Disentangling the roles of adaptive and neutral evolution in shaping HLA immune gene diversity in humans

Dissertation
in fulfilment of the requirements for the degree of
"Doctor rerum naturalium"
of the Faculty of Mathematics and Natural Sciences
at Kiel University

submitted by

Onur Özer

Kiel, September 2023

First referee: Prof. Dr. Tobias Lenz

Second referee: Prof. Dr. Hinrich Schulenburg

Date of oral examination: 29.11.2023

Table of Contents

Summary

Infectious diseases have been among the top drivers of morbidity and mortality within human populations. Consequently, many human genes that orchestrate immune responses against pathogens exhibit signatures of natural selection. The outcome of this pathogen-mediated selection depends on the function of the gene's product. Some genes, such as those targeting highly conserved structures of pathogens, exhibit very low levels of variation as they evolve under strong purifying selection. On the other hand, balancing selection might result in increased variation and maintenance of diversity. In jawed vertebrates, the extreme polymorphism of the major histocompatibility complex (MHC) genes is hypothesized to have evolved through this evolutionary process. Products of MHC genes, which are also known as human leukocyte antigen (HLA) in humans, bind and present short peptides of pathogens to the immune cells and initiate adaptive immune responses. Therefore, diverse pathogens presumably select for increased MHC diversity. Specific mechanisms of this selection process have been formulated, yet their relative roles in the maintenance of diversity continue to be debated extensively.

The aim of this thesis is to investigate the role of pathogen-mediated selection as well as other evolutionary forces in shaping the HLA diversity in humans. To this end, we employed several computational approaches, combined with the population genetics analysis of ancient and modern human populations. Our first step was to focus on the commonly held assumption that high pathogen diversity selects for different HLA alleles. We analyzed the peptide diversity of 36 human pathogens and showed that peptides that are shared between pathogens constitute only a minority of the total peptidome pool. In other words, the peptide pool of each pathogen is mostly unique to itself, exhibiting an immense diversity that our immune system faces. In order to explore the effect if this diversity on HLA genes, we used computational peptide-binding prediction algorithms and identified peptide sets that are bound by 321 common HLA class-I alleles. Our results revealed that some HLA alleles bind a larger fraction of peptides from a few pathogens, suggesting a specialization towards these particular pathogens. Furthermore, this specialization was negatively correlated with the size of the peptide pool that the HLA allele binds. These results support a scenario that HLA alleles with small peptide repertoires can be maintained in populations by virtue of specialization against a few deadly pathogens that emerge throughout human evolutionary history. On the other hand, generalist

HLA alleles with large peptide repertoires continue protecting individuals against common pathogens.

We next took a population genetics approach and analyzed two novel HLA datasets. The first dataset consists of 12 populations from sub-Saharan Africa. African populations are highly underrepresented in genomics research, hampering both the successful medical applications in many countries and the detailed characterization of humans' evolutionary history. We used both genome-wide SNP data and targeted HLA sequencing data to investigate HLA diversity patterns not only within African populations but also in a wider context by including population samples from Europe and East Asia. Our results highlight the excessive diversity across HLA genes within African populations. We observed that very old mutations constitute a significant part of the diversity within MHC, suggesting that balancing selection favors ancient variation. In line with this observation, non-African populations appear to have maintained similarly high levels of HLA diversity despite the bottlenecks associated with the colonization of Eurasia. Furthermore, contrary to what would be expected for a genomic region evolving under divergent local selective pressures, population differentiation was not higher within MHC than it is in neutral regions. In fact, population differentiation at HLA genes largely reflects differentiation patterns of neutral markers indicating population history as the main determinant of differentiation. Targeted HLA sequencing analysis also supports the dominant role of balancing selection and population history in HLA evolution. The functional diversity within HLA genes (i.e. functionally distinct HLA molecules) is maintained across different populations while the differentiation patterns based on HLA alleles follow the continental separations. Overall, these results suggest that demography and balancing selection on ancient variation play a major role in recent MHC evolution in humans while the role of adaptation to local pathogens appears much smaller. This conclusion is further supported by our analysis of the second novel HLA dataset. The dataset includes 129 ancient individuals from central Europe that lived between 5000 BCE to 2800 BCE. This period, namely the Neolithic, is associated with drastic social and cultural changes. Domestication of animals and plants brought about a sedentary lifestyle, shifts in diet towards increased consumption of carbohydrate and milk products and, inadvertently, many novel pathogens. Consequently, it is commonly hypothesized that as a result of these changes, many genes, specifically the immune genes would exhibit signatures of selection. Following that hypothesis, we used a novel pipeline to generate HLA genotypes of ancient individuals and analyzed the HLA diversity of early and late Neolithic farmers from Europe together with several modern European, East Asian and

African populations. We found that although early European farmers migrated recently from Anatolia, they exhibit high HLA diversity, potentially maintained by balancing selection. Interestingly, major shifts in frequency of some HLA alleles were observed both between early and late farmers and between late farmers and modern Europeans. In fact, the differentiation between ancient and modern Europeans is almost comparable to the differentiation between modern Europeans and Asians. Although this observation is compatible with adaptive evolution in response to changes in pathogen landscape, our analysis based on peptide-binding predictions reveals that despite differences in allele frequencies, HLA allele pools of ancient and modern populations are functionally similar. Indeed, focusing on the *measles virus* (MeV), which emerged as a human pathogen after the Neolithic, we found that modern and ancient populations bind similar numbers of MeV peptides. These results indicate that the major driver of the changes in allele frequencies was not the adaptation to novel or local pathogens. We suggest that the major differences between early and late Neolithic farmers were brought about by the admixture with local hunter-gatherers while the differences between late farmers and modern Europeans are the result of admixture with steppe pastoralists. In line with this hypothesis, most common HLA alleles in modern central Europeans follow a north to south cline similar to the steppe ancestry. Therefore, it is likely that these alleles were introduced by steppe pastoralists. We conclude that balancing selection and population history appears as the main drivers of recent HLA evolution while the effect of directional selection by pathogens is minimal.

Zusammenfassung

This German summary is translated by DeepL from English and kindly edited by Dr. Britta Meyer.

Infektionskrankheiten gehören zu den Hauptursachen für Morbidität und Mortalität in menschlichen Populationen. Folglich weisen viele menschliche Gene, die Immunreaktionen gegen Krankheitserreger steuern, Anzeichen natürlicher Selektion auf. Das Ergebnis dieser pathogenvermittelten Selektion hängt von der Funktion des Genprodukts ab. Einige Gene, z. B. solche, die auf hochkonservierte Strukturen von Krankheitserregern abzielen, weisen nur eine sehr geringe Variation auf, während sie unter starker negativer/reinigender Selektion evolvieren. Andererseits kann eine ausgleichende Selektion zu einer erhöhten Variation und zur Erhaltung der Vielfalt führen. Bei Wirbeltieren mit Kiefer wird angenommen, dass der extreme Polymorphismus der Gene des Haupthistokompatibilitätskomplexes (engl. Major histocomptaibility complex, MHC) durch diesen evolutionären Prozess entstanden ist. Die Produkte der MHC-Gene, die beim Menschen auch als Humane Leukozyten-Antigene (HLA) bezeichnet werden, binden kurze Peptide von Krankheitserregern, präsentieren sie den Immunzellen und lösen eine adaptive Immunantwort aus. Daher selektieren unterschiedliche Krankheitserreger vermutlich auf eine erhöhte MHC-Diversität. Spezifische Mechanismen dieses Selektionsprozesses sind bekannt, doch ihre relative Rolle bei der Aufrechterhaltung der Diversität wird nach wie vor heftig debattiert.

Ziel dieser Arbeit ist es, die Rolle der pathogenvermittelten Selektion sowie anderer evolutionärer Kräfte bei der Entstehung der HLA-Vielfalt beim Menschen zu untersuchen. Zu diesem Zweck haben wir mehrere rechnergestützte Ansätze mit der populationsgenetischen Analyse alter und moderner menschlicher Populationen kombiniert. In einem ersten Schritt untersuchten wir die allgemein verbreitete Annahme, dass eine hohe Erregervielfalt eine Selektion für diverse HLA-Allele bewirkt. Wir analysierten die Peptidvielfalt von 36 humanen Krankheitserregern und zeigten, dass Peptide, die zwischen den Erregern geteilt werden, nur eine Minderheit des gesamten Peptidpools ausmachen. Mit anderen Worten: Der Peptidpool jedes Erregers ist größtenteils einzigartig für ihn selbst und zeigt eine immense Vielfalt, mit der unser Immunsystem konfrontiert ist. Um die Auswirkungen dieser Vielfalt auf die HLA-Gene zu erforschen, haben wir computergestützte Algorithmen zur Peptidbindungsvorhersage eingesetzt und Peptidmengen identifiziert, die von 321 häufigen HLA-Klasse-I-Allelen gebunden werden. Unsere Ergebnisse zeigten, dass einige HLA-Allele einen größeren Anteil von Peptiden einiger weniger Krankheitserreger binden, was auf eine Spezialisierung auf diese

bestimmten Krankheitserreger hindeutet. Darüber hinaus war diese Spezialisierung negativ mit der Größe des Peptidpools korreliert, den das HLA-Allel bindet. Diese Ergebnisse unterstützen das Szenario, dass HLA-Allele mit kleinen Peptidrepertoires in Populationen durch Spezialisierung auf einige wenige tödliche Krankheitserreger, die im Laufe der menschlichen Evolutionsgeschichte aufgetaucht sind, erhalten bleiben können. Andererseits schützen generalistische HLA-Allele mit großen Peptidrepertoires die Individuen weiterhin vor häufigen Krankheitserregern.

Als nächstes wählten wir einen populationsgenetischen Ansatz und analysierten zwei neue HLA-Datensätze. Der erste Datensatz besteht aus 12 Populationen aus Afrika südlich der Sahara. Afrikanische Populationen sind in der Genomforschung stark unterrepräsentiert, was sowohl die erfolgreiche medizinische Anwendung in vielen Ländern als auch die detaillierte Charakterisierung der Evolutionsgeschichte des Menschen behindert. Wir haben sowohl genomweite SNP-Daten als auch gezielte HLA-Sequenzierungsdaten verwendet, um HLA-Diversitätsmuster nicht nur innerhalb afrikanischer Populationen zu untersuchen, sondern auch in einem breiteren Kontext, indem wir Bevölkerungsproben aus Europa und Ostasien einbezogen haben. Unsere Ergebnisse unterstreichen die übermäßige Vielfalt der HLA-Gene in afrikanischen Populationen. Wir haben festgestellt, dass sehr alte Mutationen einen bedeutenden Teil der Vielfalt innerhalb des MHC ausmachen, was darauf hindeutet, dass eine ausgleichende Selektion alte Variationen begünstigt. In Übereinstimmung mit dieser Beobachtung scheinen nicht-afrikanische Populationen trotz der mit der Besiedlung Eurasiens verbundenen Engpässe eine ähnlich hohe HLA-Diversität beibehalten zu haben. Darüber hinaus war die Populationsdifferenzierung innerhalb von MHC nicht höher als in neutralen Regionen, was bei einer genomischen Region, die sich unter unterschiedlichem lokalem Selektionsdruck entwickelt, zu erwarten wäre. Tatsächlich spiegelt die Populationsdifferenzierung bei den HLA-Genen weitgehend die Differenzierungsmuster der neutralen Marker wider, was darauf hindeutet, dass die Populationsgeschichte die Hauptdeterminante der Differenzierung ist. Die Analyse der gezielten HLA-Sequenzierung untermauert ebenfalls die dominante Rolle der ausgleichenden Selektion und der Populationsgeschichte bei der HLA-Evolution. Die funktionelle Vielfalt innerhalb der HLA-Gene (d. h. funktionell unterschiedliche HLA-Moleküle) bleibt in verschiedenen Populationen erhalten, während die Differenzierungsmuster auf der Grundlage der HLA-Allele den kontinentalen Trennungen folgen. Insgesamt deuten diese Ergebnisse darauf hin, dass die Demografie und die ausgleichende Selektion auf alte Variationen eine wichtige Rolle bei der jüngsten MHC-Evolution beim Menschen spielen,

während die Anpassung an lokale Krankheitserreger eine wesentlich geringere Rolle zu spielen scheint. Diese Schlussfolgerung wird auch durch unsere Analyse des zweiten neuen HLA-Datensatzes bestätigt. Der Datensatz umfasst 129 frühzeitliche? Menschen aus Mitteleuropa, die zwischen 5000 v. Chr. und 2800 v. Chr. lebten. Dieser Zeitraum, das Neolithikum, ist mit drastischen sozialen und kulturellen Veränderungen verbunden. Die Domestizierung von Tieren und Pflanzen führte zu einer sesshaften Lebensweise, zu einer Umstellung der Ernährung auf den vermehrten Verzehr von Kohlenhydraten und Milchprodukten und - unbeabsichtigt - zu zahlreichen neuen Krankheitserregern. Folglich wird allgemein angenommen, dass viele Gene, insbesondere die Immungene, als Folge dieser Veränderungen Selektionsmerkmale aufweisen würden. Dieser Hypothese folgend haben wir eine neuartige bioinformatische? Pipeline zur Erstellung von HLA-Genotypen alter Individuen verwendet und die HLA-Diversität früh- und spätneolithischer Bauern aus Europa zusammen mit verschiedenen modernen europäischen, ostasiatischen und afrikanischen Populationen analysiert. Wir fanden heraus, dass die frühen europäischen Bauern, obwohl sie erst kurz vorher vor kurzem aus Anatolien eingewandert sind, eine hohe HLA-Diversität aufweisen, die möglicherweise durch ausgleichende Selektion erhalten wurde. Interessanterweise werden größere Verschiebungen in der Häufigkeit einiger HLA-Allele sowohl zwischen frühen und späten Bauern als auch zwischen späten Bauern und modernen Europäern beobachtet. Tatsächlich ist die Differenzierung zwischen alten und modernen Europäern fast vergleichbar mit der Differenzierung zwischen modernen Europäern und Asiaten. Obwohl diese Beobachtung mit einer adaptiven Evolution als Reaktion auf Veränderungen in der Erregerlandschaft vereinbar ist, zeigt unsere Analyse auf der Grundlage von Peptidbindungsvorhersagen, dass die HLA-Allelpools alter und moderner Populationen trotz unterschiedlicher Allelhäufigkeiten funktionell ähnlich sind. Am Beispiel des Masernvirus (MeV), das erst nach dem Neolithikum als menschlicher Krankheitserreger auftrat, haben wir festgestellt, dass moderne und alte Populationen ähnlich viele MeV-Peptide binden. Diese Ergebnisse deuten darauf hin, dass der Hauptgrund für die Veränderungen der Allelhäufigkeiten nicht in der Anpassung an neue oder lokale Krankheitserreger liegt. Wir vermuten, dass die Hauptunterschiede zwischen frühen und späten neolithischen Bauern durch die Vermischung mit lokalen Jägern und Sammlern entstanden sind, während die Unterschiede zwischen späten Bauern und modernen Europäern das Ergebnis einer Vermischung mit Steppenhirten sind. Im Einklang mit dieser Hypothese folgen die meisten häufigen HLA-Allele in modernen Mitteleuropäern einer Nord-Süd-Kurve, die der Steppenvorfahrenschaft ähnelt. Daher ist es wahrscheinlich, dass diese Allele von Steppenhirten eingeführt wurden. Wir kommen zu dem

Schluss, dass eine ausgleichende Selektion und die Populationsgeschichte die Haupttriebkräfte der jüngsten HLA-Evolution zu sein scheinen, während der Einfluss einer gerichteten Selektion durch Krankheitserreger minimal ist.

Introduction

Every organism on our planet, be it an archaea inhabiting extremely hot and acidic springs, a deep-sea fish living under crushing pressure and complete darkness or a temperate mammal roaming wide green meadows, lives under the constant threat of pathogens (Ghielmetti et al., 2021; Ortmann et al., 2006; Quattrini & Demopoulos, 2016). From an evolutionary perspective, the relationship between pathogens and their hosts is highly dynamic. The survival and reproduction of a pathogen depend on its ability to infect a host and reproduce while the survival and reproduction of a host depends on either preventing the infection or minimizing the cost associated with it. This relationship results in an "arms-race" in which adaptations on one side trigger selection for counteracting adaptations on the other side and so on (Ebert & Hamilton, 1996). A major requirement of the process of natural selection (hence adaptation) is the existence of individual heritable differences. Charles Darwin argued in his seminal book "On the Origin of Species" that such heritable variation, combined with the fact that organisms almost always produce more offspring than their environment can support, would lead to competition between individuals over resources, space, mates, etc. The result of this competition is –in Darwin's own words– the "preservation of favoured races". Here, "favoured races" are individuals carrying heritable variation that is better suited to the environment and "preservation" refers to the differential survival and reproduction of these individuals.

Most of the abiotic factors within an organism's habitat such as temperature, oxygen pressure, humidity, etc. are stable or fluctuate predictably relative to the organism's generation time. The relative stability and predictability of abiotic factors allow organisms to accumulate random mutations over generations, some of which might be selectively advantageous and rise in frequency (Bell & Collins, 2008). Therefore, adaptation to abiotic factors usually leads to the fixation of the optimal variant within a population. On the other hand, biotic interactions, especially host-pathogen interactions, are much more dynamic (Brockhurst et al., 2014). Pathogens usually replicate much faster than their hosts do. This is especially evident considering the multicellular host organisms whose generation times are at the scale of years while some viruses or bacteria can replicate in a manner of minutes (Bremer, 1982; Fenner, 2005). A pathogen might accumulate mutations at a much faster rate and circumvent the defense mechanisms of a host before any counteracting mutations arise and are passed on to the next generation of the host (Sanjuán et al., 2010). Such imbalance in the evolutionary rates between hosts and pathogens results in the evolution of diverse host immune mechanisms (Broecker &

Moelling, 2019; Cullen, 2002; Jones & Dangl, 2006). Despite the variation in specific mechanisms, immune responses can be divided into two broad categories; innate and adaptive immune response (Murphy & Weaver, 2017). Innate immune responses are mediated by molecules recognizing structures of pathogens that are highly conserved. Examples include restriction enzymes of bacteria cutting the unmethylated DNA of invading bacteriophages or toll-like receptors (TLRs) of mammals recognizing dsRNA of viruses or lipopolysaccharide of bacteria (Koonin et al., 2017; Murphy & Weaver, 2017). These structures are functionally vital for pathogens to the extent that any change would entail a fitness cost. Innate immune responses usually form the first line of defense against invading pathogens, yet they are not specific and sometimes not enough to contain the infection. Furthermore, innate immune receptors are germline-encoded; hence suffer from the mentioned problem of lack of evolutionary time to adapt possible immune evasion strategies of pathogens. Adaptive immune responses (also known as acquired immunity) overcome that problem by shifting the generation of diversity from inter-generational times to within generation (i.e. individual's lifetime). In the case of bacteria and archaea, CRISPR-Cas systems cut a piece of nucleic acid from the invading bacteriophages and incorporate it into their own genome. This specific sequence then acts as a "guide" to recognize and inactivate the phage (Koonin & Makarova, 2019). In vertebrates, the adaptive immune system is composed of two main arms, the cell-based immune responses (T-cells) and humoral immune responses (B-cells) (Murphy & Weaver, 2017). Receptors of both arms are produced by somatic recombination of *variable* (V), *diversity* (D) and *joining* (J) genes in the genome. The germline contains multiple segments of V, D and J genes and as a result of somatic recombination, one randomly selected segment from each gene is retained in the developing lymphocyte (i.e. T- and B-cells). B-cells may further undergo a process of somatic hypermutation; introduction of point mutations into the B-cell receptor. The result of these intricate mechanisms is the generation of a large pool of lymphocytes, each carrying a unique receptor that can recognize a wide range of antigens (short peptides) originating from invading pathogens (Cooper & Alder, 2006).

Effector mechanisms of lymphocytes are highly potent such as the direct killing of infected cells by $CD8^+$ T cells or antigen-mediated neutralization of pathogens by B-cells. Therefore, it is not surprising that various mechanisms have evolved to control the activation of lymphocytes to prevent self-damage or unnecessary response to harmless antigens (O'Garra & Vieira, 2004). One of the basic limiting factors is that antigens can be recognized by T-cell receptors (TCRs) only if they are presented by a classical major histocompatibility complex (MHC) molecule in

the form of a peptide-MHC (pMHC) complex. This is achieved by a process called positive selection during the maturation of T-cells in the thymus (L. Klein et al., 2014). T-cells that survive the thymic selection are the ones that can weakly bind self pMHC complexes, ensuring that any response given by T-cells is restricted by MHC molecules (Garcia, 2012). However, strong binding to a self pMHC complex by a T-cell could result in an immune response against self tissues (i.e. autoimmunity). An accompanying process, namely the negative selection in the thymus, ensures that T-cells that strongly bind self pMHC complexes are also removed from the pool (L. Klein et al., 2014).

MHC molecules were initially discovered due to their role in determining the success of tissue transplantation and later studies revealed their role in the development of immune responses against pathogens (Gorer, 1936; McDevitt, 2000; Thorsby, 2009; Trowsdale & Knight, 2013). There are two classes of classical MHC genes, namely MHC class-I and MHC class-II (**Figure 1**) (Murphy & Weaver, 2017). Class-I proteins are expressed in all nucleated cells and bind

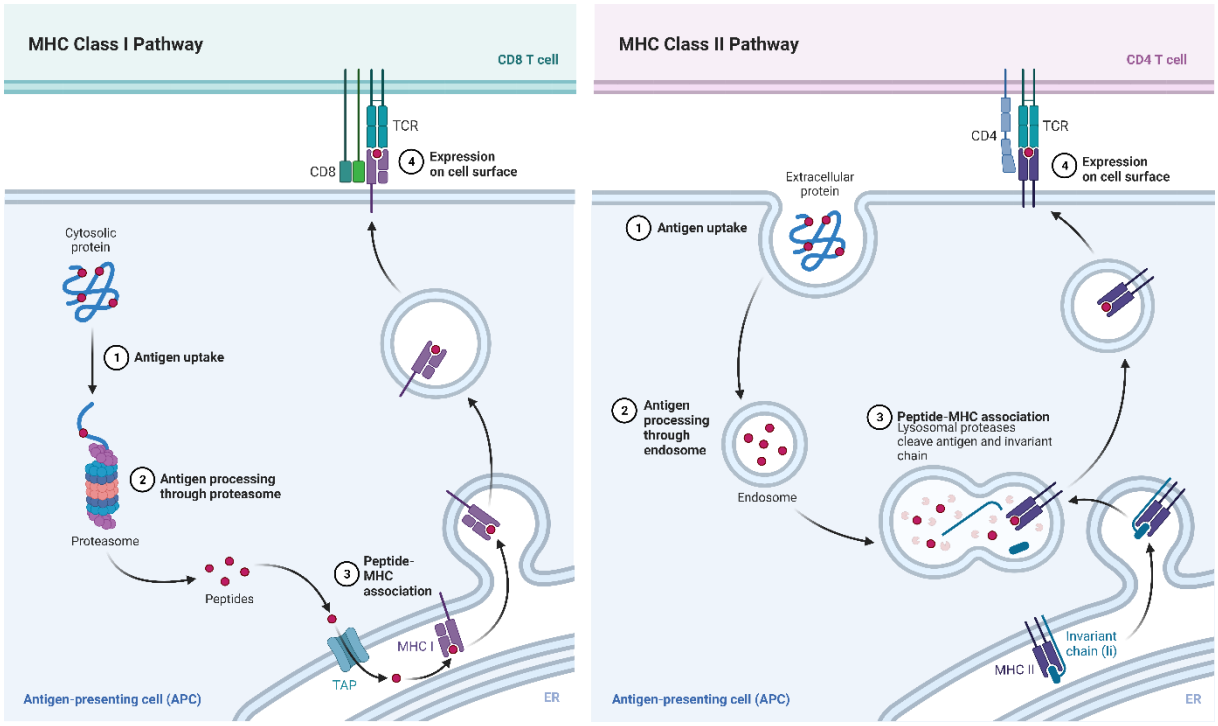


Figure 1. MHC class-I and MHC class-II antigen presentation pathways. Figure template is generated by Akiko Iwasaki and retrieved from https://app.biorender.com/biorender-templates

peptides originating from the intracellular environment. If the pMHC complex is recognized by $CD8^+$ T-cells, such as in the case of an infection, the initiated immune response kills the cell presenting the peptide (Murphy & Weaver, 2017; Reina-Campos et al., 2021). Class-II proteins are mainly expressed by specific immune cells such as B-cells or dendritic cells. Antigens presented by class-II proteins originate from the extracellular medium. If they are recognized by $CD4^+$ T cells, subsequent signaling initiates the production of antibodies by B-cells against the pathogen or orchestration of the immune response by the helper T cells (Murphy & Weaver, 2017; Ruterbusch et al., 2020).

In humans, MHC genes are named as human leukocyte antigen (HLA). HLA genes stand out within the human genome for several reasons. All HLA genes are located close to each other on the short arm of chromosome 6 (Shiina et al., 2009). Classical HLA genes are divided into two classes with three HLA class-I (HLA-A, HLA-B and HLA-C) and three HLA class-II genes (HLA-DR, HLA-DP and HLA-DQ). The so-called non-classical HLA genes are not as polymorphic as classical HLA genes. Furthermore, their primary role is not peptide binding and usually their expression is restricted to specific tissues (Kochan et al., 2013; Mellins & Stern, 2014). Classical HLA genes (hereafter simply referred to as HLA genes) are highly polymorphic and polygenic (i.e. multiple loci coding for proteins with similar functions) (Kaufman, 2018b; O'Connor et al., 2019; Yamaguchi & Dijkstra, 2019). Hundreds of different alleles were identified for each HLA loci surpassing any other region in the genome (Robinson et al., 2020). Each HLA molecule binds a specific set of peptides based on the amino acid sequence of their antigen-binding site that is encoded by exons 2 and 3 for HLA class-I and exon 2 for HLA class-II genes (Bjorkman et al., 1987; Reche & Reinherz, 2003). Most of the variation that is observed within HLA genes is concentrated in these specific exons (**Figure 2**) (Lima et al., 2019; Robinson et al., 2017). Furthermore, the number of non-synonymous mutations (i.e. mutations leading to an amino acid change in the protein) is much higher than the number of synonymous mutations (i.e. mutations that do not alter the protein sequence) across the exons coding for the antigen-binding site compared to the rest of the HLA genes (Hughes & Hughes, 1995; Hughes & Nei, 1988, 1989). Finally, an excess of intermediate frequency alleles is observed within HLA genes deviating from the neutral expectations (Brandt et al., 2018). Overall, these features of HLA genes, combined with their crucial role in adaptive immunity, have led to the hypothesis that HLA diversity is maintained by pathogen-mediated balancing selection (Bitarello et al., 2018; Hedrick & Thomson, 1983; Meyer & Thomson, 2001; Radwan et al., 2020).

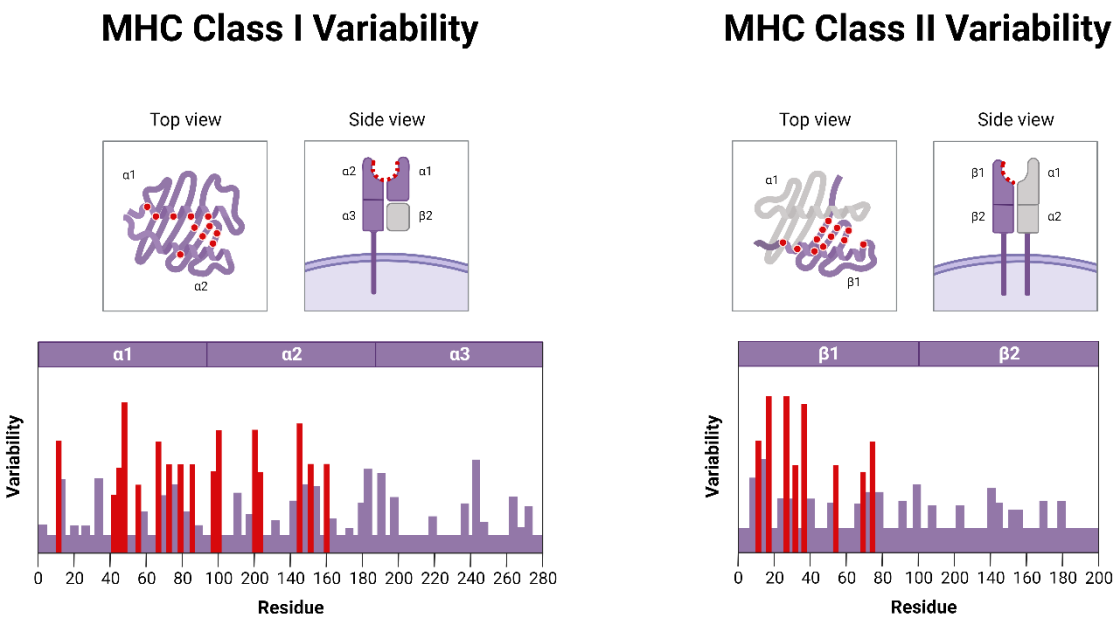


Figure 2. MHC variability is concentrated on residues of class-I and class-II proteins that interact with the presented peptide. Red points on top panel indicates these residues which form the antigen-binding site of the protein. Red bard on the bottom panel shows increased diversity within these residues. Figure template is generated by Akiko Iwasaki and retrieved from https://app.biorender.com/biorender-templates

Balancing selection is, in fact, any selective mechanism that leads to the maintenance of polymorphism within populations (Futuyma, 2009). In the case of pathogen-mediated balancing selection on MHC genes, three models of selection have been proposed. The negative-frequency dependent selection (also known as rare allele advantage) was first proposed by J. B. S. Haldane in his influential paper 'Disease and Evolution' (Haldane, 1949) and later discussed in the context of MHC diversity as well (Bodmer, 1972; Ejsmond & Radwan, 2015). The fitness of pathogens depends on their ability to infect as many hosts as possible. Therefore, pathogens that can circumvent the immune resistance of hosts carrying the most common alleles will be favored by natural selection. Consequently, hosts with rare alleles would have increased resistance and rare alleles would start increasing in frequency due to their selective advantage (Lenz, 2018). Once the rare, advantageous alleles become dominant in the host population, pathogens quickly adapt to them, impairing their selective advantage (Borghans et al., 2004). The overall result of such coevolutionary dynamics is the maintenance of diversity within the host population (Phillips et al., 2018a; Spurgin & Richardson, 2010).

Rare alleles can be novel within the population arising either by mutation or as a result of introgression. On the other hand, they can be old alleles that were once common but decreased in frequency due to selective disadvantage. The origin of the rare and presumably advantageous alleles has long been a point of discussion mainly with a focus on the observation that some MHC alleles are shared by related species despite millions of years of isolation (i.e. trans-species polymorphism) (J. Klein et al., 1998; Takahata & Nei, 1990). It has been argued that the selection on rare novel alleles would lead to rapid turnover of alleles that would prevent the long-term persistence of allelic lineages as in the case of trans-species polymorphism (Slade & McCallum, 1992; Takahata & Nei, 1990). However, if old alleles regain their selective advantage once they become rare, polymorphism can be maintained for long time (Apanius et al., 1997; Takahata & Nei, 1990). While most of the discussion on the details of the negative-frequency dependent selection is based on results of simulation studies, recent experimental work has shown that novel alleles that have introgressed from other populations or species can also provide a selective advantage (Nadachowska-Brzyska et al., 2012; Phillips et al., 2018a). Such introgression events would appear as trans-species polymorphism (Wegner & Eizaguirre, 2012). The second model of balancing selection is heterozygote advantage. This model is based on the fact that MHC heterozygous individuals can bind more peptides than MHC homozygous individuals (Doherty & Zinkernagel, 1975). Due to wider peptide binding, heterozygotes will be able to mount a stronger immune response against pathogens and will have increased relative fitness compared to homozygotes (Penn et al., 2002). As a result, a higher number of alleles can be maintained within populations (Hughes & Nei, 1988). Related to the heterozygote advantage model, the divergent allele advantage hypothesis proposes MHC alleles with higher sequence divergence would present highly different peptides while the peptide pool of similar alleles would overlap considerably. Therefore, individuals that are heterozygous with highly divergent alleles would be able to respond to a much wider set of peptides (Lenz, 2011; Pierini & Lenz, 2018; Wakeland et al., 1990). The heterozygote advantage model has received support from several computational and experimental studies (Arora et al., 2020; Carrington et al., 1999; McClelland et al., 2003; Takahata & Nei, 1990). However, it has also received critique and has been argued to be insufficient by itself to maintain a high number of alleles as observed in the MHC genes (Borghans et al., 2004; Slade & McCallum, 1992). Finally, the third model of balancing selection is the fluctuating selection. This model is based on the assumption that pathogen abundance, composition and overall selective pressure change over time and space (Dunn et al., 2010; A. V. S. Hill, 1991). In return, the subset of MHC alleles that are advantageous would change over time or across geographically distinct subpopulations and

polymorphism is maintained (Hedrick, 2002). While the three models mentioned here are not necessarily mutually exclusive, their relative roles in MHC evolution are still being debated (Meyer & Thomson, 2001; Radwan et al., 2020; Spurgin & Richardson, 2010).

Overall, models of balancing selection, especially the fluctuating selection model and the negative-frequency dependent selection model, rely on the existence of associations between MHC alleles and specific pathogen species or strains (Spurgin & Richardson, 2010). In humans, associations between HLA alleles and infectious diseases including viral (Hammer et al., 2015; McLaren & Carrington, 2015), bacterial (Tong et al., 2015) and eukaryotic (A. V. Hill et al., 1991) infections have been identified (Sanchez-Mazas, 2020). However, considering hundreds, even thousands of pathogens infecting humans ("Microbiology by Numbers," 2011), associations identified until now seem so few, usually weak and mostly inconclusive in the sense that causal explanations related to peptide-binding could not be established (Blackwell et al., 2009; Sanchez-Mazas, 2020; Trowsdale, 2011). A notable exception is the human immunodeficiency virus (HIV) infection which causes acquired immune deficiency syndrome (AIDS) if left untreated (McLaren & Carrington, 2015). It has been shown that HLA alleles that are associated with slow disease progression present specific, highly conserved HIV epitopes (Borghans et al., 2007; Kunwar et al., 2013a). Although escape mutations (i.e. mutations on the pathogen genome that prevent binding of peptides by HLA molecules) arise, they usually entail a high fitness cost for the pathogen (Kløverpris et al., 2016). However, HIV is a relatively novel pathogen that has been infecting humans only for 4-5 generations and has been subject to major disease control and treatment efforts (Korber et al., 2000). Therefore, although it is useful for establishing mechanistic explanations of HLA-mediated disease control, no effect of HIV would be visible in the human genome from an evolutionary perspective. On the other hand, older diseases such as malaria may provide opportunities to test hypotheses regarding HLA evolution. Indeed, the allele HLA-B*53 which has been suggested to be protective against malaria based on case-control studies was found to be at high frequencies in regions where malaria is highly prevalent, possibly in response to selective pressure by malaria (A. V. Hill et al., 1991; Sanchez-Mazas et al., 2017). Although encouraging, these results are still valid just for a few pathogens, missing the diversity of human pathogens. The diversity of HLA genes was shown to be correlated with the pathogen diversity, suggesting that local pathogen composition indeed shapes the evolution of HLA diversity (Prugnolle, Manica, Charpentier, et al., 2005). The role of balancing selection on HLA evolution in this context has received major research attention (Meyer et al., 2018). However, it is very well possible that selection can only

paddle against the current of genetic drift. Random fluctuations in allele frequencies across generations can easily mask the effect of selection specifically in small populations in which the effect of selection is smaller (Sanchez-Mazas, 2001). Furthermore, the extreme diversity of HLA genes is accompanied by many rare alleles (Robinson et al., 2017) that are particularly susceptible to stochastic forces of evolution. Several studies have reported that HLA diversity that is measured based on heterozygosity is negatively correlated with the distance from East Africa where anatomically modern humans have evolved (Prugnolle, Manica, Charpentier, et al., 2005; Sanchez-Mazas et al., 2012). This trend is the result of many founder effects when humans colonize the rest of the world (Prugnolle, Manica, & Balloux, 2005). Likewise, populations with a more recent history of strong genetic drift such as Amerindians or Taiwanese harbor lower HLA diversity (Buhler & Sanchez-Mazas, 2011). Although the loss of HLA diversity might be compensated by the selective maintenance of divergent alleles, genetic drift remains a major determinant of HLA diversity (Buhler et al., 2016; Pierini & Lenz, 2018).

One of the reasons why there are relatively few associations between infectious diseases and HLA might be that infectious diseases are highly understudied in genome-wide association studies. A 2018 study revealed that only 4% of studies in the GWAS Catalog were in the category of infectious diseases (Mozzi et al., 2018). HLA genes are exceptionally difficult to analyze in association studies due to high polymorphism and extreme linkage disequilibrium that can easily lead to false positive or false negative results with small sample sizes and unaccounted population structure (Mozzi et al., 2018; Sanchez-Mazas, 2020). One of the methods that can supplement association studies in revealing the mechanistic basis of HLA-pathogen interactions is computational peptide-binding prediction. These algorithms can identify the set of peptides that are likely to be bound by HLA molecules based on the amino acid sequence (Peters et al., 2020). Although many methods were developed since the 1990s, the recent advances in machine learning approaches have led to the development of highly accurate prediction algorithms (Paul et al., 2020; Reynisson et al., 2020). These algorithms were also employed to investigate hypotheses related to MHC evolution and yielded fruitful results (Arora et al., 2020; Buhler et al., 2016; Pierini & Lenz, 2018).

The fact that research on infectious diseases is lagging behind the research on non-communicable diseases is not specific to association studies but an overall trend in the biomedical sciences (Cohen, 2000). Improved nutrition and hygiene, combined with two very powerful medical interventions namely antibiotics and vaccines resulted in an immense decrease in mortality from infectious diseases (Armstrong et al., 1999). As a result, non-

infectious diseases became the major source of mortality and morbidity, which in turn attracted the attention of biomedical researchers. However, pandemics such as AIDS and much more recently COVID-19 have shown that humans are still far from eliminating the risk posed by infectious diseases (Korber et al., 2000; World Health Organization, 2020). Indeed, millions of lives are still being lost each year due to infectious diseases in parts of the world where the public health infrastructure is less developed and access to medications is limited such as many countries in sub-Saharan Africa (World Health Organization, 2021). Furthermore, climate change associated with global warming might cause changes in the distribution of pathogens or pathogen-carrying vectors and facilitate the emergence of novel pathogens in the future (Khasnis & Nettleman, 2005; Patz et al., 1996). It is clear that infectious diseases are not a book closed for humans. The best way to look forward is to learn from the past. As descendants of humans who have survived countless diseases over hundreds of thousands of years, we have a lot to learn from our genome and our evolutionary history.

Thesis Outline

My research during my PhD is centered on the mechanisms maintaining Human Leukocyte Antigen (HLA) gene diversity and their outcomes at the population level. The main hypothesis on why HLA genes are extremely diverse is built upon their role in the adaptive immune system against pathogens. I used computational methods to analyze both the pathogen diversity at the molecular level and the HLA diversity that evolves in response to pathogens. Furthermore, we employed population genetics methods to investigate how distinct evolutionary mechanisms, such as selection and genetic drift affect HLA diversity within and between populations. This work consists of three chapters.

Chapter 1

Unique Pathogen Peptidomes Facilitate Pathogen-Specific Selection and Specialization of MHC Alleles

HLA genes in humans are highly polymorphic. The excessive polymorphism observed at MHC genes is thought to result mainly from the need to recognize diverse pathogens. Indeed, HLA molecules bind and present small peptides from invading pathogens that eventually initiate the adaptive immune response against them. In this chapter, we analyzed the peptide diversity of 36 common human pathogens and found that only 1.2% of all peptides were shared between two or more pathogens while the rest were unique for each pathogen. This finding provides an empirical basis for the assumption that pathogen diversity drives HLA diversity. In order to analyze how HLA evolves in response to this diversity, we used computational peptide-binding prediction tools and calculated a specialization score for each HLA allele, based on the number of peptides that they bind from pathogens. We observe a negative correlation for HLA alleles between the specialization score and the size of the peptide repertoire (i.e. total number of peptides that alleles bind). This result supports the hypothesis that HLA alleles with small peptide repertoires are maintained within populations due to their protective effect against specific pathogens. Overall, the results of this chapter provide a mechanism through which variation in peptide-binding repertoires of HLA alleles is maintained.

Chapter 2

Balancing selection rather than local adaptation shape excessive HLA gene variation in ethnically diverse African populations

Although the balancing selection is commonly invoked to explain the high diversity observed in HLA genes, the extent and outcome of such selection on population genetic variation remain an open question. In fact, it is hypothesized that distinct mechanisms of balancing selection may cancel each other's effect and result in diversity patterns similar to neutral loci. Better characterization of HLA diversity is required to evaluate different evolutionary forces acting on HLA genes. However, the limited representation of non-European populations in immunogenomics studies hampers research efforts mainly because European populations harbor only a part of the human genetic diversity. In this chapter, we combined whole-genome sequencing (n=180) and targeted HLA data (n=514) of individuals from 12 different ethnic groups from sub-Saharan Africa with data from the 1000 Genomes Project to investigate HLA variation across the most genetically diverse human populations. We found that the genetic diversity in HLA decreases in non-African populations as a result of out-of-Africa migrations, yet this decrease is significantly smaller than it is for neutral loci. Furthermore, the MHC region is enriched with old mutations and population differentiation at HLA genes largely reflects differentiation patterns of neutral markers. These results highlight the role of balancing selection on HLA counteracting the loss of diversity through bottleneck events although genetic drift eventually leads to differentiation. Functional HLA diversity, measured at the level of distinct HLA molecules is at comparable levels in all populations. Therefore, we conclude that demography and balancing selection on ancient variation are major factors driving MHC evolution in humans while the effect of local adaptation appears to be minor.

Chapter 3

Spatio-temporal analysis of ancient HLA data reveals major effects of demography and admixture on immune gene diversity

Major changes in the lifestyle of humans following the domestication of animals and plants were suggested to be responsible for the emergence of many novel pathogens. Indeed, a significant number of pathogens that infect humans have recent origins, dating after the Neolithic. This transition raises the possibility to track the evolution of immune genes by using genomic data of ancient individuals. In the context of HLA, it was suggested that changes in the pathogen pressure would result in dynamic oscillations in allele frequencies, yet the empirical support for such evolutionary dynamics is missing. In this chapter, we employed a recently developed method to genotype HLA class-I alleles from 129 individuals that lived during the early and late Neolithic in Europe. We compared the HLA diversity of ancient individuals to modern populations from Europe, Africa and East Asia by using 1000 Genomes Project data. We observed that the most common alleles in modern central and northern Europeans were missing in the Neolithic while the most common alleles in the Neolithic populations declined in frequency. Although this pattern appears to be in line with the hypothesized dynamic allele frequency changes in response to pathogen pressures, the functional diversity of HLA alleles in both Neolithic and modern populations is similar. By using peptide-binding prediction tools on the *Measles virus*, which is a deadly and novel pathogen that did not exist during the Neolithic, we found that Neolithic populations do not differ from modern humans in terms of the capacity to bind measles antigens. Therefore, we suggest that common HLA alleles in modern European populations were unlikely to have increased due to selection, but were introduced by the steppe pastoralists who migrated to Europe at the end of the Neolithic. There were also several alleles that increased in frequency from early to late Neolithic. These alleles were probably the most common alleles within the hunter-gatherers and increased in frequency in late Neolithic farmers due to admixture. Overall, similar to the conclusions from the second chapter, which relies on spatial analysis of extant populations, our results from the temporal analysis of HLA diversity suggest that allele frequency changes are mainly driven by population admixtures with balancing selection maintaining functional diversity.

List of Papers and Manuscripts

Chapter I

Özer, O. & Lenz, T. L. (2021). Unique Pathogen Peptidomes Facilitate Pathogen-Specific Selection and Specialization of MHC Alleles. *Molecular Biology and Evolution*

Chapter II

Özer O.[*]; Harris D.[*]; McQuillan M.[*]; Hayeck T.; Mbunwe E.; Mosbruger T. L.; Duke J. L.; Azuure C.; Shaw T-W.; Nyambo T.; Mpoloka S. W.; Mokone G. G.; Belay G.; Fokunang C.; Njamnshi A. K.; Maiers M.; Monos D.[#]; Lenz T.L. [#]; Tishkoff S.[#] : Balancing selection rather than local adaptation determines HLA gene variation in ethnically diverse African populations. Unpublished manuscript. [*]equal contribution as first-authors. [#] equal contribution as senior authors

Chapter III

Özer O.; Chen Y-R.; da Silva N. A.; Haller M.; Calvignac-Spencer S.; Nebel A.; Krause-Kyora B.; Lenz T. L.: Spatio-temporal analysis of ancient HLA data reveals major effects of demography and admixture on immune gene diversity. Unpublished manuscript.

Appendix I

Pierini, F., Nutsua, M., Böhme, L., Özer, O., Bonczarowska, J., Susat, J., Franke, A., Nebel, A., Krause-Kyora, B., & Lenz, T. L. (2020). Targeted analysis of polymorphic loci from low-coverage shotgun sequence data allows accurate genotyping of HLA genes in historical human populations. *Scientific Reports*

Appendix II

The Severe Covid-19 GWAS Group. (2020). Genomewide Association Study of Severe Covid-19 with Respiratory Failure. New England Journal of Medicine, 383(16), 1522–1534.

Author Contributions

Chapter I — **OÖ** and TLL designed research; **OÖ** performed research and analyzed the data; **OÖ** and TLL interpreted the data and wrote the manuscript.

Chapter II — TLL, DM, MM and ST conceived the study. TN, SWM, GGM, GB, CF, AKN provided samples. **OÖ**, DH and MMcQ analyzed the data with input from TH, EM, TLM, JLD, CA and TWS. **OÖ**, DH, MMcQ, TLL, DM and ST interpreted the results. **OÖ** wrote the manuscript with input from DH, MMcQ. TLL, DM, and ST revised the manuscript.

Chapter III — **OÖ** and TLL designed research with input from SCS, AN and BKK. NAS and MH generated the data. **OÖ** analyzed the data with input from YRC. **OÖ** interpreted the results and wrote the manuscript with input from TLL.

Appendix I — TLL, BKK, AN and FP conceived the study. LB performed the lab work. MNu, TLL and FP developed the bioinformatic pipeline. FP analyzed the data with input from MNu, TLL, BKK, LB, **OÖ**, JB and JS. AF provided research infrastructure. FP and TLL interpreted the results and wrote the manuscript with input from AN and BKK.

Appendix II — **OÖ** contributed to data generation, interpretation and writing the manuscript. Contributions of all authors can be found in the online Supplementary Appendix.

Chapter 1

Unique pathogen peptidomes facilitate pathogen-specific selection and specialization of MHC alleles

Onur Özer[1,2], Tobias L. Lenz[1,2]

[1]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany

[2]Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany

Published in
Molecular Biology and Evolution (2021)
doi:10.1093/molbev/msab176

Abstract

A key component of pathogen-specific adaptive immunity in vertebrates is the presentation of pathogen-derived antigenic peptides by major histocompatibility complex (MHC) molecules. The excessive polymorphism observed at MHC genes is widely presumed to result from the need to recognize diverse pathogens, a process called pathogen-driven balancing selection. This process assumes that pathogens differ in their peptidomes – the pool of short peptides derived from the pathogen's proteome – so that different pathogens select for different MHC variants with distinct peptide-binding properties. Here we tested this assumption in a comprehensive dataset of 51.9 Mio peptides, derived from the peptidomes of 36 representative human pathogens. Strikingly, we found that 39.7% of the 630 pairwise comparisons among pathogens yielded not a single shared peptide and only 1.8% of pathogen pairs shared more than 1% of their peptides. Indeed, 98.8% of all peptides were unique to a single pathogen species. Using computational binding prediction to characterize the binding specificities of 321 common human MHC class-I variants, we investigated quantitative differences among MHC variants with regard to binding peptides from distinct pathogens. Our analysis showed signatures of specialization towards specific pathogens especially by MHC variants with narrow peptide-binding repertoires. This supports the hypothesis that such fastidious MHC variants might be maintained in the population because they provide an advantage against particular pathogens. Overall, our results establish a key selection factor for the excessive allelic diversity at MHC genes observed in natural populations and illuminate the evolution of variable peptide-binding repertoires among MHC variants.

Introduction

MHC molecules mediate the adaptive immune response in jawed vertebrates by binding to short peptides and presenting them on the cell surface. These peptide:MHC complexes on the cell surface are continuously surveyed by T lymphocytes to detect the presence of infectious agents. The excessively high number of alleles at the classical MHC genes, in humans for instance with several thousand for each MHC class-I locus (Robinson et al., 2020), is considered to be maintained by pathogen mediated balancing selection (Bodmer, 1972; Hedrick & Thomson, 1983; Radwan et al., 2020). In support of this hypothesis, most of the polymorphism within the MHC genes is observed in the residues forming the peptide-binding region of the MHC molecule, i.e. the region that interacts with the presented peptide (Parham, 1988; Robinson et al., 2017). The rate of nonsynonymous variation is much higher in the peptide-binding region compared to the rest of the MHC genes (Hughes & Nei, 1988) and many of these variants are observed at intermediate frequencies (Brandt et al., 2018).

Three distinct yet not mutually exclusive mechanisms of pathogen-mediated balancing selection, namely heterozygote advantage, negative frequency-dependent selection and fluctuating selection, have been proposed relatively early on and have been analyzed in a trove of different studies in various species over the last decades (Apanius et al., 1997; Radwan et al., 2020; Spurgin & Richardson, 2010). According to the heterozygote advantage hypothesis, individuals with heterozygous MHC genotype present a higher coverage of peptides and, consequently, are able to mount an immune response against a larger range of pathogens compared to homozygotes (Doherty & Zinkernagel, 1975; Hughes & Nei, 1988). The heterozygote advantage hypothesis is further extended by a divergent allele advantage model, which relies on the assumption that MHC alleles that are divergent at the sequence level would have a low overlap in their peptide repertoires (Wakeland et al., 1990). Several theoretical as well as computational studies have supported the role of the divergent allele advantage model in maintaining allelic diversity (Lenz, 2011; Pierini & Lenz, 2018; Stefan et al., 2019). The negative frequency-dependent selection hypothesis assumes that most of the pathogens evolve much faster than their hosts and adapt to evade recognition by the most common MHC alleles (Bodmer, 1972). Such adaptation provides a selective advantage to rare or novel MHC alleles and leads to cyclic fluctuations in allele frequencies (Ejsmond & Radwan, 2015; Lenz, 2018; Phillips et al., 2018b). Finally, fluctuating selection on distinct MHC alleles is expected if the prevalence or selection pressure of pathogens changes over time or across geographical locations (dos Santos Francisco et al., 2015; Dunn et al., 2010; Hedrick, 2002). All three

mechanisms are based on two main assumptions: i) that each pathogen challenges the MHC-based immune system in a different way, and ii) that MHC variants differ in the repertoire of presented peptides.

The first assumption, that each pathogen challenges the adaptive immune system in a novel way, assumes that pathogens exhibit distinct antigenic peptide composition (**Fig. 1**). Although the extent of antigenic diversity among pathogens is crucial for our understanding of the evolution of MHC genes, it has been analyzed only in few studies with a limited number of pathogen species, and mainly in the context of self/non-self overlap of peptides (Burroughs et al., 2004; Calis et al., 2012). So, while this first assumption appears widely accepted, systematic empirical evidence supporting this assumption is still lacking. The second assumption, that MHC variants differ in their repertoire of presented peptides, is empirically well supported (Pierini & Lenz, 2018; I. M. M. Schellens et al., 2015; Sidney et al., 2008) and so is the fact that different MHC alleles are associated with different infectious diseases (Sanchez-Mazas, 2020; Tian et al., 2017; Trowsdale, 2011). However, how exactly the allele-specific peptide repertoire leads to the allele's effect on disease risk is still a matter of intense research (Radwan et al., 2020). In fact, the differential ability of MHC variants to trigger an immune response against specific pathogens can be determined by both quantitative (binding many or few peptides of a given peptide pool) and qualitative (binding or not binding of specific peptides) differences among alleles. Croft et al. (2019) have shown that up to 80% of viral peptides that are presented on MHC class-I molecules can be immunogenic in mice. This suggests that selection might act more on quantitative differences among MHC alleles, i.e. binding more peptides from a specific pathogen is advantageous. In line with that idea, Arora et al. (2020) have shown that in HIV-infected individuals, the viral load is negatively correlated with the number of HIV peptides that are predicted to be presented by an individual's HLA-B variants. On the other hand, they also showed that the presence of a specific HLA-B variant (B*57:01) alone provided a stronger protective effect than the protective effect achieved by merely binding many peptides per se (Arora et al., 2019). Indeed, several studies report that HLA-B restricted T-cell responses in HIV-1infected individuals with slow disease progression tend to target conserved regions of the HIV-1 (Costa et al., 2010; Gillespie et al., 2006; Kunwar et al., 2013b). Similar observations on other pathogens such as hepatitis C virus (Rao et al., 2015) or influenza (Eickhoff et al., 2019) indicate that not only the quantity but also the quality of peptides presented on MHC molecules affects disease outcome. It thus remains an open question how quantitative and qualitative differences in peptide binding among MHC variants contribute to

their disease association, and thus to which extent each of these properties are the target of pathogen-mediated selection.

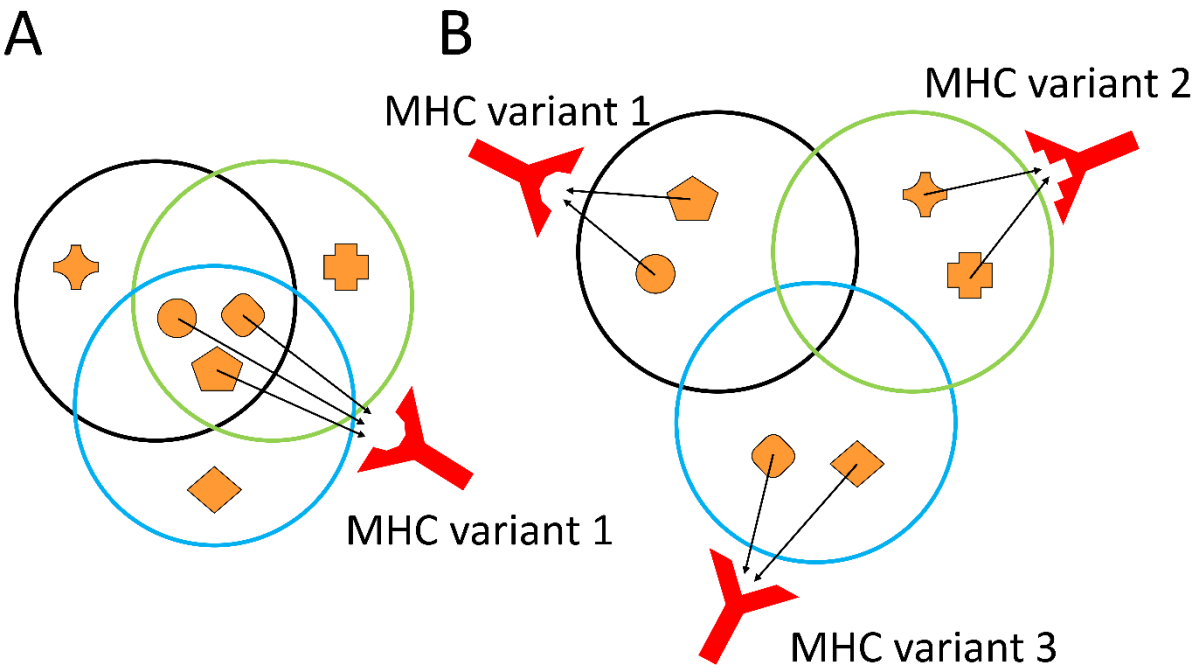


Figure 1. Conceptual representation of selection of different MHC variants depending on the peptidome compositions of pathogens. Empty circles represent peptidomes of three different pathogens while orange icons represent representative antigens. If peptide sharing among pathogens is extensive **(A)**, a single MHC variant would be enough to mount an immune response against all pathogens. If the peptide sharing among pathogens is very low **(B)**, different MHC variants would be required for effective overall pathogen control.

Intriguingly, recent studies have revealed that there is indeed substantial quantitative variation in the size of the bound peptide repertoire (i.e. total number of bound peptides, hereafter referred to as 'promiscuity') among MHC variants (Chappell et al., 2015; Paul et al., 2013). Along this promiscuity scale, promiscuous MHC variants bind a wide range of peptides while fastidious variants are much more stringent in peptide binding and exhibit narrow repertoires. Manczinger et al. (2019) showed that the frequency of promiscuous MHC class-II variants is positively correlated with the pathogen-richness across countries, possibly because more promiscuous MHC variants provide an advantage by facilitating recognition of more pathogens. However, it is yet to be determined how fastidious MHC variants, which present a smaller peptide

repertoire, are maintained in populations. One intriguing hypothesis, proposed by Kaufman (2018), postulates that fastidious alleles may have a selective advantage if they are specialized against particular pathogens, especially if the immunodominant peptides are highly conserved (Miura et al., 2009; Schneidewind et al., 2007). Previous studies focusing on modeling approaches have shown that such specialization against specific pathogens may contribute to maintenance of high polymorphism at the MHC locus (Hedrick, 2002). Although the number of experimental assays investigating peptide-MHC interactions increased rapidly in the last years (Vita et al., 2019), most of the empirical evidence is still focused on a few very common alleles and specific proteins. Therefore, it remains challenging to empirically test hypotheses on MHC specialization across a wide range of alleles and pathogens. Computational approaches for the prediction of peptide-binding by MHC variants fill this gap to some extent (Peters et al., 2020). In fact, recent advances in prediction algorithms allow relatively accurate characterization of binding specificities even for MHC variants for which there is no empirical data available (Reynisson et al., 2020).

Here we analyzed the potential antigenic diversity of 36 representative human pathogens and show that each pathogen harbors a distinct peptide pool, with only few peptides shared among pathogens. We then investigated how this antigenic diversity is reflected in the peptide binding properties among human MHC variants. We characterized the allele-specific repertoire of bound peptides for a set of 321 common HLA class-I alleles using computational binding-prediction. Our results revealed an extensive variation in peptide binding promiscuity among MHC alleles as well as signatures of specialization mainly for fastidious alleles.

Results

Human Pathogen Peptidome

Complete proteomes of a representative set of diverse human pathogen species, including viruses (N = 10), bacteria (N = 19) and eukaryotic parasites (N = 7, **Table S1**) were divided into all possible nine amino acid long peptides, shortly nine-mers, by employing a sliding window approach with a step size of 1 amino acid, resulting in 51,861,826 nine-mers reflecting a broad representation of the human pathogen peptidome. The number of peptides per pathogen species ranged from 1760 to 11,405,499 (median: 546,629.5, **Table S1**). Of all the nine-mers, 98.8% were unique to the given pathogen from which they originated, thus only 1.2% were

shared among two or more pathogens (**Fig. 2A**). Pairwise comparisons of peptide sharing among pathogens revealed that pathogens on average shared only a tiny fraction of their peptides (Median: 0.005%; Range: 0% - 8.8%), with 39.7% of the 630 pathogen pairs showing no shared peptide at all and only 11 pairs (1.8%) sharing more than 1% of their peptides (**Fig. 2B**). For some pathogens with large peptidome sizes, this can amount to large absolute numbers of peptides (**Fig. S1**), even though it remains negligible in relative terms. In a subset of bacterial (N = 14) and eukaryotic (N = 2) pathogens, for which evolutionary divergence information was available (**Table S1**), peptide sharing was found to be negatively correlated with the evolutionary distance between pathogens (Kendall's tau = -0.7, $p < 0.001$) (**Fig. 2C**), suggesting a dominant role for sequence homology based on phylogenetic relatedness as a major determinant of peptide sharing.

Peptide Binding Promiscuity of HLA variants

Having established that each pathogen is likely to challenge the adaptive immune system with a distinct set of peptides (**Fig. 1B**), we next sought to investigate whether and how HLA molecules have adapted to this extreme peptidome diversity. Here we use the term 'HLA variant' to denote distinct variants of the classical HLA molecules that are encoded by a distinct HLA allele at 2nd field resolution. In other words, each HLA variant corresponds to an HLA molecule with a distinct amino acid sequence. Prior studies have shown that the peptide binding promiscuity (i.e. size of the repertoire of bound peptides) varies markedly among HLA variants (Chappell et al., 2015; Paul et al., 2013), raising the question of how HLA alleles that encode molecule variants that bind only few peptides are maintained in the population. Kaufman (2018) suggested that such fastidious HLA variants might provide an advantage through specialization towards particular pathogens. In the light of the distinct pathogen peptidomes shown above, this possibility appears plausible.

In order to investigate this hypothesis in more detail, comprehensive data about the peptide repertoires of a large number of different HLA variants is required. The optimal data for such an analysis would be derived from in-vitro HLA:peptide binding or peptide elution assays, but such empirical data is so far only available for a very limited set of non-randomly selected HLA variants and peptide repertoires. Another possibility for approaching the variation in peptide binding promiscuity is to utilize computational peptide binding prediction algorithms. These machine-learning algorithms are developed in the context of vaccine development and cancer

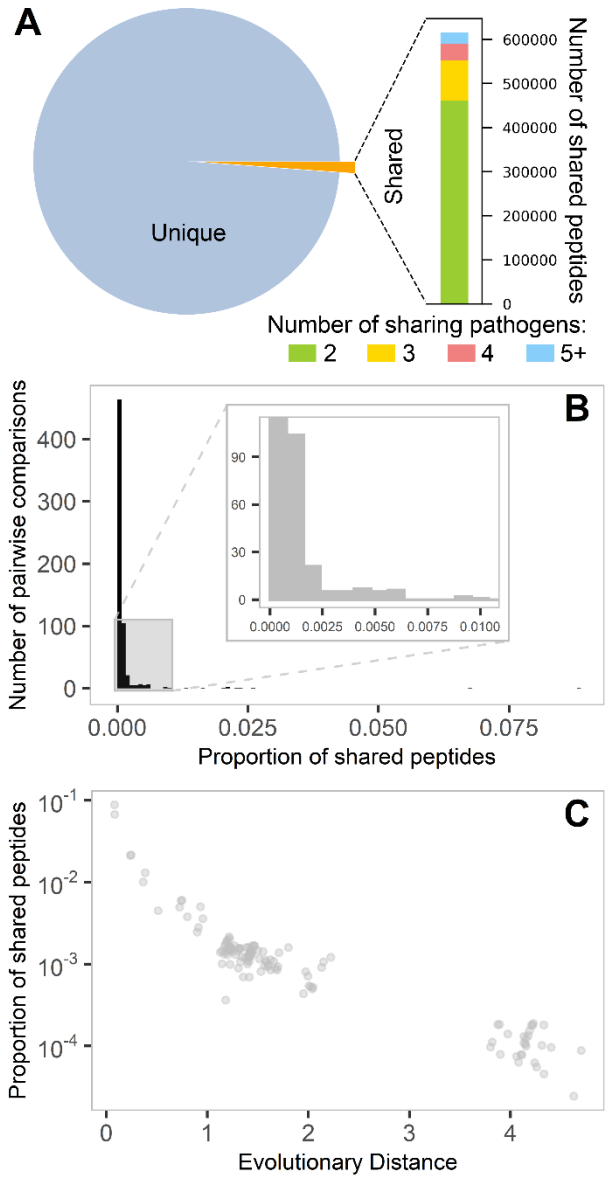


Figure 2. Peptide sharing among human pathogens. **(A)** Shared nine-mer peptides constitute a very small part of the human pathogen peptidome. The pie chart represents the proportions of shared (N = 615,904) and unique (N = 51,861,826) nine-mer peptides while the bar chart shows the extent of sharing across pathogen species for all shared peptides. **(B)** Distribution of the pairwise peptide sharing among all pathogens (N = 36). The fraction of shared peptides out of all peptides bound by either pathogen is shown for all pathogen pairs (N=630). **(C)** Pairwise peptide sharing among a subset of pathogens (N = 16) decreases with increased evolutionary distance. Evolutionary distance between pairs of pathogens was calculated as tip-to-tip distances within the tree of life. See Table S1 for organisms used in the analysis.

immunotherapy and have been improved in accuracy over the past decade (Paul et al., 2020; Schirle et al., 2001). They are now well established and used in a wide range of contexts, including evolutionary genetic studies of the MHC (Arora et al., 2020; Buhler et al., 2016; Lenz, 2011; Manczinger et al., 2019; Pierini & Lenz, 2018). These algorithms are still imperfect in accurately predicting the antigenicity of specific peptides, however, they perform relatively well in predicting overall repertoires of bound peptides for a given HLA allele (Paul et al., 2020). For the present analysis of allele-specific overall peptide repertoire sizes, we focus on HLA class I variants, because their binding motifs are more clearly defined and computational

binding prediction is considered to be more accurate for HLA class I (Reynisson et al., 2020). We thus rely on one of the most established HLA:peptide binding prediction algorithm in order to study all classical HLA class-I alleles that are classified as "common" in the CIWD alleles catalogue (Hurley et al., 2020). Binding affinities between all unique nine-mer peptides and the selected HLA class-I variants were computationally predicted. Promiscuity of an HLA variant was defined as the fraction of peptides bound by the variant (with an affinity below a defined threshold) out of the complete set of unique peptides (n = 51,861,826 nine-mers). Promiscuity values were highly correlated between 50 nM (strong binders) and 500 nM (strong and weak binders) thresholds (Kendall's tau = 0.82, $p < 0.001$). Therefore, for the rest of the analysis a threshold of 500 nM was used. Variants having the exact same binding prediction results and the same first field number as another allele with a lower second field number (representing highly related alleles with negligible sequence difference) were removed, resulting in 82 HLA-A, 180 HLA-B and 59 HLA-C alleles for subsequent analyses. The correspondence between the computational and experimental promiscuity values was tested for a subset of HLA-A (N = 19) and HLA-B (N = 15) variants, for which experimental data was available from the IEDB database (Vita et al., 2019). A moderate correlation was observed for both HLA-A (Kendall's tau = 0.51, $p = 0.002$) and HLA-B variants, although the latter was not statistically significant (Kendall's tau = 0.37, $p = 0.054$, **Fig. S2**), possibly owing to the small number of variants and the limited and non-random collection of peptides in the IEDB data.

Using the newly obtained information of predicted HLA allele-specific peptide binding, we first reanalyzed the sharing of peptides among pathogens. Peptides that were predicted to be bound by the same set of HLA variants (out of all HLA variants) were merged so that each merged peptide group represents all peptides that are equivalent from the HLA perspective (see Material and Methods). Peptide sharing based on the merged peptide groups (N = 4,157,475) showed that still 85.6% of groups were unique to a specific pathogen, with the rest shared by two or more pathogens (**Fig S3**).

We then used the HLA:peptide binding data to investigate the variation in peptide repertoire sizes (i.e. promiscuity) among HLA class I variants. Promiscuity of individual HLA variants varied greatly within and between loci (**Fig. S4**). Although both promiscuous (i.e. with large peptide repertoire) and fastidious (i.e. with small peptide repertoire) variants are observed at all loci, HLA-B and HLA-C loci appear to have narrower peptide repertoires when compared to the HLA-A locus (pairwise Wilcoxon rank sum test, HLA-A and HLA-B: $p < 0.001$, HLA-A and HLA-C: $p = 0.005$). The difference in promiscuity between HLA-A and HLA-B loci was

confirmed by two experimental datasets, the IEDB dataset (Wilcoxon rank sum test, $p = 0.026$) and a dataset curated by Abelin et al. (2017) (Wilcoxon rank sum test, $p = 0.02$). Analysis of promiscuity in a phylogenetic context showed that large differences in the peptide repertoire size of HLA alleles can evolve quickly within all loci as closely related alleles can differ markedly in promiscuity (**Fig. S5**).

Specialization of HLA variants in Peptide Binding

The observed large differences in promiscuity among HLA class-I alleles confirmed previous empirical studies on more limited sets of alleles (Chappell et al., 2015; Paul et al., 2013). However, this observation raises the question of whether the variation in promiscuity is a random byproduct of sequence evolution of these HLA allele, and also emphasizes the puzzle how fastidious alleles are maintained in the population. According to the hypothesis by Jim Kaufman (Chappell et al., 2015; Kaufman, 2018a), promiscuous variants might act as generalists, providing protection from a large set of common pathogens, while fastidious variants may be specialized against one or few pathogens. Specialization may confer a selective advantage to fastidious variants especially in times of outbreaks or persistent high pathogen pressure by these specific pathogens. In order to investigate potential specialization and test this hypothesis in a quantitative way, we calculated for each variant the normalized fraction of bound peptides from each pathogen. Overall, the peptide binding values (standardized for pathogen proteome size and HLA variant promiscuity – see Material & Methods) for each HLA class-I locus were normally distributed (**Fig. S6**). Yet, within each locus, there are HLA variant-pathogen pairs with distinct associations, potentially indicating non-random relationships (**Fig. 3A**, **Fig S7**). In order to analyze these peptide binding patterns in more detail, a specialization value was calculated for each HLA class-I variant that reflects the relative difference between the variant's ability to bind peptides of its best-covered pathogen compared to all pathogens. A high specialization score indicates that the variant binds particularly many peptides from its best-covered pathogen, compared to the number of peptides it generally binds across all pathogens. Intriguingly, this specialization value was negatively correlated with the promiscuity of variants for all HLA class-I loci (Kendall correlation, HLA-A: $tau = -0.22$, $p = 0.003$; HLA-B: $tau = -0.28$ $p < 0.001$; HLA-C: $tau = -0.33$ $p < 0.001$; **Fig. 3B**). In other words, fastidious variants tend to have higher specialization values than promiscuous variants. Stronger correlations were observed using a 50 nM threshold that includes only strong binders (Kendall correlation, HLA-A: $tau = -0.37$, $p < 0.001$; HLA-B: $tau = -0.37$ $p < 0.001$; HLA-C: $tau = -0.62$ $p < 0.001$). While we analyzed a comprehensive set of all pathogens here, we expect

that this relationship would also hold only among strictly intracellular pathogens, given that the viruses in this dataset generally exhibited the most extreme values of specialization, and that there are intracellular pathogens also among the other pathogen groups. In order to rule out the possibility that the observed negative correlation between promiscuity and specialization is driven by general variation related to both the overall differences in peptide repertoire sizes among alleles and differences in peptidome sizes of pathogens, a simulated version of the binding data for all three HLA loci was generated for each variant by using the promiscuity values of the variants as probabilities of binding a given peptide from a given pathogen (see Material & Methods). This simulation mimicked the observed data distribution and variation within and among alleles except for any potential preference towards specific pathogens. The simulated data was then analyzed in the same way as the real data. No correlation between promiscuity and specialization was observed in the simulated data (Kendall correlation, HLA-A: tau = 0.05 p = 0.54; HLA-B: tau = -0.02 p = 0.65; HLA-C: tau = 0.01 p = 0.93; **Fig. S8**).

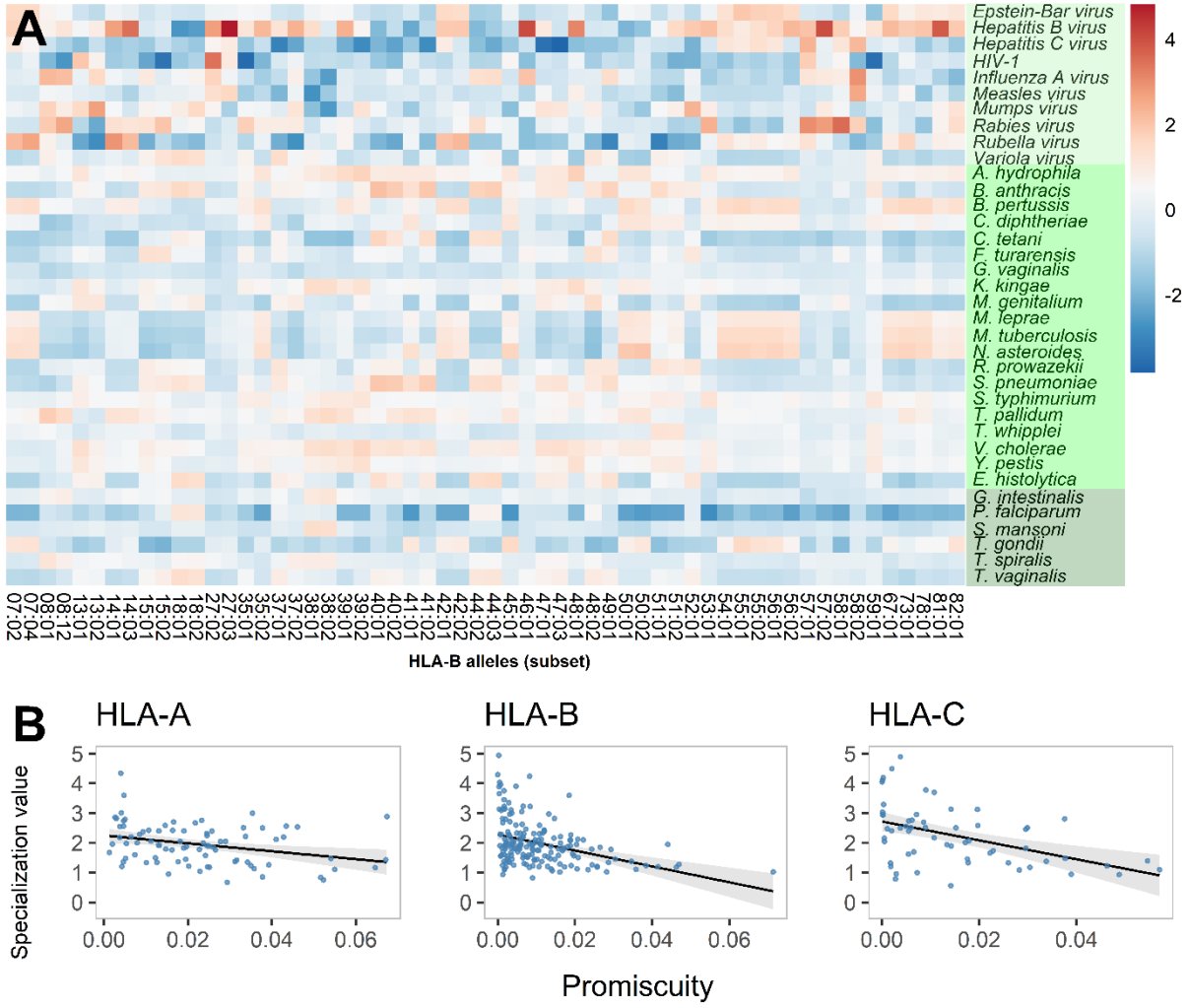


Figure 3. (A) Standardized proportions of bound peptides from distinct pathogens varies greatly for HLA-B alleles. Each cell represents the proportion of bound peptides by the HLA variant (on horizontal axis) from the corresponding pathogen (on vertical axis). Proportions are standardized within each variant to make comparisons across variants possible. Dark red color indicates high specialization, white no specialization and dark blue indicates lower than average binding of a pathogen's peptides. Green shading on vertical axis labels indicates different pathogen groups (light green; viruses, green; bacteria, dark green; eukaryotes). In this plot, only a subset of HLA-B variants were included for better visualization. The subset was selected such that a maximum of two variants with the same first field number and smallest second field number were included (e.g. only B*15:01 and B*15:02 of all the alleles of the B*15 lineage). The patterns are similar across all variants of all three HLA loci. For the complete HLA-A, HLA-B and HLA-C alleles see Fig S7. (B) Specialization is negatively correlated with promiscuity for all HLA class I loci. Each dot represents an HLA variant of the given HLA gene, shown separately for HLA-A (n = 82), HLA-B (n = 180), and HLA-C (n = 59). Specialization was calculated for each variant as the difference between the maximum and the median values of standardized proportions of bound peptides. Promiscuity was calculated for each HLA variant as the fraction of the bound peptides among the complete dataset of 51.9 Mio peptides. Linear regression line is shown in black and 95% CI around the line in gray.

Another interesting question regarding the evolution of HLA alleles and specifically the specialization towards specific peptide repertoires concerns the preferential binding of foreign (i.e. pathogen derived) and self-peptides. It was previously shown based on predicted peptide-binding data that some MHC class-I molecules, specifically HLA-A variants, preferentially present pathogen-derived peptides over self-peptides(Calis et al., 2010; Rao et al., 2009). Following that observation, we therefore used our approach to investigate the relationship between an MHC allele's promiscuity and binding of foreign over self-peptides. The ratio of self-binding fraction (fraction of bound self-peptides over all self-peptides) to foreign binding fraction was used as a proxy for an allele's preference towards self- or foreign- derived peptides (**Fig. 4**). In accordance with the results of Rao et al., (2009) all HLA-A variants were found to have self to foreign binding fractions lower than or very close to one, indicating a binding preference towards foreign peptides. Moreover, we observed a significant positive correlation between the self to foreign binding ratio and promiscuity (Kendall correlation, tau = 0.47, $p < 0.001$), indicating that fastidious HLA-A variants tend to have a significantly higher specificity towards foreign peptides than promiscuous HLA-A variants. A similar positive correlation was also observed for the HLA-C locus (Kendall correlation, tau = 0.22, p = 0.013) and with the exception of three fastidious HLA-C*01 variants, all HLA-C variants had self to foreign binding fractions lower than one. In contrast to HLA-A and HLA-C loci, a weak negative correlation between the self to foreign binding ratio and promiscuity was observed for HLA-B loci (Kendall correlation, tau = -0.14, p = 0.006). Interestingly, promiscuous HLA-B variants

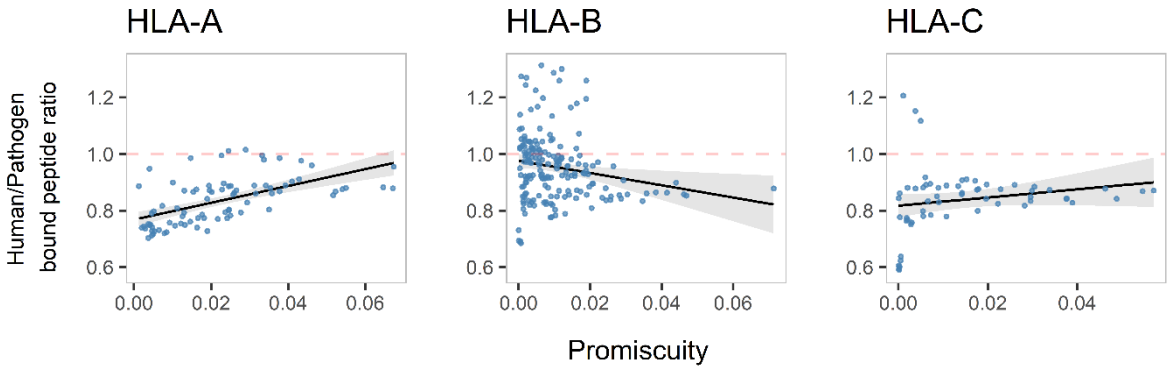


Figure 4. Self to non-self binding ratios as a function of allele promiscuity. Each dot represents an HLA allele of the given HLA gene, shown separately for HLA-A (n = 82), HLA-B (n = 180), and HLA-C (n = 59). Dashed red line indicates a 1/1 ratio, i.e. an equal tendency to bind either human or pathogen peptides. Promiscuity was calculated for each HLA variant as the fraction of the bound peptides among the complete dataset of 51.9 Mio peptides. Linear regression line is shown in black and 95% CI around the line in gray.

consistently had self to foreign binding ratio of lower than one while both high and low ratios were observed in fastidious variants (**Fig. 4**).

Despite the low proportions of shared peptides among pathogens (**Fig. 2B**) the absolute numbers of shared peptides can amount up to several thousand among pathogens with large peptidome size (**Fig. S1**). Therefore, we also investigated whether promiscuous and fastidious alleles differ in binding to peptides shared by three or more pathogens, as a specialization of fastidious variants towards shared peptides would provide a selective advantage by facilitating simultaneous recognition of multiple pathogens. For this, the ratio of the fraction of bound unique peptides to the fraction of bound shared peptides (by three or more pathogens) were compared between the most promiscuous (top 25%) variants and the least promiscuous (bottom 25%, i.e. the most fastidious) variants within each HLA locus. No significant differences were observed for HLA-B and HLA-C while fastidious HLA-A variants seem to bind more unique peptides than promiscuous ones (Wilcoxon rank sum test, $p = 0.005$; **Fig. S9**).

Discussion

Our analysis reveals the vast diversity of the human pathogen peptidome and provides insights into how different peptide binding properties of HLA molecules might have evolved to cope with this diversity. The analysis of peptidome diversity in a diverse set of pathogenic organisms showed that the overwhelming majority of the nine-mer peptides were unique to the pathogen from which they originate. The extreme diversity of peptides among pathogens strongly supports the assumption that the evolution of high allelic diversity of HLA genes is driven by the need for diverse antigen presentation. It also suggests that every pathogen has indeed a different selective effect on the HLA allele pool (**Fig. 1B**) and thus provides the empirical basis for pathogen-by-allele interaction scenarios that underlie two of the most commonly assumed balancing selection mechanisms: negative frequency-dependent selection and fluctuating selection (Radwan et al., 2020). Our results most likely represent a lower limit of peptide sharing mainly because we consider the whole length of the nine-mers for the analysis in a set of diverse pathogens. Many peptides with a few amino acid differences can be considered as equivalents from the MHC perspective especially if the differences are among amino acids with similar chemical properties and outside of the anchor residues (Rammensee et al., 1999). However, even when taking a very conservative perspective of peptide sharing by merging

peptides based on similarity in the set of HLA variants binding a given peptide, we still find the vast majority of merged peptides being unique to a given pathogen.

Several studies reported cross-reactive T-cell responses against closely related viruses (Eickhoff et al., 2019; Weiskopf et al., 2013) or bacteria (Abate et al., 2019) which is in line with our result of increased peptide sharing with smaller evolutionary divergence among pathogens. It should also be noted that even among highly unrelated pathogens, sharing of few but particularly immunogenic peptides could still lead to cross-reactivity in the host immune response. However, none of the HLA variants in our dataset had a particular binding preference towards the peptides shared by two or more pathogens. Many of the shared peptides likely originate from proteins that are highly conserved across organisms, possibly including humans. Targeting of HLA variants towards shared peptides may be ineffective, if they are also shared by humans, because T-cells recognizing those specific peptide-HLA complexes would be eliminated during the thymic selection. Alternatively, if shared peptides are not enriched in any specific sequence motif compared to unique peptides, it becomes impossible for an HLA variant to specialize on shared peptides.

The observed highly distinct peptidome composition of pathogens provides a basis for the hypothesis that some HLA variants might be specialized against particular pathogens. In fact, the computational quantification of peptide binding by common HLA variants revealed substantial variation among different pathogens and HLA variants in the proportion of bound peptides (**Fig. 3A**, **Fig S7**). Moreover, applying a specialization metric to each HLA variant, we found that the variants having higher specialization scores tend to have narrower peptide binding repertoires. This observation supports the hypothesis that specialization on one or a few pathogens might provide a selective advantage to fastidious alleles (Kaufman, 2018a).

One important question regarding our analysis is to what extent the specialization of an HLA variant towards a pathogen coincides with the protection from infection. Our specialization metric is based on comparing different pathogens with respect to the proportions of their complete peptides that are predicted to be bound by an HLA variant. Recently, Arora et al. (2019) showed that the protective effect of HLA-B alleles against HIV-1 viral load is positively correlated with the number of HIV-1 peptides that a given HLA variant is predicted to bind, an effect that was also observable at the genotype level where HIV-1-infected individuals whose HLA-B variants together were predicted to bind more HIV-1 peptides also exhibited a lower viral load and thus a slower progression towards AIDS (Arora et al., 2020). However, MHC-

related determinants of disease outcome are more complex than the mere quantity of bound peptides. It was demonstrated for several pathogens such as HIV-1 (Borghans et al., 2007) and hepatitis-C virus (Rao et al., 2015) that HLA variants that are associated with effective disease control target conserved regions of the pathogen proteome (Hertz et al., 2011). Moreover, proteins that are expressed by a pathogen throughout the infection vary greatly due to sex-specific (Lasonder et al., 2016) or stage-specific (W. Lin et al., 2016) effects, thus also affecting the potential repertoires of presented peptides. Finally, peptides need to go through the steps of the antigen processing pathways, such as proteolytic cleavage or translocation into endoplasmic reticulum before being presented by HLA molecules (Blum et al., 2013; Yewdell et al., 2003). This suggests that only a subset of all possible peptides is actually presented by HLA molecules on the cell surface and some of those presented peptides may be more important than others. It was shown for a few HLA class I alleles that promiscuity is inversely correlated with the cell surface expression of the corresponding HLA molecules (Chappell et al., 2015). If such relationship holds true in general, persistent presentation of a few immunodominant pathogen peptides on the cell surface by fastidious HLA variants would indeed allow efficient pathogen control and such variants can truly be called specialists. Therefore, we do not expect to observe a perfect correlation between quantitative specialization of an HLA variant and pathogen control by individuals carrying that allele, and the specialization parameter presented in our analysis should thus be understood as a metric for an increased probability to present immunodominant peptides from a particular pathogen.

We have observed substantial variation among HLA variants in peptide-binding promiscuity, exceeding orders of magnitude, both within and between different HLA loci. The observed variation is not correlated with the allele divergence, indicating that promiscuous or fastidious HLA class-I variants may evolve quickly in response to varying pathogen pressure. The same conclusion was also reached by Manczinger et al. (2019) for HLA class II HLA-DRB1 variants, highlighting the role of promiscuity in pathogen-mediated selection for both HLA class-I and class-II loci. The median promiscuity level was significantly higher for variants of the HLA-A locus compared to HLA-B and HLA-C loci. Multiple studies on distinct properties of HLA class-I variants revealed differences among these loci, especially between HLA-A and HLA-B loci (Di et al., 2021). Prugnolle et al. (2005) reported that the positive correlation between pathogen richness and allelic diversity is much stronger for the HLA-B locus than the HLA-A locus. dos Santos Francisco et al. (2015) also noted a similar result that when the alleles were classified into supertypes, i.e. allele groups with similar binding properties as determined by

peptide binding pockets, the effect of local adaptation is more evident for HLA-B supertypes. Based on these observations, it can be hypothesized that HLA-A alleles tend to be more promiscuous generalists while HLA-B alleles tend to be more fastidious specialists that evolve quickly in response to varying pathogen pressures. This hypothesis is further supported by the finding of Hertz et al. (2011) that HLA-B alleles effectively target conserved peptides of RNA viruses that are known to evolve very fast (Drake & Holland, 1999). Furthermore, the HLA-B locus harbors the highest number of alleles among all HLA loci, which is in line with the idea that it most closely evolves with specific pathogens. It should be noted that such hypothesis does not exclude specialist HLA-A alleles or generalist HLA-B alleles as we also show that promiscuity can evolve very quickly (by few mutational steps).

Our data reveals an interesting relationship between promiscuity and self to non-self binding ratios especially for HLA-B. Promiscuous HLA-B variants clearly show a reduced preference towards human peptides while no such preference was observed for fastidious HLA-B alleles. These differences might be explained by T-cell selection in the thymus (Takaba & Takayanagi, 2017). Chappell et al. (2015) hypothesized that low cell surface expression of promiscuous MHC variants might be an adaptation to prevent excessive depletion of T-cells that recognize a wide variety of self-peptides presented by promiscuous MHC molecules in the thymus. Following the same reasoning, promiscuous variants presenting fewer human self-peptides might be preferentially maintained as they lead to less T-cell depletion. Such depletion would not be problematic in the case of fastidious alleles due to the already small number of self-peptides presented, and no selection pressure for a decreased self-binding would be observed for fastidious variants. On the other hand, the correlation between promiscuity and self to non-self binding ratio for HLA-A was stronger, suggesting that even fastidious HLA-A variants might be under selection to bind fewer self-peptides. It is possible that different selection pressures acting on HLA-A and HLA-B loci lead to such differences. Hertz et al. (2011) noted an increased binding preference of HLA-A alleles towards conserved human peptides compared to HLA-B alleles. Whether the promiscuity has a role in such specialization towards human peptides needs to be investigated further.

In summary, we report here the first systematic characterization of the vast diversity among pathogen peptidomes and provide support for the hypothesis that fastidious MHC variants can be maintained in populations by virtue of specialization towards one or few pathogens. However, the relationship between peptide binding promiscuity and specialization, and its role for MHC evolution is complex, and involves both qualitative and quantitative aspects of peptide

binding. Our approach based on computational binding prediction can only partly capture this complexity, and focuses predominantly on the quantitative aspects of this relationship. Nevertheless, our results yield intriguing insights into pathogen diversity and the evolution of peptide promiscuity, and provide a basis for further research into the nuances of pathogen-mediated selection on the antigen-presentation pathways.

Material and Methods

Selecting pathogen species and peptide data

The rationale behind the selection of pathogens used in this study was adopted from Pierini et al. 2018, following three main criteria: a global distribution of the pathogen, high mortality and/or morbidity (World Health Organization, 2018), and an impact on the human history (Wolfe et al., 2007). 36 pathogen species that likely had an important role in shaping the current diversity of human MHC genes were selected (Pierini & Lenz, 2018), including 10 viruses, 19 bacteria and 7 eukaryotic parasites. Reference proteomes of these pathogens as well as the reference proteome of Homo sapiens were downloaded from UniProt database (The UniProt Consortium, 2019). For the specific species and accession numbers see Table S1.

Calculation of peptide sharing and evolutionary distance values among pathogens

Although the peptide-binding groove of different MHC class-I molecules can accommodate varying lengths of peptides, the median length of eluted peptides from MHC class I molecules is 9 amino acids (Abelin et al., 2017; Ritz et al., 2016). The presented analyses are therefore based on nine-mer peptides. All possible nine-mers were obtained from pathogen proteins with a sliding window approach using a step size of one amino acid. Peptides containing ambiguous amino acid calls X, U and B were removed (N = 11,457; 0.022% of total peptides) resulting in 51,861,826 non-redundant nine-mers. Peptide sharing among pathogens was analyzed with two separate approaches. With the first approach, sharing of peptides among pathogens were analyzed with respect to the complete sequence of each nine-mer. The second approach focuses only on the nine-mers bound by at least one HLA class-I variant (N = 19,222,466). Each nine-mer was assigned a code representing the HLA class-I variants that binds to it. Nine-mers having the same code were grouped together and considered as the same from the perspective of HLA molecules as they are bound by the same set of HLA class-I variants. In total 4,157,475

such groups (codes) were formed. Sharing of peptides among pathogens were analyzed with respect to these groups.

Pairwise peptide sharing among pathogens was calculated either as the proportion of shared peptides within the combined peptidome of pathogen pairs or as absolute number of shared peptides. Peptide sharing with respect to evolutionary divergence was analyzed among 14 bacteria and 2 eukaryotic parasites that were common between the dataset used in this study and the tree of life (ToL) generated by (Ciccarelli et al., 2006) (Table S1). Evolutionary divergence between pairs of pathogens was calculated as tip-to-tip distances within the tree of life by using the ape package in R (Paradis & Schliep, 2019).

HLA variant data

Three classical human MHC class-I genes (HLA-A, -B and -C) were analyzed in this study. Although thousands of different alleles has been identified for each HLA loci (Robinson et al., 2020), most of these alleles are observed at very low frequencies or defined with limited documentation. Low frequency alleles are highly informative in some specific context such as organ transplantation (Kamoun et al., 2017) but their effect on recent human evolution is likely to be negligible (Robinson et al., 2017). In order to avoid biases that can be introduced by such alleles, two main criteria were applied on allele selection. Firstly, only alleles designated as "common" in the CIWD 3.0.0 catalogue were included in the analyses (Hurley et al., 2020). CIWD 3.0.0 catalogue classifies HLA alleles into categories based on their frequency. The "common" category of the CIWD catalogue covers the most frequent alleles in populations (those that are observed at a frequency of $\geq$0.01%). Secondly, in order to capture the functional diversity of MHC class-I genes while avoiding redundancy, P group designation of HLA alleles was considered. Alleles within a P group have identical peptide binding properties as they code for the same amino acid sequence across the antigen binding domain (Marsh et al., 2010a).

Computational prediction of peptide-binding

The set of potentially bound peptides for each given MHC allele in the study was estimated by using NetMHCpan(v4.1) (Reynisson et al., 2020). NetMHCpan is an established computational binding prediction algorithm that is trained on both experimental binding affinity and mass spectrometry derived eluted ligand data. Based on the training data, it can predict the binding between any MHC molecule and peptide either as an affinity value or as a percentile rank score compared to a set of natural peptides. Although previous analysis indicates that percentile rank

score performs better than the affinity score for identification of bound peptides, the percentile rank score assumes that all MHC alleles bind same number of peptides (Nielsen & Andreatta, 2016). As the main aim of this study was to analyze differences in the size of the peptide repertoire of MHC alleles, an affinity threshold of 50 nM and 500nM were used to define bound peptides. The affinity threshold of 500nM is widely considered as the limit of weak binding between an MHC allele and peptide, hence covering both strongly and weakly bound peptides (Paul et al., 2013), while the 50 nM threshold includes only strong binders. In order to avoid pseudoreplication, alleles having the exact same binding prediction results and the same first field number were identified and 22 HLA-A, 18 HLA-B and 12 HLA-C alleles were removed by keeping only the allele having the smallest second field number. Promiscuity of an MHC allele was defined as the fraction of peptides bound by the allele from the complete set of unique peptides. In order to test whether the computational promiscuity values are in agreement with experimental results, data from the Immune Epitope Database (IEDB) were used (Vita et al., 2019). The IEDB is a collection of experimental data on T-cell and antibody responses against or MHC binding of epitopes. Complete dataset of MHC ligand assays for HLA-A, -B and –C alleles were downloaded on 15 June 2020. Assays for which the source organism of the peptide is either Homo sapiens or unidentified were removed. Furthermore, only assays with alleles having 2nd field (four-digit) or higher resolution were used. Finally, HLA alleles having assay result for less than 1000 different peptides were removed, leaving 19 HLA-A and 15 HLA-B alleles for further analysis. No HLA-C allele met the criteria. Experimental promiscuity values were calculated as the fraction of positive binding assays among the total number of assays for each HLA allele. Kendall's rank correlation test was used to analyze the relationship between experimental and computational promiscuity values of individual alleles. Another dataset of experimental binding data was also compiled using the number of peptides eluted by mass-spectrometry for 9 HLA-A and 6 HLA-B from Abelin et al. (2017). However, due to the small number of alleles, this dataset was not used for correlations of individual alleles and used only to calculate overall experimental promiscuity for HLA-A and HLA-B locus.

Calculating phylogenetic distance between MHC alleles

Complete protein sequences of HLA class-I alleles were downloaded on 25 February 2019 from IPD-IMGT/HLA Database (Robinson et al., 2020) and aligned with ClustalW software implemented in MEGA-X (Kumar et al., 2018). Positions that correspond to the peptide-binding region of HLA proteins were removed as these positions are under positive selection (Hughes & Hughes, 1995) and also most likely involved in direct interaction with the peptide,

thus defining the peptide binding properties of the HLA variant. A phylogenetic tree was built separately for each HLA class-I locus using the Maximum Likelihood method with Jones-Taylor-Thornton substitution model to calculate amino acid distances and 1000 bootstrap replicates to quantify support of nodes. Evolutionary distance between alleles was calculated as tip-to-tip distances in phylogenetic trees using ape package in R (Paradis & Schliep, 2019).

Calculating pathogen specialization of MHC variants

Quantitative differences among HLA variants with regard to binding peptides from distinct pathogens were analyzed. In order to allow unbiased comparison of variants with different promiscuity levels, fractions of bound peptides from each pathogen were standardized by converting them to z-scores within each variant. Without this normalization, variants that are more promiscuous would automatically have a higher variance of their relative peptide binding values among the different pathogens, which would bias the specialization analysis. For each variant, a specialization value was then calculated as the difference between the maximum and the median z-score. The rationale here is that a variant that is specialized to bind peptides of a specific pathogen particularly well should show a particularly high difference between the fraction of bound peptides from this pathogen and the fraction of all other pathogens (reflected by the median). A potential correlation between the specialization values and the promiscuity levels of the variants was tested using Kendall's rank correlation. In order to verify that the obtained results were not driven by random fluctuations or any methodological bias in the binding data, simulations were performed. For these simulations, an HLA variant's overall promiscuity level (fraction of peptides bound from the total number of peptides) is used as its probability of binding a peptide from a given pathogen. This probability was then used to randomly sample peptides from each pathogen peptidome and thus simulate the fraction of bound peptides from each pathogen under a no-specialization scenario. This Monte Carlo simulation approach was applied to all HLA loci separately. The difference to the real data was only that the fraction of peptide bound from each pathogen now reflected the overall promiscuity of the allele and not the pathogen-specific promiscuity. By using the same scaling approach, specialization value calculations and correlation test were then also applied to the simulated data.

Author Contributions

O.Ö. and T.L.L. designed research; O.Ö. performed research and analyzed the data; O.Ö. and T.L.L. interpreted the data and wrote the manuscript.

Acknowledgements

We thank Jim Kaufman for insightful comments on a previous version of the manuscript. This work was supported by grants from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, grant numbers 279645989, 437857095) to T.L.L.

Supplementary Material for Chapter 1 is provided in Annex III.

Chapter 2

Balancing selection rather than local adaptation determines HLA gene variation in ethnically diverse African populations

Onur Özer[1]*, Daniel Harris[2]*, Michael McQuillan[2]*, Tristan Hayeck[3,4], Eric Mbunwe[2], Timothy L. Mosbruger[3], Jamie L. Duke[3], Clinton Azuure[1], Tzun-Wen Shaw[3], Thomas Nyambo[5], Sununguko Wata Mpoloka[6], Gaonyadiwe George Mokone[7], Gurja Belay[8], Charles Fokunang[9], Alfred K. Njamnshi[10], Martin Maiers[11], Dimitri Monos[3,4#], Tobias L. Lenz[1#], Sarah Tishkoff[2#]

[1] Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany

[2] Department of Genetics, University of Pennsylvania, Philadelphia, PA

[3] Immunogenetics Laboratory, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[5] Department of Biochemistry, Kampala International University in Tanzania, P.O. Box 9790, Dar es Salaam, Tanzania.

[6] Department of Biological Sciences, Faculty of Science, University of Botswana Gaborone, Private Bag UB 0022, Gaborone, Botswana.

[7] Department of Biomedical Sciences, Faculty of Medicine, University of Botswana Gaborone, Private Bag UB 0022, Gaborone, Botswana.

[8] Department of Microbial Cellular and Molecular Biology, Addis Ababa University, Ethiopia

[9] Department of Pharmacotoxicology and Pharmacokinetics, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, P.O. Box 337, Yaoundé, Cameroon.

[10] Department of Neurology, Central Hospital Yaoundé; Brain Research Africa Initiative (BRAIN), Neuroscience Lab, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, P.O. Box 337, Yaoundé, Cameroon.

[11] CIBMTR (Center for International Blood and Marrow Transplant Research), National Marrow Donor Program/Be The Match, Minneapolis, MN, USA

* Shared first authors

[#] Shared senior and corresponding authors

Abstract

Ethnically diverse groups in Sub-Saharan Africa represent some of the greatest genetic variation in humans, while at the same time being severely underrepresented in human genetic studies. Genes of the human leukocyte antigen (HLA) are of particular importance because of their critical role in adaptive immunity, their association with a wide range of diseases, and their exceptional polymorphism across human populations. Here we investigate HLA genetic variation relative to genome-wide variation in a unique dataset from 12 ethnically diverse groups from Sub-Saharan Africa, encompassing both whole-genome sequencing and targeted HLA sequencing. An excessive level of diversity and ancient polymorphism, including several novel HLA alleles, indicated long-term balancing selection on HLA genes throughout human evolution. Among the tested populations, HLA genes did not show higher genetic differentiation than genome-wide data, indicating that ancient polymorphism and demography rather than local adaptation are the main drivers of HLA variation across modern human populations.

Introduction

African populations exhibit high levels of genetic variation within and among populations relative to non-African populations, yet they remain greatly under-represented in genetic studies (Tishkoff et al., 2009). It remains a matter of intense investigation to better understand to what extent this variation is attributable to adaptive processes or merely reflects historical migrations and demography. Current archaeological and genetic evidence indicates that anatomically modern humans originated in sub-Saharan Africa around 300,000 years ago (Hublin et al., 2017; Scerri et al., 2018). The majority of modern human evolutionary history was restricted to sub-Saharan Africa, until approximately 60,000 to 100,000 years ago when the out-of-Africa expansion occurred (Malaspinas et al., 2016; R. Nielsen et al., 2017; Pagani et al., 2016). This initial bottleneck along with subsequent migration farther away from Africa left a signature of decreasing genetic diversity in populations as a result of serial founder effects (Li et al., 2008; Prugnolle, Manica, & Balloux, 2005). On the other hand, novel selective pressures resulting from rapid colonization of regions with different environments, climates, and pathogens led to local adaptations in migrant populations (Fan et al., 2016; Rees et al., 2020). The unique evolutionary history of each population often results in marked differences in the frequency of alleles and such differences between populations are important for understanding differences in disease risk (Benton et al., 2021; Corona et al., 2013; Kim et al., 2018). In fact, pathogens are considered a major, if not the main, driver of local adaptation, potentially leading to significant differentiation in immune-related genes among populations (Brinkworth & Barreiro, 2014; Cooke & Hill, 2001; Fumagalli et al., 2011; Nédélec et al., 2016). Among the immune genes with signatures of natural selection, the HLA (human leukocyte antigen) class I and class II genes, located within the MHC (major histocompatibility complex) region on chromosome 6, particularly stand out due to their established diversity and importance for adaptive immunity (Radwan et al., 2020; Tian et al., 2017; Trowsdale & Knight, 2013). During infection, HLA molecules can present pathogen-derived peptides at the cell surface for recognition by T-lymphocytes, triggering an antigen-specific immune response. Given this key role in adaptive immunity, continuous exposure to diverse pathogens is postulated to be the major selective pressure maintaining the high diversity of HLA genes (Radwan et al., 2020; Sommer, 2005). However, the contribution of such pathogen-driven selection for population genetic variation is a matter of intense research as different mechanisms of pathogen-driven selection, such as long-term balancing selection or local adaptation, can lead to contrasting outcomes (Brandt et al., 2018; Spurgin & Richardson, 2010). While reduced population differentiation compared to

neutral markers is expected for loci under long-term balancing selection, local adaptation driven by specific associations between MHC variants and pathogens would increase population differentiation as long as pathogen communities vary among populations (Meyer et al., 2018; Schierup et al., 2000). Moreover, a population's demographic history could have substantial effects on HLA diversity, to the extent that it remains an open question whether the well-documented differences in HLA allele frequencies among human populations reflect local adaptation or simply demography (Prugnolle, Manica, Charpentier, et al., 2005; Sanchez-Mazas et al., 2012). For these reasons, combining HLA genotype data with information from suitable neutral markers is paramount to assessing the role of selective mechanisms and demographic events shaping HLA diversity, for instance by comparing SNPs within and outside the HLA genes (Brandt et al., 2018).

In humans, HLA molecules are encoded by HLA class-I (*HLA-A*, *HLA-B* and *HLA-C*) and HLA class-II genes (*HLA-DRA*, *-DRB1*, *-DQA1*, *-DQB1*, *-DPA1* and *-DPB1*). These classical HLA genes are known for their polymorphism and for harboring an elevated number of intermediate frequency variants across human populations, indicative of some form of balancing selection (Robinson et al., 2017). The unique peptide-binding properties of a given HLA molecule, i.e. the actual HLA phenotype under pathogen-mediated selection, are encoded by SNP variation along the entire sequence of a given HLA gene, defining a specific 'HLA allele'. These HLA alleles thus represent haplotypes of a given HLA gene with a unique combination of SNP variants, often differing in tens of SNPs from other alleles. Nevertheless, different HLA alleles may also share some SNP variants despite encoding very different peptide-binding properties (Maróstica et al., 2022). Therefore, it is important to complement SNP-based approaches with targeted HLA gene sequencing and allele-based analysis in order to obtain a comprehensive understanding of HLA evolutionary history.

Despite the importance of Africa in recent human evolution, populations currently living within the continent are among the least investigated in evolutionary and immunogenomics studies (Bentley et al., 2017; Hindorff et al., 2018; Peng et al., 2021; Sirugo et al., 2019). The Allele Frequency Net Database (AFND) is the most comprehensive and widely used repository for publicly available frequency data for HLA genes and several other immune-related genes (Gonzalez-Galarza et al., 2020). As of May 2023, among the 1,306 populations available in the AFND for which HLA allele frequency information is available, only 68 are from sub-Saharan Africa (5.2%), all with generally smaller sample sizes compared to populations from Europe or North America. Given that African populations harbor the highest levels of genetic diversity

and are highly underrepresented in immunogenomics research, it is not surprising that studies of African populations often report distinct allele pools and sometimes novel HLA alleles (Cao et al., 2004; Nemat-Gorgani et al., 2019; Pagkrati et al., 2023; Paximadis et al., 2012; Thorstenson et al., 2018; Tishkoff et al., 2009). The World Health Organization reports that communicable diseases that have associations with HLA polymorphism (i.e. HIV/AIDS, malaria and tuberculosis) are major sources of mortality and morbidity in Africa (Garamszegi, 2014; Kløverpris et al., 2016; Oliveira-Cortez et al., 2016; World Health Organization, 2021). Therefore, a better characterization of HLA diversity in African populations will not only improve our understanding of the evolutionary forces acting on HLA genes but also facilitate more inclusive medical applications of genomics. Here we present an in-depth characterization of genetic diversity and signatures of selection in the MHC region in comparison to genome-wide levels, based on a novel panel of ethnically diverse human populations from sub-Saharan Africa that represent some of the most genetically diverse populations in present-day humans. Here we combine SNP information from high-coverage whole genome sequencing data (also covering the entire MHC region) with data from targeted sequencing of the classical HLA genes in these diverse African populations to make inferences about the evolutionary forces influencing variation at the HLA loci.

Results

Whole Genome Sequencing and HLA dataset description

We analyzed a genomic dataset from 12 populations that represent a wide range of sub-Saharan African genetic, linguistic, and environmental diversity (**Figure 1**). These populations speak languages belonging to the main African language families (Khoesan, Nilo-Saharan, Niger-Congo, and Afroasiatic) and practice different traditional subsistence patterns (agriculturalist, hunter-gatherer, or pastoralist; **Supplementary Table S1**). Our dataset includes two main types of genetic data: 1) high coverage (~30X) whole genome sequencing (WGS) data from 180 individuals, encompassing 33.6 million single nucleotide polymorphisms (SNPs) (Fan et al., 2023) and 2) HLA class I and class II genotype data from targeted sequencing of 489 individuals (including the 180 individuals with WGS data) (Pagkrati et al., 2023). Therefore, we can not only study diversity within the entire MHC region and compare it to genome-wide levels but also the fine-scale diversity within the classical HLA genes of the same populations. To compare our samples to other African and also non-African groups, we combined our data with

1,014 individuals from the 1000 Genomes (1KG) project that have both WGS and HLA targeted sequencing data (Abi-Rached et al., 2018; Auton et al., 2015), taking the intersection of the variants. This resulted in a merged dataset with 21,201,602 SNPs for all analyses that include the 1KG data.

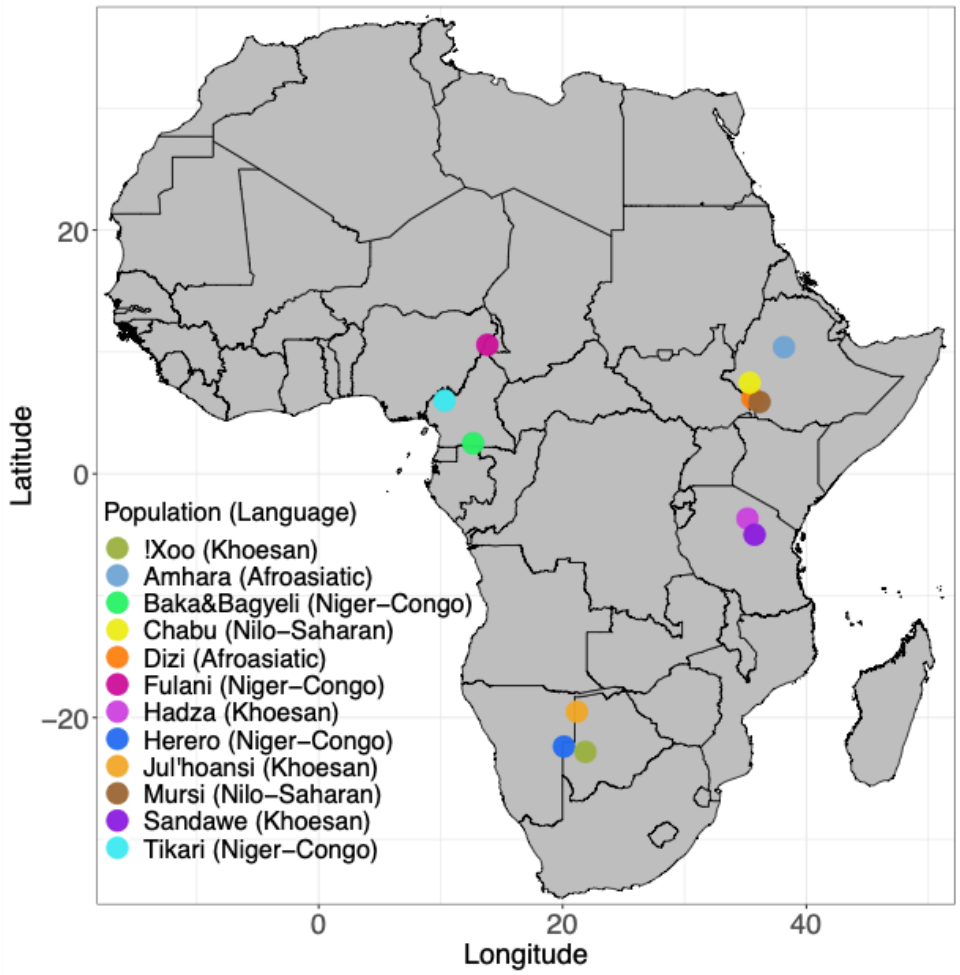


Figure 1. Map of sampling locations for the 12 populations from sub-Saharan Africa for which novel WGS (N=180) and targeted HLA sequencing (N= 489) data was analyzed.

HLA genes exhibit high levels of genetic diversity and signatures of balancing selection

Using the WGS data, we evaluated the genetic diversity and selective forces acting on SNPs within the MHC region compared to genome-wide patterns. First, we calculated nucleotide diversity (pi), i.e. the average number of pairwise differences, in 1kb windows across the genome, grouping all 180 Africans from our high coverage dataset together. Not only do we

find that the MHC region (here defined as chr6:29,000,000-34,000,000) shows the most extreme levels of nucleotide diversity compared to the rest of the genome (**Figure 2a**), but we also observe that these extreme peaks in nucleotide diversity overlap almost perfectly with the positions of the classical HLA class-I and class-II genes (**Figure 3**). We also sought to compare levels of nucleotide diversity at the MHC region to the rest of chromosome 6 for each of the African populations separately. We find that nucleotide diversity within the MHC region is higher than across the chromosome 6 background for each African population (**Figure 2c**). Noteworthy, some populations that have experienced recent bottlenecks, like the Hadza and Chabu hunter-gatherers (Fan et al., 2023; Gopalan et al., 2022; Henn et al., 2011), have a markedly lower nucleotide diversity at the MHC region compared to other populations with no history of recent bottlenecks (**Figure 2c**). Further, we compared MHC diversity in our African WGS data to other African and non-African populations from the 1KG dataset. To do this, we merged the two datasets and calculated Pi on a per-site basis for each population independently (22 populations total). We find that nucleotide diversity exhibits a drastic decrease for the chromosome 6 background in non-African populations compared to Africans, in line with the bottleneck from the out-of-Africa expansion (**Supplementary Figure S1**). We also see a decrease in MHC nucleotide diversity in non-Africans compared to Africans. However, this reduction is significantly less severe than for the chromosome 6 background (**Supplementary Figure S2**), suggesting that balancing selection not only maintained MHC diversity within Africa but also buffered the out-of-Africa bottleneck and maintained high MHC diversity in global populations. In addition, we calculated heterozygosity from the SNP data in each African and non-African population, both for the MHC region and for the remainder of the genome, and see again much higher values for the MHC region, similar to the nucleotide diversity results above (**Supplementary Figure S3**).

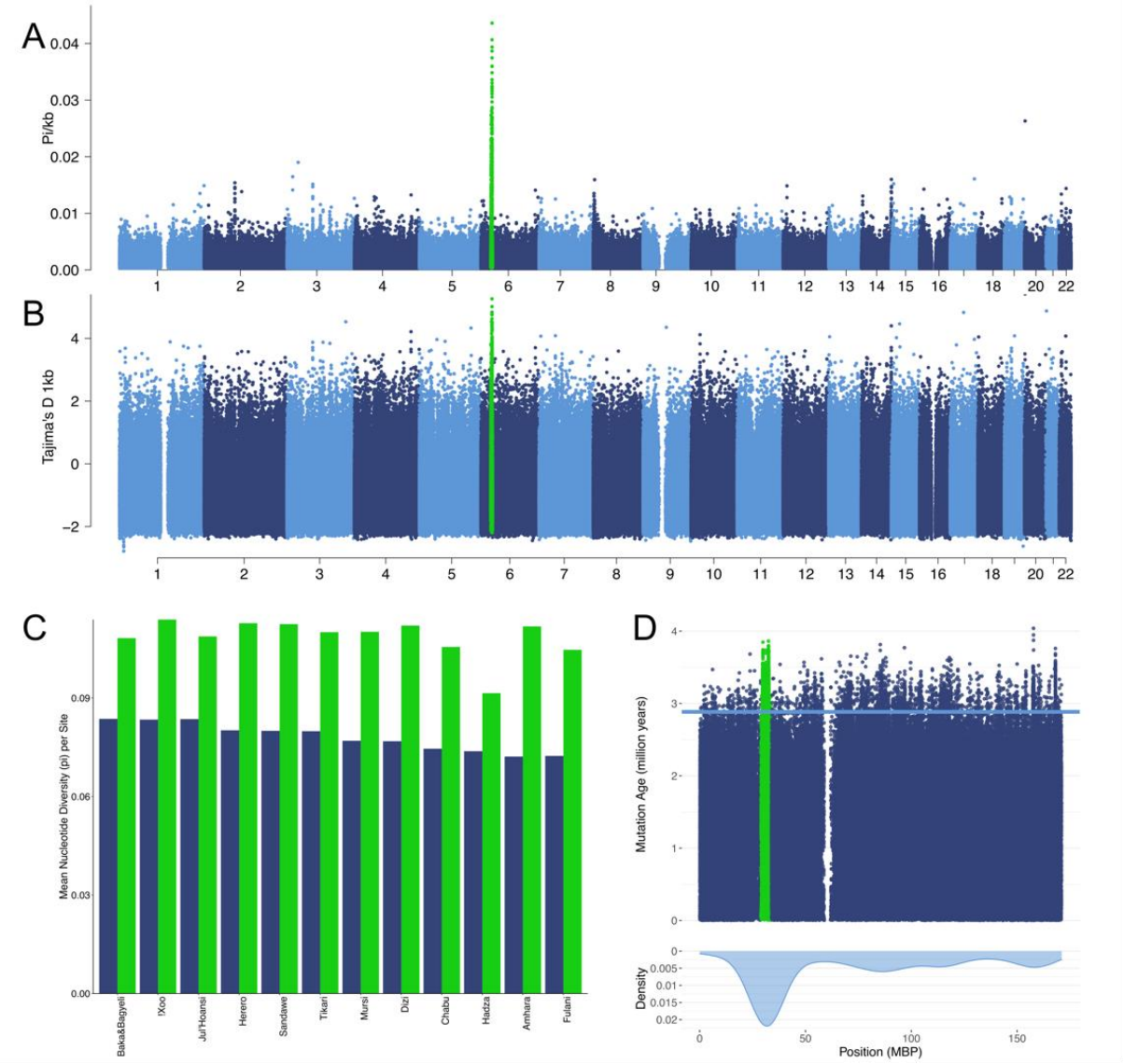


Figure 2. Patterns of diversity at the MHC region in Africans. A) Nucleotide Diversity (pi) across the genome, calculated from SNP data in 1kb windows with >20 SNPs, grouping all 180 Africans with WGS data together. Windows within the MHC region (chr6:29mb-34mb) are highlighted in green. B) Tajima's D calculated in 1kb windows from SNP data of all 180 Africans. Windows within the MHC are highlighted in green. C) Nucleotide Diversity (pi) calculated per site for each population, averaged over the MHC region (chr6:29mb-34mb) and the remainder of chromosome 6. Data were LD pruned (583,063 chr6 SNPs vs 23,179 HLA-region SNPs). 1KG populations were down-sampled to N=15/population. D) Age of Mutations across chromosome 6. Top part shows manhattan plot of mutation ages (minor allele frequency >= 5%) across chromosome 6 with the MHC region highlighted in red. The blue line represents the 99th percentile of mutation age across chromosome 6 (2,983,955 years). Bottom part shows density of mutations across chromosome 6 with an age that is >= to the 99th percentile of mutation ages. The MHC region has a high density of old mutations, which is consistent with balancing selection.

In order to explore the potential impact of balancing selection on the observed patterns of diversity in the MHC region, we calculated the Tajima's D statistic (Tajima, 1989) in 1kb windows genome-wide, grouping all 180 Africans from the WGS dataset together. Briefly, Tajima's D statistic estimates the difference in two measures of diversity: the average number of pairwise differences (pi) and the number of segregating sites. If a DNA sequence is evolving neutrally, these measures should be close to equal and Tajima's D will have a value close to zero. Large positive values of Tajima's D indicate an excess of moderate frequency polymorphisms, which is consistent with either balancing selection or a recent population contraction. In contrast, large negative values of Tajima's D indicate an excess of rare variants, suggestive of recent positive selection or a population expansion. We find that the MHC region in the African populations shows extreme positive values of Tajima's D compared to the rest of the genome (**Figure 2b and 3**), suggestive of strong balancing selection acting in this genomic region. When we calculate Tajima's D in each African population separately, we see a similar pattern, in that the highest Tajima's D values overlap with the positions of the class-I and class-II HLA genes (**Supplementary Figure S4**).

As balancing selection maintains genetic polymorphism, potentially over long periods of time, it is generally associated with older mutations. Therefore, we hypothesized that the MHC region will be enriched for exceptionally old derived mutations, compared to the rest of the genome. We thus estimated the age of all derived mutations across chromosome 6 with a minor allele frequency ≥ 0.05, using the Genealogical Estimation of Variant Age (GEVA) method (Albers & McVean, 2020). GEVA uses WGS data to construct the genealogical relationship of haplotypes with the derived allele compared to haplotypes with the ancestral allele and then employs a composite likelihood framework to estimate the posterior distribution of the derived mutation age. As many SNPs in the HLA genes have undefined ancestral and derived alleles, we estimated the age of both alleles and then conservatively selected the youngest allele as the derived one. The average age of derived mutations outside the MHC across chromosome 6 is 1.07 Mio years (95% CI: 1.06-1.07), while derived mutations in the MHC region have an average age of 1.27 Mio years (1.26-1.28). This difference is highly significant, due to the exaggerated tail of many older mutations in the MHC region (MWU-test, $P=1.59 \times 10^{-83}$). Indeed, the MHC region contains ~42% of all SNPs in the 99th percentile of derived mutation age, while representing only ~6 % of all SNPs on chromosome 6 (**Figure 2d**). Within the MHC region, the variants with an age in the 99th percentile (2,983,955 years or older) cluster into three regions, which align with the classical HLA class-I and class-II genes (**Supplementary**

Figure S5). Taking the average mutation age estimate for 1kb windows across the MHC region yields three main peaks of mutation ages that are centered around the HLA genes, a similar pattern as seen for the metrics on balancing selection (**Figure 3**).

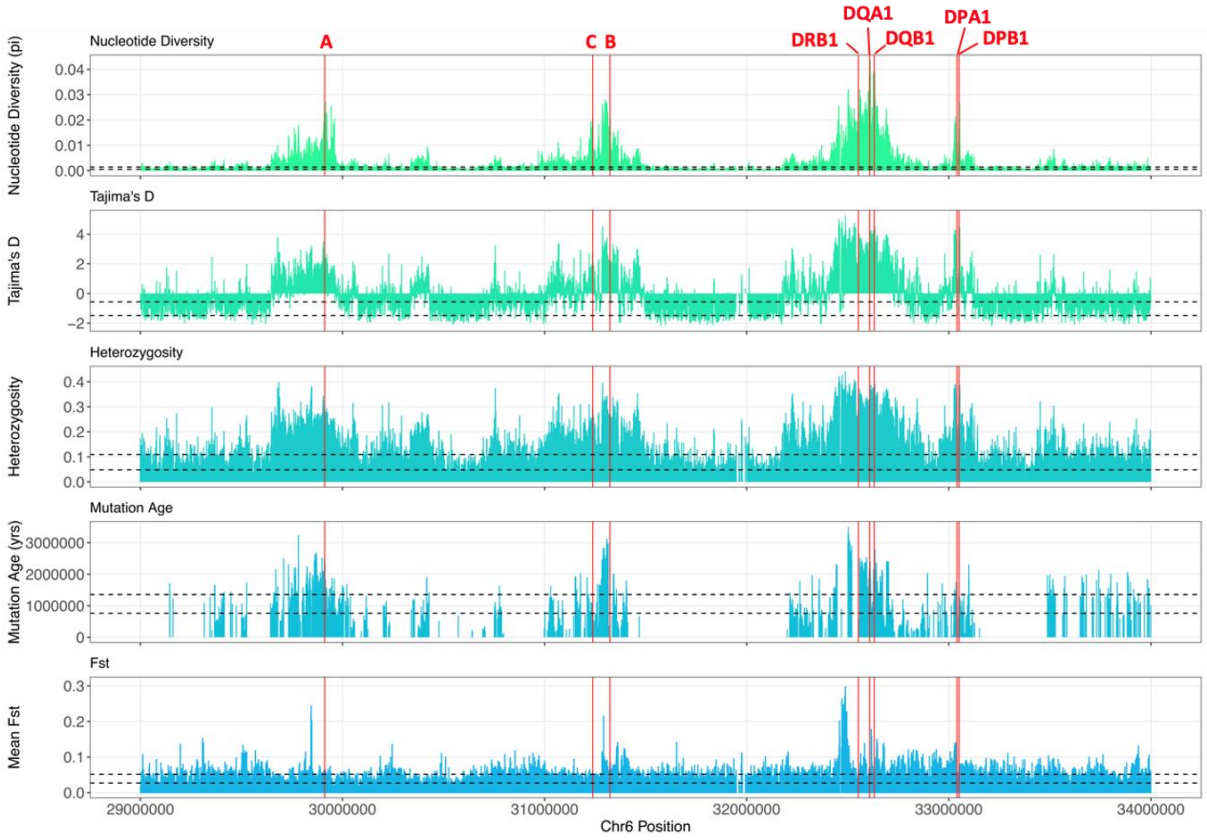


Figure 3. Summary of different measures of diversity/selection across the HLA region (chr6:29MB-34MB). Pi and Tajima's D were calculated in 1kb windows. Heterozygosity, mutation age, and F_{ST} were calculated per site, and then averaged over non-overlapping 1kb windows. Plots for all statistics show the value for each 1kb window across the HLA region. All statistics were calculated from SNP data of the African 180 WGS dataset, and vertical red lines denote the positions of class I and II HLA genes. Horizontal dashed lines indicate the 25th and 75th percentile of all 1kb windows across chromosome 6 for each statistic. Zero values for mutation age estimates reflect sites where the ancestral/derived alleles could not be confidently established.

HLA diversity does not cluster based on geographic and linguistic similarity of populations

Based on the excessive diversity and strong signatures of balancing selection observed for the MHC region in general and the HLA genes in particular, we then asked whether this genetic diversity showed population-specific clustering, as would be expected from HLA-mediated local adaptation to geographically restricted pathogens. For this, we calculated the fixation

index F_{ST} for each SNP on chromosome 6, a metric to estimate population differentiation based on relative differences in allele frequencies. First, we did this analysis on a merged dataset containing all 22 populations (12 from the African WGS dataset and 10 from the 1KG dataset). We found that, compared to the rest of chromosome 6, the MHC region does not show exceptionally high F_{ST} values (**Figure 4c**; **Supplementary Table S2**), a pattern held even when zooming to the level of the classical HLA genes (**Figure 3**). No SNPs in the class-I or class-II genes have F_{ST} values that exceed the top 0.1% of F_{ST} values across chromosome 6. In fact, when averaging F_{ST} values for each class I and class II gene, only HLA-DPA1 and HLA-DPB1 have significantly higher mean F_{ST} values than the chromosome 6 background (**Supplementary Figure S6**). This lack of elevated F_{ST} at most HLA loci suggests very limited population differentiation of HLA diversity, contrasting to what would be expected from local adaptation to divergent pathogen environments. When calculating pairwise F_{ST} between all 22 populations in the dataset, both genome-wide and at the MHC region, we find that the MHC region exhibits relatively higher differentiation only in pairwise comparisons involving historically bottlenecked populations, like the Hadza hunter-gatherers from Tanzania (Gopalan et al., 2022), or populations with evidence of inbreeding, such as the Fulani pastoralists from Cameroon (Pemberton & Rosenberg, 2014) (**Supplementary Figure S7**). In addition, most African populations show lower F_{ST} in the MHC region compared to the genome-wide average in pairwise comparisons involving non-African populations. This is in line with previous work, showing that population pairs from the same continent often have higher relative F_{ST} in the MHC region than pairs of populations from different continents (Brandt et al., 2018; Meyer et al., 2018), driven by long-term balancing selection that maintains shared polymorphism and thus counteracts isolation-by-distance effects.

In order to further explore potential genetic differentiation (or lack thereof) among the different populations at the MHC compared to genome-wide levels, we then employed principal component analysis (PCA). In line with previous studies, a PCA based on SNPs from across the whole genome demonstrates a clear clustering of populations based on geographic and linguistic similarity (**Figure 4a**). African and non-African populations differentiate across PC1, with some African populations clustering closer to the non-African ones due to recent admixture (Fan et al., 2023; McQuillan et al., 2022). The Khoisan populations (!Xoo and Ju'|Hoansi) separate from all other African populations along PC2. We observe the West African Niger-Congo speaking groups from the 1KG project clustering near the Tikari from Cameroon, who also speak a Niger-Congo language. Many of the East African populations

(Hadza, Sandawe, Mursi, Dizi, and Chabu) also cluster within PC space. In contrast, PCA with SNPs from only the classical HLA class-I and HLA class-II genes (HLA-SNPs) does not demonstrate clustering by geography or language groups (**Figure 4b**). Indeed, HLA-SNPs do not show any clear signature of the out-of-Africa expansion or recent admixture among samples either. Instead, we observe multiple clusters of individuals that originate from different populations and across continents, possibly reflecting distinct genotype combinations of common HLA alleles that are shared across populations (see below for results from the HLA-targeted sequencing data that support this hypothesis).

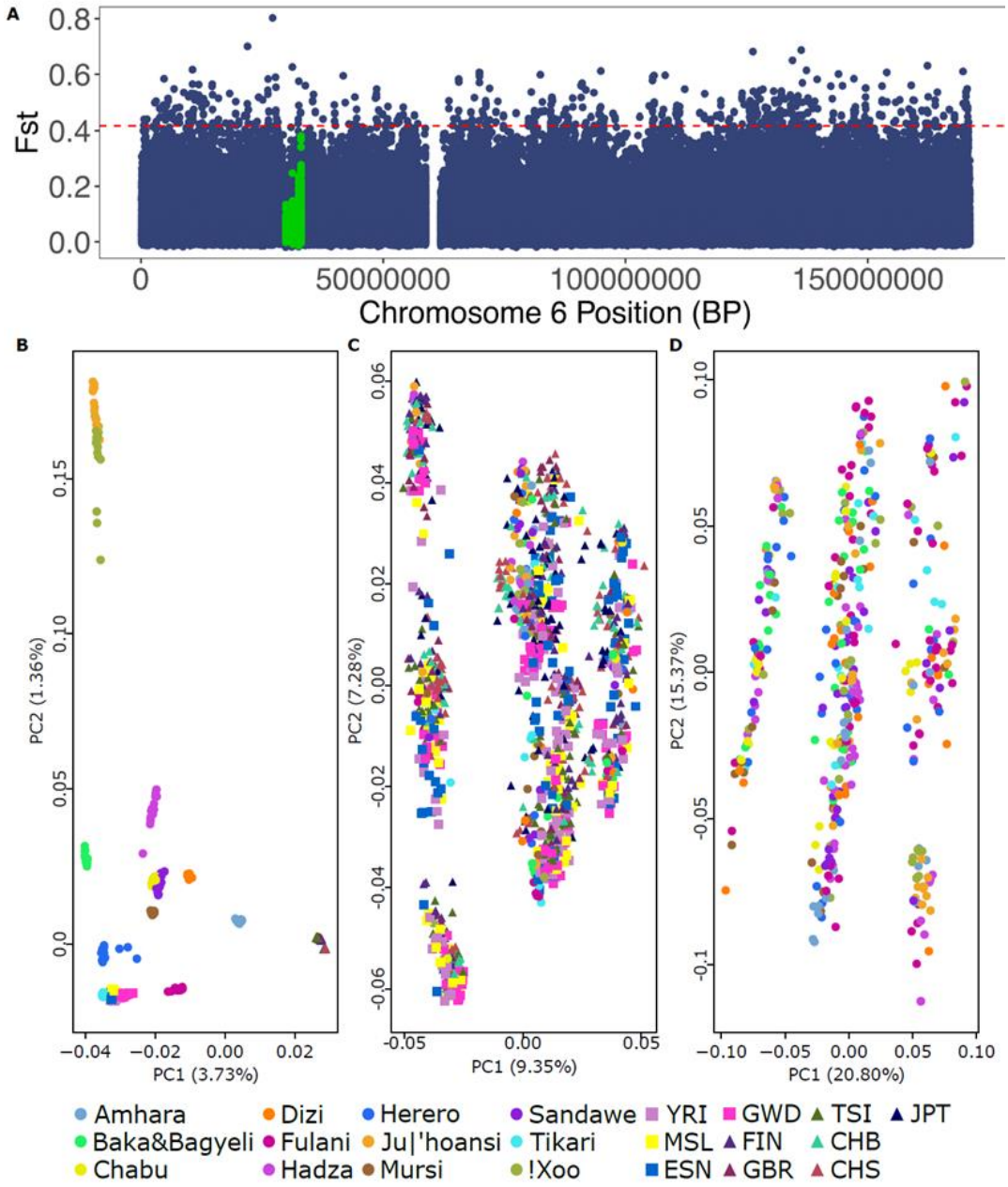


Figure 4. Genome-wide population structure vs structure at the HLA genes. A) PCA using genome-wide SNPs and B) PCA using SNPs only from the HLA Class I and Class II genes. In PC space calculated from genome-wide SNPs, individuals cluster based on population, geographic, and linguistic similarity. In contrast, individuals do not cluster based on linguistic or geographic similarity when only SNPs from the HLA Class I and Class II genes are used to calculate PCs. Circles represent the 12 sub-Saharan African populations from this study. Squares and triangles represent African and non-African populations from the 1KG dataset, respectively. C) Weir and Cockerham (1984) F_{ST} calculated for each SNP on chromosome 6. Analysis contains 22 populations total (12 from the African 180WGS dataset, 10 from the 1KG dataset). 1KG populations were down-sampled to N=15/population and data were LD-pruned. Red dashed line indicates the 99.9th percentile F_{ST} on chr6. SNPs contained within the class I and class II genes are highlighted in red.

Instead of indicating low population differentiation, the lack of geographic and linguistic clustering of individuals based on the HLA-SNP data could also be explained by the small number of SNPs (N=902) or the small genomic region (52,584 bp) represented by this dataset. We therefore tested whether random intergenic regions of similar size could identify geographic and linguistic clustering better than the PCA based on HLA-SNPs by using discriminant analysis of principal components (DAPC) that groups individuals based on the individuals' PCA coordinates and the centroid PCA coordinates of prior defined groups (Jombart et al., 2010). We used DAPC to assess the clustering of individuals into groups at the population (e.g. Amhara vs. Chabu), linguistic (e.g. Afroasiatic vs. Nilo-Saharan), and continental region (e.g. African vs. European) level. Among 100 random genomic regions containing the same number of SNPs as the HLA-SNPs (N=902), DAPC correctly assigned an average of 26% of individuals to their population label (**Supplementary Figure S8**), an average of 74% to their linguistic labels (**Supplementary Figure S9**), and an average of 82% to their continental region labels (**Supplementary Figure S10**). These values decreased to 21%, 63%, and 71%, respectively, when we used the same physical region size (52,584 bp) as the HLA genes (irrespective of SNP number in that region). In contrast, DAPC based on the HLA-SNPs assigned correctly only 16% for population labels, 44% for linguistic labels, and 50% for continental region labels, which was significantly lower than from the random regions with the same SNP number in all three cases and close to significant for random regions of the same size (**Supplementary Figures S8-S10**). Therefore, random intergenic regions of the same size or SNP number can differentiate geographic and linguistic structure in our dataset more effectively than the SNP data at the HLA genes. These analyses suggest that the relatively smaller number of SNPs in the HLA-SNP data is not sufficient to explain the lack of geographic and linguistic clustering in PC space. Instead, the lack of clustering is consistent with long term balancing selection maintaining common HLA allelic diversity in populations across the globe.

Analysis of targeted HLA genotype data

The actual 'HLA phenotype' targeted by the pathogen-mediated selection, i.e. the ability of an HLA molecule to present relevant antigens of a given pathogen, is defined by the specific combination of SNPs along the corresponding HLA gene, which is referred to as a classical 'HLA allele'. However, reliable identification of an individual's HLA alleles, i.e. its HLA genotype, is challenging from SNP data alone and is thus usually achieved by targeted HLA

sequencing. In order to reconcile the findings from the SNP-based analyses above with classical HLA allele information, we generated targeted HLA genotype data by locus-specific long-range amplicon sequencing (from hereon called 'HLA-target data') for 489 individuals from the 12 focal African populations.

In order to capture the full extent of genetic variation at the HLA genes, we first analyzed the HLA-target data at up to 4th field level of resolution, which includes variation in both exons and introns of a given HLA gene. A full report of the detected HLA alleles is provided in Pagkrati et al. (Pagkrati et al., 2023). Briefly, among the 489 individuals, a total of 694 HLA alleles were detected across the classical HLA class I loci HLA-A (70 alleles), HLA-B (76), HLA-C (68) and class II loci HLA-DRB1 (73), HLA-DQA1 (99), HLA-DQB1 (69), HLA-DPA1 (92), and HLA-DPB1 (147). Among these, 130 represented novel alleles that had not been previously reported in the IPD-IMGT/HLA database, with 11 of these representing non-synonymous exonic variants encoding for novel HLA proteins (Pagkrati et al., 2023). We then explored which alleles were unique to the different populations, first looking at populations grouped by language and then each individual population based on self-identified ethnic group. We identified a number of population-specific alleles above 10% frequency when examining populations pooled based on language group, up to the second field of resolution, including HLA-B*15:18 (10.1% frequency) and HLA-C*07:607 (11.6% frequency) in the Nilo-Saharan-speaking populations, HLA-DPB1*04:02 (11.6% frequency) in the Niger Congo-speaking populations, and HLA-DPA1*02:09 (12.6% frequency) and HLA-DPB1*763:01 (10.2% frequency) in the Khoesan-speaking populations. Population specific alleles that were unique to different language groups up to 4th field resolution with frequency greater than 10% include HLA-A*02:05:01:21 (10.9%), HLA-B*15:18:01:01 (10.1%) and HLA-C*07:607 (11.6%) in the Nilo-Saharan-speaking populations, HLA-DPA1*01:03:01:05 (11.6%) in the Niger Congo-speaking populations, and HLA-DRB1*04:01:01:01 (19.3% frequency) in the Khoesan-speaking populations. When focusing on individual ethnic groups, the Chabu exhibited two unique alleles above 15% frequency at the second and the fourth field, HLA-C*07:607 at a frequency of 20.5% and HLA-B*15:18:01:01 at a frequency of 18.0%. We also identified several novel alleles that were common in the Hadza including HLA-DPB1* 55:01:02:NEW (9.6%) and HLA-DRB1*13:16:01:NEW (8.3%) (Pagkrati et al., 2023).

We then applied PCA to the HLA-target data by converting the classical HLA allele sequences into SNP data. In agreement with the PCA shown above, based on the HLA-SNP data from whole-genome sequencing, the PCA on the HLA-target data did not show any population-

specific clustering (**Figure 4D**). Instead, this data revealed that the individuals across populations clustered according to distinct HLA-DQ haplotypes that correspond to specific DQ serotypes (**Supplemental Figure 11**), suggesting again that the genetic polymorphism at the HLA genes reflects long-term balancing selection that maintains this polymorphism across populations, rather than being the result of local adaptation.

In order to focus on functional HLA variation, the HLA-target data was then analyzed at 2nd field resolution. This level of resolution distinguishes HLA protein variants with potentially different antigen repertoires. **Supplementary Table 3** summarizes several parameters for the HLA-target data per locus, such as the number of samples, number of 2nd field alleles, observed heterozygosity and allelic richness per population. Despite some variation among individual populations, heterozygosity and allelic richness (a measure of allele pool diversity) were overall similar among continents and did not exhibit a significant decline out of Africa, except for *HLA-A*, which showed decreased allelic richness in European and East Asian populations (**Figure 5 A/B**, **Supplementary Figure S12**).

We calculated pairwise F_{ST} values to analyze genetic differentiation among populations based on HLA class-I and HLA class-II alleles, finding only weak structuring among the groups (**Supplementary Figure S13**). Principal coordinate analysis of pairwise F_{ST} values showed East Asian and African populations occupying two extreme ends of the spectrum and European populations positioned in between (**Supplementary Figure S14**). In order to analyze these patterns of differentiation among continents, we identified HLA alleles with large frequency differences ('LFD alleles') between African and non-African populations (i.e. alleles that are at least three times more common in Africa compared to other continents) (Single et al., 2020). These African LFD alleles made up a significant proportion of the allele pools of most African populations for *HLA-A* (ranging from 36% to 79%) and *HLA-B* (ranging from 43% to 83%) genes, relatively less so for *HLA-C* (ranging from 24% to 64%) and *HLA-DRB1* (ranging from 18% to 59%), and hardly for *HLA-DQB1* (ranging from 0% to 22%) (**Supplementary Figure S15**). However, despite the substantial variation in allele frequencies, especially in *HLA-A* and *HLA-B* genes, grouping of alleles with similar functional properties into HLA supertypes (Sidney et al., 2008) revealed that all populations harbor almost every HLA supertype (**Figure 5c**, **Supplementary Figure S16**).

Figure 5. Analyses of targeted classical HLA sequencing data from the 12 novel African populations and 10 1KG populations. Allelic richness for HLA-B (A) and HLA-DRB1 (B) loci is similar across populations. Decreased allelic richness in the Chabu population is likely the result of a recent bottleneck. C) Distribution of HLA-B supertypes across populations. D) Heatmap showing the distribution of the *normalized deviate of homozygosity* (F_{nd}) values. Negative values (blue) imply balancing selection while positive values (orange) imply directional selection. (*) indicates nominally significant values. No p-value remains significant after Bonferroni correction for multiple testing. E) Haplotype network of HLA-DRB1 alleles. Ju|'hoansi and !Xoo individuals were combined into the Khoesan group. All European and East Asian 1KG individuals were combined into the EUR and EAS groups, respectively. Most individuals have alleles that are common in our dataset and are present in most populations. HLA-DRB1*04:01 is of particular interest due to its high frequency in the Khoisan speaking populations (Khoisan, Hadza, Sandawe; see also **Supplementary Figure S21**).

We examined how classical HLA alleles from the HLA-targeted sequencing dataset cluster based on their exonic sequence by constructing haplotype networks for each HLA gene. The majority of 2nd field alleles of *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, and *HLA-DQB1* that were found in more than one individual were present in multiple populations in our dataset (**Figure 5E** & **Supplementary Figures S17-S20**). There is minimal evidence of HLA alleles being specific to ethnic, geographic, or linguistic groups, which is consistent with prior observations of balancing selection maintaining shared allelic diversity. However, there are a few notable exceptions of HLA alleles with strong differences in allele frequencies among African populations, as already highlighted above at up to 4th field level of resolution: HLA-DRB1*04:01 is at high frequency in all of the Khoisan speaking populations (14% in the Sandawe, 26% in the Hadza, and 27% in the Khoisan) while being at <= ~1% frequency in all other groups, except for Europeans where this allele has a frequency of 7% (**Supplementary Figure S21A**). In our dataset, *HLA-B*41:02* (**Supplementary Figure S21B**) is only present in sub-Saharan African populations and is at especially high frequency in the Hadza (27%). There are two HLA alleles (*HLA-B*27:05* and *HLA-C*02:02*) that are at high frequency (16% and 20%) in the Fulani population (**Supplementary Figure S21C-D**). These few examples are consistent with local adaptation or ancient shared ancestry, however, the broad pattern of HLA alleles in these haplotype networks is consistent with balancing selection. Following these observations, which corroborated the results from the HLA-SNP data, we also investigated signatures of selection in the HLA-targeted sequencing data. Using the Ewens-Watterson test (Slatkin, 1994; Watterson, 1977), we found that the majority of the HLA loci tend to exhibit negative F_{nd} values, indicative of balancing selection, although the results did not remain statistically significant after multiple testing correction (**Figure 5d**).

In order to compare the HLA allelic diversity among the populations at the functional level, we computed two relevant metrics, namely the HLA evolutionary allele divergence and the peptide binding promiscuity. HLA allele divergence is defined as the sequence divergence between the HLA alleles of an individual's genotype and reflects the evolution of MHC diversity via a mechanism called divergent allele advantage (Pierini & Lenz, 2018; Wakeland et al., 1990). Allele divergence was quantified as Grantham scores for the pairs of alleles of a given HLA locus for each individual across the dataset (Grantham, 1974; Pierini & Lenz, 2018). The distribution of the Grantham scores across populations was highly variable, which could be indicative of differential selective pressures and evolutionary histories. However, we did not observe any clear differences between continents or a consistent pattern within a population

across different loci (**Supplementary Figure S22**). We further analyzed HLA allele divergence by using simulations to obtain the expected distribution of divergence for each population under a no-selection scenario (**Supplementary Figure S23**). Here, we found that, when analyzing the data across all five HLA loci together (HLA-A, -B, -C, -DRB1 and -DQB1), non-African populations were more likely to exhibit larger than expected allele divergence (chi-square test, p=0.006). The other functional metric analyzed here is peptide binding promiscuity, representing another source of functional variation among HLA alleles and suggested to evolve under pathogen-mediated selection (Kaufman, 2018a; Özer & Lenz, 2021). Promiscuity is here defined as the size of the bound peptide repertoire of an HLA allele. We quantified promiscuity for each HLA allele as the predicted number of bound peptides among 1,000,000 random natural pathogen peptides by using established computational algorithms. The predicted repertoire of bound peptides for each allele was then used to calculate individual-level promiscuity values as the combined size of bound peptide repertoires that are conferred by an individual's HLA genotype. Similar to the results of evolutionary allele divergence, promiscuity values showed substantial variation across populations without any consistent pattern among loci or continents (**Supplementary Figure S24**). Finally, we hypothesized that decreased HLA diversity (i.e. allelic richness) within a population might be compensated by increased allele divergence or promiscuity so that individuals can still respond to a large variety of peptides. However, no significant correlation was observed between allele divergence and allelic richness or promiscuity and allelic richness (**Supplementary Figure S25**).

Discussion

Among polymorphic regions in the human genome, the MHC region, and particularly the classical HLA genes, are a major focus of medical and evolutionary research due to their important role in immune function (Dawkins & Lloyd, 2019). Despite the interest in studying human MHC diversity, African populations remain underrepresented in the immunogenomics field (Bentley et al., 2017; Hindorff et al., 2018; Peng et al., 2021; Sirugo et al., 2019). In this study, we analyzed the MHC diversity of 12 ethnically-diverse populations from sub-Saharan Africa, together with 10 populations from the 1KG dataset (Abi-Rached et al., 2018; Auton et al., 2015; Fan et al., 2023). Our population genetics approach complemented with the analysis of the functional diversity of classical HLA genes provides insights into the evolutionary processes that maintain high allelic diversity across human populations.

High nucleotide diversity is a widely accepted hallmark of the classical HLA genes (Robinson et al., 2017). Using whole-genome sequence data from 12 sub-Saharan African populations, we showed that this excessive diversity extends across the 4Mb of the MHC region, peaks at the classical HLA genes, and is unparalleled throughout the genome. Differences in nucleotide diversity and expected heterozygosity between the MHC region and the rest of chromosome 6 remain high within each population. In the non-African populations, we observed a relatively lower genetic variation, both in neutral regions and the MHC region, which is in line with the out-of-Africa expansion in modern humans found previously (Malaspinas et al., 2016; R. Nielsen et al., 2017; Pagani et al., 2016). However, the decrease in genetic variation due to this bottleneck is not as dramatic for the MHC region as for neutral variation, suggesting that variation at the MHC is selectively maintained, even in the face of strong demographic events. Such a high level of genetic diversity across a diverse set of populations could result from the maintenance of shared polymorphism through balancing selection or from divergent selection due to local adaptation. Indeed, HLA genes are prime candidates for local adaptation to geographically restricted pathogens due to their functional importance for antigen-specific immunity (Meyer et al., 2018; Radwan et al., 2020). However, our analyses revealed elevated Tajima's D values across the MHC region that peak at the classical HLA genes, which is an established signature of balancing selection (Buhler & Sanchez-Mazas, 2011; Tajima, 1989). Similar patterns of Tajima's D were also observed in non-African populations from Denmark (Jensen et al., 2017) and Sweden (Nordin et al., 2020), which indicates that balancing selection also maintains MHC polymorphisms outside of Africa. In line with the argument that Tajima's D is particularly suitable for detecting balancing selection of the distant past (Buhler & Sanchez-Mazas, 2011), mutation ages calculated across chromosome 6 showed that the MHC region is significantly enriched with old mutations. This excess of old mutations adds further support that long term balancing selection is maintaining very ancient polymorphisms within the MHC region.

Principal component analysis (PCA) on genome-wide SNP data clearly resembles the demographic history of the studied populations (Auton et al., 2015; Fan et al., 2023; Tishkoff et al., 2009). While the African and non-African populations were separated across PC1, the Khoisan populations were separated from other groups across PC2, in agreement with their deep divergence from all other modern human populations (Fan et al., 2019, 2023). In contrast, PCA based only on SNPs within the classical HLA genes showed no clustering of individuals with regard to populations or even continents, and this lack of clustering could not be explained

solely by the smaller set of SNP markers. This indicates that population structure reflecting the demographic history of populations is less prominent across HLA genes, a signature in agreement with the maintenance of shared MHC polymorphism throughout times of population divergence. This lack of population structure at the MHC was corroborated by similar F_{ST} values when comparing the HLA genes with the entire chromosome 6, with the exception of the HLA-DP genes. While this could indicate a unique pattern of divergent selection for HLA-DP, it has already been suggested previously that *HLA-DPA1* and *HLA-DPB1* genes might evolve under different selective constraints than the other HLA class II genes (Hollenbach et al., 2001; Meyer et al., 2018). Interestingly, when focusing on a within-Africa comparison using only the whole genome sequencing data from the 180 individuals, two additional loci (*HLA-B* and *HLA-DQA1*) stand out statistically, indicating that relatively higher population differentiation at the HLA compared to chromosome-wide levels is more pronounced when comparing groups within a continent than when comparing groups across continents. A similar observation has also been made previously in the 1KG data (Brandt et al., 2018) and likely results from the fact that most of the genome-wide variation evolves neutrally and follows a simple pattern of isolation-by-distance. Such neutral differentiation is therefore lower within continents, leaving a higher probability for weak patterns of local adaptation at the HLA to exceed genome-wide levels of differentiation. Corroborating this interpretation, none of the HLA genes exhibited average F_{ST} values exceeding the 70th percentile of the chromosome-wide F_{ST} values, indicating that while statistically significant, any effect of differentiation at the HLA due to local adaptation is small.

Overall, the patterns observed in the whole genome SNP data highlight the impact of both demography and balancing selection, the latter maintaining ancient polymorphism, as the main mechanisms of evolution for the HLA genes. This is in agreement with our results from the HLA-targeted sequencing data of the same populations. PCA plots based on classical HLA alleles (full-length sequence variants rather than SNPs) show loose clustering of individuals based on continents, but otherwise largely lack population-specific clustering, and the PCoA plots of pairwise F_{ST} values merely reflect the demographic history of populations at the continental scale. However, in contrast to the substantial variation in genome-wide genetic diversity among populations, we observed comparable levels of HLA allelic richness across populations even at the inter-continental scale. Apparently, balancing selection at the HLA has been stable and strong enough to counteract demographic events, such as the out-of-Africa bottleneck. Indeed, the Ewens-Watterson test suggests signatures of balancing selection in the

majority of the populations and loci, and the haplotype networks confirm the extensive sharing of HLA alleles among populations despite differences in allele frequencies. Similarly, the majority of the HLA-A and HLA-B supertypes (i.e. groups of functionally similar alleles) were present in almost every population suggesting that the functional breadth of HLA allele pools is conserved across populations. Notable exceptions here were the near-absence of B62 and an increased frequency of B58 in our data from African populations. Both of these supertypes were shown to have a higher between population variation (dos Santos Francisco et al., 2015) indicating that they may be targeted particularly by local selective pressures.

Although allelic richness levels were comparable across continents and also populations, there were a few populations that exhibited deviations in allelic richness. Specifically, the Chabu, a population with a known recent history of a strong population bottleneck (Gopalan et al., 2022), exhibited a lower allelic richness in several loci, indicating that the number of alleles in a population still remains sensitive to recent demographic events, despite the observed strong effect of balancing selection. One possibility of how decreased allelic richness at the population level could be compensated is by selection for more divergent alleles or alleles with higher peptide binding promiscuity so that individuals can still respond to a wide range of pathogens. However, despite substantial variation among populations in both allele divergence and promiscuity, neither was correlated with allelic richness **(Sup Fig 25),** suggesting that the recent demographic history of populations is not the main driver for variation in allele divergence and promiscuity.

The observation of PC clustering (Supp Fig 26) further supports the idea that the primary factor influencing genetic diversity among HLA alleles is not solely due to population or demographic differences, but rather linked to the physical structure (serotypes) of the HLA genes, specifically the DQA1 and DQB1 genes. Notably, when considering only the SNPs (single nucleotide polymorphisms) of all HLA genes, the observed PC clustering is not driven by ancestry, but rather by the DQ serological character, which indicates functional relevance. In essence, the functional significance of the DQ serological character outweighs the influence of local adaptation, underscoring the importance of maintaining diverse HLA alleles to combat various pathogens in populations worldwide.

As expected from high-resolution genotyping of diverse and previously under-represented ethnic groups, we found several novel HLA alleles in the studied populations from Sub-Saharan Africa (Pagkrati et al., 2023). While most of the novel variation was found in intronic or UTR

regions, eight alleles harbor variation in the functionally important peptide binding domains: 1x HLA-A, 1x HLA-B, 4x HLA-DPA1, 1x HLA-DPB1 and 1x HLA-DQB1. Interestingly, the novel allele HLA-A*43:03 is observed only in southern African Khoesan-speaking populations (!Xoo and Ju|'hoansi), highlighting the evolutionary history of these populations within Africa. In addition to novel alleles, there are also a few previously identified HLA alleles with stark frequency differences between African populations. One such allele is HLA-DRB1*04:01, which is at high frequency in all Khoesan speaking populations from southern and eastern Africa (Ju|'hoansi, !Xoo, Sandawe, and Hadza). This observation is particularly interesting since the Sandawe and Hadza do not cluster close to the other Khoesan groups in the genome-wide PCA. Therefore, this could represent an ancient shared ancestry between all Khoesan speaking populations that is not observed across the rest of the genome(Fan et al., 2023). Another interesting observation is the allele HLA-B*27:05, which is here found to be restricted mainly to one African population (Mbororo Fulani) but was found at a significant frequency in a group of ancient Europeans from the late neolithic period, apparently experiencing a continuous decline in frequency throughout European colonization from early hunter-gatherers to modern day Europeans (Immel et al., 2021). This may shed light on the origin of other alleles uniquely shared between the Fulani and Europeans including the -13910 mutation associated with lactose tolerance (Ranciaro et al., 2014). While novel alleles were expected to be identified in diverse Africans, it is also worth noting that 97.6% of the observed HLA alleles -at the 2nd field resolution- were already known in the international IMGT database. This corroborates our overall findings that most of the allelic variation at the HLA genes is shared among a wide range of populations, likely maintained by long-term balancing selection, so that even isolated ethnic groups do not harbor large amounts of private HLA alleles.

In summary, our report represents the most comprehensive characterization of HLA diversity in ethnically diverse Africans to date and provides unique insights into the mechanisms of HLA evolution in humans. We show that functional HLA diversity is not only high within Africa but also maintained at comparable levels in European and East Asian populations. We reveal signatures of very ancient balancing selection and show that the MHC region is still today enriched with old mutations, suggesting that mechanisms of balancing selection have been continuously active throughout human history. Furthermore, we find no evidence for HLA differentiation among populations exceeding that of neutral variation, which could have indicated local adaptation to specific pathogens. Selection by local pathogens thus appears to play a minor role in MHC evolution in humans, compared to demography and balancing

selection on ancient variation, such as heterozygote or divergent allele advantage (Pierini & Lenz, 2018).

Material & Methods

Populations and datasets used in analyses

We analyzed a SNP dataset of 180 sub-Saharan African individuals derived from high coverage whole genome sequencing ('the SNP data') and a dataset of 489 sub-Saharan African individuals with high-quality targeted sequencing of Class I and Class II HLA genes ('the HLA-target data'). These datasets include a wide diversity of populations across sub-Saharan Africa, representing the major African language families: 1) Khoesan: Ju|'hoansi, !Xoo, Sandawe, and Hadza; 2) Nilo-Saharan: Chabu and Mursi; 3) Niger Congo: Fulani, Herero, Baka&Bagyeli, and Tikari; and 4) Afroasiatic: Amhara and Dizi (Fan et al., 2023).

Whole Genome Sequence Data Preparation (SNP data)

The high coverage whole genome sequences were previously aligned to the human reference genome hg19 to call biallelic SNP data(Fan et al., 2023). All following genomic coordinates refer to the hg19 human reference sequence. For most of the analyses (unless indicated otherwise), we combined our SNP data from the 180 individuals with SNP data from the Phase 3 1KG project by first identifying the intersect of SNPs in both datasets. We prepared the biallelic SNPs from the 180 WGS dataset for phasing by excluding any variant that was missing in >= 10% of the sample with Plink V1.9 (Purcell et al., 2007) (PLINK v1.9: https://www.cog-genomics.org/plink/). We then phased the remaining SNPs using Eagle v2.4.1 and the Eagle distributed HG19 genetic map (Loh, Danecek, et al., 2016; Loh, Palamara, et al., 2016).

Principal Components Analysis

To prepare our data for PCA, we used plink v1.9 (Purcell et al., 2007) to filter the merged dataset by removing all singletons with "--mac 2" and LD pruning with a 50 snp sliding window, a 5 snp step size and an r^2 value of 0.5 (--indep-pairwise 50 5 0.5). We then created

two sets of variants for PCA: 1) All variants across the genome and 2) all variants in the Class I HLA genes (*HLA-A*, *HLA-B*, *HLA-C*) and Class II HLA genes (*HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*). Principal components were then calculated on each set of variants using smartpca from Eigenstrat (Patterson et al., 2006; Price et al., 2006) with 0 iterations of outlier removal.

Evaluation of PCA and small region size

The Class I and Class II HLA genes combined account for only 52,584 bp, and includes only 902 SNPs. Therefore, we evaluated if the lack of geographic and linguistic clustering observed in PCA is due to the small region size. We randomly selected 100 intergenic regions with exactly 902 SNPs and 100 intergenic regions of 52,584 bp. Discriminant analysis of principal components (DAPC) was then performed on each random region as well as the set of SNPs used in the Class I and Class II PCA (Jombart et al., 2010). DAPC was calculated using R V3.6.1 (R Core Team, 2019) and the adegenet library(Jombart, 2008; Jombart & Ahmed, 2011) with the following parameters: three principal components (n.pca=3), 10 discriminate analysis axes (n.da=10), variables scaled by their standard deviation (scale=TRUE), var.contrib=TRUE, var.loadings=FALSE, pca.info=TRUE.

Mutation Ages

We used GEVA (Albers & McVean, 2020) to calculate the age of all mutations across chromosome 6 in the 180 genomes dataset with a dataset wide minor allele frequency ≥ 5%. One complication with applying GEVA to the MHC region is that GEVA requires the correct polarization of ancestral vs. derived alleles. However, ~23% of the MHC region does not have an ancestral allele determined in the human ancestral reconstructed reference sequence (human_ancestor_GRCh37_e59) downloaded from: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/ (Auton et al., 2015). Therefore, we devised a strategy to determine the derived mutation age in regions where ancestral and derived alleles cannot be determined. We first polarized the genome by known ancestral alleles making the reference allele in the vcf file equal to the ancestral allele from the human ancestral reconstructed reference sequence. For SNPs without

a known ancestral allele, we assumed that the reference allele is ancestral. We then used GEVA to calculate the age of all mutations assuming a constant mutation rate of 1×10^{-8}, an effective population size of 20,000, and the hmm initial and emission probabilities distributed in the GEVA download from github. Once GEVA was completed we then applied the GEVA R script "estimate.R" with an effective population size of 20,000, and calculated QC scores as described by GEVA (Albers & McVean, 2020). We kept all mutation age estimates that used the joint clock with a QC score > 0.5.

We then flipped the reference and alternate alleles, for variants that passed the QC score cutoff and applied the same pipeline to recalculate variant ages. We then compared the age estimates with both runs of GEVA and took the minimum age in cases where both estimates passed the QC score cutoff and assumed that the younger age estimate is the most likely of the two that reflects the derived allele. We compared these age estimates to the age estimates of known ancestral alleles and the overall patterns of age estimates are relatively similar. Therefore, we are confident that by comparing the minimum age estimate of variants allows for the most accurate depiction of mutation ages in scenarios where we are unable to determine an ancestral vs. derived allele.

Measuring Genetic Diversity, Balancing Selection, and Differentiation

We used *VCFtools* (Danecek et al., 2011) to calculate nucleotide diversity (pi) in 1kb windows and per-site in the African 180 WGS dataset. We also calculated Pi on a per-site basis using the merged 180 WGS / 1KG data. For the per-site analysis, we removed singletons and LD pruned the dataset prior to calculating Pi using Plink (Purcell et al., 2007) (--indep-pairwise 50 5 0.5), and averaged Pi values across the entirety of chromosome 6 (583,063 SNPs) or the MHC region (23,179 SNPs). Populations from the 1KG dataset were down-sampled to n=15 per population, to match with the sample size of the African 180 WGS dataset. We calculated expected heterozygosity (2pq) for each site genome-wide in the merged 180 WGS / 1KG dataset, where we LD pruned the dataset and downsampled 1KG populations. We calculated Tajima's D in 1kb windows using *VCFtools*. For F_{ST} analyses, we ran one analysis that contained all 22 populations from the African 180 WGS / 1KG merge. For this analysis, we calculated F_{ST} using the Weir & Cockerham (Weir & Cockerham, 1984) formula implemented in the python package

scikit-allel (Miles et al., 2021). For pairwise F_{ST} calculations, we used the Weir & Cockerham estimator in *VCFtools*.

Targeted HLA sequencing data (HLA-target data)

HLA genotyping was performed at 4-field level following the method described (Pagkrati et al., 2023). Briefly, genomic DNA was isolated from white blood cells with Puregene DNA extraction kits (Qiagen, Germany). All 11 HLA class I and class II genes (HLA-A, -B, -C, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1 and -DPB1) for the aforementioned 489 individuals were sequenced on an Illumina MiSeq platform (San Diego, CA) on multiple sequencing runs using targeted amplicon-based NGS with Omixon Holotype HLA™ V2 kits (Budapest, Hungary)(Margolis et al., 2021). Fastq files were analyzed with Omixon Twin (version 3.1.3) and GenDx NGSengine (Utrecht, Netherlands, version 2.13) using IPD-IMGT/HLA database version 3.38.

In order to generate SNP data for the PCA analysis on the HLA-target data, nucleotide sequences for all 4th field alleles were obtained from the international ImMunoGeneTics (IMGT) HLA database version 3.25.0 (Robinson et al., 2020). We used plink v1.969 to filter the dataset by LD pruning with a 50 snp sliding window, a 5 snp step size and an r2 value of 0.5 (--indep-pairwise 50 5 0.5). If a locus for a particular sample did not have two fully characterized alleles at 4 fields, the sample was excluded from this analysis. This may happen if one allele is ambiguous in homopolymer or low complexity regions. Of the 489 individuals included in the study, 7 had incomplete typing for at least one locus and therefore these samples were removed.

DQA1-DQB1 haplotypes were generated using BIGDAWG (Pappas et al., 2016) then DQ serotypes were inferred based on the first field of the predicted haplotypes.

In order to be able to put the results from our HLA-target data analysis into a worldwide context, we have also used targeted genotyping data from 10 1KG Project populations with four from Africa (GWD - Gambian in Western Division, MSL - Mende in Sierra Leone, ESN - Esan in Nigeria, YRI - Yoruba in Ibadan, Nigeria), three from Europe (TSI - Toscani in Italia, FIN - Finnish in Finland, GBR - British from England and Scotland) and three from East Asia (CHB - Han Chinese in Beijing, China, CHS - Han Chinese South, China, JPT - Japanese in Tokyo,

Japan) (Abi-Rached et al., 2018). As the 1KG Project samples were typed for three HLA class-I (HLA-A, HLA-B, HLA-C) and two HLA class-II loci (HLA-DRB1 and HLA-DQB1), all the analysis were performed by using these loci unless otherwise stated.

Observed heterozygosity and allelic richness based on the rarefaction method (El Mousadik & Petit, 1996; Hurlbert, 1971) were calculated using *hierfstat* package in R environment(Goudet, 2005; R Core Team, 2019). Supertypes classification of HLA alleles was done according to Sidney et al. (Sidney et al., 2008).

Ewens-Watterson test of homozygosity was employed to test for deviations from neutrality (Watterson, 1977). This test compares the observed homozygosity (F_{obs}) and the expected homozygosity (F_{exp}) under neutrality for each population. Normalized deviate of homozygosity (F_{nd}) values were calculated by following Salamon et al. (Salamon et al., 1999) in order to account for different sample sizes and allele numbers across populations and the associated p-values were calculated by following Slatkin (Slatkin, 1994) as implemented in the PyPop software (Lancaster et al., 2007).

Principal component analysis (PCA) was performed on an allele count matrix in which rows correspond to individuals and columns to distinct HLA alleles. Values on that matrix range from 0 to 2 with 2 indicating an individual homozygous for the corresponding allele, 1 indicating the allele is found in heterozygous state and 0 showing that the allele is not observed for that individual. Individuals with missing genotypes were removed from the input data and the principal components were calculated using the *prcomp* function in R.

Genetic differentiation between populations was quantified by calculating F_{ST} values according to Weir and Cockerham (Weir & Cockerham, 1984) and principal coordinates analysis was performed on pairwise F_{ST} values by using *hierfstat* package (Goudet, 2005). In order to further analyze variation in allele frequencies between African and non-African populations, alleles with large frequency differences (LFD) were identified following a method similar to Single et al. (Single et al., 2020). Populations from Africa were merged into a group and all the other populations were merged into a non-African group. Allele frequencies were calculated within two groups and alleles were classified into four categories. African LFD alleles are at least three times more frequent in Africa while non-African LFD alleles are the ones that are at least three times more common outside Africa. Rare alleles are those observed less than two times in both groups. The remaining alleles (i.e. the ones with less than a 3-fold frequency difference) were

categorized as shared alleles. Due to the small number of populations, especially outside of Africa, and relatively small sample sizes, endemic alleles (i.e. alleles that are observed only in Africa or only outside Africa) were also considered as alleles with large frequency differences.

The evolutionary divergence between pairs of alleles of an individual's HLA genotype was calculated using the Grantham distance score (Grantham, 1974) across exon 2 and 3 for HLA class I loci and exon 2 for HLA class II loci following Pierini & Lenz (Pierini & Lenz, 2018). In order to obtain an expected distribution of Grantham scores within populations, we performed simulations. Within each population, individuals were assigned alleles randomly, Grantham distance scores were calculated and the median Grantham distance score was recorded. This procedure was repeated 1,000 times to obtain a distribution of the median Grantham distance score for each population.

HLA peptide binding promiscuity was calculated for each allele as the number of peptides that are predicted to be bound among 1,000,000 random natural pathogen peptides. Two established binding prediction algorithms, namely NetMHCpan(v4.1) for HLA class-I and NetMHCIIpan(v4.0) for HLA class-II were used to define bound peptides with the affinity threshold of 500nM (Özer & Lenz, 2021; Reynisson et al., 2020). Individual-level promiscuity values were calculated as the combined number of peptides that are bound by the pair of alleles an individual has by accounting for the overlap in alleles' peptide repertoire. Kendall's rank correlation test was used to analyze the relationship between the median promiscuity value and allelic richness as well as the median Grantham distance and allelic richness across populations.

Haplotype networks

To examine the population structure of *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1*, and *HLA-DRB1* alleles, we constructed haplotype networks from the targeted sequencing data combined with the FIN, GBR, TSI, CHB, CHS, and JPT populations from the 1KG project. The FIN, GBR, and TSI haplotypes were grouped into one European super-population, and the CHB, CHS, and JPT were grouped into one East Asian super-population. The Ju|'hoansi and !Xoo were also grouped into one Khoisan population. We limited each haplotype network to only the second and third exon of HLA class-I genes and the second exon of HLA class-II genes and grouped haplotypes based on a two-level allele designation. There were a few alleles with identical sequences based on the 2nd field-level designation, therefore for these alleles we combined them

based on the established “G-group” nomenclature (Marsh et al., 2010b). We excluded all singleton HLA alleles and constructed the haplotype networks using POPART’s (Leigh & Bryant, 2015) (http://popart.otago.ac.nz/index.shtml) implementation of the TCS algorithm (Clement et al., 2002).

Acknowledgements

This research was supported by the following funding sources: NIH Grant T32 DK07314, the Penn Training Grant in Diabetes, Endocrine and Metabolic Diseases to Eric Mbunwe, American Diabetes Association Pathway award 1-19-VSN-02 and NIH grants 1R01DK104339, 1R35GM134957, and R01AR076241 to Sarah A. Tishkoff; NIH grant R01AR070873 and institutional funds from the Children's Hospital of Philadelphia to Dimitri S. Monos and Office of Naval Research Grant (N00014-18-1-2045) to Martin Maiers. The authors acknowledge and are grateful for the technical support received by the staff of the Immunogenetics lab regarding the HLA genotyping of the samples of this study: Ioanna Pagkrati, Deborah Ferriola, Jenna Wasserman, Amalia Dinou, Nikolaos Tairis, Georgios Damianos, Ioanna Kotsopoulou, Joanna Papaioannou, Diamantoula Giannopoulos.

Supplementary Material for Chapter 2 is provided in Annex IV.

Chapter 3

Spatio-temporal analysis of ancient HLA data reveals major effects of demography and admixture on immune gene diversity

Onur Özer[1], Yan-Rong Chen[1], Nicolas Antonio da Silva[2], Magdalena Haller[2], Sébastien Calvignac-Spencer[3], Almut Nebel[2], Ben Krause-Kyora[2], Tobias L. Lenz[1]

[1] Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany

[2] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

[3] Robert Koch Institute, Berlin, Germany

Unpublished manuscript

Abstract

Understanding the extensive polymorphism observed in HLA immune genes within and among human populations is of great interest in human genetics, because of the central function of HLA molecules in adaptive immunity. High HLA diversity is assumed to evolve in response to pathogen pressures, yet empirical evidence of the expected temporal changes in HLA allele pools remains rare. Here, we analyze a unique HLA dataset generated from 129 ancient individuals from central Europe dating back to the early and late periods of the Neolithic (5000-2800 BCE). Comparison of this dataset to modern populations reveals that although ancient Europeans were more similar to modern Europeans in comparison to African and East Asian populations, there were substantial changes in frequencies of several alleles within Europe since the Neolithic. Specifically, alleles forming the haplotypes *A*03:01-B*07:02-C*07:02* and *A*01:01-B*08:01-C*07:01* were missing in Neolithic times, although they are the most common alleles in some northern and central European populations today. Based on the observation that these haplotypes exhibit a north-south cline within Europe today, we argue that they were introduced into Europe by the steppe pastoralists at the end of the Neolithic. Furthermore, comparing early and late Neolithic populations, we identify alleles *B*40:01* and *C*03:04* that were possibly part of the local hunter-gatherer allele pool and introgressed into early farmers. We investigated the effect of this allelic differentiation on antigen-binding for distinct pathogens by using peptide-binding prediction tools. Based on this computational analysis, the functional diversity of HLA allele pools in both Neolithic and modern populations appears to be similar. Overall, our results suggest a considerable role of demography and population admixture in shaping the modern distribution of HLA alleles within Europe.

Introduction

The recent evolutionary history and population structure of humans are shaped by migration, admixture and adaptation to local or changing environmental conditions (Nielsen et al., 2017; Rees et al., 2020; Rosenberg et al., 2002). Among the many events that influenced the course of human evolution, the Neolithic revolution, which is marked by the domestication of animals and plants was a significant turning point (Zeder, 2011). The transition from hunting and gathering to farming and herding is accompanied by major changes in diet, a sedentary lifestyle, increased population density within farming communities and continuous close contact with various domesticated animals (Armelagos & Harper, 2005; Bocquet-Appel, 2011; Quagliariello et al., 2022; Richards et al., 2003; Skoglund et al., 2014). These conditions have essentially laid the groundwork for increased zoonotic events and the rapid spread of pathogens (Larsen, 2018). Named as the "first epidemiological transition", drastic changes in the pathogen landscape following the Neolithic have attracted many researchers trying to understand adaptations of humans facing the challenges associated with this new lifestyle as well as the emergence of novel and deadly pathogens (Harper & Armelagos, 2010; Spyrou et al., 2019). Indeed, pathogens are considered to be a major driver of evolution and, in line with this, many immune genes were shown to exhibit signals of recent positive selection (Enard et al., 2016; Fumagalli et al., 2011; Gouy & Excoffier, 2020; Karlsson et al., 2014). Interestingly, such genes are commonly associated with an increased risk of inflammatory and autoimmune diseases due to their pleiotropic effects (Raj et al., 2013).

A significant number of disease associations have been mapped to the major histocompatibility complex (MHC) region in humans (Trowsdale & Knight, 2013). This region contains many immune-related genes including the highly polymorphic classical MHC genes (also known as human leukocyte antigen or HLA in humans). HLA class-I (i.e. HLA-A, HLA-B, HLA-C) and class-II (i.e. HLA-DR, HLA-DP, HLA-DQ) molecules bind and present peptides from the intracellular and the extracellular environment, respectively, on the cell surface. Non-self peptides, such as those derived from pathogens infecting the cell are recognized by T-cells and an immune response is initiated. HLA molecules bind different peptides based on their amino acid sequence. Differential associations of HLA alleles with infectious diseases support the hypothesis that HLA diversity evolves in response to pathogen diversity (Sanchez-Mazas, 2020). Therefore, the extreme polymorphism of HLA genes is often highlighted as a marked example of the result of pathogen-mediated balancing selection (Radwan et al., 2020). Balancing selection leads to a decreased differentiation in HLA genes between populations by

maintaining distinct allelic lineages in isolated populations (Brandt et al., 2018). However, there remain significant and well characterized differences in HLA allele frequencies between populations (Prugnolle, Manica, Charpentier, et al., 2005; Sanchez-Mazas, 2007; Sanchez-Mazas et al., 2012). Directional selection on specific alleles that is driven by local pathogens is a commonly invoked mechanism to explain HLA differentiation between populations (Hoh et al., 2020; Meyer et al., 2018). However, whether such differentiation is driven by neutral processes such as founder effects or by natural selection is a topic of ongoing discussion (Maróstica et al., 2022). Much less is known about the temporal dynamics of HLA allele frequencies within human populations. Theoretical studies suggest that under the negative frequency-dependent selection, which is one of the mechanisms of host-pathogen coevolution, pathogens rapidly adapt to evade recognition by the most common alleles in the host, essentially rendering them ineffective. As a result, rare alleles become advantageous and are expected to rise in frequency, which leads to cyclic frequency oscillations of alleles within a host population (Ebert & Fields, 2020; Rabajante et al., 2016). However, empirical evidence for such dynamic frequency changes in MHC genes over time is scarce (Lenz, 2018).

Until recently, the evolutionary history of HLA alleles was investigated mainly by theoretical modeling and simulation studies (Borghans et al., 2004; Penman & Gupta, 2018) and by comparative analysis of contemporary populations (Brandt et al., 2018; Buhler & Sanchez-Mazas, 2011). While the current empirical research focuses on the spatial distribution of HLA alleles, advances in the ancient DNA field provide a promising opportunity to compare differences between populations at the temporal level. Genetic analysis of human remains has been instrumental in identifying the migration and admixture patterns of ancient humans and signals of selection following the Neolithic revolution (Koptekin et al., 2023; Mathieson et al., 2015; Stoneking et al., 2023). Furthermore, HLA and other immune-related genes were implicated in adaptive introgression from Neanderthals and Denisovans, suggesting that humans may have benefited from locally adapted alleles of ancient hominins occupying Eurasia much earlier than humans (Abi-Rached et al., 2011; Racimo et al., 2015; Rotival & Quintana-Murci, 2020). Although HLA genes are known targets of selection, their extreme polymorphism complicates their genotyping from the low coverage ancient DNA data. Furthermore, large sample sizes are required for accurate estimation of HLA allele frequencies due to the high number of alleles (B-Rao, 2001) and most ancient DNA studies have been limited to a few individuals from a site and time interval. Therefore, a special effort is required to obtain reliable HLA data from ancient specimens. In this study, we report HLA class-1 genotypes of 129

individuals from seven excavation sites located within modern central and southern Germany dating back to the early and late Neolithic. The relatively large sample size of this ancient dataset allows comparisons between modern and ancient populations. Our results reveal the major role of balancing selection on maintaining functional HLA diversity as well as significant differences in allele frequencies, which are potentially shaped by the admixture and selection.

Results

HLA class-I genotypes of individuals within this study were determined with a method developed specifically to genotype highly polymorphic regions from low coverage sequencing data (Pierini et al., 2020). In total, 43 HLA-A, 48 HLA-B and 45 HLA-C alleles were identified at two-field resolution in ancient samples. Based on the dating, individuals from excavation sites of Fellbach – Oeffingen (n=16), Niederpoering (n=6) and Trebur (n=24) were grouped to form the Early Neolithic Farmers (ENF) metapopulation (n=46) and individuals from Altendorf (n=13), Niedertiefenbach (n=50), Rimbeck (n=3) and Warburg (n=17) were grouped to form the Late Neolithic Farmers (LNF) metapopulation (n=83). We further included HLA data from 10 population samples from the 1000 Genomes project encompassing Eurasia and sub-Saharan Africa (Abi-Rached et al., 2018) and another sample from a modern Germany population located in Kiel **(Supplementary Table 1)**. HLA diversity was analyzed based on three metrics, namely the observed heterozygosity, the Shannon Index and allelic richness. Heterozygosity declines from Early Neolithic to Late Neolithic, a pattern that is particularly marked for the HLA-A locus **(Supplementary Figure 1)**. Compared to modern populations, both ENF and LNF harbor slightly lower heterozygosity with the exception that the HLA-A heterozygosity of ENF is comparable to FIN and CHS populations. Allelic richness values of ENF and LNF are comparable to modern populations suggesting that the number of alleles maintained within populations was similar **(Figure 1)**. An interesting exception is the HLA-C locus of LNF that harbors around 40% more alleles than the population with the second highest allelic richness (LWK) **(Supplementary Figure 2)**. While this can be the result of combining several populations to form the LNF metapopulation, such an effect should also be observable for other loci and possibly for ENF. Shannon index and allelic richness values are correlated for each locus (Kendall correlation; HLA-A, tau=0.74, $p<0.001$; HLA-B, tau=0.69, $p<0.001$; HLA-C, tau=0.51, $p=0.015$). However, both ENF and LNF deviate from the modern populations towards a decreased Shannon index value in comparison to allelic richness. This is especially evident

for HLA-B where these two populations exhibit the lowest Shannon index **(Figure 1)**. Shannon index is calculated based on both the total number and the evenness of types in a sample (i.e. HLA alleles within a population) (Stirling & Wilsey, 2001). Therefore, a lower Shannon Index might indicate that the allele pool of the population is dominated by one or a few alleles. Indeed, the most common allele of both ENF and LNF exceeds 20% frequency for HLA-B. Something similar is only observed for CHS in modern populations and, similar to the Neolithic samples, CHS exhibits a slightly decreased Shannon index.

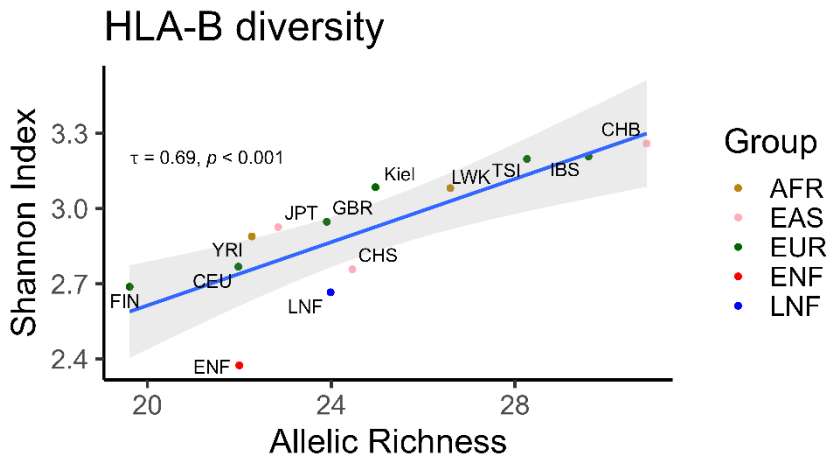


Figure 1. HLA-B diversity. Tau (τ) refers to the Kendall rank correlation coefficient. The linear regression line is shown in blue and the 95% CI around the line in gray.

Principal components analysis of ancient and modern samples revealed that both ENF and LNF cluster close to yet distinct from modern European and to some extent modern East Asian populations **(Figure 2)**. This pattern is also reflected in the pairwise population differentiation values, as the Fst between modern and ancient Europeans was on average lower than those between ancient Europeans and remaining modern populations **(Figure 2, Supplementary Figure 3)**. Several alleles in each locus remained at relatively similar frequencies within Europe from ancient to modern times. Interestingly, some of the stable alleles such as *B*44:02* and *C*05:01* are either missing or observed at very low frequencies in the East Asian and African populations. However, some Fst values between Neolithic and modern Europeans are higher than those between Neolithic Europeans and modern East Asian populations, highlighting the dynamic and high HLA variation over time. Several alleles exhibit high frequency differences, resulting in the observed differentiation between modern and ancient European populations **(Figure 3)**. These alleles include *A*24:02, A*31:01, B*27:05, B*39:01, B*51:01, C*02:02* and

*C*15:02*, which are highly common in both ENF and LNF while observed at lower frequencies in modern European populations. On the other hand, alleles *A*01:01, A*03:01, B*07:02, B*08:01, B*35:01, C*04:01, C*06:02, C*07:01* and *C*07:02* exhibit the opposite trend with low frequencies in ENF and LNF and high frequencies in modern Europe. In fact, with the exception of *A*02:01*, the most common alleles in the modern Kiel population are observed at very low frequencies in ENF and LNF for each locus **(Supplementary Figure 5, Supplementary Figure 6)**.

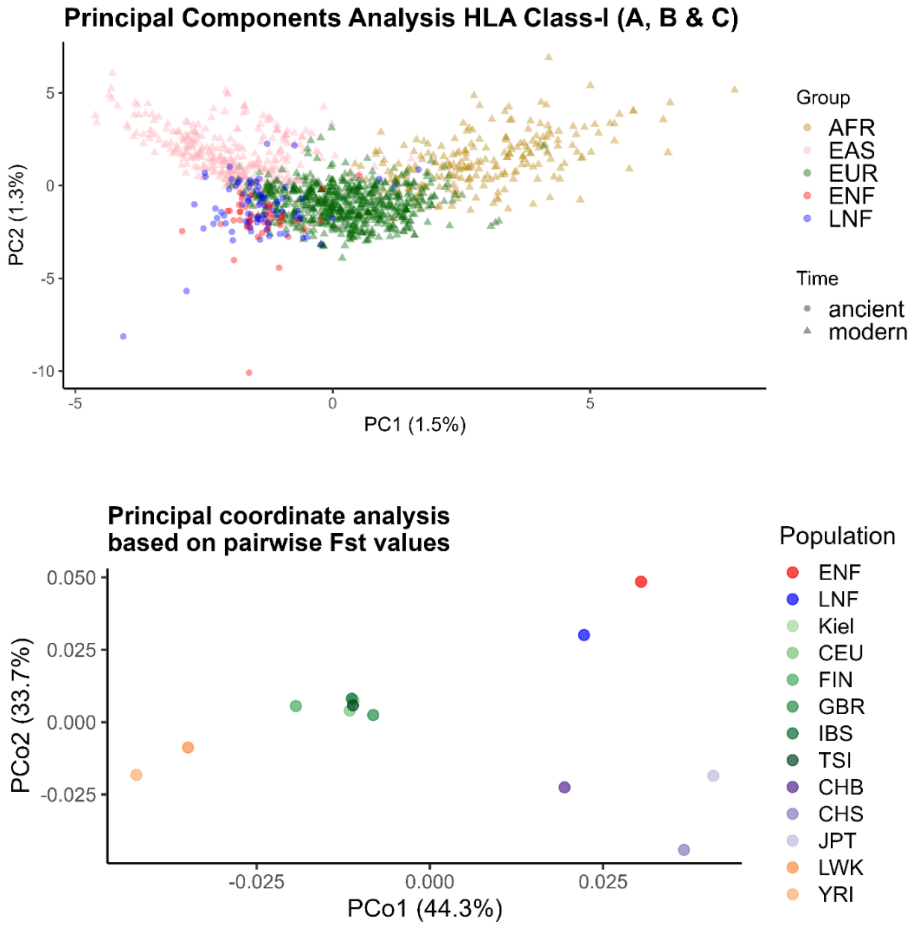


Figure 2. Population differentiation measured based on HLA alleles. Upper panel: Principal Components Analysis. Principal components were calculated using the HLA class-I genotypes of 1169 modern and 126 ancient individuals. Lower panel: Principal coordinates analysis of pairwise Fst values.

Allele divergence is calculated for each individual as the Grantham score between the protein sequences of HLA alleles of the individual's genotype. We restricted our analysis to only ENF, LNF and modern Kiel populations as the regions they occupy are the closest to each other within Europe. Homozygous individuals can skew the distribution of allele divergence values, as they essentially have zero divergence between alleles. Therefore, we also removed homozygous individuals from the analysis because of the differences in heterozygosity values between the three populations. Each HLA locus exhibits different patterns of divergence values across populations and significant differences among populations were observed for HLA-A with ENF having higher median divergence compared to the Modern Kiel population (Dunn test, $p<0.01$) and for HLA-C with ENF having lower median divergence compared to Modern Kiel population (Dunn test, $p<0.01$) **(Supplementary Figure 7a)**. No significant differences were observed between populations for the HLA-B locus. In order to analyze whether allele divergence might have an effect on the changes in allele frequencies between modern and Neolithic populations, we focused on the HLA-B locus as previous research points to HLA-B exhibiting the highest signal of divergent allele advantage (Pierini & Lenz, 2018). We compared the two most common alleles in the modern Kiel population that rose in frequency after the Neolithic, namely *B*07:02* and *B*08:01*, with the five most common alleles in Neolithic populations, *B*15:01, B*27:05, B*39:01, B*44:02* and *B*51:01*. We found that *B*44:02* was the most divergent allele among the five common Neolithic alleles and *B*39:01* was the least divergent one compared to *B*07:02* and *B*08:01* **(Supplementary Figure 7b)**. Interestingly, *B*39:01* exhibits a sharp decrease in frequency in modern populations while *B*44:02* is maintained at similar levels, suggesting an influence of allele divergence on allele frequency changes.

In order to analyze whether differences in HLA allele frequencies between Neolithic and modern European populations lead to functionally distinct allele pools, we used the computational peptide-binding prediction algorithm netMHCpan (Reynisson et al., 2020). These algorithms allow relatively accurate prediction of bound peptides by HLA alleles, even in the absence of experimental data, which is only available for a limited set of common alleles. We focused our analysis on the *Measles virus* (MeV) for several reasons. Firstly, recent analysis dates the emergence of the MeV around the first millennium BCE, long after the end of the Neolithic period (Düx et al., 2020). Therefore, if measles had any impact on human genomic variation, this should be observed in modern populations but not in Neolithic populations.

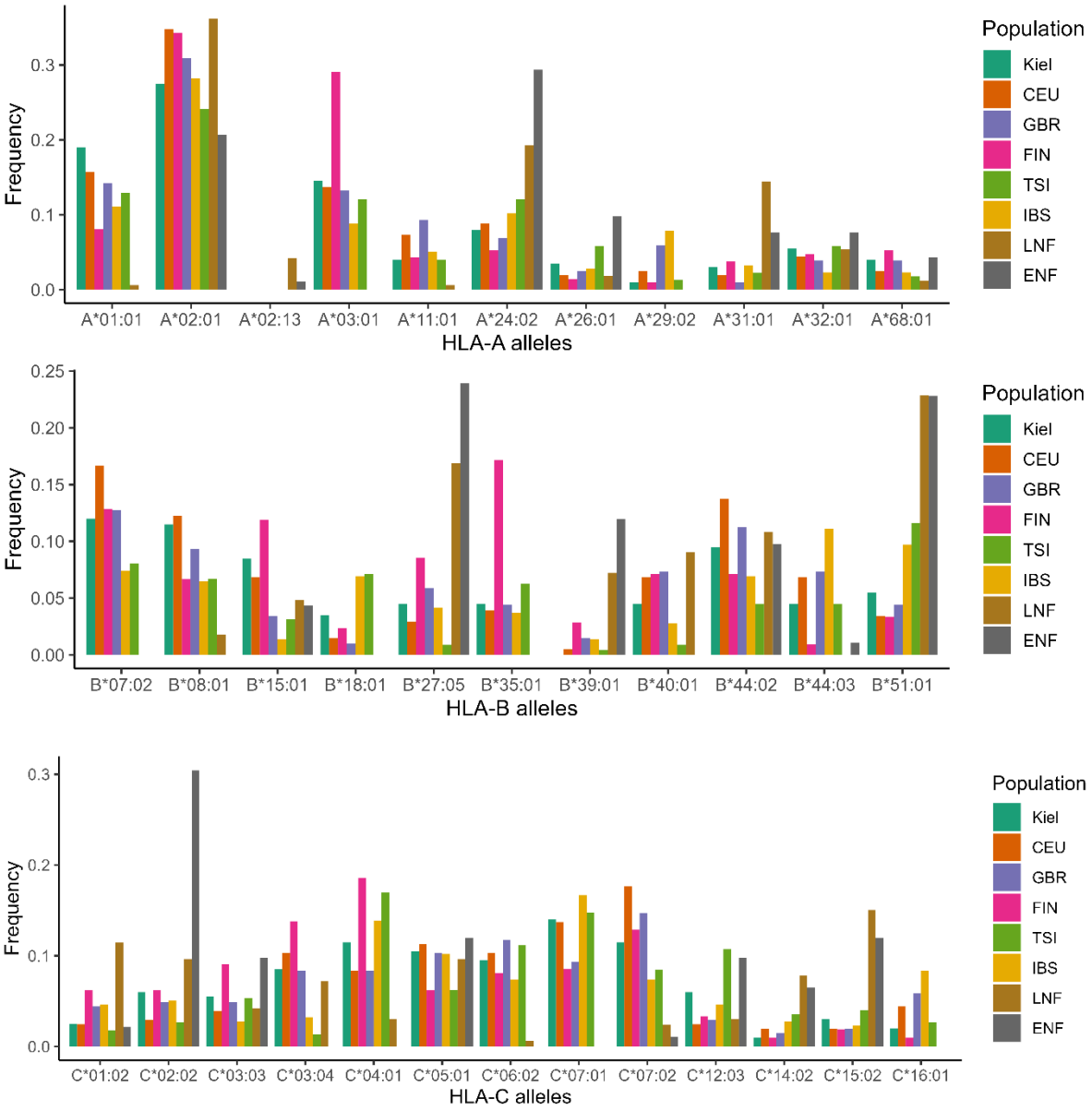


Figure 3. HLA allele frequencies of modern and ancient European populations. Only the most common five alleles in each population were plotted for clarity.

Secondly, once established in a population, measles is a childhood disease with high mortality in the absence of treatment (Rota et al., 2016). Variants associated with susceptibility or resistance to measles would be under high selective pressure because measles victims would die before reaching reproductive age. Finally, HLA class-I restricted $CD8^+$ T-cell responses were shown to be important for recovery from MeV infection and viral clearance (Lin et al., 2014; van Els & Nanan, 2002), suggesting that HLA alleles might be the target of selection by

measles. Overall, we hypothesized that the HLA allele pool of modern populations would be adapted to MeV while no such adaptation would be observed in Neolithic populations. As a control group, we also analyzed five viruses for which there exists plausible evidence of association with humans predating the Neolithic period. These viruses were *Human cytomegalovirus* (Murthy et al., 2019), *Human polyomavirus 2* (Forni et al., 2020), *Human adenovirus C* (Nielsen et al., 2021), *Varicella-zoster virus* (Weinert et al., 2015) *and Human papillomavirus 16* (Pimenoff et al., 2017), and we assumed that both the Neolithic and modern populations would be adapted to them. Finally, we have included seven bacterial pathogens into our analysis namely *Yersinia pestis*, *Vibrio cholera*, *Treponema pallidum*, *Salmonella enterica*, *Mycobacterium leprae*, *Mycobacterium tuberculosis* and *Helicobacter pylori*. While the date of emergence of most of these bacteria as human pathogens is still unresolved, they were identified in ancient remains and it was suggested that they can only be sustained in large and dense populations that arose after the Neolithic period (Spyrou et al., 2019; Wolfe et al., 2007). We used netMHCpan4.1 to predict peptides of pathogens bound by each individual in Early Neolithic Farmers, Late Neolithic Farmers and the Modern Kiel population. There were no significant differences in the number of bound peptides between populations with the exception of higher number of HLA-B bound peptides from *Human polyomavirus 2* by the modern Kiel population **(**HLA-A: **Supplementary Figure 8;** HLA-B: **Supplementary Figure 9;** HLA-C: **Supplementary Figure 10).** Next, we focused our analysis on the dominant surface proteins of the six viruses in our dataset. Although surface proteins are typically targets of antibody responses that are mediated by HLA class-II molecules (Bouche et al., 2002; de Swart et al., 2009), it is shown that MeV hemagglutinin molecule is also a main target of HLA class-I restricted immune responses (Ota et al., 2007; I. M. Schellens et al., 2015). Interestingly, a significant difference between populations in the number of HLA-B bound peptides was observed only for the MeV (Kruskal-Wallis test, $p_{adj}<0.01$, **Supplementary Figure 11**). Indeed, pairwise comparisons of populations using Dunn's test indicated that a higher number of MeV hemagglutinin peptides were bound by individuals of the Modern Kiel population than both Early Neolithic Farmers ($p<0.01$) and Late Neolithic Farmers ($p<0.01$). We further investigated this difference by generating 1000 random protein sequences having the same length with the MeV hemagglutinin and the same amino acid composition with viral surface peptides in our dataset. We analyzed these artificial proteins similarly to the real surface proteins and calculated the difference between the median number of bound peptides by Modern Kiel, ENF and LNF populations, essentially generating a randomized distribution of differences between populations against which the observed difference obtained from MEV H protein can be

compared. While the observed differences between Modern Kiel and ENF and Modern Kiel and LNF lie on the higher end of the distribution, both were below the 97.5th percentile **(Supplementary Figure 12)**.

Material and Methods

aDNA was extracted from 265 individuals and HLA genotyping was performed following a previously developed method (Pierini et al., 2020). In total 129 samples were successfully genotyped without missing data at the two-field resolution for HLA-A, HLA-B and HLA-C loci. All the analyses were performed on HLA genotypes with two-field resolution. Based on the dating, individuals from excavation sites of Altendorf (n=13), Niedertiefenbach (n=50), Rimbeck (n=3) and Warburg (n=17) were merged to form the Late Neolithic Farmers (LNF) metapopulation (n=83) and individuals from Fellbach – Oeffingen (n=16), Niederpoering (n=6) and Trebur (n=24) were merged to form the Early Neolithic Farmers (ENF) metapopulation (n=46).

The HLA dataset representing the modern populations was obtained from the 1000 Genomes Project. Five European, three East Asian and two African populations were included in the analysis for comparisons with ancient populations (**Supplementary Table 1**). We have also included an unpublished HLA typing data set of German individuals from Kiel (the Kiel cohort) as a representative of HLA profiles for the modern German population. Although 3219 individuals were genotyped within the Kiel cohort, we randomly selected 100 individuals in order to keep sample sizes of populations similar within our dataset.

HLA diversity within populations were analyzed with three metrics namely heterozygosity, the Shannon Diversity index and allelic richness. Observed heterozygosity is calculated as the proportion of heterozygous individuals within a population. Shannon index is calculated as: $H' = -\Sigma Pi \cdot \ln(Pi)$; where Pi denotes the allele frequency of the *i*th allele. *diversity* function from the *vegan* R package was used to compute Shannon index values of populations. Allelic richness based on the rarefaction method and observed heterozygosity were calculated using the *hierfstat* package in R (Goudet, 2005).

Principal components analysis were performed on an allele count table where rows correspond to individuals and the columns correspond to alleles with each cell containing a value of 0, 1 or

2 indicating how many time the allele is observed in an individual. Principal components were calculated using the *prcomp* function from the *stats* package in R version 4.1.0. However, due to the scaling of the count table during the calculation of principal components, individuals carrying multiple extremely rare alleles may distort principal components. Therefore we removed individuals whose genotype consists of alleles that are observed only once in the entire dataset (i.e. two different alleles only observed in one individual). Based on this criterion, two individuals from LNF and one from ENF were removed from the PCA.

Pairwise population divergence values were estimated as Weir-Cockerham F_{ST} using the *adegenet* and *hierfstat* packages in R. Allele frequencies were obtained by direct counting. 95% confidence intervals for the estimation of allele frequencies were calculated following Mack et al. (2012). Allele divergence was calculated as the Grantham distance between the sequences of the 2nd and the 3rd exons of two HLA alleles.

We analyzed differences in predicted peptide binding among three populations, ENF, LNF and modern Kiel cohort by using netMHCpan4.1 (Reynisson et al., 2020). Binding predictions were produced by using 0.5% RankEL threshold (i.e. strong binders) for each allele and 13 human pathogens (7 bacteria and 6 viruses). The total number of peptides bound by an individual is calculated as the combination of peptide pools of two alleles that the individual carries. Differences between populations were analyzed with the Kruskal-Wallis test with Bonferroni multiple testing adjustment for the number of different pathogens and three HLA loci. We repeated the same approach for only the surface peptides of six viruses. Surface peptides were identified using information from the literature **(Supplementary Table 2)**

Discussion

In this study, we analyzed HLA class-I genotypes of 129 ancient individuals from central Europe dating between 5000 BCE to 3300 BCE together with a population representing modern Germany and 10 modern populations of 1000 Genomes Project from Europe, East Asia and Africa. Based on the dating of the archaeological sites, ancient populations were grouped into Early Neolithic farmers (ENF) and Late Neolithic farmers (LNF). For each locus, heterozygosity was slightly decreased in Neolithic populations compared to modern Europeans. Decreased heterozygosity might indicate a genetic bottleneck, which would fit well with the history of stepwise migration of European Neolithic farmers from Anatolia (Marchi et al.,

2022). However, the results of Marchi et al. (2022) that were based on genomewide SNP markers do not indicate any significant bottleneck effect in early European farmers as they seem to harbor high neutral heterozygosity, shorter runs of homozygosity and maintain high effective population size. Furthermore, a population bottleneck is expected to decrease the number of alleles in a population much more than it does heterozygosity (Allendorf, 1986; Nei et al., 1975) but the allelic richness in Neolithic farmers is comparable to modern populations. Therefore, it is unlikely that the decreased heterozygosity is caused by a bottleneck. A decrease in heterozygosity might also be the result of a recent directional selection on a few alleles increasing their frequency. Interestingly, we observed a decreased Shannon Index value compared to allelic richness for Neolithic populations. This is possibly driven by the uneven distribution of alleles within these populations, skewed towards a few high-frequency alleles. Whether these alleles were subject to recent selection exerted by the novel farming practices or increased in frequency due to neutral events would require further sampling efforts to illuminate. It should also be noted that during the genotyping, some alleles might be missed due to the high degradation of ancient DNA, which might result in homozygous allele calls in originally heterozygous individuals. This may also contribute to the decrease in heterozygosity in Neolithic populations.

Both PCA and Fst analysis showed that ancient Europeans are more closely related to modern Europeans compared to modern African and East Asian populations. This result is not unexpected given that the time difference between the ancient and modern Europeans is around 5000 to 7000 years while the divergence between modern Europeans, Africans and East Asians in our dataset is estimated to be much earlier than that (Nielsen et al., 2017; Seguin-Orlando et al., 2014). While this observation highlights the role of the demographic history of populations in shaping the worldwide HLA genetic diversity following out-of-Africa migrations (Buhler & Sanchez-Mazas, 2011; Prugnolle, Manica, Charpentier, et al., 2005), there was considerable population differentiation between ancient and modern Europeans as well. Two non-mutually exclusive mechanisms that can account for such differentiation are population admixture and natural selection favoring specific HLA alleles. Admixture is known to play a significant role in shaping the genetic structure of European populations (Chintalapati et al., 2022; Günther & Jakobsson, 2016). In fact, population genomics studies have established that modern Europeans carry three main ancestry components namely the Western Hunter-Gatherers, occupying the continent before the arrival of the farmers, the European Neolithic farmers who were descendants of the first farmers from Anatolia and the steppe (or Yamnaya) ancestry that arrived

into Europe around the end of the Neolithic (Allentoft et al., 2015; Haak et al., 2015; Kılınç et al., 2016; Lazaridis, 2018; Lazaridis et al., 2014; Lipson et al., 2017; Marchi et al., 2022). Therefore, differentiation between the modern and Neolithic Europeans might be the result of steppe ancestry that is not observed in Neolithic farmers. Although the lack of HLA genotyping from individuals of Yamnaya culture precludes direct testing of this hypothesis, an indirect approach can be taken by comparing the modern European populations. Steppe ancestry was shown to exhibit a north to south cline within Europe and southern European populations carry lower proportions of steppe ancestry and higher proportions of Neolithic ancestry (Haak et al., 2015). Accordingly, HLA alleles introduced by the steppe migrants would be expected to follow this cline. Within that context, two sets of alleles that are either missing or observed at very low frequencies in Neolithic Europe would be of interest. The first set of alleles is A*03:01, B*07:02 and C*07:02 with lower frequencies in IBS and TSI compared to the remaining modern European populations in our dataset. These alleles are part of the conserved extended haplotype 7.1 (Dorak et al., 2006). Conserved extended haplotypes (CEHs) are large segments of conserved sequences covering classical MHC genes as well as other genes across the MHC region of chromosome 6 (Alper et al., 2006; Degli-Esposti et al., 1992). The evolution of CEHs has been a focus of discussion for a long time as some of them are observed at high frequencies with geographically restricted distributions suggesting a recent origin followed by a rapid expansion potentially due to positive selection increasing their frequency (Alper et al., 2006; Dawkins & Lloyd, 2019; Walsh et al., 2003). The second set of alleles of interest is A*01:01, B*08:01, C*07:01 that makes up the CEH 8.1. CEH 7.1 and 8.1 are among the most common haplotypes within Europe and both were shown to exhibit a correlation with latitude, highlighting the north-south cline in their frequency distribution similar to the steppe ancestry proportion within populations (Sanchez-Mazas et al., 2014). Therefore, both of these haplotypes might be introduced into Europe after the Neolithic as a part of the steppe ancestry. In support of this hypothesis, the frequencies of B*07 and B*08 alleles were significantly correlated with the steppe ancestry within Europe **(Supplementary Figure 13)**.

Genetic clines observed for HLA alleles and haplotypes were previously suggested to result from demic diffusion of Neolithic farmers into Europe with the assumption that alleles common in modern central-northern Europe were also common in Western Hunter-Gatherer populations (Menozzi et al., 1978; Nowak et al., 2008). However, common alleles of Western Hunter-Gatherers should be observed in Late Neolithic Farmers as a result of the admixture between the local hunter-gatherers and farmers which was minimal during the initial phases of Neolithic

dispersal into Europe but increased towards the late Neolithic (Lazaridis, 2018). Although appreciable frequencies of B*08:01 in Late Neolithic Farmers in our dataset might result from such admixture, none of the LNF individuals with B*08:01 carries A*01:01 or C*07:01 which are part of the modern CEH 8.1. Therefore, while it remains possible that B*08:01 constituted a small part of the HLA allele pool of European hunter-gatherers and introgressed into farmer populations, CEH 8.1 haplotype did not exist within Western Hunter-Gatherers. This observation does not support the hypothesis that the north to south cline pattern of CEH8.1 was formed due to Neolithic expansion. An interesting pair of alleles in that context is B*40:01 and C*03:04, both absent in ENF yet reaching up to 9% and 7% frequencies respectively in LNF **(Figure 4)**. This increase in frequency might be the result of the admixture of farmers with local hunter-gatherers throughout the Neolithic. Both alleles remain at high frequency in modern north and central European populations but low in southern populations. This result further supports the possibility that these alleles were introduced by local hunter-gatherers as populations from modern Spain and Italy usually harbor lower hunter-gatherer ancestry compared to northern European populations (Haak et al., 2015).

An important point to consider regarding the origins of CEH 8.1 is that this haplotype also exhibits east-west cline, mainly driven by its high frequency in the British Isles (Sanchez-Mazas et al., 2014). This observation is unexpected if CEH 8.1 had spread into Europe via steppe pastoralists from the east. However, Olalde et al., (2018) have argued that steppe-related ancestry was introduced into Britain by individuals associated with Bell Beaker Complex culture which resulted in the replacement of around 90% of the local Neolithic population. Therefore, the special demographic history of the British Isles might explain the high frequency of CEH 8.1 observed there.

Another set of alleles with intriguing diversity patterns are A*02:01, B*44:02 and C*05:01 which form the CEH 44.1 (Dorak et al., 2006). They are observed at similar frequencies in both Neolithic farmers and modern European populations. Furthermore, CEH 44.1 does not exhibit any cline pattern within Europe (Sanchez-Mazas et al., 2014) and is considered to be a Caucasian-specific haplotype (Degli-Esposti et al., 1992). Such persistence of the alleles of CEH 44.1 deserves some attention considering that several other alleles with similar frequencies in Neolithic farmers such as B*39:01 or C*15:02 exhibit a decrease in modern European populations. Interestingly, considering four diseases; AIDS, type 1 diabetes, multiple sclerosis and type 1 autoimmune hepatitis, CEH 44.1 (or one of the alleles within that haplotype) appears to have contrasting associations compared to CEH 7.1 or 8.1

(Supplementary Table 3). For example, HLA-DRB1*15:01 allele is part of the CEH 7.1 and it is commonly associated with the risk of developing multiple sclerosis whereas A*02:01, B*44:02 and C*05 alleles (all part of CEH 44.1) were independently shown to be protective (Hollenbach & Oksenberg, 2015). Similar observation was also made for HIV infection where rapid progression to AIDS was associated with CEH 8.1 and slow progression was associated with CEH 44.1 (Flores-Villanueva et al., 2003). Furthermore, we found that B*44:02 was the most divergent common Neolithic allele compared to B*07:02 and B*08:01. Functionally divergent phenotypes formed by the combination of these alleles may allow wider recognition of pathogens due to more diverse peptide presentation by HLA molecules. Based on these observations, it can be argued that CEH 44.1, CEH 7.1 and CEH 8.1 play contrasting/complementary roles in disease resistance that lead to the maintenance of CEH 44.1 within Europe via balancing selection in the face of the increased frequencies of CEH 7.1 and 8.1 since the late Neolithic.

While the admixture between Neolithic farmers and steppe pastoralists can account for the changes in frequencies of several alleles in modern European populations to some extent, the effect of selection seems to be more subtle based on our results. Although we found significant differences in allele divergence between ENF and Modern Kiel populations, these differences were not consistent across loci. When excluding the homozygous individuals within the dataset, genotypes of Early Neolithic Farmers were more divergent than modern Kiel individuals for HLA-A while the opposite was true for HLA-C. The observation that all three loci exhibit different and even opposing divergence patterns through time suggests that each locus evolves under somewhat different selective pressures. However, it is also possible that the time since the Neolithic was not long enough to observe the result of selection in the allele divergence. Interestingly, no difference between populations was observed for HLA-B. At first sight, this might seem contradictory to the results of Pierini & Lenz (2018) who showed that the HLA-B locus exhibits the strongest signal for divergent allele advantage. However, our results merely suggest that such selective force on modern populations was already acting on Neolithic populations as well and divergent HLA-B allele pools were maintained despite differences in specific alleles.

The analysis of peptide binding properties of ancient and modern allele pools supports the idea that the functional diversity of HLA alleles within populations was similar. With the exception of *Human polyomavirus 2*, we found no significant differences in the number of bound viral or bacterial peptides by individuals from Neolithic and modern Kiel populations. Interestingly

when restricting our analysis to the surface proteins of viruses, modern Kiel population appears to bind more peptides from the *Measles virus* hemagglutinin protein. However, the median difference between populations was small, only around two peptides and an independent analysis of the MeV hemagglutinin protein did not show a significantly higher binding by modern Kiel populations compared to randomized peptides. Therefore, contrary to our initial expectation, we did not observe a strong effect of measles on HLA peptide binding. Indeed, MeV is one virus among many and despite its devastating impact on humans, the signals of adaptation could be subtle and not visible with our analysis. Furthermore, our approach only relies on quantitative differences in peptide binding with the assumption that binding more peptides from a pathogen is advantageous for the host. While this might be true to some extent, as Croft et al. (2019) have shown that most peptides presented by HLA class-I molecules can be immunogenic, it is also certainly not the complete picture of adaptive immune responses. Immunodominance is a well-known phenomenon in anti-viral responses and it refers to the observation that despite the existence of many potential antigens, most of the cytotoxic T-cell responses within an individual are usually focused on a few immunodominant peptides (Yewdell, 2006). Different HLA allele pools between Neolithic and modern populations might result in different immunodominant peptides presented to T-cells. However, it is not possible to identify which peptides are immunodominant based on HLA variation or peptide binding prediction algorithms. To what extent such qualitative differences in peptide binding between modern and Neolithic populations affect disease outcomes is not clear and certainly deserves further investigation.

In summary, we generated and analyzed a large dataset of ancient HLA genotypes of the first farmers of central Europe dating back to the early and late Neolithic. Our results reveal large changes in frequencies of several alleles within Europe since the Neolithic, yet despite such differentiation, functional diversity in both Neolithic and modern populations appears to be similar. Detailed comparison of alleles with large frequency changes suggests a considerable role of population admixture with steppe pastoralists shaping the modern distribution of HLA alleles within Europe. Further ancient HLA genotypes will allow testing our conclusions presented here and illuminate the evolution of HLA diversity within Europe.

Supplementary Material for Chapter 3 is provided in Annex V.

Conclusion

The overarching aim of the work presented in this thesis was to shed further light on the effect of pathogens on human HLA diversity. The mechanistic basis of the relationship between pathogens and HLA molecules is relatively clear; presentation of pathogen-derived peptides by HLA molecules initiates the adaptive immune responses. However, the evolutionary implications of this relationship are still under investigation. Collected under the umbrella term of pathogen-mediated balancing selection, several mechanisms of selection that could account for the high HLA diversity were proposed. These mechanisms are the heterozygote advantage, fluctuating selection and negative frequency-dependent selection. All mechanisms essentially rely on the assumption that HLA alleles raise differential responses to different pathogens depending on the peptides presented by the allele. Therefore, each pathogen selects for distinct HLA alleles, resulting in higher diversity within populations than expected from neutral evolution. This assumption is supported mainly by the limited number of associations between specific HLA alleles and disease phenotypes such as viral load at a specific stage of the infection. Within that context, in the first chapter of the thesis, we analyzed a set of human pathogens and took advantage of the recent developments in the computational prediction of HLA-peptide binding in order to compare the binding properties of common HLA alleles. We showed that the peptidome of each pathogen mostly consists of peptides unique to the species. This observation highlights the immense diversity of pathogens that humans need to deal with in order to survive. Although this diversity is a commonly held prior in infectious disease research, our results are the first systematic empirical data uncovering the extent of the diversity and essentially provide the basis for the long-lasting assumption that distinct pathogens exert distinct selective pressures on HLA genes. Based on this empirical confirmation, we investigated how HLA alleles are associated with peptidomes of distinct pathogens. For each HLA allele, we revealed a vast variation in the number of bound peptides from different pathogens, further supporting the idea that HLA alleles raise differential responses to distinct pathogens. This variation was much more pronounced for HLA alleles with smaller peptide repertoires. Therefore, our results support the hypothesis that alleles with small peptide repertoires evolve as “specialists” towards one or a few pathogens while the other alleles are generalists providing resistance to a larger number of pathogens. The suggested specialization of HLA alleles for some pathogens in our analysis remains to be tested with different approaches such as genetic association studies. The specialization metric that we developed can be used to generate and test hypotheses regarding the relationship between HLA alleles and

pathogens. Nevertheless, such specialization scenario is not unlikely under a fluctuating selection regime where novel and sometimes deadly pathogens are introduced into populations occasionally. Considering the recently emerged pathogens such as HIV, Ebola virus or multiple coronaviruses within the last decades, it can be argued that spillovers from animals are not rare events (Karlsson et al., 2014). Indeed, several studies have revealed signatures within the human genome, possibly left by ancient outbreaks of viruses that are related to the ones that we face today such as coronaviruses (Souilmi et al., 2021) or retroviruses (Kaiser et al., 2007). However, these signatures are not localized on HLA, so the relationship between HLA and particular pathogens remains an open question. It is hypothesized that differences in pathogen composition in time and space may result in rapid turnover of HLA alleles within populations or differentiation in HLA genes between populations, respectively (Spurgin & Richardson, 2010). On the other hand, the well-known observation of trans-species polymorphism, i.e. the existence of highly similar MHC alleles within two species that diverged millions of years ago, suggests that the hypothesized turnover of alleles does not necessarily result in the loss of diversity. A further complication of the matter arises from the fact that suggested mechanisms of balancing selection may, in fact, result in opposing patterns of diversity within populations. For example, differentiation between two populations increases under negative frequency-dependent selection regime only when rare alleles are novel, i.e. generated after population split. If rare alleles are the old ones, either inherited from the ancestral population or introgressed from other populations, differentiation will be decreased due to maintenance of the old alleles in each population. Similarly, fluctuating selection might result in changes in allele frequencies over time, yet the population differentiation will only increase if pathogen pools that populations harbor differs.

As the next step, we attempted to investigate the complex selection regimes acting on HLA diversity in collaboration with several colleagues. We approached the problem from two angles namely the diversity across space (chapter II) and the diversity through time (chapter III). Although the HLA variation patterns within and between extant human populations are analyzed by many researchers, an overwhelming majority of these studies include mostly populations from Europe or North America. As the cradle of our species, Africa accommodates populations harboring the highest genetic diversity among humans. Underrepresentation of populations with non-European ancestry in immunogenomics studies, specifically the African populations, hinders attempts to generate appropriate evolutionary analysis while also imposing challenges on medical interventions such as vaccinations or personalized therapeutics. In the

second chapter, we investigated the HLA diversity of 12 populations from sub-Saharan Africa together with 10 populations from the 1000 Genomes Project using both the SNP data generated by whole genome sequencing and the HLA allele data generated by targeted HLA sequencing. SNP based analysis confirmed the extensive nucleotide diversity observed within HLA genes. The HLA nucleotide diversity slightly decreases in non-African populations but not to the extent that neutral genetic diversity decreases. HLA genes were enriched in old mutations and SNPs within HLA genes do not show higher differentiation than neutral SNPs. Finally, two independent tests of selective neutrality namely Tajima's D and LD-ABF, reveal strong signals of balancing selection at the HLA genes. Similar results were also obtained from the analysis of the targeted HLA data. HLA diversity is comparable across populations with the exception of a few African populations with a recent history of population bottleneck. The differentiation patterns appear to have arisen from demographic processes following the out-of-Africa migrations as populations from the same continent usually exhibit lower differentiation compared to populations from the other continents. The differentiation between populations inhabiting different continents is mainly the result of alleles that were shared yet observed at different frequencies between populations. In line with this observation, despite the immense genetic diversity within African populations, we found only a few novel HLA alleles. Furthermore, differences in allele frequencies do not change the overall functional breadth of peptide binding as each population exhibits similar levels of allele divergence and peptide-binding promiscuity. In other words, both the genetic diversity at the sequence level and the functional diversity at the phenotypic level are maintained across populations. Overall, these results show that old variants within the MHC are maintained in populations by balancing selection. It should be emphasized that the aim of this study was not to analyze worldwide diversity patterns in HLA, but rather to analyze patterns within African populations. Therefore, non-African populations were selected simply to put the diversity within Africa into a wider perspective. We claim neither that these populations represent their respective continents, nor that every other population exhibits similar diversity patterns. In fact, a previous analysis revealed HLA alleles that are observed only in a population or region (Vina et al., 2012). These alleles are more likely to be observed in isolated populations that experienced a recent genetic bottleneck such as Native Americans and possibly derive from a more common and widespread allele via point mutations. It is worth investigating whether such novel alleles increase in frequency in response to local pathogen pressures, especially if they differ from the ancestral allele across the peptide-binding region. Nevertheless, our analysis suggests that such unique alleles are exceptions rather than the norm in HLA evolution and that balancing selection

maintaining old variation together with the demographic history of each population predominantly shapes HLA diversity. Independent of the unique alleles, it can also be claimed that frequency differences among shared alleles between populations may also arise due to local adaptation. Simulations can be employed to investigate whether neutral evolution can explain these differences. However, it is not easy to generate realistic simulations without a sound knowledge of the demographic history of each population. Furthermore, identifying a "local pathogen" that presumably results in HLA adaptation is much easier said than done. As Dunn et al., (2010) put nicely, "*The fact that warbler species distributions are better understood than the distribution of human pathogens is a gap that clearly deserves research attention.*" This is one of the main reasons why it has been a daunting task to pinpoint specific pathogens that presumably drive HLA evolution.

This problem can potentially be circumvented by an approach taking the historical populations into account. Many deadly pathogens that had a major impact on our species have recent origins within a few thousand years (Wolfe et al., 2007). Standing variation in HLA genes may have resulted in a rapid rise of frequencies of alleles providing resistance against these pathogens. Therefore, if we can identify the approximate time point of the emergence of such deadly pathogens, it would be possible to track changes in HLA allele frequencies and identify any alleles associated with protective or susceptibility effects. Such an approach would be promising yet suffers from two major drawbacks. First, human populations are not frozen units living within a confined geographic area indefinitely. On the contrary, the history of populations is shaped by migrations, population bottlenecks and admixtures. Therefore, any study aiming to compare allele frequencies between two populations separated by long time intervals must take the population history into account in order to prevent false positive results. This problem is being ameliorated mainly as a result of the rapidly developing archaeogenomic field. Thanks to the efforts of geneticists, archaeologists and anthropologists, details of the mobility patterns of historical populations around the world are being uncovered, providing a baseline for testing hypotheses of selection on specific genes (Koptekin et al., 2023). The second drawback is that the ancient DNA is, by nature, marked with high degradation and contamination. This is especially problematic for HLA genes with a high concentration of SNPs across a very short span in the genome. For example, reliable mapping of reads has been challenging for HLA loci because if the allele of the sample is different from the reference allele, reads might be discarded due to the high number of differences between the read and the reference (Brandt et al., 2015). This does not only bias genotype calls towards the reference allele but also makes it more likely

to miss alleles that were common in the past but are very rare or even extinct today. Additionally, as a result of deamination, the incorporation of incorrect nucleotides during the sequencing process is commonly observed damage in ancient samples (Briggs et al., 2007; Dabney et al., 2013). Although this kind of damage usually follows a pattern such as increased C to T transitions towards the end of reads, it remains possible that damaged reads are mapped, resulting in erroneous HLA genotype calls. Therefore, accurate genotyping of HLA alleles from ancient specimens requires a special effort. The "Targeted Analysis of sequencing Reads for GenoTyping" (TARGT) pipeline is developed for analyzing low coverage shotgun sequence data, specifically to overcome the problem of HLA genotyping in ancient samples (Annex 1). The accuracy of this method was demonstrated for both modern and ancient DNA samples. In the third chapter, we employed the TARGT pipeline to genotype HLA class-I alleles of 129 individuals who lived in the early and late Neolithic periods (ENF and LNF, respectively) in central Europe. Diversity patterns suggest that despite migrating recently from Anatolia and the Balkans into central Europe, they do not show signs of genetic bottleneck and HLA diversity is comparable to modern populations. However, due to large frequency differences in alleles, Neolithic populations appear to be significantly differentiated from modern Europeans. In fact, the majority of the most common HLA alleles in modern central and northern Europe are not even observed in the Neolithic populations. It is tempting to invoke adaptation to changing pathogen landscape throughout and after the Neolithic as the reason for differences in allele frequencies. However, as mentioned above, the demographic history of populations usually has an immense effect on population structure. The alleles of the two most common HLA haplotypes, namely *A*03:01-B*07:02-C*07:02* (CEH7.1) and *A*01:01-B*08:01-C*07:01* (CEH8.1) exhibit a declining frequency cline from north to south in Europe. It could be argued that this cline is the result of the Neolithic expansion into Europe when farmers following either the land route through the Danube River or the sea route across the Mediterranean Sea pushed the local hunter-gatherers towards the north. In that case, the alleles of CEH 7.1 and 8.1 should have been part of the local hunter-gatherer allele pool and later remained within European populations as a result of admixture with farmers. However, it is known that although the admixture between the first farmers and hunter-gatherers was minimal, this has changed dramatically throughout the Neolithic. Therefore, based on this hypothesis, we should have observed CEH 7.1 and 8.1 at appreciable frequencies in the Late Neolithic when admixture with hunter-gatherers was prevalent. This is not the case with the exception of a few *B*08:01* alleles observed in LNF. Two alleles, namely *B*40:01* and *C*03:04*, conform to that pattern with a sharp increase from zero in ENF to around 8% frequency in LNF and modern Europeans. While

this may be the result of admixture with hunter-gatherers, another admixture event can better explain the high frequency of CEH7.1 and 8.1 in modern Europeans. Beginning around the end of the Neolithic, a large migration of Yamnaya pastoralists from the Pontic-Caspian Steppe into Europe, and their admixture with local farmers, reshaped the genetic structure of European populations. The traces of this admixture event within the genomes of modern Europeans reveal that steppe ancestry is observed in Europe with a north-south cline, similar to the distribution of the alleles of CEH7.1 and 8.1. Therefore, we hypothesize that these haplotypes were part of the steppe ancestry and increased in Europe as the result of the admixture with steppe pastoralists. This hypothesis remains to be tested by HLA genotyping of steppe pastoralists or their direct descendants in Europe. Meanwhile, an alternative explanation for the rise of these haplotypes could be their selective advantage against emerging pathogens during and after the Neolithic. We tested this hypothesis by using peptide-binding prediction algorithms, with a specific focus on the *Measles virus*. Interestingly, we found that when comparing the number of bound peptides from the *Measles virus*, Neolithic populations do not differ from modern Europeans. This observation remains unchanged for several other viruses and bacteria, suggesting that the functional diversity of alleles in Neolithic populations was not different from those within modern Europeans. These results do not support the hypothesis that major allele frequency differences between ancient and modern Europeans were the result of adaptation to changing pathogen patterns. It appears that the recent evolution of HLA alleles within populations is mainly driven by genetic drift and admixture events.

How can the conclusions of the first chapter, which highlights the specific associations between pathogens and HLA alleles, be conciliated with the conclusions of the second and the third chapters suggesting that population history is the major driver of allele frequencies? If the HLA diversity is mainly shaped by genetic drift and admixture, why there are HLA alleles that seem to be specialized against some pathogens? Firstly, the dominant effect of population history on HLA diversity does not exclude the possibility of local adaptation driven by resistance alleles. However, the relative selective advantage of alleles might be small due to the polygenic nature of HLA in humans. Humans can carry up to six different HLA class-I alleles in their genome. Therefore, the combined effect of different alleles might protect an individual even in the absence of one specific resistance allele. This is not the case, for example, in chickens with a single majorly expressed HLA class-I gene and alleles of this gene are strongly associated with resistance to diseases (Kaufman, 2018a). Secondly, humans are always exposed to multiple pathogens, sometimes even simultaneously. Therefore, the selective advantage of some alleles

might not be visible due to antagonistic selection pressure on the same allele by different pathogens. Unless there exists one perfect HLA allele against all pathogens, such tradeoffs may mask the effect of directional selection on distinct alleles, with the end result being the balanced polymorphism of multiple alleles within populations. An exception to this scenario may be deadly pandemics that humans have experienced several times throughout recent history. Any HLA allele providing resistance or susceptibility to the pandemic pathogen can be expected to rise or decrease, respectively, in frequency in a relatively short time. On the other hand, it is also possible that the disease outcome is simply not associated with the HLA genotype. An example is the latest pandemic that humans faced, namely the COVID-19 pandemic. We analyzed 1980 COVID-19 patients and 2205 healthy participants and did not detect any association between the HLA genotype and the COVID-19 severity (Annex II). Considering the complexity of immune responses as well as the diversity of the mechanisms of pathogen infection, it is not surprising to find cases when the HLA variation is not among the main determinants of the disease outcome. A third explanation of why major allele frequency differences appear to be the result of population history rather than adaptation to pathogens might require questioning the role of the Neolithic transition for the pathogen landscape. While it is clear that large and dense communities are necessary for the persistence of highly virulent pathogens, such as *Measles virus*, within populations, hunter-gatherers before the Neolithic were certainly not small, atomized tribes (Black, 1966; Houldcroft & Underdown, 2023). Archaeogenetic data suggest that high effective population sizes were maintained in hunter-gatherers, possibly as a result of mate exchange networks (Sikora et al., 2019). Furthermore, gatherings of large groups in the form of ritual feasting were probably a common practice to form social bonds (Liu et al., 2018). Such conditions could support the spread of pathogens albeit not at the level of a pandemic. Local outbreaks of pathogens that are similar to those affecting humans today might have been a common occurrence. In other words, the fact that a significant number of pathogens that persist in modern human populations have recent origins does not mean that similar pathogens did not infect humans before the Neolithic. Spillover events could have led to flare-ups of diseases within populations before dying out due to the lack of conditions supporting their persistence. Therefore, it can be hypothesized that the pathogen landscape and diversity that humans dealt with after the Neolithic was not completely different from what it was before. Such hypothesis may support the observation that major changes in HLA allele frequencies through time and space are the result of population admixtures and genetic drift, rather than adaptation to pathogens.

In summary, this thesis presents novel findings on the mechanisms by which the diversity of major histocompatibility complex (MHC) genes in humans is maintained and shaped. Extreme peptidome diversity of pathogens provides an empirical basis for the long-held assumption that HLA alleles evolve in response to distinct pathogen pressures. On the other hand, our results reveal that the recent evolution of HLA diversity within populations is mainly driven by balancing selection maintaining diversity and genetic drift and population admixtures resulting in differentiation through time and space. These results highlight the value of intensive work on characterizing the HLA diversity of ancient and extant populations.

References

Abate, G., Hamzabegovic, F., Eickhoff, C. S., & Hoft, D. F. (2019). BCG Vaccination Induces M. avium and M. abscessus Cross-Protective Immunity. *Frontiers in Immunology*, *10*, 234. https://doi.org/10.3389/fimmu.2019.00234

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., & Wu, C. J. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*, *46*(2), 315–326. https://doi.org/10.1016/j.immuni.2017.02.007

Abi-Rached, L., Gouret, P., Yeh, J.-H., Cristofaro, J. D., Pontarotti, P., Picard, C., & Paganini, J. (2018). Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS ONE*, *13*(10), e0206512. https://doi.org/10.1371/journal.pone.0206512

Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F. A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S. G. E., Maiers, M., Guethlein, L. A., Tavoularis, S., … Parham, P. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science (New York, N.Y.)*, *334*(6052), 89–94. https://doi.org/10.1126/science.1209202

Albers, P. K., & McVean, G. (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biology*, *18*(1), e3000586. https://doi.org/10.1371/journal.pbio.3000586

Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, *5*(2), 181–190. https://doi.org/10.1002/zoo.1430050212

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. D., Dąbrowski, P., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 7555. https://doi.org/10.1038/nature14507

Alper, C. A., Larsen, C. E., Dubey, D. P., Awdeh, Z. L., Fici, D. A., & Yunis, E. J. (2006). The Haplotype Structure of the Human Major Histocompatibility Complex. *Human Immunology*, *67*(1), 73–84. https://doi.org/10.1016/j.humimm.2005.11.006

Apanius, V., Penn, D., Slev, P. R., Ruff, L. R., & Potts, W. K. (1997). The Nature of Selection on the Major Histocompatibility Complex. *Critical Reviews™ in Immunology*, *17*(2), 179–224. https://doi.org/10.1615/CritRevImmunol.v17.i2.40

Armelagos, G. J., & Harper, K. N. (2005). Genomics at the origins of agriculture, part two. *Evolutionary Anthropology: Issues, News, and Reviews*, *14*(3), 109–121. https://doi.org/10.1002/evan.20048

Armstrong, G. L., Conn, L. A., & Pinner, R. W. (1999). Trends in infectious disease mortality in the United States during the 20th century. *JAMA*, *281*(1), 61–66. https://doi.org/10.1001/jama.281.1.61

Arora, J., McLaren, P. J., Chaturvedi, N., Carrington, M., Fellay, J., & Lenz, T. L. (2019). HIV peptidome-wide association study reveals patient-specific epitope repertoires associated with HIV control. *Proceedings of the National Academy of Sciences*, *116*(3), 944–949. https://doi.org/10.1073/pnas.1812548116

Arora, J., Pierini, F., McLaren, P. J., Carrington, M., Fellay, J., & Lenz, T. L. (2020). HLA heterozygote advantage against HIV-1 is driven by quantitative and qualitative differences in HLA allele-specific peptide presentation. *Molecular Biology and Evolution*, *37*(3), 639–650. https://doi.org/10.1093/molbev/msz249

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 7571. https://doi.org/10.1038/nature15393

Baker, T. S., Newcomb, W. W., Olson, N. H., Cowsert, L. M., Olson, C., & Brown, J. C. (1991). Structures of bovine and human papillomaviruses. Analysis by cryoelectron microscopy and three-dimensional image reconstruction. *Biophysical Journal*, *60*(6), 1445–1456.

Bell, G., & Collins, S. (2008). Adaptation, extinction and global change. *Evolutionary Applications*, *1*(1), 3–16. https://doi.org/10.1111/j.1752-4571.2007.00011.x

Bentley, A. R., Callier, S., & Rotimi, C. N. (2017). Diversity and inclusion in genomic research: Why the uneven progress? *Journal of Community Genetics*, *8*(4), 255–266. https://doi.org/10.1007/s12687-017-0316-6

Benton, M. L., Abraham, A., LaBella, A. L., Abbot, P., Rokas, A., & Capra, J. A. (2021). The influence of evolutionary history on human health and disease. *Nature Reviews Genetics*, *22*(5), 269–283. https://doi.org/10.1038/s41576-020-00305-9

Bitarello, B. D., de Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andrés, A. M. (2018). Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., & Wiley, D. C. (1987). The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature*, *329*(6139), 512–518. https://doi.org/10.1038/329512a0

Black, F. L. (1966). Measles endemicity in insular populations: Critical community size and its evolutionary implication. *Journal of Theoretical Biology*, *11*(2), 207–211. https://doi.org/10.1016/0022-5193(66)90161-5

Blackwell, J. M., Jamieson, S. E., & Burgner, D. (2009). HLA and infectious diseases. *Clinical Microbiology Reviews*, *22*(2), 370–385. https://doi.org/10.1128/CMR.00048-08

Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). Pathways of Antigen Processing. *Annual Review of Immunology*, *31*(1), 443–473. https://doi.org/10.1146/annurev-immunol-032712-095910

Bocquet-Appel, J.-P. (2011). When the World's Population Took Off: The Springboard of the Neolithic Demographic Transition. *Science*, *333*(6042), 560–561. https://doi.org/10.1126/science.1208880

Bodmer, W. F. (1972). Evolutionary Significance of the HL-A System. *Nature*, *237*(5351), 5351. https://doi.org/10.1038/237139a0

Borghans, J. A. M., Beltman, J. B., & De Boer, R. J. (2004). MHC polymorphism under host-pathogen coevolution. *Immunogenetics*, *55*(11), 732–739. https://doi.org/10.1007/s00251-003-0630-5

Borghans, J. A. M., Mølgaard, A., de Boer, R. J., & Keşmir, C. (2007). HLA Alleles Associated with Slow Progression to AIDS Truly Prefer to Present HIV-1 p24. *PLoS ONE*, *2*(9), e920. https://doi.org/10.1371/journal.pone.0000920

Bouche, F. B., Ertl, O. T., & Muller, C. P. (2002). Neutralizing B Cell Response in Measles. *Viral Immunology*, *15*(3), 451–471. https://doi.org/10.1089/088282402760312331

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3: Genes, Genomes, Genetics*, *5*(5), 931–941. https://doi.org/10.1534/g3.114.015784

Brandt, D. Y. C., 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. https://doi.org/10.1534/g3.118.200367

B-Rao, C. (2001). Sample Size Considerations in Genetic Polymorphism Studies. *Human Heredity*, *52*(4), 191–200.

Bremer, H. 1982. (1982). Variation of Generation Times in Escherichia coli Populations: Its Cause and Implications. *Microbiology*, *128*(12), 2865–2876. https://doi.org/10.1099/00221287-128-12-2865

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Brinkworth, J. F., & Barreiro, L. B. (2014). The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Current Opinion in Immunology*, *31*, 66–78. https://doi.org/10.1016/j.coi.2014.09.008

Brockhurst, M. A., Chapman, T., King, K. C., Mank, J. E., Paterson, S., & Hurst, G. D. D. (2014). Running with the Red Queen: The role of biotic conflicts in evolution. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1797), 20141382. https://doi.org/10.1098/rspb.2014.1382

Broecker, F., & Moelling, K. (2019). Evolution of Immune Systems From Viruses and Transposable Elements. *Frontiers in Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fmicb.2019.00051

Buhler, S., Nunes, J. M., & Sanchez-Mazas, A. (2016). HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*, *68*(6–7), 401–416. https://doi.org/10.1007/s00251-016-0918-x

Buhler, S., & Sanchez-Mazas, A. (2011). HLA DNA sequence variation among human populations: Molecular signatures of demographic and selective events. *PLoS ONE*, *6*(2), e14643. https://doi.org/10.1371/journal.pone.0014643

Burroughs, N. J., de Boer, R. J., & Keşmir, C. (2004). Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics*, *56*(5), 311–320. https://doi.org/10.1007/s00251-004-0691-0

Calis, J. J. A., de Boer, R. J., & Keşmir, C. (2012). Degenerate T-cell Recognition of Peptides on MHC Molecules Creates Large Holes in the T-cell Repertoire. *PLOS Computational Biology*, *8*(3), e1002412. https://doi.org/10.1371/JOURNAL.PCBI.1002412

Calis, J. J. A., Sanchez-Perez, G. F., & Keşmir, C. (2010). MHC class I molecules exploit the low G+C content of pathogen genomes for enhanced presentation. *European Journal of Immunology*, *40*(10), 2699–2709. https://doi.org/10.1002/eji.201040339

Cao, K., Moormann, A. M., Lyke, K. E., Masaberg, C., Sumba, O. P., Doumbo, O. K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., Single, R., Hartzman, R. J., Plowe, C. V., Kazura, J., Mann, D. L., Sztein, M. B., Thomson, G., & Fernández-Vina, M. A. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*, *63*(4), 293–325. https://doi.org/10.1111/j.0001-2815.2004.00192.x

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., & O'Brien, S. J. (1999). HLA and HIV-1: Heterozygote advantage and B*35-Cw*04 disadvantage. *Science (New York, N.Y.)*, *283*(5408), 1748–1752. https://doi.org/10.1126/science.283.5408.1748

Chappell, P., Meziane, E. K., Harrison, M., Magiera, L., Hermann, C., Mears, L., Wrobel, A. G., Durant, C., Nielsen, L. L., Buus, S., Ternette, N., Mwangi, W., Butter, C., Nair, V., Ahyee, T., Duggleby, R., Madrigal, A., Roversi, P., Lea, S. M., & Kaufman, J. (2015). Expression levels of mhc class i molecules are inversely correlated with promiscuity of peptide binding. *ELife*, *2015*(4), e05345. https://doi.org/10.7554/eLife.05345

Chintalapati, M., Patterson, N., & Moorjani, P. (2022). The spatiotemporal patterns of major human admixture events during the European Holocene. *ELife*, *11*, e77625. https://doi.org/10.7554/eLife.77625

Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, *311*(5765), 1283–1287. https://doi.org/10.1126/science.1123061

Clement, M., Snell, Q., Walke, P., Posada, D., & Crandall, K. (2002). TCS: Estimating gene genealogies. *Proceedings 16th International Parallel and Distributed Processing Symposium*, 7 pp. https://doi.org/10.1109/IPDPS.2002.1016585

Cohen, M. L. (2000). Changing patterns of infectious disease. *Nature*, *406*(6797), 762–767. https://doi.org/10.1038/35021206

Cooke, G. S., & Hill, A. V. S. (2001). Genetics of susceptibitlity to human infectious disease. *Nature Reviews Genetics*, *2*(12), 12. https://doi.org/10.1038/35103577

Cooper, M. D., & Alder, M. N. (2006). The Evolution of Adaptive Immune Systems. *Cell*, *124*(4), 815–822. https://doi.org/10.1016/j.cell.2006.02.001

Corona, E., Chen, R., Sikora, M., Morgan, A. A., Patel, C. J., Ramesh, A., Bustamante, C. D., & Butte, A. J. (2013). Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS Genetics*, *9*(5), e1003447. https://doi.org/10.1371/journal.pgen.1003447

Costa, A. F., Rao, X., LeChenadec, E., Baarle, D. van, & Keşmir, C. (2010). HLA-B molecules target more conserved regions of the HIV-1 proteome. *Aids*, *24*(2), 211–215. https://doi.org/10.1097/QAD.0b013e328334442e

Croft, N. P., Smith, S. A., Pickering, J., Sidney, J., Peters, B., Faridi, P., Witney, M. J., Sebastian, P., Flesch, I. E. A., Heading, S. L., Sette, A., Gruta, N. L. L., Purcell, A. W., & Tscharke, D. C. (2019). Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proceedings of the National Academy of Sciences*, *116*(8), 3112–3117. https://doi.org/10.1073/pnas.1815239116

Cullen, B. R. (2002). RNA interference: Antiviral defense and genetic tool. *Nature Immunology*, *3*(7), 597–599. https://doi.org/10.1038/ni0702-597

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7), a012567. https://doi.org/10.1101/cshperspect.a012567

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dawkins, R. L., & Lloyd, S. S. (2019). MHC Genomics and Disease: Looking Back to Go Forward. *Cells*, *8*(9), 944. https://doi.org/10.3390/cells8090944

de Swart, R. L., Yüksel, S., Langerijs, C. N., Muller, C. P., & Osterhaus, A. D. M. E. Y. 2009. (2009). Depletion of measles virus glycoprotein-specific antibodies from human sera reveals genotype-specific neutralizing antibodies. *Journal of General Virology*, *90*(12), 2982–2989. https://doi.org/10.1099/vir.0.014944-0

Degli-Esposti, M. A., Leaver, A. L., Christiansen, F. T., Witt, C. S., Abraham, L. J., & Dawkins, R. L. (1992). Ancestral haplotypes: Conserved population MHC haplotypes. *Human Immunology*, *34*(4), 242–252. https://doi.org/10.1016/0198-8859(92)90023-g

Di, D., Nunes, J. M., Jiang, W., & Sanchez-Mazas, A. (2021). Like Wings of a Bird: Functional Divergence and Complementarity between HLA-A and HLA-B Molecules. *Molecular Biology and Evolution*, *38*(4), 1580–1594. https://doi.org/10.1093/molbev/msaa325

Doherty, P. C., & Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, *256*(5512), 50–52. https://doi.org/10.1038/256050a0

Dorak, M. T., Shao, W., Machulla, H. K. G., Lobashevsky, E. S., Tang, J., Park, M. H., & Kaslow, R. A. (2006). Conserved extended haplotypes of the major histocompatibility complex: Further characterization. *Genes & Immunity*, *7*(6), 6. https://doi.org/10.1038/sj.gene.6364315

dos Santos Francisco, R., Buhler, S., Nunes, J. M., Bitarello, B. D., França, G. S., Meyer, D., & Sanchez-Mazas, A. (2015). HLA supertype variation across populations: New insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*, *67*(11–12), 651–663. https://doi.org/10.1007/s00251-015-0875-9

Drake, J. W., & Holland, J. J. (1999). Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences*, *96*(24), 13910–13913. https://doi.org/10.1073/pnas.96.24.13910

Dunn, R. R., Davies, T. J., Harris, N. C., & Gavin, M. C. (2010). Global drivers of human pathogen richness and prevalence. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1694), 2587–2595. https://doi.org/10.1098/rspb.2010.0340

Düx, A., Lequime, S., Patrono, L. V., Vrancken, B., Boral, S., Gogarten, J. F., Hilbig, A., Horst, D., Merkel, K., Prepoint, B., Santibanez, S., Schlotterbeck, J., Suchard, M. A., Ulrich, M., Widulin, N., Mankertz, A., Leendertz, F. H., Harper, K., Schnalke, T., … Calvignac-Spencer, S. (2020). Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science*, *368*(6497), 1367–1370. https://doi.org/10.1126/science.aba9411

Ebert, D., & Fields, P. D. (2020). Host–parasite co-evolution and its genomic signature. *Nature Reviews Genetics*, *21*(12), 754–768. https://doi.org/10.1038/s41576-020-0269-1

Ebert, D., & Hamilton, W. D. (1996). Sex against virulence: The coevolution of parasitic diseases. *Trends in Ecology & Evolution*, *11*(2), 79–82. https://doi.org/10.1016/0169-5347(96)81047-0

Eickhoff, C. S., Terry, F. E., Peng, L., Meza, K. A., Sakala, I. G., Van Aartsen, D., Moise, L., Martin, W. D., Schriewer, J., Buller, R. M., De Groot, A. S., & Hoft, D. F. (2019). Highly conserved influenza T cell epitopes induce broadly protective immunity. *Vaccine*, *37*(36), 5371–5381. https://doi.org/10.1016/j.vaccine.2019.07.033

Ejsmond, M. J., & Radwan, J. (2015). Red Queen Processes Drive Positive Selection on Major Histocompatibility Complex (MHC) Genes. *PLOS Computational Biology*, *11*(11), e1004627. https://doi.org/10.1371/journal.pcbi.1004627

El Mousadik, A., & Petit, R. J. (1996). High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics*, *92*(7), 832–839. https://doi.org/10.1007/BF00221895

Enard, D., Cai, L., Gwennap, C., & Petrov, D. A. (2016). Viruses are a dominant driver of protein adaptation in mammals. *ELife*, *5*, e12469. https://doi.org/10.7554/eLife.12469

Fan, S., Hansen, M. E. B., Lo, Y., & Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. *Science (New York, N.Y.)*, *354*(6308), 54–59. https://doi.org/10.1126/science.aaf5098

Fan, S., Kelly, D. E., Beltrame, M. H., Hansen, M. E. B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., Omar, S. A., Meskel, D. W., Belay, G., Froment, A., Patterson, N., Reich, D., & Tishkoff, S. A. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*, *20*, 82. https://doi.org/10.1186/s13059-019-1679-2

Fan, S., Spence, J. P., Feng, Y., Hansen, M. E. B., Terhorst, J., Beltrame, M. H., Ranciaro, A., Hirbo, J., Beggs, W., Thomas, N., Nyambo, T., Mpoloka, S. W., Mokone, G. G., Njamnshi, A. K., Fokunang, C., Meskel, D. W., Belay, G., Song, Y. S., & Tishkoff, S. A. (2023). Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell*, *186*(5), 923-939.e14. https://doi.org/10.1016/j.cell.2023.01.042

Fenner, J. N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, *128*(2), 415–423. https://doi.org/10.1002/ajpa.20188

Flores-Villanueva, P. O., Hendel, H., Caillat-Zucman, S., Rappaport, J., Burgos-Tiburcio, A., Bertin-Maghit, S., Ruiz-Morales, J. A., Teran, M. E., Rodriguez-Tafur, J., & Zagury, J.-F. (2003). Associations of MHC Ancestral Haplotypes with Resistance/Susceptibility to AIDS Disease Development. *The Journal of Immunology*, *170*(4), 1925–1929. https://doi.org/10.4049/jimmunol.170.4.1925

Forni, D., Cagliani, R., Clerici, M., Pozzoli, U., & Sironi, M. (2020). You Will Never Walk Alone: Codispersal of JC Polyomavirus with Human Populations. *Molecular Biology and Evolution*, *37*(2), 442–454. https://doi.org/10.1093/molbev/msz227

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Futuyma, D. J. (2009). *Evolution* (2nd ed.). Sinauer Associates, Inc. Publishers.

Garamszegi, L. Z. (2014). Global distribution of malaria-resistant MHC-HLA alleles: The number and frequencies of alleles and malaria risk. *Malaria Journal*, *13*(1), 349. https://doi.org/10.1186/1475-2875-13-349

Garcia, K. C. (2012). Reconciling views on T cell receptor germline bias for MHC. *Trends in Immunology*, *33*(9), 429–436. https://doi.org/10.1016/j.it.2012.05.005

Ghielmetti, G., Kupca, A. M., Hanczaruk, M., Friedel, U., Weinberger, H., Revilla-Fernández, S., Hofer, E., Riehm, J. M., Stephan, R., & Glawischnig, W. (2021). Mycobacterium microti Infections in Free-Ranging Red Deer (Cervus elaphus). *Emerging Infectious Diseases*, *27*(8), 2025–2032. https://doi.org/10.3201/eid2708.210634

Gillespie, G. M. A., Stewart-Jones, G., Rengasamy, J., Beattie, T., Bwayo, J. J., Plummer, F. A., Kaul, R., McMichael, A. J., Easterbrook, P., Dong, T., Jones, E. Y., & Rowland-Jones, S. L. (2006). Strong TCR conservation and altered T cell cross-reactivity characterize a B*57-restricted immune response in HIV-1 infection. *Journal of Immunology (Baltimore, Md.: 1950)*, *177*(6), 3893–3902. https://doi.org/10.4049/jimmunol.177.6.3893

Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. dos, Jones, J., Takeshita, L., Ortega-Rivera, N. D., Cid-Pavon, G. M. D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., Middleton, D., & Jones, A. R. (2020). Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, *48*(D1), D783–D788. https://doi.org/10.1093/nar/gkz1029

Gopalan, S., Berl, R. E. W., Myrick, J. W., Garfield, Z. H., Reynolds, A. W., Bafens, B. K., Belbin, G., Mastoras, M., Williams, C., Daya, M., Negash, A. N., Feldman, M. W., Hewlett, B. S., & Henn, B. M. (2022). Hunter-gatherer genomes reveal diverse demographic trajectories during the rise of farming in Eastern Africa. *Current Biology: CB*, *32*(8), 1852-1860.e5. https://doi.org/10.1016/j.cub.2022.02.050

Gorer, P. A. (1936). The detection of a hereditary antigenic difference in the blood of mice by means of human group a serum. *Journal of Genetics*, *32*(1), 17. https://doi.org/10.1007/BF02982499

Goudet, J. (2005). HIERFSTAT, a Package for R to Compute and Test Hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x

Gouy, A., & Excoffier, L. (2020). Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Molecular Biology and Evolution*, *37*(5), 1420–1433. https://doi.org/10.1093/molbev/msz306

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, *185*(4154), 862–864. https://doi.org/10.1126/science.185.4154.862

Günther, T., & Jakobsson, M. (2016). Genes mirror migrations and cultures in prehistoric Europe—A population genomic perspective. *Current Opinion in Genetics & Development*, *41*, 115–123. https://doi.org/10.1016/j.gde.2016.09.004

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., … Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, *522*(7555), 207–211. https://doi.org/10.1038/nature14317

Haldane, J. B. S. (1949). Disease and evolution. *Current Science*, *63*(9/10), 599–604.

Hammer, C., Begemann, M., McLaren, P. J., Bartha, I., Michel, A., Klose, B., Schmitt, C., Waterboer, T., Pawlita, M., Schulz, T. F., Ehrenreich, H., & Fellay, J. (2015). Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. *American Journal of Human Genetics*, *97*(5), 738–743. https://doi.org/10.1016/j.ajhg.2015.09.008

Harper, K., & Armelagos, G. (2010). The Changing Disease-Scape in the Third Epidemiological Transition. *International Journal of Environmental Research and Public Health*, *7*(2), 2. https://doi.org/10.3390/ijerph7020675

Healy, B. C., Liguori, M., Tran, D., Chitnis, T., Glanz, B., Wolfish, C., Gauthier, S., Buckle, G., Houtchens, M., Stazzone, L., Khoury, S., Hartzmann, R., Fernandez-Vina, M., Hafler, D. A., Weiner, H. L., Guttmann, C. R. G., & Jager, P. L. D. (2010). HLA B*44: Protective effects in MS susceptibility and MRI outcome measures. *Neurology*, *75*(7), 634–640. https://doi.org/10.1212/WNL.0b013e3181ed9c9c

Hedrick, P. W. (2002). Pathogen resistance and genetic variation at MHC loci. *Evolution*, *56*(10), 1902–1908. https://doi.org/10.1111/j.0014-3820.2002.tb00116.x

Hedrick, P. W., & Thomson, G. (1983). Evidence for Balancing Selection at Hla. *Genetics*, *104*(3), 449–456.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., Parham, P., Bustamante, C. D., Mountain, J. L., & Feldman, M. W. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(13), 5154–5162. https://doi.org/10.1073/pnas.1017511108

Hertz, T., Nolan, D., James, I., John, M., Gaudieri, S., Phillips, E., Huang, J. C., Riadi, G., Mallal, S., & Jojic, N. (2011). Mapping the Landscape of Host-Pathogen Coevolution: HLA Class I Binding and Its Relationship with Evolutionary Conservation in Human and Viral Proteins. *Journal of Virology*, *85*(3), 1310–1321. https://doi.org/10.1128/JVI.01966-10

Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., & Greenwood, B. M. (1991). Common west African HLA antigens are associated with protection from severe malaria. *Nature*, *352*(6336), 595–600. https://doi.org/10.1038/352595a0

Hill, A. V. S. (1991). HLA Associations with Malaria in Africa: Some Implications for MHC Evolution. In J. Klein & D. Klein (Eds.), *Molecular Evolution of the Major Histocompatibility Complex* (pp. 403–420). Springer. https://doi.org/10.1007/978-3-642-84622-9_33

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., & Green, E. D. (2018). Prioritizing diversity in human genomics research. *Nature Reviews. Genetics*, *19*(3), 175–185. https://doi.org/10.1038/nrg.2017.89

Hoh, B.-P., Zhang, X., Deng, L., Yuan, K., Yew, C.-W., Saw, W.-Y., Hoque, M. Z., Aghakhanian, F., Phipps, M. E., Teo, Y.-Y., Subbiah, V. K., & Xu, S. (2020). Shared Signature of Recent Positive Selection on the TSBP1–BTNL2–HLA-DRA Genes in Five Native Populations from North Borneo. *Genome Biology and Evolution*, *12*(12), 2245–2257. https://doi.org/10.1093/gbe/evaa207

Hollenbach, J. A., & Oksenberg, J. R. (2015). The Immunogenetics of Multiple Sclerosis: A Comprehensive Review. *Journal of Autoimmunity*, *64*, 13–25. https://doi.org/10.1016/j.jaut.2015.06.010

Hollenbach, J. A., Thomson, G., Cao, K., Fernandez-Vina, M., Erlich, H. A., Bugawan, T. L., Winkler, C., Winter, M., & Klitz, W. (2001). HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. *Human Immunology*, *62*(4), 378–390. https://doi.org/10.1016/S0198-8859(01)00212-9

Houldcroft, C. J., & Underdown, S. (2023). Infectious disease in the Pleistocene: Old friends or old foes? *American Journal of Biological Anthropology*, *n/a*(n/a). https://doi.org/10.1002/ajpa.24737

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., & Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, *546*(7657), 289–292. https://doi.org/10.1038/nature22336

Hughes, A. L., & Hughes, M. K. (1995). Natural selection on the peptide-binding regions of major histocompatibility complex molecules. *Immunogenetics*, *42*(4), 233–243. https://doi.org/10.1007/BF00176440

Hughes, A. L., & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, *335*(6186), 167–170. https://doi.org/10.1038/335167a0

Hughes, A. L., & Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. *Proceedings of the National Academy of Sciences*, *86*(3), 958–962. https://doi.org/10.1073/pnas.86.3.958

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology*, *52*(4), 577–586. https://doi.org/10.2307/1934145

Hurley, C. K., Kempenich, J., Wadsworth, K., Sauter, J., Hofmann, J. A., Schefzyk, D., Schmidt, A. H., Galarza, P., Cardozo, M. B. R., Dudkiewicz, M., Houdova, L., Jindra, P., Sorensen, B. S., Jagannathan, L., Mathur, A., Linjama, T., Torosian, T., Freudenberger, R., Manolis, A., … Dehn, J.

(2020). Common, intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. *HLA*, *95*(6), 516–531. https://doi.org/10.1111/tan.13811

Immel, A., Key, F. M., Szolek, A., Barquera, R., Robinson, M. K., Harrison, G. F., Palmer, W. H., Spyrou, M. A., Susat, J., Krause-Kyora, B., Bos, K. I., Forrest, S., Hernández-Zaragoza, D. I., Sauter, J., Solloch, U., Schmidt, A. H., Schuenemann, V. J., Reiter, E., Kairies, M. S., … Krause, J. (2021). Analysis of Genomic DNA from Medieval Plague Victims Suggests Long-Term Effect of Yersinia pestis on Human Immunity Genes. *Molecular Biology and Evolution*, *38*(10), 4059–4076. https://doi.org/10.1093/molbev/msab147

Jensen, J. M., Villesen, P., Friborg, R. M., Consortium, T. D. P.-G., Mailund, T., Besenbacher, S., Schierup, M. H., Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., Skov, L., Belling, K., Have, C. T., Izarzugaza, J. M. G., Grosjean, M., Bork-Jensen, J., … Schierup, M. H. (2017). Assembly and analysis of 100 full MHC haplotypes from the Danish population. *Genome Research*, *27*(9), 1597–1607. https://doi.org/10.1101/gr.218891.116

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics (Oxford, England)*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. *Nature*, *444*(7117), 323–329. https://doi.org/10.1038/nature05286

Kaiser, S. M., Malik, H. S., & Emerman, M. (2007). Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. *Science (New York, N.Y.)*, *316*(5832), 1756–1758. https://doi.org/10.1126/science.1140579

Kamoun, M., McCullough, K. P., Maiers, M., Fernandez Vina, M. A., Li, H., Teal, V., Leichtman, A. B., & Merion, R. M. (2017). HLA Amino Acid Polymorphisms and Kidney Allograft Survival. *Transplantation*, *101*(5), e170–e177. https://doi.org/10.1097/TP.0000000000001670

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, *15*(6), 6. https://doi.org/10.1038/nrg3734

Kaufman, J. (2018a). Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. *Trends in Immunology*, *39*(5), 367–379. https://doi.org/10.1016/j.it.2018.01.001

Kaufman, J. (2018b). Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates. *Annual Review of Immunology*, *36*(1), 383–409. https://doi.org/10.1146/annurev-immunol-051116-052450

Khasnis, A. A., & Nettleman, M. D. (2005). Global warming and infectious disease. *Archives of Medical Research*, *36*(6), 689–696. https://doi.org/10.1016/j.arcmed.2005.03.041

Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J., & Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. *Genome Biology*, *19*(1), 179. https://doi.org/10.1186/s13059-018-1561-7

Kılınç, G. M., Omrak, A., Özer, F., Günther, T., Büyükkarakaya, A. M., Bıçakçı, E., Baird, D., Dönertaş, H. M., Ghalichi, A., Yaka, R., Koptekin, D., Açan, S. C., Parvizi, P., Krzewińska, M., Daskalaki, E. A., Yüncü, E., Dağtaş, N. D., Fairbairn, A., Pearson, J., … Götherström, A. (2016). The Demographic Development of the First Farmers in Anatolia. *Current Biology*, *26*(19), 2659–2666. https://doi.org/10.1016/j.cub.2016.07.057

Klein, J., Sato, A., Nagl, S., & O'hUigín, C. (1998). Molecular Trans-Species Polymorphism. *Annual Review of Ecology and Systematics*, *29*(1), 1–21. https://doi.org/10.1146/annurev.ecolsys.29.1.1

Klein, L., Kyewski, B., Allen, P. M., & Hogquist, K. A. (2014). Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nature Reviews Immunology*, *14*(6), 377–391. https://doi.org/10.1038/nri3667

Kløverpris, H. N., Leslie, A., & Goulder, P. (2016). Role of HLA Adaptation in HIV Evolution. *Frontiers in Immunology*, *6*. https://doi.org/10.3389/fimmu.2015.00665

Kochan, G., Escors, D., Breckpot, K., & Guerrero-Setas, D. (2013). Role of non-classical MHC class I molecules in cancer immunosuppression. *Oncoimmunology*, *2*(11), e26491. https://doi.org/10.4161/onci.26491

Koonin, E. V., & Makarova, K. S. (2019). Origins and evolution of CRISPR-Cas systems. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1772), 20180087. https://doi.org/10.1098/rstb.2018.0087

Koonin, E. V., Makarova, K. S., & Wolf, Y. I. (2017). Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annual Review of Microbiology*, *71*, 233–261. https://doi.org/10.1146/annurev-micro-090816-093830

Koptekin, D., Yüncü, E., Rodríguez-Varela, R., Altınışık, N. E., Psonis, N., Kashuba, N., Yorulmaz, S., George, R., Kazancı, D. D., Kaptan, D., Gürün, K., Vural, K. B., Gemici, H. C., Vassou, D., Daskalaki, E., Karamurat, C., Lagerholm, V. K., Erdal, Ö. D., Kırdök, E., … Somel, M. (2023). Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. *Current Biology*, *33*(1), 41-57.e15. https://doi.org/10.1016/j.cub.2022.11.034

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., & Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science (New York, N.Y.)*, *288*(5472), 1789–1796. https://doi.org/10.1126/science.288.5472.1789

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kunwar, P., Hawkins, N., Dinges, W. L., Liu, Y., Gabriel, E. E., Swan, D. A., Stevens, C. E., Maenza, J., Collier, A. C., Mullins, J. I., Hertz, T., Yu, X., & Horton, H. (2013a). Superior control of HIV-1 replication by CD8+ T cells targeting conserved epitopes: Implications for HIV vaccine design. *PloS One*, *8*(5), e64405. https://doi.org/10.1371/journal.pone.0064405

Kunwar, P., Hawkins, N., Dinges, W. L., Liu, Y., Gabriel, E. E., Swan, D. A., Stevens, C. E., Maenza, J., Collier, A. C., Mullins, J. I., Hertz, T., Yu, X., & Horton, H. (2013b). Superior control of HIV-1 replication by CD8+ T cells targeting conserved epitopes: Implications for HIV vaccine design. *PloS One*, *8*(5), e64405. https://doi.org/10.1371/journal.pone.0064405

Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P., & Thomson, G. (2007). PyPop update – a software pipeline for large-scale multilocus population genomics. *Tissue Antigens*, *69*(s1), 192–197. https://doi.org/10.1111/j.1399-0039.2006.00769.x

Larsen, C. S. (2018). The Bioarchaeology of Health Crisis: Infectious Disease in the Past. *Annual Review of Anthropology*, *47*(1), 295–313. https://doi.org/10.1146/annurev-anthro-102116-041441

Lasonder, E., Rijpma, S. R., van Schaijk, B. C. L., Hoeijmakers, W. A. M., Kensche, P. R., Gresnigt, M. S., Italiaander, A., Vos, M. W., Woestenenk, R., Bousema, T., Mair, G. R., Khan, S. M., Janse, C. J., Bártfai, R., & Sauerwein, R. W. (2016). Integrated transcriptomic and proteomic analyses of P. falciparum gametocytes: Molecular insight into sex-specific processes and translational repression. *Nucleic Acids Research*, *44*(13), 6087–6101. https://doi.org/10.1093/nar/gkw536

Lazaridis, I. (2018). The evolutionary history of human populations in Europe. *Current Opinion in Genetics & Development*, *53*, 21–27. https://doi.org/10.1016/j.gde.2018.06.007

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 7518. https://doi.org/10.1038/nature13673

Leigh, J. W., & Bryant, D. (2015). POPART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, *6*(9), 1110–1116. https://doi.org/10.1111/2041-210X.12410

Lenz, T. L. (2011). Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution*, *65*(8), 2380–2390. https://doi.org/10.1111/j.1558-5646.2011.01288.x

Lenz, T. L. (2018). Adaptive value of novel MHC immune gene variants. *Proceedings of the National Academy of Sciences*, *115*(7), 1414–1416. https://doi.org/10.1073/pnas.1722600115

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, N.Y.)*, *319*(5866), 1100–1104. https://doi.org/10.1126/science.1153717

Lima, T. H. A., Souza, A. S., Porto, I. O. P., Paz, M. A., Veiga-Castelli, L. C., Oliveira, M. L. G., Donadi, E. A., Meyer, D., Sabbagh, A., Mendes-Junior, C. T., & Castelli, E. C. (2019). HLA-A promoter, coding, and 3′UTR sequences in a Brazilian cohort, and their evolutionary aspects. *HLA*, *93*(2–3), 65–79. https://doi.org/10.1111/tan.13474

Lin, W., de Sessions, P. F., Teoh, G. H. K., Mohamed, A. N. N., Zhu, Y. O., Koh, V. H. Q., Ang, M. L. T., Dedon, P. C., Hibberd, M. L., & Alonso, S. (2016). Transcriptional Profiling of Mycobacterium tuberculosis Exposed to In Vitro Lysosomal Stress. *Infection and Immunity*, *84*(9), 2505–2523. https://doi.org/10.1128/IAI.00072-16

Lin, W.-H. W., Pan, C.-H., Adams, R. J., Laube, B. L., & Griffin, D. E. (2014). Vaccine-induced measles virus-specific T cells do not prevent infection or disease but facilitate subsequent clearance of viral RNA. *MBio*, *5*(2), e01047. https://doi.org/10.1128/mBio.01047-14

Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., Oppenheimer, J., Broomandkhoshbacht, N., Harney, E., Nordenfelt, S., Llamas, B., Gusztáv Mende, B., Köhler, K., Oross, K., Bondár, M., … Reich, D. (2017). Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, *551*(7680), 368–372. https://doi.org/10.1038/nature24476

Liu, L., Wang, J., Rosenberg, D., Zhao, H., Lengyel, G., & Nadel, D. (2018). Fermented beverage and food storage in 13,000 y-old stone mortars at Raqefet Cave, Israel: Investigating Natufian ritual feasting. *Journal of Archaeological Science: Reports*, *21*, 783–793. https://doi.org/10.1016/j.jasrep.2018.08.008

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448. https://doi.org/10.1038/ng.3679

Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, *48*(7), 811–816. https://doi.org/10.1038/ng.3571

Ma, Y., Su, H., Yuksel, M., Longhi, M. S., McPhail, M., Wang, P., Bansal, S., Wong, G.-W., Graham, J., Yang, L., Thompson, R., Doherty, D. G., Hadzic, N., Zen, Y., Quaglia, A., Henghan, M., Samyn, M., Vergani, D., & Mieli-Vergani, G. (2021). HLA PROFILE PREDICTS SEVERITY OF AUTOIMMUNE LIVER DISEASE IN CHILDREN OF EUROPEAN ANCESTRY. *Hepatology (Baltimore, Md.)*, *74*(4), 2032–2046. https://doi.org/10.1002/hep.31893

Mack, S. J., Gourraud, P.-A., Single, R. M., Thomson, G., & Hollenbach, J. A. (2012). Analytical Methods for Immunogenetic Population Data. *Methods in Molecular Biology (Clifton, N.J.)*, *882*, 215–244. https://doi.org/10.1007/978-1-61779-842-9_13

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. https://doi.org/10.1038/nature18299

Manczinger, M., Boross, G., Kemény, L., Müller, V., Lenz, T. L., Papp, B., & Pál, C. (2019). Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLoS Biology*, *17*(1), e3000131. https://doi.org/10.1371/journal.pbio.3000131

Marchi, N., Winkelbach, L., Schulz, I., Brami, M., Hofmanová, Z., Blöcher, J., Reyna-Blanco, C. S., Diekmann, Y., Thiéry, A., Kapopoulou, A., Link, V., Piuz, V., Kreutzer, S., Figarska, S. M., Ganiatsou, E., Pukaj, A., Struck, T. J., Gutenkunst, R. N., Karul, N., … Excoffier, L. (2022). The genomic origins of the world's first farmers. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2022.04.008

Margolis, D. J., Mitra, N., Duke, J. L., Berna, R., Margolis, J. D., Hoffstad, O., Kim, B. S., Yan, A. C., Zaenglein, A. L., Chiesa Fuxench, Z., Dinou, A., Wasserman, J., Tairis, N., Mosbruger, T. L., Ferriola, D., Damianos, G., Kotsopoulou, I., & Monos, D. S. (2021). Human leukocyte antigen class-

I variation is associated with atopic dermatitis: A case-control study. *Human Immunology*, *82*(8), 593–599. https://doi.org/10.1016/j.humimm.2021.04.001

Maróstica, A. S., Nunes, K., Castelli, E. C., Silva, N. S. B., Weir, B. S., Goudet, J., & Meyer, D. (2022). How HLA diversity is apportioned: Influence of selection and relevance to transplantation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1852), 20200420. https://doi.org/10.1098/rstb.2020.0420

Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Fernández-Viña, M., Geraghty, D. E., Holdsworth, R., Hurley, C. K., Lau, M., Lee, K. W., MacH, B., Maiers, M., Mayr, W. R., Müller, C. R., Parham, P., Petersdorf, E. W., Sasazuki, T., … Trowsdale, J. (2010a). Nomenclature for factors of the HLA system, 2010. In *Tissue Antigens* (Vol. 75, Issue 4, pp. 291–455). Blackwell Publishing Ltd. https://doi.org/10.1111/j.1399-0039.2010.01466.x

Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Fernández-Viña, M., Geraghty, D. E., Holdsworth, R., Hurley, C. K., Lau, M., Lee, K. W., MacH, B., Maiers, M., Mayr, W. R., Müller, C. R., Parham, P., Petersdorf, E. W., Sasazuki, T., … Trowsdale, J. (2010b). Nomenclature for factors of the HLA system, 2010. In *Tissue Antigens* (Vol. 75). Blackwell Publishing Ltd. https://doi.org/10.1111/j.1399-0039.2010.01466.x

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrel, J., Arsuaga, J. L., de Castro, J. M. B., Carbonell, E., … Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499–503. https://doi.org/10.1038/nature16152

McClelland, E. E., Penn, D. J., & Potts, W. K. (2003). Major Histocompatibility Complex Heterozygote Superiority during Coinfection. *Infection and Immunity*, *71*(4), 2079–2086. https://doi.org/10.1128/IAI.71.4.2079-2086.2003

McDevitt, H. O. (2000). *Discovering the Role of the Major Histocompatibility Complex in the Immune Response*. https://doi.org/10.1146/annurev.immunol.18.1.1

McLaren, P. J., & Carrington, M. (2015). The impact of host genetic variation on infection with HIV-1. *Nature Immunology*, *16*(6), 6. https://doi.org/10.1038/ni.3147

McQuillan, M. A., Ranciaro, A., Hansen, M. E. B., Fan, S., Beggs, W., Belay, G., Woldemeskel, D., & Tishkoff, S. A. (2022). Signatures of Convergent Evolution and Natural Selection at the Alcohol Dehydrogenase Gene Region are Correlated with Agriculture in Ethnically Diverse Africans. *Molecular Biology and Evolution*, *39*(10), msac183. https://doi.org/10.1093/molbev/msac183

Mellins, E. D., & Stern, L. J. (2014). HLA-DM and HLA-DO, key regulators of MHC-II processing and presentation. *Current Opinion in Immunology*, *0*, 115–122. https://doi.org/10.1016/j.coi.2013.11.005

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic Maps of Human Gene Frequencies in Europeans. *Science*, *201*, 786–792. https://doi.org/10.1126/science.356262

Meyer, D., & Thomson, G. (2001). How selection shapes variation of the human major histocompatibility complex: A review. *Annals of Human Genetics*, *65*(1), 1–26. https://doi.org/10.1046/j.1469-1809.2001.6510001.x

Meyer, D., Vitor, V. R., Bitarello, B. D., Débora, D. Y., & Nunes, K. (2018). A genomic perspective on HLA evolution. *Immunogenetics*, *70*(1), 5–27. https://doi.org/10.1007/s00251-017-1017-3

Microbiology by numbers. (2011). *Nature Reviews Microbiology*, *9*(9), 628. https://doi.org/10.1038/nrmicro2644

Miles, A., bot, pyup io, R, M., Ralph, P., Harding, N., Pisupati, R., Rae, S., & Millar, T. (2021). *cggh/scikit-allel: V1.3.3*. Zenodo. https://doi.org/10.5281/zenodo.4759368

Miura, T., Brockman, M. A., Schneidewind, A., Lobritz, M., Pereyra, F., Rathod, A., Block, B. L., Brumme, Z. L., Brumme, C. J., Baker, B., Rothchild, A. C., Li, B., Trocha, A., Cutrell, E., Frahm, N., Brander, C., Toth, I., Arts, E. J., Allen, T. M., & Walker, B. D. (2009). HLA-B57/B*5801 Human Immunodeficiency Virus Type 1 Elite Controllers Select for Rare Gag Variants Associated with Reduced Viral Replication Capacity and Strong Cytotoxic T-Lymphotye Recognition. *Journal of Virology*, *83*(6), 2743–2755. https://doi.org/10.1128/JVI.02265-08

Mozzi, A., Pontremoli, C., & Sironi, M. (2018). Genetic susceptibility to infectious diseases: Current status and future perspectives from genome-wide approaches. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *66*, 286–307. https://doi.org/10.1016/j.meegid.2017.09.028

Murphy, K. M., & Weaver, C. (2017). *Janeway's Immunobiology*. Garland Science.

Murthy, S., O'Brien, K., Agbor, A., Angedakin, S., Arandjelovic, M., Ayimisin, E. A., Bailey, E., Bergl, R. A., Brazzola, G., Dieguez, P., Eno-Nku, M., Eshuis, H., Fruth, B., Gillespie, T. R., Ginath, Y., Gray, M., Herbinger, I., Jones, S., Kehoe, L., … Calvignac-Spencer, S. (2019). Cytomegalovirus distribution and evolution in hominines. *Virus Evolution*, *5*(2), vez015. https://doi.org/10.1093/ve/vez015

Nadachowska-Brzyska, K., Zieliński, P., Radwan, J., & Babik, W. (2012). Interspecific hybridization increases MHC class II diversity in two sister species of newts: HYBRIDIZATION INCREASES MHC DIVERSITY. *Molecular Ecology*, *21*(4), 887–906. https://doi.org/10.1111/j.1365-294X.2011.05347.x

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., Hebert, S., Pagé Sabourin, A., Luca, F., Blekhman, R., Hernandez, R. D., Pique-Regi, R., Tung, J., Yotova, V., & Barreiro, L. B. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, *167*(3), 657-669.e21. https://doi.org/10.1016/j.cell.2016.09.025

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The Bottleneck Effect and Genetic Variability in Populations. *Evolution*, *29*(1), 1–10. https://doi.org/10.2307/2407137

Nemat-Gorgani, N., Guethlein, L. A., Henn, B. M., Norberg, S. J., Chiaroni, J., Sikora, M., Quintana-Murci, L., Mountain, J. L., Norman, P. J., & Parham, P. (2019). Diversity of KIR, HLA class I and their Interactions in Seven Populations of sub-Saharan Africans. *Journal of Immunology (Baltimore, Md. : 1950)*, *202*(9), 2636–2647. https://doi.org/10.4049/jimmunol.1801586

Neu, U., Wang, J., Macejak, D., Garcea, R. L., & Stehle, T. (2011). Structures of the major capsid proteins of the human Karolinska Institutet and Washington University polyomaviruses. *Journal of Virology*, *85*(14), 7384–7392. https://doi.org/10.1128/JVI.00382-11

Nielsen, M., & Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine*, *8*(1), 33. https://doi.org/10.1186/s13073-016-0288-x

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. https://doi.org/10.1038/nature21347

Nielsen, S. H., van Dorp, L., Houldcroft, C. J., Pedersen, A. G., Allentoft, M. E., Vinner, L., Margaryan, A., Pavlova, E., Chasnyk, V., Nikolskiy, P., Pitulko, V., Pimenoff, V. N., Balloux, F., & Sikora, M. (2021). 31,600-year-old human virus genomes support a Pleistocene origin for common childhood infections. *BioRxiv*. bioRxiv. https://doi.org/10.1101/2021.06.28.450199

Nordin, J., Ameur, A., Lindblad-Toh, K., Gyllensten, U., & Meadows, J. R. S. (2020). SweHLA: The high confidence HLA typing bio-resource drawn from 1000 Swedish genomes. *European Journal of Human Genetics*, *28*(5), 627–635. https://doi.org/10.1038/s41431-019-0559-2

Nowak, J., Mika-Witkowska, R., Polak, M., Zajko, M., Rogatko-Koroś, M., Graczyk-Pol, E., & Lange, A. (2008). Allele and extended haplotype polymorphism of HLA-A, -C, -B, -DRB1 and -DQB1 loci in Polish population and genetic affinities to other populations. *Tissue Antigens*, *71*(3), 193–205. https://doi.org/10.1111/j.1399-0039.2007.00991.x

O'Connor, E. A., Westerdahl, H., Burri, R., & Edwards, S. V. (2019). Avian MHC Evolution in the Era of Genomics: Phase 1.0. *Cells*, *8*(10), E1152. https://doi.org/10.3390/cells8101152

O'Garra, A., & Vieira, P. (2004). Regulatory T cells and mechanisms of immune system control. *Nature Medicine*, *10*(8), 8. https://doi.org/10.1038/nm0804-801

Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., Altena, E., Lipson, M., Lazaridis, I., Harper, T. K., Patterson, N., Broomandkhoshbacht, N., Diekmann, Y., Faltyskova, Z., Fernandes, D., … Reich, D. (2018). The Beaker Phenomenon and the Genomic Transformation of Northwest Europe. *Nature*, *555*(7695), 190–196. https://doi.org/10.1038/nature25738

Oliveira-Cortez, A., Melo, A. C., Chaves, V. E., Condino-Neto, A., & Camargos, P. (2016). Do HLA class II genes protect against pulmonary tuberculosis? A systematic review and meta-analysis. *European Journal of Clinical Microbiology & Infectious Diseases*, *35*(10), 1567–1580. https://doi.org/10.1007/s10096-016-2713-x

Oliver, S. L., Yang, E., & Arvin, A. M. (2016). Varicella-Zoster Virus Glycoproteins: Entry, Replication, and Pathogenesis. *Current Clinical Microbiology Reports*, *3*(4), 204–215. https://doi.org/10.1007/s40588-016-0044-4

Ortmann, A. C., Wiedenheft, B., Douglas, T., & Young, M. (2006). Hot crenarchaeal viruses reveal deep evolutionary connections. *Nature Reviews Microbiology*, *4*(7), 7. https://doi.org/10.1038/nrmicro1444

Ota, M. O., Ndhlovu, Z., Oh, S., Piyasirisilp, S., Berzofsky, J. A., Moss, W. J., & Griffin, D. E. (2007). Hemagglutinin Protein Is a Primary Target of the Measles Virus—Specific HLA-A2—Restricted CD8+ T Cell Response during Measles and after Vaccination. *The Journal of Infectious Diseases*, *195*(12), 1799–1807. https://doi.org/10.1086/518006

Özer, O., & Lenz, T. L. (2021). Unique Pathogen Peptidomes Facilitate Pathogen-Specific Selection and Specialization of MHC Alleles. *Molecular Biology and Evolution*, *38*(10), 4376–4387. https://doi.org/10.1093/molbev/msab176

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L., Wall, J. D., Cardona, A., Mägi, R., Sayres, M. A. W., Kaewert, S., Inchley, C., Scheib, C. L., Järve, M., Karmin, M., … Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, *538*(7624), 7624. https://doi.org/10.1038/nature19792

Pagkrati, I., Duke, J. L., Mbunwe, E., Mosbruger, T. L., Ferriola, D., Wasserman, J., Dinou, A., Tairis, N., Damianos, G., Kotsopoulou, I., Papaioannou, J., Giannopoulos, D., Beggs, W., Nyambo, T., Mpoloka, S. W., Mokone, G. G., Njamnshi, A. K., Fokunang, C., Woldemeskel, D., … Monos, D. S. (2023). Genomic characterization of HLA class I and class II genes in ethnically diverse sub-Saharan African populations: A report on novel HLA alleles. *HLA*, *102*(2), 192–205. https://doi.org/10.1111/tan.15035

Pappas, D. J., Marin, W., Hollenbach, J. A., & Mack, S. J. (2016). Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline. *Human Immunology*, *77*(3), 283–287. https://doi.org/10.1016/j.humimm.2015.12.006

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Parham, P. (1988). Function and polymorphism of human leukocyte antigen-A,B,C molecules. *The American Journal of Medicine*, *85*(6A), 2–5. https://doi.org/10.1016/0002-9343(88)90369-5

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Patz, J. A., Epstein, P. R., Burke, T. A., & Balbus, J. M. (1996). Global climate change and emerging infectious diseases. *JAMA*, *275*(3), 217–223.

Paul, S., Croft, N. P., Purcell, A. W., Tscharke, D. C., Sette, A., Nielsen, M., & Peters, B. (2020). Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. *PLOS Computational Biology*, *16*(5), e1007757. https://doi.org/10.1371/journal.pcbi.1007757

Paul, S., Weiskopf, D., Angelo, M. A., Sidney, J., Peters, B., & Sette, A. (2013). HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. *The Journal of Immunology*, *191*(12), 5831–5839. https://doi.org/10.4049/jimmunol.1302101

Paximadis, M., Mathebula, T. Y., Gentle, N. L., Vardas, E., Colvin, M., Gray, C. M., Tiemessen, C. T., & Puren, A. (2012). Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. *Human Immunology*, *73*(1), 80–92. https://doi.org/10.1016/j.humimm.2011.10.013

Pemberton, T. J., & Rosenberg, N. A. (2014). Population-genetic influences on genomic estimates of the inbreeding coefficient: A global perspective. *Human Heredity*, *77*(0), 37–48. https://doi.org/10.1159/000362878

Peng, K., Safonova, Y., Shugay, M., Popejoy, A. B., Rodriguez, O. L., Breden, F., Brodin, P., Burkhardt, A. M., Bustamante, C., Cao-Lormeau, V.-M., Corcoran, M. M., Duffy, D., Fuentes-Guajardo, M., Fujita, R., Greiff, V., Jönsson, V. D., Liu, X., Quintana-Murci, L., Rossetti, M., … Mangul, S. (2021). Diversity in immunogenomics: The value and the challenge. *Nature Methods*, *18*(6), 588–591. https://doi.org/10.1038/s41592-021-01169-5

Penman, B. S., & Gupta, S. (2018). Detecting signatures of past pathogen selection on human HLA loci: Are there needles in the haystack? *Parasitology*, *145*(6), 731–739. https://doi.org/10.1017/S0031182017001159

Penn, D. J., Damjanovich, K., & Potts, W. K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11260–11264. https://doi.org/10.1073/pnas.162006499

Peters, B., Nielsen, M., & Sette, A. (2020). T Cell Epitope Predictions. *Annual Review of Immunology*, *38*(1), 123–145. https://doi.org/10.1146/annurev-immunol-082119-124838

Phillips, K. P., Cable, J., Mohammed, R. S., Herdegen-Radwan, M., Raubic, J., Przesmycka, K. J., van Oosterhout, C., & Radwan, J. (2018a). Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proceedings of the National Academy of Sciences*, *115*(7), 1552–1557. https://doi.org/10.1073/pnas.1708597115

Phillips, K. P., Cable, J., Mohammed, R. S., Herdegen-Radwan, M., Raubic, J., Przesmycka, K. J., van Oosterhout, C., & Radwan, J. (2018b). Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proceedings of the National Academy of Sciences*, *115*(7), 1552–1557. https://doi.org/10.1073/pnas.1708597115

Pierini, F., & Lenz, T. L. (2018). Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection. *Molecular Biology and Evolution*, *35*(9), 2145–2158. https://doi.org/10.1093/molbev/msy116

Pierini, F., Nutsua, M., Böhme, L., Özer, O., Bonczarowska, J., Susat, J., Franke, A., Nebel, A., Krause-Kyora, B., & Lenz, T. L. (2020). Targeted analysis of polymorphic loci from low-coverage shotgun sequence data allows accurate genotyping of HLA genes in historical human populations. *Scientific Reports*, *10*(1), 1. https://doi.org/10.1038/s41598-020-64312-w

Pimenoff, V. N., de Oliveira, C. M., & Bravo, I. G. (2017). Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16. *Molecular Biology and Evolution*, *34*(1), 4–19. https://doi.org/10.1093/molbev/msw214

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 8. https://doi.org/10.1038/ng1847

Prugnolle, F., Manica, A., & Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. *Current Biology : CB*, *15*(5), R159–R160. https://doi.org/10.1016/j.cub.2005.02.038

Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V., & Balloux, F. (2005). Pathogen-driven selection and worldwide HLA class I diversity. *Current Biology*, *15*(11), 1022–1027. https://doi.org/10.1016/j.cub.2005.04.050

Pugliese, A., Boulware, D., Yu, L., Babu, S., Steck, A. K., Becker, D., Rodriguez, H., DiMeglio, L., Evans-Molina, C., Harrison, L. C., Schatz, D., Palmer, J. P., Greenbaum, C., Eisenbarth, G. S., & Sosenko, J. M. (2016). HLA-DRB1*15:01-DQA1*01:02-DQB1*06:02 Haplotype Protects Autoantibody-Positive Relatives From Type 1 Diabetes Throughout the Stages of Disease Progression. *Diabetes*, *65*(4), 1109–1119. https://doi.org/10.2337/db15-1105

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., 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. https://doi.org/10.1086/519795

Quagliariello, A., Modi, A., Innocenti, G., Zaro, V., Conati Barbaro, C., Ronchitelli, A., Boschin, F., Cavazzuti, C., Dellù, E., Radina, F., Sperduti, A., Bondioli, L., Ricci, S., Lognoli, M., Belcastro, M. G., Mariotti, V., Caramelli, D., Mariotti Lippi, M., Cristiani, E., … Lari, M. (2022). Ancient oral microbiomes support gradual Neolithic dietary shifts towards agriculture. *Nature Communications*, *13*(1), 1. https://doi.org/10.1038/s41467-022-34416-0

Quattrini, A. M., & Demopoulos, A. W. J. (2016). Ectoparasitism on deep-sea fishes in the western North Atlantic: In situ observations from ROV surveys. *International Journal for Parasitology. Parasites and Wildlife*, *5*(3), 217–228. https://doi.org/10.1016/j.ijppaw.2016.07.004

R Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/

Rabajante, J. F., Tubay, J. M., Ito, H., Uehara, T., Kakishima, S., Morita, S., Yoshimura, J., & Ebert, D. (2016). Host-parasite Red Queen dynamics with phase-locked rare genotypes. *Science Advances*, *2*(3), e1501548. https://doi.org/10.1126/sciadv.1501548

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*(6), 6. https://doi.org/10.1038/nrg3936

Radwan, J., Babik, W., Kaufman, J., Lenz, T. L., & Winternitz, J. (2020). Advances in the Evolutionary Understanding of MHC Polymorphism. *Trends in Genetics*, *36*(4), 298–311. https://doi.org/10.1016/j.tig.2020.01.008

Raj, T., Kuchroo, M., Replogle, J. M., Raychaudhuri, S., Stranger, B. E., & De Jager, P. L. (2013). Common risk alleles for inflammatory diseases are targets of recent positive selection. *American Journal of Human Genetics*, *92*(4), 517–529. https://doi.org/10.1016/j.ajhg.2013.03.001

Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., & Stevanović, S. (1999). SYFPEITHI: Database for MHC ligands and peptide motifs. *Immunogenetics*, *50*(3), 213–219. https://doi.org/10.1007/s002510050595

Ranciaro, A., Campbell, M. C., Hirbo, J. B., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze, M. J., Ibrahim, M., Nyambo, T., Omar, S. A., & Tishkoff, S. A. (2014). Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *The American Journal of Human Genetics*, *94*(4), 496–510. https://doi.org/10.1016/j.ajhg.2014.02.009

Rao, X., Fontaine Costa, A. I. C. A., van Baarle, D., & Keşmir, C. (2009). A Comparative Study of HLA Binding Affinity and Ligand Diversity: Implications for Generating Immunodominant CD8 $^{+}$ T Cell Responses. *The Journal of Immunology*, *182*(3), 1526–1532. https://doi.org/10.4049/jimmunol.182.3.1526

Rao, X., Hoof, I., van Baarle, D., Keşmir, C., & Textor, J. (2015). HLA Preferences for Conserved Epitopes: A Potential Mechanism for Hepatitis C Clearance. *Frontiers in Immunology*, *6*, 552. https://doi.org/10.3389/fimmu.2015.00552

Reche, P. A., & Reinherz, E. L. (2003). Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. *Journal of Molecular Biology*, *331*(3), 623–641. https://doi.org/10.1016/S0022-2836(03)00750-2

Rees, J. S., Castellano, S., & Andrés, A. M. (2020). The Genomics of Human Local Adaptation. *Trends in Genetics: TIG*, *36*(6), 415–428. https://doi.org/10.1016/j.tig.2020.03.006

Reina-Campos, M., Scharping, N. E., & Goldrath, A. W. (2021). CD8+ T cell metabolism in infection and cancer. *Nature Reviews. Immunology*, *21*(11), 718–738. https://doi.org/10.1038/s41577-021-00537-8

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, *48*(W1), W449–W454. https://doi.org/10.1093/nar/gkaa379

Richards, M. P., Schulting, R. J., & Hedges, R. E. M. (2003). Sharp shift in diet at onset of Neolithic. *Nature*, *425*(6956), 6956. https://doi.org/10.1038/425366a

Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., & Fugmann, T. (2016). High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics*, *16*(10), 1570–1580. https://doi.org/10.1002/pmic.201500445

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., & Marsh, S. G. E. (2020). IPD-IMGT/HLA Database. *Nucleic Acids Research*, *48*(D1), D948–D955. https://doi.org/10.1093/nar/gkz950

Robinson, J., Guethlein, L. A., Cereb, N., Yang, S. Y., Norman, P. J., Marsh, S. G. E., & Parham, P. (2017). Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLOS Genetics*, *13*(6), e1006862. https://doi.org/10.1371/journal.pgen.1006862

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic Structure of Human Populations. *Science*, *298*(5602), 2381–2385. https://doi.org/10.1126/science.1078311

Rota, P. A., Moss, W. J., Takeda, M., de Swart, R. L., Thompson, K. M., & Goodson, J. L. (2016). Measles. *Nature Reviews Disease Primers*, *2*(1), 1–16. https://doi.org/10.1038/nrdp.2016.49

Rotival, M., & Quintana-Murci, L. (2020). Functional consequences of archaic introgression and their impact on fitness. *Genome Biology*, *21*(1), 3. https://doi.org/10.1186/s13059-019-1920-z

Ruterbusch, M., Pruner, K. B., Shehata, L., & Pepper, M. (2020). In Vivo CD4+ T Cell Differentiation and Function: Revisiting the Th1/Th2 Paradigm. *Annual Review of Immunology*, *38*, 705–725. https://doi.org/10.1146/annurev-immunol-103019-085803

Salamon, H., Klitz, W., Easteal, S., Gao, X., Erlich, H. A., Fernandez-Viña, M., Trachtenberg, E. A., McWeeney, S. K., Nelson, M. P., & Thomson, G. (1999). Evolution of HLA class II molecules: Allelic and amino acid site variability across populations. *Genetics*, *152*(1), 393–400. https://doi.org/10.1093/genetics/152.1.393

Sanchez-Mazas, A. (2001). African diversity from the HLA point of view: Influence of genetic drift, geography, linguistics, and natural selection. *Human Immunology*, *62*(9), 937–948. https://doi.org/10.1016/S0198-8859(01)00293-2

Sanchez-Mazas, A. (2007). An apportionment of human HLA diversity. *Tissue Antigens*, *69*(s1), 198–202. https://doi.org/10.1111/j.1399-0039.2006.00802.x

Sanchez-Mazas, A. (2020). A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. *Swiss Medical Weekly*, *150*(1516), w20214. https://doi.org/10.4414/smw.2020.20214

Sanchez-Mazas, A., Buhler, S., & Nunes, J. M. (2014). A new HLA map of Europe: Regional genetic variation and its implication for peopling history, disease-association studies and tissue transplantation. *Human Heredity*, *76*(3–4), 162–177. https://doi.org/10.1159/000360855

Sanchez-Mazas, A., Černý, V., Di, D., Buhler, S., Podgorná, E., Chevallier, E., Brunet, L., Weber, S., Kervaire, B., Testi, M., Andreani, M., Tiercy, J. M., Villard, J., & Nunes, J. M. (2017). The HLA-B landscape of Africa: Signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. *Molecular Ecology*, *26*(22), 6238–6252. https://doi.org/10.1111/mec.14366

Sanchez-Mazas, A., Lemaitre, J.-F., & Currat, M. (2012). Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1590), 830–839. https://doi.org/10.1098/rstb.2011.0312

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748. https://doi.org/10.1128/JVI.00694-10

Scerri, E. M. L., Thomas, M. G., Manica, A., Gunz, P., Stock, J. T., Stringer, C., Grove, M., Groucutt, H. S., Timmermann, A., Rightmire, G. P., d'Errico, F., Tryon, C. A., Drake, N. A., Brooks, A. S., Dennell, R. W., Durbin, R., Henn, B. M., Lee-Thorp, J., deMenocal, P., … Chikhi, L. (2018). Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in Ecology & Evolution*, *33*(8), 582–594. https://doi.org/10.1016/j.tree.2018.05.005

Schellens, I. M. M., Hoof, I., Meiring, H. D., Spijkers, S. N. M., Poelen, M. C. M., van Gaans-van den Brink, J. A. M., van der Poel, K., Costa, A. I., van Els, C. A. C. M., van Baarle, D., & Kesmir, C. (2015). Comprehensive Analysis of the Naturally Processed Peptide Repertoire: Differences between HLA-A and B in the Immunopeptidome. *PloS One*, *10*(9), e0136417. https://doi.org/10.1371/journal.pone.0136417

Schellens, I. M., Meiring, H. D., Hoof, I., Spijkers, S. N., Poelen, M. C. M., van Gaans-van den Brink, J. A. M., Costa, A. I., Vennema, H., Keşmir, C., van Baarle, D., & van Els, C. A. C. M. (2015). Measles Virus Epitope Presentation by HLA: Novel Insights into Epitope Selection, Dominance, and Microvariation. *Frontiers in Immunology*, *6*, 546. https://doi.org/10.3389/fimmu.2015.00546

Schierup, M. H., Charlesworth, D., & Vekemans, X. (2000). The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. *Genetical Research*, *76*(1), 63–73. https://doi.org/10.1017/s0016672300004547

Schirle, M., Weinschenk, T., & Stevanović, S. (2001). Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. *Journal of Immunological Methods*, *257*(1), 1–16. https://doi.org/10.1016/S0022-1759(01)00459-8

Schneidewind, A., Brockman, M. A., Yang, R., Adam, R. I., Li, B., Le Gall, S., Rinaldo, C. R., Craggs, S. L., Allgaier, R. L., Power, K. A., Kuntzen, T., Tung, C.-S., LaBute, M. X., Mueller, S. M., Harrer, T., McMichael, A. J., Goulder, P. J. R., Aiken, C., Brander, C., … Allen, T. M. (2007). Escape from the Dominant HLA-B27-Restricted Cytotoxic T-Lymphocyte Response in Gag Is Associated with a Dramatic Reduction in Human Immunodeficiency Virus Type 1 Replication. *Journal of Virology*, *81*(22), 12382–12393. https://doi.org/10.1128/JVI.01543-07

Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway, M., Lambert, D., Khartanovich, V., Wall, J. D., Nigst, P. R., Foley, R. A., Lahr, M. M., Nielsen, R., … Willerslev, E. (2014). Genomic structure in Europeans dating back at least 36,200 years. *Science*, *346*(6213), 1113–1118. https://doi.org/10.1126/science.aaa0114

Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: Expression, interaction, diversity and disease. *Journal of Human Genetics*, *54*(1), 1. https://doi.org/10.1038/jhg.2008.5

Sidney, J., Peters, B., Frahm, N., Brander, C., & Sette, A. (2008). HLA class I supertypes: A revised and updated classification. *BMC Immunology*, *9*(1), 1. https://doi.org/10.1186/1471-2172-9-1

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., Margaryan, A., de Barros Damgaard, P., de la Fuente, C., Renaud, G., Yang, M. A., Fu, Q., Dupanloup, I., Giampoudakis, K., Nogués-Bravo, D., Rahbek, C., Kroonen, G., Peyrot, M., McColl, H., … Willerslev, E. (2019). The population history of northeastern Siberia since the Pleistocene. *Nature*, *570*(7760), 7760. https://doi.org/10.1038/s41586-019-1279-z

Single, R. M., Meyer, D., Nunes, K., Francisco, R. S., Hünemeier, T., Maiers, M., Hurley, C. K., Bedoya, G., Gallo, C., Hurtado, A. M., Llop, E., Petzl-Erler, M. L., Poletti, G., Rothhammer, F., Tsuneto, L., Klitz, W., & Ruiz-Linares, A. (2020). Demographic history and selection at HLA loci in Native Americans. *PLOS ONE*, *15*(11), e0241282. https://doi.org/10.1371/journal.pone.0241282

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Skoglund, P., Malmström, H., Omrak, A., Raghavan, M., Valdiosera, C., Günther, T., Hall, P., Tambets, K., Parik, J., Sjögren, K.-G., Apel, J., Willerslev, E., Storå, J., Götherström, A., & Jakobsson, M. (2014). Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science (New York, N.Y.)*, *344*(6185), 747–750. https://doi.org/10.1126/science.1253448

Slade, R. W., & McCallum, H. I. (1992). Overdominant vs. Frequency-dependent selection at MHC loci. *Genetics*, *132*(3), 861–864. https://doi.org/10.1093/genetics/132.3.861

Slatkin, M. (1994). An exact test for neutrality based on the Ewens sampling distribution. *Genetical Research*, *64*(1), 71–74. https://doi.org/10.1017/s0016672300032560

Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in Zoology*, *2*(1), 16. https://doi.org/10.1186/1742-9994-2-16

Souilmi, Y., Lauterbur, M. E., Tobler, R., Huber, C. D., Johar, A. S., Moradi, S. V., Johnston, W. A., Krogan, N. J., Alexandrov, K., & Enard, D. (2021). An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Current Biology*, *31*(16), 3504-3514.e9. https://doi.org/10.1016/j.cub.2021.05.067

Spurgin, L. G., & Richardson, D. S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1684), 979–988. https://doi.org/10.1098/rspb.2009.2084

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), 6. https://doi.org/10.1038/s41576-019-0119-1

Stefan, T., Matthews, L., Prada, J. M., Mair, C., Reeve, R., & Stear, M. J. (2019). Divergent Allele Advantage Provides a Quantitative Model for Maintaining Alleles with a Wide Range of Intrinsic Merits. *Genetics*, *212*(2), 553–564. https://doi.org/10.1534/genetics.119.302022

Stirling, G., & Wilsey, B. (2001). Empirical Relationships between Species Richness, Evenness, and Proportional Diversity. *The American Naturalist*, *158*(3), 286–299. https://doi.org/10.1086/321317

Stoneking, M., Arias, L., Liu, D., Oliveira, S., Pugach, I., & Rodriguez, J. J. R. B. (2023). Genomic perspectives on human dispersals during the Holocene. *Proceedings of the National Academy of Sciences*, *120*(4), e2209475119. https://doi.org/10.1073/pnas.2209475119

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, *123*(3), 585–595.

Takaba, H., & Takayanagi, H. (2017). The Mechanisms of T Cell Selection in the Thymus. *Trends in Immunology*, *38*(11), 805–816. https://doi.org/10.1016/j.it.2017.07.010

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

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Thorsby, E. (2009). A short history of HLA. *Tissue Antigens*, *74*(2), 101–116. https://doi.org/10.1111/j.1399-0039.2009.01291.x

Thorstenson, Y. R., Creary, L. E., Huang, H., Rozot, V., Nguyen, T. T., Babrzadeh, F., Kancharla, S., Fukushima, M., Kuehn, R., Wang, C., Li, M., Krishnakumar, S., Mindrinos, M., Fernandez Viña, M. A., Scriba, T. J., & Davis, M. M. (2018). Allelic resolution NGS HLA typing of Class I and Class II loci and haplotypes in Cape Town, South Africa. *Human Immunology*, *79*(12), 839–847. https://doi.org/10.1016/j.humimm.2018.09.004

Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., & Hinds, D. A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for

multiple common infections. *Nature Communications*, *8*(1), 1. https://doi.org/10.1038/s41467-017-00257-5

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., … Williams, S. M. (2009). The Genetic Structure and History of Africans and African Americans. *Science*. https://doi.org/10.1126/science.1172257

Tong, X., Chen, L., Liu, S., Yan, Z., Peng, S., Zhang, Y., & Fan, H. (2015). Polymorphisms in HLA-DRB1 gene and the risk of tuberculosis: A meta-analysis of 31 studies. *Lung*, *193*(2), 309–318. https://doi.org/10.1007/s00408-015-9692-z

Trowsdale, J. (2011). The MHC, disease and selection. *Immunology Letters*, *137*(1–2), 1–8. https://doi.org/10.1016/j.imlet.2011.01.002

Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, *14*(1), 301–323. https://doi.org/10.1146/annurev-genom-091212-153455

van Els, C. A. C. M., & Nanan, R. (2002). T Cell Responses in Acute Measles. *Viral Immunology*, *15*(3), 435–450. https://doi.org/10.1089/088282402760312322

Vellinga, J., Van der Heijdt, S., & Hoeben, R. C. (2005). The adenovirus capsid: Major progress in minor proteins. *Journal of General Virology*, *86*(6), 1581–1588. https://doi.org/10.1099/vir.0.80877-0

Vina, M. A. F., Hollenbach, J. A., Lyke, K. E., Sztein, M. B., Maiers, M., Klitz, W., Cano, P., Mack, S., Single, R., Brautbar, C., Israel, S., Raimondi, E., Khoriaty, E., Inati, A., Andreani, M., Testi, M., Moraes, M. E., Thomson, G., Stastny, P., & Cao, K. (2012). Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1590), 820–829. https://doi.org/10.1098/rstb.2011.0320

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006

Wakeland, E. K., Boehme, S., She, J. X., Lu, C.-C., McIndoe, R. A., Cheng, I., Ye, Y., & Potts, W. K. (1990). Ancestral polymorphisms of MHC class II genes: Divergent allele advantage. *Immunologic Research*, *9*(2), 115–122. https://doi.org/10.1007/BF02918202

Walsh, E. C., Mather, K. A., Schaffner, S. F., Farwell, L., Daly, M. J., Patterson, N., Cullen, M., Carrington, M., Bugawan, T. L., Erlich, H., Campbell, J., Barrett, J., Miller, K., Thomson, G., Lander, E. S., & Rioux, J. D. (2003). An integrated haplotype map of the human major histocompatibility complex. *American Journal of Human Genetics*, *73*(3), 580–590. https://doi.org/10.1086/378101

Watterson, G. A. (1977). Heterosis or neutrality? *Genetics*, *85*(4), 789–814. https://doi.org/10.1093/genetics/85.4.789

Wegner, K. M., & Eizaguirre, C. (2012). New(t)s and views from hybridizing MHC genes: Introgression rather than trans-species polymorphism may shape allelic repertoires. *Molecular Ecology*, *21*(4), 779–781. https://doi.org/10.1111/j.1365-294X.2011.05401.x

Weinert, L. A., Depledge, D. P., Kundu, S., Gershon, A. A., Nichols, R. A., Balloux, F., Welch, J. J., & Breuer, J. (2015). Rates of vaccine evolution show strong effects of latency: Implications for varicella zoster virus epidemiology. *Molecular Biology and Evolution*, *32*(4), 1020–1028. https://doi.org/10.1093/molbev/msu406

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Weiskopf, D., Angelo, M. A., de Azeredo, E. L., Sidney, J., Greenbaum, J. A., Fernando, A. N., Broadwater, A., Kolla, R. V., De Silva, A. D., de Silva, A. M., Mattia, K. A., Doranz, B. J., Grey, H. M., Shresta, S., Peters, B., & Sette, A. (2013). Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), E2046–E2053. https://doi.org/10.1073/pnas.1305227110

Wolfe, N. D., Dunavan, C. P., & Diamond, J. (2007). Origins of major human infectious diseases. *Nature*, *447*(7142), 279–283. https://doi.org/10.1038/nature05775

World Health Organization. (2018). *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016*. https://www.who.int/healthinfo/global_burden_disease/estimates/en/

World Health Organization. (2020). *Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

World Health Organization. (2021). *World health statistics 2021: Monitoring health for the SDGs, sustainable development goals*. https://www.who.int/publications/i/item/9789240027053

Wu, H., Kropff, B., Mach, M., & Britt, W. J. (2020). Human Cytomegalovirus Envelope Protein gpUL132 Regulates Infectious Virus Production through Formation of the Viral Assembly Compartment. *MBio*, *11*(5), e02044-20. https://doi.org/10.1128/mBio.02044-20

Yamaguchi, T., & Dijkstra, J. M. (2019). Major Histocompatibility Complex (MHC) Genes and Disease Resistance in Fish. *Cells*, *8*(4), E378. https://doi.org/10.3390/cells8040378

Yewdell, J. W. (2006). Confronting Complexity: Real-World Immunodominance in Antiviral CD8+ T Cell Responses. *Immunity*, *25*(4), 533–543. https://doi.org/10.1016/j.immuni.2006.09.005

Yewdell, J. W., Reits, E., & Neefjes, J. (2003). Making sense of mass destruction: Quantitating MHC class I antigen presentation. *Nature Reviews Immunology*, *3*(12), 12. https://doi.org/10.1038/nri1250

Zeder, M. A. (2011). The Origins of Agriculture in the Near East. *Current Anthropology*, *52*(S4), S221–S235. https://doi.org/10.1086/659307

Zhang, Q., Lin, C.-Y., Dong, Q., Wang, J., & Wang, W. (2011). Relationship between HLA-DRB1 polymorphism and susceptibility or resistance to multiple sclerosis in Caucasians: A meta-analysis of non-family-based studies. *Autoimmunity Reviews*, *10*(8), 474–481. https://doi.org/10.1016/j.autrev.2011.03.003

Annex I

Targeted analysis of polymorphic loci from low-coverage shotgun sequence data allows accurate genotyping of HLA genes in historical human populations

Federica Pierini[1], Marcel Nutsua[2], Lisa Böhme[2], Onur Özer[1], Joanna Bonczarowska[2], Julian Susat[2], Andre Franke[2], Almut Nebel[2], Ben Krause-Kyora[2], Tobias L. Lenz[1]

[1]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Ploen, Germany

[2]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

Published article
Scientific Reports (2020)
https://doi.org/10.1038/s41598-020-64312-w

SCIENTIFIC REPORTS
nature research

OPEN

Targeted analysis of polymorphic loci from low-coverage shotgun sequence data allows accurate genotyping of HLA genes in historical human populations

Federica Pierini[1,3], Marcel Nutsua[2], Lisa Böhme[2], Onur Özer[1], Joanna Bonczarowska[2], Julian Susat[2], Andre Franke[2], Almut Nebel[2], Ben Krause-Kyora[2] & Tobias L. Lenz[1]✉

The highly polymorphic human leukocyte antigen (HLA) plays a crucial role in adaptive immunity and is associated with various complex diseases. Accurate analysis of HLA genes using ancient DNA (aDNA) data is crucial for understanding their role in human adaptation to pathogens. Here, we describe the TARGT pipeline for targeted analysis of polymorphic loci from low-coverage shotgun sequence data. The pipeline was successfully applied to medieval aDNA samples and validated using both simulated aDNA and modern empirical sequence data from the 1000 Genomes Project. Thus the TARGT pipeline enables accurate analysis of HLA polymorphisms in historical (and modern) human populations.

The classical human leukocyte antigen (HLA) genes play a central role in adaptive immunity. They encode for glycoproteins that present antigenic peptides on the cell surface for recognition by immune effector cells, thus enabling the immune system to distinguish between 'self' and 'non-self', eventually stimulating a specific immune response[1]. Owing to their implication in hundreds of different complex diseases[2,3], but also because of their importance in human evolution[4,5], HLA molecules have been extensively studied over the past decades.

HLA genes are among the most polymorphic loci known in the human genome[1,2]. At the molecular level, HLA genetic diversity is characterized by a remarkable amino acid sequence diversification[6,7] as well as an enhanced rate of non-synonymous substitutions[8] in the antigen binding groove of HLA molecules (i.e. the pocket where antigens are bound). The specific polymorphism patterns in the exons coding for the antigen binding groove define the several thousands of different alleles found at the classical HLA genes[9]. A complex official nomenclature has been defined for HLA to characterize the extent of its polymorphism. According to this nomenclature, alleles of an HLA gene are defined by the gene name indicating the locus (e.g. HLA-A, -B, -C, -DRB1, -DQB1, -DPB1), followed by a hierarchical numbering system[9]. The 1st field (formerly 2-digit level) defines groups of related alleles. The 2nd field (4-digit) separates alleles within 1st field groups that differ in their protein sequence. Finally, the 3rd and 4th fields define alleles harboring synonymous exonic and non-coding variations, respectively. Additionally, the G-group nomenclature has been introduced in order to merge alleles that have the same nucleotide sequence along the antigen-binding groove and differ in their sequence only outside the groove, thus binding the same repertoire of antigenic peptides.

Past and ongoing pathogen-mediated selection is proposed to be one of the major factors affecting genetic variability at HLA genes[8,10,11]. In addition, HLA genes are associated with various complex genetic disorders in contemporary humans[2], suggesting a link between historical selection by infectious agents and present prevalence of genetic disorders[12,13]. The recent development of genomic tools for the analysis of ancient DNA (aDNA) provides a unique opportunity to unravel the trajectories of alleles associated with human adaptations to newly introduced or co-evolving pathogens[14,15]. In particular, the investigation of ancient HLA genes in historical populations could shed light on the molecular signatures associated with pathogen-mediated selection and promote

[1]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306, Ploen, Germany. [2]Institute of Clinical Molecular Biology, Kiel University, 24105, Kiel, Germany. [3]Present address: Université Paris-Saclay, CNRS, Inria, Laboratoire de recherche en informatique, 91405, Orsay, France. ✉e-mail: lenz@post.harvard.edu

the identification of the exact targets of selection. Thus, reliable genotyping of the HLA genes in ancient samples is crucial to answer unresolved questions ranging from human genetics to evolutionary medicine. However, the high density of SNPs[16] as well as the paralogous organization[17] of the HLA genes makes their appropriate characterization extremely difficult. Because of the nature of such polymorphisms, SNP-based approaches have very limited applicability, while available NGS-based analysis pipelines rely on deep and homogeneous coverage, as it is readily available from modern DNA samples. DNA molecules extracted from skeletal remains, in contrast, are usually heavily fragmented and degraded through chemical modifications[18], hence, sufficient genomic coverage of the endogenous DNA can hardly be obtained, adding a further layer of complexity to reliable allele calls within the HLA region. The fragmentation of the aDNA also prevents any of the primer-based amplification approaches (e.g. SSO or SSP) that are commonly used in clinical settings. Thus, to our knowledge, none of the existing HLA genotyping tools have been proved to be suitable for aDNA samples. In this light, the development of an accurate HLA genotyping method applicable to aDNA samples is a prerequisite for studying the evolution of human resistance or susceptibility to pathogens in historical populations.

Here we present a novel aDNA-optimized analysis pipeline for low-coverage and low-quality shotgun sequence data, which we call 'TARGT' (Targeted Analysis of sequencing Reads for GenoTyping). The pipeline automatically identifies and sorts target-specific sequence reads from any kind of shotgun short-read sequence data. In principle, it can be used to analyze any targeted region in the genome, but was here applied to the HLA, the most polymorphic genes in the human genome. To identify the specific HLA allele combinations of an individual, the TARGT pipeline combines automated read selection and sorting, with highly repeatable semi-manual filtering and HLA allele identification at up to 3rd field (6-digit) resolution (Fig. 1). After pre-processing and quality control, sequence reads from genomic shotgun sequencing are aligned against a comprehensive reference file containing the exon sequences coding for the peptide-binding groove of all known classical HLA gene variants: class I (HLA-A, HLA-B, HLA-C) and class II loci (HLA-DRB1, HLA-DRB3/4/5, HLA-DQA1, HLA-DQB1, HLA-DPA, HLA-DPB1). The TARGT pipeline is highly versatile; indeed, this step could be adapted and used to target any gene or polymorphic region in the genome by providing a corresponding reference sequence for read selection. Mapped reads are then grouped by gene specificity, and saved into sample- and gene-specific FASTA files. Using a sequence alignment software, the FASTA files can be manually analyzed to genotype individual samples.

While the TARGT pipeline can be applied to any kind of shotgun sequence data, we here present it in combination with a targeted DNA capture approach to analyze HLA polymorphisms in a comprehensive set of historical human samples. Target-enrichment by hybridization, also known as DNA capture, is one of the most widely used approaches for sequencing aDNA, because of its efficiency in increasing the sequence coverage of the endogenous DNA fraction[19–22]. In this work, we used a customized DNA capture approach previously developed for modern DNA and based on sequence information from 8,159 known HLA alleles (available at the IMGT/HLA database[9])[23], to enrich a set of DNA libraries from medieval Europeans for the most polymorphic classical HLA class I (HLA-A, -B, -C) and class II (HLA-DRB1, -DQB1, -DPB1) genes. A subset of the obtained HLA class II data has already been analyzed previously in the context of medieval leprosy, yielding a leprosy-associated HLA-DRB1 risk allele[24]. We here assessed the general success of the HLA target-enrichment approach and the overall performance of the TARGT pipeline for both HLA class I and class II genes in the medieval European samples. We then employed a number of different and independent approaches to evaluate the accuracy of the pipeline in producing HLA genotypes. This evaluation includes a comparison with the independent HLA genotyping algorithm OptiType (only available for class I genes, and so far only tested on modern DNA data), with PCR-based results from a specific tag-SNP, and with simulated aDNA sequence data with known HLA alleles. Finally, we explored whether the TARGT pipeline, initially developed for aDNA sequence data, can also be applied successfully to shotgun sequence data from modern populations. For this we applied our approach to a subset of the 1000 Genomes Project samples, for which HLA typing has been performed previously[25]. The TARGT pipeline is freely available for download from SourceForge (https://targt-pipeline.sourceforge.io/).

Results

Success of the HLA target-enrichment for historical samples. One of the main purposes of this work was the accurate genotyping of HLA genes in aDNA samples. Thus, a previously described dataset of sixty-eight samples collected from the medieval cemetery of St. Jørgen (Denmark) was used to assess the success of a HLA target-enrichment approach and the subsequent performance of the TARGT pipeline. Owing to different *post mortem* degradation processes over time, DNA molecules retrieved from ancient organisms hold a high level of base pair modifications. Such DNA damages are the source of incorrect incorporation of nucleotides during DNA amplification, and might cause false SNP calling during the final step of sequencing data analysis. The majority of damage-derived miscoding lesions in aDNA sequences are caused by deamination of cytosine into uracil[26–28]. To reduce the rate of ancient DNA errors, treatment with uracil DNA glycosylase and endonuclease VIII (USER mix) is commonly used during library preparation, which generates and cleaves out abasic sites at deaminated cytosines[29]. Having a minimized amount of miscoding lesions, UDG-treated libraries, can lead to higher accuracy of aDNA sequences, and are thus more reliable for downstream population genetic analysis. On the other hand, non-UDG-treated libraries are commonly used to verify ancient DNA authenticity through the investigation of aDNA features like DNA fragmentation and the above described nucleotide misincorporation patterns[27]. We thus performed shotgun sequencing on both UDG-treated and non-UDG-treated libraries for the whole set of sixty-eight individuals. To assess the ancient origin of DNA sequences, the resulting shotgun sequencing data were aligned against the *H. sapiens* reference genome hg38 and postmortem DNA damage signatures evaluated using mapDamage v2.0.6.[30]. The analysis of damage patterns in the final nucleotide of the sequenced fragments revealed misincorporation frequencies of up to 2.6% in UDG-treated and up to 21.9% in non-UDG-treated datasets (Tables S1 and S2). This confirmed that most of the reads mapping to the human reference originate from

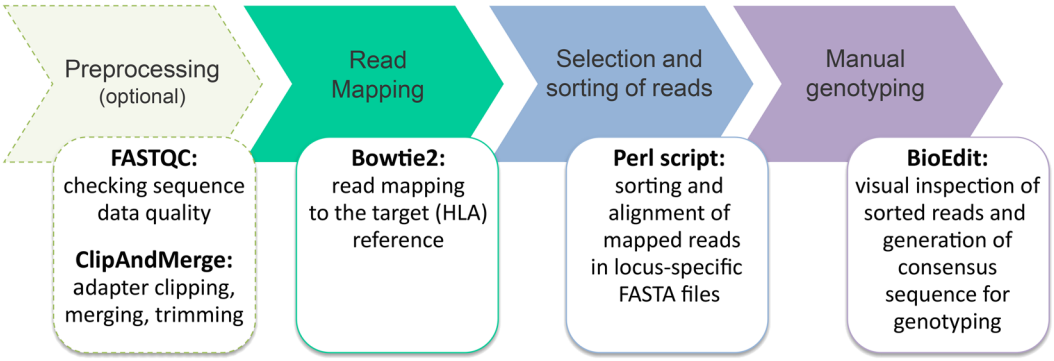


Figure 1. Different steps performed by the TARGT pipeline for HLA genotyping of ancient and modern samples. Preprocessing (optional): after quality control, genomic sequences are pre-processed (adapter clipping, merging and trimming) using ClipAndMerge (version 1.7.3) from the EAGER pipeline[55]. Mapping: performed using Bowtie2[49] against a comprehensive reference file, containing known 3rd field HLA alleles (following G-group nomenclature). Sorting: mapped reads are grouped by gene specificity and saved into gene-specific FASTA files. HLA genotyping: Sample-specific FASTA files are manually analyzed using BioEdit[50] to genotype HLA genes in ancient and modern samples.

aDNA fragments. The degree of DNA fragmentation was also explored to further authenticate ancient DNA. The average length of DNA fragments in UDG-treated datasets ranged from 47 to 101 bp (Table S3).

In response to the fragmentation and low concentration of endogenous DNA from ancient samples, hybridization capture-based target enrichment can be used to improve the yield of DNA molecules for a specific region of interest. An HLA target-enrichment approach was thus applied to the UDG-treated libraries, from here on defined as HLA-enriched UDG libraries, to investigate the highly polymorphic classical HLA class I (HLA-A, -B, -C) and class II (HLA-DRB1, -DQB1, -DPB1) genes. To evaluate the performance of the capture approach, endogenous DNA content, fold-enrichments as well as average coverage and read depth over the HLA genes were quantified on both original UDG libraries and HLA-enriched UDG libraries for a subset of 62 samples (library comparison for six samples was not possible because of technical problems unrelated to the data quality). The number of reads mapping to the human reference genome ranged from 56,925 to 137,542,840 when considering the original UDG libraries and from 244,945 to 74,490,774 when considering the HLA-enriched UDG libraries (Table S4). The corresponding median endogenous DNA content calculated over the whole set of samples was significantly higher for the HLA-enriched UDG libraries (36%) than for original UDG libraries (14%) (Mann-Whitney, $p < 0.001$; Fig. 2A and Table 1). The number of reads mapping to the HLA reference ranged from 0 to 8,452 for the original UDG libraries and was higher when considering HLA-enriched UDG libraries, ranging from 359 to 225,013 (Table S4). Consequently, the % of reads aligning to HLA genes calculated over the whole set of samples differed significantly between original (median: 0.0001%) and HLA-enriched UDG libraries (median: 0.1210%) (Mann-Whitney, $p < 0.001$; Fig. 2B and Table 1). Overall, the HLA enrichment approach performed on the historical samples yielded from 3 to 13,596-fold increases of HLA target sequences compared to the pre-capture condition i.e. the original shotgun sequence data (Table S4). Consistently, the coverage over the HLA genes calculated across the whole set of samples was significantly higher after the enrichment approach (median: 96%) compared to the original shotgun sequence data (median: 9%) (Mann-Whitney, $p < 0.001$; Fig. 2C and Table 1; for calculation at individual HLA loci see Figure S1 and Tables S5 and S6). Similarly, the read depth quantified for the whole set of samples was significantly lower when considering the UDG shotgun libraries (median: 0.2×) compared to read depth achieved after the HLA target enrichment (median: 17.4×) (Mann-Whitney, $p < 0.001$; Fig. 2D and Table 1; for calculation at individual HLA loci see Figure S2 and Tables S5 and S6).

Genotyping of the HLA genes in historical samples. To call HLA genotypes from the aDNA samples, all the sequence data generated from UDG-treated libraries were combined for each sample and processed through the TARGT pipeline. The TARGT pipeline allows HLA allele identification at up to 3rd field (6-digit) resolution; however, as most HLA typing tools and HLA genetic studies rely on 2nd field resolution, we are here only reporting results up to this level. Furthermore, when limited read coverage did not allow for 2nd field resolution, usually because several alleles of the same 1st field allele group were equally well supported, the allele call was rounded to that level of resolution (1st field) (Table S7). Of the 136 alleles (2n) investigated at each locus, we were able to call at 1st field level 83 alleles for HLA-A, 75 alleles for HLA-B, 74 alleles for HLA-C, 83 alleles for HLA-DRB1, 96 alleles for HLA-DQB1, and 46 alleles for HLA-DPB1. Of these, the allele call reached the 2nd field resolution for 45 alleles for HLA-A, 49 alleles for HLA-B, 11 alleles for HLA-C, 58 alleles for HLA-DRB1, 79 alleles for HLA-DQB1, and 46 alleles for HLA-DPB1 (Table 2). The success rate calculated across the whole dataset of ancient samples was 56% at the 1st field level and 35% at 2nd field level (for values at each locus see Fig. 3 and Table 3). As expected, a significant positive correlation between coverage and success rate (1st field, Kendall $\tau = 0.76$, $p < 0.001$; 2nd field $\tau = 0.76$, $p < 0.001$; Figure S3) as well as between read depth and success rate (1st field, $\tau = 0.75$, $p < 0.001$; 2nd field $\tau = 0.75$, $p < 0.001$; Figure S4) was observed, indicating a considerable effect of

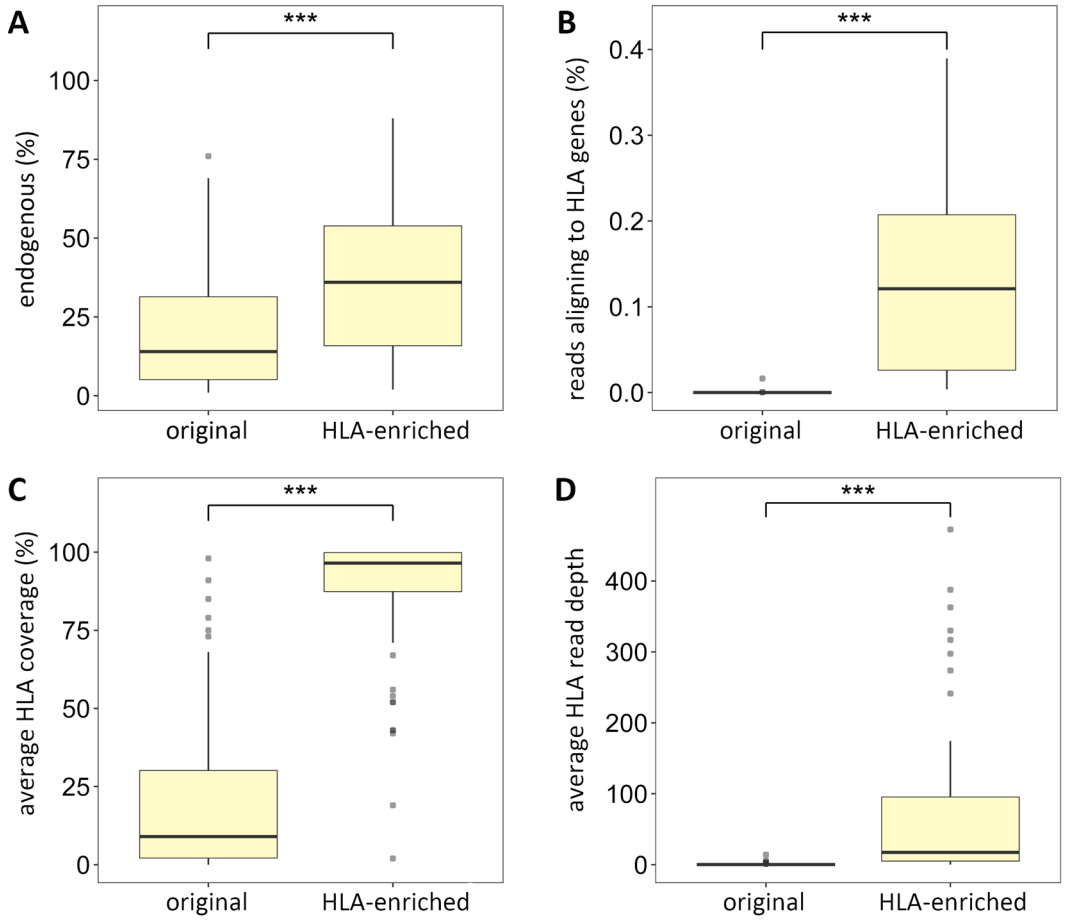


Figure 2. Performance of HLA target-enrichment experiments. Comparison between the median values of (**A**) endogenous DNA content (**B**) percentage of reads aligning to HLA genes (**C**) average HLA coverage and (**D**) average HLA read depth calculated across a subset of 62 aDNA samples before and after performing the HLA enrichment experiments. Significant differences between median values, as derived from Mann-Whitney test, are indicated by horizontal line and asterisks (***$p < 0.001$). Box plots show median, interquartile range, min-max whiskers and outliers.

	Sequence data	Min	Max	Median (95% CI)
endogenous [%]	original	1	76	14 (8–25)
	HLA-enriched	2	88	36 (24–42)
reads aligning to HLA genes [%]	original	0.00	0.02	0.00 (0.00–0.00)
	HLA-enriched	0.00	0.39	0.12 (0.05–0.17)
average HLA coverage [%]	original	0	93	7 (5–13)
	HLA-enriched	2	100	97 (94–99)
average HLA read depth (x-fold)	original	0.00	3.98	0.14 (0.11–0.22)
	HLA-enriched	0.12	472.62	18.00 (9.68–68.95)

Table 1. Performance of HLA target-enrichment experiments. Endogenous DNA content (%), percentage of reads aligning to HLA genes (%), average HLA coverage (%) and average HLA read depth (x-fold) compared between pre-capture shotgun sequence data (original) and sequence data after HLA enrichment experiments (HLA-enriched) obtained from a subset of 62 historical samples.

DNA preservation on allele call success. These associations can explain some cases where poor DNA quality and thus read coverage did not allow for more precise allele calls (Tables 2 and S7). Notably, we found no evidence for more than two alleles at any locus, supporting the notion that the vast majority of DNA fragments in each sample originate from a single individual. These results, together with the examination of misincorporation frequencies and average length of DNA fragments shown above, indicate that the human DNA analysed in each sample is likely to be endogenous.

	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1	HLA-DPB1
1st field	83	75	74	83	96	46
- of these 2nd field	45	49	11	58	79	46
NA (no call possible)	53	61	62	53	40	90

Table 2. Locus-specific allele call success for the 68 historical samples. Number of alleles called at the 1st field level and at the 2nd field level of resolution reported for each locus (2n = 136) for the 68 historical samples.

	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1	HLA-DPB1	Overall
Success 1st field (%)	61	55	54	61	71	34	56
Success 2nd field (%)	33	36	8	43	58	34	35

Table 3. Success rate for the 68 historical samples. Success rate at the 1st field level and at the 2nd field level of resolution, across the 6 investigated genes and overall, for the whole dataset of historical samples.

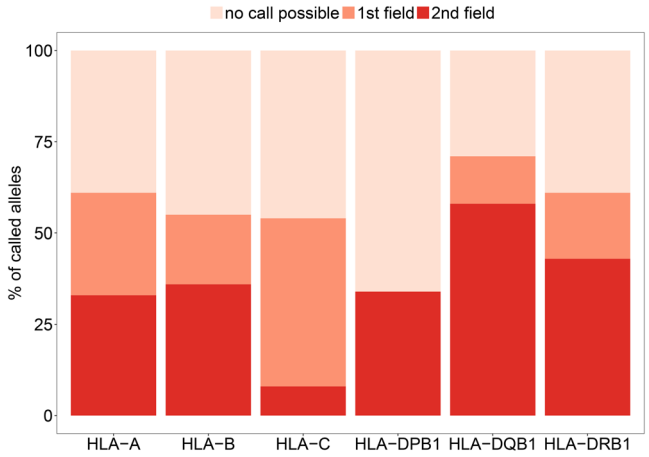


Figure 3. HLA allele call success rate for 68 historical aDNA samples. Percentage of alleles called at each individual HLA locus calculated for the whole set of historical samples (N = 68). Allele calls are reported at two different level of resolution 2nd field (4-digit) and 1st field (2-digit) levels. 'No call possible' represents the fraction of cases where the allele call was not possible or allele calls were ambiguous.

Because of the high level of linkage disequilibrium (LD) within the HLA region, it has been shown that certain SNPs outside the classical HLA genes are informative about HLA types. Such 'tag SNPs' are commonly used to test for association between HLA alleles and disease susceptibility. One example is the T allele at the SNP locus rs3135388, known to be in almost complete LD with the HLA-DRB1*15:01 allele in individuals of European (CEU) ancestry[31]. As reported in the previous study on the sixty-eight medieval samples, this SNP was assayed by PCR and the genotyping results compared to the DRB1*15:01 allele calls obtained with the TARGT pipeline[24]. The study showed perfect correspondence between the tag SNP allele rs3135388-T and the TARGT-based allele calls for the allele DRB1*15:01 where 2nd field allele resolution could be achieved (N = 13), and also with the DRB1*15 call in samples where only the broader 1st field resolution was possible (N = 20, Table S8). That previous study also revealed an interesting observation regarding the specific haplotype structure of the HLA region. Due to the fragmented nature of aDNA, and also due to a lack of intergenic sequence information in the original bait panel of the target capture approach, the haplotype structure of the HLA region cannot be resolved reliably from aDNA. However, the previous study clearly showed a co-occurrence between the allele DRB1*15:01 and the allele DQB1*06:02 (Table S7), suggesting a strong LD between these two loci[24].

HLA class I allele call comparison with OptiType. To our knowledge, OptiType[32] is currently the only HLA genotyping algorithm for which testing of both read depth and read length on prediction accuracy has shown that allele calls appear reliable even for sequence data containing short reads or only a 10x average read depth over the HLA class I loci[32]. These features make OptiType potentially suitable for studying HLA genes from aDNA. However, the algorithm has not been explicitly tested or validated for aDNA and is currently available only for the typing of HLA class I genes. Nevertheless, this tool presently appears to be the only available method to compare with our TARGT pipeline. Thus, to validate our allele calls with an independent approach, a random subset (N = 39) of the historical samples were analyzed using OptiType[32] v1.3.1, and HLA class I genotype information compared with allele calls from the TARGT pipeline. As OptiType was not designed for low quality data such as usually obtained from aDNA, it has no built-in minimum threshold for data quality, and thus always calls

two alleles for a given HLA gene, no matter how spurious the sequence information. For some of the ancient samples, the TARGT approach indicated that low DNA yield and quality, and correspondingly low read coverage, made reliable allele calls impossible at some or all of the investigated HLA genes (Tables S5 and S7). This could be confirmed by visual inspection of the read coverage at the target HLA genes, but OptiType nevertheless produced allele calls also in these cases. However, as it uses the same shotgun sequence data, the reliability of OptiType allele calls are likely to suffer similarly from such data limitations. We therefore excluded those instances from the comparisons, as we were not able to estimate the accuracy of OptiType.

Dividing the number of alleles with identical call in the two approaches by the number of total alleles called, we observed that the two approaches agreed in 93% of the calls. This high agreement rate lends further support to our genotyping approach (Table S9). The number of called alleles that differed between the two approaches (TARGT vs. OptiType) was 12 (7% of all called alleles). One advantage of the TARGT pipeline is that it allows for visual inspection of the supporting sequence reads underlying each allele call. We thus went back to those conflicting allele calls in order to explore the read support for one or the other call. In 5 out of the 12 cases, we found support for our allele call but no support for the call by OptiType. In contrast, in 4 out of the 12 cases, we could not confirm the calls from our approach, but found supporting reads for the allele call by OptiType. In the last three instances, we found that the two different allele calls by the two approaches were both supported and we could not resolve the right allele (Table S9). Note that this evaluation does not include the allele calls by OptiType made with spurious low-quality read data, which are likely to have a significantly higher error rate. Thus, while our approach has a lower success rate in the allele call compared to OptiType, it is likely providing a higher accuracy by avoiding low quality/confidence allele calls. Unfortunately, it is impossible to evaluate this point in more detail as no alternative method is available to obtain the 'true' HLA genotypes of ancient samples.

HLA molecular profile of the historical St. Jørgen samples. As the HLA class II data of these samples have been characterized already in a previous study[24], we here focus on describing the allele frequency distributions at HLA class I genes (-A, -B and -C; reported in Tables S10 and S11). Twelve distinct allele groups ('lineages') at 1st field level of resolution were observed for HLA-A. The A*02 lineage comprises 31% of the total allele pool, with the most common allele A*02:01 seen at a frequency of 0.244. The second most common linage is A*03, with the most common allele A*03:01 also found at a frequency of 0.244. The next two more common linages are A*01 and A*24, for which the most common alleles are A*01:01 (f = 0. 156) and A*24:02 (f = 0.133). The others linages (A*26, A*68, A*11, A*32) are found at frequencies of lower than 10%; while the linages A*29, A*30, A*31 and A*36 as well as the alleles A*02:06, A*31:01, A*32:01 are present as singleton copies. A total of sixteen distinct 1st field level allele lineages were observed at the HLA-B locus, four of which (B*07, B*15, B*44 and B*08) were found at frequencies of greater than 10%. The most common 2nd field HLA-B alleles are B*07:02 (f = 0.204), B*08:01 (f = 0.122), B*40:01 (f = 0.122) and B*44:02 (f = 0.122). The lineage B*42 and the alleles B*27:05, B*35:01, B*35:03, B*45:01, B*55:01 and B*56:01 are found as singleton copies. At the HLA-C locus, a total of eleven distinct lineages at 1st field level of resolution were observed. The three most common lineages (C*07, C*03 and C*04) comprise together over 70% of the total allele pool. The 2nd field level allelic resolution at HLA-C locus was lower as compared with HLA-A and –B loci, nevertheless a total of five distinct 2nd field level alleles were found, with the allele C*07:01 being the most common subtype.

A likelihood ratio test, implemented in Pypop[33], was used to test the significance of observed linkage disequilibrium (LD) between any two loci (Table S12). The analyses were performed removing individuals with NA at all loci while keeping only allele calls that reached the 2nd field level of resolution. As expect from modern day genetic data, for which high LD has been documented in the HLA region[34–36], strong linkage signals were also revealed for the historical samples, both within class I and class II as well as between class I and class II loci (Table S12). Locus HLA-A showed significant associations with locus HLA-B (D' = 0.804; W_n = 0.844), HLA-C (D' = 1; W_n = 1), HLA-DRB1 (D' = 0.703; W_n = 0.755) and HLA-DQB1 (D' = 0.764; W_n = 0.709), while no global LD was observed with the HLA-DPB1 locus. Locus HLA-B showed nonrandom associations with all loci: HLA-C (D' = 0.833; W_n = 0.829), HLA-DRB1 (D' = 0.882; W_n = 0.775), HLA-DQB1 (D' = 0.877; W_n = 0.772) and HLA-DPB1 (D' = 0.829; W_n = 0.823). Global LD was revealed also between HLA-C and HLA-DRB1 (D' = 0.833; W_n = 0.882) and between HLA-C and HLA-DQB1 (D' = 0.875; W_n = 0.913). A particularly strong associations was observed between the two adjacent HLA class II loci HLA-DRB1 and HLA-DQB1 (D' = 0.984; W_n = 0.883), while no global LD between the DP locus and the other class II genes were found. Two- and three-locus haplotype frequencies were estimated using the expectation-maximization algorithm. We confirmed the presence of the previously described class II haplotype DRB1*15:01-DQB2*06:02[24], found at a frequency of 27%. Further common haplotypes were B*07:02-DRB1*15:01 (f = 0.267) and B*07:02-DQB1*06:02 (f = 0.222), suggesting the presence of an extended class I-II haplotype, B*07:02-DRB1*15:01-DQB1*06:02 (f = 0.286). Further extended haplotypes were also observed at appreciable frequencies: A*02:01-DRB1*15:01-DQB1*06:02 (f = 0.125) and B*08:01-DRB1*03:01-DQB1*02:01 (f = 0.143). Several other common two- and three-locus haplotypic associations were observed and are reported in Table S13.

TARGT pipeline validation on simulated aDNA samples and detection of unknown HLA alleles. To assess the efficacy of the TARGT pipeline, we applied it to simulated aDNA sequence data with different read depth levels. Instead of downsampling available human genome data to match the ancient DNA read depth at HLA loci, we took advantage of the gargammel package, a software that can simulate ancient DNA fragments from provided genome sequences[37]. This approach allowed us to consider how both typical aDNA fragmentation and damage pattern can affect reliable allele calls at HLA genes, to verify if and how performance declines with lower read depth, and also to test for the detection of unknown HLA alleles. Seven haplotypes with known but distinct HLA-B and -DRB1 alleles were artificially generated and combined in six different 'diploid' combinations (Table S14). Using the program gargammel[37], typical aDNA fragmentation and damage patterns

	HLA-B	HLA-DRB1	Overall
Success 1st field (%)	62	80	71
Success 2nd field (%)	50	64	57
Accuracy 1st field (%)	100	100	100
Accuracy 2nd field (%)	100	100	100

Table 4. Success rate and accuracy rate for simulated ancient DNA samples. Success and accuracy rate of HLA allele calls, at the 1st field level and at the 2nd field level of resolution, across the 2 investigated genes and overall, for the simulated ancient samples.

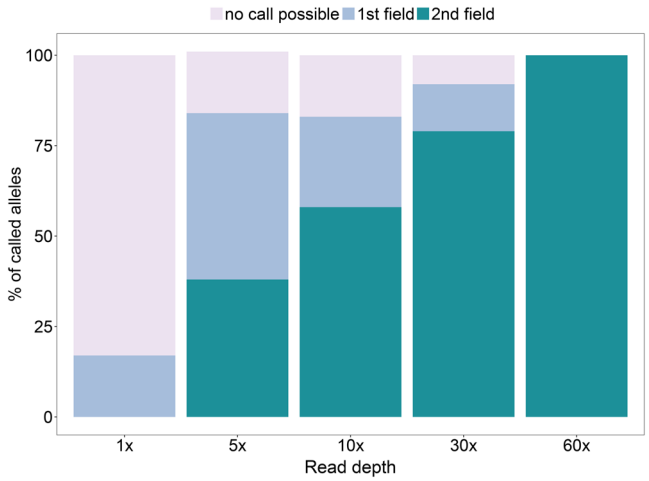


Figure 4. HLA allele call success rate for the simulated ancient samples. Percentage of allele calls calculated across the two investigated loci (HLA-B and HLA-DRB1) and across the set of samples (N = 6) investigated at different read depth (from 1x up to 60×) for a total of 30 simulated ancient samples. Allele calls are reported at two different levels of resolution, 2nd field (4-digit) and 1st field (2-digit). 'No call possible' represents the fraction of cases where the allele call was not possible or allele calls were ambiguous.

were introduced in the sequences and the TARGT pipeline tested for increasing read depth from 1x up to 60×, for a total of 30 simulated ancient samples. The HLA allele calls from the TARGT pipeline (here only run on the artificially varied genes HLA-B and HLA-DRB1) were compared to the original known HLA genotypes. If allele calls were not possible, we reported 'NA'. Allele calls naming several equally well matching 1st field allele groups were considered ambiguous and also reported as 'NA'. The success rate, defined as the proportion of times an allele call was possible with our approach across the different samples and read depth levels, was 71% at the 1st field level and 57% at the 2nd field level (Table 4). Whenever a call was possible, the allele calls were correct, thus we observed an accuracy rate of 100% for both the 1st field and 2nd field levels (Table 4). However, the 30 ancient simulated samples varied considerably in read depth (from 1× up to 60×), and as expected, the success rate was significantly positively correlated with the read depth. Association between read depth and success rate was observed at both the 1st field and 2nd field levels (Figs. 4 and S5). As the entire set of simulated ancient samples where homogeneous in terms of coverage (i.e. the proportion of covered sites at each locus) we could not test its effect on HLA allele calls in simulated aDNA samples. Intriguingly, we observed that 1st field allele calls were possible already at 1× read depth, while allele calls at 2nd field resolution were obtained starting from a read depth of 5×(Fig. 4). Notably, for some samples the resolution of HLA alleles was not possible even at moderate read depth (30×), underlining that allele call success does not depend exclusively on read depth. Indeed, the specific combination of alleles at each genotype can significantly affect the success in calling the HLA alleles. For instance, highly similar alleles that differ only by a few nucleotides would not be resolved even in samples with high read depth, if those few nucleotide positions happen to not be covered by any read. On the other hand, if a particular region that differentiates the two alleles is covered by only a few reads, an accurate allele call might be possible even if the rest of the gene is not covered at all. We further tested if the TARGT pipeline can detect unknown alleles, such as HLA alleles that were present in human history but no longer exist in modern populations, or extremely rare alleles that are not represented in the HLA reference database. To do so, we introduced three point mutations in two out of the seven artificially generated haplotypes. All of them were well detected starting from a read depth of 5× (Table S14). These results confirm that the mapping of shotgun sequence reads to a reference file containing known HLA alleles, an inherent component of the TARGT pipeline, does not prevent the detection of novel alleles.

Pipeline validation on 1000 Genomes Project samples. The TARGT pipeline was initially developed with the aim to define HLA alleles from aDNA sequence data. However, since shotgun low-coverage resequencing

	HLA-B	HLA-DRB1	Overall
Success 1st field (%)	100	100	100
Success 2nd field (%)	84	95	90
Accuracy 1st field (%)	95	100	99
Accuracy 2nd field (%)	95	100	97

Table 5. Success rate and accuracy rate for the 1000 Genomes Project samples. Success and accuracy rate of HLA allele calls, at the 1st field level and at the 2nd field level of resolution, across the 2 investigated genes and overall, for a diverse subset (N = 31) of the 1000 Genomes Project samples.

of modern population samples is becoming more and more common, our pipeline might also be useful in that context. In order to test whether the TARGT pipeline can also be successfully applied to shotgun sequence data from modern populations, we genotyped the classical HLA genes in a diverse subset of individuals from the 1000 Genomes Project, for which HLA genotype information has previously been published (N = 31; Tables S15–S17). The majority of the allele calls obtained through the TARGT pipeline were consistent with the ones obtained through independent PCR-based genotyping in Gourraud *et al.*[25], as evidenced by the high agreement rate of 96% (Tables S16–S18). The total number of called alleles that differed between the two approaches was five (4%). After careful inspection of the available read data, we found supporting evidence for the allele call by Gourraud *et al.* in three out of the five cases. In contrast, in two of the five cases, the data did not support the call by Gourraud *et al.* but instead clearly confirmed our allele calls, showing that the TARGT pipeline can rectify erroneous calls from PCR-based HLA genotyping. The success rate for allele calling obtained with the TARGT pipeline was 100% at the 1st field level and 90% at the 2nd field level. Indeed, the majority of the called alleles could be defined at the 3rd field level, using the G-group nomenclature (Table S18). When looking at individual loci, a higher success rate was achieved for allele calls at the HLA-DRB1 locus compared to HLA-B (Table 5). The accuracy rate of the called alleles was 99% at 1st field (2-digit) resolution, and 97% at 2nd field (4-digit) resolution. Also in this case, higher accuracy was obtained for the allele calls at the HLA-DRB1 locus in comparison to the HLA-B, at both 1st field and 2nd field resolution (Table 5).

Discussion

The investigation of HLA genes in ancient and modern human populations is of great interest to answer unsolved questions in the fields of biomedicine and evolutionary biology. As independent targeted HLA genotyping is very costly, the possibility to obtain HLA genotypes from the low-coverage shotgun sequence data that is now routinely generated in population genomic studies would be highly advantageous. We therefore present here the novel TARGT pipeline for HLA genotyping from low-quality shotgun data, and evaluate its accuracy for ancient and modern DNA samples. The pipeline can be applied directly to shotgun sequence data or can be combined with a target enrichment approach.

Widely used in the aDNA field, the target-enrichment approach can effectively recover endogenous DNA fractions of interest, targeting SNPs[38,39], whole chromosomes[21,40], mitochondrial[19,41,42] or nuclear genomes[20,22,43,44].

We here applied a customized DNA capture approach[23] to enrich a defined set of sixty-eight aDNA libraries for the classical HLA class I (HLA-A, -B, -C) and class II (HLA-DRB1, -DQB1, -DPB1) genes. Endogenous DNA content, percentage of reads aligning to HLA genes as well as coverage and read depth over the HLA genes were compared between shotgun libraries and HLA-enriched libraries (Fig. 2 and Table 1). The comparison between pre-capture shotgun sequence and sequence data obtained after HLA target enrichment yielded from 3-fold to 13,596-fold increases of HLA target sequences and clearly showed the efficiency of the HLA enrichment approach in increasing the number of reads mapping to the HLA genes of interest. These results highlight the advantage of the HLA target enrichment for studying HLA polymorphisms in ancient human populations.

We then applied the TARGT pipeline (Fig. 1) to the aDNA sequence data and evaluated the resulting HLA allele calls. The success rate was variable across the different loci with the highest success rate obtained at the HLA-DQB1 locus, followed by HLA-A, HLA-DRB1, HLA-B and HLA-C; while the HLA-DPB1 locus showed the lowest success rate (Fig. 3 and Table 3). As the achieved average coverage was comparable across the different loci (Figure S1 and Table S6), the differential success rate is unlikely to be due to an unbalanced bait design. The general differences in both the allelic diversity as well as in the extent of sequence divergence between alleles at the different HLA loci, together with the uneven distribution of nucleotide diversity along the investigated exons, are more plausible explanations for the observed variable resolution across loci. Furthermore, DNA preservation of individual samples can affect allele call success as shown by the strong association between success rate at each sample and both read depth and coverage. Thus, the observed success rate calculated for the historical dataset (56% at 1st field resolution; 35% at 2nd field) cannot be generalized for other ancient samples, as it will likely vary depending on the spatial and temporal scales investigated as well as on the degradation of the underlying DNA.

The allele calls obtained with the TARGT pipeline were evaluated using an independent approach, the software OptiType[32], which is currently available only for the typing of HLA class I genes. The comparison revealed a high level of agreement between the two approaches (93%) for the HLA genotyping of class I alleles. Moreover, the presence of the most frequent variant at one of the class II genes, the allele DRB1*15:01, was supported by the independent detection of a specific tag-SNP in a previous study[24]. Both comparisons provided further evidence for the accuracy of our allele calls, supporting the validity of the TARGT pipeline.

Our pipeline was also evaluated on simulated aDNA sequence data with distinct but known HLA-B and -DRB1 alleles. Here we observed a success rate of 71% at the 1st field level and 57% at 2nd field level of resolution, both with an accuracy of 100% (Table 4). We tested the pipeline for increasing read depth (from 1× up to 60×)

and observed also in this case a strong association with success rate. However, we noticed that in some cases alleles could be called at very low read depth: some 1st field alleles could be called already at 1x depth, while calls at 2nd field resolution were obtained starting from 5× read depth (Fig. 4). In contrast, in a few instances, the resolution of HLA alleles was not possible even at moderate read depth (30×), suggesting that read depth is not the only factor influencing the allele call success. Whilst additional sequence reads in terms of read depth allow the identification of sequencing errors and can provide support for specific allele calls, informative overlap among reads to build a consensus sequence of sufficient length is also essential to obtain reliable HLA genotypes. Indeed, the proportion of covered sites at each locus together with the specific combination of alleles can significantly affect the success in calling HLA alleles. For instance, if the alleles of a given genotype differ only in positions that are not covered by any reads, a full allelic resolution will be impossible, even if the rest of the sites are covered at high depth. With the allele call success depending on various properties of a given DNA sample, it would be inappropriate to define a default threshold of read depth or coverage for applicability of the TARGT pipeline, especially for ancient samples. However, the advantage of our semi-manual approach is that the experimenter receives direct visual feedback about the quality and quantity of the sequence data and can make an informed decision about the reliability of any allele call. The possibility of visual inspection of the raw data is particularly crucial in case of aDNA samples; this feature sets our pipeline apart from other more automated approaches (presently only available for modern DNA), some of which will always provide an allele call, no matter how spurious and ambiguous the underlying sequence data. During the evaluation process on simulated aDNA sequences we also observed that the HLA allele calls obtained with the TARGT pipeline achieved high accuracy even in samples whose HLA alleles are not included in the original HLA reference file. These results show that our new approach has the potential to detect unknown alleles, such as extremely rare alleles that are currently absent in the HLA reference database or alleles that were present in the past but no longer exist in modern populations.

We then also applied our pipeline to a subset of the 1000 Genomes Project samples in order to test its applicability for shotgun sequence data of modern DNA from population genomic projects. Modern DNA usually exhibits no degradation or extensive fragmentation, thus yielding much longer sequence reads and more even coverage, making allele calling much easier. Consequently, we observed a much higher success rate (100% at the 1st field; 90% at the 2nd field level of resolution) compared with both the empirical and simulated ancient datasets (Table 5). These results, together with the high accuracy rate observed (99% at the 1st field level; 97% at 2nd field level), suggest that the TARGT pipeline, originally developed to define HLA alleles in aDNA samples, can also be successfully applied to shotgun sequence data from modern DNA.

The pipeline includes a pre-processing part with several automated steps essential for the analysis of any next-generation sequencing data: quality control, adapter clipping, and merging of paired reads (Fig. 1). This part is optional and can be applied if raw sequence data is to be used for HLA analysis. However, if WGS/WES shotgun data has already been quality-checked and trimmed for other purposes, this can also be used directly as input data for the TARGT pipeline, which would then start with the mapping and sorting steps (Fig. 1). Important in the latter case is the awareness that duplicate reads will likely have been removed during quality filtering. For modern sequence data with decent coverage, this might not be a problem, but for data from aDNA, information about read abundance can be an important parameter during allele calling. In this case, it might be advisable to start the pipeline on the raw data instead, and redo the quality filtering specifically with the TARGT pipeline, not discarding duplicate reads. For the mapping step, it is also recommended to carefully consider the optimal mismatch threshold for successful mapping of reads to the HLA reference. A very stringent threshold (e.g. 0% mismatch allowed) will lead to individual reads mapping to fewer HLA loci/alleles in the reference and thus provide less ambiguous read data for allele calling. However, such a stringent threshold might also lead to missing of novel/unknown alleles in a sample, as their specific reads might not map well enough to any known allele and thus would be thrown out. This trade-off should be considered for each given dataset/project, and it might be advisable to run the pipeline multiple times with different thresholds in order to detect the presence of unknown alleles. The default threshold of 1% mismatch balances this trade-off and appears to be a good starting point for most datasets in our experience. The subsequent visual inspection of HLA sequence reads is a critical step for the correct calling of HLA alleles, especially with aDNA samples where miscoding lesions and fragmentation, in combination with the high density of SNPs and paralogous sequences naturally present in the HLA region, can easily lead to incorrect allele calls. The manual identification of HLA alleles allows detection of small differences between alleles and was successful in detecting novel HLA variants in our evaluation. Furthermore, discrepancies between different PCR-based methods routinely used for HLA typing have been observed[45], highlighting that inaccurate HLA allele calls could be a problem also in case of modern samples. In this context, we have shown that our approach successfully identifies incorrect genotypes and thus allows validation of HLA allele calls from other, even more established, methods.

Despite the advantage of reaching high accuracy by preventing low quality/confidence allele calls, we recognize that our approach can be time-consuming and that the visual inspection of the sorted reads might depend on the experience of individual researchers. However, in a previous study on medieval leprosy victims, it was shown that the results of the TARGT pipeline, including the manual allele call by different researchers, were highly reproducible, reaching >99% reproducibility at the nucleotide level[24]. Eventually, every genotyping approach has its advantages and disadvantages, and it has proven difficult to establish a single best-performing method among the growing number of available computational tools for HLA typing[45]. In such a situation, the most appropriate approach would be to use consensus information that integrates results from different complementary methods. In this context, the TARGT pipeline has a true advantage by providing an independent, non-automated allele call that includes visual inspection of the underlying sequence data and intuitive feedback about the reliability of the call (also with regard to calls from other methods that use the same sequence data).

Our results show that the TARGT pipeline is an accurate method for HLA genotyping in case of low-coverage shotgun sequence data. The pipeline also allows for the detection of unknown alleles that are not included in

the original HLA reference database. The observed solid performance demonstrates that TARGT is a reliable approach to accurately genotype HLA genes in ancient and modern DNA samples. The pipeline has already been applied successfully to a dataset of medieval European samples, associating HLA variability with susceptibility to leprosy[24], and indicating its applicability to study the evolution of human resistance or susceptibility to pathogens in historical populations. In addition, the pipeline could be employed to explore HLA allele frequency changes through time, when temporal sample series are available[46], thus providing a deeper understanding of HLA genetic variation through human history.

Methods

HLA typing from shotgun sequence data (TARGT pipeline). *HLA reference file for read mapping.* A key component of the TARGT pipeline is a comprehensive HLA reference file containing all known nucleotide sequence variants of the exons coding for the peptide-binding groove of the classical class I (exons 2 and 3; HLA-A, HLA-B, HLA-C) and class II HLA genes (exon 2; HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA, HLA-DPB1). This reference allows differentiation of HLA alleles at up to 3rd field (6-digit) resolution using the G-group nomenclature. The G-group nomenclature groups together HLA alleles whose peptide-binding domains are identical at the nucleotide level (and thus also at the protein level[47]). The reference also contains corresponding sequence variants for non-classical HLA genes, to avoid mis-mapping and misidentification of reads, due to paralogous sequence similarity. To build this reference, nucleotide coding sequences were downloaded from the IMGT/HLA database[9] (accessed 28 July 2015) for the following loci: HLA-A, -B,-C, -E, -F, -G, -H, -J, -K, -L, -U, -V, -DQA1, -DQB1, -DRA, -DRB1, -DRB3, -DRB4, -DRB5, -DRB6, -DRB7, -DRB9, -DPA1, -DPB1. Exon sequences of HLA-DQB2 from the human reference genome (not represented in the IMGT/HLA fasta files), was also included, again to preclude misidentification of its reads as HLA-DQB1 variants. Alignment of selected nucleotide sequences was then performed individually for each locus using the program muscle[48] v3.8. Gaps caused by rare non-functional alleles were removed as well as overhangs upstream and downstream of the exons of interest. Redundant sequence variants that are identical within the exons of interest were also removed (following the G-group nomenclature). One hundred nucleotides (Ns) were introduced upstream and downstream of all sequences, while 20 Ns were introduced between exon 2 and 3 for class I loci in order to allow for mapping of reads crossing the exon-intron border, as the intron sequences for most alleles are not available from the IMGT/HLA database. All aligned sequences were combined into one FASTA file, which was finally indexed using Bowtie2 to produce the final HLA reference file.

Read mapping. The standard input for the TARGT pipeline is qc-filtered and adapter-trimmed short-read sequence data in fastq format. The first step maps the sequence reads of the sample to the HLA reference file using Bowtie2[49] v2.2.7 in local alignment mode. Bowtie2 allows mapping of both merged reads and separate paired-end reads. By using the '-a reporting mode', we allow each read to map against multiple alleles in the reference, which is crucial because of the expected ambiguous mapping of most reads. To achieve a maximum mismatch threshold of 1%, the minimal alignment score was set to –score-min L,0,0.99, while keeping the local alignment matching bonus setting of –ma 1, and the maximum (MX) and minimum (MN) mismatch penalties equal to –mp 0,0. Allowing for 1% mismatch represents a balanced trade-off between mapping sensitivity and specificity, while at the same time enabling the identification of unknown alleles (not present in the reference file). However, this mismatch threshold can be varied, and should be carefully chosen with regard to the specific study question (see discussion).

Automated read sorting. Output from mapping with Bowtie2 contains reads that aligned best exactly one time to the HLA reference file as well as ambiguous reads i.e. reads that map equally well to multiple alleles in the reference. Such ambiguous reads can map to multiple alleles of the same locus and to alleles from different loci. In the latter case, multiple instances of the same read sequence, one for each distinct mapping locus, are stored in the resulting alignment. The mapping information from the Bowtie2 output is processed with a Perl script (included in the pipeline scripts on Sourceforge). During this processing procedure, identical reads, representing PCR duplicates of the same DNA fragment, are collapsed and their absolute frequency is noted. For each read, the number of duplicates as well as the mapped locus name(s) and number of alleles per locus are stored in the read's fasta tag. Read sequences are then grouped by locus and saved into a FASTA file in a specific format: sequences are ordered, in ascending order, according to the number of genes they map to and then sorted by the starting position within the corresponding locus so that they follow the sequence orientation along the locus of interest (Figure S6). The thus generated locus-specific FASTA files can then be inspected for manual allele calling using a sequence alignment editor.

Manual read sorting and allele calling. For the manual read analysis and allele calling, we use the freely available and versatile sequence alignment editor BioEdit[50] v7.2.565, which permits visual inspection and manipulation of sequence reads. However, in principle, any sequence alignment editor can be used for this purpose as long as it facilitates the steps described here. The locus-specific FASTA files generated by the above Perl script were opened in BioEdit and consensus sequences of the allele combinations present in each sample were generated. First, by visually inspecting sequence identity among reads, true SNPs (identifying the true alleles) were distinguished from PCR/sequencing errors. Sequence reads representing the possible alleles were then identified and sorted into blocks of reads belonging to the same allele (Figure S6). Here, reads were prioritized that map uniquely to the locus of interest, and/or that were represented by multiple exact PCR duplicates (less likely to represent PCR/sequencing errors). Overlapping reads that shared the same combination of variations were collapsed into a consensus sequence. To identify matching alleles, consensus sequences were compared to a reference alignment of all known 4-digit alleles of the corresponding HLA locus. Such comparison was performed first focusing on an

established set of 'common and well-documented' HLA alleles[51]. If the consensus sequences perfectly matched one or more alleles from that set, we did not look for additional matches in the rarer alleles; otherwise, the full set of all known alleles of that locus was screened for best-matching sequences. The full nucleotide sequence of the identified allele was finally compared to the read alignment to confirm that the allele call was indeed supported by all high confidence reads. In case of several equally well matching alleles belonging to the same two-digit allele group (e.g. because of incomplete coverage), only the 1st field (two-digit) allele name was reported.

Sequence data from historical samples. *Historical samples.* The human skeletal remains whose DNA was analyzed in this study were obtained from the medieval cemetery of St. Jørgen/Denmark. Sixty-eight individuals were considered for this study, ranging in age from 1270 to 1536 AD, with most of the individuals falling between 1270 and 1400 AD[24]. Sample processing, DNA extraction and DNA library preparation have previously been described in Krause-Kyora *et al.*[24]. DNA extractions and pre-PCR steps were performed in clean room facilities dedicated to aDNA research, following the guidelines on contamination control in aDNA studies[52–54]. For each sample, two different double-stranded DNA sequencing libraries (UDG-treated and non-UDG-treated) were prepared. Both UDG-treated and non-UDG-treated libraries underwent paired-end shotgun sequencing carried out on the Illumina HiSeq 2500 (2 × 125 bp) and HiSeq 4000 (2 × 75 bp) platform at the Institute of Clinical Molecular Biology, Kiel University, using the HiSeq v4 chemistry and the manufacturer's protocol for multiplex sequencing.

HLA target-enrichment for historical samples. UDG-treated libraries were enriched for DNA from the classical class I (HLA-A, HLA-B, HLA-C) and class II HLA genes (HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA, HLA-DPB1), using a custom bait library designed by Wittig *et al.*[23]. The HLA capture probes have been originally created considering the full list of available cDNA and gDNA sequences from the IMGT/HLA reference database[9] (i.e. 8,159 alleles), which resulted in a total of 16,351 distinct RNA baits, covering a cumulative target genomic sequence of 215.5 kb[23]. The in-solution targeted capture has been performed using the SureSelectXT Target Enrichment System (Illumina) for the Illumina paired-end multiplexed sequencing library (version B4, August 2015). For each capture reaction, up to four UDG-treated libraries have been pooled. The hybridization reaction required 800 ng of library DNA per pool in a volume of 3.4 µL. As the UDG-treated libraries were already indexed during library preparation, the 12 cycles of post-capture PCR was performed using 1 µL of each IS5 and IS6 primers (100 µM). According to the protocol, the resulting amplified captured libraries were purified using the AMPure XP beads, while quality assessment was performed on the Agilent 2100 Bioanalyzer with the High Sensitivity DNA Assay. Finally, sequencing was done on the Illumina HiSeq 4000 (2 × 75 cycles) platform at the Institute of Clinical Molecular Biology, Kiel University, using the HiSeq v4 chemistry and the manufacturer's protocol for multiplex sequencing.

Data preprocessing for historical samples. HTS data sets generated for the sixty-eight individuals from St. Jørgen were pre-processed (adapter clipping, merging, trimming) using ClipAndMerge (version 1.7.3) from the EAGER pipeline[55]. During the adapter clipping step, adapters were excluded when present in the sequence, while reads with fewer than 25 nucleotides after adapter clipping or containing only adapters sequences were removed. In the merging step, all remaining paired reads were merged with a minimum overlap of 10 nucleotides and at most 5% mismatches in the overlap region. In the final quality trimming phase, all nucleotides with Phred scores smaller than 20 were trimmed from the 3' end of each read, while sequences shorter than 25 nucleotides after quality trimming were removed. In order to evaluate postmortem DNA damage signatures, using mapDamage v2.0.6[30], shotgun sequencing data from both UDG-treated and non-UDG-treated libraries were aligned against the *H. sapiens* reference genome hg38 (GRCh38) using Bowtie2[49] v2.2.7, in a semi-global alignment mode and with default parameters, as described in Krause-Kyora *et al.*[24]. Read duplicates were not removed during the pre-processing and quality filtering steps, as read redundancy information is used during manual HLA allele call for identifying sequencing artifacts. Endogenous DNA content, percentage of reads aligning to HLA genes as well as coverage and read depth over the HLA genes were quantified on both the original UDG shotgun libraries and the HLA-enriched UDG shotgun libraries. Endogenous percentage was measured by the proportion of reads mapping to the human reference genome over the total amount of reads. Percentage of reads aligning to HLA genes was calculated as the proportion of reads mapping to the HLA reference over the total amount of reads. Fold-enrichment was calculated by dividing the number of on-target reads, i.e. reads mapping to the HLA reference, from HLA-enriched libraries by the number of on-target reads from pre-capture shotgun libraries; when the denominator was 0 the number of on-target HLA reads from enriched libraries has been assigned. Coverage was calculated as the proportion of covered sites at each locus. Because of the extensive number of reads mapping to multiple HLA loci, read depth (i.e. number of times a base at a given locus is sequenced) has been calculated weighing the read length for the number of loci that each read would map to. Also, to exclude reads containing PCR/technical duplicates, we considered each read with the same starting and ending position only once. Average HLA coverage and average HLA read depth are the mean of coverage and read depth calculated across the 6 investigated class I (HLA-A, -B, -C) and class II (HLA-DRB1, -DQB1, -DPB1) genes.

Statistical analysis for historical samples. Allele frequencies in historical samples for class I genes (HLA-A, -B and -C) at the 1st field level and 2nd field level of resolution were obtained by direct counting. Pairwise estimates of nonrandom associations between each pair of HLA loci, i.e. linkage disequilibrium (LD), as well as frequent haplotypes in high LD were determined using PyPop[33] with defaults settings. PyPop is a software pipeline, originally developed for the analysis of highly polymorphic human leukocyte antigen (HLA) data, and thus useful

to perform genetic statistics from multilocus genotype. Overall LD between pairwise HLA genes was defined through two measures. The first one is the normalized Hedrick's D' statistic (D') which weights the LD contribution of specific allele pairs by the product of the allele frequencies at each locus[56]. The second one is the Cramer's V statistic (W_n) which defines the normalization between zero and one of the chi-square statistic for deviations between observed and expected haplotype frequencies[57]. The normalized LD values (D' and W_n) range between 0 and 1. The permutation distribution of the likelihood-ratio test has been used to test the significance of the overall LD between pairwise HLA genes[58]. Two- and three-locus haplotype frequencies were estimated from the ancient genotypic data, using the iterative expectation-maximization (EM) algorithm[59,60]. The analyses were done removing individuals with NA at all loci while keeping only allele calls that reached the 2nd field level of resolution.

Allele call comparison to OptiType pipeline. To compare the allele call results obtained with the TARGT pipeline with an independent method, sequence data of the historical samples were also analyzed using OptiType[32] v1.3.1. The OptiType pipeline in its present form allows only analysis of HLA class I loci. Results of the genome-wide alignment against the human reference were used as input and FASTQ files were generated from aligned BAM files using samtools. OptiType was then applied in DNA mode with default settings.

Sequence data from simulated aDNA samples. To validate our HLA genotyping pipeline, we generated simulated aDNA data from genomes with known HLA variants. For this, we first created seven unique MHC haplotypes containing known HLA-B and -DRB1 alleles. The nucleotide sequence of the classical MHC region (chr6: 29,640,000–33,120,000)[61] was downloaded from the UCSC Genome Browser, using the human reference genome GRCh38. The exons forming the variable region in the peptide binding groove (i.e. exon 2 and 3 for the HLA-B locus and exon 2 for the HLA-DRB1 locus) of known alleles were first aligned and then manually edited using BioEdit[50]; thus creating haplotypes with different alleles from the reference genome. Additionally, 'de-novo mutations' were introduced in two out of seven haplotypes: in the first haplotype, a point mutation was introduced at each locus (HLA-B and -DRB1); while in the second haplotype a point mutation was introduced only at the HLA-DRB1 locus. The 7 unique HLA haplotypes were then combined in six different diploid combinations of heterozygous genotypes (Table S14). Typical bias observed in aDNA samples (fragmentation and damage patterns) were then introduced using the program gargammel[37]. For each genotype, we created five aDNA paired-end read datasets with increasing read depth (1×, 5×, 10×, 30×, 60×), for a total of 30 simulated aDNA samples. In simulating DNA fragmentation, fragment size was calculated considering the average fragment length observed in the set of medieval European samples tested in this study. In the same way, deamination patterns and base content profile were also obtained from one of the investigated ancient samples (G507). One of the advantages of the capture approach is that it can drastically reduce the extensive microbial contamination often present in ancient samples; we thus did not introduce microbial contamination while simulating the ancient HLA regions. Ideally, the assessment of the ancient origin of DNA sequences can be evaluated, and ancient samples suspected to contain human contamination should be excluded from the analysis. Because of that, human contamination was also not introduced in our simulated samples. To avoid any observer bias in the allele call of simulated aDNA samples, the HLA genotyping was performed by two independent researchers that were not aware of the specific alleles introduced in the simulated samples.

Modern sequence data from 1000 Genomes Project. In order to assess the applicability of our pipeline for shotgun sequence data from modern samples, we used whole-exome shotgun sequence data from the 1000 Genomes Project[62]. Paired-end sequencing datasets from 31 individuals of diverse ancestry (8 Africans, 8 East Asians, 7 Americans, and 9 Europeans) were downloaded from the 1000 Genomes Project database (phase 3) (Table S15). Only samples with available SBT-based HLA genotype information published in Gourraud *et al.*[25] were included (Tables S16 and S17).

Evaluation of HLA allele calling pipeline. To assess the reliability of the TARGT pipeline, three different measures were defined. The success rate quantifies the proportion of cases where an allele call was possible, in both the empirical datasets (historical samples and 1000 Genomes samples) as well as in the simulated aDNA samples. It was defined as the ratio between the number of called alleles provided by our approach and the total number of the alleles assayed (two per locus and sample). The success rate reported here can be considered conservative since ambiguous results (i.e. allele call with several equally well matching alleles belonging to different two-digit allele groups) were reported as NA (allele call 'not available'). The measure of agreement was used to compare the allele calls from two independent methods. It was calculated as the proportion of identical alleles typed using the two approaches over the total number of the alleles called, thus excluding alleles for which allele call was not possible in one or both approaches. The accuracy rate was used to assess the confidence of HLA genotypes provided with our approach, when HLA alleles were known a priori, as in the case of the simulated ancient samples, or when HLA alleles have been previously typed with different approaches in the case of 1000 Genomes samples. It was calculated as the proportion of correctly called alleles over the sum of correctly and incorrectly called alleles; also in this case non-possible allele calls were excluded.

Data availability

Pipeline scripts together with instructions and ancillary files are freely available online (https://targt-pipeline.sourceforge.io/). Sequence data were obtained from Krause-Kyora *et al.*[24] and are accessible in the European Nucleotide Archive under accession no. ERP021830 (https://www.ebi.ac.uk/ena/data/view/PRJEB19769).

Received: 12 December 2019; Accepted: 14 April 2020;
Published online: 30 April 2020

References

1. Klein, J. Natural history of the major histocompatibility complex. (John Wiley &Sons, 1986).
2. Trowsdale, J. The MHC, disease and selection. *Immunol. Lett.* **137**, 1–8, https://doi.org/10.1016/j.imlet.2011.01.002 (2011).
3. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76, https://doi.org/10.1186/s13059-017-1207-1 (2017).
4. Tishkoff, S. A. & Verrelli, B. C. Patterns of Human Genetic Diversity: Implications for Human Evolutionary History and Disease. *Annu. Rev. Genomics Hum. Genet.* **4**, 293–340, https://doi.org/10.1146/annurev.genom.4.070802.110226 (2003).
5. Meyer, D. VR, C. A., Bitarello, B. D., DY, C. B. & Nunes, K. A genomic perspective on HLA evolution. *Immunogenetics* **70**, 5–27, https://doi.org/10.1007/s00251-017-1017-3 (2018).
6. Parham, P. Function and polymorphism of human leukocyte antigen-A,B,C molecules. *The American Journal of Medicine* **85**, 2–5, https://doi.org/10.1016/0002-9343(88)90369-5 (1988).
7. Reche, P. A. & Reinherz, E. L. Sequence variability analysis of human class I and class II MHC molecules: Functional and structural correlates of amino acid polymorphisms. *J. Mol. Biol.* **331**, 623–641, https://doi.org/10.1016/S0022-2836(03)00750-2 (2003).
8. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170, https://doi.org/10.1038/335167a0 (1988).
9. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431, https://doi.org/10.1093/nar/gku1161 (2015).
10. Hughes, A. L. & Nei, M. Nucleotide substitution at major histocompatibility complex class-II loci - evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* **86**, 958–962 (1989).
11. Spurgin, L. G. & Richardson, D. S. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. B. Biol. Sci.* **277**, 979–988, https://doi.org/10.1098/rspb.2009.2084 (2010).
12. Lenz, T. L., Spirin, V., Jordan, D. M. & Sunyaev, S. R. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. *Mol. Biol. Evol.* **33**, 2555–2564, https://doi.org/10.1093/molbev/msw127 (2016).
13. Dean, M., Carrington, M. & O'Brien, S. J. Balanced polymorphism selected by genetic versus infectious human disease. *Annu. Rev. Genomics Hum. Genet.* **3**, 263–292, https://doi.org/10.1146/annurev.genom.3.022502.103149 (2002).
14. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302, https://doi.org/10.1038/nature21347 (2017).
15. Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **18**, 659, https://doi.org/10.1038/nrg.2017.65 (2017).
16. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).
17. Robinson, J. *et al.* Distinguishing functional polymorphism from random variation in the sequences of>10,000 HLA-A, -B and -C alleles. *PLoS Genet.* **13**, e1006862, https://doi.org/10.1371/journal.pgen.1006862 (2017).
18. Orlando, L., Gilbert, M. T. & Willerslev, E. Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.* **16**, 395–408, https://doi.org/10.1038/nrg3935 (2015).
19. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLOS ONE* **5**, e14004, https://doi.org/10.1371/journal.pone.0014004 (2010).
20. Burbano, H. A. *et al.* Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328**, 723–725, https://doi.org/10.1126/science.1188046 (2010).
21. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. USA* **110**, 2223–2227, https://doi.org/10.1073/pnas.1221359110 (2013).
22. Carpenter, M. L. *et al.* Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am. *J. Hum. Genet.* **93**, 852–864, https://doi.org/10.1016/j.ajhg.2013.10.002 (2013).
23. Wittig, M. *et al.* Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.* **43**, e70, https://doi.org/10.1093/nar/gkv184 (2015).
24. Krause-Kyora, B. *et al.* Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nat. Commun.* **9**, 1569, https://doi.org/10.1038/s41467-018-03857-x (2018).
25. Gourraud, P.-A. *et al.* HLA Diversity in the 1000 Genomes Dataset. *Plos One* **9**, e97282, https://doi.org/10.1371/journal.pone.0097282 (2014).
26. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Pääbo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799 (2001).
27. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. USA* **104**, 14616–14621, https://doi.org/10.1073/pnas.0704665104 (2007).
28. Brotherton, P. *et al.* Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.* **35**, 5717–5728, https://doi.org/10.1093/nar/gkm588 (2007).
29. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87, https://doi.org/10.1093/nar/gkp1163 (2010).
30. Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684, https://doi.org/10.1093/bioinformatics/btt193 (2013).
31. de Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172, https://doi.org/10.1038/ng1885 (2006).
32. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316, https://doi.org/10.1093/bioinformatics/btu548 (2014).
33. Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P. & Thomson, G. PyPop update–a software pipeline for large-scale multilocus population genomics. *Tissue Antigens* **69**(Suppl 1), 192–197, https://doi.org/10.1111/j.1399-0039.2006.00769.x (2007).
34. Carrington, M. *et al.* Major histocompatibility complex class II haplotypes and linkage disequilibrium values observed in the CEPH families. *Hum. Immunol.* **41**, 234–240, https://doi.org/10.1016/0198-8859(94)90041-8 (1994).
35. Sanchez-Mazas, A. *et al.* A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. Eur. *J. Hum. Genet.* **8**, 33, https://doi.org/10.1038/sj.ejhg.5200391 (2000).
36. Armuzzi, A. *et al.* Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* **12**, 647–656, https://doi.org/10.1093/hmg/ddg066 (2003).
37. Renaud, G., Hanghøj, K., Willerslev, E. & Orlando, L. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* **33**, 577–579, https://doi.org/10.1093/bioinformatics/btw670 (2017).
38. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211, https://doi.org/10.1038/nature14317 (2015).
39. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424, https://doi.org/10.1038/nature19310 (2016).
40. Cruz-Davalos, D. I. *et al.* In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genomics* **19**, 608, https://doi.org/10.1186/s12864-018-4945-x (2018).
41. Briggs, A. W. *et al.* Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* **325**, 318–321, https://doi.org/10.1126/science.1174462 (2009).

42. Krause, J. *et al*. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894, https://doi.org/10.1038/nature08976 (2010).
43. Enk, J. M. *et al*. Ancient whole genome enrichment using baits built from modern DNA. *Mol Biol Evol* **31**, 1292–1294, https://doi.org/10.1093/molbev/msu074 (2014).
44. Lindo, J. *et al*. Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proc. Natl. Acad. Sci. USA* **114**, 4093–4098, https://doi.org/10.1073/pnas.1620410114 (2017).
45. Bauer, D. C., Zadoorian, A., Wilson, L. O. W. & Thorne, N. P. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief. Bioinform.* **19**, 179–187, https://doi.org/10.1093/bib/bbw097 (2018).
46. Lindo, J. *et al*. A time transect of exomes from a Native American population before and after European contact. *Nat. Commun.* **7**, 13175, https://doi.org/10.1038/ncomms13175 (2016).
47. Hollenbach, J. A. *et al*. A community standard for immunogenomic data reporting and analysis: proposal for a STrengthening the REporting of Immunogenomic Studies statement. *Tissue Antigens* **78**, 333–344, https://doi.org/10.1111/j.1399-0039.2011.01777.x (2011).
48. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic. Acids. Res.* **32**, 1792–1797 (2004).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
50. Hall, T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic. Acids Symp. Ser.* **41**, 95–98 (1999).
51. Mack, S. J. *et al*. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* **81**, 194–203, https://doi.org/10.1111/tan.12093 (2013).
52. Yang, D. Y. & Watt, K. Contamination controls when preparing archaeological remains for ancient DNA analysis. *J. Archaeol. Sci.* **32**, 331–336, https://doi.org/10.1016/j.jas.2004.09.008 (2005).
53. Pilli, E. *et al*. Monitoring DNA Contamination in Handled vs. Directly Excavated Ancient Human Skeletal Remains. *PLOS ONE* **8**, e52524, https://doi.org/10.1371/journal.pone.0052524 (2013).
54. Knapp, M., Clarke, A. C., Horsburgh, K. A. & Matisoo-Smith, E. A. Setting the stage - building and working in an ancient DNA laboratory. *Ann. Anat.* **194**, 3–6, https://doi.org/10.1016/j.aanat.2011.03.008 (2012).
55. Peltzer, A. *et al*. EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60, https://doi.org/10.1186/s13059-016-0918-z (2016).
56. Hedrick, P. W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341 (1987).
57. Cramer, H. Mathematical Models of Statistics. (New Jersey: Princeton University Press, 1946).
58. Slatkin, M. & Excoffier, L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity (Edinb)* **76**(Pt 4), 377–383 (1996).
59. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927, https://doi.org/10.1093/oxfordjournals.molbev.a040269 (1995).
60. Dempster, A., Laird, N. & B. Rubin, D. Maximum Likelihood From Incomplete Data Via The EM algorithm. Vol. 39 (1977).
61. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323, https://doi.org/10.1146/annurev-genom-091212-153455 (2013).
62. Auton, A. *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74, https://doi.org/10.1038/nature15393 (2015).

Acknowledgements

This work was supported by the Max Planck Society, and the Deutsche Forschungsgemeinschaft (DFG; grant LE 2593/3-1 to T.L.L., grant 2901391021 - SFB 1266/F4 to A.N. and B.K.-K., and the DFG Cluster of Excellence "Precision Medicine in Chronic Inflammation" (PMI, EXC2167)). J.B. was funded by the International Max Planck Research School for Evolutionary Biology. We are grateful to Jesper L. Boldsen and Dorthe Dangvard Pedersen for access to the St. Jørgen specimens.

Author contributions

T.L.L., B.K.-K., A.N., M.N. and F.P. conceived the study. L.B. and F.P. performed the lab work. M.N., T.L.L. and F.P. developed the bioinformatic pipeline. F.P. analyzed the data with input from M.N., T.L.L., B.K.-K., L.B., O.O., J.B. and J.S. A.F. provided research infrastructure. F.P. and T.L.L. interpreted the results and wrote the manuscript with input from A.N. and B.K.-K.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-64312-w.

Correspondence and requests for materials should be addressed to T.L.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Annex II

Genomewide Association Study of Severe Covid-19 with Respiratory Failure

The Severe Covid-19 GWAS Group*

* The authors' full names, academic degrees,

and affiliations are listed in the Appendix.

Published article
N Engl J Med 2020;383:1522-34
DOI: 10.1056/NEJMoa2020283

ORIGINAL ARTICLE

Genomewide Association Study of Severe Covid-19 with Respiratory Failure

The Severe Covid-19 GWAS Group*

ABSTRACT

BACKGROUND

There is considerable variation in disease behavior among patients infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes coronavirus disease 2019 (Covid-19). Genomewide association analysis may allow for the identification of potential genetic factors involved in the development of Covid-19.

METHODS

We conducted a genomewide association study involving 1980 patients with Covid-19 and severe disease (defined as respiratory failure) at seven hospitals in the Italian and Spanish epicenters of the SARS-CoV-2 pandemic in Europe. After quality control and the exclusion of population outliers, 835 patients and 1255 control participants from Italy and 775 patients and 950 control participants from Spain were included in the final analysis. In total, we analyzed 8,582,968 single-nucleotide polymorphisms and conducted a meta-analysis of the two case–control panels.

RESULTS

We detected cross-replicating associations with rs11385942 at locus 3p21.31 and with rs657152 at locus 9q34.2, which were significant at the genomewide level ($P<5\times10^{-8}$) in the meta-analysis of the two case–control panels (odds ratio, 1.77; 95% confidence interval [CI], 1.48 to 2.11; $P=1.15\times10^{-10}$; and odds ratio, 1.32; 95% CI, 1.20 to 1.47; $P=4.95\times10^{-8}$, respectively). At locus 3p21.31, the association signal spanned the genes *SLC6A20*, *LZTFL1*, *CCR9*, *FYCO1*, *CXCR6* and *XCR1*. The association signal at locus 9q34.2 coincided with the *ABO* blood group locus; in this cohort, a blood-group–specific analysis showed a higher risk in blood group A than in other blood groups (odds ratio, 1.45; 95% CI, 1.20 to 1.75; $P=1.48\times10^{-4}$) and a protective effect in blood group O as compared with other blood groups (odds ratio, 0.65; 95% CI, 0.53 to 0.79; $P=1.06\times10^{-5}$).

CONCLUSIONS

We identified a 3p21.31 gene cluster as a genetic susceptibility locus in patients with Covid-19 with respiratory failure and confirmed a potential involvement of the ABO blood-group system. (Funded by Stein Erik Hagen and others.)

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Franke at the Institute of Clinical Molecular Biology and University Hospital of Schleswig-Holstein, Christian-Albrechts-University, Rosalind-Franklin-Str. 12, D-24105 Kiel, Germany, or at a.franke@mucosa.de; or to Dr. Karlsen at the Division of Surgery, Inflammatory Diseases, and Transplantation, Oslo University Hospital Rikshospitalet and University of Oslo, Postboks 4950 Nydalen, N-0424 Oslo, Norway, or at t.h.karlsen@medisin.uio.no.

*Dr. Franke serves as an author on behalf of the Covid-19 Host Genetics Initiative; members of the Initiative are listed in Supplementary Appendix 1, available at NEJM.org.

Dr. Ellinghaus and Ms. Degenhardt and Drs. Valenti, Franke, and Karlsen contributed equally to this article.

This article was published on June 17, 2020, at NEJM.org.

N Engl J Med 2020;383:1522-34.
DOI: 10.1056/NEJMoa2020283

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was discovered in Wuhan, China, in late 2019, and coronavirus disease 2019 (Covid-19), the disease caused by SARS-CoV-2, rapidly evolved into a global pandemic.[1] As of June 15, 2020, there were more than 8.03 million confirmed cases worldwide, with total deaths exceeding 436,900.[2] In Europe, Italy and Spain were severely affected early on, with epidemic peaks starting in the second half of February 2020 (Fig. 1) and 61,507 deaths reported by June 15, 2020. Covid-19 has varied manifestations,[3] with the large majority of infected persons having only mild symptoms or even no symptoms.[4] Mortality rates are driven predominantly by the subgroup of patients who have severe respiratory failure related to interstitial pneumonia in both lungs and acute respiratory distress syndrome.[5] Severe Covid-19 with respiratory failure requires early and prolonged support by mechanical ventilation.[6]

The pathogenesis of severe Covid-19 and the associated respiratory failure is poorly understood, but higher mortality is consistently associated with older age and male sex.[7,8] Clinical associations have also been reported for hypertension, diabetes, and other obesity-related and cardiovascular disease traits, but the relative role of clinical risk factors in determining the severity of Covid-19 has not yet been clarified.[7-11] Observational data on lymphocytic endotheliitis and diffuse microvascular and macrovascular thromboembolic complications suggest that Covid-19 is a systemic disease that involves injury to the vascular endothelium but provide little insight into the underlying pathogenesis.[12-14] At the peak of the epidemic in Italy and Spain in early 2020, we performed a genomewide association study (GWAS) in an attempt to delineate host genetic factors contributing to severe Covid-19 with respiratory failure. The relatively low disease burden of Covid-19 in Norway and Germany allowed for a complementary team to be set up, whereby genotyping and analysis could occur in parallel with the rapid recruitment of patients in the heavily affected Italian and Spanish epicenters.

METHODS

STUDY PARTICIPANTS AND RECRUITMENT

We recruited 1980 patients with severe Covid-19, which was defined as hospitalization with respiratory failure and a confirmed SARS-CoV-2 viral RNA polymerase-chain-reaction (PCR) test from nasopharyngeal swabs or other relevant biologic fluids, cross sectionally, from intensive care units and general wards at seven hospitals in four cities in the pandemic epicenters in Italy and Spain (Table S1A in Supplementary Appendix 1, available with the full text of this article at NEJM.org). The hospitals in Italy were the following: Fondazione IRCCS Cá Granda Ospedale Maggiore Policlinico in Milan (597 patients); Humanitas Clinical and Research Center, IRCCS, in Milan (154 patients); and UNIMIB (Università degli Studi di Milano–Bicocca) School of Medicine, San Gerardo Hospital, in Monza (a suburb of Milan) (200 patients). The hospitals in Spain were the following: Hospital Clínic and IDIBAPS (Instituto de Investigaciones Biomédicas August Pi i Sunyer) in Barcelona (56 patients), Hospital Universitario Vall d'Hebron in Barcelona (337 patients), Hospital Universitario Ramón y Cajal in Madrid (298 patients), and Donostia University Hospital in San Sebastian (338 patients).

Respiratory failure was defined in the simplest possible manner in order to ensure feasibility: the use of oxygen supplementation or mechanical ventilation, with severity graded according to the maximum respiratory support received at any point during hospitalization (supplemental oxygen therapy only, noninvasive ventilatory support, invasive ventilatory support, or extracorporeal membrane oxygenation). For severity assessments, severity was also dichotomized as no mechanical ventilation or mechanical ventilation. Whole-blood samples or buffy coats from diagnostic venipuncture were obtained for DNA extraction.

For comparison, we included 2381 control participants from Italy and Spain (Table S1B in Supplementary Appendix 1). We recruited 998 randomly selected blood donors at Fondazione IRCCS Cá Granda Ospedale Maggiore Policlinico, Milan, who underwent genotyping for the purpose of the present study. A total of 40 of these participants had evidence of the development of anti–SARS-CoV-2 antibodies, all of whom had mild or no Covid-19 symptoms. We also included two control panels with genotype data derived from previous studies and from persons with unknown SARS-CoV-2 infection status using the same genotyping array. The panels included 396 healthy volunteers, blood donors, and outpatients of gastroenterology departments in Italy[15] and 987 healthy blood donors in San Sebastian, Spain.

Figure 1. Timeline of Rapid Covid-19 Genomewide Association Study (GWAS).

The main events and milestones of the study are summarized in the plot. Samples from patients in three Italian hospitals (hospital A: Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan; hospital B: Humanitas Clinical and Research Center, IRCCS, Milan; and hospital C: UNIMIB School of Medicine, San Gerardo Hospital, Monza) and four Spanish hospitals (hospital A: Hospital Clínic and IDIBAPS, Barcelona; hospital B: Hospital Universitario Vall d'Hebron, Barcelona; hospital C: Hospital Universitario Ramón y Cajal, Madrid; and hospital D: Donostia University Hospital, San Sebastian) were obtained around the peak of the local epidemics, and ethics applications were quickly obtained by means of fast-track procedures (i.e., every local ethics review board supported studies of coronavirus disease 2019 [Covid-19] studies by providing rapid turn-around times, thus facilitating this fast de novo data generation). All the obtained blood samples were centrally isolated, genotyped, and analyzed within 8 weeks. Control data were obtained from control participants and from historical control data in Italy and Spain. The rapid workflow from patients to target identification shows the usefulness of GWAS, a standardized research tool that often relies on international and interdisciplinary cooperation. One center alone could not have completed this study, not to mention the increase in statistical power that was available because of the contribution of patients from multiple centers. The speed of data production depended heavily on laboratory automation, and the speed of analyses reflects existing analytic pipelines and the support of public so-called imputation servers (here, the Michigan imputation server of the G. Abecasis group). QC denotes quality control.

ETHICS COMMITTEE APPROVAL

The project protocol involved the rapid recruitment of patient-participants and no additional project-related procedures (we primarily used material from clinically indicated venipunctures) and afforded anonymity, owing to the minimal data set collected. Differences in recruitment and consent procedures among the centers arose because some centers integrated the project into larger Covid-19 biobanking efforts, whereas other centers did not, and because there were differences in how local ethics committees provided guidance on the handling of anonymization or deidentification of data as well as consent procedures. Written informed consent was obtained, sometimes in a delayed fashion, from the study patients at each center when possible. In some instances, informed consent was provided verbally or by the next of kin, depending on local ethics committee regulations and special policies issued for Covid-19 research. For some severely ill patients, an exemption from informed consent was obtained from a local ethics committee or according to local regulations in order to allow the use of completely anonymized surplus material from diagnostic venipuncture.

The following approvals of the project were obtained from the relevant ethics committees: Germany: Kiel (reference number, D464/20); Italy: Fondazione IRCCS Cá Granda Ospedale Maggiore Policlinico (reference numbers, 342_2020 for patients and 334-2020 for control participants), Humanitas Clinical and Research Center, IRCCS (reference number, 316/20), the University of Milano–Bicocca School of Medicine, San Gerardo Hospital, Monza (the ethics committee of the National Institute of Infectious Diseases Lazzarro Spallanzani reference number, 84/2020); Norway: Regional Committee for Medical and Health Research Ethics in South-Eastern Norway (reference number, 132550); Spain: Hospital Clínic, Barcelona (reference number, HCB/2020/0405), Hospital Universitario Vall d'Hebron, Barcelona (reference number, PR[AG]244/2020), Hospital Universitario Ramón y Cajal, Madrid (reference number, 093/20) and Donostia University Hospital, San Sebastian (reference number, PI2020064).

SAMPLE PROCESSING, GENOTYPING, AND IMPUTATION

We performed DNA extraction using a Chemagic 360 (PerkinElmer) with the use of the low-volume kit CMG-1491 and the buffy-coat kit CMG-714 (Chemagen), respectively. For genotyping, we used the Global Screening Array (GSA), version 2.0 (Illumina), which contains 712,189 variants before quality control. Details on genotyping and quality-control procedures are provided in the Supplementary Methods section in Supplementary Appendix 1. To maximize genetic coverage, we performed single-nucleotide polymorphism (SNP) imputation on genome build GRCh38 using the Michigan Imputation Server and 194,512 haplotypes generated by the Trans-Omics for Precision Medicine (TOPMed) program (freeze 5).[16]

After the exclusion of samples during quality control (the majority of which were due to population outliers; see the Supplementary Methods section and Table S1B and S1C), the final case–control data sets comprised 835 patients and 1255 control participants from Italy and 775 patients and 950 control participants from Spain. A total of 8,965,091 SNPs were included in the Italian cohort and 9,140,716 SNPs in the Spanish cohort.

STATISTICAL ANALYSIS

To take imputation uncertainty into account, we tested for phenotypic associations with allele dosage data separately for both the Italian and Spanish case–control panels with the use of the PLINK logistic-regression framework for dosage data (PLINK, version 1.9).[17] We carried out two genomewide tests of association that included covariates from principal-component analyses, with adjustments to control for potential population stratification (main analysis) and potential population stratification and age and sex bias (analysis corrected for age and sex). A fixed-effects meta-analysis was conducted with the use of the meta-analysis tool METAL[18] on 8,582,968 variants that were common to both the Italian and Spanish data sets with the use of effect-size estimates and their standard errors from the study-specific association analyses.

For the genomewide meta-analysis, we used the commonly accepted threshold of 5×10^{-8} for joint P values to determine statistical significance. Bayesian fine-mapping analysis was performed for loci reaching genomewide significance (see the Supplementary Methods section). Genomewide summary statistics of our analyses are publicly available through our web browser (www.c19-genetics.eu) and have been submitted to the European Bioinformatics Institute (www.ebi.ac

Table 1. Overview of Patients Included in the Final Analysis.*

Characteristic	Italian Hospitals			Spanish Hospitals			
	A (N = 503)	B (N = 140)	C (N = 192)	A (N = 45)	B (N = 228)	C (N = 201)	D (N = 301)
Median age (IQR) — yr	64 (54–76)	67 (57–75)	66 (56–74)	69 (59–75)	65 (56–72)	69 (60–79)	67 (57–75)
Female sex — no. (%)	159 (32)	39 (28)	51 (27)	13 (29)	78 (34)	50 (25)	124 (41)
Respiratory support — no. (%)							
Supplemental oxygen only	0	70 (50)	67 (35)	7 (16)	105 (46)	106 (53)	255 (85)
Noninvasive ventilation	399 (79)	25 (18)	89 (46)	6 (13)	7 (3)	16 (8)	0
Ventilator	104 (21)	45 (32)	33 (17)	31 (69)	116 (51)	77 (38)	46 (15)
ECMO	0	0	3 (2)	1 (2)	0	2 (1)	0
Hypertension — no./total no. (%)	166/503 (33)	71/140 (51)	109/192 (57)	26/45 (58)	113/228 (50)	112/199 (56)	114/301 (38)
Coronary artery disease — no./total no. (%)	21/503 (4)	25/140 (18)	25/192 (13)	4/45 (9)	14/228 (6)	35/199 (18)	15/301 (5)
Diabetes — no./total no. (%)	63/503 (13)	18/140 (13)	34/192 (18)	10/45 (22)	50/228 (22)	57/199 (29)	65/301 (22)

* In Italy, hospital A was Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan; hospital B Humanitas Clinical and Research Center, IRCCS, Milan; and hospital C UNIMIB School of Medicine, San Gerardo Hospital, Monza. In Spain, hospital A was Hospital Clínic and IDIBAPS, Barcelona; hospital B Hospital Universitario Vall d'Hebron, Barcelona; hospital C Hospital Universitario Ramón y Cajal, Madrid; and hospital D Donostia University Hospital, San Sebastian. The predominance of men among the patients and the advanced age (median, >63 years) were consistent across all the centers. The sample numbers provided are after quality control was conducted for the genomewide association study (Table S1C in Supplementary Appendix 1). Allele distributions for detected risk variants at loci 3p21.31 and 9q34.2 in clinical subsets are shown in Table S2. ECMO denotes extracorporeal membrane oxygenation, and IQR interquartile range.

.uk/gwas; accession numbers, GCST90000255 and GCST90000256).

On the basis of the results from the TOPMed genotype imputation, we selected three *ABO* SNPs (rs8176747, rs41302905, and rs8176719)[19,20] to infer the ABO blood type and calculated odds ratios according to blood type (A vs. B, AB, or O; B vs. A, AB, or O; AB vs. A, B, or O; and O vs. A, AB, or B) (see the Supplementary Methods). To assess in detail the HLA complex at locus 6p21, we performed sequencing-based HLA typing of seven classical HLA loci (*HLA-A*, *-C*, *-B*, *-DRB1*, *-DQA1*, *-DQB1*, and *-DPB1*) in a subgroup of 835 patients and 891 control participants from Italy and 773 patients from Spain (see the Supplementary Methods). We also assessed allelic distribution according to no mechanical ventilation (supplemental oxygen only) as compared with mechanical ventilation of any type. A similar assessment was made for lead SNPs rs11385942 and rs657152 at loci 3p21.31 and 9q34.2, respectively.

RESULTS

PATIENTS, GENOTYPING, AND QUALITY CONTROL

The milestones of the study in the context of the peak outbreaks in Italy and Spain are shown in Figure 1. Data on the age, sex, maximum respiratory support at any point during hospitalization, and relevant coexisting conditions (type 2 diabetes, hypertension, and coronary heart disease) in the patients who were included in the final analysis are shown in Table 1 and in Table S2 in Supplementary Appendix 1. Because we used the same genotyping platform (GSA) to obtain both data sets, we were able to perform a uniform quality control of the merged Italian and Spanish SNP data sets, thus reducing technical confounders to a minimum. A quantile–quantile (Q-Q) plot of the two meta-analyses (the main analysis and the analysis corrected for age and sex) showed significant associations in the tail of the distribution with minimal genomic inflation (λ_{GC}=1.015 for main analysis and λ_{GC}=1.006 for analysis corrected for age and sex) (Fig. S2 in Supplementary Appendix 1). We also carried out separate association analyses for the Italian and Spanish data sets (see the Supplementary Methods section and Fig. S3).

GENOMEWIDE ASSOCIATION ANALYSIS

We found two loci to be associated with Covid-19–induced respiratory failure with genomewide sig-

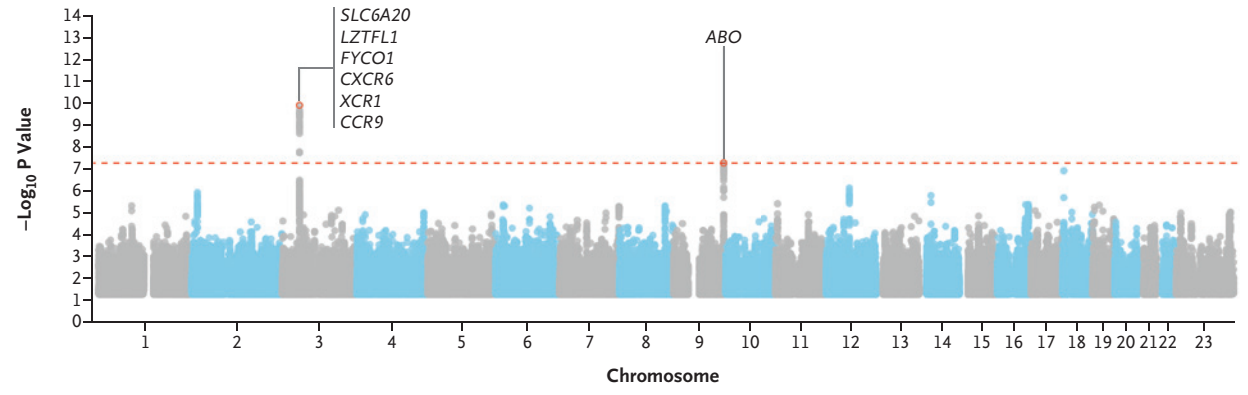


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Shown is a Manhattan plot of the association statistics from the main meta-analysis (controlled for potential population stratification). The red dashed line indicates the genomewide significance threshold of a P value less than 5×10^{-8}. Figure S6 in Supplementary Appendix 1 shows Manhattan plots that include hits passing a suggestive significance threshold of a P value less than 1×10^{-5} (total of 24 additional suggestive genomic loci) (see the Supplementary Methods section and Supplementary Appendix 4).

nificance ($P<5\times10^{-8}$) in the main meta-analysis: the rs11385942 insertion–deletion GA or G variant at locus 3p21.31 (odds ratio for the GA allele, 1.77; 95% confidence interval [CI], 1.48 to 2.11; $P=1.15\times10^{-10}$) and the rs657152 A or C SNP at locus 9q34.2 (odds ratio for the A allele, 1.32; 95% CI, 1.20 to 1.47; $P=4.95\times10^{-8}$) (Fig. 2 and Table 2 and Supplementary Appendix 2, available at NEJM.org). Both loci showed nominally significant association in both the Spanish and Italian subanalyses (Table 2). The meta-analysis association results for recessive and heterozygous genetic models for the two meta-analyses (main analysis and the analysis corrected for age and sex) are provided in Supplementary Appendix 3, available at NEJM.org. The imputation quality of the associated markers was good (Table 2 and Supplementary Appendix 2), and manual inspection of genotype cluster plots of genotyped SNPs in these regions showed distinct genotype clouds for homozygous and heterozygous calls (Fig. S4 in Supplementary Appendix 1). Furthermore, the analyses that were corrected for age and sex corroborated the observations at both rs11385942 (meta-analysis odds ratio, 2.11; 95% CI, 1.70 to 2.61; $P=9.46\times10^{-12}$) and rs657152 (meta-analysis odds ratio, 1.39; 95% CI, 1.22 to 1.59; $P=5.35\times10^{-7}$) (Table 2 and Fig. S5 in Supplementary Appendix 1).

The allele frequencies in Spanish and Italian control data sets from previously published studies[21-27] are consistent with those we report here (Supplementary Appendix 2). A further 24 different genomic loci showed suggestive evidence ($P<1\times10^{-5}$) for association with Covid-19–induced respiratory failure in the main analysis (Supplementary Appendix 4, available at NEJM.org, and Fig. S6 in Supplementary Appendix 1). Association signals at loci 3p21.31 and 9q34.2 were fine-mapped to 22 and 38 variants, respectively, with greater than 95% certainty (Fig. 3A and 3B and Supplementary Appendix 5, available at NEJM.org).

CHROMOSOME 3P21.31

The association signal at locus 3p21.31 comprised six genes (*SLC6A20*, *LZTFL1*, *CCR9*, *FYCO1*, *CXCR6*, and *XCR1*) (Fig. 3A). The risk allele GA of rs11385942 is associated with reduced expression of *CXCR6* and increased expression of *SLC6A20*, and *LZTFL1* is strongly expressed in human lung cells (Fig. S7 and Supplementary Appendix 6, available at NEJM.org). We found that the frequency of the risk allele of the lead variant at 3p21.31 (rs11385942) was higher among patients who received mechanical ventilation than among those who received oxygen supplementation only in both the main meta-analysis (odds ratio, 1.70; 95% CI, 1.27 to 2.26; $P=3.30\times10^{-4}$) and the meta-analysis corrected for age and sex (odds ratio, 1.56; 95% CI, 1.17 to 2.01; P=0.003) (Supplementary Appendix 7, available at NEJM.org). Furthermore, the 19 patients who were homozygous for the rs11385942 risk allele were younger than

Table 2. Susceptibility Loci Associated with Severe Covid-19 with Respiratory Failure.*

Chromosome and Analysis	Meta-analysis		Italian Panel			Spanish Panel			
	P Value	Odds Ratio (95% CI)	P Value	Odds Ratio (95% CI)	Allele Frequency		P Value	Odds Ratio (95% CI)	Allele Frequency	
					patient	control			patient	control
3p21.31†										
Main analysis	1.15×10^{-10}	1.77 (1.48–2.11)	1.98×10^{-7}	1.74 (1.27–2.38)	0.14	0.09	1.32×10^{-4}	1.85 (1.50–2.28)	0.09	0.05
Analysis corrected for age and sex	9.46×10^{-12}	2.11 (1.70–2.61)	7.02×10^{-8}	1.95 (1.53–2.48)	0.14	0.09	1.17×10^{-5}	2.79 (1.76–4.42)	0.09	0.05
9q34.2‡										
Main analysis	4.95×10^{-8}	1.32 (1.20–1.47)	2.90×10^{-6}	1.37 (1.20–1.57)	0.42	0.35	3.55×10^{-3}	1.26 (1.08–1.48)	0.42	0.35
Analysis corrected for age and sex	5.35×10^{-7}	1.39 (1.22–1.59)	5.31×10^{-5}	1.37 (1.17–1.60)	0.42	0.35	2.81×10^{-3}	1.45 (1.13–1.84)	0.42	0.35

* The meta-analysis included 1610 patients and 2205 control participants; the Italian analysis, 835 and 1255, respectively; and the Spanish analysis, 775 and 950, respectively. Allele frequencies of the minor or risk allele (see below) are shown among the patients and the control participants. All the association test statistics were adjusted for the top 10 principal components from the principal-component analysis. Two analyses were performed: a main analysis, which was corrected for 10 principal components, and an analysis that was corrected for age and sex in addition to 10 principal components. In the analyses that were corrected for age and sex, 25 control participants were excluded from the Spanish analysis and the meta-analysis because of missing covariate data. The P values and corresponding odds ratios and 95% confidence intervals (CIs) are shown with respect to the minor allele. Association results for the recessive and heterozygous models for both meta-analyses (main and corrected for age and sex) are shown in Supplementary Appendix 3. Covid-19 denotes coronavirus disease 2019.

† For chromosome 3p21.31, the association boundaries for each index single-nucleotide polymorphism (SNP; see the Supplementary Methods section), with the genomic positions retrieved from genome build hg38, were chr3:45800446 through 46135604. The Single Nucleotide Polymorphism database (dbSNP) identifier was rs11385942 (the rs identifier from the National Center for Biotechnology Information, rs11385942, is annotated as chr3:45834968 through 45834969:AAA:AA in dbSNP, version 153, and as chr3:45834967:GA:G in the TransOmics for Precision Medicine [TOPMed] imputation reference panel). The SNP rs11385942 was imputed according to TOPMed with high confidence (TOPMed estimated imputation accuracy, $R^2=0.94$ and $R^2=0.95$ for the Italian and Spanish panels, respectively) (Supplementary Appendix 2).The minor or risk allele was GA, and the major allele was G. The key genes (i.e., the candidate genes in the region) were *SLC6A20, LZTFL1, FYCO1, CXCR6, XCR1,* and *CCR9.*

‡ For chromosome 9q34.2, the association boundaries for each index SNP, with the genomic positions retrieved from genome build hg38, were chr9:133257521 through 133279871. The SNP rs657152 was genotyped according to the Global Screening Array (GSA) in the Italian and Spanish panels (Supplementary Appendix 2). The minor or risk allele was A, and the major allele was C. The key gene was *ABO.*

1591 patients who were heterozygous or homozygous for the nonrisk allele (median age, 59 years [interquartile range, 49 to 68] vs. 66 years [interquartile range, 56 to 75]; P=0.005). Available variant database entries suggest that the frequency of this risk allele varies among populations worldwide (Fig. S8 in Supplementary Appendix 1).

ABO LOCUS

At locus 9q34.2 the association signal coincided with the *ABO* blood group locus (Fig. 3B and Fig. S9 in Supplementary Appendix 1). Accordingly, the distribution of ABO blood groups (predicted from combinations of genotypes of three different SNPs) was skewed among patients with Covid-19 who had respiratory failure, as compared with the distribution among control participants. In the meta-analysis corrected for age and sex, we found a higher risk among persons with blood group A than among patients with other blood groups (odds ratio, 1.45; 95% CI, 1.20 to 1.75; $P=1.48\times10^{-4}$) and a protective effect for blood group O as compared with the other blood groups (odds ratio, 0.65; 95% CI, 0.53 to 0.79; $P=1.06\times10^{-5}$). Details are provided in Supplementary Appendix 8, available at NEJM.org. Both associations and effect directions were consistent in the separate Spanish and Italian case–control analyses. We found no significant difference in blood-group distribution between patients receiving supplemental oxygen only and those receiving mechanical ventilation of any kind. The ABO blood-group frequency distributions in public registries are provided for comparison in Supplementary Appendix 8, along with details of the results presented here, and corroborate our observations.

HLA ANALYSIS

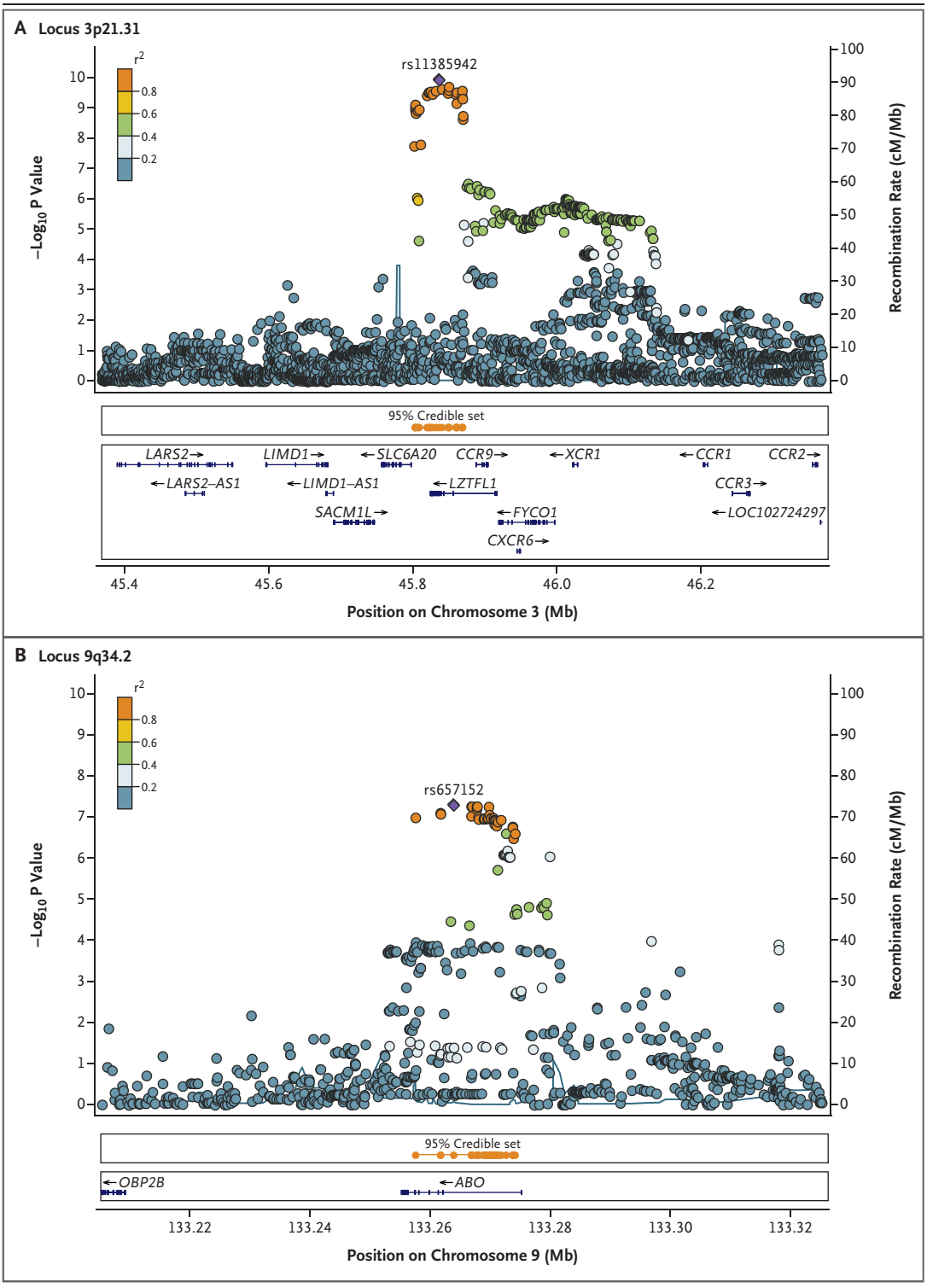
Given its important role in several viral infections, we scrutinized the extended HLA region (chromosome 6, 25 through 34 Mb). There were no SNP association signals at the HLA complex that met even the significance threshold of suggestive association: $P<1\times10^{-5}$ (Fig. S10 in Supplementary Appendix 1). Dedicated analysis of the classical HLA loci showed no significant allele associations with either Covid-19 or disease severity (oxygen supplementation only or mechanical ventilation of any kind), and further analysis of heterozygote and divergent allele advantage or predicted number of HLA-bound SARS-CoV-2 peptides did not show significant associations with Covid-19 in this data set (see the HLA Analyses section in Supplementary Appendix 1 and Supplementary Appendix 9, available at NEJM.org).

DISCUSSION

Using a pragmatic approach with simplified inclusion criteria and a complementary team of clinicians at the European Covid-19 epicenters in Italy and Spain and scientists in the less-burdened countries of Germany and Norway, we performed a GWAS that included de novo genotyping for Covid-19 with respiratory failure in approximately 2 months. We detected a novel susceptibility locus at a chromosome 3p21.31 gene cluster and confirmed a potential involvement of the ABO blood-group system in Covid-19.

On chromosome 3p21.31, the peak association signal covered a cluster of six genes (*SLC6A20*, *LZTFL1*, *CCR9*, *FYCO1*, *CXCR6*, and *XCR1*), several of which have functions that are potentially relevant to Covid-19. A causative gene cannot be reliably implicated by the present data. One candidate is *SLC6A20*, which encodes the sodium–imino acid (proline) transporter 1 (SIT1) and which functionally interacts with angiotensin-converting enzyme 2, the SARS-CoV-2 cell-surface receptor.[28,29] However, the locus also contains genes encoding chemokine receptors, including the CC motif chemokine receptor 9 (CCR9) and the C-X-C motif chemokine receptor 6 (CXCR6), the latter of which regulates the specific location of lung-resident memory CD8 T cells throughout the sustained immune response to airway pathogens, including influenza viruses.[30] Flanking genes (e.g., *CCR1* and *CCR2*) also have relevant functions,[31] and further studies will be needed to delineate the functional consequences of detected associations.

The preliminary results from the Covid-19 Host Genetics Consortium[32] include suggestive associations within the same locus at chromosome 3p21.31, which lend considerable support to our findings (Fig. S11 in Supplementary Appendix 1). The consortium analysis also used population-based controls, but the patients included persons with mild Covid-19 and those with severe Covid-19. The parallel findings nevertheless underscore an important point about the ascertainment of patients and controls in genetic studies of Covid-19. Because the majority of patients with SARS-CoV-2 infection are asymp-



tomatic, any sample involving patients with a positive nasopharyngeal RNA test is likely to hold a bias toward some degree of symptomatic burden. Two of the identifiers for inclusion in the current study were a positive result for the presence of SARS-CoV-2 according to PCR testing and receipt of respiratory support (an extreme Covid-19 phenotype). As such, it seems

Figure 3 (facing page). Regional Association Plots of Susceptibility Loci Associated with Severe Covid-19 with Respiratory Failure.

Bayesian fine-mapping analysis prioritized 22 and 38 variants for loci 3p21.31 (Panel A) and 9q34.2 (Panel B), respectively, with greater than 95% certainty. The linkage disequilibrium values were calculated on the basis of genotypes of the merged Italian and Spanish data sets derived from TOPMed (Trans-Omics for Precision Medicine) imputation. The positions in the genome assembly hg38 are plotted. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The plot shows the names and locations of the genes; the transcribed strand is indicated with an arrow. Genes are represented with intronic and exonic regions. The purple diamond in each panel represents the variant most strongly associated with severe Covid-19 and respiratory failure.

reasonable to conclude that the chromosome 3p21.31 locus is involved in Covid-19 susceptibility per se, with a possible enrichment in patients with severe disease. This latter interpretation is supported by the significantly higher frequency of the risk allele among patients who received mechanical ventilation than among those who received supplemental oxygen only as well as by the finding of younger age among patients who were homozygous for the risk allele than among patients who were heterozygous or homozygous for the nonrisk allele.

Nongenetic studies that were reported as preprints[33,34] have previously implicated the involvement of ABO blood groups in Covid-19 susceptibility, and ABO blood groups have also been implicated in susceptibility to SARS-CoV-1 infection.[35] Our genetic data confirm that blood group O is associated with a risk of acquiring Covid-19 that was lower than that in non-O blood groups, whereas blood group A was associated with a higher risk than non-A blood groups.[33,34] The biologic mechanisms undergirding these findings may have to do with the ABO group per se (e.g., with the development of neutralizing antibodies against protein-linked N-glycans)[36] or with other biologic effects of the identified variant,[37-39] including the stabilization of von Willebrand factor.[40,41] The *ABO* locus holds considerable risk for population stratification,[42] which is increased by the inclusion of randomly selected blood donors in the current study (for which there is an inherent risk of blood group O enrichment). Alignment of the allele frequencies at the *ABO* locus in our control population with those in several non–blood-donor control populations would suggest that this is not a major bias, and at least one study[34] that tested for association with blood type used disease controls with no affiliation to blood donors.

The pragmatic aspects leading to the feasibility of this massive undertaking in a very short period of time during the extreme clinical circumstances of the pandemic imposed limitations that will be important to explore in follow-up studies. For example, to enable the recruitment of study participants, a bare minimum of clinical metadata was requested. For this reason, extensive genotype–phenotype elaboration of current findings could not be conducted, and adjustments for all potential sources of bias (e.g., underlying cardiovascular and metabolic factors relevant to Covid-19) could not be performed. Furthermore, we have limited information about the SARS-CoV-2 infection status in the control participants; this concern is mitigated by the fact that the presence of susceptible persons in the control group would only bias the tests toward the null. In addition, few restrictions were imposed during inclusion, which led to genotyped samples having to be excluded owing to differing ethnic groups (population outliers). Further exploration of current findings, both as to their usefulness in clinical risk profiling of patients with Covid-19 and toward a mechanistic understanding of the underlying pathophysiology, is warranted.

Supported by a philanthropic donation from Stein Erik Hagen and Canica; by a grant from the Deutsche Forschungsgemeinschaft Cluster of Excellence "Precision Medicine in Chronic Inflammation" (EXC2167); by a Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico Covid-19 Biobank grant (to Dr. Valenti); by grants from the Italian Ministry of Health (RF-2016-02364358, to Dr. Valenti) and Ministero dell'Istruzione, dell'Università e della Ricerca project "Dipartimenti di Eccellenza 2018–2022" (D15D18000410001 to the Department of Medical Sciences, University of Turin; by a grant from the Spanish Ministry of Science and Innovation JdC fellowship (IJC2018-035131-I, to Dr. Acosta-Herrera); and by the GCAT Cession Research Project PI-2020-01. HLA typing was performed and supported by the Stefan-Morsch-Stiftung.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank all the patients who consented to participate in this study, and we express our condolences to the families of patients who died from Covid-19. We also thank the entire clinical staff during the outbreak situation at the different centers who were able to work on this scientific study in parallel with their clinical duties; all the members of the Humanitas Covid-19 Task Force for contributions to the recruitment of patients (see the Supplementary Notes section in Supplementary Appendix 1); Sören Brunak and Karina Banasik for discussions on the *ABO* associa-

tion; Goncalo Abecasis and his team for providing the Michigan imputation server; Fabrizio Bossa and Francesca Tavano for contributions to control-sample acquisition; Maria Reig for help in the case-sample acquisition; the staff of the Basque Biobank in Spain for assistance in the acquisition of samples; the staff of GCAT|Genomes for Life, a cohort study of the Genomes of Catalonia, Institute for Health Science Research Germans Trias i Pujol, for data contribution; Alexander Eck, Jenspeter Horst, and Jens Scholz for supporting the HLA typing in the project; and the members of the ethics commissions, review boards, and consortia who fast-track reviewed our applications and enabled this rapid genetic discovery study.

APPENDIX

The authors' full names and academic degrees are as follows: David Ellinghaus, Ph.D., Frauke Degenhardt, M.Sc., Luis Bujanda, M.D., Ph.D., Maria Buti, M.D., Ph.D., Agustín Albillos, M.D., Ph.D., Pietro Invernizzi, M.D., Ph.D., Javier Fernández, M.D., Ph.D., Daniele Prati, M.D., Guido Baselli, Ph.D., Rosanna Asselta, Ph.D., Marit M. Grimsrud, M.D., Chiara Milani, Ph.D., Fátima Aziz, B.S., Jan Kässens, Ph.D., Sandra May, Ph.D., Mareike Wendorff, M.Sc., Lars Wienbrandt, Ph.D., Florian Uellendahl-Werth, M.Sc., Tenghao Zheng, M.D., Ph.D., Xiaoli Yi, Raúl de Pablo, M.D., Ph.D., Adolfo G. Chercoles, B.S., Adriana Palom, M.S., B.S., Alba-Estela Garcia-Fernandez, B.S., Francisco Rodriguez-Frias, M.S., Ph.D., Alberto Zanella, M.D., Alessandra Bandera, M.D., Ph.D., Alessandro Protti, M.D., Alessio Aghemo, M.D., Ph.D., Ana Lleo, M.D., Ph.D., Andrea Biondi, M.D., Andrea Caballero-Garralda, M.S., Ph.D., Andrea Gori, M.D., Anja Tanck, Anna Carreras Nolla, B.S., Anna Latiano, Ph.D., Anna Ludovica Fracanzani, M.D., Anna Peschuck, Antonio Julià, Ph.D., Antonio Pesenti, M.D., Antonio Voza, M.D., David Jiménez, M.D., Ph.D., Beatriz Mateos, M.D., Ph.D., Beatriz Nafria Jimenez, B.S., Carmen Quereda, M.D., Ph.D., Cinzia Paccapelo, M.Sc., Christoph Gassner, Ph.D., Claudio Angelini, M.D., Cristina Cea, B.S., Aurora Solier, M.D., David Pestaña, M.D., Ph.D., Eduardo Muñiz-Diaz, M.D., Ph.D., Elena Sandoval, M.D., Elvezia M. Paraboschi, Ph.D., Enrique Navas, M.D., Ph.D., Félix García Sánchez, Ph.D., Ferruccio Ceriotti, M.D., Filippo Martinelli-Boneschi, M.D., Ph.D., Flora Peyvandi, M.D., Ph.D., Francesco Blasi, M.D., Ph.D., Luis Téllez, M.D., Ph.D., Albert Blanco-Grau, B.S., M.S., Georg Hemmrich-Stanisak, Ph.D., Giacomo Grasselli, M.D., Giorgio Costantino, M.D., Giulia Cardamone, Ph.D., Giuseppe Foti, M.D., Serena Aneli, Ph.D., Hayato Kurihara, M.D., Hesham ElAbd, M.Sc., Ilaria My, M.D., Iván Galván-Femenia, M.Sc., Javier Martín, M.D., Ph.D., Jeanette Erdmann, Ph.D., Jose Ferrusquía-Acosta, M.D., Koldo Garcia-Etxebarria, Ph.D., Laura Izquierdo-Sanchez, B.S., Laura R. Bettini, M.D., Lauro Sumoy, Ph.D., Leonardo Terranova, Ph.D., Leticia Moreira, M.D., Ph.D., Luigi Santoro, M.S., Luigia Scudeller, M.D., Francisco Mesonero, M.D., Luisa Roade, M.D., Malte C. Rühlemann, Ph.D., Marco Schaefer, Ph.D., Maria Carrabba, M.D., Ph.D., Mar Riveiro-Barciela, M.D., Ph.D., Maria E. Figuera Basso, Maria G. Valsecchi, Ph.D., María Hernandez-Tejero, M.D., Marialbert Acosta-Herrera, Ph.D., Mariella D'Angiò, M.D., Marina Baldini, M.D., Marina Cazzaniga, M.D., Martin Schulzky, M.A., Maurizio Cecconi, M.D., Ph.D., Michael Wittig, M.Sc., Michele Ciccarelli, M.D., Miguel Rodríguez-Gandía, M.D., Monica Bocciolone, M.D., Monica Miozzo, Ph.D., Nicola Montano, M.D., Ph.D., Nicole Braun, Nicoletta Sacchi, Ph.D., Nilda Martínez, M.D., Onur Özer, M.Sc., Orazio Palmieri, Ph.D., Paola Faverio, M.D., Paoletta Preatoni, M.D., Paolo Bonfanti, M.D., Paolo Omodei, M.D., Paolo Tentorio, M.S., Pedro Castro, M.D., Ph.D., Pedro M. Rodrigues, Ph.D., Aaron Blandino Ortiz, M.D., Rafael de Cid, Ph.D., Ricard Ferrer, M.D., Roberta Gualtierotti, M.D., Rosa Nieto, M.D., Siegfried Goerg, M.D., Salvatore Badalamenti, M.D., Ph.D., Sara Marsal, Ph.D., Giuseppe Matullo, Ph.D., Serena Pelusi, M.D., Simonas Juzenas, Ph.D., Stefano Aliberti, M.D., Valter Monzani, M.D., Victor Moreno, Ph.D., Tanja Wesse, Tobias L. Lenz, Ph.D., Tomas Pumarola, M.D., Ph.D., Valeria Rimoldi, Ph.D., Silvano Bosari, M.D., Wolfgang Albrecht, Wolfgang Peter, Ph.D., Manuel Romero-Gómez, M.D., Ph.D., Mauro D'Amato, Ph.D., Stefano Duga, Ph.D., Jesus M. Banales, Ph.D., Johannes R Hov, M.D., Ph.D., Trine Folseraas, M.D., Ph.D., Luca Valenti, M.D., Andre Franke, Ph.D., and Tom H. Karlsen, M.D., Ph.D.

The authors' affiliations are as follows: the Institute of Clinical Molecular Biology, Christian-Albrechts-University (D.E., F.D., J.K., S. May, M. Wendorff, L.W., F.U.-W., X.Y., A.T., A. Peschuck, C.G., G.H.-S., H.E.A., M.C.R., M.E.F.B., M. Schulzky, M. Wittig, N.B., S.J., T.W., W.A., M. D'Amato, A.F.), and University Hospital Schleswig-Holstein, Campus Kiel (N.B., A.F.), Kiel, the Institute for Cardiogenetics, University of Lübeck, Lübeck (J.E.), the German Research Center for Cardiovascular Research, partner site Hamburg–Lübeck–Kiel (J.E.), the University Heart Center Lübeck (J.E.), and the Institute of Transfusion Medicine, University Hospital Schleswig-Holstein (S.G.), Lübeck, Stefan-Morsch-Stiftung, Birkenfeld (M. Schaefer, W.P.), and the Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön (O.O., T.L.L.) — all in Germany; Novo Nordisk Foundation Center for Protein Research, Disease Systems Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen (D.E.); the Department of Liver and Gastrointestinal Diseases, Biodonostia Health Research Institute–Donostia University Hospital–University of the Basque Country (L.B., K.G.-E., L.I.-S., P.M.R., J.M.B.), Osakidetza Basque Health Service, Donostialdea Integrated Health Organization, Clinical Biochemistry Department (A.G.C., B.N.J.), and the Department of Liver and Gastrointestinal Diseases, Biodonostia Health Research Institute (M. D'Amato), San Sebastian, Centro de Investigación Biomédica en Red en Enfermedades Hepáticas y Digestivas, Instituto de Salud Carlos III (L.B., M. Buti, A. Albillos, A. Palom, F.R.-F., B.M., L. Téllez, K.G.-E., L.I.-S., F.M., L.R., M.R.-B., M. Rodríguez-Gandía, P.M.R., M. Romero-Gómez, J.M.B.), the Departments of Gastroenterology (A. Albillos, B.M., L. Téllez, F.M., M. Rodríguez-Gandía), Intensive Care (R.P., A.B.O.), Respiratory Diseases (D.J., A.S., R.N.), Infectious Diseases (C.Q., E.N.), and Anesthesiology (D. Pestaña, N. Martínez), Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria, University of Alcalá, and Histocompatibilidad y Biologia Molecular, Centro de Transfusion de Madrid (F.G.S.), Madrid, the Liver Unit, Department of Internal Medicine, Hospital Universitari Vall d'Hebron, Vall d'Hebron Barcelona Hospital Campus (M. Buti, A. Palom, L.R., M.R.-B.), Hospital Clinic, University of Barcelona, and the August Pi i Sunyer Biomedical Research Institute (J.F., F.A., E.S., J.F.-A., L.M., M.H.-T., P.C.), the European Foundation for the Study of Chronic Liver Failure (J.F.), Vall d'Hebron Institut de Recerca (A. Palom, F.R.-F., A.J., S. Marsal), and the Departments of Biochemistry (A.-E.G.-F., F.R.-F., A.C.-G., C.C., A.B.-G.), Intensive Care (R.F.), and Microbiology (T.P.), University Hospital Vall d'Hebron, the Immunohematology Department, Banc de Sang i Teixits, Autonomous University of Barcelona (E.M.-D.), Catalan Institute of Oncology, Bellvitge Biomedical Research Institute, Consortium for Biomedical Research in Epidemiology and Public Health and University of Barcelona, l'Hospitalet (V. Moreno), and Autonoma University of Barcelona (T.P.), Barcelona, Universitat Autònoma de Barcelona, Bellatera (M. Buti, F.R.-F., M.R.-B.), GenomesForLife–GCAT Lab Group, Germans Trias i Pujol Research Institute (A.C.N., I.G.-F., R.C.), and High Content Genomics and Bioinformatics Unit, Germans Trias i Pujol Research Institute (L. Sumoy), Badalona, Institute of Parasitology and Biomedicine Lopez-Neyra, Granada (J.M., M.A.-H.), the Digestive Diseases Unit, Virgen del Rocio University Hospital, Institute of Biomedicine of Seville, University of Seville, Seville (M. Romero-Gómez), and Ikerbasque, Basque Foundation for Science, Bilbao (M. D'Amato, J.M.B.) — all in Spain; the Division of Gastroenterology, Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milan Bicocca (P.I., C.M.),

Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (D. Prati, G.B., A.Z., A. Bandera, A.G., A.L.F., A. Pesenti, C.P., F.C., F.M.-B., F.P., F.B., G.G., G. Costantino, L. Terranova, L. Santoro, L. Scudeller, M. Carrabba, M. Baldini, M.M., N. Montano, R.G., S.P., S. Aliberti, V. Monzani, S. Bosari, L.V.), the Department of Biomedical Sciences, Humanitas University (R.A., A. Protti, A. Aghemo, A. Lleo, E.M.P., G. Cardamone, M. Cecconi, V.R., S.D.), Humanitas Clinical and Research Center, IRCCS (R.A., A. Protti, A. Aghemo, A. Lleo, A.V., C.A., E.M.P., H.K., I.M., M. Cecconi, M. Ciccarelli, M. Bocciolone, P.P., P.O., P.T., S. Badalamenti, S.D.), University of Milan (A.Z., A. Bandera, A.G., A.L.F., A. Pesenti, F.M.-B., F.P., F.B., G.G., G. Costantino, M.M., N. Montano, R.G., S.P., S. Aliberti, S. Bosari, L.V.), and the Center of Bioinformatics, Biostatistics, and Bioimaging (M.G.V.) and the Phase 1 Research Center (M. Cazzaniga), School of Medicine and Surgery, and the Departments of Emergency, Anesthesia, and Intensive Care (G.F.), Pneumologia (P.F.), and Infectious Diseases (P.B.); University of Milano–Bicocca, Milan, the European Reference Network on Hepatological Diseases (P.I., C.M.) and the Infectious Diseases Unit (P.B.), San Gerardo Hospital, Monza, the Pediatric Departement and Centro Tettamanti–European Reference Network PaedCan, EuroBloodNet, MetabERN–University of Milano–Bicocca–Fondazione MBBM–Ospedale, San Gerardo (A. Biondi, L.R.B., M. D'Angiò), the Gastroenterology Unit, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (A. Latiano, O.P.), the Department of Medical Sciences, Università degli Studi di Torino, Turin (S. Aneli, G.M.), and the Italian Bone Marrow Donor Registry, E.O. Ospedali Galliera, Genoa (N.S.) — all in Italy; the Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Surgery, Inflammatory Diseases, and Transplantation, and the Research Institute for Internal Medicine, Division of Surgery, Inflammatory Diseases, and Transplantation, Oslo University Hospital Rikshospitalet and University of Oslo (M.M.G., J.R.H., T.F., T.H.K.), and the Section for Gastroenterology, Department of Transplantation Medicine, Division for Cancer Medicine, Surgery, and Transplantation, Oslo University Hospital Rikshospitalet (J.R.H., T.F., T.H.K.), Oslo; the School of Biological Sciences, Monash University, Clayton, VIC, Australia (T.Z., M. D'Amato); Private University in the Principality of Liechtenstein (C.G.); the Institute of Biotechnology, Vilnius University, Vilnius, Lithuania (S.J.); and the Unit of Clinical Epidemiology, Department of Medicine Solna, Karolinska Institutet, Stockholm (M. D'Amato).

REFERENCES

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727-33.
2. Dong E, Du H, Gardner L. An interactive Web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020; 20:533-4.
3. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 2020;5:536-44.
4. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA 2020 February 24 (Epub ahead of print).
5. Berlin DA, Gulick RM, Martinez FJ. Severe Covid-19. N Engl J Med. DOI: 10.1056/NEJMcp2009575.
6. Marini JJ, Gattinoni L. Management of COVID-19 respiratory distress. JAMA 2020 April 24 (Epub ahead of print).
7. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395:1054-62.
8. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol 2020 April 12 (Epub ahead of print).
9. Chen R, Liang W, Jiang M, et al. Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest 2020 April 15 (Epub ahead of print).
10. Docherty AB, Harrison EM, Green CA, et al. Features of 20133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ 2020;369:m1985.
11. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA 2020; 323(20):2052-9.
12. Levi M, Thachil J, Iba T, Levy JH. Coagulation abnormalities and thrombosis in patients with COVID-19. Lancet Haematol 2020;7(6):e438-e440.
13. Varga Z, Flammer AJ, Steiger P, et al. Endothelial cell infection and endotheliitis in COVID-19. Lancet 2020;395:1417-8.
14. Ackermann M, Verleden SE, Kuehnel M, et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2015432.
15. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118-25.
16. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. March 6, 2019 (https://www.biorxiv.org/content/10.1101/563866v1). preprint.
17. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.
18. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26:2190-1.
19. Bugert P, Rink G, Kemp K, Klüter H. Blood group ABO genotyping in paternity testing. Transfus Med Hemother 2012;39: 182-6.
20. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA database. Nucleic Acids Res 2020;48(D1):D948-D955.
21. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 2010;42:295-302.
22. Bentham J, Morris DL, Graham DSC, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet 2015;47:1457-64.
23. Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 2009;41: 334-41.
24. Julià A, González I, Fernández-Nebro A, et al. A genome-wide association study identifies SLC8A3 as a susceptibility locus for ACPA-positive rheumatoid arthritis. Rheumatology (Oxford) 2016; 55:1106-11.
25. López-Isac E, Acosta-Herrera M, Kerick M, et al. GWAS for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. Nat Commun 2019;10:4955.
26. Obón-Santacana M, Vilardell M, Carreras A, et al. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. BMJ Open 2018;8(3):e018324.
27. Galván-Femenía I, Obón-Santacana M, Piñeyro D, et al. Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. J Med Genet 2018;55:765-78.
28. Vuille-dit-Bille RN, Camargo SM, Emmenegger L, et al. Human intestine luminal ACE2 and amino acid transporter

expression increased by ACE-inhibitors. Amino Acids 2015;47:693-705.
29. Kuba K, Imai Y, Ohto-Nakanishi T, Penninger JM. Trilogy of ACE2: a peptidase in the renin-angiotensin system, a SARS receptor, and a partner for amino acid transporters. Pharmacol Ther 2010; 128:119-28.
30. Wein AN, McMaster SR, Takamura S, et al. CXCR6 regulates localization of tissue-resident memory CD8 T cells to the airways. J Exp Med 2019;216:2748-62.
31. Hickey MJ, Held KS, Baum E, Gao JL, Murphy PM, Lane TE. CCR1 deficiency increases susceptibility to fatal coronavirus infection of the central nervous system. Viral Immunol 2007;20:599-608.
32. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur J Hum Genet 2020;28:715-8.
33. Zhao J, Yang Y, Huang H, et al. Relationship between the ABO blood group and the COVID-19 susceptibility. March 27, 2020 (https://www.medrxiv.org/content/10.1101/2020.03.11.20031096v2). preprint.
34. Zietz M, Tatonetti NP. Testing the association between blood type and COVID-19 infection, intubation, and death. April 11, 2020 (https://www.medrxiv.org/content/10.1101/2020.04.08.20058073v1). preprint.
35. Cheng Y, Cheng G, Chui CH, et al. ABO blood group and susceptibility to severe acute respiratory syndrome. JAMA 2005;293:1450-1.
36. Breiman A, Ruvën-Clouet N, Le Pendu J. Harnessing the natural anti-glycan immune response to limit the transmission of enveloped viruses such as SARS-CoV-2. PLoS Pathog 2020;16(5):e1008556.
37. Comuzzie AG, Cole SA, Laston SL, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS One 2012; 7(12):e51954.
38. Aziz M, Fatima R, Assaly R. Elevated interleukin-6 and severe COVID-19: a meta-analysis. J Med Virol 2020 April 28 (Epub ahead of print).
39. Naitza S, Porcu E, Steri M, et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet 2012; 8(1):e1002480.
40. Franchini M, Crestani S, Frattini F, Sissa C, Bonfanti C. ABO blood group and von Willebrand factor: biological implications. Clin Chem Lab Med 2014;52:1273-6.
41. Murray GP, Post SR, Post GR. ABO blood group is a determinant of von Willebrand factor protein levels in human pulmonary endothelial cells. J Clin Pathol 2020;73:347-9.
42. Thomson G, Bodmer WF. Letter: population stratification as an explanation of IQ and ABO association. Nature 1975; 254:363-4.

Annex III

Supplementary Material for Chapter 1

Unique pathogen peptidomes facilitate pathogen-specific selection and specialization of MHC alleles

Onur Özer[1,2] & Tobias L. Lenz[1,2,*]

[1] Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany

[2] Research Unit for Evolutionary Immunogenomics, Department of Biology, Universität Hamburg, 20146 Hamburg, Germany

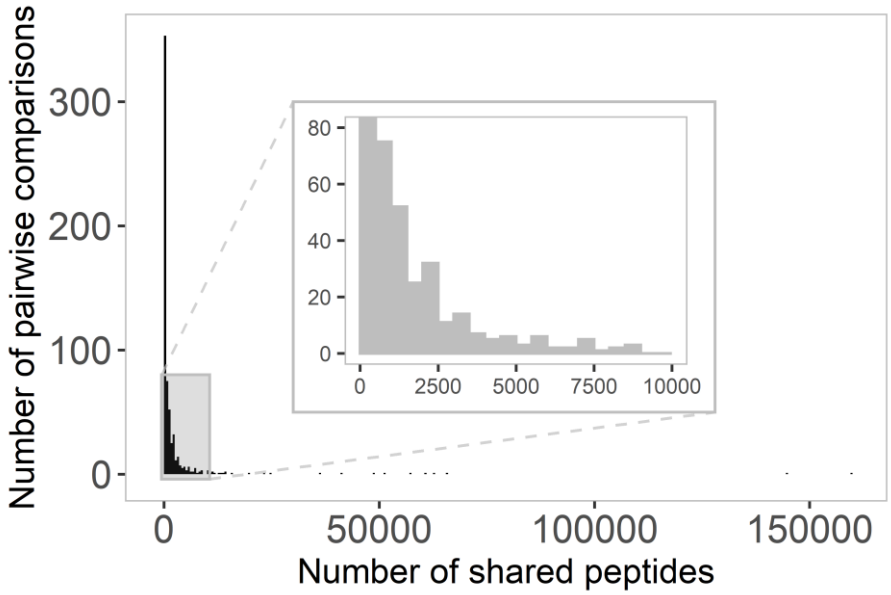


Figure S1. Distribution of the number of shared peptides among all pairs of pathogens (N=630). For each pathogen pair, shared peptides represent the overlap between both pathogen peptidomes.

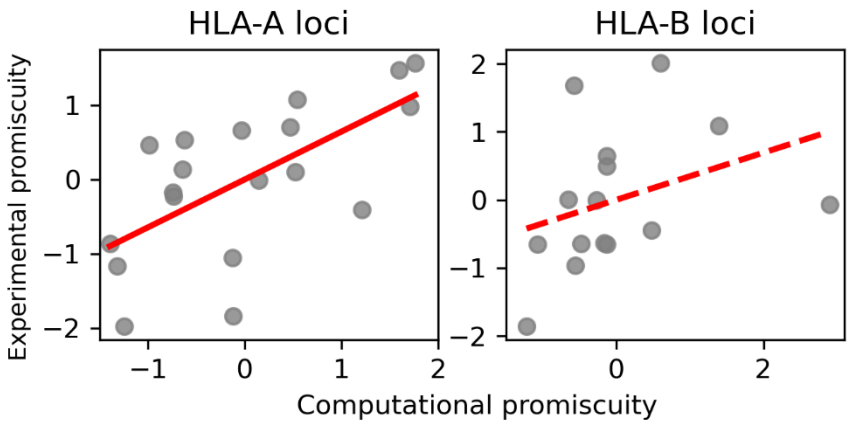


Figure S2. Experimental and computational promiscuity values are correlated for HLA-A (Kendall's tau = 0.51, p = 0.002) and HLA-B (Kendall's tau = 0.37, p = 0.054) loci. Each dot represents HLA-A (N = 19) and HLA-B (N = 15) variants for which sufficient experimental data is available on the Immune Epitope Database. Computational promiscuity was calculated as the fraction of the bound peptides among the complete dataset of 51.9 Mio peptides. Experimental promiscuity was calculated based on the data from the Immune Epitope Database as the fraction of positive binding assays among the total number of assays for each HLA allele. Computational and experimental promiscuity values were normalized for comparison. Solid red line represents significant correlation while dashed line represents positive trend. The HLA alleles included in the analysis are A*01:01, A*02:01, A*02:02, A*02:03, A*02:06, A*03:01, A*11:01, A*23:01, A*24:02, A*24:03, A*26:01, A*30:01, A*30:02, A*31:01, A*33:01, A*68:01, A*68:02, A*69:01, A*80:01, B*07:02, B*08:01, B*15:01, B*15:17, B*18:01, B*27:05, B*35:01, B*39:01, B*40:01, B*44:02, B*46:01, B*51:01, B*53:01, B*57:01, B*58:01.

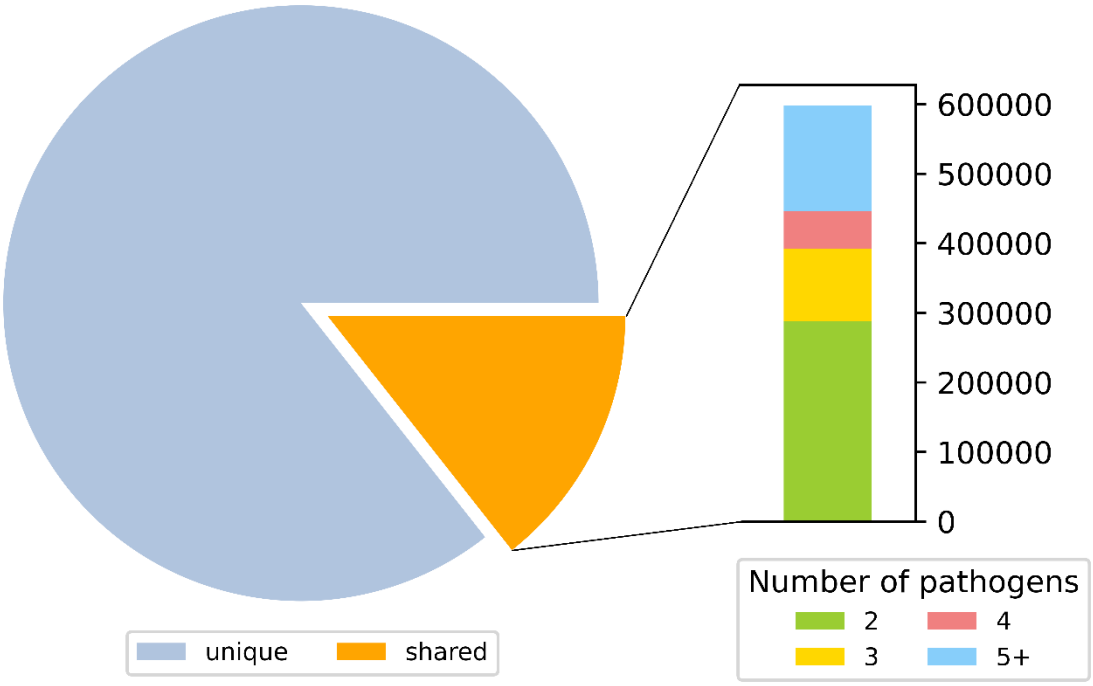


Figure S3. Peptide sharing among human pathogens based on groups of peptides that are bound by same set of HLA alleles. A total of 4,157,475 groups were analyzed and only 14.4% of these groups were shared among pathogens. The pie chart represents the proportions of shared (N = 597,700) and unique (N = 3,559,775) groups of peptides while the bar chart shows the extent of sharing across pathogen species for all shared peptide groups.

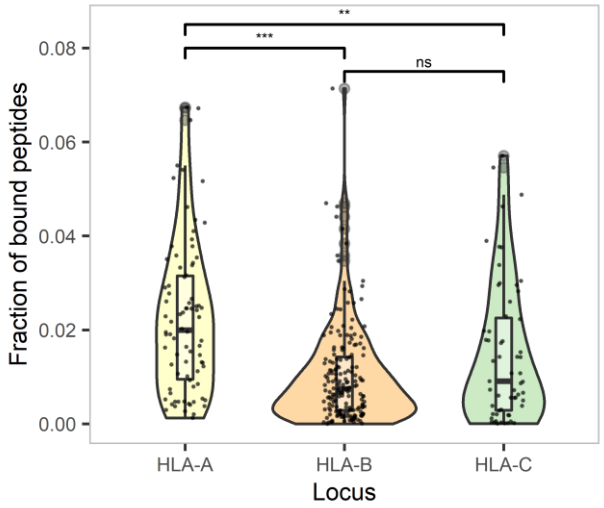


Figure S4. Variation of promiscuity within and among MHC loci. The fraction of bound peptides out of the complete set of unique peptides (N = 51,861,826 nine-mers) is shown for common variants of the three HLA class I loci (HLA-A; n = 82, HLA-B; n = 180 and HLA-C; n = 59). Each dot represents an HLA variant. Upper and lower edges of boxes correspond to the first and the third quartiles of the data while whiskers extend up to the data at most 1.5 IQR away from the edges of the box. Statistical significance from Wilcoxon rank sum test is indicated: *** - $p<0.001$, ** - $p<0.01$, ns - $p>0.05$.

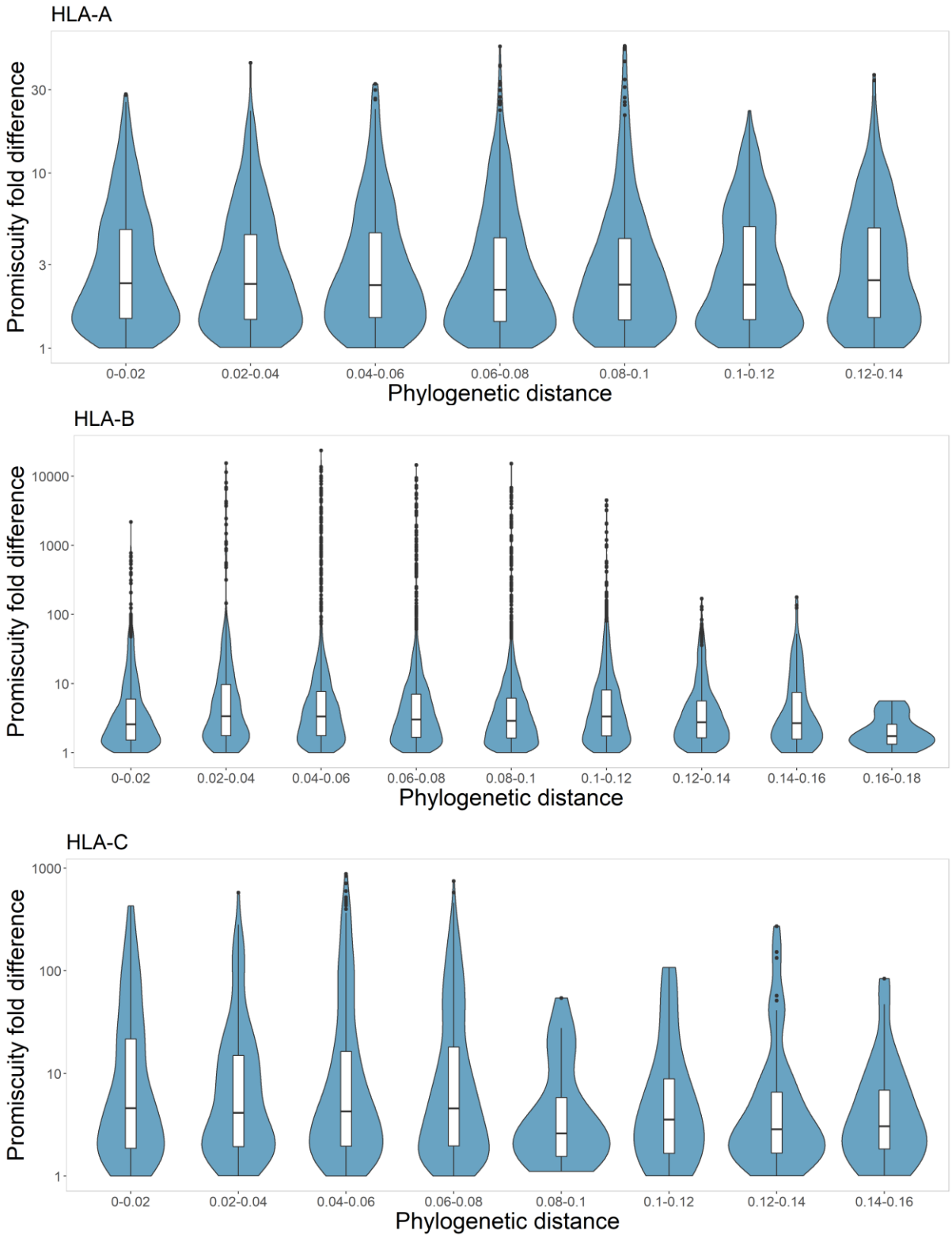


Figure S5. Differences in promiscuity in comparison to phylogenetic distance between pairs of HLA variants (HLA-A; n = 82, HLA-B; n = 180 and HLA-C; n = 59). Phylogenetic distance between each pair is calculated as tip-to-tip distance in a phylogenetic tree. Promiscuity differences were calculated for each pair as the ratio of the number of bound peptides of more promiscuous variant to the number of bound peptides of the less promiscuous variant. Variant pairs were binned based on phylogenetic distance for better visualization.

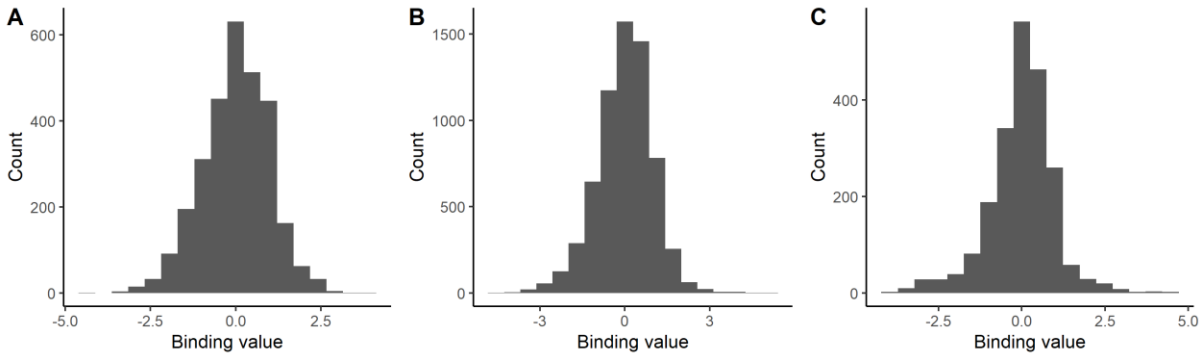


Figure S6. Distribution of standardized binding proportions for (A) HLA-A (n = 82), (B) HLA-B (n = 180) and (C) HLA-C (n = 59) loci. The fraction of bound peptides from each pathogen was normalized for each allele to obtain binding values (i.e. standardized binding proportions).

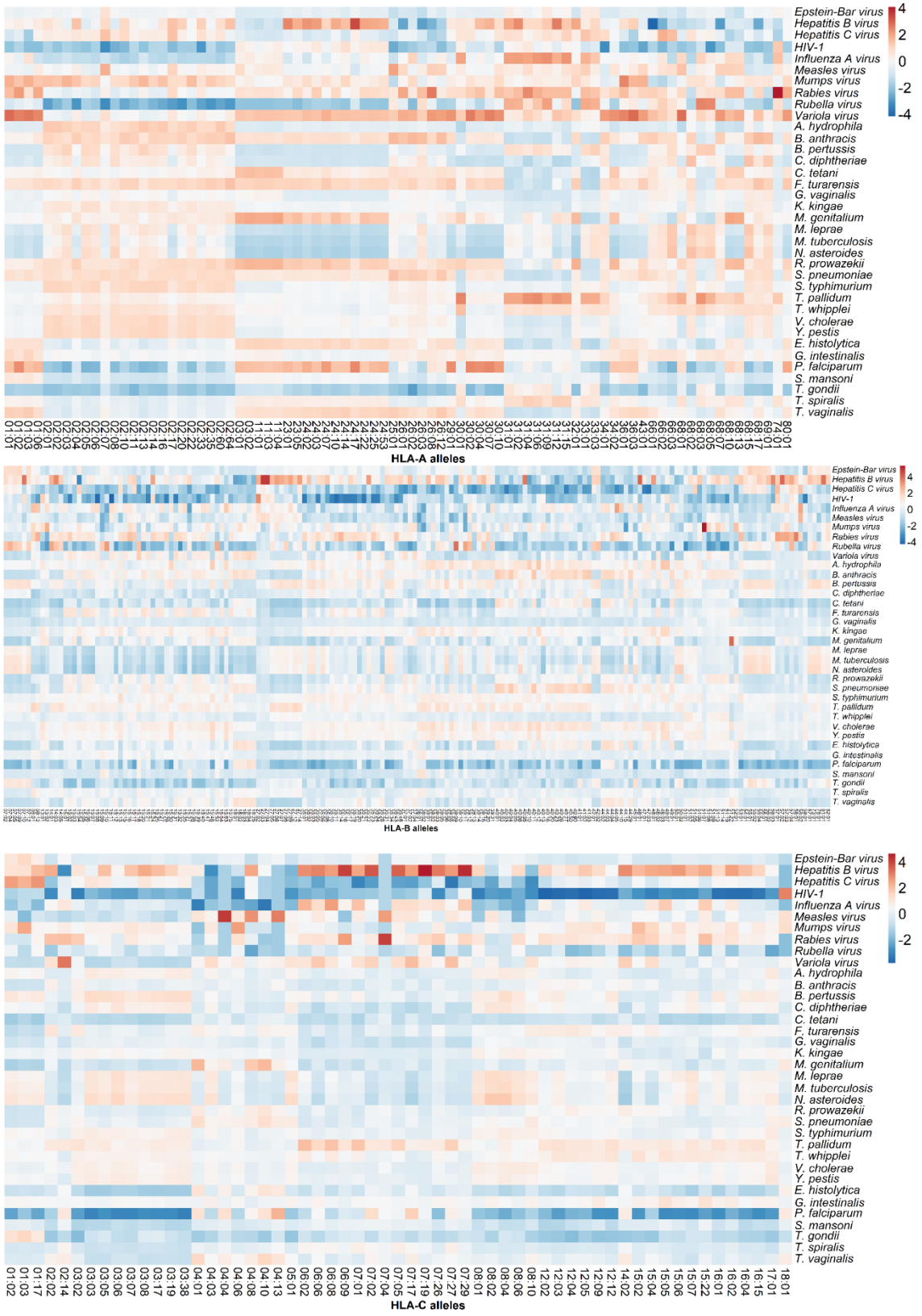


Figure S7. Standardized proportions of bound peptides by each HLA variant (x-axis) (HLA-A: n = 82; HLA-B: n = 180; HLA-C: n = 59) from each pathogen (y-axis)(n = 36). Corresponding HLA locus of each heatmap is written on the x-axis label.

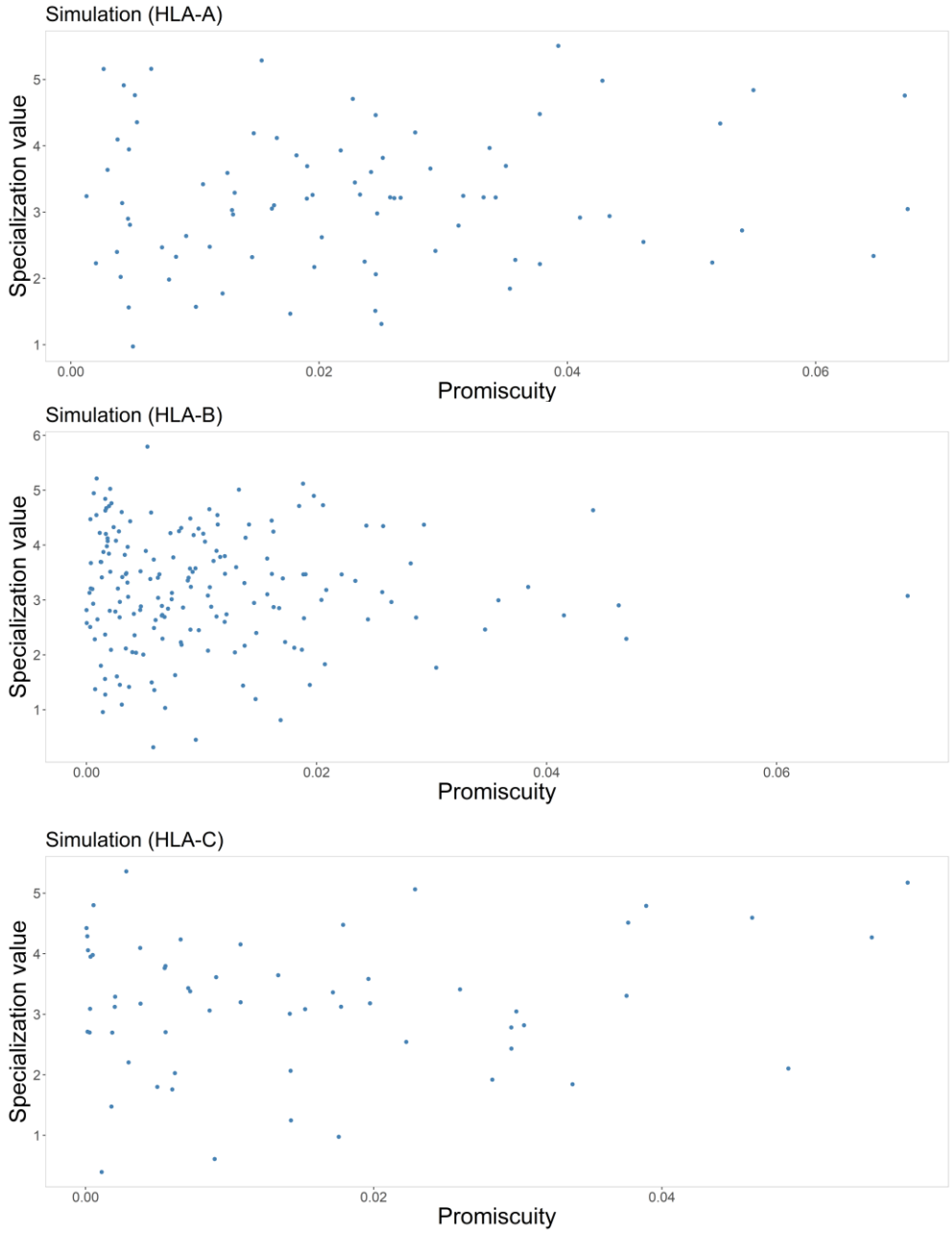


Figure S8. Specialization as a function of promiscuity in the simulated data. Each dot corresponds to a simulated HLA variant. The number of bound peptides by each HLA variant from each pathogen were simulated by using the observed promiscuity of an HLA variant as the probability of binding a peptide. Specialization value of simulated HLA variants were calculated in the same way as the real data (i.e. difference between the maximum and the median values of standardized proportions of bound peptides). No significant correlation between specialization and promiscuity were observed for any locus in the simulated data (Kendall correlation, HLA-A: tau = 0.05 p = 0.54; HLA-B: tau = -0.02 p = 0.65; HLA-C: tau = 0.01 p = 0.93).

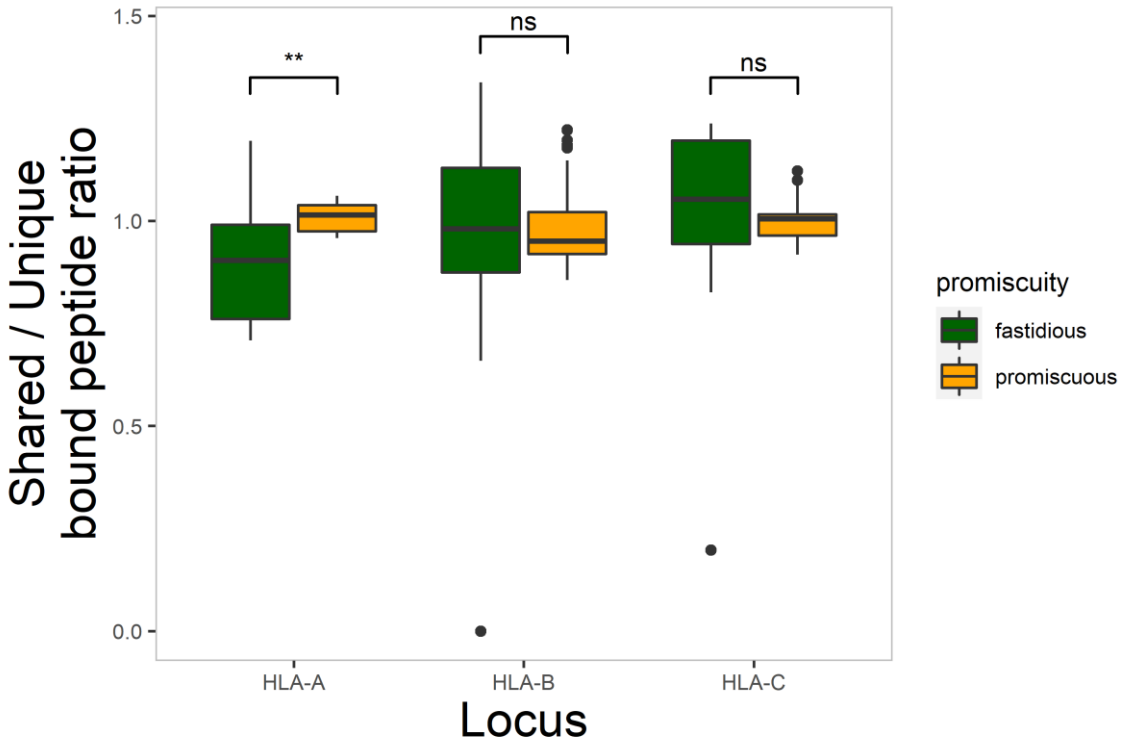


Figure S9. Comparison of the most promiscuous (top 25%) and the most fastidious (bottom 25%) HLA variants regarding binding to shared peptides. Shared peptides are the peptides observed in at least three different pathogens. The ratio on the y-axis is expected to be one if there is no tendency of alleles to bind either shared or unique peptides. ** - $p<0.01$, ns - $p>0.05$ (Wilcoxon rank sum test)

Table S1. Human pathogens used for the analysis. The asterisk (*) indicates that the pathogen is used for the pairwise peptide sharing analysis in Figure 2.

Organism	UniProt Proteome ID	Date accessed	Number of proteins	Number of nine-mers
Aeromonas hydrophila	UP000000756	11.08.2018	4121	1330710
Bacillus anthracis (*)	UP000000594	11.12.2018	5490	1375030
Bordetella pertussis (*)	UP000002676	3.07.2018	3258	1013488
Clostridium tetani (*)	UP000001412	3.07.2018	2415	780947
Corynebacterium diphtheriae (*)	UP000002198	3.07.2018	2265	701220
Entamoeba histolytica	UP000001926	3.07.2018	7959	2972321
Epstein-Barr virus	UP000153037	3.07.2018	92	39553
Francisella turarensis	UP000001174	11.08.2018	1528	459802
Gardnerella vaginalis	UP000001453	11.08.2018	1365	470742
Giardia intestinalis (*)	UP000001548	3.07.2018	7154	3048781
Hepatitis B virus	UP000007930	3.07.2018	7	1760
Hepatitis C virus	UP000000518	3.07.2018	2	3154
HIV1	UP000002241	11.12.2018	9	3062
Influenza A virus	UP000009255	3.07.2018	13	4508
Kingella kingae	UP000004207	11.08.2018	2102	551780
Measles virus	UP000008699	3.07.2018	8	4907

Mumps virus	UP000002331	11.12.2018	8	4759
Mycobacterium leprae (*)	UP000000806	11.12.2018	1603	521894
Mycobacterium tuberculosis (*)	UP000001584	12.06.2018	3993	1263777
Mycoplasma genitalium (*)	UP000000807	11.12.2018	483	172194
Nocardia asteroids	UP000017048	11.08.2018	6459	2018795
Plasmodium falciparum (*)	UP000001450	16.06.2018	5449	3806032
Rabies virus	UP000008649	3.07.2018	5	3570
Rickettsia prowazekii (*)	UP000002480	11.08.2018	834	271251
Rubella virus	UP000000571	3.07.2018	2	3162
Salmonella typhimurium (*)	UP000001014	11.08.2018	4431	1341710
Schistosoma mansoni	UP000008854	3.07.2018	11723	4939504
Streptococcus pneumonia (*)	UP000002642	2.07.2018	2823	541479
Toxoplasma gondii	UP000002226	3.07.2018	8404	6440952
Treponema pallidum (*)	UP000000811	11.08.2018	1027	337608
Trichinella spiralis	UP000006823	3.07.2018	16041	4319328
Trichomonas vaginalis	UP000001542	3.07.2018	50190	11405499
Tropheryma whpplei	UP000002200	11.08.2018	805	251906
Variola virus	UP000002060	3.07.2018	199	52593
Vibrio cholera (*)	UP000000584	2.07.2018	3783	1110752

Yersinia pestis (*)	UP000000815	3.07.2018	3909	1193604

Annex IV

Supplementary Material for Chapter II

Balancing selection rather than local adaptation determines HLA gene variation in ethnically diverse African populations

Onur Özer[1]*, Daniel Harris[2]*, Michael McQuillan[2]*, Tristan Hayeck[3,4], Eric Mbunwe[2], Timothy L. Mosbruger[3], Jamie L. Duke[3], Clinton Azuure[1], Tzun-Wen Shaw[3], Thomas Nyambo[5], Sununguko Wata Mpoloka[6], Gaonyadiwe George Mokone[7], Gurja Belay[8], Charles Fokunang[9], Alfred K. Njamnshi[10], Martin Maiers[11], Dimitri Monos[3,4#], Tobias L. Lenz[1#], Sarah Tishkoff[2#]

[1] Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany

[2] Department of Genetics, University of Pennsylvania, Philadelphia, PA

[3] Immunogenetics Laboratory, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[5] Department of Biochemistry, Kampala International University in Tanzania, P.O. Box 9790, Dar es Salaam, Tanzania.

[6] Department of Biological Sciences, Faculty of Science, University of Botswana Gaborone, Private Bag UB 0022, Gaborone, Botswana.

[7] Department of Biomedical Sciences, Faculty of Medicine, University of Botswana Gaborone, Private Bag UB 0022, Gaborone, Botswana.

[8] Department of Microbial Cellular and Molecular Biology, Addis Ababa University, Ethiopia

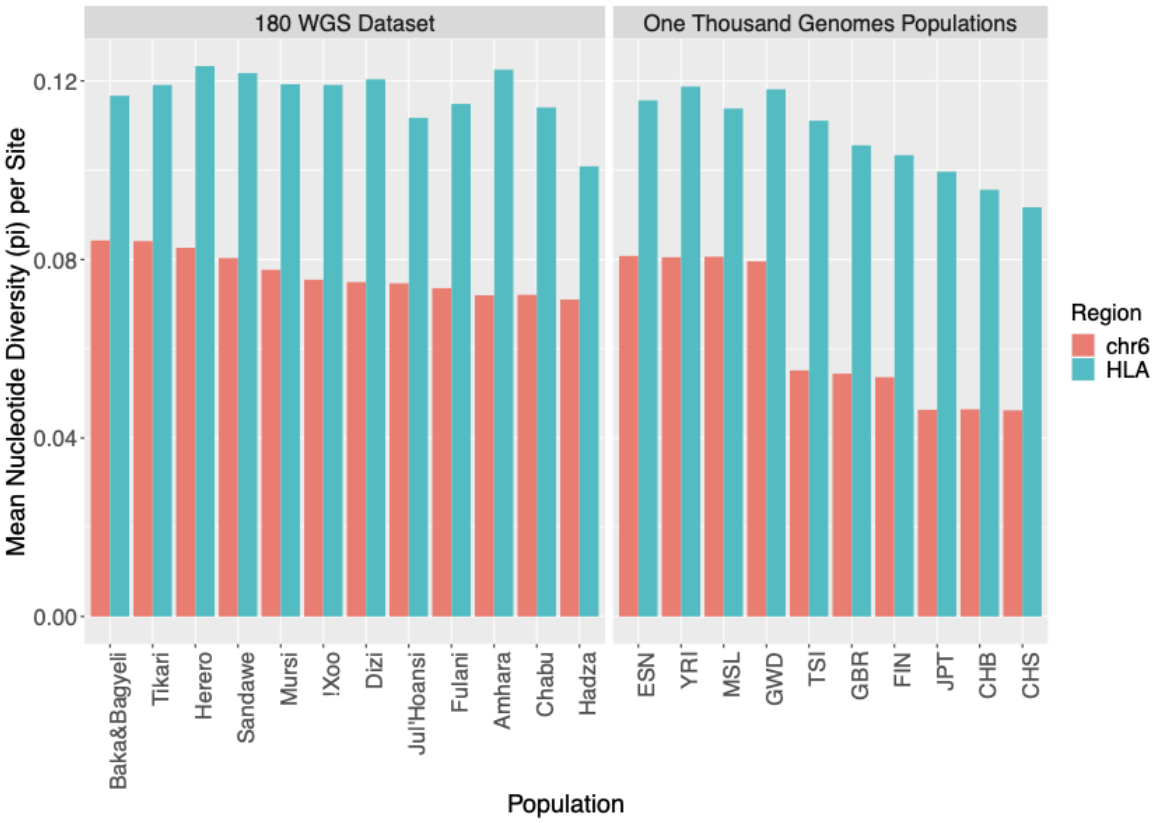
[9] Department of Pharmacotoxicology and Pharmacokinetics, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, P.O. Box 337, Yaoundé, Cameroon.

[10] Department of Neurology, Central Hospital Yaoundé; Brain Research Africa Initiative (BRAIN), Neuroscience Lab, Faculty of Medicine and Biomedical Sciences, The University of Yaoundé I, P.O. Box 337, Yaoundé, Cameroon.

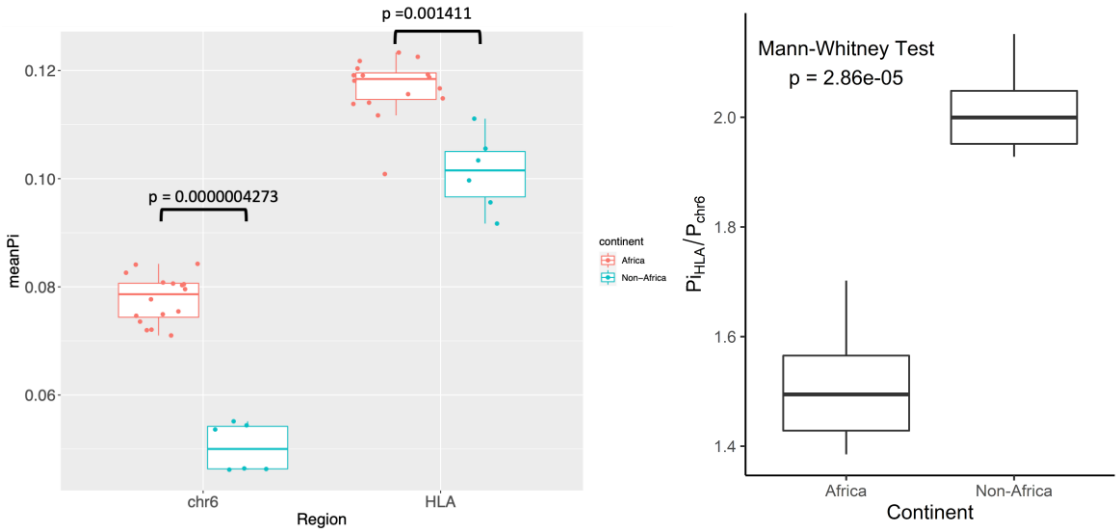
[11] CIBMTR (Center for International Blood and Marrow Transplant Research), National Marrow Donor Program/Be The Match, Minneapolis, MN, USA

*** Shared first authors**

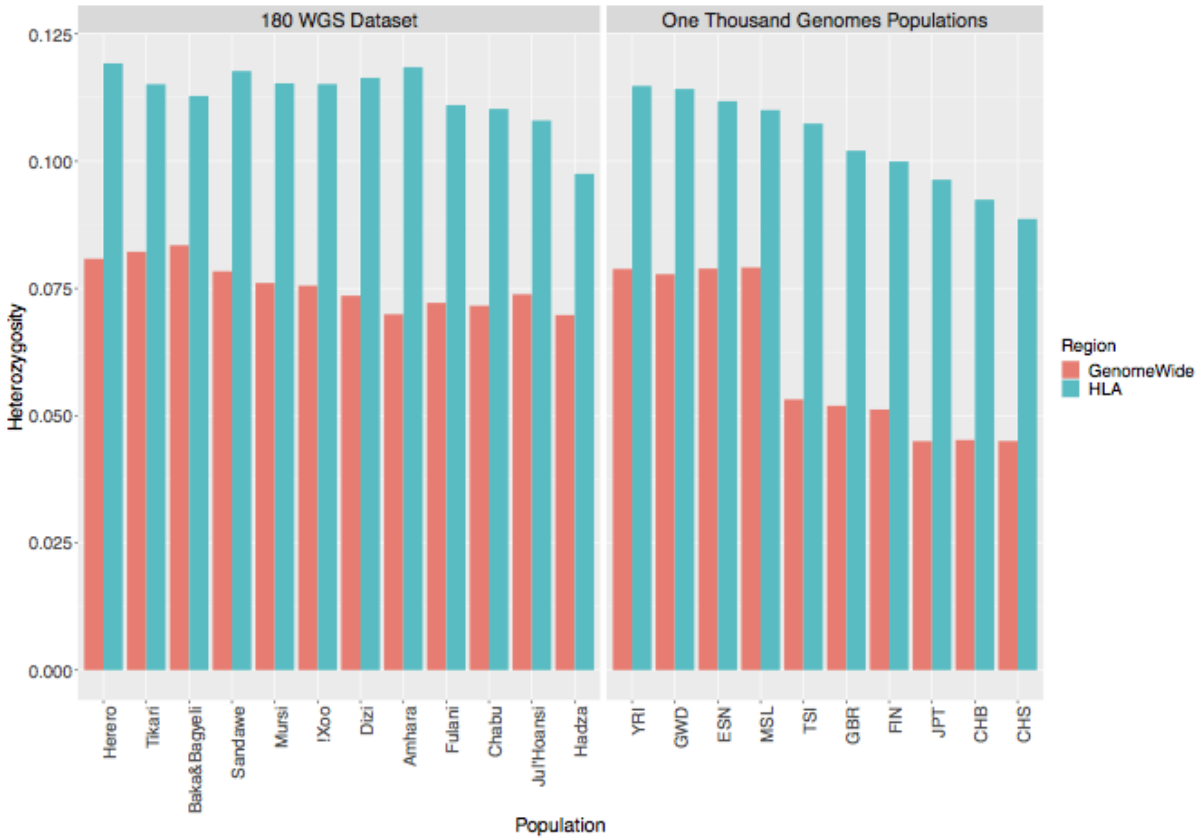
[#] Shared senior and corresponding authors



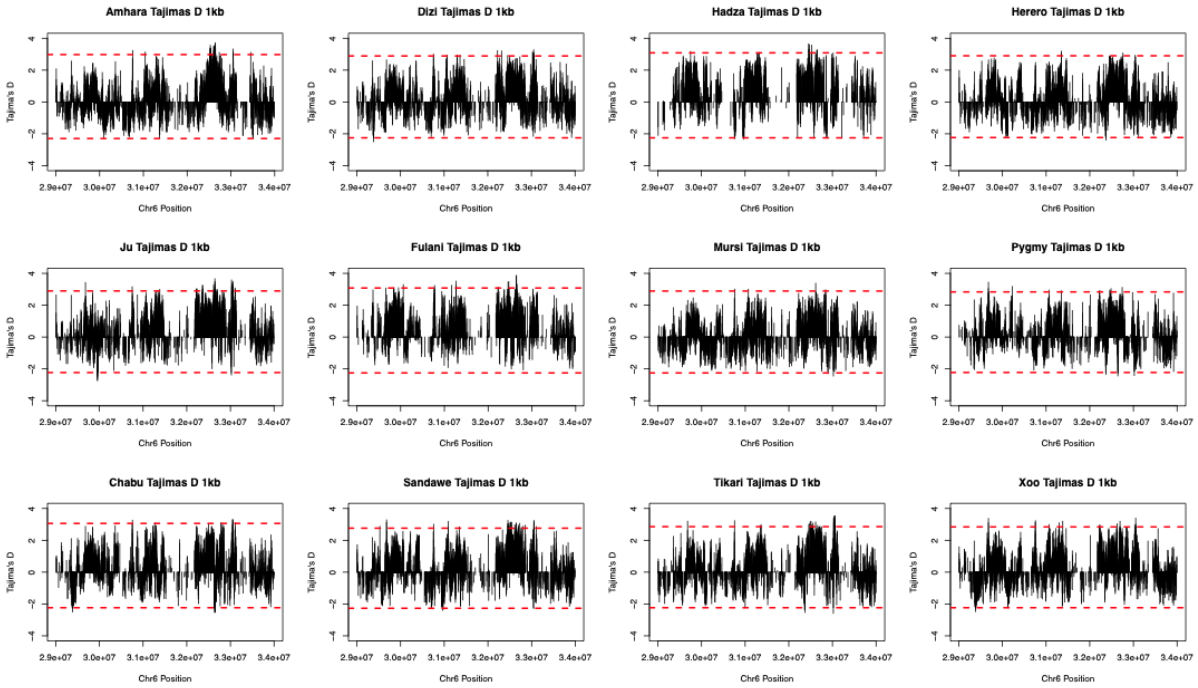
Supplementary Figure S1. Nucleotide Diversity (pi) calculated per site for each population, averaged over the HLA region (chr6:29mb-34mb) and the remainder of chromosome 6. SNP data were LD pruned (583,063 chr6 SNPs vs 23,179 HLA-region SNPs). One thousand genomes (1KG) populations were downsampled to n=15/population.



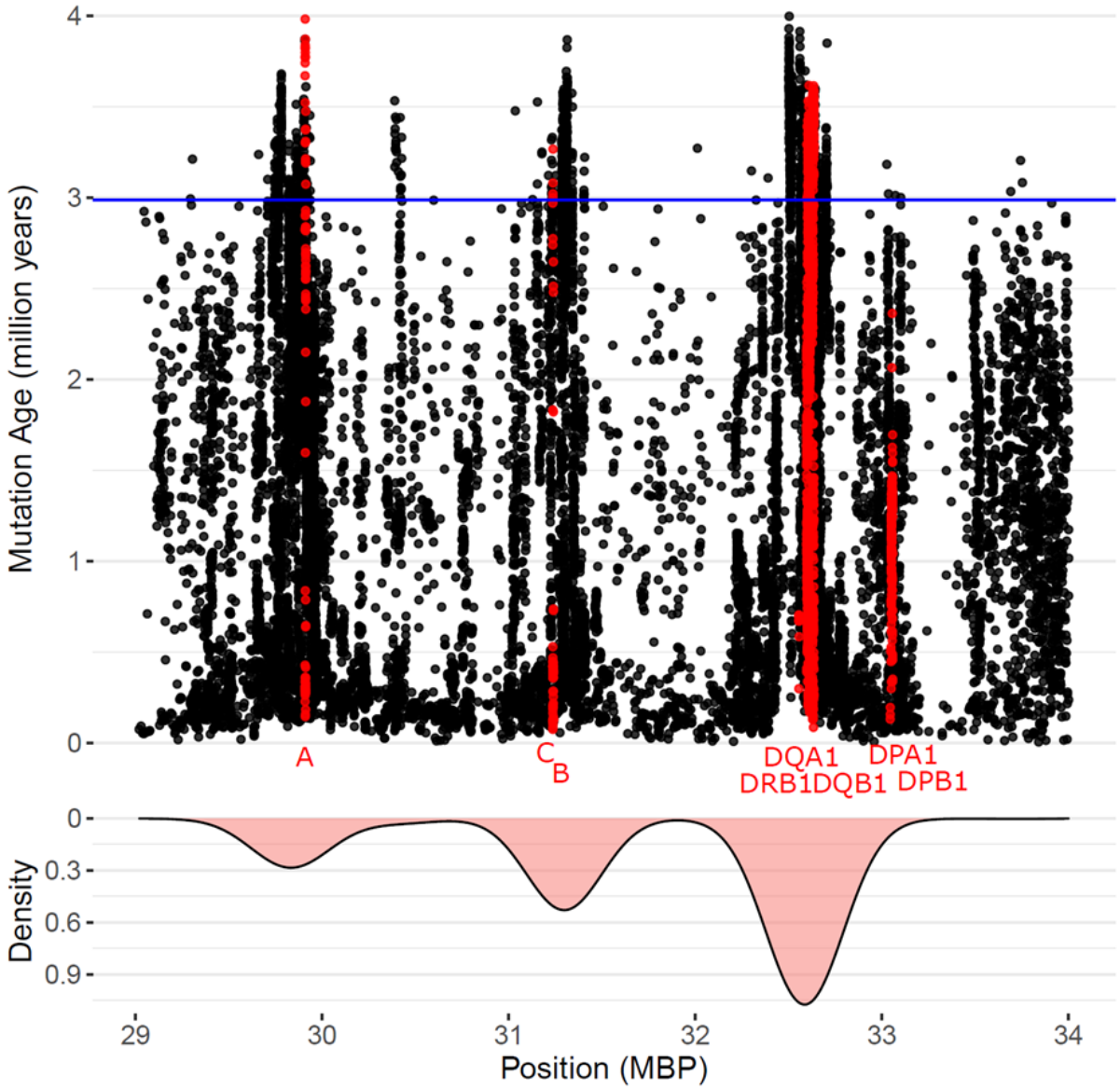
Supplementary Figure S2: Loss of genetic diversity due to out-of-Africa bottleneck is less severe for the MHC region compared to the entire chromosome 6. Left panel; Average per site nucleotide diversity (pi) over the MHC region (chr6:29mb-34mb) or the remainder of chromosome 6, calculated for each population. P-values denote two-sample t-tests (N=22). Right panel; ratio of the average nucleotide diversity within the MHC region to the average diversity within the remainder of chromosome 6, calculated for each population (N=22). The ratio is significantly higher in non-African populations, indicating that the out-of-Africa bottleneck has affected general genetic diversity (here calculated for chromosome 6) more dramatically than the MHC region.



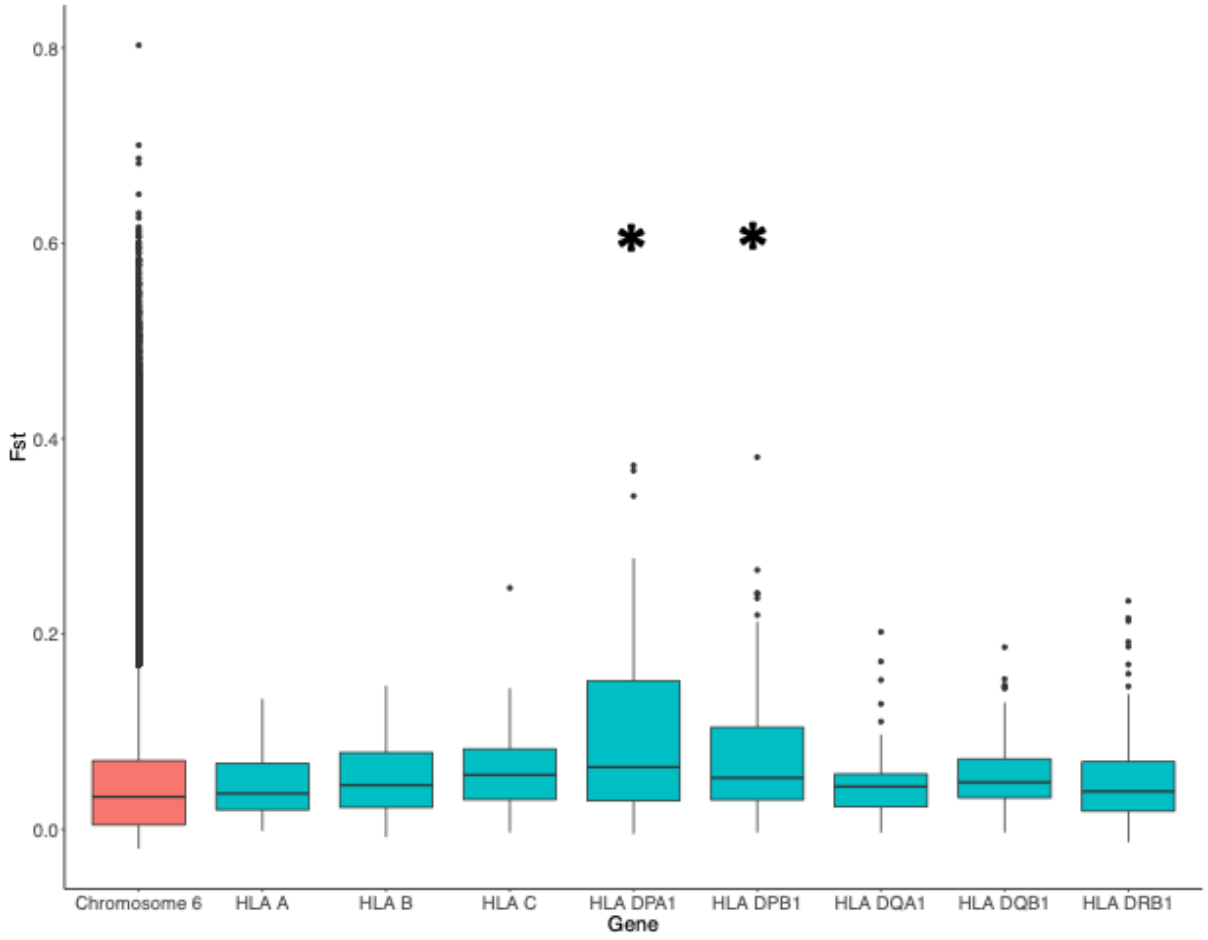
Supplementary Figure S3. Expected heterozygosity (2*p*q) genome-wide vs. the MHC region (chr6:29,000,000-34,000,000; hg19) in all 180WGS populations and relevant 1000 genomes populations. 1000 genomes populations downsampled to n=15/population and dataset LD-pruned to include 23,179 LD-pruned SNPs within the HLA region vs. 9,783,767 LD-pruned SNPs genome-wide. Heterozygosity was calculated per SNP and averaged over the entire genome or the HLA region.



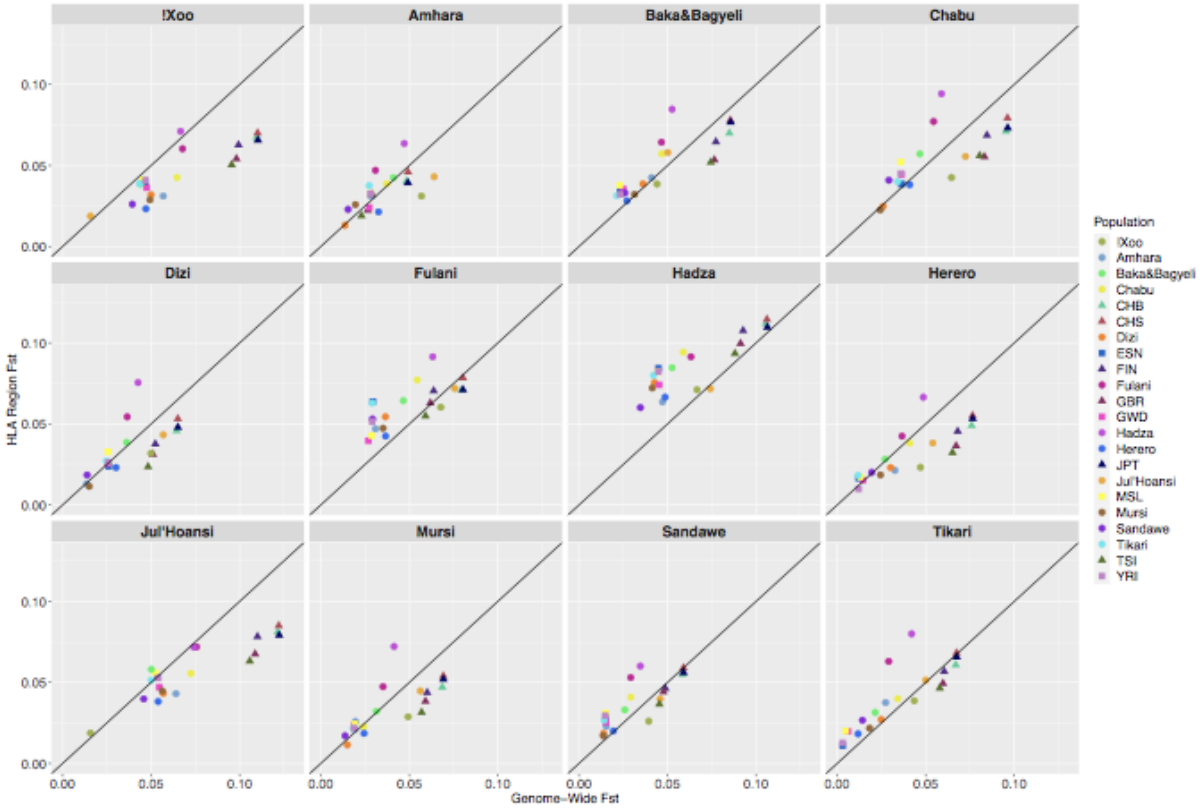
Supplementary Figure S4. Tajima's D (1kb windows) calculated for each population in the 180 whole-genome sequencing dataset (n=15 per population), zoomed in on HLA region.



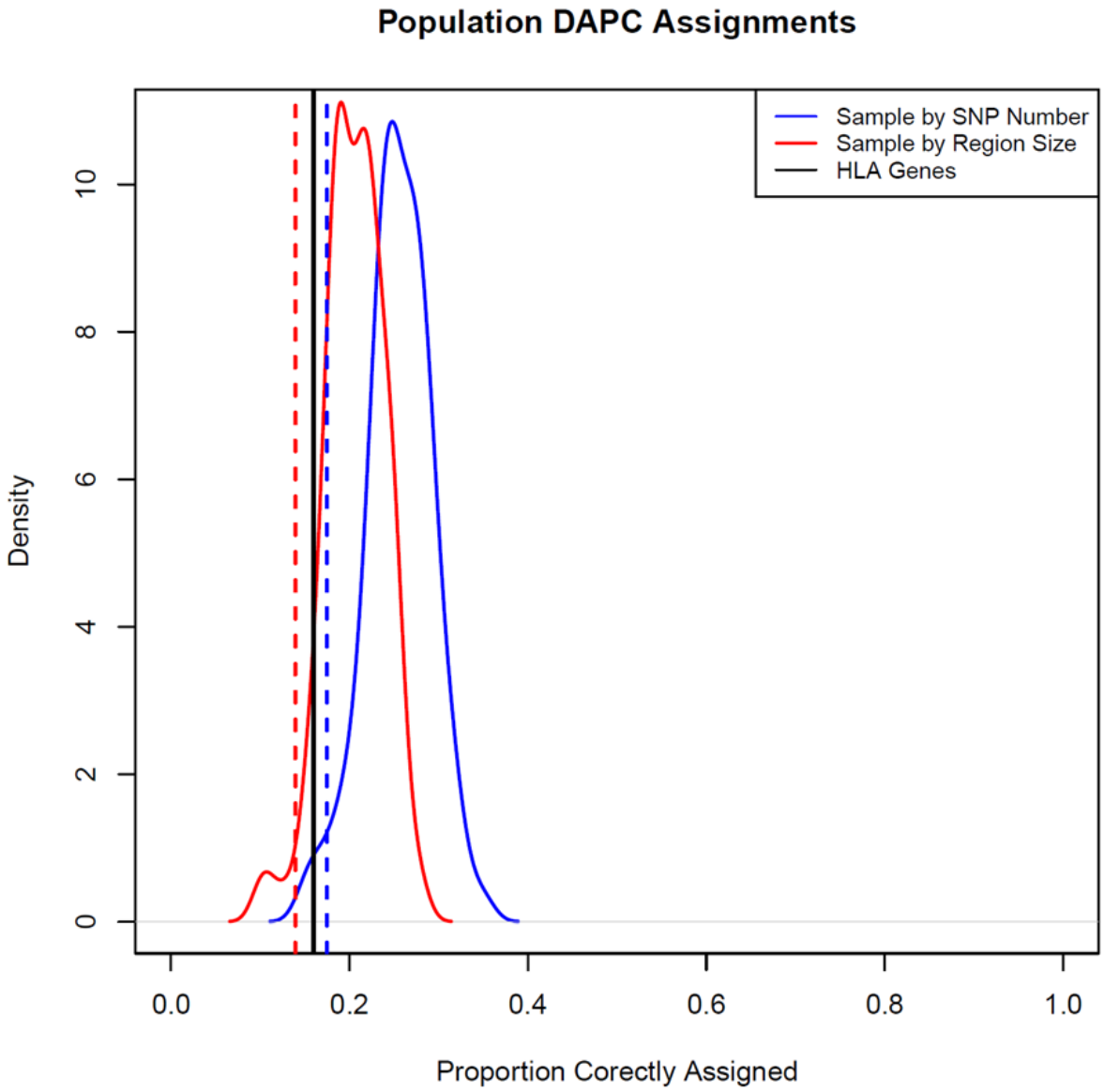
Supplementary Figure S5. Age of Mutations in the MHC region (chr6:29,000,000 - 34,000,000). Top) Manhattan plot of mutation ages across the MHC, with class I and II HLA genes highlighted in red. The blue line represents the 99th percentile of mutation age across chromosome 6 (2,983,955 years). Bottom) Density of mutations in the MHC region that are greater than or equal to the 99th percentile of mutation ages across chr6. There are three main peaks of old mutations, which generally correlates with HLA class I and class II genes.



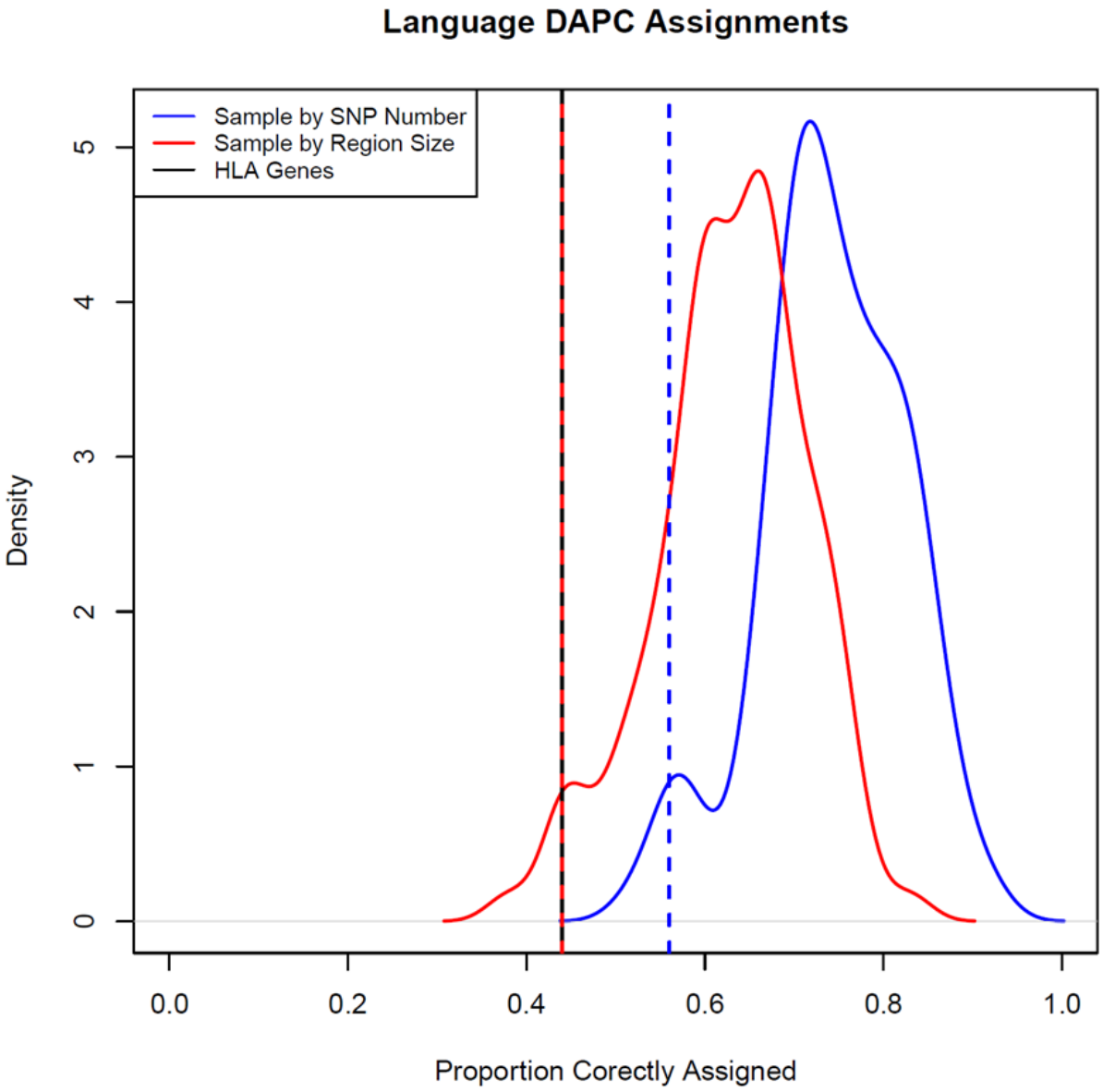
Supplementary Figure S6. Weir and Cockerham (1984) Fst distributions for Class 1 and 2 HLA genes compared to all SNPs on chromosome 6. Fst analysis contains 22 populations total (12 from the African 180WGS dataset, 10 from the one thousand genomes dataset). One thousand genomes populations were downsampled to n=15/population. Fst for genes marked with asterisk are significantly different than the Fst of all SNPs on chromosome 6, based on two-sample t-tests after bonferroni correction ($p<0.00625$).



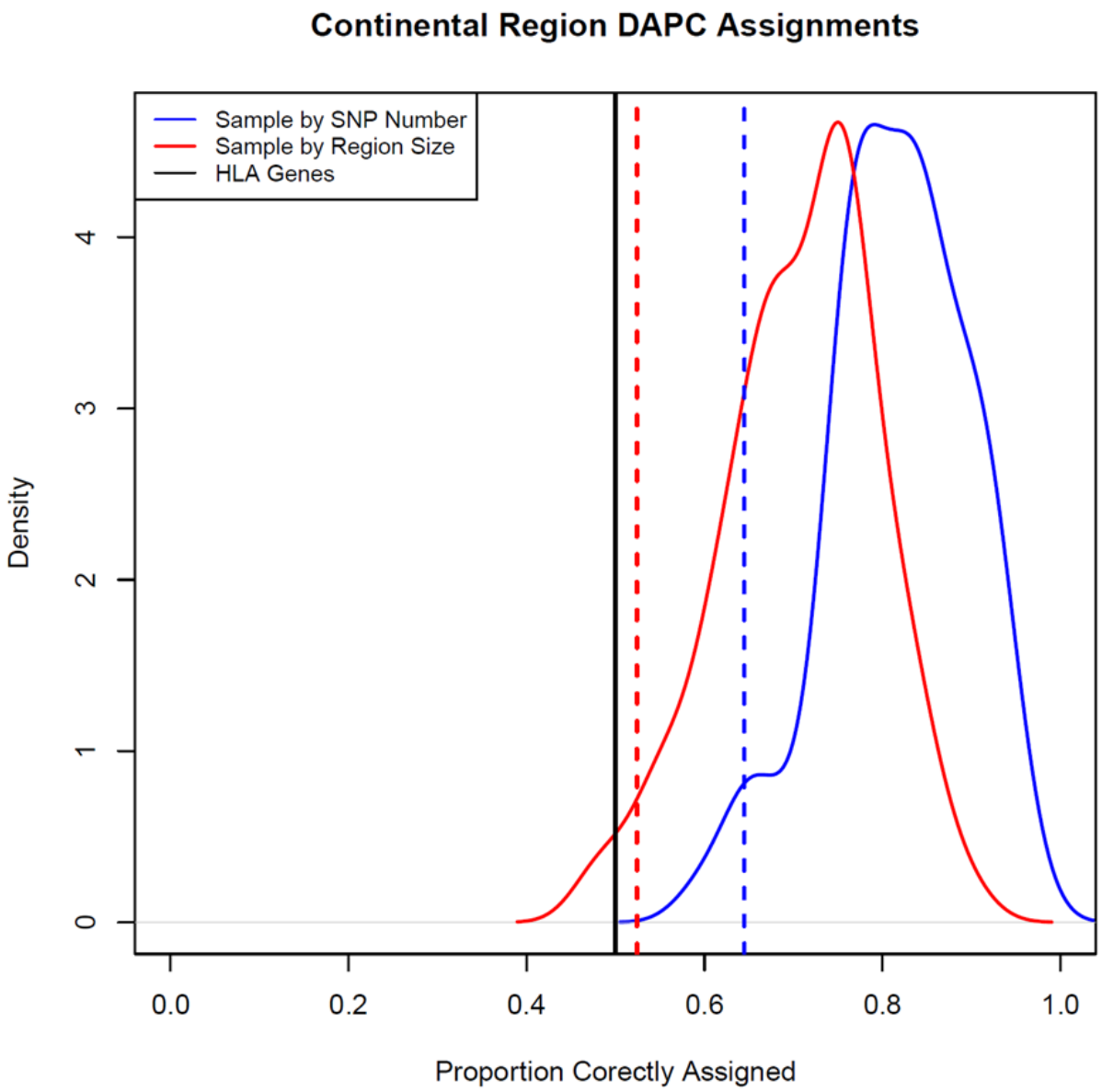
Supplementary Figure S7. Pairwise Fst calculated between all populations in the 180WGS and select 1000 genomes populations. Genome-wide Fst is the average pairwise Fst for 9,783,767 LD-pruned SNPs. HLA region Fst is the average pairwise Fst for 23,179 LD-pruned SNPs between chr6:29,000,000-34,000,000. 1000 genomes populations down-sampled to n=15 per population. Triangles indicate non-African populations, circles indicate African populations from the 180WGS dataset, and squares indicate African population from one thousand genomes.



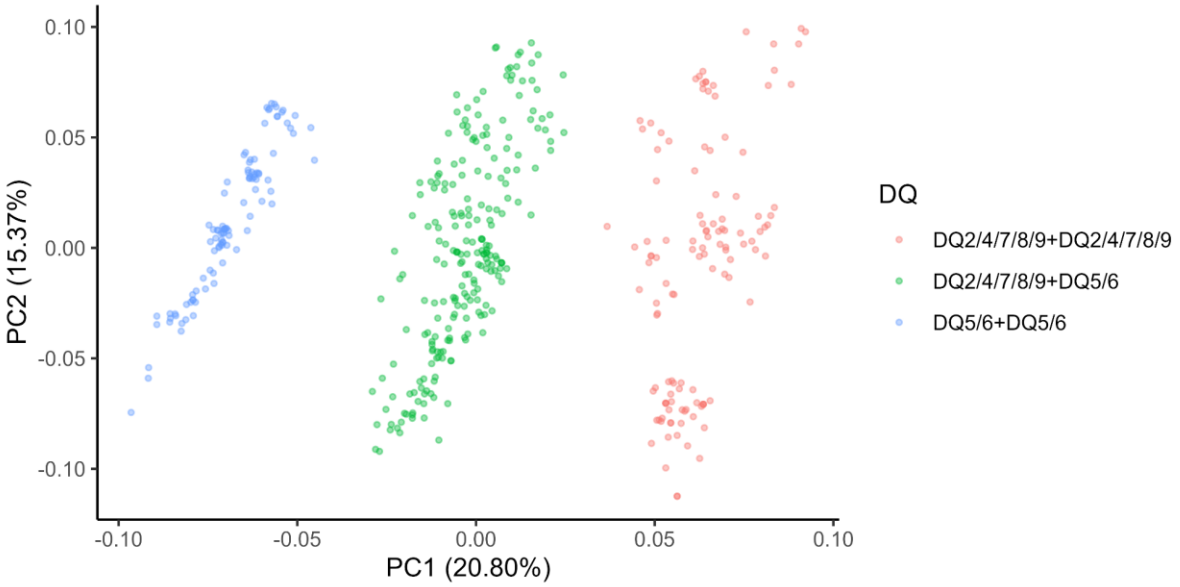
Supplementary Figure S8. Density plot of the proportion of correctly assigned population labels by DAPC in randomly sampled regions of the same number of base pairs (red) and the same number of SNPs (blue) as the Class I and Class II HLA genes. The dashed red and blue lines represent the 2.5 percentile of the same number of base pairs and same number of SNPs distributions, respectively. The solid black line represents the proportion of correctly assigned population labels by DAPC based on SNPs from the Class I and Class II HLA genes.



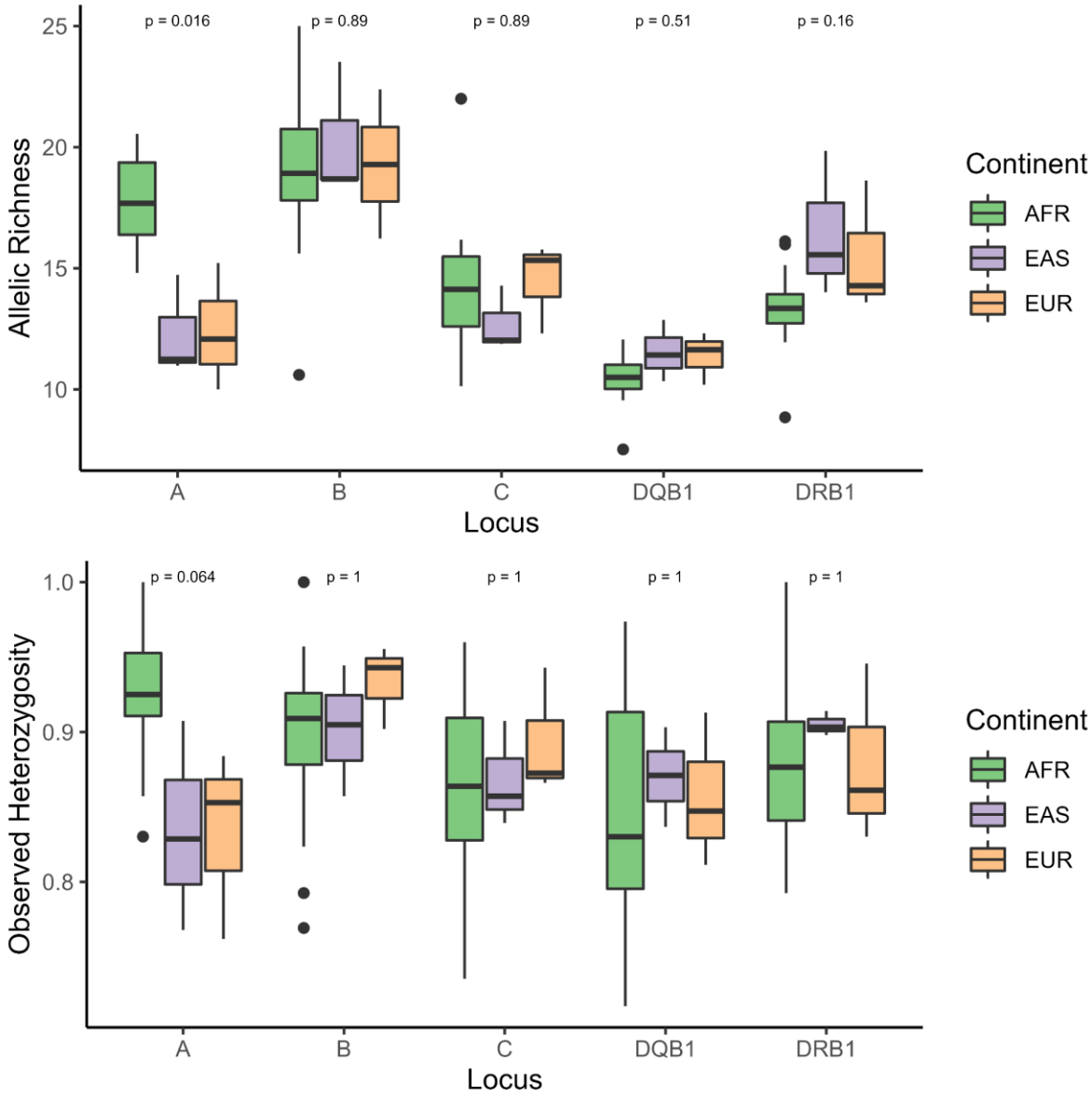
Supplementary Figure S9. Density plot of the proportion of correctly assigned language labels by DAPC in randomly sampled regions of the same number of base pairs (red) and the same number of SNPs (blue) as the Class I and Class II HLA genes. The dashed red and blue lines represent the 2.5 percentile of the same number of base pairs and same number of SNPs distributions, respectively. The solid black line represents the proportion of correctly assigned language labels by DAPC based on SNPs from the Class I and Class II HLA genes.



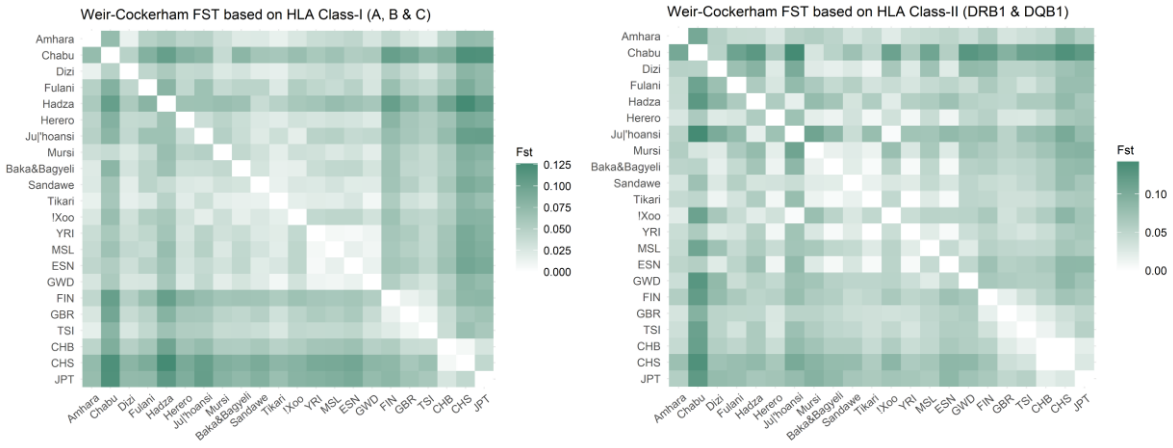
Supplementary Figure S10. Density plot of the proportion of correctly assigned continental region labels by DAPC in randomly sampled regions of the same number of base pairs (red) and the same number of SNPs (blue) as the Class I and Class II HLA genes. The dashed black line represents the proportion of correctly assigned continental region labels by DAPC based on SNPs from the Class I and Class II HLA genes.



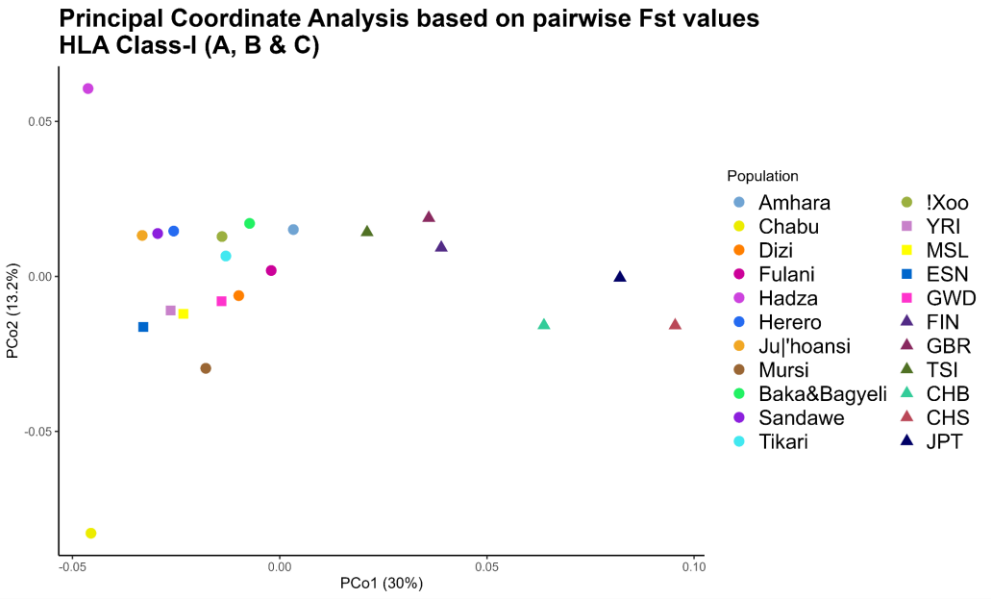
Supplementary Figure (11). PCA using SNPs obtained from the targeted HLA sequencing data. Dots represent individuals from 12 sub-Saharan African populations. Individuals cluster based on their HLA-DQ genotypes that correspond to distinct serotypes.

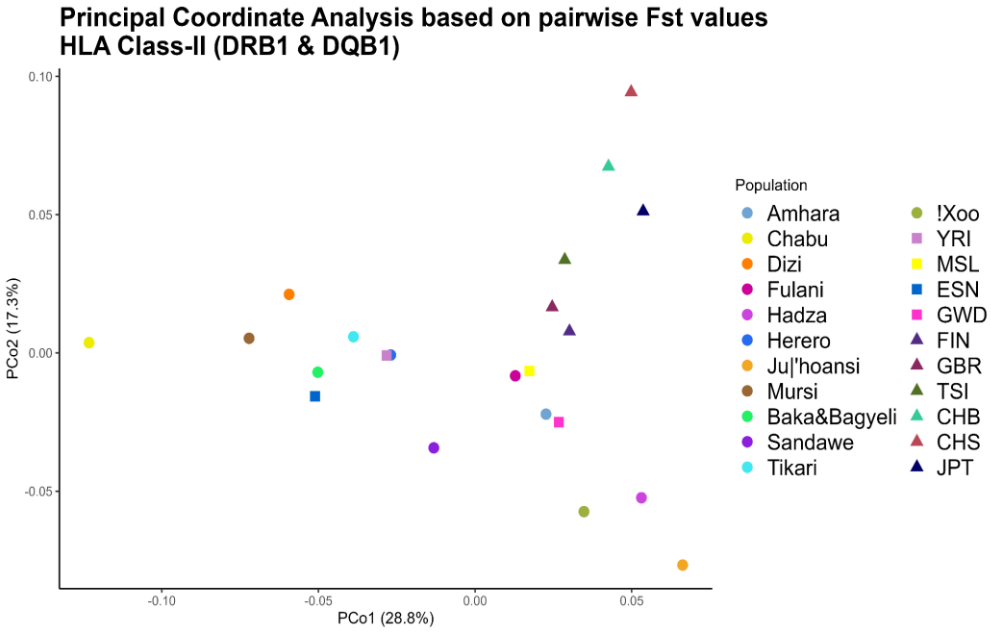


Supplementary Figure S12. Allelic richness and observed heterozygosity values across continents. p values denote Kruskal-Wallis-test between continents within each locus, including Bonferroni-adjustment for testing 5 loci. AFR: Africa, EAS: East Asia, EUR: Europe

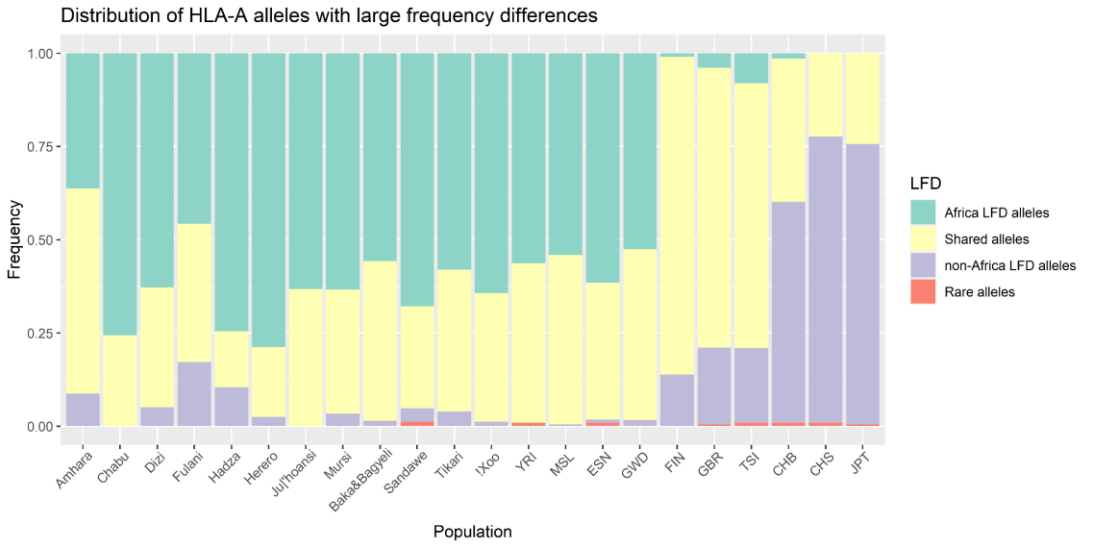


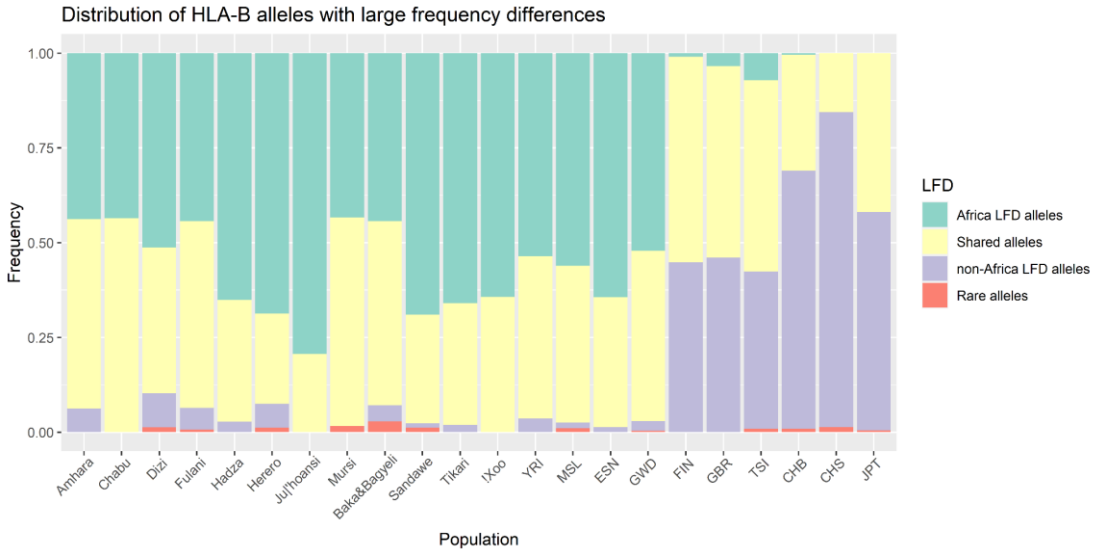
Supplementary Figure S13. Pairwise F_{ST} values based on targeted sequencing data of HLA class-I (HLA-A, HLA-B and HLA-C) and class-II (HLA-DRB1 and HLA-DQB1)

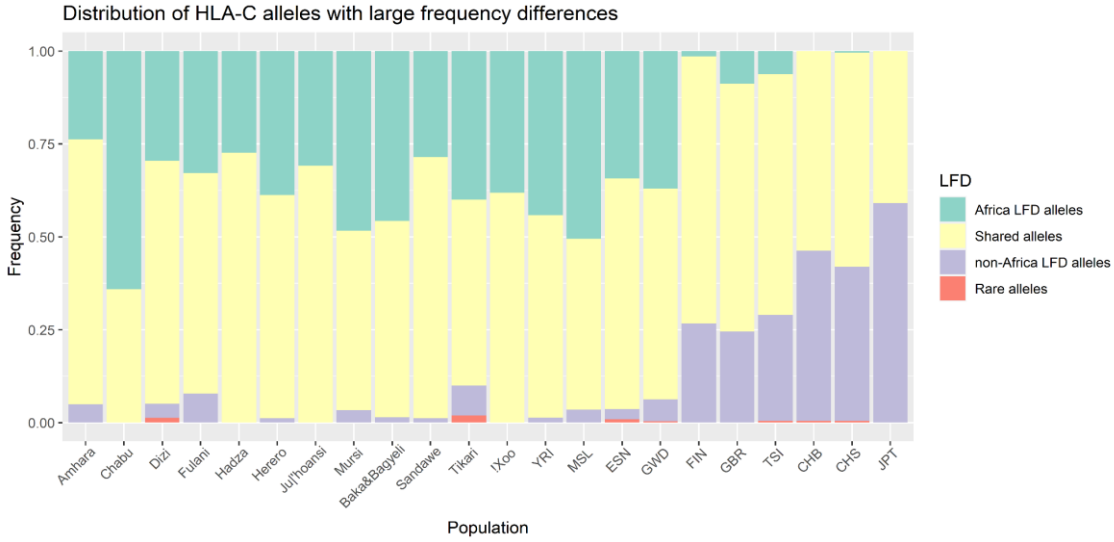


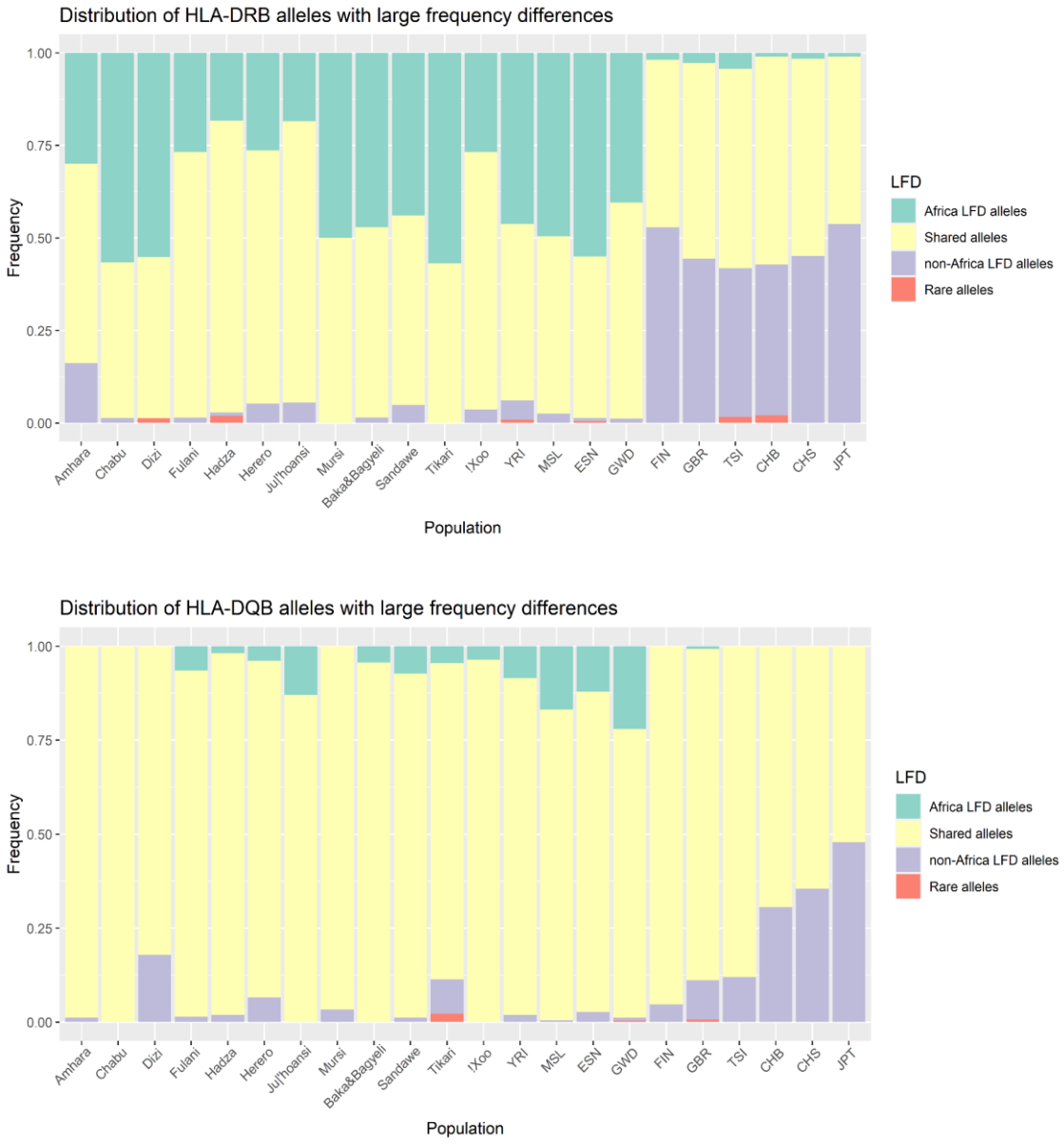


Supplementary Figure S14. Principle coordinate analysis of population differentiation. PCo1 and PCo2 were calculated from the pairwise Fst values (based on targeted HLA sequencing data) among the 12 novel African and 10 1000G populations.

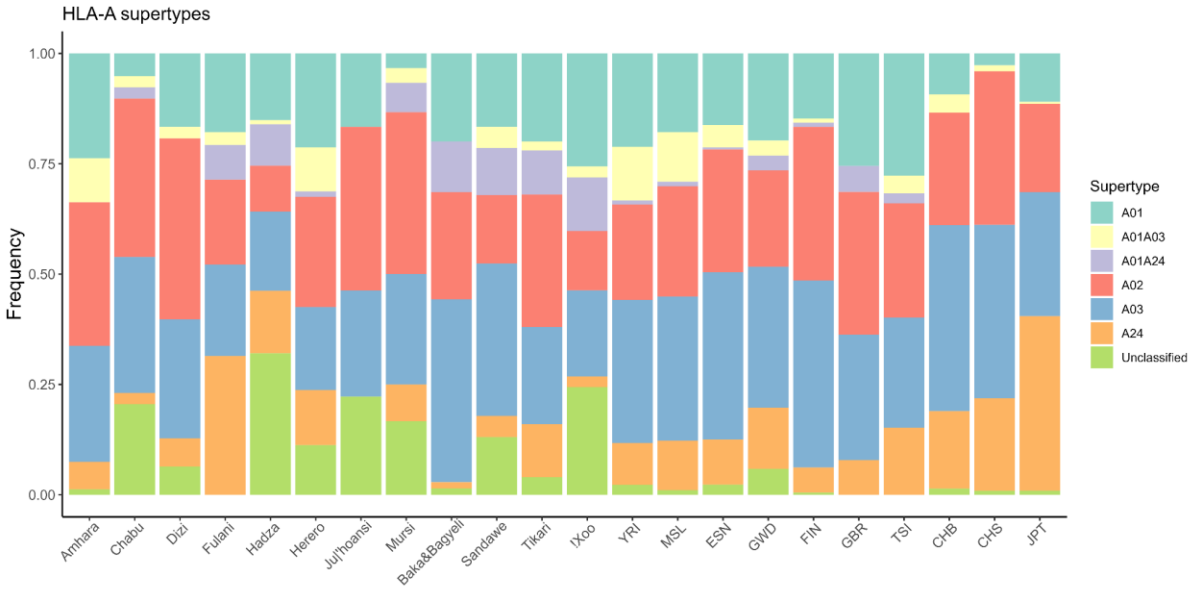

Distribution of HLA-A alleles with large frequency differences
Frequency
1.00
0.75
0.50
0.25
0.00
Amhara
Chabu
Dizi
Fulani
Hadza
Herero
Ju|'hoansi
Mursi
Baka&Bagyeli
Sandawe
Tikari
!Xoo
YRI
MSL
ESN
GWD
FIN
GBR
TSI
CHB
CHS
JPT
Population
LFD
Africa LFD alleles
Shared alleles
non-Africa LFD alleles
Rare alleles


Distribution of HLA-B alleles with large frequency differences
Frequency
1.00
0.75
0.50
0.25
0.00
Amhara
Chabu
Dizi
Fulani
Hadza
Herero
Ju|'hoansi
Mursi
Baka&Bagyeli
Sandawe
Tikari
!Xoo
YRI
MSL
ESN
GWD
FIN
GBR
TSI
CHB
CHS
JPT
Population
LFD
Africa LFD alleles
Shared alleles
non-Africa LFD alleles
Rare alleles

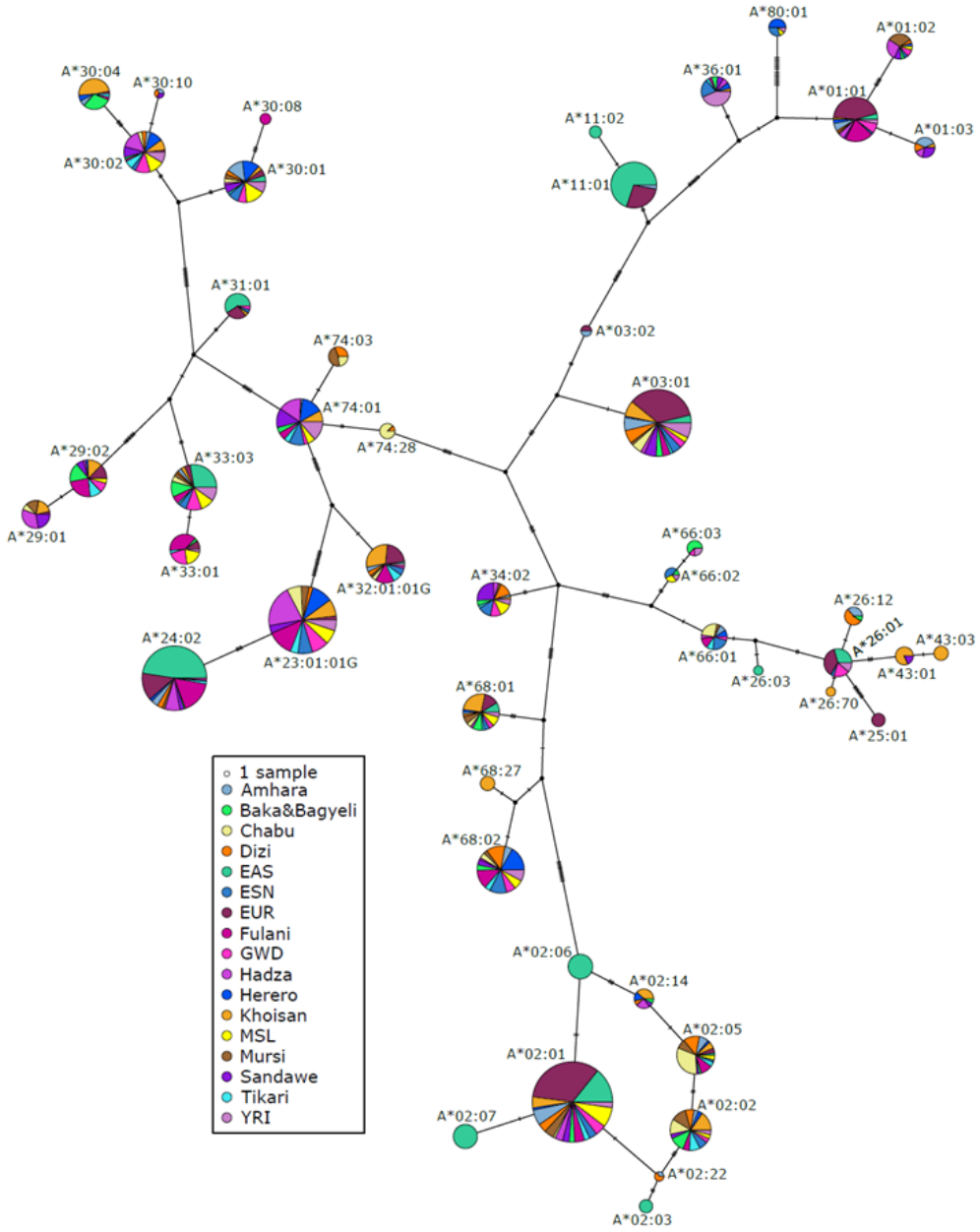

Distribution of HLA-C alleles with large frequency differences
Frequency
1.00
0.75
0.50
0.25
0.00
Amhara
Chabu
Dizi
Fulani
Hadza
Herero
Ju|'hoansi
Mursi
Baka&Bagyeli
Sandawe
Tikari
!Xoo
YRI
MSL
ESN
GWD
FIN
GBR
TSI
CHB
CHS
JPT
Population
LFD
Africa LFD alleles
Shared alleles
non-Africa LFD alleles
Rare alleles



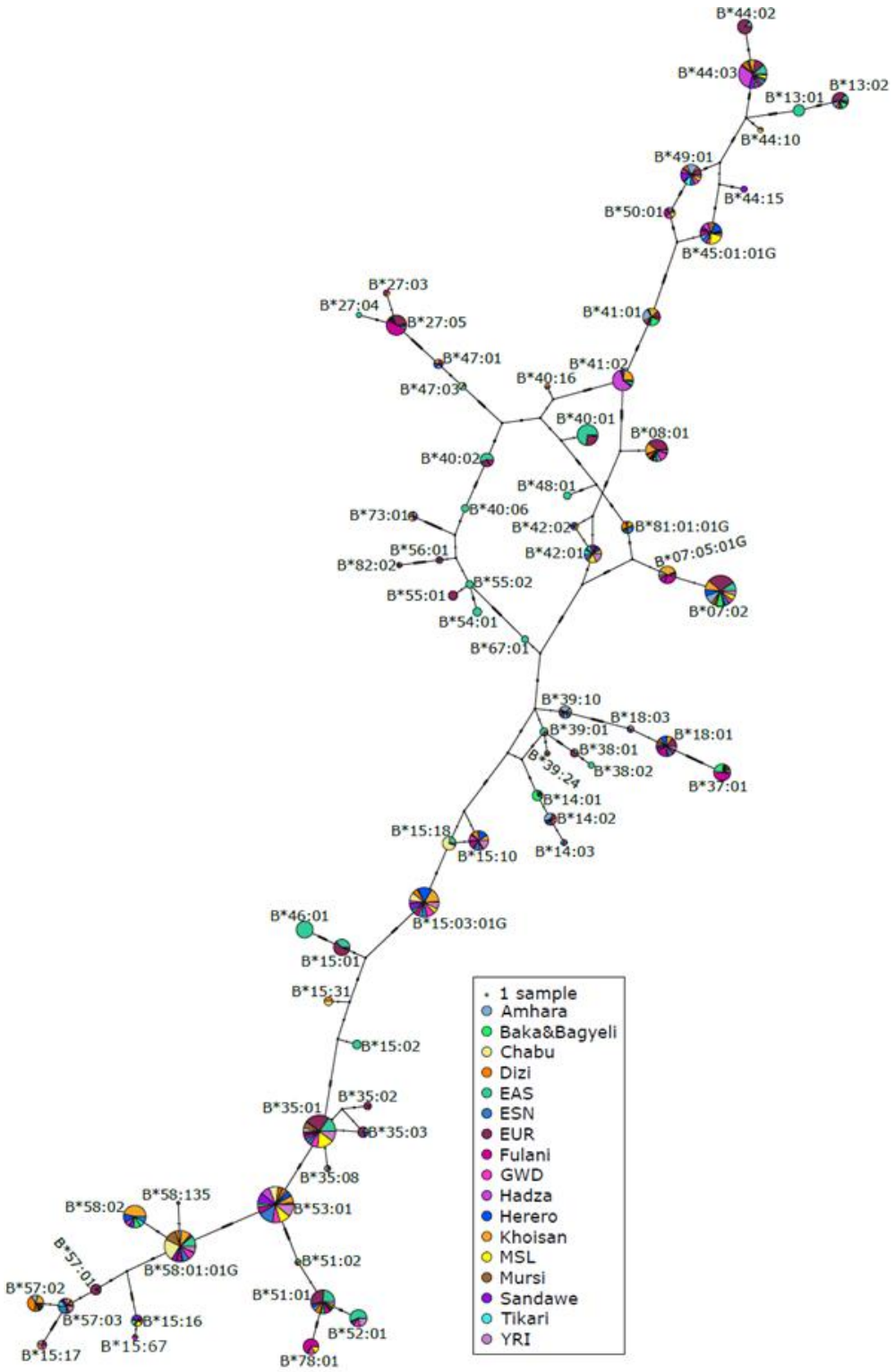
Supplementary Figure S15. Alleles with large frequency differences between African and non-African populations. Alleles with "large frequency difference" are defined (following Single et al 2020) as those having at least 3-fold difference in frequency between African and Eurasian populations. Due to the small number of populations and relatively small sample sizes, endemic alleles (i.e. alleles that are observed in Africa but not in Eurasia or vice versa) were also considered as alleles with large frequency differences. Shared alleles are those with less than 3-fold frequency differences. Rare alleles are those observed less than two times in both groups.



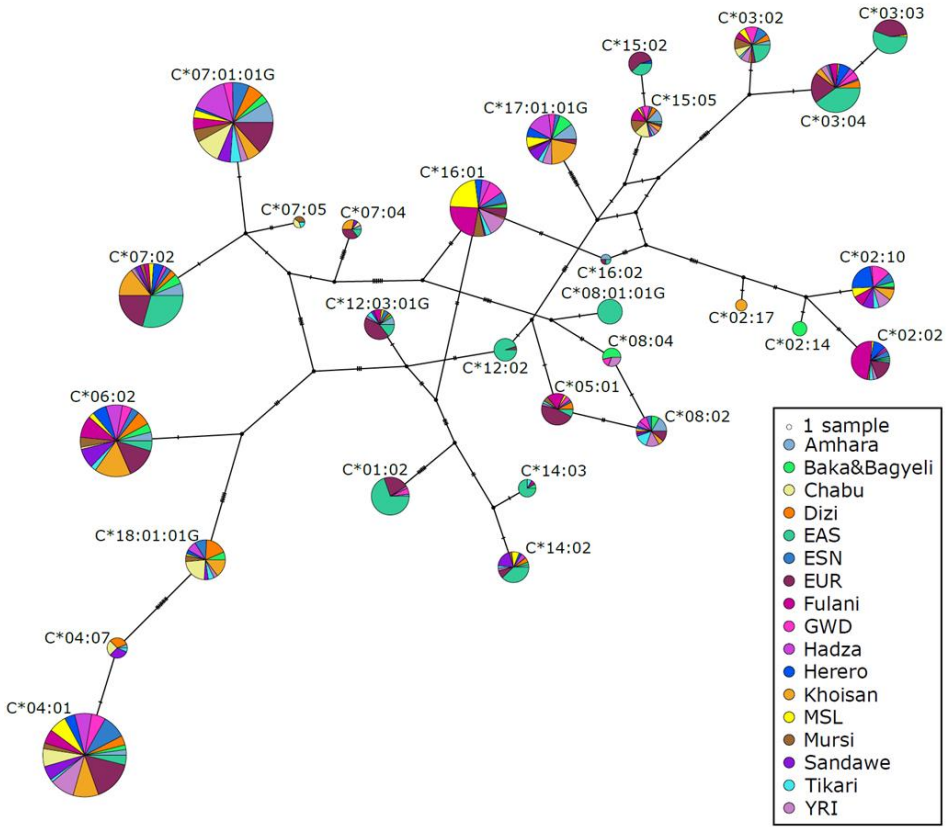
Supplementary Figure S16. Distribution of HLA-A supertypes within the 12 novel African and 10 1KG populations. Supertypes are defined as groups of HLA alleles with largely similar and overlapping peptide repertoires (Sidney et.al. 2009).



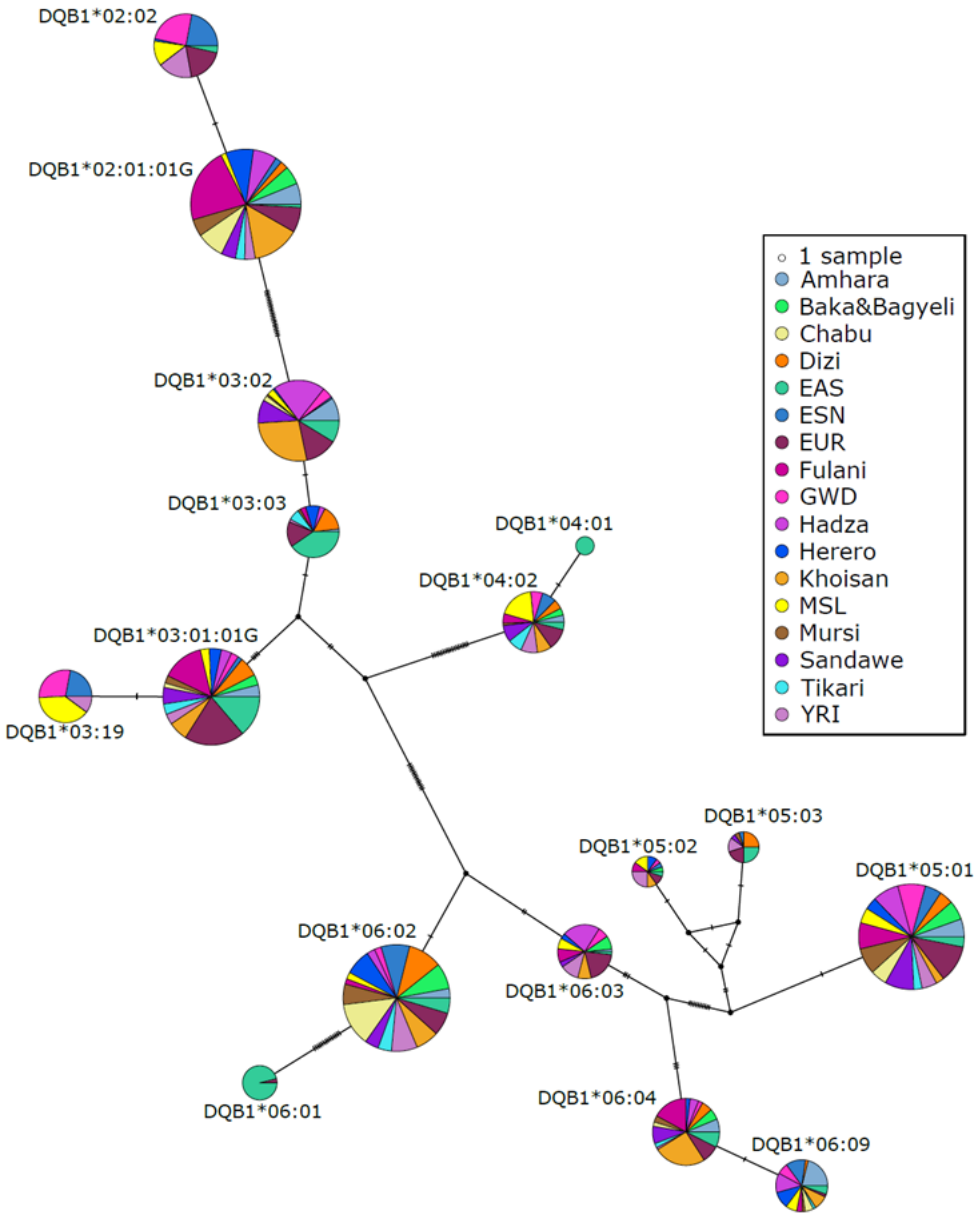
Supplementary Figure S17. Haplotype network of HLA-A alleles. Ju|'hoansi and !Xoo individuals were combined into the Khoisan group. All European and East Asian 1KG individuals were combined into the EUR and EAS groups, respectively.



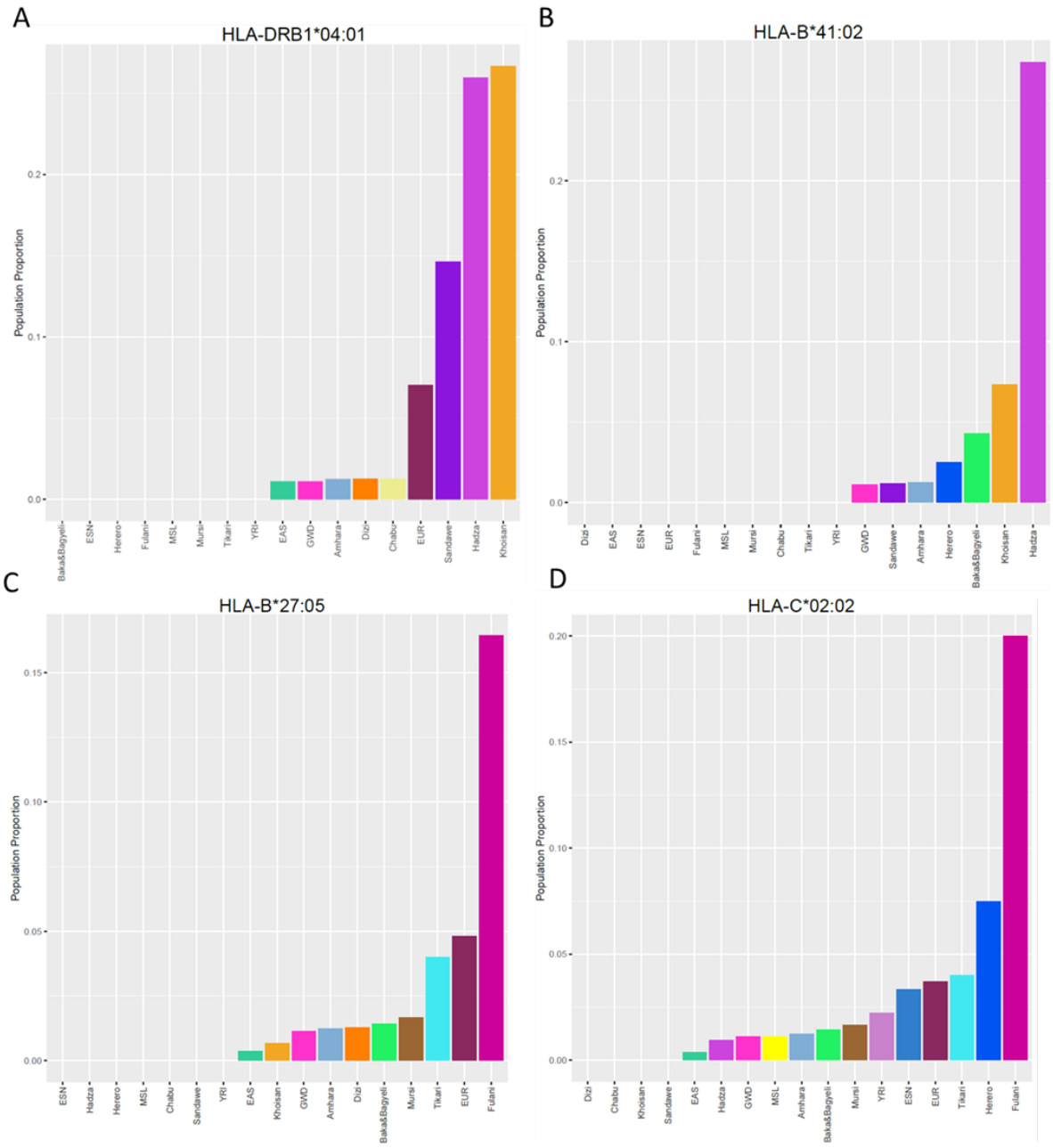
Supplementary Figure S18. Haplotype network of HLA-B alleles. Ju|'hoansi and !Xoo individuals were combined into the Khoisan group. All European and East Asian 1KG individuals were combined into the EUR and EAS groups, respectively.



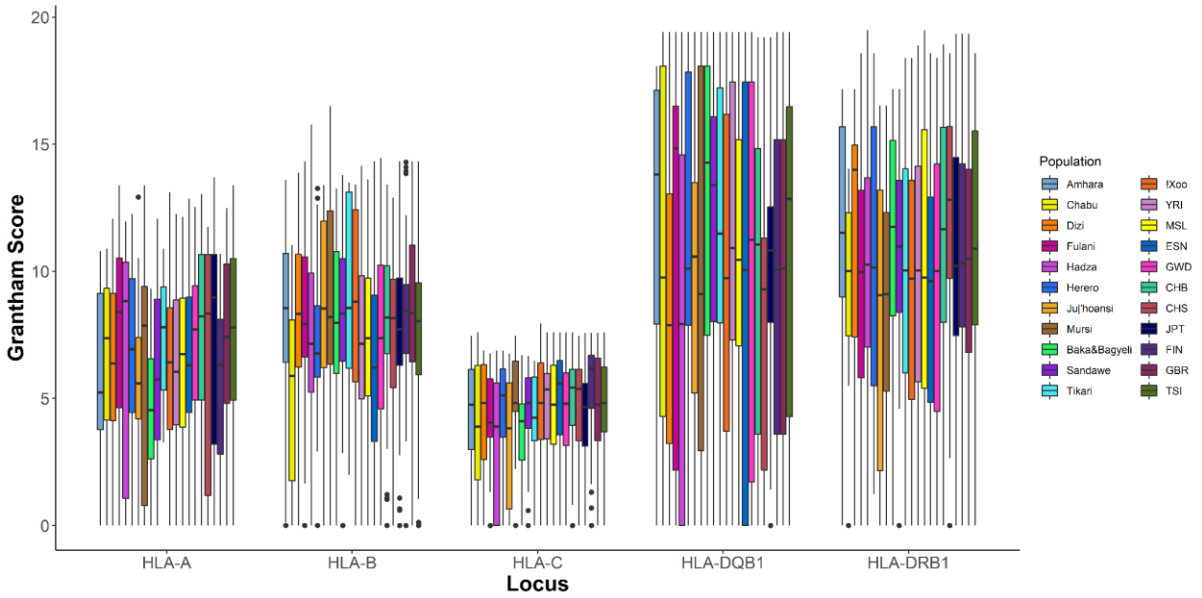
Supplementary Figure S19. Haplotype network of HLA-C alleles. Ju|'hoansi and !Xoo individuals were combined into the Khoisan group. All European and East Asian 1KG individuals were combined into the EUR and EAS groups, respectively.



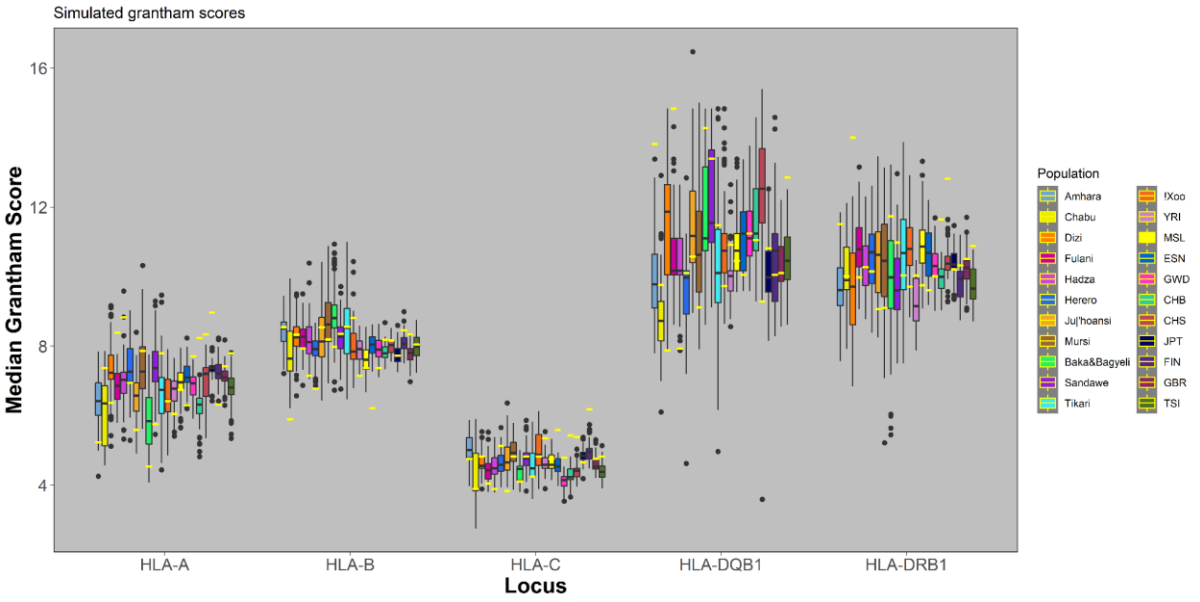
Supplementary Figure S20. Haplotype network of HLA-DQB1 alleles. Ju|'hoansi and !Xoo individuals were combined into the Khoisan group. All European and East Asian 1KG individuals were combined into the EUR and EAS groups, respectively.



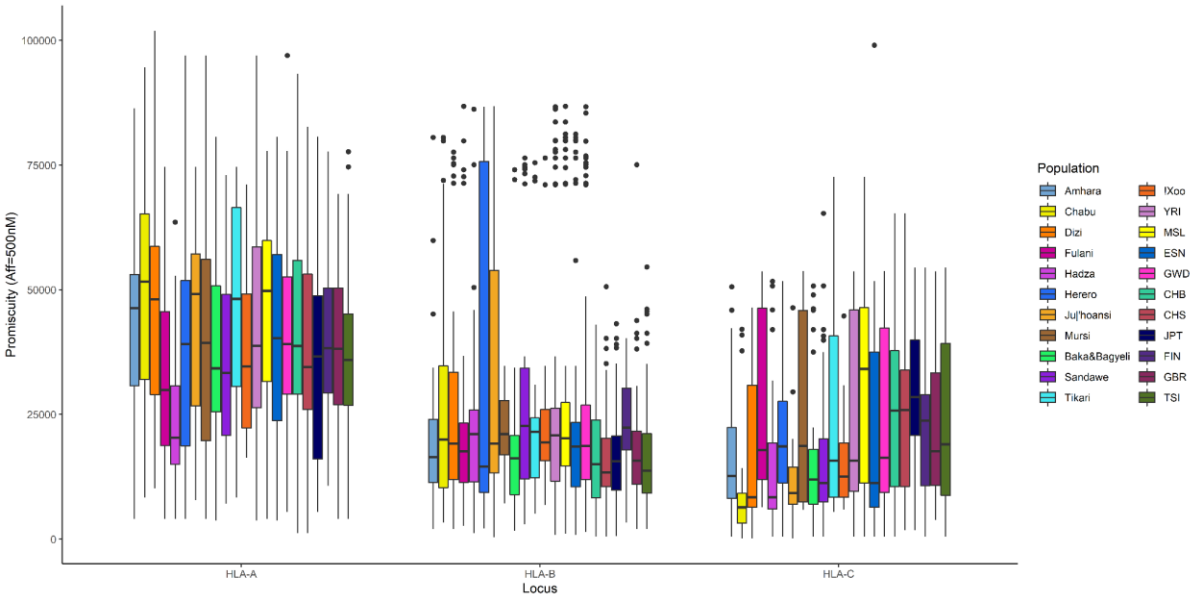
Supplementary Figure S21. HLA allele frequency in all populations included in the haplotype network. A) HLA-DRB1*04:01 is at high frequency in populations that speak Khoisan languages (Sandawe, Hadza, and Khoisan). B) HLA-B*41:02 is at high frequency in the Hadza. C) HLA-B*27:05 is at high frequency in the Fulani. D) HLA-C*02:02 is at high frequency in the Fulani. Ju|'hoansi and !Xoo individuals were combined into the Khoisan group.



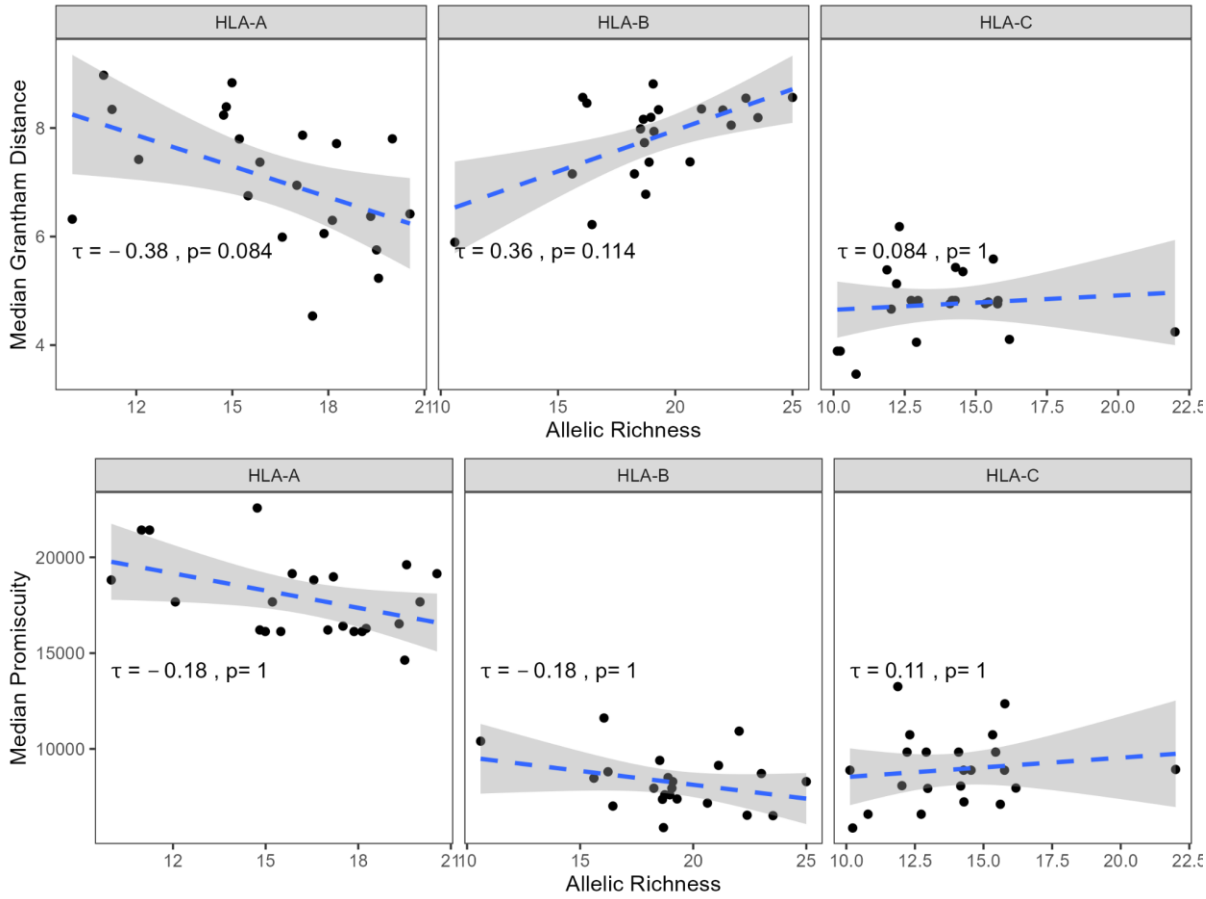
Supplementary Figure S22. Distribution of the evolutionary allele divergence of HLA genotypes (i.e. amino acid sequence divergence between two alleles of an individual) within populations.



Supplementary Figure S23. Observed (short yellow lines) and simulated (boxplots) individual HLA allele divergence (median grantham scores). Simulations were produced by assigning alleles randomly to each individual within populations. Grantham scores of randomly assigned genotypes were calculated and the median grantham score was recorded. Boxplots shows the distribution of median grantham scores of 1000 simulations for each population. When analyzing across all five loci, non-African populations were more likely to exhibit larger than expected allele divergence (chi-square test, p=0.006).



Supplementary Figure S24. Distribution of individual-level promiscuity values within populations. Promiscuity of each HLA allele was calculated as the number of peptides that are predicted to be bound among 1,000,000 random natural pathogen peptides. Individual-level promiscuity was calculated as the number of non-redundant peptides that are predicted to be bound by an individual's genotype.



Supplementary Figure S25. Correlation plot between median allele divergence and allelic richness (upper panels) and promiscuity and allelic richness (lower panels). Allele divergence was calculated as the Grantham distance between the pairs of alleles of an individual's HLA genotype and the promiscuity was calculated as the combined number of peptides bound by the pair of alleles an individual carries. The median divergence and promiscuity values were used to represent populations. p-values denote the Kendall's rank correlation test after multiple testing correction. Linear regression lines are shown in blue and 95% CI around the line in gray shade.

Supplementary Tables

Supplementary Table S1. Dataset summary. The sample size column includes the number of samples in each population with targeted HLA sequencing data and the number of samples with whole genome sequencing SNP data in parentheses. The region column describes the continental region of each sample as either sub-Saharan African (SSA), East Asian (EAS), or European (EUR). The subsistence column lists the primary subsistence pattern that the population follows. All 1KG project subsistence patterns are listed as Not Available (NA). The Language column lists the language family of each population's primary language.

Population	Sample Size	Region	Subsistence	Language
Amhara	40 (15)	SSA	Agriculturalist	Afroasiatic
Baka&Bagyeli	35 (15)	SSA	Hunter-gatherer	Niger Congo
Chabu	39 (15)	SSA	Hunter-gatherer	Nilo-Saharan
Dizi	39 (15)	SSA	Agropastoralist	Afroasiatic
Hadza	53 (15)	SSA	Hunter-gatherer	Khoesan
Herero	40 (15)	SSA	Pastoralist	Niger Congo
Ju|'hoansi	34 (15)	SSA	Hunter-gatherer	Khoesan
Fulani	70 (15)	SSA	Pastoralist	Niger Congo

Mursi	30 (15)	SSA	Pastoralist	Nilo-Saharan
Sandawe	42 (15)	SSA	Hunter-gatherer	Khoesan
Tikari	25 (15)	SSA	Agriculturalist	Niger Congo
!Xoo	42 (15)	SSA	Hunter-gatherer	Khoesan
GWD	120 (113)	SSA	NA	Niger Congo
ESN	111 (99)	SSA	NA	Niger Congo
MSL	98 (85)	SSA	NA	Niger Congo
YRI	111 (108)	SSA	NA	Niger Congo
CHB	108 (103)	EAS	NA	East Asian
CHS	112 (105)	EAS	NA	East Asian
JPT	105 (104)	EAS	NA	East Asian
FIN	105 (99)	EUR	NA	European
GBR	102 (91)	EUR	NA	European
TSI	112 (107)	EUR	NA	European

Supplementary Table S2. F_{ST} percentile for SNPs in HLA class I and class II genes. F_{ST} was calculated for each SNP on chromosome 6 and assigned a percentile. Percentiles were averaged for all SNPs contained within the listed genes. F_{ST} analysis contains 22 population total (12 from the novel African 180WGS dataset, 10 from the 1KG dataset). 1KG populations were downsampled to N=15/population, and data were LD-pruned.

Gene	N SNPs	Mean Percentile
HLA A	77	55.92442
HLA B	84	58.8716
HLA C	58	63.75039
HLA DPA1	116	68.32376
HLA DPB1	104	67.37026
HLA DQA1	94	55.52156
HLA DQB1	122	62.10997
HLA DRB1	146	54.85313

Supplementary Table S3. Summary statistics for all loci and populations. Samples sizes are the number of individuals within a population typed for the corresponding HLA loci. Number of alleles within each population was counted at the two-field resolution level. Allelic richness values were calculated as rarefied allele counts (based on the lowest sample size for each locus). Ewens-Watterson test results were reported as the normalized deviate of homozygosity (F_{nd}) and the associated p-values were calculated with Slatkin's exact test implementing a Markov-chain Monte Carlo method.

Population	Continent / Region	Locus	Sample Size	Number of Alleles	Allelic Richness	Observed Heterozygosity	Hardy-Weinberg Equilibrium	Ewens-Watterson Test (Fnd)	Ewens-Watterson Test p-value (non-adjusted)
Amhara	Africa	A	40	23	19.567	0.925	TRUE	-0.6333	0.2749
Amhara	Africa	B	40	28	23.014	0.9	TRUE	-0.743	0.2051
Amhara	Africa	C	40	19	15.765	0.825	TRUE	-0.3976	0.4153
Amhara	Africa	DPA1	40	7	5.939	0.475	TRUE	0.8083	0.8105
Amhara	Africa	DPB1	40	15	13.026	0.825	FALSE	-0.6347	0.2804
Amhara	Africa	DQA1	40	8	7.790	0.85	FALSE	-1.229	0.0364
Amhara	Africa	DQB1	40	11	9.893	0.925	TRUE	-1.3153	0.0135
Amhara	Africa	DRB1	40	15	13.268	0.9	TRUE	-0.8307	0.1687
Chabu	Africa	A	39	17	15.861	0.9231	TRUE	-1.0162	0.0776
Chabu	Africa	B	39	12	10.600	0.7692	FALSE	-0.4285	0.4032

Chabu	Africa	C	39	11	10.233	0.8974	TRUE	-0.9686	0.1156
Chabu	Africa	DPA1	39	5	4.826	0.7895	TRUE	-1.1538	0.0908
Chabu	Africa	DPB1	39	12	11.169	0.9211	TRUE	-0.8183	0.1838
Chabu	Africa	DQA1	39	7	6.826	0.7105	TRUE	-0.4451	0.4037
Chabu	Africa	DQB1	38	8	7.521	0.7632	TRUE	-0.4284	0.4097
Chabu	Africa	DRB1	39	10	8.846	0.8158	TRUE	-0.5176	0.353
Dizi	Africa	A	39	23	19.323	0.9231	TRUE	-0.8284	0.1703
Dizi	Africa	B	39	26	22.030	0.9231	TRUE	-1.0384	0.0721
Dizi	Africa	C	39	16	14.169	0.9231	TRUE	-0.9775	0.0985
Dizi	Africa	DPA1	39	8	7.337	0.6667	TRUE	-0.3341	0.455
Dizi	Africa	DPB1	39	18	15.278	0.9231	TRUE	-0.6573	0.2588
Dizi	Africa	DQA1	39	10	8.577	0.7949	TRUE	-0.5864	0.3197
Dizi	Africa	DQB1	39	12	10.265	0.8205	TRUE	-0.6281	0.2866
Dizi	Africa	DRB1	39	14	12.026	0.8718	TRUE	-0.4396	0.3921
Fulani	Africa	A	70	18	14.811	0.9143	TRUE	-1.2633	0.0177
Fulani	Africa	B	70	28	19.093	0.9571	TRUE	-0.6156	0.2848
Fulani	Africa	C	70	16	12.913	0.8571	TRUE	-0.8794	0.144

Fulani	Africa	DPA1	70	6	4.782	0.6232	TRUE	-0.2485	0.4846
Fulani	Africa	DPB1	69	17	12.698	0.913	TRUE	-0.6953	0.2401
Fulani	Africa	DQA1	70	9	8.612	0.8696	TRUE	-1.6022	0.0007
Fulani	Africa	DQB1	70	13	11.640	0.8571	TRUE	-0.9431	0.1249
Fulani	Africa	DRB1	70	17	13.398	0.8571	TRUE	-1.1039	0.0485
Hadza	Africa	A	53	19	14.989	0.8302	TRUE	-0.3752	0.4279
Hadza	Africa	B	53	23	15.609	0.7925	FALSE	1.8081	0.9453
Hadza	Africa	C	53	13	10.133	0.7358	TRUE	-0.2652	0.4852
Hadza	Africa	DPA1	53	11	9.121	0.6604	TRUE	0.2755	0.7003
Hadza	Africa	DPB1	53	16	12.972	0.8302	TRUE	-1.2225	0.0244
Hadza	Africa	DQA1	53	9	7.899	0.6604	FALSE	-0.53	0.3512
Hadza	Africa	DQB1	53	12	10.279	0.717	FALSE	-0.9508	0.1179
Hadza	Africa	DRB1	53	17	13.381	0.7925	TRUE	-0.5524	0.3262
Herero	Africa	A	40	20	17.015	0.925	FALSE	-0.7684	0.1951
Herero	Africa	B	40	23	18.741	0.925	TRUE	-0.376	0.414
Herero	Africa	C	40	14	12.214	0.9	TRUE	-1.0243	0.0783
Herero	Africa	DPA1	40	9	8.386	0.7368	TRUE	-1.1615	0.0527

Herero	Africa	DPB1	40	14	11.836	0.8421	TRUE	0.0096	0.6175
Herero	Africa	DQA1	40	8	7.693	0.8947	TRUE	-1.0445	0.0975
Herero	Africa	DQB1	38	12	11.071	0.9737	TRUE	-1.0255	0.0851
Herero	Africa	DRB1	40	14	12.664	1	TRUE	-1.2343	0.0185
Ju|'hoansi	Africa	A	34	18	16.561	0.9706	TRUE	-1.4153	0.0046
Ju|'hoansi	Africa	B	34	18	16.055	0.8235	TRUE	-0.2046	0.5201
Ju|'hoansi	Africa	C	34	12	10.790	0.7353	FALSE	-0.3731	0.4313
Ju|'hoansi	Africa	DPA1	34	8	7.464	0.7941	TRUE	-0.6336	0.3062
Ju|'hoansi	Africa	DPB1	34	15	12.786	0.8529	TRUE	-0.2033	0.5159
Ju|'hoansi	Africa	DQA1	34	8	7.731	0.7059	TRUE	-0.0758	0.5701
Ju|'hoansi	Africa	DQB1	34	11	10.088	0.8235	TRUE	-0.5804	0.3185
Ju|'hoansi	Africa	DRB1	34	15	13.304	0.8235	FALSE	0.0404	0.6278
Mursi	Africa	A	30	18	17.195	0.9	TRUE	-1.1182	0.0509
Mursi	Africa	B	30	21	18.961	0.9	TRUE	0.5305	0.7852
Mursi	Africa	C	30	15	14.275	0.9333	TRUE	-1.1461	0.0423
Mursi	Africa	DPA1	30	8	7.420	0.5667	FALSE	-0.5373	0.3546
Mursi	Africa	DPB1	30	17	14.876	0.8333	TRUE	-0.4529	0.3863

Mursi	Africa	DQA1	30	12	10.777	0.7667	TRUE	-0.1307	0.5547
Mursi	Africa	DQB1	30	11	9.812	0.7667	TRUE	-0.1496	0.5411
Mursi	Africa	DRB1	30	13	11.949	0.8	TRUE	-0.6237	0.2895
Baka & Bagyeli	Africa	A	35	20	17.506	0.8571	FALSE	-0.938	0.1102
Baka & Bagyeli	Africa	B	35	21	18.521	0.9143	TRUE	-1.0403	0.0665
Baka & Bagyeli	Africa	C	35	18	16.180	0.8286	TRUE	-1.2304	0.0222
Baka & Bagyeli	Africa	DPA1	35	7	6.756	0.6286	TRUE	0.1205	0.6365
Baka & Bagyeli	Africa	DPB1	35	10	8.326	0.4857	TRUE	2.7398	0.9762
Baka & Bagyeli	Africa	DQA1	33	10	9.250	0.8571	TRUE	-0.5904	0.3152
Baka & Bagyeli	Africa	DQB1	35	10	9.553	0.9143	TRUE	-1.0179	0.1013
Baka & Bagyeli	Africa	DRB1	35	15	13.758	0.9143	TRUE	-0.9036	0.1286
Sandawe	Africa	A	42	23	19.505	0.9524	TRUE	-1.1173	0.049
Sandawe	Africa	B	42	27	21.120	0.9286	TRUE	-0.2207	0.5053
Sandawe	Africa	C	42	16	12.968	0.9286	TRUE	-0.5918	0.3016

Sandawe	Africa	DPA1	42	8	7.288	0.7143	TRUE	-0.6527	0.2957
Sandawe	Africa	DPB1	42	16	13.606	0.8333	TRUE	-1.1431	0.0441
Sandawe	Africa	DQA1	42	9	8.298	0.8571	TRUE	-0.8456	0.186
Sandawe	Africa	DQB1	42	12	10.955	0.9524	TRUE	-1.1192	0.053
Sandawe	Africa	DRB1	42	15	12.750	0.9048	TRUE	-0.8742	0.1506
Tikari	Africa	A	25	20	20.000	1	TRUE	-1.2551	0.0177
Tikari	Africa	B	25	25	25.000	1	TRUE	-1.2002	0.0298
Tikari	Africa	C	25	22	22.000	0.96	TRUE	-0.7772	0.206
Tikari	Africa	DPA1	25	5	5.000	0.8333	TRUE	-1.05	0.121
Tikari	Africa	DPB1	24	13	13.000	0.8696	TRUE	-0.3621	0.4378
Tikari	Africa	DQA1	25	10	10.000	0.913	TRUE	-0.576	0.315
Tikari	Africa	DQB1	23	11	11.000	0.913	TRUE	-1.2116	0.0283
Tikari	Africa	DRB1	25	14	14.000	0.913	TRUE	-0.9956	0.0852
!Xoo	Africa	A	42	24	20.549	0.881	FALSE	-1.231	0.0251
!Xoo	Africa	B	42	22	19.057	0.9524	FALSE	-1.3297	0.011
!Xoo	Africa	C	42	14	12.729	0.9048	FALSE	-1.062	0.0632
!Xoo	Africa	DPA1	42	9	8.290	0.8293	TRUE	-1.0882	0.0796

!Xoo	Africa	DPB1	42	13	10.479	0.7805	TRUE	0.3415	0.7261
!Xoo	Africa	DQA1	42	10	8.609	0.6585	TRUE	0.07	0.6299
!Xoo	Africa	DQB1	41	14	12.061	0.8049	TRUE	-0.7363	0.2195
!Xoo	Africa	DRB1	41	16	13.270	0.878	TRUE	-0.6678	0.259
YRI	Africa	A	111	27	17.864	0.9459	TRUE	-1.2499	0.0199
YRI	Africa	B	111	28	18.259	0.9099	TRUE	-1.1221	0.0433
YRI	Africa	C	111	21	14.546	0.8288	TRUE	-0.4432	0.3855
YRI	Africa	DQB1	106	15	11.481	0.8585	TRUE	-1.2922	0.0144
YRI	Africa	DRB1	111	23	15.984	0.9151	TRUE	-1.3001	0.0124
MSL	Africa	A	98	21	15.491	0.949	FALSE	-1.2491	0.0178
MSL	Africa	B	98	30	18.886	0.9082	TRUE	-0.6606	0.2552
MSL	Africa	C	98	20	14.094	0.8469	TRUE	-0.5801	0.3099
MSL	Africa	DQB1	98	12	10.798	0.8367	TRUE	-1.519	0.0009
MSL	Africa	DRB1	98	20	16.115	0.8878	TRUE	-1.3956	0.0049
ESN	Africa	A	111	29	18.125	0.9537	TRUE	-1.2461	0.0192
ESN	Africa	B	108	27	16.449	0.8796	TRUE	-0.1503	0.5449
ESN	Africa	C	111	26	15.613	0.8704	TRUE	0.5101	0.7838

ESN	Africa	DQB1	111	13	10.064	0.7477	FALSE	-0.9164	0.1407
ESN	Africa	DRB1	111	20	13.903	0.8468	TRUE	-1.0524	0.0675
GWD	Africa	A	120	25	18.256	0.9664	TRUE	-1.4729	0.0027
GWD	Africa	B	119	35	20.630	0.8739	FALSE	-0.9825	0.0986
GWD	Africa	C	120	22	15.442	0.8151	FALSE	-1.2318	0.0219
GWD	Africa	DQB1	120	14	10.712	0.8167	TRUE	-0.8637	0.1579
GWD	Africa	DRB1	120	23	15.129	0.875	TRUE	-1.0771	0.059
CHS	East Asia	A	112	17	11.242	0.7679	FALSE	-0.489	0.3674
CHS	East Asia	B	112	35	18.641	0.8571	TRUE	1.0692	0.8812
CHS	East Asia	C	112	18	11.877	0.8393	TRUE	-0.571	0.3163
CHS	East Asia	DQB1	31	11	10.339	0.871	TRUE	-0.8281	0.1816
CHS	East Asia	DRB1	112	25	14.013	0.9032	TRUE	-0.8012	0.1843
CHB	East Asia	A	108	21	14.728	0.9074	FALSE	-0.6827	0.2447
CHB	East Asia	B	108	41	23.527	0.9444	TRUE	-0.8105	0.1774

CHB	East Asia	C	108	21	14.287	0.9074	TRUE	-1.0299	0.0717
CHB	East Asia	DQB1	49	14	12.866	0.8367	FALSE	-1.1476	0.0388
CHB	East Asia	DRB1	108	31	19.845	0.898	TRUE	-1.1004	0.0519
JPT	East Asia	A	105	17	10.979	0.8286	TRUE	0.1342	0.6634
JPT	East Asia	B	105	28	18.685	0.9048	TRUE	-1.2281	0.0207
JPT	East Asia	C	105	16	12.027	0.8571	TRUE	-1.2889	0.0152
JPT	East Asia	DQB1	93	14	11.417	0.9032	TRUE	-1.3633	0.0049
JPT	East Asia	DRB1	105	22	15.560	0.914	TRUE	-1.2615	0.0182
FIN	Europe	A	105	15	10.001	0.7619	TRUE	0.0613	0.6376
FIN	Europe	B	105	23	16.232	0.9429	TRUE	-1.0393	0.0717
FIN	Europe	C	105	15	12.313	0.9429	TRUE	-1.4003	0.0058
FIN	Europe	DQB1	53	11	10.193	0.8113	TRUE	-1.2584	0.0224
FIN	Europe	DRB1	105	21	13.592	0.8302	TRUE	-0.8202	0.1663
GBR	Europe	A	102	18	12.075	0.8529	TRUE	-0.3403	0.4508

GBR	Europe	B	102	28	19.287	0.902	TRUE	-1.1975	0.0273
GBR	Europe	C	102	21	15.332	0.8725	TRUE	-1.3007	0.0106
GBR	Europe	DQB1	72	14	11.640	0.8472	TRUE	-1.3534	0.0054
GBR	Europe	DRB1	102	23	14.284	0.8611	TRUE	-0.9792	0.097
TSI	Europe	A	112	25	15.220	0.8839	TRUE	-0.2013	0.5138
TSI	Europe	B	112	36	22.384	0.9554	TRUE	-1.2729	0.0213
TSI	Europe	C	112	25	15.778	0.8661	TRUE	-0.7622	0.2051
TSI	Europe	DQB1	92	14	12.313	0.913	TRUE	-1.2815	0.013
TSI	Europe	DRB1	112	31	18.624	0.9457	TRUE	-0.7172	0.2221

Annex V

Supplementary Material for Chapter III

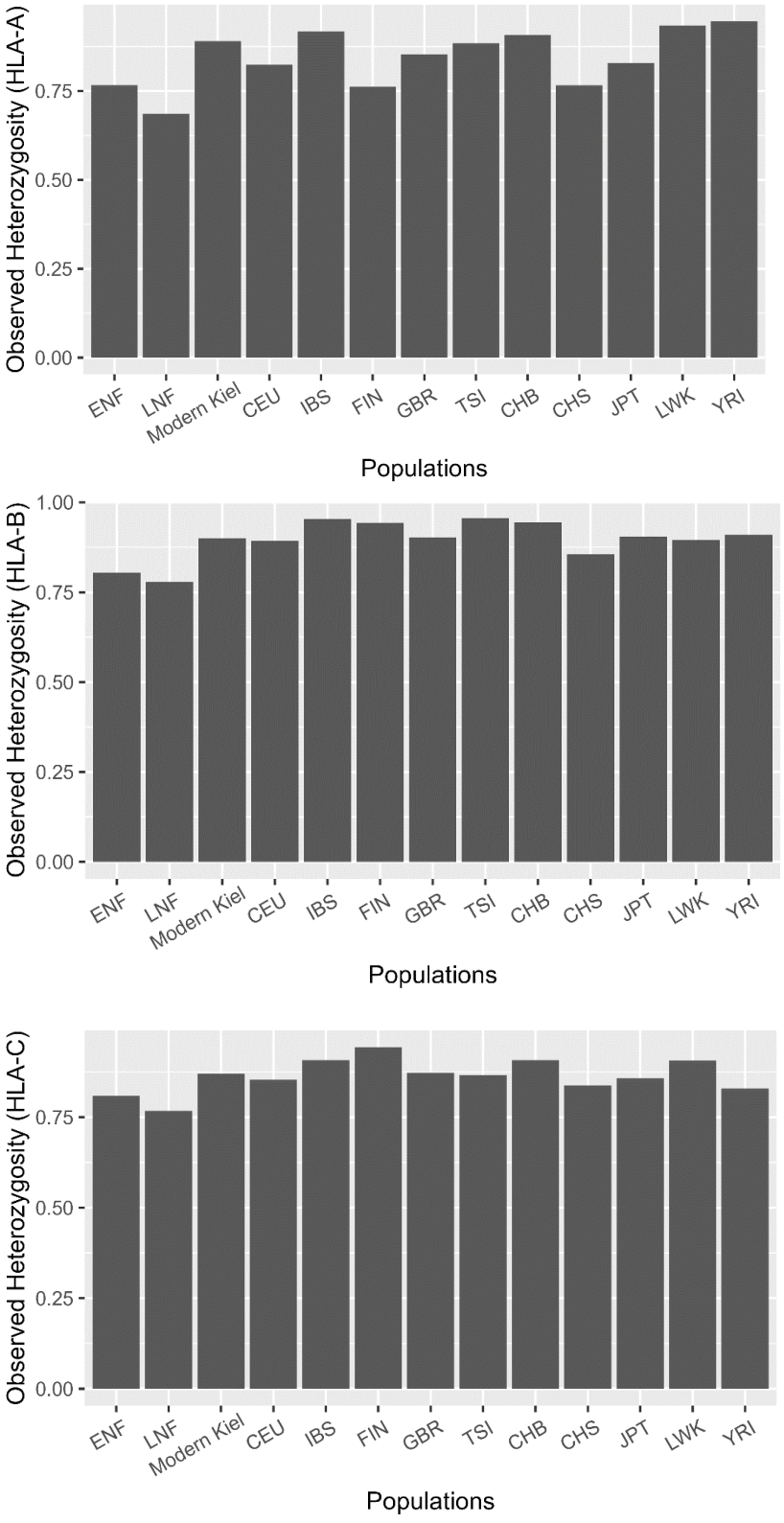
Spatio-temporal analysis of ancient HLA data reveals major effects of demography and admixture on immune gene diversity

Onur Özer[1], Yan-Rong Chen[1], Nicolas Antonio da Silva[2], Magdalena Haller[2], Sébastien Calvignac-Spencer[3], Almut Nebel[2], Ben Krause-Kyora[2], Tobias L. Lenz[1]

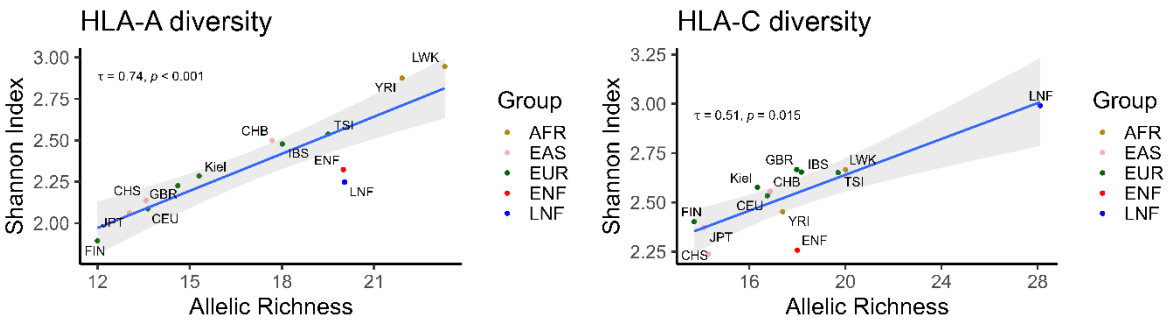
[1] Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany

[2] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

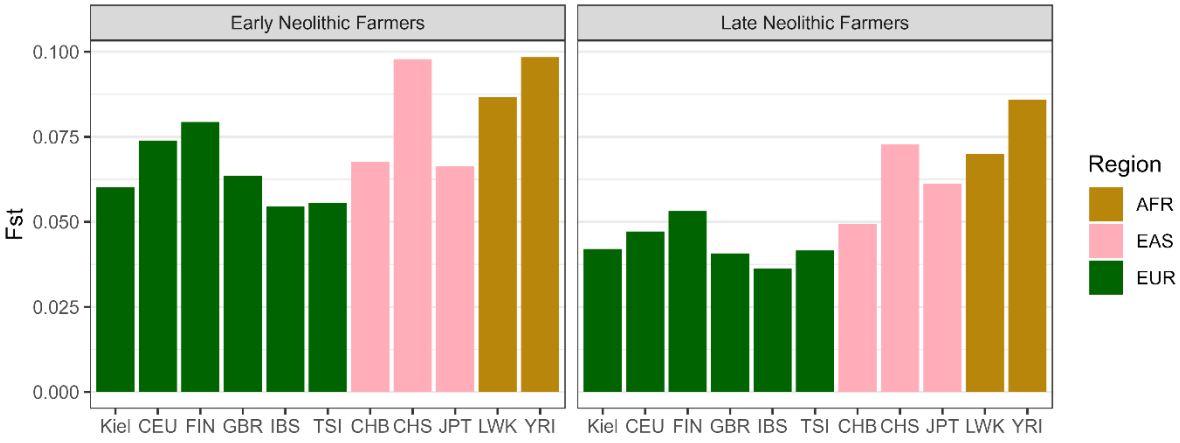
[3] Robert Koch Institute, Berlin, Germany



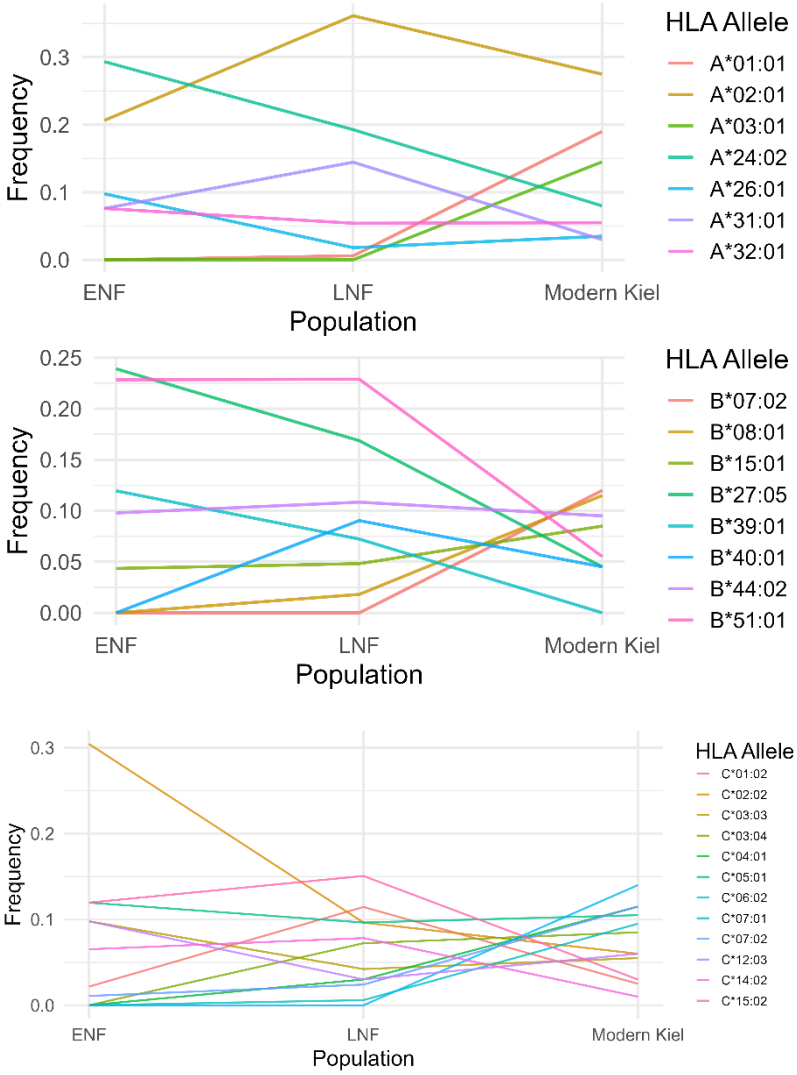
Supplementary Figure 1. Observed heterozygosity. Heterozygosity is calculated as the proportion of heterozygous individuals within each population.



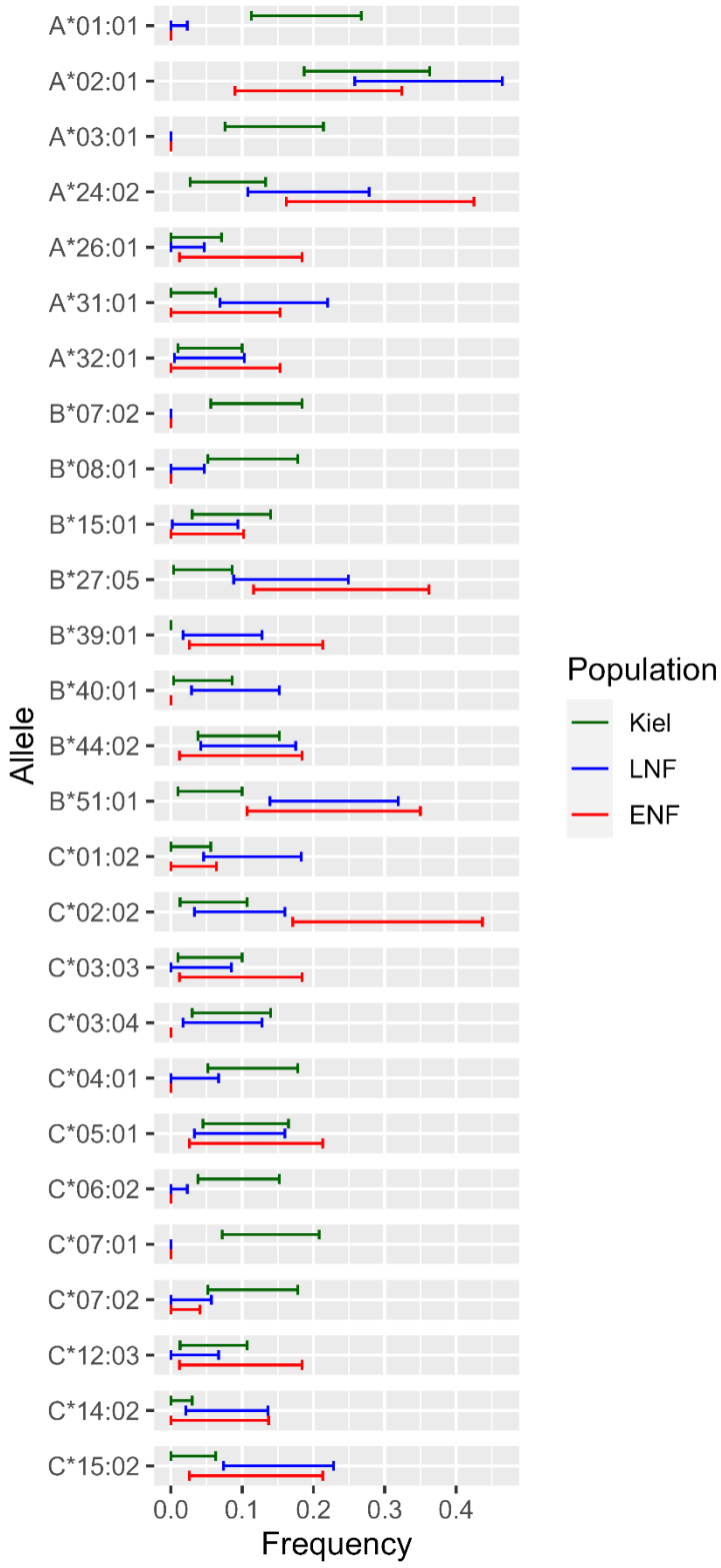
Supplementary Figure 2. HLA diversity. Tau (τ) refers to the Kendall rank correlation coefficient. The linear regression line is shown in blue and the 95% CI around the line is in gray.



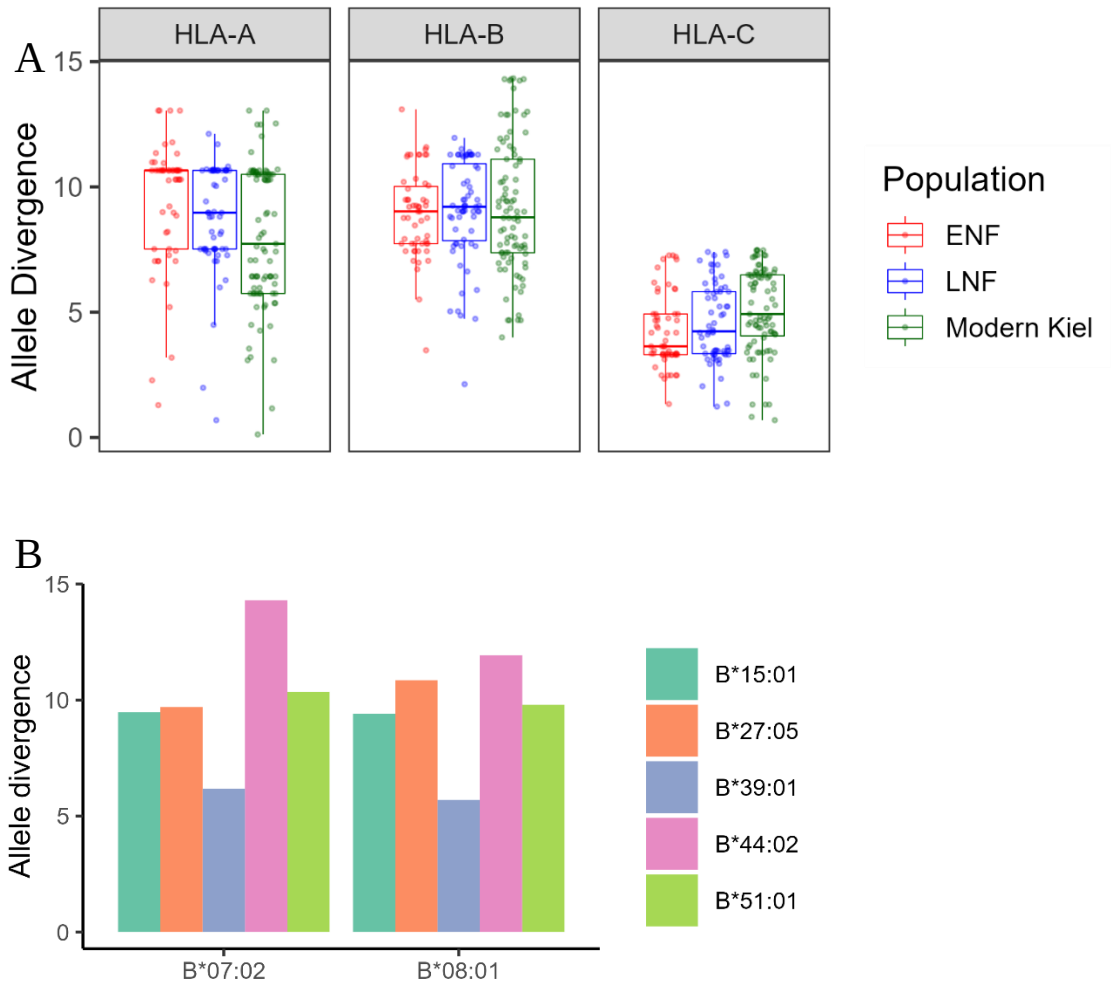
Supplementary Figure 3. Population differentiation based on Weir-Cockerham F_{ST}.



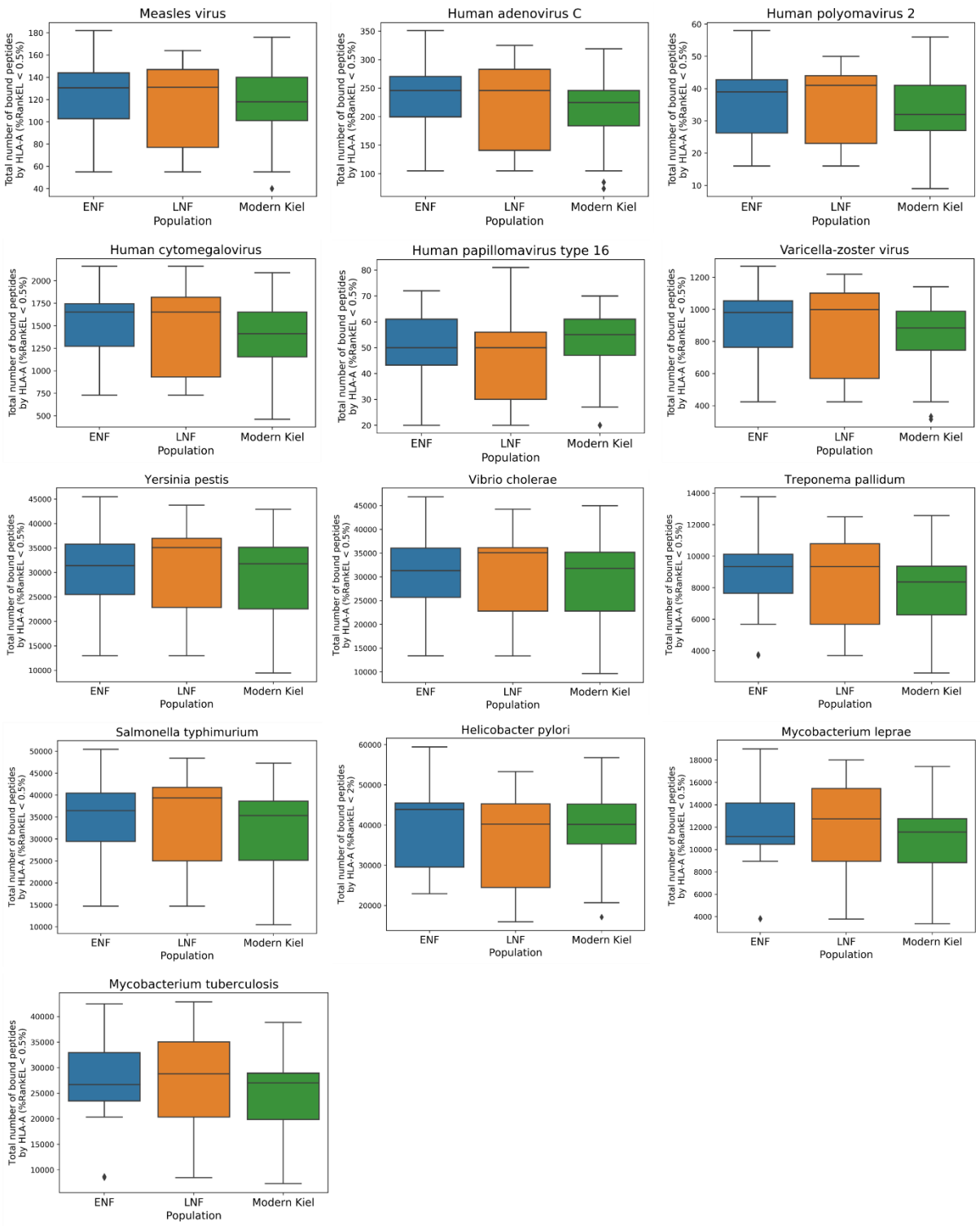
Supplementary Figure 5. Allele frequency changes since the early Neolithic. Only alleles with higher than 5% frequency in at least one population were plotted. Note that the distances across the X-axis do not correspond to the time difference between populations.



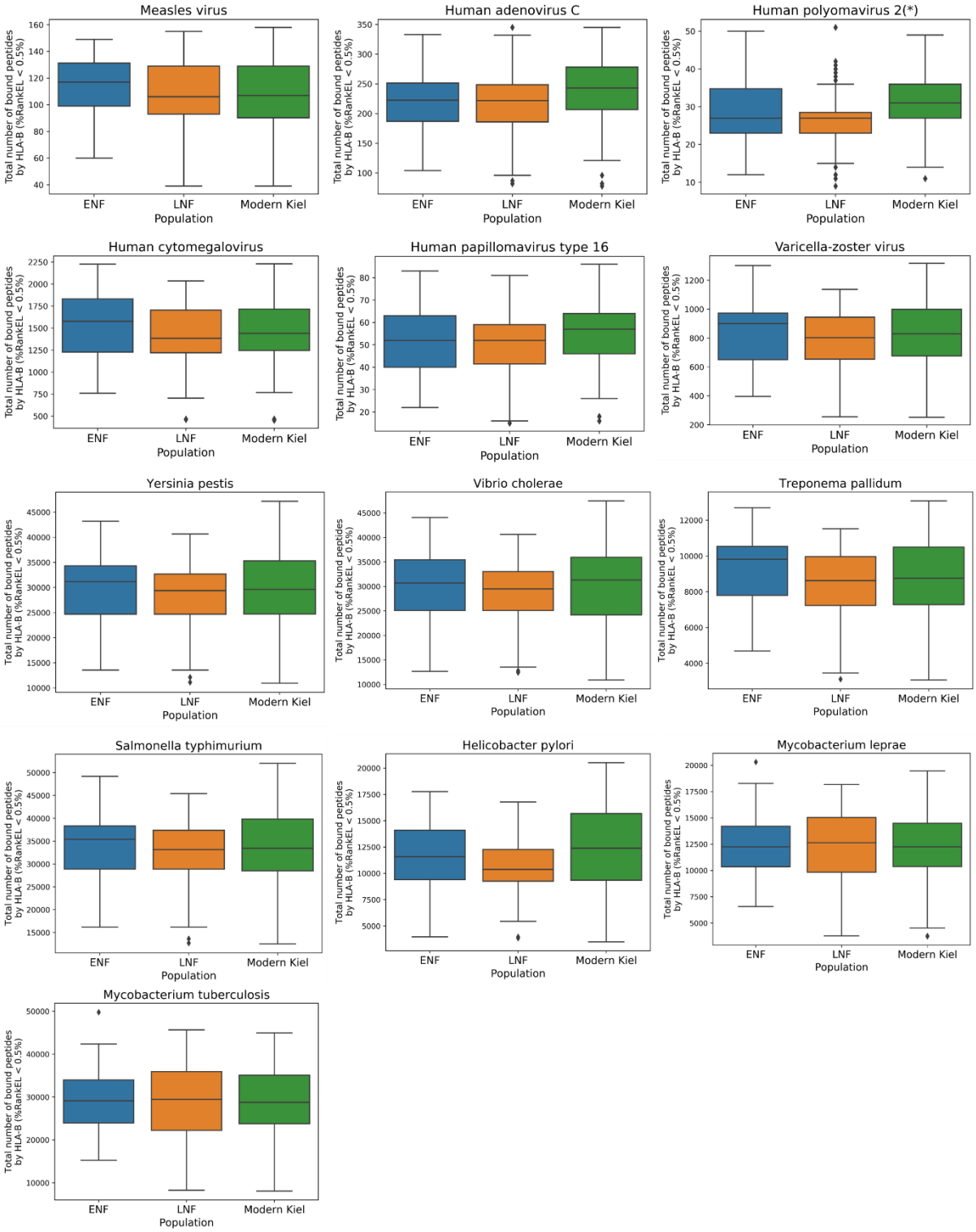
Supplementary Figure 6. Allele frequency differences between Neolithic farmers and modern Kiel population. Each line represents the upper and lower estimates of the 95% confidence interval of the frequency of the corresponding allele within populations.



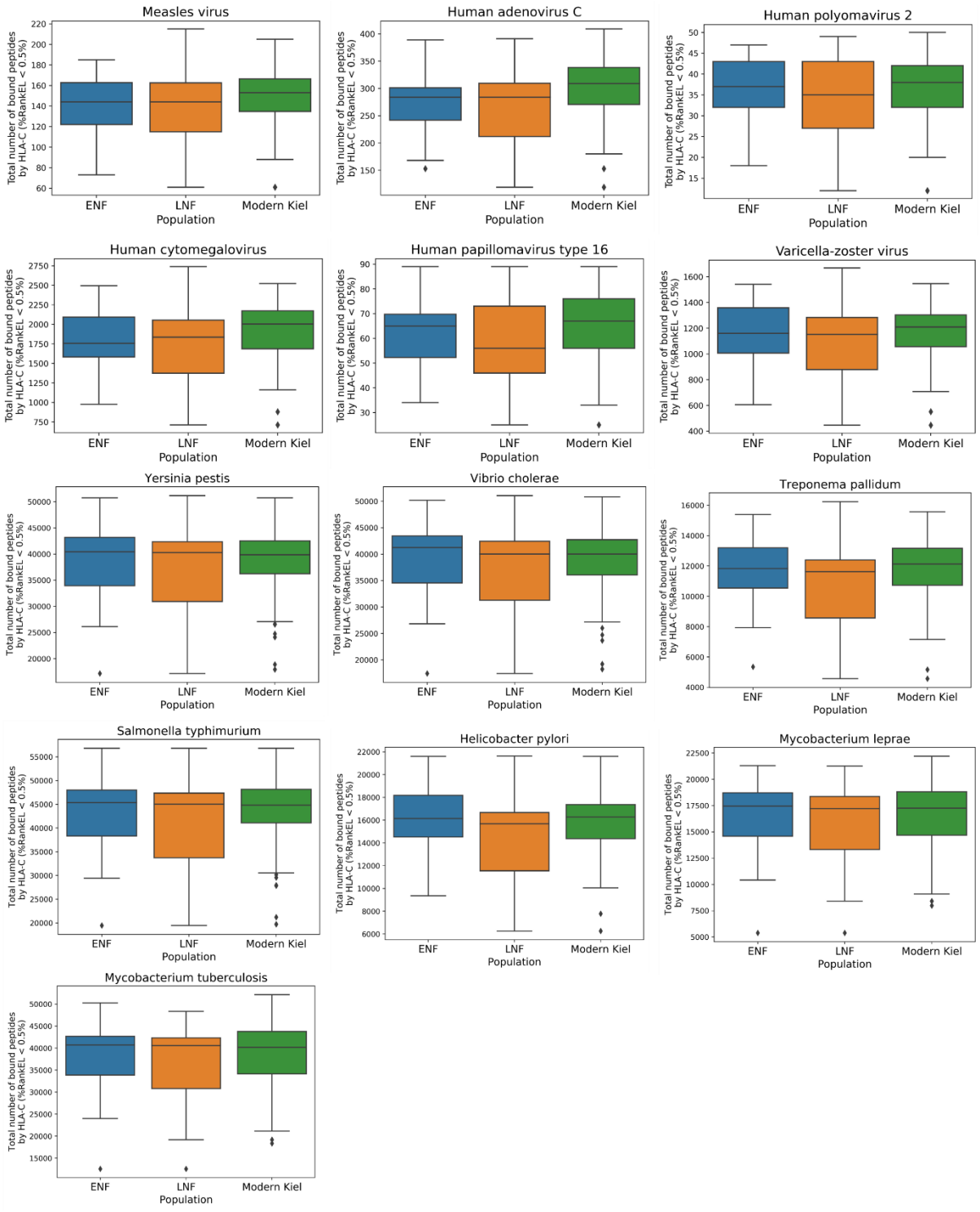
Supplementary Figure 7. (A) Distribution of the allele divergence within populations. Each dot is the Grantham score between the alleles of an individual. Homozygous individuals were excluded from the analysis. (B) Grantham score between the two most common HLA-B (*B*07:02* and *B*08:01*) alleles in the modern Kiel population and the five most common HLA-B alleles in the ancient populations. *B*44:02* is the most divergent allele compared to both of them.



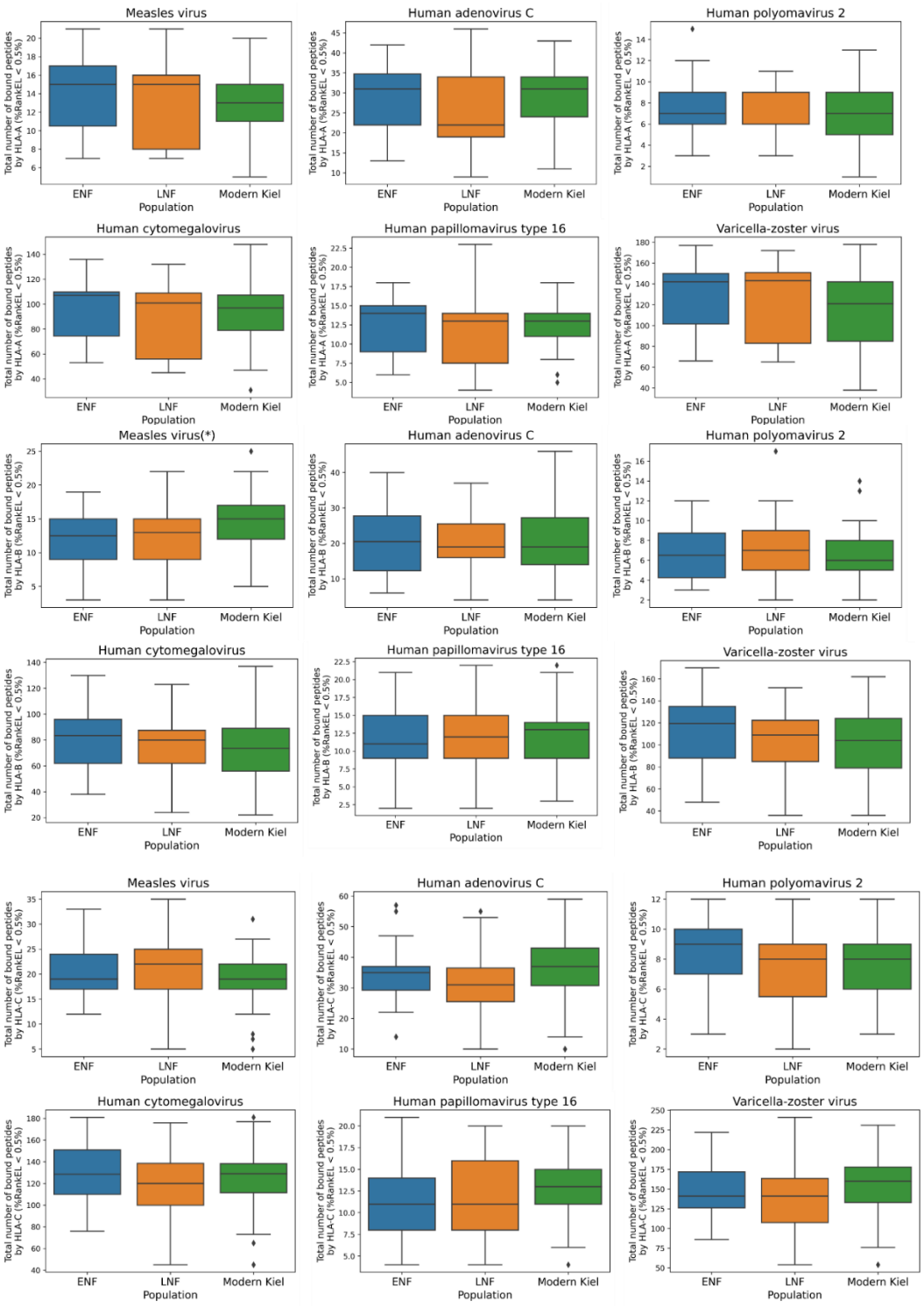
Supplementary Figure 8. Peptide binding analysis of HLA-A alleles and complete proteomes of pathogens. Boxplots represent the distribution of the number of bound peptides from corresponding pathogens by the alleles carried by individuals. Populations were compared by the Kruskal-Wallis test. An asterisk (*) in the header indicates significant differences ($p < 0.05$) between populations.



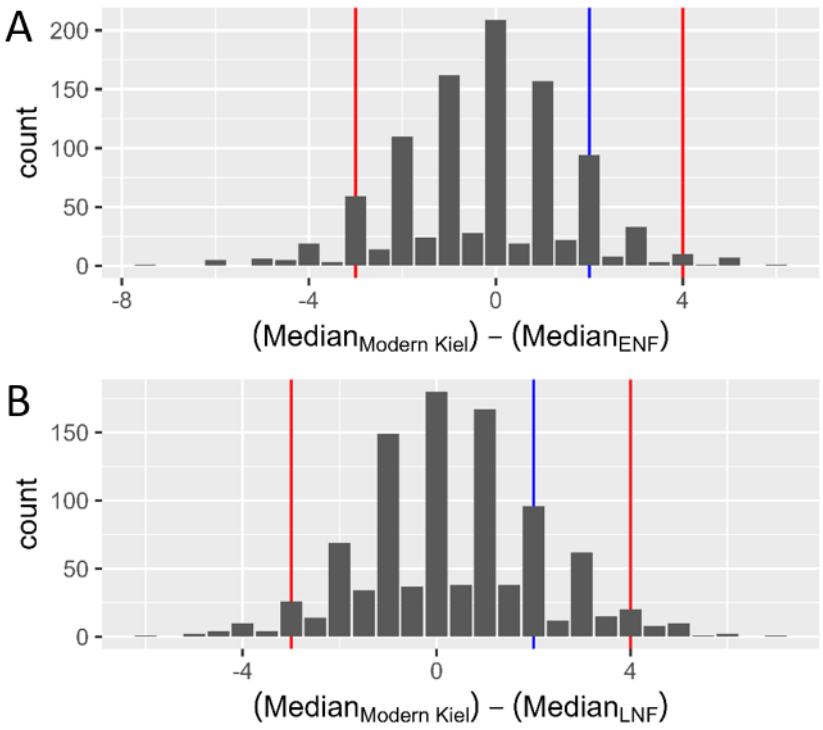
Supplementary Figure 9. Peptide binding analysis of HLA-B alleles and complete proteomes of pathogens. Boxplots represent the distribution of the number of bound peptides from corresponding pathogens by the alleles carried by individuals. Populations were compared by the Kruskal-Wallis test. An asterisk (*) in the header indicates significant differences ($p < 0.05$) between populations.



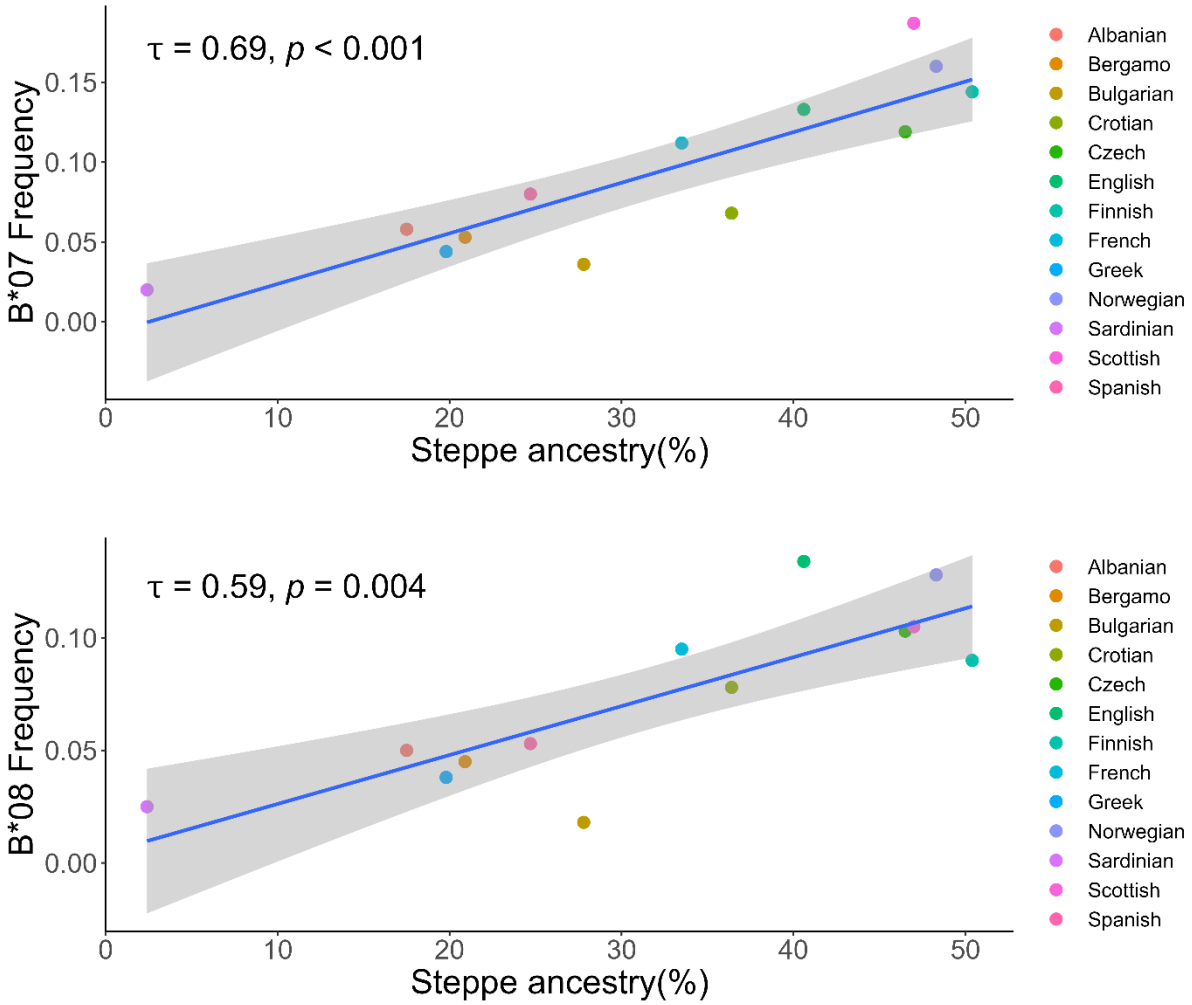
Supplementary Figure 10. Peptide binding analysis of HLA-C alleles and complete proteomes of pathogens. Boxplots represent the distribution of the number of bound peptides from corresponding pathogens by the alleles carried by individuals. Populations were compared by the Kruskal-Wallis test. An asterisk (*) in the header indicates significant differences ($p < 0.05$) between populations.



Supplementary Figure 8. Peptide binding analysis of HLA alleles and surface proteins of viruses. Boxplots represent the distribution of the number of bound peptides from corresponding pathogens by the alleles carried by individuals. Populations were compared by the Kruskal-Wallis test. An asterisk (*) in the header indicates significant differences ($p < 0.05$) between populations.



Supplementary Figure 12. Differences in the number of bound MeV hemagglutinin peptides by the HLA-B alleles of modern and ancient populations. Binding predictions were generated on 1000 randomly generated peptides and the median value within each population was recorded. Bar plot shows the distribution of the differences between median values of (A) modern Kiel and Early Neolithic Farmers (ENF) and (B) modern Kiel and Late Neolithic Farmers (LNF). Blue line shows the observed difference for the MeV hemagglutinin protein. Red lines show the limits of upper and lower 2.5 percentiles.



Supplementary Figure 13. Correlation between B*07 and B*08 allele frequencies and steppe ancestry within Europe. Tau (τ) refers to the Kendall rank correlation coefficient. The linear regression line is shown in blue and the 95% CI around the line is in gray.

Supplementary Tables

Supplementary Table 1. Modern populations used in the analysis

Population	Sample Size
LWK - Luhya in Webuye, Kenya	105
YRI - Yoruba in Ibadan, Nigeria	111
CHB - Han Chinese in Beijing, China	108
CHS - Han Chinese South, China	111
JPT - Japanese in Tokyo, Japan	105
CEU - Utah residents (CEPH) with Northern and Western European ancestry	102
IBS - Iberian Populations in Spain	108
TSI - Toscani in Italia	112
FIN - Finnish in Finland	105
GBR - British from England and Scotland	102
Kiel cohort	100

Supplementary Table 2. Surface proteins of viruses that were used in the peptide-binding analysis.

Pathogen	Surface Proteins	Reference
Human cytomegalovirus	gB, gH, gL, gM, gO, gN	(Wu et al., 2020)
Human polyomavirus 2	VP-1	(Neu et al., 2011)
Human adenovirus C	Hexon protein	(Vellinga et al., 2005)
Varicella-zoster virus	gB, gH, gL, gM, gC, gN, gK, gE, gI,	(Oliver et al., 2016)
Human papillomavirus 16	Major capsid protein L1	(Baker et al., 1991)

Supplementary Table 3. Disease associations of three conserved extended haplotypes

Disease	CEH 44.1 (DRB1*04:01)	CEH 8.1 (DRB1*03:01)	CEH 7.1 (DRB1*15:01)	References
AIDS	Protective	Susceptibility		(Flores-Villanueva et al., 2003)
Type 1 diabetes	Susceptibility	Susceptibility	Protective	(Pugliese et al., 2016)
Multiple sclerosis	Protective	Susceptibility	Susceptibility	(Healy et al., 2010; Zhang et al., 2011)
Type 1 autoimmune hepatitis	Susceptibility	Susceptibility	Protective	(Ma et al., 2021)

Acknowledgements

I have always considered myself a lucky person in life and the fact that my path crossed with Prof. Tobias Lenz certainly supports this belief. I would like to thank him for his patience and supportive supervision throughout my doctoral research. His enthusiasm kept me going in times when I almost lost mine and his scientific expertise guided me through many projects.

I would like to acknowledge all the collaborators, specifically Andre Franke, Almut Nebel, Ben Krause-Kyora, Daniel Harris, Dimitri Monos, Federica Pierini, Magdalena Haller, Michael McQuillan, Nicolas Antonio da Silva, Sébastien Calvignac-Spencer, Sarah Tishkoff and Yan-Rong Chen. Thank you for generously sharing your ideas, your time and your expertise.

I would like to thank all the previous and current members of the Evolutionary Immunogenomics Research Unit Alejandro, Alexey, Ana, Arnaud, Artemis, Britta, Clinton, Federica, Jatin, Joanna, Malavi and Reem. I benefitted a lot from all our discussions and surely enjoyed my time in five (or more, I lost count) different offices that we shared.

I met many great people in the long journey that culminated in this thesis. Many thanks to my colleagues from the International Max Planck Research School for Evolutionary Biology who were an inspiration for me since the beginning. I am especially grateful to Gökçe, Çağdaş, Rooi and their newest family member, as well as my best friend Gece.

I would like to thank the Max Planck Institute for Evolutionary Biology, the German Research Foundation (DFG) and the University of Hamburg for providing financial support that allowed me to pursue my research and attend several meetings and conferences.

I would like to express my deepest gratitude to my parents, Gülten Özer and Mehmet Ali Özer. Their support has been the best comfort I could get in difficult times.

Finally, thanks to my partner, benim hanım, Pelin Kasap for her wholehearted love. As Nazım said: “Our most beautiful days: we haven't seen yet.”

Curriculum vitae

Onur Özer

Date of birth: 18.4.1992
Nationality: Turkish
Address: Greifstraße 11, 24143 Kiel, Germany
e-mail: ozer@evolbio.mpg.de

Education

2018 – present	**Max Planck Institute for Evolutionary Biology, Plön, University of Hamburg, Hamburg, GERMANY** Evolutionary Immunogenomics, Ph.D. Supervisor: Prof. Dr. Tobias Lenz
2016 June – 2016 September	**Institute of Evolutionary Medicine, University of Zurich, Zürich, SWITZERLAND** Graduate Research Project, Supervisor: Dr. Abigail Bouwman
2015 – 2017	**Middle East Technical University, Ankara, TURKEY** Molecular Biology and Genetics, M.Sc. CGPA: 3.86 / 4.00 Supervisor: Prof. Dr. Inci Togan
2014 June - 2014 September	**Danish Archaea Center, University of Copenhagen, Copenhagen, DENMARK** Undergraduate Internship, Supervisor: Prof. Dr. Qunxin She
2010 – 2015	**Middle East Technical University, Ankara, TURKEY** Molecular Biology and Genetics, B.Sc. CGPA: 3.35 / 4.00

Skills

Languages: Turkish (Native), English (Advanced)

Computer skills: Python, R, bash, SQL (hands-on experience), Git

Teaching Experience

Teaching Assistant, University of Hamburg (2021-2022)
Course: Introduction to NextGen Sequencing (Einführung in die NextGen Sequenzierungswelt)

Teaching Assistant, Middle East Technical University (2016-2017)
Courses: General Biology, Cell Biology, Physiology

Publications

Özer, O.,& Lenz, T. L. (2021). **Unique Pathogen Peptidomes Facilitate Pathogen-Specific Selection and Specialization of MHC Alleles**. *Molecular Biology and Evolution*

Pierini, F., Nutsua, M., Böhme, L., **Özer, O.**, Bonczarowska, J., Susat, J., Franke, A., Nebel, A., Krause-Kyora, B., & Lenz, T. L. (2020). **Targeted analysis of polymorphic loci from low-coverage shotgun sequence data allows accurate genotyping of HLA genes in historical human populations.** *Scientific Reports,10*(1),7339

The Severe Covid-19 GWAS Group. (2020). **Genomewide Association Study of Severe Covid-19 with Respiratory Failure.** *New England Journal of Medicine*, *383*(16), 1522–1534.

Yurtman, E. *, **Özer, O.** *, Yüncü, E. *, Dağtaş, N. D., Koptekin, D., Çakan, Y. G., Özkan, M., Akbaba, A., Kaptan, D., Atağ, G., Vural, K. B., Gündem, C. Y., Martin, L., Kılınç, G. M., Ghalichi, A., Açan, S. C., Yaka, R., Sağlıcan, E., Lagerholm, V. K., … Özer, F. (2021). **Archaeogenetic analysis of Neolithic sheep from Anatolia suggests a complex demographic history since domestication.** Communications Biology, 4(1), 1–11. https://doi.org/10.1038/s42003-021-02794-8

*Co-first authors

Degenhardt, F., Ellinghaus, D., Juzenas, S., Lerga-Jaso, J., Wendorff, M., Maya-Miles, D., Uellendahl-Werth, F., ElAbd, H., Rühlemann, M. C., Arora, J., **Özer, O.**, Lenning, O. B., Myhre, R., Vadla, M. S., Wacker, E. M., Wienbrandt, L., Blandino Ortiz, A., de Salazar, A., Garrido Chercoles, A., … Franke, A. (2022). **Detailed stratified GWAS analysis for severe COVID-19 in four European populations.** Human Molecular Genetics, 31(23), 3945–3966. https://doi.org/10.1093/hmg/ddac158

Courses and Workshops

- **Evolutionary Genomics Winter School**, 30.01 – 04.02.2017,
  Hacettepe University, Department of Biology
- **Ancient DNA Techniques for Zoonosis Research Workshop**, 15 – 16.02.2018,
  Robert Koch Institute
- **Sequence Analysis on the UNIX Command Line**, 19 – 23.03.2018,
  Max Planck Institute for Evolutionary Biology, Instructor: Prof. Bernhard Haubold
- **Genomic Signatures of Selection & Association Studies**, 22 – 26.10.2018
  Instructors: Dr. Pablo Orozco-terWengel & Dr. Filippo Biscarini
- **Research Software Development Workshop,** 10 – 11.12.2020
  Max Planck Institute for Evolutionary Biology,
  Instructors: Dr. Carsten Fortmann-Grote & Dr. Nikoleta E. Glynatsi

Declaration

Hereby I declare that:

i. apart from my supervisor's guidance, the content and design of this dissertation is the product of my own work. The co-author's contributions are listed in the Author Contributions section;

ii. this thesis has not already been submitted either partially or wholly as part of a doctoral degree to another examination body, and no other materials are published or submitted for publication than indicated in the thesis;

iii. the preparation of the thesis has been subjected to the Rules of Good Scientific Practice of the German Research Foundation.

iv. I have not had any academic degree withdrawn.

Kiel, 12.09.2023

Onur Özer