

Evolution and function of adaptive immunogenetic diversity in humans



Dissertation

in fulfilment of the requirements for the degree
“Doctor rerum naturalium”
of the Faculty of Mathematics and Natural Sciences
at the Christian Albrechts University of Kiel

Submitted by

Federica Pierini

Emmy Noether group for Evolutionary Immunogenomics
Max Planck Institute for Evolutionary Biology

Plön, June 2019

First referee: Dr. Tobias Lenz

Second referee: Prof. Hinrich Schulenburg

Date of oral examination: 12.07.2019

*“It takes a thousand voices to tell a single story”
Native American proverb*

Table of Contents

Summary	1
Zusammenfassung	4
Introduction	8
Host-pathogen coevolution	8
The human immune system	9
The major histocompatibility complex (MHC).....	10
Genetic variation at HLA loci	11
Evolution of HLA genes	13
Pathogen-mediated selection at HLA genes in humans.....	16
Studying pathogen-mediated selection at HLA genes using ancient DNA	18
Thesis outline.....	20
List of papers / manuscripts	26
Author Contributions	27
<i>Chapter I</i>	29
Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection	
<i>Chapter II</i>	44
Accurate genotyping of HLA immune genes from shotgun sequence data of modern and ancient DNA	
<i>Chapter III</i>	100
Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans	
<i>Chapter IV</i>	112
Ancient history of HLA genes in the America	
Conclusion and perspectives	148
References	153
Annex I.....	168
HLA heterozygote advantage against HIV-1 is driven by quantitative and qualitative differences in allele-specific peptide presentation	
Annex II.....	201
Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy	
Acknowledgements.....	237
Curriculum vitae.....	239
Declaration.....	241

Summary

Natural selection is the process by which heritable traits, that influence adaptation to the environment, contribute to the differential survival and reproduction of organisms. Selection has been one of the main forces influencing human evolution and its role in maintaining genetic variation in different populations is currently one of the most fascinating topics in human evolutionary studies. At the same time, infectious diseases have had an overwhelming effect along the human history; they indeed have been, and still are, one of the leading causes of mortality in human populations. Parasite are thus considered one of the most powerful forces of selection, shaping patterns of human genetic variation and promoting adaptive changes in biological functions such as immunity. The major histocompatibility complex (MHC), and specifically the highly polymorphic human leucocyte antigen (HLA) genes, play a key role in adaptive immunity and are a prime candidate to investigate mechanisms of pathogen selection throughout human history. Decades of studies on MHC genetic diversity revealed that selection at the MHC genes is a dynamic process that involves parallel and not mutually exclusive mechanisms acting at different time scales. With the aim of providing new insight on the nature of long-term coevolution between humans and their pathogens, this thesis addresses the genetic diversity of the immunologically important major histocompatibility complex (MHC), and its evolutionary significance, in historical and present-day human populations. The dynamic action of pathogen-mediated selection and the parallel mechanisms through which it can act are explored in the different chapters, to reconstruct past selection events and to eventually understand, where possible, the functional consequences of pathogen selection for modern humans.

The first chapter focused on one of the mechanisms of balancing selection proposed to contribute to the exceptional polymorphism observed at the classical HLA genes: The divergent allele advantage. Using computational antigen-binding prediction on a large data set of potentially antigenic pathogen peptides, we explored whether pairs of highly diverged MHC alleles together bind more different antigens than more similar alleles. The positive correlation between the genetic distance of two alleles and the combined number of peptides they bind together, detected across the five key classical human MHC genes (HLA-A, -B, -C, -DRB1, and -DQB1), supported the hypothesis that enhanced sequence diversity between alleles in a heterozygous MHC genotype increases the range of potential MHC-presented peptides. Furthermore, a significant correlation between an allele's population frequency and its average pairwise sequence divergence observed for specific HLA loci, suggested still ongoing selection

for divergent HLA genotypes, at least in some human populations. The subsequent investigation of HLA allelic divergence in biomedical datasets revealed the relevance of this mechanism of enhanced antigen presentation on HIV disease progression as well as on cancer immunotherapy response. First, the quantitative effects of both heterozygosity and allele divergence was explored in a large cohort of 6,311 HIV-1–infected individuals, with known HLA genotypes and set point viral load (spVL) information (Annex I). A lower level of viral load in heterozygous individuals compared to homozygous at the HLA-B and HLA-C loci, and a negative correlation between pairwise allele divergence and viral load across individuals at the HLA-B locus confirmed the effects of both heterozygosity and allele divergence on AIDS outcomes. Computational prediction of HLA-presented antigenic HIV peptides revealed that heterozygous patients can bind a broader array of HIV-1 peptides than do homozygotes, at all three classical HLA loci. Also, the number of peptides bound by a pair of alleles was positively correlated with the sequence divergence between the alleles at the HLA-B locus, suggesting that heterozygote advantage at HLA-B is in part mediated by quantitative peptide presentation. Further, the effect of sequence divergence at HLA class I genes on cancer immunotherapy efficacy was examined by studying a dataset of cancer patients treated with immune checkpoint inhibitors (ICI), for which clinical response information as well as cancer exome sequencing data were available (Annex II). We observed that sequence divergence was a strong determinant of survival after ICI. Our results suggested a link between the divergent allele advantage and immunotherapy: highly divergent HLA-I genotypes can present higher diversity of the neopeptide repertoire, influencing T cell clonal expansion, thus facilitating tumor control during immunotherapy.

The study of HLA variability in historical human populations during specific epidemiological events, thus using ancient DNA samples, is of strong interest to explore molecular signatures resulting from transient and fluctuating pathogen selection. Towards this goal, reliable genotyping of the HLA genes in ancient samples is crucial. However, HLA genes exhibit exceptional genetic variability that defies standard HLA genotyping pipelines, especially when it comes to low-coverage and highly degraded ancient DNA (aDNA) sequences. The second chapter described a new HLA genotyping pipeline ('aHLA-Seq') optimized for low coverage shotgun sequence data and showed its accuracy for modern and ancient DNA samples. The approach was subsequently applied to two datasets of ancient samples, from different historical context and geographical area, illustrated in chapters three and four. The third chapter described the first genetic association study on aDNA to date, performed to test whether

variability at HLA genes in medieval Europeans was associated with susceptibility to leprosy in those times. The allele DRB1*15:01, known to be a risk factor for leprosy in different modern human populations, was significantly more frequent in the *M. leprae* DNA-positive cases collected from the St. Jørgen leprosarium in Denmark than in both contemporary and medieval controls. Accordingly, computational antigen-binding prediction on *M. leprae* peptides highlighted limited HLA-presentation capacity of *M. leprae* antigens for the DRB1*15:01 allele. Results were corroborated applying the aHLA-Seq pipeline, which further revealed the presence of the class II haplotype DRB1*15:01-DQB2*06:02, known to be a strong risk factor for inflammatory diseases in present-day populations. In the fourth chapter, we presented the first spatiotemporal characterization of genetic variability of class II HLA loci (HLA-DRB1 and HLA-DQB1) performed to date in ancient Native American populations. A specific HLA target-enrichment approach in combination with the aHLA-Seq pipeline, have been used to explore HLA polymorphisms in ancient and contemporary residents of the town of Xaltocan in central Mexico, and thus to investigate potential HLA allele frequency shifts from pre- to post-European Native American populations. The aHLA-Seq pipeline was further applied to available ancient whole-genome data of samples collected from different sites across the American continent, to eventually describe HLA polymorphisms in ancient Native American populations at a broad geographical and temporal scale.

Zusammenfassung

Natürliche Selektion ist der Prozess, durch den vererbte Merkmale, welche die Anpassung an die Umwelt beeinflussen, weiter gegeben werden und so zum unterschiedlichen Überleben und Reproduzieren von Organismen beitragen. Selektion war eine der Hauptkräfte, die die Evolution des Menschen beeinflussten, und ihre Rolle bei der Aufrechterhaltung der genetischen Variation in verschiedenen Populationen ist aktuell eines der faszinierendsten Themen in Studien zur Evolution des Menschen. Gleichzeitig haben Infektionskrankheiten einen überwältigenden Einfluss auf die menschliche Geschichte gehabt. In der Tat waren und sind sie nach wie vor eine der häufigsten Todesursachen in menschlichen Populationen. Parasiten gelten somit als eine der mächtigsten Faktoren der Selektion. Sie gestalten Muster in der genetischen Variation des Menschen und fördern adaptive Veränderungen in biologischen Funktionen wie der Immunität. Der Haupthistokompatibilitätskomplex (MHC) und insbesondere die hochpolymorphen humanen Leukozytenantigen (HLA) – Gene, spielen eine Schlüsselrolle in der adaptiven Immunität und sind ein Hauptkandidat für die Untersuchung von Mechanismen der Selektion durch Erreger im Verlauf der Menschheitsgeschichte. Jahrzehntelange Studien zur genetischen Vielfalt von MHC zeigten, dass die Selektion an den MHC-Genen ein dynamischer Prozess ist, der parallele und sich nicht gegenseitig ausschließende Mechanismen umfasst, welche auf verschiedenen Zeitskalen wirken. Mit dem Ziel, neue Erkenntnisse über die Natur der langfristigen Koevolution zwischen Menschen und ihren Krankheitserregern zu gewinnen, widmet sich diese Dissertation der genetischen Vielfalt des immunologisch wichtigen Haupthistokompatibilitätskomplex (MHC) und seiner evolutionären Bedeutung in der historischen und heutigen menschlichen Population. Die dynamische Wirkung der pathogenvermittelten Selektion und die parallelen Mechanismen, über die sie wirken kann, wird in den verschiedenen Kapiteln untersucht, um vergangene Selektionsereignisse zu rekonstruieren und um schließlich, soweit möglich, die funktionellen Konsequenzen der Selektion durch Pathogene für den modernen Menschen zu verstehen.

Das erste Kapitel konzentriert sich auf einen der Mechanismen der ausgleichenden Selektion, welche mutmaßlich zum außergewöhnlichen Polymorphismus, welcher bei klassischen HLA Genen beobachtet werden kann, beiträgt. Der Vorteil divergierender Allele. Durch die computergestützte Vorhersage von Antigenbindung, basierend auf einer großen Datenbank von potentiellen Antigen-Peptiden aus Pathogenen, untersuchten wir, ob Paare von hochdivergenten MHC-Allelen zusammen mehr verschiedene Antigene binden konnten als Kombinationen ähnlicher Allele. Die positive Korrelation zwischen der genetischen Distanz von

zwei Allelen und der kombinierten Anzahl von Peptiden, die sie binden konnten, gemessen für die fünf klassischen humanen MHC-Gene (HLA-A, -B, -C, -DRB1 und -DQB1) unterstützten die Hypothese, dass eine erhöhte Sequenzdiversität bei Allelen in einem heterozygoten MHC-Genotyp die Anzahl der potentiell durch MHC-präsentierten Peptide erhöht. Darüber hinaus deutete eine signifikante Korrelation zwischen der Frequenz eines Allels in einer Population und seiner durchschnittlichen paarweisen Sequenzdivergenz, die für bestimmte HLA-Loci beobachtet wurde, darauf hin, dass, zumindest in einigen menschlichen Populationen, die Selektion für divergierende HLA-Genotypen noch andauert. Die anschließende Untersuchung der HLA-Alleldivergenz in biomedizinischen Datensätzen ergab, dass dieser Mechanismus der verbesserten Antigenpräsentation sowohl für das Fortschreiten von HIV-Erkrankungen als auch für das Ansprechen auf Krebsimmuntherapie relevant ist. Zunächst wurden die quantitativen Effekte von Heterozygotie und Alleldivergenz in einer großen Kohorte von 6.311 HIV-1-infizierten Personen mit bekannten HLA-Genotypen und Soll-Viruslast-Informationen (set point viral load = spVL) untersucht (Anhang I). Ein geringeres Ausmaß an Viruslast bei heterozygoten Personen im Vergleich zu homozygoten Personen an den HLA-B- und HLA-C-Loci und eine negative Korrelation zwischen der paarweisen Alleldivergenz und der Viruslast zwischen Personen am HLA-B-Locus bestätigten die Wirkungen sowohl der Heterozygotie als auch der Alleldivergenz bei AIDS-Verläufen. Die computergestützte Vorhersage von HLA-präsentierten Antigen-HIV-Peptiden zeigte, dass heterozygote Patienten an allen drei klassischen HLA-Loci eine breitere Palette von HIV-1-Peptiden binden können als homozygote. Die Anzahl der durch ein Allelpaar gebundenen Peptide korrelierte dabei positiv mit der Sequenzdivergenz zwischen den Allelen am HLA-B-Locus, was darauf hindeutet, dass der Vorteil heterozygoter bei HLA-B teilweise durch quantitative Peptidpräsentation vermittelt wird. Ferner wurde die Auswirkung der Sequenzdivergenz bei HLA-Klasse-I-Genen auf die Wirksamkeit von Krebsimmuntherapie untersucht, indem ein Datensatz von Krebspatienten, welche mit Immun-Checkpoint-Inhibitoren (ICI) behandelt wurden, für die klinische Daten sowie Sequenzdaten des Krebs-Exoms verfügbar waren (Anhang II), untersucht wurde. Wir beobachteten, dass die Sequenzdivergenz dabei stark mit dem Überleben nach ICI zusammenhing. Unsere Ergebnisse legen eine Verbindung zwischen dem Vorteil divergierender Allele und der Immuntherapie nah: Hoch divergente HLA-I Genotypen können eine höhere Neopeptid-Diversität präsentieren, beeinflussen die klonale T-Zell Expansion und vermitteln so verbesserte Tumorkontrolle während der Immuntherapie.

Die Untersuchung der Variabilität der HLA-Gene in historischen menschlichen Populationen während spezifischer epidemiologischer Ereignisse unter Verwendung alter DNA-Proben ist von großem Interesse, um molekulare Signaturen zu untersuchen, die sich aus der transienten und fluktuierenden Selektion durch Krankheitserreger ergeben. Um dieses Ziel zu erreichen, ist eine zuverlässige Genotypisierung der HLA-Gene in historischen Proben von entscheidender Bedeutung. HLA-Gene weisen jedoch eine außergewöhnliche genetische Variabilität auf, welche die standardmäßigen HLA-Genotypisierungs-Pipelines vor große Probleme stellt. Dies ist insbesondere bei Sequenzen mit geringer Abdeckung und stark abgebauter alter DNA (aDNA) der Fall. Im zweiten Kapitel dieser Arbeit wurde eine neue HLA-Genotypisierungs-Pipeline („aHLA-Seq“) beschrieben, welche für Shotgun-Sequenzierung mit geringer Abdeckung optimiert ist. Dabei wurde die Genauigkeit der Methode für neue und historische DNA-Proben gezeigt. Der Ansatz wurde anschließend auf zwei Datensätze antiker Proben aus unterschiedlichen historischen Kontexten und geografischen Gebieten angewendet, dargestellt in den Kapiteln drei und vier. Das dritte Kapitel beschrieb die bisher erste genetische Assoziationsstudie zu aDNA, um zu testen, ob die Variabilität von HLA-Genen bei mittelalterlichen Europäern mit der Anfälligkeit für Lepra in Verbindung stand. Das Allel DRB1*15:01, das als Risikofaktor für Lepra in verschiedenen modernen menschlichen Populationen bekannt ist, trat bei den *M. leprae* DNA-positiven Fällen vom Leprosarium St. Jørgen in Dänemark signifikant häufiger auf als in neuzeitlichen und mittelalterlichen Kontrolldaten. Dementsprechend zeigte die rechnergestützte Vorhersage der Antigenbindung an *M. leprae*-Peptide die begrenzte HLA-Präsentationskapazität von *M. leprae*-Antigenen für das DRB1*15:01-Allel. Diese Ergebnisse wurden durch die aHLA-Seq-Pipeline bestätigt, welche zusätzlich das Vorhandensein des Klasse-II-Haplotyps DRB1*15:01-DQB2*06:02 zeigte, welcher als starker Risikofaktor für entzündliche Erkrankungen in heutigen Populationen bekannt ist. Im vierten Kapitel präsentierten wir die bisher erste raumzeitliche Charakterisierung der genetischen Variabilität von HLA-Lokussen der Klasse II (HLA-DRB1 und HLA-DQB1) in historischen Populationen indigener amerikanischer Völker. Ein spezifischer Ansatz zur HLA-Anreicherung in Kombination mit der aHLA-Seq-Pipeline wurde verwendet, um HLA-Polymorphismen bei historischen und heutigen Bewohnern der Stadt Xaltocan in Zentralmexiko aufzudecken. Damit wurde insbesondere eine mögliche Veränderung der HLA-Allel-Frequenzen zwischen vor- und nachkolonialen indigenen amerikanischen Populationen untersucht. Die aHLA-Seq-Pipeline wurde ferner auf vollständige Genome angewendet, die aus Proben von verschiedenen Orten auf dem gesamten amerikanischen Kontinent gewonnen wurden. Dies

ermöglichte schließlich HLA-Polymorphismen in historischen indigenen Populationen auf einer breiten geografischen und zeitlichen Skala zu beschreiben.

Introduction

Host-pathogen coevolution

The intuition that infectious diseases should be regarded as a major selective pressure in our species has been firstly proposed by JBS Haldane, who claimed that “...one of the principal characters possessing survival value is immunity to disease. Unfortunately, this is not a very permanent acquisition, because the agents of disease also evolve, and on the whole more rapidly than their victims” (Haldane 1932). Pointing out that pathogens also evolve, and in a faster way than their hosts, he pioneered the human host–pathogen coevolution field. The cyclical process of adaptation and counter-adaptation between hosts and parasites is generally summarized under the “Red Queen hypothesis” proposed by Leigh Van Valen in 1973 (Van Valen 1973). Inspired by the Red Queen's race imagery in Lewis Carroll's novel *Through the Looking-Glass*, representing the Queen in chess and Alice constantly running but remaining in the same spot (**Figure 1**) (Carroll 1900), Leigh Van Valen described the core idea of his hypothesis as the perpetual motion of biotic forces underlying the evolution of species affected by it (Van Valen 1973).

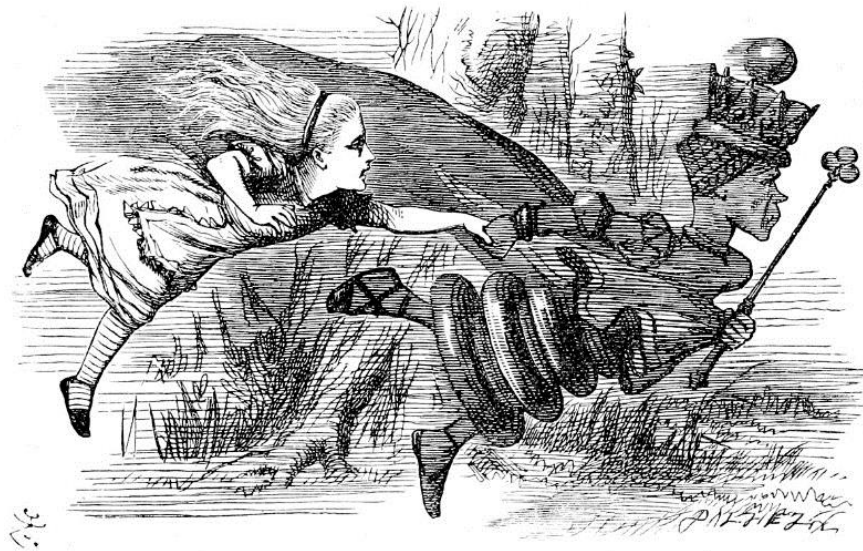


Figure 1 | Illustration by Sir John Tenniel from Lewis Carroll's *Through the Looking-Glass*, 1871 Carroll (1900).

Coevolution can be defined as the process of reciprocal adaptive genetic change in two or more species (Woolhouse et al. 2002). Because of the intimate nature of the association, it is

particularly relevant in host–pathogen systems, where hosts are under pressure to evolve resistance to pathogens, while pathogens attempt to evade host surveillance in order to achieve a successful infection (Sironi et al. 2015). Due to the transient nature of the advantage conferred by specific genetic variants, detecting human adaptations to pathogens at genetic level might be challenging. Nevertheless, several cases have been described mostly at the level of host defense genes (reviewed in Barreiro and Quintana-Murci (2010)).

The human immune system

The immune system is a complex system of cells, tissues, and organs that work together within an organism, to protect the body from pathogenic microbes present in the environment, but also to recognize and remove dysfunctional cells present in the body itself (Murphy and Weaver 2017). In humans, two complementary categories of host defense reactions work closely together, taking on different tasks. An immediate and nonspecific defense mechanisms, present in all animals and known as the innate immunity, is readily able to limit the presence of invading pathogens (Murphy and Weaver 2017). It uses innate generic receptors, called pattern recognition receptors (PRRs), constitutively produced, which are able to recognize conserved molecular patterns on different classes of pathogens as well as common pathological changes in self-cells (Murphy and Weaver 2017). Along with the innate immunity, a more sophisticated defense mechanism has evolved in jawed vertebrates, including humans: the adaptive immunity. The adaptive receptors are exposed on the surface of both T and B lymphocytes (TCR and BCR respectively), and are not conservatively encoded in the genome (Murphy and Weaver 2017). They are generated in each individual by the strategy of receptor diversification, i.e. by somatic recombination and diversification of gene segments. B-cell receptors are able to recognize antigens directly, while T-cell receptors can bind only those antigens associated to the major histocompatibility complex (MHC) proteins (Danilova 2012). The adaptive response is antigen-specific since only receptors able to recognize antigenic configurations of specific pathogens are activated, thus promoting proliferation of immune cells, and leading to a targeted immune response. As a consequence, each individual has a unique set of adaptive immune receptors, which depends on its life history (Danilova 2012). Moreover, during adaptive response an antigen-specific memory component is generated, which allows a faster response upon a second exposure to the same pathogen (Murphy and Weaver 2017).

The major histocompatibility complex (MHC)

A key component of the adaptive immune system, common to all jawed vertebrates, is the major histocompatibility complex (MHC) (Klein and Figueroa 1986). Also known in humans as the human leukocyte antigen (HLA) locus, this gene-dense region of the genome spans ~4 Mb on the short arm of chromosome 6 and comprises over 200 genes, many of which are involved in adaptive immunity (Murphy and Weaver 2017). Among these genes, the classical HLA genes (class I and class II) encode cell-surface glycoproteins, with homologous structures and complementary functions in binding peptide antigens from degraded self and non-self-proteins and presenting them to lymphocyte receptors (i.e. TCRs). If non-self antigenic peptides bound to HLA molecules are recognized by the antigen receptors of T lymphocytes, this eventually stimulates an immune response (Jensen 2007).

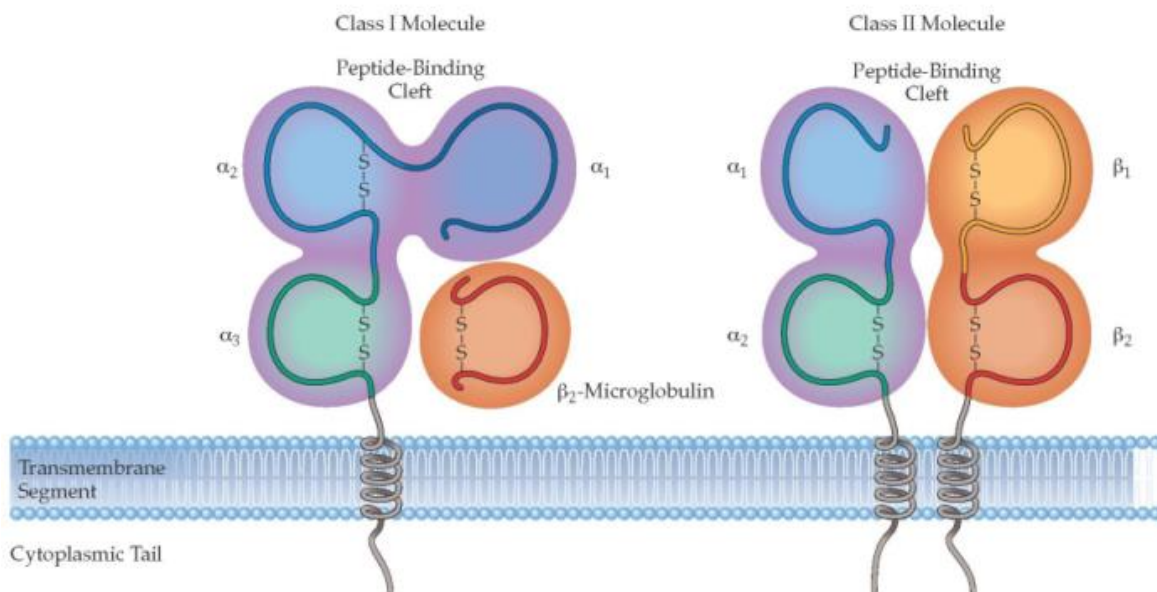


Figure 2 | Human Leukocyte Antigen (HLA) Class I and II Domain Organization.

HLA class I glycoproteins are expressed on the surface of all nucleated somatic cells. They are heterodimer consisting of a heavy chain, made up of three extracellular domains (α_1 , α_2 , α_3), a transmembrane region and a cytoplasmic domain, and a light chain called β_2 -microglobulin, which consists of a single domain (Hughes and Yeager 1998). In humans, the alpha chains are

encoded inside the HLA complex by the three polymorphic class I loci HLA-A, HLA-B, HLA-C; while β 2-microglobulin is encoded outside the HLA region, on chromosome 15. Two of the extracellular domain (α 1 and α 2) form a peptide-binding cleft, known as the peptide-binding region (PBR) (**Figure 2**) (Jensen 2007). Peptides resulting from proteasome degradation and mainly 9 amino acids in length are presented within the peptide-binding region of class I molecules (Neefjes et al. 2011). In this way, the internal proteome is made available for surveillance by cytotoxic T lymphocytes (CTL), which can detect and kill cells presenting pathogen-derived peptides or tumor antigens (Jensen 2007).

HLA class II glycoproteins are expressed in specialized antigen-presenting cells. They consist of two heavy chains, α and β , both of them including two extracellular domains, encoded in the class II region of the HLA complex. The class II region is divided into DR, DP, DQ subregions, each of which contains a functional α chain gene and one or more β chain genes (Hughes and Yeager 1998). In class II molecules the peptide-binding region is formed by α 1 and β 1 domains. Peptides presented by HLA class II molecules can vary substantially in length between 11 and 17 residues; they are usually derived from exogenous proteins that have entered the cells via endocytosis and processed through lysosomal protease degradation (Neefjes et al. 2011). The exposed non-self-peptides are recognized by helper T lymphocytes (TH cells), leading to a complex cascade of specific immune responses (Jensen 2007).

Genetic variation at HLA loci

The hallmark of HLA genes is the extensive degree of structural and functional polymorphism which can be characterized at different levels (Clarke and Kirby 1966). In humans, each individual holds several adjacent related HLA genes encoding proteins with same or similar functions. There are 3 genes for the classical HLA-I α -chain, (-A, -B and -C) and three pairs of classical HLA class II α - and β -chain genes (-DR, -DP, and -DQ) which are all co-expressed. Consistently, each individual produces on the cell surface a number of different HLA molecules with different ranges of peptide binding specificities (Murphy and Weaver 2017). Moreover, for all of these loci, the chance for an individual to have the same allele on both the homologous chromosomes is very small and most individual will be heterozygous (Murphy and Weaver 2017). Finally, the classical HLA genes are highly polymorphic, with hundreds of different alleles identified at some of these loci (Klein and Figueroa 1986; Trowsdale 2011). Polygeny, heterozygosity and allelic polymorphism together produce the extensive variation observed within individuals, families and populations.

The analysis of both nucleotide and amino acid sequences has clarified how this exceptional polymorphism is expressed at allelic level. HLA alleles are deeply divergent i.e. they show a large number of differences compared with other alleles at the same locus. Indeed, nucleotide diversity within HLA class I and II genes has been shown to be much higher than elsewhere in the genome (Li and Sadler 1991; Satta et al. 1998). The high level of diversity found at the HLA loci is mainly located within peptide-binding region (i.e. the pocket where antigens are bound) (Hughes and Yeager 1998), where the molecular differences between alleles vary up to 50 nucleotides and more (Buhler and Sanchez-Mazas 2011). Further, in codons specifying amino acid involved in peptide binding, called antigen-binding sites (ABS), the rate of non-synonymous substitutions is greater than the rate of synonymous substitutions (Hughes and Nei 1988). Since the majority of the polymorphic sites are located in functionally important site, they are more likely to be maintained in the population because of their functional effect in increasing the binding ability of HLA molecules towards a high variety of antigens, thus optimizing immune protection against pathogens (Robinson et al. 2017). Humans share similar MHC allelic lineages with closely related species. The persistence of allelic diversity across multiple species is a common feature observed at MHC genes. This observation has been described as trans-species polymorphism (Klein 1987): ancestral lineages present in the common ancestor are inherited through successive speciation events, thus persisting over long periods of time (Klein et al. 2007). The features of HLA polymorphism are inconsistent with neutral expectation, therefore, the ancestral and highly diverged HLA variants are assumed to be adaptive and selectively maintained by balancing selection (Hughes and Nei 1988).

Along with the persistence of ancestral variants, new MHC alleles arise in natural populations by point mutation as well as through the mechanism of recombination and gene conversion (Parham and Ohta 1996). HLA alleles that stem from point mutations differ by one nucleotide substitution from the nucleotide sequences of an older allele. They are of more recent origin and commonly considered as rare alleles. Indeed, they are assumed to show limited differentiation from their original established alleles in terms of peptides binding specificities (Robinson et al. 2017), thus rarely providing a selective advantage and mostly being affected by drift. In contrast, HLA variants created through the mechanisms of recombination and/or gene conversion often show new functional binding properties (Robinson et al. 2017); such alleles are more likely to be advantageous and consequently rise in frequency in human populations, at least temporarily.

Evolution of HLA genes

How selection acts to maintain genetic variation in different populations is currently one of the most interesting issues in human evolutionary studies. Pathogens have been a major selective force in human evolution (Haldane 1932; Fumagalli et al. 2011); their constant burden over evolutionary time has likely shaped the genetic variation found at a large number of immune genes within and among present day populations. Because of the fundamental role of HLA genes in immunity, their exceptional polymorphism probably reflects the selective pressures imposed by the diversity of pathogens (Dean et al. 2002; Quintana-Murci et al. 2007). Accordingly, past and ongoing pathogen-mediated selection is proposed to be one of the major factors affecting the genetic variability at those genes (Apanius et al. 1997). Decades of studies on MHC genetic diversity in natural population as well as in model species under laboratory conditions have revealed that selection at the MHC genes is a dynamic process that involves parallel and not mutually exclusive mechanisms, acting at different time scales (Spurgin and Richardson 2010).

The heterozygote advantage hypothesis was first proposed by Doherty and Zinkernagel (1975b). The initial hypothesis assumes that heterozygous individuals at MHC loci can present a greater range of pathogen peptides than homozygotes. They show increased resistance to pathogens, and are more likely to have higher relative fitness, resulting in an enhanced persistence of MHC alleles in the population (Hughes and Yeager 1998). Heterozygote advantage at MHC genes has been supported by experimental infections studies of laboratory-bred organisms (Penn et al. 2002; O'Connor et al. 2010), genetic association studies in domesticated species (Takekuma et al. 2008), investigation of parasitic bacterial community in wild populations (Evans and Neff 2009; Niskanen et al. 2014) as well as investigation of human disease (Carrington et al. 1999). The excessive sequence divergence frequently observed among MHC alleles has favored a further explanation of the heterozygote advantage, known as divergent allele advantage (Potts and Wakeland 1990). Under this form of asymmetric overdominant selection diverged alleles are selectively favored. Heterozygous individuals with more divergent MHC allele combinations (i.e. larger number of sequence differences along the antigen-binding domains) are thought to encode glycoproteins which differ proportionally in the repertoire of antigens they can bind. Because of their divergent peptide binding specificities, they are able to present a wider array of antigens to immune effector cells, conferring an advantage against pathogen infections. Overall, this mode of selection can be seen as quantitative mechanism, which does not respond to specific pathogen species or strains but

rather implies an effective immunity against a wide range of parasites. It can act over long evolutionary time scales promoting the maintenance of ancestral allelic lineages in natural populations (Lenz 2011). Examples of divergent allele advantage acting on MHC loci have been demonstrated in a number of natural and laboratory populations (Landry et al. 2001; Richman et al. 2001; Lenz et al. 2009; Schwensow et al. 2010; Lenz et al. 2013), and supported by several computer-based binding prediction studies (Lenz 2011; Lau et al. 2015; Buhler et al. 2016). A plausible alternative hypothesis has been later proposed for the heterozygote advantage, which can help in explaining those cases where heterozygous genotypes may not constantly be more resistant to specific pathogen infection (Langefors et al. 2000; Grimholt et al. 2003). It is well established that specific MHC alleles provide resistance to specific parasite, this observation implies qualitative differences between different MHC alleles (Piertney and Oliver 2006; Trowsdale 2011). Consistently, the fitness advantage of heterozygous genotypes might be conferred by the higher probability to carry the needed specific allele able to cope with a specific pathogen, rather than a general quantitative advantage conferred by the wider number of antigens presented to the immune system.

The negative frequency-dependent selection or rare-allele advantage (Slade and McCallum 1992) is the second mechanism of balancing selection which has been proposed to promote high genetic diversity at MHC genes. Pathogens adapt to infect the most common host genotype, operating a strong selection to overcome its resistance. As a result, the relative fitness of common host genotypes decreases. At the same time, rare genotypes providing resistance against an invading disease are less infected, which provides them with a selective advantage. Rare alleles can hence spread through the population until they become common. The negative frequency-dependent selection (NFDS) can be seen as a qualitative mechanism in which the frequencies of specific MHC alleles are influenced by the selection pressure imposed by single pathogens. It has often been defined as a cyclical, coevolutionary arms race which leads to a cycling of fitness values of both pathogens and MHC alleles (Slade and McCallum 1992). Because of the nature of epidemic diseases, which tend to be episodic, this mechanism likely works on a shorter time scale, resulting in a transient impact of both MHC and pathogen evolution. Correlative and experimental support for the negative frequency-depend selection at MHC genes has been provided in humans (Trachtenberg et al. 2003), reed warbler population (Westerdahl et al. 2004), laboratory mice (Kubinak et al. 2012), stickleback (Eizaguirre et al. 2012; Bolnick and Stutz 2017) and guppies (Phillips et al. 2018).

Finally, the diversity at the MHC loci may also be maintained through fluctuating selection pressures (Hill 1991). Adaptive frequency shifts of MHC alleles can result from temporally and spatially varying parasite pressures. In this case, different subsets of MHC alleles will be selected at different points in space and time, depending on spatial and temporal heterogeneity in the type and abundance of pathogens. Pathogen abundance is determined by the biotic and abiotic environment as well as stochasticity in host-parasite coevolution. Similarly to the negative frequency-dependent selection, fluctuating selection concerns individual alleles rather than allele combinations or genotypes, which are instead important under the heterozygote advantage mechanism. Further, while negative frequency-dependent selection promotes polymorphism within populations, fluctuating selection promotes polymorphism between different populations. The concurrent difference in parasite community and MHC diversity of host populations between habitat types (Wegner, Reusch, et al. 2003; Eizaguirre et al. 2011), the geographical structure of MHC variability in natural populations (Ekblom et al. 2007; Alcaide et al. 2008; Babik et al. 2008) and the existence of population-specific alleles associated with increased resistance to infection diseases (Hill et al. 1991; Bonneaud et al. 2006; Loiseau et al. 2009; Sanchez-Mazas et al. 2017) are among the evidence supporting the effect of locally varying parasite pressures on MHC diversity.

The role of pathogen-mediated selection in the maintenance of genetic variability at the MHC loci does not preclude the synergic action of stochastic factors associated with the demographic history of species. Accordingly, several studies have elucidated how geography can describe to some extent patterns of MHC genetic variation either worldwide or within continents (Prugnolle et al. 2005b; Meyer et al. 2006; Currat et al. 2010; Buhler and Sanchez-Mazas 2011; Di and Sanchez-Mazas 2011; Sanchez-Mazas et al. 2013). For instance, the reduction of variation associated with the consecutive migration events of modern humans throughout the world, which is well established for non-HLA loci (Prugnolle et al. 2005a), has also been detected at HLA genes. Indeed, while higher values of HLA heterozygosity have been observed in population from sub-Saharan Africa, heterozygosity at HLA loci is lower for populations at greater geographic distances from Africa (i.e. North and South America) (Prugnolle et al. 2005b). However, geographic distance, and thus demographic events, explains only a part of HLA genetic diversity, while the remaining proportion of HLA heterozygosity is significantly associated with local pathogen richness of different world regions, hence ascribed to pathogen selection.

Among the factors influencing the high level of diversity observed at MHC, mate choice during sexual selection has been shown to play a role (Milinski 2006; Chaix et al. 2008). MHC-disassortative mating strategy, achieved through detection by the choosing sex of the so-called “honest signals”, can promote an offspring with either optimal or maximum MHC genetic diversity, optimizing its resistance to pathogens and thus its fitness (Penn and Potts 1999). For instance, in sticklebacks females choose a mate to complement their own set of MHC genes, which results in an optimal number of different MHC alleles in their offspring (Aeschlimann et al. 2003; Milinski 2003).

Pathogen-mediated selection at HLA genes in humans

Coevolution between humans and their pathogens has led to the development of an effective host immune response achieved through extensive HLA genetic diversity. As described before, there is abundant evidence consistent with host-pathogen coevolution, and its effects on MHC variability, from studies of pathogens infecting non-human host. However, convincing empirical validations for pathogen-mediated selection in humans are difficult to obtain. On one hand, comprehensive experimental tests using controlled human infections are not feasible, because of obvious ethical reasons. On the other hand, the dynamic nature of host-pathogen coevolution makes its detection at the genetic level in natural context difficult (Penman and Gupta 2018). Consistently, while the molecular signatures of continuous and directional selection on HLA genes might be intuitive and potentially detectable in human populations, identifying signatures resulting from transient and fluctuating selection might be challenging. Nevertheless, the different modes of past pathogen selection are suspected to have functional consequences detectable in present-day human populations. Indeed, HLA alleles play a key role in the resistance to infectious diseases but are also associated with various complex genetic disorders in contemporary humans, potentially reflecting a legacy of historical coevolution between humans and their pathogens (Dean et al. 2002; Lenz et al. 2016).

Among the few examples of infectious diseases that have been clearly associated with the HLA (Trowsdale 2011) are human immunodeficiency virus (HIV) (Moore et al. 2002; Fellay et al. 2007; International HIV Controllers Study et al. 2010), human papilloma virus (HPV) (Chen et al. 2015), hepatitis (Kamatani et al. 2009; Duggal et al. 2013; Zhu et al. 2016), malaria (Hill et al. 1991), leprosy (Zhang, Huang, et al. 2009) and tuberculosis (Sveinbjornsson et al. 2016). While exploring the biology of the above mentioned infectious diseases, it has been shown that HLA variation affects pathogen peptides presentation to T lymphocytes: a defective antigen

presentation can often cause a more severe or a persistent form of the disease (Yang et al. 2016; Zhu et al. 2016), while the specific presentation of conserved pathogen epitopes can result in a slow disease progression (International HIV Controllers Study et al. 2010; Rao et al. 2015). However, not only the domains involved in peptides binding have been shown to be under pathogen selection but also sites with regulatory function that can modulate the expression of HLA molecules (Kulkarni et al. 2011; Thomas et al. 2012). A large number of disparate complex genetic disorders, other than infectious diseases, have been associated with the HLA region (Trowsdale 2011). The most striking associations have been found with autoimmune conditions, for some of which the HLA genes are among the strongest risk factors (i.e. diabetes, arthritis, celiac disease, lupus, ankylosing spondylitis, multiple sclerosis, psoriasis and Crohn's disease) (Matzaraki et al. 2017). Among the other diseases associated with HLA variants are cancer as well as some neurological disorders (Trowsdale 2011). The main hypothesis proposed to explain the extensive number of diseases associated with HLA is that selection for resistance to infection drives the evolution of MHC polymorphism while promoting a higher risk of developing other complex genetic disorders. According to this hypothesis, HLA variants, which are maintained in human populations because advantageous against some pathogens, are at the same time able to stimulate an immune response against host antigens being thus associated with autoimmune conditions (Matzaraki et al. 2017). Pathogen selection, which has largely shaped the repertoire of pathogen derived sequence bound by HLA molecules, has been proposed to also play a role in cancer-immune reaction, and neoantigens (i.e. cancer-specific peptides derived from somatically mutated self-proteins) are thought to be more likely detected as non-self when resembling pathogen sequences.

Despite the large number of studies carried out, several features of HLA disease associations make its characterization challenging, and the underlying functional mechanisms remain partially unexplored. Some of the described diseases are polygenic showing associations with various alleles of classical class I and class II loci as well as interaction with non-HLA genes (Meyer et al. 2018). Alleles at adjacent loci are often in strong linkage disequilibrium within the HLA region, and variants are commonly inherited in linked clusters or haplotypes. The epistatic interaction between variants has often been invoked to explain why advantageous haplotypes can be selectively maintained at high frequency in populations (Penman et al. 2013). HLA genes are also among the most pleiotropic genes, and the same variant can often affect more than one phenotype (Matzaraki et al. 2017). Further, the same HLA allele or haplotype might have antagonistic fitness effects, conferring resistance against one disease and susceptibility to

another (Matzaraki et al. 2017). Finally, the observed HLA disease associations can, in some cases, differ between different populations (Matzaraki et al. 2017). Notably, the majority of the studies have been carried on populations of European ancestry, with an evident under-representation of ethnically diverse populations. Therefore, to fully perceive the nature of pathogen-mediated selection at HLA genes it is necessary to bridge this gap by including under-studied populations in the genetic studies.

Studying pathogen-mediated selection at HLA genes using ancient DNA

The recent development of genomic tools for the analysis of ancient DNA provides a unique opportunity to unravel the trajectories of alleles associated with human adaptations to newly introduced or co-evolving pathogens (Marciniak and Perry 2017). The characterization of allelic variants associated with immunity using ancient genomes already gave promising results, providing a first glimpse into the spatiotemporal distribution of specific immune system genes in ancient human populations (Olalde et al. 2014; Mathieson et al. 2015; Hofmanova et al. 2016; Lindo et al. 2016; Gelabert et al. 2017). Among those studies, genome-wide scans for selection applied to ancient genomes have identified MHC genes among the candidate loci. A strong signal of selection at the major histocompatibility complex (MHC) on chromosome 6 was detected by Mathieson et al. (2015) performing a genome-wide scan for selection with 230 West Eurasians ancient samples (~8,500 – 2,300 years BP). Furthermore, Lindo et al. (2016) found a substantial change in the frequency of a HLA-DQA1 allele when comparing the genetic diversity of an ancient indigenous North American population and the descendant population living in the region today. Such difference has been linked to European-borne epidemics affecting Native American population after European contact. Nevertheless, in both studies the link between the signals of selection and their functional consequences remains unresolved. One of the first issues comes from the difficulties in genotyping HLA genes using ancient DNA. Indeed, the extensive genetic variability found at these genes together with the profound degradation observed in ancient DNA sequences make their characterization challenging. Moreover, despite the increasingly large number of ancient-pathogen genomes now available (Spyrou et al. 2019), the geographical and temporal spread of the infectious agents along human history remains largely unexplored. Thus, the exact links between HLA genetic variants and disease causative agents have been so far difficult to detect using ancient DNA. In this light, the development of accurate HLA genotyping tools is a prerequisite for studying the evolution of human resistance or susceptibility to pathogens in historical populations. Such advances, in conjunction with the increasing amount of information available from ancient pathogen genomics and human

population genomics, together with historical and archaeological data concerning epidemiological events, will provide a deeper understanding of infectious disease dynamics, shedding light on the nature of long-term coevolution between humans and their pathogens.

Thesis outline

I focused my PhD research on the genetic diversity of the immunologically important major histocompatibility complex (MHC), and its evolutionary significance in historical and present-day human populations. This genomic region, and specifically the highly polymorphic human leucocyte antigen (HLA) genes, play a key role in the resistance to infectious diseases but are also associated with various complex genetic disorders in contemporary humans, potentially reflecting a legacy of historical coevolution between humans and their pathogens. Using a wide spectrum of approaches I studied the dynamic action of pathogen-mediated selection, which is proposed to be a major driver of HLA diversity. I thus explored the parallel mechanisms through which pathogen selection can act, using datasets of contemporary human populations as well as various datasets of ancient samples belonging to different historical periods and spanning different geographical regions. Furthermore, using biomedical datasets, I investigated the implication of pathogen-mediated selection, and thus the potential functional consequence of adaptive immunogenetic variation, on HIV disease progression and cancer immunotherapy response. My thesis consists of four main chapters and two annexes, which are outlined below. All projects have been conducted in cooperation with colleagues; therefore, a list of publications included in this thesis, together with a detailed overview of each author's contribution, is also reported.

Chapter I

Divergent allele advantage at human MHC genes: signatures of past and ongoing selection

The divergent allele advantage, a mechanism of balancing selection, is proposed to contribute to the exceptional polymorphism observed at the classical HLA genes. It assumes that MHC genotypes with more divergent alleles allow for broader antigen-presentation to immune effector cells, by that increasing immunocompetence (Potts and Wakeland 1990; Wakeland et al. 1990). In this chapter, we investigated the direct correlation between pairwise sequence divergence and the corresponding repertoire of bound peptides across the five key classical human MHC genes (human leukocyte antigen; HLA-A,-B, -C, -DRB1, and -DQB1). Using computational antigen-binding prediction on a large data set of potentially antigenic pathogen peptides we first explored whether pairs of highly diverged MHC alleles together bind more different antigens

than more similar alleles. The positive correlation between the genetic distance of two alleles and the combined number of peptides they bind together supported the hypothesis that enhanced sequence diversity between alleles in a heterozygous MHC genotype increases the range of potential MHC-presented peptides. The observed functional advantage was reinforced when addressing the frequency of divergent HLA alleles in different human populations. The significant correlation between an allele's population frequency and its average pairwise sequence divergence observed for specific HLA loci, suggested still ongoing selection for divergent HLA genotypes, at least in some human populations. Overall, our results support the divergent allele advantage as a meaningful quantitative mechanism through which pathogen-mediated selection leads to the evolution of MHC diversity.

Annex I (collaborative project with fellow PhD student Jatin Arora)

HLA heterozygote advantage against HIV-1 is driven by quantitative and qualitative differences in allele-specific peptide presentation

The relationship between HIV and HLA is currently one of the best studied pathogen–HLA interactions, and the strong effect of HLA class I loci on AIDS outcomes is well established. HLA heterozygosity at class I loci (A, B, and C) has been associated with delayed AIDS progression, prolonged survival of HIV-infected individuals (Carrington et al. 1999) and decreased set point viral load (spVL) (McLaren et al. 2015). In this work, we explored the potential mechanism through which heterozygotes patients can obtain an advantage against HIV-1, using a large cohort of 6,311 HIV-1–infected individuals, with known HLA genotypes and set point viral load (spVL) information. A lower level of viral load in heterozygous individuals compared to homozygous at the HLA-B and HLA-C loci, and a negative correlation between pairwise allele divergence and viral load across individuals at the HLA-B locus confirmed the effects of both heterozygosity and allele divergence on AIDS outcomes. Computational prediction of HLA-presented antigenic HIV peptides was performed to characterize the quantitative effects of both heterozygosity and allele divergence. Heterozygotes were predicted to bind a broader array of HIV-1 peptides than do homozygotes at all three classical HLA loci, while the number of peptides bound by a pair of alleles was positively correlated with the sequence divergence between the alleles at the HLA-B locus, suggesting that heterozygote advantage at HLA-B is in part mediated by quantitative peptide presentation. Nevertheless, alleles significantly associated with HIV control were enriched in heterozygous patient, suggesting that HLA heterozygote

advantage might results also from a qualitative effect in antigen presentation, conferred by the higher probability to carry certain protective HLA alleles (i.e. HLA-B*57:01).

Annex II (collaborative project with group at MSKCC, NY)

Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy

The increasing interest in immune-based therapies for cancer has been recently driven by the success of checkpoint blockade immunotherapy. This innovative therapy uses immune checkpoint inhibitors (ICI), antibody able to target inhibitory receptors on T cells, in order to reinvigorate antitumor immune responses. A crucial step for antitumor immune responses is the recognition of HLA-peptide complexes on cancer cells by T-cell receptors. It has been recently established that HLA heterozygosity at class I loci (A, B, and C) is associated with increased survival of cancer patients treated with ICIs, possibly owing to the ability in presenting a broader range of tumor antigens to T cells (Chowell et al. 2018). In this work, we explored the effect of sequence divergence at HLA class I genes on immune checkpoint inhibitors (ICI) efficacy in patients with melanoma and non-small cell lung cancer, treated with immune checkpoint inhibitors (ICI), for which clinical response information as well as cancer exome sequencing data were available. We observed that patients with high mean sequence divergence (i.e. mean of the three pairwise divergences at the three class I loci: HLA-A, -B, and -C) respond significantly better to ICI than patients with low sequence divergence. Moreover, increased sequence divergence of HLA-I genotypes was significantly associated with: i) number of candidate neopeptides, defined based on cancer exome sequencing and predicted via computational antigen-binding prediction algorithms; ii) clonality of TCR CDR3s. Our results suggest a link between the divergent allele advantage and immunotherapy: highly divergent HLA-I genotypes can present higher diversity of the neopeptide repertoire, influencing T cell clonal expansion, thus facilitating tumor control during immunotherapy.

Chapter II

Accurate genotyping of HLA immune genes from shotgun sequence data of modern and ancient DNA

Past and ongoing pathogen-mediated selection has been proposed to be one of the major factors affecting the genetic variability at HLA genes (Prugnolle et al. 2005b) but unlike for several other species, convincing evidence for pathogen-driven selection in humans is still awaited. In this light, the possibility to obtain HLA genotypes from modern and ancient low-coverage shotgun sequence data, which are nowadays routinely generated in population genomic studies, is of strong interest and will enable investigating the molecular signature associated with pathogen-mediated selection. However, HLA genes exhibit exceptional genetic variability that defies standard HLA genotyping pipelines, especially when it comes to low-coverage and highly degraded ancient DNA (aDNA) samples. In this chapter, we have established a new HLA genotyping pipeline ('aHLA-Seq') optimized for low coverage shotgun sequence data and evaluated its accuracy for modern and ancient DNA. The pipeline has been initially used in combination with an HLA target-enrichment approach, to analyze HLA polymorphism in aDNA from medieval samples. After assessing the performance of the HLA target-enrichment approach applied to the medieval samples, we evaluated the accuracy of our HLA genotyping pipeline using a number of different and independent approaches, including the comparison with one independent HLA genotyping algorithm and with PCR-based results from a specific tag-SNP. Both comparisons provided pieces of evidence for the reliability of our allele calls; we thus described the HLA molecular profile for the medieval samples. The aHLA-Seq pipeline was further evaluated using simulated aDNA data with known HLA-B and -DRB1 alleles as well as 1000 Genomes Project data of modern humans for which HLA typing has been performed in an independent study. Overall, we show the reliability of the aHLA-Seq pipeline in genotyping HLA genes accurately in ancient and modern DNA samples, which might serve to disentangle the mechanisms of pathogen-driven selection throughout human history.

Chapter III

Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans

Leprosy is a chronic infectious disease caused by *Mycobacterium leprae*. It was endemic in Europe during the Middle Ages, from where it disappeared in the 16th century. Leprosy

nowadays remains a major public health problem affecting more than 200,000 people worldwide (World Health Organization 2016). The genomic comparison of ancient and modern strains of *M. leprae* revealed a remarkable genomic conservation during the last 1000 years. Because of that, the severity of the disease in the past, largely documented in historical records, has not been ascribed to changes in pathogen virulence but rather linked to other factors including host genetics (Andam et al. 2016). HLA genes are among the immune-related genes thought to influence susceptibility as well as disease progression of leprosy, with the allele DRB1*15:01 known to be a risk factor in different modern human populations (Zhang et al. 2009; Escamilla-Tilch et al. 2013; Zhang et al. 2016). This chapter describes the first genetic association study carried out using aDNA to date, which was performed to test whether the allele DRB1*15:01 also predisposed medieval Europeans to the disease. The frequency of the allele DRB1*15:01 was investigated using a tagSNP approach in 69 *M. leprae* DNA-positive cases collected from the St. Jørgen leprosarium in Denmark and compared with both contemporary medieval and modern controls. A statistically significant enrichment of DRB1*15:01 was detected in the medieval cases, suggesting that the risk allele predisposed also medieval Europeans to leprosy. Results were corroborated applying the HLA target-enrichment approach in combination with the aHLA-Seq pipeline described in chapter 2 to the 69 St. Jørgen samples. This analysis revealed the presence of the class II haplotype DRB1*15:01-DQB2*06:02, known to be a strong risk factor for inflammatory diseases in present-day populations. Furthermore, using a computational antigen-binding prediction approach, we explored the binding properties of common DRB1 alleles toward *M. leprae* peptides, which revealed limited relative HLA-presentation capacity of *M. leprae* antigens for DRB1*15:01 allele.

Chapter IV

Ancient history of HLA genes in the Americas

Archeological, historical and genetic studies indicate that Native Americans experienced a strong population bottleneck following the first European contact. It has been proposed that Native Americans' HLA genes may have lacked both general genetic polymorphism and specific resistance alleles to a variety of new pathogens introduced by European colonizers, resulting in an increased susceptibility to new diseases. In this chapter, we present the first spatiotemporal characterization of genetic variability of class II HLA loci (HLA-DRB1 and HLA-DQB1) performed to date in ancient Native American populations. The HLA target-enrichment approach in combination with the aHLA-Seq pipeline, both described in chapter 2, have been used to

explore HLA polymorphisms in ancient and present-day residents of the town of Xaltocan in central Mexico, and thus to investigate potential HLA allele frequency shifts from pre- to post-European contact period. The aHLA-Seq pipeline was further applied to available ancient whole-genome data of samples collected from different sites across the American continent, to eventually describe HLA polymorphisms in ancient Native American populations at a broad geographical and temporal scale.

List of papers / manuscripts

- Chapter I** **Pierini, F;** Lenz TL.: Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. *Molecular biology and evolution* (2018)
- Chapter II** **Pierini, F.;** Nutsua, M.; Boehme, L.; Özer, O.; Bonczarowska, J.; Susat, J.; Franke, A.; Nebel, A.; Krause-Kyora, B.; Lenz, TL.: Accurate genotyping of HLA immune genes from shotgun sequence data of modern and ancient DNA. Unpublished manuscript
- Chapter III** Krause-Kyora, B.; Nutsua, M.*; Boehme, L.*; **Pierini, F.*;** Pedersen, DD.*; Kornell, SC.; Drichel, D.; Bonazzi, M.; Möbus, L.; Tarp, P.; Susat, J.; Bosse, E.; Willburger, B.; Schmidt, HA.; Sauter, J.; Franke, A.; Wittig, M.; Caliebe, A.; Nothnagel, M.; Schreiber, S.; Boldsen, LJ.; Lenz, TL.; Nebel, A.: Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nature communications* (2018) *: equal contribution.
- Chapter IV** **Pierini, F.;** Reynolds, WA.; Balentine, MC.; Mata-Míguez, J.; Franke, A.; Nebel, A.; Krause-Kyora, B.; Bolnick, AD., Lenz, TL. Ancient history of HLA genes in the Americas. Unpublished manuscript
- Annex I** Arora, J.; **Pierini, F.;** McLaren, P.J.; Carrington, M.; Fellay, J.; Lenz, TL.: Quantitative and qualitative differences in allele-specific antigen repertoires underlie HLA heterozygote advantage against HIV-1. Unpublished manuscript
- Annex II** Chowell, D.*; Krishna, C.*; **Pierini, F.*;** Makarov, V; Rizvi, NA; Kuo, F; Riaz, N; Lenz, TL.; Chan TA. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. Manuscript in revision at *Nature Medicine*. *: equal contribution.

Author Contributions

- Chapter I** TLL and **FP** conceived the study. **FP** performed the bioinformatic analysis. **FP** analyzed the data with input from TLL. **FP** and TLL wrote the paper.
- Chapter II** 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, OO, JB and JS. AF provided research infrastructure. **FP** and TLL interpreted the results and wrote the manuscript with input from AN and BKK.
- Chapter III** BKK, AN, JLB and TLL designed the experiment. BKK, LB, DDP, SCK, MB, LM, JS, PT, EB, **FP** and MW performed the experiment. BKK, MNu, LB, **FP**, TLL, MB, DD, JLB, MNo and AC, analyzed the data. BW, AS and JS provided comparative data. AF, SS provided research infrastructure. BKK, JLB, DD, MNo, AC, TLL and AN interpreted the results. BKK, MNu, LB, JLB, DD, MNo, JS, AC, AF, **FP**, TLL and AN wrote the manuscript. BKK, TLL, and AN revised the paper.
- Chapter IV** TLL, BAD, **FP** and AWR conceived the study. **FP**, AWR, JMM and CMB performed the lab work. **FP** analyzed the data. BKK, AN and AF provided research infrastructure. **FP** interpreted the results and wrote the manuscript.
- Annex I** TLL, JA conceived the study. JF, PJM and MC provided the data. JA and **FP** analyzed the data. JA, TLL, **FP**, JF, PJM and MC interpreted the results. JA, TLL and **FP** wrote the manuscript.
- Annex II** DC, CK, **FP**, TLL and TAC conceived the study. DC, CK, **FP**, VM, TLL, TAC, FK, LGTM and NR performed data acquisition and analyses. DC, CK, **FP**, TLL and TAC wrote the manuscript with input from all authors.

Authors are given in alphabetical order:

AC: Amke Caliebe, AF: Andre Franke, AN: Almut Nebel, AS: Alexander H. Schmidt, AWR: Austin W. Reynolds, BAD: Deborah A. Bolnick, BKK: Ben Krause-Kyora, BW: Beatrix

Willburger, CK: Chirag Krishna, CMB: Christina M. Balentine, DC: Diego Chowell, DDP: Dorthe Dangvard Pedersen, EB: Beatrix Willburger, FK: Fengshen Kuo, **FP: Federica Pierini**, JA: Jatin Arora, JF: Jacques Fellay, JLB: Jesper L. Boldsen, JMM: Jaime Mata-Míguez, JS: Jürgen Sauter, JS: Julian Susat, LB: Lisa Boehme, LGTM: Luc G. T. Morris, LM: Lena Möbus, MB: Marion Bonazzi, MC: Mary Carrington, MNo: Michael Nothnagel, MNu: Marcel Nutsua, MW: Michael Wittig, NAR: Naiyer A. Rizvi, NR: Nadeem Riaz, PJM: Paul J. McLaren, PT: Peter Tarp, SCK: Sabin-Christin Kornell, SS: Stefan Schreiber, TAC: Timothy A. Chan, TLL: Tobias L. Lenz, VM: Vladimir Makarov

Chapter I

Divergent allele advantage at human MHC genes: signatures of past and ongoing selection

Federica Pierini¹ and Tobias L. Lenz¹

¹Max Planck Institute for Evolutionary Biology, Plön, Germany

Published in
Molecular Biology and Evolution (2018)
doi: 10.1093/molbev/msy116

Supporting information with supplementary figures and tables are available online.

Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection

Federica Pierini¹ and Tobias L. Lenz^{*,1}

¹Research Group for Evolutionary Immunogenomics, Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Ploen, Germany

*Corresponding author: E-mail: lenz@post.harvard.edu.

Associate editor: Claus Wilke

Abstract

The highly polymorphic genes of the major histocompatibility complex (MHC) play a key role in adaptive immunity. Divergent allele advantage, a mechanism of balancing selection, is proposed to contribute to their exceptional polymorphism. It assumes that MHC genotypes with more divergent alleles allow for broader antigen-presentation to immune effector cells, by that increasing immunocompetence. However, the direct correlation between pairwise sequence divergence and the corresponding repertoire of bound peptides has not been studied systematically across different MHC genes. Here, we investigated this relationship for five key classical human MHC genes (human leukocyte antigen; *HLA-A*, *-B*, *-C*, *-DRB1*, and *-DQB1*), using allele-specific computational binding prediction to 118,097 peptides derived from a broad range of human pathogens. For all five human MHC genes, the genetic distance between two alleles of a heterozygous genotype was positively correlated with the total number of peptides bound by these two alleles. In accordance with the major antigen-presentation pathway of MHC class I molecules, *HLA-B* and *HLA-C* alleles showed particularly strong correlations for peptides derived from intracellular pathogens. Intriguingly, this bias coincides with distinct protein compositions between intra- and extracellular pathogens, possibly suggesting adaptation of MHC I molecules to present specifically intracellular peptides. Eventually, we observed significant positive correlations between an allele's average divergence and its population frequency. Overall, our results support the divergent allele advantage as a meaningful quantitative mechanism through which pathogen-mediated selection leads to the evolution of MHC diversity.

Key words: HLA, balancing selection, heterozygote advantage, pathogen-mediated selection, human evolution.

Introduction

Pathogens are suspected to be one of the strongest selective forces in human evolution and the constant exposure to parasites over evolutionary time has likely contributed to the genetic variation found at a large number of genes within and among present day populations (Fumagalli et al. 2011). In this light, the genes of the major histocompatibility complex (MHC) with their exceptional genetic diversity and immune function are a prime candidate to investigate the exact mechanisms through which pathogen-mediated selection has contributed to human evolution.

The MHC is a key component of the adaptive immune system common to all jawed vertebrates (Klein 1986). In humans, it is known as a gene-dense region that spans ~4 Mb on the short arm of chromosome 6. It comprises over 200 genes, many of which are involved in immunity (Beck and Trowsdale 2000). Among these genes, the classical MHC genes (also called human leukocyte antigen, HLA) encode for cell-surface glycoproteins with a key role in adaptive immunity (Hughes and Yeager 1998; Trowsdale 2011). In cells infected by intracellular parasites, MHC class I molecules can present parasite-derived peptides to cytotoxic T lymphocytes (CTL). Upon recognition of these foreign peptides, the

infected cells are destroyed. The MHC class II molecules present antigens, mainly derived from extracellular pathogens, on the surface of specialized antigen-presenting cells. The exposed peptides are recognized by helper T lymphocytes (T_H cells), leading to a complex cascade of specific immune responses (Hughes and Yeager 1998; Jensen 2007; Neefjes et al. 2011).

The classical MHC genes are among the most polymorphic genes in the human genome and thousands of different alleles have been identified at some of these loci (Klein 1986; Trowsdale 2011). This polymorphism is characterized by a remarkable sequence variation in the peptide-binding grooves of MHC molecules (i.e., the pocket where antigens are bound) (Parham 1988; Reche and Reinherz 2003) as well as an enhanced rate of nonsynonymous substitutions (Hughes and Nei 1988). MHC polymorphisms are often ancient and allele lineages whose origin predates species divergence are retained across multiple species, an observation described as transspecies polymorphism (Klein 1987). The general action of balancing selection in enhancing both the rate of nonsynonymous substitutions in codons forming the peptide binding groove (Hughes and Nei 1988, 1989) and the persistence of allelic diversity over extremely long time

periods is strongly supported (Klein et al. 1998, 2007). However, the exact mechanisms of balancing selection are still disputed. Accordingly, three main mechanisms of pathogen-mediated selection have been suggested (Spurgin and Richardson 2010) which are potentially not mutually exclusive and may interact with one another: heterozygote advantage (Doherty and Zinkernagel 1975), rare-allele advantage (Bodmer 1972), and fluctuating selection (Hill 1991).

The heterozygote advantage was first proposed by Doherty and Zinkernagel (1975). Heterozygous individuals at MHC loci are assumed to present a broader range of pathogen-derived peptides than homozygotes, thus increasing the probability of triggering a specific immune response. They show increased resistance to pathogens, and are more likely to have higher relative fitness, resulting in an increased persistence of different MHC alleles in the population (Hughes and Yeager 1998; Penn et al. 2002). The heterozygote advantage hypothesis has been further extended by taking into account the sequence level, leading to the idea of a divergent allele advantage (Potts and Wakeland 1990; Wakeland et al. 1990). The high sequence divergence observed at MHC genes results in structural polymorphism that may impact the functional properties of MHC molecules. Heterozygous individuals with more divergent MHC allele combinations (i.e., larger number of amino acid differences along the sequence of the antigen-binding domains) are thought to encode glycoproteins that differ more in the repertoire of antigens they can bind. Those individuals may thus be able to present a wider array of antigens to immune effector cells, conferring an advantage against pathogen infections. In contrast, alleles more similar at the sequence level presumably exhibit more similar peptide binding specificities, thus leading to recognition of a lower overall number of peptides when co-occurring in a heterozygous individual (Lenz 2011).

Because of the extremely high number of pathogen proteins to which each host might be exposed throughout its lifetime, comprehensively measuring the relevant repertoire of MHC-bound peptides is impractical in humans and impossible in nonmodel species. Consequently, the divergent allele advantage hypothesis has been difficult to test. However, different measures of MHC sequence divergence are increasingly being used as a proxy for the potential MHC-bound peptide repertoire diversity, leading to correlative evidence that highlights how selection has favored the evolution of multiple MHC loci with divergent alleles in natural populations (She et al. 1990; Landry et al. 2001; Richman et al. 2001; Forsberg et al. 2007; Neff et al. 2008; Lenz et al. 2009; Schwensow et al. 2010; Lenz, Eizaguirre, et al. 2013; Lenz, Mueller, et al. 2013). In humans, the development of computational MHC antigen-binding prediction algorithms has enabled a more direct test of the divergent allele advantage. With this approach it has been shown previously that more divergent *HLA-DRB1* allele pairs experience less overlap in the antigenic peptides they can bind, that is, they are able to present a broader range of potential antigens (Lenz 2011), thus supporting the divergent allele advantage hypothesis at this locus. The investigation of the *DRB1* locus has been

further extended by considering two distinct phylogenetic groups of alleles, denoted as group A and B (Yasukochi and Satta 2014). The same pattern of increased pathogen recognition capacity was observed only for those alleles that in the phylogenetic tree cluster together with primate alleles forming a polyphyletic group (group B) (Lau et al. 2015). Recently, a mechanism of joint divergent asymmetric selection acting on *HLA-A* and *B* as a whole was suggested, which has potentially evolved to counter-balance the lack of diversity at individual HLA loci often found in small-sized and isolated human populations (Buhler et al. 2016).

In order to evaluate the divergent allele advantage hypothesis more systematically across all key classical human MHC genes, we here investigated the relationship between sequence divergence and peptide binding properties for three class I genes (*HLA-A*, *-B*, *-C*) and two class II genes (*HLA-DRB1* and *-DQB1*). Focusing on “common” alleles for each locus, as defined by the CWD catalogue (Mack et al. 2013) (supplementary table S1, Supplementary Material online), we evaluated different estimates of amino acid sequence divergence as proxies for the functional divergence among different alleles. Functional divergence was characterized by allele-specific computational binding prediction for a broad range of representative human pathogens (supplementary table S2, Supplementary Material online). Considering a larger and more comprehensive data set of pathogen-derived peptides compared with the set of pathogenic peptides that has been used in previous studies, we were also able to describe the functional features of the divergent allele advantage, by investigating the differential pattern of antigenic presentation between MHC class I and class II loci. Finally, the frequency distribution of HLA allele pools was investigated in several European populations in order to explore ongoing selection for divergent MHC alleles in modern humans.

Results

Functional Characterization of Common Human MHC Alleles

The set of 232 proteins from a broad collection of relevant human pathogens ($N = 27$), including macroparasites, bacteria, and viruses, resulted in a total of 118,097 unique pathogen-derived peptides. These peptides are meant to represent a comprehensive repertoire of potential antigens to which humans may have been exposed to throughout their evolutionary history, and which may thus have contributed to the exceptional MHC diversity that we see in present-day human populations.

The number of alleles defined as *common* varied among the different MHC loci: *HLA-A*: 63, *HLA-B*: 123, *HLA-C*: 40, *HLA-DRB1*: 73, *HLA-DQB1*: 21 (supplementary table S1, Supplementary Material online), reflecting general differences in allelic diversity among the loci (Trowsdale 2011). The proportion of peptides predicted to be bound by a given allele varied substantially within and among the different loci (table 1 and supplementary fig. S1, Supplementary Material online). For each locus, the proportions of common (shared among different alleles) and private (allele-specific) peptides

Table 1. Proportion of Bound Peptides across the Five Classical MHC Loci.

Locus	Proportion of Peptides Bound By At Least One Allele	Proportion of Bound Peptides Per Allele		
		Min	Max	Median (95% CI)
<i>HLA-A</i>	0.185	0.014	0.074	0.017 (0.0159, 0.0178)
<i>HLA-B</i>	0.192	0.012	0.018	0.017 (0.0168, 0.0172)
<i>HLA-C</i>	0.079	0.014	0.019	0.017 (0.0165, 0.0171)
<i>HLA-DRB1</i>	0.025	0.002	0.005	0.003 (0.0026, 0.0031)
<i>HLA-DQB1</i>	0.011	0.001	0.004	0.002 (0.0015, 0.0027)

NOTE.—MHC allele-specific peptide binding was predicted computationally for each locus. Total number of peptides: 118,097.

bound by each allele are reported in [supplementary figure S2, Supplementary Material](#) online. The overlap in bound peptides between different loci was significantly higher for class I genes ($A \cap B = 8,253$, $B \cap C = 5,930$, $A \cap C = 5,692$, $A \cap B \cap C = 3,545$) than for class II gene ($DRB1 \cap DQB1 = 133$) (χ^2 test, $P < 0.001$) ([supplementary fig. S3, Supplementary Material](#) online). Interestingly, we also found a remarkable amount of bound peptides that was shared between class I and class II genes ($A \cap DRB1 = 1,377$, $A \cap DQB1 = 385$, $B \cap DRB1 = 933$, $B \cap DQB1 = 403$, $C \cap DRB1 = 689$, $C \cap DQB1 = 251$) ([supplementary fig. S4, Supplementary Material](#) online). However, since the binding prediction algorithms do not account for different antigen processing pathways, it remains to be explored whether in reality these shared peptides are presented by alleles from both classes of MHC. The number of combined peptides bound by any two alleles of a given locus (equivalent to a heterozygous genotype) differed significantly among the genes (Kruskal–Wallis test, $P < 0.001$) ([supplementary fig. S5a, Supplementary Material](#) online). Generally, class II genes showed a lower median number of peptides bound by allele pairs (*HLA-DRB1* = 587, *HLA-DQB1* = 446) compared with class I genes (*HLA-A* = 3,936, *HLA-B* = 3,753, *HLA-C* = 3,349).

Sequence-Based Divergence Parameters and Functional Differences among MHC Alleles

A growing number of studies are investigating the fitness consequences of MHC allele divergence both in natural populations and in model species by using some estimate of sequence dissimilarity as a proxy for functional divergence among alleles. These estimates range from simple nucleotide differences (only partially relevant at the protein level) to sophisticated methods that take into account different physicochemical properties of amino acids at the protein sequence level.

For instance, the pairwise amino acid p-distance simply counts the relative number of differences along the amino acid sequence, but information on amino acid properties and relationships are not incorporated, and all nonidentical amino acids are treated as equivalent (Henikoff 1996). However, as the substitution rate usually varies among amino acid site, methods including information about different mutation rates for each amino acid as well as scores that take into account residue-specific properties have been introduced (May 1999). DayHoff (Dayhoff et al. 1978) and JTT (Jones

et al. 1992) are two examples of the most popular methods mainly used to investigate mutational trajectories and evolutionary distances between amino acids. Additionally, quantitative measures of pairwise distance have been developed, in which the physicochemical properties of the amino acids, and thus the functional similarity between sequences are considered. Among the different physicochemical features, the molecular volumes of amino acid residues might be particularly meaningful for the question whether a peptide fits into the various pockets of the peptide-binding groove of an MHC molecule. Grantham (Grantham 1974) and Sandberg (Sandberg et al. 1998) distances are two examples of sequence-based measures of sequence divergence where the molecular volume of the different amino acids is taken into account.

So far, a comprehensive evaluation has been lacking as to which sequence parameters are most suitable, that is, most strongly correlated with functional divergence. Thus, in order to identify the most relevant sequence-based parameter, the allele-specific functional binding properties at each MHC locus were correlated with five different, commonly used measures of sequence divergence: pairwise amino acid p-distance (Henikoff 1996), DayHoff (Dayhoff et al. 1978), JTT (Jones et al. 1992), Grantham (Grantham 1974) and Sandberg (Sandberg et al. 1998).

For each possible allele pair at a given HLA locus, we calculated the total number of unique pathogen-derived peptides (obtained by computational binding prediction as described earlier) bound by both alleles together (meant to reflect the MHC-presented antigen repertoire conferred by a heterozygote genotype). The pairwise number of bound peptides was then correlated with the sequence divergence of the two given alleles, estimated by the different measures. As expected, the five measures of genetic distance were highly correlated with each other ([supplementary fig. S6 and table S3, Supplementary Material](#) online). Accordingly, the correlation values between allele divergence and the combined number of bound peptides by all possible allele pairs at each locus under investigation were largely consistent across the different parameters of sequence divergence ([supplementary table S4, Supplementary Material](#) online). Nevertheless, despite similar correlation values, a rank analysis across the five human MHC gene (*HLA-A*, *-B*, *-C*, *-DRB1*, and *-DQB1*) revealed that the Grantham distance measure consistently ranked at the top, that is, showed the strongest correlation values ([table 2](#)). Therefore, for subsequent analyses, we focused on one parameter only, the Grantham distance.

Table 2. Rank analysis between different measures of sequence divergence.

	Total		Extracellular		Intra-Extra		Intracellular	
	Average Tau	Rank	Average Tau	Rank	Average Tau	Rank	Average Tau	Rank
P-distance	0.321	2	0.265	3	0.297	2	0.297	2
Dayhoff	0.313	3	0.257	4	0.285	5	0.297	2
JTT	0.309	5	0.256	5	0.286	4	0.287	4
Grantham	0.327	1	0.272	1	0.298	1	0.305	1
Sandberg	0.310	4	0.269	2	0.289	3	0.282	5

Note - Average correlation values (Kendall's tau coefficient) across the five human MHC genes (*HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1*) and rank analysis across the five parameters of sequence divergence.

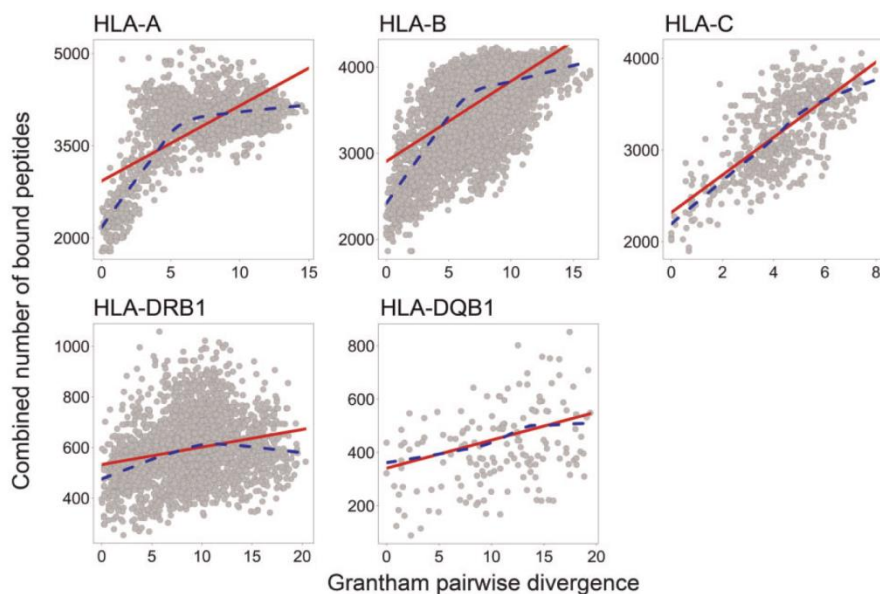


Fig. 1. In silico evidence for divergent allele advantage across five classical human MHC genes. Correlation between pairwise genetic distances reported as Grantham distance (x axes) and number of bound peptides (y axes) counted for all possible pairs of common HLA alleles. Each dot represents an allele pair. Binding prediction analyses performed on the complete data set of pathogen proteins ($n = 232$). Linear model (red line) and smoothed loess curve (dashed blue line), describing the association between the combined number of bound peptides and pairwise Grantham sequence divergence. Note the different axis scales.

Relation between Sequence Divergence and Functional Divergence

The median distance between allele pairs of a given locus differed significantly between the five genes (Kruskal–Wallis test, $P < 0.001$) with class II genes showing larger median Grantham distances (*HLA-DRB1* = 9, *HLA-DQB1* = 11) compared with class I genes (*HLA-A* = 7, *HLA-B* = 8, *HLA-C* = 5) (supplementary fig. S5b, Supplementary Material online).

For the sake of completeness, we first used our data to confirm the intuitive assumption of the heterozygote advantage hypothesis that allele pairs (representing heterozygous genotypes) together bind a larger number of peptides than single alleles (representing homozygous genotypes). This was generally true for all five loci (Kruskal–Wallis test, $P < 0.001$; supplementary fig. S7, Supplementary Material online), even though there were some rare cases where certain alleles alone bound

more peptides than certain allele combinations, suggesting interesting variation in peptide promiscuity among alleles. This general result is in line with a large body of empirical studies showing higher pathogen resistance for MHC heterozygotes (Carrington et al. 1999; Penn et al. 2002).

Subsequently, we focus all our analyses on allele pairs with two different alleles (reflecting heterozygous genotypes). According to the divergent allele advantage hypothesis, we expect that the number of peptides bound by heterozygote genotypes increases with increasing sequence divergence between the two given HLA alleles. Following this expectation, all five HLA genes revealed a significant positive correlation between the pairwise genetic distance and the combined number of bound peptides across all possible allele pairs (fig. 1 and table 3). Interestingly, for *HLA-A*, *-B*, and *-DRB1*, the rate at which the number of bound peptides increases in

Table 3. Divergent Allele Advantage and Different Pathogen Groups.

Pathogen Groups	HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1	
	Tau	P_{adj}	Tau	P_{adj}	Tau	P_{adj}	Tau	P_{adj}	Tau	P_{adj}
Total	0.361	<0.001	0.397	<0.001	0.507	<0.001	0.157	<0.001	0.210	0.001
Extracellular	0.345	<0.001	0.192	<0.001	0.392	<0.001	0.130	<0.001	0.303	<0.001
Intra-Extra	0.351	<0.001	0.289	<0.001	0.475	<0.001	0.166	<0.001	0.210	0.024
Intracellular	0.293	<0.001	0.377	<0.001	0.544	<0.001	0.137	<0.001	0.172	<0.001

NOTE.—Correlation values (Kendall's tau) between combined number of bound peptide and Grantham genetic distance between all possible allele pairs across the five key classical MHC genes. Binding prediction was performed on the complete data set of pathogen proteins ($n = 232$) as well as considering proteins separately within three groups of pathogens: extracellular ($n = 57$), intracellular ($n = 100$), and intra-extracellular ($n = 75$).

Tau, Kendall's tau coefficient; P_{adj} , P value after Bonferroni-correction across multiple alleles tested at each locus and number of loci.

response to larger allele divergence appears to slow down after a certain point, seemingly approaching a maximum. This can be explained by the fact that, for some loci, even allele pairs with only intermediate sequence divergence do not share any bound peptides anymore (supplementary fig. S8, Supplementary Material online). As the combined number of bound peptides cannot be larger than the sum of peptides bound by each allele, as soon as zero overlap is reached, any further sequence divergence cannot increase the combined number of bound peptides any further. This suggests that alleles at some loci can diverge functionally with only a small number of sequence changes, possibly at sites located in the peptide binding region.

Furthermore, a significant negative correlation could be observed between pairwise genetic distance and the proportion of shared peptides (peptides bound by both alleles of a given combination; supplementary fig. S8, Supplementary Material online), revealing a decreasing proportion of peptides shared between more divergent alleles. Of note, for the two class II loci, the correlations between genetic distance and peptide sharing were stronger than between genetic distance and the total number of bound peptides. This is owing to the fact that the latter measure includes additional variation from differences in the size of the bound peptide repertoire among HLA class II alleles (supplementary fig. S2, Supplementary Material online), an allele-specific property that is independent of the divergence between alleles.

Phylogenetic analysis of the *HLA-DRB1* gene has revealed two subgroups of allelic lineages: a human-specific monophyletic group (Group A), and a polyphyletic group with primates (Group B) (Yasukochi and Satta 2014). It has been proposed that group A and B allele lineages have evolved with contrasting binding capacity, and only the alleles from the polyphyletic group B showed increased presentation of pathogen peptides with increasing sequence divergence (Lau et al. 2015). In contrast with previous findings, our binding prediction analysis revealed a significant positive correlation for the whole set of *DRB1* alleles (fig. 1 and table 3) as well as for both groups of alleles separately (supplementary fig. S9, Supplementary Material online). This discrepancy to the earlier results might be due to the much larger and more comprehensive data set of pathogen-derived peptides used in our analysis (here 118,097 peptides vs. 265 peptides in the previous study).

Antigen Processing and Different Origins of Pathogenic Peptides

Up to this point, our analysis treated all pathogen peptides as equally likely targets for each given MHC locus. However, in reality some of those peptides will never be in contact with certain MHC molecules, due to the different processing pathways by which eukaryotic cells degrade proteins: the proteasome and lysosomal proteases. Peptides resulting from proteasome degradation, generally derived from intracellular proteins, are presented by MHC class I molecules, whereas peptides presented by MHC class II molecules are usually of extracellular origin and processed through lysosomal protease degradation in antigen-presenting cells (Jensen 2007). If the exceptional sequence divergence among MHC alleles evolved at least partly as a consequence of pathogen-mediated selection for divergent alleles, we would expect to see a stronger signature of selection at a given MHC locus when focusing on biologically meaningful pathogens that are likely to actually be encountered by a given MHC molecule. That is, we expect a stronger correlation between sequence divergence and functional divergence when only focusing on peptides originating from the locus-specific antigen-processing pathway. We thus divided the pathogen proteins used for binding prediction analysis into three groups. The three groups were based on their agent's lifestyle in the host: "extracellular" ($n = 58$ proteins), "intracellular" ($n = 100$) and a third group of pathogen proteins belonging to those agents whose life cycle involves both intracellular and extracellular stages inside the host (Silva 2012), here named as "intra-extracellular" ($n = 75$) (supplementary table S2, Supplementary Material online). Within each of the three groups of pathogen proteins, and for all five investigated HLA genes, we again observed a significant positive correlation between the pairwise sequence divergence and the combined number of bound peptides across all possible allele pairs. Interestingly however, in some cases, the strength of correlation differed among the three pathogen groups: at two of the MHC class I genes, *HLA-B* and *HLA-C*, a stronger positive correlation was observed for the group of intracellular pathogens, compared with extracellular pathogens, while this bias was not observed at *HLA-A* (fig. 2 and table 3). Conversely, a stronger correlation for peptides derived from extracellular pathogens was detected at one MHC class II locus, *HLA-DQB1*; while correlation values between pairwise sequence divergence and the combined

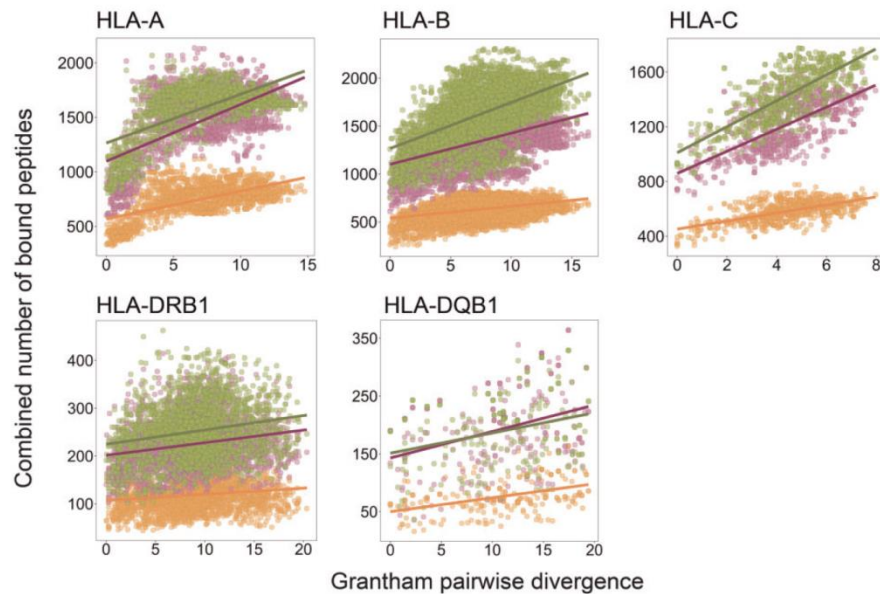


Fig. 2. Different origins of pathogenic peptides. Correlation between pairwise genetic distances reported as Grantham distance (x axes) and number of bound peptides (y axes) calculated for all possible pairs of common HLA alleles. Each dot represents an allele pair. Binding prediction analyses performed considering proteins within three groups of pathogens: extracellular (orange, $n = 57$ proteins), intracellular (green, $n = 100$), and intra-extracellular (purple, $n = 75$).

number of bound peptides at the *HLA-DRB1* locus were comparable across the three groups of pathogen proteins (fig. 2 and table 3). In the three cases where a bias across the groups of pathogen proteins was detected, the intra-extracellular proteins showed intermediate correlation values between the intra- and the extracellular proteins (table 3). However, permutation of the proteins among the three groups revealed that only for *HLA-B* and *HLA-C*, the observed difference between extracellular and intracellular correlation values were indeed larger than expected by chance (both $P < 0.001$; supplementary fig. S10, Supplementary Material online), while the other three loci did not show a statistically significant deviation from random expectations.

Our main analyses were performed considering pathogen-derived peptides that are all of the same length (9 aa). However, MHC class II molecules allow binding longer peptides than class I. Thus, for alleles at the two class II loci, binding prediction of 15mer peptides from the same set of pathogen proteins were considered. This analysis showed stronger correlations for *DRB1* and weaker correlations for *DQB1*, compared with the 9mer predictions, but overall support the main conclusions (supplementary table S5, Supplementary Material online). Furthermore, to test if our results were sensitive to the choice of the binding threshold, we additionally repeated the primary analysis using a different binding threshold (%rank of 0.5, indicating strong binding). The comparable results suggest that our main conclusions hold across a range of established binding thresholds (supplementary table S6, Supplementary Material online).

Distinct Amino Acid Composition among Pathogen Groups

The observed bias in the correlation between sequence divergence and functional divergence (peptide binding) toward a specific group of pathogens suggests distinct differences in the peptide repertoires among these groups. This could be due either to certain group-specific peptide sequences or to a more general difference in the amino acid composition of proteins among the pathogen groups. Amino acid usage has changed over evolutionary time in different species, and proteins have evolved in terms of physico-chemical and structural properties, reflecting adaptations to specific environmental conditions (Bogatyeva et al. 2006; Tekaiia and Yeramian 2006). Intriguingly, intra- and extracellular environments exhibit significant differences, including different pH value (Casey et al. 2010) and availability of different nutrients (Goetz et al. 2001; O’Riordan and Portnoy 2002; Ross 2014). It thus appears plausible that intra- and extracellular pathogens may have evolved proteomes with distinct amino acid compositions.

In order to test if the observed bias in correlation values across the groups of pathogen proteins was the results of group-specific peptide sequences or due to more general differences in the amino acid composition within each group, we created four different data sets of artificial proteins. These four data sets were then analyzed in the same way as above, again assessing the strength of correlation between allele divergence and functional divergence (here based on bound peptides from the artificial proteins) for the three pathogen groups and across the five HLA loci. First, amino acids forming

each pathogen protein sequence were randomly shuffled, maintaining the same amino acid composition of a given protein, but changing its actual sequence. If the stronger correlation with intracellular pathogen proteins by *HLA-B* and *HLA-C* were due to group-specific peptide sequences, we would expect this bias to disappear when reshuffling the protein sequences. However, correlation values resulting from shuffled proteins did not differ substantially from the true observed correlation values obtained with the real data set of pathogen proteins (supplementary table S7, Supplementary Material online). For the second set of artificial proteins, we created random protein sequences but maintained amino acid frequencies as they occurred within each group of pathogen proteins. Again, correlation values did not differ substantially from the observed true correlation values obtained in the initial test and the specific bias across the three groups was still observed (supplementary table S7, Supplementary Material online). For the third set of artificial proteins, we again created random protein sequences but this time maintained amino acid frequencies as they occurred in the whole data set of pathogen proteins. While a general positive correlation was also observed in this data set, the specific bias across the groups of pathogen protein was not detected anymore (supplementary table S7, Supplementary Material online). Finally, the amino acid composition computed from UniProtKB/Swiss-Prot data bank (Gasteiger et al. 2005; Boutet et al. 2016) was used to create the fourth set of artificial proteins. Again, the specific bias across the three groups of pathogen proteins was not detected anymore (supplementary table S7, Supplementary Material online). Thus, the observed bias in correlation values persisted only when amino acid frequencies mirrored the specific frequencies observed within each group of pathogen proteins. One of the possible explanations of our results could be that the observed stronger correlation of *HLA-B* and *HLA-C* alleles with peptides from intracellular pathogens is not due to specific peptides but is the results of adaptation of MHC alleles to the differences in amino acid composition between groups of pathogens. To further explore this hypothesis, a nonparametric multivariate analysis of variance was performed to quantify the similarity among proteins with regard to their amino-acid composition. The two groups of intracellular and extracellular pathogen proteins indeed differed significantly in their amino acid composition (PerMANOVA test, $P < 0.001$), with 9% of the total variance associated with the divisions in intracellular and extracellular proteins (R^2 estimate from PerMANOVA) (fig. 3). Accordingly, when average amino acid compositions were compared between the two groups of pathogen proteins, significant variations in the mean amino acid composition were observed for specific amino acids (one-way ANOVA, $P < 0.05$) (supplementary fig. S11 and table S8, Supplementary Material online). The observed differences in amino acid composition could be linked to glycosylation patterns, which differ between extracellular and intracellular peptides (Marshall 1972). N-glycosylation is one of the forms of protein glycosylation in eukaryotic organisms which is mainly targeting extracellular and secreted proteins. It has been shown that N-glycosylation sites are specific to the

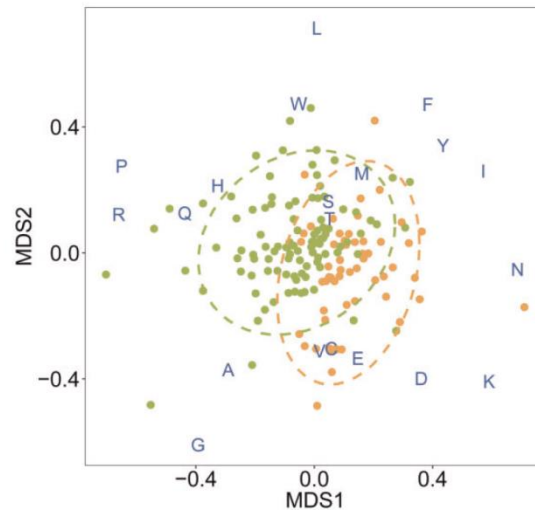


Fig. 3. Multidimensional scaling plot of amino acid composition in intracellular and extracellular pathogen proteins. Multidimensional scaling (MDS) based on amino acid frequencies indicates similarity in amino acid composition among individual proteins (dots). Intracellular proteins ($n = 100$) are reported in green while extracellular proteins ($n = 57$) in orange. MDS enables a standardized unit-less representation of variation among data points in 2D space (along perpendicular axes MDS1 and MDS2): location of proteins within the plot is indicative of potential bias toward specific amino acids (blue characters in one letter code), proteins with more similar amino acid composition are displayed closer to each other. The dashed circles indicate 95% confidence intervals for each group. Stress for 2D representation: 0.21.

consensus sequence Asn-Xaa-Ser/Thr and that the presence of proline between Asn and Ser/Thr inhibits N-glycosylation (Bause 1983). Accordingly, in our analysis proteins of extracellular pathogens show low proline concentration which is instead prevalent in proteins of intracellular pathogens. These exploratory analyses suggest that the amino acid composition might be different between the two groups of intracellular and extracellular pathogen proteins and that MHC alleles might have potentially adapted their binding specificities accordingly, at least at the *HLA-B* and *HLA-C* loci. However, further research is necessary to validate this conclusion and to exclude other potential causes, such as taxonomy, driving the observed difference in amino acid composition.

Population Frequency of Divergent HLA Alleles

The above-described results support historical pathogen-mediated selection through divergent allele advantage at the human MHC. However, we were also interested in exploring whether the divergent allele advantage was still maintaining diverse HLA allele pools in present day human populations. We hypothesized that, under the divergent allele advantage, alleles that on an average yield a more divergent genotype (conferring higher fitness) when paired with another allele in a heterozygote individual, would be selected for and thus exhibit higher frequencies in a given population.

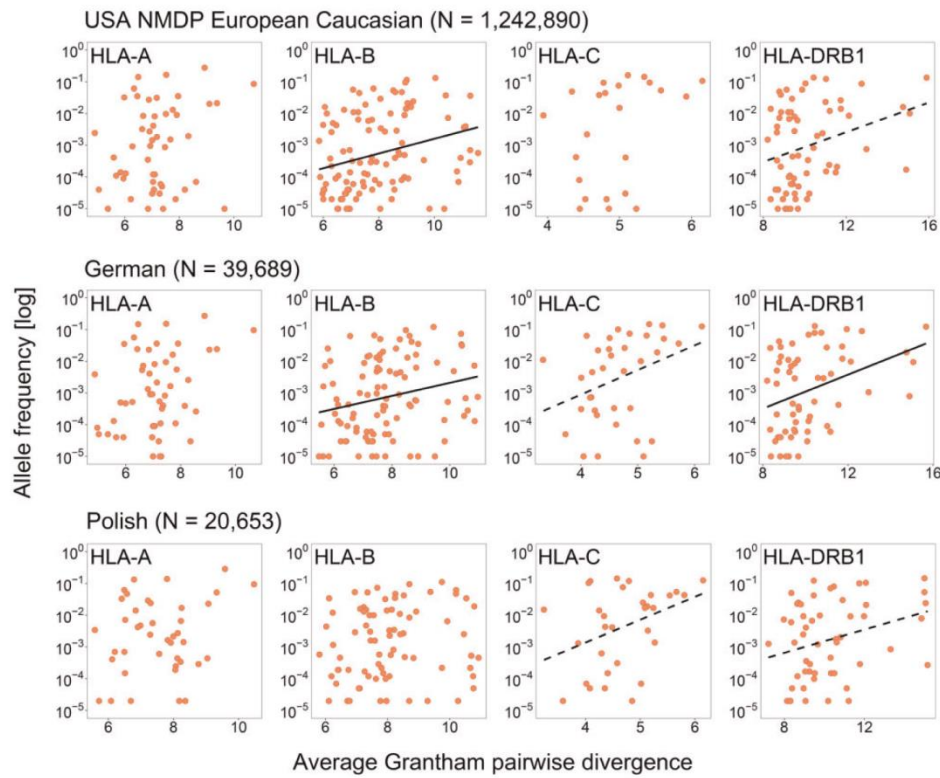


Fig. 4. Population frequency of divergent HLA alleles. Correlation between the average Grantham pairwise divergence to the most common alleles and the allele frequency in the USA NMDP European Caucasian ($N = 1,242,890$), German ($N = 39,689$), and Polish ($N = 20,653$) populations, for four classical HLA loci with available allele frequency data in AlleleFrequencies.net. Significant associations that persisted after Bonferroni correction across populations are reported with a solid line, while dashed lines indicate associations that are only nominally significant ($P < 0.05$ before Bonferroni correction; for exact values see supplementary table S9, Supplementary Material online).

A similar observation had been made in the allele pool of a social marine mammal whose reproductive success is partly predicted by the divergence of its MHC genotype (Lenz, Mueller, et al. 2013). In order to test this hypothesis across the five classical HLA loci, we calculated for each allele of a given locus the average pairwise amino acid sequence divergence to the most common alleles at this locus ($\geq 5\%$ allele frequency, representing alleles most likely to be forming a heterozygote with the allele in question). This average sequence divergence was then correlated with the allele's population frequency. For this analysis, we focused again on alleles defined as “common” in the CWD catalogue (Mack et al. 2013), assuming that very rare alleles are more susceptible to stochasticity and neutral demographic processes. In line with our expectation, a significant positive correlation between the average pairwise sequence divergence of an allele and its population frequency was observed for a number of HLA genes (*HLA-B*, *-C*, and *-DRB1*) across different European populations (USA European Caucasian, German, and Polish) (fig. 4 and supplementary table S9, Supplementary Material online).

Discussion

Here, we used computational antigen-binding prediction on a large data set of potentially antigenic pathogen peptides to investigate whether pairs of highly diverged MHC alleles together bind more different antigens than more similar alleles. Such an association is predicted by the *divergent allele advantage* hypothesis (Wakeland et al. 1990) and had previously been investigated only for the *HLA-DRB1* gene in humans (Lenz 2011; Lau et al. 2015). The observed positive correlation between the genetic distance of two alleles and the combined number of peptides they bind together confirmed and extended the predictions of the divergent allele advantage for all five investigated human MHC genes. These results support the hypothesis that enhanced sequence diversity between alleles in a heterozygous MHC genotype increases the range of potential MHC-presented peptides, thus raising the chance to recognize specific antigens and consequently enhance immune surveillance.

Our observation that *HLA-B* and *HLA-C* (and potentially *HLA-DQB1*) exhibit the strongest associations when considering antigens originating from their most plausible targets is intriguing and lends further support to the biological

relevance of this mechanism. It might indicate that the exceptional sequence divergence commonly observed among alleles of a given MHC locus has evolved specifically in response to selection by pathogens that are processed through the major protein degradation and antigen-presentation pathways that this given locus is associated with: alleles at *HLA-B* and *-C* loci have apparently evolved to bind specifically peptides derived from intracellular pathogens, while alleles at the *HLA-DQB1* locus may have evolved to bind a wider array of peptides from extracellular pathogens. The fact that we did not observe such a pathogen group-specific bias at the *HLA-A* and *HLA-DRB1* loci might indicate either that divergent allele advantage has not played a significant role in their evolution, or that they are less specific with regard to the pathogen origin of the peptides they present. It has indeed been shown that autophagy of intracellular components can promote the presentation of endogenous antigens by MHC class II molecules (Paludan et al. 2005; Levine and Deretic 2007; Münz 2012; Roche and Furuta 2015). Furthermore, several studies have reported a potential role for DRB1 molecules in viral infections (Martin and Carrington 2005). For instance, HLA-DR variants have been associated with spontaneous clearance of HBV and HCV infections (Thursz et al. 1995; McKiernan et al. 2004), with protective effect against dengue shock syndrome (DSS) development (Nguyen et al. 2008) and with HIV suppression (Malhotra et al. 2001). In this light, the lack of a particular bias by DRB1 alleles toward either of the pathogen groups may indicate that this locus evolves under selection by both intra- and extracellular pathogens.

In addressing the frequency of divergent HLA alleles in different human populations, we observed in some human populations, and for specific HLA loci, significant correlations between an allele's population frequency and its average pairwise sequence divergence. These results might suggest still ongoing selection for divergent HLA genotypes, at least in some modern human populations, possibly depending on population-specific differences in historical pathogen communities. However, allelic age may also contribute to the observed pattern, as, in principle, older alleles are both more likely to have reached high frequencies, even under neutrality, and to have accumulated more point mutations (thus being more divergent). On the other hand, HLA genes are known to undergo frequent recombination and gene conversion events, yielding novel alleles with high divergence from their origin at the very start (i.e., at low frequency). It is thus unclear to what extent novel alleles contribute to the observed pattern, warranting further research to explore the effect of genetic drift on the frequency of divergent HLA alleles. Furthermore, HLA alleles that on an average form more divergent allele combination, and which have been maintained in the population because of their increased capacity in presenting pathogen-derived peptides, might also be advantageous in case of newly emerging and fast-evolving pathogens (i.e., HIV).

Humans share similar MHC allelic lineages with closely related species (Klein 1987; Lawlor et al. 1988). This observation is a typical feature of MHC genes, compatible with the theory of transspecies evolution: ancestral lineages present in the common ancestor are inherited through successive

speciation events, persisting over long periods of time (Klein et al. 2007). The ancestral and highly diverged MHC variants are assumed to be adaptive and selectively maintained as a polymorphism by balancing selection (Hughes and Nei 1988; Lenz 2011). Recently, the role of adaptive introgression has been proposed to contribute to the exceptional level of polymorphism at the MHC (Abi-Rached et al. 2011; Wegner and Eizaguirre 2012). So far, explanations for the maintenance of introgressed MHC alleles have largely relied on the idea that such alleles were somehow locally adapted and thus beneficial. For instance, it has been suggested that modern humans might have maintained introgressed archaic HLA variants because they conferred an advantage against local pathogens (Abi-Rached et al. 2011). However, another explanation appears also plausible: MHC alleles from another species are, on average, likely to have diverged significantly from the species' own allele pool. Thus, any allele that introgresses from another species is likely to lead to highly divergent MHC genotypes. Following the divergent allele advantage hypothesis (and our results), such introgressed alleles should then confer a significant advantage and should consequently be selected for in the new species. This scenario would easily explain the maintenance of introgressed MHC alleles, but further research is necessary to support this hypothesis.

While the present analysis focuses exclusively on the divergent allele advantage, in reality, selection at MHC genes is a dynamic process that involves additional mechanisms apart from the divergent allele advantage. Conceptually, the divergent allele advantage can be considered a quantitative mechanism, which works independent of specific pathogen species or strains. It can act over long evolutionary time scales, promoting the maintenance of ancient allelic lineages in natural populations (Lenz 2011) and facilitating immunity against the constant simultaneous barrage by many different pathogens. In contrast, negative frequency-dependent selection (NFDS) is a qualitative mechanism in which specific alleles can be selected by specific pathogens (Slade and McCallum 1992; Lenz 2018). This mechanism likely works on a shorter time scale, for instance affecting MHC evolution in humans in very recent history (Lindo et al. 2016; Krause-Kyora et al. 2018). Both mechanisms, the divergent allele advantage and NFDS might also act in parallel, but at different time scales, creating an intriguing combination of shared polymorphism but distinct allele pools among populations and possibly even species (Lighten et al. 2017). Local adaptation plays another significant role in MHC evolution and might modulate the effect of the above mechanisms (Eizaguirre and Lenz 2010). The simultaneous action of these additional mechanisms might occasionally mask the effect of the divergent allele advantage and potentially explain the only sporadic evidence for this mechanism in the population frequency analysis reported here. Nevertheless, our results strongly support the divergent allele advantage as a meaningful quantitative mechanism through which pathogen-mediated selection contributes to the evolution of MHC diversity.

Materials and Methods

MHC Loci and Alleles Included in Analyses

Five key classical human MHC genes (*HLA-A*, *-B*, *-C*, *-DRB1*, and *-DQB1*) were analyzed in this study. Alleles at each locus were defined at second field (four-digit) resolution and only alleles annotated as “common” in the CWD catalogue (Mack et al. 2013) were included in the analyses. The allele annotation “common” in the CWD catalogue does not specifically indicate a high population frequency but more the extent and quality of documentation available for the given allele. This category indicates that there is universal agreement about the identity of this allele because it has been observed in multiple populations and there is sufficient data for robust frequency estimation (Mack et al. 2013). These criteria resulted in the analysis of 63 alleles for *HLA-A*, 123 for *HLA-B*, 40 for *HLA-C*, 73 for *HLA-DRB1*, and 21 for *HLA-DQB1* (supplementary table S1, Supplementary Material online).

Pathogen Proteins

Binding prediction analyses were performed on a data set of representative human pathogen proteins. Pathogens were selected from the Gideon database (Berger 2005) based on the following criteria: a global distribution, a potential for high mortality and/or morbidity, and a significant impact over the course of human history (Wolfe et al. 2007). The rationale for these criteria was that such pathogens are likely to have contributed significantly to human evolution in general and to the evolution of MHC genes in particular. Wolfe et al. (2007) provided a comprehensive list of infectious diseases with the greatest evolutionary and historical significance. From that list, we have taken the majority of pathogens in our data set. However, to assess mortality and morbidity, epidemiological data were also collected from two published reports: the Annual report of the European Centre for Disease Prevention and Control (European Centre for Disease Prevention and Control 2013) and the WHO Global Health Estimates (World Health Organization 2016). First, pathogens with the highest current mortality were included. However, not just mortality, but also nonfatal morbidity can be historically and evolutionarily significant. Indeed, morbid pathogens can reduce the fitness of their host in different ways (e.g., by increasing the sterility), thus pathogens considered morbid were also included. Finally, eradicated pathogens known to be important in human history were taken into account. Here, we used protein sequences of present day pathogens to explore signatures of historical selection, even though ancient pathogen strains might have differed slightly in their antigen repertoires. While we do not expect an effect on the general patterns observed here, it might be interesting to explore subtle differences in future work. We further aimed for a balanced representation of different groups of pathogens (i.e., viruses, bacteria, parasites). Based on these criteria, we identified 27 pathogens (10 viruses, 10 bacteria, 7 macro-parasites) that were classified into three groups: extracellular, intracellular, and intra-extracellular, based on their primary environment in the human body (supplementary table S2, Supplementary Material online). Then, for the selected

pathogens, amino acid sequences of 232 pathogen proteins (8.5 ± 5.8 per pathogen) known to be antigenic (Vita et al. 2015) and/or likely exposed to the host immune system (mostly secreted and surface proteins) (Rana et al. 2016) were obtained from GenBank (for accession numbers see supplementary table S2, Supplementary Material online).

Peptide Binding Prediction Algorithms

Computational antigen-binding prediction algorithms for MHC molecules were used to determine pathogen peptides potentially bound by the MHC alleles under investigation. Binding prediction was computed for all alleles at each of the five human MHC genes. Furthermore, as prediction analysis are likely to be more accurate for the core of the binding groove, which is known to be nine residues long and contributes the most to the recognition of the antigens, binding prediction was performed considering all possible 9mer pathogen-derived peptides. The data set of 232 representative human pathogen proteins described above resulted in a total of 118,097 unique pathogen-derived 9mer peptides that were analyzed using two different algorithms: NetMHCpan (v2.8) (Hoof et al. 2009) for the alleles at class I loci (*HLA-A*, *-B*, *-C*) and NetMHCIpan (v3.0) (Karosiene et al. 2013) for the alleles at class II loci (*HLA-DRB1*, *-DQB1*). For alleles at the two class II loci (*HLA-DRB1* and *HLA-DQB1*), we repeated the binding prediction analysis considering all possible 15mer pathogen-derived peptides. The predicted binding affinity between pathogen peptides and MHC molecule variants (defined in nanomolar IC₅₀, i.e., half maximal inhibitory concentration) are ranked by the respective software, based on comparison with a large pool of naturally occurring peptides, and a rank percentage score (%rank) is assigned to each peptide. To define “bound” peptides, we used the default %rank threshold of 2, which includes weak and strong binders. All analysis were also repeated using another established binding threshold (%rank of 0.5) which includes only strong binders. The allele *HLA-A*30:04* was predicted to bind about four times as many peptides as the other 62 *HLA-A* alleles (supplementary fig. S2, Supplementary Material online) and was thus excluded as an outlier from subsequent analysis in order to prevent distortion of results. The binding prediction analyses were performed first on the complete data set of pathogen proteins ($n = 232$), and then considering proteins within three groups separately: extracellular ($n = 58$), intracellular ($n = 100$), and intra-extracellular ($n = 75$).

Sequence Divergence

Allele divergence was computed on the same set of alleles used in the binding prediction analysis reported in supplementary table S1, Supplementary Material online. Protein sequences of HLA alleles were obtained from IMGT/HLA database (Robinson et al. 2015). Exons forming the variable region in the peptide binding groove (i.e., exon 2 and 3 for class I alleles and exon 2 for class II alleles) were selected following the annotation obtained from Ensemble database (Aken et al. 2016). Amino-acid sequence alignments were performed using MUSCLE (Edgar 2004), and sites containing alignment gaps at the beginning or the end of sequences were

removed. Genetic distances between alleles for all possible allele pairs at each locus were determined removing missing sites in pairwise comparisons and using five different pairwise parameters of allele divergence: p-distance (Henikoff 1996), DayHoff (Dayhoff et al. 1978), JTT (Jones et al. 1992), Grantham (Grantham 1974), and Sandberg (Sandberg et al. 1998). Pairwise amino acid p-distance, DayHoff and JTT distances were calculated in MEGA 7 (Kumar et al. 2016). Grantham and Sandberg sequence distances were calculated using a custom Perl script that required two input files: a FASTA file with aligned HLA alleles and a specific amino acid distance matrix. Grantham amino acid distance matrix was constructed from Grantham (1974). Sandberg amino acid distance matrix was calculated based on Euclidian distances between all 20 amino acids, using the Euclidian distance method in R version 3.4.1 (R Development Core Team 2017) according to the five physicochemical z-descriptors described in Sandberg et al. (1998): z1 (hydrophobicity), z2 (steric bulk), z3 (polarity), z4, and z5 (electronic effects). Our perl script (together with the Grantham amino acid similarity matrix) is freely available for download from SourceForge (<https://granthamdist.sourceforge.io/>). It can be used for calculation of pairwise Grantham divergence for any set of aligned MHC alleles of any species.

Allele Frequencies

Information about HLA allele frequencies in different human populations where obtained from the Allele Frequency Net Database (AFND) (Gonzalez-Galarza et al. 2015). We considered only populations of European ancestry with large sample sizes and for which frequencies of alleles at second field resolution were available: USA NMDP European Caucasian ($N = 1,242,890$), German ($N = 39,689$), and Polish ($N = 20,653$) populations. Furthermore, as with the analyses above, we focused on alleles defined as “common” in the CWD catalogue, which led to exclusion of some alleles with a frequency $< 1\%$. For each population, we first determined the most common alleles (allele frequency $\geq 5\%$) and for all the alleles under investigation in a given population, we calculated the average Grantham pairwise divergence to the most common alleles, considering all possible heterozygote genotypes.

Statistical Analyses

Correlation Tests

The Shapiro–Francia test was performed for all the parameters under investigation (i.e., measures of genetic distance, combined number of bound peptides and average Grantham pairwise amino acid divergence to the most common alleles) to explore samples’ distribution. As parameters were not normally distributed and tied ranks could be detected within our data, the nonparametric Kendall correlation was used to test for associations between parameters. When testing the association between sequence divergence and functional divergence, all P values were adjusted for multiple testing using a sequential Bonferroni correction across the number of alleles tested at each locus as well as across the number of different loci tested. When testing the association

between the allele’s average divergence and its population frequency, P values were corrected across the number of populations tested. Correlations were performed in R version 3.4.1 (R Development Core Team 2017).

Permutation Tests

To test for significant differences in the strength of correlation between allele divergence and the binding to pathogen group-specific peptides, we performed permutation tests. For this analysis, the set of 232 representative human pathogen proteins were randomly shuffled among the three groups of pathogens, maintaining the same number of proteins as observed in the original data (extracellular $n = 57$, intracellular $n = 100$ and intra-extracellular $n = 75$). For each group of pathogens, permuted proteins were used to perform binding prediction analyses and compute correlation values between genetic distances and combined number of bound peptides counted for all possible allele pairs for the five HLA genes (analogous to original analysis). Each permutation was run 1,000 times, and the difference between correlation coefficients for intracellular and extracellular proteins for the five HLA genes was recorded. If there was no significant bias for intracellular or extracellular pathogens, on average this difference should be zero. The distribution of permuted differences was then used to infer the significance of our initial observations using a one-tailed test with a 0.05 cut-off.

Artificial Proteins

Four sets of artificial proteins were created and analyzed to test for potential differentiation of the amino acid composition (AAC) among the three groups of pathogens. The first set of artificial proteins was created by randomly shuffling amino acids within each pathogen protein by using the Shuffle Protein program (Stothard 2000), thus maintaining the AAC of each protein intact. Three more sets of artificial proteins were created in R version 3.4.1 (R Development Core Team 2017) by assembling random amino acids while maintaining several features as they occurred within each of the three pathogen groups used in the initial test (i.e., the number of proteins, the average length of sequences, the SD of the length and the minimum and maximum length). The second set of artificial proteins was created from random amino acids but maintaining the AAC as it occurred within each group of pathogen proteins. The third set of artificial proteins was created from random amino acids, while maintaining amino acid frequencies as they occur in the whole data set of pathogen proteins. Finally, amino acid composition computed from UniProtKB/Swiss-Prot data bank (Gasteiger et al. 2005; Boutet et al. 2016) was used to create the fourth set of artificial proteins.

Multivariate Analysis of Variance

Multidimensional scaling is a multivariate statistical technique that can be used to display and summarize a high-dimensional data set in 2D graphical form. The technique was here applied to explore associations between subsets of pathogen proteins and amino acids. A nonparametric,

permutational multivariate analysis of variance (PerMANOVA) was used to test for differences in the amino acid composition between pathogen groups. The PerMANOVA, based on a Bray–Curtis dissimilarity distance matrix, was run with 999 permutations to tests for statistical significance. Both procedures are implemented in the vegan package (Oksanen et al. 2012) in R version 3.4.1 (R Development Core Team 2017).

Comparison of Average Amino Acid Compositions

Comparison of mean amino acid compositions between the two groups of pathogen proteins (extracellular and intracellular) were performed using one-way analysis of variance; all *P* values were adjusted for multiple testing using Bonferroni correction across the number of amino acids tested.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Martin Kalbe and Jamie Winternitz for their precious input and suggestions. We also thank Matteo Fumagalli and an anonymous reviewer for very constructive comments on an earlier version of this article. This work was supported by the Max Planck Society and an Emmy Noether grant from the German Research Foundation (DFG; LE 2593/3-1 to T.L.L.).

References

- Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA. 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 334(6052):89–94.
- Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, et al. 2016. The Ensembl gene annotation system. *Database* 2016:baw093.
- Bause E. 1983. Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem J*. 209(2):331–336.
- Beck S, Trowsdale J. 2000. The human major histocompatibility complex: lessons from the DNA sequence. *Annu Rev Genomics Hum Genet*. 1(1):117–137.
- Berger SA. 2005. GIDEON: a comprehensive Web-based resource for geographic medicine. *Int J Health Geogr*. 4(1):10.
- Bodmer WF. 1972. Evolutionary significance of the HL-A system. *Nature* 237(5351):139–183.
- Bogatyeva NS, Finkelstein AV, Galzitskaya OV. 2006. Trend of amino acid composition of proteins of different taxa. *J Bioinform Comput Biol*. 4(2):597–608.
- Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol Biol*. 1374:23–54.
- Buhler S, Nunes JM, Sanchez-Mazas A. 2016. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenet*. 68(6–7):401–416.
- Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ. 1999. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283(5408):1748–1752.
- Casey JR, Grinstein S, Orłowski J. 2010. Sensors and regulators of intracellular pH. *Nat Rev Mol Cell Biol*. 11:50–61.
- Dayhoff MO, Schwartz RM, Orcutt BC. 1978. [A model of evolutionary change in proteins]. *Atlas Protein Seq Struct*. 5:345–351.
- Doherty PC, Zinkernagel RM. 1975. Enhanced immunological surveillance in mice heterozygous at H-2 gene complex. *Nature* 256:50–52.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.
- Eizaguirre C, Lenz TL. 2010. Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. *J Fish Biol*. 77:2023–2047.
- European Centre for Disease Prevention and Control. 2013. Annual Epidemiological Report 2013. Reporting on 2011 surveillance data and 2012 epidemic intelligence data. Stockholm: ECDC.
- Forsberg LA, Dannewitz J, Petersson E, Grahm M. 2007. Influence of genetic dissimilarity in the reproductive success and mate choice of brown trout – females fishing for optimal MHC dissimilarity. *J Evol Biol*. 20:1859–1869.
- Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*. 7:e1002355.
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. *The proteomics protocols handbook*. Totowa (NJ): Humana Press. p. 571–607.
- Goetz M, Buber A, Wang G, Chico-Calero I, Vazquez-Boland JA, Beck M, Slaghuis J, Szalay AA, Goebel W. 2001. Microinjection and growth of bacteria in the cytosol of mammalian host cells. *Proc Natl Acad Sci U S A*. 98:12221–12226.
- Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, da Silva AL, Teles e Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR. 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res*. 43:D784–D788.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.
- Henikoff S. 1996. Scores for sequence searches and alignments. *Curr Opin Struct Biol*. 6:353–360.
- Hill AVS. 1991. HLA associations with malaria in Africa: some implications for MHC evolution. In: Klein J, Klein D, editors. *Molecular evolution of the major histocompatibility complex*. Heidelberg (Berlin): Springer Berlin Heidelberg. p. 403–420.
- Hoof I, Peters B, Sidney J, Pedersen L, Sette A, Lund O, Buus S, Nielsen M. 2009. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 61:1–13.
- Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170.
- Hughes AL, Nei M. 1989. Nucleotide substitution at major histocompatibility complex class-II loci – evidence for overdominant selection. *Proc Natl Acad Sci U S A*. 86:958–962.
- Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet*. 32:415.
- Jensen PE. 2007. Recent advances in antigen processing and presentation. *Nat Immunol*. 8:1041–1048.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 8:275–282.
- Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. 2013. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 65:711–724.
- Klein J. 1986. Natural history of the major histocompatibility complex. New York: John Wiley and Sons.
- Klein J. 1987. Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. *Hum Immunol*. 19:155–162.
- Klein J, Sato A, Nagl S, O'hUigin C. 1998. Molecular trans-species polymorphism. *Annu Rev Ecol Syst*. 29:1–21.

- Klein J, Sato A, Nikolaidis N. 2007. MHC, TSP, and the origin of species: from immunogenetics to evolutionary genetics. *Annu Rev Genet.* 41:281–304.
- Krause-Kyora B, Nutsua M, Boehme L, Pierini F, Pedersen DD, Kornell S-C, Drichel D, Bonazzi M, Möbus L, Tarp P, et al. 2018. Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nat Commun.* 9(1):1569.
- Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol.* 33(7):1870–1874.
- Landry C, Garant D, Duchesne P, Bernatchez L. 2001. 'Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). *Proc R Soc Lond B Biol Sci.* 268(1473):1279–1285.
- Lau Q, Yasukochi Y, Satta Y. 2015. A limit to the divergent allele advantage model supported by variable pathogen recognition across HLA-DRB1 allele lineages. *Tissue Antigens* 86(5):343–352.
- Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P. 1988. Hla-a and Hla-B polymorphisms predate the divergence of humans and chimpanzees. *Nature* 335(6187):268–271.
- Lenz TL. 2011. Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution* 65(8):2380–2390.
- Lenz TL. 2018. Adaptive value of novel MHC immune gene variants. *Proc Natl Acad Sci U S A.* 115(7):1414.
- Lenz TL, Eizaguirre C, Kalbe M, Milinski M. 2013. Evaluating patterns of convergent evolution and trans-species polymorphism at MHC immunogenes in two sympatric stickleback species. *Evolution* 67(8):2400–2412.
- Lenz TL, Mueller B, Trillmich F, Wolf JBW. 2013. Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. *Proc R Soc B Biol Sci.* 280(1762):20130714.
- Lenz TL, Wells K, Pfeiffer M, Sommer S. 2009. Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the Long-tailed giant rat (*Leopoldamys sabanus*). *BMC Evol Biol.* 9:269.
- Levine B, Deretic V. 2007. Unveiling the roles of autophagy in innate and adaptive immunity. *Nat Rev Immunol.* 7(10):767–777.
- Lighten J, Papadopoulos AST, Mohammed RS, Ward BJ, Paterson G, I, Baillie L, Bradbury IR, Hendry AP, Bentzen P, van Oosterhout C. 2017. Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nat Commun.* 8(1):1294.
- Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, Malhi RS. 2016. A time transect of exomes from a Native American population before and after European contact. *Nat Commun.* 7:13175.
- Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE, Kempenich JH, Reed EF, et al. 2013. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* 81(4):194–203.
- Malhotra U, Holte S, Dutta S, Berrey MM, Delpit E, Koelle DM, Sette A, Corey L, McElrath MJ. 2001. Role for HLA class II molecules in HIV-1 suppression and cellular immunity following antiretroviral treatment. *J Clin Invest.* 107(4):505–517.
- Marshall RD. 1972. Glycoproteins. *Annu Rev Biochem.* 41:673–702.
- Martin MP, Carrington M. 2005. Immunogenetics of viral infections. *Curr Opin Immunol.* 17(5):510–516.
- May AC. 1999. Towards more meaningful hierarchical classification of amino acid scoring matrices. *Protein Eng.* 12(9):707–712.
- McKiernan SM, Hagan R, Curry M, McDonald GS, Kelly A, Nolan N, Walsh A, Hegarty J, Lawlor E, Kelleher D. 2004. Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. *Hepatology* 40(1):108–114.
- Münz C. 2012. Antigen processing for MHC class II presentation via autophagy. *Front Immunol.* 3:9.
- Neeffes J, Jongsma MLM, Paul P, Bakke O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol.* 11(12):823–836.
- Neff BD, Garner SR, Heath JW, Heath DD. 2008. The MHC and non-random mating in a captive population of Chinook salmon. *Heredity* 101(2):175–185.
- Nguyen TP, Kikuchi M, Vu TQ, Do QH, Tran TT, Vo DT, Ha MT, Vo VT, Cao TP, Tran VD, et al. 2008. Protective and enhancing HLA alleles, HLA-DRB1*0901 and HLA-A*24, for severe forms of dengue virus infection, dengue hemorrhagic fever and dengue shock syndrome. *PLoS Negl Trop Dis.* 2(10):e304.
- Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, et al. 2012. vegan: community ecology package. Version R package version 2.0-3. <https://CRAN.R-project.org/package=vegan>.
- O'Riordan M, Portnoy DA. 2002. The host cytosol: front-line or home front? *Trends Microbiol.* 10(8):361–364.
- Paludan C, Schmid D, Landthaler M, Vockerodt M, Kube D, Tuschl T, Münz C. 2005. Endogenous MHC class II processing of a viral nuclear antigen after autophagy. *Science* 307(5709):593–596.
- Parham P. 1988. Function and polymorphism of human leukocyte antigen-A, B, C molecules. *Am J Med.* 85(6A):2–5.
- Penn DJ, Damjanovich K, Potts WK. 2002. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci U S A.* 99(17):11260–11264.
- Potts WK, Wakeland EK. 1990. Evolution of diversity at the major histocompatibility complex. *Trends Ecol Evol.* 5(6):181–187.
- R Development Core Team. 2017. R: a language and environment for statistical computing. Version Version 3.4.1. Vienna (Austria): R Foundation for Statistical Computing.
- Rana A, Thakur S, Bhardwaj N, Kumar D, Akhter Y. 2016. Excavating the surface-associated and secretory proteome of *Mycobacterium leprae* for identifying vaccines and diagnostic markers relevant immunodominant epitopes. *Pathog Dis.* 74:ftw110.
- Reche PA, Reinherz EL. 2003. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *J Mol Biol.* 331(3):623–641.
- Richman AD, Herrera LG, Nash D. 2001. MHC class II beta sequence diversity in the deer mouse (*Peromyscus maniculatus*): implications for models of balancing selection. *Mol Ecol.* 10(12):2765–2773.
- Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh Steven GE. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 43(Database issue):D423–D431.
- Roche PA, Furuta K. 2015. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat Rev Immunol.* 15(4):203–216.
- Ross AC. 2014. Modern nutrition in health and disease. Philadelphia: Wolters Kluwer Health/Lippincott Williams and Wilkins.
- Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. 1998. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem.* 41(14):2481–2491.
- Schwensow N, Eberle M, Sommer S. 2010. Are there ubiquitous parasite-driven major histocompatibility complex selection mechanisms in gray mouse lemurs? *Int J Primatol.* 31(4):519–537.
- She JX, Boehme S, Wang TW, Bonhomme F, Wakeland EK. 1990. The generation of MHC class II gene polymorphism in the genus *Mus*. *Biol J Linn Soc.* 41(1–3):141–161.
- Silva MT. 2012. Classical labeling of bacterial pathogens according to their lifestyle in the host: inconsistencies and alternatives. *Front Microbiol.* 3:71.
- Slade RW, McCallum HI. 1992. Overdominant vs. frequency-dependent selection at MHC loci. *Genetics* 132(3):861–862.
- Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc R Soc B Biol Sci.* 277(1684):979–988.
- Stothard P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* 28:1102, 1104.
- Tekaia F, Yeramian E. 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genomics* 7:307.

- Thursz MR, Kwiatkowski D, Allsopp CE, Greenwood BM, Thomas HC, Hill AV. 1995. Association between an MHC class II allele and clearance of hepatitis B virus in the Gambia. *N Engl J Med.* 332(16):1065–1069.
- Trowsdale J. 2011. The MHC, disease and selection. *Immunol Lett.* 137(1–2):1–8.
- Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, et al. 2015. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 43(Database issue):D405–D412.
- Wakeland EK, Boehme S, She JX, Lu CC, McIndoe RA, Cheng I, Ye Y, Potts WK. 1990. Ancestral polymorphisms of MHC class-II genes – divergent allele advantage. *Immunol Res.* 9(2):115–122.
- Wegner KM, Eizaguirre C. 2012. New(t)s and views from hybridizing MHC genes: introgression rather than trans-species polymorphism may shape allelic repertoires. *Mol Ecol.* 21(4):779–781.
- Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature* 447(7142):279–283.
- World Health Organization. 2016. Global health estimates 2015: deaths by cause, age, sex, by country and by region, 2000-2015. Geneva: World Health Organization.
- Yasukochi Y, Satta Y. 2014. A human-specific allelic group of the MHC *DRB1* gene in primates. *J Physiol Anthropol.* 33:14.

Chapter II

Accurate genotyping of HLA immune genes from shotgun sequence data of modern and ancient DNA

Federica Pierini¹, Marcel Nutsua², Lisa Böhme², Onur Özer¹, Joanna Bonczarowska², Julian Susat², Andre Franke², Almut Nebel², Ben Krause-Kyora², Tobias L. Lenz¹

¹Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Ploen, Germany, ²Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

Abstract

The human leukocyte antigen (HLA) genes play a crucial role in adaptive immunity and are associated with a vast number of diseases. Past and ongoing pathogen-mediated selection is assumed to be driving the extensive diversity at these genes, and a better understanding of these processes will shed light on the genetic architecture and prevalence of many present-day diseases. Shotgun genomic sequence data is becoming increasingly available from diverse population studies, including historical populations. Its use could facilitate the analysis of HLA genes in ancient and modern populations, which is otherwise prevented by prohibitive costs of targeted HLA genotyping. However, HLA genes exhibit exceptional genetic variability that defies standard mapping and genotyping approaches. Available HLA genotyping pipelines thus rely on deep and homogeneous coverage, and none of them has yet been adapted to deal with the small amount of DNA, the risk of contamination and the profound degradation observed in ancient DNA (aDNA) samples. Here, we present a new HLA genotyping pipeline ('aHLA-Seq') optimized for low coverage shotgun sequence data. In combination with an HLA target-enrichment approach, the pipeline has been used to analyze HLA polymorphism in aDNA from medieval samples. Further, the aHLA-Seq pipeline has been evaluated for use with ancient and modern sequence data using simulated aDNA and 1000 Genomes Project data. Our pipeline now enables association studies between HLA variability and complex genetic disorders in historical human populations, and could be further applied to explore HLA allele frequency changes through time when temporal sample series are available.

Introduction

Owing to their crucial role in tissue transplantation, molecules related to histocompatibility have been extensively studied over the last decades (Clarke and Kirby 1966). As discoveries regarding HLA molecules accumulated, their immunological function in presenting antigenic peptides on the cell surface became apparent, thus enabling the immune system to distinguish between 'self' and 'non-self', eventually stimulating an immune response (McDevitt and Tyan 1968; Doherty and Zinkernagel 1975a). At present, HLA genes are still among the most studied loci of the human genome for their implication in hundreds of different complex diseases (Trowsdale 2011; Matzaraki et al. 2017), but also because of their importance in human evolution (Tishkoff and Verrelli 2003; Meyer et al. 2018).

Initial classifications of HLA variants were based on antigenic characterization by either cytotoxicity or serological assays, which allowed discrimination between groups of related alleles. They have later been substituted by more exact classifications, based on the DNA sequence of corresponding genes. DNA-based typing techniques have further undergone significant development, starting from PCR-RFLP methods, SSOP immobilized probes, PCR-SSP, and Sanger sequencing (reviewed in Erlich (2012)). Since the advent of next-generation sequencing (NGS) technologies, various high-throughput HLA-typing protocols have been described (reviewed in Hosomichi et al. (2015) and Carapito et al. (2016)). HLA data produced through next-generation sequencing technologies cannot be addressed without sophisticated bioinformatics approaches, and thus several automatic analysis pipelines have been developed (Bauer et al. 2018). These new approaches promise a shift towards higher resolution and throughput, and raise the hope to prevent typing ambiguities associated with traditional methods (Sanchez-Mazas and Nunes 2018).

The application of HLA typing methodologies over the last three decades has revealed the extensive variability of HLA genes within and among different human populations (Solberg et al. 2008). At the molecular level, HLA genetic diversity is characterized by a remarkable amino acid sequence diversification (Parham 1988; Reche and Reinherz 2003) as well as an enhanced rate of non-synonymous substitutions (Hughes and Nei 1988) in the antigen binding groove of HLA molecules (i.e. the pocket where antigens are bound). Past and ongoing pathogen-mediated selection is proposed to be one of the major factors affecting this genetic variability (Hughes and Nei 1988, 1989; Spurgin and Richardson 2010).

The genetic diversity at HLA genes is also associated with various complex genetic disorders in contemporary humans (Trowsdale 2011), suggesting a link between historical selection by infectious agents and present prevalence of genetic disorders (Dean et al. 2002; Lenz et al. 2016). In this light, the recent development of genomic tools for the analysis of ancient DNA (aDNA) provides a unique opportunity to unravel the selection processes shaping the human genome (Marciniak and Perry 2017; Nielsen et al. 2017). 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 the identification of the exact targets of selection. A reliable genotyping of the HLA genes in ancient and modern samples is thus crucial to answer unresolved questions from human genetics to evolutionary medicine.

However, the high density of SNPs (Shiina et al. 2009) as well as the paralogous organization (Robinson et al. 2017) of the HLA genes make 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. However, DNA molecules extracted from skeletal remains are normally heavily fragmented and degraded through chemical modifications (Orlando et al. 2015), hence, sufficient genomic coverage of the endogenous DNA is often difficult to obtain, adding a further layer of complexity to reliable allele calls within the HLA region. To overcome these obstacles, we developed a novel aDNA-optimized analysis pipeline for low-coverage and low-quality shotgun sequence data, which we call 'aHLA-Seq'. In combination with a targeted DNA capture approach, the pipeline was successfully used to analyze HLA polymorphisms in a comprehensive set of historical human samples.

Target-enrichment by hybridization, also known as DNA capture approach, is now one of the most widely used sequencing approaches to study aDNA, because of its efficiency in increasing the sequence coverage of the endogenous fraction (Burbano et al. 2010; Maricic et al. 2010; Carpenter et al. 2013; Fu et al. 2013). Owing to this method, DNA fragments of interest can be selected by using specific DNA or RNA baits, so that the fraction of sequenced molecules aligning to the region of interest is enhanced (Cappellini et al. 2018). In this study, 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 (Robinson et al. 2015))(Wittig et al. 2015), was used to enrich a defined set of medieval European DNA libraries for the most polymorphic classical HLA class I (HLA-A, -B, -C) and class II (HLA-DRB1, -DQB1, -DPB1) genes.

To identify specific HLA allele combinations, our novel aDNA-optimized aHLA-Seq pipeline combines automated read selection and sorting, with highly repeatable semi-manual filtering and allele identification at up to 3rd field (6-digit) resolution (**Figure 1**). After pre-processing and quality control, genomic sequence reads 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). Mapped reads are then grouped by gene specificity, and saved into sample-specific FASTA files. Using a sequence alignment software, the FASTA files can be manually analyzed to genotype individual samples. A subset of the HLA class II data has already been analyzed in the context of medieval leprosy, yielding a leprosy-associated HLA-DRB1 risk allele (Krause-Kyora et al. 2018).

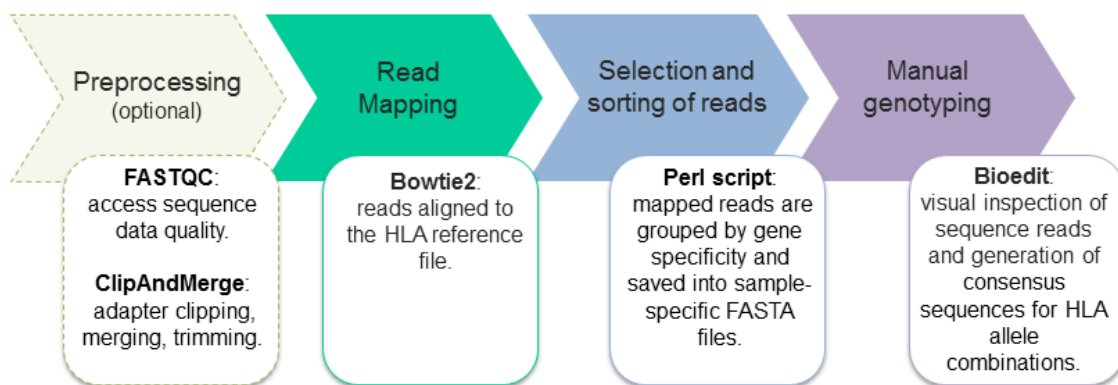


Figure 1 | Different steps performed by the aHLA-Seq pipeline for HLA genotyping of ancient and modern samples. Preprocessing (optional): after quality control, genomic sequences are pre-processed (adapter clipping, merging and trimming) using ClipAndMerge (version 1.7.3) from the EAGER pipeline (Peltzer et al. 2016). Mapping: performed using Bowtie2 (Langmead and Salzberg 2012) against a comprehensive reference file, containing known 3rd field HLA alleles (following G-group nomenclature). Sorting: mapped reads are grouped by gene specificity using a Perl script and saved into a FASTA file. HLA genotyping: Sample-specific FASTA files are manually analyzed using BioEdit (Hall 1999) to genotype HLA genes in ancient and modern samples.

After assessing the performance of the HLA target-enrichment approach applied to the medieval European samples, we employed a number of different and independent approaches to evaluate the accuracy of our HLA genotyping pipeline. This evaluation includes a comparison with the most applicable independent HLA genotyping algorithm currently available, with PCR-

based results from a specific tag-SNP, and with simulated aDNA sequence data with known HLA-B and -DRB1 alleles. Finally, we explored whether our HLA genotyping pipeline, initially developed for aDNA sequence data, can also be applied successfully to shotgun sequence data from modern populations. We thus applied our aHLA pipeline to a subset of the 1000 Genome Project samples, for which HLA typing has been performed previously (Gourraud et al. 2014).

Material and Methods

Sequence data from historical samples

Historical samples

The human skeletal remains whose DNA was analyzed in this study have been described in a previous study (Krause-Kyora et al. 2018). Briefly, the specimens 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 (Krause-Kyora et al. 2018). Sample processing, DNA extraction and DNA library preparation have previously been described Krause-Kyora et al. (2018). 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. (2015). The HLA capture probes have been originally created considering the full list of available cDNA and gDNA sequences from the IMGT/HLA reference database (Robinson et al. 2015) (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 (Wittig et al. 2015). 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 μ L. As the UDG-treated libraries were already indexed during library preparation, the 12 cycles of post-capture PCR was performed using 1 μ L of each IS5 and IS6 primers (100 μ M). 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 \times 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 (Peltzer et al. 2016). 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 (Jonsson et al. 2013), shotgun sequencing data from both UDG-treated and non-UDG-treated libraries were aligned against the *H. sapiens* reference genome hg38 (GRCh38) using Bowtie2 (Langmead and Salzberg 2012) v2.2.7, in a semi-global alignment mode and with default parameters, as described in Krause-Kyora et al. (2018). 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.

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) (Trowsdale and Knight 2013) 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 (Hall 1999); 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 heterozygous genotypes (**Table S10**). Typical bias observed in aDNA samples (fragmentation and damage patterns) were then introduced using the program gargammel (Renaud et al. 2017). For each genotype, we created five aDNA paired-ends read datasets with increasing coverages (1x, 5x, 10x, 30x, 60x), 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.

Sequence data from modern samples 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 (Auton et al. 2015). 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 S11**). Only samples with available SBT-based HLA genotype information published in Gourraud et al. (2014) were included (**Table S12-S13**).

HLA typing from shotgun sequence data (aHLA-Seq pipeline)

HLA reference file for read mapping

A key component of the aHLA 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) (Hollenbach et al. 2011). 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 (Robinson et al. 2015) (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 (Edgar 2004) 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 aHLA-Seq pipeline is qc-filtered and adapter-trimmed sequence data in fastq format. The first step maps the sequence reads of the sample to the HLA reference file using Bowtie2 (Langmead and Salzberg 2012) 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 (see **Figure S1**). 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(Hall 1999) 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 S1**). 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 (Mack et al. 2013). 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.

Allele call comparison to Optitype pipeline

To compare the allele call results of our aHLA-Seq pipeline with an independent method, sequence data of the historical samples were also analyzed using OptiType (Szolek et al. 2014) 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.

Statistical analysis

Allele frequencies 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 (Lancaster et al. 2007) 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 (Hedrick 1987). 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 (Cramer 1946). 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 (Slatkin and Excoffier 1996). Two- and three-locus haplotype frequencies were estimated from the ancient genotypic data, using the iterative expectation-maximization (EM) algorithm (Dempster et al. 1977; Excoffier and Slatkin 1995). The analyses were done removing individuals with NA at all loci while keeping only allele calls that reached the 2nd field level of resolution.

Evaluation of HLA allele calling pipeline

To assess the reliability of the aHLA-Seq 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.

Results

Performance of the HLA target-enrichment for historical samples

A dataset of sixty-eight medieval European samples was used for the investigation of ancient HLA polymorphisms. 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 (Hofreiter et al. 2001; Briggs et al. 2007; Brotherton et al. 2007). 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 (Briggs et al. 2010). 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 mis-incorporation patterns (Briggs et al. 2007).

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. (Jonsson et al. 2013). The analysis of damage patterns in the final nucleotide of the sequenced fragments revealed mis-incorporation frequencies of up to 2.6% in UDG-treated and up to 21.9% in non-UDG-treated datasets. This confirmed that most of the reads mapping to the human reference originate from aDNA fragments.

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. To investigate the highly polymorphic classical HLA class I (HLA-A, -B, -C) and class II (HLA-DRB1, -DQB1, -DPB1) genes, an HLA target-enrichment approach was applied to the UDG-treated libraries, here defined as HLA-enriched UDG libraries. 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 have been quantified on both original UDG libraries and HLA-enriched UDG libraries for a subset of 62 samples. 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 S1**). 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, $p < 0.001$; **Figure 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 S3**). Also, the median % of reads aligning to HLA genes calculated over the whole set of samples differed significantly between original (0.0001%) and HLA-enriched UDG libraries (0.1210%) (Mann-Whitney, $p < 0.001$; **Figure 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 S1**). The median of coverage over the HLA genes calculated across the whole set of samples was significantly higher after the enrichment approach (96%) compared to the median coverage observed in the original shotgun sequence data (9%) (Mann-Whitney, $p < 0.001$; **Figure 2C** and **Table 1**; for calculation at individual HLA loci see **Figure S2** and **Table S2-S3**). Similarly, the median read depth quantified for the whole set of samples was significantly lower when considering the UDG shotgun libraries (0.2x) compared to sequence data obtained after the HLA target enrichment (17.4x) (Mann-Whitney, $p < 0.001$; **Figure 2D** and **Table 1**; for calculation at individual HLA loci see **Figure S3** and **Table S2-S3**).

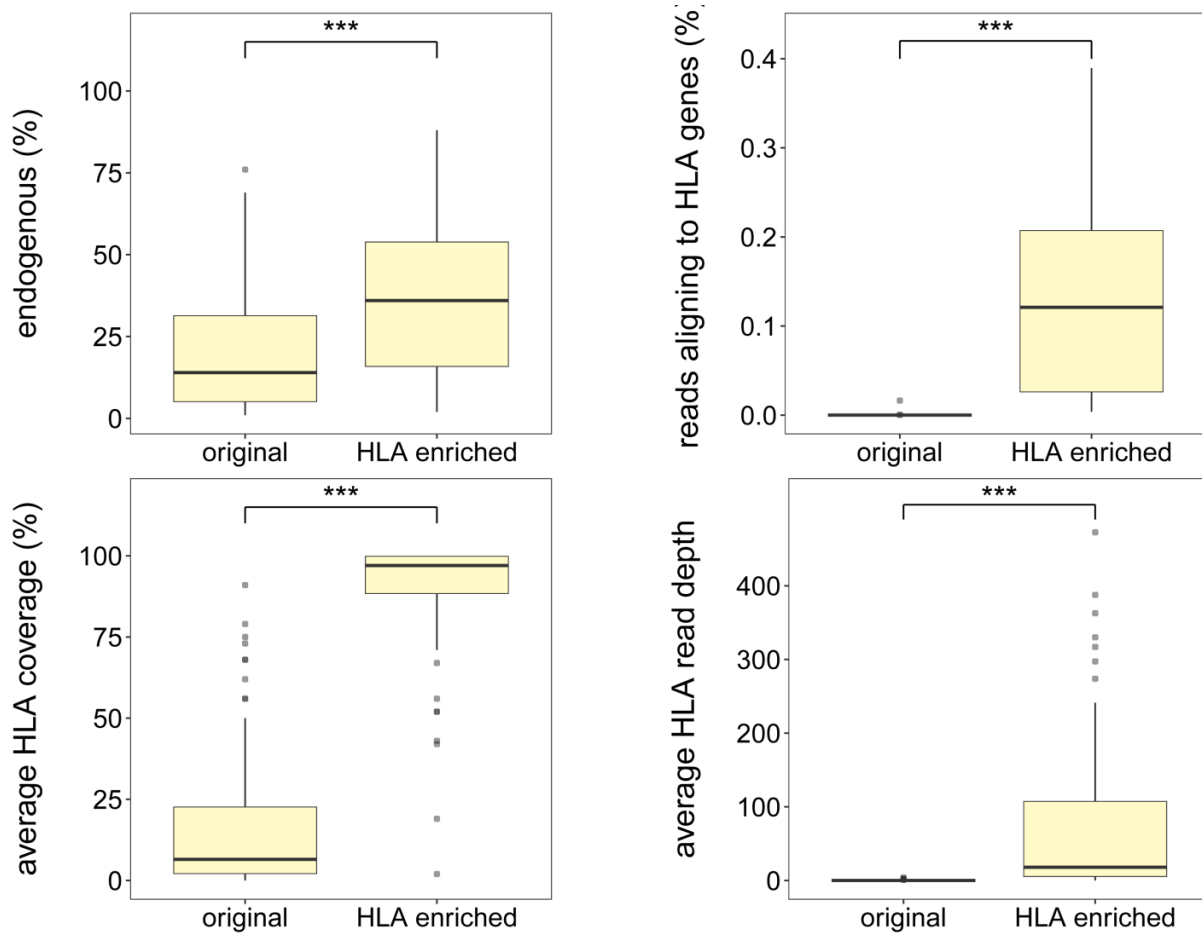


Figure 2 | Performance of HLA target-enrichment experiments. Comparison between the median values of (a) endogenous DNA content (b) percentage of reads aligning to HLA genes c) average HLA coverage and (d) average HLA read depth calculated across a subset of 62 aDNA samples before and after performing the HLA enrichment experiments. Significant differences between median values, as derived from Mann-Whitney test, are indicated by horizontal line and asterisks (***) $p < 0.001$.

Table 1 | Performance of HLA target-enrichment experiments

	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)

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.

HLA typing from aDNA of historical samples

To call HLA genotypes of the historical samples, all the sequence data generated from UDG-treated libraries were combined for each sample and processed through our aHLA-Seq pipeline. According to the official HLA nomenclature, allele definition at HLA gene is defined by the gene name indicating the locus (HLA-A, -B, -C, -DR, -DQ, -DP), followed by a hierarchical numbering system (Robinson et al. 2015). The 1st field (2-digit) defines different allele groups, resembling the traditional classification provided by serotyping. The 2nd field (4-digit) separates alleles differing 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 group alleles that differ in their nucleotide sequence only outside the peptide-binding domains and thus bind the same peptide repertoires. Using this G-group nomenclature, our aHLA-Seq pipeline allows HLA typing at up to 3rd field 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. 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 S4**). Of the 136 alleles investigated at each locus, we were able to call at the 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 **Figure 3** and **Table 3**). As expected, a significant positive correlation between coverage and success rate (1st field, $\tau = 0.76$, $p < 0.001$; 2nd field $\tau = 0.76$, $p < 0.001$; **Figure S4**) 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 S5**) was observed, indicating a considerable effect of DNA preservation on allele call success. These associations can explain some cases where DNA quality and thus read coverage did not allow for more precise allele calls (**Table 2**). 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 evidences, together with the examination of mis-incorporation frequencies shown above, indicate that the human DNA analysed in each sample is likely to be endogenous.

Table 2 | Locus-specific allele call success for the 68 historical samples

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

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.

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 the CEU population (de Bakker et al. 2006). 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 our aHLA-Seq pipeline (Krause-Kyora et al. 2018). The study confirmed that the tag SNP allele rs3135388-T always co-occurred with the allele DRB1*15:01 (N = 13), when 2nd field resolution was reached, but also co-occurred with the DRB1*15 allele call (N = 20) in samples where only the broader 1st field resolution was possible. An interesting observation could be done in that

study regarding the specific haplotype structure. 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 S4**), suggesting a strong LD between these two loci (Krause-Kyora et al. 2018).

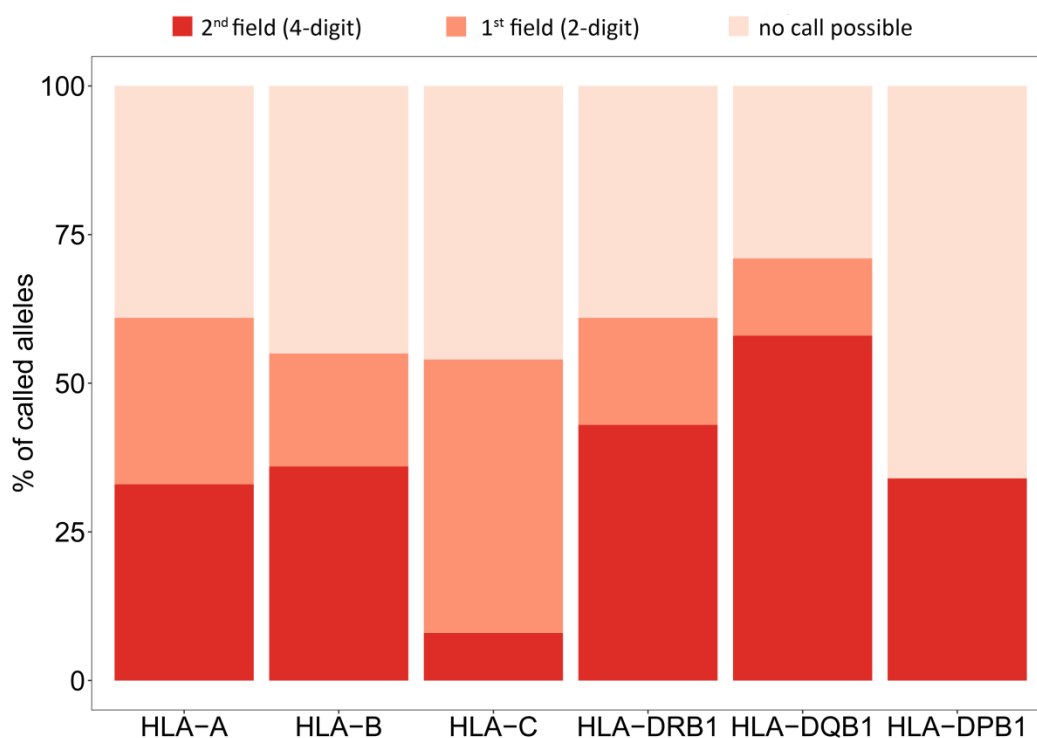


Figure 3 | HLA allele call success rate for 68 historical aDNA samples. Percentage of alleles called at each individual HLA locus calculated for the whole set of historical samples (N=68). Allele calls are reported at two different level of resolution 2nd field (4-digit) and 1st field (2-digit) levels.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

Table 3 | Success rate for the 68 historical samples

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

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.

HLA class I allele call comparison with OptiType

To our knowledge, OptiType (Szolek et al. 2014) is currently the only HLA genotyping algorithm for which the influence of both coverage depth and read length on prediction accuracy has been tested, and for which allele calls appear reliable even for sequence data containing short reads or only a 10x average coverage depth over the HLA class I loci (Szolek et al. 2014). These features make OptiType potentially suitable for studying HLA genes from aDNA. However, the algorithm is currently available only for the typing of HLA class I genes (Szolek et al. 2014). To validate our allele calls with an independent method, a random subset (N = 39) of the historical samples were analyzed using OptiType (Szolek et al. 2014) v1.3.1, and HLA class I genotype information was compared with allele calls from our aHLA-Seq pipeline. For some of the ancient samples, our approach indicated that low DNA yield and quality, and correspondingly low read coverage, made reliable allele calls impossible at any of the investigated HLA genes (**Table S2** and **Table S4**). While OptiType always calls an allele, no matter how spurious the sequence information, the reliability of its allele calls is likely to suffer similarly from such data limitations. As we were not able to estimate the accuracy of OptiType, we therefore excluded those instances from the comparisons.

Dividing the number of alleles with identical call from 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 S5**). The number of called alleles that differed between the two approaches was 12. One advantage of our aHLA-Seq 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 (note **Table S5**). We observed that although our approach has a lower success rate in the allele call compared to OptiType, it can likely provide a higher accuracy by avoiding low quality/confidence allele calls.

HLA molecular profile of the historical St. Jørgen samples

As HLA class II data have been already analyzed in a previous work, (Krause-Kyora et al. 2018) we here focus on describing the allele frequency distributions at HLA class I genes (-A, -B and -

C), reported in **Table S6** and **Table S7**. Twelve distinct alleles with a 1st field level of resolution have been observed at HLA-A. The A*02 lineage comprises 31% of the total, with the most common subtype A*02:01 seen at a frequency of 0.244. The second most common lineage is represented by A*03, with the most common subtype A*03:01 also found at a frequency of 0.244. The next two more common lineages are A*01 and A*24, for which the most common subtypes are A*01:01 ($f = 0.156$) and A*24:02 ($f = 0.133$). The others lineages (A*26, A*68, A*11, A*32) are found at frequencies of lower than 10%; while the lineages A*29, A*30 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 alleles have been 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 1st field level allele B*42 and the 2nd field level 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 alleles with a 1st field level of resolution were observed. The three more common lineages (C*07, C*03 and C*04) comprise together over 70% of the total. 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.

The permutation distribution of the likelihood ratio test, implemented in Pypop (Lancaster et al. 2007), was used to test the significance of the overall LD between any two loci (**Table S8**). 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 higher than expected levels of LD have been documented in the HLA region (Carrington et al. 1994; Sanchez-Mazas et al. 2000; Armuzzi et al. 2003), 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 S8**). Locus A showed significant association with locus B ($D' = 0.804$; $W_n = 0.844$), C ($D' = 1$; $W_n = 1$), DRB1 ($D' = 0.703$; $W_n = 0.755$) and DQB1 ($D' = 0.764$; $W_n = 0.709$), while no global LD was observed with the DPB1 locus. Locus B showed a nonrandom association with all loci: C ($D' = 0.833$; $W_n = 0.829$), DRB1 ($D' = 0.882$; $W_n = 0.775$), DQB1 ($D' = 0.877$; $W_n = 0.772$) and DPB1 ($D' = 0.829$; $W_n = 0.823$). Global LD has been revealed between locus C and DRB1 ($D' = 0.833$; $W_n = 0.882$) and between locus C and DQB1 ($D' = 0.875$; $W_n = 0.913$). A strong nonrandom associations between DRB1 and DQB1 genes ($D' = 0.984$; $W_n = 0.883$) have been observed, while no global LD between the DP locus and the other class II

genes have been found. Two- and three-locus haplotype frequencies were estimated using the expectation-maximization algorithm. For loci showing significant association the allele names at each analyzed locus, the maximum likelihood estimate for the haplotypes frequencies sorted in decreasing order and the corresponding approximate number of haplotypes are reported in **Table S9**. As previously described, we confirmed the presence of the class II haplotype DRB1*15:01-DQB2*06:02, found at a frequency of 27%. Common haplotypes are 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 II haplotype, B*07:02-DRB1*15:01-DQB1*06:02 ($f = 0.286$). Other two extended class II haplotype involving HLA-A and HLA-B loci were also observed at high 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 two- and three-locus common haplotypic associations were observed reported in **Table S9**.

Pipeline validation on simulated aDNA samples

To evaluate our HLA genotyping pipeline, we simulated aDNA sequence data. Seven haplotypes, with known HLA-B and -DRB1 alleles, were generated and combined in six different genotypes (**Table S10**). Using the program gargammel (Renaud et al. 2017) typical aDNA fragmentation and damage pattern were introduced in the sequences and the pipeline tested for increasing depth of coverages from 1x up to 60x, for a total of 30 simulated ancient samples. The HLA allele calls from our aHLA-Seq pipeline (here only run on the most polymorphic genes HLA-B and HLA-DRB1) were compared to the original known HLA genotypes. If allele calls were not possible because of limited sequence coverage, we reported 'NA'. Allele calls with several equally well matching alleles from different 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, was 71% at the 1st field level and 57% at the 2nd field level (**Table 4**). In this test, all of the allele calls were correct, thus we observed an accuracy rate of 100% for both the 1st field and 2nd field levels (**Table 4**). We further tested if our aHLA-Seq pipeline can detect unknown alleles, such as HLA alleles that were present in the past but no longer exist in modern populations or extremely rare alleles that are not represented in the HLA reference database. The three point mutations introduced in the sixth genotype were all well detected starting from a coverage depth of 5x (**Table 4**). Considering that the 30 ancient simulated samples were quite variable in the range of coverage depth (from 1x up to 60x), we tested if the success rate was correlated with the coverage depth. Association between

coverage depth and success rate was observed at both the 1st field and 2nd field levels (**Figure S6**); as expected, success in defining HLA allele increases with higher read depth (**Figure 4**).

Table 4 | Success rate and accuracy rate for simulated ancient DNA samples

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

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.

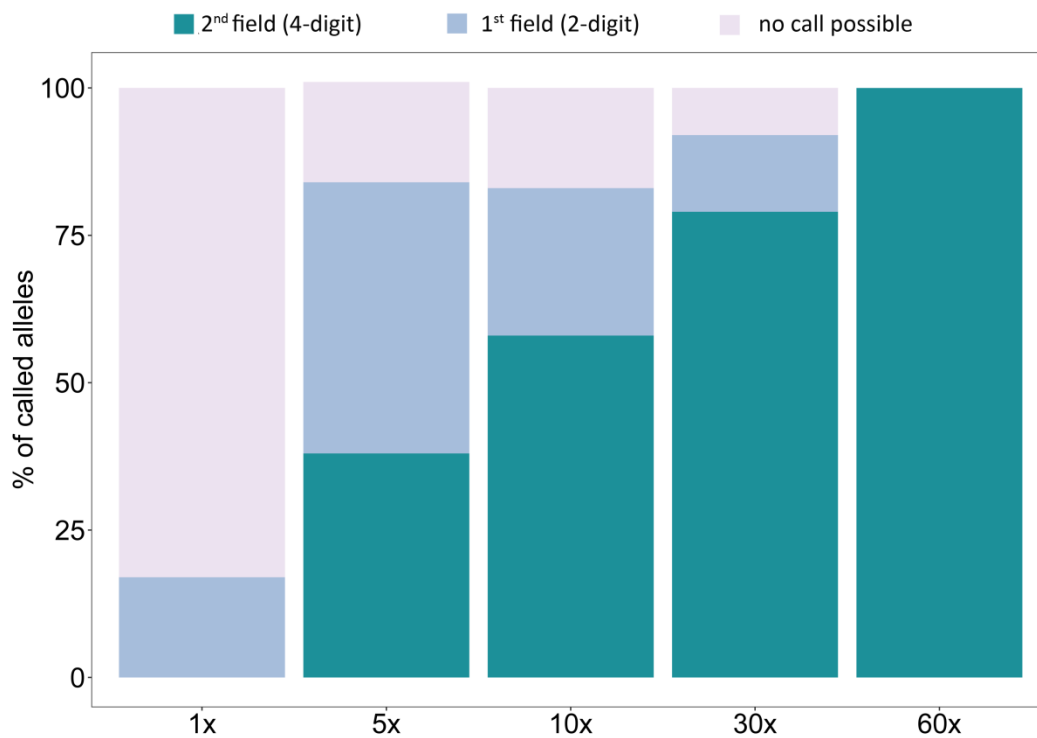


Figure 4 | HLA allele call success rate for the simulated ancient samples. Percentage of allele calls calculated across the two investigated loci (HLA-B and HLA-DRB1) and across the set of samples (N=6) investigated at different coverage depths (from 1x up to 60x) for a total of 30 simulated ancient 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.

Pipeline validation on 1000 Genomes Project samples

The aHLA-Seq pipeline was initially developed with the aim to define HLA alleles from aDNA sequence data. However, since shotgun low-coverage resequencing 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 aHLA-Seq 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 was available from a previous study (N = 31; **Table S11, S12 and S13**). The success rate for allele calling 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 G-group level (**Table S14**). 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**). The majority of the allele calls obtained through our pipeline were consistent with the ones obtained through independent PCR-based genotyping in Gourraud et al. (2014), as evidenced by the high agreement rate of 96% (**Table S12, S13 and S14**). 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. (2014) 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. (2014) but instead clearly confirmed our allele calls, showing that our pipeline can rectify erroneous calls from PCR-based HLA genotyping.

Table 5 | Success rate and accuracy rate for the 1000 Genomes Project samples

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

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.

Discussion

The investigation of HLA genes in ancient and modern human populations is of strong 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 aHLA-Seq pipeline for HLA genotyping from low-quality shotgun data, and evaluate its accuracy for modern and ancient DNA. 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 (Haak et al. 2015; Lazaridis et al. 2016), whole chromosomes (Fu et al. 2013; Cruz-Davalos et al. 2018), mitochondrial (Briggs et al. 2009; Krause et al. 2010; Maricic et al. 2010) or nuclear genomes (Burbano et al. 2010; Carpenter et al. 2013; Enk et al. 2014; Lindo et al. 2017). We here applied a customized DNA capture approach (Wittig et al. 2015) 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 (**Figure 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 experiments 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 aHLA-Seq pipeline (**Figure 1**) to the aDNA dataset 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 (**Figure 3 and Table 3**). As the achieved average coverage was comparable across the different loci (**Figure S2 and Table S2**), 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 HLA allele calls observed at the class I genes were evaluated using the HLA genotyping algorithm of OptiType, revealing a high level of agreement between the two approaches (93%). Moreover, the presence of the allele DRB1*15:01 was supported by the independent detection of a specific tag-SNP (Krause-Kyora et al. 2018). Both comparisons provided further pieces of evidence for the accuracy of our allele calls, supporting the validity of the aHLA-Seq pipeline.

Our pipeline was also evaluated on simulated aDNA sequence data with distinct but known HLA-B and -DRB1 alleles, produced using the software gargammel (Renaud et al. 2017). For the simulated ancient samples we observed a success rate of 71% at the 1st field level and 57% at 2nd field level of resolution. As all allele calls were classified as correct, we observed an accuracy rate of 100% (**Table 4**). We tested the pipeline for increasing depth of coverages (from 1x up to 60x) and observed also in this case a strong association between success rate and read depth. However, we noticed that in some cases allele calls were possible at very low coverage: 1st field allele call could be obtained already at 1x coverage, while allele calls 2nd field resolution were obtained starting from 5x coverage (**Figure 4**). In contrast, in a few instances, the resolution of HLA alleles was not possible even at high depth of coverage, suggesting that depth of coverage is not the only factor influencing the allele call success. Whilst additional sequence reads in terms of coverage 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 aHLA-Seq 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. This opportunity sets our pipeline apart from other more automated approaches, some of which will always provide an allele call, no matter how spurious and ambiguous the underlying sequence data.

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 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 our aHLA-Seq pipeline, originally developed to define HLA alleles in aDNA samples, can 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 (**Figure 1**). This part is optional and can be applied if the 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 aHLA-Seq pipeline, which would then start with the mapping and sorting steps (**Figure 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 aHLA-Seq pipeline. 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 fewer reads that map to multiple HLA loci 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. Setting a less stringent threshold will ensure detection of reads from unknown alleles, but on the other hand will lead to more ambiguous read mapping, which complicates allele calling. 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 reads. The default threshold of 1% mismatch balances this trade-off and appears to be a good starting point for most datasets. The subsequent visual inspection of HLA sequence reads is a critical step for a correct definition of HLA alleles, especially with aDNA samples where miscoding lesion 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 the HLA alleles allowed the detection of small differences between alleles and was successful in detecting novel HLA variants differing by single point mutations. Furthermore, discrepancies between different PCR-based methods routinely used for HLA typing have been observed (Bauer et al. 2018); these inconsistencies highlight 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 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 aHLA-Seq pipeline, including the manual allele call, were highly reproducible. Comparison of calls for 28 alleles between two different researchers yielded only two disagreements, and in those two cases, the alleles differed only by one and two nucleotides, respectively, resulting in >99% reproducibility at the nucleotide level (Krause-Kyora et al. 2018). 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 (Bauer et al. 2018). In such a situation, the most appropriate approach would be to use consensus information that integrates results from different complementary methods. In this context, our aHLA-Seq 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).

Overall, we have presented a reliable pipeline to genotype HLA genes accurately in ancient and modern DNA samples. The aHLA-Seq pipeline has already been applied successfully to a dataset of medieval European samples, associating HLA variability with susceptibility to leprosy (Krause-Kyora et al. 2018), 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 (Lindo et al. 2016), thus providing a deeper understanding of HLA genetic variation through human history.

Acknowledgments

We are grateful to Jesper L. Boldsen and Dorthe Dangvard Pedersen for access to the St. Jørgen specimens. This work was supported by the Max Planck Society, an Emmy Noether grant from the German Research Foundation (DFG; LE 2593/3-1 to T.L.L.), the Graduate School Human Development in Landscapes and the Cluster of Excellence Inflammation at Interfaces. J.B. was funded by the International Max Planck Research School for Evolutionary Biology. We acknowledge financial support by Land Schleswig-Holstein within the funding programme Open Access Publikationsfonds.

Supplementary Materials

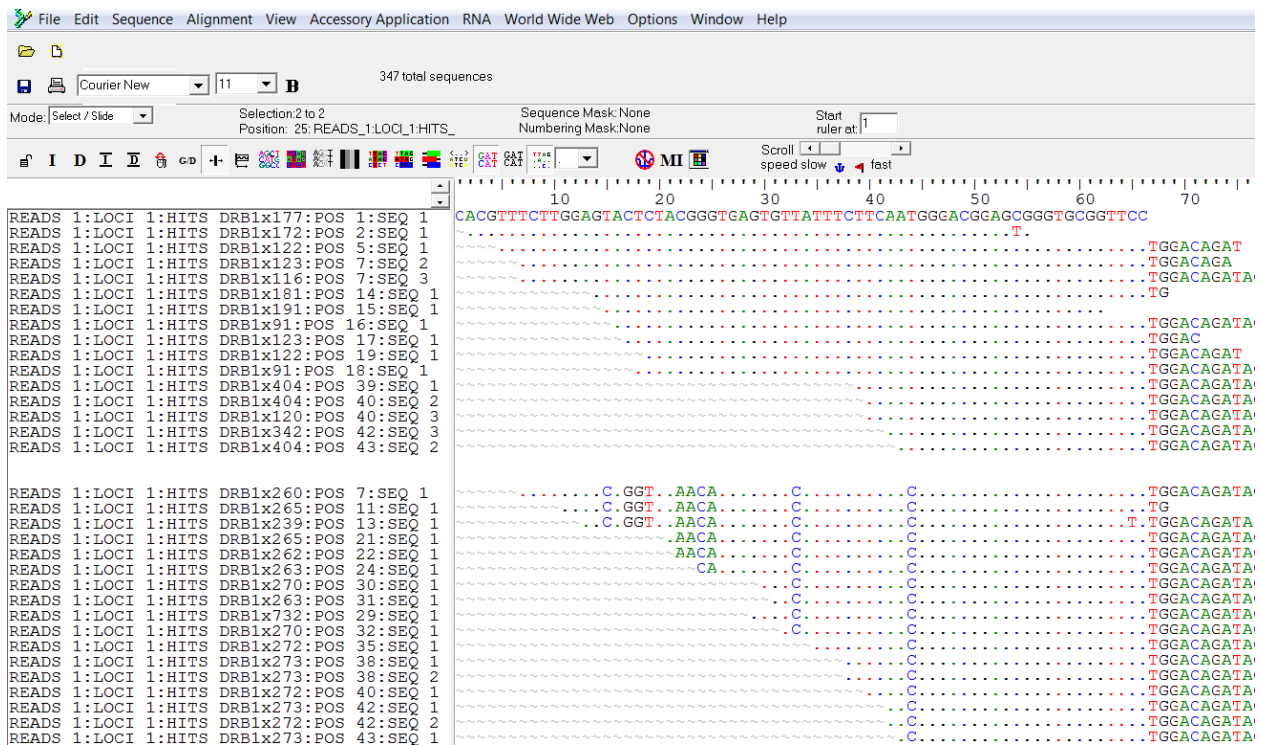


Figure S1 | Screen-shot of sequence alignment editor BioEdit. Screen-shot of sequence alignment editor BioEdit. The picture shows the distinction between SNPs specific of each individual allele, typical aDNA base modifications or PCR/sequencing errors.

Note - the ID tag of each sequence in the FASTA file summarizes information from the sorting procedure performed by the Perl script. In the first domain, READS, is stored the observed number of reads of the exact same sequence fragment (representing PCR duplicates), while the actual sequence read is represented in the FASTA file only once. The second domain, LOCI, stores the number of HLA loci that this read is mapping to. The third domain, HITS, stores the name and allele number of all loci to which the read is mapping to. The fourth domain, POS, indicates the starting position of the reads along the locus. The fifth domain, SEQ, provides a unique number for reads starting from the same position.

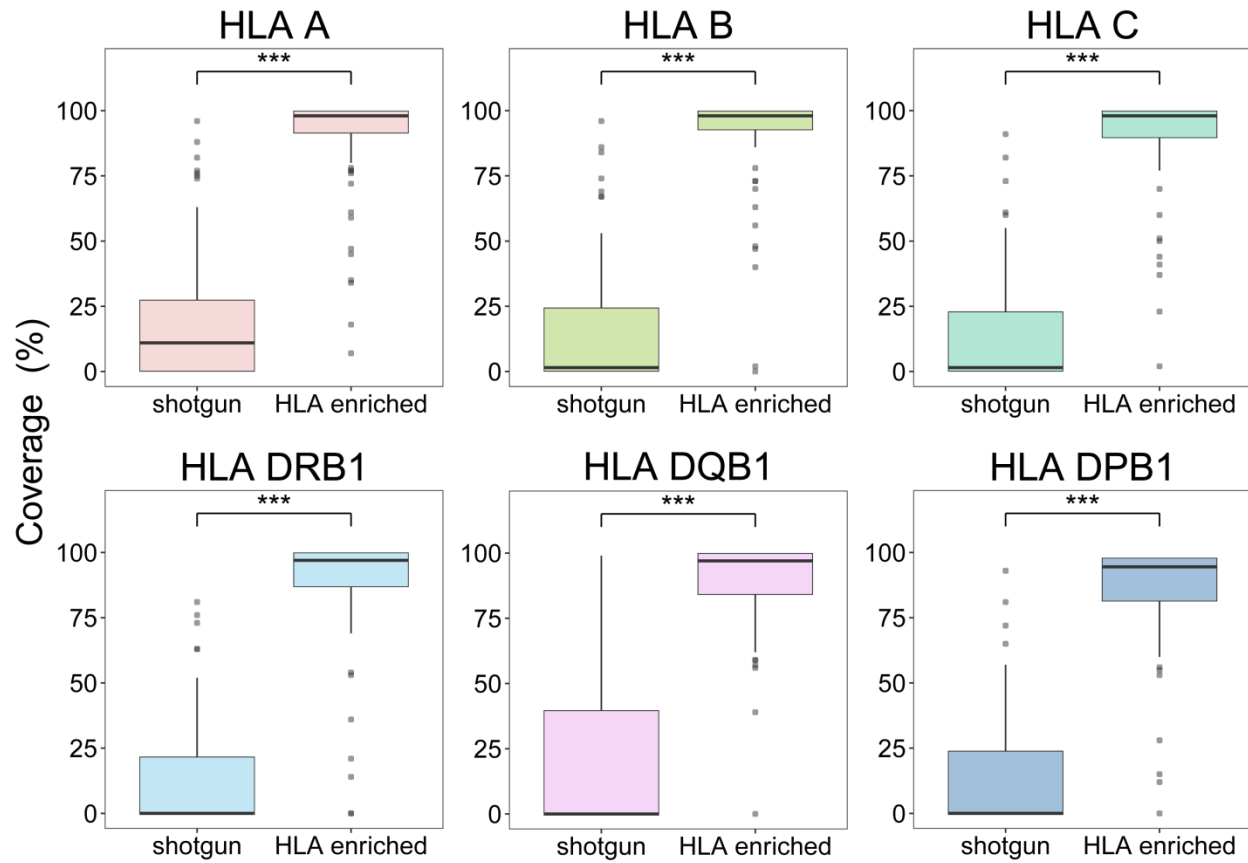


Figure S2 | Median of coverage compared between pre-capture shotgun sequence data (original) and sequence data after HLA enrichment experiments (HLA-enriched) across a subset of 62 historical samples and reported for the individual HLA genes. Significant differences between median values, as derived from Mann-Whitney test, are indicated by horizontal line and asterisks (** $p < 0.001$).

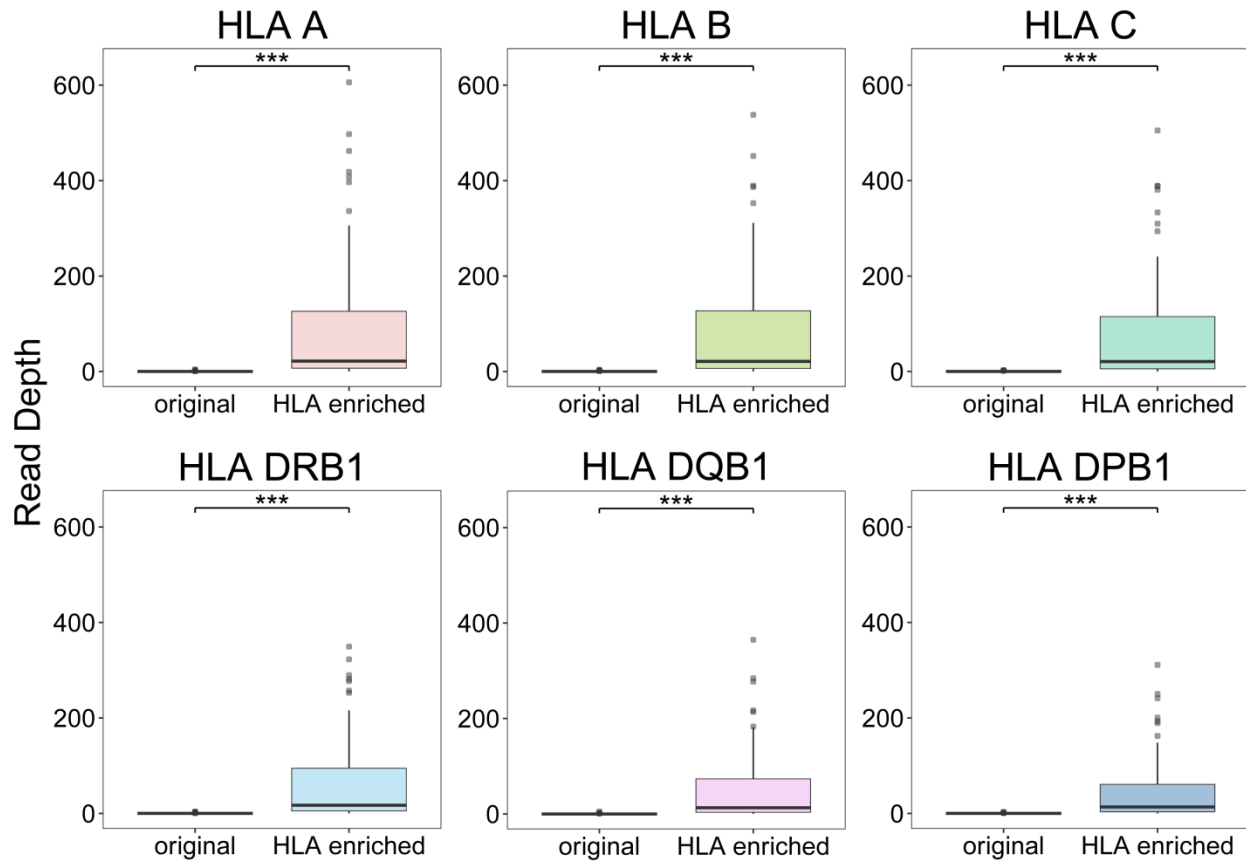


Figure S3 | Median of read depth compared between pre-capture shotgun sequence data (original) and sequence data after HLA enrichment experiments (HLA-enriched) across a subset of 62 historical samples and reported for the individual HLA genes. Significant differences between median values, as derived from Mann-Whitney test, are indicated by horizontal line and asterisks (** $p < 0.001$).

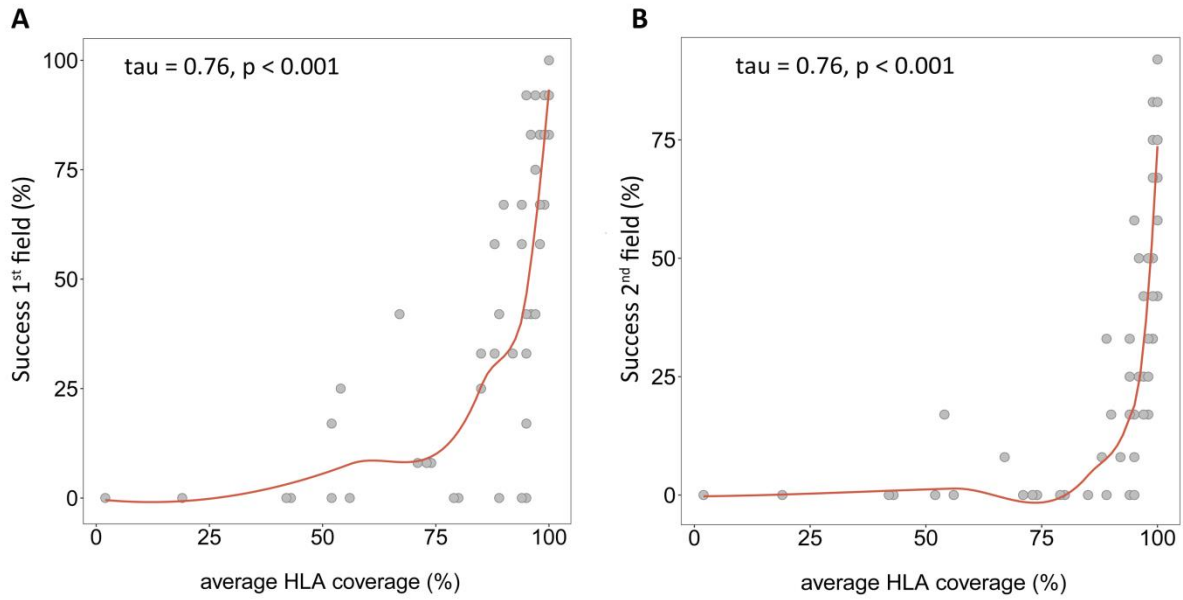


Figure S4 | Correlation between average coverage over the HLA genes and success rate at two different levels of resolutions (A: 1st field; B: 2nd field) across the 68 historical samples.

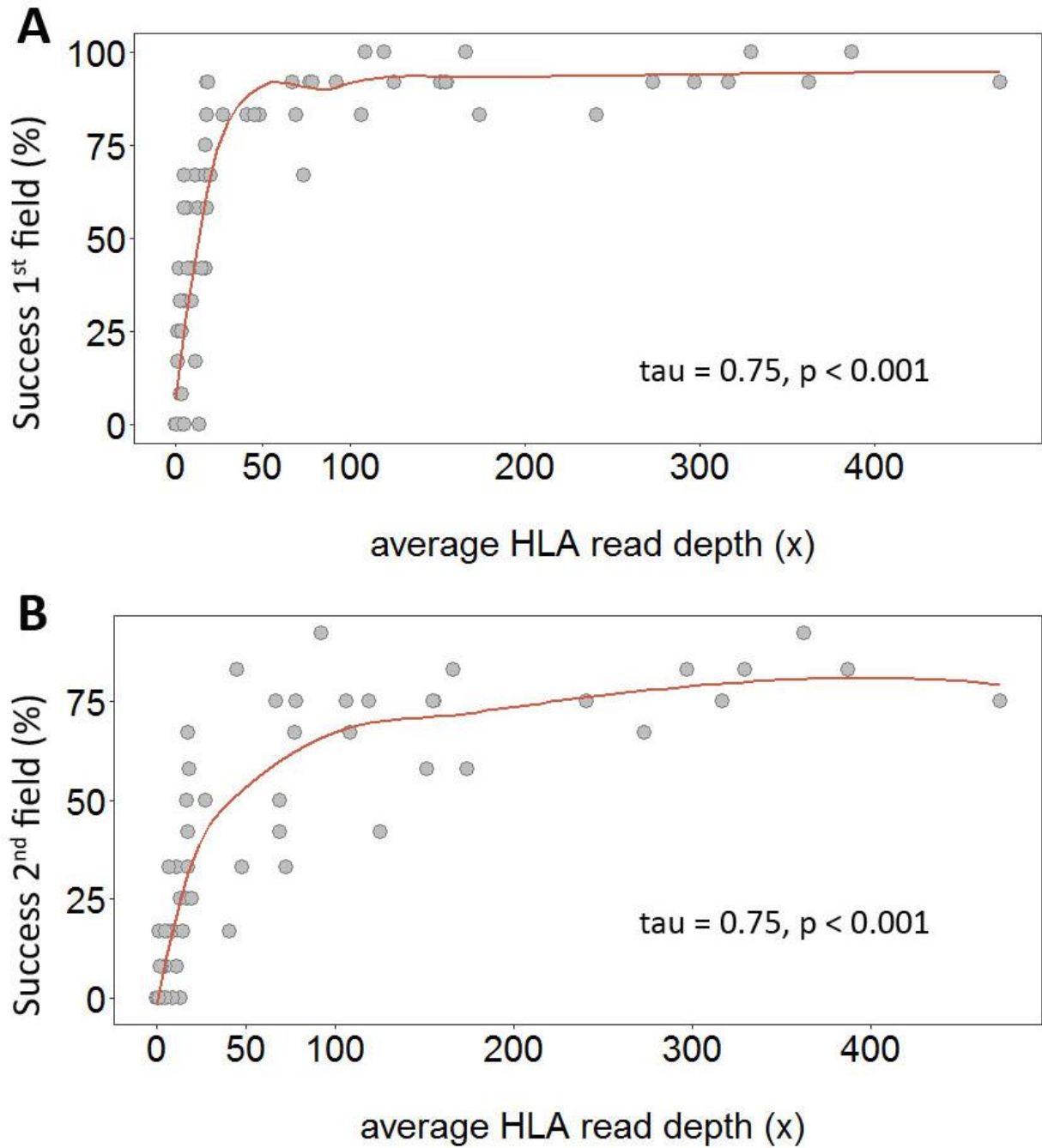


Figure S5 | Correlation between average read depth over the HLA genes and success rate at two different levels of resolutions (A: 1st field; B: 2nd field) across the 68 historical samples.

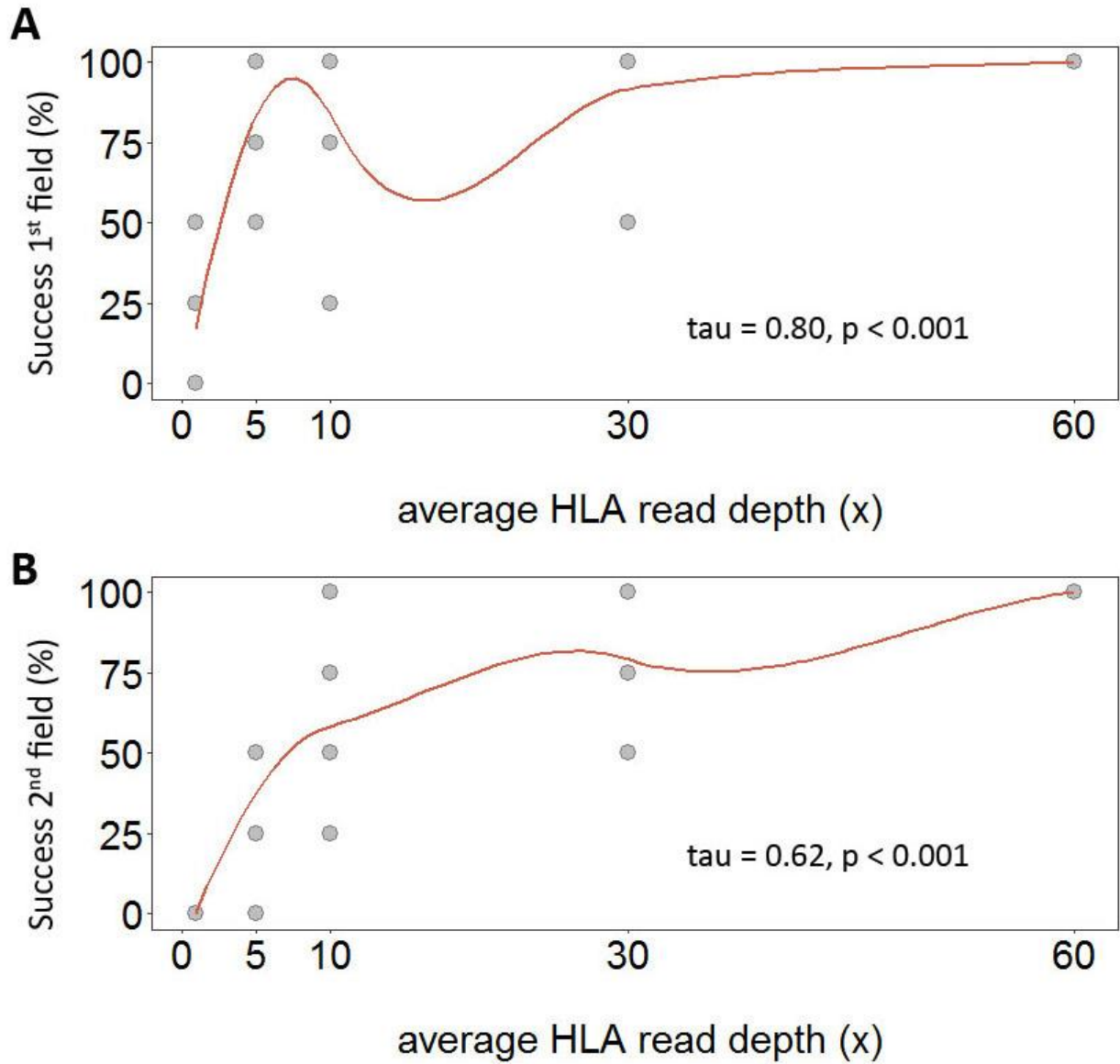


Figure S6 | Correlation between average read depth over the HLA genes and success rate at two different levels of resolutions (A: 1st field; B: 2nd field) for the simulated aDNA samples.

Table S1 | Comparison of sequence data before and after HLA enrichment experiments reported for a subset of 62 historical samples

ID	sequence data	total reads	reads aligning to hg19	endogenous (%)	reads aligning to HLA genes	% of reads aligning to HLA genes	HLA fold-enrichments
G022	original	8599437	56925	1	0	0.000	493.00
	enriched	10851598	278504	3	493	0.005	
G102	original	29473512	7055498	24	32	0.000	2429.75
	enriched	48872766	16524082	34	77752	0.159	
G104	original	16762492	2273849	14	14	0.000	1159.21
	enriched	18296360	4331589	24	16229	0.089	
G1044	original	11939954	1474786	12	4	0.000	11717.50
	enriched	30191012	12147837	40	46870	0.155	
G1049	original	19490520	4863184	25	8	0.000	9708.25
	enriched	42578983	18618541	44	77666	0.182	
G1065	original	36441574	23593147	65	109	0.000	2064.34
	enriched	65461148	44052176	67	225013	0.344	
G1083	original	408892110	16764367	4	57	0.000	163.00
	enriched	18962338	2923148	15	9291	0.049	
G1137	original	174340757	10896459	6	49	0.000	244.51
	enriched	56265338	12331989	22	11981	0.021	
G1149	original	414092343	49259685	12	137	0.000	1341.17
	enriched	85890554	60138238	70	183740	0.214	
G117	original	10235336	147869	1	0	0.000	2836.00
	enriched	15701994	976619	6	2836	0.018	
G118	original	12907976	219935	2	0	0.000	4521.00
	enriched	29259916	906802	3	4521	0.015	
G119	original	14841003	2686014	18	9	0.000	13596.11
	enriched	117183091	32594583	28	122365	0.104	
G120	original	18132253	5381710	30	23	0.000	5577.00
	enriched	63535664	26550293	42	128271	0.202	
G131	original	17278424	176996	1	0	0.000	359.00
	enriched	9515007	253355	3	359	0.004	
G140	original	7628920	106989	1	1	0.000	3484.00
	enriched	14312395	596777	4	3484	0.024	
G149	original	9196823	443414	5	5	0.000	1157.80
	enriched	36304032	7004939	19	5789	0.016	
G154	original	240299649	91470953	38	378	0.000	400.35
	enriched	62627578	34744640	55	151333	0.242	
G164	original	16248305	624401	4	0	0.000	3084.00
	enriched	28650945	1201882	4	3084	0.011	
G165	original	19328959	4569589	24	19	0.000	7032.21
	enriched	45755967	22552165	49	133612	0.292	
G189	original	51568721	5879989	11	8452	0.016	2.97
	enriched	37770246	13518812	36	25095	0.066	
G208	original	286129178	70862102	25	259	0.000	616.76
	enriched	56621036	36900228	65	159741	0.282	
G21	original	23972210	12964863	54	48	0.000	2511.29
	enriched	49630066	35396253	71	120542	0.243	
G24	original	30498430	8071032	26	34	0.000	2080.82
	enriched	57474432	15081426	26	70748	0.123	
G255	original	13384857	340867	3	2	0.000	5851.00
	enriched	28720784	3716194	13	11702	0.041	
G274	original	14859656	393268	3	1	0.000	4427.00
	enriched	15030111	1112193	7	4427	0.029	
G289	original	16927524	160918	1	1	0.000	386.00
	enriched	9978233	244945	2	386	0.004	
G300	original	15439492	10677785	69	35	0.000	1512.71

	enriched	28579922	15262686	53	52945	0.185	
G314	original	31398694	14536475	46	40	0.000	2290.55
	enriched	39074393	24212018	62	91622	0.234	
G33	original	246906567	13095876	5	47	0.000	956.36
	enriched	37557603	14434462	38	44949	0.120	
G34	original	382282413	137542840	36	519	0.000	163.24
	enriched	48638831	19843954	41	84721	0.174	
G348	original	16754756	2295115	14	12	0.000	1044.08
	enriched	18109228	4175796	23	12529	0.069	
G393	original	35355816	15875665	45	61	0.000	2028.85
	enriched	64103915	32163347	50	123760	0.193	
G397	original	12425806	590819	5	2	0.000	12547.50
	enriched	25663966	5379042	21	25095	0.098	
G404	original	219936423	14371069	7	49	0.000	169.57
	enriched	40563956	11014098	27	8309	0.020	
G427	original	245763525	9384556	4	35	0.000	349.29
	enriched	171289835	13167034	8	12225	0.007	
G43	original	35150927	17599123	50	57	0.000	3581.75
	enriched	77091157	51187499	66	204160	0.265	
G472	original	243418738	27042316	11	132	0.000	183.96
	enriched	64528676	23686463	37	24283	0.038	
G48	original	21187364	9869247	47	26	0.000	2470.00
	enriched	25368973	13697638	54	64220	0.253	
G507	original	237551028	97709253	41	409	0.000	117.97
	enriched	86310325	58291258	68	48248	0.056	
G533	original	28125125	2864347	10	13	0.000	1246.54
	enriched	41730888	10265836	25	16205	0.039	
G658	original	548165791	46394838	8	162	0.000	37.37
	enriched	57576755	7410760	13	6054	0.011	
G669	original	40030680	18633086	47	66	0.000	2642.26
	enriched	83187634	56356467	68	174389	0.210	
G708	original	29202942	6861354	23	27	0.000	2942.22
	enriched	49400523	17902389	36	79440	0.161	
G712	original	29076810	7897978	27	37	0.000	1314.62
	enriched	25073439	10471184	42	48641	0.194	
G722	original	412635307	57950390	14	156	0.000	40.13
	enriched	35231822	8704397	25	6261	0.018	
G730	original	30979029	8017931	26	39	0.000	2122.05
	enriched	32166123	17368044	54	82760	0.257	
G738	original	36353460	1437603	4	3	0.000	3812.67
	enriched	55485529	2720910	5	11438	0.021	
G749	original	22577271	5329614	24	12	0.000	1096.17
	enriched	40918094	16020855	39	13154	0.032	
G750	original	16027731	1129349	7	7	0.000	754.71
	enriched	5496339	1016297	18	5283	0.096	
G860	original	20727017	15675250	76	99	0.000	1752.87
	enriched	53408873	36828601	69	173534	0.325	
G864	original	18458498	12683368	69	54	0.000	1411.33
	enriched	19567241	13942742	71	76212	0.389	
G870	original	17782756	1009246	6	7	0.000	4071.14
	enriched	23287681	5015827	22	28498	0.122	
G876	original	30275792	14274672	47	69	0.000	1439.75
	enriched	41966479	22663531	54	99343	0.237	
G911	original	12122972	3271201	27	12	0.000	4809.08
	enriched	26377084	14688225	56	57709	0.219	
G912	original	16458021	1682533	10	3	0.000	8077.33
	enriched	17875627	5299164	30	24232	0.136	
G936	original	18568653	3038829	16	21	0.000	9137.67

	enriched	84300687	74490774	88	191891	0.228	
G939	original	14134206	4505953	32	11	0.000	8090.55
	enriched	47798424	26595569	56	88996	0.186	
G942	original	18190020	2211172	12	7	0.000	11819.43
	enriched	54209829	22938407	42	82736	0.153	
G943	original	170193606	2110175	1	5	0.000	484.20
	enriched	19445661	418647	2	2421	0.012	
G951	original	16358004	1243961	8	3	0.000	3349.00
	enriched	24554661	3164285	13	10047	0.041	
G973	original	15462662	5926620	38	17	0.000	3889.47
	enriched	39044885	22012088	56	66121	0.169	
G978	original	221493132	3906293	2	17	0.000	243.35
	enriched	109771423	5684949	5	4137	0.004	

Note - Endogenous DNA content and percentage of reads aligning to HLA genes calculated from original UDG shotgun libraries and HLA-enriched UDG libraries are reported for each individual sample, for a total of 62 ancient samples. Fold-enrichment was obtained by dividing the number of reads mapping to the HLA reference obtained from enriched HLA libraries by the number of reads mapping to the HLA reference calculated from pre-capture shotgun libraries; when the denominator was 0 the number of on-target HLA reads from enriched libraries has been assigned.

Table S2 | Coverage and read depth at each locus and averaged across the 6 investigated genes before and after HLA enrichment experiments reported for a subset of 62 historical samples

ID	sequence data	HLA-A		HLA-B		HLA-C		HLA-DRB1		HLA-DQB1		HLA-DPB1		average HLA cov [%]	average HLA read depth
		cov [%]	read depth	cov [%]	read depth	cov [%]	read depth	cov [%]	read depth	cov [%]	read depth	cov [%]	read depth		
G022	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	18	0.23	0	0.1	23	0.36	14	0.33	39	0.49	0	0.00	19	0.31
G102	original	8	0.22	6	0.2	0	0.17	0	0.00	14	0.14	21	0.21	6	0.14
	enriched	98	53.42	99	59.4	100	52.29	99	45.17	94	29.88	98	27.76	98	48.03
G1042	original	0	0.02	0	0.0	0	0.04	0	0.00	0	0.00	0	0.00	0	0.02
	enriched	7	0.18	2	0.2	2	0.20	0	0.09	0	0.05	15	0.24	2	0.14
G1044	original	16	0.19	0	0.0	0	0.03	0	0.00	0	0.00	0	0.00	3	0.05
	enriched	98	28.98	97	15.0	96	16.21	94	11.72	92	14.61	94	12.04	96	17.30
G1049	original	0	0.08	0	0.1	0	0.02	22	0.29	0	0.00	14	0.14	4	0.10
	enriched	100	127.86	100	133.2	100	119.43	99	77.36	100	73.55	98	61.28	100	106.29
G1065	original	77	1.62	52	1.2	38	0.77	23	0.64	62	0.71	15	0.26	50	0.98
	enriched	100	606.03	100	537.8	100	504.98	100	349.45	100	364.80	98	311.15	100	472.62
G1083	original	33	0.33	27	0.5	0	0.24	0	0.00	0	0.00	0	0.00	12	0.22
	enriched	76	4.75	91	5.7	82	5.50	99	10.29	74	2.15	12	0.25	85	5.67
G1137	original	15	0.31	23	0.4	6	0.28	36	0.46	32	0.31	27	0.33	22	0.36
	enriched	72	1.37	73	4.2	70	3.19	84	2.67	56	1.36	60	0.94	71	2.56
G1149	original	75	1.56	38	0.8	43	0.88	52	0.78	70	1.39	53	0.59	56	1.08
	enriched	100	462.14	100	389.4	100	388.26	100	289.73	100	284.29	98	250.32	100	362.77
G117	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	61	1.99	63	1.3	41	1.20	36	1.15	57	1.40	55	0.95	52	1.41
G118	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	80	4.53	88	4.4	94	4.06	75	1.21	65	1.38	66	1.41	80	3.11
G119	original	0	0.09	15	0.3	0	0.14	40	0.10	10	0.10	0	0.00	13	0.14
	enriched	100	201.52	100	201.4	100	184.83	100	116.40	100	127.56	98	95.18	100	166.34
G120	original	0	0.08	3	0.2	3	0.17	20	0.27	0	0.00	0	0.00	5	0.14
	enriched	100	91.42	100	85.6	100	73.98	97	67.85	100	46.67	95	33.21	99	73.11
G131	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	47	1.11	40	0.8	44	0.92	21	0.70	62	1.32	28	0.39	43	0.97
G140	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	98	18.64	97	19.2	97	17.04	95	11.12	89	3.26	93	11.84	95	13.84
G149	original	0	0.03	0	0.0	0	0.00	0	0.00	12	0.12	36	0.36	2	0.03
	enriched	34	1.24	48	0.9	37	1.16	84	2.10	77	1.31	82	1.97	56	1.35
G154	original	82	2.60	84	2.6	82	1.68	48	0.92	77	2.04	41	0.92	75	1.97
	enriched	98	56.65	99	46.4	99	44.51	96	41.66	98	17.26	97	15.09	98	41.29
G164	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	96	7.12	95	6.2	91	4.47	96	4.76	91	2.77	61	1.25	94	5.07
G165	original	22	0.42	8	0.1	0	0.09	0	0.00	0	0.00	19	0.19	6	0.12
	enriched	100	93.40	100	83.8	100	84.92	99	49.04	99	35.77	98	41.78	99	69.38
G189	original	59	1.07	69	1.6	61	1.27	63	1.20	87	1.27	0	0.04	68	1.29
	enriched	96	10.85	94	13.9	96	11.31	96	8.93	93	3.42	88	6.79	95	9.68
G208	original	74	1.61	74	1.8	60	2.29	63	1.09	94	1.52	57	0.88	73	1.66
	enriched	100	140.98	100	138.9	100	142.55	100	103.31	100	100.74	98	83.59	100	125.30
G21	original	28	0.45	6	0.2	18	0.27	0	0.17	50	0.50	30	0.36	20	0.31
	enriched	100	188.99	100	172.5	100	159.81	100	118.05	100	118.89	97	92.83	100	151.65
G24	original	17	0.35	6	0.1	13	0.22	21	0.23	0	0.02	0	0.02	11	0.18
	enriched	100	225.47	100	222.0	100	199.88	100	128.56	100	95.55	98	100.75	100	174.29
G255	original	8	0.11	0	0.0	0	0.03	0	0.00	0	0.00	0	0.00	2	0.03
	enriched	77	5.06	78	3.7	82	5.00	54	2.00	78	2.73	92	3.75	74	3.69
G274	original	8	0.09	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	2	0.02
	enriched	91	6.21	70	3.2	60	2.58	86	2.29	87	3.52	56	2.32	79	3.56
G289	original	12	0.12	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	2	0.02
	enriched	45	0.77	56	0.8	51	0.62	0	0.02	59	0.64	53	0.99	42	0.58
G300	original	28	0.41	17	0.4	19	0.39	0	0.18	0	0.00	24	0.24	13	0.27
	enriched	100	306.00	100	262.5	100	240.71	100	215.92	100	181.40	98	148.22	100	241.30
G314	original	8	0.17	0	0.2	15	0.41	0	0.21	26	0.26	23	0.23	10	0.24
	enriched	100	142.28	100	130.8	100	134.82	100	101.30	100	88.67	98	78.65	100	119.58
G33	original	15	0.22	19	0.3	0	0.09	0	0.00	0	0.00	0	0.00	7	0.12

	enriched	99	87.80	99	81.4	100	66.98	97	60.40	100	48.24	97	41.41	99	68.95
G34	original	96	4.40	96	3.7	91	3.00	76	3.60	99	5.22	93	3.39	91	3.98
	enriched	100	124.39	100	119.4	100	105.55	100	118.37	100	74.45	98	61.76	100	108.43
G348	original	16	0.20	0	0.0	0	0.08	15	0.22	0	0.00	0	0.07	6	0.11
	enriched	96	11.29	96	11.4	95	11.52	85	8.27	96	13.45	85	12.16	94	11.19
G393	original	45	0.63	25	0.5	23	0.47	0	0.13	41	0.53	0	0.06	27	0.45
	enriched	100	113.52	100	106.0	100	105.14	100	71.93	100	65.56	98	54.91	100	92.43
G397	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	93	15.40	98	16.8	98	14.10	97	14.86	83	5.49	84	3.99	94	13.32
G404	original	21	0.33	22	0.4	4	0.20	0	0.00	39	0.41	57	0.57	17	0.26
	enriched	96	7.35	95	7.0	88	7.24	95	4.44	97	5.12	78	5.47	94	6.24
G427	original	42	0.71	0	0.1	23	0.35	36	0.45	31	0.31	0	0.09	26	0.39
	enriched	89	3.75	86	3.6	91	4.45	80	5.08	78	2.33	81	2.25	85	3.84
G43	original	20	0.42	28	0.5	23	0.45	0	0.12	42	0.44	0	0.05	23	0.38
	enriched	100	497.30	100	451.7	100	389.18	100	322.74	100	277.25	98	241.24	100	387.63
G472	original	50	0.96	67	1.0	43	0.88	50	0.76	22	0.30	57	1.09	46	0.78
	enriched	92	7.64	96	5.7	96	6.37	97	3.84	80	1.29	70	3.13	92	4.98
G48	original	12	0.28	0	0.2	0	0.27	0	0.00	14	0.14	0	0.00	5	0.17
	enriched	99	55.35	100	58.8	99	51.10	98	36.73	97	24.19	97	19.97	99	45.24
G507	original	88	3.21	86	2.8	73	2.19	73	3.54	76	1.18	72	1.70	79	2.58
	enriched	97	17.43	98	21.7	98	18.45	99	20.38	95	9.15	94	5.19	97	17.42
G533	original	6	0.12	17	0.3	0	0.04	46	0.67	0	0.00	0	0.00	14	0.22
	enriched	78	6.23	95	6.3	89	4.79	90	3.62	89	3.84	65	2.09	88	4.95
G658	original	76	1.58	67	1.2	51	1.02	81	2.70	67	0.99	65	1.09	68	1.50
	enriched	98	6.18	92	4.2	80	3.56	89	7.84	92	3.77	92	7.01	90	5.11
G669	original	17	0.24	34	0.5	24	0.37	20	0.45	53	0.71	22	0.32	30	0.46
	enriched	100	396.57	100	311.3	100	309.92	100	252.61	100	216.81	98	192.93	100	297.44
G708	original	12	0.27	0	0.1	0	0.17	0	0.00	0	0.00	0	0.00	2	0.12
	enriched	100	77.52	99	79.9	100	72.18	99	56.53	100	49.37	97	37.03	100	67.09
G712	original	20	0.39	0	0.2	8	0.27	0	0.23	0	0.00	0	0.08	6	0.21
	enriched	100	211.86	100	161.1	100	177.42	100	132.34	100	96.72	97	109.71	100	155.89
G722	original	63	1.75	53	1.4	54	0.94	42	0.62	99	1.78	31	0.31	62	1.31
	enriched	77	3.85	92	3.9	87	2.61	85	2.66	99	2.12	79	2.43	88	3.02
G730	original	11	0.22	14	0.3	9	0.19	0	0.09	22	0.22	18	0.18	11	0.20
	enriched	100	100.89	100	91.5	100	86.16	100	54.41	99	53.78	97	43.96	99	77.35
G738	original	2	0.07	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.01
	enriched	99	13.04	97	11.6	93	10.59	91	7.48	96	5.35	86	7.73	95	9.61
G749	original	0	0.00	0	0.0	47	0.48	16	0.16	17	0.17	15	0.15	16	0.16
	enriched	59	1.76	73	1.4	77	2.90	69	3.87	59	0.66	84	4.53	67	2.12
G750	original	0	0.07	0	0.0	11	0.15	0	0.12	0	0.00	0	0.08	2	0.07
	enriched	97	22.68	98	20.1	99	18.07	97	15.82	97	12.68	96	9.88	98	17.87
G860	original	54	1.31	52	1.1	55	0.93	52	0.87	65	1.13	81	1.01	56	1.06
	enriched	100	418.27	100	386.6	100	380.91	100	281.45	100	183.29	98	200.72	100	330.11
G864	original	11	0.26	19	0.4	23	0.44	21	0.34	0	0.00	51	0.97	15	0.28
	enriched	100	336.17	100	310.7	100	293.96	100	257.25	100	171.12	98	162.23	100	273.85
G870	original	5	0.05	0	0.0	0	0.03	16	0.20	0	0.00	16	0.16	4	0.06
	enriched	99	19.93	97	15.3	97	16.51	96	12.80	97	9.56	94	9.19	97	14.83
G876	original	26	0.43	39	0.7	41	0.68	0	0.00	50	0.71	0	0.00	31	0.50
	enriched	100	407.98	100	352.7	100	333.26	100	277.28	100	213.86	98	189.30	100	317.02
G911	original	7	0.14	0	0.1	0	0.08	0	0.13	0	0.00	20	0.20	1	0.09
	enriched	95	32.83	98	26.6	98	26.16	99	31.71	90	20.35	89	15.00	96	27.53
G912	original	0	0.00	0	0.0	0	0.04	0	0.00	0	0.00	0	0.00	0	0.02
	enriched	96	20.79	97	20.7	97	25.32	98	11.52	96	9.97	96	13.07	97	17.67
G936	original	0	0.12	11	0.3	0	0.13	20	0.20	0	0.00	12	0.11	6	0.15
	enriched	100	115.71	100	100.8	100	91.91	100	45.62	100	39.41	98	30.54	100	78.69
G939	original	0	0.04	0	0.0	0	0.02	0	0.07	40	0.40	0	0.00	8	0.11
	enriched	96	19.34	98	24.8	98	21.44	99	23.31	98	10.52	96	16.30	98	19.88
G942	original	0	0.00	14	0.2	9	0.12	0	0.05	0	0.00	24	0.30	5	0.07
	enriched	98	27.77	95	22.4	96	20.27	96	17.73	91	4.31	97	13.76	95	18.49
G943	original	0	0.00	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	enriched	96	6.53	98	5.7	94	5.04	76	4.36	81	2.34	91	2.28	89	4.79
G951	original	6	0.06	0	0.0	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01
	enriched	99	20.26	97	17.9	99	21.83	99	16.36	99	14.39	93	14.04	99	18.14
G973	original	0	0.04	0	0.1	7	0.15	11	0.19	12	0.12	0	0.06	6	0.11
	enriched	100	177.29	100	157.3	100	164.78	100	162.15	100	112.80	98	92.54	100	154.86
G978	original	0	0.06	0	0.1	19	0.25	0	0.00	37	0.37	29	0.29	11	0.15
	enriched	35	0.51	47	1.6	50	1.03	53	0.81	75	2.45	79	1.35	52	1.28

Table S3 | Median of coverage and read depth at each locus before and after HLA enrichment experiments reported for a subset of 62 historical samples

	Coverage (%)						Read Depth (x)					
	original			HLA enriched			original			HLA enriched		
	Min	Max	Median (95% CI)	Min	Max	Median (95% CI)	Min	Max	Median (95% CI)	Min	Max	Median (95% CI)
HLA A	0	96	11 (7-17)	7	100	98 (96-100)	0	4.4	0.19 (0.09-0.31)	0.18	606.03	21.74 (11.29-87.80)
HLA B	0	96	1 (0-15)	0	100	98 (96-100)	0	3.67	0.16 (0.06-0.28)	0.13	537.84	21.21 (11.58-81.35)
HLA C	0	91	1 (0-13)	2	100	98 (96-100)	0	3	0.17 (0.08-0.27)	0.2	504.98	20.85 (11.31-72.18)
HLA DRB1	0	81	0 (0-16)	0	100	97 (96-99)	0	3.6	0.12 (0.00-0.20)	0.02	349.45	17.04 (8.93-49.04)
HLA DQB1	0	99	0 (0-22)	0	100	97 (92-99)	0	5.22	0.00 (0.00-0.22)	0.05	364.8	13.06 (4.31-39.41)
HLA DPB1	0	93	(0-18)	0	98	95(89-97)	0	3.39	0.08 (0.00-0.19)	0	311.15	13.42 (6.79-33.21)

Table S4 | Allele calls at HLA class I and class II genes for the 68 historical samples

ID	A1	A2	B1	B2	C1	C2	DRB1 1	DRB1 2	DQB1 1	DQB1 2	DPB1 1	DPB1 2
G022	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G102	A*02	NA	B*18	B*44	C*07	C*03	DRB1*04:03	DRB1*11	DQB1*03:01	DQB1*03:02	DPB1*04:02	NA
G104	NA	NA	NA	NA	NA	NA	DRB1*04	NA	NA	NA	DPB1*422:01	NA
G1042	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G1044	A*03:01	NA	NA	NA	C*03	NA	NA	NA	DQB1*03:02	DQB1*06	DPB1*03:01	NA
G1049	A*03:01	A*03:01	B*07:02	B*07:02	C*07	NA	DRB1*14:01	DRB1*15:01	DQB1*05:03	DQB1*06:02	DPB1*452:01	NA
G1065	A*03:01	A*11:01	B*07:02	B*35:01	C*07	C*04	DRB1*01:01	DRB1*04:01	DQB1*03:01	DQB1*05:01	DPB1*04:01	NA
G1083	NA	NA	NA	NA	NA	NA	DRB1*01	DRB1*08	DQB1*04	DQB1*05	NA	NA
G1137	NA	NA	NA	NA	NA	NA	NA	NA	DQB1*02	NA	NA	NA
G1149	A*02:01	A*03:01	B*07:02	B*44:02	C*05:01	C*07:02	DRB1*04:01	DRB1*15:01	DQB1*03:01	DQB1*06:02	DPB1*04:01	NA
G117	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G118	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G119	A*24:02	A*24:02	B*07:02	B*07:02	C*07	C*07	DRB1*15:01	DRB1*15:01	DQB1*06:02	DQB1*06:02	DPB1*04:01	DPB1*240:01
G120	A*11	A*68	NA	NA	C*03	NA	DRB1*13	DRB1*15:01	DQB1*06:02	DQB1*06:04	DPB1*03:01	NA
G131	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G140	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G149	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G154	A*03	A*02	B*35	B*42	C*04	C*07	DRB1*04	DRB1*15	DQB1*06:02	DQB1*03:02	NA	NA
G164	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G165	A*03:01	A*02	B*15:01	B*44:03	C*03	C*16	DRB1*04:01	DRB1*15	DQB1*06:02	DQB1*03:02	NA	NA
G166	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G189	A*02	A*01	B*57:01	NA	NA	NA	DRB1*04	DRB1*07:01	NA	NA	NA	NA
G208	A*24	A*32	B*40	B*44:02	C*03	C*07	DRB1*01:01	DRB1*04	DQB1*03:02	DQB1*05:01	DPB1*04:01	NA
G21	A*68:01	A*03:01	B*38	B*15	C*12	C*03	DRB1*04:01	DRB1*13:01	DQB1*06:03	DQB1*03:02	DPB1*131:01	NA
G24	A*01:01	A*24:02	B*57:01	B*39:06	C*07	C*06	DRB1*07:01	DRB1*08:01	DQB1*04	DQB1*03:03	NA	NA
G255	NA	NA	NA	NA	C*07	NA	NA	NA	NA	NA	NA	NA
G274	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G28	A*68	NA	NA	NA	NA	NA	NA	NA	DQB1*06:02	NA	DPB1*04:01	NA

G911	A*02:01	NA	B*15	B*44:02	C*03	C*05	DRB1*04:01	DRB1*04:01	DQB1*03:01	DQB1*03	DPB1*04:01	NA
G912	A*26	A*03	B*40:01	B*44:03	C*03	C*16	DRB1*07	DRB1*09	DQB1*02:01	DQB1*03:03	DPB1*11:01	NA
G914	A*02:06	NA	NA	NA	NA	NA	NA	NA	DQB1*02	DQB1*05:01	DPB1*109:01	DPB1*136:01
G936	A*01:01	A*03:01	B*07:02	B*51:01	C*07	C*15	DRB1*04:04	DRB1*15:01	DQB1*03:02	DQB1*06:02	DPB1*04:01	NA
G939	A*03	A*26	B*15	NA	C*03	C*14	DRB1*01:01	DRB1*01:01	NA	NA	DPB1*04:01	NA
G942	A*11	NA	B*35	B*40:01	C*03	C*04	DRB1*04:03	DRB1*15:01	DQB1*03:10	DQB1*06:02	DPB1*16:01	DPB1*442:01
G943	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
G951	A*02:01	A*03:01	B*15	B*44:02	C*05	C*03:07	DRB1*04:01	DRB1*04:01	DQB1*03:01	DQB1*03:02	NA	NA
G973	A*02:01	A*31:01	B*40:01	B*44:02	C*03	C*05	DRB1*13:01	DRB1*13:01	DQB1*06:03	DQB1*06:04	DPB1*03:01	NA
G978	NA	NA	NA	NA	NA	NA	NA	NA	DQB1*02	DQB1*06	NA	NA

Table S5 | Validation of historical allele calls at class I HLA genes using OptiType

ID	aHLAseq											
	OptiType						aHLAseq					
	A1	A2	B1	B2	C1	C2	A1	A2	B1	B2	C1	C2
G102	A*02:01	A*68:01	B*18:01	B*44:02	C*03:03	C*07:01	A*02	NA	B*18	B*44	C*03	C*07
G1044	A*03:01	A*31:01	B*27:08	B*40:14	C*02:08	C*03:04	A*03:01	NA	NA	NA	NA	C*03
G120	A*11:01	A*68:01	B*07:33	B*40:01	C*03:04	C*07:02	A*11	A*68	NA	NA	C*03	NA
G154	A*02:01	A*03:01	B*07:02	B*35:02	C*04:01	C*07:02	A*02	A*03	B*42	B*35	C*04	C*07
G165	A*02:01	A*03:01	B*15:38	B*44:03	C*03:04	C*16:01	A*02	A*03:01	B*15:01	B*44:03	C*03	C*16
G189	A*01:01	A*02:01	B*40:01	B*57:01	C*03:04	C*06:02	A*01	A*02	NA	B*57:01	NA	NA
G24	A*01:01	A*24:02	B*39:06	B*57:01	C*06:03	C*07:02	A*01:01	A*24:02	B*39:06	B*57:01	C*06	C*07
G255	A*24:04	A*69:01	B*07:33	B*18:20	C*03:04	C*07:02	NA	NA	NA	NA	NA	C*07
G28	A*03:01	A*68:01	B*07:02	B*44:07	C*07:04	C*07:04	NA	A*68	NA	NA	NA	NA
G300	A*01:01	A*02:01	B*07:02	B*08:01	C*07:01	C*07:02	A*01:01	A*02:01	B*07:02	B*08:01	C*07	NA
G314	A*24:02	A*32:01	B*27:05	B*44:02	C*02:27	C*05:01	A*24:02	A*32:01	B*27:05	B*44	C*02	C*05
G33	A*26:01	A*68:01	B*07:02	B*51:01	C*07:02	C*15:02	A*26:01	A*68	B*07	B*51:01	C*07	C*15
G348	A*30:01	A*33:03	B*07:02	B*35:24	C*04:01	C*07:02	NA	NA	NA	NA	C*04	C*07
G393	A*01:01	A*24:02	B*08:01	B*45:04	C*06:02	C*07:01	A*01:01	A*24:02	B*08:01	B*45:01	NA	C*07:01
G397	A*01:23	A*02:01	B*45:01	B*57:01	C*07:17	C*12:03	A*29	A*02	B*45	NA	NA	NA
G404	A*02:13	A*26:01	B*08:01	B*15:38	C*03:04	C*07:02	A*02	NA	B*08	B*15	C*03	C*07
G427	A*02:01	A*03:07	B*15:05	B*35:14	C*03:04	C*04:01	NA	NA	B*15	NA	C*03	C*04
G472	A*02:01	A*68:01	B*40:01	B*44:02	C*03:67	C*05:01	A*02	A*68	B*40:01	B*44	NA	NA
G48	A*02:01	A*24:56	B*39:06	B*57:01	C*06:02	C*07:02	A*02:01	A*24:02	B*39:06	B*57:01	NA	NA
G533	A*02:01	A*03:01	B*15:39	B*40:01	C*03:04	C*03:04	A*02	NA	B*15	NA	C*03	NA
G658	A*24:02	A*30:02	B*07:02	B*18:01	C*05:01	C*07:02	A*24	A*30	B*07	B*18	NA	C*07
G708	A*02:01	A*02:01	B*15:38	B*57:10	C*03:04	C*07:01	A*02:01	NA	B*15:01	B*57:01	C*03	C*07:01
G712	A*02:01	A*03:01	B*07:02	B*08:01	C*07:01	C*07:02	A*02	A*03	B*07:02	B*08:01	C*07:01	C*07:01
G722	A*01:01	A*02:01	B*07:02	B*35:21	C*04:01	C*07:123	A*36	NA	NA	NA	NA	NA
G730	A*02:01	A*03:01	B*40:01	B*51:01	C*02:02	C*03:04	A*02	A*03:01	B*40:01	B*51:01	C*02	C*03
G738	A*02:01	A*03:22	B*07:02	B*15:01	C*03:04	C*07:02	A*02	A*03	NA	NA	NA	NA
G749	A*02:01	A*03:22	B*07:03	B*07:03	C*03:04	C*07:29	A*02	NA	NA	NA	C*03	NA

G750	A*03:01	A*24:02	B*35:33	B*56:01	C*01:02	C*04:01	NA	NA	B*35:03	B*56:01	C*01	C*04
G870	A*01:01	A*02:13	B*07:02	B*08:01	C*07:01	C*07:01	A*01	A*02	NA	B*08	NA	NA
G876	A*02:01	A*03:01	B*15:38	B*55:01	C*03:03	C*03:03	A*02:01	A*03:01	B*15:01	B*55:01	C*03	C*03
G896	A*01:01	A*02:01	B*07:02	B*08:01	C*07:01	C*07:01	A*01:01	A*02	NA	NA	C*07	NA
G911	A*02:01	A*03:01	B*15:07	B*44:02	C*03:04	C*05:01	A*02:01	NA	B*15	B*44:02	C*03	C*05
G912	A*03:01	A*26:01	B*40:01	B*44:03	C*03:04	C*16:01	A*03	A*26	B*40:01	B*44:03	C*03	C*16
G914	A*02:01	A*02:01	B*40:01	B*44:03	C*03:04	C*16:01	A*02:06	NA	NA	NA	NA	NA
G936	A*01:01	A*03:01	B*07:02	B*51:01	C*07:02	C*15:02	A*01:01	A*03:01	B*07:02	B*51:01	C*07	C*15
G939	A*03:01	A*26:01	B*15:01	B*51:13	C*03:03	C*14:04	A*03	A*26	B*15	NA	C*03	C*14
G942	A*11:01	A*33:03	B*35:24	B*40:01	C*03:04	C*04:01	A*11	NA	B*35	B*40:01	C*03	C*04
G951	A*02:01	A*03:01	B*15:01	B*44:02	C*03:04	C*05:01	A*02:01	A*03:01	B*15	B*44:02	C*03:07	C*05
G973	A*02:01	A*31:01	B*40:01	B*44:02	C*03:04	C*05:01	A*02:01	A*31:01	B*40:01	B*44:02	C*03	C*05

Note – Alleles reported in black are those with identical call from the two approaches. In blue are reported the cases for which we found support for our allele call but we could not support the call provided by OptiType. Alleles reported in red are instead those for which we could not confirm the calls from our approach, but found supporting reads for the allele call provided by OptiType. In green are reported the allele calls we could not resolve. Called alleles that differed between the two approaches:

G154 - Optitype: B*07:02; aHLA-Seq: B*42. B*07:02 supported by 3 reads mapping to the second exon.

G165 - Optitype: B*15:38; aHLA-Seq: B*15:01. B*15:01 and B*15:38 differ in one position at the end of the second exon where the allele B*15:01 is supported by many reads while we did not find reads supporting B*15:38.

G393 - Optitype: B*45:04; aHLA-Seq: B*45:01. B*45:04 and B*45:01 differ in one position at the end of the second exon where B*45:01 is supported by many reads while we did not find reads supporting B*45:04.

G397 - Optitype: A*01:23; aHLA-Seq: A*29. Both the allele calls were supported by many reads and we could not resolve the allele calls.

G48 - Optitype: A*24:56; aHLA-Seq: A*24:02. A*24:02 supported by many reads while we did not find reads supporting A*24:56

G708 - Optitype: B*15:38-B*57:10; aHLA-Seq: B*15:01-B*57:01. B*15:01 supported by some reads while no reads supporting B*15:38. B*57:01 supported by some reads while no reads supporting B*57:10

G712 - Optitype: C*07:02; aHLA-Seq: C*07:01. C*07:02 supported by a small number of unique reads.

- G722** - Optitype: A*01:01; aHLA-Seq: A*36. Both the allele calls were supported by many reads and was not and we could not resolve the allele calls.
- G750** - Optitype: B*35:33; aHLA-Seq: B*35:03. Both the allele calls were supported by many reads and we could not resolve the allele calls.
- G876** - Optitype: B*15:38; aHLA-Seq: B*15:01. B*15:01 supported by some reads while we did not find reads supporting B*15:38.
- G914** - Optitype: A*02:01; aHLA-Seq: A*02:06. The two alleles differ in two position at the beginning of the first exon where A*02:01 supported by some reads but we did not find unique reads supporting A*02:06.
- G951** - Optitype: C*03:04; aHLA-Seq: C*03:07. The allele C*03:04 is supported by more reads than C*03:07.

Table S6 | HLA class I and class II allele frequencies at 1st field level for the historical samples (2n = 136)

HLA-A			HLA-B			HLA-C			HLA-DRB1			HLA-DQB1			HLA-DPB1		
Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n
02	0.313	26	07	0.160	12	07	0.338	25	04	0.268	22	06	0.375	36	04	0.413	19
03	0.192	16	15	0.147	11	03	0.311	23	15	0.244	20	03	0.322	31	03	0.109	5
01	0.120	10	44	0.147	11	04	0.095	7	01	0.110	9	02	0.145	14	422	0.065	3
24	0.096	8	08	0.107	8	05	0.081	6	13	0.098	8	05	0.135	13	16	0.043	2
26	0.072	6	40	0.093	7	12	0.041	3	03	0.085	7	04	0.02	2	02	0.022	1
68	0.072	6	35	0.053	4	15	0.041	3	07	0.061	5				109	0.022	1
11	0.048	4	51	0.053	4	02	0.027	2	08	0.024	2				11	0.022	1
32	0.036	3	57	0.053	4	16	0.027	2	09	0.024	2				131	0.022	1
29	0.012	1	38	0.040	3	01	0.014	1	11	0.024	2				136	0.022	1
30	0.012	1	18	0.027	2	06	0.014	1	14	0.024	2				138	0.022	1
31	0.012	1	39	0.027	2	14	0.014	1	16	0.024	2				169	0.022	1
36	0.012	1	45	0.027	2				10	0.012	1				20	0.022	1
			55	0.027	2										234	0.022	1
			27	0.013	1										240	0.022	1
			42	0.013	1										258	0.022	1
			56	0.013	1										452	0.022	1
															46	0.022	1
															52	0.022	1
															54	0.022	1
															57	0.022	1
															90	0.022	1

Total number of 2-digit typed alleles at each locus: HLA-A = 83, HLA-B = 75, HLA-C = 74, HLA-DRB1 = 82, HLA-DQB1 = 96; HLA-DPB1 = 46.
 Distinct 2-digit alleles at each locus: HLA-A = 12, HLA-B = 16, HLA-C = 11, HLA-DRB1 = 12, HLA-DQB1 = 5; HLA-DPB1 = 21.

Table S7 | HLA class I and class II allele frequencies at the 2nd field level for the historical samples (2n = 136)

HLA-A			HLA-B			HLA-C			HLA-DRB1			HLA-DQB1			HLA-DPB1		
Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n	Allele	Frequency	n
02:01	0.244	11	07:02	0.204	10	07:01	0.545	6	04:01	0.228	13	06:02	0.266	21	04:01	0.348	16
03:01	0.244	11	08:01	0.122	6	05:01	0.182	2	15:01	0.228	13	03:02	0.203	16	03:01	0.109	5
01:01	0.156	7	40:01	0.122	6	03:07	0.091	1	01:01	0.105	6	02:01	0.127	10	04:02	0.065	3
24:02	0.133	6	44:02	0.122	6	07:02	0.091	1	03:01	0.088	5	05:01	0.101	8	422:01	0.065	3
26:01	0.067	3	51:01	0.082	4	12:03	0.091	1	13:01	0.088	5	03:01	0.089	7	16:01	0.043	2
11:01	0.044	2	57:01	0.082	4				07:01	0.053	3	03:03	0.051	4	02:02	0.022	1
68:01	0.044	2	15:01	0.061	3				04:03	0.035	2	06:03	0.051	4	109:01	0.022	1
02:06	0.022	1	39:06	0.041	2				04:04	0.035	2	06:04	0.051	4	11:01	0.022	1
31:01	0.022	1	44:03	0.041	2				13:02	0.035	2	05:02	0.025	2	131:01	0.022	1
32:01	0.022	1	27:05	0.020	1				14:01	0.035	2	05:03	0.025	2	136:01	0.022	1
			35:01	0.020	1				08:01	0.018	1	03:10	0.013	1	138:01	0.022	1
			35:03	0.020	1				09:01	0.018	1			1	169:02	0.022	1
			45:01	0.020	1				16:01	0.018	1			1	20:01	0.022	1
			55:01	0.020	1				16:02	0.018	1			1	234:01	0.022	1
														1	240:01	0.022	1
														1	258:01	0.022	1
														1	452:01	0.022	1
														1	46:01	0.022	1
														1	52:01	0.022	1
														1	54:01	0.022	1
														1	57:01	0.022	1
														1	90:01	0.022	1

Total number of 4-digit typed alleles at each locus: HLA-A = 45, HLA-B = 49, HLA-C = 11, HLA-DRB1 = 57, HLA-DQB1 = 79; HLA-DPB1 = 46.
 Distinct 4-digit alleles at each locus: HLA-A = 10, HLA-B = 15, HLA-C = 5, HLA-DRB1 = 14, HLA-DQB1 = 11; HLA-DPB1 = 22.

Table S8 | Pairwise global linkage disequilibrium estimates for the historical samples

Locus pair	D'	Wn	p-value
A:B	0.804	0.844	0.0040*
A:C	1	1	0.0000*
A:DPB1	0.858	0.927	0.083
A:DQB1	0.764	0.709	0.0160*
A:DRB1	0.703	0.755	0.0430*
B:DPB1	0.829	0.823	0.0000*
B:DQB1	0.877	0.772	0.0000*
B:DRB1	0.882	0.775	0.0000*
C:B	0.833	0.829	0.0000*
C:DPB1	0	nan	0.0000*
C:DQB1	0.875	0.913	0.0000*
C:DRB1	0.833	0.882	0.0000*
DQB1:DPB1	0.85	0.882	0.459
DRB1:DPB1	0.929	0.917	0.186
DRB1:DQB1	0.984	0.883	0.0000*

Table S9 | HLA two and three-locus haplotype frequencies ($f > 0.03$) for the historical samples calculated using the expectation-maximization algorithm

Loci	Haplotype	Frequency	Copy number
A:B	A*0301:B*0702	0.19231	5
A:B	A*0101:B*0801	0.15385	4
A:B	A*0201:B*0702	0.07692	2
A:B	A*0201:B*4402	0.07692	2
A:B	A*2402:B*0702	0.07692	2
A:B	A*2402:B*3906	0.07692	2
A:C	A*0201:C*0501	0.33333	2
A:C	A*0101:C*0701	0.33333	2
A:DQB1	A*0201:0602	0.11765	4
A:DQB1	A*0101:0201	0.11765	4
A:DQB1	A*0301:0301	0.08824	3
A:DQB1	A*0301:0503	0.05882	2
A:DQB1	A*2402:0602	0.05882	2
A:DQB1	A*0301:0302	0.05882	2
A:DQB1	A*6801:0603	0.05882	2
A:DQB1	A*2402:0303	0.05882	2
A:DQB1	A*2601:0502	0.05882	2
A:DQB1	A*0201:0302	0.05882	2
A:DRB1	A*0301:0401	0.10969	3.7
A:DRB1	A*0201:1501	0.10969	3.7
A:DRB1	A*0301:1501	0.06678	2.3
A:DRB1	A*0201:0401	0.06678	2.3
A:DRB1	A*2402:1501	0.05882	2
A:DRB1	A*6801:1301	0.05882	2
A:DRB1	A*2402:0701	0.05882	2
A:DRB1	A*0101:0301	0.05882	2
A:DRB1	A*2402:0401	0.05882	2
A:DRB1	A*0301:1401	0.05882	2
B:DPB1	B*0801:DPB1*0401	0.16667	2
B:DPB1	B*5701:DPB1*0401	0.16667	2
B:DQB1	B*0702:DQB1*0602	0.22222	8
B:DQB1	B*0801:DQB1*0201	0.16667	6
B:DQB1	B*1501:DQB1*0302	0.08333	3
B:DQB1	B*5101:DQB1*0302	0.05556	2
B:DQB1	B*4001:DQB1*0604	0.05556	2
B:DRB1	B*0702:DRB1*1501	0.26667	8
B:DRB1	B*0801:DRB1*0301	0.13333	4
B:DRB1	B*3906:DRB1*0701	0.06667	2
B:DRB1	B*5101:DRB1*0404	0.06667	2
C:B	C*0701:B*0801	0.5	3
C:DQB1	C*0701:DQB1*0201	0.375	3
C:DQB1	C*0501:DQB1*0602	0.25	2
C:DRB1	C*0501:DRB1*1501	0.33333	2
DRB1:DQB1	DRB1*1501:0602	0.27273	12
DRB1:DQB1	DRB1*0401:0301	0.11364	5
DRB1:DQB1	DRB1*0301:0201	0.11364	5
DRB1:DQB1	DRB1*0401:0302	0.09091	4
DRB1:DQB1	DRB1*0101:0501	0.06818	3
DRB1:DQB1	DRB1*1301:0603	0.06818	3

DRB1:DQB1	DRB1*1401:0503	0.04545	2
DRB1:DQB1	DRB1*0404:0302	0.04545	2
A:B:C	A*0101:B*0801:C*0701	0.5	2
A:DRB1:DQB1	A*0201:DRB1*1501:DQB1*0602	0.125	4
A:DRB1:DQB1	A*0301:DRB1*0401:DQB1*0301	0.09375	3
A:DRB1:DQB1	A*0301:DRB1*1401:DQB1*0503	0.0625	2
A:DRB1:DQB1	A*2402:DRB1*1501:DQB1*0602	0.0625	2
A:DRB1:DQB1	A*6801:DRB1*1301:DQB1*0603	0.0625	2
A:DRB1:DQB1	A*0101:DRB1*0301:DQB1*0201	0.0625	2
B:DRB1:DQB1	B*0702:DRB1*1501:DQB1*0602	0.28571	8
B:DRB1:DQB1	B*0801:DRB1*0301:DQB1*0201	0.14286	4
B:DRB1:DQB1	B*5101:DRB1*0404:DQB1*0302	0.07143	2
C:DRB1:DQB1	C*0501:DRB1*1501:DQB1*0602	0.33333	2

Table S10 | Allele calls at HLA –B and –DRB1 genes for 30 simulated aDNA samples

Simulation ID	Read depth	B1	B2	DRB1 1	DRB1 2
genotype 1		B*07:02	B*40:08	DRB1*04:01	DRB1*15:01
simulation 1	1x	NA	NA	DRB1*04	NA
simulation 1	5x	B*07	B*40:08	DRB1*04	DRB1*15
simulation 1	10x	B*07	B*40:08	DRB1*04	DRB1*15
simulation 1	30x	B*07	B*40:08	DRB1*04:01	DRB1*15:01
simulation 1	60x	B*07:02	B*40:08	DRB1*04:01	DRB1*15:01
genotype 2		B*07:02	B*51:01	DRB1*10:04	DRB1*15:01
simulation 2	1x	NA	NA	NA	NA
simulation 2	5x	B*07	B*51	DRB1*10	DRB1*15
simulation 2	10x	B*07	NA	DRB1*10:04	DRB15:01
simulation 2	30x	B*07:02	B*51:01	DRB1*10:04	DRB15:01
simulation 2	60x	B*07:02	B*51:01	DRB1*10:04	DRB15:01
genotype 3		B*13:04	B*45:01	DRB1*01:01	DRB1*08:04
simulation 3	1x	NA	NA	DRB1*01	DRB1*08
simulation 3	5x	B*13:04	B*45:01	DRB1*01	NA
simulation 3	10x	B*13:04	B*45:01	DRB1*01:01	DRB1*08
simulation 3	30x	B*13:04	B*45:01	DRB1*01:01	DRB1*08:04
simulation 3	60x	B*13:04	B*45:01	DRB1*01:01	DRB1*08:04
genotype 4		B*51:01	B*13:04	DRB1*10:04	DRB1*01:01
simulation 4	1x	NA	NA	NA	NA
simulation 4	5x	B*51	B*13	DRB1*10:04	DRB1*01:01
simulation 4	10x	B*51:01	B*13	DRB1*10:04	DRB1*01:01
simulation 4	30x	B*51:01	B*13	DRB1*10:04	DRB1*01
simulation 4	60x	B*51:01	B*13:04	DRB1*10:04	DRB1*01:01
genotype 5		B*45:01	B*40:08	DRB1*08:04	DRB1*04:01
simulation 5	1x	NA	NA	NA	NA
simulation 5	5x	B*45:01	NA	NA	DRB1*04:01
simulation 5	10x	NA	NA	DRB1*08:04	NA
simulation 5	30x	NA	NA	DRB1*08:04	DRB1*04:01
simulation 5	60x	B*45:01	B*04:08	DRB1*08:04	DRB1*04:01
genotype 6		B*45:01	B*51:01 (153 T > C)	DRB1*08:04 (217 C > G)	DRB1*10:04 (164 G > T)
simulation 6	1x	NA	NA	NA	DRB*10
simulation 6	5x	NA	B*51 (153 T > C)	DRB1*08:04 (217 C > G)	DRB1*10:04 (164 G > T)
simulation 6	10x	B*45:01	B*51:01 (153 T > C)	DRB1*08:04 (217 C > G)	DRB1*10:04 (164 G > T)
simulation 6	30x	B*45:01	B*51:01 (153 T > C)	DRB1*08:04 (217 C > G)	DRB1*10:04 (164 G > T)
simulation 6	60x	B*45:01	B*51:01 (153 T > C)	DRB1*08:04 (217 C > G)	DRB1*10:04 (164 G > T)

Table S11 | 31 individuals from the 1000 Genomes Project used in the validation test

ID	File1	File2	Population
NA19093	SRR100033_1	SRR100033_2	Yoruba (Africa)
NA19098	SRR077453_1	SRR077453_2	Yoruba (Africa)
NA19334	SRR100001_1	SRR100001_2	Luhya (Africa)
NA19332	SRR099997_1	SRR099997_2	Luhya (Africa)
HG00452	ERR031838_1	ERR031838_2	Han (East Asia)
HG00457	ERR031839_1	ERR031839_2	Han (East Asia)
NA20504	SRR748294_1	SRR748294_2	Toscani (Europe)
NA20505	SRR766033_1	SRR766033_2	Toscani (Europe)
HG00736	SRR099974_1	SRR099974_2	Puerto_Rico (Americas)
HG00737	SRR099984_1	SRR099984_2	Puerto_Rico (Americas)
HG01205	SRR098489_1	SRR098489_2	Puerto_Rico (Americas)
HG01241	SRR099990_1	SRR099990_2	Puerto_Rico (Americas)
HG01242	SRR098493_1	SRR098493_2	Puerto_Rico (Americas)
NA18501	SRR100022_1	SRR100022_2	Yoruba (Africa)
NA18504	SRR100028_1	SRR100028_2	Yoruba (Africa)
NA18516	SRR100026_1	SRR100026_2	Yoruba (Africa)
NA18517	ERR034551_1	ERR034551_2	Yoruba (Africa)
NA07000	SRR766039_1	SRR766039_2	Utah (European ancestry)
NA07037	ERR034542_1	ERR034542_2	Utah (European ancestry)
NA07048	SRR099452_1	SRR099452_2	Utah (European ancestry)
NA10851	SRR766044_1	SRR766044_2	Utah (European ancestry)
NA18939	SRR766031_1	SRR766031_2	Japan (East Asia)
NA18940	ERR034596_1	ERR034596_2	Japan (East Asia)
NA19794	SRR748785_1	SRR748785_2	Mexico (Americas)
NA19795	SRR708374_1	SRR708374_2	Mexico (Americas)
HG00097	SRR765989_1	SRR765989_2	Great_Britain (Europe)
HG00099	SRR765993_1	SRR765993_2	Great_Britain (Europe)
HG00100	SRR099966_1	SRR099966_2	Great_Britain (Europe)
NA18534	ERR034577_1	ERR034577_2	China (East Asia)
NA18536	ERR034578_1	ERR034578_2	China (East Asia)
NA18542	ERR031855_1	ERR031855_2	China (East Asia)

Table S12 | SBT-based HLA-B genotypes for the 31 individuals from the 1000 Genomes Project used in the validation test.
Information obtained from Gourraud et al. 2014.

ID	B 1	B 2
NA19093	B*35:01:00	B*53:01:00
NA19098	B*51:01:00	B*53:01:00
NA19334	B*15:10	B*45:01/45:07
NA19332	B*15:10	B*58:02:00
HG00452	B*40:01:01/40:01:02/40:55	B*55:02:01
HG00457	B*39:01:01/39:01:01/02L/39:01:03/39:46	B*46:01:01/46:15N
NA20504	B*55:01:01/55:01:03	B*57:01:01
NA20505	B*35:03:01/35:70	B*57:01:01
HG00736	B*08:01:01/08:19N	B*58:01:01/58:11
HG00737	B*35:01:01/35:01:03/35:40N/35:42/35:57/35:94	B*51:01:01/51:01:05/51:01:07/51:11N/51:30/51:32/51:48/51:51
HG01205	B*07:02:01/07:02:06/07:02:09/07:44/07:49N/07:58/07:59/07:61	B*35:01:01/35:01:03/35:40N/35:42/35:57/35:94
HG01241	B*07:02:01/07:02:06/07:02:09/07:44/07:49N/07:58/07:59/07:61	B*15:03:01/15:103
HG01242	B*35:01:01/35:01:03/35:40N/35:42/35:57/35:94	B*44:03:01/44:03:03/44:03:04
NA18501	B*14:01	B*78:01:00
NA18504	B*15:03	B*39:10:00
NA18516	B*15:10	B*53:01:00
NA18517	B*49:01:00	B*51:01:00
NA07000	B*44:02:00	B*40:01:00
NA07037	B*15:10	B*40:01:00
NA07048	B*44:02:01:01	B*07:02
NA10851	B*40:01:00	B*08:01
NA18939	B*27:04:01	B*67:01:01
NA18940	B*46:01:00	B*52:01:00
NA19794	B*51:01:01/51:01:05/51:01:07/51:11N/51:30/51:32/51:48/51:51	B*56:01/56:24
NA19795	B*35:17:00	B*48:01:01/48:09
HG00097	B*07:02:01/07:02:06/07:02:09/07:44/07:49N/07:58/07:59/07:61	B*07:02:01/07:02:06/07:02:09/07:44/07:49N/07:58/07:59/07:61
HG00099	B*08:01:01/08:19N	B*44:02:01:01/44:02:01:02S/44:19N/44:27/44:66
HG00100	B*08:01:01/08:19N	B*57:01:01
NA18534	B*15:01:01/15:01:02N/15:01:06/15:01:07/15:102/15:104/15:140/15:146	B*58:01:01/58:11
NA18536	B*48:01:01/48:09	B*51:02:01
NA18542	B*46:01:00	B*58:01:00

Table S13 | SBT-based HLA-DRB1 genotypes for the 31 individuals from the 1000 Genomes Project used in the validation test. Information obtained from Gourraud et al. 2014.

ID	DRB1 1	DRB1 2
NA19093	DRB1*15:03	DRB1*13:01
NA19098	DRB1*01:02	DRB1*07:01
NA19334	DRB1*03:01:01/03:01:01:02	DRB1*08:04
NA19332	DRB1*03:01:01/03:01:01:02	DRB1*13:01:01
HG00452	DRB1*01:01:01	DRB1*11:01:01/11:01:08
HG00457	DRB1*15:01:01/15:01:01:02	DRB1*09:01:02
NA20504	DRB1*14:01:01/14:54	DRB1*04:04
NA20505	DRB1*16:01:01	DRB1*07:01:01/07:01:01:02
HG00736	DRB1*16:01:01	DRB1*03:01:01/03:01:01:02
HG00737	DRB1*13:01:01	DRB1*14:01:01/14:54
HG01205	DRB1*01:01:01	DRB1*13:01:01
HG01241	DRB1*15:01:01/15:01:01:02	DRB1*13:02:01
HG01242	DRB1*13:01:01	DRB1*13:01:01
NA18501	DRB1*01:02	DRB1*13:01
NA18504	DRB1*03:01	DRB1*01:02
NA18516	DRB1*08:04	DRB1*03:02
NA18517	DRB1*12:01	DRB1*13:03
NA07000	DRB1*03:01	DRB1*11:01:01/11:01:08
NA07037	DRB1*04:04	DRB1*13:02
NA07048	DRB1*04:01	DRB1*15:01
NA10851	DRB1*04:04:01	DRB1*07:01
NA18939	DRB1*15:01:01/15:01:01:02	DRB1*15:01:01/15:01:01:02
NA18940	DRB1*15:02	DRB1*08:02
NA19794	DRB1*01:01:01	DRB1*11:04:01
NA19795	DRB1*16:02:01	DRB1*08:02
HG00097	DRB1*15:01:01/15:01:01:02	DRB1*13:03:01
HG00099	DRB1*03:01:01/03:01:01:02	DRB1*11:01:01/11:01:08
HG00100	DRB1*03:01:01/03:01:01:02	DRB1*07:01:01/07:01:01:02
NA18534	DRB1*03:01:01/03:01:01:02	DRB1*04:06:01/04:06:02
NA18536	DRB1*15:01:01/15:01:01:02	DRB1*08:03:02
NA18542	DRB1*09:01	DRB1*03:01

Table S14 | HLA genotypes obtained with our approach at HLA-B and HLA-DRB1 loci, for the 31 individuals from the 1000 Genomes Project used in the validation test

ID	B 1	B 2	DRB1 1	DRB1 2
NA19093	B*53:01:01G	B*53:01:01G	DRB1*13:02:01G	DRB1*15:03:01G
NA19098	B*51:01:01G	B*53:01:01G	DRB1*01:02:01G	DRB1*7:01:01G
NA19334	B*15:10:01	B*45:xx	DRB1*03:01:01G	DRB1*08:04:01
NA19332	B*15: xx	B*58:02:01	DRB1*03:01:01G/03:07:01G	DRB1*13:01:01G/13:27:01
HG00452	B*40:xx	B*55:02:01G	DRB1*01:01:01G	DRB1*11:01:01G
HG00457	B*39:01:01G	B*46:01:01G	DRB1*09:01:02G	DRB1*15:01:01G
NA20504	B*55:01:01G	B*57:01:01G	DRB1*04:04:01	DRB1*14:01:01G
NA20505	B*35:03:01G	B*57:01:01G	DRB1*07:01:01G	DRB1*16:01:01
HG00736	B*08:01:01G	B*58:01:xx	DRB1*03:01:01G	DRB1*16:01:01
HG00737	B*35:01:01G	B*51:01:01G	DRB1*13:01:01G	DRB1*14:01:01G
HG01205	B*07:02:01G	B*35:01:01G	DRB1*01:01:01G	DRB1*13:01:01G
HG01241	B*07:02:01G	B*15:03:01G	DRB1*13:02:01G	DRB1*15:01:01G
HG01242	B*35:01:01G/35:37:01	B*44:03:01G	DRB1*13:01:01G	DRB1*13:01:01G
NA18501	B*14:01:01G	B*78:01:01G	DRB1*01:02:01G	DRB1*13:01:01G
NA18504	B*15:03:01G	B*39:10:01	DRB1*01:23:01	DRB1*03:01:01G
NA18516	B*15:10:01	B*53:01:01G	DRB1*03:02:01	DRB1*08:04:01
NA18517	B*49:xx	B*51:01:01G	DRB1*12:01:01G	DRB1*13:03:01G
NA07000	B*40:01:01G	B*44:02:01G	DRB1*03:01:01G	DRB1*11:01:01G
NA07037	B*15:10:01	B*40:01:01G/40:30:01	DRB1*04:04:01	DRB1*13:02:01G
NA07048	B*07:02:01G	B*44:02:01G	DRB1*04:01:01G	DRB1*15:01:01G
NA10851	B*08:12:01	B*40:80:01	DRB1*04:04:01	DRB1*07:01:01G
NA18939	B*27:04:01G	B*67:01:01	DRB1*15:01:01G	DRB1*15:01:01G
NA18940	B*46:01:01G/46:32:01	B*52:01:01G	DRB1*08:02:01G	DRB1*15:02:01G
NA19794	B*51:01:01G	B*56:01:01G	DRB1*01:01:01	DRB1*11:04:01
NA19795	B*35:17:01	B*48:01:01G	DRB1*08:02:01G	DRB1*16:02:01G
HG00097	B*07:02:01G	B*07:02:01G	DRB1*13:03:01G	DRB1*15:01:01G
HG00099	B*08:01:01G	B*44:02:01G/44:20:01	DRB1*03:01:01G	DRB1*11:01:01G
HG00100	B*08:xx	B*57:01:01G	DRB1*03:01:01G	DRB1*07:01:01G
NA18534	B*15:xx	B*58:01:01G	DRB1*03:01:01G/03:05:01	DRB1*04:06:01G
NA18536	B*48:01:01G	B*51:02:01G	DRB1*08:03:02G	DRB1*15:01:01G
NA18542	B*46:01:01G	B*58:01:01G	DRB1*03:01:01G	DRB1*09:01:02G

Note – Alleles reported in black are those with identical call from the two approaches. In blue are reported the cases for which we found support for our allele call and we could not support the call provided in Gourraud et al. 2014. Alleles reported in red are instead those for which we could not confirm the calls from our approach. Called alleles that differed between the two approaches:

NA19093 - Gourraud et al. 2014: B*35:01:00; aHLA-Seq: B*53:01:01G, the former supported by a few reads.

NA19093 - Gourraud et al. 2014: DRB1*13:01; aHLA-Seq: DRB1*13:02:01G, the latter supported by six reads.

NA18504 - Gourraud et al. 2014: DRB1*01:02; aHLA-Seq: DRB1*01:23:01, the latter supported by 19 reads.

NA10851 - Gourraud et al. 2014: B*08:01; aHLA-Seq: B*08:12:01, the former supported by a few reads.

NA10851 - Gourraud et al. 2014: B*40:01:00; aHLA-Seq: B*40:80:01, the former supported by a few reads.

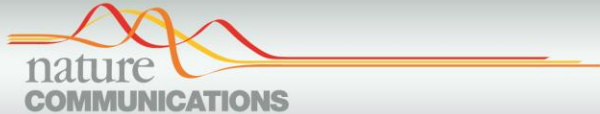
Chapter III**Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans**

Ben Krause-Kyora¹, Marcel Nutsua¹, Lisa Boehme¹, Federica Pierini², Dorthe Dangvard Pedersen³, Sabin-Christin Kornell¹, Dmitriy Drichel⁴, Marion Bonazzi¹, Lena Möbus¹, Peter Tarp³, Julian Susat¹, Esther Bosse¹, Beatrix Willburger⁵, Alexander H. Schmidt⁵, Jürgen Sauter⁵, Andre Franke¹, Michael Wittig¹, Amke Caliebe⁶, Michael Nothnagel⁴, Stefan Schreiber^{1,7}, Jesper L. Boldsen³, Tobias L. Lenz², Almut Nebel¹

¹Institute of Clinical Molecular Biology, Kiel University, 24105 Kiel, Germany, ²Research Group for Evolutionary Immunogenomics, Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany, ³Unit of Anthropology (ADBOU), Department of Forensic Medicine, University of Southern Denmark, 5260 Odense S, Denmark, ⁴Cologne Center for Genomics (CCG), Department of Statistical Genetics and Bioinformatics, University of Cologne, 50931 Cologne, Germany, ⁵DKMS, 72072 Tübingen, Germany, ⁶Institute of Medical Informatics and Statistics, Kiel University, 24105 Kiel, Germany, ⁷Clinic for Internal Medicine I, University Hospital of Schleswig-Holstein, 24105 Kiel, Germany.

Published in
Nature Communications (2018)
doi: 10.1038/s41467-018-03857-x

Supporting information with supplementary notes, figures and tables are available online.



ARTICLE

DOI: 10.1038/s41467-018-03857-x

OPEN

Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans

Ben Krause-Kyora^{1,2}, Marcel Nutsua¹, Lisa Boehme¹, Federica Pierini³, Dorthe Dangvard Pedersen⁴, Sabin-Christin Kornell¹, Dmitriy Drichel⁵, Marion Bonazzi¹, Lena Möbus¹, Peter Tarp⁴, Julian Susat¹, Esther Bosse¹, Beatrix Willburger⁶, Alexander H. Schmidt⁶, Jürgen Sauter⁶, Andre Franke¹, Michael Wittig¹, Amke Caliebe⁷, Michael Nothnagel⁵, Stefan Schreiber^{1,8}, Jesper L. Boldsen⁴, Tobias L. Lenz³ & Almut Nebel¹

Leprosy, a chronic infectious disease caused by *Mycobacterium leprae* (*M. leprae*), was very common in Europe till the 16th century. Here, we perform an ancient DNA study on medieval skeletons from Denmark that show lesions specific for lepromatous leprosy (LL). First, we test the remains for *M. leprae* DNA to confirm the infection status of the individuals and to assess the bacterial diversity. We assemble 10 complete *M. leprae* genomes that all differ from each other. Second, we evaluate whether the human leukocyte antigen allele DRB1*15:01, a strong LL susceptibility factor in modern populations, also predisposed medieval Europeans to the disease. The comparison of genotype data from 69 *M. leprae* DNA-positive LL cases with those from contemporary and medieval controls reveals a statistically significant association in both instances. In addition, we observe that DRB1*15:01 co-occurs with DQB1*06:02 on a haplotype that is a strong risk factor for inflammatory diseases today.

¹Institute of Clinical Molecular Biology, Kiel University, Kiel 24105, Germany. ²Max Planck Institute for the Science of Human History, Jena 07745, Germany. ³Department of Evolutionary Ecology, Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön 24306, Germany. ⁴Department of Forensic Medicine, Unit of Anthropology (ADBOU), University of Southern Denmark, Odense S 5260, Denmark. ⁵Department of Statistical Genetics and Bioinformatics, Cologne Center for Genomics (CCG), University of Cologne, Cologne 50931, Germany. ⁶DKMS, Tübingen 72072, Germany. ⁷Institute of Medical Informatics and Statistics, Kiel University, Kiel 24105, Germany. ⁸Clinic for Internal Medicine, University Hospital of Schleswig-Holstein, Kiel 24105, Germany. These authors contributed equally: Marcel Nutsua, Lisa Boehme, Federica Pierini, Dorthe Dangvard Pedersen, Sabin-Christin Kornell. These authors jointly supervised this work: Jesper L. Boldsen, Tobias L. Lenz, Almut Nebel. Correspondence and requests for materials should be addressed to B.K.-K. (email: b.krause-kyora@ikmb.uni-kiel.de)

Leprosy is a chronic infectious disease caused by *Mycobacterium leprae* (*M. leprae*). It was very common during the Middle Ages in Europe, from where it almost completely disappeared in the 16th century¹. A recent ancient DNA (aDNA) analysis has revealed a high level of *M. leprae* genome conservation over the past 1000 years, indicating that the leprosy epidemic during the European Middle Ages was unlikely to be due to particularly virulent strains². Instead, other factors such as malnutrition, co-infections, and host genetics may have increased disease susceptibility in the medieval period. At present, leprosy is virtually absent in Europe, but still remains a big health problem in South-East Asia (e.g., India), North and Central Africa (Central African Republic, Democratic Republic of the Congo), Oceania (Indonesia, Papua New Guinea) and the Americas (Brazil, Mexico)³. All 10 modern human *M. leprae* genomes sequenced up to now fall in five distinct phylogenetic branches that show a specific geographic distribution pattern².

In addition to environmental factors, predisposition to leprosy is considerably influenced by variation in immune-related genes⁴. Very strong disease associations were reported for the human leukocyte antigen (HLA) region^{4–6}. Although the involvement of both HLA class I and class II genes was intensively studied in leprosy⁶, class II alleles, particularly in the DRB1 locus, were shown to be most consistently associated with the disease^{5–8}. Of these, DRB1*15:01 is the most notable risk factor for lepromatous leprosy (LL) in India, China, and Brazil today^{7–9}. LL represents a severe form of the disease, characterized by a high bacterial load and specific lesions that can reliably be diagnosed on bones^{1,10}. DRB1*15:01 influences LL susceptibility in very diverse populations, therefore, it remains an interesting open question whether the same association also existed in Europeans of the Middle Ages. We tested this hypothesis by an aDNA analysis in medieval skeletons with signs of LL.

Here, we isolated aDNA from remains with LL lesions from one locale, the St. Jørgen leprosarium in Odense, Denmark¹¹ (historically dated 1270–1550 AD; Fig. 1). First, we analyzed extracts for the presence of *M. leprae* DNA to confirm the infection status of the individuals and to assess the bacterial diversity. We assembled 10 complete *M. leprae* genomes that all differ from each other and represent three different branches. Second, we performed an aDNA association study to evaluate whether the risk allele DRB1*15:01 predisposed medieval Europeans to LL. The comparison of genotype data from 69 *M. leprae* DNA-positive cases with those from contemporary and medieval controls revealed a statistically significant association in both instances. In addition, we studied the DRB1 locus and its haplotype structure in more detail in the LL cases. We observed that DRB1*15:01 co-occurred on a haplotype that also carried DQB1*06:02; this haplotype is a strong risk factor for inflammatory diseases in present-day populations.

Results

Ancient association study with LL. DNA extracts from 85 specimens with LL-specific bone lesions from the St. Jørgen cemetery were investigated for the DRB1*15:01 allele (Supplementary Note 1). Direct Sanger sequencing of the DRB1 exon is not possible from aDNA, because the length of the relevant exon coding for the antigen-binding domain far exceeds the average DNA fragment size of ancient samples and the polymorphic exon sequence precludes the use of intra-exon primers. However, the SNP allele rs3135388-T is an established marker for high-throughput genotyping of DRB1*15:01 in disease studies^{12,13}. Genotyping a single SNP requires only a very short amplicon size, which renders this marker much more suitable for analysis of the highly fragmented aDNA. When we performed PCR-based

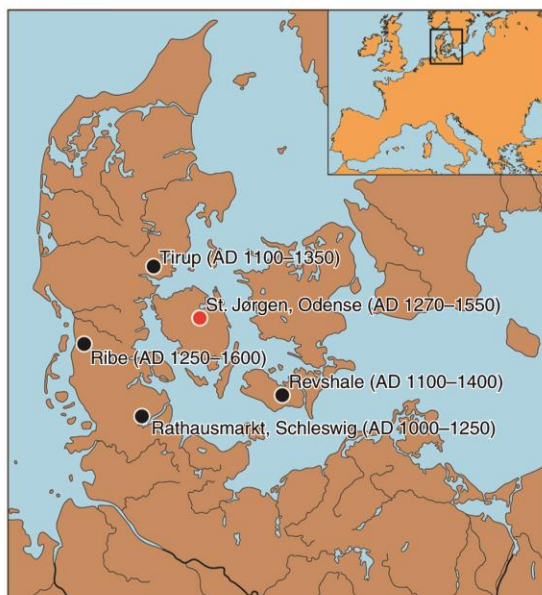


Fig. 1 Origin of medieval samples. Geographic location of the medieval cemeteries in southern Denmark and northern Germany from where the specimens in this study were obtained. Dates in parentheses indicate the time span of active use as cemeteries. Red dot marks the cemetery with LL-positive samples. LL: lepromatous leprosy. The software CorelDRAW was used to create the map (designed by B.K.-K.)

Sanger sequencing for the SNP allele in the 85 St. Jørgen specimens, 69 of them contained sufficient endogenous nuclear DNA to yield genotype data (Supplementary Data 1, Supplementary Note 2, Supplementary Fig. 1, Supplementary Table 1). These 69 LL samples also tested positive for the presence of *M. leprae* DNA by shotgun high-throughput sequencing (HTS) (see below and Supplementary Table 2, 3) and/or screening PCR (Supplementary Data 1).

We first performed an association test for rs3135388 and medieval leprosy considering the 69 *M. leprae*-positive LL individuals from St. Jørgen as the case group and a large sample of contemporary northern Germans as controls, using DRB1 genotype information from an extensive bone marrow database. The rs3135388-T allele frequency in the LL individuals (0.283) was significantly higher than in the controls (0.138; $p = 9.49 \times 10^{-06}$, OR = 2.46; two-sided Fisher's exact test, Fig. 2, Table 1). Subsequently, we compared the rs3135388-T frequency in the 69 St. Jørgen individuals with genetic information from those in medieval controls. Such a case-control study would require—besides the herein-studied 138 alleles of our cases—another 250 alleles in controls to obtain an a priori power of 75% (see Methods). We were able to generate genotype data from 152 randomly sampled individuals (i.e., 304 alleles) excavated from four medieval sites in Denmark and northern Germany (Fig. 1, Table 1, Supplementary Note 1). Although the rs3135388-T frequencies did not differ significantly between the four cemeteries ($p = 0.932$, two-sided Fisher's exact test), LL cases from St. Jørgen showed a significant enrichment of this allele compared to the combined set of medieval controls (0.283 vs. 0.184; $p = 0.024$; OR = 1.74; two-sided Fisher's exact test, Table 1, Fig. 2).

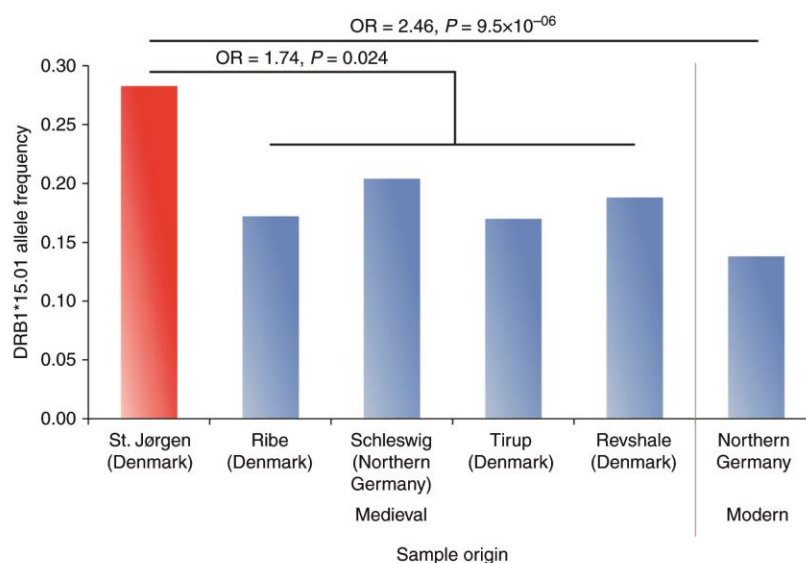


Fig. 2 Frequency of rs3135388-T variant. Allele frequency of rs3135388-T in 69 LL-positive medieval St. Jørgen individuals (red bar) and in four sample sets from medieval unaffected controls, compared with DRB1*15:01 data in a contemporary cohort of northern Germany from the DKMS bone marrow database (see Table 1 and Supplementary Information for sample sizes and other details). Odds ratio (OR) and *P* value from Fisher's exact test are reported. LL: lepromatous leprosy

***M. leprae* genome analysis and metagenomic screening.** DNA extracts from 68 of the 69 specimens with LL-specific bone lesions were successfully subjected to HTS, without any prior *M. leprae* or human genome enrichment (Supplementary Table 4). Damage patterns of both human and *M. leprae* reads were consistent with an ancient origin of the DNA (Supplementary Note 2, 3, Supplementary Table 2, 5–9). Metagenomic analysis of the HTS reads using MALT¹⁴ showed the typical spectrum of soil bacteria and no signs of bacterial co-infection (Supplementary Note 3, Supplementary Fig. 2, 3). For ten specimens, the *M. leprae* genome coverage was sufficient to allow for de novo assemblies of the HTS data without any enrichment bias (Supplementary Note 3, Supplementary Table 2, 10–13). Genome-wide comparisons and SNP effect analysis did not show any variants that would indicate a change in virulence or function in these ten genomes (Supplementary Note 3, Supplementary Data 2). We then included the ten well-covered *M. leprae* genomes together with another genome from the St. Jørgen cemetery (Jørgen_625), four additional medieval genomes and 12 modern samples, all published elsewhere², in a phylogenetic analysis. The eleven genomes from St. Jørgen fell into 3 branches (Fig. 3, Supplementary Note 4, Supplementary Fig. 4–7, Supplementary Data 3).

HTS-based HLA typing. Given the association between LL and rs3135388-T, which is known to be statistically associated with the presence of DRB1*15:01 in Europeans, we subsequently focused on the HLA class II region in more detail in the medieval LL individuals. We adapted an existing DNA capture¹⁵ method and optimized it for short aDNA fragments. After HTS, we called HLA alleles using a new aDNA-optimized analysis pipeline (Supplementary Fig. 8). Depending on each sample's aDNA quality and resulting sequence coverage, HLA alleles were called at two different levels of resolution, following the official HLA nomenclature¹⁶. The broader two-digit (now also called 1st-field) resolution defines different functional lineages of alleles, yielding a level of separation largely equivalent to the classical serotyping

used in early transplantation medicine (e.g., DRB1*15 vs. DRB1*01). The finer four-digit (now also called 2nd field) resolution separates alleles that share functional properties but differ in their protein sequence (e.g., DRB1*15:01 vs. DRB1*15:02, differing in 1 out of 89 codons of the molecule's antigen-binding groove). Using our target capture and HTS approach, we were able to determine two- or four-digit resolution for HLA class II genes (Supplementary Table 14, 15). Of the 136 alleles investigated (= 68 screened individuals), we succeeded in calling 82 DRB1 alleles at the two-digit level and of these, 57 alleles at the four-digit resolution (Supplementary Table 14). In some samples, the DNA quality and thus read coverage did not allow for more precise allele calls. Among the 82 DRB1 alleles at two-digit resolution, 20 were determined to belong to the DRB1*15 lineage. In modern Europeans, by far the most common DRB1*15 molecule variant is DRB1*15:01 (allele frequency in northern Germany: 0.138, compared with 0.007 of the second most common allele *15:02), rendering it highly likely that those DRB1*15 calls represent the allele DRB1*15:01. This was confirmed as 13 of the 20 DRB1*15 alleles could also be called at four-digit resolution, in each case showing DRB1*15:01. If all 20 DRB1*15 alleles were to consist of the DRB1*15:01 allele, this would result in a frequency of 0.24, very similar to 0.28 as measured by rs3135388 in the 69 individuals. The deviation between the two frequency estimates is owing to the much smaller number of LL cases that could successfully be genotyped by HLA sequencing compared with SNP genotyping. In the 43 LL samples for which allele calls were available from both PCR- and HTS-based methods, we confirmed that rs3135388-T always co-occurred with DRB1*15:01 ($r^2 = 1.0$) (Supplementary Table 15, Supplementary Data 1), verifying the suitability of the T allele for detecting the presence of DRB1*15:01. Interestingly, in all carriers of DRB1*15:01 (representing 12 individuals, including one homozygote), we also observed the allele DQB1*06:02. The same was true in reverse, i.e., all individuals for which DQB1*06:02 could be called ($n = 20$) also had either DRB1*15:01, DRB1*15 (not

Table 1 Frequency of the rs3135388-T allele

	LL cases	LL controls				
		Medieval			Modern	
Site	Odense, St. Jørgen (Denmark)	Ribe (Denmark)	Schleswig, Rathausmarkt (northern Germany)	Tirup (Denmark)	Revshale (Denmark)	Northern Germany*
Dating	1270–1550 AD	1250–1600 AD	1000–1250 AD	1100–1350 AD	1100–1400 AD	modern
Number of individuals	69	32	49	47	24	129,336
Number of rs3135388-T alleles	39	11	20	16	9	35,681
Allele frequency	0.283	0.172	0.204	0.170	0.188	0.138

LL lepromatous leprosy. *HLA DRB1*15:01 data provided by DKMS (Supplementary Note 6)

allowing for more precise allele call), or had an incomplete allele call (because of insufficient aDNA quality). This observation suggests strong linkage disequilibrium (LD) between the two loci (Supplementary Note 5, Supplementary Table 15). The two alleles indeed define a DRB1-DQB1 haplotype that is still found in modern Europeans at a considerable frequency (13.3%, provided by DKMS, Supplementary Note 6).

Binding properties of DRB1 alleles. As the binding and presentation of antigenic peptides to immune effector cells is the key function of HLA molecules, we also explored the relative binding properties of the detected HLA alleles¹⁷ with regard to potential *M. leprae* antigens. We found that among 18 contemporarily common DRB1 alleles, DRB1*15:01 is predicted to bind the second-smallest number of potential *M. leprae* antigens (DRB1*15:01: 11 out of 5345 peptides, mean of 18 DRB1 alleles: 64.7; Fig. 4). HLA binding-prediction for the entire *M. leprae* proteome (516,303 unique peptides) still revealed limited relative presentation capacity for DRB1*15:01, but to a lesser extent (Supplementary Fig. 9). The fact that peptide binding of DRB1*15:01 is relatively more limited when focusing on potential antigenic proteins suggests that it might be particularly ineffective in the context of antigen presentation. Limited antigen presentation could impair specific immunity against *M. leprae* infections and thus confer susceptibility to its carriers, which is exactly what we found in the association analysis above.

Discussion

Leprosy was endemic during the Middle Ages in Europe, where it reached its greatest prevalence between AD 1200 and 1400¹. The disease was greatly feared because it caused visible disfigurement, was incurable and contagious. Stigmatization and social segregation of patients were common in the medieval period; especially those affected with the severe form of LL were quarantined in so-called leproseries or “leper houses”. Deceased LL patients were buried in special leproseries-associated cemeteries as, for instance, in St. Jørgen in the Danish town of Odense^{10,11}. In the present aDNA study, we successfully analyzed 69 human remains from St. Jørgen. All specimens showed typical LL bone lesions and were *M. leprae* DNA-positive.

From 10 samples with the best *M. leprae* DNA content, complete bacterial genomes were generated without any enrichment, which—together with a previously published one²—resulted in a total of eleven genomes from St. Jørgen alone. This high success rate is remarkable and could be due to the very good preservation of *M. leprae* DNA in general² or certain favorable environmental conditions at this site. It could also be a consequence of the high bacterial load associated with the severe LL form of leprosy¹⁸. Of note, the number of medieval *M. leprae* genomes published so far ($n = 15$, including this study) exceeds that of modern bacteria

and has led to a considerable increase in the phylogenetic and temporal resolution. The 11 St. Jørgen genomes differed from each other and fell into three branches (Fig. 3, Supplementary Note 3, 4, Supplementary Fig. 4–7). In our relatively small data set, we did not observe any correlation between *M. leprae* strains and the HLA allele DRB1*15:01. We identified an ancient strain in branch 0 with the SNP type 3K. This type has previously been reported in medieval Hungarians and Byzantines^{19,20}; today it shows a wide geographic distribution ranging from the Near East to Oceania^{2,19,20}. Furthermore, 10 of the 11 St. Jørgen genomes clustered in branches 2 and 3, representing two types that were already described in skeletons from the European Middle Ages^{2,18,19}. Among the St. Jørgen samples, branch 3-strains were most abundant ($n = 9$) and very similar to extant strains recently identified in European red squirrels²¹ (Fig. 3). Interestingly, modern branch 3-bacteria have the ability to infect at least three different hosts including humans and squirrels²¹. The presence of so many different strains in one branch and in a rather remote locale over a few hundred years is surprising. In view of the very low mutation rate of the pathogen and the relatively short time window, one would expect more similar, even identical, genomes. This interesting observation indicates multiple sources of infection with different strains. In agreement with previous observations^{2,22}, no variants in the *M. leprae* genomes were identified that would suggest a change in bacterial virulence or function. In particular, there were no apparent changes in the outer membrane peptides that could lead to different binding affinities to HLA class II proteins. This finding highlights the importance of investigating host genetic factors to explain the high disease prevalence in the medieval period.

As the HLA allele DRB1*15:01 is the most notable risk factor for LL in various modern populations worldwide^{4,5,7,8}, we wanted to test the hypothesis whether this HLA allele also predisposed Europeans to the disease during the medieval leprosy epidemic. To this end, we conducted a two-stage association analysis for SNP rs3135388. Its T allele is an established and reliable marker for DRB1*15:01^{12,13}. First, we compared rs3135388 data from 69 St. Jørgen LL samples, which were all *M. leprae* DNA-positive, with those from contemporary controls. Such an analysis between medieval and modern allele frequencies is valid under the assumption that the common genetic make-up of the Danish population has not significantly changed since the 11th century. This assumption is corroborated by the observation that those 53 St. Jørgen samples that were of sufficient data quality fell within the variability of modern northern Europeans (Fig. 5, Supplementary Note 7). In addition, mitochondrial DNA (mtDNA) haplogroup frequencies at St. Jørgen were comparable to those of northern Europeans today (Supplementary Data 1). These findings indicate no major genome-wide changes in the Danish population structure in the past 1000 years. Furthermore, based

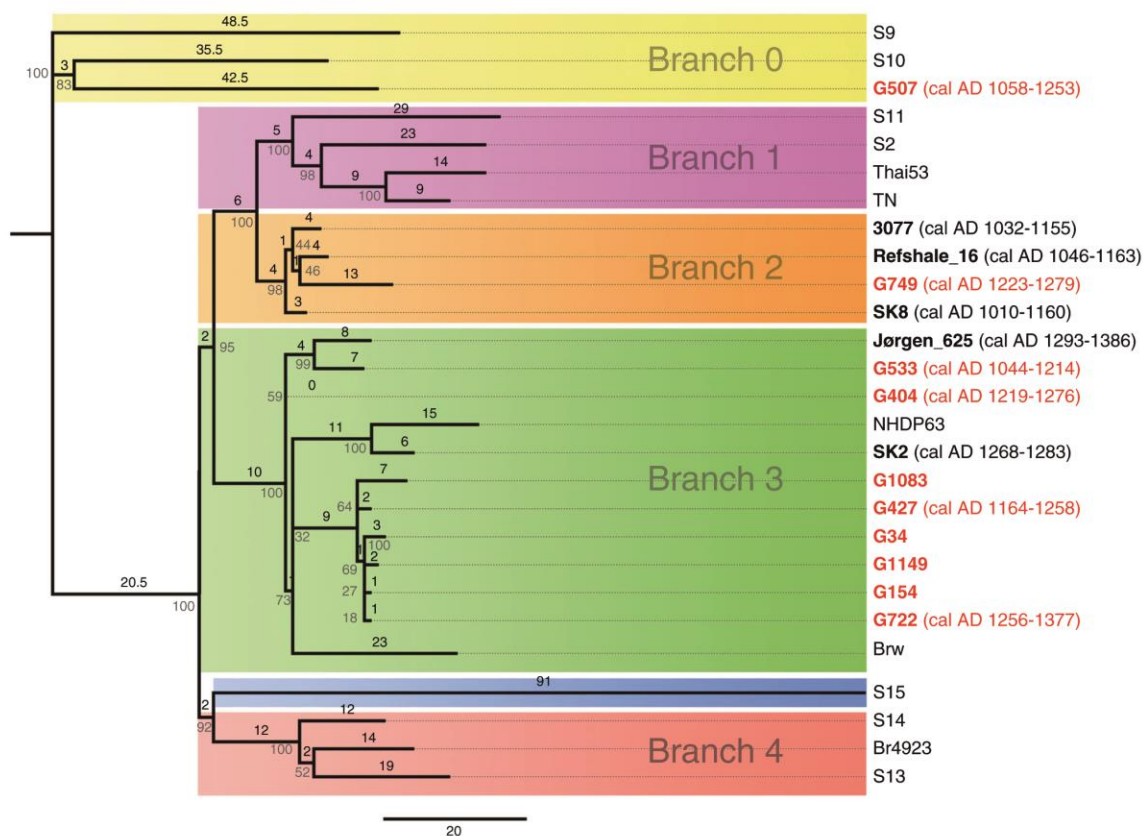


Fig. 3 Phylogeny of medieval and extant *M. leprae* strains. Phylogeny of medieval and modern *M. leprae* genomes using a maximum parsimony tree. Labeling colors: red bold (ancient genomes generated in this study), black bold (ancient genomes generated in a previous study (2)) and black (modern genomes generated in previous studies (2, 13)). Calibrated radiocarbon dates (2 sigma range) given in parentheses refer to branch tips. Bootstrap node support (in percent, from 500 replicates) is shown in gray numbers next to the branches, whereas the number of nucleotide substitutions on each branch is set in black. Color shading highlights different phylogenetic branches

on estimates for identical-by-descent (IBD) sharing we determined that the St. Jørgen individuals were not directly related to each other (Supplementary Data 4, Supplementary Table 16). The comparison of the rs3135388-T frequencies from the 69 LL cases with the DRB1*15:01 data from modern controls revealed a statistically significant association. To strengthen this finding, we further analyzed whether the allele frequency in the cases was different from that of medieval controls. The controls (i) were selected from sites that were geographically close and dated earlier or contemporaneous to St. Jørgen, (ii) showed no osteological evidence of the severe LL form of the disease and (iii) were *M. leprae* DNA-negative (Supplementary Data 1). The rs3135388-T frequency differed significantly between ancient cases and controls with an odds ratio similar in size to previous reports about DRB1*15:01^{4,5}. Notably, the medieval control frequency was slightly but significantly higher than today ($p = 0.024$, OR = 1.41, Table 1), possibly indicating weak selection against this allele during or since medieval times. This observation raises the question of how leprosy might have led to reduced reproductive fitness given that affected people, at least in modern populations, rarely die from the disease. Leprosy patients in the Middle Ages had to endure rejection and isolation in leprosaria and were not allowed to marry²³. In addition, leprosy is known today to result in a hormone-related decrease in fertility^{24,25} and a higher

vulnerability to other infections^{26–29}. By contrast, the fact that this allele is still the most common DRB1 allele in contemporary northern Germany suggests that it is (and was) associated with further and likely antagonistic fitness effects that prevented stronger frequency declines in medieval times.

Taken together, our results demonstrate a significant association between the HLA class II region and LL susceptibility in medieval leprosy patients from northern Europe, involving the DRB1*15:01 allele. As this allele is predicted to bind only a very small number of potential *M. leprae* antigens, this observation lends support to the hypothesis that limited HLA-presentation of relevant antigens may have impaired an *M. leprae*-specific immune response, leading to increased susceptibility to LL. One could speculate that, since DRB1*15:01 apparently was one of the most common DRB1 alleles in medieval northern Europe (or at least in Denmark, according to the control samples), *M. leprae* proteins might have evolved to evade presentation by this common allele, following a process of negative frequency-dependent selection. However, further molecular data and comparative work is required to address this hypothesis.

Interestingly, the DRB1*15:01-DQB1*06:02 haplotype represents an allele combination that is still common in contemporary Europeans^{30,31}. Furthermore, in modern populations, it is a strong risk factor for ulcerative colitis, sarcoidosis, and multiple

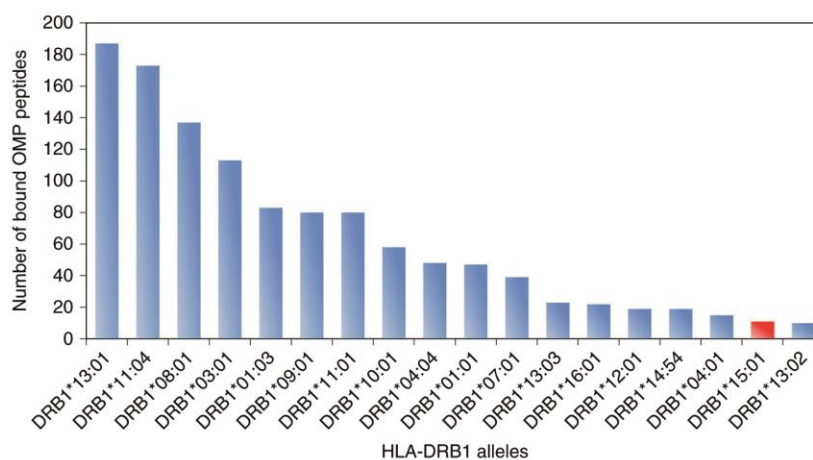


Fig. 4 Functional association between DRB1*15:01 and *M. leprae*. Computational prediction of HLA-presented antigenic *M. leprae* peptides for common HLA-DRB1 alleles revealed that DRB1*15:01 (red bar) presents one of the smallest *M. leprae* antigen repertoires. Binding predictions were run for the 18 HLA-DRB1 alleles with an allele frequency of >1% in representative contemporary samples from Schleswig-Holstein/Germany ($n = 129,336$) and a Danish minority population from northern Germany ($n = 918$). OMP—outer membrane proteins

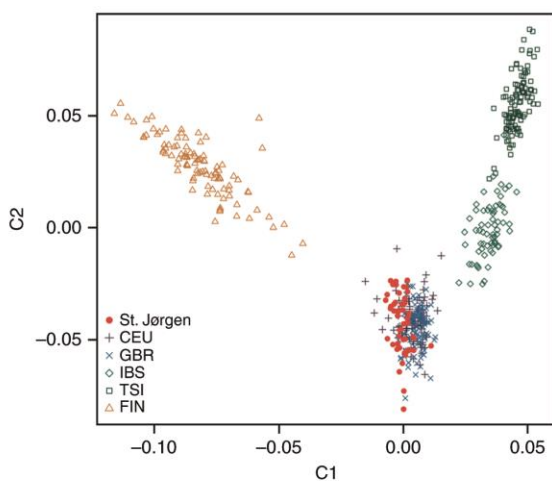


Fig. 5 Relationship of 53 medieval leprosy-positive Danes to contemporary Europeans. Principal component analysis plot for 53 medieval St. Jørgen individuals in relation to European population samples from the 1000 Genomes project. (CEU, Northern Europeans from Utah; GBR, British in England and Scotland; IBS, Iberian population in Spain; TSI, Tuscans in Italy; FIN, Finnish in Finland)

sclerosis^{32–34}, whereas being protective against type-1 diabetes³⁵. These different disease associations highlight the well-known pleiotropy of HLA variants that affect the population frequency of specific haplotypes and contribute to the genetic diversity in the HLA region in general. More generally, our findings provide a new, temporal layer of evidence for the hypothesis that ancient epidemics such as leprosy have influenced the present-day frequency of genetic factors associated with modern inflammatory diseases^{36,37}.

Methods

Selection of archeological specimens. In the current study, human skeletal remains were obtained from five medieval cemeteries in Denmark and northern

Germany (Fig. 1, Supplementary Note 1). The specimens were analyzed with permission from the respective museums and collections (Horsens Museum, Museum Lolland-Falster, Odense Bys Museer, Sydvestjyske Museer, Arch-äologisches Landesmuseum der Stiftung Schleswig Holsteinische Landesmuseen Schloss Gottorf). Osteological information (sex, age at death, leprosy status) of all individuals was collected. Individuals who suffered from LL and were buried at the leprosarium cemetery of St. Jørgen in Odense are considered cases in our study. Given the high infection prevalence, people interred at ordinary cemeteries might also have been infected. However, if this was the case, they were affected with a milder form of the disease, as evidenced from osteological analysis. These individuals were here regarded as controls. They were selected from the four sites Ribe, Revshale, Tirup, and Schleswig/Rathausmarkt that were geographically close and dated earlier or contemporaneous to Odense/St. Jørgen.

Sample processing. Material from each selected skeleton was collected in person by a member of the Kiel aDNA group or a member of the Odense ADBOU group to minimize the risks of contamination. Two different types of samples were taken: teeth and/or petrous bones (Supplementary Data 1). The DNA extractions and pre-PCR steps were carried out in clean room facilities dedicated to aDNA research. Following the guidelines on contamination control in aDNA studies^{38–40}, all surfaces and re-usable utensils were extensively cleaned with bleach before and after work. Besides, UV lights were used to improve the decontamination process. In addition, negative controls were included in each step involving non-indexed DNA molecules. Finally, access of staff and objects to the laboratories was restricted to a one-direction route from the pre-PCR rooms to the post-PCR facility. The post-PCR room was located in an independent building. This room was used to run the screening and indexing PCRs (the reaction mixes were prepared beforehand in the dedicated pre-PCR rooms) and to process the indexed PCR products (purification, amplification). As for the pre-PCR laboratories, no modern leprosy studies had ever been conducted there prior to the start of this project.

Whole teeth were cleaned in pure bleach solution (sodium hypochlorite) and rinsed with purified water. After drying at 37 °C overnight, the samples were ground in a ball mill homogenizer for 45 s at maximum speed. Petrous bones were cleaned in pure bleach solution (sodium hypochlorite), rinsed with purified water and dried overnight at 37 °C. The inner ear (cochlea and vestibule) was cut out with a bone saw. Subsequently, the cleaned (with bleach analogous to the petrous bone) and dried inner ear piece was ground in a ball mill homogenizer for 45 s at maximum speed.

DNA extraction for PCR assays. After grinding, 50 mg of bone powder were incubated in 500 μ L ethylenediaminetetraacetic acid (EDTA; pH 8, 0.5 M) and 20 μ L proteinase K (0.25 mg/mL) under gentle rotation at 37 °C overnight. The suspension was centrifuged for 3 min at 6000 rpm and 200 μ L of the supernatant were used for DNA extraction with the Qiagen EZ1 Advanced Investigator Kit (program setting “Trace”). Purified DNA was eluted in 50 μ L TE buffer and stored at –20 °C until further use. One negative control was processed per five samples⁴¹.

DNA extraction for HTS. A total of 50 mg of bone powder were incubated in 960 μ L EDTA (pH 8, 0.5 M) and 40 μ L proteinase K (0.25 mg/mL) under gentle

rotation at 37 °C overnight. The suspension was further processed according to a published protocol by Dabney et al.⁴²

PCR-based experiments. For this study, the following three PCR-based setups were carried out: *M. leprae* DNA screening (RLEP and 18-kDa), nuclear DNA SNP rs3135388 and mtDNA (hypervariable regions I and II) analysis. PCRs were performed in a 25- μ L volume containing 1 \times Immolase buffer (Bioline), 0.04 U/ μ L Immolase DNA Polymerase (Bioline), 1.5 mM MgCl₂ (Bioline), 4% DMSO, 200 μ M dNTP mix (Bioline) and 0.4 μ M of each primer (see Supplementary Table 1 and Supplementary Fig. 1 for primer and amplicon details). For the *M. leprae* and the SNP screening, 5 μ L of aDNA extract were used as template. For the mtDNA PCR, the amount of extract added was 1 μ L. The annealing temperatures were 52 °C (*M. leprae* DNA screening), 58 °C (SNP PCR) or 60 °C (mtDNA screening), respectively. PCR success was evaluated by gel electrophoresis. Subsequent Sanger sequencing was performed following standard procedures.

PCR-based *M. leprae* DNA screening. All 85 samples from the St. Jørgen cemetery and all 223 controls were subjected to PCR-based *M. leprae* DNA screening^{2,18}. Two primer pairs were used to amplify regions coding for RLEP (130 bp) and the 18-kDa protein (98 bp). A sample was considered positive for *M. leprae* when at least one PCR yielded a product of the expected length and sequence. If the amplicon did not match the expected sequence, a nucleotide BLAST was performed to identify possible contamination sources.

PCR-based genotyping of rs3135388. According to de Bakker et al. (2006)¹², the T allele at SNP locus rs3135388 is in almost complete LD with the HLA-DRB1*15:01 allele in the CEU population. With the designed SNP primers (Supplementary Table 1, Supplementary Fig. 1), the T/C variant rs3135388 was detected on the forward and reverse strand by Sanger sequencing. Three hundred and eight specimens (85 cases from St. Jørgen and 223 controls) were tested for this locus in the following manner: (1) For the St. Jørgen samples that provided HLA-DRB1 data ($n=43$), PCRs were performed independently at least two times per extract. (2) For the 26 St. Jørgen samples for which no HLA-DRB1 data were available but SNP amplicons could be generated, PCRs were performed between 7 and 11 times per specimen using a minimum of two independent extracts (exceptions: sample G417 only five replicates, sample G859 only six replicates). (3) For the 16 remaining St. Jørgen samples, neither HLA-DRB1 data nor SNP genotypes could be generated. (4) For the controls, PCRs were performed independently at least two times per extract. For 43% of control samples, PCRs were performed independently at least four times.

Only individuals with consistent results were included in the statistical analysis. However, allelic drop-out is a common problem when amplifying aDNA⁴¹, and a heterozygote might erroneously be determined as a homozygote^{43,44}. We therefore randomly selected 32 specimens (marked in Supplementary Data 1) that showed homozygous genotypes (either TT or CC) after the first round of PCR. For each of them, we performed between 7 and 11 PCRs as recommended^{43,44} using a minimum of two independent extracts. For each individual, the homozygous state was confirmed in all replicates.

PCR-based human mtDNA analysis. MtDNA studies were performed with all 308 samples by sequencing parts of the hypervariable regions I (150 bp and 183 bp) and II (253 bp)⁴⁵. For a random subset of 27 samples, new aDNA extracts were generated and used for a replication. In these cases, the haplogroup classification was confirmed. Haplogroups were assigned with Haplogrep 2.0 (<http://haplogrep.uibk.ac.at/>). Haplogroups showing a quality score below 100% were manually re-evaluated by consulting the established mtDNA phylogeny (www.phylotree.org).

Statistical association analysis. For the recruitment of medieval controls, the required sample size was calculated with the software G*Power, v3.1.9.2. For the available 138 alleles of our medieval LL samples and under the assumption of an odds ratio of 2.0⁴⁵, another 250 alleles in controls are required to obtain an a priori power of 75% at a significance level of 0.05. The comparisons of allele frequencies of DRB1*15:01/rs3135388 between different populations were performed with Fisher's exact test with the software R, v3.2.2⁴⁶.

Shotgun HTS. For each sample, two double-stranded DNA sequencing libraries were prepared according to an established, but slightly modified protocol for multiplex high-throughput sequencing⁴⁷.

First library: UDG-treated libraries were prepared in a 50- μ L volume containing 20 μ L of DNA extract, 1 \times NEB buffer 2 (New England Biolabs), 300 μ M dNTPs (each), 0.005 mg/mL BSA, 1mM ATP, 20 U T4 Polynucleotide Kinase (Thermo Fisher Scientific) and 3 U USER enzyme (Uracil-Specific Excision Reagent, New England Biolabs). The reaction mix was incubated at 37 °C for 3 h. Subsequently, 6 U T4 DNA Polymerase (Thermo Fisher Scientific) were added and the reaction mix was incubated at 25 °C for 30 min and at 10 °C for 5 min.

Second library: non-UDG-treated libraries were prepared in a 50- μ L volume containing 20 μ L of DNA extract, 1 \times NEB buffer 2 (New England Biolabs), 300 μ M dNTPs (each), 0.005 mg/mL BSA, 1mM ATP, 20 U T4 Polynucleotide Kinase

(Thermo Fisher Scientific) and 1.2 U T4 Polymerase (New England Biolabs). The mixture was incubated at room temperature for 30 min.

Both library preparations were continued as follows: After purification with the MinElute PCR Purification Kit (Qiagen) (elution volume 18 μ L), adapter ligation was done in a 40- μ L volume containing 18 μ L DNA, 1 \times Quick Ligase buffer (New England Biolabs), 2.5 μ M adapter mix (Solexa) and 0.5 U Quick Ligase (New England Biolabs). The mix was incubated at room temperature for 20 min. After another MinElute purification step (elution volume 20 μ L), adapter fill-in was performed in a 40- μ L volume containing 1 \times ThermoPol buffer (Thermo Fisher Scientific), 125 μ M dNTPs and 16 U BSM DNA Polymerase (Thermo Fisher Scientific). The reaction mix was incubated at 37 °C for 20 min and then at 80 °C for 20 min.

Sample-specific indices were added to both library adapters via amplification with two index primers (i7, i5). Extraction and library blanks were treated in the same manner.

For UDG-treated libraries, indexing PCRs were performed in a 50- μ L volume containing 10 μ L template DNA, 1 \times AccuPrime Pfx reaction mix (Thermo Fisher Scientific), 1.25 U AccuPrime Pfx DNA Polymerase (Thermo Fisher Scientific), 0.3 μ M P5 DNA primer and 0.3 μ M P7 DNA primer. PCR conditions were as follows: 95 °C for 2 min, 10 cycles of (95 °C for 15 seconds, 60 °C for 30 seconds, 68 °C for 70 seconds).

For non-UDG-treated libraries, indexing PCRs were performed in a 50 μ L volume containing 10 μ L template DNA, 1 \times PfuTurbo Cx reaction buffer, 200 μ M dNTPs, 2.5 U PfuTurbo Cx Hotstart DNA Polymerase (Agilent Technologies), 0.3 μ M i7 DNA primer and 0.3 μ M i5 DNA primer. PCR conditions were as follows: 95 °C for 2 min, 10 cycles of (95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 60 seconds), elongation step of 72 °C for 10 min.

Both indexed DNA libraries were purified with the MinElute PCR Purification Kit (Qiagen) according to the manufacturer's instructions and eluted in 50 μ L elution buffer.

A second amplification step was performed for all indexed libraries in 50- μ L reactions containing 5 μ L indexed library template, 1.25 U AccuPrime Pfx DNA Polymerase (Thermo Fisher Scientific), 1 \times AccuPrime Pfx reaction mix (Thermo Fisher Scientific) and 0.3 μ M IS5 (5'-AATGATACGGCAGCCACCGA-3') and IS6 (5'-CAAGCAGAAGACGGCATACGA-3') primers that bind to the adapters of the indexed libraries. Amplified products were purified with the MinElute PCR Purification Kit (Qiagen) (elution volume 53 μ L) and quantified using the Agilent 2100 Bioanalyzer DNA 1000 chip. The sequencing was carried out on the Illumina HiSeq 2500 (2 \times 125 bp) and HiSeq 4000 (2 \times 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 capture and HTS. In this study, HLA regions were enriched with an in-solution bait capture and the SureSelectXT Target Enrichment System (Illumina) for the Illumina paired-end multiplexed sequencing library. Using a UDG-treated sequencing library and a custom bait library designed by Michael Wittig et al.¹⁵, the classical class I (HLA-A, HLA-B, HLA-C) and class II HLA genes (HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA, and HLA-DPB1) were enriched in 68 St. Jørgen samples. Samples were pooled (up to four) per capture with 800 ng of library DNA per pool in a volume of 3.4 μ L. Hybridization buffer and blocking reagent were handled according to the manufacturer's instructions. As the captured samples were indexed already, 1 μ L of each IS5 and IS6 primers (100 μ M) were used for indexing/amplification instead of the primers provided in the capture kit. The post-capture PCR cycle number was set to 12. Purification of the amplified captured libraries was performed according to the protocol using AMPure XP beads. The quality of the captured library pools was assessed on the Agilent 2100 Bioanalyzer with the High Sensitivity DNA Assay. The sequencing was carried out on the Illumina HiSeq 4000 (2 \times 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.

Bioinformatic analysis. Multiple HTS data sets were generated for the 68 individuals from St. Jørgen. The data sets were pre-processed (adapter clipping, merging, trimming) according to published protocols specific for aDNA using the EAGER pipeline⁴⁸.

Genome-wide analysis of *M. leprae*. Multiple HTS data sets were generated for 68 of the 85 individuals from St. Jørgen. These data sets were pre-processed according to published protocols specific for aDNA⁴⁸. The reads were then aligned to the *M. leprae* TN reference genomes applying established algorithms. Post-processing included the identification of genomic variation, effect prediction of SNPs, the de novo assembly of the bacterial genomes and metagenomic screening as well as phylogenetic analyses and the identification of HLA allele combinations. A detailed description of each step is given below.

Pre-processing of ancient genomes: The data sets produced for all ancient samples contained paired-end reads with varying numbers of overlapping nucleotides as well as artificial adapter sequences. We used ClipAndMerge v1.7.3, a module of the EAGER pipeline⁴⁸, to clip adapter sequences, merge corresponding paired-end reads in overlapping regions and to trim the resulting reads. These steps

are explained in detail below. We used the default options with the following command:

```
java -jar ClipAndMerge.jar -in1 $FASTQ1 -in2 $FASTQ2 \
-f AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
-r AGATCGGAAGAGCGTCGTGAGGAAAGAGTGTGA \
-l 25 -qt -q 20 -o Soutput_file
where $FASTQ1 and $FASTQ2 are the two gzipped FASTQ input files.
```

ClipAndMerge uses an overlap alignment of the respective forward or reverse adapter with the 3' end of each read in order to remove sequencing adapter sequences. Regions at the 3' end of each read that were contained in the alignment were clipped. Reads that were shorter than 25 nucleotides after adapter clipping or contained only adapter sequences (adapter dimmers) were removed. All remaining reads were then used in the merging step. Merging was performed for all remaining paired reads with a minimum overlap of 10 nucleotides and at most 5% mismatches in the overlap region. The algorithm selected the maximal overlap fulfilling these criteria. The consensus sequence was generated using the nucleotides in the overlap regions from the read with the higher PHRED quality score, maximizing the quality of the resulting read. In a final step, ClipAndMerge performed quality trimming of the reads and all nucleotides with PHRED scores smaller than 20 were trimmed from the 3' end of each read. Finally, all reads with fewer than 25 nucleotides after quality trimming were removed. The resulting high-quality reads were used for the alignment against the *M. leprae* TN reference genome (NC_002677) with Bowtie2⁴⁹ v2.2.7. In this step, all reads were treated as single-end reads and mapping was performed using semi-global alignment mode and default parameters. The following command was used:

```
bowtie2 -t -mp 1,1 --ignore-quals --score-min L,0,-0.05 \
--no-unal -x $REF -U $FASTQ -S $SAM
```

where \$REF is the reference FASTA file, \$FASTQ is the gzipped input FASTQ file and \$SAM is the output SAM file.

After the alignment, the SAM files were converted into BAM files, which were sorted and indexed using SAMtools v1.3 with default parameters and the following commands:

```
samtools view -h -q 0 -bS $SAM -o $BAM
samtools sort -o $OUT -T $TMP $BAM
samtools index $OUT
```

where \$SAM is the bowtie2 output, \$BAM is the converted BAM file, \$TMP is a temporary file and \$OUT is the final sorted BAM file.

All individual-specific final BAM files were concatenated using the MergeSamFiles algorithm of picard tools v1.139 (<http://broadinstitute.github.io/picard/>) and indexed using SAMtools. We used the default parameters with the following commands:

```
java -jar picard.jar MergeSamFiles I=$BAMs O=$OUT
samtools index $OUT
```

where \$BAMs is a string containing all individual-specific BAM files to be concatenated, separated by whitespace, and \$OUT is the concatenated BAM file.

We used DeDup v0.9.9, part of the EAGER pipeline⁴⁸, to identify and remove all duplicate reads in the individual-specific BAM files with the default options and the following command:

```
java -jar DeDup.jar -i $IN -o $OUT
where $IN is the input BAM file and $OUT is the output BAM file.
```

To authenticate aDNA data sets, we evaluated the presence of postmortem DNA damage signatures from read alignments using mapDamage⁵⁰ v2.0.6 with default parameters and the following command:

```
mapDamage -v -i $BAM -r $REF -l 100 -d $PREFIX
```

where \$BAM is the input BAM file containing only merged reads, \$REF is the reference FASTA file used for the alignment and \$PREFIX is a string containing the full path and an optional prefix for the output.

After alignment and duplicate removal, genomic variation was identified with the Genome Analysis Toolkit (GATK)⁵¹ v3.6. First, a local realignment of the individual-specific BAM files was performed with the RealignerTargetCreator and the IndelRealigner modules of GATK⁵¹. Subsequently, the UnifiedGenotyper module was applied to call reference bases and variants from the alignment. Default parameters were used in the following commands:

```
java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
-R $REF -I $BAM -o $INTERVALS
java -jar GenomeAnalysisTK.jar -T IndelRealigner \
-R $REF -I $BAM -targetIntervals $INTERVALS -o $REALIGNED
java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper \
-R $REF -I $REALIGNED -o $VCF -mbq 15 -rf MappingQuality \
--mmq 20 --strand_call_conf 50 --sample_ploidy 2 -dcov 250 \
--output_mode EMIT_ALL_CONFIDENT_SITES
```

where the variable \$REF is the FASTA file of the used reference genome, \$BAM is the final alignment BAM file after duplicate removal, \$INTERVALS are the target intervals for the local realignment, \$REALIGNED is the output from the local realignment and \$VCF is the output variant call set in VCF format.

To annotate an estimated effect of identified genomic variation, the software SnpEff⁵² v4.2 was applied. We used the default parameters and the following command line:

```
java -jar snpEff.jar eff $GFF $VCF
```

where \$GFF is a previously set up database based on a GFF file and \$VCF is the output VCF file. We used the GFF file from NCBI for *M. leprae* (NC_002677) to build the database for SnpEff⁵².

To analyze the effects of single SNPs, the VCF files containing the lists of annotated SNPs for each sample were merged using the following command available from VCFtools⁵³:

```
vcf-merge sample1.vcf.gz [...] sampleX.vcf.gz | \
bgzip -c > merged.all.emit.confident.sites.vcf.gz
```

Variant annotations of upstream and downstream SNPs located further than 100 bases away from a gene were ignored. The annotations were manually completed with information about transcript type and function available on the NCBI and Mycobrowser databases. The results were compiled into a table containing information for each SNP regarding its effect on the genes in the strains in which the SNP occurs (Supplementary Data 2). In vivo phenotypic studies confirming the in silico annotations are rare because it is currently impossible to grow *M. leprae* in culture. Therefore, most of the predicted gene functions are based on comparisons with *M. tuberculosis* and numerous genes remain annotated as hypothetical proteins whose functions are unknown. Variants for which there is enough information available to perform in silico evaluation of the variant effects were selected by removing all variants with the following annotated effects: (1) Intergenic, upstream, downstream: The understanding of *M. leprae* intergenic regions is insufficient at present to estimate the effects of the intergenic variants on gene expression genes. (2) Hypothetical protein and pseudogene: No sufficient information was available on the transcripts. (3) Stable RNAs: The variants in annotated stable RNAs were removed so that the focus was on protein variants. (4) Variants present in only one genome: This filtering step was performed to focus on common variants.

Extracted mapping reads of ten high-coverage data sets (G34, G154, G404, G427, G507, G533, G722, G749, G1083, and G1149), covering the complete *M. leprae* reference genome (NC_002677.1) at least 10-fold, were de novo assembled using the SPAdes genome assembler⁵⁴ v3.5.0 with the following settings:

```
spades.py -t 8 -m 60 -k 121 --careful -s $IN -o $OUT
where $IN is a FASTQ file containing the mapped reads and $OUT is the output folder for SPAdes54.
```

Other samples were not assembled as their respective data sets did not match the previously described threshold of a 10-fold coverage of the complete *M. leprae* reference genome. All possible values for the *k* parameter were tried with *k* = 121 yielding the best result with respect to the contig mean, N50, maximal contig size and number of produced contigs. The multiple genome alignment software Mauve v2.4.0^{55,56} was used to reorder the resulting contigs relative to the *M. leprae* TN reference genome for subsequent calculation of the genomic coverage. The reordering in Mauve^{55,56} was executed with default parameters for the Mauve Contig Mover (MCM).

Shotgun sequencing data of multiple sequencing runs was pooled after pre-processing for the samples G34, G154, G404, G427, G507, G533, G722, G749, G1083, and G1149. The pooled reads were used to carry out a metagenomic de novo assembly with the Megahit assembler v1.0.3-8-g4b5271e⁵⁷. The following command line was used:

```
megahit --presets meta -r IN -o OUT, where IN is a file containing the pre-processed reads, OUT is the output directory.
```

For subsequent calculation of the genome coverage, contigs from the assemblies with a length of > 500 bases were blasted (v2.2.30) against the *M. leprae* strain TN genome. The following command line was used:

```
blastn -evaluate 10e-6 -perc_identity 95 -outfmt "6 qseqid" \
-query $IN -out $OUT -db $REF
```

where \$IN is a file containing the contigs, \$OUT is the output file of blast and \$REF is the blast database. Subsequently, the MCM was used as described above to calculate the genomic coverage for the contigs that mapped to the *M. leprae* TN genome.

Metagenome screening. All UDG-treated samples derived from teeth (*n*=68) were screened for their metagenomic content with the alignment tool MALT¹⁴ and the metagenome analyzer MEGAN [58]. After pre-processing, as described above, MALT¹⁴ v0_3_6 was used to align all samples against a collection of all complete bacterial genomes in FASTA format downloaded from the NCBI FTP server (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/>). We used MALT¹⁴ in BlastN mode with the following command line:

```
malt-run -inFile $IN -index $REF -output $OUT -id 85.0 -v \
-m BlastN -at SemiGlobal -top 1 -supp 0 \
-mq 100 -ssc -sps
```

where \$IN is a FASTQ file after pre-processing, \$REF is the MALT index, \$OUT is output folder for MALT¹⁴.

Subsequently, MEGAN⁵⁸ v6_4_15 was used to compute the taxonomical content, while pathogen screening was performed manually. Potentially interesting bacteria were selected for a specific alignment against the respective reference genome with Bowtie2⁴⁹ v2.2.7. The following parameters were used:

```
bowtie2 -t -mp 1,1 --ignore-quals --score-min L,0,-0.05 \
--no-unal --rg-id --rg SM:blank --rg PL:illumine
```

Genomic variation was identified in the genome-wide analysis of *M. leprae*. Variant positions (SNPs) as well as reference alleles were called in each data set if the quality was at least 40. The resulting VCF files were filtered to contain only

35. Hu, X. Y. et al. Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).
36. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15**, 379–393 (2014).
37. Khor, C. C. & Hibberd, M. L. Host-pathogen interactions revealed by human genome-wide surveys. *Trends Genet.* **28**, 233–243 (2012).
38. Yang, D. Y. & Watt, K. Contamination controls when preparing archaeological remains for ancient DNA analysis. *J. Archaeol. Sci.* **32**, 331–336 (2005).
39. Pilli, E. et al. Monitoring DNA contamination in handled vs. directly excavated ancient human skeletal remains. *PLoS ONE* **8**, 1–6 (2013).
40. 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 (2012).
41. Taberlet, P. et al. Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res.* **24**, 3189–3194 (1996).
42. Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. USA* **110**, 15758–15763 (2013).
43. Krause, J. et al. The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr. Biol.* **17**, 1908–1912 (2007).
44. Morin, P. A., Chambers, K. E., Boesch, C. & Vigilant, L. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). *Mol. Ecol.* **10**, 1835–1844 (2001).
45. Lee, E. J. et al. Emerging genetic patterns of the European Neolithic: perspectives from a late Neolithic Bell Beaker burial site in Germany. *Am. J. Phys. Anthropol.* **148**, 571–579 (2012).
46. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/> (2015).
47. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, <https://doi.org/10.1101/pdb.prot5448> (2010).
48. Peltzer, A. et al. EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
51. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
52. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2011).
53. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
54. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
55. Darling, A. E., Mau, B. & Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
56. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
57. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
58. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
59. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger data sets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
60. Tamura, K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol. Biol. Evol.* **9**, 678–687 (1992).
61. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
62. Lipatov, M., Sanjeev, K., Patro, R. & Veeramah, K. R. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. Preprint at *bioRxiv* <https://doi.org/10.1101/023374> (2015).
63. Robinson, J. et al. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).
64. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
65. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
66. Wang, P. et al. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics.* **11**, 568 (2010).
67. Rana, A., Thakur, S., Bhardwaj, N., Kumar, D. & Akhter, Y. Excavating the surface-associated and secretory proteome of *Mycobacterium leprae* for identifying vaccines and diagnostic markers relevant immunodominant epitopes. *Pathog. Dis.* **74**, ftw110 (2016).

Acknowledgements

We are grateful to the following people and institutions for providing samples, support, and advice: Johannes Krause, Alexander Herbig, Verena Schüenemann, Horsens Museum, Museum Lolland-Falster, Odense Bys Museer, Sydvestjyske Museer, Archäologisches Landesmuseum der Stiftung Schleswig-Holsteinische Landesmuseen Schloss Gottorf. This work was supported by the Graduate School Human Development in Landscapes, Cluster of Excellence Inflammation at Interfaces, and the Medical Faculty of Kiel University. We acknowledge financial support by Land Schleswig-Holstein within the funding programme Open Access Publikationsfonds. T.L.L. was supported by the German Research Foundation (DFG, grant LE2593/3-1).

Author contributions

B.K.-K., J.L.B., T.L.L., and A.N. designed the experiment. B.K.-K., L.B., D.D.P., S.-C.K., M.B., L.M., J.S., P.T., E.B., M.W. performed the experiment. B.K.-K., M.Nu., L.B., M.B., D. D., J.L.B., M.No., A.C. F.P., T.L.L. analyzed the data. B.W., A.S., J.S. provided comparative data. A.F., S.S. provided research infrastructure. B.K.-K., J.L.B., D.D., M.No., A.C., T.L.L., and A.N. interpreted the results. B.K.-K., M.Nu., L.B., J.L.B., D.D., M.No., J.S., A.C., A.F., T.L.L., and A.N. wrote the manuscript. B.K.-K., T.L.L., and A.N. revised the paper.


Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-018-03857-x>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

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

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

distinct mapping locus. All instances of the same sequence fragment (representing PCR duplicates) were noted and then collapsed per gene. The name and number of all mapping loci and alleles, respectively, were also noted. The sequences were then grouped by gene specificity—for each gene, the resulting set of sequences was inversely ordered according to the number of genes they would map to and then sorted by the starting position within the corresponding allele (Supplementary Fig. 7). This set of sequences was saved to a FASTA file for each sample.

Sample-specific FASTA files were used for each HLA gene to generate consensus sequences of the allele combination present in the sample. The GUI-based sequence alignment editor BioEdit v7.2.5⁶⁵ was used to display and sort the sequence reads representing the two alleles and identify sequences with PCR/sequencing errors (Supplementary Fig. 7). Overlapping reads that shared the same combination of variation were collapsed into a consensus sequence. If a read showed variants that were not supported by other overlapping reads, most likely representing PCR or sequencing errors, the read was discarded. Most emphasis was given to reads that were specific to the evaluated locus (i.e., did not map to other HLA genes) and to reads that were represented by multiple exact PCR duplicates (making it less likely that they represent sequencing errors).

The consensus sequences were manually compared with a reference alignment of all known four-digit alleles of the corresponding HLA gene and matching alleles were identified. Here we made use of the allele frequency information implicitly included in the HLA nomenclature: As four-digit alleles were named with a consecutive numbering scheme in the order of their discovery, the second set of digits roughly indicates their frequency (common alleles were discovered earlier than rare alleles). We therefore focused first on the first 10 four-digit alleles of each two-digit allele group, and if we identified one or more alleles that perfectly matched the given consensus sequences, we did not look for additional matches in the rarer alleles. In case of multiple equally well matching known alleles, we first recorded all of them, but eventually reported only the two-digit allele name (we never found equally well matching alleles from different two-digit allele groups). The identified full-length allele sequences were then again compared to a sample's given read alignment to verify that the identified alleles were supported by all high confidence reads. As there was an a priori expectation to find certain DRB1 alleles among the leprosy cases, we recoded the names of all known alleles in the DRB1 reference alignment to allow for an observer-blinded allele call. The true allele names were only revealed after allele calls for DRB1 had been completed.

Replication of manual allele call for DRB1. The manual read filtering and allele call for each sample was performed by one of three different researchers. In order to evaluate reproducibility of our allele call approach, we obtained consistent DRB1 genotype calls by two different researchers for 14 of the 68 samples. Of the 28 allele calls, 28 (100%) matched at the two-digit level (representing functional serotypes) and 14 (50%) matched exactly at the four-digit level (identical protein sequences). In 12 of the 14 called alleles that did not match exactly, the two replicate calls could not identify a unique four-digit allele sequence but overlapped in the range of possible four-digit alleles. Only in two cases did the allele calls lead to two different four-digit alleles, resulting in one and two nucleotide mismatches, respectively, over the length of the entire typed exon sequence (270 bp). Overall, this results in 99.96% reproducibility at the nucleotide level (3/7560 mismatches).

Prediction of *M. leprae* peptide binding by HLA-DRB1 variants. We used the IEDB-AR consensus algorithm⁶⁶ to computationally predict the repertoire of bound *M. leprae* peptides for all DRB1 variants (at 2nd field resolution) with a frequency of >1% in representative population samples from Schleswig-Holstein/Germany ($n = 129,336$) and a Danish minority population from northern Germany ($n = 918$), both included in the DKMS database. We first predicted DRB1 allele-specific binding (rank threshold for binding: ≤ 0.5) to all possible 15-mer peptides ($n = 5345$) of 19 *M. leprae* outer membrane proteins that were previously identified as likely antigenic, harboring known T- and B-cell epitopes⁶⁷ (Fig. 3). We subsequently also predicted DRB1 allele-specific binding to all possible 15-mer peptides using the entire *M. leprae* proteome ($n = 516,303$ peptides; Ensembl accession: ASM19585v1, Supplementary Fig. 9).

Data availability. Raw sequence read files have been deposited at the European Nucleotide Archive under accession no. ERP021830 (<https://www.ebi.ac.uk/ena/data/view/PRJEB19769>).

Received: 14 July 2017 Accepted: 16 March 2018
Published online: 01 May 2018

References

- Boldsen, J. Epidemiological approach to the paleopathological diagnosis of leprosy. *Am. J. Phys. Anthropol.* **115**, 380–387 (2001).
- Schuenemann, V. J. et al. Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* **341**, 179–183 (2013).
- WHO. Global leprosy update, 2016: accelerating reduction of disease burden. *Wkly. Epidemiol. Rec.* **35**, 501–520 (2017).
- Liu, H. et al. Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. *Nat. Genet.* **47**, 267–271 (2015).
- Zhang, F. R. et al. Genomewide association study of leprosy. *N. Engl. J. Med.* **361**, 2609–2618 (2009).
- Jarduli, L. R. et al. Role of HLA, KIR, MICA, and cytokines genes in leprosy. *Biomed. Res. Int.* **2013**, 989837 (2013).
- Zhang, F. R. et al. Evidence for an association of HLA-DRB1*15 and DRB1*09 with leprosy and the impact of DRB1*09 on disease onset in a Chinese Han population. *BMC Med. Genet.* **10**, 133 (2009).
- Zhang, F. R., Wang, D., Li, Y.-Y. & Yao, Y.-G. Integrative analyses of leprosy susceptibility genes indicate a common autoimmune profile. *J. Dermatol. Sci.* **82**, 18–27 (2016).
- Escamilla-Tilch, M. et al. Association of genetic polymorphism of HLA-DRB1 antigens with the susceptibility to lepromatous leprosy. *Biomed. Rep.* **1**, 945–949 (2013).
- Boldsen, J. L. Leprosy in Medieval Denmark—osteological and epidemiological analyses. *Anthropol. Anz.* **67**, 407–425 (2009).
- Boldsen, J. L. & Møllerup, L. Outside St. Jørgen: leprosy in the medieval Danish city of Odense. *Am. J. Phys. Anthropol.* **130**, 344–351 (2006).
- de Bakker, P. I. et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
- Goris, A. et al. A Taqman assay for high-throughput genotyping of the multiple sclerosis-associated HLA-DRB1*1501 allele. *Tissue Antigens* **72**, 401–403 (2008).
- Vågène, Å. J. et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2**, 520–528 (2018).
- Wittig, M. et al. Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.* **43**, e70 (2015).
- Marsh, S. G. E. et al. Nomenclature for factors of the HLA system. *Tissue Antigens* **75**, 291–455 (2010).
- Lenz, T. L. Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution* **65**, 2380–2390 (2011).
- Ridley, D. S. & Jopling, W. H. Classification of leprosy according to immunity. A five-group system. *Int. J. Lepr. Other Mycobact. Dis.* **34**, 255–273 (1966).
- Donoghue, H. D. et al. A migration-driven model for the historical spread of leprosy in medieval Eastern and Central Europe. *Infect. Genet. Evol.* **31**, 250–256 (2015).
- Monot, M. et al. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat. Genet.* **41**, 1282–1289 (2009).
- Avanzi, C. et al. Red squirrels in the British Isles are infected with leprosy bacilli. *Science* **354**, 744–747 (2016).
- Mendum, T. A. et al. *Mycobacterium leprae* genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics* **15**, 270 (2014).
- Grzybowski, A. et al. Leprosy: social implications from antiquity to the present. *Clin. Dermatol.* **34**, 8–10 (2016).
- Leal, A. M. & Foss, N. T. Endocrine dysfunction in leprosy. *Eur. J. Clin. Microbiol. Infect. Dis.* **28**, 1–7 (2009).
- Smith, D. G. & Guinto, R. S. Leprosy and fertility. *Hum. Biol.* **50**, 451–460 (1978).
- van Brakel, W. H. Measuring leprosy stigma - a preliminary review of the leprosy literature. *Int. J. Lepr. Other Mycobact. Dis.* **71**, 190–197 (2003).
- Rao, P. S. et al. Disability adjusted working life years (DAWLYs) of leprosy affected persons in India. *Indian J. Med. Res.* **137**, 907–910 (2013).
- Guinto, R. S., Doull, J. A. & De Guia, L. Mortality of persons with leprosy before sulphone therapy, Cordova and Talisay, Cebu, Philippines. *Int. J. Lepr.* **22**, 273–284 (1954).
- Saporta, L. & Yuksel, A. Androgenic status in patients with lepromatous leprosy. *Br. J. Urol.* **74**, 221–224 (1994).
- Klitz, W. et al. New HLA haplotype frequency reference standards: high-resolution and large sample typing of HLA DR-DQ haplotypes in a sample of European Americans. *Tissue Antigens* **62**, 296–307 (2003).
- Schipper, R. F., Schreuder, G. M. T., D'Amato, J. & Oudshoorn, M. HLA gene and haplotype frequencies in Dutch blood donors. *Tissue Antigens* **48**, 562–574 (1996).
- International Multiple Sclerosis Genetics Consortium et al. Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851–862 (2007).
- Gregersen, J. W. et al. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature* **443**, 574–577 (2006).
- Fischer, A. et al. Genetics of sarcoidosis. *Semin. Respir. Crit. Care Med.* **35**, 296–306 (2014).

positions that were covered by at least five reads. Genome drafts were generated using the respective variant or reference alleles of these loci. Variant alleles were only used if the fraction of mapped reads containing the variation was at least 90%, otherwise the reference allele was used instead. If a reference locus was not sufficiently covered, the character "N" was inserted at the respective position.

A multiple genome alignment of 33 *M. leprae* genomes was computed using the progressive Mauve algorithm⁵⁵ integrated in the whole genome alignment software Mauve⁵⁶ v2.4.0. The sequences of the ten medieval genomes G34, G154, G404, G427, G507, G533, G722, G749, G1083, and G1149, five previously published medieval genomes (3077, Jorgen_625, Refshale_16, SK2, and SK8) and seventeen previously published modern genomes^{2,21} (S2, S9, S10, S11, S13, S14, S15, Brw-01, Brw-05, Brw-10, Brw-20, Brw-25, TN, BR4923, Thai53, and NHDP63) were included in the alignment. Mauve's SNP export function was used to generate FASTA files for each data set that contained all sites that were variable among all genomes. These SNP assemblies were merged into a multi FASTA file and then used as input for MEGA⁷⁵⁹ to perform the phylogenetic analyses.

MEGA⁷⁵⁹ v7.0.18 was used to create maximum parsimony (MP), neighbor-joining (NJ) and maximum likelihood (ML) trees. All sequences included in the multiple alignment were used in all three tree reconstructions. In a separate analysis rooted MP, NJ, and ML trees were generated using *M. avium* 104 (NC_008595.1) as outgroup.

Close-Neighbor-Interchange algorithm (search level 1) was applied to construct the MP tree. Random addition of sequence (10 replicates) was used to obtain the initial trees. Maximum likelihood model selection analysis of the MEGA package (default parameters) was used to assess the best model for the evolutionary distances in the NJ and ML tree. The model with the lowest BIC value was chosen (Supplementary Data 3). The Tamura 3-parameter model⁶⁰ and uniform rate for all sites was the best model for the alignment without the outgroup. This model was also used for the alignment including the outgroup. The NJ and ML trees were then constructed using the Tamura 3-parameter model⁶⁰, uniform rates and default settings in MEGA⁷⁵⁹. Bootstrap values were inferred from 500 replicates in all three reconstruction methods.

The analysis involved 32 (33 with outgroup) nucleotide sequences. In the alignments, positions were not used that had less than 95% site coverage. The final data set comprised a total of 955 (547,078 with outgroup) informative positions.

Genome-wide analysis of *Homo sapiens* and population genetics. Each of the 68 individual-specific HTS data sets was pre-processed during the genome-wide analysis of *M. leprae*, as described above. After adapter clipping, merging, quality trimming, and quality control, the reads were additionally aligned to the human reference genome hg38. Post-processing included the identification of genomic variation, mtDNA haplotyping, sex assignment, principal component analysis and the estimation of IBD. A detailed description of these methods follows below.

After pre-processing, the resulting high-quality reads were mapped for each data set individually against the *H. sapiens* reference genome hg38 (GRCh38), downloaded from UCSC (<http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/>) using Bowtie2⁴⁹ v2.2.7. In this step, all reads were treated as single-end reads and mapping was performed using semi-global alignment mode and default parameters. The following command was used:

```
bowtie2 -t -mp 1,1 --ignore-quals -score-min L,0,-0.05 \
-no-unal -x $REF -U $FASTQ -S $SAM
```

where \$REF is the reference FASTA file, \$FASTQ is the gzipped input FASTQ file and \$SAM is the output SAM file.

After the alignment, the SAM files were converted into BAM files which were sorted and indexed using SAMtools v1.3 with default parameters and the following commands:

```
samtools view -h -q 0 -bS $SAM -o $BAM
samtools sort -o $OUT -T $TMP $BAM
samtools index $OUT
```

where \$SAM is the bowtie2⁴⁹ output, \$BAM is the converted BAM file, \$TMP is a temporary file, and \$OUT is the final sorted BAM file.

All individual-specific final BAM files were concatenated using the MergeSamFiles algorithm of picard tools v1.139 (<http://broadinstitute.github.io/picard/>) and indexed with SAMtools. We used the default parameters with the following commands:

```
java -jar picard.jar MergeSamFiles I=$BAMs O=$OUT
samtools index $OUT
```

where \$BAMs is a string containing all individual-specific BAM files to be concatenated, separated by whitespace, and \$OUT is the concatenated BAM file.

We used DeDup v0.9.9, part of the EAGER pipeline⁴⁸, to identify and remove all duplicate reads in the individual-specific BAM files with the default options and the following command:

```
java -jar DeDup.jar -i $IN -o $OUT
where $IN is the input BAM file and $OUT is the output BAM file.
```

To authenticate aDNA data sets, we evaluated the presence of postmortem DNA damage signatures from read alignments using mapDamage⁵⁰ v2.0.6 with default parameters and the following command:

```
mapDamage -v -i $BAM -r $REF -l 100 -d $PREFIX
```

where \$BAM is the input BAM file containing only merged reads, \$REF is the reference FASTA file used for the alignment, and \$PREFIX is a string containing the full path and an optional prefix for the output.

The alignment of high-quality reads against the *H. sapiens* reference genome hg38 was performed analogous to the *M. leprae* alignment using bowtie2⁴⁹ and the same settings as described above. The removal of duplicates and the evaluation of damage patterns with DeDup and mapDamage, respectively, were also performed according to the protocols described above. Genomic variation was identified using the haplotype caller module of GATK⁵¹ v3.6 based on the alignment data sets.

First, a local realignment of the individual-specific BAM files was performed with the RealignerTargetCreator and the IndelRealigner modules of GATK⁵¹. Subsequently, the UnifiedGenotyper module was applied to call reference bases and variants from the alignment. Default parameters were used in the following commands:

```
java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
-R $REF -I $BAM -o $INTERVALS
java -jar GenomeAnalysisTK.jar -T IndelRealigner \
-R $REF -I $BAM -targetIntervals $INTERVALS -o $REALIGNED
java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper \
-R $REF -I $REALIGNED -o $VCF -mbq 15 -rf MappingQuality \
-mmq 20 -strand_call_conf 50 -sample_ploidy 2 -dcov 250 \
-output_mode EMIT_ALL_CONFIDENT_SITES
```

where the variable \$REF is the FASTA file of the used reference genome, \$BAM is the final alignment BAM file after duplicate removal, \$INTERVALS are the target intervals for the local realignment, \$REALIGNED is the output from the local realignment, and \$VCF is the output variant call set in VCF format.

For determination of sex, we computed the "read densities" dX, dY for the X and Y chromosomes, respectively, as the ratio of the number of reads mapped to the respective chromosome by its total chromosomal length (156,040,895 and 57,227,415 bp, respectively).

We applied principal component analysis (PCA) to pseudohomozygous calls obtained by randomly sampling alleles at genomic positions that are biallelic in the European individuals in the 1000 Genomes data set (release 20130502)⁶¹. Genotypes with alleles other than those observed in the European samples from the 1000 Genomes project were excluded as likely errors. We used the smartpca software (v13050) with default parameters for outlier detection (numoutliervec 10, outliersigmathresh 6) to conduct the PCA. Overall, 463 of 503 European samples from the 1000 Genomes project and 53 of 68 individuals from St. Jørgen remained for analysis. The lsproject parameter was used for projection of samples with differentially missing genotypes onto a single plot.

To assess the degree of relatedness between individuals, we used the software lcmkin⁶² v0.5.0 for estimating expected IBD sharing, π , between pairs of individuals and the accompanying script SNPbam2vcf for computing genotype likelihoods. The set of SNPs included in these calculations was restricted to positions that were reported as being biallelic in the European 1000 Genomes populations. In addition, only those SNPs that were covered by at least four reads (after removal of duplicates) in at least two individuals were considered for the computations. Finally, default quality filters were applied using SNPbam2vcf (MQ ≥ 20 , BQ ≥ 5 , GQ ≥ 0.1).

Identification of HLA allele combinations. Even after enrichment and deep sequencing, the highly degraded and fragmented nature of aDNA led to a low and incomplete coverage of the HLA region. This prevents the use of most computational HLA genotyping algorithms that are designed for high-coverage and high-quality sequence data. Furthermore, the allele combinations present in the samples might include unknown alleles that are undetectable by approaches that compare the reads only to modern reference sequences. We therefore used a semi-manual approach, where we performed an automated read selection and sorting procedure, followed by manual filtering and allele identification. These steps are described in detail below.

To select all reads from DNA fragments belonging to the HLA class II region in our data sets, we first generated a comprehensive reference file, containing all known four-digit (2nd field resolution) alleles of the exon coding for the peptide-binding groove of the classical HLA class II genes (HLA-DPA1, -DPB1, -DQA1, -DQB1, -DRA, -DRB1). In order to prevent mis-mapping and the resulting mis-identification of reads due to paralogous sequence similarity, we also included all known corresponding alleles for the non-classical genes HLA-DRB3, -DRB4, -DRB5, -DRB6, -DRB7, and -DRB9. Sequences were downloaded from the IMGT/HLA database⁶³ (accessed 28 July 15). In addition, we included a corresponding exon sequence of HLA-DQB2 from the human reference genome (not represented in the IMGT/HLA database), again to prevent their reads to be mis-identified as HLA-DQB1 variants. We did not consider intron sequences, as these evolve rather neutrally and may contain ancient variation that could be misleading during the identification of functional HLA alleles. All sequence variants for a given locus were initially aligned using MUSCLE⁶⁴ and then evaluated manually in BioEdit⁶⁵.

The actual read alignment was performed using Bowtie2⁴⁹ v2.2.7 in local alignment and "a reporting mode", allowing each read to map against multiple alleles. We used the following command:

```
bowtie2 -a -t --ma 1 --mp 1,1 --local --ignore-quals \
--score-min L,0,1 --no-unal -x $REF -U $FASTQ -S $SAM
where $REF is the reference FASTA file, $FASTQ is the gzipped input FASTQ file and $SAM is the output SAM file.
```

The resulting alignment contained all mapped reads and, in case of an unspecific mapping, multiple instances of the same read sequence, one for each

Chapter IV

Ancient history of HLA genes in the Americas

Federica Pierini¹, Austin W. Reynolds², Christina M. Balentine², Jaime Mata-Míguez², Andre Franke³, Almut Nebel³, Ben Krause-Kyora³, Deborah A. Bolnick² and Tobias L. Lenz¹

¹Max Planck Institute for Evolutionary Biology, Plön, Germany, ²University of Texas at Austin, Austin, United States, ³Institute of Clinical Molecular Biology, Kiel University, Kiel

Abstract

A strong population bottleneck following the first European contact in Native American populations has been largely documented. The spread of epidemic diseases caused by European-borne pathogens is considered one of the main drivers of the Native American population decline. In humans, a number of studies suggest that specific alleles of the human leukocyte antigen (HLA) system are associated with susceptibility or resistance to infectious diseases. Consistently, it has been proposed that Native Americans' HLA genes may have lacked genetic polymorphism and/or specific resistance HLA alleles to a variety of new pathogens introduced by European colonizers, resulting in an increased susceptibility to new diseases. While the majority of the studies concerning variability at HLA genes in Native American populations have been conducted on contemporary humans, the genetic variation at such genes in pre-European contact Native American populations remains unexplored. We here investigate the polymorphisms of HLA class II loci (HLA-DRB1 and HLA-DQB1) in ancient Native American populations. We first described the HLA molecular profile of pre-European contact archaeological samples as well as present-day residents of the town of Xaltocan in central Mexico. By studying the same population through time, we aimed to explore potential HLA allele frequency shifts from pre- to post-European contact period. To reconstruct HLA polymorphism in ancient Native American populations at a large geographical and temporal scale, we further characterized HLA variability in available ancient whole-genome data of samples collected from different sites across the American continent. We present the first spatiotemporal characterization of genetic variability of class II HLA loci in ancient Native American populations, which revealed allelic lineages currently present in Amerindian and/or Asian populations but also unknown alleles present in ancient Native American populations but no longer found in modern-day humans.

Introduction

Archeological, historical and genetic studies indicate a widespread collapse of the Native Americans population over the century following the European conquest of the Americas (McNeill 1976; Livi-Bacci 2006; O'Fallon and Fehren-Schmitz 2011). The magnitude, the dynamics and the potential causes of such population decline remain under debate. While some studies suggest significant reduction of populations localized to specific geographical area (Larsen 1994), others argue in favor of a more intense and widespread demographic process, with loss rate estimates of the total Native American population size ranging from 40% to 95% (O'Fallon and Fehren-Schmitz 2011; Koch et al. 2019). Factors like warfare, enslavement, disruption of social structure and spread of epidemic diseases have been linked to Native American population decline. Undoubtedly, population collapses following disease outbreaks caused by European-borne pathogens have been extensively documented across the American continent (Koch et al. 2019), and epidemics are often considered the main driver behind the majority of the deaths (Black 1994). Multiple pathogens caused multiple waves of epidemics, strictly depending on the differential transmission mechanisms of causal agents. Pathogens that can be transmitted directly from one person to another led to sudden and broad diffusion of diseases such as smallpox and measles; while a slower distribution of vector-borne diseases like malaria and yellow fever, in which the pathogens had to adapt to vectors in the new continent, has been also documented (Berlinguer 1992). Accordingly, the simultaneous action of several non-endemic pathogens causing epidemics in rapid succession, led to a quick rise in the mortality rate.

Modern Native Americans exhibit less genomic diversity than do other worldwide populations (Bolnick et al. 2016). Such low genetic diversity has been attributed to the founder effect occurred $\sim 25,000 \pm 1100$ years ago, when a small ancestral Native American population diverged from Siberians and East Asians, later radiating rapidly across the American continent (Moreno-Mayar et al. 2018b) and evolving in relatively isolated groups. The remarkable genetic homogeneity among the Native Americans has been linked to the lowered resistance against the new infectious diseases. Indeed, in population with limited diversity, pathogens can be fully adapted to their successive hosts, which facilitate the spread of more virulent pathogens (Black 1992). Such hypothesis has been further supported by the reduced diversity at several immune-related genes, including HLA genes, observed in contemporary Native American populations (Cadavid and Watkins 1997; Lindenau et al. 2013).

A number of studies suggest that specific alleles of the human leukocyte antigen (HLA) system are associated with susceptibility or resistance to infectious diseases. In this light, it has been proposed that Native American HLA genes may have lacked both general genetic polymorphisms and/or specific resistance alleles, due to the lack of historical contact with the introduced pathogens, resulting in an increased susceptibility to the new diseases (Dean et al. 2002; Penman and Gupta 2018). Studies on HLA genetic variability in modern Native American populations have shown a remarkably small number of alleles (Belich et al. 1992; Watkins et al. 1992). However, the persistence of founding alleles probably brought by the first migrants as well as new alleles arisen from recombination events have been documented in South Amerindian populations. The new recombinant alleles replaced the older ones, under a process known as allele turn over, which resulted in an enhanced differentiation between populations, without increasing allelic diversity within populations (Parham et al. 1997). It has been also shown that such small numbers of alleles still provide good coverage of the types of peptide binding motifs, and do not correspond to functional deficiency (Parham et al. 1997). The differential pattern of HLA polymorphism between North and South Native American populations have been linked to different pathogen selective pressures acting in the two different geographical area, underling how also small isolated populations might be source of genetic diversity under pathogen-driven natural selection (Parham and Ohta 1996). Nevertheless, all such studies are focused on present-day Native American populations, while the allelic diversity at HLA genes in Native American populations before the first European contact remains largely unexplored. By comparing genome-wide diversity of an indigenous North American population before and after the arrival of European, Lindo et al. (2016) found a significant shift in the frequency of a single nucleotide polymorphism (SNP) falling in the 5' untranslated region of the gene HLA-DQA1. These results suggest that the changing in immune-related selection pressures associated with the European contact had an effect on the genetic makeup of Native Americans and highlight the need of further studying the genetic variability at immune-related genes in ancient Native American populations.

We here characterize the human leukocyte antigen (HLA) polymorphisms in ancient Native American populations with a particular focus on pre-European contact samples. First, HLA target-enrichment approach in combination with the aHLA-Seq pipeline, described in chapter 2, have been used to explore HLA polymorphism in pre-European contact samples as well as present-day residents of the town of Xaltocan in central Mexico (**Figure 1**). We thus performed a temporal comparison on the same population through time, to explore potential HLA allele

frequency shifts from pre- to post-European contact period. Further, using the same approach we explored the HLA variability in available ancient whole-genome sequence data of samples collected from different sites across the American continent (ranging in age from 11,500 to 500 BP) (**Figure 3**), to eventually describe HLA polymorphism in ancient Native American populations at a broader geographical and temporal scale. We present the first spatiotemporal characterization of genetic variability of class II HLA loci (HLA-DRB1 and HLA-DQB1) in pre-European contact Native American populations which revealed allelic lineages currently present in Amerindian and/or Asian populations but also unknown alleles present in ancient Native Americans that are no longer found in modern-day humans.

Material and Methods

Archaeological site and samples information

The archaeological samples analyzed in this study come from individual remains recovered from the following sites: Xaltocan (Mexico), Fallen Tree site and Santa Catalina de Gual cemetery (St. Catherines Island, Georgia) (**Table S1**).

Xaltocan (Mexico)

The town of Xaltocan is located 35 km north of Mexico City. When founded by a group of Otomi-speaking people around 900 C.E., it was a small island in the middle of the now drained Lake Xaltocan (**Figure 1**) (Rodríguez-Alegría 2010). The town has a complex history characterized by a series of sociopolitical transitions that occurred over the past thousand years. The 18 skeletal remains (teeth or bone) were collected from different houses across the archaeological site. Archaeological studies at Xaltocan were carried out with support and appropriate permissions from the Mexican National Institute for Anthropology and History, the town's delegados (delegates), and local property owners. Eleven individuals were directly radiocarbon dated, while the approximate ages for the remaining individuals were defined based on stratigraphic analyses and the style of pottery. Further, all of the individuals that were associated with deposits containing Aztec II Black-on-Orange, radiocarbon dated between 1240 and 1350, were assigned in the pre-1395 Xaltocan group. Pre-1395 Xaltocan individuals are divided into two spatial subgroups based on excavation location: pre-1395 interior (n=3) and periphery (n=3) Xaltocan. Six individual were assigned to the post-1395 Tepanec period, two individuals to the Aztec period (1428-1521) while the remaining four samples are post-1395 individuals that could not be ascribed to either Tepanec or Aztec Xaltocan (**Table S1**). The demographic changes

resulting from past political transitions have been extensively studied in the dataset of ancient Xatocan samples using mitochondrial, Y-chromosome, and autosomal DNA markers and are reported elsewhere (Mata-Miguez et al. 2012; Mata-Míguez 2016).

St. Catherines Island (Georgia)

Located off the coast of Georgia (US), St. Catherines Island, was inhabited by Indigenous peoples for thousands of years. Eight of the archaeological samples analyzed in this study come from two sites on St. Catherines Island (SCI) (**Table S1**):

- Guale-Fallen Tree (n = 5): a cemetery discovered in 2013 where approximately 70 skeletal remains have been recovered. Radiocarbon dating suggests that the cemetery was used as late as the early Spanish mission period, while archaeological remains indicate that the cemetery was already used before the Spanish missionization on the island.
- Mission Santa Catalina de Guale (n = 3): discovered in 1970, it was an important Spanish mission during the 17th century. Over 400 skeletal remains have been excavated, representing one of the most extensive series of human remains from early contact site in North America.

Samples from St. Catherines Island belong to the pre-contact and early post-contact period (**Table S1**). However, as in this study, we were mainly interested in detecting changes in allele frequency compared with modern day populations, and because at the moment we could not establish when specific epidemics occurred in the island, the St. Catherines samples were considered in our analysis together with all the other pre-contact samples.

Permissions for the analysis of samples collected from St. Catherines Island were obtained from the Georgia Council on American Indian Concerns, the Franciscan Order, and the Bishop of Savannah.

Laboratory procedures for archaeological samples

Sample processing have been performed following strict procedures commonly adopted for aDNA (Gilbert et al. 2005; Willerslev and Cooper 2005). Extractions of ancient DNA from the skeletal samples and library preparation have been done in dedicated clean lab facilities in Austin (University of Texas at Austin) following published protocols (Rohland and Hofreiter 2007; Bolnick et al. 2012). Extraction negative controls as well as library negative controls were

processed to test for reagent contamination. Library preparation was performed using partial uracil – DNA – glycosylase treatment in which the enzyme uracil–DNA–glycosylase (UDG), cleaves deaminated cytosines (i.e. the common damage pattern observed in aDNA sequences) reducing the rate of ancient DNA errors (Rohland et al. 2014). Being a partial treatment, the damage signals are preserved at the terminal bases of the molecules so that the authenticity of the DNA in the library can be assessed. Each ancient DNA library was prepared including a unique barcode combination of 6 bp length. Contamination rates in the ancient libraries analyzed in this study were assessed in a parallel study (Reynolds 2018). A first evaluation was done based on mitochondrial DNA by estimating the fraction of reads that don't map to the determined endogenous mitochondrial genome. For male individuals the contamination rates were evaluated determining the rate of heterozygosity observed on the X chromosome. Post-mortem damage signatures, assessed using MAPDAMAGE v. 2.0 (Jonsson et al. 2013), were also considered and consistent with an ancient origin of the DNA. Few samples showed potential signal of contamination but did not appear as outliers in population genetic analyses, suggesting only a very small amount of contaminant reads when present. As none of the libraries failed the contamination assessments they were all included in our study. All ancient DNA libraries were then PCR amplified at the ancient DNA laboratory in Kiel. PCRs were performed in 50- μ L reactions containing 5 μ L barcoded library template DNA, 1 \times AccuPrime Pfx reaction mix (Thermo FisherScientific), 1.25 U AccuPrime Pfx DNA Polymerase (Thermo Fisher Scientific), 0.3 μ M of each of the two primers PreHyb-F, PreHyb-R. PCR conditions were as follows: 95 °C for 2 min, 20 cycles of (95 °C for 15 seconds, 55 °C for 30 seconds, 68 °C for 15 seconds), while performing a final extension at C 68 °C for 5 min. Amplified products were purified with the MinElute PCR Purification Kit (Qiagen) (elution volume 28 μ L) and quantified using the Agilent 2100 Bioanalyzer DNA 1000 chip.

Modern Xaltocan samples

29 present-day residents of Xaltocan town were included in this study. Saliva samples were collected together with written individual informed consent, after obtaining permits from the Ayuntamiento Constitucional (Town Council), the Consejo de Participación Ciudadana (literally, Board of Citizen Participation), and the town's delegados (delegates). The collection and analyses of samples from this community was approved by the IRB of the University of Texas at Austin (protocol #2012–05-0105). Samples from individuals whose maternal grandmother and paternal grandfather were born in Xaltocan were preferentially collected to minimize the effect of gene flow into Xaltocan over the last generations. Samples from individuals who are not closely

related (i.e., parents and their children, or siblings) were collected to avoid confound effect during population allele frequency estimates. Furthermore, a larger dataset of Xaltocan individuals have been previously studied at genetic level and relatedness between individuals as well as genetic ancestries estimated (Mata-Míguez 2016; Reynolds 2018; Reynolds et al. 2019). Based on previous results we could select for our study only unrelated individuals that did not show evidence of European and African ancestry. In the same study genome-wide data from ancient and modern residents of Xaltocan were explored. Results suggested that ancient and modern Xaltocan, and other central Mexican populations, are genetically similar overall, and can therefore be considered part of the same regional population. Furthermore, the community has social connections through time, which is consistent with being the same social population. These findings confirmed that modern Xaltocan are the descendants of the ancient Xaltocan population here explored, and we could thus define the dataset as the same population through time.

Laboratory procedures for modern Xaltocan samples

DNA was extracted from the saliva samples of 29 present-day residents of Xaltocan using the prepIT L2P kit (DNA Genotek) and following the manufacturer's guidelines. Library preparation of modern samples was performed following the same protocol used for the ancient samples, without applying the partial uracil-DNA glycosylase (UDG) treatment (Rohland et al. 2014). Each modern DNA library was prepared including a unique barcode combination of 7 bp length. Modern libraries were PCR amplified at the ancient DNA laboratory in Kiel. PCRs were performed in 50- μ L reactions containing 5 μ L barcoded library template DNA, 1 \times AccuPrime Pfx reaction mix (Thermo FisherScientific), 1.25 U AccuPrime Pfx DNA Polymerase (Thermo Fisher Scientific), 0.3 μ M of each of the two primers PreHyb-F, PreHyb-R. PCR conditions were as follows: 95 °C for 2 min, 10 cycles of (95 °C for 15 seconds, 55 °C for 30 seconds, 68 °C for 15 seconds), while performing a final extension at C 68 °C for 5 min. Amplified products were purified with the MinElute PCR Purification Kit (Qiagen) (elution volume 28 μ L) and quantified using the Agilent 2100 Bioanalyzer DNA 1000 chip.

HLA target-enrichment for ancient and modern samples

Both ancient partial UDG-treated and modern libraries were enriched for sequences 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 the custom bait library approach designed by Wittig et al. (2015) described in chapter 2. 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). A single capture reaction was performed for each of the ancient partial UDG-treated libraries, for which we prepared 750 ng of library DNA in a volume of 3.4 μ L. In the case of modern samples, up to four libraries have been pooled, for which 800 ng of library DNA per pool in a volume of 3.4 μ L has been prepared. To increase the specificity of target enrichment via hybridization, both ancient and modern libraries have been built using short (incomplete) barcoded adapters. Because of that, during hybridization reaction we used specific oligonucleotide blockers for each of the respective adapter lengths, as reported in the original protocol (Rohland et al. 2014). Once target enrichment is completed the short barcoded adapters need to be completed for sequencing by an indexing PCR. Thus, indexing post-capture PCRs were performed using a unique combination and 1 μ L of each P5 and P7 DNA indices (100 μ M) for each reaction. The post-capture PCR cycle number was set to 12. 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. Sequencing was done on the Illumina HiSeq 4000 (2 \times 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 ancient and modern samples

HTS data sets generated for the ancient and modern samples were pre-processed (adapter clipping, merging, trimming) using ClipAndMerge (version 1.7.3) from the EAGER pipeline (Peltzer et al. 2016). Adapters were excluded when present in the sequence, while reads with fewer than 25 nucleotides after adapter clipping or containing only adapter sequences were removed. All remaining paired-end reads were merged with a minimum overlap of 10 nucleotides and at most 5% mismatches in the overlap region. 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. As a single capture reaction was performed for each of the ancient partial UDG-treated libraries, the 6 mer barcode sequences were trimmed from both ends of the merged molecules after verifying that the reads contained the expected unique barcode combination. Up to four modern libraries were instead pooled in the same hybridization reaction, thus for each modern samples we selected the reads that matched the expected 7mer barcodes allowing up to two mismatches. As read redundancy can be useful information to identifying sequencing artifacts during manual HLA allele call (see chapter 2), we maintained read duplicates during the pre-processing and quality filtering steps.

Whole genome sequencing dataset of ancient Native American samples

We compiled a whole genome sequencing dataset ($n = 52$) of ancient Native American samples for comparison with the ancient sequence data generated in this study. Ancient human genome data of samples collected from different sites across the American continent (**Figure 3**), produced and analyzed in previous studies and showing genomic depth of coverage higher than 1x were selected (**Table S2**). Fastq files of samples were downloaded from NCBI Sequence Reads Archive (SRA) (Leinonen et al. 2011) using the accession numbers provided in each of the studies. When ancient genome data were available in SAM/BAM format, the corresponding files have been converted into fastq format.

HLA typing using the aHLA-Seq pipeline

All the sequence data generated in this study from ancient and modern libraries as well as ancient whole genome sequencing data compiled from published studies were processed through our aHLA-Seq pipeline and HLA genotypes at class II HLA loci (HLA-DRB1 and HLA-DQB1) manually defined following the steps described in chapter 2.

Statistical analysis

All the allele frequencies for class II genes (DRB1 and DQB1) reported in this work were obtained considering two different levels of resolution (1st field and 3rd field). The software PyPop (Lancaster et al. 2007) was applied to evaluate nonrandom associations, i.e. linkage disequilibrium (LD), between the two investigated loci DRB1 and DQB1 as well as frequent haplotypes with high LD in the modern Xaltocan samples. The normalized Hedrick's D' statistic (D') as well as the Cramer's V statistic (W_n) are the two measures provided for estimating the overall LD between DRB1 and DQB1 genes, while the permutation distribution of the likelihood-ratio test was used to test the significance of the overall LD between the genes. Haplotype frequencies were estimated using the iterative expectation-maximization (EM) algorithm (Dempster et al. 1977; Excoffier and Slatkin 1995).

Results

HLA class II molecular profile of pre-European contact Xaltocan samples

HLA molecular profiles at two of the class II HLA loci (HLA-DRB1 and HLA-DQB1) were defined for a set of historical samples (n=18) recovered from the archaeological sites of Xaltocan (Mexico) (**Figure 1**) and reported in **Table S3**.

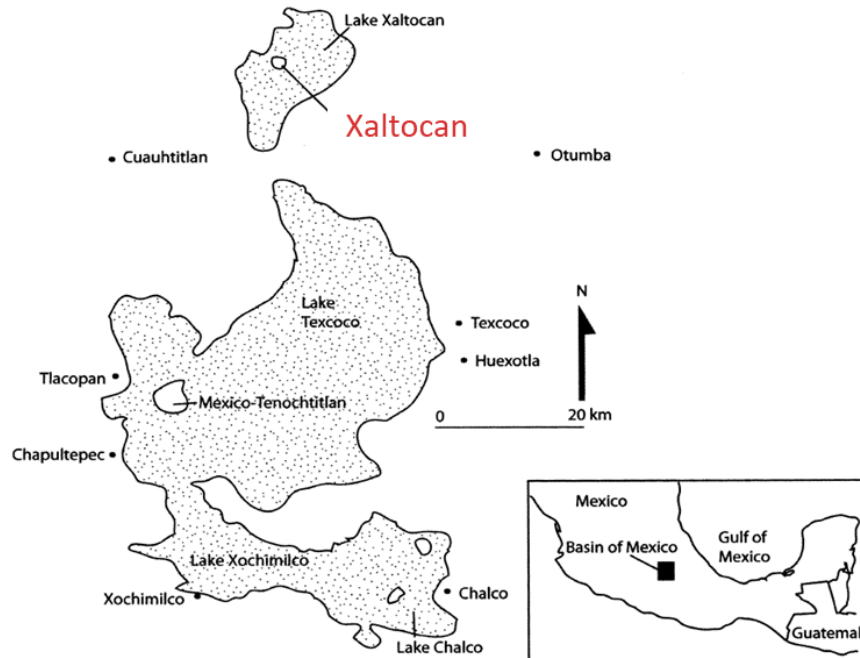


Figure 1 | Map of Xaltocan within the Basin of Mexico in ancient times. Adapted from Rodríguez-Alegría (2010).

The aHLA-Seq pipeline adopts the G-group nomenclature, a categorization of alleles that focuses only on nucleotide sequence differences inside the peptide-binding domains, and allows HLA typing at up to 3rd field resolution. Thus, the approach here used can in principle differentiate between alleles belonging to different allele groups (1st field); alleles differing in their protein sequence (2nd field) as well as alleles harboring synonymous exonic variants (3rd field). However, limited DNA quality and thus incomplete read coverage often found in ancient DNA sequences do not always allow to reach the 2nd field allele resolution, when for example the distinguishing SNP between two alleles of the same 1st field allele group is not covered by any read. In such cases, allele calls were rounded at 1st field level of resolution. In the same

way, when more than one allele defined at the 3rd field, but sharing the same 2nd field allele group, were equally well supported, allele calls were rounded at 2nd field level of resolution (**Table S3**). Of the 36 alleles investigated at each locus (N = 18 individuals), we were able to call at the 1st field resolution 19 alleles for HLA-DRB1 and 24 alleles for HLA-DQB1. Of these the allele call reached the 2nd field resolution for 10 alleles for HLA-DRB1 and 17 alleles for HLA-DQB1 (**Table 1**).

Table 1 | Locus-specific allele call success for the ancient Xaltocan samples (N =18)

	HLA-DRB1	HLA-DQB1
1 st field	19	24
- of these 2 nd field	10	17
NA (no call possible)	17	12

Number of alleles called at the 1st field and at the 2nd field level of resolution reported for each locus (2n = 36).

We thus quantified the ratio between the number of called alleles and the total number of the alleles assayed (two per locus and sample) and defined this as the success rate. As the 2nd field allele call success is equivalent with the success rate quantified for the 3rd field allele call, for the purpose of this study, we report only the success rate at the 1st field and at the 2nd field levels. The overall success rate calculated across the 2 investigated genes (DRB1 and DQB1) and for the whole dataset of historical samples was 60% at the 1st field level and 38% at 2nd field resolution (for values at each locus see **Figure 2** and **Table 2**). No more than two alleles at each locus have been found, indicating that the majority of DNA fragments in each sample originate from a single individual, thus supporting the validity of the contamination assessments previously mentioned.

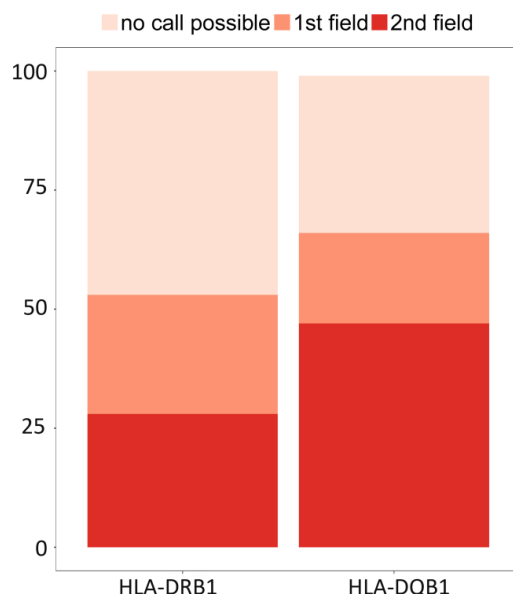


Figure 2 | HLA allele call success rate for 18 ancient Xaltocan samples. Percentage of alleles called at each individual HLA locus calculated for the whole set of ancient samples (N=18). Allele calls are reported at two different levels of resolution: 2nd field (4-digit) and 1st field (2-digit). 'No call possible' represents the fraction of cases where the allele call was not possible or allele calls were ambiguous.

Table 2 | Success rate for the ancient Xaltocan samples (N =18)

	HLA-DRB1	HLA-DQB1	Overall
Success 1 st field (%)	53	67	60
Success 2 nd field (%)	28	47	38

Success rate at the 1st field and at the 2nd field level of resolution, across the 2 investigated genes and overall.

We then attempted to describe the allele frequency distributions at the two investigated class II genes (DRB1 and DQB1) reported in **Table S4** and **Table 3**. At the HLA-DRB1 locus we described seven distinct allelic lineages defined at the 1st field resolution, with the DRB1*04 lineage observed at a frequency of 0.526 being thus the most common. The second most common lineage was represented by DRB1*08 found with a frequency of 0.211 (**Table S4**). The most common subtype observed was the allele DRB1*04:51:01, detected in 3 distinct samples. All the other alleles were found as singleton copies (**Table 3**). Two distinct allele lineages at 1st field resolution have been observed at HLA-DQB1 (**Table S4**). The DQB1*03 was observed with

a frequency of 0.667, with the most common subtype, the allele DQB1*03:02:01G, seen at a frequency of 0.529. The next two more common alleles were DQB1*04:02:01G and DQB1*03:01:01G both found at a frequency of 0.176 (**Tables 3**). We noted that the limited DNA quality allowed a good but incomplete reconstruction of the molecular profile at the DRB1 and DQB1 loci. Thus, given also the small samples size of the historical samples here investigated we could not statistically test the significance of the overall linkage disequilibrium between the two investigated loci. However, we noticed that in two out of three cases in which we could define the allele DRB1*04:51:01, it was observed together with the allele DQB1*03:02:01G (**Table S3**).

Table 3 | HLA class II allele frequencies at the 3rd field level for the ancient Xaltocan samples (N = 18)

HLA-DRB1			HLA-DQB1		
Allele	Frequency	n	Allele	Frequency	n
DRB1*04:51:01	0.300	3	DQB1*03:02:01G	0.529	9
DRB1*03:42:01	0.100	1	DQB1*04:02:01G	0.176	3
DRB1*04:07:01G	0.100	1	DQB1*03:01:01G	0.176	3
DRB1*04:11:01	0.100	1	DQB1*04:02:02 (1m 69 A >G)	0.059	1
DRB1*08:04:02	0.100	1	DQB1*04:18:01	0.059	1
DRB1*11:34:01	0.100	1			
DRB1*14:06:01	0.100	1			
DRB1*16:04:01	0.100	1			

Total number of alleles typed at 3rd field at each locus: HLA-DRB1 = 10, HLA-DQB1 = 17.

Distinct 3rd field alleles at each locus: HLA-DRB1 = 8, HLA-DQB1 = 5.

HLA class II molecular profile of modern Xaltocan samples

We next characterized the HLA molecular profile at the HLA class II loci (HLA-DRB1 and HLA-DQB1) for a dataset of 29 present-day residents of Xaltocan town (**Figure 1**), reported in **Table S5**. We included in this study unrelated individuals for which evidence of European and African ancestries have been excluded based on previous genetic ancestries estimates (Reynolds et al. 2019). As all alleles were defined at the G-group level **Table S5**, the success rate for allele calling was 100% at both the 1st and the 2nd field level of resolution. We thus described the allele

frequency distributions at HLA class II genes (DRB1 and DQB1), reported in **Table S6** and **Table 4**.

Table 4 | HLA class II allele frequencies at 3rd field resolution for the modern Xaltocan samples (N = 29)

HLA-DRB1			HLA-DQB1		
Allele	Frequency	n	Allele	Frequency	n
DRB1*04:07:01G	0.241	14	DQB1*03:02:01G	0.534	31
DRB1*04:04:01	0.190	11	DQB1*03:01:01G	0.293	17
DRB1*14:06:01	0.190	11	DQB1*04:02:01G	0.086	5
DRB1*16:02:01G	0.086	5	DQB1*03:03:02G	0.052	3
DRB1*04:11:01	0.052	3	DQB1*03:04:01G	0.017	1
DRB1*04:54:01	0.052	3			
DRB1*08:04:01	0.052	3			
DRB1*03:02:02	0.034	2			
DRB1*08:02:01G	0.034	2			
DRB1*14:02:01G	0.034	2			
DRB1*03:03:01	0.017	1			
DRB1*04:03:02	0.017	1			

Total number of alleles typed at 3rd field at each locus: HLA-DRB1 = 58, HLA-DQB1 = 58.

Distinct 3rd field alleles at each locus: HLA-DRB1 = 12, HLA-DQB1 = 5.

Five distinct allelic lineages (1st field resolution) were observed at HLA-DRB1 (**Table S6**). The most common lineage observed was DRB1*04 (f = 0.552), with the most common subtype DRB1*04:07:01G seen at a frequency of 0.241, followed by the allele DRB1*04:04:01 observed at a frequency of 0.190 (**Table 4**). The second most common lineage was represented by DRB1*14 (f = 0.224), with the most common subtype DRB1*14:06:01 found at a frequency of 0.190. The remaining HLA-DRB1 alleles were all observed at frequencies of less than 10% (**Table S6** and **Table 4**), while the alleles DRB1*03:03:01 and DRB1*04:03:02 were found as singleton copies. Two distinct allelic lineages were described at the DQB1 locus: the DQB1*03 was observed with a frequency of 0.914 while the DQB1*04 occurred with a frequency of f = 0.086. The most common subtypes were DQB1*03:02:01G (f = 0.534) and DQB1*03:01:01G (f = 0.293), while the remaining HLA-DQB1 alleles were observed at frequencies of lower than 10%. The allele DQB1*03:04:01G was the only one found as a single copy. The significance of the overall LD between the DRB1 and DQB1 loci was tested using the permutation distribution of the likelihood ratio test, implemented in Pypop (Lancaster et al. 2007). The two loci DRB1 and DQB1 showed significant nonrandom association (D' = 0.859; Wn = 0.846; p < 0.001).

Thus, haplotype frequencies were quantified using the expectation-maximization algorithm (Dempster et al. 1977; Excoffier and Slatkin 1995). The corresponding maximum likelihood estimates sorted in decreasing order, together with approximate number of haplotypes are reported in **Table S7**. Common haplotypes observed among the modern Xaltocan individual were DRB1*04:07:01-DQB1*03:02:01 ($f = 0.224$), DRB1*04:04:01-DQB1*03:02:01 ($f = 0.172$), DRB1*14:06:01-DQB1*03:01:01 ($f = 0.172$). Several other haplotypic associations were observed with frequencies of lower than 10% (**Table S7**).

Temporal comparison of HLA polymorphism between pre-European contact and contemporary Xaltocan samples

While we could describe 7 distinct allelic lineages at 1st field level of resolution at the DRB1 locus in the ancient Xaltocan samples, only 5 of them were still observed in the modern dataset. Two allelic lineages (DRB1*11 and DRB1*15) were not observed anymore in present-day Xaltocan, indicating a potential reduction in the number of allelic lineages through time. On the contrary, at the DQB1 locus the same allelic lineages were detected in both the ancient and modern datasets (DQB1*03 and DQB1*04). Some of the alleles defined at the 3rd field level of resolution that were described in the ancient dataset were not detected anymore in the modern Xaltocan samples: DRB1*04:51:01, DRB1*03:42:01, DRB1*11:34:01 and DRB1*16:04:01. We thus explored in the Allele Frequency Net Database (AFND) (González-Galarza et al. 2015) the worldwide allele distribution of variants in modern-day human populations. The allele DRB1*04:51:01 that was detected in three different ancient samples, is an allele currently present at low frequencies in Chinese ($f = 0.001$) and South Korean ($f = 0.0002$) populations, but not found in present-day Native Americans. The other three alleles DRB1*03:42:01, DRB1*11:34:01 and DRB1*16:04:01 have been described at low frequencies in populations from Central and South America, Asia, Israel and Europe. One of the variants observed in the ancient dataset but not in the modern one showed an A→G transition in position 69: DQB1*04:02:02 (1m 69 A >G). When translating the coding sequence into the corresponding protein sequence this was equivalent to a change from the amino acid threonine (T) into alanine (A) occurring in one of the antigen binding sites (ABS) as they are defined by Brown et al. (1993) (**Figure S1**).

HLA polymorphism in pre-European contact Native American populations

Genetic variability at HLA class II loci was also explored in the subset of historical samples collected from two archaeological sites on St. Catherines Island (Guale-Fallen Tree and Mission

Santa Catalina de Guale, $n = 8$) as well as in available ancient whole-genome data of samples collected from different sites across the American continent (**Figure 3**). Allele calls obtained at the two loci (HLA-DRB1 and HLA-DQB1) are reported in **Table S8** and **Table S9**. To characterize HLA polymorphism in ancient pre-European contact Native American populations at a broader geographical scale, results were combined together with allele calls obtained for the ancient Xaltocan dataset and the allele frequency distributions in the whole set of historical Native American samples are described (**Table S10** and **Table 5**).

Nine distinct allelic lineages defined at the 1st field level of resolution were described at the HLA-DRB1 locus, with the DRB1*04 lineage observed at a frequency of 0.314 being thus the most common. The second most common lineage was represented by DRB1*14 ($f = 0.257$), followed by the DRB1*08 lineage ($f = 0.157$). The remaining allelic lineages were all observed at frequencies lower than 10% (**Table S10**). According with the allele frequency distributions characterized in ancient Xaltocan, the allele DRB1*04:51:01 appeared as the most common 3rd field HLA-DRB1 allele also in the whole dataset of ancient samples, observed in 9 instances widespread across the continent. The second most common allele was DRB1*14:02:01G ($f = 0.135$) which has been observed in five samples all collected from the archaeological site of Prince Rupert Harbor in British Columbia. The allele DRB1*08:02:01G was observed three times ($f = 0.081$) in samples from California Channel Islands and St. Catherines Island; while the allele DRB1*12:05:01 ($f = 0.054$) has been observed in two samples from archaeological sites located on the Pacific Northwest coast of America (Lovelock Cave and Prince Rupert Harbor). All the other 2nd field HLA-DRB1 alleles were found as singleton copies (**Table 5** and **Table S8-S9**).



Figure 3 | Overview of sampling locations for the ancient samples. Sampling locations for ancient whole genome data from previous studies are labeled in dark blue, while archaeological locations for ancient sequence data generated in this study are labeled in red.

Table 5 | HLA class II allele frequencies at the 2nd field level for whole set of ancient Native American samples (N = 78)

HLA-DRB1			HLA-DQB1		
Allele	Frequency	n	Allele	Frequency	n
DRB1*04:51:01	0.243	9	DQB1*03:01:01G	0.340	18
DRB1*14:02:01G	0.135	5	DQB1*03:02:01G	0.245	13
DRB1*08:02:01G	0.081	3	DQB1*04:02:01G	0.170	9
DRB1*12:05:01	0.054	2	DQB1*03:03:02G	0.094	5
DRB1*03:01:02	0.027	1	DQB1*04:02:02 (69 A >G)	0.057	3
DRB1*03:02:02	0.027	1	DQB1*03:03:02G (162 G>A)	0.019	1
DRB1*03:04:01	0.027	1	DQB1*03:04:02	0.019	1
DRB1*03:42:01	0.027	1	DQB1*03:133:01	0.019	1
DRB1*04:04:01	0.027	1	DQB1*03:255:01	0.019	1
DRB1*04:07:01G	0.027	1	DQB1*04:18:01	0.019	1
DRB1*04:10:01G	0.027	1			
DRB1*04:11:01	0.027	1			
DRB1*08:04:02	0.027	1			
DRB1*08:38:01	0.027	1			
DRB1*11:34:01	0.027	1			
DRB1*14:03:02 (129 C>T)	0.027	1			
DRB1*14:06:01	0.027	1			
DRB1*14:106:01 (135 G>A)	0.027	1			
DRB1*14:48:01	0.027	1			
DRB1*15:64:01	0.027	1			
DRB1*16:02:01G	0.027	1			
DRB1*16:04:01	0.027	1			

Total number of alleles typed at 1st field resolution at each locus: HLA-DRB1 = 37, HLA-DQB1 = 53. Distinct 1st field alleles at each locus: HLA-DRB1 = 22, HLA-DQB1 = 10.

At the HLA-DQB1 locus we observed two distinct allelic lineages defined at the 1st field level of resolution (DQB1*03 and DQB1*04) with a frequency distribution resembling ancient Xaltocan samples (**Table S10**). The most common 3rd field HLA-DQB1 allele was DQB1*03:01:01G (f = 0.340) observed in eighteen instances widespread across the continent. The second most common allele DQB1*03:02:01G (f = 0.245) was observed in thirteen samples also well distributed across the continent. The allele DQB1*04:02:01G showed a frequency of 0.170, and was detected in nine samples mainly from Central and North America. The allele DQB1*03:03:02G (f = 0.094) was instead observed five times in samples collected from sites located on the Pacific Northwest coast of America (Big Bar, Lovelock Cave and Spirit Cave) as well as from the site of Guale-Fallen Tree in St. Catherines Island. Finally the mutated variant

DQB1*04:02:02 (69 A >G) observed in the dataset of ancient Xaltocan samples was also found in two other different historical samples from California Channel Islands and St. Catherines Island. All the other 3rd field HLA-DQB1 alleles were found as singleton copies (**Table 5** and **Table S8-S9**).

Discussion

Pathogen-mediated selection pressures are suspected to be among the strongest forces in human evolution, likely affecting the immunogenetic variation of ancient and present-day human populations (Fumagalli et al. 2011). The highly polymorphic human leukocyte antigen (HLA) genes play a key role in adaptive immunity and are a prime candidate to investigate mechanisms of pathogen selection mechanisms throughout human history (Klein and Figueroa 1986). Furthermore, the recent development of genomic tools for analysis of ancient DNA provides a unique opportunity to unravel the trajectories of genetic variants associated with human adaptations to newly introduced or co-evolving pathogens (Marciniak and Perry 2017). In this light, the study of HLA variability in historical human populations during epidemiological events is of strong interest to explore the genetic determinants of infection susceptibility or severity, which might be relevant in explaining the emergence of infectious disease.

The arrival of Europeans in the Americas marked the beginning of Native American population decline. The spread of epidemic diseases caused by European-borne pathogens into immunologically naive indigenous populations is considered one of the main drivers behind the quick rise in mortality rate documented among Native Americans (Thornton 1997). This hypothesis has been extensively studied by exploring the genetic variability at immune-related genes in present-day Native American populations (Black 1975; Belich et al. 1992; Watkins et al. 1992; Black and Hedrick 1997; Lindenau et al. 2013), however genetic diversity at immune genes in ancient Native American populations remains largely unexplored.

In this work we explored the polymorphisms of HLA class II loci (HLA-DRB1 and HLA-DQB1) in ancient pre-European contact Native American individuals, providing the first spatiotemporal characterization of genetic variability at HLA class II genes in ancient Native American populations.

HLA polymorphisms were first investigated in eighteen historical samples collected from the archaeological site of the town of Xaltocan (Central Mexico). The samples, ranging in age from 1240 to 1521 CE, belong to different historical periods and are characterized by a complex

history. Although genetic analysis at uniparental markers (Y-chromosome and mitochondrial DNA) in a larger set of ancient Xaltocan samples revealed genetic discontinuities associated with the Tepanec, Aztec, and Spanish conquests (Mata-Míguez 2016), the same discontinuities between the different ancient time periods was not confirmed when exploring genome-wide diversity (Reynolds 2018). When looking at the HLA molecular profiles of ancient Xaltocan at the HLA-DRB1 locus we observed the occurrence of DRB1*04 alleles across the different ancient time periods. However, distinct alleles with a 1st field level of resolution have been also observed between the different ancient times. In pre-1395 Xaltocan samples, which belong to different subgroups based on excavation location, the allele DRB1*15 was uniquely detected in the Early Interior samples while the DRB1*14 allele was only observed in the Early Periphery samples. These alleles were not detected anymore in samples from the subsequent Tepanec period, for which we instead observed DRB1*03 and DRB1*11 lineages. Furthermore, it may appear that these lineages were subsequently replaced by DRB1*08 alleles in post-1395 samples from the Aztec and late unknown periods, for which the allele DRB1*16 was also detected. As the sample size for each subgroup here investigated is very small, we cannot exclude that the alleles observed in samples from previous periods were also present in the subsequent phases, and further investigations are required. The most common subtype at the DRB1 locus, the allele DRB1*04:51:01, was detected in 3 distinct samples from both pre-1395 and post-1395 periods. When exploring HLA polymorphisms at the HLA-DQB1 locus, DQB1*03 alleles were observed across the different ancient time periods, while the DQB1*04 alleles was detected only in pre-1395 Xaltocan samples from the Early Periphery subgroup and in all subsequent post-1395 samples. Our preliminary results support discontinuities between the different ancient time periods as revealed by uniparental markers (Y-chromosome and mitochondrial DNA), however further work using a larger samples size are required to validate the observed differences. Notably, we were able to detect HLA alleles that were present in the ancient Native American populations but no longer exist in modern human populations. We indeed observed in one of the ancient Xaltocan samples an HLA variant differing by a single point mutations (A to G) to the present-day DQB1*04:02:02 allele (**Figure S1**). The non-synonymous mutation result in the substitution of the amino acid threonine (T) into alanine (A) occurring in one of the antigen binding sites (ABS). ABS sites are located within the peptide binding pocket of HLA molecules, thus in contact with the antigen side chains, and are characterized by an excess of non-synonymous substitutions (Furlong and Yang 2008). Changes in these positions result in modifications of binding properties of alleles and are likely adaptive (Brown et al. 1993).

We next characterized the HLA genetic diversity in 29 present-day residents of Xaltocan town. Thus, focusing on the same population through time, we explored the changes in the HLA molecular profiles between pre- and post-European contact periods. A potential reduction in the number of allelic lineages through time was revealed at the DRB1 locus. Nevertheless, the allelic lineages DRB1*11 and DRB1*15, which were detected in the ancient but not in the modern Xaltocan, are still widespread at intermediate frequency in modern Native American populations other than Xaltocan. The most common subtype at the DRB1 locus in the ancient dataset, the allele DRB1*04:51:01, was not observed either in modern Xaltocan samples or other modern Native American populations. HLA polymorphism were next defined for a dataset of 52 available ancient Native American samples with whole genome sequencing data collected from different sites across the American continent, and results combined with the ancient sequence data generated in this study (18 samples from Xaltocan site and 8 samples from St. Catherines Island sites). Thus, by defining the allele frequency distributions in the whole set of historical Native American samples ($n = 78$) we were able to describe the HLA polymorphism in ancient pre-European contact populations at a broad geographical and temporal scale.

In line with previous results, the DRB1*04 lineage was detected across different geographical areas as well as different time points in the past. The allele DRB1*04:51:01, appeared as the most common HLA-DRB1 allele also in the whole dataset of ancient samples, observed in 9 instances widespread across the continent and from different periods. It was indeed detected in samples 9000 years old and located in the northern part of the continent, like in the case of USR1 (Alaska), but also in samples from more recent periods and located in the extreme south, like in the case of Aconcagua (Argentina) and Ayayema (Chile). These results suggest that the ancestral allele DRB1*04:51:01 was probably quite widespread in ancient pre-European Native American populations, possibly reflecting local adaptation to pathogens present in the ancient environment. Despite it still being present in 500 years old Native American samples, the allele is not found anymore in present-day Native American populations, but it persists at low frequencies in Chinese ($f = 0.001$) and South Korean ($f = 0.0002$) populations. Notably, the mutated variant DQB1*04:02:02 (69 A >G) observed in the dataset of ancient Xaltocan samples was successfully detected in two more historical samples from California Channel Islands and St. Catherines Island. Thus, the allele was quite distributed in ancient pre-European Native American populations of Central and North America but is no longer found in modern-day humans.

Both the alleles, DRB1*04:51:01 and DQB1*04:02:02 (69 A >G), could have been thus removed under negative selection in recent history because they conferred susceptibility to one or more of the newly introduced European-borne pathogens. However, the observed change in allele frequency might also be the consequence of the strong population bottleneck documented in Native American populations, resulting from stochastic factors associated with their demographic history. Further works are thus required, in which, for instance, the information about HLA diversity here obtained could be contrasted with information about genome-wide diversity from neutral SNPs, to disentangle the differential contribution of neutral processes and potential pathogen-mediated selection for the evolution of genetic diversity in indigenous populations during and after European contact.

Supplementary Materials

Table S1 | Summary data for the archaeological samples analyzed in this study

ID	Archaeological sites	Date (CE)	Period
G.1	Xaltocan (central Mexico)	1240-1395	Early Interior
ZC.1	Xaltocan (central Mexico)	1240-1395	Early Interior
ZA.1	Xaltocan (central Mexico)	1240-1395	Early Interior
Y3.6	Xaltocan (central Mexico)	1240-1395	Early Periphery
E14.5	Xaltocan (central Mexico)	1290-1350	Early Periphery
Y2.7	Xaltocan (central Mexico)	1200-1500	Early Periphery
E8.5	Xaltocan (central Mexico)	1390-1440	Tepanec
E7.1	Xaltocan (central Mexico)	1390-1460	Tepanec
E8.1	Xaltocan (central Mexico)	1390-1460	Tepanec
E8.3	Xaltocan (central Mexico)	1410-1490	Tepanec
E10.1	Xaltocan (central Mexico)	1400-1470	Tepanec
E8.4	Xaltocan (central Mexico)	1390-1490	Tepanec
E14.6	Xaltocan (central Mexico)	1430-1500	Aztec
E6.1	Xaltocan (central Mexico)	1410-1450	Aztec
E30.3	Xaltocan (central Mexico)	1395-1521	Late unknown
E34.1	Xaltocan (central Mexico)	1330-1430	Late unknown
E14.1	Xaltocan (central Mexico)	1395-1521	Late Unknown
E5.1	Xaltocan (central Mexico)	1390-1520	Late Unknown
FT16	Guale-Fallen Tree, St. Catherines Island (Georgia)	1500-1600	pre-contact / contact period
FT28	Guale-Fallen Tree, St. Catherines Island (Georgia)	1500-1600	pre-contact / contact period
FT32	Guale-Fallen Tree, St. Catherines Island (Georgia)	1500-1600	pre-contact / contact period
FT34	Guale-Fallen Tree, St. Catherines Island (Georgia)	1500-1600	pre-contact / contact period
FT49	Guale-Fallen Tree, St. Catherines Island (Georgia)	1500-1600	pre-contact / contact period
SCDG32b	Mission Santa Catalina de Guale, St. Catherines Island (Georgia)	1600-1700	post-contact mission period
SCDG58b	Mission Santa Catalina de Guale, St. Catherines Island (Georgia)	1600-1700	post-contact mission period
SCDG60b	Mission Santa Catalina de Guale, St. Catherines Island (Georgia)	1600-1700	post-contact mission period

Table S2 | Summary data for the 52 available ancient whole-genome data collected from different sites across the American continent, reported together with their references

ID	Archaeological sites	Radiocarbon age BP	Ref.
Saqqaq	Greenland	4,000	(Rasmussen et al. 2010)
USR1	Upward Sun River site, Alaska	9970 ± 30	(Moreno-Mayar et al. 2018a)
XVII-B-167	Prince Rupert Harbor, British Columbia	n/a	(Lindo et al. 2016)
XVII-B-322	Prince Rupert Harbor, British Columbia	2050±50	(Lindo et al. 2016)
XVII-B-516	Prince Rupert Harbor, British Columbia	n/a	(Lindo et al. 2016)
XVII-B-412	Prince Rupert Harbor, British Columbia	1940±40	(Lindo et al. 2016)
XVII-B-158	Prince Rupert Harbor, British Columbia	2290±50	(Lindo et al. 2016)
XVII-B-507	Prince Rupert Harbor, British Columbia	2320± 65	(Lindo et al. 2016)
XVII-B-470	Prince Rupert Harbor, British Columbia	1600±40	(Lindo et al. 2016)
XVII-B-468	Prince Rupert Harbor, British Columbia	1940±45	(Lindo et al. 2016)
XVII-B-532	Prince Rupert Harbor, British Columbia	n/a	(Lindo et al. 2016)
XVII-B-406	Prince Rupert Harbor, British Columbia	n/a	(Lindo et al. 2016)
XVII-B-939	Prince Rupert Harbor, British Columbia	5710±40	(Lindo et al. 2016)
XVII-B-365	Prince Rupert Harbor, British Columbia	2270±65	(Lindo et al. 2016)
XVII-B-443	Prince Rupert Harbor, British Columbia	1820±55	(Lindo et al. 2016)
XVII-B-525	Prince Rupert Harbor, British Columbia	1860±40	(Lindo et al. 2016)
XVII-B-125	Prince Rupert Harbor, British Columbia	2260±40	(Lindo et al. 2016)
XVII-B-318	Prince Rupert Harbor, British Columbia	1550±50	(Lindo et al. 2016)
XVII-B-413	Prince Rupert Harbor, British Columbia	1970±42	(Lindo et al. 2016)
XVII-B-386	Prince Rupert Harbor, British Columbia	1060±40	(Lindo et al. 2016)
XVII-B-168	Prince Rupert Harbor, British Columbia	2650±75	(Lindo et al. 2016)
XVII-B-357	Prince Rupert Harbor, British Columbia	n/a	(Lindo et al. 2016)
XVII-B-181	Prince Rupert Harbor, British Columbia	2620±40	(Lindo et al. 2016)
XVII-B-163	Prince Rupert Harbor, British Columbia	n/a	(Lindo et al. 2016)
XVII-B-302	Prince Rupert Harbor, British Columbia	2440±75	(Lindo et al. 2016)
XVII-B-311	Prince Rupert Harbor, British Columbia	2090±60	(Lindo et al. 2016)
XVII-B-300	Prince Rupert Harbor, British Columbia	1650±75	(Lindo et al. 2016)
BigBar	Northwest Canada	5110 ± 40	(Lindo et al. 2016)
Kennewick Man	Washington State	8,340–9,200	(Moreno-Mayar et al. 2018b)
Lovelock1	Lovelock Cave, Nevada, US	2001 ± 28	(Rasmussen et al. 2015)
			(Moreno-Mayar et al. 2018b)

Lovelock2	Lovelock Cave, Nevada, US	1925 ± 29	(Moreno-Mayar et al. 2018b)
Lovelock3	Lovelock Cave, Nevada, US	703 ± 26	(Moreno-Mayar et al. 2018b)
Lovelock4	Lovelock Cave, Nevada, US	1836 ± 28	(Moreno-Mayar et al. 2018b)
Spirit Cave	Spirit Cave, Nevada, US	9615 ± 50	(Moreno-Mayar et al. 2018b)
CK-13	Lucier, Aouthwestern Ontario	4267 ± 22	(Scheib et al. 2018)
CR-01	Santa Cruz Island, California Channel Islands	1111 ± 39	(Scheib et al. 2018)
CT-01	Santa Catalina Island, California Channel Islands	387 ± 38	(Scheib et al. 2018)
NC	New Cuyama, California Channel Islands	1450 ± 30	(Scheib et al. 2018)
PS-06	Point Sal, California Channel Islands	1570 ± 41	(Scheib et al. 2018)
SC-05	San Clemente Island, California Channel Islands	1101 ± 41	(Scheib et al. 2018)
SM-02	San Miguel Island, California Channel Islands	826 ± 26	(Scheib et al. 2018)
SN-11	San Nicolas Island, California Channel Islands	1172 ± 39	(Scheib et al. 2018)
SN-17	San Nicolas Island, California Channel Islands	4517 ± 51	(Scheib et al. 2018)
Lucayan Taino	Preacher's Cave, Bahamas, Carabian	1,000	(Schroeder et al. 2018)
Sumidouro4	Caverna do Sumidouro, Lagoa Santa, Brazil	9240 ± 50	(Moreno-Mayar et al. 2018b)
Sumidouro5	Caverna do Sumidouro, Lagoa Santa, Brazil	8785 ± 50	(Moreno-Mayar et al. 2018b)
Sumidouro6	Caverna do Sumidouro, Lagoa Santa, Brazil	9070 ± 45	(Moreno-Mayar et al. 2018b)
Sumidouro7	Caverna do Sumidouro, Lagoa Santa, Brazil	9015 ± 45	(Moreno-Mayar et al. 2018b)
Sumidouro8	Caverna do Sumidouro, Lagoa Santa, Brazil	8990 ± 50	(Moreno-Mayar et al. 2018b)
Aconcagua	Aconcagua, Mendoza province, Argentina	500	(Moreno-Mayar et al. 2018b)
Ayayema	Ayayema Cave, Patagonia, Chile	4,520±60BP	(Moreno-Mayar et al. 2018b)
Punta Santa Ana	Punta Santa Ana, Patagonia, Chile	7200	(Moreno-Mayar et al. 2018b)

Table S3 | Individual allele call at HLA class II genes for the ancient Xaltocan samples (N = 18)

ID	DRB1 1	DRB1 2	DQB1 1	DQB1 2
G.1	NA	NA	DQB1*03:02:01G	NA
ZC.1	DRB1*04:07:01G	DRB1*15	DQB1*03:01:01G	DQB1*03:02:01G
ZA.1	DRB1*04:11:01	DRB1*04:51:01	DQB1*03:02:01G	DQB1*03:02:01G
Y3.6	DRB1*14:06:01	NA	DQB1*04:02:01G	DQB1*03:01:01G
E14.5	DRB1*04:51:01	NA	DQB1*04:18:01	NA
Y2.7	NA	NA	NA	NA
E8.5	DRB1*04	NA	DQB1*03:02:01G	DQB1*03:02:01G
E7.1	DRB1*03:42:01	NA	DQB1*03	NA
E8.1	DRB1*04	NA	NA	NA
E8.3	DRB1*11:34:01	NA	DQB1*03	NA
E10.1	DRB1*04:51:01	DRB1*04	DQB1*03:02:01G	DQB1*03:02:01G
E8.4	NA	NA	DQB1*04	DQB1*04
E14.6	DRB1*08	NA	DQB1*03	DQB1*04:02:01G
E6.1	NA	NA	DQB1*03	DQB1*04
E30.3	DRB1*08	DRB1*04	DQB1*04:02:01G	DQB1*03:02:01G
E34.1	DRB1*08	DRB1*16:04:01	DQB1*04:02:02 (1m 69 A >G)	NA
E14.1	DRB1*04	NA	DQB1*03:01:01G	NA
E5.1	DRB1*08:04:02	NA	NA	NA

Table S4 | HLA class II allele frequencies at 1st field level for the ancient Xaltocan samples (N = 18)

HLA-DRB1			HLA-DQB1		
Allele	Frequency	n	Allele	Frequency	n
DRB1*04	0.526	10	DQB1*03	0.667	16
DRB1*08	0.211	4	DQB1*04	0.333	8
DRB1*03	0.053	1			
DRB1*11	0.053	1			
DRB1*14	0.053	1			
DRB1*15	0.053	1			
DRB1*16	0.053	1			

Total number of 2-digit typed alleles at each locus: HLA-DRB1 = 19, HLA-DQB1 = 22. Distinct 2-digit alleles at each locus: HLA-DRB1 = 7, HLA-DQB1 = 2.

Table S5 | Individual allele call at HLA class II genes for the modern Xaltocan samples (N = 29)

ID	DRB1 1	DRB1 2	DQB1 1	DQB1 2
XalMod2	DRB1*04:04:01	DRB1*04:04:01	DQB1*03:02:01G	DQB1*03:02:01G
XalMod3	DRB1*16:02:01G	DRB1*03:02:02	DQB1*03:02:01G	DQB1*03:01:01G
XalMod5	DRB1*04:07:01G	DRB1*14:02:01G	DQB1*03:01:01G	DQB1*03:02:01G
XalMod6	DRB1*04:04:01	DRB1*03:02:02	DQB1*03:01:01G	DQB1*03:03:02G
XalMod8	DRB1*14:02:01G	DRB1*04:04:01	DQB1*03:01:01G	DQB1*03:02:01G
XalMod9	DRB1*16:02:01G	DRB1*04:04:01	DQB1*03:02:01G	DQB1*03:02:01G
XalMod14	DRB1*08:04:01	DRB1*04:04:01	DQB1*04:02:01G	DQB1*03:02:01G
XalMod15	DRB1*14:06:01	DRB1*14:06:01	DQB1*03:03:02G	DQB1*03:01:01G
XalMod17	DRB1*04:07:01G	DRB1*14:06:01	DQB1*03:03:02G	DQB1*03:01:01G
XalMod19	DRB1*16:02:01G	DRB1*04:03:02	DQB1*03:04:01G	DQB1*03:02:01G
XalMod22	DRB1*04:11:01	DRB1*04:11:01	DQB1*03:02:01G	DQB1*03:02:01G
XalMod25	DRB1*04:54:01	DRB1*08:04:01	DQB1*04:02:01G	DQB1*03:02:01G
XalMod28	DRB1*14:06:01	DRB1*04:07:01G	DQB1*03:02:01G	DQB1*03:01:01G
XalMod30	DRB1*08:02:01G	DRB1*04:54:01	DQB1*04:02:01G	DQB1*03:02:01G
XalMod31	DRB1*14:06:01	DRB1*14:06:01	DQB1*03:01:01G	DQB1*03:01:01G
XalMod33	DRB1*04:07:01G	DRB1*04:07:01G	DQB1*03:02:01G	DQB1*03:02:01G
XalMod36a	DRB1*16:02:01G	DRB1*14:06:01	DQB1*03:01:01G	DQB1*03:01:01G
XalMod38	DRB1*14:06:01	DRB1*14:06:01	DQB1*03:01:01G	DQB1*03:01:01G
XalMod43	DRB1*04:04:01	DRB1*14:06:01	DQB1*03:01:01G	DQB1*03:02:01G
XalMod47	DRB1*04:07:01G	DRB1*04:07:01G	DQB1*03:02:01G	DQB1*03:02:01G
XalMod13	DRB1*04:04:01	DRB1*03:03:01	DQB1*03:02:01G	DQB1*03:01:01G
XalMod18	DRB1*04:04:01	DRB1*08:04:01	DQB1*03:02:01G	DQB1*04:02:01G
XalMod20	DRB1*04:07:01G	DRB1*04:07:01G	DQB1*03:02:01G	DQB1*03:02:01G
XalMod21	DRB1*04:07:01G	DRB1*16:02:01G	DQB1*03:02:01G	DQB1*03:01:01G
XalMod23	DRB1*04:07:01G	DRB1*04:07:01G	DQB1*03:02:01G	DQB1*03:02:01G
XalMod37	DRB1*14:06:01	DRB1*04:04:01	DQB1*03:02:01G	DQB1*03:01:01G
XalMod40	DRB1*04:11:01	DRB1*04:07:01G	DQB1*03:02:01G	DQB1*03:02:01G
XalMod41	DRB1*04:07:01G	DRB1*04:04:01	DQB1*03:02:01G	DQB1*03:02:01G
XalMod45	DRB1*04:54:01	DRB1*08:02:01G	DQB1*04:02:01G	DQB1*03:02:01G

Table S6 | HLA class II allele frequencies at 1st field level for the modern Xaltocan samples (N = 29)

HLA-DRB1			HLA-DQB1		
Allele	Frequency	n	Allele	Frequency	n
DRB1*04	0.552	32	DQB1*03	0.914	53
DRB1*14	0.224	13	DQB1*04	0.086	5
DRB1*08	0.086	5			
DRB1*16	0.086	5			
DRB1*03	0.052	3			

Total number of 2-digit typed alleles at each locus: HLA-DRB1 = 58, HLA-DQB1 = 58. Distinct 2-digit alleles at each locus: HLA-DRB1 = 5, HLA-DQB1 = 2.

Table S7 | Haplotype frequencies ($f > 0.03$) for the modern Xaltocan samples ($N = 29$) calculated using the expectation-maximization algorithm

Haplotype	Frequency	Copy number
DRB1*040701:DQB1*030201	0.224	13
DRB1*040401:DQB1*030201	0.172	10
DRB1*140601:DQB1*030101	0.172	10
DRB1*080401:DQB1*040201	0.052	3
DRB1*041101:DQB1*030201	0.052	3
DRB1*045401:DQB1*030201	0.052	3
DRB1*160201:DQB1*030201	0.052	3
DRB1*160201:DQB1*030101	0.034	2
DRB1*140201:DQB1*030101	0.034	2
DRB1*080201:DQB1*040201	0.034	2
DRB1*030202:DQB1*030101	0.034	2

Global linkage disequilibrium estimates: $D' = 0.859$; $Wn = 0.846$; $p\text{-value} = 0.0000^*$

Table S8 | Individual allele call at HLA class II genes for the St. Catherines Island (Georgia) samples

ID	DRB1 1	DRB1 2	DQB1 1	DQB1 2
FT16	DRB1*09	NA	DQB1*04:02:01G	NA
FT28	DRB1*16:02:01G	DRB1*08	DQB1*04:02:01G	DQB1*03:01:01G
FT32	NA	NA	DQB1*03:01:01G	DQB1*03:02
FT34	DRB1*03:01:02	NA	DQB1*03:02	DQB1*03
FT49	DRB1*08:02:01G	DRB1*04:04:01	DQB1*03:02:01G	DQB1*03:02:01G
SCDG32b	NA	NA	NA	NA
SCDG58b	NA	NA	DQB1*04:02:02 (69 A >G)	NA
SCDG60b	DRB1*08:02:01G	NA	DQB1*03	DQB1*04

Table S9 | Individual allele call at HLA class II genes for the 52 available ancient whole-genome data

ID	DRB1 1	DRB1 2	DQB1 1	DQB1 2
Saqqaq	DRB1*04	NA	DQB1*03:01	NA
USR1	NA	DRB1*04:51:01	DQB1*03	DQB1*03:04:02
XVII-B-167	NA	NA	DQB1*03	NA
XVII-B-322	NA	NA	DQB1*03:133:01	NA
XVII-B-516	DRB1*14:106:01 (135 G>A)	NA	DQB1*03	NA
XVII-B-412	NA	NA	DQB1*03	NA
XVII-B-158	DRB1*14	DRB1*14:03:02 (129 C>T)	DQB1*03	NA
XVII-B-507	DRB1*12:05:01	DRB1*14	DQB1*03:02:01G	NA
XVII-B-470	DRB1*14:02:01G	DRB1*14:02:01G	DQB1*03:01:01G	DQB1*03:255:01
XVII-B-468	DRB1*14:02:01G	DRB1*14	DQB1*04:02:01G	DQB1*03:01:01G
XVII-B-532	NA	NA	NA	NA
XVII-B-406	NA	NA	NA	DQB1*03
XVII-B-939	NA	NA	NA	NA
XVII-B-365	DRB1*14	NA	DQB1*03:01:01G	DQB1*03:01:01G
XVII-B-443	DRB1*04	DRB1*14	DQB1*04:02:01G	DQB1*03:01:01G
XVII-B-525	NA	NA	NA	NA
XVII-B-125	NA	NA	DQB1*03	NA
XVII-B-318	NA	NA	DQB1*03:01:01G	NA
XVII-B-413	DRB1*14:02:01G	DRB1*14	DQB1*04:02:01G	DQB1*03:01:01G
XVII-B-386	NA	NA	NA	NA
XVII-B-168	DRB1*14:48:01	DRB1*14	DQB1*03	DQB1*03:01:01G
XVII-B-357	NA	NA	DQB1*03:01:01G	DQB1*04
XVII-B-181	DRB1*04:10:01G	NA	DQB1*03:01:01G	DQB1*04
XVII-B-163	NA	NA	NA	NA
XVII-B-302	DRB1*14	NA	DQB1*03:01:01G	DQB1*03:01:01G
XVII-B-311	DRB1*14	NA	DQB1*03	NA
XVII-B-300	DRB1*14	DRB1*09	DQB1*03:03:02G (162 G>A)	NA
BigBar	DRB1*04:51:01	NA	DQB1*03:03:02G	NA
Kennewick Man	DRB1*04	na	NA	NA
Lovelock1	NA	NA	DQB1*03	NA
Lovelock2	NA	DRB1*12:05:01	DQB1*03:03:02G	DQB1*03:03:02G

Lovelock3	DRB1*03:04:01	DRB1*03:02:02	DQB1*03:01	DQB1*03:01
Lovelock4	NA	NA	DQB1*03	NA
Spirit Cave	DRB1*09	DRB1*04:51:01	DQB1*03:03:02G	DQB1*03:03:02G
CK-13	DRB1*09	NA	DQB1*03	NA
CR-01	NA	NA	DQB1*03	NA
CT-01	DRB1*08	NA	DQB1*04	NA
NC	DRB1*08	NA	DQB1*04	NA
PS-06	NA	NA	DQB1*04	NA
SC-05	DRB1*08:02:01G	NA	DQB1*03:01:01G	DQB1*04:02:02 (69 A>G)
SM-02	DRB1*08:38:01	NA	DQB1*04:02:01G	NA
SN-11	DRB1*15:64:01	NA	NA	NA
SN-17	DRB1*09	DRB1*16	DQB1*03	DQB1*03
Lucayan Taino	DRB1*04	NA	DQB1*03	NA
Sumidouro4	NA	NA	NA	NA
Sumidouro5	DRB1*04:51:01	DRB1*15	NA	DQB1*03
Sumidouro6	NA	NA	DQB1*03	NA
Sumidouro7	NA	NA	DQB1*03	NA
Sumidouro8	NA	NA	NA	NA
Aconcagua	DRB1*04:51:01	DRB1*15	DQB1*03	NA
Ayayema	DRB1*04:51:01	NA	DQB1*03:02:01G	NA
Punta Santa Ana	NA	NA	DQB1*03	NA

Table S10 | HLA class II allele frequencies at 1st field level for whole set of ancient Native America samples (N = 78)

HLA-DRB1			HLA-DQB1		
Allele	Frequency	n	Allele	Frequency	n
DRB1*04	0.314	22	DQB1*03	0.766	72
DRB1*14	0.257	18	DQB1*04	0.234	22
DRB1*08	0.157	11			
DRB1*09	0.071	5			
DRB1*03	0.057	4			
DRB1*15	0.057	4			
DRB1*16	0.043	3			
DRB1*12	0.029	2			
DRB1*11	0.014	1			

Total number of 2-digit typed alleles at each locus: HLA-DRB1 = 70, HLA-DQB1 = 94.
 Distinct 2-digit alleles at each locus: HLA-DRB1 = 9, HLA-DQB1 = 2.

Nucleotide sequence

DQB1*04:02:02

AGGATTTTCGTGTTCCAGTTTAAGGGCATGTGCTACTTCACCAACGGGA
 CAGAGCGCGTGCGGGGTGTG**A**CCAGATACATCTATAACCGAGAGGAG
 TACGCGCGCTTCGACAGCGACGTGGGGGTGTATCGGGCGGTGACGC
 CGCTGGGGCGGCTTGACGCCGAGTACTGGAATAGCCAGAAGGACATC
 CTGGAGGAGGACCGGGCGTTCGGTGGACACCGTATGCAGACACAATA
 CCAGTTGGAGCTCCGCACGACCTTGCAGCGGCGAG

DQB1*04:02:02 (1m 69 A >G)

AGGATTTTCGTGTTCCAGTTTAAGGGCATGTGCTACTTCACCAACGGGA
 CAGAGCGCGTGCGGGGTGTG**G**CCAGATACATCTATAACCGAGAGGAG
 TACGCGCGCTTCGACAGCGACGTGGGGGTGTATCGGGCGGTGACGC
 CGCTGGGGCGGCTTGACGCCGAGTACTGGAATAGCCAGAAGGACATC
 CTGGAGGAGGACCGGGCGTTCGGTGGACACCGTATGCAGACACAATA
 CCAGTTGGAGCTCCGCACGACCTTGCAGCGGCGAG

Protein sequence

DQB1*04:02:02

DFVFQFKGMCYFTNGTERVRGV**T**RYIYNREEYARFSDVGVYRAVTPLG
 RLDAEYWNSQKDILEEDRASVDTVCRHNYQLELRITTLQRR

DQB1*04:02:02 (1m 69 A >G)

DFVFQFKGMCYFTNGTERVRGV**A**RYIYNREEYARFSDVGVYRAVTPLG
 RLDAEYWNSQKDILEEDRASVDTVCRHNYQLELRITTLQRR

ABS (antigen binding sites)

NNNANANANNNNNNNNNNNNN**A**NANANNNNAANNNNNNNNNANNNNN
 NNNANNNNAANNNANNNANAANNANNNANNAANNAANAANNNNN

Figure S1 | Nucleotide and proteins sequences of the new variant DQB1*04:02:02 (1m 69 A >G) described in the ancient samples, reported together with the ABS (antigen binding sites) as defined by Brown et al. (1993).

Conclusion and perspectives

In this thesis I have explored the genetic diversity of the immunologically important major histocompatibility complex (MHC), known in humans as human leukocyte antigen complex (HLA), and its evolutionary significance, in historical and present-day human populations. This work, conducted in cooperation with several colleagues, provides new insights into the nature of the historical coevolution between humans and their pathogens, disentangling, on the one hand, the parallel mechanisms through which pathogen-mediated selection affected genetic diversity at the HLA, and exploring, on the other hand, the functional consequences of such adaptive immunogenetic variation for present-day populations.

According to the divergent allele advantage hypothesis, individuals with two divergent HLA alleles are provided with a better coverage of the spectrum of pathogenic peptides bound by HLA molecules. This functional advantage reduces the “immune response void”, as defined by Wakeland et al. (1990), the global immune response defects resulting from a lack of antigen binding capacity by HLA molecules for antigens of a pathogen. It therefore follows that divergent HLA allelic lineages might be maintained over long evolutionary time in natural populations, to facilitate immunity against the constant and simultaneous barrage imposed by many different pathogens. Overall, the divergent allele advantage can be considered a quantitative mechanism of pathogen-mediated selection, which does not respond to specific pathogen species or strains but rather implies an effective immunity against a wide range of parasites. In the first chapter we characterized, using a computational approach, the functional advantage of HLA sequence divergence by describing how enhanced genetic diversity between alleles in a heterozygous MHC genotype increases the range of potential MHC-presented peptides. When focusing on different groups of pathogens, i.e. intracellular and extracellular, we observed that sequence divergence at some HLA loci (HLA-B and HLA-C, and potentially HLA-DQB1) has evolved specifically in response to selection by pathogens that are processed through the corresponding major protein degradation and antigen-presentation pathways. These results co-occurred with distinct protein compositions between intra- and extracellular pathogens, possibly suggesting adaptation of MHC I molecules to present specifically intracellular peptides. In the final part of this first chapter, we observed for specific HLA loci a significant correlation between an allele’s population frequency and its average pairwise sequence divergence, indicating that

Conclusion and perspectives

heterozygotes with divergent alleles are more likely to be maintained in some human populations over heterozygotes with genetically closer alleles. Overall, we have established signatures of historical and potentially still ongoing selection for functional divergence in HLA allele pools, supporting the hypothesis that pathogen selection has led to the persistence of divergent HLA allelic lineages.

Past pathogen selection is suspected to have functional consequences detectable in present-day human populations. Driven by this notion, and intrigued by the potential implication of the divergent allele advantage in different biomedical scenarios, we explored the effect of allelic divergence at HLA class I genes on HIV disease progression as well as on cancer immunotherapy response. We explored a large cohort of 6,311 HIV-1–infected individuals, for which HLA genotypes and the set-point viral load (a variable associated with HIV-1 control and disease progression) were available (Annex I). We detected a lower level of viral load in heterozygous individuals compared to homozygous at the HLA-B and HLA-C loci, and a negative correlation between pairwise allele divergence and viral load across individuals at the HLA-B locus, confirming the effects of both heterozygosity and allele divergence on AIDS outcomes. Using computational prediction of HLA-presented antigenic HIV peptides we observed that both heterozygosity and higher sequence divergence at HLA-B genotypes allow for the presentation of a broader array of HIV-1 peptides, suggesting that heterozygote advantage at HLA-B is in part mediated by quantitative peptide presentation. Notably, this study enabled for the first time to test whether the evolutionarily advantageous sequence divergence between HLA alleles could have an effect on an empirical disease phenotype (the set point viral load in this case) in humans, thus to explore the functional basis of the divergent allele advantage hypothesis. Despite both mechanisms of pathogen selection, the heterozygote advantage and the divergent allele advantage, are thought to confer benefit mainly in the case of selection from multiple pathogens simultaneously, we here observed that also single pathogen species can lead to selection for both HLA heterozygosity and excessive allele divergence. These results also suggested that the evolutionary signatures of particular pathogens are recognizable on the HLA genes; this might be particularly true in case of pathogen species which are likely to have caused high levels of mortality throughout the human history (Penman and Gupta 2018). We next explored the effect of sequence divergence at HLA class I genes on cancer immunotherapy efficacy (Annex II). Studying a dataset of patients with

Conclusion and perspectives

melanoma and non-small cell lung cancer, treated with immune checkpoint inhibitors (ICI), for which clinical response information as well as cancer exome sequencing data were available, we observed that HLA sequence divergence was a strong determinant of survival after ICI. Patients with high mean sequence divergence (i.e. mean of the three pairwise divergences at the three class I loci: HLA-A, -B, and -C) responded significantly better to ICI than patients with low sequence divergence. We proposed a link between the divergent allele advantage and immunotherapy efficacy: highly divergent HLA-I genotypes can present a higher diversity of the neopeptide repertoire, influencing T cell clonal expansion, thus facilitating tumor control during immunotherapy. Overall, these results support the notion that immune checkpoint inhibitor (ICI) therapy relies on the evolved efficiency of HLA-mediated immunity. These two projects highlighted how the study of evolutionary processes can have important implication in the field of human medicine and, in general, can serve to answer unsolved questions in the field of evolutionary medicine. Future studies might indeed explore the effect of sequence divergence at HLA genes on other complex genetic disorders. Increased diversity at MHC genes entails a larger repertoire of MHC-presented peptides. On the one hand, this might be advantageous in conferring resistance against pathogen infections, because more pathogen peptides are exposed to the immune system, but at the same time, it might cause a higher presentation of self-peptides, increasing the possibility of autoreactive T cells activation and thus autoimmune diseases. As results of these two opposing selective forces affecting MHC genes, it has been empirically shown, in non-model species, that intermediate rather than maximal genetic diversity confers the highest level of fitness (Wegner et al. 2003; Woelfing et al. 2009). In this light, it would be interesting to explore how sequence divergence at HLA genes affects empirical phenotypes in the case of autoimmune diseases.

Selection at HLA genes can occur through additional mechanisms apart from the heterozygote advantage and the divergent allele advantage. Specific HLA variants can be selected on short time scale by the transient selection pressure imposed by specific pathogens. Under fluctuating selection, host exposures and thus selection at HLA can vary significantly across space and time, depending on pathogens' spatio-temporal variations. Since empirical validation of such modes of selection requires the examination of population dynamics and changes in allele frequencies over time, finding empirical evidence for transient and fluctuating selection on HLA genes has proved

elusive. In this light, the study of HLA variability in historical human populations during specific epidemiological events, thus using ancient DNA (aDNA) samples, is of strong interest. However, the exceptional genetic variability observed at HLA genes together with the high level of degradation and the low amount of DNA often found in aDNA samples make HLA genotyping using ancient DNA extremely challenging. In the second chapter we described a new HLA genotyping pipeline ('aHLA-Seq') optimized for low coverage shotgun sequence data and showed its accuracy for modern and ancient DNA samples. As described in the last two chapters, the newly developed approach can be applied to study the evolution of human resistance or susceptibility to pathogens in historical populations, but also to explore HLA allele frequency changes through time, when temporal sample series are available, thus increasing our knowledge of HLA genetic variation over time in humans. In the third chapter, we revealed a significant association between the allele DRB1*15:01 and medieval samples collected from a leprosarium in Denmark (1270–1550 AD) that tested positive for the *M. leprae* bacterium. These results suggested that the allele DRB1*15:01, known to be a risk factor for leprosy in modern populations, predisposed also medieval Europeans to leprosy. Further, a limited HLA-presentation capacity of the DRB1*15:01 allele for *M. leprae* antigens was predicted performing computational antigen-binding prediction on *M. leprae* peptides. To our knowledge this is the first association study carried out using aDNA to date, which showed the potential of our approach in characterizing the impact of disease in historical populations. In the fourth chapter, we investigated the polymorphisms of HLA class II loci (HLA-DRB1 and HLA-DQB1) in ancient Native American populations. With the aim of exploring potential HLA allele frequency shifts from pre- to post-European contact populations, we described the HLA molecular profile of pre-European contact archaeological samples as well as present-day residents of the town of Xaltocan in central Mexico. We next explored the HLA variability at a broader geographical and temporal scale, defining the HLA molecular profiles in a large dataset of available ancient whole genome data of samples collected from different sites across the American continent. Interestingly, our analyses revealed allelic lineages quite widespread in ancient pre-European Native American populations that are no longer found in present-day Native American populations. Further, we observed also unknown alleles present in ancient Native Americans that are no longer found in modern-day humans. It is tempting to speculate that such allelic lineages might have conferred susceptibility to one or more of the newly introduced European-borne pathogens, and

Conclusion and perspectives

thus might have been removed under the effect of negative selection. However, the observed change in allele frequency might also result from stochastic factors associated with the complex demographic history of Native Americans and further research is necessary to unravel the differential contribution of neutral processes and potential pathogen-mediated selection.

Overall, the results presented in this thesis provide new insight into the mechanisms of pathogen-mediated selection and their role in maintaining the genetic diversity at the major histocompatibility complex (MHC) genes in humans. Our findings revealed signatures of continuous and directional selection on HLA genes in modern human populations with evidence for the divergent allele advantage. Furthermore, the results here exposed have shed light on the functional consequences of adaptive immunogenetic variation in present-day human populations, highlighting how the study of past and ongoing pathogen selection might have important implications in the field of human medicine. Finally, thanks to the development of a new accurate HLA genotyping tool optimized for low coverage shotgun sequence data and thanks to its successful application in two different studies here exposed, this work gives directions for future investigations of HLA genetic variation through human history.

References

- Aeschlimann PB, Haberli MA, Reusch TBH, Boehm T, Milinski M. 2003. Female sticklebacks *Gasterosteus aculeatus* use self-reference to optimize MHC allele number during mate selection. *Behavioral Ecology and Sociobiology* 54:119-126.
- Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL. 2008. Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). *Molecular Ecology* 17:2652-2665.
- Andam CP, Worby CJ, Chang Q, Campana MG. 2016. Microbial Genomics of Ancient Plagues and Outbreaks. *Trends in Microbiology* 24:978-990.
- Apanius V, Penn D, Slev PR, Ruff LR, Potts WK. 1997. The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology* 17:179-224.
- Armuzzi A, Jewell DP, Sato H, Crawshaw J, Welsh KI, Ling K-L, Mulcahy-Hawes K, Barnardo M, Neville M, Bunce M, et al. 2003. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Human Molecular Genetics* 12:647-656.
- Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* 526:68-74.
- Babik W, Pabijan M, Radwan J. 2008. Contrasting patterns of variation in MHC loci in the Alpine newt. *Molecular Ecology* 17:2339-2355.
- Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics* 11:17-30.
- Bauer DC, Zadoorian A, Wilson LOW, Thorne NP. 2018. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Briefings in Bioinformatics* 19:179-187.
- Belich MP, Madrigal JA, Hildebrand WH, Zemmour J, Williams RC, Luz R, Petzl-Erler ML, Parham P. 1992. Unusual HLA-B alleles in two tribes of Brazilian Indians. *Nature* 357:326-329.
- Berlinguer G. 1992. The interchange of disease and health between the Old and New Worlds. *American Journal of Public Health* 82:1407-1413.
- Black F. 1992. Why did they die? *Science* 258:1739-1740.
- Black FL. 1994. An explanation of high death rates among New World peoples when in contact with Old World diseases. *Perspectives in Biology and Medicine* 37:292-307.
- Black FL. 1975. Infectious diseases in primitive societies. *Science* 187:515-518.
- Black FL, Hedrick PW. 1997. Strong balancing selection at HLA loci: Evidence from segregation in South Amerindian families. *Proceedings of the National Academy of Sciences of the United States of America* 94:12452-12456.

- Bolnick DA, Bonine HM, Mata-Miguez J, Kemp BM, Snow MH, LeBlanc SA. 2012. Nondestructive sampling of human skeletal remains yields ancient nuclear and mitochondrial DNA. *American Journal of Physical Anthropology* 147:293-300.
- Bolnick DA, Raff JA, Springs LC, Reynolds AW, Miró-Herrans AT. 2016. Native American Genomics and Population Histories. *Annual Review of Anthropology* 45:319-340.
- Bolnick DI, Stutz WE. 2017. Frequency dependence limits divergent evolution by favouring rare immigrants over residents. *Nature advance online publication*.
- Bonneaud C, Perez-Tris J, Federici P, Chastel O, Sorci G. 2006. Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution* 60:383-389.
- Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajković D, Kučan Ž, et al. 2009. Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* 325:318-321.
- Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences* 104:14616-14621.
- Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Paabo S. 2010. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research* 38:e87.
- Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. 2007. 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 Research* 35:5717-5728.
- Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC. 1993. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364:33-39.
- Buhler S, Nunes JM, Sanchez-Mazas A. 2016. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics Online early*.
- Buhler S, Sanchez-Mazas A. 2011. HLA DNA Sequence Variation among Human Populations: Molecular Signatures of Demographic and Selective Events. *PLoS ONE* 6:e14643.
- Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PLF, Xuan Z, et al. 2010. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* 328:723-725.

- Cadauid LF, Watkins DI. 1997. Heirs of the jaguar and the anaconda: HLA, conquest and disease in the indigenous populations of the Americas. *Tissue Antigens* 50:209-218.
- Cappellini E, Prohaska A, Racimo F, Welker F, Pedersen MW, Allentoft ME, de Barros Damgaard P, Gutenbrunner P, Dunne J, Hammann S, et al. 2018. Ancient Biomolecules and Evolutionary Inference. *Annual Review of Biochemistry* 87:1029-1060.
- Carapito R, Radosavljevic M, Bahram S. 2016. Next-Generation Sequencing of the HLA locus: Methods and impacts on HLA typing, population genetics and disease association studies. *Human Immunology* 77:1016-1023.
- Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, et al. 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American Journal of Human Genetics* 93:852-864.
- Carrington M, Claiborne Stephens J, Klitz W, Begovich AB, Erlich HA, Mann D. 1994. Major histocompatibility complex class II haplotypes and linkage disequilibrium values observed in the CEPH families. *Human Immunology* 41:234-240.
- Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ. 1999. HLA and HIV-1: Heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283:1748-1752.
- Carroll L. 1900. *Through the looking-glass and what Alice found there*: Chicago : W.B. Conkey Co., [1900] ©1900.
- Chaix RI, Cao C, Donnelly P. 2008. Is mate choice in humans MHC-dependent? *PLoS Genetics* 4:e1000184.
- Chen D, Gaborieau V, Zhao Y, Chabrier A, Wang H, Waterboer T, Zaridze D, Lissowska J, Rudnai P, Fabianova E, et al. 2015. A systematic investigation of the contribution of genetic variation within the MHC region to HPV seropositivity. *Human Molecular Genetics* 24:2681-2688.
- Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, Kuo F, Kendall SM, Requena D, Riaz N, et al. 2018. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* 359:582-587.
- Clarke B, Kirby DRS. 1966. Maintenance of histocompatibility polymorphisms. *Nature* 211:999-1000.
- Cramer H. 1946. *Mathematical Models of Statistics*: New Jersey: Princeton University Press.
- Cruz-Davalos DI, Nieves-Colon MA, Sockell A, Poznik GD, Schroeder H, Stone AC, Bustamante CD, Malaspinas AS, Avila-Arcos MC. 2018. In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genomics* 19:608.

- Currat M, Poloni E, Sanchez-Mazas A. 2010. Human genetic differentiation across the Strait of Gibraltar. *BMC Evolutionary Biology* 10:237.
- Danilova N. 2012. The evolution of adaptive immunity. *Advances in Experimental Medicine and Biology* 738:218-235.
- de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, et al. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics* 38:1166-1172.
- Dean M, Carrington M, O'Brien SJ. 2002. Balanced polymorphism selected by genetic versus infectious human disease. *Annual Review of Genomics and Human Genetics* 3:263-292.
- Dempster A, Laird N, Rubin D. 1977. Maximum Likelihood From Incomplete Data Via The EM algorithm.
- Di D, Sanchez-Mazas A. 2011. Challenging views on the peopling history of East Asia: the story according to HLA markers. *American Journal of Physical Anthropology* 145:81-96.
- Doherty PC, Zinkernagel RM. 1975a. A biological role for the major histocompatibility antigens. *Lancet* 1:1406-1409.
- Doherty PC, Zinkernagel RM. 1975b. A biological role for the major histocompatibility antigens. *The Lancet* 305:1406-1409.
- Duggal P, Thio CL, Wojcik GL, Goedert JJ, Mangia A, Latanich R, Kim AY, Lauer GM, Chung RT, Peters MG, et al. 2013. Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. *Annals of Internal Medicine* 158:235-245.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792-1797.
- Eizaguirre C, Lenz TL, Kalbe M, Milinski M. 2012. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nature Communications* 3:621.
- Eizaguirre C, Lenz TL, Sommerfeld R, Harrod C, Kalbe M, Milinski M. 2011. Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. *Evolutionary Ecology* 25:605-622.
- Eklblom R, Saether SA, Jacobsson PAR, Fiske P, Sahlman T, Grahn M, Kalas JA, Hoglund J. 2007. Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Molecular Ecology* 16:1439-1451.
- Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN. 2014. Ancient whole genome enrichment using baits built from modern DNA. *Mol Biol Evol* 31:1292-1294.
- Erlich H. 2012. HLA DNA typing: past, present, and future. *Tissue Antigens* 80:1-11.

- Escamilla-Tilch M, Torres-Carrillo NM, Payan RR, Aguilar-Medina M, Salazar MI, Fafutis-Morris M, Arenas-Guzman R, Estrada-Parra S, Estrada-Garcia I, Granados J. 2013. Association of genetic polymorphism of HLA-DRB1 antigens with the susceptibility to lepromatous leprosy. *Biomedical reports* 1:945-949.
- Evans ML, Neff BD. 2009. Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* 18:4716-4729.
- Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921-927.
- Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C, Castagna A, Cossarizza A, et al. 2007. A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science* 317:944-947.
- Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Pääbo S. 2013. DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences* 110:2223-2227.
- Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettlla A, Pattini L, Nielsen R. 2011. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLoS Genetics* 7:e1002355.
- Furlong R, Yang Z. 2008. Diversifying and purifying selection in the peptide binding region of DRB in mammals. *Journal of Molecular Evolution* 66:384-394.
- Gelabert P, Olalde I, de-Dios T, Civit S, Lalueza-Fox C. 2017. Malaria was a weak selective force in ancient Europeans. *Scientific Reports* 7:1377.
- Gilbert MT, Bandelt HJ, Hofreiter M, Barnes I. 2005. Assessing ancient DNA studies. *Trends in Ecology & Evolution* 20:541-544.
- González-Galarza Faviel F, Takeshita Louise YC, Santos Eduardo JM, Kempson F, Maia Maria Helena T, Silva Andrea Luciana Soares d, Silva André Luiz Teles e, Ghattaoraya Gurpreet S, Alfirevic A, Jones Andrew R, et al. 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research* 43:D784-D788.
- Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, D. Rioux J, Hauser S, Oksenberg J. 2014. HLA Diversity in the 1000 Genomes Dataset. *PLoS ONE* 9:e97282.
- Grimholt U, Larsen S, Nordmo R, Midtlyng P, Kjoeglum S, Storset A, Saebø S, Stet RJM. 2003. MHC polymorphism and disease resistance in Atlantic salmon (*Salmo salar*); facing pathogens with single expressed major histocompatibility class I and class II loci. *Immunogenetics* 55:210-219.
- Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207-211.

- Haldane JBS. 1932. The causes of evolution. London; New York: Longmans, Green and Co.
- Hall T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.
- Hedrick PW. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117:331-341.
- Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM. 1991. Common west African HLA antigens are associated with protection from severe malaria. *Nature* 352:595-600.
- Hill AVS. 1991. HLA Associations with Malaria in Africa: Some Implications for MHC Evolution. In: Klein J, Klein D, editors. *Molecular Evolution of the Major Histocompatibility Complex*. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 403-420.
- Hofmanova Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Diez-Del-Molino D, van Dorp L, Lopez S, Kousathanas A, Link V, et al. 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc Natl Acad Sci U S A* 113:6886-6891.
- Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* 29:4793-4799.
- Hollenbach JA, Mack SJ, Gourraud PA, Single RM, Maiers M, Middleton D, Thomson G, Marsh SG, Varney MD. 2011. A community standard for immunogenomic data reporting and analysis: proposal for a Strengthening the Reporting of Immunogenomic Studies statement. *Tissue Antigens* 78:333-344.
- Hosomichi K, Shiina T, Tajima A, Inoue I. 2015. The impact of next-generation sequencing technologies on HLA research. *Journal of Human Genetics* 60:665-673.
- Hughes AL, Nei M. 1989. Nucleotide substitution at major histocompatibility complex class-II loci - evidence for overdominant selection. *Proceedings of the National Academy of Sciences of the United States of America* 86:958-962.
- Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167-170.
- Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annual Review of Genetics* 32:415.
- International HIV Controllers Study, Pereyra F, Jia X, McLaren P, Telenti A, de B, PI, Walker B, Ripke S, Brumme C, Pulit S, et al. 2010. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 330:1551-1557.

- Jensen PE. 2007. Recent advances in antigen processing and presentation. *Nat Immunol* 8:1041-1048.
- Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29:1682-1684.
- Kamatani Y, Wattanapokayakit S, Ochi H, Kawaguchi T, Takahashi A, Hosono N, Kubo M, Tsunoda T, Kamatani N, Kumada H, et al. 2009. A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet* 41:591-595.
- Klein J. 1987. Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. *Human Immunology* 19:155-162.
- Klein J, Figueroa F. 1986. The evolution of class I MHC genes. *Immunology Today* 7:41-44.
- Klein J, Sato A, Nikolaidis N. 2007. MHC, TSP, and the origin of species: From immunogenetics to evolutionary genetics. *Annual Review of Genetics* 41:281-304.
- Koch A, Brierley C, Maslin MM, Lewis SL. 2019. Earth system impacts of the European arrival and Great Dying in the Americas after 1492. *Quaternary Science Reviews* 207:13-36.
- Krause-Kyora B, Nutsua M, Boehme L, Pierini F, Pedersen DD, Kornell S-C, Drichel D, Bonazzi M, Möbus L, Tarp P, et al. 2018. Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nature Communications* 9:1569.
- Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Pääbo S. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 464:894.
- Kubinak JL, Ruff JS, Hyzer CW, Slev PR, Potts WK. 2012. Experimental viral evolution to specific host MHC genotypes reveals fitness and virulence trade-offs in alternative MHC types. *Proceedings of the National Academy of Sciences* 109:3422-3427.
- Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, Martin MP, Hunt P, Deeks SG, Telenti A, et al. 2011. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature* 472:495-498.
- Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. 2007. PyPop update--a software pipeline for large-scale multilocus population genomics. *Tissue Antigens* 69 Suppl 1:192-197.
- Landry C, Garant D, Duchesne P, Bernatchez L. 2001. 'Good genes as heterozygosity': The major histocompatibility complex and mate choice in Atlantic salmon (*Salmo*

- salar*). Proceedings of the Royal Society of London B: Biological Sciences 268:1279-1285.
- Langefors A, Lohm J, von Schantz T, Grahn M. 2000. Screening of Mhc variation in Atlantic salmon (*Salmo salar*): a comparison of restriction fragment length polymorphism (RFLP), denaturing gradient gel electrophoresis (DGGE) and sequencing. *Molecular Ecology* 9:215-219.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357-359.
- Larsen CS. 1994. In the wake of Columbus: Native population biology in the postcontact Americas. *American Journal of Physical Anthropology* 37:109-154.
- Lau Q, Yasukochi Y, Satta Y. 2015. A limit to the divergent allele advantage model supported by variable pathogen recognition across HLA-DRB1 allele lineages. *Tissue Antigens* 86:343-352.
- Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. *Nature* 536:419-424.
- Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011. The sequence read archive. *Nucleic Acids Research* 39:D19-D21.
- Lenz TL. 2011. Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution* 65:2380-2390.
- Lenz TL, Mueller B, Trillmich F, Wolf JB. 2013. Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. *Proc Biol Sci* 280:20130714.
- Lenz TL, Spirin V, Jordan DM, Sunyaev SR. 2016. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. *Molecular Biology and Evolution* 33:2555-2564.
- Lenz TL, Wells K, Pfeiffer M, Sommer S. 2009. Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the Long-tailed giant rat (*Leopoldamys sabanus*). *BMC Evolutionary Biology* 9:269.
- Li WH, Sadler LA. 1991. Low nucleotide diversity in man. *Genetics* 129:513-523.
- Lindenau JD, Salzano FM, Guimaraes LS, Callegari-Jacques SM, Hurtado AM, Hill KR, Petzl-Erler ML, Tsuneto LT, Hutz MH. 2013. Distribution patterns of variability for 18 immune system genes in Amerindians--relationship with history and epidemiology. *Tissue Antigens* 82:177-185.
- Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, Mitchell J, Worl R, Dixon EJ, Fifield TE, et al. 2017. Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *Proceedings of the National Academy of Sciences* 114:4093-4098.

- Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, Malhi RS. 2016. A time transect of exomes from a Native American population before and after European contact. *Nature Communications* 7:13175.
- Livi-Bacci M. 2006. The Depopulation of Hispanic America after the Conquest. *Population and Development Review* 32:199-232.
- Loiseau C, Richard M, Garnier S, Chastel O, Julliard R, Zoorob R, Sorci G. 2009. Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Molecular Ecology* 18:1331-1340.
- Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE, Kempenich JH, Reed EF, et al. 2013. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* 81:194-203.
- Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* 18:659.
- Maricic T, Whitten M, Pääbo S. 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* 5:e14004.
- Mata-Míguez J. 2016. Assessing the Demographic and Genetic Impact of State Expansion in Pre-Hispanic and Colonial Mexico. [Ph.D. dissertation]. [Austin (US)]: University of Texas at Austin.
- Mata-Miguez J, Overholtzer L, Rodriguez-Alegria E, Kemp BM, Bolnick DA. 2012. The genetic impact of Aztec imperialism: ancient mitochondrial DNA evidence from Xaltocan, Mexico. *American Journal of Physical Anthropology* 149:504-516.
- Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499-503.
- Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. 2017. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol* 18:76.
- McDevitt HO, Tyman ML. 1968. Genetic control of the antibody response in inbred mice. Transfer of response by spleen cells and linkage to the major histocompatibility (H-2) locus. *Journal of Experimental Medicine* 128:1-11.
- McLaren PJ, Coulonges C, Bartha I, Lenz TL, Deutsch AJ, Bashirova A, Buchbinder S, Carrington M, Cossarizza A, Dalmau J, et al. 2015. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proceedings of the National Academy of Sciences of the United States of America* 112:14658-14663.
- McNeill WH. 1976. *Plagues and Peoples*. Garden City, N.Y: Anchor Press.
- Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. 2006. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics* 173:2121-2142.

- Meyer D, VR CA, Bitarello BD, DY CB, Nunes K. 2018. A genomic perspective on HLA evolution. *Immunogenetics* 70:5-27.
- Milinski M. 2003. The function of mate choice in sticklebacks: optimizing Mhc genetics. *Journal of Fish Biology* 63:1-16.
- Milinski M. 2006. The major histocompatibility complex, sexual selection, and mate choice. *Annual Review of Ecology Evolution and Systematics* 37:159-186.
- Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. 2002. Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science* 296:1439-1443.
- Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspina A-S, Sikora M, et al. 2018a. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* 553:203.
- Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, et al. 2018b. Early human dispersals within the Americas. *Science* 362:2621.
- Murphy K, Weaver C. 2017. *Janeway's immunobiology*.
- Neefjes J, Jongsma MLM, Paul P, Bakke O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology* 11:823-836.
- Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. *Nature* 541:302.
- Niskanen AK, Kennedy LJ, Ruokonen M, Kojola I, Lohi H, Isomursu M, Jansson E, Pyhäjärvi T, Aspi J. 2014. Balancing selection and heterozygote advantage in major histocompatibility complex loci of the bottlenecked Finnish wolf population. *Molecular Ecology* 23:875-889.
- O'Connor SL, Lhost JJ, Becker EA, Detmer AM, Johnson RC, Macnair CE, Wiseman RW, Karl JA, Greene JM, Burwitz BJ, et al. 2010. MHC heterozygote advantage in simian immunodeficiency virus-infected Mauritian cynomolgus macaques. *Sci Transl Med* 2:22ra18.
- O'Fallon BD, Fehren-Schmitz L. 2011. Native Americans experienced a strong population bottleneck coincident with European contact. *Proc Natl Acad Sci U S A* 108:20444-20448.
- Olalde I, Allentoft ME, Sanchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, Prado-Martinez J, Rodriguez JA, Rasmussen S, Quilez J, et al. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* 507:225-228.
- Organization WH. 2016. *Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2015*. Geneva, World Health Organization; 2016.

- Orlando L, Gilbert MT, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics* 16:395-408.
- Parham P. 1988. Function and polymorphism of human leukocyte antigen-A,B,C molecules. *The American Journal of Medicine* 85:2-5.
- Parham P, Arnett KL, Adams EJ, Little AM, Tees K, Barber LD, Marsh SG, Ohta T, Markow T, Petzl-Erler ML. 1997. Episodic evolution and turnover of HLA-B in the indigenous human populations of the Americas. *Tissue Antigens* 50:219-232.
- Parham P, Ohta T. 1996. Population biology of antigen presentation by MHC class I molecules. *Science* 272:67-74.
- Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K. 2016. EAGER: efficient ancient genome reconstruction. *Genome Biology* 17:60.
- Penman BS, Ashby B, Buckee CO, Gupta S. 2013. Pathogen selection drives nonoverlapping associations between HLA loci. *Proceedings of the National Academy of Sciences* 110:19645-19650.
- Penman BS, Gupta S. 2018. Detecting signatures of past pathogen selection on human HLA loci: are there needles in the haystack? *Parasitology* 145:731-739.
- Penn DJ, Damjanovich K, Potts WK. 2002. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America* 99:11260-11264.
- Penn DJ, Potts WK. 1999. The evolution of mating preferences and major histocompatibility complex genes. *American Naturalist* 153:145-164.
- Phillips KP, Cable J, Mohammed RS, Herdegen-Radwan M, Raubic J, Przesmycka KJ, van Oosterhout C, Radwan J. 2018. Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proceedings of the National Academy of Sciences* 115:1552-1557.
- Piertney SB, Oliver MK. 2006. The evolutionary ecology of the major histocompatibility complex. *Heredity* 96:7-21.
- Potts WK, Wakeland EK. 1990. Evolution of diversity at the major histocompatibility complex. *Trends in Ecology & Evolution* 5:181-187.
- Prugnolle F, Manica A, Balloux F. 2005a. Geography predicts neutral genetic diversity of human populations. *Current Biology* 15:R159-160.
- Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. 2005b. Pathogen-driven selection and worldwide HLA class I diversity. *Current Biology* 15:1022-1027.
- Quintana-Murci L, Alcais A, Abel L, Casanova J-L. 2007. Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. *Nature Immunology* 8:1165-1171.

- Rao X, Hoof I, van Baarle D, Kesmir C, Textor J. 2015. HLA Preferences for Conserved Epitopes: A Potential Mechanism for Hepatitis C Clearance. *Frontiers in Immunology* 6:552.
- Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757.
- Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CPE, Ponce de León MS, Allentoft ME, Moltke I, et al. 2015. The ancestry and affiliations of Kennewick Man. *Nature* 523:455.
- Reche PA, Reinherz EL. 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* 331:623-641.
- Renaud G, Hanghøj K, Willerslev E, Orlando L. 2017. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33:577-579.
- Reynolds AW. 2018. Investigating Regional Human Population Histories in North America Using Genomics. [Ph.D. dissertation]: University of Texas at Austin.
- Reynolds AW, Mata-Míguez J, Miró-Herrans A, Briggs-Cloud M, Sylestine A, Barajas-Olmos F, Garcia-Ortiz H, Rzhetskaya M, Orozco L, Raff JA, et al. 2019. Comparing signals of natural selection between three Indigenous North American populations. *Proceedings of the National Academy of Sciences* 116:9312-9317.
- Richman AD, Herrera LG, Nash D. 2001. MHC class II beta sequence diversity in the deer mouse (*Peromyscus maniculatus*): Implications for models of balancing selection. *Molecular Ecology* 10:2765-2773.
- Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE, Parham P. 2017. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genetics* 13:e1006862.
- Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh Steven GE. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research* 43:D423-D431.
- Rodríguez-Alegría E. 2010. Incumbents and Challengers: Indigenous Politics and the Adoption of Spanish Material Culture in Colonial Xaltocan, Mexico. *Historical Archaeology* 44:51-71.
- Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. 2014. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 370.
- Rohland N, Hofreiter M. 2007. Comparison and optimization of ancient DNA extraction. *BioTechniques* 42:343-352.

- Sanchez-Mazas A, Buhler S, Nunes JM. 2013. A New HLA Map of Europe: Regional Genetic Variation and Its Implication for Peopling History, Disease-Association Studies and Tissue Transplantation. *Human Heredity* 76:162-177.
- Sanchez-Mazas A, Cerny V, Di D, Buhler S, Podgorna E, Chevallier E, Brunet L, Weber S, Kervaire B, Testi M, et al. 2017. The HLA-B landscape of Africa: Signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. *Molecular Ecology* 26:6238-6252.
- Sanchez-Mazas A, Djoulah S, Busson M, Le Monnier de Gouville I, Poirier J-C, Dehay C, Charron D, Excoffier L, Schneider S, Langaney A, et al. 2000. A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. *European Journal of Human Genetics* 8:33.
- Sanchez-Mazas A, Nunes JM. 2018. Does NGS typing highlight our understanding of HLA population diversity?: Some good reasons to say yes and a few to say be careful. *Human Immunology*.
- Satta Y, Li YJ, Takahata N. 1998. The neutral theory and natural selection in the HLA region. *Frontiers in Bioscience* 3:d459-467.
- Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Mörseburg A, Johnson JR, Potter A, et al. 2018. Ancient human parallel lineages within North America contributed to a coastal expansion. *Science* 360:1024-1027.
- Schroeder H, Sikora M, Gopalakrishnan S, Cassidy LM, Maisano Delser P, Sandoval Velasco M, Schraiber JG, Rasmussen S, Homburger JR, Avila-Arcos MC, et al. 2018. Origins and genetic legacies of the Caribbean Taino. *Proc Natl Acad Sci U S A* 115:2341-2346.
- Schwensow N, Eberle M, Sommer S. 2010. Are there ubiquitous parasite-driven major histocompatibility complex selection mechanisms in gray mouse lemurs? *International Journal of Primatology* 31:519-537.
- Shiina T, Hosomichi K, Inoko H, Kulski JK. 2009. The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of Human Genetics* 54:15-39.
- Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nature Reviews Genetics* 16:224-236.
- Slade RW, McCallum HI. 1992. Overdominant vs. frequency-dependent selection at MHC loci. *Genetics* 132:861-862.
- Slatkin M, Excoffier L. 1996. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity (Edinb)* 76 (Pt 4):377-383.
- Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G. 2008. Balancing selection and heterogeneity across the classical human

- leukocyte antigen loci: a meta-analytic review of 497 population studies. *Human Immunology* 69:443-464.
- Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences* 277:979-988.
- Spyrou MA, Bos KI, Herbig A, Krause J. 2019. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature Reviews Genetics*.
- Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV, Kristinsson KG, Gottfredsson M, Barrett JC, Gudmundsson LJ, Blondal K, Gylfason A, Gudjonsson SA, et al. 2016. HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nature Genetics* 48:318-322.
- Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. 2014. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 30:3310-3316.
- Takehima S, Matsumoto Y, Chen J, Yoshida T, Mukoyama H, Aida Y. 2008. Evidence for cattle major histocompatibility complex (BoLA) class II DQA1 gene heterozygote advantage against clinical mastitis caused by *Streptococci* and *Escherichia* species. *Tissue Antigens* 72:525-531.
- Thomas R, Thio CL, Apps R, Qi Y, Gao X, Marti D, Stein JL, Soderberg KA, Moody MA, Goedert JJ, et al. 2012. A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection. *Journal of Virology* 86:6979-6985.
- Thornton R. 1997. Aboriginal North American Population and Rates of Decline, ca. A.D. 1500-1900. *Current Anthropology* 38:310-315.
- Tishkoff SA, Verrelli BC. 2003. Patterns of Human Genetic Diversity: Implications for Human Evolutionary History and Disease. *Annual Review of Genomics and Human Genetics* 4:293-340.
- Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E, Funkhouser R, Fugate M, Theiler J, Hsu YS, et al. 2003. Advantage of rare HLA supertype in HIV disease progression. *Nature Medicine* 9:928-935.
- Trowsdale J. 2011. The MHC, disease and selection. *Immunology Letters* 137:1-8.
- Trowsdale J, Knight JC. 2013. Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics* 14:301-323.
- Van Valen L. 1973. A new evolutionary law. *Evolutionary Theory* 1:1-30.
- Wakeland EK, Boehme S, She JX, Lu CC, McIndoe RA, Cheng I, Ye Y, Potts WK. 1990. Ancestral polymorphisms of MHC class-II genes - divergent allele advantage. *Immunologic Research* 9:115-122.
- Watkins DI, McAdam SN, Liu X, Strang CR, Milford EL, Levine CG, Garber TL, Dogon AL, Lord CI, Ghim SH, et al. 1992. New recombinant HLA-B alleles in a tribe of

- South American Amerindians indicate rapid evolution of MHC class I loci. *Nature* 357:329-333.
- Wegner KM, Kalbe M, Kurtz J, Reusch TBH, Milinski M. 2003. Parasite selection for immunogenetic optimality. *Science* 301:1343.
- Wegner KM, Reusch TBH, Kalbe M. 2003. Multiple parasites are driving major histocompatibility complex polymorphism in the wild. *Journal of Evolutionary Biology* 16:224-232.
- Westerdahl H, Hansson B, Bensch S, Hasselquist D. 2004. Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? *Journal of Evolutionary Biology* 17:485-492.
- Willerslev E, Cooper A. 2005. Review Paper. Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences* 272:3-16.
- Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, Ellinghaus E, Hov JR, Sauer S, Schimpler M, Ziemann M, et al. 2015. Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Research* 43:e70.
- Woelfing B, Traulsen A, Milinski M, Boehm T. 2009. Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philosophical Transactions of the Royal Society of London B Biological Sciences* 364:117-128.
- Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics* 32:569-577.
- Yang D, Shui T, Miranda JW, Gilson DJ, Song Z, Chen J, Shi C, Zhu J, Yang J, Jing Z. 2016. Mycobacterium leprae-Infected Macrophages Preferentially Primed Regulatory T Cell Responses and Was Associated with Lepromatous Leprosy. *PLoS Negl Trop Dis* 10:e0004335.
- Zhang DF, Wang D, Li YY, Yao YG. 2016. Integrative analyses of leprosy susceptibility genes indicate a common autoimmune profile. *Journal of Dermatological Science* 82:18-27.
- Zhang F, Liu H, Chen S, Wang C, Zhu C, Zhang L, Chu T, Liu D, Yan X, Liu J. 2009. Evidence for an association of HLA-DRB1*15 and DRB1*09 with leprosy and the impact of DRB1*09 on disease onset in a Chinese Han population. *BMC Medical Genetics* 10:133.
- Zhang FR, Huang W, Chen SM, Sun LD, Liu H, Li Y, Cui Y, Yan XX, Yang HT, Yang RD, et al. 2009. Genomewide association study of leprosy. *New England Journal of Medicine* 361:2609-2618.
- Zhu M, Dai J, Wang C, Wang Y, Qin N, Ma H, Song C, Zhai X, Yang Y, Liu J, et al. 2016. Fine mapping the MHC region identified four independent variants modifying susceptibility to chronic hepatitis B in Han Chinese. *Human Molecular Genetics* 25:1225-1232.

Annex I

HLA heterozygote advantage against HIV-1 is driven by quantitative and qualitative differences in allele-specific peptide presentation

Jatin Arora¹, Federica Pierini¹, Paul J. McLaren^{2, 3}, Mary Carrington^{4, 5}, Jacques Fellay^{6, 7, 8} & Tobias L. Lenz¹

¹Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany, ²JC Wilt Infectious Diseases Research Center, National HIV and Retrovirology Laboratory, Public Health Agency of Canada, R3E 0W3, Winnipeg, Canada, ³Department of Medical Microbiology and Infectious Diseases, University of Manitoba, R3E 0J9, Winnipeg Canada, ⁴Cancer and Inflammation Program, Leidos Biomedical Research, Frederick National Laboratory, Frederick, MD 21702, USA, ⁵Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02139-3583, USA, ⁶Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, ⁷Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, ⁸Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, 1011 Lausanne, Switzerland

Unpublished manuscript

Abstract

Pathogen-mediated balancing selection is regarded as a key driver of host immunogenetic diversity. A hallmark for balancing selection in humans is the heterozygote advantage at genes of the human leukocyte antigen (HLA), resulting in improved HIV-1 control. However, the actual mechanism of the observed heterozygote advantage is still elusive. HLA heterozygotes may present a broader array of antigenic viral peptides to immune cells, possibly resulting in a more efficient cytotoxic T cell response. Alternatively, heterozygosity may simply increase the chance to carry the most protective HLA alleles, as individual HLA alleles are known to differ substantially in their association with HIV-1 control. Here we used data from 6,311 HIV-1-infected individuals to explore the relative contribution of quantitative and qualitative aspects of peptide presentation in HLA heterozygote advantage against HIV. Screening the entire HIV-1 proteome, we observed that HLA class I alleles bound a broader array of HIV-1 peptides in heterozygous individuals. In addition, viral load was negatively correlated with the breadth of the HIV-1 peptide repertoire bound by an individual's HLA variants, particularly at HLA-B. This suggests that heterozygote advantage at HLA-B is at least in part mediated by quantitative peptide presentation. We also observed higher HIV-1 sequence diversity in HLA-B heterozygous individuals, suggesting stronger evolutionary pressure from HLA heterozygosity. However, HLA heterozygotes were also more likely to carry certain protective HLA alleles, including the highly protective HLA-B*57:01 variant, indicating that HLA heterozygote advantage ultimately results from a combination of quantitative and qualitative effects in antigen presentation.

Introduction

Human Leucocyte Antigen (HLA) genes, located in the Major Histocompatibility Complex (MHC) region on chromosome 6, play a central role in immune-recognition of pathogens. They encode cell-surface molecules that present intra- and extra-cellular peptides to T-cells, which, upon recognizing them as non-self, initiate specific immune responses (Neefjes et al. 2011). Each HLA molecule binds a specific peptide-repertoire which is largely defined by the amino-acid composition of its peptide-binding groove (Rao et al. 2011; Neefjes & Ovaa 2013). Due to their key role in adaptive immunity and unparalleled allelic diversity within and across vertebrate species, MHC genes have become a paradigm for studying the effect of genetic diversity on immuno-competence and fitness (Nei & Hughes 1991; Sommer 2005; Trowsdale & Knight 2013). The dynamic action of pathogen-mediated balancing selection is widely regarded as a key driver of MHC diversity (Solberg et al. 2008; Hedrick & Thomson 1983; Hughes & Yeager 1998; Apanius et al. 1997; Lenz 2018). Three main mechanisms of pathogen-mediated balancing selection have been proposed: *heterozygote advantage*, *rare-allele advantage*, and *fluctuating selection*, all of which have received empirical support (Sommer 2005; Spurgin & Richardson 2010; Eizaguirre & Lenz 2010; Lenz 2018). It is also largely established that these mechanisms are not mutually exclusive and likely act in parallel to shape the MHC allele pool of a population. However, the relative contribution of each of these mechanisms is still debated, and may indeed depend on the specific conditions of a given population or species (Stutz & Bolnick 2017; Ejsmond & Radwan 2015).

First proposed by Doherty & Zinkernagel (Doherty & Zinkernagel 1975), the heterozygote advantage hypothesis assumes that heterozygous MHC genotypes confer a higher probability of triggering a specific immune response upon infection. This would result in enhanced pathogen resistance for MHC heterozygous individuals, compared to MHC homozygotes, promoting the persistence of different MHC alleles in the population (Penn et al. 2002; Hughes & Yeager 1998). One possible explanation for heterozygote advantage is the presumed ability of HLA heterozygous individuals to present a broader array of pathogen-derived peptides, thus increasing the probability of inducing a targeted response. This would result in overdominance, where a heterozygote does better than either homozygote. This quantitative explanation for heterozygote advantage has been expanded to the sequence level, triggered by the frequently observed excessive

sequence divergence among MHC alleles: the *divergent allele advantage* hypothesis (Potts & Wakeland 1990; Wakeland et al. 1990). It assumes that heterozygous individuals with more divergent MHC allele combinations (i.e. higher number of pairwise amino acid differences along the peptide-binding domains) would encode for MHC molecules with greater difference in their presented peptide-repertoires. This would result in a more diverse array of presented peptides at the cell surface, conferring increased immune-surveillance against pathogens (Lenz 2011; Pierini & Lenz 2018). An alternative explanation for heterozygote advantage, based on qualitative differences between MHC alleles, stipulates that heterozygosity increases the probability of carrying specific protective MHC alleles. Such qualitative differences have indeed been observed in a number of species, including humans (Piertney & Oliver 2006; Blackwell et al. 2009; Trowsdale 2011). However, it is unclear whether these qualitative differences among MHC alleles result from unique peptide-binding properties (i.e. the ability to present critical peptides) or whether they are also due to quantitative differences in the size of the allele-specific antigen repertoires (Chappell et al. 2015; Manczinger et al.).

A number of studies across a range of species have provided empirical support for a phenotypic advantage conferred by general MHC heterozygosity (Takeshima et al. 2008; Connor et al. 2010; Evans & Neff 2009; Niskanen et al. 2014; Penn et al. 2002) as well as higher sequence divergence between MHC alleles (Lenz et al. 2013; Landry et al. 2001; Lenz et al. 2009; Richman et al. 2001; Neff et al. 2008; Schwensow et al. 2010). Humans have thousands of known alternative HLA alleles (Solberg et al. 2008; Robinson et al. 2017, 2015), yet empirical evidence for pathogen-mediated selection is surprisingly sparse. Owing to the growing number of individuals included in immunogenetic studies and to denser genotyping approaches, multiple significant associations have been identified between infectious or immune phenotypes and variation in the MHC region (Meyer et al. 2018; Trowsdale 2011; Matzaraki et al. 2017). However, most association studies assume simple additive genetic contributions and do not explore the evolutionary implications of potential findings, due to a general focus on the underlying biology and disease mechanisms. One of the few exceptions is the seminal study on HIV control by Carrington *et al.* (Carrington et al. 1999), which demonstrated a slower progression to HIV-related outcomes (AIDS-defining conditions, < 200 CD4+ T cell count and/or death) in HLA heterozygous individuals. Indeed, the role of MHC genes in modulating spontaneous HIV control and progression to AIDS is now

well established (Carrington & O'Brien 2003). However, a recent fine mapping study on the association between MHC and HIV, while confirming substantial additive associations between various MHC alleles and HIV viral load, showed only a very small independent protective effect of HLA-B heterozygosity (McLaren et al. 2015). Here we are therefore revisiting this hallmark example for MHC heterozygote advantage in humans and explore the relative effect of specific MHC alleles versus a general effect of zygosity on HIV control. We take advantage of the well-established association between MHC and HIV control to test whether MHC heterozygote advantage results from quantitative or qualitative differences among MHC alleles. We use genotyping data from 6,311 HIV-1 infected individuals and antigen-binding prediction algorithms for HLA class-I proteins to define the individual repertoires of HIV-1 peptides bound by HLA-A, HLA-B and HLA-C. We show that heterozygosity at HLA-B and HLA-C but not at HLA-A is associated with viral control. While at HLA-B, the heterozygote advantage is potentially mediated by quantitative CTL mediated immune response, another mechanism seems to operate at HLA-C. Furthermore, we show that specific HLA alleles with very strong effects exceed the general heterozygote advantage against HIV-1.

Results

The available data comprised HLA genotypes and alleles (imputed at 4-digit resolution, see methods) and pre-treatment set point viral load (spVL), an established correlate of HIV-1 control and disease progression (Mellors et al. 1996), for 6,311 HIV-infected individuals of European descent. We focused on classical HLA class-I genes (*HLA-A*, *HLA-B* and *HLA-C*), as they are the only genes within the HLA region reported to be independently associated with HIV-1 progression (Carrington et al. 1999; McLaren et al. 2015). A total of 37 HLA-A, 69 HLA-B and 27 HLA-C alleles were represented in the dataset (**Table S1-S3**). We screened all possible 9mer HIV-1 peptides (N = 3,252) across the HIV-1 proteome, and identified 409, 491 and 223 distinct peptides predicted to be bound by at least one of the represented HLA-A, HLA-B and HLA-C alleles, respectively (see methods) (**Table S1-S3**).

HLA heterozygote advantage

We first tested whether heterozygosity at any of the classical HLA class-I loci was associated with better HIV-1 control (i.e. lower spVL), as reported previously (Carrington et al. 1999). Indeed, we observed a lower level of viral load in both HLA-B and HLA-C

heterozygous individuals, compared to homozygous individuals (Wilcox rank sum test, $P = 1.3 \times 10^{-6}$ and 2.8×10^{-6} for HLA-B and HLA-C, respectively, after correcting for multiple testing), while heterozygosity at HLA-A showed no statistically significant effect on spVL (Wilcox rank sum test $P = 0.16$) (**Figure 1**). The associations for HLA-B and HLA-C were highly significant, even though the actual effect was quite small (effect size = -0.25 and -0.21, respectively).

The HLA region is characterized by strong linkage disequilibrium (LD) (Stenzel et al. 2004; Blomhoff et al. 2006; Ahmad et al. 2003). To test whether the observed effects of HLA-B and HLA-C heterozygosity were independent, we calculated the association of HLA-B heterozygosity with spVL among HLA-C heterozygotes only ($N = 5,748$), thus controlling for HLA-C zygosity. This test demonstrated the effect of HLA-B heterozygosity to be independent of HLA-C heterozygosity ($P = 0.01$). The same was true for HLA-C zygosity ($P = 0.007$ within HLA-B heterozygotes, $N = 5,933$). We also observed that heterozygosity at two loci led to better HIV-1 control compared to heterozygosity at only one locus (**Figure S1**). Overall, these results suggest that our dataset is appropriate to explore the effect of HLA heterozygote advantage against HIV and its underlying mechanism in more detail.

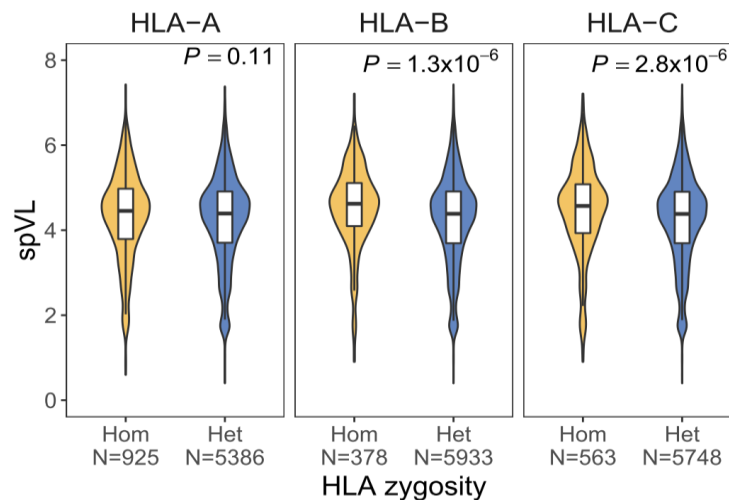


Figure 1 | Viral load in HLA homozygotes and heterozygotes. Comparison of the set point viral load (spVL) between HLA homozygote and heterozygote individuals for the three HLA class I loci. Outlier values are not shown for better visual comparison. N indicates the number of individuals. Bonferroni-corrected P -value from Wilcoxon rank sum test is shown.

Divergent allele advantage

We sought to test the hypothesis of divergent allele advantage by evaluating the relation between sequence divergence of an individual's allele pair and viral load. We calculated pairwise sequence divergence as a measure of sequence distance between both alleles of a given HLA class-I locus in each individual. Sequence distance was calculated at the amino acid level, using Grantham scores, which proved to be the most suitable proxy for functional divergence in an earlier study (Pierini & Lenz 2018). First, we saw that, while zygosity correlated strongly between the HLA-B locus and the other two HLA class I loci, HLA-C and HLA-A ($\tau = 0.55$ and $\tau = 0.16$, respectively), the sequence divergence between an individual's HLA-B alleles was only weakly correlated with the divergence between HLA-C ($\tau = 0.09$, $P < 0.0001$) and HLA-A alleles ($\tau = 0.02$, $P = 0.002$) (**Figure S2**). This might suggest independent selection for sequence divergence at these loci and is in line with high recombination in this region as well as selection usually targeting very specific residues involved in peptide binding (Reche & Reinherz 2003). Following the predictions of the divergent allele advantage hypothesis, HLA-B, the locus with by far the strongest association to HIV control, showed a negative correlation between pairwise allele divergence and spVL across individuals ($\tau = -0.08$, $P = 8.6 \times 10^{-20}$) (**Figure 2**). While we found no such correlation for HLA-A ($P = 0.77$), HLA-C surprisingly showed a positive association between allele divergence and viral load ($\tau = 0.03$, $P = 4.9 \times 10^{-4}$). We observed similar correlations after excluding homozygous individuals (**Figure S3**).

This would essentially indicate a divergent allele disadvantage at HLA-C, for which it is difficult to conceive a plausible mechanistic scenario, and it is also at odds with the observed advantage for HLA-C heterozygotes described above. However, it has been shown that HLA-C coevolves with KIR genes (Parham et al. 2012) and does not seem to be under selection for contributing to diverse immunological surveillance (Buhler et al. 2016). It is thus possible that this positive correlation is not due to quantitative differences in antigen presentation among HLA-C alleles, but rather due to interactions between HLA-C and KIR. Notably, even though HLA-B and HLA-C loci are in strong LD, the pairwise Grantham distance between individual's HLA-B alleles was significantly higher than between HLA-C alleles ($P < 0.001$) (**Figure S4**), supporting the notion that HLA-C is not evolving under the same selective constraints as HLA-B.

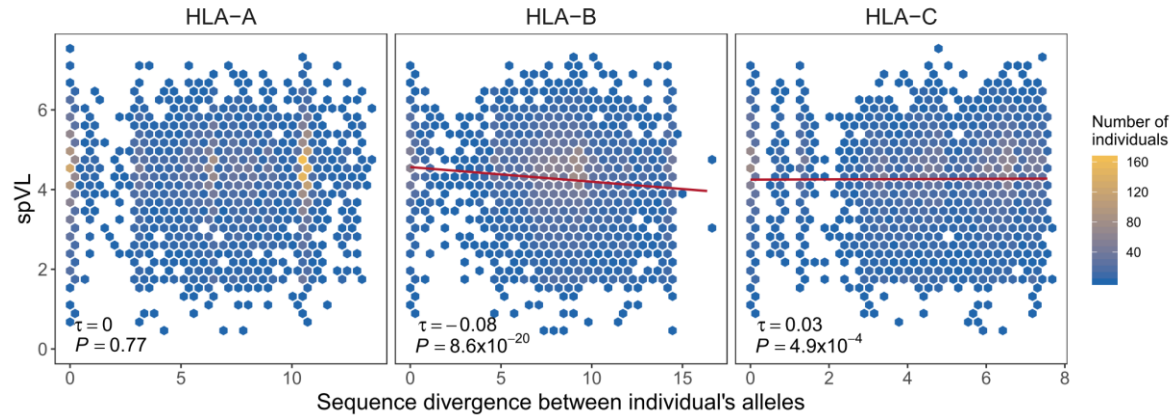


Figure 2 | Sequence divergence between individual's HLA alleles and viral load. Correlation between set point viral load (spVL) and sequence divergence between individual's HLA-A, HLA-B and HLA-C alleles is shown (including homo- and heterozygotes; $N = 6,311$). Individuals with similar parameter values are binned for better visualization, with bin color indicating the number of individuals per bin. Kendall's estimate of correlation τ and Bonferroni-corrected P -value are shown.

Functional consequence of heterozygosity and allele divergence

Having observed the effects of both heterozygosity and allele divergence in our data, we then asked whether these measures of genetic variability would indeed allow for the presentation of a broader array of HLA-bound peptides, as hypothesized by the quantitative explanation for MHC heterozygote advantage. Using computational peptide-binding prediction, we found that heterozygosity on average resulted in a broader array of bound peptides for all three classical HLA loci (**Figure S5**). Furthermore, the number of peptides bound by a pair of HLA-B alleles was positively correlated with the sequence divergence between the alleles (**Figure S6**). It was true for HLA-A and HLA-C as well (**Figure S6**). This association between sequence divergence at the HLA loci and predicted functional divergence among HLA molecule variants has been reported before (Lenz 2011; Pierini & Lenz 2018). However, our present HIV dataset allowed us to evaluate this association using an empirical disease phenotype. We thus tested whether the ability to bind more HIV-1 peptides was associated with HIV control (i.e. spVL). Indeed, the individual-specific number of HIV-1 peptides predicted to be bound by the individual's HLA-B molecules was negatively associated with viral load (Kendall correlation, $\tau = -0.12$, $P = 1.0 \times 10^{-47}$) (**Figure 3**). This was also true for HLA-A and HLA-C, but correlation coefficients were much smaller (Kendall correlation, HLA-A: $\tau = -0.04$,

$P = 1.7 \times 10^{-5}$; HLA-C: $\tau = -0.05$, $P = 8.7 \times 10^{-8}$). Interestingly, the association between viral load and the breadth of individual-specific HLA-B bound peptides was stronger ($\tau = -0.12$) than the association between viral load and allele divergence ($\tau = -0.07$), suggesting that allele divergence is a useful but imperfect proxy for functional divergence among HLA-B alleles, at least in the case of HIV-1 with its limited peptide repertoire.

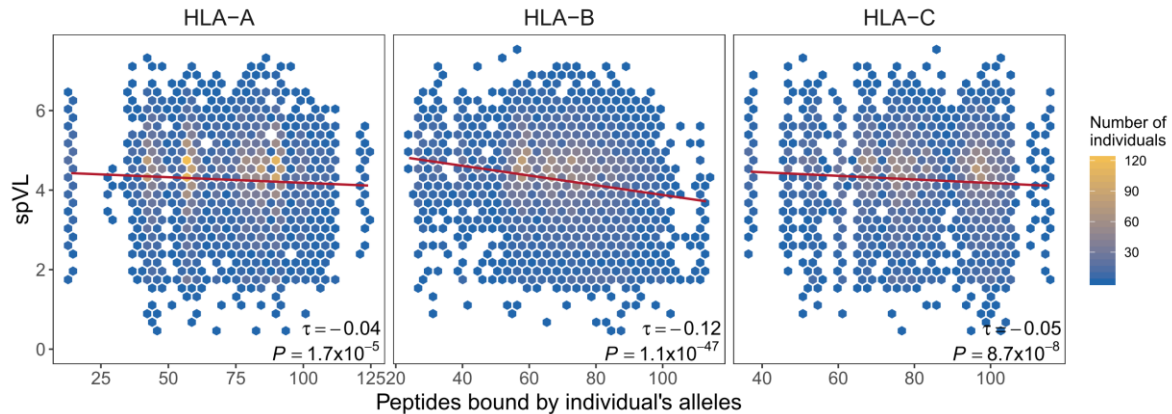


Figure 3 | HLA-bound peptides and viral load. Correlation between individual's set point viral load (spVL) and the breadth of HIV-1 peptides predicted to be bound by HLA-A, HLA-B and HLA-C alleles is shown (including homo- and heterozygotes; $N = 6,311$). Individuals with similar parameter values are binned for better visualization, with bin color indicating the number of individuals per bin. Kendall's estimate of correlation τ and Bonferroni-corrected P -value are shown.

HLA heterozygosity and within-host evolution of HIV

Presentation of pathogenic peptides is thought to increase the likelihood and efficiency of a pathogen-specific immune response. Consequently, such HLA restriction is a potential factor that influences the evolutionary landscape of pathogens. Specifically for HIV-1, the virus has been shown to acquire mutations within HLA-bound peptides that can help it to escape immune recognition (Bronke et al. 2013; Leslie et al. 2004; Arora et al. 2019). In this context, following the predictions of the quantitative heterozygote advantage hypothesis, heterozygous HLA genotypes should exert a broader selective pressure on the virus, leading to a larger number of escape mutants. Taking advantage of our unique dataset, which also comprised a limited set of autologous HIV-1 sequences from a subset of individuals (see methods), we performed a phylogenetic

comparison of these autologous HIV-1 sequences to test whether HLA heterozygosity leads to more pronounced within-individual evolution, possibly because of the broader HLA restriction. For this analysis we focused on HLA-B, the locus with strongest association with HIV control, and observed that autologous virus sequences in HLA-B heterozygous individuals (N = 36) were indeed more divergent than the ones in homozygous individuals (N = 4) (Difference in distance from the root to tips; Wilcoxon rank sum test $P = 0.0004$, **Figure 4 A, B**). In order to account for the unbalanced sample size, we permuted the individuals across zygosity groups for 1,000 times and each time re-calculated the average divergences. This analysis showed that the observed difference was unlikely to be due to chance (one-tailed $P = 0.028$) (**Figure S7**). We extrapolated this finding to the observed protective effect of the overall breadth of HLA-presented peptides in individuals. We observed a positive correlation between the total number of HLA-bound peptides and the divergence of HIV sequence within individuals for which autologous HIV sequence data was available (N = 40) (**Figure 4C**). Nonetheless, more autologous HIV-1 sequences, particularly from HLA homozygous individuals, will help to corroborate these findings.

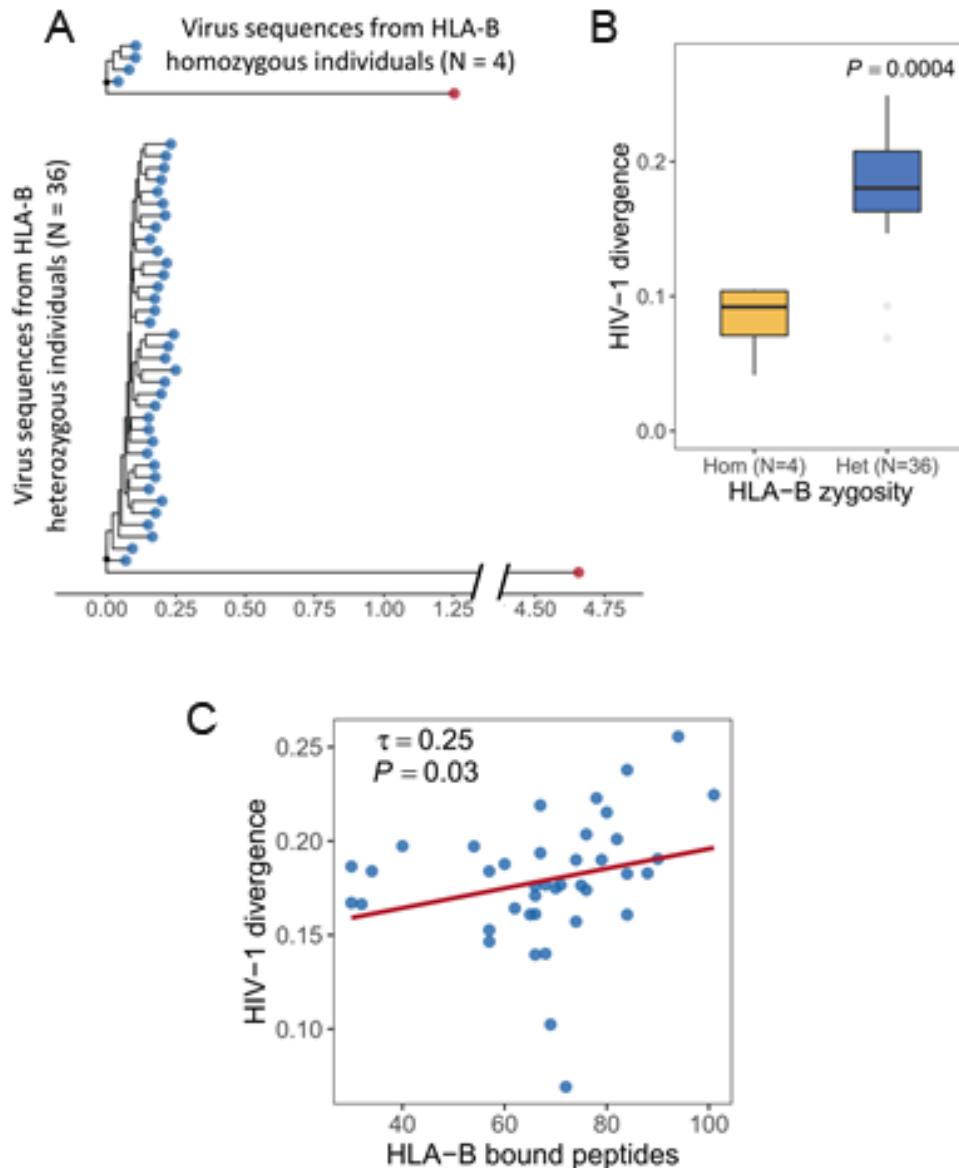


Figure 4 | Within-host evolution of HIV in response to HLA-B heterozygosity. (A) Rooted trees of autologous virus sequences (blue dots) from HLA-B homozygous (N = 4) and heterozygous (N = 36) individuals. We used HIV-2 (red dot) as outgroup to identify the root during tree construction. **(B)** The comparison of divergence (measured as root-to-tip distance) of virus sequences in HLA-B homozygous and heterozygous individuals. N indicates the number of individuals. *P*-value from Wilcoxon rank sum test is shown. **(C)** The correlation between the number of HLA-bound peptides and the divergence of HIV sequence in individuals (N = 40).

Allele-specific effects versus general heterozygote advantage

HLA alleles are known to show differential association with HIV-1 control ranging from risk to protection (McLaren et al. 2015). We therefore tested the alternative hypothesis for heterozygote advantage, which suggests that heterozygosity might simply increase the chances of carrying a particular HLA allele that binds immunogenic peptides and through that provides better viral control. Of 7 HLA-B alleles significantly associated with HIV control, 4 (including B*57:01) were indeed enriched in the heterozygous state (**Figure S8**), making this hypothesis a viable explanation. The non-significant enrichment for the other 3 alleles could be due to their low frequency in our dataset. We then asked what property made individual alleles more protective. Following the same intuition as for the quantitative heterozygote advantage above, they could confer protection simply by presenting more peptides to T-cells compared to other alleles. Alternatively, they could confer a qualitative advantage by presenting very specific peptides that are particularly difficult (i.e. costly in terms of fitness) for the virus to mutate. We tested this point by focusing on HLA-B*57:01, the allele known to confer the strongest resistance against disease progression (Bailey et al. 2006; Migueles et al. 2000; Pohlmeier et al. 2013; McLaren et al. 2015). We found that B*57:01+/- heterozygous individuals exhibited a lower viral load than B*57:01-/- heterozygotes (carrying any two HLA-B alleles except B*57:01) (Wilcoxon rank sum test $P < 0.001$; **Figure 5A**). However, B*57:01+/- individuals are predicted to bind a greater breadth of peptides compared to B*57:01-/- heterozygous individuals (Wilcoxon rank sum test $P < 0.001$; **Figure 5B**), making it difficult to discern quantitative and qualitative effects of B*57:01. Yet, the allele B*57:01 alone was predicted to bind fewer HIV-1 peptides ($N = 54$) than the median number of the peptides ($N = 68$) bound by B*57:01-/- heterozygous individuals (Wilcoxon rank sum test $P = 0.003$; **Figure 5B**). This allowed us to evaluate the qualitative effect of binding specific HIV peptides on viral load while excluding any quantitative advantage of binding more HIV peptides. We found that individuals homozygous for B*57:01 also exhibited a lower viral load than B*57:01-/- heterozygotes (Wilcoxon rank sum test $P = 0.011$) (**Figure 5A**), suggesting that binding of specific HIV-1 peptides provides a qualitative advantage to B*57:01. Nevertheless, B*57:01 is also predicted to bind the largest number of HIV-1 peptides among all tested HLA-B alleles (Arora et al. 2019), maintaining the possibility that both qualitative and quantitative aspects of peptide binding are contributing to the protective effect of this allele.

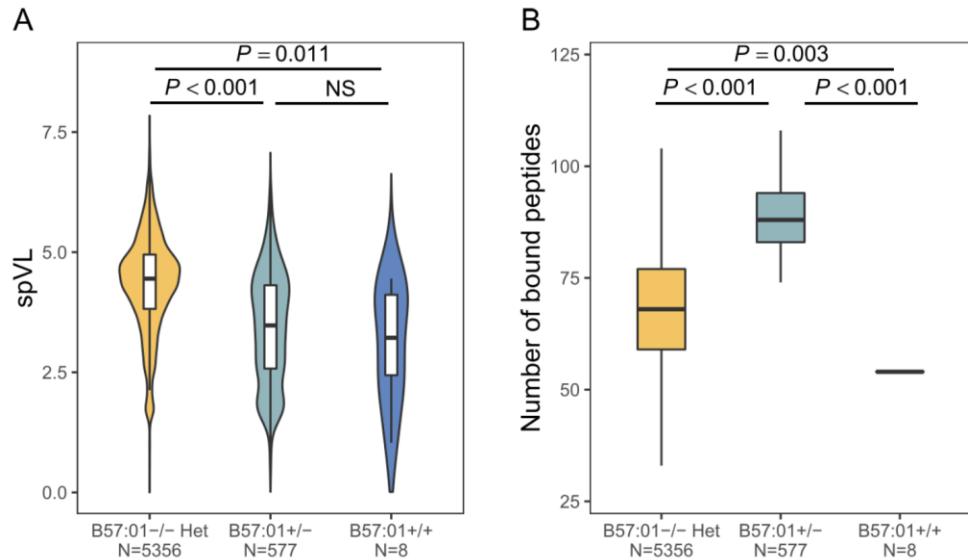


Figure 5 | Heterozygote advantage vs. allele-specific effect for HLA-B*57:01. Comparison of set point viral load (spVL) **(A)** and the number of HLA-bound HIV-1 peptides **(B)** in HLA heterozygous individuals not carrying HLA-B*57:01 allele, individuals carrying one copy of HLA-B*57:01 allele and individuals homozygous for HLA-B*57:01. N indicates the number of individuals. Bonferroni-corrected P -values from Wilcoxon rank sum test are shown.

Having established that the effect of individual HLA alleles can significantly exceed the effect of general zygosity, we next aimed to explore more generally the relative role of additive effects of HLA-B alleles versus HLA-B heterozygosity. We therefore tested whether the observed associations of HLA heterozygosity, sequence divergence and the number of bound HIV-1 peptides with spVL remained significant after accounting for the additive effect of individual HLA-B alleles. Using a regression model that included allele-specific effects in an additive manner, we observed qualitatively equivalent effects of HLA heterozygosity, sequence divergence and the number of bound peptides on viral load as in the absence of allele-specific effects, with a variation in viral load associated with these compound parameters ranging from 0.06 to 0.09 % (**Table 1**). However, it is also apparent that the independent effect of these compound parameters is substantially lower than the combined additive effects of all individual HLA-B alleles, which account for 11.4 % of the variation in viral load (McLaren et al. 2015).

Table 1 | Variation in set point viral load (spVL) associated with HLA-B heterozygosity, sequence divergence between individuals' HLA-B alleles, and individual-specific breadth of HLA-B bound peptides before and after accounting for allele-specific additive effects.

HLA-B	Associated variation in spVL in % (P-value)	
	Without allele-specific effects	With allele-specific effects
Heterozygosity	0.3 (6.6×10^{-6})	0.06 (0.016)
Sequence divergence	1.3 (1.0×10^{-20})	0.09 (0.005)
Bound peptides	3.3 (2.6×10^{-50})	0.09 (0.006)

Discussion

Of all three classical HLA class-I genes tested here, heterozygosity at HLA-B and HLA-C was independently associated with viral control. The absence of any significant association between HLA-A heterozygosity and viral load suggests that the previously observed contribution of HLA-A heterozygosity to disease resistance (Carrington et al. 1999) might be the consequence of its LD with neighboring HLA-B and/or HLA-C loci (Stenzel et al. 2004; Blomhoff et al. 2006; Ahmad et al. 2003). While we recapitulated the general observation that higher sequence divergence between an HLA allele pair could lead to a larger number of bound peptides (Pierini & Lenz 2018; Lenz 2011) (HIV peptides in this case), a weak association of sequence divergence with HIV-1 viral load suggests that the number of bound peptides could be a better proxy for immunocompetence when focusing on a specific pathogen.

The negative correlation between the number of HLA-B-bound peptides and viral load in individuals suggests that HLA-B heterozygote advantage is significantly mediated via quantitative CTL response to a broad set of HLA-presented HIV-1 peptides, though empirical validation would substantiate the finding. Interestingly, a relatively weak negative correlation between HLA-C bound peptides and viral load suggests that an effector mechanism other than CTL-mediated quantitative immune response might be responsible for HLA-C heterozygote advantage. This suggestion gains additional support from the fine mapping study by McLaren *et al.* 2015 (McLaren et al. 2015), where unlike for HLA-B, there were no HIV-associated amino-acid residues found for HLA-C. Moreover, with only about half the peptide-repertoire size of HLA-C alleles, relative to HLA-A and HLA-B, HLA-C might be evolving not to interact with the vast diversity of CTLs, but other relatively less diverse cell types. One such cell type could be Natural

Killer (NK) cells which express Killer-cell immunoglobulin-like receptors (KIRs) on their cell surface (Parham 2005). HLA-C molecules are thought to be a potent ligand of KIRs (Colonna et al. 1993; Parham 2005; Hilton et al. 2015), and specific interactions between HLA-C molecules and KIRs have been associated with multiple diseases, including HIV infection (Rajagopalan & Long 2005; Zipeto & Beretta 2012; Körner et al. 2017).

Pereyra *et al.* 2014 have shown that CD8+ T-cell targeting of specific HLA-presented peptides could confer viral control (Pereyra et al. 2014). A broader array of HLA-bound peptides in heterozygous individuals might increase the possibility that such peptides are presented on the cell surface. However, the particular case of B*57:01 allele conferring superior viral control compared to general HLA-B heterozygote advantage suggests that HLA allele-specific effects might arise from binding specific immuno-dominant peptides. However, since B*57:01 also bound the largest number of peptides among all HLA-B alleles, we cannot completely rule out that the quantitative advantage of binding a large number of peptides contributes to its strong protective effect in HIV-1 control.

Together, these results suggest that HLA heterozygosity in an individual might confer advantage in multiple, possibly additive ways. One is the quantitative advantage through presentation of a larger number of viral peptides, which might generate a broad immune response. This appears to exert stronger evolutionary pressure on the virus to evolve, as shown by elevated sequence diversity of the virus in HLA heterozygous individuals, possibly resulting in replicative fitness cost. In addition, HLA heterozygosity might provide an advantage by making it more likely to carry certain protective HLA alleles that can present immuno-dominant peptides to T-cells and thus lead to disease-control.

In conclusion, our findings shed light on the functional basis of the protective association between HLA heterozygosity and HIV progression. Interestingly, heterozygote advantage is generally thought to be more important in a multi-parasite context, with HLA heterozygosity assumed to enable hosts to recognize and fight more different parasites (Penn et al. 2002). It is certainly conceivable that in such a multi-parasite context, quantitative aspects of antigen presentation become more important than qualitative aspects due to the sheer number of peptides. Nevertheless, our study demonstrates and characterizes HLA gene-specific heterozygote advantage even against a single pathogen. The findings disentangle the role of quantitative and qualitative features of the HLA's peptide-repertoire in mediating the immune response,

and suggest that even a single pathogen can lead to selection for both HLA heterozygosity (including excessive allele divergence) and specific HLA alleles. Moreover, they lend support to HIV vaccine programs aiming to impart antiviral immunity using a broad, yet specific array of HIV peptides.

Materials and Methods

Samples and Genotype data

We used genotyping and clinical data from 6,311 chronically HIV-1 infected individuals that were previously analyzed and described in detail in McLaren *et al.* (McLaren *et al.* 2015). Briefly, genome-wide genotype data were collected from cohorts participating in the International Collaboration for the Genomics of HIV. SNP genotypes absent in original genotyping platforms were imputed (McLaren *et al.* 2015). Imputed SNPs with low imputation quality (r^2 score < 0.3) or minor allele frequency of $< 0.5\%$ were discarded. HLA alleles (at 4-digit resolution) for classical class-I loci (*HLA-A*, *HLA-B*, *HLA-C*) were imputed from genome-wide genotype data using best-guess genotypes. Pre-treatment HIV-1 set point viral load (spVL; log₁₀ HIV-1 RNA copies/ μ l of plasma) was used as quantitative disease phenotype (McLaren *et al.* 2015).

HLA binding affinity for HIV-1 peptides

Following Arora *et al.* (Arora *et al.* 2019), we used the reference proteome of HIV-1 M group subtype B (NCBI accession NC_001802.1) that comprised 10 proteins. The *Gag-Pol* protein is a precursor protein resulting from a -1 ribosomal frameshifting event in upstream *Gag* (Jacks *et al.* 1988), and then cleaved by the virus-encoded protease to produce the mature Pol protein. In order to avoid redundancy with the separate *Gag* protein in our analysis, we manually trimmed the *Gag-Pol* protein sequence to Pol. HLA class-I molecules preferentially bind and present 9mer peptides (Falk *et al.* 1991; York *et al.* 2002). We used a computational method called NetMHCpan v4.0 (Jurtz *et al.* 2017) to predict the binding affinity of all possible 9mer peptides derived from the entire HIV-1 proteome to individual HLA class I alleles represented in our dataset. The method reports the rank of predicted binding affinity of HLA-peptide complexes against predicted affinity of random natural peptides. HLA-peptide complexes with predicted binding affinity rank < 0.5 were retained (corresponding to 'strongly bound' peptides) (Jurtz *et al.*

2017). The breadth of peptides bound by an individual's HLA allele pair was taken as the total number of unique peptides predicted to be bound by both alleles.

Sequence divergence between alleles of HLA genotype per individual

Sequence divergence between alleles was computed for all HLA allele pairs (genotypes) of HLA-A, HLA-B and HLA-C loci. Protein sequences of HLA alleles were taken from IMGT/HLA database (Robinson et al. 2015). Exons 2 and 3, which encode the variable region in the peptide binding groove of HLA class I molecules, were obtained following the exon annotation reported in Ensemble database (Aken et al. 2016). The alignment of amino acid sequences was performed using MUSCLE (Edgar 2004). The genetic distances between aligned allele pairs were calculated based on the Grantham distance matrix (Grantham 1974) using a custom Perl script freely available online (Pierini & Lenz 2018). The non-parametric Kendall correlation was used to test for the associations of the sequence divergence between individual's HLA alleles with: (i) set point viral load (spVL), and (ii) the combined number of bound peptides. All p-values were adjusted for multiple testing across the number of loci tested.

Phylogenetic comparison of autologous virus sequences

Autologous sequences of 6 HIV-1 proteins, namely Gag, Pol, Vif, Vpr, Vpu and Nef, were available for 65 individuals. Due to fragmented coverage of sequences, we concatenated these 6 proteins in order to obtain better resolution and statistical power, and aligned them using MAFFT v7 with default parameters (Kato et al. 2002). The alignment was optimized for maximum gap-free area using MaxAlign-1.1 (Gouveia-Oliveira et al. 2007), which resulted into 4 sequences for HLA-B homozygous and 36 for heterozygous individuals. Maximum likelihood trees were made using PhyML-3.1 (Guindon & Gascuel 2003) with default parameters. The tip-to-tip distances were extracted using Ape-3.5 package (Paradis et al. 2004) in R 3.5.1. Using HIV-2 as an outgroup, the root-to-tip distance was used as a proxy for evolutionary divergence of host-specific HIV-1 clones. Difference in the mean root-to-tip distance in each group of individuals was taken as the observed difference in divergence. We obtained statistical significance of the observed difference by permuting the individuals across the groups and repeating the above procedure 1000 times. *P-value* was taken as the fraction of permutations where the difference in mean divergence was equal to or more than the observed difference.

Association with viral load while controlling for allele-specific effects

The association of a variable with spVL was calculated using a linear regression model following McLaren *et al.* (McLaren et al. 2015). Variation in spVL attributable to a given variable (heterozygosity, sequence divergence or the breadth of bound peptides) while controlling for allele-specific effects was taken as the difference between adjusted-R2 values of the model with variable, alleles and covariates (Equation 1) and the model with alleles and covariates only (Equation 2). We did the analysis for each classical HLA class-I locus separately. Models contained all imputed alleles (N = 69 for HLA-B; N = 37 for HLA-A; N = 27 for HLA-C), the first five principle components of SNP variation, and the cohort identity (all adopted from McLaren *et al.* (McLaren et al. 2015)) as the covariates. The significance of the variable's association with viral load was calculated by comparing these two models using chi-square test.

$$spVL = \beta_1 * variable + \sum_{i=1}^N \beta_i * allele_i + \beta_2 * covariates + \varepsilon_1$$

[Equation 1]

$$spVL = \sum_{i=1}^N \beta_i * allele_i + \beta_2 * covariates + \varepsilon_2$$

[Equation 2]

All analyses were performed in R v3.5.1 and data was visualized using the ggplot2 v2.2.1 package (Wickham & Wickham 2007).

Acknowledgements

Individual and HIV sequence data was collected and generously provided by the International Collaboration for the Genomics of HIV. This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was

supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research as well as the Emmy Noether Programme of the Deutsche Forschungsgemeinschaft (grant LE 2593/3-1 to T.L.L.). We also thank Pleuni Pennings and Bernhard Haubold for advice on analyzing within-individual evolution of HIV-1 virus.

Supplementary Materials

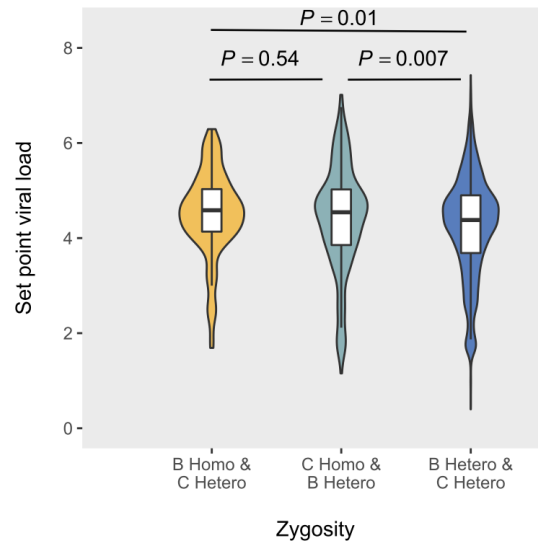


Figure S1 | Comparison of the set point viral load between individuals who are heterozygous for HLA-C only (N = 110), for HLA-B only (N = 295) and for both HLA-B and HLA-C (N = 5,368). P values from Wilcoxon rank sum test are shown.

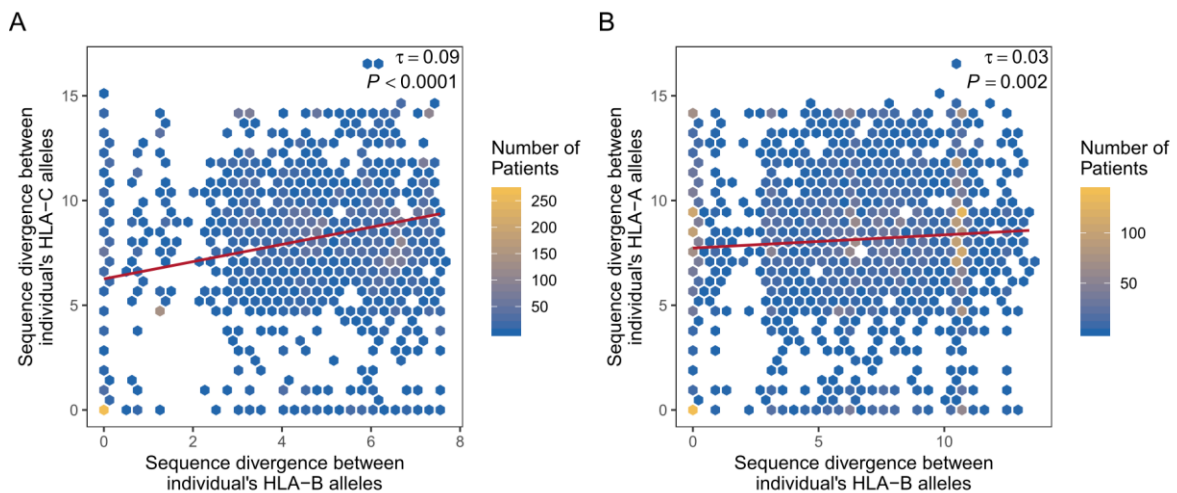


Figure S2 | Correlation between the sequence divergence between HLA-B alleles and the sequence divergence between HLA-C alleles (A), and HLA-A (B) of an individual is shown. Individuals with similar parameter values are binned for better visualization. Kendall's estimate of correlation and Bonferroni-corrected *P*-value is shown.

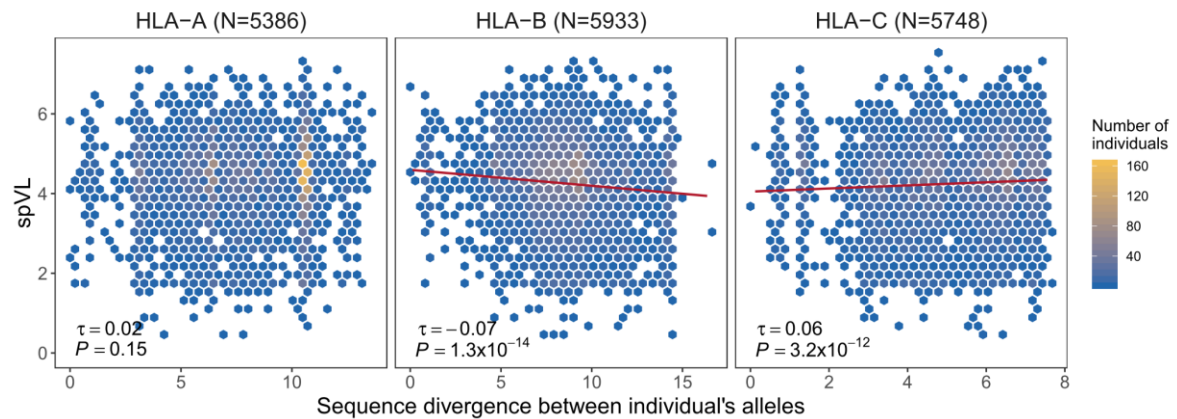


Figure S3 | Sequence divergence between individual's HLA alleles and viral load. Correlation between set point viral load (spVL) and sequence divergence between individual's HLA-A, HLA-B and HLA-C alleles is shown for heterozygous individuals. Individual with similar parameter values are binned for better visualization, with bin color indicating the number of individuals per bin. N indicates the number of heterozygous individuals. Kendall's estimate of correlation τ and Bonferroni-corrected P -value are shown.

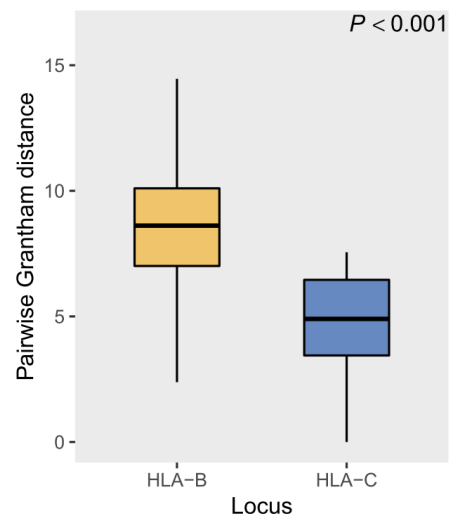


Figure S4 | Comparison of pairwise Grantham distance between individual's HLA-B and HLA-C alleles. P -value from Wilcoxon rank sum test are shown.

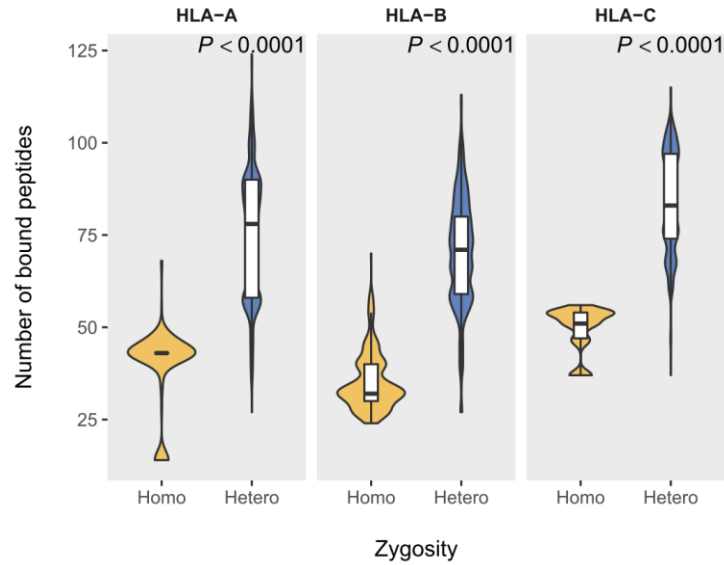


Figure S5 | HLA heterozygous individuals, whether for HLA-A, HLA-B or HLA-C, bound a significantly greater breadth of HIV-1 peptides compared to homozygotes.

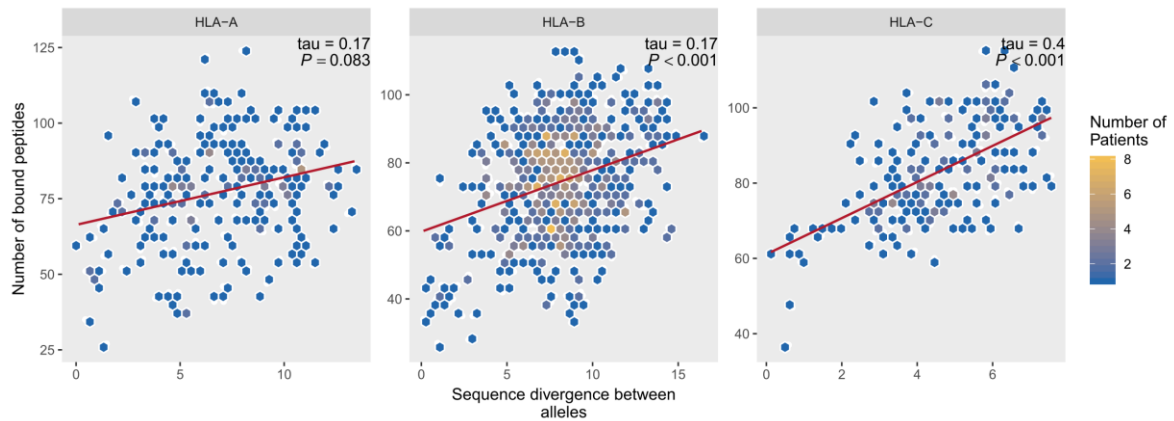


Figure S6 | Correlation between pairwise sequence divergence (Grantham distance) between an individual's HLA alleles and the breadth of bound HIV-1 peptides for HLA-A, HLA-B and HLA-C.

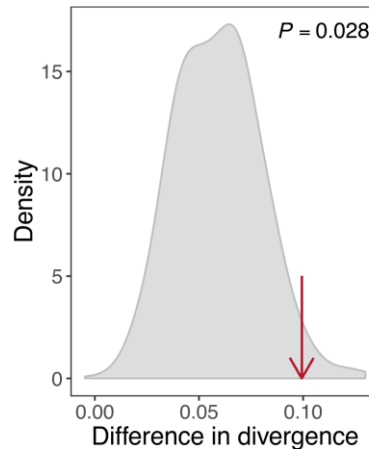


Figure S7 | The observed difference in the mean root-to-tip distance in groups of HLA-B heterozygous and homozygous individuals resided within the top 2.8% of distance distribution of 1000 tree-pairs generated by permuting the individuals across zygosity groups. P -value is one-tailed.

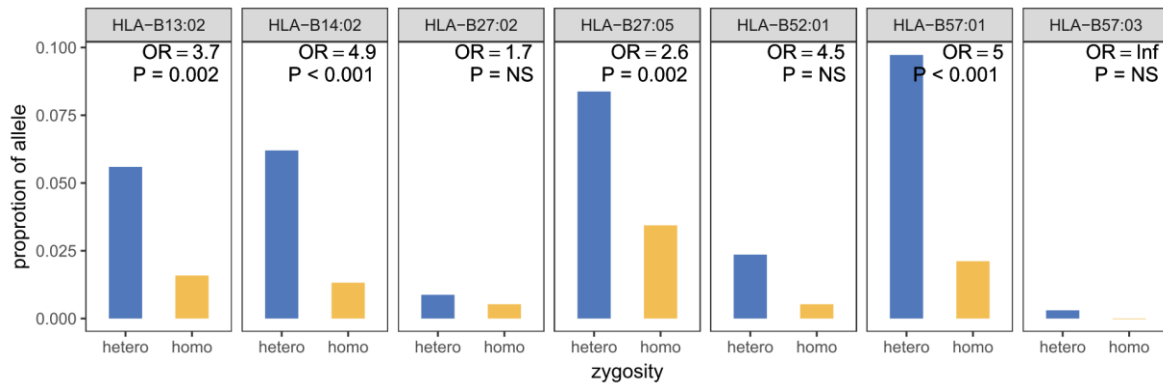


Figure S8 | Enrichment for protective HLA-B alleles in HLA-B heterozygous individuals compared to homozygous individuals. Odds ratio (OR) and P -value from Fisher exact test are shown.

Table S1 | The HLA-B alleles represented in the dataset. There were 69 alleles for HLA-B gene represented in our dataset, which bound varied number of HIV-1 peptides (41 ± 10).

HLA allele Number of bound HIV peptides

HLA-B07:02	32
HLA-B07:04	37
HLA-B07:05	34
HLA-B08:01	40
HLA-B13:01	65
HLA-B13:02	48
HLA-B14:01	45
HLA-B14:02	45
HLA-B15:01	45
HLA-B15:03	53
HLA-B15:05	48
HLA-B15:07	47
HLA-B15:08	38
HLA-B15:10	56
HLA-B15:16	49
HLA-B15:17	54
HLA-B15:18	51
HLA-B15:25	46
HLA-B15:27	49
HLA-B18:01	34
HLA-B27:02	28
HLA-B27:03	31
HLA-B27:04	31
HLA-B27:05	35
HLA-B27:07	27
HLA-B35:01	34
HLA-B35:02	41
HLA-B35:03	41
HLA-B35:08	33
HLA-B35:12	41
HLA-B35:17	33
HLA-B37:01	43
HLA-B38:01	58
HLA-B39:01	54
HLA-B39:06	38
HLA-B39:09	55
HLA-B39:10	42
HLA-B40:01	24

HLA-B40:02	26
HLA-B40:06	36
HLA-B41:01	44
HLA-B41:02	35
HLA-B42:01	42
HLA-B42:02	49
HLA-B44:02	27
HLA-B44:03	27
HLA-B44:04	25
HLA-B44:05	31
HLA-B45:01	43
HLA-B46:01	47
HLA-B47:01	49
HLA-B48:01	70
HLA-B49:01	35
HLA-B50:01	39
HLA-B51:01	30
HLA-B51:02	33
HLA-B51:05	35
HLA-B51:06	36
HLA-B51:08	36
HLA-B52:01	51
HLA-B53:01	48
HLA-B55:01	33
HLA-B56:01	39
HLA-B56:04	41
HLA-B57:01	54
HLA-B57:02	57
HLA-B57:03	56
HLA-B58:01	57
HLA-B73:01	33

Table S2 | The HLA-A alleles represented in the dataset. There were 37 alleles for HLA-A gene represented in our dataset, which bound varied number of HIV-1 peptides (40 ± 11).

HLA allele Number of bound HIV peptides	
HLA-A01:01	14
HLA-A01:02	26
HLA-A01:03	15
HLA-A02:01	43
HLA-A02:02	40
HLA-A02:03	47
HLA-A02:05	47
HLA-A02:06	44
HLA-A02:11	51
HLA-A03:01	47
HLA-A03:02	41
HLA-A11:01	41
HLA-A11:02	41
HLA-A11:03	44
HLA-A23:01	39
HLA-A24:02	43
HLA-A24:07	47
HLA-A25:01	38
HLA-A26:01	28
HLA-A26:08	29
HLA-A29:01	35
HLA-A29:02	35
HLA-A30:01	62
HLA-A30:02	37
HLA-A30:04	45
HLA-A31:01	55
HLA-A32:01	68
HLA-A33:01	41
HLA-A33:03	46
HLA-A34:02	49
HLA-A36:01	21
HLA-A66:01	43
HLA-A68:01	34
HLA-A68:02	38
HLA-A69:01	41
HLA-A74:01	52
HLA-A80:01	23

Table S3 | The HLA-C alleles represented in the dataset. There were 27 alleles for HLA-C gene represented in our dataset, which bound varied number of HIV-1 peptides (50 ± 5).

HLA allele	Number of bound peptides
HLA-C01:02	51
HLA-C02:02	56
HLA-C02:06	55
HLA-C03:02	52
HLA-C03:03	47
HLA-C03:04	47
HLA-C04:01	37
HLA-C04:03	48
HLA-C04:07	37
HLA-C05:01	46
HLA-C06:02	54
HLA-C07:01	54
HLA-C07:02	51
HLA-C07:04	51
HLA-C08:01	52
HLA-C08:02	54
HLA-C12:02	48
HLA-C12:03	54
HLA-C14:02	50
HLA-C15:02	55
HLA-C15:04	53
HLA-C15:05	51
HLA-C16:01	53
HLA-C16:02	53
HLA-C16:04	47
HLA-C17:01	65
HLA-C18:01	43

References

- Ahmad T et al. 2003. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* 12:647–656. doi: 10.1093/hmg/ddg066.
- Aken BL et al. 2016. The Ensembl gene annotation system. *Database (Oxford)*. 2016:1–19. doi: 10.1093/database/baw093.
- Apanius V, Penn D, Slev PR, Ruff LR, Potts WK. 1997. The Nature of Selection on the Major Histocompatibility Complex. *Crit. Rev. Immunol.* doi: 10.1615/CritRevImmunol.v17.i2.40.
- Arora J et al. 2019. HIV peptidome-wide association study reveals patient-specific epitope repertoires associated with HIV control. *Proc. Natl. Acad. Sci.* 201812548. doi: 10.1073/pnas.1812548116.
- Bailey JR, Williams TM, Siliciano RF, Blankson JN. 2006. Maintenance of viral suppression in HIV-1-infected HLA-B*57+ elite suppressors despite CTL escape mutations. *J. Exp. Med.* 203:1357–69. doi: 10.1084/jem.20052319.
- Blackwell JM, Jamieson SE, Burgner D. 2009. HLA and Infectious Diseases. *Clin. Microbiol. Rev.* 22:370–385. doi: 10.1128/CMR.00048-08.
- Blomhoff A et al. 2006. Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype specific manner assessed mainly by DRB1*03 and DRB104* haplotypes. *Genes Immun.* 7:130–140. doi: 10.1038/sj.gene.6364272.
- Bronke C et al. 2013. HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes. *Aids.* 27:899–905. doi: 10.1097/QAD.0b013e32835e1616.
- Buhler S, Nunes JM, Sanchez-Mazas A. 2016. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics.* 68:401–416. doi: 10.1007/s00251-016-0918-x.
- Carrington M et al. 1999. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage.
- Carrington M, O'Brien SJ. 2003. The Influence of *HLA* Genotype on AIDS. *Annu. Rev. Med.* 54:535–551. doi: 10.1146/annurev.med.54.101601.152346.
- Chappell PE et al. 2015. Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife.* 4:1–22. doi: 10.7554/eLife.05345.
- Colonna M, Borsellino G, Falco M, Ferrara GB, Strominger JL. 1993. HLA-C is the inhibitory ligand that determines dominant resistance to lysis by NK1- and NK2-specific natural killer cells. *Proc. Natl. Acad. Sci. U. S. A.* 90:12000–4. doi: 10.1073/pnas.90.24.12000.
- Connor SLO et al. 2010. MHC Heterozygote Advantage in Simian Immunodeficiency Virus – Infected Mauritian Cynomolgus Macaques. 2.

- Doherty PC, Zinkernagel RM. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*. 50–52. doi: doi:10.1038/256050a0.
- Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797. doi: 10.1093/nar/gkh340.
- Eizaguirre C, Lenz TL. 2010. Major histocompatibility complex polymorphism: Dynamics and consequences of parasite-mediated local adaptation in fishes. *J. Fish Biol*. 77:2023–2047. doi: 10.1111/j.1095-8649.2010.02819.x.
- Ejsmond MJ, Radwan J. 2015. Red Queen Processes Drive Positive Selection on Major Histocompatibility Complex (MHC) Genes. *PLoS Comput. Biol*. 11:1–14. doi: 10.1371/journal.pcbi.1004627.
- Evans ML, Neff BD. 2009. Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Mol. Ecol*. 18:4716–4729. doi: 10.1111/j.1365-294X.2009.04374.x.
- Falk K, Röttschke O, Stevanović S, Jung G, Rammensee HG. 1991. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*. 351:290–296. doi: 10.1038/351290a0.
- Gouveia-Oliveira R, Sackett PW, Pedersen AG. 2007. MaxAlign: Maximizing usable data in an alignment. *BMC Bioinformatics*. 8:1–8. doi: 10.1186/1471-2105-8-312.
- Grantham R. 1974. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* (80-). 185:862–864. doi: 10.1126/science.185.4154.862.
- Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol*. 52:696–704. doi: 10.1080/10635150390235520.
- Hedrick PW, Thomson G. 1983. Evidence for balancing selection at HLA. *Genetics*. 104:449–456. doi: 10.1038/387138a0.
- Hilton HG et al. 2015. Polymorphic HLA-C Receptors Balance the Functional Characteristics of KIR Haplotypes. *J. Immunol*. 195:3160–3170. doi: 10.4049/jimmunol.1501358.
- Hughes AL, Yeager M. 1998. Natural Selection At Major Histocompatibility Complex Loci of Vertebrates. *Annu. Rev. Genet*. 32:415–435. doi: 10.1146/annurev.genet.32.1.415.
- Jacks T et al. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*. 331:280–283.
- Jurtz V et al. 2017. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol*. 199:3360–3368. doi: 10.4049/jimmunol.1700893.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple

- sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066. doi: 10.1093/nar/gkf436.
- Körner C et al. 2017. HIV-1-Mediated Downmodulation of HLA-C Impacts Target Cell Recognition and Antiviral Activity of NK Cells. *Cell Host Microbe.* 22:111–119.e4. doi: 10.1016/j.chom.2017.06.008.
- Landry C, Garant D, Duchesne P, Bernatchez L. 2001. ‘Good genes as heterozygosity’: The major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). *Proc. R. Soc. B Biol. Sci.* 268:1279–1285. doi: 10.1098/rspb.2001.1659.
- Lenz TL. 2018. Adaptive value of novel MHC immune gene variants. *Proc. Natl. Acad. Sci.* 201722600. doi: 10.1073/pnas.1722600115.
- Lenz TL. 2011. Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution (N. Y.)* 65:2380–2390. doi: 10.1111/j.1558-5646.2011.01288.x.
- Lenz TL, Mueller B, Trillmich F, Wolf JBW. 2013. Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. *Proc. R. Soc. B Biol. Sci.* 280. doi: 10.1098/rspb.2013.0714.
- Lenz TL, Wells K, Pfeiffer M, Sommer S. 2009. Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the long-tailed giant rat (*Leopoldamys sabanus*). *BMC Evol. Biol.* 9:1–13. doi: 10.1186/1471-2148-9-269.
- Leslie AJ et al. 2004. HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.* 10:282–289. doi: 10.1038/nm992.
- Manczinger M et al. Pathogen diversity drives the evolution of generalist antigen presentation in human populations. *PLoS Biol.* (In press).
- Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. 2017. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* 18. doi: 10.1186/s13059-017-1207-1.
- McLaren PJ et al. 2015. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci.* 112:14658–14663. doi: 10.1073/pnas.1514867112.
- Mellors JW et al. 1996. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science (80-.)*. 272:1167–1170. doi: 10.1126/science.272.5265.1167.
- Meyer D, C. Aguiar VR, Bitarello BD, C. Brandt DY, Nunes K. 2018. A genomic perspective on HLA evolution. *Immunogenetics.* 70:5–27. doi: 10.1007/s00251-017-1017-3.
- Migueles SA et al. 2000. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci.* 97:2709–2714. doi: 10.1073/pnas.050567397.
- Neefjes J, Jongsma ML, Paul P, Bakke O. 2011. Towards a systems understanding of

- MHC class I and MHC class II antigen presentation. *Nat Rev Immunol.* 11:823–36. doi: 10.1038/nri3084.
- Neefjes J, Ovaa H. 2013. A peptide's perspective on antigen presentation to the immune system. *Nat. Chem. Biol.* 9:769–775. doi: 10.1038/nchembio.1391.
- Neff BD, Garner SR, Heath JW, Heath DD. 2008. The MHC and non-random mating in a captive population of Chinook salmon. *Heredity (Edinb).* 101:175–185. doi: 10.1038/hdy.2008.43.
- Nei M, Hughes AL. 1991. Polymorphism and evolution of the major histocompatibility complex loci in mammals. *Evol. Mol. Lev.* 222–247.
- Niskanen AK et al. 2014. Balancing selection and heterozygote advantage in major histocompatibility complex loci of the bottlenecked Finnish wolf population. *Mol. Ecol.* 23:875–889. doi: 10.1111/mec.12647.
- Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics.* 20:289–290. doi: 10.1093/bioinformatics/btg412.
- Parham P. 2005. MHC class I molecules and KIRS in human history, health and survival. *Nat. Rev. Immunol.* 5:201–214. doi: 10.1038/nri1570.
- Parham P, Norman PJ, Abi-Rached L, Guethlein LA. 2012. Human-specific evolution of killer cell immunoglobulin-like receptor recognition of major histocompatibility complex class I molecules. *Philos. Trans. R. Soc. B Biol. Sci.* 367:800–811. doi: 10.1098/rstb.2011.0266.
- Penn DJ, Damjanovich K, Potts WK. 2002. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc. Natl. Acad. Sci. U. S. A.* 99:11260–11264. doi: 10.1073/pnas.162006499.
- Pereyra F et al. 2014. HIV Control Is Mediated in Part by CD8+ T-Cell Targeting of Specific Epitopes. *J. Virol.* 88:12937–12948. doi: 10.1128/JVI.01004-14.
- Pierini F, Lenz TL. 2018. Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection. *Mol. Biol. Evol.* 1–14. doi: 10.1093/molbev/msy116.
- Piertney SB, Oliver MK. 2006. The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb).* 96:7–21. doi: 10.1038/sj.hdy.6800724.
- Pohlmeyer CW, Buckheit RW, Siliciano RF, Blankson JN. 2013. CD8+T cells from HLA-B*57 elite suppressors effectively suppress replication of HIV-1 escape mutants. *Retrovirology.* 10:27–32. doi: 10.1186/1742-4690-10-152.
- Potts WK, Wakeland EK. 1990. Evolution of diversity at the major histocompatibility complex. *Trends Ecol. Evol.* 5:181–187. doi: [http://dx.doi.org/10.1016/0169-5347\(90\)90207-T](http://dx.doi.org/10.1016/0169-5347(90)90207-T).
- Rajagopalan S, Long EO. 2005. Understanding how combinations of HLA and KIR genes influence disease. *J. Exp. Med.* 201:1025–1029. doi: 10.1084/jem.20050499.

- Rao X, Hoof I, Fontaine Costa AICA, Van Baarle D, Keşmir C. 2011. HLA class I allele promiscuity revisited. *Immunogenetics*. 63:691–701. doi: 10.1007/s00251-011-0552-6.
- Reche PA, Reinherz EL. 2003. 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. doi: 10.1016/S0022-2836(03)00750-2.
- Richman AD, Herrera LG, Nash D. 2001. MHC class II beta sequence diversity in the deer mouse (*Peromyscus maniculatus*): Implications for models of balancing selection. *Mol. Ecol.* 10:2765–2773. doi: 10.1046/j.0962-1083.2001.01402.x.
- Robinson J et al. 2017. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles Keating, BJ, editor. *PLOS Genet.* 13:e1006862. doi: 10.1371/journal.pgen.1006862.
- Robinson J et al. 2015. The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* 43:D423–D431. doi: 10.1093/nar/gku1161.
- Schwensow N, Eberle M, Sommer S. 2010. Are there ubiquitous parasite-driven major histocompatibility complex selection mechanisms in gray mouse lemurs? *Int. J. Primatol.* 31:519–537. doi: 10.1007/s10764-010-9411-9.
- Solberg OD et al. 2008. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Hum. Immunol.* 69:443–464. doi: 10.1016/j.humimm.2008.05.001.
- Sommer S. 2005. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* 2:1–18. doi: 10.1186/1742-9994-2-16.
- Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. B Biol. Sci.* 277:979–988. doi: 10.1098/rspb.2009.2084.
- Stenzel A et al. 2004. Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum. Genet.* 114:377–385. doi: 10.1007/s00439-003-1075-5.
- Stutz WE, Bolnick DI. 2017. Natural selection on MHC II β in parapatric lake and stream stickleback: Balancing, divergent, both or neither? *Mol. Ecol.* 26:4772–4786. doi: 10.1111/mec.14158.
- Takehima S et al. 2008. Evidence for cattle major histocompatibility complex (BoLA) class II DQA1 gene heterozygote advantage against clinical mastitis caused by *Streptococci* and *Escherichia* species. *Tissue Antigens.* 72:525–531. doi: 10.1111/j.1399-0039.2008.01140.x.
- Trowsdale J. 2011. The MHC, disease and selection. *Immunol. Lett.* 137:1–8. doi: 10.1016/j.imlet.2011.01.002.
- Trowsdale J, Knight JC. 2013. Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genomics Hum. Genet.* 14:301–323. doi: 10.1146/annurev-

- genom-091212-153455.
- Wakeland E et al. 1990. Ancestral polymorphisms of MHC class II genes: Divergent allele advantage. *Immunol. Res.* 9:115–122.
- Wickham H, Wickham MH. 2007. The ggplot package.
- York IA et al. 2002. The Er aminopeptidase ERAP I enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat. Immunol.* 3:1177–1184. doi: 10.1038/ni860.
- Zipeto D, Beretta A. 2012. HLA-C and HIV-1 : friends or foes ? 1–9.

Annex II

Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy

Diego Chowell^{1,2}, Chirag Krishna³, Federica Pierini⁴, Vladimir Makarov^{1,2}, Naiyer A. Rizvi⁵, Fengshen Kuo², Nadeem Riaz^{2,6}, Tobias L. Lenz⁴, Timothy A. Chan^{1,2,6,7}

¹Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA, ²Immunogenomics and Precision Oncology Platform, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

³Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA, ⁴Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany

⁵Dept. of Medicine, Columbia University Medical Center, New York, NY, USA,

⁶Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA, ⁷Weill Cornell School of Medicine, New York, NY 10065, USA

Submitted manuscript

Abstract

Functional diversity of major histocompatibility complex class I (MHC-I) molecules, encoded by the highly polymorphic human leukocyte antigen class I (HLA-I) genes, underlies successful immunologic control of both infectious disease and cancer. The divergent allele advantage hypothesis dictates that a HLA-I genotype with two alleles whose sequences are more divergent enables presentation of a more diverse immunopeptidome. However, the effect of sequence divergence between HLA-I alleles—a quantifiable measure of HLA-I evolution—on the efficacy of immune checkpoint inhibitor (ICI) treatment for cancer remains unknown. Here, we determined the germline HLA-I evolutionary divergence (HED) of patients with melanoma and non-small cell lung cancer treated with ICI, by quantifying the physiochemical sequence divergence between HLA-I alleles of each patient's genotype. Strikingly, HED was a strong determinant of survival after ICI. Even among patients fully heterozygous at HLA-I, patients with high HED respond better to ICI than patients with low HED. We further show that HED strongly impacts the diversity of tumor and human immunopeptidomes and intratumoral TCR clonality. Much like tumor mutation burden, HED is a fundamental metric of diversity at the MHC-peptide complex, which influences ICI efficacy. Our data link divergent HLA allele advantage to immunotherapy efficacy and unveil how ICI response relies on the evolved efficiency of HLA-mediated immunity.

Main

Checkpoint blockade immunotherapy has revolutionized the treatment of patients with advanced-stage cancers. Agents of ICI include monoclonal antibodies that target cytotoxic T lymphocyte–associated protein 4 (CTLA-4) or programmed cell death protein 1 (PD-1) or its ligand (PD-L1) (Callahan et al. 2016). However, durable benefit from these agents is limited to a minority of patients. Among the critical determinants of ICI response is tumor mutational burden, a proxy for the number of tumor-derived neoantigens that can be presented on the cell surface by MHC molecules (Snyder et al. 2014; Rizvi et al. 2015; Van Allen et al. 2015; McGranahan et al. 2016; Goodman et al. 2017; Le et al. 2017; Yarchoan et al. 2017; Samstein et al. 2019). These neoantigens facilitate anti-tumor immunity through recognition by cytotoxic T-cells (van Rooij et al. 2013; Snyder et al. 2014; Rizvi et al. 2015; Schumacher and Schreiber 2015; Tran et al. 2015; Tran et al. 2016). Another related genetic factor that determines the success of ICI is heterozygosity at the highly polymorphic HLA-I loci (Chowell et al. 2018). It has been suggested that compared to homozygotes, heterozygous HLA-I genotypes can facilitate presentation of a more diverse set of tumor antigens to T-cells (Chowell et al. 2018). This concept, known as HLA-I heterozygote advantage, is well supported by studies in the context of infectious diseases where presentation of a more diverse set of pathogen-derived antigens facilitates pathogen recognition (Doherty and Zinkernagel 1975a, b; Thursz et al. 1997; Carrington et al. 1999; Penn et al. 2002).

The primary function of the major histocompatibility complex class I (MHC-I) molecules, which are encoded by HLA-I, is the presentation of self and non-self peptides for recognition by cytotoxic T-cells (Parham and Ohta 1996; Hughes and Yeager 1998). Each individual's HLA-I genotype consists of a pair of alleles at each of the classical class I genes—HLA-A, B, and C, and their polymorphism is concentrated within their peptide-binding domains (Hughes and Nei 1988; Parham 1988; Hughes and Yeager 1998; Robinson et al. 2017). The set of peptides bound by each MHC-I molecule is collectively referred to as its immunopeptidome, and different HLA-I alleles have different peptide-binding specificities with varying degrees of overlap according to the amount of physiochemical sequence divergence between them (Rammensee et al. 1995; Parham and Ohta 1996; Sette and Sidney 1999; Paul et al. 2013; Gfeller and Bassani-Sternberg 2018; La Gruta et al. 2018). The concomitant diversity of HLA-I genotypes and peptide-binding specificities yields marked variability in the diversity of peptide repertoires that

can be displayed by the MHC-I complex across individuals (Parham and Ohta 1996; Paul et al. 2013). Accordingly, this variation may affect the ability of each individual's immune system to recognize tumor antigens and consequently may influence response to ICI.

Motivated by the divergent allele advantage proposed three decades ago (Wakeland et al. 1990; Parham and Ohta 1996), here we hypothesized that the effect of HLA-I heterozygosity on response to ICI may be modulated by the amount of sequence divergence between the peptide-binding domains of a patient's two HLA-I alleles. High sequence divergence between the alleles' peptide-binding domains strongly affects the combined peptide-binding properties of the corresponding MHC-I molecules (Potts and Wakeland 1990; Wakeland et al. 1990; Lenz 2011; Pierini and Lenz 2018). Thus, heterozygous patients with more divergent alleles may present a broader set of peptides for T cell recognition than heterozygous patients with less divergent HLA-I alleles (Wakeland et al. 1990; Lenz 2011; Pierini and Lenz 2018).

We first determined HLA-I evolutionary divergence (HED) using HLA-I genotypes across multiple patient cohorts with metastatic melanoma or non-small cell lung cancer (NSCLC) treated with CTLA-4 blockade or PD-1/PD-L1 blockade (**Fig. 1a**). For each patient, we calculated HED at each of HLA-A, HLA-B, and HLA-C by measuring the Grantham distance (Grantham 1974; Pierini and Lenz 2018) between the peptide-binding domains of the two alleles. The Grantham distance is a classic metric that allows quantification of physiochemical differences between protein amino acid sequences, taking into account composition, polarity, and volume. To explore the landscape of HEDs in our dataset, we performed hierarchical clustering of HED per locus for all pairwise allele combinations across HLA-A, HLA-B, and HLA-C. Hierarchical clustering of HEDs from each locus demonstrated distinct clusters of high and low divergence between alleles (**Fig. 1b, Extended Data Fig. 1 a-c**), as expected and consistent with known relationships between HLA-A, HLA-B, and HLA-C loci^{26,35}. It also showed that HLA-B pairwise divergences are higher relative to HLA-A and HLA-C (**Fig. 1c**), consistent with prior reports that HLA-B is the oldest and most diverse of the three HLA-I loci (McKenzie et al. 1999; Robinson et al. 2017). Moreover, HLA-C alleles had the lowest pairwise divergences, in line with prior studies that HLA-C has evolved most recently (McKenzie et al. 1999; Robinson et al. 2017) (**Fig. 1c**). For each patient, we next calculated the mean HED as the mean of the three pairwise divergences of HLA-A, HLA-B, and HLA-C.

Mean HED distributions in patients from our cohorts were similar to those observed in the TCGA cohorts (**Fig. 1d,e**). A prior comparison of the Grantham distance to other common metrics of sequence divergence showed that the Grantham distance best captured the functional properties of HLA-I molecules (Pierini and Lenz 2018). The Grantham distance is a well-recognized metric that has been applied to measure amino acid polymorphism in many studies of comparative evolution, cancer, infectious disease, and immunity (Subramanian and Kumar 2006; Wain et al. 2007; Grueber et al. 2014; International Wheat Genome Sequencing 2014; Rentoft et al. 2016; Sundaram et al. 2018; Feng et al. 2019). Taken together, these data verify that the Grantham distance is a suitable measure of HLA-I polymorphism in our patient cohorts.

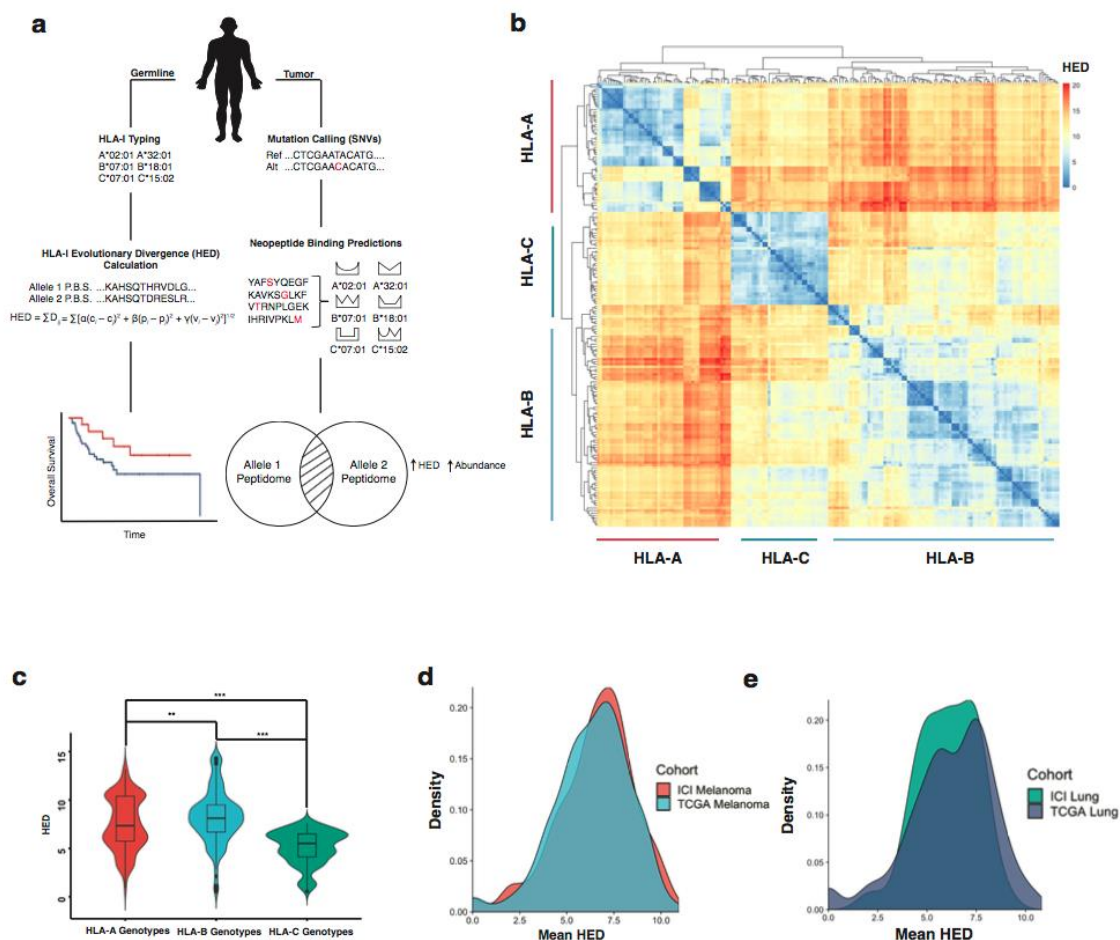


Fig. 1 | Landscape of HLA-I evolutionary divergences at HLA-A, B, and C. **a**, Schematic of experimental design. HLA-I evolutionary divergences (HED) are calculated between peptide-binding domains using the Grantham distance and used to stratify patients treated with immune

checkpoint inhibitors. Predicted neopeptides are called using whole-exome sequencing from the patient's tumor, counted, and correlated with genotype divergence. **b**, Hierarchical clustering of HED at HLA-A, HLA-B, and HLA-C (HLA-I). Heat map shows z-score normalized HED across all alleles across all patient cohorts used for downstream analyses. **c**, Comparison of HED at HLA-A, HLA-B, and HLA-C. ** $P < 0.01$; *** $P < 0.001$; Mann-Whitney test. **d**, Distribution of patient mean HED across all melanoma cohorts treated with ICI (ICI Melanoma) and TCGA (TCGA Melanoma). **e**, Distribution of patient mean HED across all melanoma cohorts treated with ICI (ICI Lung) and TCGA (TCGA Lung).

We next asked whether HED is associated with response to ICI. We stratified patients by mean HED in a cohort of 100 patients with melanoma treated with anti-CTLA-4 (Van Allen et al. 2015) (hereafter called cohort 1), and observed improved overall survival after ICI therapy in patients with high mean HED ($P = 0.0072$, HR = 0.47, 95 % confidence interval (C.I.) = 0.26 – 0.82) (**Extended Data Fig 2a**). We also found that the effect of mean HED on survival was independent of tumor mutational burden (TMB) and other relevant genomic and clinical variables, when these were included in a multivariable Cox regression model of survival (**Extended Data Fig. 2d**). Finally, we found that the effect of both high mean HED and high TMB on overall survival after ICI was more pronounced than the effect of either alone, as reflected by the reduction in hazard ratio when considering both variables (**Extended Data Fig. 2a-c**).

Prior studies of divergent allele advantage have suggested that the diversity of peptide repertoires of fully heterozygous HLA-I genotypes varies with sequence divergence (Wakeland et al. 1990; Parham and Ohta 1996; Pierini and Lenz 2018). Therefore, we hypothesized that even among patients fully heterozygous at HLA-I, response to ICI may also vary with HED. Strikingly, we found that high mean HED was associated with improved survival after ICI in the 78 fully heterozygous patients from cohort 1 (Van Allen et al. 2015) ($P = 0.0094$, HR = 0.43, 95% C.I. = 0.22 – 0.83) (**Fig. 2a**). In a second cohort of 76 fully heterozygous patients with NSCLC treated with anti-PD-1 (Rizvi et al. 2015; Chowell et al. 2018), we also found that high mean HED was associated with better overall survival ($P = 0.049$, HR = 0.32, 95% C.I. = 0.10-1.06) (**Fig. 2b**). We observed the same in an additional third cohort of 95 fully heterozygous patients with metastatic melanoma treated with anti-PD-1 (Zehir et al. 2017; Chowell et al. 2018) ($P = 0.025$) (**Fig. 2c**). Beyond survival, clinical response to ICI was also associated with high mean HED when considering all patients (HLA-I homozygotes or heterozygotes) (57.4%

vs. 32.0%, $P = 0.003$, OR = 0.35) (**Fig. 2d**), or only fully heterozygous patients (55.6% vs. 35.3%, $P = 0.03$, OR = 0.44) (**Fig. 2e**) in all cohorts.

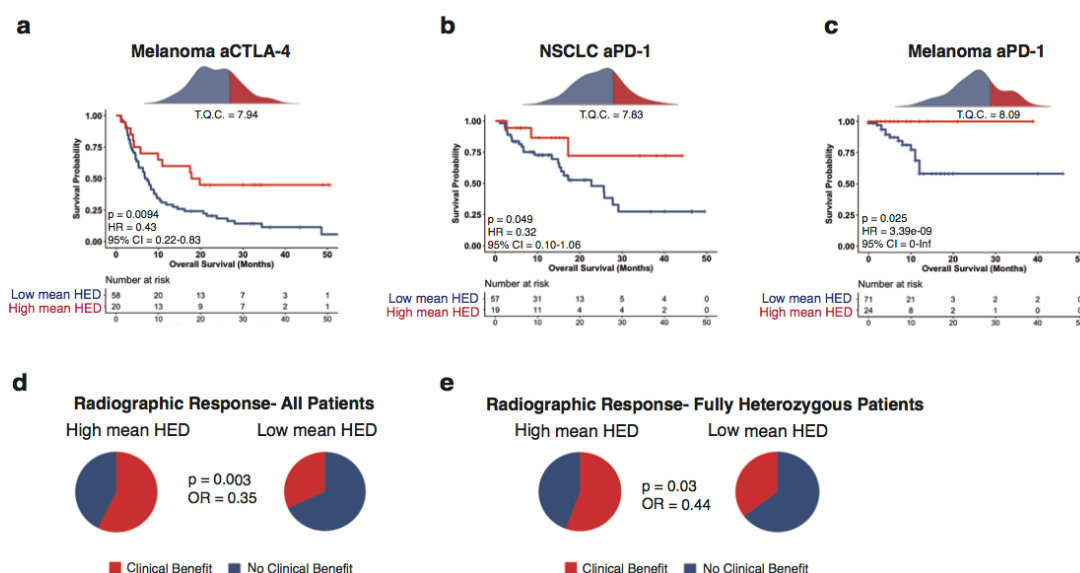


Fig. 2 | High mean HLA-I evolutionary divergence is associated with improved response and survival to immune checkpoint inhibitors. **a**, Association of high mean HLA-I evolutionary divergence (HED) (red) with improved survival after anti-CTLA-4 treatment in a cohort of metastatic melanoma patients fully heterozygous at HLA-I; $P = 0.0094$; log-rank test. Density plots indicate the distribution and cutoff for mean HED used in the survival curves. T.Q.C. = top quartile cutoff, HR=hazard ratio, CI=confidence interval. **b**, Association of high mean HED (red) with improved survival after anti-PD-1 treatment in an independent cohort of patients with non-small cell lung cancer fully heterozygous at HLA-I; $P = 0.049$; log-rank test. **c**, Association of high mean HED (red) with improved overall survival in an independent cohort of patients with melanoma fully heterozygous at HLA-I treated with anti-PD1; $P = 0.025$; log-rank test. **d**, Association of high patient mean HED with clinical response (red) to ICI including all patients (both homozygous and heterozygous at HLA-I) for whom clinical response data were available from Fig. 2a-c; $P = 0.003$; OR = 0.35; Fisher's exact test. **e**, Association of high mean HED with clinical response to (red) ICI including only patients fully heterozygous at HLA-I for whom clinical response data were available from Fig. 2a-c; $P = 0.03$, OR = 0.44; Fisher's exact test.

These data indicate that HED is predictive of survival and response among cancer patients treated with ICI. To determine whether HED might simply reflect a general prognostic factor in cancer, we examined the association of HLA-I heterozygosity or HED with overall survival among patients with melanoma and non-small cell lung cancer who did not receive ICI therapy, and whose tumors were profiled with exome sequencing, and observed no effect (**Extended Data Fig. 3, 4**). This suggests that mean HED is predictive of response to ICI, and not prognostic.

We examined all cohorts from Fig. 2 to investigate the combined effect of mean HED and TMB on response to ICI. First, we found that the effect of mean HED on survival after ICI (**Fig. 3a**) was independent of other clinical variables in multivariable Cox regression analysis (**Extended Data Fig. 5a**), and that high HED did not co-occur with known mutations in genes that have been reported to impact response to ICI (Gao et al. 2016; Zaretsky et al. 2016; Zhao et al. 2016; Patel et al. 2017; Sade-Feldman et al. 2017) (**Extended Data Fig. 6**). We found that the combined effect of high HED and high TMB on overall survival after ICI was stronger than the effect of either alone (**Fig. 3a-c**). This effect was also observed when analyzing only fully heterozygous patients (**Fig. 3d-f, Extended Data Fig. 5b**). Furthermore, this result remained robust across a wide range of cutpoints for HED and TMB (**Fig. 3g**) used to stratify patients into groups for survival analysis. Interestingly, we found that high HED at HLA-A and HLA-B were each associated with improved response to ICI, when considering all patients or only fully heterozygous patients (**Fig. 3h**), suggesting that divergence at individual class I loci may differentially affect ICI efficacy.

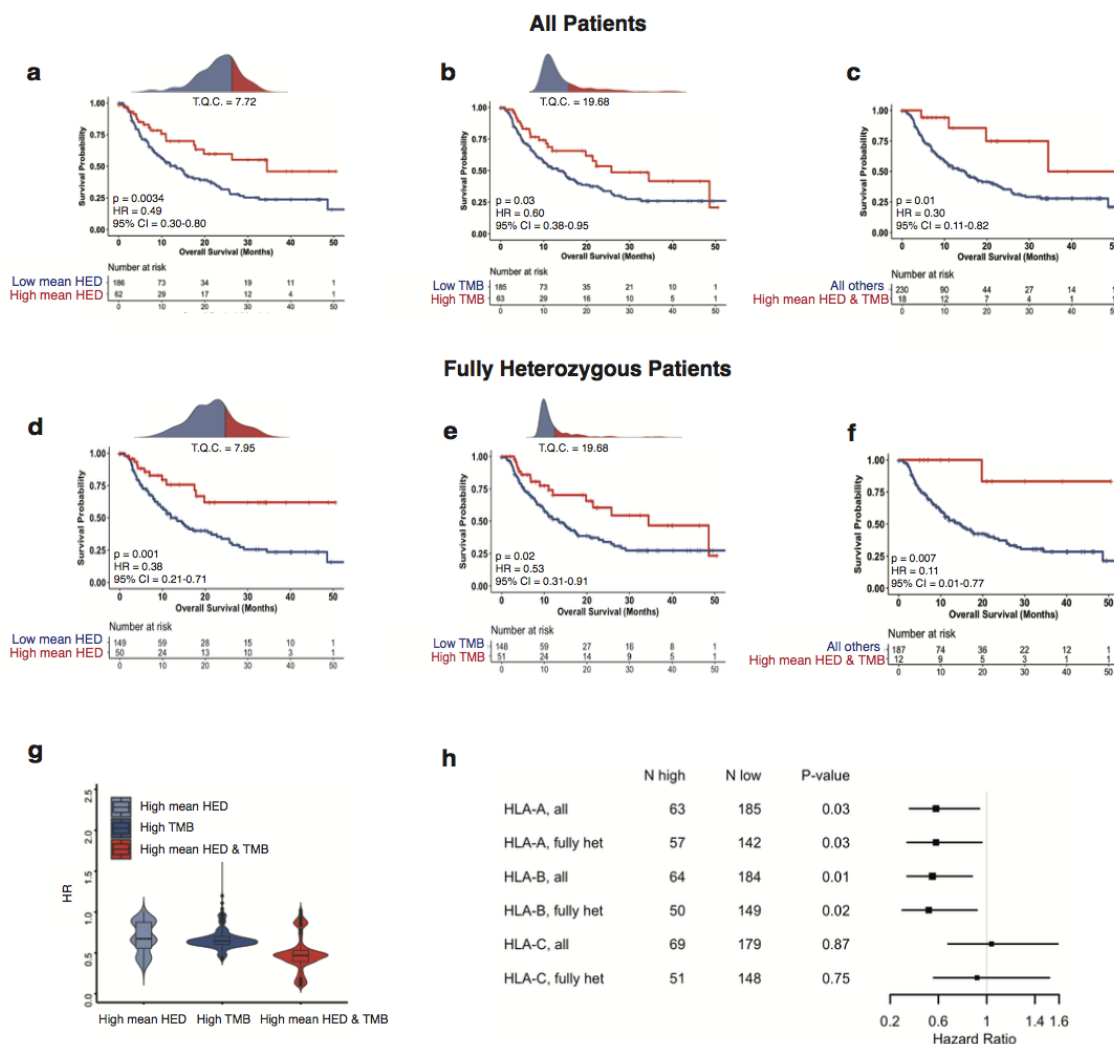


Fig. 3 | Effect of mean HLA-I evolutionary divergence and tumor mutational burden on efficacy of immune checkpoint inhibitor treatment. **a**, Association of high mean HED (red) with improved overall survival after ICI in all patients (HLA-I homozygous or heterozygous) from Fig. 2 for whom tumor mutational burden (TMB) were available; $P = 0.0034$; log-rank test. Density plot indicates the distribution and cutoff for mean HED used in the survival curves. T.Q.C. = top quartile cutoff, HR=hazard ratio, CI=confidence interval. **b**, Association of high TMB (red) with improved overall survival after ICI in patients from Fig. 3a; $P = 0.03$; log-rank test. Density plot indicates the distribution and cutoff for TMB used in the survival curves. T.Q.C. = top quartile cutoff. **c**, Survival of patients with both high mean HED and high TMB (red) after ICI treatment in patients from Fig. 3a; $P = 0.01$; log-rank test **d**, Association of high mean HED (red) with improved overall survival after ICI in patients fully heterozygous at HLA-I from Fig. 2 for whom tumor mutational burden (TMB) were available; $P = 0.001$; log-rank test. **e**, Association of high TMB with improved overall survival after ICI in patients from Fig. 3d; $P = 0.02$; log-rank test. **f**, Survival of patients with both high mean HED and high TMB after ICI treatment in patients from Fig. 3a; $P = 0.007$; log-rank test **g**, Cutpoint analysis showing the association of both high mean HED and high TMB with improved survival after ICI. Data show that the effect is present across

all cutpoints (see Methods). **h**, Univariable Cox regression analysis showing the association of high HED (top quartile) at individual HLA-I loci with improved survival after ICI in patients from Fig. 3a (HLA-I homozygous or heterozygous, “all”) and Fig. 3d (fully heterozygous at HLA-I, “fully het”).

A possible explanation for the improved survival after ICI among patients who were fully heterozygous at HLA-I, and had high HED, is that higher HED may be associated with increased diversity of the neopeptide repertoire presented by HLA-I. Such an effect would lead to a higher probability of T-cell responses after administration of immunotherapy, due to the larger and more diverse immunopeptidome presented to cytotoxic T cells. This hypothesis is supported by the idea that pathogen-driven selection has maintained HLA-I polymorphism worldwide (Prugnolle et al. 2005; Buhler et al. 2016; Pierini and Lenz 2018), and the idea that neoantigen recognition is driven in part by sequence similarity to pathogenic epitopes (Balachandran et al. 2017; Luksza et al. 2017; Kim et al. 2018). In an exploratory analysis limited to patients fully heterozygous at each locus, we found that the number of candidate neopeptides bound by heterozygous genotypes is correlated with mean HED (**Fig. 4a**). Moreover, mean HED was not correlated with tumor mutational burden (**Fig. 4b**), indicating that the diversity in HLA-I peptide-binding domains specifically reflects the diversity of neopeptide repertoire binding to the HLA-I molecules, rather than overall tumor mutation burden. We further detected associations between HED and diversity of the neopeptide repertoire at individual class loci (**Extended Data Fig. 7a-c**).

We next examined whether HED is also associated with human immunopeptidome diversity. We computationally generated all unique peptides of length nine from the entire human proteome using a common reference, and performed HLA-I binding predictions. We found that HED was correlated with diversity of the predicted self immunopeptidome (**Fig. 4c, Extended Data Fig. 7d-f**). We further determined HED in an independent cohort of 12 individuals for which naturally eluted peptide data were available (Pearson et al. 2016), and observed an association between HED at HLA-B and self immunopeptidome diversity, but not at HLA-A (**Extended Data Fig. 8**), recapitulating our findings from peptide binding predictions. Altogether, these data suggest for the first time that increased sequence divergence of a HLA-I genotype is associated with increased diversity of human and tumor immunopeptidomes.

We further investigated whether HED could be associated with the clonality of the intratumoral TCR repertoire. We hypothesized that the association of high HED with a broader neopeptide repertoire would increase the probability of neoantigen recognition by tumor-infiltrating T-cells, and subsequently influence T cell clonal expansion. Accordingly, in a subset of patients treated with ICI therapy for whom next-generation deep sequencing of T cell receptor CDR3 regions (TCR-seq) were available (Riaz et al. 2017), we found a positive correlation between mean HED and clonality of TCR CDR3s (**Fig. 4d**). However, additional patient cohorts with whole-exome and TCR-sequencing will be required to confirm this result.

Taken together, these data show that HLA-I evolutionary divergence—as measured by sequence divergence between alleles of a HLA-I genotype— is associated with response to checkpoint blockade immunotherapy in patients treated for cancer, and the diversity of tumor and human immunopeptidomes. One explanation for these findings is that highly divergent HLA-I genotypes can present more diverse antigenic epitopes (Fig. 4e), which may lead to the expansion of T cell clones that facilitate superior tumor control. In such cases, it may be more challenging for a tumor clone to evade recognition by the host immune system during immunotherapy. Further studies may investigate the effect of HED on tumor evolution (Greaves and Maley 2012) and specificity of the host T cell receptor repertoire (Sharon et al. 2016).

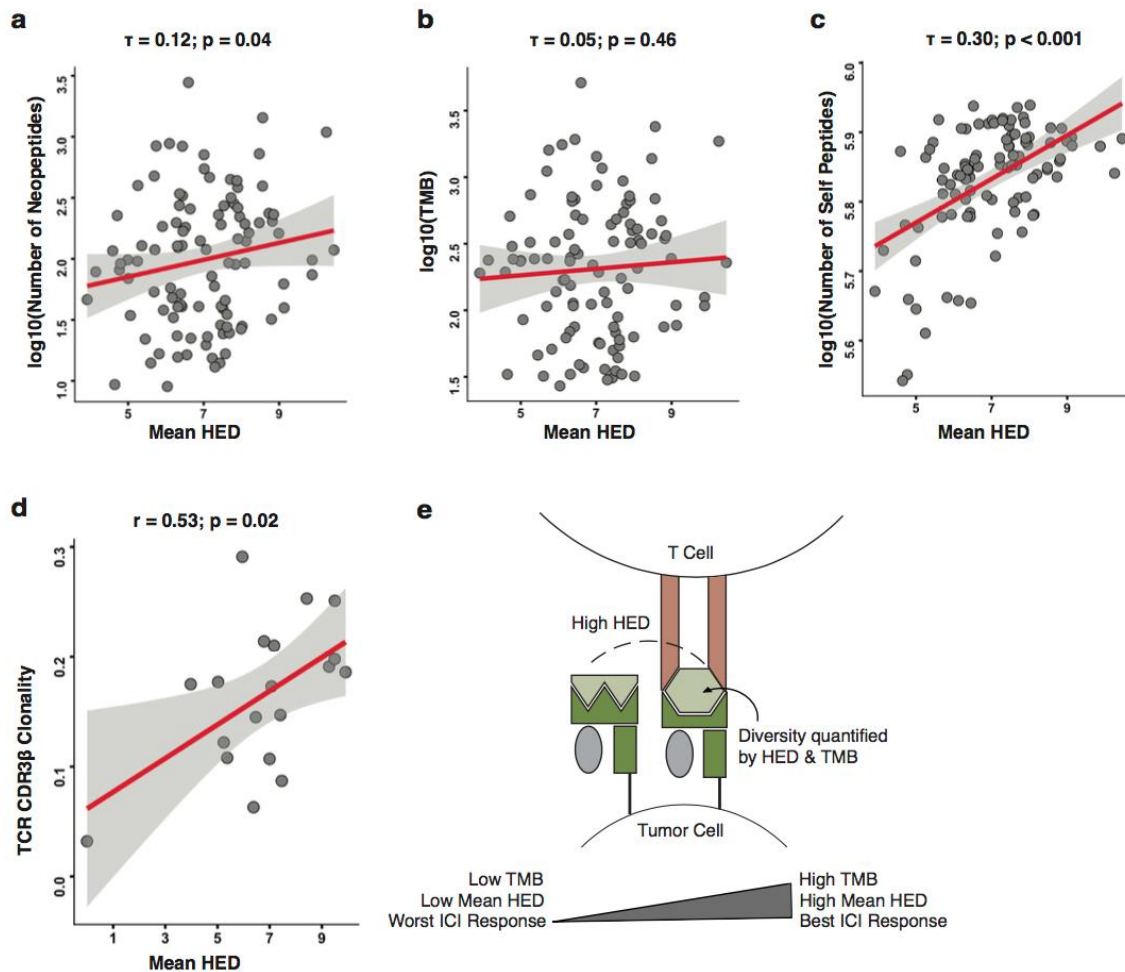


Fig. 4 | Mean HLA-I evolutionary divergence is positively correlated with diversity of the tumor and human immunopeptidome.. **a**, Correlation of mean HED with number of unique neopeptides bound to alleles of each patient genotype using all patients fully heterozygous at HLA-I from Fig. 2 for whom neopeptide data were available; $P = 0.04$; Kendall's rank correlation. Each point represents a patient HLA-I genotype (HLA-A, B, and C); y-axis depicts the mean number of neopeptides bound across HLA-A, B, and C (see Methods). **b**, Correlation of mean HED with TMB; $P = 0.46$; Kendall's rank correlation. **c**, Correlation of mean HED with number of unique self-peptides from the human proteome bound to alleles of each HLA-I genotype ; $P < 0.001$; Kendall's rank correlation. Y-axis depicts the mean number of self-peptides bound across HLA-A, B, and C (see Methods). **d**, Association of mean HED with intratumoral TCR CDR3 β clonality; $P = 0.02$; Pearson's correlation. Red line indicates line of best linear fit. **e**, Schematic depicting the effects of HLA-I evolutionary divergence and TMB on immunopeptidome diversity and response to ICI. One representative HLA-I locus with high HED between alleles is depicted.

Methods

Description of Patient Cohorts

We used four previously published cohorts of patients with late-stage melanoma and non-small cell lung cancer (NSCLC) treated with anti-CTLA-4, or PD-1/PD-L1 blockade (Rizvi et al. 2015; Van Allen et al. 2015; Zehir et al. 2017; Chowell et al. 2018). Ten patients from the Van Allen et al. cohort were excluded, since they achieved long-term survival after anti-CTLA-4 treatment with early tumor progression (Van Allen et al. 2015). The NSCLC data are from patients with metastatic disease treated mainly with anti-PD-1 monotherapy. The patients are from a prospective trial that we reported previously (Rizvi et al. 2015) and from New York-Presbyterian /Columbia University Medical Center (Chowell et al. 2018). In these NSCLC cohorts, only patients with exome sequencing data were included (mutation data not available for 67 patients with NSCLC). All patients were treated under institutional review approved prospective protocols. Clinical characteristics of patient cohorts are provided in the original studies. The TCGA exome data for the patients with melanoma (N = 457) and lung cancer (N = 545) was obtained from The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov>).

Overall Survival and Clinical Response

Overall survival was defined as the length of time from treatment start to time to event (survival or censor). Response data was available for some cohorts (Rizvi et al. 2015; Van Allen et al. 2015; Chowell et al. 2018); clinical benefit was defined as complete response, partial response, or stable disease as indicated in previous studies (Rizvi et al. 2015; Van Allen et al. 2015; Chowell et al. 2018). No clinical benefit was defined as progressive disease. All clinical data, including overall survival and clinical response data, were obtained from the original studies. Clinical data for the TCGA patients with melanoma and non-small cell lung cancer were obtained through the TCGA data portal.

HLA Class I Genotyping

We performed HLA-I genotyping as described previously (Chowell et al. 2018). Briefly, we performed high-resolution HLA class I genotyping from germline normal DNA exome sequencing data directly or using a clinically validated HLA typing assay (LabCorp). Patient exome data or targeted gene panels were obtained and the well-validated tool

Polysolver was used to identify HLA class I alleles with default parameter settings (Shukla et al. 2015). For the 67 patients with NSCLC with no available exome sequencing data, HLA class I typing was done at LabCorp. For quality assurance of HLA-I genotyping using MSK-IMPACT (CLIA-approved hybridization-capture based assay) with melanoma samples from anti-PD1-treated patients, we compared HLA class I typing by Polysolver between 37 samples that we sequenced with MSK-IMPACT and whole exome. The MSK-IMPACT panel successfully captured HLA-A, -B, and -C reads and validation was previously performed (Chowell et al. 2018). The overall concordance of class I typing between the MSK-IMPACT samples and their matched WES samples was 96%. To make sure that HLA class I genes have adequate coverage in MSK-IMPACT bam files, we also applied bedtools multicov tool (<http://bedtools.readthedocs.io/en/latest/content/tools/multicov.html>), which reports the count of alignments from multiple position-sorted and indexed BAM files that overlap with targets intervals in a BED format. Only high quality reads were counted and only samples with sufficient coverage were used.

Calculation of Patient HLA-I Evolutionary Divergence (HED)

HLA-I genotype divergence was calculated as described in Pierini and Lenz (Pierini and Lenz 2018). Briefly, we first extracted the protein sequence of exons 2 and 3 of each allele of each patient's HLA-I genotype, which correspond to the peptide-binding domains. Protein sequences were obtained from the IMGT/HLA database (Robinson et al. 2015), and exons coding for the variable peptide-binding domains were selected following the annotation obtained from Ensembl database (Zerbino et al. 2018). Sequence divergences between allele sequences were calculated using the Grantham distance metric (Grantham 1974), as implemented in Pierini and Lenz (Pierini and Lenz 2018). The Grantham distance is a quantitative pairwise distance in which the physiochemical properties of amino acids, and hence the functional similarity between sequences are considered (Grantham 1974). Given a particular HLA-I locus with two alleles, the sequences of the peptide-binding domains of each allele are aligned (Edgar 2004), and the Grantham distance is calculated as the sum of amino acid differences (taking into account the biochemical composition, polarity, and volume of each amino acid) along the sequences of the peptide-binding domains, following the formula by R. Grantham (Grantham 1974):

$$(1) \text{ Grantham Distance} = \sum D_{ij} = \sum [\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2]^{1/2}$$

where i and j are the two homologous amino acids at a given position in the alignment, c , p , and v represent composition, polarity, and volume of the amino acids respectively, and α , β , and γ are constants; all values are taken from the original study (Grantham 1974). The final Grantham distance is calculated by normalizing the value from (1) by the length of the alignment between the peptide-binding domains of a particular HLA-I genotype's two alleles. A prior analysis of multiple common sequence divergence measures showed that the correlation of Grantham distance with the number of peptides bound by both alleles of a heterozygous genotype exceeded that of the other distance measures (Pierini and Lenz 2018). Patient mean HED was calculated as the mean of divergences at HLA-A, HLA-B, and HLA-C.

Tumor Mutational Analysis Pipeline

For cohorts which received whole exome sequencing, reads in FASTQ format were aligned to the reference human genome GRCh37 using the Burrows–Wheeler aligner (BWA; v0.7.10) (Li and Durbin 2009). Local realignment was performed using the Genome Analysis Toolkit (GATK 3.7) (McKenna et al. 2010). Duplicate reads were removed using Picard version 2.13. To identify somatic single nucleotide variants (SNVs), we used a validated pipeline that integrates mutation calls from four different mutation callers: MuTect 1.1.7, Strelka 1.0.15, SomaticSniper 1.0.4, and VarScan 2.4.3 (Koboldt et al. 2012; Larson et al. 2012; Saunders et al. 2012; Cibulskis et al. 2013). SNVs with an alternative allele read count of less than 4, total coverage of less than 10 or with corresponding normal coverage of less than 7 reads were filtered out.

Computational Identification of HLA-I Restricted Neopeptides

Each non-synonymous SNV was translated into a 17-mer peptide sequence, centered on the mutated amino acid. Adjacent SNVs were corrected using MAC (Wei et al. 2015). Subsequently, the 17-mer was then used to create 9-mers via a sliding window approach for determination of HLA-I binding predictions for neopeptides using NetMHCpan-4.0 (Jurtz et al. 2017). All peptides with a rank <2% were considered for further analyses.

Computational Identification of HLA-I Restricted Peptides from the Human Proteome

We identified peptides from the entire human proteome that bind to patient-specific HLA-I alleles. The human peptidome was downloaded from Ensembl (Zerbino et al. 2018) (ftp://ftp.ensembl.org/pub/grch37/update/fasta/homo_sapiens/pep//Homo_sapiens.GRCh37.pep.all.fa). Only sequences annotated as `gene_biotype:protein_coding` and `transcript_biotype:protein_coding` were kept. Transcripts with identical sequences were de-duplicated. The resulting FASTA file was submitted to NetMHCpan-4.0 (Jurtz et al. 2017) to determine HLA-I binding predictions. All peptides from the human proteome with a rank <2% were considered for further analyses. For the peptide-divergence correlation analyses in Extended Data Fig. 8, we used self-peptides identified via mass spectrometry and HLA-I genotypes from Pearson et al (Pearson et al. 2016). All correlation analyses were limited to peptides of length 9.

TCR β -Chain Sequencing and Analysis

We employed next-generation sequencing of TCR β -chain complementarity determining regions (CDR3s) (TCR-seq) (Adaptive Biotechnologies) (Robins et al. 2009; Carlson et al. 2013) from a subset of tumor samples collected pre-therapy from responders (CR/PR/SD) in the Riaz et al cohort (Riaz et al. 2017). We subsequently calculated the clonality of the TCR CDR3 repertoire, defined as the complement of evenness (i.e., $1 - \text{evenness}$). Evenness is defined as the observed Shannon entropy (H) divided by the maximum possible H , given the number of unique elements in a population. Data for individual TCR sequences, were obtained from Adaptive Biotechnologies for customized analysis of T cell repertoire. Correlation analyses were performed using Pearson's r .

Genomic Oncoprint

The oncoprint displays mutated genes that have been reported to impact response to ICI. The genes in the IFNG gene cluster on 9p are: IFNA1, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA2, IFNA21, IFNA22P, IFNA4, IFNA, IFNA6, IFNA7, IFNA8, IFNB1, IFNE, IFNW1. The Loss events were identified in the following manner: 1. Rounded FACETS ploidy value to the nearest whole number; 2. Used rounded ploidy value to correct Total Copy Number (tcn.em) with: $\text{Corrected_TCN} = \text{tcn.em} - \text{rounded_ploidy}$; 3. If $\text{Corrected_TCN} \leq -1$, then marked as a "Loss" Event. Note: this

computation was performed for each FACETS (Shen and Seshan 2016) segment on chromosome 9 and was assigned to individual genes with coordinates within the FACETS segment. Homozygous loss events were identified if $tcn.em = 0$ (did not use ploidy-corrected TCN). All losses were manually verified. For assessing loss of heterozygosity (LOH) of HLA class I, copy number variation analysis was performed using FACETS 0.5.6 to determine allele specific copy number (Shen and Seshan 2016). Segments within the chromosome 6p locus were identified containing the HLA-A, HLA-B and HLA-C loci. Loss of heterozygosity (LOH) was defined as a minor allele copy number estimate of 0 for any of the HLA-I loci using the expectation-maximization model (Shen and Seshan 2016).

Correlation Analyses

We limited all HLA-I evolutionary divergence (HED)- peptide correlation analyses to patients heterozygous at each locus only. For Fig. 4a and c, the y-axis shows the mean number of neo- or self-peptides bound uniquely to each allele for each of HLA-A, B, and C. For the analysis shown in Extended Data Fig. 7f, two patients had an HLA-C genotype (C*03:03, C*03:04) that bound a total of 0 neopeptides. These patients were excluded from the plot for visualization purposes only. The correlation displayed in Extended Data Fig. 7f is significant regardless of whether these patients are included. We used the nonparametric Kendall correlation as shown in Pierini and Lenz (2018), since parameters were not normally distributed. For Fig. 4a and Extended Data Fig. 7a-c, we used one-sided p-values—given the prior association of genotype divergence with diversity of nonself pathogenic peptides shown by Pierini and Lenz (2018); we hypothesized that a similar association would be observed for the neopeptide correlations. For Fig. 4c and Extended Data Fig. 7d-f, we had no prior hypothesis regarding the direction of the association between genotype divergence and diversity of self-peptides from the human proteome; thus, two-sided p-values were used.

Statistical Analyses

Comparisons of HED distributions across individual HLA-I loci were calculated using the Mann-Whitney test. Survival analyses were performed using the Kaplan-Meier estimator. All cutoffs for germline HLA-I evolutionary divergence (HED) and tumor mutational burden (TMB) were determined using the top quartile. For analyses in Fig. 3 combining cohorts with whole-exome and targeted panel sequencing, the TMB of the

whole-exome cohorts was divided by 30 to normalize per megabase (Zehir et al. 2017). We performed the survival analysis in Extended Data Fig. 2a using the mean of divergences at HLA-A, B, and C as well as the sum, median, and geometric mean. Results were similar across all metrics used (Extended Data Table 1). For the analysis in Fig. 3g, we used each value of mean HED in the data as a cutpoint for high HED, and did the same for TMB. When combining mean HED and TMB, patients were in the high group if their mean HED and TMB were both greater than the cutpoints for mean HED and TMB, and in the low group if either variable was less than its respective cutpoint (i.e. a patient with high HED but low TMB could be in the low group). For all multivariable analyses, P-values and hazard ratios were calculated using the Cox regression. For all survival analyses, P-values were calculated using the log-rank test, and hazard ratios were calculated using the univariable Cox regression. For the analyses of clinical response data, P-values and odds ratios (OR) were calculated using Fisher's exact test (two-sided). All survival and correlation analyses were performed in the R Statistical Computing Environment version 3.5.0 (<http://www.r-project.org>)

Data availability

The data from prior studies are available at the following accession numbers: dbGAP, phs000452.v2.p1; dbGAP, phs000980.v1.p1; SRA, PRJNA419415, PRJNA419422, and PRJNA419530; TCGA data are available from the cBioPortal for Cancer Genomics, <http://cbioportal.org/msk-impact>, <http://cancergenome.nih.gov>.

Acknowledgements

We thank the Chan lab and members of the Immunogenomics and Precision Oncology Platform for advice and input. This work was supported in part by NIH R35 CA232097 (T.A.C.), NIH RO1 CA205426 (T.A.C., N.A.R.), the PaineWebber Chair (T.A.C.), NIH/NCI Cancer Center Support Grant (P30 CA008748), and the Deutsche Forschungsgemeinschaft grant LE 2593/3-1 (T.L.L.).

Author information

These authors contributed equally: Diego Chowell, Chirag Krishna and Federica Pierini.

These authors jointly directed this work: Tobias L. Lenz and Timothy A. Chan.

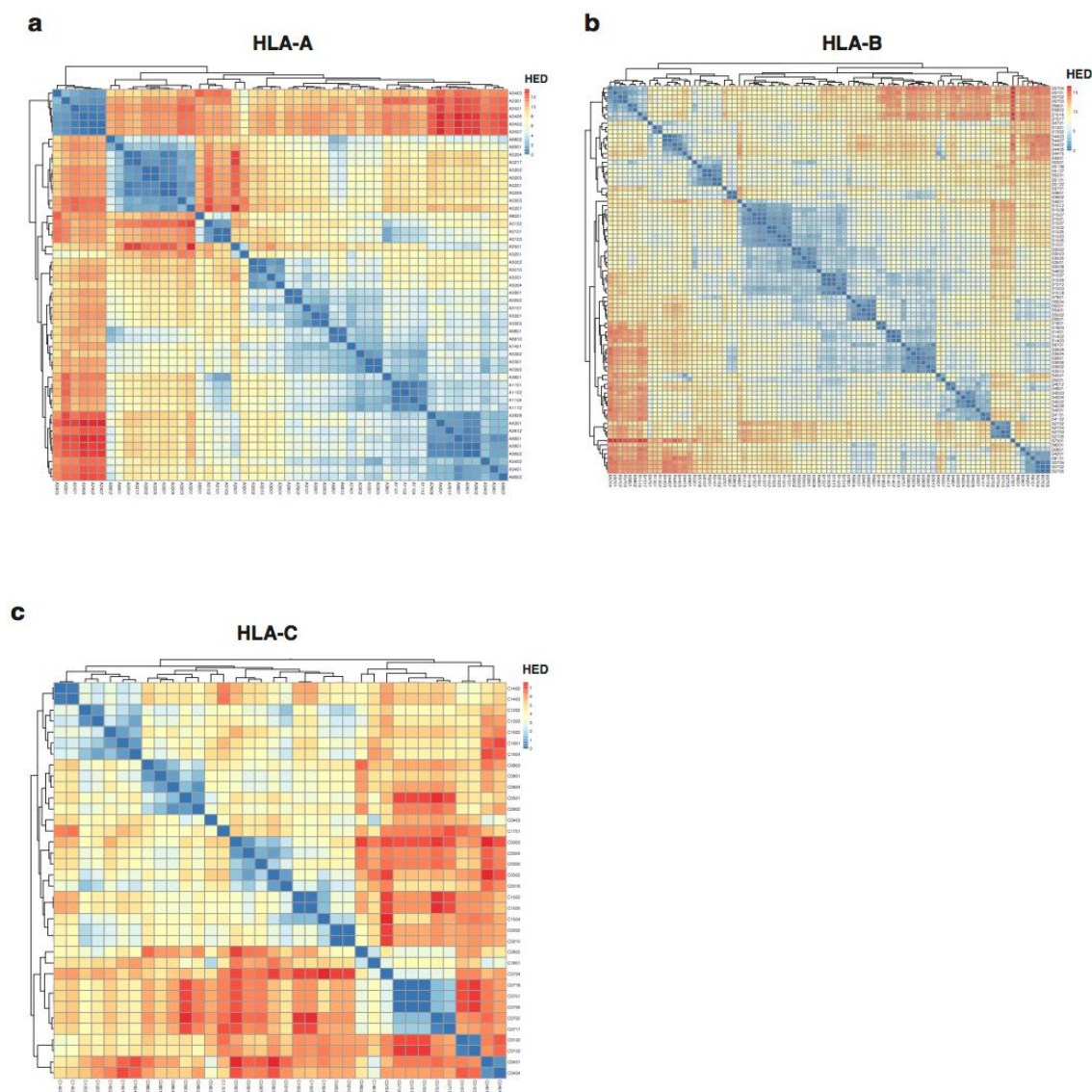
Contributions

Data acquisition and analyses were performed by D.C., C.K., F.P., V.M., T.L.L., T.A.C., F.K., N.R.. The manuscript was written by D.C., C.K., F.P., T.L.L., and T.A.C. with input from all authors.

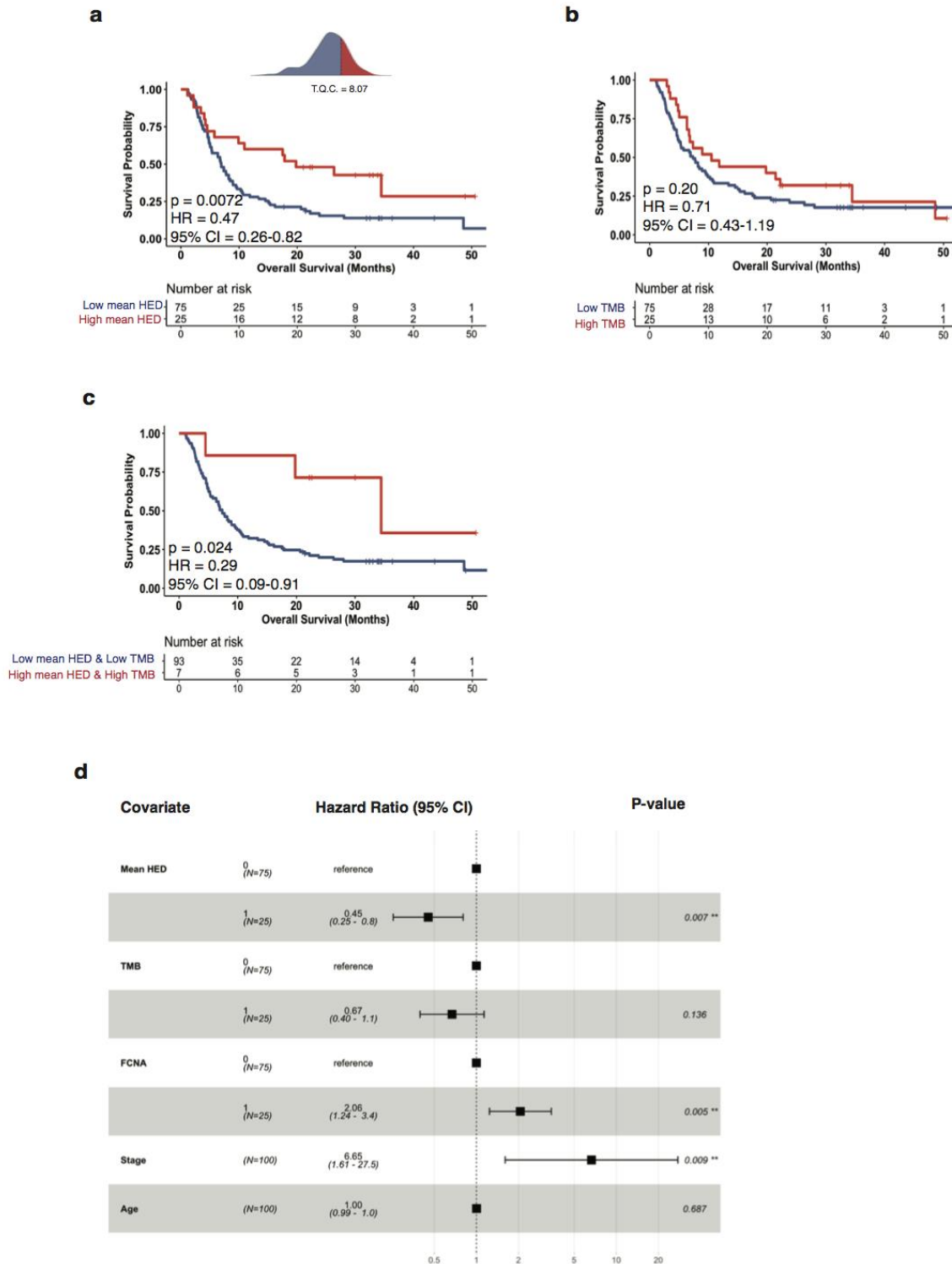
Competing Interests

T.A.C. is a co-founder of Gritstone Oncology and holds equity. T.A.C. holds equity in An2H. T.A.C. acknowledges grant funding from Bristol-Myers Squibb, AstraZeneca, Illumina, Pfizer, An2H, and Eisai. T.A.C. has served as an advisor for Bristol-Myers, MedImmune, Squibb, Illumina, Eisai, AstraZeneca, and An2H. T.A.C. and D.C. hold ownership of intellectual property on using tumor mutation burden to predict immunotherapy response, with pending patent, which has been licensed to PGDx.

Extended Data

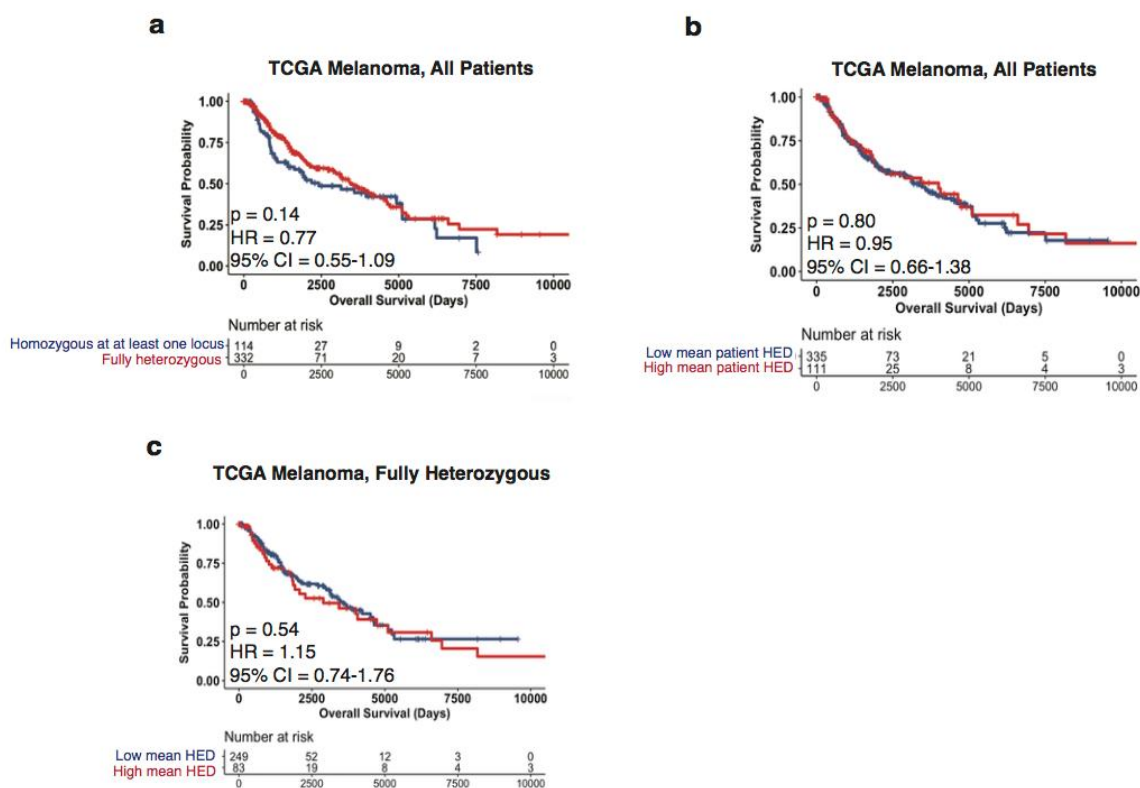


Extended Data Fig. 1 | Hierarchical clustering of HLA-I evolutionary divergences at individual HLA class I loci. a, Hierarchical clustering of HED at HLA-A using all HLA-A alleles from all patient cohorts used for downstream analyses. **b,** Hierarchical clustering of HED at HLA-B using all HLA-B alleles. **c,** Hierarchical clustering of HED at HLA-C using all HLA-C alleles. Heat maps shows z-score normalized HED across all alleles. Red indicates high HED; blue indicates low HED.

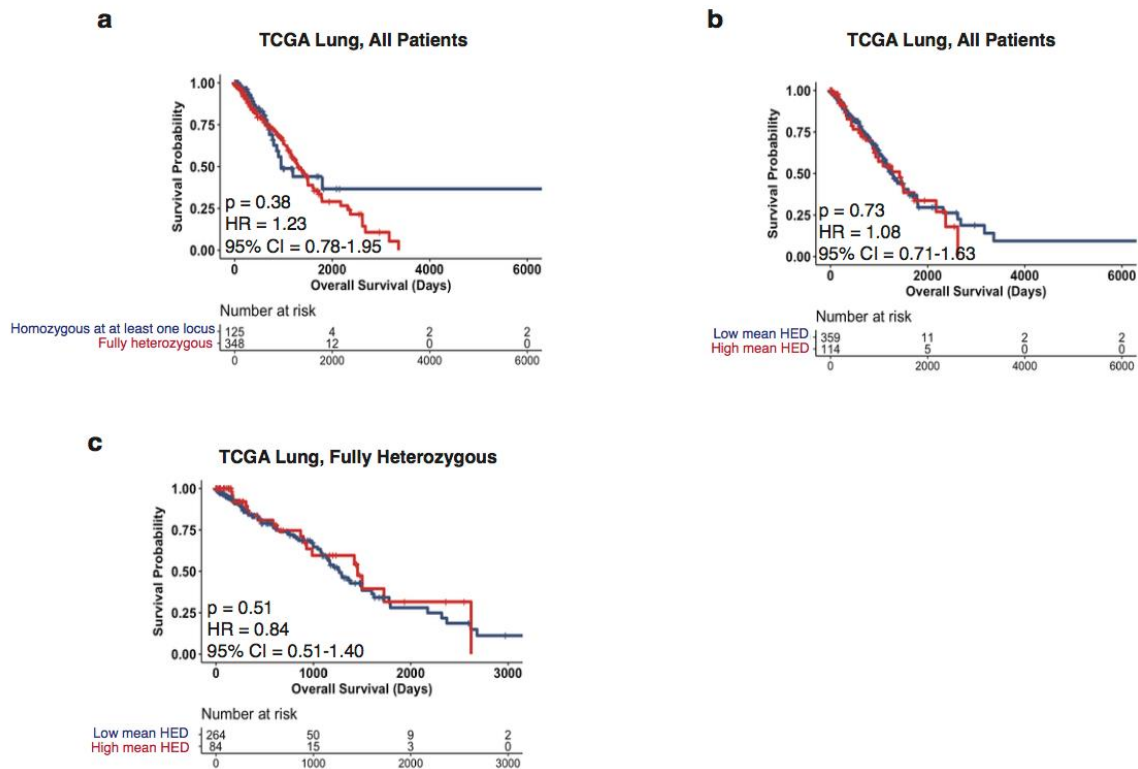


Extended Data Fig. 2 | Mean HLA-I evolutionary divergence is associated with improved response to immune checkpoint inhibitors. **a**, Association of high mean HED (red) with improved efficacy of anti-CTLA-4 treatment in a cohort of patients with metastatic melanoma; $P = 0.0072$; log-rank test. Density plots indicate the distribution and cutoff used in the survival curves.

T.Q.C. = top quartile cutoff, HR=hazard ratio, CI=confidence interval. **b**, Association of high (top quartile) tumor mutational burden (TMB) with overall survival after anti-CTLA-4 treatment in patients from Extended Data Fig. 2a; $P = 0.20$; log-rank test. **c**, Association of high mean HED and high TMB (red) with improved overall survival after anti-CTLA4 treatment in patients from Extended Data Fig. 2a; $P = 0.024$; log-rank test. **d**, Multivariable Cox proportional-hazards model including mean HED and other clinical variables. Data show independent effect of mean HED associated with improved survival after anti-CTLA-4. HED, TMB, and fraction of copy number alterations (FCNA) are dichotomized into high (1) and low (0) groups based on the top quartile for each variable.



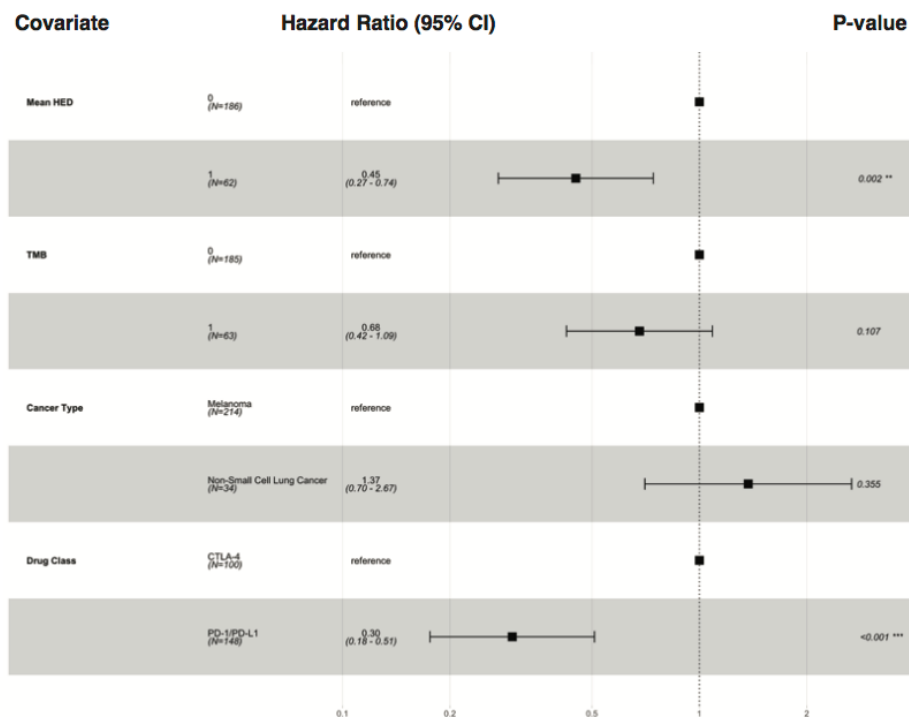
Extended Data Fig. 3 | Neither HLA-I heterozygosity nor HLA-I evolutionary divergence is associated with prognosis in TCGA melanoma patients. **a**, Full heterozygosity at HLA-I (red) is not associated with prognosis in TCGA melanoma patients; $P = 0.14$, log-rank test. **b**, High patient mean HED (red) is not associated with prognosis in patients from Extended Data Fig. 3a; $P = 0.80$, log-rank test. **c**, High mean HED (red) is not associated with prognosis in TCGA melanoma patients fully heterozygous at HLA-I; $P = 0.54$; log-rank test.



Extended Data Fig. 4 | Neither HLA-I heterozygosity nor HLA-I evolutionary divergence is associated with prognosis in TCGA lung cancer patients. a, Full heterozygosity at HLA-I (red) is not associated with prognosis in TCGA lung cancer patients; $P = 0.38$, log-rank test. **b**, High mean HED is not associated with prognosis in patients from Extended Data Fig. 4a; $P = 0.73$, log-rank test. **c**, High mean HED (red) is not associated with prognosis in TCGA lung cancer patients fully heterozygous at HLA-I; $P = 0.51$, log-rank test.

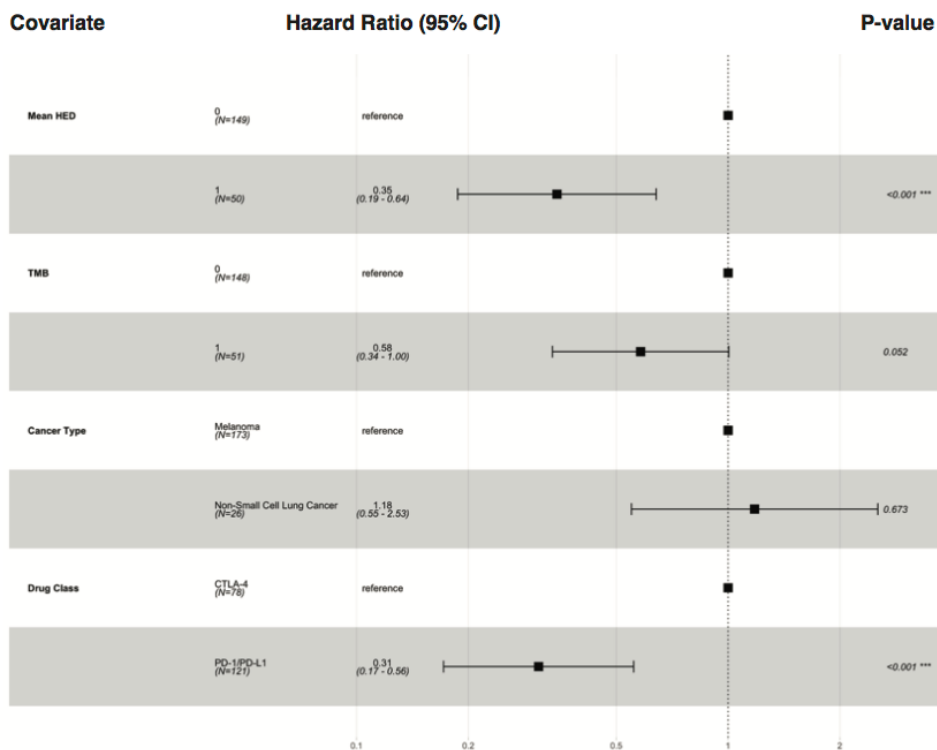
a

All Patients

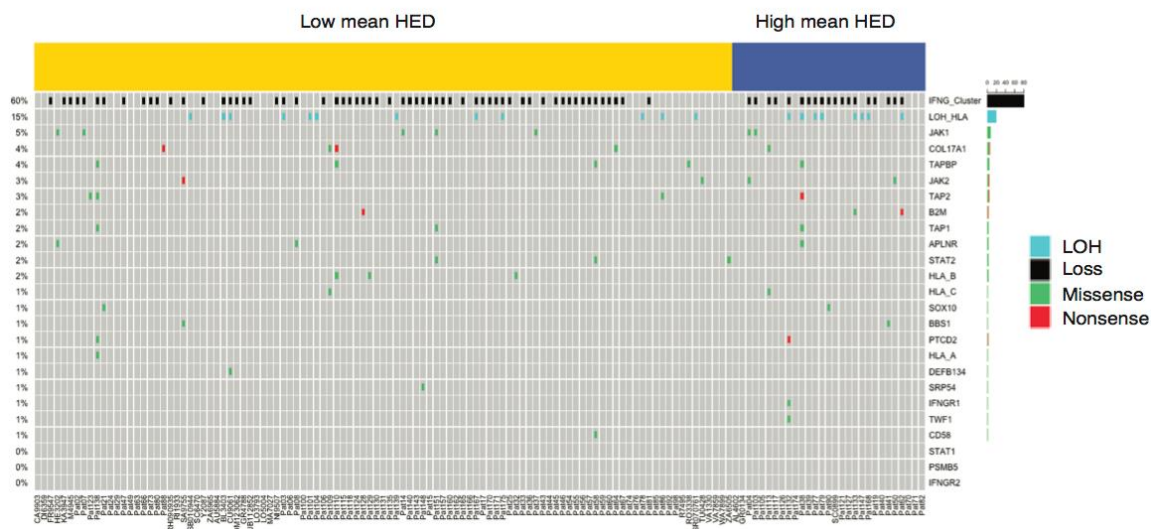


b

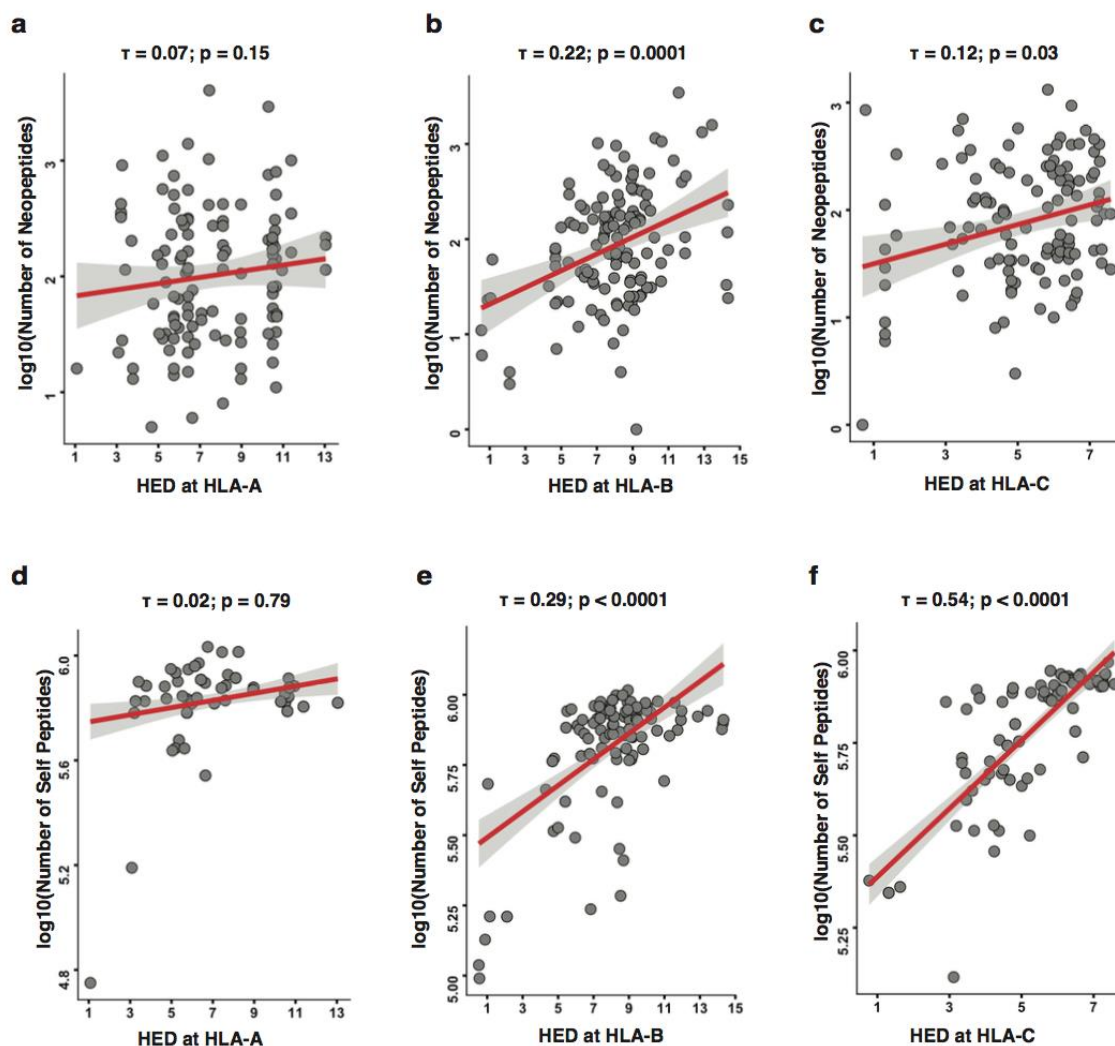
Fully Heterozygous Patients



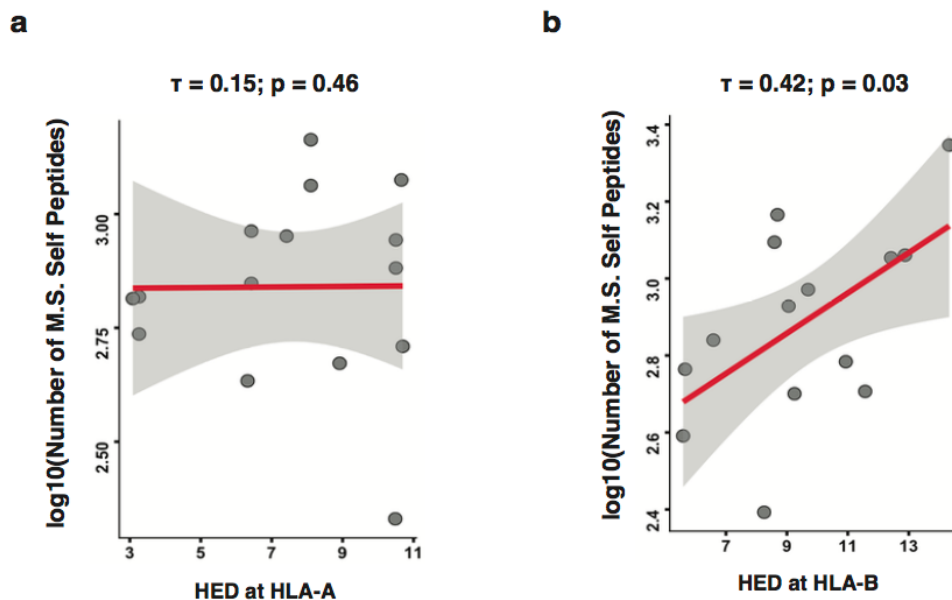
Extended Data Fig. 5 | The effects of mean HLA-I evolutionary divergence and tumor mutational burden are independent of cancer type and drug class. **a**, Multivariable Cox proportional-hazards model including mean HED and other clinical variables using patients from Fig. 3a (all patients, i.e. either HLA-I homozygous or heterozygous). Data show independent effect of mean HED in predicting response to ICI. **b**, Multivariable Cox proportional-hazards model including mean HED and other clinical variables using patients from Fig 3d (patients fully heterozygous at HLA-I). Data show independent effect of mean HED associated with improved survival after ICI therapy. HED and TMB are dichotomized into high (1) and low (0) groups using the top quartile for each variable.



Extended Data Fig. 6 | Oncoprint showing mutations in genes in our patient cohorts. Data show no difference in proportion of patients with mutations in the presented genes between patients with high mean HLA-I evolutionary divergence (HED) and low mean HED. LOH = loss of heterozygosity at HLA-I.



Extended Data Fig. 7 | Association of HLA-I evolutionary divergence at each class I locus with diversity of tumor and human immunopeptidomes. **a**, Correlation of HED at HLA-A with number of unique neopeptides bound to HLA-A alleles of each patient genotype using all patients heterozygous at HLA-A from Fig. 2 for whom neopeptide data were available; $P = 0.15$. Each point represents a patient HLA-A genotype **b**, Correlation of HED at HLA-B with number of unique neopeptides bound to HLA-B alleles of each patient genotype using patients heterozygous at HLA-B; $P = 0.001$ **c**, Correlation of HED at HLA-C with number of unique neopeptides bound to HLA-C alleles of each patient genotype using patients heterozygous at HLA-C; $P = 0.03$. **d**, Correlation of HED at HLA-A with number of unique self-peptides bound to HLA-A alleles of each patient genotype using patients heterozygous at HLA-A; $P = 0.79$; Kendall's rank correlation. **e**, Correlation of HED at HLA-B with number of unique self-peptides bound to HLA-B alleles of each patient genotype using patients heterozygous at HLA-B; $P < 0.0001$ **f**, Correlation of HED at HLA-C with number of unique self-peptides bound to HLA-C alleles of each patient genotype using patients heterozygous at HLA-C; $P = 0.79$. All p-values were calculated using Kendall's rank correlation. Red line indicates line of best linear fit.



Extended Data Fig. 8 | Association of HLA-I evolutionary divergence at HLA-A and HLA-B with diversity of the self immunopeptidome generated by mass spectrometry. a, Correlation of HED at HLA-A with number of unique naturally processed self-peptides bound to alleles of each HLA-A genotype from patients with metastatic melanoma patients heterozygous at each HLA-I from Pearson *et al*; $P = 0.46$; Data recapitulate results derived from computational peptide-HLA binding predictions shown in Extended Data Fig. 7d. **b**, Correlation of HED at HLA-B with number of unique naturally processed self-peptides bound to alleles of each HLA-B genotype; $P = 0.03$. Data recapitulate results derived from computational peptide-HLA binding predictions shown in Extended Data Fig. 7e. All p-values were calculated using Kendall's rank correlation. Red line indicates line of best fit.

Extended Data Table 1 | Survival analysis from Extended Data Fig. 2a using different metrics for aggregating divergence across HLA-A, B, and C. Table shows results from log-rank test for data from Extended Data Fig 2a when using various metrics (mean, sum, median, or geometric mean) to aggregate HLA-I evolutionary divergences across HLA-A, B, and C. Table shows that results from Extended Data Fig 2a do not depend on the metric chosen.

Metric	Log-rank p-value	Hazard Ratio	95% Confidence Interval
Mean	0.0072	0.47	0.26-0.82
Sum	0.0072	0.47	0.26-0.82
Median	0.067	0.6	0.35-1.04
Geometric Mean	0.015	0.5	0.29-0.89

References

- Balachandran VP, Luksza M, Zhao JN, Makarov V, Moral JA, Remark R, Herbst B, Askan G, Bhanot U, Senbabaoglu Y, et al. 2017. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* 551:512-516.
- Buhler S, Nunes JM, Sanchez-Mazas A. 2016. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics* 68:401-416.
- Callahan MK, Postow MA, Wolchok JD. 2016. Targeting T Cell Co-receptors for Cancer Therapy. *Immunity* 44:1069-1078.
- Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, et al. 2013. Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Communications* 4.
- Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ. 1999. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283:1748-1752.
- Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, Kuo F, Kendall SM, Requena D, Riaz N, et al. 2018. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* 359:582-587.
- Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* 31:213-219.
- Doherty PC, Zinkernagel RM. 1975a. A biological role for the major histocompatibility antigens. *Lancet* 1:1406-1409.
- Doherty PC, Zinkernagel RM. 1975b. Enhanced immunological surveillance in mice heterozygous at H-2 gene complex. *Nature* 256:50-52.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.
- Feng S, Fang Q, Barnett R, Li C, Han S, Kuhlwilm M, Zhou L, Pan H, Deng Y, Chen G, et al. 2019. The Genomic Footprints of the Fall and Recovery of the Crested Ibis. *Curr Biol* 29:340-349 e347.
- Gao J, Shi LZ, Zhao H, Chen J, Xiong L, He Q, Chen T, Roszik J, Bernatchez C, Woodman SE, et al. 2016. Loss of IFN-gamma Pathway Genes in Tumor Cells as a Mechanism of Resistance to Anti-CTLA-4 Therapy. *Cell* 167:397-404 e399.
- Gfeller D, Bassani-Sternberg M. 2018. Predicting Antigen Presentation-What Could We Learn From a Million Peptides? *Front Immunol* 9:1716.

- Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, Stephens PJ, Daniels GA, Kurzrock R. 2017. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* 16:2598-2608.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862-864.
- Greaves M, Maley CC. 2012. Clonal evolution in cancer. *Nature* 481:306-313.
- Grueber CE, Wallis GP, Jamieson IG. 2014. Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. *PLoS ONE* 9:e89632.
- Hughes AL, Nei M. 1988. Pattern of Nucleotide Substitution at Major Histocompatibility Complex Class-I Loci Reveals Overdominant Selection. *Nature* 335:167-170.
- Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annual Review of Genetics* 32:415-435.
- International Wheat Genome Sequencing C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1251788.
- Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of Immunology* 199:3360-3368.
- Kim S, Kim HS, Kim E, Lee MG, Shin E, Paik S, Kim S. 2018. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol*.
- Koboldt DC, Zhang QY, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22:568-576.
- La Gruta NL, Gras S, Daley SR, Thomas PG, Rossjohn J. 2018. Understanding the drivers of MHC restriction of T cell receptors. *Nature Reviews Immunology* 18:467-478.
- Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28:311-317.
- Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, et al. 2017. Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*.
- Lenz TL. 2011. Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution* 65:2380-2390.

- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
- Luksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, Rizvi NA, Merghoub T, Levine AJ, Chan TA, et al. 2017. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* 551:517-520.
- McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, et al. 2016. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351:1463-1469.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297-1303.
- McKenzie LM, Pecon-Slattery J, Carrington M, O'Brien SJ. 1999. Taxonomic hierarchy of HLA class I allele sequences. *Genes and Immunity* 1:120-129.
- Parham P. 1988. Function and polymorphism of human leukocyte antigen-A,B,C molecules. *Am J Med* 85:2-5.
- Parham P, Ohta T. 1996. Population biology of antigen presentation by MHC class I molecules. *Science* 272:67-74.
- Patel SJ, Sanjana NE, Kishton RJ, Eidizadeh A, Vodnala SK, Cam M, Gartner JJ, Jia L, Steinberg SM, Yamamoto TN, et al. 2017. Identification of essential genes for cancer immunotherapy. *Nature* 548:537-542.
- Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. 2013. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *Journal of Immunology* 191:5831-5839.
- Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, Rodenbrock A, Laverdure JP, Cote C, Mader S, et al. 2016. MHC class I-associated peptides derive from selective regions of the human genome. *Journal of Clinical Investigation* 126:4690-4701.
- Penn DJ, Damjanovich K, Potts WK. 2002. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci U S A* 99:11260-11264.
- Pierini F, Lenz TL. 2018. Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. *Mol Biol Evol*.
- Potts WK, Wakeland EK. 1990. Evolution of diversity at the major histocompatibility complex. *Trends in Ecology & Evolution* 5:181-187.

- Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol* 15:1022-1027.
- Rammensee HG, Friede T, Stevanović S. 1995. MHC ligands and peptide motifs: first listing. *Immunogenetics* 41:178-228.
- Rentoft M, Lindell K, Tran P, Chabes AL, Buckland RJ, Watt DL, Marjavaara L, Nilsson AK, Melin B, Trygg J, et al. 2016. Heterozygous colon cancer-associated mutations of SAMHD1 have functional significance. *Proc Natl Acad Sci U S A* 113:4723-4728.
- Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, Hodi FS, Martin-Algarra S, Mandal R, Sharfman WH, et al. 2017. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*.
- Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, et al. 2015. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348:124-128.
- Robins HS, Campregher PV, Srivastava SK, Wachter A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099-4107.
- Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE, Parham P. 2017. Distinguishing functional polymorphism from random variation in the sequences of > 10,000 HLA-A, -B and -C alleles. *PLoS Genetics* 13.
- Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research* 43:D423-D431.
- Sade-Feldman M, Jiao YXJ, Chen JH, Rooney MS, Barzily-Rokni M, Eliane JP, Bjorgaard SL, Hammond MR, Vitzthum H, Blackmon SM, et al. 2017. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nature Communications* 8.
- Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, Barron DA, Zehir A, Jordan EJ, Omuro A, et al. 2019. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet*.
- Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28:1811-1817.
- Schumacher TN, Schreiber RD. 2015. Neoantigens in cancer immunotherapy. *Science* 348:69-74.
- Sette A, Sidney J. 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50:201-212.

- Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK. 2016. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet* 48:995-1002.
- Shen RL, Seshan VE. 2016. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research* 44.
- Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, Stevens J, Lane WJ, Dellagatta JL, Steelman S, et al. 2015. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* 33:1152-1158.
- Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, et al. 2014. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* 371:2189-2199.
- Subramanian S, Kumar S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* 7:306.
- Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, et al. 2018. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* 50:1161-1170.
- Thursz MR, Thomas HC, Greenwood BM, Hill AV. 1997. Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nat Genet* 17:11-12.
- Tran E, Ahmadzadeh M, Lu YC, Gros A, Turcotte S, Robbins PF, Gartner JJ, Zheng Z, Li YF, Ray S, et al. 2015. Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 350:1387-1390.
- Tran E, Robbins PF, Lu YC, Prickett TD, Gartner JJ, Jia L, Pasetto A, Zheng Z, Ray S, Groh EM, et al. 2016. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N Engl J Med* 375:2255-2262.
- Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Foppen MH, Goldinger SM, et al. 2015. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 350:207-211.
- van Rooij N, van Buuren MM, Philips D, Velds A, Toebes M, Heemskerk B, van Dijk LJ, Behjati S, Hilkmann H, El Atmioui D, et al. 2013. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J Clin Oncol* 31:e439-442.
- Wain LV, Bailes E, Bibollet-Ruche F, Decker JM, Keele BF, Van Heuverswyn F, Li Y, Takehisa J, Ngole EM, Shaw GM, et al. 2007. Adaptation of HIV-1 to its human host. *Mol Biol Evol* 24:1853-1860.

- Wakeland EK, Boehme S, She JX, Lu CC, McIndoe RA, Cheng I, Ye Y, Potts WK. 1990. Ancestral polymorphisms of MHC class II genes: divergent allele advantage. *Immunol Res* 9:115-122.
- Wei L, Liu LT, Conroy JR, Hu Q, Conroy JM, Morrison CD, Johnson CS, Wang J, Liu S. 2015. MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics* 16:569.
- Yarchoan M, Hopkins A, Jaffee EM. 2017. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N Engl J Med* 377:2500-2501.
- Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, Hu-Lieskovan S, Torrejon DY, Abril-Rodriguez G, Sandoval S, Barthly L, et al. 2016. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N Engl J Med* 375:819-829.
- Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, et al. 2017. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* 23:703-713.
- Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* 46:D754-D761.
- Zhao F, Sucker A, Horn S, Heeke C, Bielefeld N, Schrors B, Bicker A, Lindemann M, Roesch A, Gaudernack G, et al. 2016. Melanoma Lesions Independently Acquire T-cell Resistance during Metastatic Latency. *Cancer Res* 76:4347-4358.

Acknowledgement

I would like to thank my supervisor Dr. Tobias Lenz for his encouragement, enthusiasm, and patient during my doctoral research, but above all, for allowing me to pursue under his supervision the eclectic and fascinating projects presented in this thesis. I am thankful for all the scientific discussions which provided precious insights into my research. Thanks for leaving me alone at the right moment, because I have learned how to face and overcome difficulties. Thank you for the honest supervision. I am proud of our work.

I would like to acknowledge all the collaborators, Almut Nebel, Austin W. Reynolds, Ben Krause-Kyora, Chirag Krishna, Christina M. Balentine, Deborah A. Bolnick, Diego Chowell, Jacques Fellay, Jaime Mata-Míguez, Jatin Arora, Jesper L. Boldsen, Lisa Boehme, Marcel Nutsua, Mary Carrington, Paul J. McLaren and Timothy A. Chan. None of this would have been possible without their hard work, availability, and valuable comments.

Thank you to all my colleagues from the Emmy Noether Group for Evolutionary Immunogenomics Alejandro, Ana, Artemis, Clinton, Jatin, Leo, Malavi, Onur and Reem for the stimulating discussions and for creating a positive and inspiring work environment.

I would like to thank my thesis committee members Dr. Almut Nebel, Dr. John Baines and Dr. Philip Rosenstiel for their interest and support.

I am thankful to the International Max Planck Research School for financial support which allowed me to present my research at various meetings and conferences.

I would like to acknowledge Britta and Petra, for their care and for their kind help with bureaucratic troubles.

Thank you to Marc, Michael, and Britta for translating in German the summary of this thesis, in a very short time.

I would like to thank my parents, Catia and Massimo for their love and support. For always letting me free to be myself, for their constant presence despite the distance. For having believed in me even when I no longer thought I could make it. You will always be my favorite and unique flatmates!

Thanks to my grandparents Jolanda and Mario, for their encouragement and enthusiasm. Thank you for being part of this.

Acknowledgement

I am deeply grateful to my sweet sister, Melinda. The timetable of your first semester at the uni is still written on my window. Each exam, each smile, each message I received from you, each moment we have spent together are written in my heart.

Thanks to Nicolas, for the first funny video you sent me four years ago on WhatsApp. Since then I felt good knowing that the one I love the most was in safe and funny hands. Thanks for sharing the flat Coke and for the long chats about life every time I get home.

Special thanks to Annelore, for all the work we have done together, for all the golden words I will preserve along the way. Thank you to my Tuesday's fellows, for all the moments we have shared, none of this would have been possible without them.

I am thankful to Francesco, for our international and hilarious “pandacalls”; in the middle of the storm, we knew there would always be a friend.

Thank you to Francesca, for having been present during the hard moments but, above all, for being always present in the happiest ones.

I would like to thank the great friends I have met in Plön. Ana Isabel, for being the brave woman she is, for sharing so much in a very short time. Jelena, my sister from another mother, for her constant support. Ana, for all the happiness she brings to the office and outside. Thanks to my perseverant, leftist, clever, dreamer friend Ezgi. Malavi for her funny stories that cheer up the days at work. Chaitanya, for our coffees at whatever time. Çağdaş and Gökçe for making me feel at home. Neva, for our Sunday walks into the forest. Jatin, for the Indian songs he used to sing at the door.

Thank you to all EvolBio crew!

Heartfelt thanks to Roberto, for his courage and persistence. Destiny, it could not have been more perfectly timed. Thanks for finding the teaspoons in the kitchen. *Tutto il mare che c'è tra i miei e tuoi occhi, lo custodisce la notte ora che ti aspetto lontana.*

Curriculum vitae

Personal data

Name: Federica Pierini
Date and place of birth: 20.06.1984 in Rome, Italy
Nationality: Italian
Place of residence: Gänsemarkt 6, 24306 Plön, Germany

Education

- 2015 - Present **PhD student**
Emmy Noether Group Evolutionary Immunogenomics
Max Planck Institute for Evolutionary Biology, Plön, Germany
- 2013 **MSc in Biology and Human Evolution**
University of Rome Tor Vergata, Rome, Italy
Final grade: 110/110 *cum laude*
- 2009 **BSc in Human Biology**
University of Rome Tor Vergata, Rome, Italy
Final grade: 108/110

Research experience

- 2014 Institute of Evolutionary Biology, Barcelona, Spain
Internship at the Paleogenomics Lab
Supervisor: Carles Lalueza-Fox
- 2011 University of York, York, United Kingdom
Research trainee at the BioArCh center
Master thesis: "Aminostratigraphy of Mediterranean sites based on intra-crystalline protein diagenesis in *Pecten* shells: a pilot study"
Supervisors: Olga Rickards, Beatrice Demarchi, Kirsty Penkman
- 2009 University of Rome Tor Vergata, Rome, Italy
Research trainee at the Molecular Anthropology Lab
Bachelor thesis: "Complete sequencing of human mitochondrial DNA of an indigenous population of Ecuador, the "Tsachila": analysis of haplogroup D"
Supervisor: Olga Rickards

Grants

- 2017 Conference Participation Grant Wellcome Genome Campus Conference Centre

2015	Travel grant from Boehringer Ingelheim Fonds
2013	Leonardo da Vinci grant
2011	Travel Grant IUCA Environmental Sciences Institute, University of Zaragoza
2011	Erasmus Placement grant

Publications

Chowell, D.*; Krishna, C.*; **Pierini, F.***; Makarov, V; Rizvi, NA; Kuo, F; Riaz, N; Lenz, TL.; Chan TA. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Manuscript in revision at Nature Medicine*. *: equal contribution.

Krause-Kyora, B.; Nutsua, M.*; Boehme, L.*; **Pierini, F.***; Pedersen, DD.*; Kornell, SC.; Drichel, D.; Bonazzi, M.; Möbus, L.; Tarp, P.; Susat, J.; Bosse, E.; Willburger, B.; Schmidt, HA.; Sauter, J.; Franke, A.; Wittig, M.; Caliebe, A.; Nothnagel, M.; Schreiber, S.; Boldsen, LJ.; Lenz, TL.; Nebel, A.: Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nature communications* (2018) *: equal contribution.

Pierini, F; Lenz TL.: Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. *Molecular biology and evolution* (2018)

Pierini, F.; Demarchi, B.; Turner, J.; Penkman, K.: Pecten as a new substrate for IcPD dating: The quaternary raised beaches in the Gulf of Corinth, Greece. *Quaternary Geochronology* (2016)

Gómez-Sánchez, D.; Olalde, I.; **Pierini, F.**; Matas-Lalueza, L.; Gigli, E.; Lari, M.; Civit, S.; Lozano, M.; Vergès, J. M.; Caramelli, D.; Ramírez, O.; Lalueza-Fox, C.: Mitochondrial DNA from El Mirador Cave (Atapuerca, Spain) reveals the heterogeneity of Chalcolithic populations. *PLoS One* (2014)

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 to specific paragraphs are listed in the thesis outline 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.

Place, Date

(Federica Pierini)