimmune and infectious diseases', *Genome Biology*, 18(1), pp. 1–21. Available at: https://doi.org/10.1186/s13059-017-1207-1.

Mayer, W.E. *et al.* (1988) 'Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution.', *The EMBO journal*, 7(9), pp. 2765–2774. Available at: https://doi.org/10.1002/j.1460-2075.1988.tb03131.x.

McArthur, E. *et al.* (2022) 'Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence', *bioRxiv* [Preprint].

McArthur, E., Rinker, D.C. and Capra, J.A. (2021) ‘Quantifying the contribution of Neanderthal introgression to the heritability of complex traits’, *Nature Communications*, 12(1), pp. 1–14. Available at: https://doi.org/10.1038/s41467-021-24582-y.

McCoy, R.C., Wakefield, J. and Akey, J.M. (2017) ‘Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression’, *Cell*, 168, pp. 916–927. Available at: https://doi.org/10.1016/j.cell.2017.01.038.

McLaren, W. *et al.* (2016) ‘The Ensembl Variant Effect Predictor’, *Genome Biology*, 17(1). Available at: https://doi.org/10.1186/s13059-016-0974-4.

McPherron, S.P. *et al.* (2010) ‘Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia’, *Nature*, 466(7308), pp. 857–860. Available at: https://doi.org/10.1038/nature09248.

Meyer, M. *et al.* (2012a) ‘A high-coverage genome sequence from an archaic Denisovan individual’, *Science*, 338(6104), pp. 222–226. Available at: https://doi.org/10.1126/science.1224344.

Meyer, M. *et al.* (2012b) ‘A high-coverage genome sequence from an archaic Denisovan individual’, *Science*, 338(6104), pp. 222–226. Available at: https://doi.org/10.1126/science.1224344.

Meyer, M. *et al.* (2014) ‘A mitochondrial genome sequence of a hominin from Sima de los Huesos’, *Nature*, 505(7483). Available at: https://doi.org/10.1038/nature12788.

Meyer, M. *et al.* (2016) ‘Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins’, *Nature*, 531(7595), pp. 504–507. Available at: https://doi.org/10.1038/nature17405.

Meyer, M.R. *et al.* (2023) ‘Knuckle-walking in Sahelanthropus? Locomotor inferences from the ulnae of fossil hominins and other hominoids’, *Journal of Human Evolution*, 179, p. 103355. Available at: https://doi.org/10.1016/j.jhevol.2023.103355.

Michael, T.P. *et al.* (2003) ‘Enhanced Fitness Conferred by Naturally Occurring Variation in the Circadian Clock’, *Science*, 302(5647), pp. 1049–1053. Available at: https://doi.org/10.1126/science.1082971.

Michel Brunet *et al.* (2005) ‘New material of the earliest hominid from the Upper Miocene of Chad’, *Nature*, 434(7034), pp. 749–752. Available at: https://doi.org/10.1038/nature03433.1.

Moore, J.E. *et al.* (2020) ‘Expanded encyclopaedias of DNA elements in the human and mouse genomes’, *Nature*, 583(7818), pp. 699–710. Available at: https://doi.org/10.1038/s41586-020-2493-4.

Morrow, E.M. *et al.* (2008) ‘Identifying autism loci and genes by tracing recent shared ancestry’, *Science*, 321(5886), pp. 218–223. Available at: https://doi.org/10.1126/science.1157657.

Moyano, D.B., Sola, Y. and González-Lezcano, R.A. (2020) ‘Blue-Light levels emitted from portable electronic devices compared to sunlight’, *Energies*, 13(6). Available at: https://doi.org/10.3390/en13164276.

Nagai, A. *et al.* (2017) ‘Overview of the BioBank Japan Project: Study design and profile’, *Journal of Epidemiology*, 27(3), pp. S2–S8. Available at: https://doi.org/10.1016/j.je.2016.12.005.

Nielsen, R. *et al.* (2017) ‘Tracing the peopling of the world through genomics’, *Nature*, 541(7637), pp. 302–310. Available at: https://doi.org/10.1038/nature21347.

Núñez-Lahuerta, C. *et al.* (2022) ‘A bird assemblage across the MIS 9/8 boundary: The Middle Pleistocene of Galería (Atapuerca)’, *Quaternary Science Reviews*, 293, p. 107708. Available at: https://doi.org/10.1016/j.quascirev.2022.107708.

Ollé, A. *et al.* (2013) ‘The Early and Middle Pleistocene technological record from Sierra de Atapuerca (Burgos, Spain)’, *Quaternary International*, 295, pp. 138–167. Available at: https://doi.org/10.1016/j.quaint.2011.11.009.

Ollier, W., Sprosen, T. and Peakman, T. (2005) ‘UK Biobank: From concept to reality’, *Pharmacogenomics*, 6(6), pp. 639–646. Available at: https://doi.org/10.2217/14622416.6.6.639.

O’Malley, K.G. and Banks, M.A. (2008) ‘A latitudinal cline in the Chinook salmon (Oncorhynchus tshawytscha) Clock gene: Evidence for selection on PolyQ length variants’, *Proceedings of the Royal Society B: Biological Sciences*, 275(1653), pp. 2813–2821. Available at: https://doi.org/10.1098/rspb.2008.0524.

O’Malley, K.G., Ford, M.J. and Hard, J.J. (2010) ‘Clock polymorphism in Pacific salmon: Evidence for variable selection along a latitudinal gradient’, in *Proceedings of the Royal Society B: Biological Sciences*. Royal Society, pp. 3703–3714. Available at: https://doi.org/10.1098/rspb.2010.0762.

Orlando, L. *et al.* (2005) ‘Revisiting Neandertal diversity with a mtDNA sequence’, *Current Biology*, 16(11), pp. 400–402. Available at: https://doi.org/10.1016/j.cub.2006.05.019.

Orlando, L. *et al.* (2021) ‘Ancient DNA analysis’, *Nature Reviews Methods Primers*. Springer Nature, pp. 1–14. Available at: https://doi.org/10.1038/s43586-020-00011-0.

Ovchinnikov, I. V. *et al.* (2000) ‘Molecular analysis of Neanderthal DNA from the northern Caucasus’, *Nature*, 404(6777), pp. 490–493. Available at: https://doi.org/10.1038/35006625.

Pääbo, S. (1984) ‘Über den Nachweis von DNA in altägyptischen Mumien’, *Das Altertum*, 30, pp. 213–218.

Pääbo, S. (1985) ‘Preservation of DNA in ancient Egyptian mummies’, *Journal of Archaeological Science*, 12(6), pp. 411–417. Available at: https://doi.org/10.1016/0305-4403(85)90002-0.

Pääbo, S. (1989) ‘Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification’, *Proceedings of the National Academy of Sciences*, 86(6), pp. 1939–1943. Available at: https://doi.org/10.1073/pnas.86.6.1939.

Page, S.L. and Goodman, M. (2001) ‘Catarrhine phylogeny: Noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade’, *Molecular Phylogenetics and Evolution*, 18(1), pp. 14–25. Available at: https://doi.org/10.1006/mpev.2000.0895.

Papantoniou, K. *et al.* (2016) ‘Breast cancer risk and night shift work in a case–control study in a Spanish population’, *European Journal of Epidemiology*, 31(9), pp. 867–878. Available at: https://doi.org/10.1007/s10654-015-0073-y.

Papantoniou, K. *et al.* (2017) ‘Shift work and colorectal cancer risk in the MCC-Spain case–control study’, *Scandinavian Journal of Work, Environment and Health*, 43(3), pp. 250–259. Available at: https://doi.org/10.5271/sjweh.3626.

Passey, B.H. *et al.* (2010) ‘High-temperature environments of human evolution in East Africa based on bond ordering in paleosol carbonates’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(25), pp. 11245–11249. Available at: https://doi.org/10.1073/pnas.1001824107.

Patterson, N. *et al.* (2006) ‘Genetic evidence for complex speciation of humans and chimpanzees’, *Nature*, 441(7097), pp. 1103–1108. Available at: https://doi.org/10.1038/nature04789.

Petr, M. *et al.* (2019) ‘Limits of long-term selection against Neandertal introgression’, *Proceedings of the National Academy of Sciences*, 116(5), pp. 1639–1644. Available at: https://doi.org/10.1073/pnas.1814338116.

Piel, F.B. *et al.* (2013) ‘Global epidemiology of Sickle haemoglobin in neonates: A contemporary geostatistical model-based map and population estimates’, *The Lancet*, 381(9861), pp. 142–151. Available at: https://doi.org/10.1016/S0140-6736(12)61229-X.

Plagnol, V. and Wall, J.D. (2006) ‘Possible ancestral structure in human populations’, *PLoS Genetics*, 2(7), p. e105. Available at: https://doi.org/10.1371/journal.pgen.0020105.

Poinar, H.N. *et al.* (1996) ‘Amino Acid Racemization and the Preservation of Ancient DNA’, *Science*, 272(5263), pp. 864–866. Available at: https://doi.org/10.1126/science.272.5263.864.

Porter, C.A. *et al.* (1995) ‘Evidence on primate phylogeny from ε-globin gene sequences and flanking regions’, *Journal of Molecular Evolution*, 40(1), pp. 30–55. Available at: https://doi.org/10.1007/BF00166594.

Porter, C.A., Page, S.L., *et al.* (1997) ‘Phytogeny and evolution of selected primates as determined by sequences of the ε-globin locus and 5′ flanking regions’, *International Journal of Primatology*, 18(2), pp. 261–295. Available at: https://doi.org/10.1023/A:1026328804319.

Porter, C.A., Czelusniak, J., *et al.* (1997) ‘Sequences of the primate ε-globin gene: Implications for systematics of the marmosets and other New World primates’, *Gene*, 205(1–2), pp. 59–71. Available at: https://doi.org/10.1016/S0378-1119(97)00473-3.

Potts, R. (1996) ‘Evolution and climate variability’, *Science*, 273(5277), pp. 922–923. Available at: https://doi.org/10.1126/science.273.5277.922.

Potts, R. (1998a) ‘Environmental Hypotheses of Hominin Evolution’, *Yearbook of Physical Anthropology*, 41, pp. 93–136. Available at: https://doi.org/10.1002/(sici)1096-8644(1998)107:27+<93::aid-ajpa5>3.0.co;2-x.

Potts, R. (1998b) ‘Variability selection in hominid evolution’, *Evolutionary Anthropology*, 7(3), pp. 81–96. Available at: https://doi.org/10.1002/(SICI)1520-6505(1998)7:3<81::AID-EVAN3>3.0.CO;2-A.

Potts, R. (2012a) ‘Environmental and behavioral evidence pertaining to the evolution of early Homo’, *Current Anthropology*, 53(S6), pp. S299–S317. Available at: https://doi.org/10.1086/667704.

Potts, R. (2012b) ‘Evolution and environmental change in early human prehistory’, *Annual Review of Anthropology*, 41, pp. 151–167. Available at: https://doi.org/10.1146/annurev-anthro-092611-145754.

Prado-Martinez, J. *et al.* (2013) ‘Great ape genetic diversity and population history’, *Nature*, 499(7459), pp. 471–475. Available at: https://doi.org/10.1038/nature12228.

Prentice, M.L. and Denton, G.H. (1988) ‘The deep-sea oxygen isotope record, the global ice sheet system and hominid evolution’, in *The evolutionary history of the robust Australopithecines*. Aldine de Gruyter, pp. 383–403.

Prüfer, K. *et al.* (2014) ‘The complete genome sequence of a Neanderthal from the Altai Mountains’, *Nature*, 505(7481), pp. 43–49. Available at: https://doi.org/10.1038/nature12886.

Prüfer, K. *et al.* (2017) ‘A high-coverage Neandertal genome from Vindija Cave in Croatia’, *Science*, 358(6363), pp. 655–658. Available at: https://doi.org/10.1126/science.aao1887.

Prüfer, K. *et al.* (2021) ‘A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia’, *Nature Ecology and Evolution*, 5(6), pp. 820–825. Available at: https://doi.org/10.1038/s41559-021-01443-x.

Pruim, R.J. *et al.* (2011) ‘LocusZoom: Regional visualization of genome-wide association scan results’, in *Bioinformatics*. Oxford University Press, pp. 2336–2337. Available at: https://doi.org/10.1093/bioinformatics/btq419.

Putilov, A.A. *et al.* (2019) ‘Genetic-based signatures of the latitudinal differences in chronotype’, *Biological Rhythm Research*, 50(2), pp. 255–271. Available at: https://doi.org/10.1080/09291016.2018.1465249.

Putilov, A.A., Dorokhov, V.B. and Poluektov, M.G. (2018) ‘How have our clocks evolved? Adaptive and demographic history of the out-of-African dispersal told by polymorphic loci in circadian genes’, *Chronobiology International*, 35(4), pp. 511–532. Available at: https://doi.org/10.1080/07420528.2017.1417314.

Quinlan, A.R. (2014) ‘BEDTools: The Swiss-Army tool for genome feature analysis’, *Current Protocols in Bioinformatics*, 47(1), pp. 11–12. Available at: https://doi.org/10.1002/0471250953.bi1112s47.

Racimo, F. *et al.* (2015) ‘Evidence for archaic adaptive introgression in humans’, *Nature Reviews Genetics*, 16(6), pp. 359–371. Available at: https://doi.org/10.1038/nrg3936.

Racimo, F. *et al.* (2017) ‘Archaic adaptive introgression in TBX15/WARS2’, *Molecular Biology and Evolution*, 34(3), pp. 509–524. Available at: https://doi.org/10.1093/molbev/msw283.

Racimo, F., Marnetto, D. and Huerta-Sánchez, E. (2017) ‘Signatures of archaic adaptive introgression in present-day human populations’, *Molecular Biology and Evolution*, 34(2), pp. 296–317. Available at: https://doi.org/10.1093/molbev/msw216.

Randler, C. and Rahafar, A. (2017) ‘Latitude affects Morningness- Eveningness: evidence for the environment hypothesis based on a systematic review’, *Nature Publishing Group*, 7(39976), pp. 1–6. Available at: https://doi.org/10.1038/srep39976.

Rasmussen, M.D. *et al.* (2014) ‘Genome-Wide Inference of Ancestral Recombination Graphs’, *PLoS Genetics*, 10(5). Available at: https://doi.org/10.1371/journal.pgen.1004342.

Reich, D. *et al.* (2010) ‘Genetic history of an archaic hominin group from Denisova cave in Siberia’, *Nature*, 468(7327), pp. 1053–1060. Available at: https://doi.org/10.1038/nature09710.

Richter, D. *et al.* (2017) ‘The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age’, *Nature*, 546(7657), pp. 293–296. Available at: https://doi.org/10.1038/nature22335.

Rightmire, G.P. (1979) ‘Cranial remains of Homo erectus from Beds II and IV, Olduvai Gorge, Tanzania’, *American Journal of Physical Anthropology*, 51(1), pp. 99–115. Available at: https://doi.org/10.1002/ajpa.1330510113.

Rightmire, G.P. (1996) ‘The human cranium from Bodo, Ethiopia: Evidence for speciation in the Middle Pleistocene?’, *Journal of Human Evolution*, 31(1), pp. 21–39. Available at: https://doi.org/10.1006/jhev.1996.0046.

Rinker, D.C. *et al.* (2020) ‘Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations’, *Nature Ecology and Evolution*, 4(10), pp. 1332–1341. Available at: https://doi.org/10.1038/s41559-020-1261-z.

Rito, T. *et al.* (2019) ‘A dispersal of Homo sapiens from southern to eastern Africa immediately preceded the out-of-Africa migration’, *Scientific Reports*, 9(1), p. 4728. Available at: https://doi.org/10.1038/s41598-019-41176-3.

Roenneberg, T. *et al.* (2004) ‘A marker for the end of adolescence’, *Current Biology*, 14(24), pp. 1038–1039. Available at: https://doi.org/10.1016/j.cub.2004.11.039.

Roenneberg, T., Daan, S. and Merrow, M. (2003) ‘The art of entrainment’, *Journal of Biological Rhythms*, 18(3), pp. 183–194. Available at: https://doi.org/10.1177/0748730403018003001.

Rogers, A., Iltis, D. and Wooding, S. (2004) ‘Genetic variation at the MC1R locus and the time since loss of human body hair’, *Current Anthropology*, 45(1), pp. 105–108. Available at: https://doi.org/10.1086/381006.

Saiki, R.K. *et al.* (1985) ‘Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia’, *Science*, 230(4732), pp. 1350–1354. Available at: https://doi.org/10.1126/science.2999980.

Sala, N. *et al.* (2016) ‘The Sima de los Huesos Crania: Analysis of the cranial breakage patterns’, *Journal of Archaeological Science*, 72, pp. 25–43. Available at: https://doi.org/10.1016/j.jas.2016.06.001.

Sanchez-Roige, S. *et al.* (2019) ‘Genome-wide association study meta-analysis of the alcohol use disorders identification test (AUDIT) in two population-based cohorts’, *American Journal of Psychiatry*, 176(2), pp. 107–118. Available at: https://doi.org/10.1176/appi.ajp.2018.18040369.

Sandrelli, F. *et al.* (2007) ‘A molecular basis for natural selection at the timeless locus in Drosophila melanogaster’, *Science*, 316(5833), pp. 1898–1900. Available at: https://doi.org/10.1126/science.1138426.

Sankararaman, S. *et al.* (2012) ‘The date of interbreeding between Neandertals and modern humans’, *PLoS Genetics*, 8(10), p. e1002947. Available at: https://doi.org/10.1371/journal.pgen.1002947.

Sankararaman, S. *et al.* (2014) ‘The genomic landscape of Neanderthal ancestry in present-day humans’, *Nature*, 507(7492), pp. 354–359. Available at: https://doi.org/10.1038/nature12961.

Sankararaman, S. *et al.* (2016) ‘The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans’, *Current Biology*, 26(9), pp. 1241–1247. Available at: https://doi.org/10.1016/j.cub.2016.03.037.

Santos Ganges, L. (2003) *El ferrocarril minero de la Sierra de Burgos y las sociedades de Mr. Richard Preece Williams*. Fundación de los Ferrocarriles Españoles. Available at: http://uvadoc.uva.es/handle/10324/26439.

Sato, D.X. and Kawata, M. (2018) ‘Positive and balancing selection on SLC18A1 gene associated with psychiatric disorders and human-unique personality traits’, *Evolution Letters*, 2(5), pp. 499–510. Available at: https://doi.org/10.1002/evl3.81.

Schaefer, N.K., Shapiro, B. and Green, R.E. (2021) ‘An ancestral recombination graph of human, Neanderthal, and Denisovan genomes’, *Science advances*, 7(29), pp. 776–792. Available at: https://doi.org/10.1126/sciadv.abc077.

Schmitz, R.W. *et al.* (2002) ‘The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany’, *Proceedings of the National Academy of Sciences*, 99(20), pp. 13342–13347. Available at: https://doi.org/10.1073/pnas.192464099.

Schoetensack, O. (1908) ‘Der Unterkiefer des Homo heidelbergensis’, *Links*, 5, p. 5.

Schrider, D.R. and Kern, A.D. (2017) ‘Soft sweeps are the dominant mode of adaptation in the human genome’, *Molecular Biology and Evolution*, 34(8), pp. 1863–1877. Available at: https://doi.org/10.1093/molbev/msx154.

Schwarcz, H.P. *et al.* (1988) ‘ESR dates for the hominid burial site of Qafzeh in Israel’, *Journal of Human Evolution*, 17(8), pp. 733–737. Available at: https://doi.org/10.1016/0047-2484(88)90063-2.

Seguin-Orlando, A. *et al.* (2014) ‘Genomic structure in Europeans dating back at least 36,200 years’, *Science*, 346(6213), pp. 1113–1118. Available at: https://doi.org/10.1126/science.aaa0114.

Ségurel, L. *et al.* (2012) ‘The ABO blood group is a trans-species polymorphism in primates’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(45), pp. 18493–18498. Available at: https://doi.org/10.1073/pnas.1210603109.

Semaw, S. *et al.* (1997) ‘2.5-million-year-old stone tools from Gona, Ethiopia’, *Nature*, 385(6614), pp. 333–336. Available at: https://doi.org/10.1038/385333a0.

Semaw, S. (2000) ‘The world’s oldest stone artefacts from Gona, Ethiopia: Their implications for understanding stone technology and patterns of human evolution between 2.6-1.5 million years ago’, *Journal of Archaeological Science*, 27(12), pp. 1197–1214. Available at: https://doi.org/10.1006/jasc.1999.0592.

Semaw, S. *et al.* (2003) ‘2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia’, *Journal of Human Evolution*, 45(2), pp. 169–177. Available at: https://doi.org/10.1016/S0047-2484(03)00093-9.

Senut, B. *et al.* (2001) ‘First hominid from the Miocene (Lukeino Formation, Kenya)’, *Comptes Rendus de l’Academie de Sciences - Serie IIa: Sciences de la Terre et des Planetes*, 332(2), pp. 137–144. Available at: https://doi.org/10.1016/S1251-8050(01)01529-4.

Serjeant, G.R. (2013) ‘The natural history of sickle cell disease’, *Cold Spring Harbor Perspectives in Medicine*, 3(10), p. a011783. Available at: https://doi.org/10.1101/cshperspect.a011783.

Serre, D. *et al.* (2004) ‘No evidence of Neandertal mtDNA contribution to early modern humans’, *PLoS Biology*, 2(3), pp. 313–317. Available at: https://doi.org/10.1371/journal.pbio.0020057.

Shen, G. *et al.* (2009) ‘Age of Zhoukoudian Homo erectus determined with 26Al/ 10Be burial dating’, *Nature*, 458(7235), pp. 198–200. Available at: https://doi.org/10.1038/nature07741.

Shi, Y. *et al.* (2020) ‘Night-shift work duration and risk of colorectal cancer according to IRS1 and IRS2 expression’, *Cancer Epidemiology Biomarkers and Prevention*, 29(1), pp. 133–140. Available at: https://doi.org/10.1158/1055-9965.EPI-19-0325.

Shoshani, J. *et al.* (1996) ‘Primate phylogeny: Morphological vs molecular results’, *Molecular Phylogenetics and Evolution*, 5(1), pp. 102–154. Available at: https://doi.org/10.1006/mpev.1996.0009.

Siewert, K.M. and Voight, B.F. (2017) ‘Detecting long-term balancing selection using allele frequency correlation’, *Molecular Biology and Evolution*, 34(11), pp. 2996–3005. Available at: https://doi.org/10.1093/molbev/msx209.

Siewert, K.M. and Voight, B.F. (2020) ‘BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data’, *Genome Biology and Evolution*, 12(2), pp. 3873–3877. Available at: https://doi.org/10.1093/gbe/evaa013.

Simonti, C.N. *et al.* (2016) ‘The phenotypic legacy of admixture between modern humans and Neanderthals’, *Science*, 351(6274), pp. 737–741. Available at: https://doi.org/10.1126/science.aad2149.

Simpson, S.W. *et al.* (2008) ‘A female Homo erectus pelvis from gona, Ethiopia’, *Science*, 322(5904), pp. 1089–1092. Available at: https://doi.org/10.1126/science.1163592.

Sirugo, G., Williams, S.M. and Tishkoff, S.A. (2019) ‘The Missing Diversity in Human Genetic Studies’, *Cell*. Cell Press, pp. 26–31. Available at: https://doi.org/10.1016/j.cell.2019.02.048.

Skoglund, P. and Mathieson, I. (2018) ‘Ancient genomics of modern humans: The first decade’, *Annual Review of Genomics and Human Genetics*. Annual Reviews Inc., pp. 381–404. Available at: https://doi.org/10.1146/annurev-genom-083117-021749.

Skov, L. *et al.* (2018) ‘Detecting archaic introgression using an unadmixed outgroup’, *PLoS Genetics*, 14(9), p. e1007641. Available at: https://doi.org/10.1371/journal.pgen.1007641.

Skov, L. *et al.* (2020) ‘The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes’, *Nature*, 582(7810), pp. 78–83. Available at: https://doi.org/10.1038/s41586-020-2225-9.

Skov, L. *et al.* (2022) ‘Genetic insights into the social organization of Neanderthals’, *Nature*, 610(7932), pp. 519–525. Available at: https://doi.org/10.1038/s41586-022-05283-y.

Slon, V. *et al.* (2017) ‘A fourth Denisovan individual’, *Science Advances*, 3(7), p. e1700186. Available at: https://doi.org/10.1126/sciadv.1700186.

Slon, V. *et al.* (2018) ‘The genome of the offspring of a Neanderthal mother and a Denisovan father’, *Nature*, 561, pp. 113–117. Available at: https://doi.org/10.1038/s41586-018-0455-x.

Sollis, E. *et al.* (2023) ‘The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource’, *Nucleic Acids Research*, 51(1 D), pp. D977–D985. Available at: https://doi.org/10.1093/nar/gkac1010.

Souilmi, Y. *et al.* (2022) ‘Admixture has obscured signals of historical hard sweeps in humans’, *Nature Ecology and Evolution*, 6(12), pp. 2003–2015. Available at: https://doi.org/10.1038/s41559-022-01914-9.

Srinivasan, V. *et al.* (2006) ‘Melatonin in mood disorders’, *World Journal of Biological Psychiatry*, pp. 138–151. Available at: https://doi.org/10.1080/15622970600571822.

Steele, J., Clegg, M. and Martelli, S. (2013) ‘Comparative morphology of the Hominin and African ape hyoid bone, a possible marker of the evolution of speech’, *Human Biology*, 85(5), pp. 639–672. Available at: https://doi.org/10.3378/027.085.0501.

Steinrücken, M. *et al.* (2018) ‘Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans’, *Molecular Ecology*, 27(19), pp. 3873–3888. Available at: https://doi.org/10.1111/mec.14565.

Steinrücken, M. *et al.* (2019) ‘Inference of complex population histories using whole-genome sequences from multiple populations’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(34), pp. 17115–17120. Available at: https://doi.org/10.1073/pnas.1905060116.

Stepanchuk, V.N. *et al.* (2017) ‘The last Neanderthals of Eastern Europe: Micoquian layers IIIa and III of the site of Zaskalnaya VI (Kolosovskaya), anthropological records and context’, *Quaternary International*, 428, pp. 132–150. Available at: https://doi.org/10.1016/j.quaint.2015.11.042.

Stotz, M. *et al.* (2014) ‘Evaluation of uric acid as a prognostic blood-based marker in a large cohort of pancreatic cancer patients’, *PLoS ONE*, 9(8). Available at: https://doi.org/10.1371/journal.pone.0104730.

Stout, D. *et al.* (2015) ‘Cognitive demands of lower Paleolithic toolmaking’, *PLoS ONE*, 10(4), p. e0121804. Available at: https://doi.org/10.1371/journal.pone.0121804.

Strait, D.S. (2010) ‘The Evolutionary History of the Australopiths’, *Evolution: Education and Outreach*, 3, pp. 341–352. Available at: https://doi.org/10.1007/s12052-010-0249-6.

Stringer, C. (2016) ‘The origin and evolution of Homo sapiens’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698), p. 20150237. Available at: https://doi.org/10.1098/rstb.2015.0237.

Stringer, C.B., Howell, F.C. and Melentis, J.K. (1979) ‘The significance of the fossil hominid skull from Petralona, Greece’, *Journal of Archaeological Science*, 6(3), pp. 235–253. Available at: https://doi.org/10.1016/0305-4403(79)90002-5.

Sudlow, C. *et al.* (2015) 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLoS Medicine*, 12(3), pp. 1–10. Available at: https://doi.org/10.1371/journal.pmed.1001779.

Sudmant, P.H. *et al.* (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75–81. Available at: https://doi.org/10.1038/nature15394.

Susman, R.L. (2017) 'Who Made the Oldowan Tools? Fossil Evidence for Tool Behavior in Plio-Pleistocene Hominids', *Journal of Anthropological Research*, 47(2), pp. 129–151.

Suwa, G. *et al.* (2007) 'Early Pleistocene Homo erectus fossils from Konso, southern Ethiopia', *Anthropological Science*, 115, pp. 133–151. Available at: https://doi.org/10.1537/ase.061203.

Tajima, F. (1989) 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics*, 123(3), pp. 585–595. Available at: https://doi.org/10.1093/genetics/123.3.585.

Takahata, N. and Nei, M. (1990) 'Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci', *Genetics*, 124(4), pp. 967–978. Available at: https://doi.org/10.1093/genetics/124.4.967.

Takahata, N. and Satta, Y. (1998) 'Footprints of intragenic recombination at HLA loci', *Immunogenetics*, 47(6), pp. 430–441. Available at: https://doi.org/10.1007/s002510050380.

Tauber, E. *et al.* (2007) 'Natural selection favors a newly derived timeless allele in Drosophila melanogaster', *Science*, 316(5833), pp. 1895–1898. Available at: https://doi.org/10.1126/science.1138412.

Taylor, B.J. and Hasler, B.P. (2018) 'Chronotype and Mental Health: Recent Advances', *Current Psychiatry Reports*, 20(8). Available at: https://doi.org/10.1007/s11920-018-0925-8.

Taylor, S.M. and Fairhurst, R.M. (2014) 'Malaria parasites and red cell variants: When a house is not a home', *Current Opinion in Hematology*, 21(3), pp. 193–200. Available at: https://doi.org/10.1097/MOH.0000000000000039.

Teixeira, J.C. *et al.* (2015) 'Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos', *Molecular Biology and Evolution*, 32(5), pp. 1186–1196. Available at: https://doi.org/10.1093/molbev/msv007.

Telis, N., Aguilar, R. and Harris, K. (2020) 'Selection against archaic hominin genetic variation in regulatory regions', *Nature Ecology and Evolution*, 4(11), pp. 1558–1566. Available at: https://doi.org/10.1038/s41559-020-01284-0.

Tierney, J.E., deMenocal, P.B. and Zander, P.D. (2017) 'A climatic context for the out-of-Africa migration', *Geology*, 45(11), pp. 1023–1026. Available at: https://doi.org/10.1130/G39457.1.

Tin, A. *et al.* (2019) 'Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels', *Nature Genetics*, 51(10), pp. 1459–1474. Available at: https://doi.org/10.1038/s41588-019-0504-x.

Todorova, T., Bock, F.J. and Chang, P. (2015) 'Poly(ADP-ribose) polymerase-13 and RNA regulation in immunity and cancer', *Trends in Molecular Medicine*. Elsevier Ltd, pp. 373–384. Available at: https://doi.org/10.1016/j.molmed.2015.03.002.

Toh, K.L. *et al.* (2001) 'An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome', *Science*, 291(5506), pp. 1040–1043. Available at: https://doi.org/10.1126/science.1057499.

Tordjman, S. *et al.* (2017) 'Melatonin: Pharmacology, Functions and Therapeutic Benefits', *Current Neuropharmacology*, 15, pp. 434–443. Available at: https://doi.org/10.2174/1570159X14666161228122.

Trauth, M.H. *et al.* (2005) 'Climate change: Late cenozoic moisture history of east Africa', *Science*, 309(5743), pp. 2051–2053. Available at: https://doi.org/10.1126/science.1112964.

Trauth, M.H. *et al.* (2007) 'High- and low-latitude forcing of Plio-Pleistocene East African climate and human evolution', *Journal of Human Evolution*, 53(5), pp. 475–486. Available at: https://doi.org/10.1016/j.jhevol.2006.12.009.

Trauth, M.H. *et al.* (2010) ‘Human evolution in a variable environment: The amplifier lakes of Eastern Africa’, *Quaternary Science Reviews*, 29(23–24), pp. 2981–2988. Available at: https://doi.org/10.1016/j.quascirev.2010.07.007.

Trauth, M.H., Larrasoaña, J.C. and Mudelsee, M. (2009) ‘Trends, rhythms and events in Plio-Pleistocene African climate’, *Quaternary Science Reviews*, 28(5–6), pp. 399–411. Available at: https://doi.org/10.1016/j.quascirev.2008.11.003.

Trinkaus, E. *et al.* (2003) ‘An early modern human from the Peştera cu Oase, Romania’, *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), pp. 11231–11236. Available at: https://doi.org/10.1073/pnas.2035108100.

University of Helsinki (2017) *FinnGen, a global research project focusing on genome data of 500,000 Finns, launched*. Available at: https://www.eurekalert.org/pub_releases/2017-12/uoh-fag121917.php (Accessed: 24 June 2023).

U.S. Department of Health & Human Services (2016) *Chapter 2: The neurobiology of substance use, misuse, and addiction*, *Facing Addiction in America: The Surgeon General’s Report on Alcohol, Drugs, and Health*.

Uyenoyama, M.K. (1997) ‘Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants’, *Genetics*, 147(3), pp. 1389–1400. Available at: https://doi.org/10.1093/genetics/147.3.1389.

Uyenoyama, M.K. (2005) ‘Evolution under tight linkage to mating type’, *New Phytologist*, 165(1), pp. 63–70. Available at: https://doi.org/10.1111/j.1469-8137.2004.01246.x.

Vernot, B. *et al.* (2016) ‘Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals’, *Science*, 352(6282), pp. 235–239. Available at: https://doi.org/10.1126/science.aad9416.

Vernot, B. and Akey, J.M. (2014) ‘Resurrecting Surviving Neandertal Lineages from Modern Human Genomes’, *Science*, 343(6174), pp. 1017–1021. Available at: https://doi.org/10.1126/science.1245938.

Le Veve, A. *et al.* (2023) ‘Long-term balancing selection and the genetic load linked to the self-incompatibility locus in Arabidopsis halleri and A. lyrata’, *Molecular Biology and Evolution*, 40(6), p. msad120. Available at: https://doi.org/10.1093/molbev/msad120.

Vidal, C.M. *et al.* (2022a) ‘Age of the oldest known Homo sapiens from eastern Africa’, *Nature*, 601(7894), pp. 579–583. Available at: https://doi.org/10.1038/s41586-021-04275-8.

Vidal, C.M. *et al.* (2022b) ‘Age of the oldest known Homo sapiens from eastern Africa’, *Nature*, 601(7894), pp. 579–583. Available at: https://doi.org/10.1038/s41586-021-04275-8.

Vignaud, P. *et al.* (2002) ‘Geology and palaeontology of the upper Miocene Toros-Menalla hominid locality, Chad’, *Nature*, 418(6894), pp. 152–155. Available at: https://doi.org/10.1038/nature00880.

Villanea, F.A. and Schraiber, J.G. (2019) ‘Multiple episodes of interbreeding between Neanderthal and modern humans’, *Nature Ecology and Evolution*, 3(1), pp. 39–44. Available at: https://doi.org/10.1038/s41559-018-0735-8.

Viscardi, L.H. *et al.* (2018) ‘Searching for ancient balanced polymorphisms shared between Neanderthals and modern humans’, *Genetics and Molecular Biology*, 41(1), pp. 67–81. Available at: https://doi.org/10.1590/1678-4685-gmb-2017-0308.

Võsa, U. *et al.* (2021) ‘Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression’, *Nature Genetics*, 53(9), pp. 1300–1310. Available at: https://doi.org/10.1038/s41588-021-00913-z.

Vrba, E.S. (1988) ‘Late Pliocene climatic events and hominid evolution’, in *Evolutionary history of the “robust” australopithecines*. Aldine de Gruyter, pp. 405–426.

Vrba, E.S. (1994) ‘An hypothesis of heterochrony in response to climatic cooling and its relevance to early hominid evolution’, in *Integrative Paths to the Past*. Prentice Hall, pp. 345–376.

Vrba, E.S. (1995a) ‘On the connections between paleoclimate and evolution’, in *Paleoclimate and evolution, with emphasis on human origins*. Yale University Press, pp. 24–45.

Vrba, E.S. (1995b) ‘The fossil record of African antelopes (Mammalia, Bovidae) in relation to human evolution and paleoclimate’, in *Paleoclimate and Evolution with Emphasis on Human Origins*. Yale University Press, pp. 385–424.

Wall, J.D. *et al.* (2013) ‘Higher levels of Neanderthal ancestry in east Asians than in Europeans’, *Genetics*, 194(1), pp. 199–209. Available at: https://doi.org/10.1534/genetics.112.148213.

Wang, H. *et al.* (2019) ‘Genome-wide association analysis of self-reported daytime sleepiness identifies 42 loci that suggest biological subtypes’, *Nature Communications*, 10(1), p. 3503. Available at: https://doi.org/10.1038/s41467-019-11456-7.

Wang, K., Li, M. and Hakonarson, H. (2010) ‘ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data’, *Nucleic Acids Research*, 38(16). Available at: https://doi.org/10.1093/nar/gkq603.

Watanabe, K. *et al.* (2019) ‘A global overview of pleiotropy and genetic architecture in complex traits’, *Nature Genetics*, 51(9). Available at: https://doi.org/10.1038/s41588-019-0481-0.

Watterson, G.A. (1975) ‘On the Number of Segregating Sites in Genetical Models without Recombination’, *Theoretical Population Biology*, 7(2), pp. 256–276.

Weidenreich, F. (1938) ‘Discovery of the femur and the humerus of Sinanthropus Pekinensis’, *Nature*, 141(3570), pp. 614–617. Available at: https://doi.org/10.1038/141614a0.

Weinstein, J.N. *et al.* (2013) ‘The cancer genome atlas pan-cancer analysis project’, *Nature Genetics*, 45(10), pp. 1113–1120. Available at: https://doi.org/10.1038/ng.2764.

White, T.D. *et al.* (1993) ‘New discoveries of Australopithecus at Maka in Ethiopia’, *Nature*, 366(6452), pp. 261–265. Available at: https://doi.org/10.1038/366261a0.

White, T.D. *et al.* (2003) ‘Pleistocene Homo sapiens from Middle Awash, Ethiopia’, *Nature*, 423, pp. 742–747. Available at: https://doi.org/10.1038/nature01669.

White, T.D. *et al.* (2009) ‘Ardipithecus ramidus and the paleobiology of early hominids’, *Science*, 326(5949), pp. 75–86. Available at: https://doi.org/10.1126/science.1175802.

White, T.D., Suwa, G. and Asfaw Berhane (1994) ‘Australopithecus ramidus, a new species of early hominid from Aramis, Ethiopia.’, *Nature*, 371(6495), pp. 306–312. Available at: https://doi.org/10.1038/371306a0.

Wiuf, C. *et al.* (2004) ‘The probability and chromosomal extent of trans-specific polymorphism’, *Genetics*, 168(4), pp. 2363–2372. Available at: https://doi.org/10.1534/genetics.104.029488.

Wong, N.A. and Bahmani, H. (2022) ‘A review of the current state of research on artificial blue light safety as it applies to digital devices’, *Heliyon*. Elsevier Ltd. Available at: https://doi.org/10.1016/j.heliyon.2022.e10282.

Woodward, A.S. (1921) ‘A New Cave Man from Rnodesia, South Africa’, *Nature*, 108, pp. 371–372.

World Health Organization (WHO) (2022) *World malaria report*. Geneva.

Wright, S. (1931) ‘Evolution in mendelian populations’, *Genetics*, 16(2), p. 97. Available at: https://doi.org/10.1007/BF02459575.

Wright, S.I. and Charlesworth, B. (2004) ‘The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model’, *Genetics*, 168(2), pp. 1071–1076. Available at: https://doi.org/10.1534/genetics.104.026500.

Yang, M.A. *et al.* (2017) ‘40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia’, *Current Biology*, 27(20), pp. 3202–3208. Available at: https://doi.org/10.1016/j.cub.2017.09.030.

Yousef, E. *et al.* (2020) ‘Shift work and risk of skin cancer: A systematic review and meta-analysis’, *Scientific Reports*, 10(1), pp. 1–11. Available at: https://doi.org/10.1038/s41598-020-59035-x.

Yu, N. *et al.* (2002) ‘Larger genetic differences within Africans than between Africans and Eurasians’, *Biological Research*, 161(May), pp. 269–274.

Yue, L. *et al.* (2004) ‘Paleomagnetic age and palaeobiological significance of hominoid fossil strata of Yuanmou Basin in Yunnan’, *Science in China, Series D: Earth Sciences*, 47(5), pp. 405–411. Available at: https://doi.org/10.1360/02yd0217.

Zaim, Y. *et al.* (2011) 'New 1.5 million-year-old Homo erectus maxilla from Sangiran (Central Java, Indonesia)', *Journal of Human Evolution*, 61, pp. 363–376. Available at: https://doi.org/10.1016/j.jhevol.2011.04.009.

Zeitoun, V. *et al.* (2010) 'Solo man in question: Convergent views to split Indonesian Homo erectus in two categories', *Quaternary International*, 223, pp. 281–292. Available at: https://doi.org/10.1016/j.quaint.2010.01.018.

Zerbino, D.R. *et al.* (2015) 'The Ensembl Regulatory Build', *Genome Biology*, 16(1). Available at: https://doi.org/10.1186/s13059-015-0621-5.

Zhang, Q. *et al.* (2008) 'Association of the circadian rhythmic expression of GmCRY1a with a latitudinal cline in photoperiodic flowering of soybean', *Proceedings of the National Academy of Sciences of the United States of America*, 105(52), pp. 21028–21033. Available at: https://doi.org/10.1073/pnas.0810585105.

Zhang, X. *et al.* (2023) 'MaLAdapt Reveals Novel Targets of Adaptive Introgression From Neanderthals and Denisovans in Worldwide Human Populations', *Molecular Biology and Evolution*, 40(1), p. msad001. Available at: https://doi.org/10.1093/molbev/msad001.

Zhao, B. *et al.* (2019) 'Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n = 17,706)', *Molecular Psychiatry* [Preprint]. Available at: https://doi.org/10.1038/s41380-019-0569-z.

Zietkiewicz, E., Richer, C. and Labuda, D. (1999) 'Phylogenetic Affinities of Tarsier in the Context of Primate Alu Repeats', *Molecular Phylogenetics and Evolution*, 11(1), pp. 77–83. Available at: https://doi.org/10.1006/mpev.1998.0564.