**Distribution Agreement**

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Cora Hirst April 12, 2022

# Evidence of Selection on Circadian Regulation of the Immune System in Ancient Iberia

by

Cora Hirst

Dr. John Lindo
Adviser

Department of Anthropology

Dr. John Lindo

Adviser

Dr. Dietrich Stout

Committee Member

Dr. Michal Arbilly

Committee Member

2022

# Evidence of Selection on Circadian Regulation of the Immune System in Ancient Iberia

By

Cora Hirst

Dr. John Lindo

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Anthropology

2022

Abstract

Evidence of Selection on Circadian Regulation of the Immune System in Ancient Iberia
By Cora Hirst

Various components of the immune system oscillate in a time-of-day dependent manner, a phenomenon that has been attributed to the circadian clock. Circadian regulation of the immune system is one of many mechanisms that allows the gradation of immune responses, likely timing immune robusticity to meet the demands of infection while limiting damage to tissues. Increased exposure to pathogens as human settlements adopted Bronze age technologies likely placed a distinct selective pressure on this regulatory system. Utilizing genome-wide SNP data of ancient Iberians during the Bronze age shift, this study aims to identify evidence of selection upon genes participating in immune system/circadian rhythm crosstalk. Specifically, allele frequencies at these sites of interest are compared before and after the historical shift through two differentiation statistics, $F_{ST}$ and *DAnc.* Of six significantly differentiated SNPs, five observed frequencies fail to be explained by a Wright-Fisher model of evolution under genetic drift. The Wright-Fisher models are re-implemented with varying selection coefficients to determine the average strength of selection acting at these loci. Signals of positive selection are reported for two Class II MHC variants, suggesting potential constraints on foreign antigen presentation. Signals of negative selection are reported for a regulatory variant of *KIR3DL2* known to decrease susceptibility to autoimmunity. Notably, this thesis reveals likely selection on two variants, one of the clock gene *BMAL1* and the other of chemokine receptor *CCR7,* known to be involved in the time-of-day dependent trafficking of lymphocytes to the lymph nodes. Future studies into the functional relationship between these two variants may provide insight into disease phenotypes correlated with differences in the timing of peak T-cell activation. Overall, this thesis demonstrates how archaeological data and advancements in ancient DNA sequencing can complement human niche construction hypotheses during times of large-scale cultural change, helping us understand how cultural environments can be embodied in our genomes and identify potential therapeutic targets.

# Evidence of Selection on Circadian Regulation of the Immune System in Ancient Iberia

By

Cora Hirst

Dr. John Lindo

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Anthropology

2022

Acknowledgements

I would like to extend the utmost gratitude toward the adviser of my thesis, Dr. John Lindo, for his time, effort, and patience as I grapple with this new world of ancient DNA. This thank you extends to the members of my thesis committee, Dr. Dietrich Stout and Dr. Michal Arbilly, for curating my interest in evolutionary theory and for their commitment to my project.

I would also like to thank Dr. Debra Vidali, Heather Carpenter, and the Department of Anthropology for helping me see this project through to its end; the graduate and undergraduate students of the Lindo Lab, for their kindness and vivaciousness; to Dr. William Shapiro, for an unending source of inspiration and a reminder not to take myself too seriously (or to split infinitives); to my family, for their understanding, support, and enthusiasm; and lastly, to my friends, in Atlanta and elsewhere, for keeping me calm and caffeinated.

Table of Contents

Tables



Equations

# Introduction

In 1794, a well-known physician, philosopher, and naturalist argued that evolutionary principles (as they would soon be known) would revolutionize medicine, helping mankind "unravel the theory of diseases" (Darwin, 1809). Only a generation later, his son would establish his evolutionary axioms - associating the inheritance of disease to the inheritance of variation - in his magnum opus, *On the Origin of Species.*

In recent years, the application of evolutionary principles to medicine has launched astounding and promising research programs, restructuring exhausted paradigms in fields diverse as oncology, virology, epidemiology, and immunology. Certainly, the logic of evolution has informed the genesis of robust hypotheses regarding the inheritance of disease and evolution of senescence. In combination with ecology, it has also influenced our perception of the interactions between pathogens and mechanisms of host defense, modeling our bodies as distinct ecosystems in which immune cells and pathogens co-evolve (Ebert and Fields, 2020).

John Maynard Smith's evaluation of evolution as "information transfer" is particularly interesting when looking at host-pathogen interactions in an evolutionary light (Smith, 2000). In brief, there is some validity in the assertion that selective forces act to mediate the transfer of information from one generation to the next. Catalytic binding on its own may be a purely physical process; the initiation and timing of that binding, differential and affinity-based chemical response, and the activation of appropriate downstream pathways, however, all serve as contextual clues that binding was instigated *to achieve some effect.* Even the use of the word "appropriate" reveals that the binding is based on its effect*,* perhaps regulated by a *necessary* effect, and thus prompts two important questions: under what circumstances is this binding necessary, and what is the necessity?

When applied to evolution, we must ask ourselves: what trait is being inherited and under what conditions is inheritance occurring? With these questions in mind, our genes and their functions allow us to look into the past and determine how the inheritance of environmental conditions (e.g culture) acts as a constraint on our evolutionary trajectories.

In recent years, evolutionary theorists (namely Laland and Pigliucci) have engaged these questions and proposed a shift in perspective when designing evolutionary research programs. The Extended Evolutionary Synthesis (ESS), though arguably not novel in *content,* emphasizes observations that the extragenetic (e.g. environmental, developmental, Lamarckian) information inherited by children from the parent generation shapes their evolutionary trajectory (Pigliucci and Mueller, 2010; Laland et al, 2015). And we can clearly see how that may be the case for niche construction, especially during times of expedited cultural change. Long-lasting changes in the environment, like those that emerge with the introduction of new agricultural technologies, very likely introduce new selective pressures into a population; consequently, interactions between the population and its new guild continue to place constraints on genetic information transfer through generations.

While niche construction is hardly a new consideration in the design of evolutionary research programs (even Rousseau's version of the history of mankind in his *Second Discourse* follows a certain theory of evolution under constraint), it becomes vastly more valuable in its application to the evolution of communication between body systems. Genes that evolve together may very likely not interact with one another (and may even be the result of hitchhiking or background selection). When the interactions between variants are unknown, knowledge of the selective pressures faced by a population can drive hypotheses regarding the potential phenotypic

fitness relative to the fluctuating variants; from these, functional genetic studies can be executed to evaluate the reason behind the variant disequilibrium.

This thesis aims to engage the theoretical framework of the EES by utilizing the concept of niche construction to develop a robust research question investigating the evolution of time-of-day dependent variation in immune response. The ancient DNA (aDNA) data used is publicly available, revealing that pilot studies under this computational framework need not be an expensive enterprise.

A literature review of bio/archaeological explorations into the material culture of a single, continuous population during the Neolithic-Bronze Age transition (~ 4250 - 1850 BCE) will allow for inference into related disease-related pressures while evidence from reviewed functional genetic studies will reveal mechanisms upon which natural selection may have acted. In particular, this study will explore variation in genes implicated in time-of-day dependent immune responses. Statistical techniques will be employed to take advantage of open-source aDNA data in a pilot study evaluating variant differentiation during the cultural shift. A null hypothesis of genetic drift will then be established for greatly differentiated variants during the time transect; for those variants whose observed differentiation cannot be explained under the neutral evolutionary scenario, a Fisher-Wright-based model will be built to test other potential selection scenarios.

The niche construction hypothesis will lastly be revisited to suggest the extension of the research question into a more rigorous aDNA capture technique and analysis, and to further recommend potential mechanistic research pathways to further investigate the therapeutic potential of identified targets.

As such, my research is a pilot study situated neatly at the intersection of archaeology, evolutionary biology, and functional genetics, and engages the theory of niche construction to pursue three distinct objectives:

1) To demonstrate how the EES can be applied to research projects utilizing new methods made possible by aDNA sequencing;
2) to evaluate the robusticity of the above methodology utilizing a case study of circadian rhythm - immune system regulation during a time of new pathogen exposure; and
3) to identify immune system / circadian rhythm targets for future study to further address the niche construction hypothesis.

That is, this study hypothesizes that selective pressures on circadian rhythm - immune system crosstalk emerged with the changes in diet, routine, and population size characterizing the shift to the Bronze age, and address the first of a series of questions that must be asked to tell that story : is there evidence of selection on genes that mediate this regulation? And if so, would a more robust, targeted amplification of these sights be valuable?

# Literature Review

## A. Circadian clocks carry temporal information

### i. The body is home to many clocks

Organisms among various taxa exhibit time-of-day dependent alterations to biochemical and physiological processes, known as *circadian clocks*. The preservation of circadian timekeeping across diverse clades - from cyanobacteria to *Homo sapiens* - provides strong evidence that daily biological rhythmicity evolved early on in the history of living organisms and suggests a distinct evolutionary advantage to the circadian regulation of biological activity (Bell-Peterson et al, 2005).

Comparative studies of circadian clocks across species reveal the different evolutionary trajectories of this "primordial" circadian clock in different clades, likely adapted to anticipate and optimize reactivity to time-of-day dependent variation in a species' niche. Among multicellular organisms, the specificity of circadian rhythms to distinct tissues prepares the tissue for habitual, environmentally-mediated changes. Together, these peripheral and autonomous clocks compose a *multi-oscillator* system in the organism (Bell-Peterson et al, 2005).

While the individual mechanisms that comprise these systems vary drastically across species, the conservation of the multi-oscillator phenomenon highlights the clear advantage that tissue-specific clocks confer. Elements of the feedback loops that auto-regulate circadian clocks receive environmental input to entrain a clock to environmental time, known as *zeitgebers,* or time-setters (Cermakian and Sassone-Corsi, 2002; Belle-Peterson et al, 2005; Lowrey and Takahashi, 2011). In doing so, these regulatory factors carry vital *temporal* information, instructing tissues to meet the demands of a niche by timing the expression of genes necessary for adaptive biological activity (Li and Zhang, 2015). Genes under the regulatory supervision of clock transcription factors are known as *clock-controlled genes* (CCGs), and their effects are diverse and curated to agree with both the intra- and extra-organismal environments (Li and Zhang, 2015).

**ii. Clocks are entrained by environmental factors**

Virtually all cells of a tissue express their own circadian clocks that drive the expression of CCGs specific to organ function. Of course, cell-autonomous clocks do not exist in isolation. They affect and respond to elements of other systems, often factors that themselves vary over the course of the day due to the effects of clocks in other cells and tissues. How, then, do cell-autonomous clocks syncronize throughout the tissue, the organ, and the body?

The central clock is housed in the *Suprachiasmatic Nucleus* (SCN) of the human brain, which receives signals directly from the environment and, through the oscillatory expression of CCGs, provides neurohormonal "cues" to dial peripheral clocks into environmental time (Cermakian and Sassone-Corsi, 2002; Bell-Pederson et al, 2005). The SCN receives temporal information from metabolites (food intake) and changes in temperature, but the main entrainer of the central circadian pacemaker is sunlight (Mohawk et al, 2012). Ganglion cells in the retina act as an interface between environmental time and internal activity, receiving and transferring daylight information via the retinohypothalamic tract to the master clock in the SCN (Lowrey and Takahashi, 2011).

At least three feedback loops interact with one another to autoregulate the master clock and are composed of the following core clock genes: *Cryptochrome (Cry1,2)* and *Period (Per1,2,3),* whose expression is dependent upon retinohypothalamic stimulus, and the constitutively expressed *Clock* and *Bmal1* (*ARNTL*) (Lowrey and Takahashi, 2011). Input from supporting genes, like the *Rev-erb* group, form ancillary loops that stabilize the primary circuit (Lowrey and Takahashi, 2011; Mohawk et al, 2012). Environmentally-regulated, rhythmic alterations to the expression and activity of *Bmal1* and *Clock,* mainly through *Cry* and *Per,* elicit changes in the expression of CCGs that stimulate the representation and relay of temporal information to the peripheral clocks (figure 1a; Bell-Perderson et al, 2005; Abele et al, 2019). In essence, this nucleus exerts top-down control on the body's clocks, directly responding to the environment and relaying that information to clocks in the "dark" periphery.

The core transcription/translation feedback loop regulating the rhythmic output of temporal information in the SCN is the same mechanism timing physiological activity in the peripheral clocks of individual cells. The main difference from the core clock is, of course, that

entrainment occurs in response to neuroendocrine signals received from the SCN. However, recent studies demonstrate that, in conjunction with nervous and endocrine input from the SCN, peripheral clocks can also respond indirectly to alternative zeitgebers (Husse et al, 2015). Though unsustainable without the "window to the outside" provided by the SCN, these second-order oscillations bypass the SCN by directly altering the activity of the core clock gene feedback system. For instance, changes in an organism's level of activity, food intake, and exterior temperature can be translated to information within the cell and alter the expression, form, and function of clock proteins (figure 1b; Husse et al, 2011). Curiously, time-of-day dependent changes in social demands may also act to entrain human peripheral clocks (Husse et al, 2011). For the purpose of this study, I will explore the zeitgeber information gleaned from pathogens to entrain the clocks of the immune system - and how they may take advantage of them.

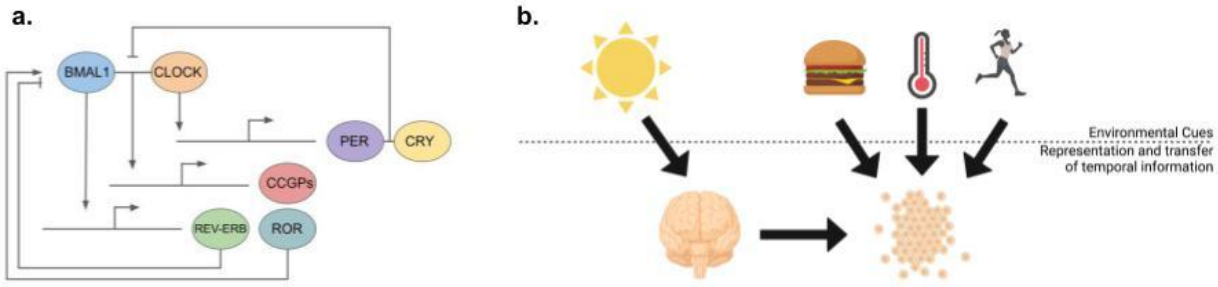


**Figure 1. Levels of peripheral clock entrainment.** A. The transcriptional/translational feedback mechanism exists in most cells of our tissues and times the expression of clock controlled genes (CCGs). Constitutively expressed *CLOCK and BMAL1* proteins form a heterodimer, active as a transcription factor promoting the expression of clock genes *Per, Cry, Rev-erba,b* and *rora,* and tissue-specific CCGs. PER and CRY dimerize to simulate the release of BMAL1 from CLOCK, inhibiting CCG expression. REVERBs and RORs similarly act to inhibit and stimulate BMAL expression, respectively; relative amounts of these two gene products vary due to internalized zeitgebers. B. Environmental zeitgebers time peripheral clocks both directly, through transformation of environmental information into internalized zeitgebers, and indirectly, through synchronization mechanisms originating in the SCN.

**B. Autonomous circadian clocks exist in cells of the immune system**

As discussed, the clock gene transcription/translation feedback system is not only conserved in some fashion across taxa, but is also present in most cells of almost all tissues of the human body. Entrainment of these tissues to environmental time thus creates an association between temporal information and the distinct time-dependent demands upon a tissue. By "keeping time", cells within tissues can effectively take advantage of this association through the expression of CCGs and prepare for predictable daily challenges to homeostasis.

Since virtually all tissues in the body benefit from this daily preparation, it's no wonder that oscillations have been described in many functional components of the immune system (Abele et al, 2019; Downtown et al, 2020). In the same way that the immune system must bring about a swift, efficient response to infection while limiting collateral damage done to the body's tissues during inflammation, it is likely that immune-specific circadian rhythms must strike a balance between effective preparation for habitual peaks in exposure to pathogens and adequate time for tissue repair after inflammation. In both scenarios, this balance requires communication between the two main networks of the immune system: the *innate* and *adaptive* responses.

**i. Clocks regulate information transfer between networks of the immune system**

Distinguished on the bases of specificity, speed, and rigor necessary to combat a pathogen, the innate and adaptive systems are interlocked, mutually coaching and stimulating one another throughout the course of infection. The two networks exchange information about the type and degree of danger posed by a pathogen and momentary needs at a site of infection through chemical signals called *cytokines.* Immune systems utilize another class of chemicals, *chemokines,* to traffic cells to the locations where they are needed.

Studies on the dysregulation of circadian clocks have revealed that circadian control of phagocytosis, immune cell trafficking, and cytokine expression is highly correlated with the day/night cycle and plays a significant role in mounting successful, graded responses to pathogens (Arjona and Sarkar, 2005; Druzd et al, 2017; Keller et al, 2009). Moreover, peripheral clocks in immune cells may play a role in carefully timed immunosuppression to allow for tissue rest and repair (Amir et al, 2018; Early et al, 2018, Kitchen, 2020).

**ii. Circadian clocks in the innate immune system**

Elements of the innate immune system are the first to be alerted that a pathogen has breached our skin and mucosal membranes. First, proteins of a general *complement system* circulate the blood in search of hydroxyl and amino group residues. Once bound, a cascade mechanism begins to recruit more complement system proteins and establish a formal cellular attack. Certainly, our own cells express surface proteins with these molecular patterns, and the complement system makes no discrimination between "self" cells and the cells of a pathogen. Instead, our cells express enzymes to inhibit the complement attack, essentially hiding them from the complement response. The complement system, then, carries quite general information: by-products of the attack act as signals to cells of the innate immune system, priming them to effectuate a more robust response, trafficking them to the site of infection, and tagging (opsonizing) pathogens for phagocytosis.

Many innate immune cells also express a class of *pattern recognition receptors (PRRs),* called *toll-like receptors (TLRs),* which recognize molecules that are essential to the survival of a pathogen (Pathogen Associated Molecular Patterns, or *PAMPS*). While TLRs cannot distinguish specific differences between similar pathogens, they are specific to certain PAMPS that distinguish clades of pathogens. And, because these PAMPS are obligatory to the survival of a

certain class of pathogens (i.e the lipopolysaccharide coating of gram-negative bacteria (LPS) or double-stranded RNA of some viruses), they are highly conserved. Activation of a toll-like receptor thus carries more specific information that there is *an active infection of a certain class in the body.*

The innate immune response gauges the extent of the infection by the extent to which its cells receive signals from the complement system and the pathogen directly. It then integrates this information to determine the necessary robusticity for attack, including phagocytosis (engulfing and digesting pathogens) and the secretion of inflammatory cytokines.

This integration of information is important for maintaining the balance between effectuating a powerful response against infection and limiting the amount of damage done to the body by powerful cells (i.e neutrophils) recruited during inflammation. It also allows the innate immune system to determine at which "point" of the infection the adaptive immune response should get involved - via the transfer of information to adaptive immune cells (figure 2).

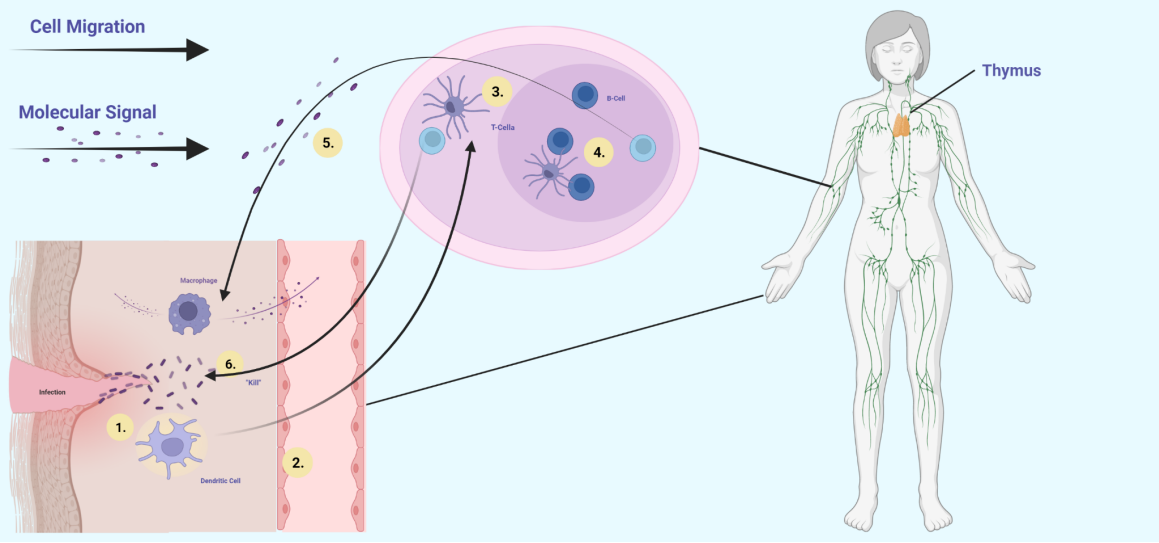


**Figure 2. Simplified stages of information transfer during recruitment of adaptive immune response.** Communication between innate and adaptive immune cells occurs in the secondary lymphoid organs: 1. APC activation via binding of a PAMP to a PRR delivers information to the phagocyte that the body has been infected. 2. As more PRRs are stimulated, the APC presents pathogen antigen on MHC molecules and migrates to the lymph nodes. Dendritic cell migration means it has received enough

stimulation from the infection to warrant an adaptive renic antigenesponse. 3. Antigen presentation, costimulation, and cytokines specific to the type of infection delivers specific information from dendritic cells to T cells about the identity, location, and strategies of the pathogen. 4. The identity of the pathogen is transferred from stationary dendritic cells to B cells that can recognize it, which present the identity to activated Th cells. When they "agree" upon an antigen, the B cell matures and produces antibodies. 5. The adaptive immune system delivers chemical signals to help the innate immune system target the right pathogens for phagocytosis, to alert the local immune response that the infection is ongoing, and to deliver the best strategy to eliminate the pathogen. 6. Activated T cells migrate to the site of infection, where some identify and eliminate infected cells based on the pathogen information they present, and others provide information to the innate immune system that the infection is ongoing and requires a certain strategy. 6. Lastly, in response to input from the adaptive immune system and the pathogen itself, the innate immune system delivers chemical signals to itself to maintain an inflammatory (cell-recruiting) state.

Oscillations in expression of immune system CCGs were initially described in elements of innate immunity, including macrophages, natural killer and dendritic cells, and the complement system (Labreque and Cermakian, 2015; Silver et al, 2012). Mouse *in vivo* studies have implicated the SCN in the regulation of peak phagocytic activity of macrophages, neutrophils, and dendritic cells (Hriscu, 2005; Silver et al, 2012). In humans, the effects of complement system byproducts on the attraction and activation of innate immune cells also appear to be regulated by circadian control. In particular, the complement fragment C5a, which mainly serves as a chemoattractant and cytokine received by macrophages, demonstrated sleep-dependent and circadian-correlated efficiency in humans (Reis et al, 2011).

A study by Keller and colleagues revealed that ~8% of the macrophage transcriptome (expressed genes) oscillates in a circadian manner. The same study also suggested that the clock transcription/translation feedback system operated within macrophages without SCN input (*in vitro)* (Keler et al, 2009). Since, cell-autonomous clocks that play a role in daily oscillations in phagocytosis, trafficking, and cytokine secretion have been identified in many cells of the innate immune system (Silver et al, 2012).

**iii. Circadian clocks in the adaptive immune system**

To transfer information to the adaptive immune system, dendritic cells and other antigen-presenting cells (known as APCs, including macrophages) must first receive stimulus from the complement system and direct binding of PAMPs to toll-like receptors. These stimuli encourage the APCs to upregulate the expression of a class of major histocompatibility complexes (MHCs). These molecules, class II MHCs, act as "podiums" that present pathogen-specific antigens (fragments of partially-digested proteins) to the adaptive immune system, recruiting only those immune cells that can recognize and target small, unique characteristics of the infectious agent. The APCs are *also* encouraged, through increased stimulation during an infection, to upregulate the expression of *co-stimulatory molecules*. These molecules are only expressed during an active infection (binding of PAMPS to PRRs), and thus serve to transfer a second important piece of information to activate the powerful adaptive response. That is, both the innate and adaptive systems **must agree that there is an active infection.**

The transfer of these two pieces of information - the identity of the pathogen (via it's presented antigen) and the innate-decided call to engage the adaptive immune system (via costimulation) - occurs mainly in the secondary lymphoid organs (i.g the lymph nodes). Not surprisingly, the migration of dendritic cells to the lymph nodes oscillates in a time-of-day fashion, alongside oscillations in the core clock genes of the transcription/translation clock mechanism (Holtkamp et al, 2021). So, it appears the circadian rhythm plays a role in regulating information transfer not only between cells of the innate immune system, but also between the cells of the innate and adaptive networks.

While the mechanisms through which circadian clocks regulate the innate immune system have been explored in great detail, it is only recently that studies have emerged about the mechanisms controlling the oscillatory behavior of T- and B- cells. Perhaps the hesitance to investigate the existence of cell-autonomous clocks in adaptive lymphocytes is an artifact of the information transfer between the two networks: because the adaptive immune system does not generally receive information directly from pathogens to be activated (except in the rare cases of T-cell independent B cell activation), daily oscillatory activity of the adaptive response could be explained by the innate-driven oscillations in information transfer.

Regardless of the reason for the delay in identifying autonomous clocks in T- and B-cells, the obligatory role of the circadian transcriptional feedback loops in the lymphocytes of the adaptive response is now well established (Downton et al, 2020; Scheiermann et al, 2018; Druzd et al, 2018).

During the adaptive response, helper T (Th) cell activation by APCs occurs mainly in the secondary lymphoid organs and is typically required to activate B cells, to stimulate cytotoxic T cells (CTLs), and to maintain an active innate immune response in the peripheral tissues. The Th cell surface protein CCR7 plays a key role in shuttling the cells from the blood into the lymph node. Naive Th cells appear to traffic to the lymph nodes in a rhythmic fashion correlated with their expression of core clock genes. The number of circulating Th cells that visit the lymph nodes does *not* oscillate throughout the day in knockout mice deficient in CCR7, nor knockout mice deficient in T-cell specific BMAL1 (Druzd et al, 2018). Cell-specific clocks within T-cells appear entrained to express CCGs related to trafficking - but because T-cells do not receive information directly from pathogens during activation, what environmental factor entrains them?

It has been proposed that the selective advantage to rhythmic T-cell movement is a response to the rhythmicity of antigen-presenting dendritic cell migration to the lymph nodes, increasing the abundance and diversity of activated Th cells and mounting a faster (and more diverse) adaptive response to infection (Druzd et al, 2018). In a positive feedback loop, the Th cells go on to re-stimulate elements of the innate system during infection, until the infection has been cleared below the threshold warranting the adaptive response. In this sense, the relationship between the temporal information carried by the circadian rhythm and the infection-specific information transferred between the innate and adaptive responses is clear. Entrainment of the adaptive response may be the result of indirect zeitgeber information transmitted both via the SCN and through interaction with the daily-oscillating innate immune system.

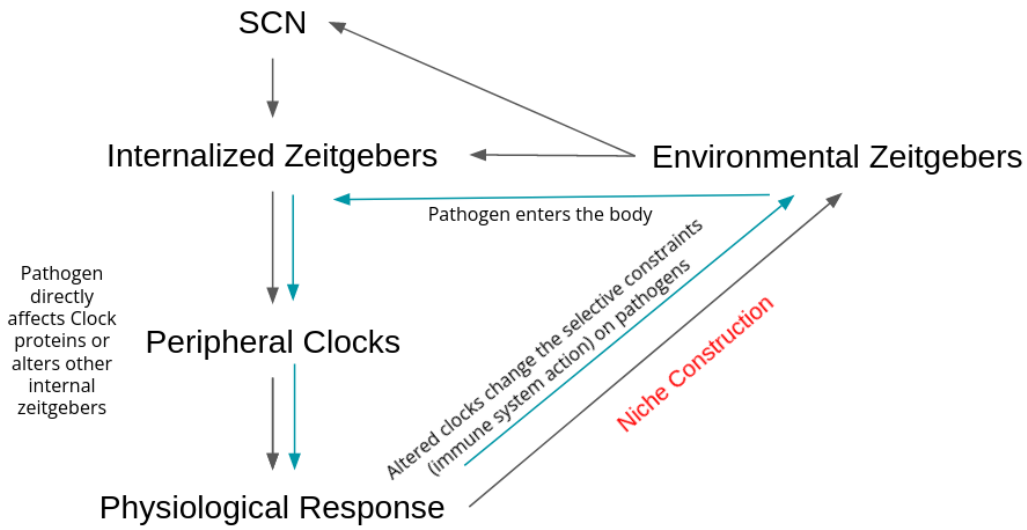


**Figure 3. Pathogens alter information transfer in the immune system through disruption of the circadian clock.** Transfer of environmental information to the SCN and internalized zeitgebers (Clock gene transcription factors, inhibitors of clock activity, etc) times and synchronizes peripheral clocks, which invoke time-of-day dependent physiological responses. These responses themselves either reaffirm the environmental zeitgebers (i.e the maintenance of a sleep schedule) or alter them, as is the case of niche construction. Pathogens can act as environmental zeitgebers, as well, or exploit this pathway by altering information transfer between the environment and peripheral clocks (blue arrows).

## C. Pathogens affect circadian control of immunity

### i. Hosts and their pathogens evolve together

The existence of immune system mechanisms across taxa reveals that hosts and pathogens have been engaged in a longstanding battle. The conservation of immune systems across extant species certainly suggests that they are descendents of a small few primordial systems. Perhaps more importantly, however, differences in their mechanisms provide us a window to the past. The immune systems we observe today are the results of different evolutionary trajectories shaped by different wars in branching lineages from the primordial system. This main driver of differential evolutionary trajectories is termed *host-pathogen coevolution.*

In some cases, hosts and pathogens can evolve strategies to maintain homeostasis in a reciprocal relationship, like the mutualistic bacteria that exist in the human gut microbiome (Quintana-Merci, 2019). In other cases, successful replication of a pathogen within a host occurs at the cost of the host's homeostasis; the immune system is a valiant effort to return the body to that state. But, by wielding weapons against the infection, the body's immunity places a selective pressure upon the microbial population, which usually has much shorter generation times than the average host. Individual pathogens who, by chance, develop a character that allows them to circumvent the immune system proliferate, replacing the majority population of the infection with a new, resistant population (Eberts and Fields, 2020).

The human immune system houses myriad techniques to accommodate for the speed with which its pathogens can accumulate insidious characters. Not only can the adaptive immune system recognize just about *any* organic molecule in existence (Sompayrac 20 ), and thus mount a response to a mutated antigen relatively quickly, but specialized cells like, natural killer cells,

can target infected cells that have "shut off" the information presentation systems that allow adaptive recognition. That is, when the default mode during infection is information transfer, a cell that presents *no information* becomes a target.

These strategies are not unique to humans, but are artifacts of selection upon previous host survival strategies. Still, numerous examples of selection on uniquely human characters against pathogens can be cited. A powerful example of evolutionary axioms at work in host survival strategies is the resistance conferred by sickle cell anemia against the malaria-causing *Plasmodium malariae,* known as Haldane's *Malaria Hypothesis* (Haldane, 1949). More recently, alleles that influence disease phenotype have been the revelations of genetic studies in populations who vary significantly in disease susceptibility from the global average. For instance, unique polymorphisms (namely Single Nucleotide Polymorphisms, or *SNPs*) in the PRRs of the innate immune system appear to confer resistance to tuberculosis in highly admixed lineages in West Africa (Azad et al, 2012; Mukherjee et al, 2019). Genetic studies on the population level have also demonstrated that polymorphisms in CCR5 are highly correlated with variability in HIV-1 susceptibility across populations (Hladik, 2005; Naicker et al, 2009). Evidence of the latter led to the success of the "Berlin Patient", a man whose bone marrow transplant from a donor deficient in the chemokine receptor CCR5 effectively cured his HIV-1 infection (Brown, 2015).

**ii. The circadian immune system is an advantageous disadvantage**

It is likely that the addition of temporal information into the immune system is yet another strategy evolved within animals during their war against pathogens. *In vivo* clock gene knockout models in mice have shown that dysregulation of the circadian rhythm can greatly increase disease susceptibility, viral load, and speed of pathogenesis to some pathogens *while*

*greatly reducing the efficiency of other pathogens* (Edgar et al, 2015; Majumdar et al, 2017; Ehlers et al, 2018). Variation in host susceptibility to different infections according to relative levels of different clock genes (which fluctuate in a time-dependent manner) is a clear indication that hosts may be *more prepared to battle certain kinds of infections during certain times of day;* what sort of evolutionary advantage might this convey?

There is, too, plenty of evidence to suggest that pathogens have developed their own mechanisms to take advantage of host circadian machinery. A study exploring the influence of PAMPs on the circadian rhythm demonstrated that these conserved pathogenic molecules, including LPS, flagellin (a common bacterial component), and ssRNA40 (a viral RNA), altered splenocyte (secondary lymphoid cells) clock gene expression *ex vivo* (Silver, 2017). In an *Arabidopsis* model, crosstalk between the plant's circadian rhythm and *Pseudomonas syringae* demonstrated that pathogens serve as another environmental cue to the entrainment of the immunological clocks via their modulatory effect on clock gene expression (Li et al, 2018). A growing body of evidence extends this crosstalk to mammals, implicating the gut microbiome in the regulation of clock gene expression in the *host genome* (Ku et al, 2020; Mattenchuk, 2020).

While the mechanisms of pathogen-induced changes in circadian rhythm are largely unknown, this evidence suggests that communication between microbes and human clock gene expression is hardly out of question. Considering the likelihood that circadian clocks in immune cells are artifacts of the evolutionary arms race between hosts and their pathogens, this communication may help maintain mutualistic relationships (i.e the gut microbiome), entrain the immunological clock to prepare for time-of-day dependent spikes in infection, and/or disrupt the information transfer necessary to coordinate immune response among the networks of the host immune system.

## D. Hypotheses regarding evolution in the human immune system

### i. Niche construction and Balancing Selection

The human immune system was evolving alongside pathogens long before *Homo sapiens* first emerged from Africa, and continued to do so as pathogens accompanied their migrations across continents (Karlsson et al, 2014). While paleolithic technologies helped early modern humans access new resources, they also exposed our ancestors to the myriad microbial populations extant in their expanding niche (Brennan, 1991; Karlsson et al, 2014). Indeed, signals of selection upon human immune system variants are strongest in those regions of the genome that directly influence susceptibility to diseases that emerged early on in the history of human migration, including malaria, tuberculosis, and smallpox (Anderson et al, 1982; Cohen, 2000).

As humans continued to colonize new environments - and construct niches within them - their immune systems faced everchanging selective pressures. In some cases, existing variation decreases at a genetic region, or *locus,* within a population. This *purifying selection* is evidence that the gene encoded serves a non-redundant, obligatory function in the new niche, and variation at this region cannot be tolerated (Quintana-Murci, 2019). In one population genetics study, intracellular toll-like receptors (those TLRs that bind patterns in viral nucleic acid sequences) were highly constrained, suggesting a history of purifying selection acting upon the functional activation of the innate immune system in response to viruses (Barreiro et al, 2009).

Curiously, the same study revealed that variation in extracellular TLRs was significantly less constrained; this variation may be reminiscent of independent moments of *positive selection.* The emergence of a *new* variant in an extracellular TLR may have conferred a reproductive advantage to the individual during a time of pandemic; the disproportionate increase in the

reproduction of this individual and his lineage increases the proportion of his variant in the population, but it's advantage is not *essential* to survival (Bareirro et al, 2009; Quintana-Murci, 2019).

Toll-like receptors like those explored by Bareirro and colleagues recognize molecular patterns on pathogens that *must* be conserved in the pathogen genome, because they are constrained by their own purifying selection (i.e are essential to the pathogen's survival). But pathogens can incur and establish high frequencies of beneficial mutations much faster than humans, due to short generation times and higher mutation rates. As discussed**,** strategies (like NK cells) used by the innate immune system to overcome previous attempts by pathogens to inhibit information transfer are evidence of a historic fixation of the beneficial phenotype somewhere in the ancestral line. Furthermore, the immense diversity of the adaptive immune system *within* the individual - via the variability in lymphocyte receptors - is a canonized strategy to recognize diverse antigens, even those emergent during an existing infection.

Population genetic studies have revealed *yet another* immune system strategy, this time at the population level: phenotypic polymorphism. In some instances, potentially deleterious variants are maintained under neutrality - what evolutionary biologists consider *drift* - because their functions are fulfilled by some other mechanism (Quintana-Murci, 2019; ). In some cases, the deleterious polymorphism actually confers a selective advantage during a future infection in the population; thus, the benefit of maintaining a storehouse of variability through functional redundancy is evident (Kraukauer and Nowak, 1999; Mills et al, 2001). In other cases, some immune system components, like the human MHCs, HLA I and II, are highly variable (i.e polymorphic) in humans (Hedrick and Thompson, 1983). These polymorphisms are maintained within populations at higher frequencies than would be expected if they were to drift

stochastically, suggesting that evolutionary constraint acts on these regions to *maintain* variation (Brandt et al, 2018; Eberts and Fields, 2020). By stocking the population with individuals employing unique toolsets, the "population immune system" can prepare for diverse threats that change from epidemic to epidemic. This strategy confers the survival of the *population* (as opposed to the individual) through a selective mode known as *balancing selection,* and host-parasite coevolution may be one of the main drivers for immune genes to evolve under a balancing selection model (Croze et al, 2016; Quintana-Murci, 2019).

**iii. Limitations to common evolutionary approaches**

Direct modeling of host-pathogen coevolution rarely happens in humans due to concerns with expense, ethical, and reproducibility. Most often, the studies occur initially in model organisms (like *Drosophila)* that have short enough generation times to directly observe forward changes in allele frequencies throughout the course of a selective pressure (Anju et al, 2020). While these models provide valuable insight into likely mechanisms that drive selection, they suffer consequences of sampling from a larger population, like inflating the count of rare alleles and masking selective sweeps (i.e a reduction in variation at neutral sites linked to a gene under selection) (Dehasque et al, 2020). While functional studies of the *phenotypic impacts* of selection targets can take place *ex vivo,* both functional redundancies and interactions with other systems in the host can result in dramatically different results *in vivo* (for an interesting study testing the incongruencies between *in vivo, in vitro,* and *ex vitro* approaches to modeling NK cell responses to stress hormones, see Gotlieb et al, 2015). Even when *in vivo* approaches can be made, commonly via transmutation and knockout studies in mammalian models, they are greatly limited in scope and reveal little about the dynamics of selection that may have occurred historically (Anju et al, 2020).

In a more descriptive methodology, recent selection in humans can be detected using population genetic data. Each mode of selection - purifying, positive, and balancing - leaves a unique signature on the locus of selection and its neighboring genomic regions. Positive selection is easily detected by comparing variation within loci between populations: differentiation between the communities that exceeds what would be expected of two different trajectories of neutral evolution is evidence of local adaptivity (Krakauer and Lewontin, 1973). Purifying selection leaves a slightly different signature, greatly reducing differentiation at the locus between populations but potentially increasing measured variation in surrounding regions (due to a decrease in population size) (Bareirro et al, 2008).

Balancing selection is a trickier evolutionary mode to glean from time-independent population data. Because balancing selection often maintains alleles at intermediate frequencies that would not be expected of a neutral evolutionary model, detecting it requires inference about the frequencies of polymorphisms and trends in mutation in an ancestor common to the modern populations (Brandt et al, 2018). What's more, balancing selection may result from different mechanisms that are difficult to distinguish from one another with data from a single moment in time (Brant et al, 2018; Quintana-Murci, 2019).

It should also be noted that momentary selective pressures driving frequency change at a locus under balancing selection may be mistaken for positive selection when utilizing a population genetics approach (Dehasque et al, 2020). And, since selection upon standing variation allows a more rapid response to a selective pressure than selection upon a *de novo* advantageous mutation (which must emerge in the population before it can be seen by selection), it is likely that much of the non-neutral evolution of the human genome is the result of

momentary selective pressures on standing variation as human populations rapidly built and amended their niches (Duhasque et al, 2020).

Genome-wide association studies (GWAS) have dramatically improved the identification of polymorphisms associated with observable phenotypes. These studies collect genomic data from populations exhibiting the focus phenotype - often disease - and compare it with genomic data from a population that does not exhibit the phenotype. Differences in the frequencies of known SNPs at each site in the genome are compared between the phenotype- and non-phenotype presenting groups, and those SNPs which occur at significantly different frequencies are distinguished as potential biomarkers for the trait. Thus, GWAS provides a powerful tool for inferring risk and tolerance factors associated with disease.

Unfortunately, GWAS studies are limited in their explanatory power. Some regions of the genome are "linked" and more likely to be inherited together, whether it be a functional linkage - as is the case for gene products that serve multiple purposes - or linkage due to physical distance (Smith and Haigh, 1974). Moreover, neutral locations in the genome may lose diversity if they are linked to regions undergoing purifying selection (Brandt et al, 2018). These two cases, *hitchhiking* and *background selection,* are fundamental tenets upon which GWAS identifies biomarkers associated with disease susceptibility. However, due to the indirect nature of selection on these markers, few claims can be made about the functionality of the SNPs.

Genome-Wide Selection Scans (GWSS) suffer a similar limitation. These studies look for differentiation between two closely related populations whose divergence (and subsequent occupation of different niches) is hypothesized to have introduced new selective pressures (Oleksyk, 2010). The process uses a *population branch statistic* (a variant of the Fst, see below) that measures the significance of divergence at genomic loci between populations. The statistic

relies upon the allele frequencies exhibited in the two *modern* groups to identify alleles whose frequencies vary significantly; therefore, SNPs that do not vary significantly are not identified. This tells us little about SNPs which may have evolved in parallel, due to similar selective pressures, and SNPs which may have greatly diverged in frequency or emerged in one lineage, but were masked by recent admixture between the two populations (figure 3a) (Oleksyk, 20210; Duhasque et al, 2020).

**iv. Insights from ancient DNA**

Ideally, the limitations to inferring selection from contemporary human populations would be circumvented by studying human genetic information from multiple time transects. In this approach, the effects of selective modes could be measured directly by identifying changes in allele frequencies through a time of hypothesized selective pressure (Quintana-Muci, 2019). Ancient DNA (aDNA) studies allow us to sample from ancient generations of a population to directly measure allele frequencies at multiple points in a lineages' evolutionary history (figure 3b).

aDNA studies of the last decade have provided tremendous insight into the modes *and* targets of selection in the human genome through time transects of human allele frequencies in known ecological contexts. These studies typically utilize the Wright-Fisher model to estimate the expected frequency change of a neutrally-evolving allele in a population through time, measure the observed frequency change gleaned from sequencing and genotyping ancient samples from multiple time points, and identify sites whose observed modern frequencies differ significantly from the expected drift model (Duhasque et al, 2020). Various statistical techniques can be applied to estimate the age of a variant, the strength of selection (i.e $S$ coefficient), and

even infer instances of selection upon variants introduced during population admixture (Malaspinas et al, 2012; Schraiber et al, 2015).

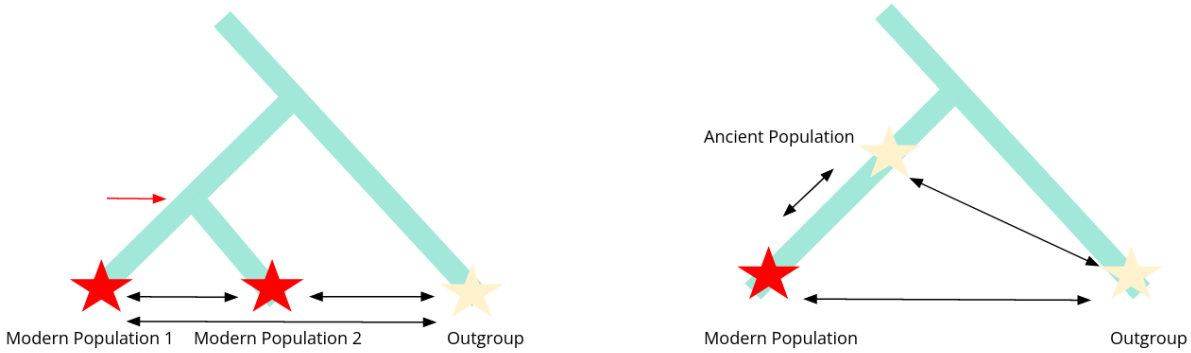


**Figure 4. Inferences of natural selection require fewer assumptions about evolutionary trajectories when using a time transect.** A. Measuring the divergence (black arrows) at a locus (yellow star) between two modern populations sometimes reveals variants which may have undergone selection, but can also mask positive selection (for instance, adaptation to a new niche). This is because an assumption must be made about the frequency of the allele of interest in the common ancestor population (red arrow; Garland et al, 1994): if both populations exhibit similar allele frequencies of the derived allele (red star), the most parsimonious explanation for divergence is drift. However, it may be that the same selective force acted on this locus, or that recent admixture between the groups has masked the difference in their evolutionary trajectories. B. This problem can be circumvented by isolating aDNA to sample allele frequencies directly from the ancient population, and a branching statistic can be calculated by a measure of its divergence from the modern population.

## E. Previous aDNA studies into immune system regulation

### i. Evolutionary modes can be inferred from a Wright-Fisher model

**In part C,** we explored how differences in immune system variants are a window to the past, the results of a long history of host-pathogen coevolution spanning numerous instances of inter- and intra-specific evolutionary events. Advancements in our abilities to excavate human remains (and make inferences about their material culture), to accurately isolate and sequence ancient human DNA, and to reconstruct genomic regions with increasing statistical power may allow us to make sense of the world outside that window, reconstructing stories from the past.

For one thing, aDNA has allowed us to explore paleoanthropological mysteries, uncovering, for instance, ancient interactions between Neanderthals, Denisovans, and early modern humans (Wall and Brandt, 2016). The field has also given insight into ancient population demographics, driving hypotheses about the migrational activity, admixture, and even cultural exchange in early human history (Ratkin and Racimo, 2016).

Statistical methods stemming from the Wright-Fisher model can be utilized to measure selective forces directly, magnifying changes in variant frequency from ancient populations down to the base pair (i.e SNPs). The logic of the Wright-Fisher model is that of a *discrete Markov chain,* a stochastic process where the value of an outputted variable is only dependent upon the value of the previously outputted variable (Tataru et al, 2017). When applied to biallelic (two alleles per gene) population genetics, the simplest form of the Markov chain models the evolution of a gene under genetic drift and is described by the transition matrix (listed probabilities of moving from one state to another) below:

$$P_{ij} = \binom{2k}{j} \left(\frac{i}{2k}\right)^{j} \left(1 - \frac{i}{2k}\right)^{2k-j}, \; for \; 0 \leq i, j \leq 2k$$

**Equation 1. The Fisher-Wright model of neutral evolution as the transition matrix of a Markov chain.** The count of one allele of a biallelic site in a population of size *k* is represented by *i;* the same count in the subsequent generation, a discrete time step forward, is represented by *j.* The possible counts of the allele fall between 0 (non-existent in the population) and 2*k* (fixed within the population, where each individual has two copies of the allele). Taking into account that an individual's genotype in the second generation is the result of inheriting a single allele from two different parents, random sampling with replacement from the first generation gives the probability of different allele counts in the subsequent generation. The dependence of the probability of a value of *j* on the value of *i* gives the Wright-Fisher model its Markov character.

Alterations to this chain can be made to account for forces of selection, changes in population size, admixture, and other processes that likely occurred as humans constructed their niches (Dehasque et al, 2020; Haller, 2016).

The Messer lab's forward genetic simulation framework, *SLiM1-3,* is a powerful computational tool allowing evolutionary biologists to model expected changes in a variant's frequency under set evolutionary modes (Haller and Messer, 2019). The Wright-Fisher based framework can be used in complement with ancient DNA data, where a modeled, expected change in a variant's frequency can be directly compared with the observed frequency change in a population along a time transect (Dehasque et al, 2020; Haller and Messer, 2019). Thus, by choosing to evaluate the change in SNP frequencies in a population *during* a time of suspected evolutionary pressure (e.g pandemic, technological innovation, or in the case of this study, agricultural revolution), hypotheses can be made regarding the functional mechanism of variants under selection. The Wright-Fisher model recipe was utilized for the purposes of this thesis (see Methods), and an example of the Eidos code is available in Figure 5.

Because evolutionary models are made on a site by site basis (a fantastic resolution brought to us by sequence data), there are simply too many polymorphisms to model directly. In the same way that selection scans and GWAS look to establish significant differences in allele frequencies between populations, we can utilize statistics that reveal variants whose total diversity cannot be accounted for by population genetic structure (predictable differences in allele frequencies within subpopulations). To distinguish variants with the possibility of selection in this study, two such statistics are leveraged for each SNP: the fixation index, Fst, and a Descent from Ancestry Score, DAnc.

**ii. Identifying candidate SNPs with Fst and DAnc**

The fixation index is a classical F-statistic proposed by Wright in 1949 to evaluate genetic structure differences between populations (Wright, 1949). Defined as "the correlation between random gametes, drawn from the same subpopulation, relative to the total", Wright's Fst score measures the amount of genetic variation at a site that cannot be explained by variance within subpopulations (Wright, 1943). In this sense, Fst can be thought of the ratio of the variance at a site between subpopulations to the variance at that site within the total population, given below:

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2}$$

**Equation 2. The fixation index, Fst, is a theoretical measurement of population differentiation due to genetic substructure.** Variation is conceptualized as the probability of randomly selecting two of the same allele from a subpopulation, superscript *S,* or total population, superscript *P*. Variance between subpopulations, $\sigma_s^2$, is weighted by the size of each subpopulation to mitigate potential bias (Hollsinger and Weir, 2009).

In practice, measuring diversity within subpopulations and within the total can be challenging. Estimators for Fst are often used, famously branching from Weir and Cockerman's use of jacknife procedures to estimate variances (Weir and Cockerman, 1984; Holsinger and Weir, 2009). Fst scores are thus necessarily positive, and range from 0 to 1: small values of Fst suggest that the frequencies of the allele within subpopulations are similar, while large values conversely result from a large $\sigma_S^2$ and indicate a disproportionately large variance for the allele frequency between subpopulations.

Descent from Ancestry (DAnc) scores are a branching statistic also calculated per site (Key et al, 2016). As a branching statistic, the DAnc score conceptualizes population differentiation as the difference between each population's allele frequency with an outgroup (thus, from a common ancestor), as given below:

$$DAnc(P_1,\ P_2,\ A) = \Delta(P_1, A) - \Delta(P_2, A)$$

**Equation 3. DAnc is a branching statistic that measures population differentiation for a given polymorphism.** The extent of differentiation of an allele between two populations, $P_1$ and $P_2$, due to different evolutionary trajectories is inferred by utilizing an outgroup, *A,* to establish a common ancestor. Allele frequency is estimated from the *derived ancestral allele,* the alternative of a shared ancestral allele.

DAnc scores thus range from -1 to 1, and can be used in conjunction with time-delimited population data to infer possible cases of selection:

$$DAnc(M,\ A,\ O) = \Delta(M, O) - \Delta(A, O)$$

**Equation 4. DAnc can be used to infer extreme allele differentiation along a time transect.** In this case, the populations under consideration are different generations (a more recent, *M,* and an ancient, *A*) from a continuous population, and are compared with an outgroup, O, to infer differentiation from a common ancestral population.

An extreme negative score suggests that the allele frequency of the ancient population differentiates from the modern and outgroup populations, while an extreme positive score suggests that the modern population differentiates from the ancient and outgroup populations. The latter scores are inferred signatures of recent positive selection, while the former may indicate the emergence and cession of a selective pressure around the ancient population.

**iii. Ancient DNA studies on the evolution of immune system genes**

Genome-wide selection scans utilizing aDNA have implicated numerous immune-system SNPs in recent selective regimes. A scan of 230 ancient West Eurasians living between 6500 and

300 BC revealed strong signals of selection on the TLR10-TLR1-TLR6 cluster of toll-like receptors (Mathieson et al, 2015). Because toll-like receptors activate downstream cascades involved in the development of efficient immunity upon binding to a conserved pathogenic pattern, they are curious markers of long-term (multi-generational) memory in the immune system, retaining environmental information (i.e *this* pattern is pathogenic) in the form of a ready-to-activate receptor. Thus, the Mathieson study, corroborated by others examining the selection upon different HLA (human MHC) phenotypic variants during the Mesolithic, have scratched the surface of pathogen-mediated immune system evolution in mankind during the Holocene (Olalde et al, 2014).

Even now, very few aDNA genomes could be completely isolated and sequenced with certainty. The result of aDNA sequencing and genome reconstruction techniques is often rather the collection of SNP data, allowing for the genotyping of various (but not all) known SNP locations in the human genome. Utilizing theories of niche construction and gene-culture coevolution - where humans act to change their environments, and the change in environment is inherited by and places new selective pressures on the subsequent generation - hypotheses can be made about potential genetic targets of selection during a time of known cultural evolution (Oldling-smee et al, 2013; Deschamps et al, 2016; Quintana-Murci, 2019).

The utilization of feed-forward simulations has had a profound impact on uncovering the stories of host-pathogen coevolution through a time transect. A selection scan using ancient exomes from a pacific northwest Native American population, the Tsimshian, revealed both signals of recent positive selection on an *HLA-DQA1* allele near fixation before European contact, *and* intense frequency decline to the state of the modern population (Lindo et al, 2016). A feed-forward simulation model with a positive-to-negative selection shift was the only

selection regime to agree with the observed modern variant frequency, suggesting that previous positive selection upon the HLA variant may have conferred a stronger susceptibility of the population to pathogenesis during the changing disease-landscape of the colonial era (Lindo et al, 2016).

Studies like the latter provide strong evidence for the potential selective benefit of maintaining phenotypic polymorphism within a population like that seen of *HLA,* and may streamline the identification of functional therapeutic targets in the context of diseases susceptibility, pathogenesis, and outcome. Perhaps more significantly, they demonstrate how aDNA has revolutionized our ventures into uncovering the nuanced evolutionary trajectories of a population lineage. In the Lindo study, a scan of the modern Tsimshian (utilizing a branching statistic) without aDNA failed to identify the *HLA-DQAI* variant as a candidate of selection; similarly, a demographic history model accounting for european admixture failed to account for the significant frequency change from the ancient to modern population (Lindo et al, 2016). Simply put, without an aDNA-mediated time transect, the Tsimshian story of positive to negative selection would be lost to history, and so too our obligation to investigate its cause.

A different publication utilizes a distinct probe-mediated gene capture protocol to amplify the sequence count of immune genes of particular interest during a time of hypothesized pathogenic pressure. The study evaluated the changes in frequency to the modern community of 488 immune-related genes from a mass grave site associated with the Black Death in Ellwangen, Germany (Immel et al, 2021). In lieu of a whole-genome selection scan, the scan was conducted on these 488 genes in particular, using DAnc and Fst scores to identify SNPs whose differentiation may be the result of selection. Selection upon variants of both class I and II human MHC complexes, as well as a NK cell receptor, was inferred from SLiM simulations with

varying *S* parameters, corroborating the hypothesis that epidemic eras may place selective pressures on standing variation within the immunological exome (Immel et al, 2021; Quintana-Murci, 2019).

Studies like the above have only begun to break into bio anthropological literature, providing a direct molecular lens into the histories of human evolution during large scale cultural and environmental change. That is, the resolution of our "window to the outside" is increasing rapidly, allowing us to examine the evolutionary processes by which humans embody the environmental information they help to shape. To the author's knowledge, no studies have yet been conducted utilizing an aDNA time transect to identify the influences of cultural change on the evolution of the circadian rhythm. This study aims to highlight that gap with an example of circadian rhythm evolution, possibly mediated by changes in pathogen zeitgeber information, during Bronze-age agricultural revolution in the Iberian Peninsula.

**F. Archaeology of the Iberian Peninsula**

"Change in pathogen zeitgeber information", as I have previously called it, is just one of many examples of selective pressures that may emerge as humans change their environments. It is therefore crucial to establish the environmental and cultural changes that characterize the MLN-BA shift in Iberia, particularly as they relate to exposure to new pathogens. The SNP data used in this study is a subset of the open source dataset built by Olalde et al in the paper "The genomic history of the Iberian Peninsula over the past 8000 years"(Olalde et al, 2019). Specifically, a time transect of known material culture change was established between two distinct capture points: the mid-to-late Neolithic (~5500-3000 BCE) and the Bronze Age (~2200 - 1500 BCE). It is therefore crucial to establish the environmental and cultural changes that

characterize the MLN-BA shift in Iberia, particularly as they relate to exposure to new pathogens and pathogenic strategies.

**i. Changes in environment and material culture defined Bronze Age Iberia**

The genome-wide SNP data collected in the Iberian dataset spans the entirety of the peninsula, though scattered mainly along the coastline with inward migration during the BA (Oldalde et al, 2019; Figure 5). Pre-agricultural paleo-diets were typically constrained by the availability of resources in a given environment, as evidenced by the diversification of diets detected in pre-farming mesolithic Iberians (Cubas et al, 2019). Not surprisingly, this constraint becomes greater as populations engage in ritualized agricultural practices, homogenizing their diets to crops that can be dependably farmed each season. Isotopes collected from Neolithic dental enamel near the Mediterranean strongly suggest a homogenized diet of C3 plants (cereal grains) with little animal meat intake (Cubas et al, 2019). These stable foods were likely prevalent at most sites throughout the continent, with little evidence of C4 (grain, millet, sorghum) plant consumption until the late Bronze age (Lopez-Costas, 2019). Curiously, it appears that marine resources provided little to no supplement to the Neolithic diet, despite their proximity to the coast (Diaz-Zorita et al, 2019).

The Bronze age transition in the Southwest can be characterized by the establishment of hilltop settlements scattered among fertile lowlands with conditions particularly well-suited for rain-fed barley (Knipper et al, 2020). Throughout the north, landscape dryness during the growing season and absence of irrigational artifacts likely limited the cultivation of wheat, similarly implicating the less-laborious barely as the source of recollected C3 plant isotopes (Knipper et al, 2019). Meat and dairy foodstuffs were likely consumed much more during the Bronze age than the Neolithic, comparing time-separated remains from sites in close proximity

across the peninsula (Plantinga et al, 2012; Lopez-Costas et al, 2019). Meat consumption has been attributed to both wild and domesticated animals, though the former is thought to account for a greater proportion of meat-based foodstuffs (Knipper et al, 2020).

While the shift to primary food production and sedentism associated with neolithization certainly contributed to population growth, neolithic communities appear to have occurred in smaller demes (Alt et al, 2016). However, the emergence of large-scale funerary practices, evidenced by megalithic tombs like Alto de Reinosos (Burgos, Spain) may have been a cross-community unifying factor. Dietary analysis of group burials reveals high variation in staple foodstuffs, suggesting non-local burials, and food preferences of groups >600km apart shared striking resemblance despite variation in substencence resources (Alt et al, 2016; Fontanas-Coll, 2017). Taken together, there is not only evidence of homogeneity in metabolites for Neolithic individuals throughout the peninsula, but also evidence of cross-region exchange of subsistence strategies and, likely, gene flow.

Conversely, highly varied intrapopulation diets indicated by isotope analyses of the Bronze age hilltop/valley communities suggests both substantially larger population sizes and a greater complexity of food distribution than observed in the egalitarian neolithic, likely associated with stratified socio political structures (Lopez-Costas, 2019). Furthermore, due to the topographic isolation of a potentially emergent farming class from their hilltop contemporaries, new strategies for the distribution of foodstuffs was likely of high necessity. In light of both the increased meat and dairy consumption and the necessity of grain delivery and storage during the Bronze age, it is fair to conjecture that subsistence strategies were dramatically altered during the MLN-BA shift.

Direct evidence of endemic pathogens in ancient communities is difficult to establish, as the decay of soft tissue greatly restricts techniques to the identification of pathogens in calcified blood, bone and bone lesions, and targeted arrays on microbial DNA; evidence of ancient viral infections is perhaps even more of a challenge (Callaway, 2015; Spyrou et al, 2019). While the literature on the peninsula is lacking targeted pathogen studies, population growth, inferior or absent waste-removal systems, and increased interaction with domesticated animals introduce reservoirs, vectors, and direct contact points between humans and human pathogens (Armelagos et al, 1991). Evidently, these conditions became much more established during the MLN-BA shift, with static population sizes grand enough to theoretically accommodate endemic disease (Armelagos et al, 1991)

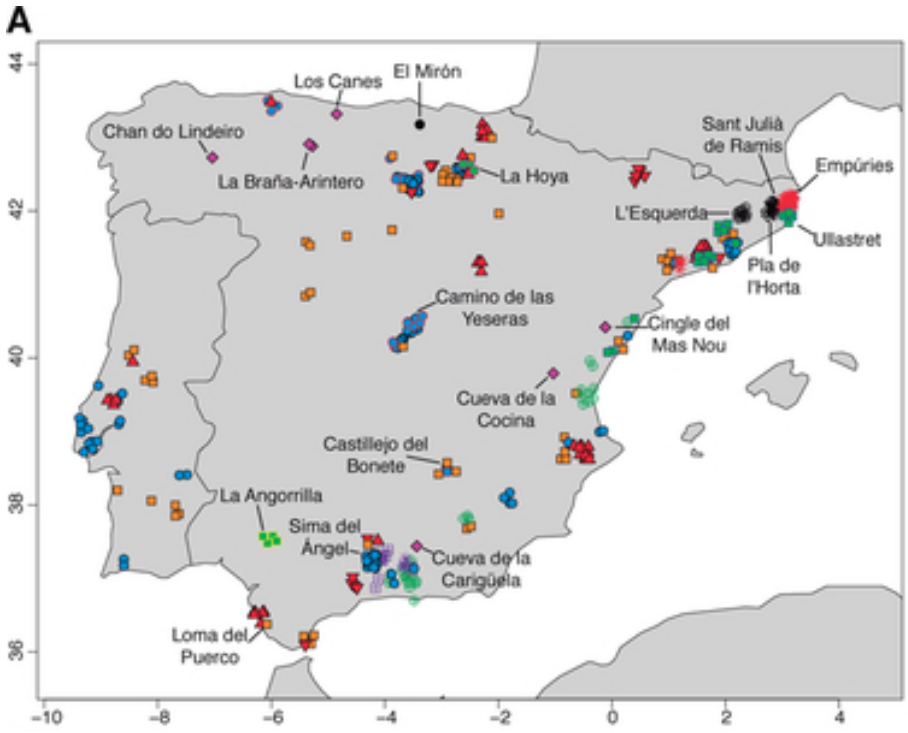


**Figure 5. Geographic distribution of collection sites in the Iberian Peninsula (Olalde et al, 2019, Figure 1A).** The data used in this study was collected from human remains at sites dated back to the mid-to-late Neolithic (MLN, red, upright triangles) and the Bronze age (BA, orange squares). The MLN-BA shift brought with it inland migration.

**ii. Extent of gene flow during the Bronze age is disputed**

The neolithic Iberian Peninsula demonstrates significantly lower genetic diversity than the rest of western Europe during the period, documenting only initial admixture events with northern Africa during the mesolithic (Serrano et al, 2021). However, Bell-beaker pottery and ivory culture from North Africa dating to the early Bronze age have been excavated from Northern and Central sites, like Camino de las Yeseras (Madrid, central Iberia), proposing long-distance exchange and a potential source of gene flow (Schuhmacher et al, 2009). Indeed, a single haplogroup of North African origin, Lb1, was contemporaneously present in West-central Iberia (Szécsényi-Nagy et al, 2017). In conjunction with the identification of Steppe-related haplogroups in Northern Iberian samples from the late Bronze age, the continuity of the population has been called into question (Olalde et al, 2019). A slew of population genetic analyses across the region subsequently emerged, launching a debate into the relative contributions of Steppe haplotypes to the overall genetic diversity of the BA Iberian peninsula.

According to an ancestry analysis previously conducted through mitochondrial haplotypes, Bronze age interactions only slightly increased the Steppe-related ancestry of the Neolithic population, an observation also common to contemporary Mediterranean Bronze age populations (Valdiosera, 2018). What's more, the genetic structure of southwestern settlements, like La Bastida (Totana, Murcia), remained particularly similar to its neolithic counterpart. Of particular relevance to this study, it has also been proposed that continuous and *increasing* Admixture between particularly continuous populations in the southwest with populations in the North and Central regions diluted the influence of Steppe ancestry in the population genetic structure (Szécsényi-Nagy et al, 2017; Villalba-Mouco, 2021).

While this admixture may have introduced novel variants into the Iberian gene pool and influenced observed changes in standing variation during the MLN-BA shift, the minimal contribution of Steppe ancestry to the population genetic structure will likely only play a minimal role in the identification of potentially selected variants (and the suggestion of a robust future study design) pursued here.

# Methods

## A. Description of dataset

This study utilized an open-source genome-wide SNP dataset of 271 ancient Iberians collected by *Olalde et al* (Olalde et al, 2019). The next-generation sequencing libraries were enriched for ~1.2 million SNPs previously documented in Eurasian genomes, and individuals showing evidence of contamination, first-degree kin, or low coverage (<10,000 SNPs) were filtered from the resulting filesets. While the 271 individuals ranged from the mesolithic to the historical period, only those individuals dated to the MLN (n = 44) and BA (n = 53) were analyzed for the purpose of this thesis.

## B. Establishing genetic continuity in temporally separated populations

The genomic toolset *PLINK* was utilized to subset SNP data from MLN and BA individuals and filter SNPs at sites with a genotyping rate below 50% (Purcell et al, 2007). The study also utilized SNP data from the Yoruba genomes included in phase 3 of the *1000 Genomes* project as an outgroup (1000 Genomes Project Consortium, 2015). Reference alleles for both datasets were inferred from the *1000 Genomes* Yoruba subset.

Before identifying candidate alleles for evolution simulations, it must be established that the mid-to-late Neolithic (MLN) and Bronze Age (BA) individuals form a continuous population. Principal component analysis, an orthogonal transformation-based dimensionality

reduction algorithm, was utilized to contrast the genetic similarity between the two populations with their similarities to 65 published West-Eurasian populations. The genomic data for these modern populations was extracted from the *Simons Genome Diversity Project* (Mallick et al, 2016). PCA was conducted using EIGENSOFT's *smartpca* executable (Galinsky et al, 2016), and single genomes were mapped onto PCA space according to the first two principal components utilizing the Novembre lab's *PCAViz* r-based package (Williams et al, 2017).

**C. Identifying candidate variants for selection simulation**

To determine allele frequency changes through time, only the alleles present in both Iberian populations *and* the Yoruba outgroup were considered. A list of known variants for each gene of interest was downloaded from the *NCBI dbSNP* database and called from the three populations; the union of variants among the three populations was used to calculate the differentiation statistics.

Two statistics were calculated and considered when determining variants with potentially significant (non-neutral) frequency changes. *DAnc* scores were calculated per site for the 18 variants present in all datasets using the derived allele frequencies (DAF) obtained from the - - freq command of the *PLINK* (v2.0) software (Purcell et al, 2007). It should be noted that the DAF was inferred from the minor allele frequency of the outgroup and set as the derived (e.g alternate, minor) allele in both Iberian datasets. $F_{st}$ scores were obtained using the Weir-Cockerham estimator implemented in the *VCFtools* - - weir-fst-pop command (Danecek et al, 2011). An empirical cumulative distribution function (Ecdf) was estimated from the observed $F_{st}$ scores for all genome-wide shared variants (n=19655). The Ecdf was evaluated for each variant's $F_{st}$, and due to the one-tailed nature of $F_{st}$ score distribution, its empirical p-value taken as *1-p.*

**D. Establishing evolutionary scenarios using SLiM3**

1000 neutral forward genetic simulations were run using Haller and Messer's *SLiM3* to establish a null distribution of frequency changes expected for each variant under drift (Haller and Messer, 2019). Population size was estimated to be ~5,000 at the start with a population growth rate of ~1.006 per generation, taken as the average of published estimated growth rates for late Neolithic Iberia (Armelagos, 1991). A time transect was established by estimating the MLN Iberia as 63 kya (4250 BCE) and the BA Iberia as 39 kya (1850 BCE). Assuming a generation time of ~25 years, selected for reproducibility in accordance with the human demographic model published in Gravel et al (2011), 95 generations were simulated. Genomic elements were modeled as complete genes and mutations introduced at the SNP site relative to the first 5' base of the gene. Lastly, the sampled SNP frequencies from the MLN were used as input. A complete list of parameters for each of the six candidate SNPs is provided in Table 1.

The observed BA SNP frequencies were compared to the expected distribution of frequencies under drift; frequencies > 97.5% and < 2.5% quantiles of their expected neutral distribution were selected for natural selection simulation.

To simulate natural selection, a non-zero *S* parameter was added to the SLiM3 model described above, controlling the strength of selection. Positive *S* values represent positive selection, while negative *S* values represent purifying selection. A range of ± *S* parameters was tested (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1), and the selected *S* parameter was determined as the mean value under which the observed frequency fell within the inner 95% of the expected frequency distribution.

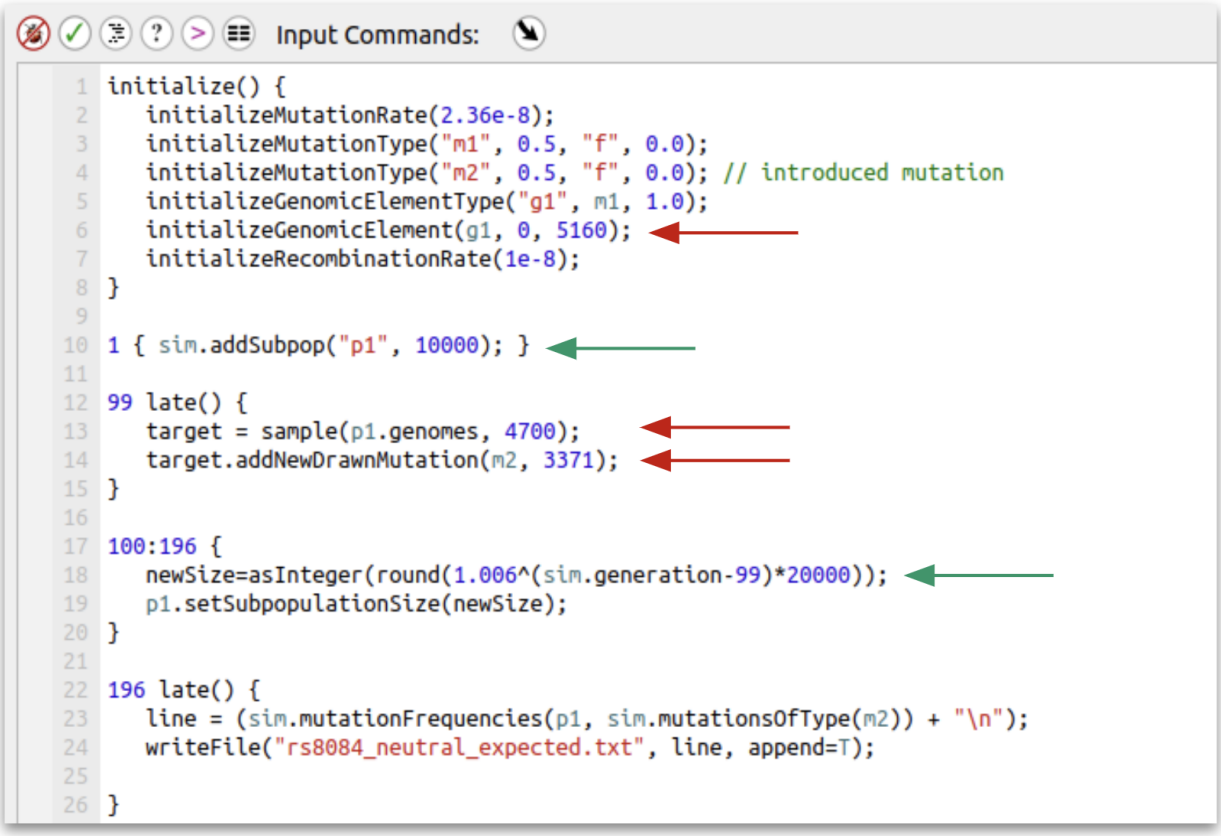


**Figure 6. SLiMgui interface and Eidos script for a neutral selection scenario.** Green arrows indicate the non-allele-specific parameters of initial population size and growth and neutral selection coefficient (0.0). Red arrows indicate parameters that were changed to model distinct evolutionary trajectories of alleles of interest, including starting frequency, gene length, and location on the gene. While mutations emergent in SLiM simulations do not stack, it is important to include the latter two parameters in future models of potential haplotype and linkage scenarios.

**Table 1. Input Parameters for *SLiM3* Neutral Selection Simulation**

| dbSNP rsID | Gene Length (nt) | Variant Location (nt) | Initial Frequency | Population Size* | Population Growth Rate* |
|---|---|---|---|---|---|
| **rs3128970** | *13,707* | *13,466* | *0.2022* | *10,000* | *1.006* |
| **rs8084** | *5,160* | *3,371* | *0.2350* | *10,000* | *1.006* |

| | | | | | |
|---|---|---|---|---|---|
| **rs3745902** | *16,890* | *16,233* | *0.5000* | *10,000* | *1.006* |
| **rs2853950** | *6,000* | *5,000* | *0.0120* | *10,000* | *1.006* |
| **rs72867447** | *110,717* | *3,776* | *0.4091* | *10,000* | *1.006* |
| **rs3136685** | *11,704* | *3,500* | *0.2308* | *10,000* | *1.006* |

***static parameters for all SLiM3 recipes**

# Results

## A. Mid-to-late Neolithic and Bronze Age Iberia forms a single continuous population

Considering the possibility of admixture indicated by Bell-beaker artifacts and potential Steppe-related ancestral haplotypes documented by *Olalde et al*, PCA was conducted to test the genetic similarity of the two populations (MLN and BA). When plotted according to the first to principal components, the MLN population formed a tight cluster in PCA space, overlapping the cluster of BA genomes; these results suggest that the two cohorts form a single continuous population (figure 7).

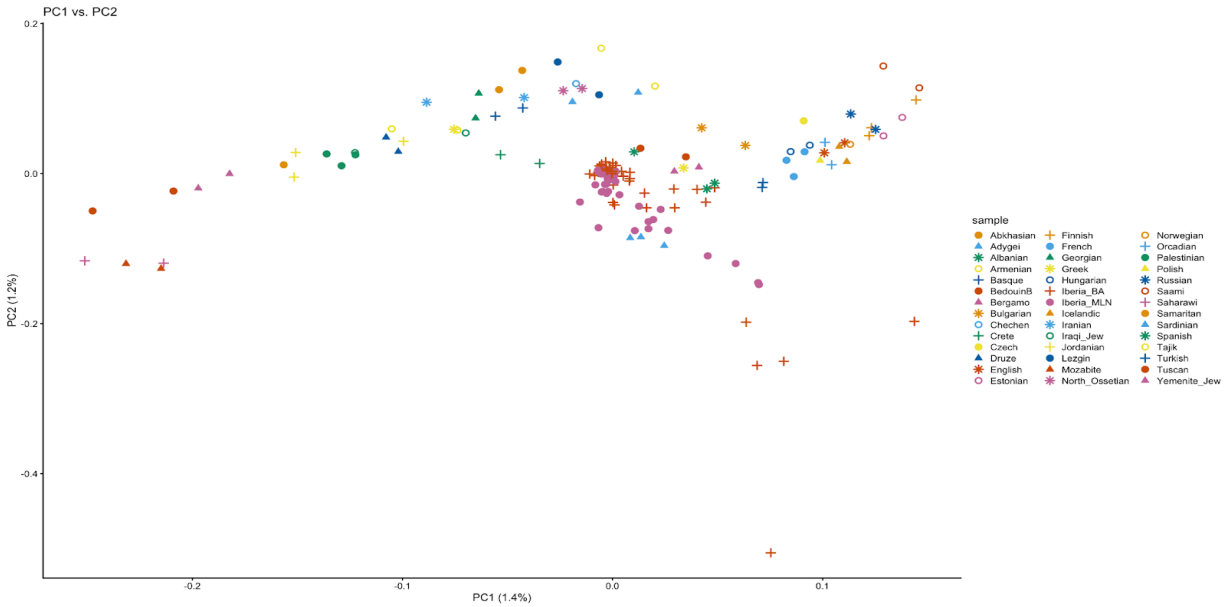


**Figure 7. PCA plot reveals tight clustering of MLN and BA single genomes in PCA space.** Bronze age genomes are visualized as orange crosses; MLN genomes are visualized as purple circles.

## B. One circadian rhythm and five immune system polymorphisms are highly differentiated

*NCBI dbSNP* - registered polymorphisms for 38 genes involved in circadian rhythm-immune system crosstalk were extracted from the two Iberian subpopulations, as well as an outgroup from west-African Yoruba. 32 polymorphisms were shared among the three cohorts. A cumulative probability dentistry function was estimated from the observed distribution of per-site $F_{st}$ statistics calculated per-site across the genome. Variants of five immunity related genes (KIR3DL2, HLA-DPB1, HLA-DRA, CCR7, and HLA-C) and one core clock gene (BMAL-1) had empirical $p < 0.05$, suggesting with 95% confidence these polymorphisms are strongly differentiated between MLN and BA subpopulations. This inference is bolstered by the per-site calculation of DAnc scores for each of the 32 shared polymorphisms, as the six highest DAnc scores correlated with the same variants (Table 2).

**Table 2. Significantly Differentiated Immune System and Clock Variants during the MLN-BA Shift**

| dbSNP rsID | Gene | Type* | Ref | Alt | Yoruba | MLN | BA | Danc | Fst P-value |
|---|---|---|---|---|---|---|---|---|---|
| **rs3745902** | *KIR3DL2* | *missense_variant* | *C* | *T* | *0.0110* | *0.5000* | *0.2110* | *-0.2890* | *0.03109* |
| **rs3128970** | *HLA-DPB1* | *3_prime_UTR_ variant* | *T* | *G* | *0.9676* | *0.2022* | *0.4885* | *-0.2863* | *0.02966* |
| **rs8084** | *HLA-DRA* | *Coding_sequence_ variant* | *A* | *C* | *0.5278* | *0.2350* | *0.4815* | *-0.2465* | *0.03063* |
| **rs3136685** | *CCR7* | *upstream_transcript_ variant* | *C* | *T* | *0.5278* | *0.2308* | *0.0189* | *0.2119* | *0.04424* |
| **rs72867447** | *BMAL-1* | *upstream_transcript_ variant* | *C* | *G* | *0.0509* | *0.4091* | *0.2114* | *-0.1977* | *0.04452* |
| **rs2853950** | *HLA-C* | *downstream_transcript _variant* | *C* | *T* | *0.4444* | *0.0106* | *0.1073* | *-0.1053* | *0.03185* |

***variant types were annotated from their *NCBI dbSNP* entries**

## C. Selection likely explains differentiation of HLA, KIR3DL, and CCR7 variants

Because the data ascertained is not whole-genome, but rather restricted to SNPs, strong differentiation statistics cannot alone be used as evidence for selection. Instead, the observed MLN frequencies for each of the six variants of interest were used as input for a SLiM3 forward genetic simulation. 1000 forward genetic simulations were run for each variant to establish a null distribution of frequencies expected under neutrality.

All but one SNP fell outside the 95% confidence interval of their null distributions, suggesting that the neutral model of genetic drift could not explain their differentiation (figure 8A). The observed frequencies of two class II HLA variants were identified in the extreme tails of their distributions: **rs3128970,** an *HLA-DPB1* 3' UTR variant (T to G), and **rs8084,** an *HLA-DRA* synonymous variant (A to C). The observed frequency of the *KIR3DL2* missense

variant **rs3745902** was identified in the tail of it's expected neutral distribution, as were the upstream transcript variants **rs3136685** (C to T) and **rs72867447** (C to G) for *CCR7* and *BMAL-1,* respectively.

Forward genetic simulations with varying *S* parameters revealed that selection regimes better fit the changes in frequency of these five variants than the genetic drift scenario (figure 8B). The *HLA-DPB1* and *HLA-DRA* variants best fit a positive selection model with a strength of 0.025, rising in frequency from 20% to 49% and 24% to 48%, respectively. The *KIR3DL2* variant, which dropped in frequency from 50% to 21%, best fits a forward model with a purifying selection force of - 0.020, the same selection force best modeling the frequency change of the *BMAL-1* variant. The frequency change of the CCR7 variant best fits a purifying model with a selection force of - 0.06, the highest estimated *S* parameter of the five simulated variants. Variant types and their modeled selection coefficients are listed in Table 3.

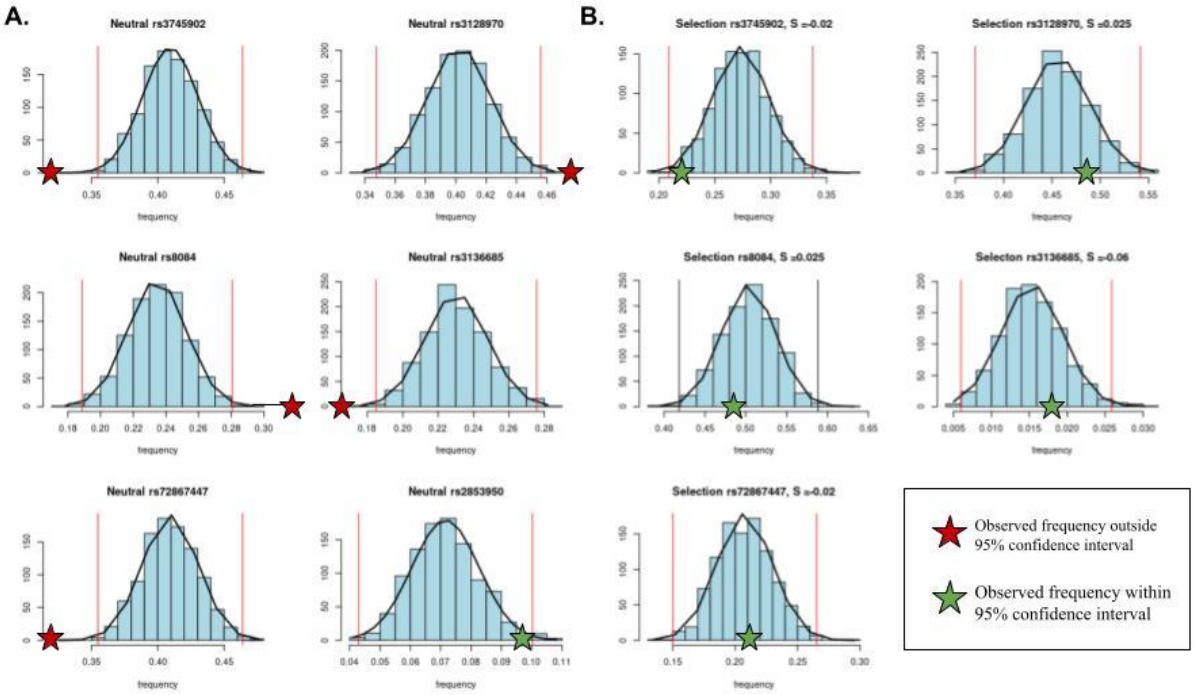

**Figure 8. Expected distributions of BA allele frequencies under a drift and selection scenario.** Vertical lines distinguish the bounds of a 95% confidence interval. Stars show the relative placement of observed BA allele frequencies. A. Five observed allele frequencies fall outside the interval of confidence ( $p < 0.05$), and the null hypothesis of neutral evolution at these loci was rejected. The drift scenario for rs2853950 could not be rejected. B. Selection coefficients are listed by variant names. A selection scenario for each of the five alleles was identified to fit observed BA frequencies.

**Table 3. Potential Ranges of Selection Coefficient in Purifying and Positive Selection Models**

| | rs3745902 | rs3128970 | rs8084 | rs3136685 | rs72867447 |
|---|---|---|---|---|---|
| **Gene** | *KIR3DL2* | *HLA-DPB1* | *HLA-DRA* | *CCR7* | *BMAL1* |
| **Variant*** | *missense* | *3_prime_UTR* | *synonymous* | *genic_upstream_ transcript_variant* | *genic_upstream_ transcript_variant* |
| **Function*** | *Activation of NK and CD8+ T-cells* | *Antigen presentation by APCs* | *Antigen presentation by APCs* | *Naive CD4+ T-cell trafficking* | *CCG transcription factor* |
| **Selection Coefficient Range** | *(-0. 02, -0.04)* | *(0. 02, 0.03)* | *(0. 02, 0.03)* | *(-0. 05, -0.07)* | *(-0. 02, -0.04)* |

***Variant types and functions were annotated based on their *NCBI dbSNP* entries**

# Discussion

## A. Selection on immune system/circadian rhythm variants is a window to the past

### i. Selection on HLA genes fits published models of balancing selection … for different reasons

Multiple allotypes within the two classes of HLA complexes are strategically maintained within the population at intermediate frequencies (Ebert and Fields, 2020). The two allotypes identified in this study, rs8084 and rs3128970, belong to the second class HLA complexes.

Class II MHC complexes are notorious hotspots for extreme polymorphism, due to their role in mediating information transfer (via antigen presentation) to elicit an adaptive immune response(Brandt et al, 2018). Polymorphisms in the beta chain of *HLA-DP,* to which the variant rs3128970 belongs, have been well documented in coding regions that mediate the loading and binding of pathogenic peptides (Petersdorf et al, 2001; Anczurowski and Hirano, 2018). However, their correlation with disease phenotypes is often unexplained by function studies. For instance, while highly polymorphic at position 84, there is no clear evidence of differential binding specificity *in silico* for *HLA-DP*$^{84}$ variants; changes in the amino acid at this position may influence the *conformation* of the pocket into which peptides are loaded, but do not directly influence the ease of peptide presentation specific to a certain pathogen class (Doytchinova and Flower, 2005). Thus, it has been proposed that the extreme polymorphism at the regions may be maintained by linkage to coding regions *actually seen* by balancing selective pressures (Anczurowski and Hirano, 2018).

This same logic may be applicable to the apparent selection scenario modeled for rs 3128970 in this study, as well as the synonymous (no phenotypic change) variant rs8084. By nature of aDNA SNP capture data, some regions of the genome degrade through time to a point of unrecognizability, wherein SNP identification cannot be called from typical alignment protocols. It may be that the evidence of selection on this variant is an artifact of selection upon a linked, advantageous polymorphism in a peptide-binding region.

However, as is the case with circadian control of gene expression, 3' UTR variants may play a role in monitoring and regulating *HLA-DP* expression in an infection scenario to optimize the adaptive response. For instance, variants ~2,000 bp upstream of rs3128970 in the 3' UTR of *HLA-DPB1* have been correlated with varying phenotypes in recovery from Hepatitis B Virus infection, and functional studies have linked this variance to alterations in the gene's expression (Thomas et al, 2012). rs3128970 may play, or have played during the MLN-BA transition, a similar role in varying the robusticity in response to pathogenic infection, and high polymorphism maintained by balancing selection within the population could be explained by the advantage of testing population variants' elicitation of too *great* or too *little* a response during a pandemic event.

**ii. Selection on KIR3DL2 variant may have been better of two evils**

KIR3DL2 is a killer cell immunoglobulin-like receptor (KIR) central to the activation of NK cells - which target cells that have been infected and do *not* express MHC I - and altering the threshold of activation of CD8 + cells - which target infected cells that *do* express MHC I. NK cells receive inhibitory signals upon binding MHC I complexes via extracellular receptors like *KIR3DL2;* this binding is key to the immune system's distinguishing of self and non-self (or self and infected) in situations where viruses may downregulate the cell's expression of MHC I (Shaw et al, 2012).

Binding affinity of KIR3DL2 to the HLA is quite possibly specific to the peptide sequence of the expressed *HLA-A* exon (Hansasuta et al, 2004; Graef, 2009). Thus, observed negative selection on the identified *KIR3DL2* variant in this study, rs3745902, may be evidence of a rise or fall in frequency of a functionally linked *HLA-A* exon variant. Again, while no

evidence of selection was determined for *HLA-A* exon variants, poor coverage as their loci in the Iberian dataset may obscure differentiation at this region.

However, selection at this site may also be evidence of disadvantageous interactions between rs3745902 and its ligands, like those contributing to CD8 + T - cell activation - without simultaneous selection on the ligand. This would be a clear example of niche construction changing the constraints on the transfer of standing variation from one population to the next; from this abstract point of view, a decreased abundance of rs3745902 in subsequent populations carries a large quantum of information. Namely, what *was* the functional result of binding between the KIR3DL2 allele and its ligand has somehow become deleterious, instead of evolving to maintain the status quo (as would be expected if the variant decreased in frequency alongside another functionally linked variant).

Interestingly, the missense variant rs3745902, which results in a base substitution of threonine-to-methionine at position 376, has been linked to a *decreased susceptibility* to Pemphigus foliaceus (Augusto et al, 2015). This autoimmune response is one of the only documented diseases known to be endemic, with a strong correlation to environmental factors like hematophagous insects and infections microorganisms (Aoki et al, 2004). Infectious microorganisms may be linked with pemphigoid autoimmunity through the repeat triggering of adaptive responses and chance establishment of immunological memory against self antigens similar to the pathogens'. Regardless, the selection against the low-susceptibility *KIR3DL2* is curious, but may be explained by its coincidental correlation with significantly lower expression rates (Augusto et al, 2015. Selection would therefore play a balancing game, weighing the disadvantage of low-expression during a time of high pathogen exposure with the disadvantage of potentially developing pemphigoid autoimmunity from that same exposure.

**iii. Selection of CCR7 and BMAL1 may be due to clock-related, functional linkage**

CCR7 is a chemokine receptor implicated in the trafficking of naive CD4+ T cells to the lymph nodes, where they meet APCs and search for co-stimulatory signs of infection. Its variant rs3136685, which experienced the strongest purifying selection force ($S$=0.06), has been associated with predisposition to the development of myeloma and prostate cancers in men while also disproportionately low in women with developed breast cancers (Purdue et al, 2011; Lin et al, 2009; Martino et al, 2014). The location, whether a linked, hidden biomarker or functionally relevant due to its regulation of CCR7, has thus been recommended as a risk-factor marker and, moreover, has been pinpointed as a possible mechanistic component for targeting T cell trafficking early on in cancer development (Marino et al, 2016). However, the extent of cancers in the MLN-BA transect is unknown and likely particularly low; what other hypotheses may explain the strong selection against the rs3136685 variant?

Independent events of selection on *CCR7* and *BMAL1* variants could easily occur; *BMAL1,* of course, controls the expression of tissue-specific CCGs in many tissues of the body, and without evidence of specific pathogens and clock-altering mechanisms present during the time transect, pinning down the story of *CCR7/BMAL1* co-inheritance is a tricky business.

However, it should be noted that rs3136685 occurs at site -1853 relative to the transcription start site of *CCR7;* the upstream variant has, as previously discussed, been implicated in the regulation of *CCR7* expression. Curiously, potential E-box sequences (CANNTG) have been identified at positions -1937 and -1670 of the *CCR7* gene, placing rs3136685 within ~100 and ~200 bp of potential CLOCK-BMAL1 heterodimer binding site (Ou et al, 2008). The negatively selected *BMAL1* variant, rs72867447, occurs at an upstream

regulatory site near the site of repressive activity of clock genes *Reverb-alpha* and *rora* (Mohawk et al, 2012).

It may be the case that variation in the expression of *BMAL1,* due to the upstream regulatory variant, when inherited with the potential Ebox-influencing *CCR7* variant, resulted in a trafficking of lymphocytes inappropriate for the the degree of an emergent infection. Determining whether the occurrence of the rs3136685~rs72867447 haplotype was itself purified during the MLN-BA shift would be an appropriate study to begin to address this hypothesis, particularly when accompanied by functional studies exploring the impact of the mutant alleles on the time-of-day dependent trafficking.

Previous studies have demonstrated *BMAL1* dependency of both dendritic cell trafficking and CD4+ T-cell trafficking via CCR7 (Silver et al, 2012; Druzd et al, 2017), suggesting that information transfer between the innate and adaptive immune systems occurs under the constraint of circadian time. Thus, studying the interaction between these polymorphisms may be a place to start unraveling the genetic elements that predispose individuals to launching inappropriate immune responses to pathogenic load.

**B. Future directions**

Missing data due to aDNA degradation can artificially alter the counts of rare alleles, skewing observed frequency changes and potentially hiding alleles under selection from detection via differentiation statistics. While the dataset was filtered for a genotyping rate of > 50%, poor capture may have filtered potentially selected variants. This study aimed to determine whether evidence of selection on interacting components of the circadian rhythm and immune system could be inferred from open source genomic data. The evidence observed here makes the case for a more targeted, probe-mediated capture technique on related regulatory components, in

order to significantly amplify sequencing reads from these regions of interest and improve the statistical power of calculated differentiation.

A composite likelihood simulation could further be conducted to establish likely demographic history of the BA population and evaluate the proportion of Steppe-related ancestry directly (Lindo et al, 2016). Utilizing these methods, corrections to allele frequencies post-admixture can increase the robusticity of downstream analysis with probe-mediated capture data.

Various pathogen screening techniques, including metagenomic computational pipelines like *MALT* (Spyrou et al, 2019; Herbig et al, 2017), may similarly allow for correlation between observed genetic changes and specific pandemic events, allowing us to narrow the potential constraints placed on the immune system - circadian rhythm crosstalk even further.

Lastly, the selection models employed in this study generalized selection events to overall selection trends, a sort of average of the total forces acting on the variants through the time transect. For one, establishing distinct historical instances of wide-spread infection would allow for the establishment of an even shorter time transect, allowing for higher temporal resolution in selection scenarios and eliminating models requiring more time to establish observed late-term frequencies. Also, a clearer picture of demographic and material changes during the time transect, through both computation studies (like *TreeAnalysis* and ancestry proportion-based demographic models) and sociopolitical reconstructions of the ancient communities in question, can provide additional parameters for complex interactions between competing selective pressures.

This study revealed the likelihood of selective pressures acting on five distinct variants of the immune system and circadian rhythm. While three of the five variants cannot be directly

linked to pathogenic strategies of evading immune response, the study nonetheless identifies them as potential targets for future functional studies.

It should be noted that, while attributing the regulation of *HLA-DP* expression to circadian regulation factors like the CLOCK-BMAL1 heterodimer would be incredibly premature (and likely wrong, as the heterodimer typically acts upstream of the transcription pathway), evidence of selection on a potential regulatory region of *HLA-DP* begs the question of *what factor* does *regulate it's expression?* This is a prime example of how niche construction may inform further genetic analyses and functional studies: what conditions present during this transect may have placed a selective pressure on the expression of the antigen-presenting complex component?

Similarly, the strange observation of selection against a *KIR3DL2* variant known to reduce susceptibility to autoimmunity may be linked to its coincidentally evolved downregulation. The phenotypic consequence of this downregulation - that is, a decreased capacity for information transfer to NK and CD8 + T cells - may have placed an overwhelmingly larger pressure on the population, resulting in a net negative selection scenario. To evaluate this hypothesis, it would be necessary to identify potential instances of pandemic during the time transect via pathogen screening techniques, and to seek evidence of increased instances of the autoimmune pemphigoid disease.

Lastly, while a functional relationship between BMAL1 and CCR7 has been established in coordinating the time-of-day dependent tracking of lymphocytes to the lymph nodes, an ontological study of the BMAL1 and CCR7 proteins may reveal the degree to which BMAL1 directly influences CCR7 activity (Druzd et al, 2017). A haplotype based analysis of rs3136685~rs72867447 co-inheritance may further support the crosstalk theory. In the field of

functional genetics, transgenic models of EBOX$^-$ *CCR7* may help determine whether it is the activity of BMAL1 on an upstream EBox of *CCR7* or instead indirect activity of BMAl-CLOCK through other CCGs on CCR7, that is responsible for these daily oscillations.

## Conclusion

This thesis aimed to engage with niche construction theory to three distinct ends:

1. To demonstrate how the EES can be applied to research projects utilizing new methods made possible by aDNA sequencing;
1) to evaluate the robusticity of the above methodology utilizing a case study of circadian rhythm - immune system regulation during a time of new pathogen exposure; and
2) to identify immune system / circadian rhythm targets for future study to further address the niche construction hypothesis.

The increasing availability of open-source aDNA datasets potentiates the development of various pilot studies such as this, aiming to test hypotheses of selection on mechanisms of particular relevance to altered environments and encourage further, invested investigation. The identification of immune-system and circadian rhythm variants under selection during the MLN-BA shift should encourage more both direct approaches of aDNA recapture in these samples, and coincidental pathogen screening to improve temporal resolution and better associate significantly differentiated variants with pathogen types. As such, it is clear that objectives *1* and *2* have been met.

While the third objective may be further addressed by pursuing the projects outlined above, this pilot study provided the fodder for the following allele-specific questions:

1) Is the purifying selection on master clock gene *BMAL1* and T-cell trafficking *CCR7* functionally linked?

2) Did the population suffer increased events of pemphigoid autoimmunity as the low-expression *KIR3DL2* variant was purified?
3) What factors, if any, interact with rs3128970 to regulate the expression of *HLA-DPB1*?

Furthermore, it is evident that anthropological and archaeological methods are not only necessary, but a great gift to evolutionary biologists engaging with niche construction hypotheses. It is my hope that this thesis can demonstrate how a shift in perspective toward niche construction during hypothesis is particularly congruent with the temporal resolution of aDNA-mediated natural selection modeling, and may help guide functional studies of mechanistic and clinical significance.

## References


1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68.

Abele, S. H., Meadows, K. E., Medeiros, D., & Silver, A. C. (2019). Focus: Clocks and Cycles: Time is on the Immune System's Side, Yes it is. *The Yale Journal of Biology and Medicine*, *92*(2), 225.

Alt, K. W., Zesch, S., Garrido-Pena, R., Knipper, C., Szécsényi-Nagy, A., Roth, C., ... & Rojo-Guerra, M. A. (2016). A community in life and death: the Late Neolithic megalithic tomb at Alto de Reinoso (Burgos, Spain). *PloS one*, *11*(1), e0146176.

Amir, M., Chaudhari, S., Wang, R., Campbell, S., Mosure, S. A., Chopp, L. B., ... & Solt, L. A. (2018). REV-ERBα regulates TH17 cell development and autoimmunity. *Cell reports*, *25*(13), 3733-3749.

Anderson, R. M. & May, R. M. Co-evolution of hosts and parasites. *Parasitology* 85, 411–426 (1982).

Anju, V. T., Siddhardha, B., & Dyavaiah, M. (2020). Animal Models to Understand Host–Pathogen Interactions. In *Model Organisms for Microbial Pathogenesis, Biofilm Formation and Antimicrobial Drug Discovery* (pp. 393-411). Springer, Singapore.

Aoki, V., Millikan, R. C., Rivitti, E. A., Hans-Filho, G., Eaton, D. P., Warren, S. J., ... & Diaz, L. A. (2004, January). Environmental risk factors in endemic pemphigus foliaceus (fogo selvagem). In *Journal of Investigative Dermatology Symposium Proceedings* (Vol. 9, No. 1, pp. 34-40). Elsevier.

Arjona, A., & Sarkar, D. K. (2005). Circadian oscillations of clock genes, cytolytic factors, and cytokines in rat NK cells. *The Journal of Immunology*, *174*(12), 7618-7624.

Armelagos, G. J., Goodman, A. H., & Jacobs, K. H. (1991). The origins of agriculture: Population growth during a period of declining health. *Population and environment*, *13*(1), 9-22.

Augusto, D. G., O'Connor, G. M., Lobo‐Alves, S. C., Bass, S., Martin, M. P., Carrington, M., ... & Petzl Erler, M. L. (2015). Pemphigus is associated with KIR3DL2 expression levels and provides evidence that KIR3DL2 may bind HLA‐A3 and A11 in vivo. *European journal of immunology*, *45*(7), 2052-2060.

Azad, A. K., Sadee, W., & Schlesinger, L. S. (2012). Innate immune gene polymorphisms in tuberculosis. *Infection and immunity*, *80*(10), 3343-3359.

Barik, S. (2019). Molecular interactions between pathogens and the circadian clock. *International Journal of Molecular Sciences*, *20*(23), 5824.

Barreiro, L. B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J. K., ... & Quintana-Murci, L. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS genetics*, *5*(7), e1000562.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature genetics*, *40*(3), 340-345.

Bell-Pedersen, D., Cassone, V. M., Earnest, D. J., Golden, S. S., Hardin, P. E., Thomas, T. L., & Zoran, M. J. (2005). Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Reviews Genetics*, *6*(7), 544-556

Brandt, D. Y., César, J., Goudet, J., & Meyer, D. (2018). The effect of balancing selection on population differentiation: a study with HLA genes. *G3: Genes, Genomes, Genetics*, *8*(8), 2805-2815.

Brennan, M. U. (1991). *Health and disease in the Middle and Upper Paleolithic of southwestern France: A bioarcheological study* (Doctoral dissertation, New York University).

Callaway, E. (2015). Bronze Age skeletons were earliest plague victims. *Nature*, *22*.

Cermakian, N., & Sassone-Corsi, P. (2002). Environmental stimulus perception and control of circadian clocks. *Current opinion in neurobiology*, *12*(4), 359-365.

Cohen, M. L. (2000). Changing patterns of infectious disease. *Nature* 406, 762–767 .

Croze, M., Živković, D., Stephan, W., & Hutter, S. (2016). Balancing selection on immunity genes: review of the current literature and new analysis in Drosophila melanogaster. *Zoology*, *119*(4), 322-329.

Cubas, M., Peyroteo-Stjerna, R., Fontanals-Coll, M., Llorente-Rodríguez, L., Lucquin, A., Craig, O. E., & Colonese, A. C. (2019). Long-term dietary change in Atlantic and Mediterranean Iberia with the introduction of agriculture: a stable isotope perspective. *Archaeological and Anthropological Sciences*, *11*(8), 3825-3836.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Darwin, E. (1809). *Zoonomia; or, the Laws of Organic Life: In Three Parts: Complete in Two Volumes* (Vol. 1).

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J. L., ... & Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *The American Journal of Human Genetics*, *98*(1), 5-21.

Downton, P., Early, J. O., & Gibbs, J. E. (2020). Circadian rhythms in adaptive immunity. *Immunology*, *161*(4), 268-277.

Doytchinova, I. A., & Flower, D. R. (2005). In silico identification of supertypes for class II MHCs. *The Journal of Immunology*, *174*(11), 7085-7095.

Druzd, D., Matveeva, O., Ince, L., Harrison, U., He, W., Schmal, C., ... & Scheiermann, C. (2017). Lymphocyte circadian clocks control lymph node trafficking and adaptive immune responses. *Immunity*, *46*(1), 120-132.

Early, J. O., Menon, D., Wyse, C. A., Cervantes-Silva, M. P., Zaslona, Z., Carroll, R. G., ... & Curtis, A. M. (2018). Circadian clock protein BMAL1 regulates IL-1β in macrophages via NRF2. *Proceedings of the National Academy of Sciences*, *115*(36), E8460-E8468.

Ebert, D., & Fields, P. D. (2020). Host–parasite co-evolution and its genomic signature. *Nature Reviews Genetics*, *21*(12), 754-768.

Edgar, R. S., Stangherlin, A., Nagy, A. D., Nicoll, M. P., Efstathiou, S., O'Neill, J. S., & Reddy, A. B. (2016). Cell autonomous regulation of herpes and influenza virus infection by the circadian clock. *Proceedings of the National Academy of Sciences, 113*(36), 10085-10090.

Ehlers, A., Xie, W., Agapov, E., Brown, S., Steinberg, D., Tidwell, R., ... & Haspel, J. A. (2018). BMAL1 links the circadian clock to viral airway pathology and asthma phenotypes. *Mucosal immunology,* 11(1), 97-111.

Fontanals‐Coll, M., Eulàlia Subirà, M., Díaz‐Zorita Bonilla, M., & Gibaja, J. F. (2017). First insight into the Neolithic subsistence economy in the north‐east Iberian Peninsula: paleodietary reconstruction through stable isotopes. *American Journal of Physical Anthropology*, *162*(1), 36-50.

Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East *American Journal of Human Genetics*, *98*(3), 456-472.

Garland Jr, T., & Adolph, S. C. (1994). Why not to do two-species comparative studies: limitations on inferring adaptation. *Physiological Zoology*, *67*(4), 797-828.

Gotlieb, N., Rosenne, E., Matzner, P., Shaashua, L., Sorski, L., & Ben-Eliyahu, S. (2015). The misleading nature of in vitro and ex vivo findings in studying the impact of stress hormones on NK cell cytotoxicity. *Brain, behavior, and immunity*, *45*, 277-286.

Graef, T., Moesta, A. K., Norman, P. J., Abi-Rached, L., Vago, L., Older Aguilar, A. M., ... & Parham, P. (2009). KIR2DS4 is a product of gene conversion with KIR3DL2 that introduced specificity for HLA-A*11 while diminishing avidity for HLA-C. *Journal of Experimental Medicine*, *206*(11), 2557-2572.

Haller, B. C., & Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular biology and evolution*, *36*(3), 632-637.

Hansasuta, P., Dong, T., Thananchai, H., Weekes, M., Willberg, C., Aldemir, H., ... & Braud, V. M. (2004). Recognition of HLA‐A3 and HLA‐A11 by KIR3DL2 is peptide‐specific. *European journal of immunology*, *34*(6), 1673-1679.

Hedrick, P. W., & Thomson, G. (1983). Evidence for balancing selection at HLA. *Genetics*, *104*(3), 449 456.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2017). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman.

Hladik, F., Liu, H., Speelmon, E., Livingston-Rosanoff, D., Wilson, S., Sakchalathorn, P., ... & McElrath, M. (2005). Combined effect of CCR5-Δ32 heterozygosity and the CCR5 promoter polymorphism− 2459 A/G on CCR5 expression and resistance to human immunodeficiency virus type 1 transmission. *Journal of virology*, *79*(18), 11677-11684.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, *10*(9), 639-650.

Holtkamp, S. J., Ince, L. M., Barnoud, C., Schmitt, M. T., Sinturel, F., Pilorz, V., ... & Scheiermann, C. (2021). Circadian clocks guide dendritic cells into skin lymphatics. Nature immunology, 22(11), 1375-1381.

Hriscu, M. L. (2005). Modulatory factors of circadian phagocytic activity. *Annals of the new York Academy of Sciences*, *1057*(1), 403-430

Husse, J., Eichele, G., & Oster, H. (2015). Synchronization of the mammalian circadian timing system: light can control peripheral clocks independently of the SCN clock: alternate routes of entrainment optimize the alignment of the body's circadian clock network with external time. *BioEssays*, *37*(10), 1119 1128.

Husse, J., Zhou, X., Shostak, A., Oster, H., & Eichele, G. (2011). Synaptotagmin10-Cre, a driver to disrupt clock genes in the SCN. *Journal of biological rhythms*, *26*(5), 379-389.

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, *15*(6), 379-393.

Keller, M., Mazuch, J., Abraham, U., Eom, G. D., Herzog, E. D., Volk, H. D., ... & Maier, B. (2009). A circadian clock in macrophages controls inflammatory immune responses. *Proceedings of the National Academy of Sciences*, *106*(50), 21407-21412.

Kitchen, G. B., Cunningham, P. S., Poolman, T. M., Iqbal, M., Maidstone, R., Baxter, M., ... & Ray, D. W. (2020). The clock gene Bmal1 inhibits macrophage motility, phagocytosis, and impairs defense against pneumonia. *Proceedings of the National Academy of Sciences*, *117*(3), 1543-1551.

Knipper, C., Rihuete-Herrada, C., Voltas, J., Held, P., Lull, V., Micó, R., ... & Alt, K. W. (2020).

Reconstructing Bronze Age diets and farming strategies at the early Bronze Age sites of La Bastida and Gatas (southeast Iberia) using stable isotope analysis. *PLoS One*, *15*(3), e0229398.

Krakauer, D. C., & Nowak, M. A. (1999, October). Evolutionary preservation of redundant duplicated genes. In *Seminars in cell & developmental biology* (Vol. 10, No. 5, pp. 555-559). Academic Press.

Ku, K., Park, I., Kim, D., Kim, J., Jang, S., Choi, M., ... & Kim, K. (2020). Gut microbial metabolites induce changes in circadian oscillation of clock gene expression in the mouse embryonic fibroblasts. *Molecules and Cells*, *43*(3), 276.

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., ... & Odling-Smee, J. (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the royal society B: biological sciences*, *282*(1813), 20151019.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175-195.

Lin, G. T., Tseng, H. F., Yang, C. H., Hou, M. F., Chuang, L. Y., Tai, H. T., ... & Chang, H. W. (2009). Combinational polymorphisms of seven CXCL12-related genes are protective against breast cancer in Taiwan. *OMICS A Journal of Integrative Biology*, *13*(2), 165-172.

López-Costas, Olalla, and Michelle Alexander. "Paleodiet in the Iberian Peninsula: exploring the connections between diet, culture, disease and environment using isotopic and osteoarchaeological evidence." *Archaeological and Anthropological Sciences* 11.8 (2019): 3653-3664.

Lowrey, P. L., & Takahashi, J. S. (2011). Genetics of circadian rhythms in Mammalian model organisms. *Advances in genetics*, *74*, 175-230.

Majumdar, T., Dhar, J., Patel, S., Kondratov, R., & Barik, S. (2017). Circadian transcription factor BMAL1 regulates innate immunity against select RNA viruses. *Innate immunity, 23*(2), 147-154.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., ... & Reich, D. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201-206.

Martino, A., Campa, D., Jurczyszyn, A., Martínez-López, J., Moreno, M. J., Varkonyi, J., ... & Canzian, F. (2014). Genetic variants and multiple myeloma risk: IMMEnSE validation of the best reported associations—an extensive replication of the associations from the candidate gene era. *Cancer Epidemiology and Prevention Biomarkers*, *23*(4), 670-674.

Matenchuk, B. A., Mandhane, P. J., & Kozyrskyj, A. L. (2020). Sleep, circadian rhythm, and gut microbiota. *Sleep medicine reviews*, *53*, 101340.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., ... & Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499-503.

Mills, M., Yang, N., Weinberger, R., Vander Woude, D. L., Beggs, A. H., Easteal, S., & North, K. (2001). Differential expression of the actin-binding proteins, α-actinin-2 and-3, in different species: implications for the evolution of functional redundancy. *Human molecular genetics*, *10*(13), 1335-1346.

Mohawk, J. A., Green, C. B., & Takahashi, J. S. (2012). Central and peripheral circadian clocks in mammals. *Annual review of neuroscience*, *35*, 445-462.

Mukherjee, S., Huda, S., & Sinha Babu, S. P. (2019). Toll‐like receptor polymorphism in host immune response to infectious diseases: A review. *Scandinavian journal of immunology*, *90*(1), e12771.

Naicker, D. D., Werner, L., Kormuth, E., Passmore, J. A., Mlisana, K., Karim, S. A., ... & CAPRISA

Acute Infection Study Team. (2009). Interleukin-10 promoter polymorphisms influence HIV-1 susceptibility and primary HIV-1 pathogenesis. *Journal of Infectious Diseases*, *200*(3), 448-452.

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2013). *Niche construction*. Princeton university press.

Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., Silva, M., ... & Reich, D. (2019). The genomic history of the Iberian Peninsula over the past 8000 years. *Science*, *363*(6432), 1230-1234.

Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 185-205.

Ou, D. L., Chien, H. F., Chen, C. L., Lin, T. C., & Lin, L. I. (2008). Role of Twist in head and neck carcinoma with lymph node metastasis. *Anticancer research*, *28*(2B), 1355-1359.

Petersdorf, E. W., Gooley, T., Malkki, M., Anasetti, C., Martin, P., Woolfrey, A., ... & Hansen, J. A. (2001). The biological significance of HLA‐DP gene variation in haematopoietic cell transplantation. *British journal of haematology*, *112*(4), 988-994.

Pigliucci, M., & Müller, G. B. (2010). Elements of an extended evolutionary synthesis. *Evolution: The extended synthesis*, 3-17.

Plantinga, T. S., Alonso, S., Izagirre, N., Hervella, M., Fregel, R., Van Der Meer, J. W., ... & De La Rúa, C. (2012). Low prevalence of lactase persistence in Neolithic South-West Europe. *European journal of human genetics*, *20*(7), 778-782.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics, 81(3), 559–575.

Purdue, M. P., Lan, Q., Menashe, I., Zheng, T., Zhang, Y., Yeager, M., ... & Baris, D. (2011). Variation in innate immunity genes and risk of multiple myeloma. *Hematological oncology*, *29*(1), 42-46.

Quintana-Murci, L. (2019). Human immunology through the lens of evolutionary genetics. *Cell*, *177*(1), 184-199.

Reis, E. S., Lange, T., Köhl, G., Herrmann, A., Tschulakow, A. V., Naujoks, J., ... & Köhl, J. (2011). Sleep and circadian rhythm regulate circulating complement factors and immunoregulatory properties of C5a. *Brain, behavior, and immunity*, *25*(7), 1416-1426.

Schuhmacher, T. X., Cardoso, J. L., & Banerjee, A. (2009). Sourcing african ivory in Chalcolithic Portugal. *Antiquity*, *83*(322), 983-997.

Serrano, J. G., Ordóñez, A. C., & Fregel, R. (2021). Paleogenomics of the prehistory of Europe: human migrations, domestication and disease. *Annals of Human Biology*, *48*(3), 179-190.

Shaw, J., & Kollnberger, S. (2012). New perspectives on the ligands and function of the killer cell immunoglobulin-like receptor KIR3DL2 in health and disease. *Frontiers in immunology*, *3*, 339.

Silver, A. C. (2017). Pathogen-associated molecular patterns alter molecular clock gene expression in mouse splenocytes. *PLoS One*, *12*(12), e0189949.

Silver, A. C., Arjona, A., Hughes, M. E., Nitabach, M. N., & Fikrig, E. (2012). Circadian expression of clock genes in mouse macrophages, dendritic cells, and B cells. *Brain, behavior, and immunity*, *26*(3), 407 413.

Slatkin, M., & Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, *113*(23), 6380-6387.

Smith, J. M. (2000). The concept of information in biology. *Philosophy of science*, *67*(2), 177-194.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, *23*(1), 23-35.

Spyrou, M. A., Bos, K. I., Herbig, A., & Krause, J. (2019). Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature Reviews Genetics*, *20*(6), 323-340.

Szécsényi-Nagy, A., Roth, C., Brandt, G., Rihuete-Herrada, C., Tejedor-Rodríguez, C., Held, P., ... & Alt, K. W. (2017). The maternal genetic make-up of the Iberian Peninsula between the Neolithic and the Early Bronze Age. *Scientific Reports*, *7*(1), 1-15.

Tataru, Paula, et al. "Statistical inference in the Wright–Fisher model using allele frequency data." *Systematic biology* 66.1 (2017): e30-e46.

Anczurowski, M., & Hirano, N. (2018). Mechanisms of HLA-DP antigen processing and presentation revisited. *Trends in immunology*, *39*(12), 960-964.

Thomas, R., Thio, C. L., Apps, R., Qi, Y., Gao, X., Marti, D., ... & Carrington, M. (2012). A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection. *Journal of virology*, *86*(12), 6979-6985.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., ... & Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*, *39*(1), 31-40.

Valdiosera, C., Günther, T., Vera-Rodríguez, J. C., Ureña, I., Iriarte, E., Rodríguez-Varela, R., ... & Jakobsson, M. (2018). Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proceedings of the National Academy of Sciences*, *115*(13), 3428-3433.

Villalba-Mouco, V., Oliart, C., Rihuete-Herrada, C., Childebayeva, A., Rohrlach, A. B., Fregeiro, M. I., ... & Haak, W. (2021). Genomic transformation and social organization during the Copper Age–Bronze Age transition in southern Iberia. *Science advances*, *7*(47), eabi7038.

Wall, J. D., & Brandt, D. Y. C. (2016). Archaic admixture in human history. *Current opinion in genetics & development*, *41*, 93-97.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *evolution*, 1358-1370.

Williams R., Pourreza H., Wang Y., Carbonetto P., & Novembre J. PCAviz: visualizing principal components analysis. (2017). http://github.com/NovembreLab/PCAviz

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114. Formula

Wright, S. (1949). The genetical structure of populations. *Annals of eugenics*, *15*(1), 323-354. Definition

Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., ... & Jin, L. (2011). A genome-wide search for signals of high-altitude adaptation in Tibetans. *Molecular biology and evolution*, *28*(2), 1003-1011.

Li, Z., Bonaldi, K., Uribe, F., & Pruneda-Paz, J. L. (2018). A localized *Pseudomonas syringae* infection triggers systemic clock responses in Arabidopsis. *Current Biology*, 28(4), 630-639.