Functional characterization of antigen repertoires in HLA-associated complex diseases to investigate antagonistic selection on HLA genes

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

Plön, December 2018

First referee: Dr. Tobias Lenz

Second referee: Prof. Tal Dagan

Date of oral examination: 22.02.2019

Table of content

Summary ____ *5*
Zusammenfassung ____ *7*
Introduction ____ *10*
Thesis outline ____ *18*
Chapter 1 ____ *21*
Introduction ____ **21**
Results ____ **23**
Discussion ____ **31**
Material and methods ____ **33**
Chapter 2 ____ *40*
Chapter 3 ____ *70*
Introduction ____ **70**
Results ____ **72**
Discussion ____ **78**
Material and methods ____ **81**
Chapter 4 ____ *85*
Introduction ____ **85**
Results ____ **86**
Discussion ____ **92**
Material and methods ____ **94**
Conclusion ____ *113*
Perspectives ____ *117*
References ____ *118*
Annex ____ *126*
Acknowledgements ____ *143*
Curriculum vitae ____ *144*
Declaration ____ *145*

Summary

The interaction between hosts and pathogens represents a major driver of their evolution, and a core interest in the evolutionary biology. In vertebrate hosts, the classical genes of Major Histocompatibility Complex (MHC) make a central component of their adaptive immune system. The cell-surface molecules encoded by the MHC genes present peptide fragments (derived from both self-antigens and pathogens) to T-cells, which upon recognizing them as foreign, initiate a specific immune response. To explore the general MHC evolution, I used humans as a study system where MHC is designated as Human Leukocyte Antigen (HLA). Although a vast allelic diversity of HLA genes exists at the population level, potentially maintained by pathogen-mediated balancing selection, only a fraction of that is seen at the individual level in the form of a limited number of HLA genes and their alleles. This setting has been proposed as an optimum between HLA-conferred resistance to pathogens and the risk of autoimmunity, which represent major antagonistic selection forces on HLA genes. The fine-mapping of various HLA's association with different infectious and autoimmune diseases has suggested a major role of the peptide-repertoire of HLA alleles in determining their specific effect on diseases. However, the exact mechanisms that underlie HLA's association with the diseases, and by that modulate the antagonistic selection on HLA genes remain elusive.

In order to elucidate them, I started by investigating the functional basis of the previously known protective effect of HLA heterozygosity at the HLA class-I genes on HIV-1 progression. I used a dataset of 6,311 HIV-1 infected individuals and predicted the HLA-bound peptides derived from HIV-1 proteome for each HLA allele represented in the dataset. The individual-specific repertoire of HLA-bound peptides suggested that HLA heterozygote advantage against HIV-1 could be mediated by both a broader array of HLA-bound peptides and a higher likelihood of carrying specific protective alleles in heterozygotes compared to homozygotes. The comparison of the peptide-repertoire of risk and protective alleles suggested that individual alleles could confer disease control by binding either a large number of peptides or specific immunodominant peptides. The separate analysis of the individual HLA genes indicated that different mechanism for the heterozygote advantage might work at different genes, such as either T-cell or NK-cell-mediated immune attack on the virus, possibly resulting in different evolutionary constraints on different HLA genes. Overall, the findings suggested that the pathogen-mediated selection might favor both HLA heterozygosity and individual alleles.

Further, hypothesizing that not all HLA-bound peptides would be relevant for disease control, we developed a new approach, named Peptidome-wide association study (PepWAS), that can predict HLA-bound disease-associated epitopes from a given peptidome. The PepWAS-predicted HIV-1-associated epitopes accounted for as much variation (12%) in HIV-1 viral load as by the genetic variants in HLA class-I genes, providing a functional basis for the association between HLA and HIV-1 control.

I then focused on the association between HLA class-II genes and Type 1 Diabetes (T1D). Using a case-control dataset of 16,029 individuals, I first showed that heterozygosity at the HLA class-II genes conferred T1D risk. To investigate functional basis of this HLA heterozygote disadvantage, I predicted individual-specific repertoires of HLA-bound peptides from 17 T1D-relevant human proteins. The comparison of individual-specific peptide-repertoires between HLA heterozygous and homozygous individuals suggested that both a broader array of HLA-bound self-peptides and higher odds of carrying risk alleles might contribute to HLA heterozygote disadvantage. The characterization of the allele-specific peptide-repertoire suggested that an allele might confer T1D risk due to its low peptide-binding affinity possibly contributing to inefficient removal of autoreactive T-cells in thymus and (or) by binding specific disease-causing peptides, e.g. post-translationally deamidated peptides. The PepWAS-predicted T1D-associated epitopes accounted for even more deviance in T1D status than HLA class-II haplotypes (33.1 vs. 29.6%). Moreover, the sequence homology between predicted T1D-associated epitopes and pathogenic peptides suggested pathogens as the potential trigger of autoimmunity.

Overall, the insights from this thesis shed light on different mechanisms that possibly underlie the differential association of HLA genes with infectious and autoimmune diseases, which, in turn, potentially shape the antagonistic selection on the classical HLA genes.

Zusammenfassung

This summary was kindly translated from English to German by Marc Ritter and Sina Schirmer.

Die Wechselwirkungen zwischen Wirten und Pathogenen sind ein wesentlicher Treiber ihrer Evolution und von besonderem Interesse in der Evolutionsbiologie. In Wirbeltieren bilden die klassischen Gene des Haupthistokompatibilitätskomplex (MHC von engl. „major histocompatibility complex“) eine zentrale Komponente ihres adaptiven Immunsystems. Die Moleküle der Zelloberfläche, die von den MHC-Genen kodiert werden, präsentieren T-Zellen Peptidfragmente (sowohl die eigenen Antigene als auch von Krankheitserregern), die wenn sie als fremd erkannt werden, eine spezifische Immunantwort auslösen. Um die allgemeine MHC-Evolution zu untersuchen, habe ich den Menschen als Studiensystem verwendet, in dem MHC als Human Leukocyte Antigen (HLA) bezeichnet wird. Obwohl auf Populationsebene eine große Allel-Diversität von HLA-Genen vorhanden ist, die möglicherweise durch pathogenvermittelte balancierte Polymorphismen erhalten wird, ist nur ein Bruchteil davon auf der individuellen Ebene in Form einer begrenzten Anzahl von HLA-Genen und Allelen zu sehen. Dieser Zustand wurde als ein Optimum zwischen der durch HLA vermittelten Resistenz gegen Krankheitserreger und dem Risiko der Autoimmunität vorgeschlagen, welche die stärksten gegensätzlichen Selektionskräfte für HLA-Gene darstellen. Die genauere Untersuchung der Verbindung von HLA mit verschiedenen Infektions- und Autoimmunkrankheiten legt nahe, dass das Peptid-Repertoire von HLA-Allelen eine wichtige Rolle bei der Bestimmung ihrer spezifischen Wirkung auf Krankheiten spielt. Die genauen Mechanismen, die der Verbindung von HLA mit den Krankheiten zugrunde liegen und so die antagonistische Selektion auf HLA-Gene beeinflussen bleiben dabei schwer fassbar.

Daher untersuchte ich zunächst die funktionale Basis der schon bekannten schützenden Wirkung von HLA-Klasse-I Heterozygosität auf HIV-1-Progression. Ich verwendete dazu einen Datensatz von 6.311 HIV-1-infizierten Individuen und prognostizierte die HLA-gebundenen Peptide, welche ich vom HIV-1-Proteom abgeleitetet hatte, für jedes im Datensatz dargestellte HLA-Allel voraus. Das individuelle spezifische Repertoire an HLA-gebundenen Peptiden legt nahe, dass der Vorteil von HLA-Heterozygoten gegenüber HIV-1 sowohl durch ein breiteres Spektrum an HLA-gebundenen Peptiden als auch durch eine

höhere Wahrscheinlichkeit in Heterozygoten im Vergleich zu Homozygoten, Träger eines spezifische Schutzallele zu sein. Der Vergleich des Peptid-Repertoires von Risiko und schützenden Allelen legt nahe, dass einzelne Allele die Krankheit durch das Binden entweder einer großen Anzahl von Peptiden oder aber von spezifischen immundominanten Peptiden kontrollieren. Die getrennte Analyse der einzelnen HLA-Gene zeigte, dass unterschiedliche Mechanismen beim Vorteil von heterozygoten bei verschiedenen Genen wirken können, beispielsweise bei T-Zellen- oder NK-Zell-vermittelten Immunangriffen auf das Virus, was möglicherweise zu unterschiedlichen evolutionären Einschränkungen bei verschiedenen HLA-Genen führt. Insgesamt deuteten die Ergebnisse darauf hin, dass die durch Pathogene vermittelte Selektion sowohl HLA-Heterozygosität als auch einzelne Allele begünstigen könnte.

Mit der Hypothese, dass nicht alle HLA-gebundenen Peptide für die Krankheitsbekämpfung relevant sind, entwickelten wir einen neuen Ansatz, Peptidome-wide Association Study (PepWAS), welcher HLA-gebundene krankheitsassoziierte Epitope aus einem gegebenen Peptidom vorhersagen kann. Die von PepWAS-vorhergesagten HIV-1-assoziierten Epitope machten soviel der Variation (12 %) in der HIV-1 Viruslast aus, wie die genetischen Varianten in HLA-Klasse-I-Genen, und bildeten eine funktionelle Grundlage für die Verbindung zwischen HLA und HIV-1 Kontrolle.

Im Anschluss konzentrierte ich mich auf die Assoziation zwischen HLA-Klasse-II-Genen und Typ-1-Diabetes (T1D). An einem Fall-Kontroll-Datensatz von 16.029 Individuen zeigte ich zuerst, dass Heterozygosität bei den HLA-Klasse-II-Genen, T1D-Risiko mit sich brachte. Um die funktionelle Basis dieses HLA-Heterozygoten-Nachteils zu untersuchen, habe ich aus 17 T1D-relevanten humanen Proteinen individuelle spezifische Repertoires an HLA-gebundenen Peptiden vorhergesagt. Der Vergleich von individualspezifischen Peptid-Repertoires zwischen HLA-heterozygoten und homozygoten Individuen legt nahe, dass sowohl ein breiteres Spektrum an HLA-gebundenen Selbstpeptiden als auch höhere Chancen für das Tragen von Risiko-Allelen zum Nachteil für HLA-Heterozygote beitragen können. Die Charakterisierung des Allel-spezifischen Peptid-Repertoires legt nahe, dass ein Allel aufgrund seiner geringen Peptidbindungsaffinität ein T1D-Risiko verursachen kann, das möglicherweise zu einer ineffizienten Entfernung autoreaktiver T-Zellen im Thymus und (oder) durch Bindung spezifischer krankheitsverursachender Peptide, z. posttranslational desamidierte Peptide. Die mit PepWAS vorhergesagten T1D-assoziierten Epitope erklärten im T1D-Status sogar noch mehr Abweichungen als die HLA-Klasse-II-Haplotypen (33,1 vs. 29,6%). Darüber hinaus deutete die

Sequenzhomologie zwischen vorhergesagten T1D-assoziierten Epitopen und pathogenen Peptiden auf Krankheitserreger als möglichen Auslöser der Autoimmunität hin.
Insgesamt werfen die Erkenntnisse aus dieser Arbeit Licht auf verschiedene Mechanismen, die möglicherweise der unterschiedlichen Assoziation von HLA-Genen mit Infektions- und Autoimmunkrankheiten zugrunde liegen, welche wiederum die antagonistische Selektion von klassischen HLA-Genen prägen können.

Introduction

The evolution of species is strongly influenced by their environment. Changes in the environment exert pressure on species to adapt (*1*). Those individuals of a species whose phenotype allows them to cope up with the changes are more likely to survive, and their genetic variants are thus selected and passed on to the next generations. Pathogens represent an important part of the environment (*1*). Being both dynamic and extremely diverse, they are regarded as amongst the strongest selective pressure to drive the evolution of species (*2*, *3*). Since they can inflict harm to their host, defense mechanisms like the immune system have evolved in order to contain them. In vertebrates, the immune system is generally divided into innate and adaptive and comprises of a wide range of cells specialized in various aspects of the defense such as pathogen recognition, immune signaling, killing the pathogens etc. (*4*). The innate immune system makes generic responses through a limited repertoire of germline-encoded receptors on certain immune-cells (*4*, *5*). These receptors recognize fixed sequence patterns broadly shared by pathogens (*5*). The innate immune response does not need a long start-up phase and can respond very quickly. It makes sure, for example, that bacteria that have entered the skin are detected and destroyed (completely or partly) within a few hours. Over the life time, an individual could get exposed to a wide variety of pathogens, which could differ in several properties e.g. the mode of transmission, the mode of infection e.g. intracellular (e.g. HIV) or extracellular (e.g. *M.* tuberculosis), virulence etc. Despite these differences, an important feature shared by almost all of the pathogens is their tendency to evolve and adapt to the immune defense of their host (*6*–*8*). This has stimulated a lot of research on host-parasite co-evolution that has significant implication in health and diseases (*3*, *9*, *10*). It has been observed that pathogens can adapt to the host's immune defense in multiple ways. For example, they can acquire mutations at key locations in their proteins which are targeted by the immune system and thus evade it (*11*). They can acquire changes in their antigen sequence to mimic the self-antigens of the host, so as to deceive the immune system and not be detected by it (*12*). Due to low specificity of innate immune responses, the innate immune system is capable of controlling co-evolving pathogens only to a limited extent. Therefore, the effective control of pathogens would require a system that learn and make the specific immune responses against pathogens (*13*–*17*). This kind of defense is represented by the adaptive immune system that has evolved in vertebrate species (*18*). As in innate immune system, adaptive immune system is also comprised of

certain types of immune-cells. However, a vast diversity of these cell types in the host allows making the specific immune response to individual pathogens. If the body's first line of defense – the innate immune system – is unsuccessful in destroying the pathogens, then adaptive immune system kicks in. In comparison to the innate immune system, the adaptive defense could take longer to start, but it targets the pathogen in a specific and accurate manner. Besides the specificity, other major advantage of adaptive immune system is that it can remember the encountered pathogen and respond very quickly upon re-infection with the same.

Major Histocompatibility complex and its role in the immune system

A fundamental component of the adaptive immune system is represented by the genes of the **Major Histocompatibility Complex (MHC)** locus, which was first discovered in 1936 in mouse (*19*). The genes of the MHC locus are categorized into classical and non-classical ones (*20*). Classical MHC genes encode for the molecules that took up the peptide fragments inside the cell, bring them to cell surface and present them to surveilling T-cells. Upon recognizing the MHC-presented peptides as foreign or non-self-peptides, through T-cell receptors, the T-cells can become activated and initiate a specific immune response. All jawed vertebrate species have MHC molecules and use T-cell receptors (TCRs) and B-cell receptors (BCRs) as antigen receptors, whereas some jawless vertebrates, e.g. lampreys and hagfish, lack MHC molecules and use variable lymphocyte receptors as antigen receptors (*18*). In my thesis, I focused on human MHC, designated as **Human Leukocyte Antigen (HLA),** for studying the general evolution of MHC. It lies on the short arm of chromosome 6 (*21*). The classical HLA genes are broadly divided into two classes which different in certain aspects. HLA class-I genes (HLA-A, -B and -C) are expressed on almost all nucleated cells and present short (~9mer) peptides of intracellular origin to CD8+ T-cells (Cytotoxic T-cells; CTLs), while HLA class-II genes (HLA-DP, -DQ and -DR) are expressed mostly on professional antigen presenting cells, e.g. macrophages, dendritic cells, and typically present relatively longer (~15mer) peptides of extracellular origin to CD4+ T-cells (Helper T-cells) **(Fig. I1)** (*22*).

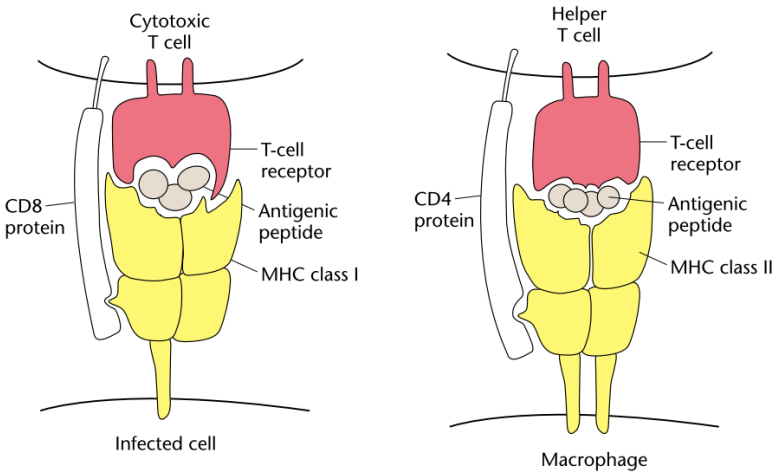


Fig. I1. Interaction between peptide-MHC class-I complex and Cytotoxic T-cell (Left), and peptide-MHC class-II complex and Helper T-cell (Right). The image is taken from Penn, 2002 (*22*).

Notably, HLA molecules bind both self and pathogen-derived peptides, and the distinction between them is critical to the success of adaptive immune response. This distinction between self and pathogen-derived peptides is achieved by optimizing the repertoire of T-cells in an individual. Naïve T-cells that bind self-peptide-HLA complexes with strong affinity are removed during their maturation process in the thymus (*23*) so as to deplete those T-cells that could possibly get activated by HLA-presented self-peptides and cause autoimmunity (so called **autoreactive T-cells**). However, this process is not flawless and some autoreactive T-cells can still escape negative selection (*23*). Studies have shown the presence of autoreactive T-cells also in healthy individuals, though they remain either inactivated or kept in check by regulatory immune cells (*24*).

HLA's diversity and balancing selection

A remarkable feature of HLA genes is their unparalleled allelic diversity in all human populations **(Table I1)** (*25*). It is considered to arise from mutation, recombination and gene conversion (*26*, *27*). The presence all HLA alleles at nominal frequency suggests that they are selective maintained (*28*). Pathogen-mediated balancing selection is widely regarded as the major selective agent that leads to HLA's allelic diversity (*13*, *29*–*32*). Several observations like the correlation between HLA's allelic diversity and pathogen richness across geographical regions (*33*, *34*), and the enrichment for membrane glycoproteins in haplotypes shared between chimps and human (*35*) lend indirect

empirical support to that. Two major mechanistic models for balancing selection at HLA have been proposed, (1) *Heterozygote advantage* and (2) *Negative frequency dependent selection* (NFDS) (*36*). The HLA heterozygote advantage model proposes that an individual with two different alleles (heterozygous) of an HLA gene would bind broader array of the peptides than an individual with the same alleles (homozygous). This would increase the breadth of the immune response in an individual, conferring stronger resistance against single or multiple pathogens. Besides allelic diversity, another remarkable feature of HLA genes is their very high inter-allelic divergence defined by the variation at the sequence level (*37*, *38*). Notably, the variation is concentrated inside the peptide-binding groove of HLA molecules and characterized by the high ratio of non-synonymous to synonymous polymorphism, suggesting that it has been selectively maintained (*39*). This observation has led to another hypothesis named, *Divergent allele advantage* (*40*, *41*), which assumes that heterozygote individuals with more divergent HLA alleles would encode for HLA molecules with a larger difference in their peptide-repertoires. This would result in a broader array of HLA-presented peptides, conferring higher immune-surveillance against pathogens and thus stronger resistance. The negative frequency dependent selection model, also usually referred as *rare-allele advantage*, proposes that the selective advantage of an allele increases as it becomes rare, because the pathogens are more likely to adapt to common HLA alleles in the population (*42*, *43*).

These mechanisms of balancing selection do not allow selection for the loss or fixation of an allele, thus enable the retention of a large number of HLA alleles for a very long time in a population. Moreover, it is widely accepted that these models are not mutually exclusive and likely act in parallel to shape the overall HLA's allelic diversity in a population. Although the majority of the known HLA alleles are rare (> 80%) (*44*), it might still allow diverse immune response in populations and by that ensure that no single pathogen can wipe out the entire populations. As HLA alleles are co-dominantly expressed and each encoded HLA molecule binds a specific peptides largely defined by the amino acid composition of its peptide-binding groove (*45*, *46*), a large number of HLA loci and HLA alleles can be speculated to be beneficial in individuals. However, an evolutionary paradox is that HLA diversity is expressed as several alleles at the population level rather than at individual level. Each individual carries a limited number of HLA gene and their alleles, which is a small fraction of HLA's total allelic diversity in the population. Several theoretical and empirical studies have suggested that this configuration has been selected

for an optimum between the resistance to pathogens and the risk of autoimmunity as HLA molecules present both self and pathogenic peptides to T-cells (*13*, *47–49*).

Table I1. Number of HLA alleles in IMGT/HLA database as of November 2018.

Number of HLA alleles	
HLA class I alleles	14,800
HLA class II alleles	5,288
Total HLA alleles	20,088

Individual HLA alleles are designated at 4-digit level including in this thesis, e.g. HLA-B*57:01 or HLA-DRB1*02:01, which represents a specific protein.

Association between HLA and diseases

It is possible that the HLA genotype of an individual might not fit to the given environment, e.g. HLA alleles of the individual bind only a few peptides of the pathogen prevalent in the individual's environment or those peptides that do not induce immune response, that could lead to disease traits. GWASs have associated HLA genes with a wide variety of diseases, be it infectious, e.g. HIV (*50*), Leprosy (*51*), Hepatitis (*52*), Tuberculosis (*53*) etc., or autoimmune, e.g. Type 1 diabetes (*54*), Rheumatoid arthritis (*55*), Alopecia areata (*56*) etc. In fact, HLA genes have been associated with the outcome of more number of diseases than any other locus of the human genome **(Fig. I2)** (*57–62*). These associations have brought an interesting observation that a single HLA allele could have an effect on the outcome of multiple diseases. For example, the HLA-DRB1 locus is involved in immune response against *M.* tuberculosis infection and has also been linked to the risk of Rheumatoid arthritis and Inflammatory bowel diseases (*63*). Indeed, it has been observed that HLA, which has been the prime target of pathogen-mediated selection, is enriched for the association with autoimmune diseases (*64*). Often, the alleles of the HLA gene associated with a disease show differential effect, ranging from risk to protective (*53*, *58*, *65*, *66*). For example, whereas DRB1*0301 confers susceptibility to Type 1 Diabetes (T1D), DRB1*13:01 provides protection from T1D in Caucasian population. Studies have shown that individual HLA alleles could also exhibit differential effect on different diseases (*67*). For example, HLA-DRB1*15:01 allele has been associated with the protection from Type 1 Diabetes and Rheumatoid arthritis but confers susceptibility to Leprosy (*68*) and Systemic lupus (*62*). It could be speculated that differential effect of HLA alleles could be an artefact of different cohorts across association studies. However, the

replication of the association of individual HLA alleles with several common infectious and autoimmune diseases between different Caucasian cohorts by Karnes *et al.* 2017 (*62*) suggested that the differential effect of HLA alleles is not an artefact of cohort-effect.

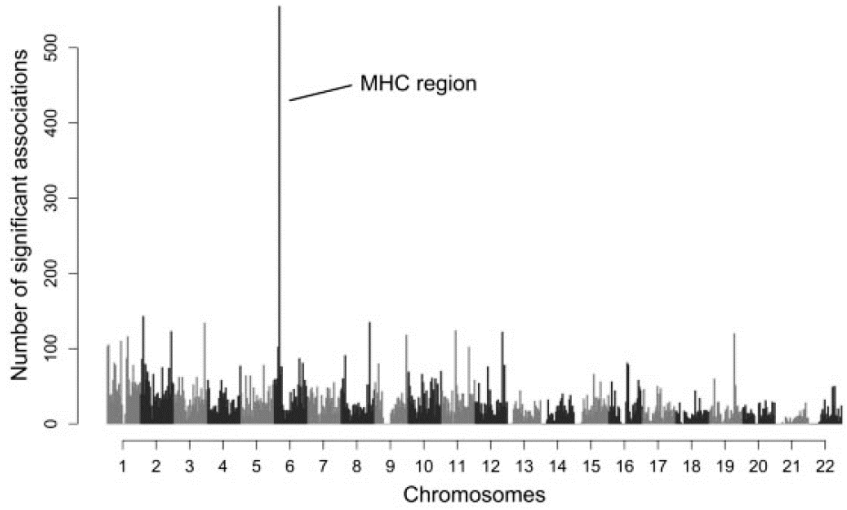


Fig. I2. Number of significant GWAS association with common human diseases across all human autosomes. The figure is taken from Lenz *et al.* 2016 (*61*).

Thesis topic

As mentioned above, HLA-presentation of a broader array of peptides in an individual might confer resistance against single or multiple pathogens. However, at the same time, it might make it more likely to activate autoreactive T-cells by HLA-presentation of self-peptides, that can result into autoimmune diseases (*69*, *70*). This conceptual trade-off has received some indirect support from studies in non-model organisms showing the maximum fitness for those individuals that carried an optimal number of MHC genes (*13*, *49*, *71–73*). This trade-off has facilitated the consideration of pathogens and autoimmunity as two antagonistic selection forces operating on HLA genes (*47*, *48*, *74*), where the former would favor broader peptide presentation while the later would limit that. This also suggests that the mechanisms of balancing selection that are thought to operate on the breadth of HLA-presented peptides, such as heterozygote advantage, should behave oppositely in infections and autoimmune conditions. This is supported by the observed heterozygote advantage against pathogens like HIV (*75*), HBV (*76*), HCV (*77*), and possible effect of heterozygosity on the risk of autoimmune diseases like T1D and RA (*58*). While it has been hypothesized and widely considered that the functional basis of HLA heterozygote (dis)advantage lies in the breadth of HLA-bound peptides, the exact mechanisms remain unclear. Consequently, this renders it unclear whether

pathogen- and autoimmunity-mediated antagonistic selection forces operates on the breadth of the HLA-bound peptides or any other variables related to HLA are also involved.

Moreover, since individual HLA alleles have also been associated with a wide variety of infectious and autoimmune diseases with differential effect (*62*, *78*), an individual could be resistant or susceptible to a disease due to carrying specific HLA alleles, irrespective of HLA zygosity. The specific association of individual alleles with a disease might modulate the overall selection on them. For example, if an allele confers very large risk of a high mortality infection, but protects from another low mortality autoimmune disease, there would be strong negative selection on that allele. Therefore, in order to elucidate the antagonistic selection on HLA alleles, it would be important to understand the basis of their differential association across diseases. However, the functional basis of alleles' differential association across diseases is not known. The fine mapping of HLA's association with some diseases have suggested peptide-properties of HLA alleles as the basis of their differential association (*65*, *66*, *79*), but what exactly those properties are remains unknown. There could be several possibilities. For example, the breadth of the allele-specific peptide repertoire; an allele might confer resistance against a pathogen because it binds a large number of pathogen-derived peptides, which might generate a broad immune response. Or it could be peptide-binding affinity; an allele can bind peptides with high affinity which enables T-cells to interact with peptide-HLA complex efficiently and make strong immune response. It is also possible that an allele selectively binds those peptides which are highly immunogenic and where the pathogen cannot evolve due to fitness constraints. In addition, it is likely that not one variable modulates an allele-mediated immune response, but multiple variables together, e.g. the number of bound peptides and the ability to bind specific peptides, because if the former was the case, an allele would have similar effect across all associated diseases. Moreover, while an HLA allele can bind several peptides derived from a given pathogen, it is possible that not all of them might induce an immune response. This suggests that it would be beneficial for the individual if HLA alleles bind those peptides which are relevant to the disease control. Likewise, it would be beneficial for the pathogen to evolve in such peptide regions in order to evade the immune response.

The HLA alleles which confer resistance and susceptibility to a disease might be under positive and negative selection, respectively. Therefore, the characterization of the

peptide-repertoires of HLA alleles associated with infectious and autoimmune diseases would be important to get a deep insight into the molecular basis of the antagonistic selection that possibly operates on HLA genes and shapes the HLA allelic diversity in populations. However, the peptide-repertoires of most HLA alleles remain so-far poorly defined. There are several reasons for that. First, until not long ago, most of the linkage and association studies had focussed on the common alleles only, which provided a handful of associated HLA alleles and thus did not allow a broad comparison of their peptide-repertoires. Secondly, while a very large number of HLA alleles exist in all human populations (*25*), each allele can potentially bind several thousand of peptides (*37*, *80*). A fully-factorial experimental assay to screen such a vast repertoire of peptides for the binding by individual HLA alleles is unmanageable to-date. However, several recent advances in computational and large scale genomics, e.g. the availability of computational methods for predicting binding-affinity between individual HLA alleles and any set of peptides, large multi-ethnic panels that allow the inference of HLA genotype (combination of alleles) of an individual based on combination of SNP markers, large genomic case-control datasets consisting of a large number of HLA alleles for several diseases etc., are providing promising opportunities to address the above discussed issues. In this thesis work, I took advantage of these advances and characterized the peptide-repertoire of several common and rare HLA alleles in the context of an infectious disease (HIV/AIDS) and an autoimmune disease (Type 1 Diabetes). In addition, we have developed a new approach named **Peptidome-wide Association Study (PepWAS)** that allows screening the entirety of a peptidome and predict those HLA-bound peptides which are associated with a given disease. I used PepWAS to predict the disease-associated peptide-repertoire of HLA alleles associated with HIV and Type 1 Diabetes that provided insights into mechanism through which the antagonistic selection could operate on classical HLA genes.

Thesis outline

In this thesis, I focused on elucidating the mechanism through which pathogen and autoimmunity-mediated antagonistic selection forces might operate on the classical HLA genes. For that, I characterized the peptide-repertoire of several common and rare HLA alleles differentially associated with HIV-1 infection and Type 1 Diabetes (T1D; an autoimmune disease). I investigated the effect of the variation in HLA-bound peptide-repertoire between HLA heterozygous and homozygous individuals on the outcome of HIV-1 and T1D. I also analyzed the variation in allele-specific peptide-repertoire to unravel the basis of the specific association of HLA alleles with HIV-1 and T1D. In addition, using our new approach, Peptidome-wide Association Study, I predicted the disease-associated peptides bound by individual HLA alleles. The whole work is split into four chapters outlined below.

Chapter 1: Quantitative and qualitative differences in allele-specific antigen repertoires underlie HLA heterozygote advantage against HIV-1.

HLA heterozygote advantage against HIV-1 has been shown by Carrington *et al.* 1999 and McLaren *et al*. 2015. While it is thought to be conferred by the ability of HLA heterozygotes to present a broader array of peptides to immune cells than HLA homozygotes, the actual mechanism remained elusive. I characterized HLA heterozygote advantage against HIV-1 by using HLA genotype and HIV-1 viral load of 6,311 individuals. First, I showed that heterozygosity at HLA-B and HLA-C but not at HLA-A was associated with lower viral load. Then, I predicted individual-specific repertoire of HIV-1 peptides bound by individual's HLA-A, HLA-B and HLA-C alleles. While at HLA-B, heterozygote advantage was potentially conferred by quantitative CTL-mediated immune response, another mechanism seemed to operate at HLA-C. Moreover, HLA-B heterozygosity resulted into elevated sequence evolution of the virus, suggesting HLA heterozygosity might incur replicative fitness cost to virus.

Chapter 2: HIV Peptidome-Wide Association Study Reveals Patient-Specific Epitope Repertoires Associated with HIV Control.

Specific amino-acid residues inside the peptide-binding groove of HLA-B and HLA-A have been associated with HIV-1 progression, and together account for12.3% of variation in

the viral load, suggesting a major role of HLA-presented HIV peptides in disease control. In order to identify HIV-1-associated peptides (epitopes), we created Peptidome-wide Association Study (PepWAS) approach that can interrogate the pool of all possible peptides derived from a given proteome and predict HLA-bound disease-associated epitopes. I applied PepWAS on the dataset of HIV-1 infected individuals from Chapter 1, and predicted a core set of HLA-B and HLA-A bound HIV-1-associated epitopes. Individual-specific repertoire of these epitopes accounted for nearly entire variation in viral load previously attributed to amino-acid residues, providing a functional basis of the association between HLA and HIV-1. Notably, HLA alleles that bound a broader array of predicted epitopes had stronger protective effect on disease progression.

Chapter 3: Does broader antigen presentation underlie HLA-mediated risk for Type 1 Diabetes?

Lenz e*t al.* 2015 have suggested a possible effect of heterozygosity at HLA-DRB1 and HLA-DQ genes on the risk of Type 1 Diabetes (T1D). I used the dataset from Hu *et al.* 2015 that contained HLA genotype of 6,651 T1D cases and 9,378 controls. First, I showed that HLA heterozygosity at DRB1 and DQ conferred T1D risk. Then, I predicted DRB1 and DQ allele-specific repertoire of bound peptides derived from 17 T1D-relevant β-cell proteins and tested multiple mechanisms hypothesized to underlie heterozygote disadvantage. The results suggested that both a broader array of HLA-bound self-peptides and higher likelihood of carrying specific risk alleles in heterozygotes compared to homozygotes could contribute to HLA heterozygote disadvantage. I also investigated the functional basis of specific association of HLA alleles with T1D. The comparison of the peptide-promiscuity and the peptide-binding affinity of T1D risk and protective alleles of DRB1 and DQ genes suggested that the functional basis of the effect of risk alleles is not the size of their peptide-repertoire but low peptide-binding affinity, that might contribute to inefficient removal of autoreactive T-cell in thymus. In additional, DQ risk alleles bound a larger number of deamidated peptides than protective alleles, lending support to the other possibility that HLA risk alleles confer risk by binding specific disease-causing peptides.

Chapter 4: Individual-specific analysis of HLA-presented peptide repertoires reveals novel epitopes associated with Type 1 Diabetes.

Hu *et al.* 2015 fine mapped that association between HLA and T1D outcome to specific amino-acid residues inside the peptide-binding groove of HLA-DRB1 and HLA-DQ genes, which together accounted for 26.9% of deviance in disease status, strongly suggesting the role of HLA-presented self-epitopes in disease pathology. However, the identity of the majority of T1D-associated epitopes remained missing. I used T1D dataset from Chapter 3 and applied PepWAS on all possible peptides (N = 8,018) from 17 T1D-relevant β-cell proteins. It predicted a core set of DRB1, cis and trans encoded DQ-bound peptides that were associated with the T1D outcome (T1D-associated epitopes). I also simulated deamidation, a stress-induced post-translational modification, of candidate proteins and predicted T1D-associated deamidated epitopes. The difference in the individual-specific repertoires of predicted epitopes between cases and controls accounted for 33.1% deviance in diseases, significantly exceeding the one attributed to amino-acid residues in DRB1 and DQ genes. Moreover, sequence similarity between predicted T1D-associated epitopes and *P.* tuberculosis-derived peptides suggested pathogen as a potential trigger of autoimmunity.

Table: Individual contribution to the chapters.

Chapters	I	II	III	IV
Study design	TLL, JA	TLL, JA	TLL, JA	TLL, JA
Provided data	JF, PJM, NC	JF, PJM, NC	SOG, SR, SSR, WMC	SOG, SR, SSR, WMC
Data analysis	JA, FP	JA	JA	JA
Discussion	JA, TLL, JF, PJM, MC	JA, TLL, JF, PJM, MC	JA, TLL	JA, TLL AK, SR, SSR
Manuscript writing	JA, TLL, FP	JA, TLL	JA	JA

The authors' name in alphabetical order.
AK: Åke Lernmark, FP: Federica Perini, JF: Jacques Fellay, JA: Jatin Arora, MC: Mary Carrington, NC: Nimisha Chaturvedi, PJM: Paul J. McLaren, SR: Soumya Raychaudhuri, SSR: Stephen S. Rich, SOG: Suna Onengut-Gumuscu, TLL: Tobias L. Lenz, WMC: Wei-Min Chen

Chapter 1

Quantitative and qualitative differences in allele-specific antigen repertoires underlie HLA heterozygote advantage against HIV-1

Jatin Arora [a], Federica Pierini [a], Paul J. McLaren [b, c], Mary Carrington [d, e], Jacques Fellay [f, g] & Tobias L. Lenz [a, *]

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

[b] JC Wilt Infectious Diseases Research Center, National HIV and Retrovirology Laboratory, Public Health Agency of Canada, R3E 0W3, Winnipeg, Canada

[c] Department of Medical Microbiology and Infectious Diseases, University of Manitoba, R3E 0J9, Winnipeg Canada

[d] Cancer and Inflammation Program, Leidos Biomedical Research, Frederick National Laboratory, Frederick, MD 21702, USA.

[e] Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02139-3583, USA.

[f] Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[g] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Introduction

Genes of the Major Histocompatibility Complex (MHC) play a central role in immune-recognition of pathogens. They encode for cell-surface molecules that present intra- and extra-cellular peptides to T-cells, which, upon recognizing them as non-self, initiate appropriate immune responses (*81*). Each MHC molecule binds a specific peptide repertoire which is largely defined by the amino-acid composition of its peptide-binding groove (*82*, *83*). Due to a key role in adaptive immunity and an 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 (*57*, *84*, *85*). The dynamic action of pathogen-mediated balancing selection is widely regarded as a key driver of MHC diversity (*29–31*, *36*, *42*). 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 (*85–88*). 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 (*89*, *90*).

First proposed by Doherty & Zinkernagel (*91*), 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 (*36*, *92*). One possible explanation why MHC heterozygotes would have a higher probability to trigger a specific immune response, compared to homozygous MHC genotypes, could be their presumed ability to present a broader array of antigenic peptides, thus increasing the probability to recognize at least one of them as foreign and consequently induce a targeted response. 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 (*40*, *41*). It assumes that heterozygote 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 elevated immune-surveillance against pathogens (*37*, *38*). An alternative explanation for heterozygote advantage that is not relying on quantitative differences in antigen presentation among MHC alleles has also been put forward. It is based purely on qualitative differences among MHC alleles, suggesting that heterozygosity simply increases the probability of carrying specific protective MHC alleles. Such qualitative differences among individual MHC alleles have indeed been observed in a number of species, including humans(*32*, *93*, *94*). However, it is unclear whether these qualitative differences among MHC alleles result from unique binding properties and presentation of critical peptides or whether they are ultimately also due to quantitative differences in the size of the allele-specific antigen repertoires (*95*).

A significant number of studies across a range of species have provided empirical support for a phenotypic advantage conferred by general MHC heterozygosity (*92*, *96*–*99*) as well as higher sequence divergence between MHC alleles (*100*–*104*). In humans, with thousands of alternative alleles and by far the most comprehensively studied species with regard to MHC polymorphism (*29*, *44*), empirical evidence for pathogen-mediated selection is surprisingly sparse. Owing to growing individual cohorts and denser

genotyping approaches, the number of infectious diseases for which statistical associations in the MHC have been identified is increasing (*69*, *94*, *105*). However, most association studies assume a simple additive genetic contribution and rarely consider evolutionary implications of their findings, due to their focus on the underlying disease mechanics. One of the few exceptions is the seminal study on HIV control by Carrington et al. (*75*), showing that MHC heterozygote individuals progressed significantly slower to AIDS-defining conditions, CD4+ T-cell count and/or death. Indeed, the role of MHC genes in slowing the progression to AIDS is now well established (*106*). 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 (*65*). 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 address the question whether MHC heterozygote advantage results from quantitative or qualitative differences among MHC alleles. For these goals, we use data from an unprecedented cohort of 6,311 HIV-1 infected Caucasian individuals and take advantage of well-established antigen-binding prediction algorithms for the human MHC (called Human Leukocyte Antigen or HLA). We predict individual-specific repertoires of HIV-1 peptides bound by classical HLA class-I molecules (HLA-A, -B and –C), and 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 genotype (at 4-digit allele resolution) and pre-treatment set point viral load (spVL), an established correlate of HIV-1 control and disease progression (*107*), for 6,311 HIV-infected individuals of European descent. We focused on classical HLA class-I genes as they are the only genes within the MHC region reported to be independently associated with HIV-1 progression (*65*, *75*). Overall 37, 69, and 27

alleles for HLA-A, HLA-B and HLA-C respectively were represented in the dataset **(Table S1)**. We screened all possible 9mer HIV-1 peptides (N = 3,252) across the entire HIV-1 proteome, and identified 409, 491 and 223 distinct peptides likely to be bound by one or more of the represented HLA-A, HLA-B and HLA-C alleles, respectively.

HLA heterozygote advantage

Initially, we aimed to replicate the HLA heterozygote advantage effect observed by Carrington et al. (*75*), since they had focused on time of progression to AIDS or death, while our data comprised only a proxy for the progression to AIDS (set point viral load). We thus tested whether heterozygosity at any of the classical HLA class-I loci was associated with lower viral load. 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 $P = 2.2 \times 10^{-6}$, respectively, after correcting for multiple testing), while heterozygosity at HLA-A showed no statistically significant effect on spVL (Wilcox rank sum test, $P = 0.16$) **(Fig. 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) (*108*–*110*) which can overshadow the individual effect of HLA loci on disease progression. In order to test whether the observed effects of HLA-B and HLA-C heterozygosity were independent of each other, we calculated the association of HLA-B heterozygosity with viral load only among HLA-C heterozygotes (N = 5748), thus controlling for HLA-C zygosity. This test demonstrated the effect of HLA-B heterozygosity to be independent of HLA-C heterozygosity ($P = 0.01$). Moreover following Carrington *et al*. (*75*), we tested if heterozygosity at multiple loci adds up to the conferred protection and observed that heterozygosity at two loci led to stronger resistance to disease progression compared to heterozygosity at only one locus **(Fig. S1)**. Overall, these results suggested that our dataset was appropriate to explore the effect of HIV heterozygote advantage against HIV and its underlying mechanism in more detail.

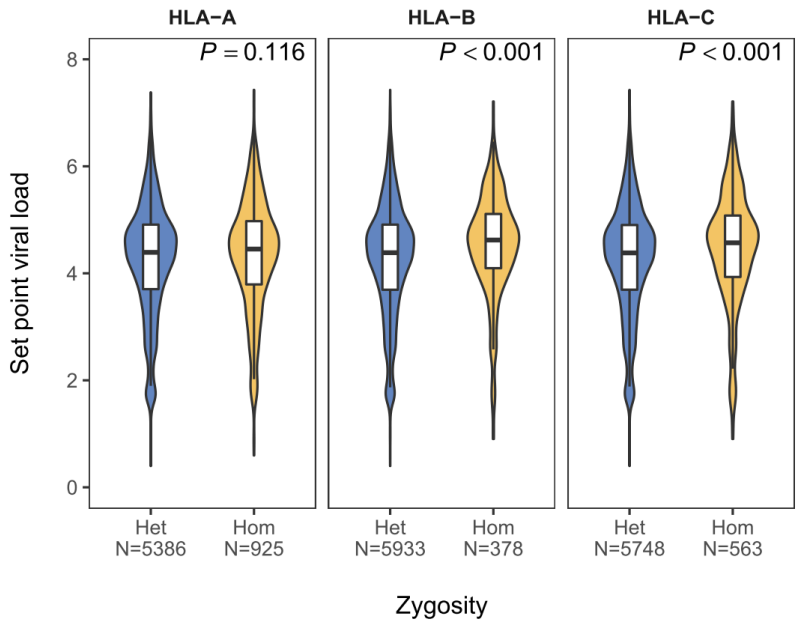


Fig. 1. Set point viral load in HLA heterozygotes and homozygotes. Comparison of the set point viral load between HLA homozygous and heterozygous individuals for the three classical HLA class I loci. Outlier values are not shown for better visual comparison. N indicates the number of individuals. Bonferroni-corrected *P*-values from Wilcoxon rank sum test are shown.

Divergent allele advantage

We sought to test the hypothesis of divergent allele advantage by evaluate the relation between sequence divergence of an individual's allele pair and viral load. We calculated pairwise sequence distance as a measure of sequence divergence 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 (*38*). First, we saw that the sequence divergence between an individual's HLA-B alleles was correlated with the divergence between HLA-C (Pearson correlation, r = 0.26, $P < 0.001$) and HLA-A alleles (r = 0.07, $P = 0.001$) **(Fig. S2)**, reflecting the gene organisation and LD pattern across the HLA class-I region. 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 viral load across individuals ($\tau = -0.07$, $P = 1.3 \times 10^{-14}$ in HLA-B heterozygous individuals) **(Fig. 2)**. While we found no such correlation for HLA-A ($P = 0.11$), HLA-C surprisingly showed a positive association between allele divergence and viral load ($\tau = 0.06$, $P = 3.2 \times 10^{-12}$ in HLA-C heterozygous individuals). 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 (*111*) and does not seem to be under selection for contributing to diverse immunological surveillance (*112*). 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 ($P < 0.001$) **(Fig. S3)**, supporting the notion that HLA-C is not evolving under the same selective constraints as HLA-B.

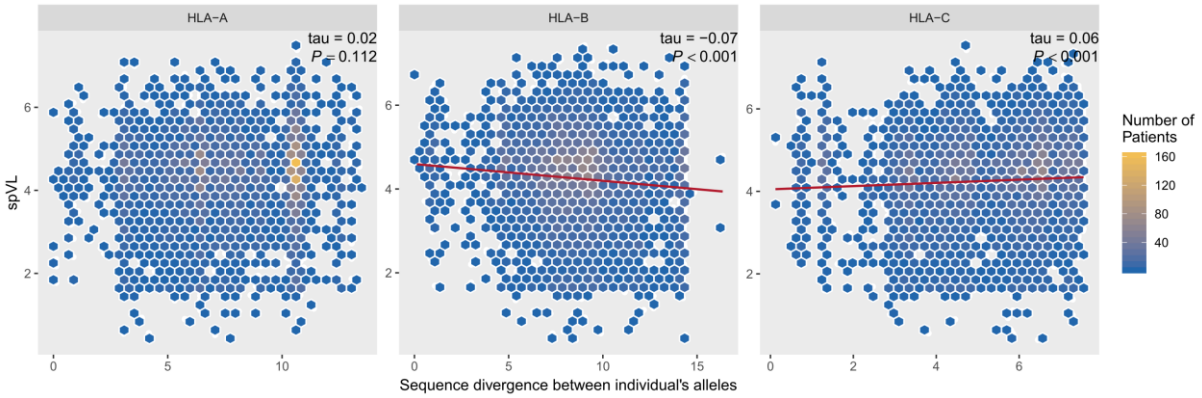


Fig. 2. Sequence divergence between HLA alleles and set point viral load. Correlation between set point viral load (spVL) and sequence divergence between individual's HLA-A **(Left),** HLA-B **(Center)** and HLA-C **(Right)** alleles. Each dot represents a heterozygous individual (N = 5386, 5933 and 5742 for HLA-A, -B and -C, respectively). Kendall's estimate of correlation τ and Bonferroni-corrected *P*-values are shown.

Functional consequence of heterozygosity and allele divergence

Observing effects of both heterozygosity and allele divergence in our data, we then asked whether these measures of genetic variability would indeed confer presentation of a broader array of HLA-bound pathogenic peptides, as hypothesized by the quantitative explanation for MHC heterozygote advantage. Using computational peptide-binding prediction, we found that heterozygosity always resulted in a broader array of bound peptides for all three classical HLA loci **(Fig. S4)**. Furthermore, the number of peptides bound by a pair of HLA-B alleles was positively correlated with the sequence divergence alleles **(Fig. S5)**. It was true for HLA-A and HLA-C as well **(Fig. S5)**. This association between sequence divergence at the HLA loci and predicted functional divergence among HLA alleles had been reported before (*37*, *38*). 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, τ = -0.12, P = 1.0x10^{-47}) **(Fig. 3)**. This was also true for HLA-A and HLA-C, though the correlation coefficient was relatively smaller (Kendall correlation, HLA-A: τ = -0.04, P = 1.7x10^{-5}; HLA-C: τ = -0.05, P = 8.8x10^{-8}). Interestingly, the association between viral load and the breadth of individual-specific HLA-B bound peptides was stronger (τ = -0.12) than the association between viral load and allele divergence (τ = -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.

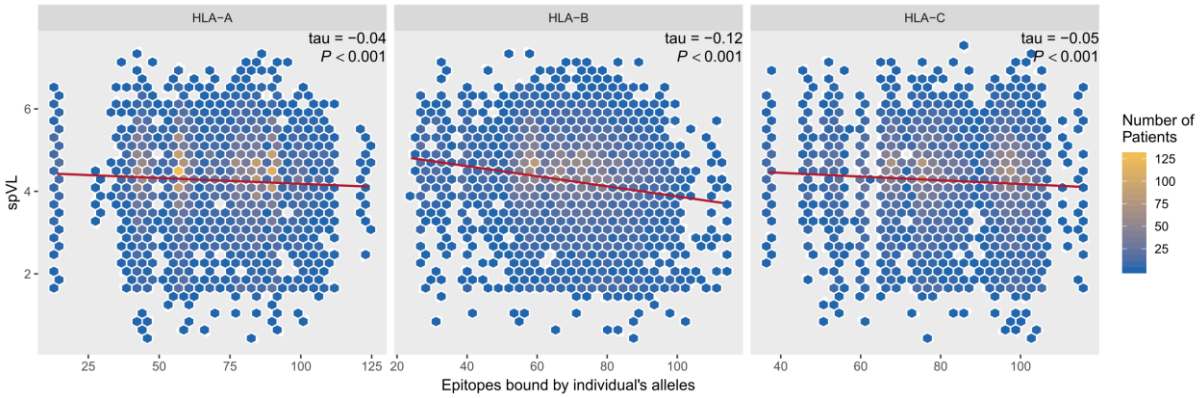


Fig. 3. Correlation between HLA-bound peptides and set point viral load. Correlation between set point viral load and individual-specific breadth of HIV-1 peptides bound by HLA-A **(Left),** HLA-B **(Center)** and HLA-C **(Right)** alleles. Each dot represents an individual (N = 6,311). Kendall's estimate of correlation (tau) and Bonferroni-corrected *P*-values are shown.

HLA heterozygosity and within-individual 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 (*7*, *113*). In this context, following the predictions of the quantitative heterozygote advantage hypothesis, heterozygote HLA genotypes should pose 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 limit set of autologous HIV-1 sequences from each individual, 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 to HIV control, and observed that virus sequences in HLA-B heterozygous individuals (N = 36) were indeed more diverged compared to the ones in homozygous individuals (N = 4) **(Fig. 4A)**. 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 is unlikely to be due to chance (one-tailed P = 0.028) **(Fig. 4B)**. Nonetheless, more autologous HIV sequences, particularly from HLA homozygous individuals will help to corroborate this finding.

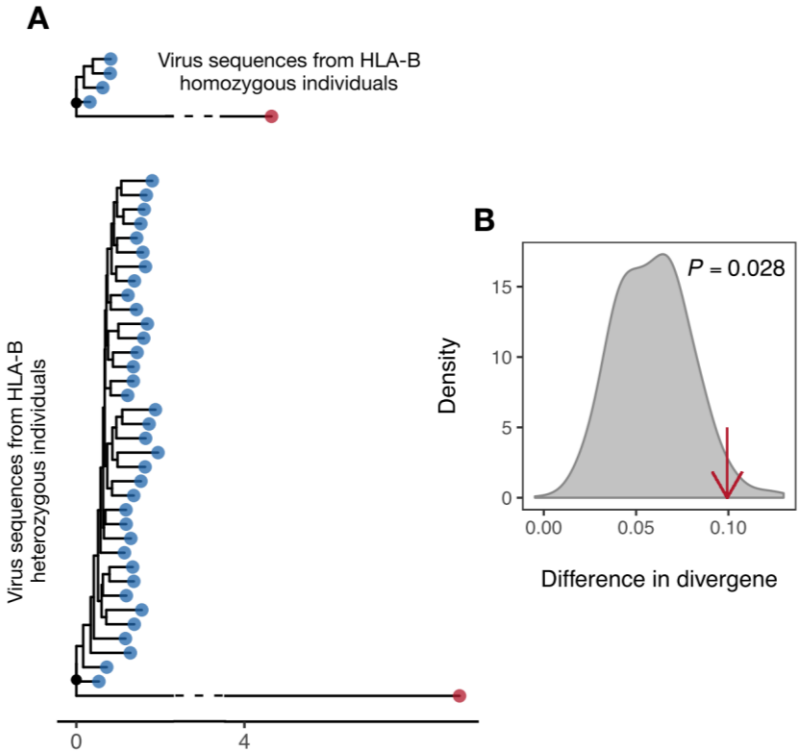


Fig. 4. Within-individual evolution of the virus in HLA homozygotes and heterozygotes (A) Rooted trees of the autologous virus sequences from HLA-B homozygous (N = 4) and heterozygous (N = 36) individuals. We used HIV-2 as outgroup to identify the root during tree construction. **(B)** The observed difference 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.

Allele-specific effects versus general heterozygote advantage

HLA alleles are known to show differential association with HIV control ranging from risk to protection (*65*). 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 **(Fig. S6)**, making this hypothesis a viable explanation. The insignificant 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) for the virus to mutate. We tested this point by focusing on HLA-B*57:01, the allele known to confer strongest resistance against disease progression (*65*, *114–116*). 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 $P < 0.001$; **Fig. 5 Right**). 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 $P < 0.001$; **Fig 5 Left**), 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 $P = 0.003$; **Fig. 5 Left**). 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 $P = 0.011$) **(Fig. 5 Right)**, 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.* In press; Chapter 2), maintaining the possibility that both qualitative and quantitative aspects of peptide binding are contributing to the protective effect of this allele.

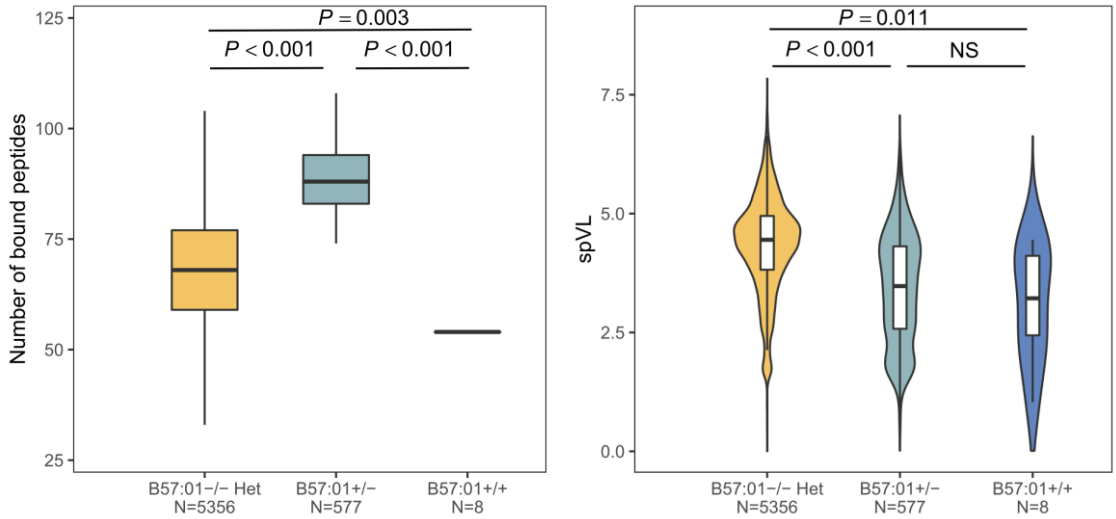


Fig. 5. Heterozygote advantage vs. allele-specific effect. Comparison of the number of bound HIV-1 peptides **(Left)** and set point viral load (spVL) **(Right)** in HLA heterozygous individuals not carrying HLA-B*57:01 allele, individuals carrying one copy of HLA-B*57:01 allele and individuals homozygous 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 to viral load remained significant after accounting for additive effect of individual 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 % for HLA-B **(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 (*65*).

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

	Associated variation in spVL in % (*P* value)		
Locus	**Heterozygosity**	**Sequence divergence**	**Bound peptides**
HLA-B	0.06 (0.016)	0.09 (0.005)	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 (*75*) might be the consequence of its linkage disequilibrium with neighboring HLA-B and/or HLA-C loci (*108*–*110*). While we recapitulated the general observation that higher sequence divergence between an HLA allele pair could lead to a larger number of bound peptides (*37*, *38*) (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 suggest that HLA-B heterozygote advantage is probably mediated via quantitative CTL response to a broad set of 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 (*65*), where unlike for HLA-B, there were no HIV-associated amino-acid residues found for HLA-C. Moreover, nearly only half the peptide-repertoire size of HLA-C alleles, relative to HLA-A and HLA-B, allows to speculate that HLA-C might be evolving not to work 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 (*117*). HLA-C molecules have been proposed to be potent ligand of KIRs (*117*, *118*), and specific interactions between HLA-C molecules and KIRs have been associated with multiple diseases, including HIV control (*119*–*121*).

Pereyra *et al.* 2014 have shown that CD8+ T-cell targeting of specific HLA-presented peptides could confer viral control (*122*). A broader array of HLA-bound peptides in heterozygous individuals might increase the possibility that such peptides are presented on the cell surface. Moreover, even though the B*57:01 allele bound fewer HIV-1 peptides compared to B*57:01-/- heterozygous individuals, the superior viral control conferred by

B*57:01 compared to general HLA-B heterozygote advantage suggests that an HLA allele might bind specific disease-relevant peptides and control the disease without HLA binding of a broader array of peptides and being heterozygous at HLA. However, since B*57:01 also bound the largest number of peptides among all HLA-B alleles, we cannot completely rule out that the possibility of quantitative advantage of binding a large number of peptides contributes to its association with HIV-1 control.

Together, these results suggest that HLA heterozygosity in an individual might confer advantage in multiple ways. One is the quantitative advantage through a broader HLA-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 evolution of the virus in HLA heterozygous individuals, possibly resulting in replicative fitness cost. Additionally, HLA heterozygosity provides an advantage by making it more likely to carry certain protective HLA alleles that can present specific peptides to T-cells and thus confer disease-control.

In conclusion, these results shed light on the functional basis of the protective association between HLA heterozygosity and HIV progression. They disentangle the role of quantitative and qualitative features of the HLA's peptide-repertoire in mediating the immune response. These results suggest that even a single pathogen can lead to selection for both, HLA heterozygosity (including excessive allele divergence) as well as 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.

Material and methods

Samples and genotype data

We analyzed data of 6,311 chronically HIV-1 infected individuals. The original data and thorough quality check are described in detail in McLaren *et al.* (*65*). Briefly, genome-wide genotype data had been collected from 8 independent GWAS studies and combined as part of the International Collaboration for the Genomics of HIV. Principal component analysis was used to infer ancestry using HapMap 3 (*123*) as a reference. Only samples grouping with HapMap Europeans were retained to ensure consistent ancestry for across samples. The first five principal components were used as covariates to account for residual population stratification in downstream analysis. SNP genotypes that had not been covered in original genotyping platforms were imputed with Minimac (*124*) using haplotypes from the 1,000 Genomes Project Phase 1 v3 reference panel. Imputed genotypes from other imputation protocols, Shapeit (*125*) and Impute2 (*126*), yielded highly concordant results. Imputed SNPs with low r2 score (< 0.3) or minor allele frequency of < 0.5% were discarded. HLA alleles (4-digit resolution) for classical class I loci (HLA-A, -B, -C) were imputed from genome-wide genotype data using best-guess genotypes. Available measurement of pre-treatment set point viral load (spVL; log10 HIV-1 RNA copies/μl of plasma) was used as quantitative disease phenotype for all individuals (*65*).

HLA binding affinity for HIV-1 peptides

We used the NCBI accession NC_001802.1 as the reference sequence for the HIV-1 proteome (M group subtype B). It comprised 10 proteins with sequence length ranging from 82 to 1435 amino acids. The *Gag-Pol* protein is a precursor protein that results from a −1 ribosomal frameshifting event in upstream *Gag* (*127*). It is cleaved by virus-encoded protease to produce the mature Pol protein. In our analysis, we manually trimmed the *Gag-Pol* protein sequence to Pol in order to avoid redundancy with the separate *Gag* protein. HLA class-I molecules preferentially bind and present 9mer peptides (*128*, *129*). NetMHCpan v4.0 is a computational method that predicts binding affinity of 9mer peptides to any known HLA class I molecule (*130*). The method has been trained on naturally eluted ligands and binding affinity data. It reports the rank of predicted binding affinity of HLA-peptide complexes against predicted affinity of random natural peptides. Using this tool, we predicted HLA allele-specific binding affinities for all 9mer peptides

generated from the entire HIV-1 proteome. HLA-peptide complexes with predicted binding affinity rank less than 0.5 were retained (corresponding to 'strongly bound' peptides) (*130*). 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 but accounting for multiple occurrences of peptides in the HIV-1 proteome.

Sequence divergence between alleles of HLA genotype per individual

Sequence divergence between alleles was computed for all HLA allele pairs (genotypes) of HLA-A, -B and -C genes. Protein sequences of HLA alleles were taken from IMGT/HLA database (*25*). Exons 2 and 3, which are known to encode for the variable region in the peptide binding groove of HLA class I molecules, were obtained following the annotation reported in Ensemble database (*131*). The alignment of amino acid sequences was performed using MUSCLE (*132*), and the sites containing alignment gaps at the beginning or the end of sequences were removed. The genetic distances between an aligned allele pair were calculated based on the Grantham distance matrix (*133*) using custom script (*38*). The non-parametric Kendall correlation was used to test for the associations of the sequence divergence between individual's HLA alleles with: (i) the set point viral load (spVL), (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 poor quality of sequences, we concatenated these 6 proteins in order to obtain better resolution and statistical power, and aligned using MAFFT v7 with default parameters (*134*). The alignment was optimized for maximum gap-free area using MaxAlign-1.1 (*135*), which resulted into 5 sequences for HLA-B homozygote and 37 for heterozygotes individuals. Maximum likelihood trees were made using PhyML-3.1 (*136*) with default parameters. The tip-to-tip distances were extracted using Ape-3.5 package (*137*) in R 3.5.1. Difference in the mean tip-to-tip distance in each group of individuals was taken as the observed difference in diversity. We obtained the 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 number of times the difference in mean diversity of permutations was equal to more than the observed difference.

Association with the viral load while controlling for allele-specific effects

The association of a variable with viral load (spVL) was calculated using linear regression model following McLaren *et al.* (*65*). Variation in viral load attributable to a given variable (heterozygosity, sequence divergence or the breadth of bound peptides) while controlling for allele-specific effects was calculated 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 HLA class-I locus separately that contained all imputed alleles (N = 69 for HLA-B; N = 37 for HLA-A; N = 27 for HLA-C) and first five principle components of SNP variation and the cohort identity (all adopted from McLaren *et al.* (*65*)) 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 v3.3.3 and data was visualized using the ggplot2 v2.2.1 package (*138*).

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.).

Supplementary data

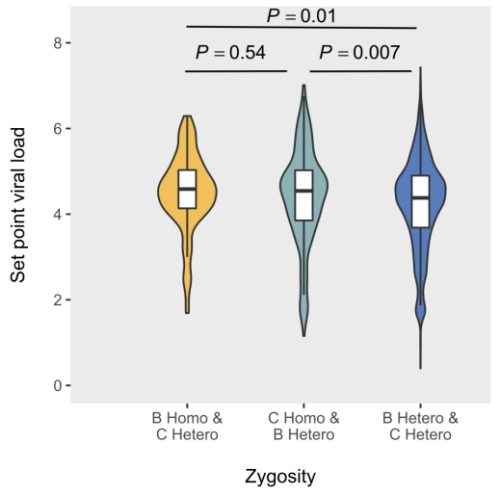


Fig. S1. Comparison of the set point viral load between individuals who are heterozygous for HLA-C only (N = 110), for HLA-B only (N = 295) and for both HLA-B and HLA-C (N = 5368). *P* values from Wilcoxon rank sum test are shown.

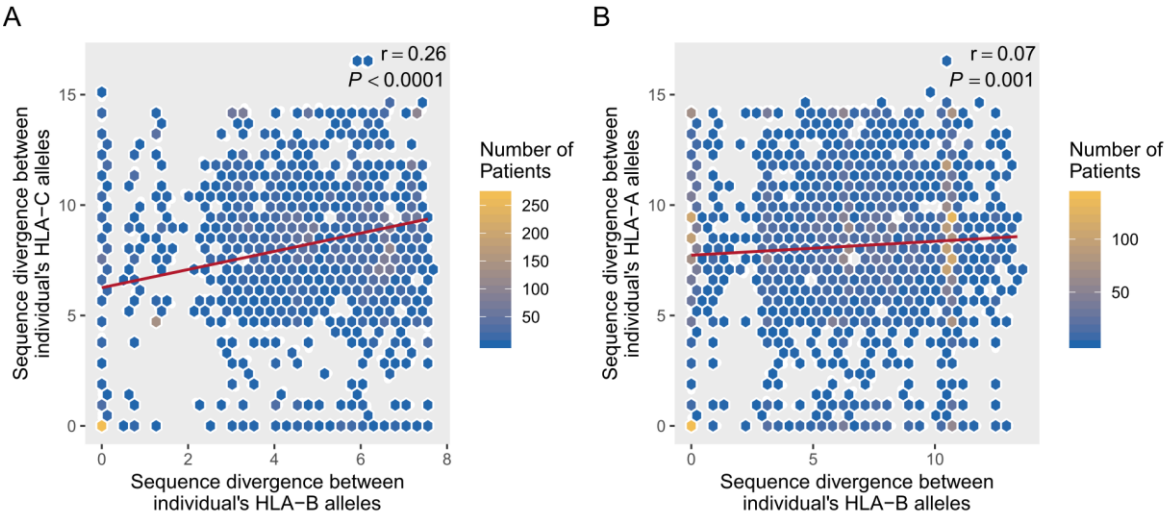


Fig. S2. (A) Correlation between the pairwise sequence divergence between HLA-B alleles and the pairwise sequence divergence between HLA-C alleles of an individual. **(B)** Alike correlation between the pairwise sequence divergence between HLA-B alleles and the pairwise sequence divergence between HLA-A alleles of an individual. Pearson's estimate of correlation (r) and *P*-value are shown.

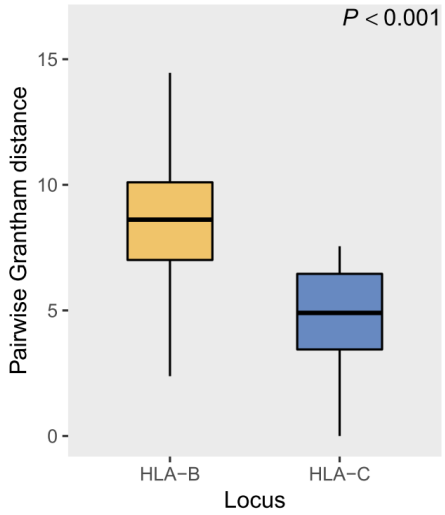


Fig. S3. Comparison of pairwise Grantham distance between individual's HLA-B and HLA-C alleles. *P* value from Wilcoxon rank sum test are shown.

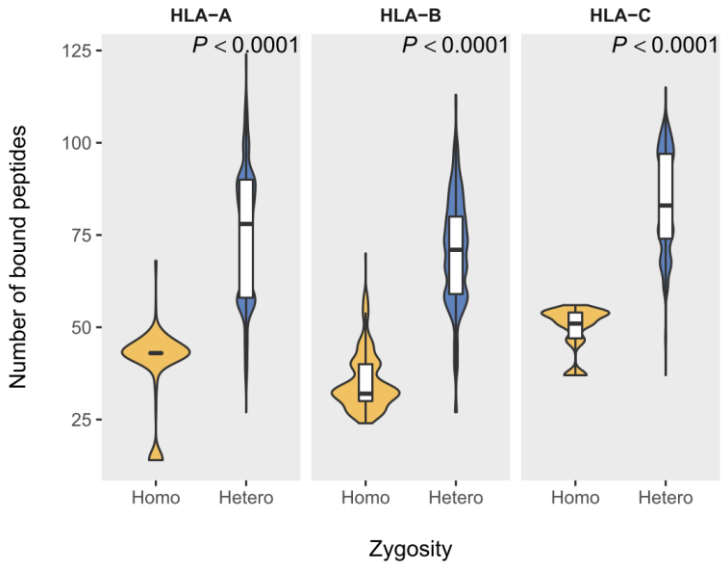


Fig. S4. HLA heterozygous individuals, whether for HLA-A, HLA-B or HLA-C, bound a significantly greater breadth of HIV-1 peptides compared to homozygotes.

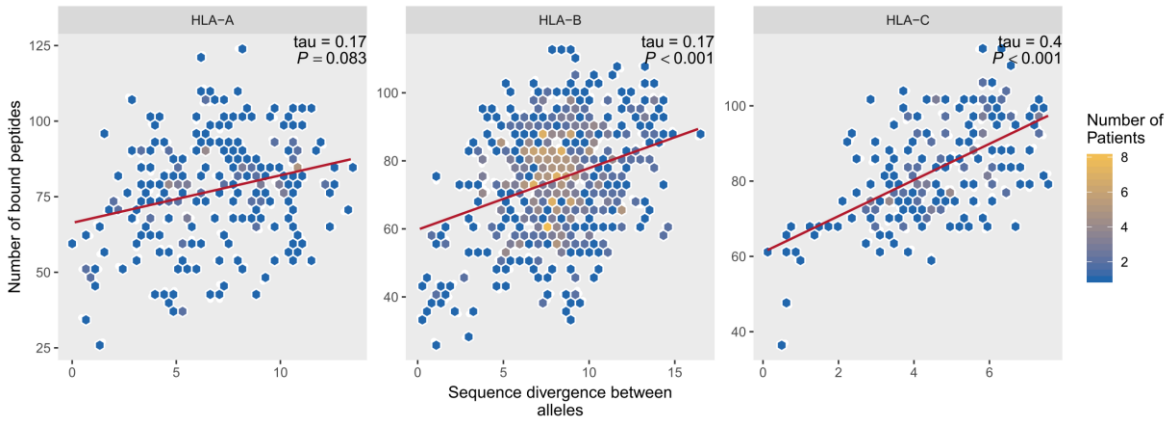


Fig. S5. Correlation between pairwise Grantham distance between individual's HLA alleles and the breadth of bound HIV-1 peptides for HLA-A **(left),** HLA-B **(center)** and HLA-C **(right)**.

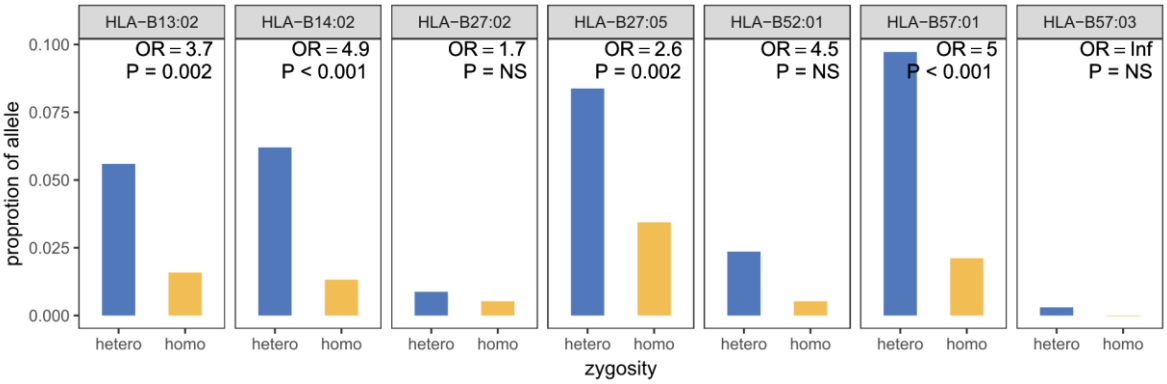


Fig. S6. Enrichment for protective HLA-B alleles in heterozygous individuals compared to homozygous individuals. Odds ratio (OR) and *P*-value from Fisher exact test are shown.

Table S1. The HLA alleles represented in the dataset. There were 37, 69, and 27 alleles for HLA-A, HLA-B and HLA-C genes, respectively, that were represented in the dataset.

HLA-B	HLA-A	HLA-C
B07:02	A01:01	C01:02
B07:04	A01:02	C02:02
B07:05	A01:03	C02:06
B08:01	A02:01	C03:02
B13:01	A02:02	C03:03
B13:02	A02:03	C03:04
B14:01	A02:05	C04:01
B14:02	A02:06	C04:03
B15:01	A02:11	C04:07
B15:03	A03:01	C05:01
B15:05	A03:02	C06:02
B15:07	A11:01	C07:01
B15:08	A11:02	C07:02
B15:10	A11:03	C07:04
B15:16	A23:01	C08:01
B15:17	A24:02	C08:02
B15:18	A24:07	C12:02
B15:25	A25:01	C12:03

B15:27	A26:01	C14:02
B18:01	A26:08	C15:02
B27:02	A29:01	C15:04
B27:03	A29:02	C15:05
B27:04	A30:01	C16:01
B27:05	A30:02	C16:02
B27:07	A30:04	C16:04
B35:01	A31:01	C17:01
B35:02	A32:01	C18:01
B35:03	A33:01	
B35:08	A33:03	
B35:12	A34:02	
B35:17	A36:01	
B37:01	A66:01	
B38:01	A68:01	
B39:01	A68:02	
B39:06	A69:01	
B39:09	A74:01	
B39:10	A80:01	
B40:01		
B40:02		
B40:06		
B41:01		
B41:02		
B42:01		
B42:02		
B44:02		
B44:03		
B44:04		
B44:05		
B45:01		
B46:01		
B47:01		
B48:01		
B49:01		
B50:01		
B51:01		
B51:02		
B51:05		
B51:06		
B51:08		
B52:01		
B53:01		
B55:01		
B56:01		
B56:04		
B57:01		
B57:02		
B57:03		
B58:01		
B73:01		

Chapter 2

HIV Peptidome-Wide Association Study Reveals Patient-Specific Epitope Repertoires Associated with HIV Control

Jatin Arora[1], Paul McLaren[2], Nimisha Chaturvedi[3], Mary Carrington[4], Jacques Fellay[3], Tobias L Lenz[1]

[1]Max Planck Institute for Evolutionary Biology, [2]Public Health Agency of Canada, [3]EPFL, [4]SAIC-Frederick

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Genetic variation in the peptide-binding groove of the highly polymorphic human leukocyte antigen (HLA) class I molecules has repeatedly been associated with HIV-1 control and progression to AIDS, accounting for up to 12% of the variation in HIV-1 set point viral load (spVL). This suggests a key role in disease control for HLA presentation of HIV-1 epitopes to cytotoxic T cells. However, a comprehensive understanding of the relevant HLA-bound HIV epitopes is still elusive. Here we describe a peptidome-wide association study (PepWAS) approach that integrates HLA genotypes and spVL data from 6,311 HIV-infected patients to interrogate the entire HIV-1 proteome (3,252 unique peptides) for disease-relevant peptides. This PepWAS approach predicts a core set of epitopes associated with spVL, including previously characterized epitopes but also several novel disease-relevant peptides. More importantly, each patient presents only a small subset of these predicted core epitopes through their individual HLA-A and -B variants. Eventually, the individual differences in these patient-specific epitope repertoires account for the variation in spVL that was previously associated with HLA genetic variation. PepWAS thus enables a comprehensive functional interpretation of the robust but little understood association between HLA and HIV-1 control, prioritizing a short list of disease-associated epitopes for the development of targeted therapy.

HLA | HIV-1 | Epitope prediction | Antigen presentation | Evolution

HLA class I proteins are thought to play a critical role in immune recognition of HIV-1 by presenting endogenously processed viral peptides at the surface of infected cells to cytotoxic T cells, in order to trigger destruction of the infected cells (1). Indeed, genetic variation in the HLA region has repeatedly been identified as the major genetic determinant of HIV-1 control in genome-wide association studies (2, 3). Most recently, McLaren *et al.* (4) fine-mapped the entire HLA's association with HIV-1 control and disease progression to five independent amino acid residues in the peptide binding groove of the HLA-B and HLA-A molecules. These five residues alone accounted for 12.3% of the variation in viral load, suggesting a major role for specific HLA-presented viral epitopes in HIV-1 control. However, our understanding of the disease-relevant viral epitopes is still incomplete, hampered by the economically hardly feasible challenge of employing a full-factorial experimental assay to screen the entirety of the HIV-1 peptidome for binding by all relevant HLA alleles. Therefore, we developed a novel computational analysis approach that identifies and prioritizes disease-associated peptides based on individual HLA genotype and disease phenotype information. Our approach uses established computational algorithms to predict for each individual whether a given peptide is bound by the individual's HLA variants, and then uses regression analysis on the disease phenotype (here HIV set point viral load) to estimate whether the ability to bind the peptide is non-randomly associated with the disease phenotype. This approach is analogous to a genetic association study, except that it incorporates one additional layer by translating genetic variation into functional variation (HLA variant-specific peptide binding). Importantly, this approach does not simply define all peptides bound by a risk HLA variant as risk peptides. *Instead, for each peptide it integrates the disease effect of all HLA variants that are able to bind the peptide and thus estimates a peptide-specific association with disease.* Since most peptides are bound by several HLA variants, integrating the effect of all binding HLA variants is essential (**Fig. 1**). For instance, a peptide can have no association with disease, even if it is bound by the highest risk variant, simply because it is also bound by several other non-associated (or even protective) variants. Ultimately, our approach identifies a list of peptides with varying associations to disease, which can directly inform therapy development by prioritizing global as well as patient-specific candidate epitopes. As a proof-of-concept, we analyze here a unique dataset of 6,311 individuals of European ancestry with chronic HIV-1 infection (***SI Appendix*, Table S1**). Screening the entire HIV-1 peptidome for candidate epitopes, we identify a comprehensive list of peptides that explain the well-established association between HLA genetic variation and HIV-1 control, including several previously uncharacterized epitopes as novel candidates for targeted therapy.

Results and Discussion:

Our analyses are based on a large dataset of HIV-infected individuals (4) that includes both pre-treatment level of set point viral load (spVL) as a correlate of disease progression (5) and imputed HLA genotypes (4-digit allele resolution). We focused on the two HLA loci (HLA-B and HLA-A) reported to have independent associations with HIV-1 control and disease progression (4). Po-

Significance

Individual differences in HIV-1 control and progression to AIDS have been pinpointed to genetic variation in the Human Leukocyte Antigen (HLA), coding for antigen-presenting molecules. However, our understanding of the corresponding antigens is still incomplete. Here we developed a new approach that combines HLA genotypes and viral load data of HIV infected individuals to screen the entire HIV proteome for disease-relevant peptides. Our PepWAS approach identified a limited manageable core set of peptides, accounting for the entire variation in viral load previously associated with genetic variation in the HLA. This core set of disease-relevant antigens thus provides a functional link between HLA genetic variation and HIV-1 control, confirming several known antigens, but also prioritizing novel antigens as new therapeutic targets.

Reserved for Publication Footnotes

This chapter has been accepted for publication in PNAS (the email from PNAS is attached in Annex)

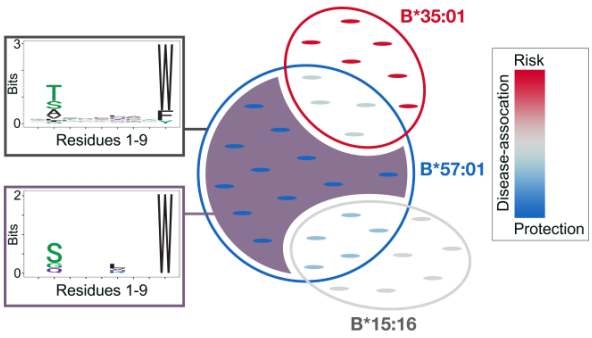
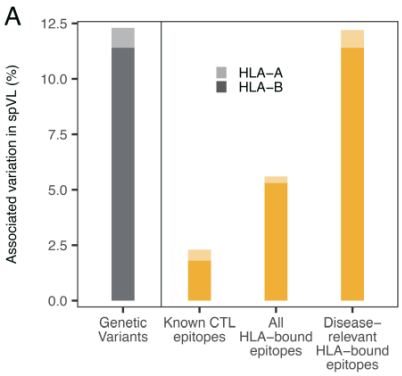


Fig. 1. Schematic for determining peptide-specific associations through PepWAS. Disease-associated peptides are identified by integrating the different disease-associations of the different HLA alleles that are predicted to bind them. Some peptides will only be bound by one HLA allele, thus drawing their disease-association directly from the disease-association of that allele (e.g. peptides in the purple shaded area, bound only by HLA-B*57:01). However, many peptides will be bound by several HLA alleles, which can have quite distinct, possibly even opposing disease associations (e.g. peptides in overlap of *B57:01 and *B35:01). In this case, the disease-association of the peptide derives from the disease-associations of each of the binding HLA alleles as well as their frequencies in the dataset. The novel peptidome-wide association study (PepWAS) approach differentiates these distinct sets of peptides and identifies both specific peptides and epitope motifs with distinct disease-association (e.g. distinct motif of purple shaded peptides, corresponding to the dark purple cluster in Fig. 5). Circles depict repertoires of peptides (small pointed ovals) predicted to be bound by the given HLA allele. Overlap of circles defines sets of peptides bound by both HLA alleles. Color of circles and peptides depicts disease-association of corresponding HLA alleles and peptides, respectively, from blue (protective) to red (risk). The number of peptides in this schematic does not correspond to the actual number of peptides observed for these HLA alleles. In reality, the overlap among HLA alleles is substantially more complex than depicted in this simplified schematic.



B

Model parameters	HLA-B	HLA-A	HLA-B + HLA-A
Covariates	–	–	–
Covariates + Known CTL epitopes	1.8 ($4.1x10^{-27}$)	0.5 ($1.7x10^{-8}$)	2.3 ($1.8x10^{-33}$)
Covariates + All HIV-1 epitopes	5.3 ($2.1x10^{-77}$)	0.3 ($1.8x10^{-6}$)	5.7 ($4.9x10^{-82}$)
Covariates + Disease-associated epitopes	11.4 ($5.1x10^{-170}$)	1.4 ($2.9x10^{-20}$)	12.2 ($5.3x10^{-180}$)

Fig. 2. Variation in viral load associated with predicted epitope repertoires bound by HLA-B and HLA-A. Among HIV patients (N = 6,311), the proportion of variation (estimated as adjusted ΔR^2) in set point viral load (spVL) associated with the patient-specific number of predicted HLA-bound HIV-1 epitopes is shown separately for HLA-B and HLA-A, and for different epitope sets. (**A**) Previously, 11.4% and 0.9% of the variation in spVL had been associated with independent genetic variants in HLA-B and HLA-A, respectively (grey bars; data from ref. 4). Here we instead calculated the variation in spVL associated with individual HLA-bound HIV epitope repertoires (yellow bars), based on known CTL epitopes from Los Alamos HIV Molecular Immunology Database, all HLA-bound HIV epitopes, and only the disease-associated HIV epitopes (the latter corresponding to 99.2% of the variation previously associated with HLA genetic variation). (**B**) Variation associated with different sets of predicted epitopes. *P*-values (in parentheses) indicate the improvement over null model (covariates only: first five PCs and cohort group). Number of disease-associated predicted epitopes is 132 for HLA-B, and 74 for HLA-A, respectively.

tential HLA-bound peptides were identified using an established computational algorithm that is based on empirical training data (6) and integrates several complementary prediction methods in a consensus approach, outperforming comparable algorithms (6, 7). Such algorithms have been used in a wide spectrum of HLA-related studies ranging from vaccine design to cancer evolution and HIV disease genetics (8–10). Without *a-priori* selection, we screened all possible 9mer HIV-1 peptides (N = 3,252) in a sliding window across the entire HIV-1 M group subtype B reference proteome (11) against all represented HLA-B and HLA-A alleles (344,712 HLA:peptide complexes), and identified 214 and 173 distinct HIV-1 peptides predicted to be bound by one or more of the represented HLA-B and HLA-A alleles, respectively.

In order to evaluate the significance of the predicted peptide repertoires, we interrogated several layers of empirical evidence (see ***SI Appendix*, Supporting Text**). We observed an enrichment for previously known HIV-1 epitopes (***SI Appendix*, Fig. S1A**), a correlation between an HLA-B allele's effect on viral load and the number of HIV-1 peptides it is predicted to bind (***SI Appendix*, Fig. S1B**), and detected previously reported viral escape mutations (***SI Appendix*, Fig. S1C**). Following these independent layers of evidence that our analysis pipeline predicts disease-relevant binding of HLA to HIV-1 peptides, we subsequently refer to the entire predicted set of HLA-bound peptides as *predicted epitopes*, highlighting the point that not all of them have been experimentally validated.

Next, we tested whether the patient-specific repertoire of predicted HIV-1 epitopes, defined by the number of peptides predicted to be bound by the specific HLA allele combination of the patient, was associated with spVL. For this, we ran a linear regression across the 6,311 HIV-1 patients, with spVL as dependent variable and the patient-specific number of bound peptides as predicting variable, together with other covariates (see methods). We first focused on the effect of peptides bound by HLA-B, and used only the known CTL epitopes from the Los Alamos HIV Molecular Immunology Database (12), of which 80 were represented among the 214 predicted HLA-B bound epitopes. The individual number of these known CTL epitopes bound by patient-specific HLA-B variants accounted for only 1.8% of the individual variation in spVL (**Fig. 2**). In order to evaluate this association, we then included all predicted HLA-bound HIV-1 epitopes (N = 214) in the analysis, including the previously known CTL epitopes as well as any other HLA-B-bound peptide from the HIV-1 proteome. Interestingly, the total number of all predicted HLA-B-bound epitopes per patient accounted for 5.3% variation in spVL (**Fig. 2**), suggesting that the Los Alamos CTL epitope dataset is not yet fully saturated with regard to disease-relevant peptides. However, the accounted variation was still lower than the 11.4% variation associated with genetic variation at HLA-B in previous genotype-based studies, suggesting that the total predicted epitope repertoire still included peptides irrelevant for the association between HLA and HIV-1. This is supported by a previous study, which showed that not all HLA-bound peptides are epitopes targeted by CD8+ T-cells (13). We thus aimed to refine the repertoire of predicted HLA-bound HIV-1 epitopes further to comprise only disease-relevant epitopes. For this, we calculated the epitope-specific association with spVL by running a separate linear regression for each predicted

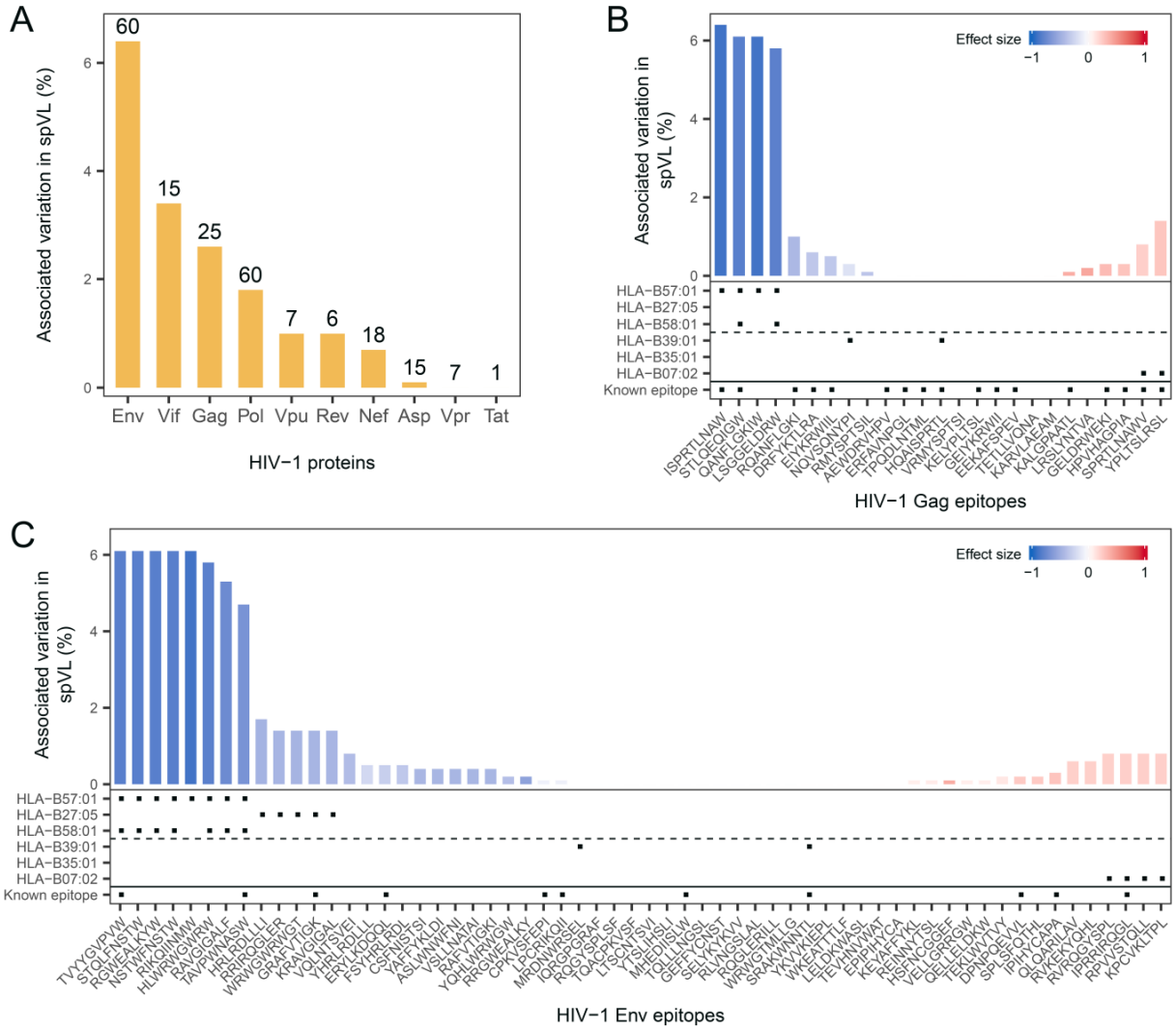


Fig. 3. Epitope- and protein-specific association with viral load. (**A**) Percent of variation in spVL associated with all predicted epitopes of a given HIV-1 protein. Absolute number of predicted HLA-B bound epitopes per protein is shown above the bars. (**B-C**)Predicted HLA-B-bound epitopes accounted for varied levels of variation in set point viral load (spVL). Height of the bar represents the fraction of variation in spVL associated with each epitope, while the color reflects each epitope's effect on spVL, ranging from protection (blue) to risk (red). Note that epitope effects are estimated separately and are thus not independent. *Gag* (**B**) and *Env* (**C**) proteins are shown as representative examples, together with information on predicted binding for 3 protective and 3 risk HLA-B alleles highlighted in a recent review (24) and whether peptides are known epitopes in Los Alamos HIV database. All other HIV-1 proteins are shown in ***SI Appendix*, Fig. S8**.

epitope and recording R^2 and β-coefficient as measures of the epitope's effect on spVL. This is analogous to the approach of a genome-wide association study (GWAS), where each genetic variant is tested for its association with a given trait, except that here we focus on functional protein variation (peptide binding by a patient's HLA molecules) rather than genetic variation. Following this analogy, we term our approach *peptidome-wide association study* (PepWAS). Of 214 HIV-1 epitopes predicted to be bound by HLA-B, 132 accounted for nominal variation (adjusted R^2 value > 0) in spVL, 74 of which were negatively and 58 positively associated with spVL (β-coefficients ranging from -0.1 to 0.77; ***SI Appendix*, Table S2**). Importantly, we do not require statistical significance at this point as this is a candidate screen and we thus aim to minimize the number of false negatives. Subsequently, we designate the nominally associated epitopes as *disease-associated* predicted epitopes, even though their effects are not necessarily independent as they were tested with separate regression models. An analogous investigation of peptide binding by HLA-A alleles revealed an additional 74 disease-associated epitopes (***SI Appendix*, Table S3**).

Having refined the predicted HIV-1 epitope repertoire to only disease-associated predicted epitopes, we then tested whether this subset accounted for a larger fraction of the variation in spVL than the total predicted HIV-1 epitope repertoire. Indeed, the patients' ability to bind a smaller or larger fraction of the HLA-B-specific disease-associated predicted epitopes accounted for 11.4% of the variation in spVL (**Fig. 2**). Similarly, for HLA-A, the total number of predicted HIV-1 epitopes bound by individual HLA-A genotypes accounted for 0.3% of the variation, while

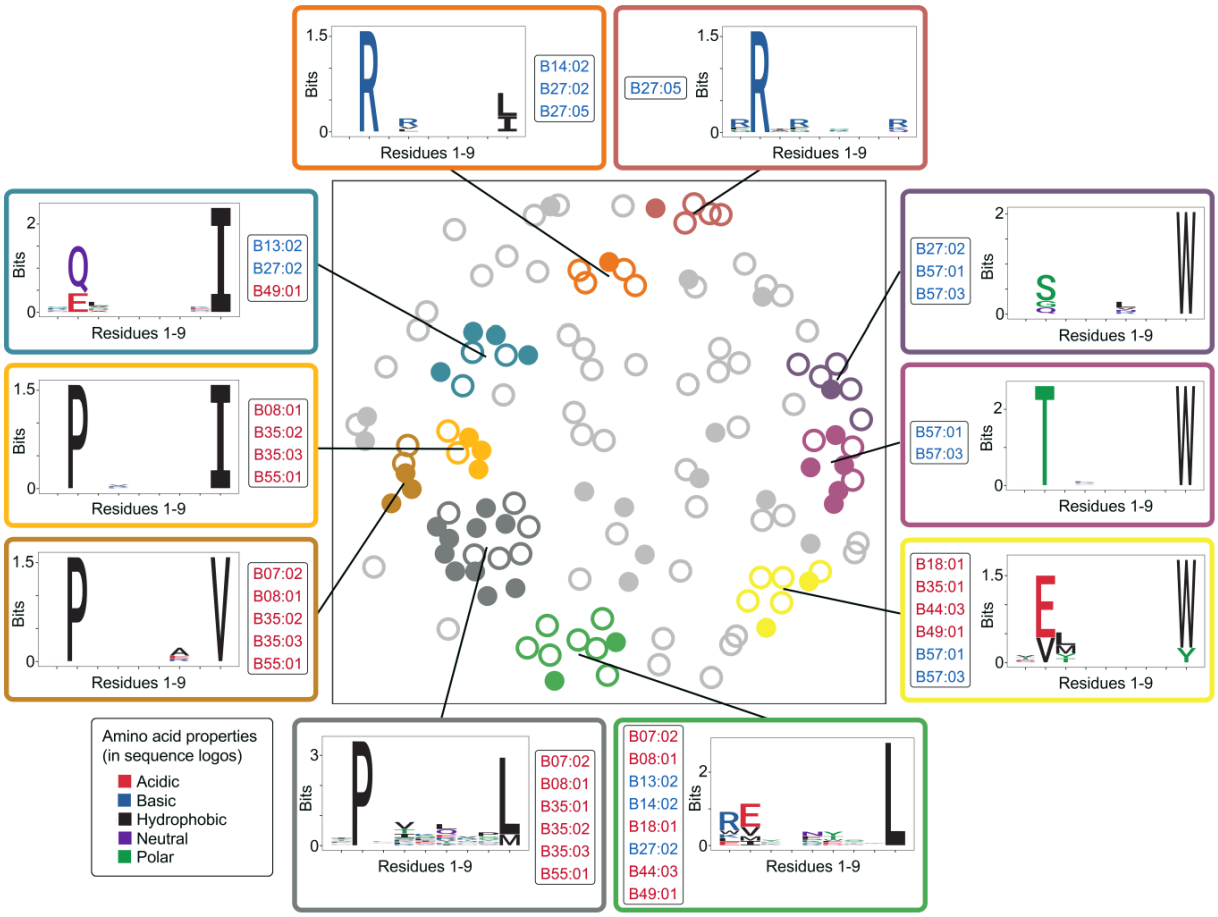


Fig. 4. Clusters of disease-associated epitopes. Non-metric multidimensional scaling (NMDS) was used to visualize the pairwise distance between predicted HLA-B-bound disease-associated epitopes, which revealed 10 dominant clusters. Each circle represents an HLA-B bound disease- associated epitope (N = 132). Filled circles represent known CTL epitopes from the Los Alamos HIV Molecular Immunology Database (N = 45), while open circles represent previously uncharacterized disease-associated predicted epitopes. Cluster-specific motif and HIV-1 associated HLA-B alleles (N = 16) binding the cluster's epitopes are shown. The coloring of the allele names indicates disease-association of the specific alleles.

disease-associated predicted epitopes accounted for 1.4% of the variation in spVL. On average, a patient's HLA-B allele pair bound 16.2 ±7 (SD) disease-associated predicted HIV-1 epitopes, while its HLA-A alleles bound significantly less (6.6 ±6.5; Paired Wilcox rank sum test, $P < 0.0001$; ***SI Appendix*, Fig. S6**). This quantitative difference in peptide presentation might contribute to the stronger spVL-association of HLA-B compared to HLA-A, as a larger number of presented peptides should more likely lead to a more efficient CD8 T cell response, as has indeed been observed for HLA-B compared to HLA-A (14). HLA-C-bound epitopes did not show any significant association with spVL, mirroring the lack of independent genetic associations for HLA-C in the latest GWAS (4). Predicted disease-associated epitopes of HLA-B and HLA-A together accounted for 12.2% of the variation in HIV-1 viral load, approximately corresponding to the 12.3% variation previously attributed to all independent genetic associations in the entire HLA (**Fig. 2A**).

Interestingly, the *Env* protein showed the largest number of disease-associated predicted epitopes, with both positive and negative effects. Among the disease-associated predicted HLA-B-bound epitopes, *Env*-derived epitopes alone accounted for 6.4% of variation in spVL, the highest among all HIV-1 proteins (**Fig. 3A**). In addition to already known *Env*-derived CD8+ T-cell targeted epitopes associated with lower viral load and disease control e.g. RIKQIINMW, HRLRDLLLI (13), ERYLKDQQL (15), our analysis revealed previously undescribed HLA-epitope complexes e.g. B*57:01-STQLFNSTW, -NSTWFNSTW, or -RGWEALKYW showing strong associations with lower viral load (**Fig. 3C**). The potential importance of the predicted *Env* epitopes is quite surprising, since the high genetic variability of the *Env* protein across different HIV-1 isolates suggests that the virus could readily evolve escape variants in this protein. However, a previous study has already established that sequence conservation alone is not a reliable predictor of protective epitopes, instead highlighting structural conservation as the more important feature (13). More intriguingly, we found that the protective *Env* epitopes predicted through our PepWAS approach are significantly enriched for residues that are associated with broadly neutralizing antibodies (bNAbs; OR = 1.5, $P = 0.036$, ***SI Appendix*, Fig. S7**), suggesting that they represent parts of the Env protein that can be efficiently targeted in both antibody therapy as well as in HLA-mediated CTL response.

Notably, several of the represented HLA alleles were predicted to bind both negatively and positively disease-associated epitopes (***SI Appendix*, Tables S4 and S5**), i.e. epitopes bound by the same HLA allele did not necessarily have the same effect

on viral load. This can be explained by the fact that a given epitope can be bound by several different HLA alleles with very distinct disease association (see schematic in **Fig. 1**). This is also in agreement with a previous study showing that viral control is mediated by specific immunogenic epitopes which could be restricted by HLA alleles other than already known ones (13).

HLA molecule variants are known to bind peptide repertoires with distinct anchor motifs, based on the composition of their peptide-binding groove (16). This entailed the possibility that our PepWAS approach is merely identifying distinct groups of peptides per HLA variant, thus translating the known HLA variant-specific effect on viral load into peptide group-specific effects. While still helpful in guiding epitope research, this would provide only limited knowledge-gain compared to the HLA allele-specific associations known from previous work (4). In order to test for this possibility, we performed a cluster analysis on the predicted disease-associated epitopes bound by HLA-B (N = 132) and analyzed cluster-specific motifs and HLA allele binding patterns. Intriguingly, among the ten most dominant epitope clusters, each exhibiting a distinct peptide motif, nine were defined by multiple HLA-B alleles (**Fig. 4**), some of them even belonging to different supertypes (***SI Appendix*, Table S7**). All of these clusters included both novel and previously described epitopes, and three of them were defined by both risk- and protection-conferring alleles. Furthermore, all HLA variants bound peptides of multiple dominant clusters; e.g. B*57:01 is associated with 3 dominant clusters, each showing a distinct peptide motif, but all showing a strong preference for amino acid 'W' at anchor position 9 (**Fig. 4**). Overall, the cluster analysis shows that our PepWAS approach identifies groups of peptides with distinct motifs that are different from HLA variant-specific binding motifs (see also schematic in **Fig. 1**). Generally, the 24 disease-associated epitopes predicted to be bound by HLA-B*57:01 (but some of these also by other alleles), accounted for the highest level of variation in spVL, even though they derived from 5 different HIV-1 proteins (**Fig. 3B, C and *SI Appendix*, Fig. S8**). One of these epitopes, the well characterized HIV-1 *Gag* epitope ISPRTLNAW (belonging to the dark purple cluster in **Fig. 4**), slightly exceeded the effect of all other epitopes (**Fig. 3B**), in concordance with experimental evidence (17). Other HLA-B alleles, including the B*08, B*44, and B*51 types, were also included in our dataset, and their predicted epitope repertoires roughly followed their disease-association known from previous studies (***SI Appendix*, Fig. S1B;** allele-specific associations and number of bound peptides are given in ***SI Appendix*, Table S4**).

Mechanistically, a negative association between predicted HIV-1 epitopes and viral load is intuitive and likely resulting from the peptides' immunodominant role in CTL response and their escape mutations leading to significant fitness costs for the virus. However, a number of the predicted HIV-1 epitopes exhibited a positive association with viral load, indicating that they confer lower disease protection relative to the bulk of the peptides. They likely represent peptide variants that fail to elicit an efficient CTL response or can readily mutate with negligible fitness effects, thus allowing viral escape from HLA presentation at no cost for the virus. Indeed, the most risk-associated predicted *Vpu* epitope, IPIVAIVAL (***SI Appendix*, Fig. S8F;** belonging to the largest, grey cluster in **Fig. 4**), includes an anchor residue that exhibits significant variation in primary HIV-1 clones and is involved in mediating immune-evasion through down-regulation of HLA-C (18), whose high expression has been implicated in HIV control (19). The lack of significant associations between predicted HLA-C bound epitopes and viral load in our analysis might indicate that previously observed viral control associated with HLA-C is not mediated through specific peptide presentation of HLA-C. However, more research is required to fully understand the role of HLA-C in viral control (18).

So far, our analysis was based on the HIV-1 genome reference sequence. Though widely used for research, focusing on this sequence accession may restrict our findings. We thus repeated the entire analysis using the HIV-1 proteome consensus sequence from the Los Alamos database, which incorporates major variation across different HIV-1 strains. The results remained qualitatively the same (***SI Appendix*, Fig. S9 and Table S8**). However, HIV is well known to exhibit substantial within-host evolution (20, 21) and it is easily conceivable that the ability of a patient's HLA variants to bind HIV epitopes is significantly affected by genetic variation in the patient's HIV population (22). We therefore also analyzed patient-specific autologous HIV-1 sequence information, which was available for a small subset of patients, covering 8 of the 10 HIV-1 proteins (***SI Appendix*, Table S6**). For 4 of the 8 proteins (*Gag, Pol, Vif* and *Nef*) we found that the proportion of variation in spVL associated with HLA-bound epitope repertoires changed when predicting epitopes from autologous sequences instead of from the reference sequence. In all 4 cases, the variation associated with predicted autologous epitopes was higher than when using their homologs from the reference sequence (***SI Appendix*, Table S6**), suggesting that our PepWAS approach might be able to explain more variation in spVL than a standard GWAS if autologous sequences were available for a larger fraction of infected individuals.

PepWAS relies on computational algorithms for the prediction of binding affinities between HLA variants and peptides, and is thus inherently limited by their accuracy and specificity. For instance, the empirical data used to train currently established HLA class I algorithms contains mainly 9mer peptides, even though HLA class I molecules can occasionally bind slightly shorter or longer peptides. Such peptides might therefore be missed by current prediction algorithms. On the other hand, the current setup does in fact identify 9mer cores of larger known epitopes. For instance, the here predicted protective 9mer *Gag* epitope 'STLQEQIGW' resides within the previously described 10mer *Gag* epitope TW10 (**Fig. 3B**). Furthermore, this limitation is likely to be alleviated as more training data is becoming available.

Overall, our findings reveal a functional basis of the robustly established association between HLA genes and HIV-1 infection outcome. We show that both quantity and quality of HLA-bound HIV epitopes contribute to controlling a patient's viral load. Our data also suggests a more important role for *Env* protein-derived epitopes than previously thought. Ultimately, our PepWAS approach of combining computational HLA-specific epitope prediction with disease phenotype validation provides a promising avenue for identification and prioritization of novel epitopes. As such, it complements existing empirical essays for the development of targeted therapy. Noteworthy, by involving a functional layer (peptide binding), the PepWAS approach enables the detection of disease-relevant properties that are shared among several genetic variants (overlap in peptide binding among HLA alleles). Such shared properties would be undetectable by GWAS, because of its focus on distinct genetic variants instead of function, and should therefore lead to higher sensitivity in the PepWAS approach compared to GWAS. Furthermore, the PepWAS approach allows to account for individual variation in the pathogen proteome if autologous sequence information is available, potentially further increasing sensitivity. As such it may be applied to any HLA-associated complex disease.

Material and Methods:

For detailed information on Material and Methods see *SI Appendix* Supporting Methods available online.

Samples and Genotype data:

We analyzed HLA genotype data and set point viral load (spVL) measurements of 6,311 subjects chronically infected with HIV-1. The original data and thorough quality check are de-

scribed in detail in McLaren *et al.* (4) and explained briefly in Supporting Methods.

HLA binding affinity for HIV-1 epitopes:

We used the NCBI accession NC_001802.1 as the reference sequence for the HIV-1 proteome (M group subtype B). The algorithm NetMHCcons-1.1 was used to predict HLA allele-specific binding affinities for all 9mer peptides generated from the entire HIV-1 proteome, applying the default affinity rank threshold for 'strongly bound' peptides (rank < 0.5).

Association with viral load:

The association of an allele or a peptide with viral load (spVL) was calculated using a linear regression model corrected for population covariates following McLaren *et al.* (4). Covariates included the first five principle components of SNP variation and the cohort identity (all adopted from McLaren *et al.* (4)). Variation in viral load attributable to a given variable (allele or epitope) was calculated as the difference between adjusted-R^2 values of the model with variable and covariates and the model with covariates only, following McLaren et al. (4). The variable's regression coefficient was used as the measure of its effect on viral load.

Clustering of HLA-B-specific predicted epitopes:

Position-associated entropy was calculated for all HLA-B-bound disease-associated epitopes (N = 132) and used for visualization in a non-metric multidimensional scaling plot as well as for density-based clustering.

HLA binding of peptides from autologous HIV-1 sequences:

We analyzed autologous HIV-1 sequences from Bartha et al. (23). Autologous sequences were available for 8 of 10 HIV-1 proteins (only Gp41 segment for *Env*), and only for a small subset of patients in our cohorts (***SI Appendix*, Table S6**).

Acknowledgements:

Patient and HIV sequence data was collected and generously provided by the International Collaboration for the Genomics of HIV. This project was supported by the Emmy Noether Programme of the Deutsche Forschungsgemeinschaft (DFG grant LE 2593/3-1 to T.L.L.). Furthermore, 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 also supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research. **Competing interests:** The authors declare no conflict of interest.

1. Goulder PJR, Walker BD (2012) HIV and HLA Class I: An Evolving Relationship. *Immunity* 37(3):426–440.
2. Fellay J, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science (80-)* 317(5840):944–947.
3. Pereyra F, et al. (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science (80-)*. doi:10.1126/science.1195271.
4. 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(47):14658–14663.
5. Mellors JW, et al. (1996) Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science (80-)* 272(5265):1167–1170.
6. Karosiene E, Lundegaard C, Lund O, Nielsen M (2012) NetMHCcons: A consensus method for the major histocompatibility complex class i predictions. *Immunogenetics* 64(3):177–186.
7. Zhang H, Lundegaard C, Nielsen M (2009) Pan-specific MHC class I predictors: A benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* 25(1):83–89.
8. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N (2015) Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160(1–2):48–61.
9. Strønen E, et al. (2016) Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science (80-)* 352(6291):1337–1341.
10. Košmrlj A, et al. (2010) Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* 465(7296):350–354.
11. Martoglio B (1997) Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. *EMBO J* 16(22):6636–6645.
12. Yusim K, et al. (2009) HIV molecular immunology. *Los Alamos, New Mex Los Alamos Natl Lab Theor Biol Biophys*:3–24.
13. Pereyra F, et al. (2014) HIV Control Is Mediated in Part by CD8+ T-Cell Targeting of Specific Epitopes. *J Virol* 88(22):12937–12948.
14. Kiepiela P, et al. (2004) Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 432(7018):769–775.
15. Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MBA (1994) Virus-Specific CD8+ Cytotoxic T-Lymphocyte Activity Associated with Control of Viremia in Primary Human Immunodeficiency. 68(9):6103–6110.
16. Falk K, Rötzschke O, Stevanović S, Jung G, Rammensee HG (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351(6324):290–296.
17. Llano A, Williams A, Olvera A, Silva-Arrieta S, Brander C (2013) Best-Characterized HIV-1 CTL Epitopes: The 2013 Update. *HIV Mol Immunol 2013*:3–25.
18. Apps R, et al. (2016) HIV-1 Vpu Mediates HLA-C Downregulation. *Cell Host Microbe* 19(5):686–695.
19. Thomas R, et al. (2009) HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet* 41(12):1290–1294.
20. Cotton LA, et al. (2014) Genotypic and Functional Impact of HIV-1 Adaptation to Its Host Population during the North American Epidemic. *PLoS Genet* 10(4). doi:10.1371/journal.pgen.1004295.
21. Li G, et al. (2015) An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology* 12(1). doi:10.1186/s12977-015-0148-6.
22. Kawashima Y, et al. (2009) Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458(7238):641–645.
23. Bartha I, et al. (2013) A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* 2013(2):1–16.
24. McLaren PJ, Carrington M (2015) The impact of host genetic variation on infection with HIV-1. *Nat Immunol* 16(6):577–583.

Supplementary data

Supporting information for

HIV Peptidome-Wide Association Study Reveals Patient-Specific Epitope Repertoires Associated with HIV Control

Jatin Arora, Paul J. McLaren, Nimisha Chaturvedi, Mary Carrington, Jacques Fellay & Tobias L. Lenz

Content:	Page
Supporting Methods	2
Supporting Text	6
Supporting Figures S1 – S10	9
Supporting Tables S1 – S9	19
Supporting References	33

Supporting Methods

Samples and Genotype data:

We analyzed data of 6,311 chronically HIV-1 infected patients. The original data and thorough quality check are described in detail in McLaren *et al.* (4). Briefly, genome-wide genotype data had been collected from 8 independent GWAS studies and combined as part of the International Collaboration for the Genomics of HIV. Principal component analysis was used to infer ancestry using HapMap 3 (25) as a reference. Only samples grouping with HapMap Europeans were retained to ensure consistent ancestry across samples. The first five principal components were used as covariates to account for residual population stratification in downstream analysis. SNP genotypes that had not been covered in original genotyping platforms were imputed with Minimac (26) using haplotypes from the 1,000 Genomes Project Phase 1 v3 reference panel. Imputed genotypes from other imputation protocols, Shapeit (27) and Impute2 (28), yielded highly concordant results. Imputed SNPs with low r2 score (< 0.3) or minor allele frequency of < 0.5% were discarded. HLA alleles (best-guess genotypes at 4-digit resolution) for classical class I loci (HLA-A, -B, -C) were imputed from genome-wide genotype data using SNP2HLA and a reference panel of 5,225 individuals of European ancestry (29). Overall, 69 and 37 classical alleles for HLA-B and HLA-A, respectively, were represented in the data (**Fig. S10)**. Available measurement of pre-treatment set point viral load (spVL; log10 HIV-1 RNA copies/μl of plasma) was used as quantitative disease phenotype for all patients (4). All participants were HIV-1–infected adults, and written informed consent for genetic testing was obtained from all individuals as part of the original study in which they were enrolled (4).

HLA binding affinity for HIV-1 epitopes:

We used the NCBI accession NC_001802.1 as the reference sequence for the HIV-1 proteome (M group subtype B). It comprised 10 proteins with sequence length ranging from 82 to 1435 amino acids (Table S9). For the consensus sequence-based analysis, the most recent Consensus HIV-1 proteome (subtype B, year 2004, ID 104CP2) was taken from Los Alamos HIV sequence database. The *Gag-Pol* protein is a precursor protein that results from a −1 ribosomal frameshifting event in upstream *Gag* (30). It is cleaved by virus-encoded protease to produce the mature *Pol* protein. In our analysis, we manually trimmed the *Gag-Pol* protein sequence to *Pol* in order to avoid redundancy with the separate *Gag* protein. HLA class-I molecules preferentially bind and present

9mer peptides (16, 31). NetMHCcons-1.1 is a computational method that predicts binding affinity of 9mer peptides to any known HLA class I molecule (6). It is a consensus approach which integrates three prediction methods, namely NetMHC 3.4, NetMHCpan 2.8 and PickPocket 1.1. The analyzed dataset includes some HLA alleles with no available training data for the prediction methods, which makes the consensus approach more favorable over either of the underlying methods (6). It also reports the rank of predicted binding affinity of HLA-peptide complexes against predicted affinity of 200,000 random natural peptides. Using this tool, we predicted HLA allele-specific binding affinities for all 9mer peptides generated from the entire HIV-1 proteome. HLA-peptide complexes with predicted binding affinity rank less than 0.5 were retained (corresponding to the default threshold for 'strongly bound' peptides (6)). The breadth of peptides bound by a patient's HLA allele pair was taken as the total number of unique peptides predicted to be bound by both alleles but accounting for multiple occurrences of peptides in the HIV-1 proteome. The specificity of the correlation between HLA-B allele-specific effect on spVL and HLA-B-bound HIV-1 peptides was confirmed by running the same correlation on predicted HLA-B-bound peptides from four randomly selected viruses, namely Dengue (NC_002640.1), Rhinovirus-A (NC_001617.1), Hepatitis-B (NC_003977.2) and Rubella (NC_001545.2). Information about known HIV-1 epitopes was obtained from LANL HIV database [accessed 01/28/16].

Association with viral load:

The association of an allele or a peptide with viral load (spVL) was calculated using a linear regression model corrected for population covariates following McLaren *et al.* (4). Covariates included the first five principle components of SNP variation and the cohort identity (all adopted from McLaren *et al.* (4)). Variation in viral load attributable to a given variable (allele or epitope) was calculated as the difference between adjusted-R2 values of the model with variable and covariates (Equation 1) and the model with covariates only (Equation 2), following McLaren et al. (4). The variable's regression coefficient (β1) was used as the measure of its effect on viral load. ε, residual term, is the difference between predicted and true values of spVL given the value of variable and covariates.

$$spVL = \beta 1 * variable + \beta 2 * covariates + \varepsilon 1 \quad \text{[Equation 1]}$$

$$spVL = \beta 2 * covariates + \varepsilon 2 \quad \text{[Equation 2]}$$

We selected 6 HLA-B alleles, highlighted by McLaren & Carrington (24), for representation in epitope-specific association plots (**Fig. 3B, C, S8 and S9)**. Three of them, namely B*57:01, B*58:01, and B*27:05, are associated with lower spVL, while the 3 others, namely B*07:02, B*35:01, and B*39:01 are associated with increased spVL. All analyses were performed in R v3.3.3 and data was visualized using the ggplot2 v2.2.1 package (32).

Clustering of HLA-B bound epitopes:

Position-associated entropy was calculated for all HLA-B-bound disease-associated epitopes (N = 132) using HDMD v1.2 package (33). Subsequently, entropy-weighted pairwise distance between those epitopes was calculated by using PMBEC similarity matrix as defined in Kim *et al.* (34) (Equation 3). Non-metric multidimensional scaling was performed on a pairwise distance matrix using the ecodist v2.0.1 package (35). Density-based clustering was performed using the dbscan v1.1-1 package (36). For each cluster, HLA-B alleles with an overall significant association to viral load (*P*-value < 0.05/69; corrected for multiple testing) are shown (N = 16). Sequence logo plots to represent cluster motifs were made using the ggseqlogo v0.1 package (37).

$$distance(entropy_1, entropy_2) = \frac{1}{9}\sum_{i=1}^{9} distPMBEC(residue_{1,i}, residue_{2,i}) * (1 - entropy_i)$$

[Equation 3]

HLA binding of peptides from autologous HIV-1 sequences:

We analyzed autologous HIV-1 sequences from Bartha et al. (23). Autologous sequences were available for 8 of 10 HIV-1 proteins (only Gp41 segment for *Env*) but only for a small subset of patients in our cohorts (**Table S6)**. Due to the bulk sequencing of viral RNA from pre-treatment stored plasma with limited coverage, there were large segments of missing sites in individual sequences, further reducing the available sequence information for epitope prediction. For comparison of autologous and reference epitopes, for each individual we extracted the corresponding parts of the reference sequence for which autologous sequence data was available. HLA-B binding affinities were calculated for autologous sequences and homologous reference parts using NetMHCcons-1.1 (6). HLA-peptide complexes with binding affinity rank < 0.5 were retained (same criterion as used for the general analysis). The stronger association of autologous epitopes with viral load, compared to reference-based epitopes, further substantiates our epitope prediction pipeline and corroborates that within-host variation of HIV-1 can significantly alter the efficiency of HLA-based immunity.

HLA-B allele-specific escape mutations in HIV-1 epitopes:

Of a total of 24 epitopes from the HIV-1 reference sequence that were predicted to be bound by HLA-B*57:01, autologous sequence information was only available for 16. For each of these 16 epitopes, we created a contingency table with the number of patients who are carriers and non-carriers of the HLA-B*57:01 allele, and the number of patients in whom the autologous version of the given epitope is and is not bound by B*57:01. Fischer exact test was performed on the contingency table to calculate the significance.

Supporting Text

In order to evaluate the significance of the predicted peptide repertoires, we interrogated several layers of empirical evidence. Firstly, we observed that the identified subsets of predicted HLA-B- and HLA-A-bound HIV-1 peptides were strongly enriched for experimentally tested cytotoxic T-lymphocyte (CTL) epitopes from the Los Alamos HIV Molecular Immunology Database (12) (HLA-B: OR = 5.8, $P < 0.0001$; HLA-A: OR = 5.0, $P < 0.0001$; **Fig. S1A**; **Table S2 and S3**). Several predicted HLA-peptide complexes, e.g. B*57:01-IW9$_{15\text{-}23}$, B*58:01-SW9$_{241\text{-}249}$, in *Gag*, B*14:01-EL9$_{584\text{-}592}$ in *Env* and B*57:01-HW9$_{116\text{-}124}$, B*35:01–YY9$_{135\text{-}143}$ in *Nef* were previously shown to be either frequently recognized by $CD8^+$ T-cells (38) or associated with low viremia and slow disease progression (39, 40). Secondly, we observed a negative correlation between an allele's effect on spVL and the number of HIV-1 peptides that are predicted to be bound by that allele (Kendall's tau = -0.26, $P = 0.002$; **Fig. S1B**; **Table S4**). The observed correlation remains significant upon excluding the HLA-B*57:01 allele, which was predicted to bind the largest number of peptides (Kendall's tau = -0.24, $P = 0.004$; **Fig. S2**). Interestingly, at first sight, this observation appears to be in conflict with a previous report showing that protective HLA-B alleles bound less human self-peptides (10). In that study it was shown that certain HLA-B alleles associated with HIV control (i.e. with a negative effect on spVL) were predicted to bind a smaller number of self-peptides. The authors explained this observation by invoking the HLA-mediated negative T cell selection in the thymus: HLA-B alleles that present a smaller number of human self-peptides lead to the deletion of less $CD8^+$ T-cells, thus allowing for a broader T cell repertoire in the periphery that in turn allows for a more competent immune response to HIV. The key point here is that they predicted allele-specific binding of self-peptides (i.e. human peptides), while in our analyses we predicted the binding of HIV peptides. When we explored this further, we indeed found an interesting pattern that is consistent with both the previous result by Kosmrlj et al. on self-peptide binding (10) and our result on HIV peptide binding. Specifically, the HLA-B alleles with the strongest association to HIV protection (B*57:01) and HIV risk (B*07:02), show very antagonistic binding properties (**Fig. S3**): B*57:01 binds the largest number of HIV peptides, while binding only an average number of self-peptides (blue circle at top edge of **Fig. S3**), whereas B*07:02 binds an average number of HIV peptides while binding the largest number of self-peptides (red circle at right edge of **Fig. S3**). At present, it is not clear whether these quantitative differences in binding HIV peptides versus human self-peptides are coincidental, possibly owing

to the comparatively smaller set of HIV peptides, or reflect intrinsic differences between HIV and human peptides, e.g. in the amino acid composition of the proteins. However, if these results were a true biological phenomenon, this could suggest that both mechanisms (possibly independently) might contribute to an allele's association with HIV control, where a protective effect results from binding many HIV peptides and/or binding few self-peptides (thus allowing for a broad T cell response). It is outside the scope of this manuscript to explore this interesting pattern further, but a first glimpse at the data suggests that alleles might be affected by the two mechanisms differently. For instance, B*27:05, which is also quite protective (but much less than B*57:01), binds only an average number of HIV peptides but also few self-peptides, suggesting that its protective effect results mainly from an undiminished T cell repertoire. The results for B*35:01 (the fourth alleles specifically analyzed in Kosmrlj et al.) are less conclusive. In our analysis, it binds neither many HIV peptides nor many self-peptides, while it was predicted to bind many self-peptides in Kosmrlj et al. (10). However, computational binding prediction for this allele was less accurate than for the other ones (10), and B*35:01 showed also a substantially lower effect on HIV than B*35:02 in the largest association study to date (4), which in turn is predicted to bind more self-peptides than B*35:01 (**Fig. S3**).

In the context of our analysis, this correlation between an allele's effect on viral load and its HIV peptide presentation indicates that our computationally predicted peptides are enriched for epitopes that contribute to the well-established association between HLA and viral load. No such association was observed when correlating the HIV-specific allele effect with binding to peptides from other viruses, confirming an HIV-1 peptide-specific effect (**Fig. S4**). Thirdly, we used our prediction pipeline to identify HLA allele-specific *escape variants* in autologous HIV-1 sequences (available for a subset of patients, **Table S6**). Escape variants exhibit sequence variation that impedes either HLA binding or TCR binding to a given epitope and thus allow the virus to avoid CTL recognition. Here we focused on variants escaping HLA binding, which are expected to evolve in response to patient-specific HLA restriction and of which several have previously been reported (22, 23, 41). Replicating these findings with our prediction pipeline would confirm that we are capturing disease-relevant variation among HIV-1 peptides. Focusing on the allele HLA-B*57:01, associated with strongest protection against HIV-1 (23, 42, 43), we thus tested whether we could identify HIV-1 peptides whose autologous versions in patients carrying B*57:01 harbored more escape variants specific to this HLA allele than those in patients who did not carry this allele. Our autologous sequence data covered 16 B*57:01-bound peptides, 10 of which showed no significant difference in the proportion of escape variants between B*57:01 carriers and non-carriers (**Fig. S5**). Of the six HIV-1 peptides that did show a statistically significant difference, all showed a higher proportion of escape variants in B*57:01 carriers (Chi-squared test, all $P < 0.01$ after Bonferroni correction; **Fig. S1C**), and four of them had previously been characterized (23). Understanding the intrinsic differences between these peptides with and without escape variation would provide important insights into HLA-mediated control of HIV-1 replication. However, due to the limited amount of autologous sequence data and the resulting small number of escape candidates, we could here only speculate on possible explanations, which might include critical fitness consequences (preventing escape mutants) or lack of T cell response (absence of selection for escape mutants). Nevertheless, this finding confirms that our approach is able to identify previously described B*57:01-specific restriction on HIV-1.

Supporting Figures

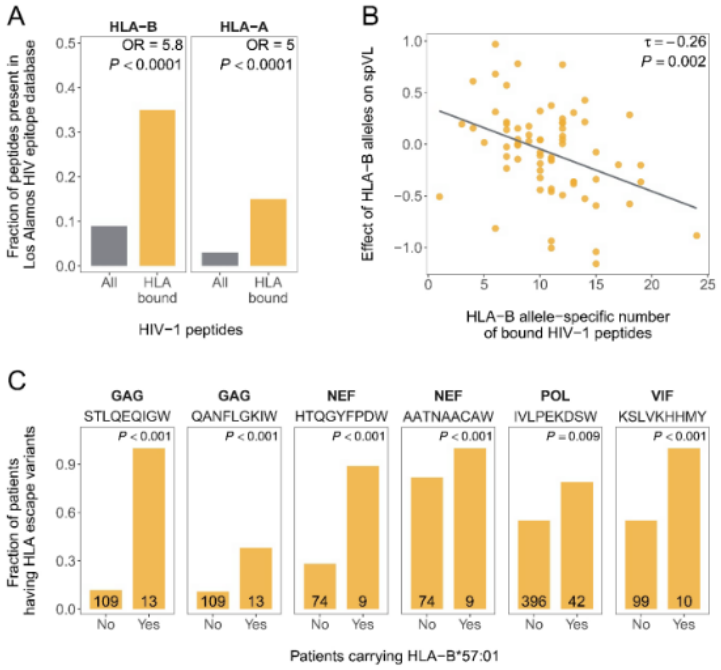


Fig. S1. Characterization of predicted HLA-bound HIV-1 epitopes.

(**A**) In comparison to all possible HIV-1 proteome-wide peptides (N = 3,252), the subsets of peptides predicted to be bound by HLA-B (N = 214) and HLA-A (N = 173), respectively, were strongly enriched for known epitopes defined in the Los Alamos HIV Molecular Immunology Database. Odds ratios (OR) and *P* values from Fisher exact test are shown. (**B**) HLA allele-specific effects on set point viral load (spVL) were calculated using linear regression and correlated with the predicted number of bound HIV-1 peptides. Each dot represents a distinct HLA-B allele (N = 69). Kendall correlation coefficient and p-value are shown. (**C**) For 6 of the 16 HLA-B*57:01-bound reference peptides with autologous data, the proportion of HLA escape variants in autologous sequence data differed significantly between B*57:01 carriers and non-carriers. The total number of patients with available autologous HIV-1 sequence data is given inside each bar. Bonferroni-corrected *P* values from Fisher exact test are shown.

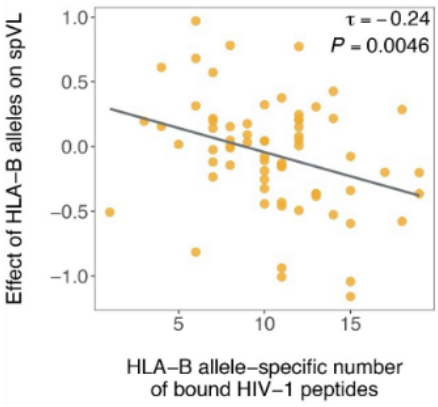


Fig. S2. HIV-1-specific HLA allele effect excluding B*57:01 and bound peptides. HLA allele-specific effects on set point viral load (spVL) were correlated with the predicted number of bound HIV-1 peptides. We have excluded HLA-B*57:01 allele here. Each dot represents a distinct HLA-B allele (N = 68). Kendall correlation coefficient and p-value are shown.

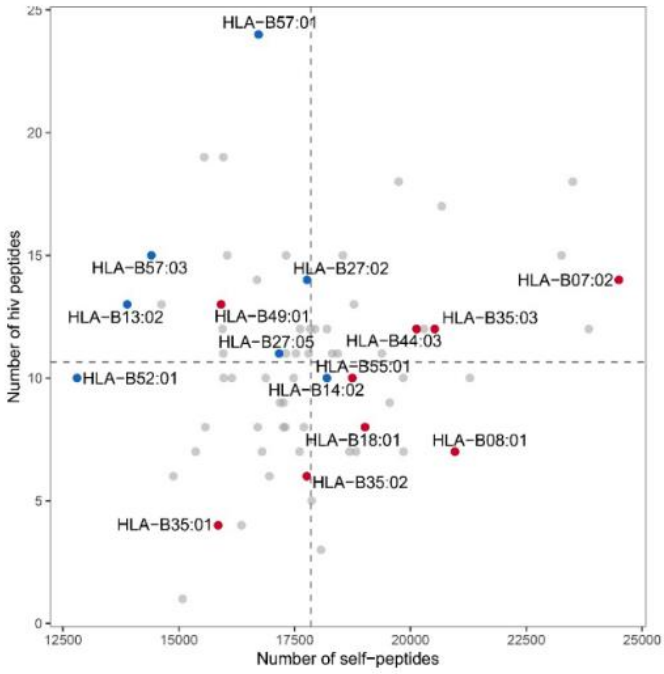


Fig. S3. Number of HIV and human self-peptides bound by HLA-B alleles. Relationship between the number of human self-peptides and the number of HIV-1 peptides predicted to be bound by individual HLA-B alleles (N = 69). Self-peptides were derived from 10,000 randomly selected human proteins from UniProt database (44). HIV-1 peptides were derived from the HIV-1 reference proteome. The common protection- and risk-associated alleles are highlighted in blue and red respectively. The horizontal and vertical dashed lines represent the average number of bound HIV-1 and self-peptides respectively.

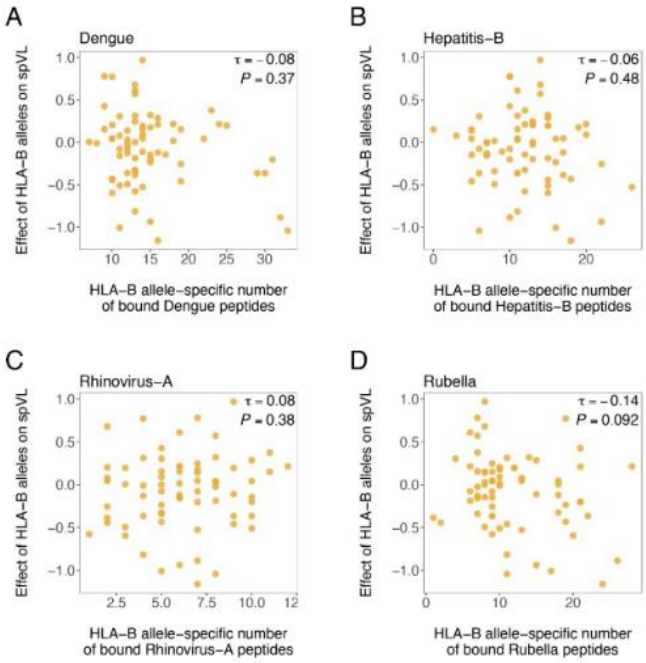


Fig. S4. HIV-1-specific HLA allele effect and epitopes derived from random viruses. No significant correlation was observed between HIV-1-specific effect of HLA-B alleles on viral load (spVL) and the number of HLA-B bound epitopes derived from four random viruses; namely Dengue (**A**), Hepatitis-B (**B**), Rhinovirus-A (**C**) and Rubella (**D**). Each dot represents a distinct HLA-B allele (N = 69).

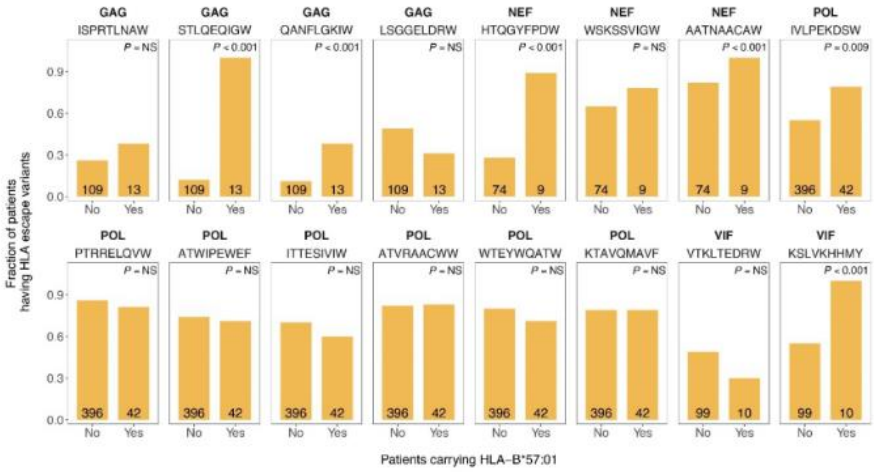


Fig. S5. Escape variants in HLA-B*57:01-bound peptides. The proportion of HIV-1 escape variants in autologous sequence data for 16 of 24 reference peptides bound by B*57:01 allele is shown. For the other 8 reference peptides, no autologous virus sequence was available. The total number of patients with available autologous HIV-1 sequence data is given inside each bar. Bonferroni-corrected *P* values from Fisher exact test are shown.

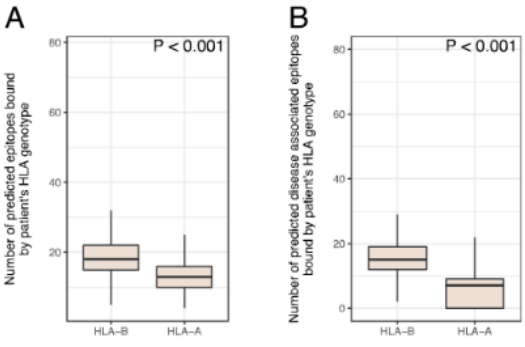


Fig. S6. Breadth of HIV-1 peptides bound by HLA-B and HLA-A genotypes. The breadth of all (**A**) and disease-associated (**B**) HLA bound HIV-1 peptides is shown. A patient's two HLA-B variants together bind a broader array of peptides than the HLA-A variants. Outliers are not shown for better visual comparison. *P* value from paired Wilcoxon rank sum test is shown.

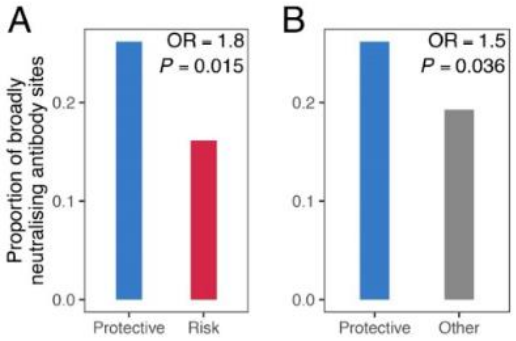


Fig. S7. Proportion of broadly neutralizing antibody sites in predicted Env epitopes. The proportion of sites associated with broadly neutralizing HIV antibodies (bNAbs) in predicted *Env*-derived epitopes is compared between protective epitopes (N = 26) and risk epitopes (N = 15) (**A**) as well as between protective and all non-protective epitopes (other, also includes risk epitopes, N = 34) (**B**). Odds ratio (OR) and *P* value from two-sided Fisher exact test are shown. bNAbs data was downloaded from Neutralizing Antibody Context track from genome browser of Los Alamos HIV Database [accessed 10/22/18].

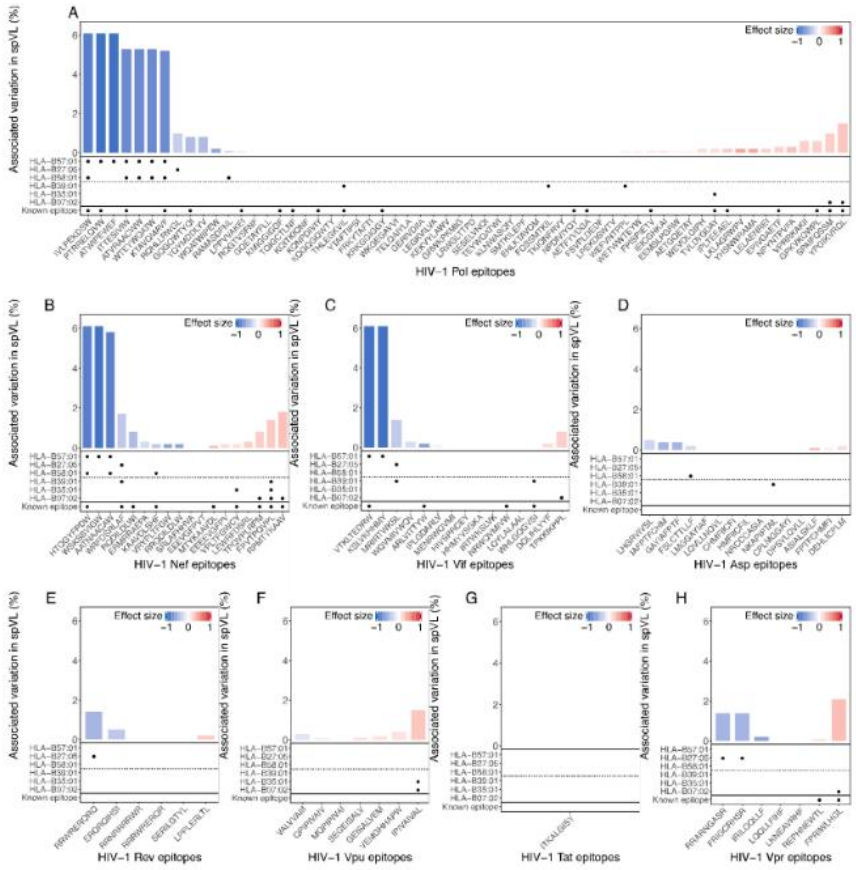


Fig. S8. Epitope-specific association with viral load. Epitope-specific effect on set point viral load (spVL) was calculated using linear regression. X-axis shows the HLA-B bound epitopes derived from Pol (**A**), Nef (**B**), Vif (**C**), Asp (**D**), Rev (**E**), Vpu (**F**), Tat (**G**) and Vpr (**H**) HIV-1 proteins. The height of an epitope's bar represents the fraction of associated variation in spVL, and the color reflects its effect on spVL ranging from protective (blue) to risk (red). For each epitope is also shown if it is a known epitope in Los Alamos HIV database and predicted binding information for 3 protective and 3 risk HLA-B alleles highlighted in a recent review (24).

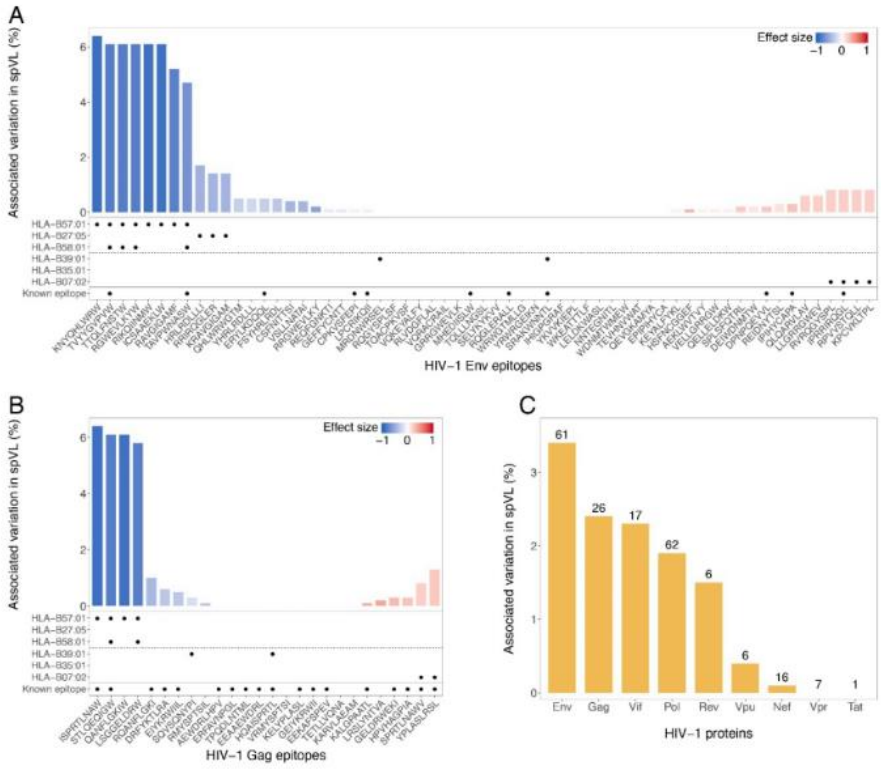


Fig. S9. Epitope- and protein-specific association with viral load using consensus HIV-1 sequence. (**A-B**) Height of the bars represents the percentage of variation in spVL associated with each epitope, while the color reflects each epitope's effect on spVL, ranging from protective (blue) to risk (red). *Env* (**A**) and *Gag* (**B**) proteins are shown as representative examples, together with information on predicted binding for 3 protective and 3 risk HLA-B alleles highlighted in a recent review and whether peptides are known epitopes in Los Alamos HIV database. (**C**) Variation in spVL accounted for by all predicted epitopes of a given HIV-1 protein in consensus sequence. Consensus HIV-1 proteome did not contain Asp protein. Absolute number of predicted HLA-B bound epitopes per protein is shown above the bars.

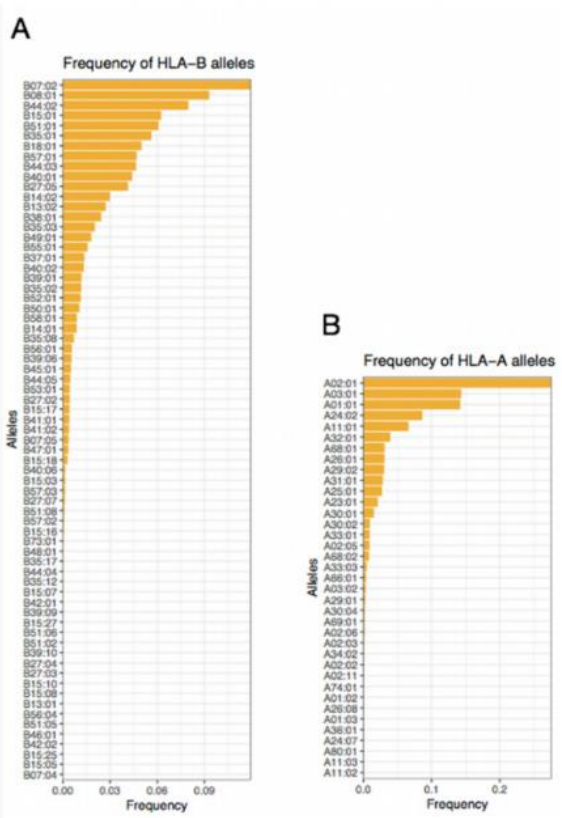


Fig. S10. Frequency of HLA alleles in our dataset. Distribution of the frequencies of (**A**) HLA-B alleles (N = 69) and (**B**) HLA-A alleles (N = 37) in our dataset of 6,311 HIV-1 patients of European ancestry. Note the different allele frequency scales on the x-axis.

Supporting Tables

Table S1. Summary of the represented sample cohorts.

Cohort	Number of samples	Origin	Reference
1	1304	Europe / Australia	(45)
2	1034	USA	(46, 47)
3	581	France	(48, 49)
4	729	USA	(46, 47)
5	576	USA	(3)
6	576	USA	(3)
7	503	USA	(3)
8	364	USA	(46, 47)
9	383	Netherlands	(50)
10	261	USA	NA

The dataset includes a total of 6,311 samples. Shown are the different cohorts, their origin, and the corresponding references. For more detailed information see McLaren et al. 2015 (4).

Table S2. Epitope-specific association of HLA-B bound epitopes (N = 214) with viral load.

Protein	Epitope	Associated variation in spVL	Effect on spVL	Binding HLA-B alleles	Present in Los Alamos HIV Database
Nef	RPMTYKAAV	1.8	0.295	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B08:01, HLA-B39:10, HLA-B42:01, HLA-B42:02, HLA-B55:01, HLA-B81:01	X
Nef	FPVTPQVPL	1.4	0.254	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B15:09, HLA-B15:10, HLA-B35:01, HLA-B35:02, HLA-B35:03, HLA-B35:05, HLA-B35:08, HLA-B35:12, HLA-B35:17, HLA-B35:21, HLA-B39:01, HLA-B39:06, HLA-B39:09, HLA-B39:10, HLA-B42:01, HLA-B42:02, HLA-B51:01, HLA-B51:02, HLA-B51:05, HLA-B51:06, HLA-B51:08, HLA-B53:01, HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B56:03, HLA-B56:04, HLA-B78:01, HLA-B81:01	X
Nef	TPQVPLRPM	0.8	0.228	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B42:01, HLA-B42:02, HLA-B56:03, HLA-B81:01	X
Nef	RQDILDLWI	0.8	-0.412	HLA-B13:01, HLA-B13:02, HLA-B39:08	X
Nef	WRFDSRLAF	1.7	-0.332	HLA-B14:01, HLA-B14:02, HLA-B15:03, HLA-B15:10, HLA-B15:18, HLA-B18:02, HLA-B27:02, HLA-B27:03, HLA-B27:04, HLA-B27:05, HLA-B38:01, HLA-B38:02, HLA-B39:01, HLA-B73:01	X
Nef	ERMRRAEPA	0.3	-0.22	HLA-B14:02, HLA-B39:06, HLA-B73:01	
Nef	MTYKAAVDL	0.1	0.392	HLA-B15:17	X
Nef	KAAVDLSHF	0.2	-0.33	HLA-B15:24, HLA-B57:02, HLA-B57:03, HLA-B58:01	X
Nef	EEEEVGFPV	0.2	0.159	HLA-B18:01, HLA-B45:01	
Nef	VRYPLTFGW	0.2	-0.587	HLA-B27:02	
Nef	RRQDILDLW	0.2	-0.587	HLA-B27:02	
Nef	SRLAFHHVA	0	0.201	HLA-B27:03, HLA-B27:07, HLA-B39:06, HLA-B73:01	
Nef	YPLTFGWCY	0.2	0.15	HLA-B35:01, HLA-B35:05, HLA-B35:08, HLA-B35:17, HLA-B35:21, HLA-B53:01	X
Nef	LEWRFDSRL	0.3	0.236	HLA-B37:01, HLA-B47:01, HLA-B49:01	
Nef	EEEVGFPVT	0	0.079	HLA-B45:01	
Nef	HTQGYFPDW	6.1	-0.824	HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01	X
Nef	WSKSSVIGW	6.1	-0.897	HLA-B57:01, HLA-B57:02	
Nef	AATNAACAW	5.8	-0.811	HLA-B57:01, HLA-B57:02, HLA-B58:01	
Env	RVRQGYSPL	0.8	0.227	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B15:30, HLA-B42:02	
Env	IPRRIRQGL	0.8	0.228	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B42:01, HLA-B42:02, HLA-B81:01	X
Env	RPVVSTQLL	0.8	0.228	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B42:01, HLA-B81:01	
Env	KPCVKLTPL	0.8	0.228	HLA-B07:02, HLA-B07:04, HLA-B42:01, HLA-B42:02, HLA-B81:01	
Env	QLQARILAV	0.6	0.225	HLA-B08:01	
Env	RVKEKYQHL	0.6	0.225	HLA-B08:01	
Env	KEYAFFYKL	0.1	0.078	HLA-B13:01, HLA-B13:02, HLA-B18:01, HLA-B27:02, HLA-B37:01, HLA-B40:01, HLA-B40:02, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B40:10, HLA-B40:23, HLA-B41:01, HLA-B41:02, HLA-B44:03, HLA-B44:05, HLA-B45:01, HLA-B47:01, HLA-B48:01, HLA-B49:01, HLA-B50:01, HLA-B50:02	
Env	YQHLWRWGW	0.2	-0.399	HLA-B13:01, HLA-B15:13, HLA-B18:02, HLA-B27:02, HLA-B47:01	
Env	REINNYTSL	0.1	0.094	HLA-B13:01, HLA-B37:01, HLA-B39:02, HLA-B40:01, HLA-B40:02, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B40:10, HLA-B40:23, HLA-B41:01, HLA-B41:02, HLA-B44:05, HLA-B47:01, HLA-B48:01, HLA-B49:01, HLA-B50:01	
Env	VQLNTSVEI	0.8	-0.412	HLA-B13:02, HLA-B39:02	
Env	SRAKWNNTL	0	0.064	HLA-B14:01, HLA-B27:07, HLA-B39:01, HLA-B39:06, HLA-B39:09	X
Env	YHRLRDLLL	0.5	-0.238	HLA-B14:01, HLA-B14:02, HLA-B15:10, HLA-B38:01	
Env	MRDNWRSEL	0	-0.016	HLA-B14:01, HLA-B38:01, HLA-B39:01, HLA-B39:06, HLA-B39:09	
Env	ERYLKDQQL	0.5	-0.293	HLA-B14:01, HLA-B14:02	X
Env	FSYHRLRDL	0.5	-0.335	HLA-B14:02	
Env	IQRGPGRAF	0	-0.043	HLA-B15:01, HLA-B15:02, HLA-B15:03, HLA-B15:06, HLA-B15:07, HLA-B15:24, HLA-B15:25, HLA-B15:27	
Env	RQGYSPLSF	0	-0.042	HLA-B15:01, HLA-B15:03, HLA-B15:05, HLA-B15:06, HLA-B15:25	
Env	TQACPKVSF	0	-0.427	HLA-B15:03	
Env	YKVVKIEPL	0	0.262	HLA-B15:10, HLA-B39:02, HLA-B39:09	
Env	TVYYGVPVW	6.1	-0.823	HLA-B15:13, HLA-B15:16, HLA-B57:01, HLA-B57:03, HLA-B58:01	X
Env	TAVPWNASW	4.7	-0.684	HLA-B15:13, HLA-B15:16, HLA-B15:17, HLA-B53:01, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	X
Env	HSFNCGGEF	0.1	0.398	HLA-B15:13, HLA-B15:15, HLA-B15:17, HLA-B15:21, HLA-B15:24, HLA-B46:01	
Env	STQLFNSTW	6.1	-0.82	HLA-B15:16, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	
Env	CSFNISTSI	0.4	-0.453	HLA-B15:16, HLA-B52:01	
Env	RAVGIGALF	5.3	-0.74	HLA-B15:16, HLA-B15:17, HLA-B15:24, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	
Env	LTSCNTSVI	0	-0.365	HLA-B15:16	
Env	YTSLIHSLI	0	-0.365	HLA-B15:16	
Env	WKEATTTLF	0	0.195	HLA-B15:18	
Env	TEKLWVTVY	0.2	0.113	HLA-B18:01, HLA-B18:02, HLA-B44:02, HLA-B44:03, HLA-B44:04	
Env	HRLRDLLLI	1.7	-0.477	HLA-B27:02, HLA-B27:03, HLA-B27:05, HLA-B27:07	
Env	RRGWEALKY	0.2	-0.587	HLA-B27:02	
Env	RRIRQGLER	1.4	-0.464	HLA-B27:03, HLA-B27:04, HLA-B27:05	
Env	WRWGWRWGT	1.4	-0.466	HLA-B27:03, HLA-B27:04, HLA-B27:05, HLA-B73:01	
Env	GRAFVTIGK	1.4	-0.461	HLA-B27:03, HLA-B27:05	X
Env	KRAVGIGAL	1.4	-0.459	HLA-B27:03, HLA-B27:04, HLA-B27:05, HLA-B27:07	
Env	DPNPQEVVL	0.2	0.252	HLA-B35:03	X
Env	SPLSFQTHL	0.2	0.253	HLA-B35:03, HLA-B42:01	
Env	LELDKWASL	0	0.02	HLA-B37:01, HLA-B40:02, HLA-B40:09, HLA-B41:02	
Env	MHEDIISLW	0	-0.107	HLA-B38:01, HLA-B38:02	X
Env	TQLLLNGSL	0	-0.342	HLA-B39:02, HLA-B39:08, HLA-B39:09, HLA-B48:01	
Env	GEFFYCNST	0	-0.093	HLA-B40:02, HLA-B40:06, HLA-B41:02, HLA-B50:01, HLA-B50:02	
Env	SELYKYKVV	0	-0.077	HLA-B41:02	
Env	CPKVSFEPI	0.1	-0.086	HLA-B42:01, HLA-B42:02, HLA-B51:01, HLA-B51:02, HLA-B51:05, HLA-B51:08, HLA-B78:01	X

Env	IPIHYCAPA	0.3	0.281	HLA-B42:01, HLA-B42:02, HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B56:04, HLA-B78:01	X
Env	VELLGRRGW	0.1	0.08	HLA-B44:02, HLA-B44:03, HLA-B44:04	
Env	QELLELDKW	0.1	0.08	HLA-B44:02, HLA-B44:03, HLA-B44:04	
Env	TEVHNVWAT	0	0.079	HLA-B45:01	
Env	RLVNGSLAL	0	-0.625	HLA-B48:01	
Env	RQGLERILL	0	-0.625	HLA-B48:01	
Env	LPCRIKQII	0.1	-0.092	HLA-B51:01, HLA-B51:08	X
Env	EPIPIHYCA	0	0.156	HLA-B51:08, HLA-B54:01, HLA-B56:01, HLA-B78:01	
Env	YAFFYKLDI	0.4	-0.457	HLA-B52:01	
Env	ASLWNWFNI	0.4	-0.457	HLA-B52:01	
Env	VSLLNATAI	0.4	-0.457	HLA-B52:01	
Env	RAFVTIGKI	0.4	-0.457	HLA-B52:01	
Env	RGWEALKYW	6.1	-0.824	HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	
Env	HLWRWGWRW	5.8	-0.815	HLA-B57:01, HLA-B58:01, HLA-B58:02	
Env	NSTWFNSTW	6.1	-0.824	HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	
Env	RIKQIINMW	6.1	-0.904	HLA-B57:01	
Env	WRWGTMLLG	0	-0.595	HLA-B73:01	
Gag	YPLTSLRSL	1.4	0.274	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B35:02, HLA-B35:03, HLA-B35:05, HLA-B35:12, HLA-B35:17, HLA-B39:10, HLA-B42:01, HLA-B42:02, HLA-B51:06, HLA-B55:01, HLA-B56:03, HLA-B56:04, HLA-B78:01, HLA-B81:01	X
Gag	SPRTLNAWV	0.8	0.228	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B42:01, HLA-B42:02	X
Gag	AEWDRVHPV	0	-0.022	HLA-B13:01, HLA-B13:02, HLA-B37:01, HLA-B39:02, HLA-B40:02, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B41:01, HLA-B41:02, HLA-B44:02, HLA-B44:04, HLA-B44:05, HLA-B45:01, HLA-B47:01, HLA-B49:01, HLA-B50:01, HLA-B50:02	X
Gag	NQVSQNYPI	0.3	-0.177	HLA-B13:01, HLA-B13:02, HLA-B15:10, HLA-B38:01, HLA-B38:02, HLA-B39:01, HLA-B39:02, HLA-B39:06, HLA-B39:08, HLA-B39:09	
Gag	RQANFLGKI	1	-0.437	HLA-B13:02, HLA-B27:02, HLA-B27:07, HLA-B48:01	X
Gag	ERFAVNPGL	0	-0.148	HLA-B14:01, HLA-B73:01	X
Gag	EIYKRWIIL	0.5	-0.335	HLA-B14:02	X
Gag	DRFYKTLRA	0.6	-0.341	HLA-B14:02, HLA-B73:01	X
Gag	RMYSPTSIL	0.1	-0.353	HLA-B15:03, HLA-B15:04, HLA-B15:06, HLA-B15:07, HLA-B15:24, HLA-B15:30, HLA-B27:07, HLA-B39:02, HLA-B40:05, HLA-B48:01	
Gag	HQAISPRTL	0	0.033	HLA-B15:09, HLA-B15:10, HLA-B39:01, HLA-B39:02, HLA-B39:08, HLA-B48:01	X
Gag	KALGPAATL	0.1	0.392	HLA-B15:17	X
Gag	VRMYSPTSI	0	0.201	HLA-B27:04, HLA-B27:07, HLA-B39:06, HLA-B73:01	
Gag	LRSLYNTVA	0.2	0.427	HLA-B39:06	
Gag	TPQDLNTML	0	-0.235	HLA-B39:10, HLA-B81:01	X
Gag	KELYPLTSL	0	0.032	HLA-B40:01, HLA-B40:02, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B40:10, HLA-B40:23, HLA-B41:01, HLA-B41:02, HLA-B47:01, HLA-B48:01	X
Gag	GEIYKRWII	0	0.08	HLA-B40:02, HLA-B40:06, HLA-B40:09, HLA-B41:02, HLA-B49:01	X
Gag	EEKAFSPEV	0	0.079	HLA-B45:01, HLA-B50:02	X
Gag	TETLLVQNA	0	0.079	HLA-B45:01, HLA-B50:02	
Gag	KARVLAEAM	0	0.681	HLA-B46:01	
Gag	GELDRWEKI	0.3	0.324	HLA-B49:01	X
Gag	HPVHAGPIA	0.3	0.282	HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B56:04, HLA-B78:01	X
Gag	STLQEQIGW	6.1	-0.824	HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01	X
Gag	ISPRTLNAW	6.4	-0.908	HLA-B57:01, HLA-B57:02, HLA-B57:03	X
Gag	LSGGELDRW	5.8	-0.811	HLA-B57:01, HLA-B57:02, HLA-B58:01	
Gag	QANFLGKIW	6.1	-0.904	HLA-B57:01	
Pol	SPAIFQSSM	1	0.233	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B35:03, HLA-B35:08, HLA-B42:01, HLA-B42:02, HLA-B56:03, HLA-B56:04, HLA-B81:01	X
Pol	YPGIKVRQL	1.5	0.274	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B08:01, HLA-B42:01, HLA-B42:02, HLA-B81:01	X
Pol	VPRRKAKII	0.6	0.223	HLA-B07:04, HLA-B08:01, HLA-B42:01, HLA-B42:02	
Pol	GPKVKQWPL	0.6	0.224	HLA-B08:01, HLA-B42:02	X
Pol	WQATWIPEW	0.2	-0.572	HLA-B13:01, HLA-B15:05, HLA-B15:13, HLA-B15:24, HLA-B27:02	
Pol	WEFVNTPPL	0.1	0.066	HLA-B13:01, HLA-B13:02, HLA-B15:10, HLA-B18:01, HLA-B18:02, HLA-B37:01, HLA-B38:02, HLA-B39:01, HLA-B39:02, HLA-B39:06, HLA-B39:08, HLA-B39:09, HLA-B40:01, HLA-B40:02, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B40:10, HLA-B40:23, HLA-B41:01, HLA-B41:02, HLA-B44:05, HLA-B47:01, HLA-B48:01, HLA-B49:01, HLA-B50:01, HLA-B73:01	
Pol	RQGTVSFNF	0	-0.34	HLA-B13:01	
Pol	SESELVNQI	0	0.019	HLA-B13:01, HLA-B13:02, HLA-B37:01, HLA-B40:02, HLA-B41:01, HLA-B41:02, HLA-B44:02, HLA-B44:03, HLA-B44:04, HLA-B44:05, HLA-B49:01	
Pol	GQGQWTYQI	0.8	-0.412	HLA-B13:01, HLA-B13:02	X
Pol	WETWWTEYW	0.1	0.079	HLA-B13:01, HLA-B44:02, HLA-B44:03, HLA-B44:04	
Pol	GQETAYFLL	0	-0.542	HLA-B13:01, HLA-B39:08, HLA-B40:10, HLA-B48:01	
Pol	TEYWQATWI	0	0.014	HLA-B13:02, HLA-B44:03, HLA-B44:04, HLA-B44:05, HLA-B49:01	
Pol	YQYMDDLYV	0.8	-0.412	HLA-B13:02	
Pol	KMIGGIGGF	0	-0.033	HLA-B15:01, HLA-B15:05, HLA-B15:06, HLA-B15:07, HLA-B15:24, HLA-B15:25, HLA-B15:27	X
Pol	TQIGCTLNF	0	-0.044	HLA-B15:01, HLA-B15:03, HLA-B15:05, HLA-B15:06, HLA-B15:13, HLA-B15:24, HLA-B15:27	X
Pol	KQITKIQNF	0	-0.043	HLA-B15:01, HLA-B15:03, HLA-B15:06, HLA-B15:07, HLA-B15:24, HLA-B15:25, HLA-B15:27	
Pol	KQNPDIVIY	0	-0.415	HLA-B15:03, HLA-B15:05	X
Pol	KLNWASQIY	0	0.042	HLA-B15:04, HLA-B15:06, HLA-B15:07, HLA-B15:25	
Pol	SMTKILEPF	0	0.117	HLA-B15:05, HLA-B15:06, HLA-B15:27, HLA-B46:01	
Pol	KQGQGQWTY	0	-0.187	HLA-B15:05	
Pol	KTAVQMAVF	5.2	-0.733	HLA-B15:07, HLA-B15:16, HLA-B15:17, HLA-B15:24, HLA-B15:27, HLA-B46:01, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	X
Pol	THLEGKVIL	0	-0.061	HLA-B15:09, HLA-B15:10, HLA-B38:01, HLA-B38:02, HLA-B39:01, HLA-B39:09	X
Pol	EHLKTAVQM	0	0.772	HLA-B15:09, HLA-B15:10, HLA-B38:02	
Pol	FQSSMTKIL	0	0.033	HLA-B15:09, HLA-B15:10, HLA-B15:30, HLA-B39:01, HLA-B39:02, HLA-B39:08, HLA-B39:09, HLA-B48:01	

Pol	ITTESIVIW	5.3	-0.74	HLA-B15:16, HLA-B15:17, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	X
Pol	YTAFTIPSI	0	-0.365	HLA-B15:16	
Pol	ATVRAACWW	5.3	-0.74	HLA-B15:16, HLA-B15:17, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	
Pol	WTEYWQATW	5.3	-0.74	HLA-B15:16, HLA-B15:17, HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01, HLA-B58:02	
Pol	TKIQNFRVY	0	0.195	HLA-B15:18	
Pol	WEVQLGIPH	0.2	0.158	HLA-B18:01, HLA-B18:02	
Pol	RQHLLRWGL	1	-0.292	HLA-B27:02, HLA-B27:04, HLA-B27:05, HLA-B27:07, HLA-B37:01, HLA-B39:02, HLA-B39:08, HLA-B40:02, HLA-B40:05, HLA-B40:09, HLA-B40:10, HLA-B41:02, HLA-B47:01, HLA-B48:01	
Pol	FRKYTAFTI	0	-0.074	HLA-B27:02, HLA-B27:07, HLA-B39:06, HLA-B73:01	
Pol	KRKGGIGGY	0	-0.937	HLA-B27:04	X
Pol	TVLDVGDAY	0.2	0.159	HLA-B35:01	X
Pol	FPISPIETV	0.1	0.097	HLA-B35:02, HLA-B35:03, HLA-B35:12, HLA-B39:10, HLA-B42:01, HLA-B51:01, HLA-B51:02, HLA-B51:05, HLA-B51:06, HLA-B51:08, HLA-B53:01, HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B56:03, HLA-B56:04, HLA-B78:01, HLA-B81:01	X
Pol	EPIVGAETF	0.3	0.213	HLA-B35:02, HLA-B35:03, HLA-B35:08, HLA-B35:12, HLA-B51:06, HLA-B53:01, HLA-B56:03	X
Pol	IPLTEEAEL	0.2	0.252	HLA-B35:03	X
Pol	NPDIVIYQY	0	0.003	HLA-B35:08	X
Pol	WKGEGAVVI	0	-0.107	HLA-B38:01	
Pol	LKLAGRWPV	0.2	0.427	HLA-B39:06	
Pol	YHSNWRAMA	0.2	0.427	HLA-B39:06	
Pol	IEICGHKAI	0.1	0.116	HLA-B40:01, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B40:23, HLA-B41:01, HLA-B41:02, HLA-B44:05, HLA-B49:01, HLA-B50:01	
Pol	TELQAIYLA	0	-0.049	HLA-B40:06, HLA-B45:01, HLA-B50:02	
Pol	GERIVDIIA	0	-0.493	HLA-B40:06	
Pol	LEGKVILVA	0	-0.025	HLA-B40:06, HLA-B45:01, HLA-B50:01, HLA-B50:02	
Pol	KEKVYLAWV	0	-0.077	HLA-B41:02	
Pol	EEMSLPGRW	0.1	0.068	HLA-B44:02, HLA-B44:03, HLA-B44:04, HLA-B44:05	
Pol	AETGQETAY	0.1	0.08	HLA-B44:02, HLA-B44:03, HLA-B44:04	
Pol	LELAENREI	0.2	0.226	HLA-B44:05, HLA-B49:01	
Pol	AETFYVDGA	0	0.079	HLA-B45:01, HLA-B50:02	X
Pol	FSVPLDEDF	0	0.681	HLA-B46:01	
Pol	RAMASDFNL	0.1	-0.235	HLA-B48:01, HLA-B57:02, HLA-B58:01	
Pol	LPPVVAKEI	0.1	-0.088	HLA-B51:01, HLA-B51:02, HLA-B51:05, HLA-B51:08	X
Pol	LPEKDSWTV	0	0.166	HLA-B51:02, HLA-B51:05, HLA-B51:08	
Pol	NPYNTPVFA	0.3	0.282	HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B78:01	
Pol	IVLPEKDSW	6.1	-0.824	HLA-B57:01, HLA-B57:02, HLA-B57:03, HLA-B58:01	X
Pol	PTRRELQVW	6.1	-0.904	HLA-B57:01	X
Pol	ATWIPEWEF	6.1	-0.904	HLA-B57:01, HLA-B58:02	
Pol	GRWKPKMIG	0	-0.595	HLA-B73:01	
Pol	LRWGLTTPD	0	-0.595	HLA-B73:01	
Vpr	FPRIWLHGL	2.1	0.312	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B08:01, HLA-B35:03, HLA-B39:10, HLA-B42:01, HLA-B42:02, HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B56:03, HLA-B56:04, HLA-B78:01, HLA-B81:01	X
Vpr	LQQLLFIHF	0	-0.044	HLA-B15:01, HLA-B15:03, HLA-B15:05, HLA-B15:27	
Vpr	LKNEAVRHF	0	-0.427	HLA-B15:03	
Vpr	IRILQQLLF	0.2	-0.587	HLA-B27:02	
Vpr	RRARNGASR	1.4	-0.464	HLA-B27:03, HLA-B27:04, HLA-B27:05	
Vpr	FRIGCRHSR	1.4	-0.461	HLA-B27:03, HLA-B27:05	
Vpr	REPHNEWTL	0.1	0.143	HLA-B40:01, HLA-B40:10	X
Vpu	IPIVAIVAL	1.5	0.27	HLA-B07:02, HLA-B07:04, HLA-B07:05, HLA-B35:01, HLA-B35:02, HLA-B35:03, HLA-B35:05, HLA-B35:12, HLA-B35:17, HLA-B39:10, HLA-B42:01, HLA-B42:02, HLA-B51:06, HLA-B56:03, HLA-B56:04, HLA-B78:01, HLA-B81:01	
Vpu	VEMGHHAPW	0.4	0.138	HLA-B13:01, HLA-B18:01, HLA-B18:02, HLA-B44:02, HLA-B44:03, HLA-B44:04, HLA-B44:05, HLA-B47:01, HLA-B49:01	
Vpu	SEGEISALV	0.1	0.17	HLA-B37:01, HLA-B45:01	
Vpu	GEISALVEM	0.2	0.124	HLA-B40:01, HLA-B40:02, HLA-B40:05, HLA-B40:09, HLA-B40:10, HLA-B40:23, HLA-B44:03, HLA-B47:01	
Vpu	MQPIPIVAI	0	-0.625	HLA-B48:01	
Vpu	VALVVAIII	0.3	-0.16	HLA-B51:01, HLA-B52:01	
Vpu	QPIPIVAIV	0.1	-0.089	HLA-B51:01, HLA-B51:02, HLA-B51:08	
Vif	TPKKIKPPL	0.8	0.228	HLA-B07:02, HLA-B07:04, HLA-B42:01, HLA-B42:02, HLA-B81:01	
Vif	MENRWQVMI	0	-0.003	HLA-B13:01, HLA-B13:02, HLA-B18:02, HLA-B37:01, HLA-B38:02, HLA-B40:05, HLA-B40:06, HLA-B40:09, HLA-B40:10, HLA-B40:23, HLA-B41:01, HLA-B41:02, HLA-B44:02, HLA-B44:03, HLA-B44:04, HLA-B44:05, HLA-B45:01, HLA-B47:01, HLA-B49:01, HLA-B50:01, HLA-B50:02, HLA-B52:01	
Vif	WQVMIVWQV	0.3	-0.206	HLA-B13:02, HLA-B37:01	
Vif	MRIRTWKSL	1.4	-0.318	HLA-B14:01, HLA-B14:02, HLA-B27:02, HLA-B27:03, HLA-B27:04, HLA-B27:05, HLA-B27:07, HLA-B39:01, HLA-B39:06, HLA-B39:09	
Vif	LQYLALAAL	0	0.123	HLA-B15:03, HLA-B15:04, HLA-B15:09, HLA-B15:27, HLA-B15:30, HLA-B37:01, HLA-B39:02, HLA-B39:08, HLA-B40:05, HLA-B40:10, HLA-B48:01	
Vif	HIVSPRCEY	0	-0.508	HLA-B15:08, HLA-B15:21, HLA-B35:14, HLA-B35:43	
Vif	WHLGQGVSI	0	0.01	HLA-B15:09, HLA-B15:10, HLA-B38:01, HLA-B38:02, HLA-B39:01, HLA-B39:06, HLA-B39:09	X
Vif	HHMYVSGKA	0	-0.595	HLA-B15:09, HLA-B73:01	
Vif	DQLIHLYYF	0.2	0.158	HLA-B18:01, HLA-B18:02	
Vif	ARLVITTYW	0.2	-0.587	HLA-B27:02	X
Vif	IRTWKSLVK	0	-1.007	HLA-B27:03	
Vif	IPLGDARLV	0.1	-0.091	HLA-B51:01	
Vif	VTKLTEDRW	6.1	-0.897	HLA-B57:01, HLA-B57:02	X
Vif	KSLVKHHMY	6.1	-0.904	HLA-B57:01, HLA-B58:02	
Vif	NRWQVMIVW	0	-0.595	HLA-B73:01	X
Rev	ERQRQIHSI	0.5	-0.335	HLA-B14:02	

Rev	RRNRRRRWR	0	-0.937	HLA-B27:04
Rev	RRRWRERQR	0	-0.937	HLA-B27:04
Rev	RRWRERQRQ	1.4	-0.458	HLA-B27:05
Rev	LPPLERLTL	0.2	0.252	HLA-B35:03
Rev	SERILGTYL	0	-0.077	HLA-B40:09, HLA-B41:02
Asp	LHGRVIVSL	0.5	-0.253	HLA-B14:02, HLA-B38:01
Asp	LMGGAYIAF	0	-0.043	HLA-B15:01, HLA-B15:02, HLA-B15:03, HLA-B15:05, HLA-B15:06, HLA-B15:07, HLA-B15:15, HLA-B15:25, HLA-B15:27
Asp	FSLCTTLLF	0.2	-0.239	HLA-B15:03, HLA-B15:05, HLA-B15:13, HLA-B15:16, HLA-B15:17, HLA-B46:01, HLA-B52:01, HLA-B58:01, HLA-B58:02
Asp	IAFPTFCHM	0.4	-0.457	HLA-B15:04, HLA-B15:13, HLA-B52:01
Asp	NKAPIPTAL	0	0.054	HLA-B15:09, HLA-B15:10, HLA-B39:01, HLA-B39:09
Asp	ASIALSKLF	0.1	0.392	HLA-B15:17, HLA-B15:24
Asp	GAYIAFPTF	0.4	-0.457	HLA-B15:24, HLA-B52:01
Asp	LQVLLNQVL	0	-0.625	HLA-B15:30, HLA-B39:02, HLA-B39:08, HLA-B40:05, HLA-B40:10, HLA-B48:01
Asp	DEHLICPLM	0.2	0.158	HLA-B18:01, HLA-B18:02
Asp	FPTFCHMFI	0.1	0.096	HLA-B35:02, HLA-B35:03, HLA-B35:12, HLA-B51:01, HLA-B51:02, HLA-B51:05, HLA-B51:06, HLA-B51:08, HLA-B53:01, HLA-B54:01, HLA-B55:01, HLA-B56:01, HLA-B56:04, HLA-B78:01
Asp	CHMFIICFI	0	-0.107	HLA-B38:01, HLA-B38:02
Asp	HMFIICFIL	0	-0.625	HLA-B48:01
Asp	CPLMGGAYI	0	0.022	HLA-B51:01, HLA-B51:02, HLA-B51:05, HLA-B51:06, HLA-B51:08, HLA-B53:01, HLA-B55:01, HLA-B56:01, HLA-B78:01
Asp	DPSVLQVLL	0	0.047	HLA-B51:06
Asp	NRCCCASIA	0	-0.595	HLA-B73:01
Tat	ITKALGISY	0	-0.144	HLA-B15:04, HLA-B15:07

Each row represents the epitope, the protein it is derived from, variation in spVL associated with it, its effect size on spVL, HLA-B alleles its binds to and its presence in the list of experimentally tested CTL epitopes from Los Alamos HIV Database.

Table S3. Epitope-specific association of HLA-A bound epitopes (N = 173) with viral load.

Protein	Epitope	Associated variation in spVL	Effect on spVL	Binding HLA-A alleles	Present in Los Alamos HIV Database
Vif	LADQLIHLY	0	0.036	HLA-A01:01, HLA-A01:02, HLA-A01:03, HLA-A30:04, HLA-A36:01, HLA-A36:03	
Vif	SLVKHHMYV	0	0.09	HLA-A02:02, HLA-A02:03, HLA-A02:11, HLA-A02:16, HLA-A02:17	X
Vif	WQVMIVWQV	0	-0.134	HLA-A02:05, HLA-A02:06	
Vif	ALAALITPK	0	-0.034	HLA-A03:01, HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05	
Vif	QYLALAALI	0.3	0.147	HLA-A23:01, HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10	
Vif	KIKPPLPSV	0.1	-0.19	HLA-A30:01, HLA-A30:04	
Vif	RGWFYRHHY	0	0.036	HLA-A30:02, HLA-A30:04	
Vif	KSLVKHHMY	0	0.045	HLA-A30:02, HLA-A30:04, HLA-A80:01	
Vif	SGKARGWFY	0	-0.194	HLA-A30:04	
Vif	HIVSPRCEY	0	-0.194	HLA-A30:04	
Vif	LLGHIVSPR	0	-0.337	HLA-A74:01	
Gag	GSEELRSLY	0	0.036	HLA-A01:01, HLA-A01:02, HLA-A01:03, HLA-A30:04, HLA-A36:01, HLA-A36:03	
Gag	HSNQVSQNY	0	0.046	HLA-A01:01, HLA-A01:02, HLA-A30:02, HLA-A30:04, HLA-A36:01	
Gag	VLAEAMSQV	0	0.013	HLA-A02:01, HLA-A02:02, HLA-A02:03, HLA-A02:05, HLA-A02:06, HLA-A02:07, HLA-A02:11, HLA-A02:16	X
Gag	HQAAMQMLK	0	-0.042	HLA-A03:01, HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05, HLA-A30:04, HLA-A34:02	X
Gag	EVIPMFSAL	0.2	-0.142	HLA-A25:01, HLA-A26:01, HLA-A26:08, HLA-A34:01, HLA-A66:01, HLA-A66:02, HLA-A68:02, HLA-A69:01	X
Gag	DIAGTTSTL	0.4	-0.289	HLA-A25:01	
Gag	FLGKIWPSY	0.2	0.209	HLA-A29:01, HLA-A29:02, HLA-A30:04, HLA-A80:01	
Gag	IVKCFNCGK	0.1	-0.201	HLA-A30:01, HLA-A34:02	
Gag	KIRLRPGGK	0.1	-0.187	HLA-A30:01	X
Gag	RLRPGGKKK	0.1	-0.187	HLA-A30:01	X
Gag	LYNTVATLY	0	0.036	HLA-A30:02, HLA-A30:04	
Gag	RMYSPTSIL	0.1	-0.138	HLA-A30:04, HLA-A32:01	X
Gag	RQANFLGKI	0	-0.194	HLA-A30:04	
Gag	RTLNAWVKV	0	-0.194	HLA-A30:04	
Gag	IMMQRGNFR	0.4	-0.241	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:01, HLA-A33:03, HLA-A74:01	
Gag	TARNCRAPR	0.4	-0.239	HLA-A31:01, HLA-A31:12, HLA-A33:01, HLA-A33:03	
Gag	ATLYCVHQR	0.3	-0.272	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A74:01	
Gag	MVHQAISPR	0.2	-0.138	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:01, HLA-A33:03, HLA-A34:01, HLA-A34:02, HLA-A66:01, HLA-A66:02, HLA-A68:01, HLA-A68:03	
Gag	SLYNTVATL	0.1	-0.134	HLA-A32:01	X
Gag	ELYPLTSLR	0	-0.033	HLA-A33:01, HLA-A33:03, HLA-A34:01, HLA-A66:01, HLA-A66:02, HLA-A68:01, HLA-A68:03	
Gag	NSATIMMQR	0	0.021	HLA-A33:03, HLA-A66:02, HLA-A68:01, HLA-A68:03	
Gag	MTNNPPIPV	0	-0.035	HLA-A34:01, HLA-A66:02, HLA-A68:02, HLA-A69:01	
Gag	NTVATLYCV	0	-0.035	HLA-A68:02, HLA-A69:01	
Gag	KIWPSYKGR	0	-0.337	HLA-A74:01	
Rev	ISERILGTY	0	0.043	HLA-A01:01, HLA-A01:02, HLA-A01:03, HLA-A30:02, HLA-A30:04, HLA-A36:01, HLA-A36:03	
Rev	TVRLIKLLY	0.2	0.209	HLA-A29:01, HLA-A29:02, HLA-A30:04, HLA-A80:01	
Rev	LIRTVRLIK	0.1	-0.187	HLA-A30:01	
Rev	RQIHSISER	0.3	-0.271	HLA-A31:01, HLA-A31:03, HLA-A31:12	
Rev	RQARRNRRR	0.3	-0.271	HLA-A31:01	
Rev	RILGTYLGR	0	-0.337	HLA-A31:03, HLA-A74:01	
Env	PIDNDTTSY	0	0.045	HLA-A01:01, HLA-A01:03, HLA-A36:01	
Env	FNITNWLWY	0.2	0.191	HLA-A01:02, HLA-A29:01, HLA-A29:02, HLA-A30:04	
Env	NITNWLWY1	0.1	-0.448	HLA-A02:06, HLA-A69:01	
Env	LLNATAIAV	0	0	HLA-A02:11	
Env	ITNWLWYIK	0.3	-0.152	HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05, HLA-A31:01, HLA-A31:03, HLA-A74:01	
Env	SLCLFSYHR	0.3	-0.237	HLA-A03:02, HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A74:01	
Env	ILAVERYLK	0	0.07	HLA-A03:02	
Env	CLFSYHRLR	0.3	-0.22	HLA-A03:02, HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:01, HLA-A33:03, HLA-A74:01	
Env	SVITQACPK	0.1	-0.113	HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05, HLA-A34:02	
Env	SSGRMIMEK	0.1	-0.108	HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05	
Env	VTENFNMWK	0.1	-0.108	HLA-A11:01, HLA-A11:02, HLA-A11:05	
Env	FYCNSTQLF	0.3	0.141	HLA-A23:01, HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10, HLA-A30:04	
Env	YWWNLLQYW	0.3	0.147	HLA-A23:01, HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10	
Env	KWASLWNWF	0.3	0.141	HLA-A23:01, HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10, HLA-A30:04	
Env	NWLWYIKLF	0.3	0.291	HLA-A23:01	
Env	WYIKLFIMI	0.3	0.147	HLA-A23:01, HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10	
Env	RYLKDQQLL	0.1	0.104	HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10	
Env	ASLWNWFNI	0	-0.296	HLA-A24:07	
Env	VTIGKIGNM	0	-0.027	HLA-A26:01, HLA-A26:08	
Env	EVHNVWATH	0	-0.028	HLA-A26:01	
Env	FNCGGEFFY	0.2	0.206	HLA-A29:01, HLA-A29:02, HLA-A30:04	
Env	KYWWNLLQY	0.2	0.191	HLA-A29:01, HLA-A29:02, HLA-A30:02, HLA-A30:04, HLA-A80:01	
Env	RAKWNNTLK	0.1	-0.187	HLA-A30:01	
Env	RVRQGYSPL	0.1	-0.19	HLA-A30:01, HLA-A30:04	
Env	RVKEKYQHL	0.1	-0.187	HLA-A30:01	

Env	VQKEYAFFY	0	0.045	HLA-A30:02, HLA-A30:04, HLA-A80:01	
Env	IVNRVRQGY	0	0.036	HLA-A30:02, HLA-A30:04	
Env	NWRSELYKY	0	-0.194	HLA-A30:04	
Env	KVQKEYAFF	0	-0.194	HLA-A30:04	
Env	RAVGIGALF	0	-0.194	HLA-A30:04	
Env	SFEPIPIHY	0	-0.194	HLA-A30:04	
Env	HSFNCGGEF	0	-0.194	HLA-A30:04	
Env	RDNWRSELY	0	-0.194	HLA-A30:04	
Env	RIKQIINMW	0	-0.194	HLA-A30:04	X
Env	YTSLIHSLI	0	-0.047	HLA-A30:04, HLA-A68:02	
Env	GTMLLGMLM	0	-0.194	HLA-A30:04	
Env	CSSNITGLL	0	-0.194	HLA-A30:04	
Env	WQKVGKAMY	0	-0.194	HLA-A30:04	
Env	RWGWRWGTM	0	-0.194	HLA-A30:04	
Env	RSLCLFSYH	0	-0.194	HLA-A30:04	
Env	KAKRRVVQR	0.3	-0.271	HLA-A31:01, HLA-A31:03, HLA-A31:12	
Env	RAIRHIPRR	0.3	-0.271	HLA-A31:01, HLA-A31:03, HLA-A31:12	
Env	RQAHCNISR	0.3	-0.272	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A74:01	
Env	HLWRWGWRW	0.1	-0.134	HLA-A32:01	
Env	RIVFAVLSI	0.1	-0.134	HLA-A32:01	
Env	STQLFNSTW	0.1	-0.134	HLA-A32:01	
Env	KVSFEPIPI	0.1	-0.134	HLA-A32:01	
Env	HTTWMEWDR	0	-0.036	HLA-A33:01, HLA-A33:03, HLA-A68:01, HLA-A68:03	
Env	ELYKYKVVK	0	-0.383	HLA-A34:02	
Env	NTLKQIASK	0	-0.383	HLA-A34:02	
Env	TSVEINCTR	0	0.026	HLA-A68:01, HLA-A68:03	
Env	FAVLSIVNR	0	0.026	HLA-A68:01, HLA-A68:03	
Env	NVWATHACV	0	-0.149	HLA-A69:01	
Asp	YLYNSLLQL	0	0.018	HLA-A02:01, HLA-A02:02, HLA-A02:03, HLA-A02:07, HLA-A02:11, HLA-A02:16, HLA-A02:17	
Asp	RVIVSLPSV	0.1	-0.453	HLA-A02:06, HLA-A30:04	
Asp	LISPPPGLK	0	0.007	HLA-A03:01, HLA-A11:03, HLA-A34:02	
Asp	SIFTTLYLY	0	0.011	HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:05, HLA-A29:01, HLA-A29:02, HLA-A30:02, HLA-A30:04, HLA-A34:02, HLA-A36:01, HLA-A36:03, HLA-A80:01	
Asp	IICFILHGR	0.3	-0.22	HLA-A03:02, HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:01, HLA-A33:03, HLA-A68:03, HLA-A74:01	
Asp	AFPTFCHMF	0.1	0.104	HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10	
Asp	NGSIFTTLY	0	-0.194	HLA-A30:04	
Asp	IVSLPSVLF	0	-0.194	HLA-A30:04	
Asp	FSLCTTLLF	0	-0.194	HLA-A30:04	
Asp	KAPIPTALF	0	-0.194	HLA-A30:04	
Asp	GSIFTTLYL	0	-0.194	HLA-A30:04	
Asp	ASIALSKLF	0	-0.194	HLA-A30:04	
Asp	EAAPIVLPA	0	-0.01	HLA-A68:02	
Asp	CTTLLFALV	0	-0.01	HLA-A68:02	
Asp	SVIEAAPIV	0	-0.149	HLA-A69:01	
Pol	FLREDLAFL	0	0.013	HLA-A02:01, HLA-A02:02, HLA-A02:03, HLA-A02:05, HLA-A02:06, HLA-A02:07, HLA-A02:11, HLA-A02:16, HLA-A02:17	
Pol	YQYMDDLYV	0	0.01	HLA-A02:01, HLA-A02:06, HLA-A02:16	
Pol	YLALQDSGL	0	0.117	HLA-A02:02	
Pol	ILKEPVHGV	0	0.108	HLA-A02:03	X
Pol	IVGAETFYV	0	-0.007	HLA-A02:05	
Pol	YQLEKEPIV	0.2	-0.755	HLA-A02:06	
Pol	AIFQSSMTK	0	-0.037	HLA-A03:01, HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05, HLA-A34:02	X
Pol	KLAGRWPVK	0	0.013	HLA-A03:01	
Pol	KLVSAGIRK	0	0.013	HLA-A03:01	
Pol	YLAWVPAHK	0	0.007	HLA-A03:01, HLA-A03:02, HLA-A34:02	
Pol	MAVFIHNFK	0	-0.062	HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:05, HLA-A30:04, HLA-A33:03, HLA-A34:01, HLA-A34:02, HLA-A66:02, HLA-A68:01, HLA-A68:03, HLA-A74:01	
Pol	AVFIHNFKR	0.2	-0.112	HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:05, HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:03, HLA-A34:02, HLA-A68:01, HLA-A68:03, HLA-A74:01	
Pol	IQNFRVYYR	0.3	-0.22	HLA-A03:02, HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:01, HLA-A33:03, HLA-A74:01	
Pol	IVIWGKTPK	0.1	-0.112	HLA-A11:01, HLA-A34:02	
Pol	TYQIYQEPF	0.3	0.147	HLA-A23:01, HLA-A24:02, HLA-A24:03, HLA-A24:07, HLA-A24:10	
Pol	LWQRPLVTI	0.1	0.104	HLA-A24:02, HLA-A24:07	
Pol	WWAGIKQEF	0	-0.194	HLA-A24:03, HLA-A30:04	
Pol	ETAYFLLKL	0.2	-0.142	HLA-A25:01, HLA-A26:01, HLA-A26:08, HLA-A34:01, HLA-A66:01, HLA-A66:02, HLA-A68:02, HLA-A69:01	
Pol	ETPGIRYQY	0.2	-0.152	HLA-A25:01, HLA-A26:01, HLA-A26:08, HLA-A30:04, HLA-A34:01, HLA-A66:01	
Pol	ETKLGKAGY	0.2	-0.162	HLA-A25:01, HLA-A26:01, HLA-A26:08	
Pol	YTAFTIPSI	0.3	-0.244	HLA-A25:01, HLA-A30:04, HLA-A66:01, HLA-A69:01	
Pol	ETGQETAYF	0.2	-0.163	HLA-A25:01, HLA-A26:01	
Pol	KVYLAWVPA	0.1	-0.19	HLA-A30:01, HLA-A30:04	
Pol	KLNWASQIY	0	0.045	HLA-A30:02, HLA-A30:04, HLA-A80:01	
Pol	KIQNFRVYY	0	0.045	HLA-A30:02, HLA-A30:04, HLA-A80:01	X
Pol	KQNPDIVIY	0	0.036	HLA-A30:02, HLA-A30:04	
Pol	KQGQGQWTY	0	0.045	HLA-A30:02, HLA-A30:04, HLA-A80:01	
Pol	IYQYMDDLY	0	-0.194	HLA-A30:04	
Pol	KTAVQMAVF	0.1	-0.138	HLA-A30:04, HLA-A32:01	
Pol	QMAVFIHNF	0	-0.194	HLA-A30:04	
Pol	SMTKILEPF	0	-0.194	HLA-A30:04	

Pol	TWETWWTEY	0	-0.194	HLA-A30:04	
Pol	RQHLLRWGL	0	-0.194	HLA-A30:04	
Pol	KMIGGIGGF	0.1	-0.138	HLA-A30:04, HLA-A32:01	
Pol	RQGTVSFNF	0	-0.194	HLA-A30:04	
Pol	MTKILEPFR	0.2	-0.139	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:01, HLA-A33:03, HLA-A68:01, HLA-A68:03	
Pol	YFLLKLAGR	0.1	-0.241	HLA-A33:01	
Pol	QITKIQNFR	0	0.021	HLA-A33:03, HLA-A68:01, HLA-A68:03	
Pol	ETFYVDGAA	0	-0.012	HLA-A34:01, HLA-A66:01, HLA-A66:02, HLA-A68:02	
Pol	EVIPLTEEA	0	-0.012	HLA-A34:01, HLA-A66:01, HLA-A66:02, HLA-A68:02	
Pol	DVGDAYFSV	0	-0.149	HLA-A69:01	
Pol	TVSFNFPQV	0	-0.149	HLA-A69:01	
Pol	TVLVGPTPV	0	-0.149	HLA-A69:01	
Vpr	WTLELLEEL	0.2	-0.755	HLA-A02:06	
Vpr	RQRRARNGA	0.1	-0.187	HLA-A30:01	
Vpr	QQLLFIHFR	0.3	-0.272	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A74:01	
Vpr	QLLFIHFRI	0.1	-0.134	HLA-A32:01	
Vpr	DTWAGVEAI	0	-0.149	HLA-A69:01	
Nef	RLAFHHVAR	0.2	-0.208	HLA-A03:02, HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A33:03, HLA-A74:01	
Nef	AVDLSHFLK	0.1	-0.097	HLA-A03:02, HLA-A11:01, HLA-A11:02, HLA-A11:03, HLA-A11:05	X
Nef	RMRRAEPAA	0.1	-0.187	HLA-A30:01	
Nef	LWIYHTQGY	0	-0.194	HLA-A30:04	
Nef	GYFPDWQNY	0	-0.194	HLA-A30:04	X
Nef	WIYHTQGYF	0	-0.194	HLA-A30:04	
Nef	VARELHPEY	0	-0.194	HLA-A30:04	
Nef	YTPGPGVRY	0	-0.194	HLA-A30:04	
Nef	LTFGWCYKL	0.1	-0.134	HLA-A32:01	
Nef	SVIGWPTVR	0	0.009	HLA-A33:03, HLA-A68:03	
Tat	ITKALGISY	0	0.036	HLA-A30:02, HLA-A30:04	
Tat	KALGISYGR	0.3	-0.272	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A74:01	
Vpu	VWSIVIIEY	0	-0.194	HLA-A30:04	
Vpu	RLIDRLIER	0.3	-0.272	HLA-A31:01, HLA-A31:03, HLA-A31:12, HLA-A74:01	
Vpu	EYRKILRQR	0.1	-0.241	HLA-A33:01	
Vpu	WSIVIIEYR	0	0.026	HLA-A68:01, HLA-A68:03	

Each row represents the epitope, the protein it is derived from, variation in spVL associated with it, its effect size on spVL, HLA-A alleles its binds to and its presence in the list of experimentally tested CTL epitopes from Los Alamos HIV Database.

Table S4. HLA-B allele-specific association with viral load.

Allele	β estimate of association with spVL	Total number of bound epitopes	Number of protective epitopes	Number of risk epitopes
HLA-B07:02	0.217	14	0	14
HLA-B07:04	-1.159	15	0	15
HLA-B07:05	0.211	12	0	12
HLA-B08:01	0.216	7	0	7
HLA-B13:01	-0.34	15	5	5
HLA-B13:02	-0.385	13	7	2
HLA-B14:01	-0.119	7	4	0
HLA-B14:02	-0.325	10	10	0
HLA-B15:01	-0.025	7	0	0
HLA-B15:03	-0.363	13	3	0
HLA-B15:05	-0.187	10	2	0
HLA-B15:07	-0.144	8	2	0
HLA-B15:08	-0.508	1	0	0
HLA-B15:10	0.772	12	3	2
HLA-B15:16	-0.365	13	10	0
HLA-B15:17	0.375	11	7	4
HLA-B15:18	0.195	3	1	0
HLA-B15:25	0.97	6	0	0
HLA-B15:27	0.03	9	1	0
HLA-B18:01	0.154	8	0	8
HLA-B27:02	-0.527	14	12	1
HLA-B27:03	-1.007	11	9	0
HLA-B27:04	-0.937	11	7	0
HLA-B27:05	-0.431	11	11	0
HLA-B27:07	-0.254	10	6	0
HLA-B35:01	0.156	4	0	4
HLA-B35:02	0.314	6	0	6
HLA-B35:03	0.246	12	0	12
HLA-B35:08	0.017	5	0	4
HLA-B35:12	-0.815	6	0	6
HLA-B35:17	0.611	4	0	4
HLA-B37:01	0.204	12	2	5
HLA-B38:01	-0.109	10	4	0
HLA-B39:01	0.045	12	3	2
HLA-B39:06	0.427	14	3	5
HLA-B39:09	0.006	12	2	2
HLA-B39:10	-0.235	7	0	6
HLA-B40:01	0.141	7	0	6
HLA-B40:02	-0.129	11	1	4
HLA-B40:06	-0.493	12	0	4
HLA-B41:01	0.001	8	0	4
HLA-B41:02	-0.077	15	1	4
HLA-B42:01	0.286	18	1	17
HLA-B42:02	-0.199	17	1	16
HLA-B44:02	0.043	10	0	7
HLA-B44:03	0.152	12	0	9
HLA-B44:04	-0.46	11	0	7
HLA-B44:05	-0.147	11	0	7
HLA-B45:01	0.079	12	0	3
HLA-B46:01	0.681	6	2	1
HLA-B47:01	-0.157	11	2	6
HLA-B48:01	-0.578	18	4	3
HLA-B49:01	0.306	13	0	8
HLA-B50:01	-0.01	8	0	4
HLA-B51:01	-0.078	10	6	3
HLA-B51:02	0.781	8	3	3
HLA-B51:05	0.572	7	2	3
HLA-B51:06	0.047	8	0	6
HLA-B51:08	0.042	10	4	3
HLA-B52:01	-0.442	10	9	0
HLA-B53:01	0.198	7	1	5
HLA-B55:01	0.322	10	0	9
HLA-B56:01	0.176	9	0	7
HLA-B56:04	0.09	9	0	9
HLA-B57:01	-0.886	24	24	0
HLA-B57:02	-0.365	19	19	0
HLA-B57:03	-1.042	15	15	0
HLA-B58:01	-0.202	19	19	0
HLA-B73:01	-0.595	15	4	1

Association of each HLA-B allele was calculated using linear regression model. Each row represents a HLA-B allele represented in the dataset (N = 69). The allele's regression coefficient, β estimate, represents its effect on viral load. The total number of predicted HIV-1 epitopes bound by HLA-B alleles (column bound epitopes), and of which how many were protective and risk epitopes are shown.

Table S5. HLA-A allele-specific association with viral load.

Allele	β estimate of association with spVL	Total number of bound epitopes	Number of protective epitopes	Number of risk epitopes
HLA-A01:01	0.055	5	0	0
HLA-A01:02	-0.838	5	0	1
HLA-A01:03	-1.265	4	0	0
HLA-A02:01	0.016	4	0	0
HLA-A02:02	0.117	5	0	0
HLA-A02:03	0.108	5	0	0
HLA-A02:05	-0.013	4	0	0
HLA-A02:06	-0.724	8	4	0
HLA-A02:11	0	5	0	0
HLA-A03:01	0.018	7	0	0
HLA-A03:02	0.07	15	8	0
HLA-A11:01	-0.107	12	7	0
HLA-A11:02	-0.954	11	6	0
HLA-A11:03	-1.165	8	4	0
HLA-A23:01	0.281	7	0	7
HLA-A24:02	0.109	9	0	9
HLA-A24:07	-0.296	10	0	9
HLA-A25:01	-0.282	7	7	0
HLA-A26:01	-0.03	7	5	0
HLA-A26:08	0.058	5	4	0
HLA-A29:01	0.34	6	0	5
HLA-A29:02	0.225	6	0	5
HLA-A30:01	-0.182	11	11	0
HLA-A30:02	0.091	14	0	1
HLA-A30:04	-0.194	69	9	7
HLA-A31:01	-0.265	20	20	0
HLA-A32:01	-0.154	10	10	0
HLA-A33:01	-0.234	11	9	0
HLA-A33:03	0.009	15	9	0
HLA-A34:02	-0.383	13	5	0
HLA-A36:01	-0.126	6	0	0
HLA-A66:01	-0.016	8	5	0
HLA-A68:01	0.034	11	3	0
HLA-A68:02	-0.01	9	2	0
HLA-A69:01	-0.149	12	4	0
HLA-A74:01	-0.337	17	13	0
HLA-A80:01	1.42	9	0	3

Association of each HLA-A allele was calculated using linear regression model. Each row represents a HLA-A allele represented in the dataset (N = 37). The allele's regression coefficient, β estimate, represents its effect on viral load. The total number of predicted HIV-1 epitopes bound by HLA-A alleles (column bound epitopes), and of which how many were protective and risk epitopes are shown.

Table S6. Improvement in model fit by predicting epitopes from autologous HIV-1 sequences.

HIV Protein	Number of Samples	% of non-missing sites in autologous sequences	Variation associated with epitopes from Reference sequence	Variation associated with epitopes from Autologous sequence
Gag	122	86.8	0	0.016
Pol	438	51.1	0.002	0.003
Vpu	103	88.7	0	0
Vpr	102	97.3	0	0
Nef	83	87.0	0.012	0.040
Gp41	83	86.0	0	0
Vif	110	94.5	0.050	0.051
Rev	109	84.1	0	0

In 4 out of 8 HIV-1 proteins for which autologous sequences were available, the variation (estimated as adjusted ΔR^2) in set point viral load (spVL) associated with individual HLA-bound epitope repertoires increased when we predicted epitopes from autologous sequences instead of the HIV-1 reference sequence. Autologous sequence information was only available for a small subset of patients and coverage of the given protein sequences was incomplete. Estimates for both reference and autologous epitopes are therefore based on the same subset of patients and incomplete sequence data in order to make them comparable. No sequence data was available for the *Env* protein.

Table S7. Supertype information for HLA-B alleles.

Allele	Supertype
B07:02	B07
B08:01	B08
B13:02	B13
B14:02	B27
B18:01	B44
B27:02	B27
B27:05	B27
B35:01	B07
B35:02	B07
B35:03	B07
B44:03	B44
B49:01	B49
B52:01	B62
B55:01	B07
B57:01	B58
B57:03	B58

Supertype information for HLA-B alleles (N = 16) that were nominally associated with spVL ($P < 0.05$) is listed. Supertype information was taken from Sidney *et al.* (51). Two-digit resolution was used to infer supertype for alleles that had not been assigned to any supertypes in Sidney *et al.* (51).

Table S8. Comparison of the associated variation in viral load using reference and consensus HIV-1 proteome.

Model parameters	HLA-B	HLA-A	HLA-B + HLA-A
Covariates + Disease-associated epitopes from reference proteome	11.3 (3.3×10^{-168})	1.4 (1.4×10^{-20})	12.0 (1.3×10^{-178})
Covariates + Disease-associated epitopes from consensus proteome	11.0 (6.9×10^{-162})	1.3 (1.4×10^{-19})	11.8 (2.1×10^{-173})

The percentage of variation (estimated as adjusted R^2) in spVL associated with the patient-specific number of predicted HLA-bound HIV-1 epitopes is shown separately for HLA-B and HLA-A, and for different epitope categories and HIV-1 proteome sequences. Consensus HIV-1 sequence did not contain Asp protein, and so we excluded it from reference to do this comparison. *P*-values (in parentheses) indicate the improvement over null model (covariates only: first five PCs and cohort group). Number of disease-associated predicted epitopes using reference sequence is 125 and 71 and using consensus sequence is 123 and 69 for HLA-B and HLA-A respectively.

Table S9. Summary of HIV-1 reference genome.

Protein	Length	Number of HLA-B bound epitopes	Number of HLA-A bound epitopes
Pol	1003	60	43
Gag	500	25	24
Vif	192	15	11
Vpr	96	7	5
Tat	86	1	2
Rev	116	6	6
Vpu	82	7	4
Env	856	60	53
Asp	189	15	15
Nef	206	18	10

HIV-1 genome having accession: NC_001802.1 (11) was used for epitope prediction. It comprises 10 proteins. For each protein, its sequence length along with the number of predicted epitopes bound by HLA-B (N = 214 in total) and HLA-A (N = 173 in total) are shown.

Supporting References

25. The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58.
26. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 44(8):955–959.
27. Delaneau O, Zagury J-F, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10(1):5–6.
28. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5(6):e1000529.
29. Jia X, et al. (2013) Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* 8(6). doi:10.1371/journal.pone.0064683.
30. Jacks T, et al. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* 331(6153):280–283.
31. 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(12):1177–1184.
32. Wickham H, Wickham MH (2007) The ggplot package.
33. McFerrin L, McFerrin ML (2013) Package HDMD. *Stazeno z*.
34. Kim Y, Sidney J, Pinilla C, Sette A, Peters B (2009) Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* 10:1–11.
35. Goslee S, Urban D (2007) The ecodist Package: Dissimilarity-based functions for ecological analysis. *J Stat Softw* 22:1–19.
36. Hahsler M, Piekenbrock M (2017) dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. *R Packag version*:0–1.
37. Wagih O (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 33(22):3645–3647.
38. Addo MM, et al. (2003) Comprehensive Epitope Analysis of Human Immunodeficiency Virus Type 1 (HIV-1)-Specific T-Cell Responses Directed against the Entire Expressed HIV-1 Genome Demonstrate Broadly Directed Responses, but No Correlation to Viral

Load. *J Virol* 77(3):2081–2092.

39. Goulder PJ, et al. (1996) Novel, cross-restricted, conserved, and immunodominant cytotoxic T lymphocyte epitopes in slow progressors in HIV type 1 infection. *AIDS Res Hum Retroviruses* 12(18):1691–1698.
40. Kiepiela P, et al. (2007) CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat Med* 13(1):46–53.
41. Phillips RE, et al. (1991) Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* 354(6353):453–459.
42. Roberts HE, et al. (2015) Structured Observations Reveal Slow HIV-1 CTL Escape. *PLoS Genet* 11(2):1–22.
43. Kaslow RA, et al. (1996) Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med* 2(4):405–411.
44. Bateman A, et al. (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 45(D1):D158–D169.
45. Fellay J, et al. (2009) Common genetic variation and the control of HIV-1 in humans. *PLoS Genet* 5(12). doi:10.1371/journal.pgen.1000791.
46. 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(6):2709–2714.
47. Dean M, et al. (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science (80-)* 273(5283):1856–1862.
48. Limou S, et al. (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J Infect Dis* 199(3):419–26.
49. Le Clerc S, et al. (2009) Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J Infect Dis* 200(8):1194–201.
50. van Manen D, et al. (2011) Genome-Wide association scan in HIV-1-infected individuals identifying variants influencing disease course. *PLoS One* 6(7):2–7.
51. Sidney J, Peters B, Frahm N, Brander C, Sette A (2008) HLA class I supertypes: A revised and updated classification. *BMC Immunol* 9:1–15.

Chapter 3

Does broader antigen presentation underlie HLA-mediated risk for Type 1 Diabetes?

Jatin Arora [1], Suna Onengut-Gumuscu [2], Wei-Min Chen [2], Åke Lernmark [3], Stephen S. Rich [2], Soumya Raychaudhuri [4], Tobias L. Lenz [1,*]

[1] Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany
[2] Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA
[3] Department of Clinical Sciences, Lund University/CRC, Skåne University Hospital SUS, Malmö, Sweden
[4] Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, USA

Introduction

The classical genes of the Human Leukocyte Antigen (HLA) complex play a central role in the adaptive immune system. They encode for cell-surface molecules that present peptides to T-cells. Upon recognizing HLA-presented peptides, T-cells can initiate a specific immune response (*139*). It is important to note that HLA molecules do not discriminate between self- and non-self-peptides while presenting them to T-cells (*23*). Therefore, whereas the higher peptide promiscuity of an HLA molecule makes a stronger immune response against pathogens more likely, it also increases the possibility of activating autoreactive T-cells with immunogenic self-peptides. This trade-off lays down the basis for the concept that the advantage gained against pathogens by HLA-presentation of a broader array of peptides comes at the risk of autoimmunity (*47*, *48*, *69*, *70*). This concept has received some indirect empirical support from the studies in non-model organisms (*13*, *49*, *71–73*), and facilitated the consideration of pathogens and autoimmunity as two antagonistic selection forces operating on HLA genes, where the former would favor broader peptide presentation while the later would limit that. In a host, HLA-presentation of a broader array of peptides could be achieved in different ways. One is by having two different alleles of an HLA gene (heterozygous) instead of the same alleles (homozygous). This mechanism has been named HLA heterozygote advantage hypothesis (*91*) since it is thought to provide stronger resistance against pathogens. This

mechanism is considered to contribute to the unparalleled allelic diversity of classical HLA genes across human populations (*25*) (*86*, *105*, *140*, *141*). It has been studied in multiple infections (*76*, *77*, *96*, *102*, *142*), of which one famous example is HIV (*75*). We previously characterized HLA heterozygote advantage against HIV and showed that, in an HLA heterozygous individual, HLA binding of a larger number of HIV peptides leads to lower viral load (Chapter 1).

In case of autoimmune condition, Lenz *et al.* 2015 have suggested a possible effect of HLA heterozygosity on the risk of multiple autoimmune diseases including Type 1 Diabetes, and Rheumatoid Arthritis (*58*), though a thorough analysis of this effect remains to be done. HLA heterozygosity is considered to result in a broader array of HLA-presented peptides (*38*). However, it is not clear whether HLA heterozygosity confer the risk of autoimmune diseases through a broader array of HLA-bound self-peptides. Alternatively, HLA heterozygosity in an individual might simply make it more likely to carry those HLA alleles that confer risk of autoimmune diseases, since individual HLA alleles have also been associated with different autoimmune diseases. But, it is also not known how an HLA allele confer risk of autoimmune diseases. The fine-mapping of the association between HLA and Type 1 Diabetes (T1D) by Hu *et al.* 2015 has suggested a potential role of the peptides bound by individual alleles in determining their specific effect on T1D outcome. Therefore, considering that HLA-presentation of a broader array of self-peptides might make it more likely to activate autoreactive T-cells, we hypothesize that risk alleles bind a broader array of peptides than protective alleles.

Alternatively, another hypothesis proposes that an HLA allele might confer T1D risk by binding specific disease-causing peptides irrespective of the overall size of its peptide-repertoire. Indeed, antibodies against Insulin and GAD65 antigens can be observed long before the disease onset (*143*, *144*), which implicates specific CD4+ T-cell epitopes in breaking the immune tolerance. In addition, CD4+ T-cell clones reactive to certain β-cell epitopes have been found in individuals with T1D (*145*, *146*), and the adoptive transfer of specific autoantigen reactive CD4+ T-cells, such as the one reactive to Insulin B:9-23 epitope (*147*), has been shown to induce diabetes in the mouse model of T1D.

Here, in order to address the above mentioned issues, we took advantage of a large dataset on Type 1 Diabetes (T1D), an autoimmune disease that is thought to result from CD4+ T-cell-mediated destruction of pancreatic β-cells, from Hu *et al.* 2015 (*66*). It consisted of HLA genotype (at 4-digit level) of 6,651 T1D cases and 9,378 controls. We first tested whether there is a general HLA heterozygote disadvantage in T1D. Then, we tested its

functional basis; whether it is due to a larger breadth of HLA-bound self-peptides in HLA heterozygous individuals compared to homozygous individuals, or higher likelihood of having risk alleles. Then we investigated the functional basis of allele-specific association with T1D risk or protection. We tested whether it is the size of peptide-repertoire or binding specific disease-causing peptides that contributes to their specific effect on T1D.

Results

In this study, we focused on HLA-DRB1 and HLA-DQ genes since the variation in these genes remain the strongest associate of T1D risk and protection to-date (*66*). Overall, 35 DRB1 and 39 DQ alleles were represented in our dataset **(Table S1)**. We first tested whether HLA heterozygosity at DRB1 and DQ genes was associated with T1D. We regressed disease status on zygosity at DRB1 and DQ genes while accounting for allele-specific additive effects. Heterozygosity at both DRB1 and DQ was observed to be associated with T1D risk **(Table 1)**. Notably, while DRB1 heterozygosity had a slightly larger effect on the disease than DQ heterozygosity when tested separately, the effect of DRB1 heterozygosity was not independent of DQ heterozygosity **(Table 1)**, potentially because of high linkage disequilibrium (LD) across HLA region (*108*).

Table 1: The association between HLA heterozygosity and T1D status. The association of DRB1 and DQ heterozygosity with T1D status was calculated using logistic regression model while accounting for the allele-specific additive effects. The coefficient represents the size of the effect of heterozygosity on T1D risk.

Heterozygous locus	Coefficient	*P*-value
DRB1	1.55	4.2×10^{-10}
DQ	1.36	1.1×10^{-5}
DRB1 + DQ	0.9, 1.45	0.65, 0.004

We subsequently investigated different mechanisms that could possibly underlie this heterozygote disadvantage. First, we tested whether HLA-heterozygosity conferred risk because of a broader array of HLA-bound self-peptides in heterozygous individuals than homozygous individuals. We computationally predicted the binding affinity of individual DRB1 and DQ alleles for all possible 8,018 overlapping 15mer self-peptides derived from 17 T1D-relevant candidate human proteins (*148*). We then used the default threshold on

the predicted affinity to select for the HLA-bound peptides, which resulted in 719 and 1,030 peptides bound by one or more DRB1 and DQ alleles, respectively. It was observed that compared to all possible peptides, the predicted HLA-bound peptides were significantly enriched for already known T1D-associated CD4+ T-cell epitopes present in the IEDB database (OR = 1.6 and $P < 0.001$). This suggested that the predicted HLA-bound peptides were relevant to T1D **(Fig. 1 Left)**. We then compared the size of individual-specific peptide-repertoires between HLA heterozygous and homozygous individuals. Heterozygosity at DRB1 or DQ in an individual led to a broader peptide-repertoire than homozygosity at these genes (Wilcoxon rank sum test $P < 0.0001$ for DRB1 and DQ; **Fig. 1 Center**), which remained true even after controlling for heterozygosity at the other of these two genes (Wilcoxon rank sum test $P < 0.0001$ for DRB1 and DQ). Now, given that HLA heterozygosity confers T1D risk and leads to a broader array of HLA-bound peptides in an individual, we hypothesized that T1D cases should have a broader array of HLA-bound peptides than controls. We tested this and found that individual-specific peptide-repertoire bound by individual's DRB1 alleles was larger in cases than controls (Wilcoxon rank sum test $P < 0.0001$ for DRB1 and DQ; **Fig. 1 Right**). However, individual-specific peptide-repertoire bound by DQ alleles was significantly smaller in cases than controls (Wilcoxon rank sum test $P < 0.0001$ for DRB1 and DQ; **Fig. 1 Right**). The opposite effect of the breadth of peptides bound by DRB1 and DQ on the disease risk might not conflict as it has been suggested that DQ molecules preferentially bind those peptides which induce immune tolerance (*149*), implying that the smaller the number of DQ-bound peptides is, the less likely the immune tolerance to self-peptides would be.

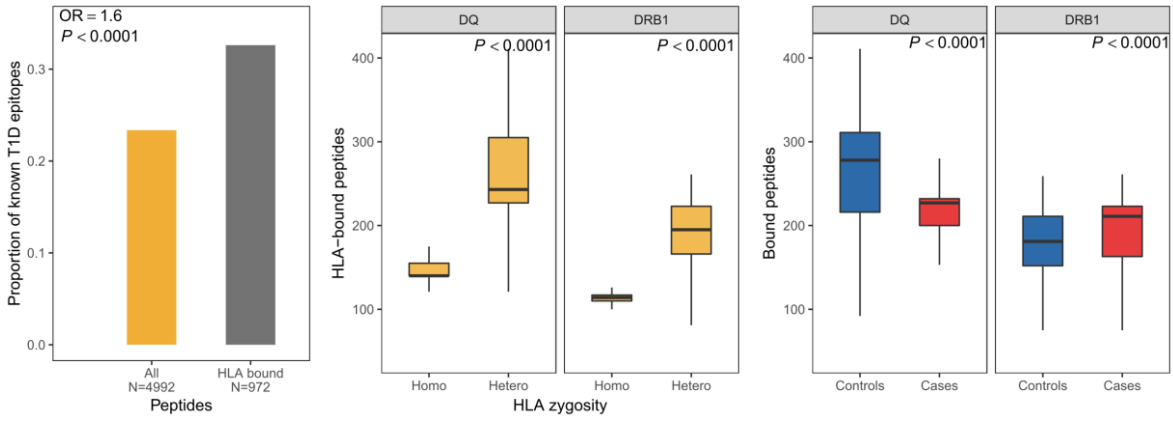


Fig. 1: HLA-bound peptides. (Left) Predicted HLA-bound peptides were enriched for known T1D-associated epitopes present in IEDB database. Odds ratio (OR) and *P*-value from Fisher exact test are shown. **(Center)** Both DRB1 and DQ heterozygosity resulted in a broader individual-specific peptide-repertoire. **(Right)** Individual-specific peptide-

repertoire defined by individual's DRB1 alleles was significantly larger in T1D cases than in controls, while an opposite trend was observed for DQ. *P*-values from Wilcoxon rank-sum test is shown.

Next, we tested the alternative hypothesis that HLA heterozygosity confers risk by making it more likely to carry T1D risk alleles. We first calculated the association of individual HLA alleles with T1D. There were 6 DRB1 and 4 DQ alleles that were nominally associated with the T1D risk **(Fig. S1)**, of which 5 DRB1 and 2 DQ risk alleles were enriched in HLA heterozygosity **(Fig. S1)**, supporting that the hypothesis was true. Next, we investigated the functional basis of T1D risk conferred by DRB1 and DQ alleles. We first tested whether risk alleles bind a broader array of peptides compared to protective alleles. We observed no significant difference in the size of peptide-repertoire of risk and protective alleles of both DRB1 and DQ genes **(Fig. 2 Left)**. Then we sought which other properties of risk alleles could explain their association with T1D and distinguish them from protective alleles. T1D is thought to result from action of β-cell antigen-specific autoreactive T-cells activated by recognizing HLA-presented self-peptides (*150*). In an ideal situation, autoreactive T-cells should not exist in the mature peripheral T-cell repertoire due to negative T-cell selection in the thymus, though their presence have been shown even in healthy individuals (*24*), indicating that autoreactive T-cells can escape their negative selection in thymus. Several reasons linked to the peptide-affinity of HLA molecules have been put forward e.g. unstable HLA-self-peptide complex (*151*), or poor HLA-presentation of self-peptides to naïve T-cells. So, we tested if risk alleles bind peptides with different affinity than protective alleles. The risk alleles of both DRB1 and DQ genes exhibited lower peptide-binding affinity than protective alleles (Wilcoxon rank sum test $P < 0.0001$ for DRB1 and DQ; **Fig. 2 Right**). We further tested if the observed insignificant difference in size of the peptide-repertoire and significantly different peptide-binding affinity of risk and protective alleles were specific to T1D-relevant candidate proteins used in this study. We randomly sampled human, bacterial and viral proteins (1,000 for each) and calculated allele-specific peptide-repertoire from them. In all three cases of human, bacterial and viral proteins, the size of the peptide-repertoire of risk alleles was not significantly different from protective alleles ($P > 0.05$ for DRB1 and DQ, and for human, bacterial and viral proteins; **Fig. S2**), and risk alleles exhibited lower affinity than protective alleles ($P < 0.0001$ for DRB1 and DQ, and for human, bacterial and viral proteins; **Fig. S3**). This suggests that indifferent size of peptide-repertoire of risk and protective alleles and the

lower peptide-binding affinity of risk alleles than protective alleles are their general characteristics.

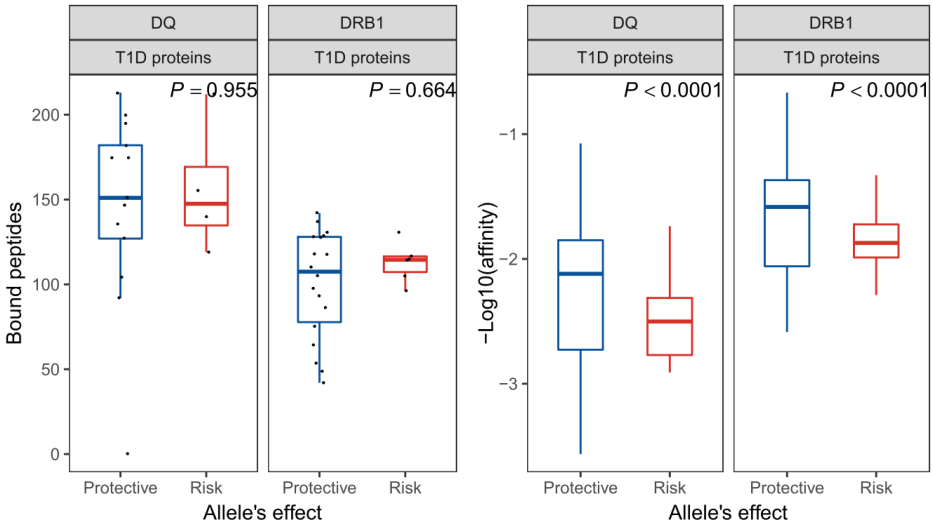


Fig. 2: Peptide-repertoire of T1D-associated HLA alleles. (Left) The number of peptides derived from candidate T1D-relevant proteins (N = 17) and bound by risk and protective alleles of DRB1 and DQ genes. **(Right)** The binding-affinity of risk and protective alleles of DRB1 and DQ genes for the bound peptides. *P*-values from Wilcoxon rank-sum test is shown.

The aggregation of multiple autoimmune diseases even within individuals is not an uncommon phenomenon (*152*). Genome wide association studies (GWASs) have linked variation in HLA class-II region with multiple autoimmune disease including T1D, Rheumatoid Arthritis (RA) and Systemic Lupus Erythematosus (SLE) (*62*, *153*). In addition, disease-associated variation across these autoimmune diseases have been shown to lie inside the peptide-binding groove of HLA molecules (*55*, *66*, *79*), which has resulted in the association of multiple DRB1 and DQ alleles diseases with more than one autoimmune disease. Although the alleles have been associated with differential effect, possibly due to different disease-causing antigens they bind, the sharing of alleles provides an evidence of shared immunological characteristics. Therefore, we tested if our findings from T1D could be extended to RA and SLE, and if the differential association of alleles could be explained by the size of their peptide-repertoire or peptide-binding affinity. Here, we focused on the alleles of DRB1 gene only, because the alleles of DQA1 and DQB1 chains have mostly been separately associated across these diseases (*62*) and the prediction of DQ-bound peptides requires both chains, complicating the analysis. Moreover, since the identity of the self-proteins containing epitopes relevant for RA and SLE is not known, we took HLA class-II restricted antigens with positive T-cell assay for

T1D, RA and SLE present in IEDB database and predicted DRB1-bound 15mer peptides from them (designated as *epitopes* here onwards). Considering that the common disease-associated HLA alleles are more likely to be studied than others and they had larger empirical affinity data for training in peptide-binding prediction method that was used here (*154*) **(Fig. S3)**, we restricted our analysis to the highly significantly disease-associated alleles that were shared between at-least 2 diseases and predicted to bind at-least 5 epitopes. On average, DRB1 alleles associated with risk of RA (DRB1*04:01), SLE (DRB1*03:01) or T1D (DRB1*03:01, DRB1*04:01) had lower peptide-affinity than the alleles associated with protection from these diseases (DRB1*13:01 and DRB1*15:01 for RA and T1D; Wilcoxon rank sum test $P = 4.6 \times 10^{-12}$; **Fig. 3 Left**). The allele-specific comparison across these autoimmune diseases also suggested that an allele might confer risk of an autoimmune disease by binding the epitopes of that disease with low affinity (**Fig. 3 Right**). In order to make sure that predicted affinity between DRB1 alleles and IEDB epitopes is not specific to these epitopes, we predicted the affinity of these alleles for the peptides derived from randomly sampled human proteins (N = 1,000), which suggested that variation in the predicted affinity between DRB1 alleles and IEDB epitopes reflects their general affinity range (**Fig. 3 Right**). The peptide-affinity of DRB1:1501 across these diseases is of particular interest here. It confers SLE risk and bound SLE epitopes with lower affinity than RA and T1D epitopes which it conferred protection (ANOVA $P = 0.018$; **Fig. 3 Right**).

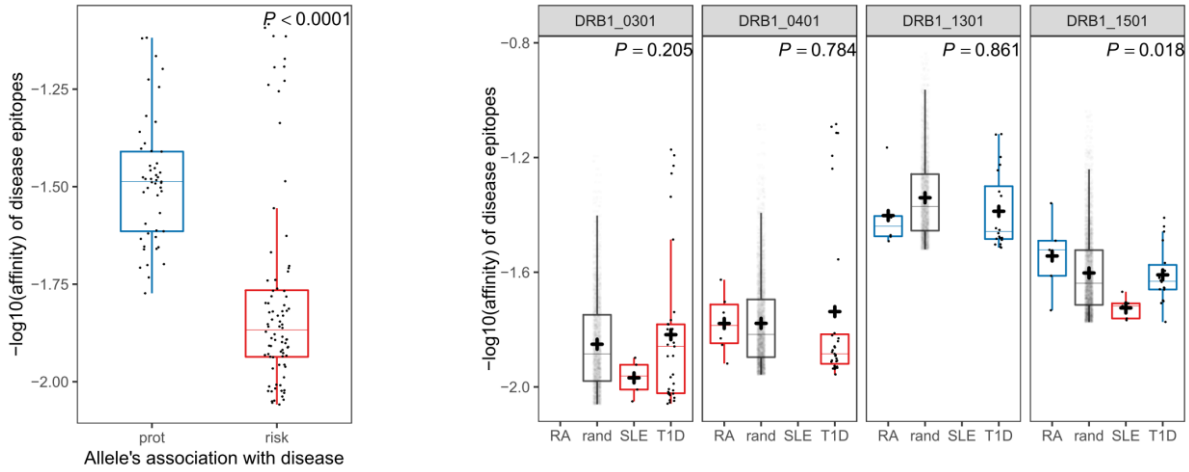


Fig. 3: Peptide-binding affinity of DRB1 alleles associated with multiple autoimmune diseases. (Left) The comparison of predicted peptide-binding affinity between DRB1 alleles associated with risk and protection from Rheumatoid Arthritis (RA), Systemic lupus (SLE) and Type 1 Diabetes (T1D). *P*-value from Wilcoxon rank sum test is shown. **(Right)** The peptide-binding affinity of individual common disease-

associated DRB1 alleles for the epitopes associated with RA, SLE, T1D and the peptides derived from randomly taken human proteins (N = 1,000) is shown (rand). The disease-associated epitopes were taken from the IEDB database and the human proteins were taken from UniProt database. Plus "**+**" sign shows the mean values. *P*-values from ANOVA is shown.

Next, we tested the hypothesis that risk alleles confer risk because they bind specific disease-causing peptides. Post-translational modification (PTM) of proteins can generate neo-peptides for which central and peripheral tolerance might not exist. Deamidation is a type of PTM which is triggered by cellular stress and takes places in endoplasmic reticulum (ER) (*155*). Deamidation is considered not to take place in thymus during T-cell maturation, suggesting that deamidated peptides (neo-peptides) might be the potential target of immune response (*156*). Pancreatic β-cells specialize in their function of insulin production. Due to being not very large in number and high turn-over rate, they are prone to ER stress (*157*). Deamidated peptides from β-cell antigens have been shown to be bound by T1D risk alleles(*158*, *159*) and recognized by the T-cell from individuals with T1D (*160*). Therefore, we tested whether risk alleles differ from protective alleles in their binding of deamidated-peptides. Following Vader *et al*. (*161*), we simulated the deamidation of candidate T1D-relevant proteins and predicted 120 and 240 deamidated peptides that were bound by DRB1 and DQ alleles, respectively. The binding of a larger number of deamidated-peptides by DQ alleles reflects their known strong preference for the negatively charged deamidated residues (*162*, *163*). While DRB1 risk and protective alleles did not differ significantly in the number of bound deamidated-peptides and the affinity for them **(Fig. 4)**, DQ risk alleles bound a larger number of deamidated peptides and with slightly lower affinity than protective alleles (Wilcoxon rank sum test $P = 0.008$ and $P < 0.0001$ for the number of peptides and affinity, respectively; **Fig. 4**).

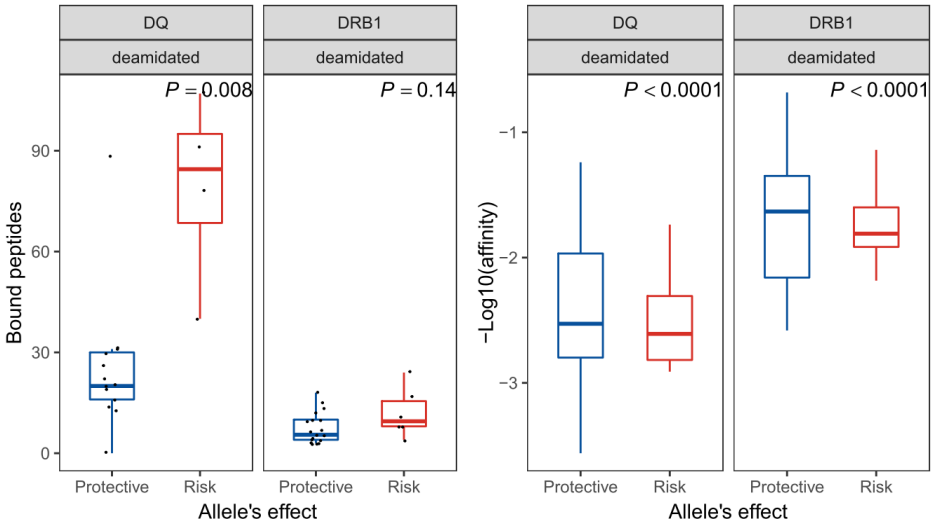


Fig. 4: Deamidated peptide-repertoire of T1D-associated HLA alleles. (Left) The number of deamidated peptides simulated from candidate T1D-relevant proteins (N = 17) and bound by risk and protective alleles of DRB1 and DQ genes. **(Right)** The binding affinity of risk and protective alleles of DRB1 and DQ genes for deamidated peptides. *P*-values from Wilcoxon rank-sum test is shown.

Discussion

We tested whether the heterozygosity at HLA-DRB1 and HLA-DQ genes confer T1D risk in addition to allele-specific effects. We observed HLA heterozygote disadvantage for both DRB1 and DQ genes. Both DRB1 and DQ heterozygosity resulted in a broader array of HLA-bound self-peptides. A broader array of DRB1-bound peptides in T1D cases than controls suggests the quantity of DRB1-bound peptides might contribute to DRB1 heterozygote disadvantage. Interestingly, DQ genotype (combination of both alleles) bound a smaller number of peptides in T1D cases than controls. This is in agreement with the previously shown association between the induction of immune tolerance and DQ-presented peptides (*149*). The opposite trend for the number of DQ-bound peptides in heterozygotes and cases suggests that the basis of DQ heterozygote disadvantage might be more complex than mere the number of HLA-bound self-peptides. As Lenz *et al.* 2015 have shown that the interactions between the alleles of DQ gene in heterozygote individuals might confer additional risk of T1D (*58*), there might be additional factors contributing to HLA heterozygote disadvantage.

In addition, DRB1 and DQ heterozygosity in an individual led to the higher likelihood of carrying T1D risk-associated alleles. This might, at-least partly, also be due to the presence of a large number of HLA alleles at intermediate frequency and the excess of

heterozygosity. Further, in order to understand the functional basis of the specific association of risk and protective alleles with T1D, we compared the size of peptide-repertoire and the peptide-binding affinity of risk and protective alleles. The insignificant difference in the size of peptide-repertoire derived from candidate T1D-relevant beta-cell proteins does not lend support to the hypothesis that alleles confer T1D risk because they bind a broader array of self-peptides. Rather, the lower peptide-binding affinity of risk alleles compared to protective alleles suggests the peptide-binding affinity as a potential determinant of the effect of individual alleles on T1D risk. Some studies have suggested that the affinity between a peptide and an HLA molecule might affect the interaction of peptide-HLA complex with T-cells (*164*, *165*), and low avidity interaction of immature T-cells with self-peptide-HLA complex is required for their survival in thymus (*23*).

This leads to the speculation that HLA alleles with low peptide-binding affinity might form unstable complexes with self-peptides that do not allow efficient removal of autoreactive T-cells in thymus. Such autoreactive T-cells in periphery might get activated by non-genetic triggers e.g. the infections with pathogens (*166*–*168*). We found the sequence homology between predicted T1D-associated epitopes and the Heat shock proteins in *P.* tuberculosis. This bring forward the possibility that the immune response initiated against the pathogens containing self-mimicking epitopes might spread to self-epitopes (*166*, *169*), possibly leading to an overt autoimmune disease.

Furthermore, the comparison of peptide-binding affinity of DRB1 alleles across multiple autoimmune diseases shed light on the functional basis of their differential association. The particular case of DRB1*15:01 illustrates how an allele, which confer protection from RA and T1D by binding their epitopes with higher affinity, could have opposite outcome in SLE by binding SLE epitopes with lower affinity. In addition, we tested the alternative hypothesis that alleles confer T1D risk by binding specific disease-causing peptides. One such type of peptides could be represented by deamidated peptides, which results from the enzymatic post-translational modification of the proteins in endoplasmic reticulum triggered by the cellular stress. As deamidation is considered not to take place in thymus, deamidated peptides could be viewed as foreign by the immune system (*156*), making them a potential target of the immune response. We observed that the risk alleles bound a large number of deamidated peptides. As discussed above for the self-mimicking pathogen-derived peptides, the sequence homology between self and deamidated peptides might allow the immune response to get triggered by deamidated peptides and later shift to native self-peptides over the course of the disease (*170*).

Overall, the findings in this study provided an empirical evidence for the HLA heterozygote disadvantage in T1D. Considering the similarities in the etiology among autoimmune diseases (*62*, *153*), it is probable that HLA heterozygosity contribute to the risk of other autoimmune diseases with the similar mechanisms as in T1D, but, of course, additional studies focused on individual diseases are required to validate this proposition. The findings also suggest that multiple factors, a large breadth of HLA-bound self-peptides and higher likelihood of carrying specific risk alleles, can contribute to the HLA heterozygote disadvantage. Additional work would be needed to elucidate other possible factors. In addition, the findings also provided low peptide-binding affinity and particular disease-causing peptides as the possible basis for the association of individual HLA alleles with T1D risk.

Material and methods

Samples and genotype data

We analyzed data of 16,086 samples (6,670 T1D cases and 9416 controls) from Type 1 Diabetes Genetics Consortium (T1DGC). The original data and the detailed quality check are described in detail in Hu *et al.* (*66*). Briefly, the samples were collected from 13 regions and genotyped on Immunochip array. HLA alleles (best-guess genotypes at 4-digit resolution) for classical class I (HLA-A, -B, -C) and class II loci (HLA-DRB1, -DQA1, -DQB1, -DPA1, -DPB1) were imputed from genome-wide genotype data using SNP2HLA and a reference panel of 5,225 individuals of European ancestry (*171*). Overall, 35 and 39 alleles of HLA-DRB1 and HLA-DQ genes were represented in our dataset, respectively **(Table S1)**.

Affinity between HLA and peptides

The sequences of all human proteins were taken from UniProt database (*172*). We used NetMHCIIpan v3.2 (*154*) for calculating the binding affinity between HLA class-II alleles and peptides. It is a neural network based computational method that predicts binding affinity of peptides of variable length to any known HLA class II molecule. It also reports the rank of predicted binding affinity of HLA-peptide complexes against predicted affinity of 200,000 random natural peptides. Using this method, we predicted HLA DRB1 and DQ allele-specific binding affinities for all 15mer peptides generated from given set of proteins. We used the default cutoff of 2 on the %rank of predicted affinity to select for HLA-bound peptides. 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.

Supplementary data

Table S1. HLA-DRB1 and HLA-DQ alleles represented in our dataset.

DRB1 (N = 35)	DQ (N = 39)
HLA-DRB1-0101	HLA-DQA10501-DQB10201
HLA-DRB1-0102	HLA-DQA10201-DQB10402
HLA-DRB1-0103	HLA-DQA10501-DQB10301
HLA-DRB1-0301	HLA-DQA10101-DQB10501
HLA-DRB1-0401	HLA-DQA10301-DQB10301
HLA-DRB1-0402	HLA-DQA10301-DQB10303
HLA-DRB1-0403	HLA-DQA10103-DQB10603
HLA-DRB1-0404	HLA-DQA10102-DQB10603
HLA-DRB1-0405	HLA-DQA10401-DQB10402
HLA-DRB1-0406	HLA-DQA10301-DQB10302
HLA-DRB1-0407	HLA-DQA10201-DQB10202
HLA-DRB1-0408	HLA-DQA10102-DQB10602
HLA-DRB1-0701	HLA-DQA10101-DQB10503
HLA-DRB1-0801	HLA-DQA10201-DQB10303
HLA-DRB1-0802	HLA-DQA10102-DQB10504
HLA-DRB1-0803	HLA-DQA10102-DQB10609
HLA-DRB1-0804	HLA-DQA10103-DQB10601
HLA-DRB1-0901	HLA-DQA10102-DQB10604
HLA-DRB1-1001	HLA-DQA10101-DQB10602
HLA-DRB1-1101	HLA-DQA10501-DQB10302
HLA-DRB1-1102	HLA-DQA10102-DQB10502
HLA-DRB1-1103	HLA-DQA10601-DQB10301
HLA-DRB1-1104	HLA-DQA10102-DQB10501
HLA-DRB1-1201	HLA-DQA10102-DQB10201
HLA-DRB1-1301	HLA-DQA10201-DQB10201
HLA-DRB1-1302	HLA-DQA10301-DQB10201
HLA-DRB1-1303	HLA-DQA10301-DQB10304
HLA-DRB1-1401	HLA-DQA10401-DQB10301
HLA-DRB1-1402	HLA-DQA10102-DQB10601
HLA-DRB1-1404	HLA-DQA10501-DQB10202
HLA-DRB1-1501	HLA-DQA10101-DQB10502
HLA-DRB1-1502	HLA-DQA10103-DQB10602
HLA-DRB1-1503	HLA-DQA10301-DQB10402
HLA-DRB1-1601	HLA-DQA10201-DQB10301
HLA-DRB1-1602	HLA-DQA10301-DQB10603
	HLA-DQA10103-DQB10503
	HLA-DQA10301-DQB10602
	HLA-DQA10201-DQB10302
	HLA-DQA10301-DQB10502

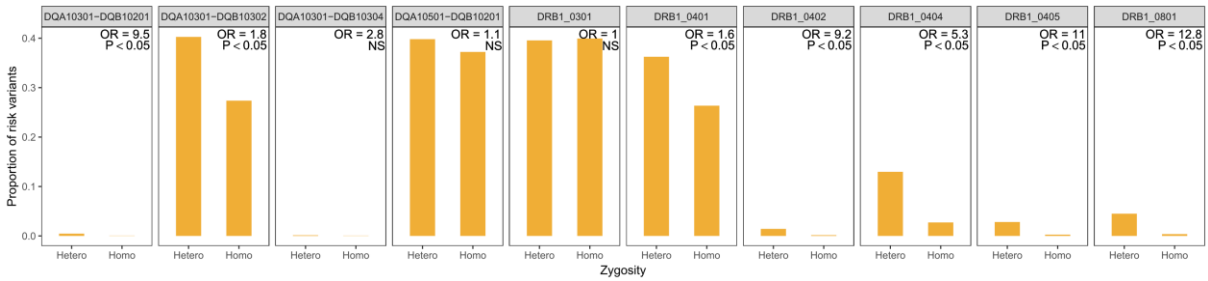


Fig S1. Enrichment for T1D risk-associated HLA-DRB1 and HLA-DQ alleles in heterozygous individuals compared to homozygous individuals. Odds ratio (OR) and *P*-value from Fisher exact test are shown.

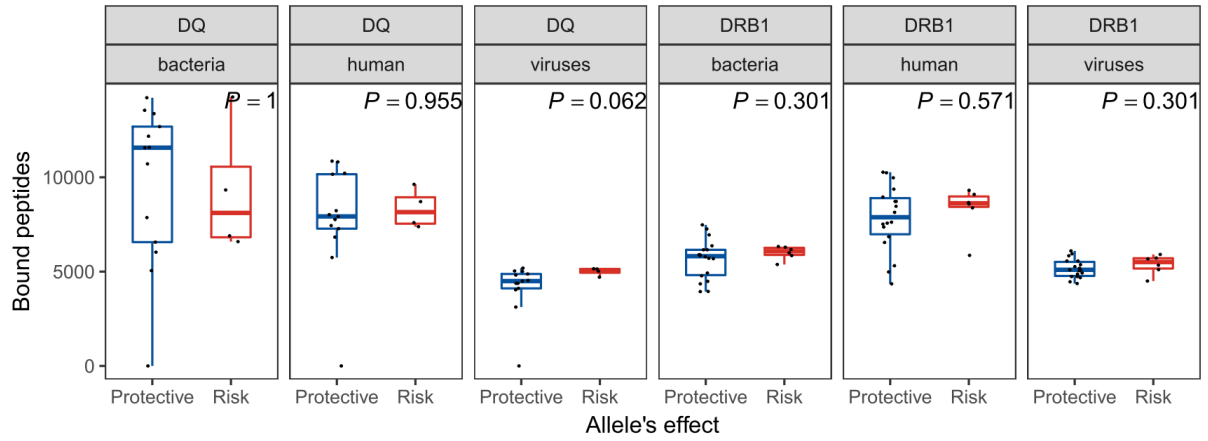


Fig. S2: Peptide-repertoire of T1D-associated HLA alleles derived from random proteins. The number of peptides derived from randomly samples human, bacterial and viral proteins and bound by risk and protective alleles of DRB1 and DQ genes is shown. *P*-value from Wilcoxon rank-sum test and without correction for multiple testing is shown.

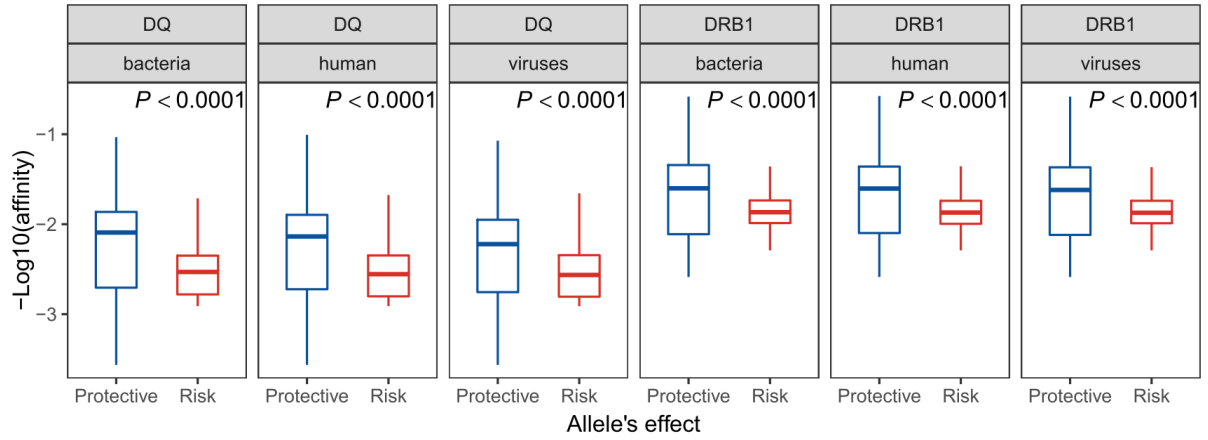


Fig. S3: Peptide-binding affinity of T1D-associated HLA alleles. The binding affinity of risk and protective alleles of DRB1 and DQ genes for the peptides derived from randomly samples human, bacterial and viral proteins is shown. *P*-value from Wilcoxon rank-sum test is shown.

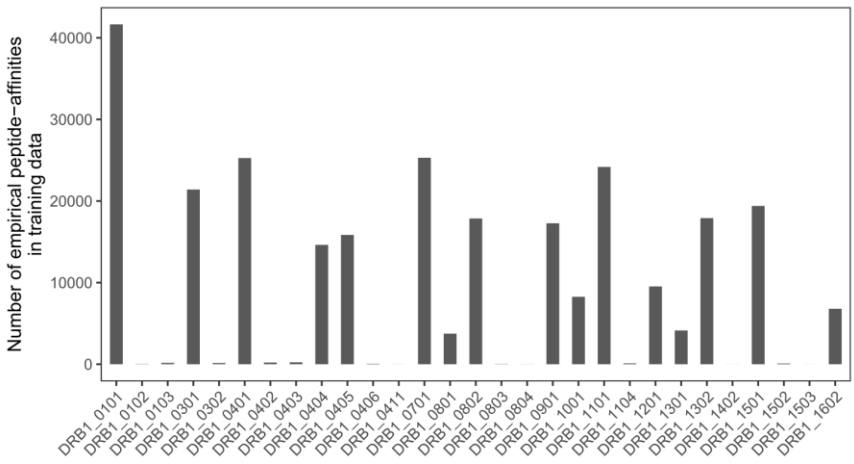


Fig S3. The count of empirical affinity data for DRB1 alleles in training data of NetMHCIIpan v3.2.

Chapter 4

Individual-specific analysis of HLA-presented peptide repertoires reveals novel epitopes associated with Type 1 Diabetes

Jatin Arora [1], Suna Onengut-Gumuscu [2], Wei-Min Chen [2], Åke Lernmark [3], Stephen S. Rich [2], Soumya Raychaudhuri [4], Tobias L. Lenz [1,*]

[1] Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany
[2] Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA
[3] Department of Clinical Sciences, Lund University/CRC, Skåne University Hospital SUS, Malmö, Sweden
[4] Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, , USA

Introduction

Type-1 Diabetes (T1D) is a highly heritable autoimmune disease that results from the chronic loss of insulin-producing β-cells in the islets of pancreas. According to the 2015 report of International Diabetes Foundation (*173*), T1D is increasing by 3% every year, particularly in the age group of 0-14 years, and there is no available cure to date. The situation represents a major health challenge. It is thought that T1D involves T cell-mediated destruction of β-cells (*150*). T1D patients commonly show antibodies against islet proteins months to years before disease symptoms (*143*). These autoantibodies are used as the biomarkers of T1D-associated autoimmunity (*143*, *150*). Notably, the appearance of autoantibodies reflects the presentation of autoantigens by HLA molecules on B-cells and antigen presenting cells, and the subsequent activation of autoantigen-specific T-cells. In fact, genome wide association studies (GWASs) have associated T1D susceptibility to >50 loci (*54*), with particularly the largest association seen for Human Leukocyte Antigen (HLA) locus. HLA genes encode for cell surface molecules that present intra- and extra-cellular peptides to T-cells and mediate the immune response. Besides the strong association of these loci, we know little about the functional basis of T1D pathology (*81*). The fine mapping of GWAS association between T1D and HLA has linked T1D susceptibility to three independent amino acid residues in HLA-DRB1 and HLA-DQ genes (*66*). These residues were localized inside the peptide-binding groove and accounted for 26.9% of variation in the disease status (*66*), lending a strong support to a

major contribution of DRB1 and DQ bound epitopes in disease risk. The existing methods like profiling autoantibodies targeting islet antigens and islet reactive CD4+ T-cells in T1D patients, non-obese diabetic (NOD) mice models etc. have enabled the discovery of a number of the candidate proteins that contain potential epitopes (Di Lorenzo et al. 2007). However, the precise identity and significance of disease-relevant epitopes remains largely unknown (*174*). One major reason is that an HLA molecule can bind a very large number of peptides (*37*), and therefore it would require a full-factorial experimental assay to screen those peptides that can bind disease-relevant HLA alleles. Such an assay has been unmanageable to date that puts the hurdle to precisely identify disease-relevant epitopes derived from given proteins. To this end, we used a computational approach to find T1D-associated epitopes and performed our previously established Peptidome-wide Association Study (PepWAS) on 6,651 T1D cases and 9,378 controls to interrogate 8,018 15mer peptides derived from 17 candidate β-cell proteins. PepWAS identified and prioritized a core set of T1D-associated epitopes bound by DRB and DQ alleles. The difference between cases and controls in the breadth of individual-specific repertoire made of these epitopes accounted for even more deviance in disease status compared to previously T1D-associated genetic variants.

Results

Prediction of HLA-bound peptides

The fate of naïve T-cells depends on their interaction with self-peptide-HLA complex in thymus. T-cells that interact strongly with self-peptide-HLA complex are removed so as to avoid auto-reactivity in the later stage of life, though the process is not foolproof. The strength of the interaction depends on the stability of self-peptide-HLA complex, which itself is affected by the binding affinity between self-peptides and HLA (*164*, *165*). Since the HLA-binding affinity range that allows the survival of naïve T-cell is not known and it is quite permissive, we started without any *a-priori* assumption over the strength of the binding affinity between HLA and potential T1D-associated epitopes. We predicted the binding affinity for each of 35 HLA-DRB1 and 39 HLA-DQ alleles represented in the dataset **(Table S1)** with all 8,018 possible 15mer peptides derived from 17 β-cell proteins with potential implication in T1D (*148*). In order to select for HLA-bound peptides, we used a model that regressed the disease status on the number of bound peptides in each individual that lie within a given window of the rank of predicted binding affinity against

random peptides (see methods). We screened the windows of different sizes to select for the one that maximized the enrichment for previously known T1D-associated epitopes taken from Lorenzo *et al.* (*174*) and the IEDB database, and the fit of the model. This revealed the rank window of 0 to 1 as the best suited to these criteria **(Fig. S1)**. We then took those peptides to be bound by DRB1 and DQ alleles within this rank range, resulting in 408 and 639 peptides to be bound by one or more DRB1 and DQ alleles, respectively. The selected peptides were significantly enriched for known T1D-associated epitopes compared to all possible 8,018 peptides (Fisher exact test OR = 1.44, $P = 0.001$, **Fig. 1 Left**).

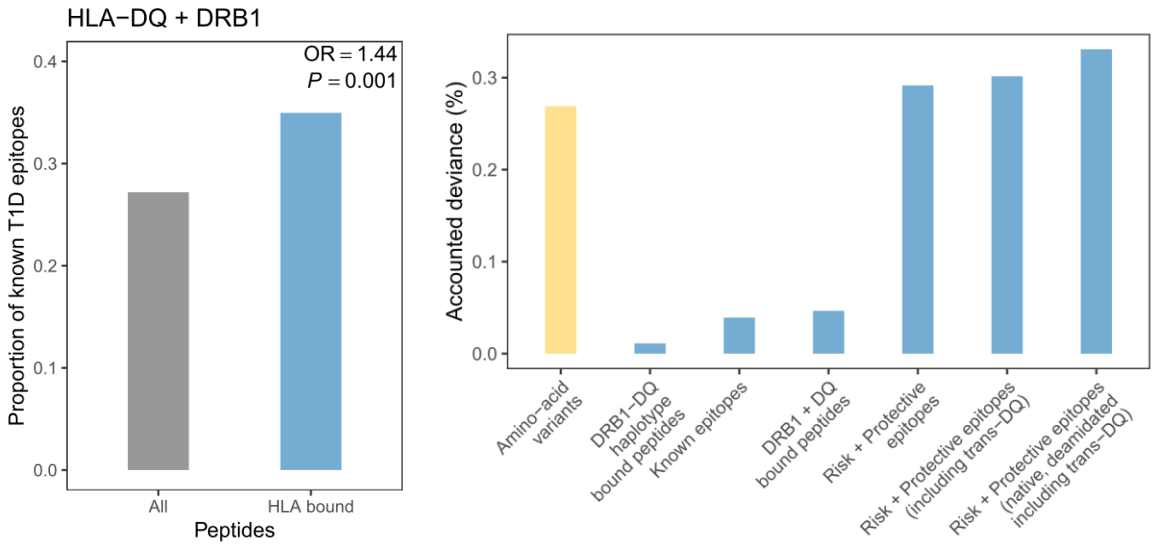


Fig. 1. HLA-bound peptides and accounted deviance in disease. (Left) The predicted HLA-DRB1- and HLA-DQ-bound peptides were significantly enriched for known T1D-associated epitopes taken from previous literature and the IEDB database. The odds ratio (OR) and *P*-value from fisher exact test are shown. **(Right)** The deviance in T1D status accounted for by T1D-assocaited amino-acid residues in DRB1 and DQ in Hu et *al*. 2015, predicted individual-specific repertoires of all peptides bound by DRB1-DQ haplotype, DRB1 and DQ genes additively, known T1D-associated epitopes from the IEDB database, predicted T1D risk and protection-associated epitopes bound by DRB1, DQ and trans-encoded DQ, as well as with predicted deamidated epitopes.

Deviance in disease status associated with bound peptides

We tested whether the individual-specific repertoire of HLA-bound peptides, which is taken as the total number of unique peptides predicted to be bound by the HLA alleles of an individual, was associated with disease. We first focused on previously known T1D-associated antigens taken from Lorenzo et al. (*174*) and IEDB database, of which 102 and 151 were represented in the predicted DRB1- and DQ-bound peptides. The individual-specific peptide-repertoire derived from known T1D-associated epitopes bound by DRB1 and DQ alleles additively accounted for 3.9% deviance in disease **(Fig. 1 Right)**. It was

significantly lower than 26.9% associated with T1D-associated amino-acid residues in these genes, suggesting that a majority of T1D-associated epitopes that underlie genetic association remains missing. Majority of studies have so-far associated T1D risk with specific HLA-DRB1-DQ haplotypes (*54*, *66*) as the extensive linkage disequilibrium (LD) across the HLA region hampers the estimation of gene-specific contribution in the disease (*175*, *176*). The predicted HLA-specific peptide-repertoire allowed us to approximate this. The individual-specific peptide-repertoire defined by HLA-DRB1-DQ haplotypes accounted for 1.1% of deviance in disease, while the one defined by individual's DRB1 and DQ alleles separately accounted for 1.9% and 3.7% of deviance in disease status, respectively **(Fig. 1 Right**). Interestingly, additively DRB1 and DQ accounted for 4.6% of deviance.

Prediction of T1D-associated epitopes

The accounted deviance was still lower than 26.9% associated with amino acid residues ins DRB1 and DQ genes (*66*), suggesting that the predicted peptide-repertoires still contained peptides irrelevant for the association between HLA and T1D. In order to select for T1D-associated peptides, we used our Peptidome-wide Association Study (PepWAS) approach that allows screening any set of HLA-bound peptides for their relevance with disease phenotype (see methods and chapter 2). Of 408 and 639 predicted DRB1- and DQ-bound peptides, respectively, PepWAS predicted 346 and 399 to be significantly associated with T1D. Subsequently, we merged the nested peptides, narrowing down the predicted repertoire to 233 and 322 for DRB1 and DQ, respectively **(Fig. 2, Table S2 and S3)**. Here onwards, we will designate these peptides as *predicted T1D-associated epitopes.* Notably, on average, an individual's DQ alleles bound nearly twice as many T1D-associated epitopes as bound by DRB1 alleles (138 ±27 vs. 74 ±16; Wilcoxon rank sum test $P < 0.001$) **(Fig. S2)**, which potentially explains the stronger association of DQ amino-acid residues with T1D compared to DRB1. We then evaluated the goodness of predicted T1D-associated epitopes. First, upon permuting the disease status of individuals, none of HLA-bound peptides showed significant association with T1D. Second, we hypothesized that candidate β-cell proteins should contain more T1D risk-associated epitopes than any randomly sampled human proteins. We compared the proportion of risk-associated epitopes derived from 7 candidate proteins, which did not contain any already known epitopes, with the predicted risk-associated epitopes from 7 proteins sampled from 300 random human proteins. Upon repeating this for 10,000 times, we observed that the

observed proportion of risk-associated epitopes in candidate proteins was unlikely to be observed in random human proteins ($P < 0.01$ for both DRB1 and DQ) **(Fig. S3)**. Moreover, our approach was able to predict four epitopes within recently discovered PTPRN$_{142\text{-}159}$ islet autoantigen (*177*). While the autoantigen was eluted from high risk conferring DQ alleles, T-cell response was observed against it in only 2 of 21 T1D patients (*177*), which is in agreement with the observed odds ratio of the predicted epitopes (OR < 1). Together, these evidences confirmed the specificity of our approach and the relevance of predicted epitope-repertoires for T1D.

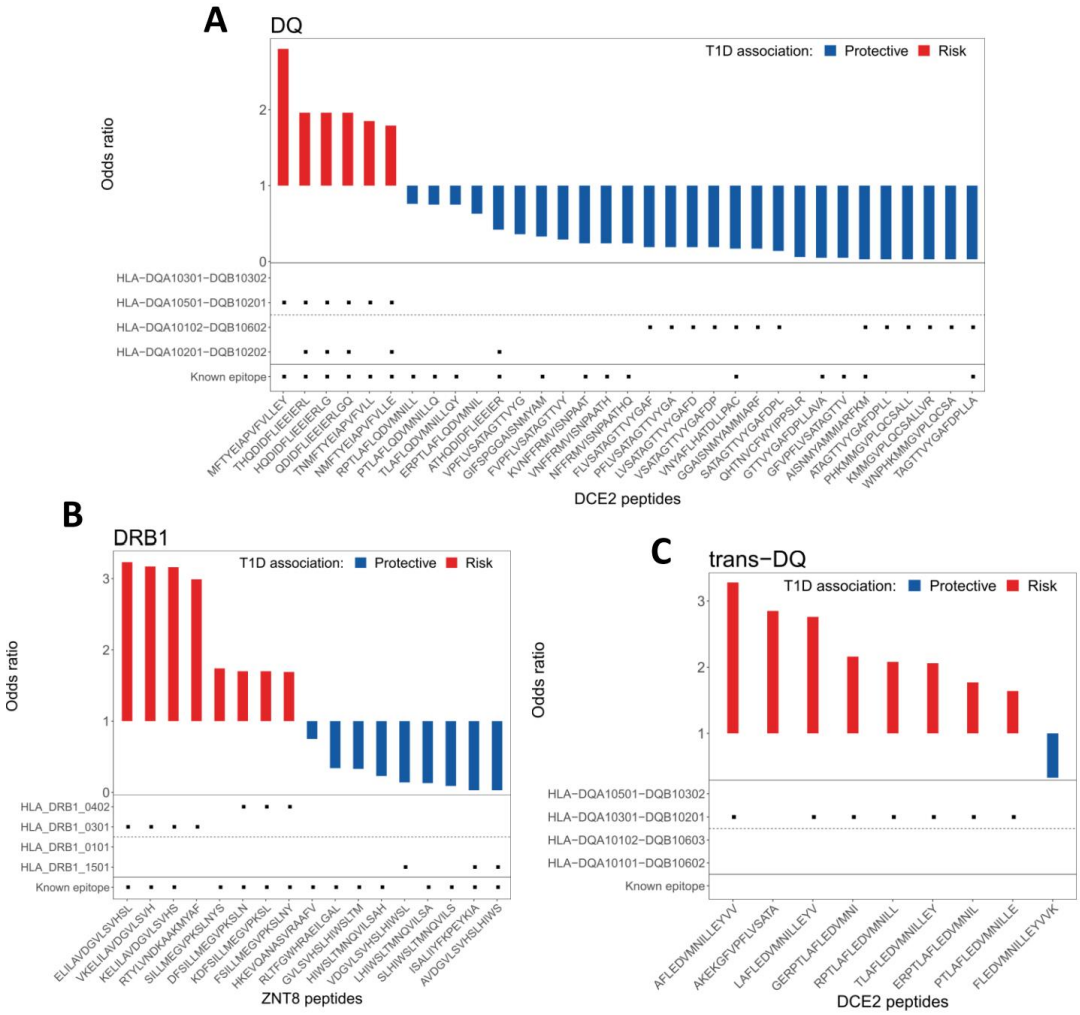


Fig. 2. Epitope-specific association with T1D. (A-C) Predicted DRB1-, DQ- and trans-encoded DQ-bound epitopes were differentially associated with T1D. Height of the bar represents epitope's odds ratio, while the color reflects its effect on T1D, either risk (red) or protection (blue). Note that epitopes' effect is estimated separately and, therefore, are not independent. DCE2 (also known as GAD2) **(A)**, ZNT8 **(B)** and deamidated DCE2 **(C)** proteins are shown as representative examples, together with information on predicted binding for the top 2 protective and 2 risk HLA alleles and whether predicted epitopes are known epitopes in the IEDB database. The predicted epitopes for other candidate proteins are listed in **tables S2-S7**.

Deviance in disease status associated with predicted T1D-associated epitopes

We then tested whether predicted T1D-associated epitopes accounted for a larger deviance in disease than previously known epitopes and all HLA-bound peptides. In an additive manner, individual-specific repertoires of T1D risk and protection-associated epitopes bound by DQ accounted for 22.2% of deviance in disease **(Fig. 1 Right)**. Similarly, individual-specific T1D-associated epitopes bound by DRB1 additively accounted for 24% of deviance **(Fig. 1 Right)**. The risk- and protection-associated epitopes bound by DRB1 and DQ together accounted for 29.2% of the deviance in disease, which was nearly 2% more than T1D-associated amino-acid residues in DRB1 and DQ. Interestingly, among predicted DRB1-bound T1D-associated epitopes, the highest risk (OR > 5) was associated with the peptides derived from G6PC2, Islet cell autoantigen-1 and Heat-shock 70 proteins, while among DQ-bound T1D-associated epitopes, the highest risk (OR > 6) was associated with the peptides derived from PDX1, GFAP, Insulin and PTPR2 **(Table S2 and S3)**. Moreover, specific combinations of DQ alleles have been shown to exert non-additive and interaction effect which confer additional T1D risk (*58*, *66*). One proposed mechanism for that was the presentation of self-epitopes by trans-encoded DQ molecules in DQ heterozygous individuals. We therefore screened all peptides bound by trans-encoded DQ alleles (N = 112) represented in our dataset. PepWAS predicted 500 T1D-associated epitopes bound by one or more trans-encoded DQ alleles **(Table S4)**. In an additive manner, predicted individual-specific repertoires of T1D-associated epitopes bound by DRB1, DQ and trans-encoded DQ molecules together accounted for 30.2% of deviance in disease **(Fig. 1 Right)**.

Prediction of T1D-associated deamidated epitopes

Post-translational modification (PTM) of proteins can generate neo-peptides for which central and peripheral tolerance might not exist, making them potential trigger of autoimmunity in several diseases (*178*). One particular PTM of interest in T1D is deamidation, a cell stress-induced PTM by Tissue transglutaminase (tTG) enzyme in endoplasmic reticulum (*155*). HLA-DQ molecules, specifically DQ2/DQ8 alleles which are associated with the high risk of celiac disease and T1D, have been shown to preferentially bind deamidated peptides (*159*, *161*). Moreover, certain DQ2/DQ8-bound deamidated peptides recognized by autoreactive T-cells have been eluted from a limited number of islet autoantigens (*159*). In order to complement these previous efforts, we screened for all T1D-associated deamidated peptides derived from candidate β-cell proteins.

Following Vader *et al*. (*161*), we simulated deamidation of candidate proteins with 85% concordance with previously discovered deamidated peptides (24 out of 28 peptides from Lummel *et al.*) from islet autoantigens (*159*). Overall, 36, 151 and 167 deamidated peptides were predicted to be bound by DRB1, DQ and trans-encoded DQ, respectively. Of them, PepWAS predicted 20, 94 and 124 T1D-associated epitopes bound by DRB1, DQ and trans-encoded DQ, respectively **(Fig. 2C, Tables S5, S6 and S7)**. Of all cis and trans-encoded DQ-bound T1D-associated deamidated epitopes, 60% (56/94) and 51% (63/124) were bound by cis- and trans-encoded DQ2/DQ8 alleles respectively, corresponding with their preference for negatively charged deamidated residues (*159*, *162*). Together with native T1D-associated epitopes, the individual-specific repertoire of T1D-associated deamidated-epitopes accounted for 33.1% of deviance in disease **(Fig. 1 Right)**, notably, which even exceeded 29.6% associated with HLA-DRB1-DQ haplotype.

Molecular mimicry between pathogen-derived peptides and predicted epitopes

Pathogens have been linked to T1D for more than a century (*93*, *179*). They have been postulated as the trigger of autoimmunity because of their molecular homogeneity with the self-epitopes, though the evidences remain circumstantial (*166*, *168*). We tested 16 pathogens **(Table S8)** with potentially association with T1D for the signature of molecular mimicry with T1D-associated epitopes. Of all the tested pathogens, peptides derived from Bovine insulin and heat shock proteins 60 and 70 (Hsp60 and Hsp70) of *P.* tuberculosis showed complete similarity with the full-length T1D-associated epitopes. HSPs are housekeeping proteins that are evolutionary conserved across eukaryotes and prokaryotes. Both Hsp60 and Hsp70 are immunogenic proteins and have been linked to a variety of autoimmune diseases (*180*). In addition, these genes are over-expressed in β-cells relative to other cells in the islets (*181*). We compared the median HLA affinity of pathogen-derived peptides and their matched T1D-assocaiged epitopes. In case of DQ, all mimicking pathogen-derived peptides had higher HLA-binding affinity than corresponding T1D risk-associated self-epitopes, while no unidirectional trend for those mimicking protection-associated epitopes **(Fig. 3)**. On the other side, all DRB1-bound pathogen-derived peptides mimicking self-epitopes showed lower HLA-binding affinity, the magnitude of the difference was very small relative to DQ. We also tested for the molecular mimicry between pathogen-derived peptides and the proteins overexpressed in alpha-cells (N = 572) (*181*) but did not find the tested level of similarity, confirming that

the observed molecular mimicry occurs specifically between candidate pathogens and β-cells peptides.

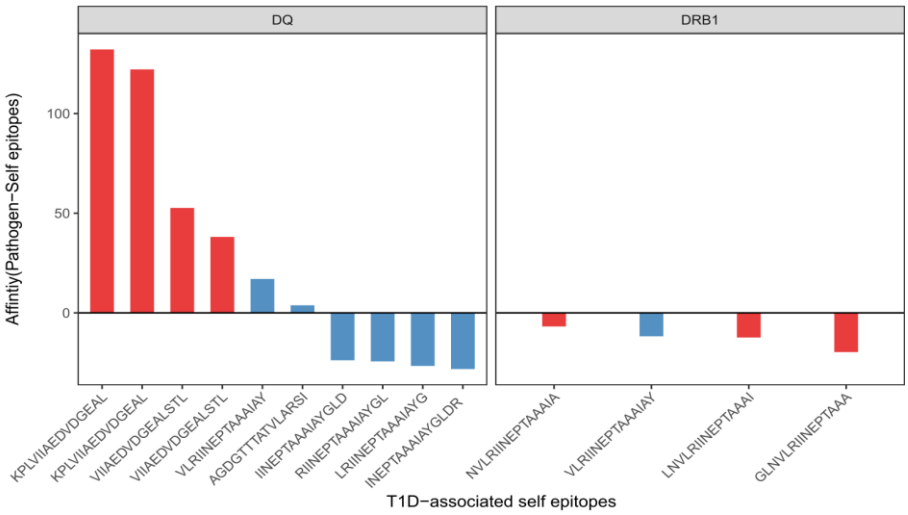


Fig. 3. Molecular mimicry between pathogenic peptides and self-epitopes. The difference in the HLA-binding affinity of T1D-associated epitopes bound by DQ and DRB1 and the mimicking pathogen-derived peptides. The T1D-risk associated epitopes are colored in red and protection-associated ones are in blue. On x-axis, each predicted self-epitope is shown for all pathogen-derived peptides mimicking it.

Discussion

The precise identity and complete repertoire of the HLA-bound T1D-relevant epitopes had been missing. To this goal, we applied our previously established PepWAS approach to certain β-cell proteins that had been proposed to contain the epitopes relevant for the T1D pathogenesis. Of all peptides that could be derived from these proteins, PepWAS predicted a core set of T1D-associated epitopes bound by DRB1, cis and trans-encoded DQ alleles. The difference in the breadth of individual-specific epitope-repertoire defined by DRB1 and DQ (cis and trans) alleles between cases and controls accounted for even more deviance in disease status than the one previously attributed to HLA-DRB1-DQ haplotypes.

The limited knowledge about the non-genetic triggers of islet autoimmunity and the underlying mechanism poses a significant challenge in the understanding of the etiology of T1D. Our results shed light on some of them. One is the deamidation, a kind of post-translation modification, of the proteins in β-cells, which are known to often undergo cellular stress due to their limited number and high turn-over rate (*182*). The deamidated epitopes, which are thought to lack both immune tolerance since they are not encoded in

thymus (*158*), represent a potential trigger of T-cell mediated autoimmunity. Our results showed that the predicted T1D-associated deamidated epitopes were preferentially bound by certain cis- and trans-encoded DQ alleles which conferred high risk of T1D. The individual-specific deamidated epitope-repertoire added a significant proportion to the total accounted deviance in the T1D status. Whereas these results provide an amenable set of novel deamidated epitopes, they also substantiate the previous proposition that once islet autoimmunity has been triggered by deamidated epitopes, it can potentially spread to self-epitopes due to their sequence homology, possibly leading to overt T1D (*183*). Pathogenic infections have also been proposed as the potential trigger of islet autoimmunity (*183*, *184*). The high degree of sequence similarity between predicted T1D-associated epitopes and the peptides derived from heat-shock proteins in *P.* tuberculosis suggests the same. The out-competition of T1D risk-associated self-epitopes by mimicking pathogen-derived peptides for the binding with HLA provides a mechanistic basis of how immune-response to the infections could prime T-cells for self-peptides that might lead to overt T1D in genetically susceptible individuals.

Overall, the predicted T1D-associated epitope repertoires not only provide a functional basis for the robust association between T1D and HLA, but also represent a refined form of the association. In the next step, we have planned the experimental validation of the predicted epitopes.

Material and methods

Samples and genotype data

We analyzed data of 16,086 samples (6,670 T1D cases and 9416 controls) from Type 1 Diabetes Genetics Consortium (T1DGC). The original data and the detailed quality check are described in detail in Hu *et al.* (*66*). Briefly, the samples were collected from 13 regions and genotyped on Immunochip array. HLA alleles (best-guess genotypes at 4-digit resolution) for classical class I (HLA-A, -B, -C) and class II loci (HLA-DRB1, -DQA1, -DQB1, -DPA1, -DPB1) were imputed from genome-wide genotype data using SNP2HLA and a reference panel of 5,225 individuals of European ancestry (*171*). Overall, 35 and 39 alleles for HLA-DRB1 and HLA-DQ were represented in our dataset, respectively **(Table S1)**.

Affinity between HLA and peptides

The sequences of all human proteins were taken from UniProt database (*172*). We used NetMHCIIpan v3.2 (*154*) for calculating the binding affinity between a given HLA class-II molecule and peptide. It is a neural network based computational method that predicts binding affinity of peptides of variable length to any known HLA class II molecule. It also reports the rank of predicted binding affinity of HLA-peptide complexes against predicted affinity of 200,000 random natural peptides. Using this method, we predicted HLA allele-specific binding affinities for all 15mer peptides generated from all human and pathogenic proteins. 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.

Selection of HLA-bound peptides

In order to select for the peptides to be bound by DRB1 and DQ alleles represented in our dataset, we used a regression model that included disease-status as predicted variable and the number of HLA-bound peptide in each individual that lie within a binding affinity range as the predictor variable along with individual's sex and 12 regions of sample collection for population stratification. Using three window sizes (1, 2 and 3) over the rank of predicted binding affinity calculated against random peptides, we sought a range that maximizes (1) the enrichment for the known HLA class-II restricted T1D-associated epitopes combined from Lorenzo *et al.* (*174*) and IEDB database, and (2) the deviance in disease status accounted by individual-specific peptide-repertoire defined by peptides

within focal rank window **(Fig. S1)**. The default rank threshold for HLA-binding peptides in NetMHCIIpan v3.2 was 2 (*154*).

Association with T1D disease

The association of a variable (be it an HLA allele, a peptide or the number of HLA-bound peptides in an individual) with the T1D status was calculated using a logistic regression model corrected for covariates following Hu *et al.* (*66*). Deviance associated with the variable was taken as the difference in the residual deviance of the model with the variable (Equation 1) and null model (Equation 2), which follows a $\chi 2$ distribution with m – 1 degrees of freedom. The covariates included the individual's sex and 12 regions of sample collection for population stratification (adopted from Hu *et al.*).

$$LogOdds(T1D) \sim \sum_{locus}^{DRB1,DQ,transDQ} variable_{locus} + covariates \quad \text{[Equation 1]}$$

$$LogOdds(T1D) \sim covariates \quad \text{[Equation 2]}$$

We used variable's regression coefficient as the measure of its association with disease status. Predicted disease-associated epitopes with one frameshift but with the same binding cores and HLA-alleles were singled out. We selected top 2 risk and 2 protection-associated DRB1, cis- and trans-encoded DQ alleles for representation in peptide-specific association plots **(Fig. 2)**. All analyses were performed in R v3.5.1 and data was visualized using the ggplot2 v2.2.1 package (*138*).

Molecular mimicry

Predicted DRB1 and DQ-bound disease-associated epitopes were blasted against the proteomes of 16 pathogens (*166*, *169*, *185*) **(Table S8),** which were taken from UniProt database (*172*), using Protein Blast from NCBI with default parameters (*186*). The 15mer sequences from pathogens with complete positive match (100%) with disease-associated epitopes were taken and their binding affinity with the DRB1 and DQ alleles were predicted as described above.

Supplementary data

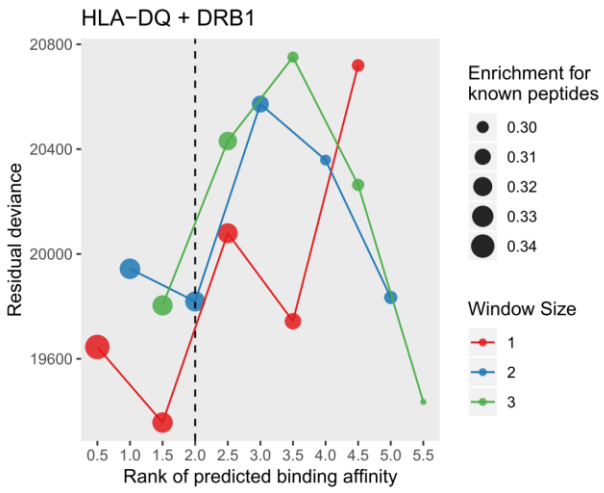


Fig. S1. Selection of HLA-bound peptides. The predicted DRB1 and DQ-bound peptides within each affinity rank window were evaluated for their enrichment for known T1D-associated epitopes and the fitness of the model. We tested affinity rank window of three sizes (1,2 and 3). The size of each dot represents the enrichment for known T1D epitopes combined from Lorenzo *et al.* (*174*) and IEDB database. The Y-axis shows the residual deviance from the model that included disease status as predicted variable and number of individual-specific peptides bound by DRB1 and DQ alleles as predictor variable along with covariates (see methods). The dashed line represents the default rank threshold for binders of the NetMHCIIpan v3.2. Based on the above-mentioned criteria, the affinity rank window of 0-1 was selected.

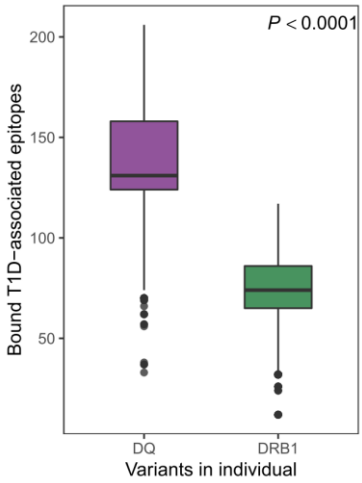


Fig. S2. Individual-specific epitope repertoires of DRB1 and DQ. The individual-specific repertoire of PepWAS predicted T1D-associated epitopes bound by individual's DRB1 and DQ alleles is shown. *P*-value from Wilcoxon rank sum test is shown.

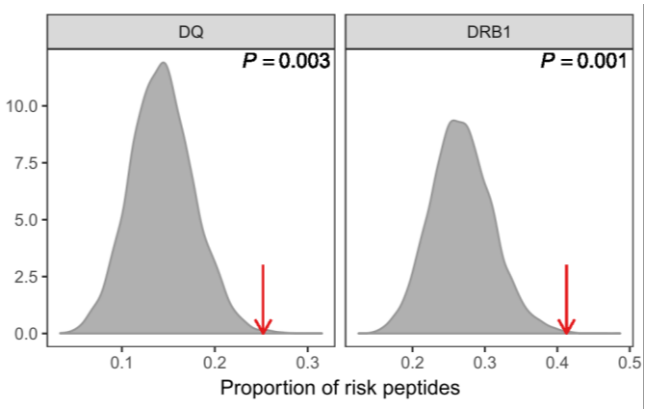


Fig. S3. T1D risk-associated epitopes in random proteins. The distribution of the proportion of predicted T1D-risk associated epitopes in randomly sampled as many human proteins as the number of candidate proteins is shown. This was iterated for 10,000 times. The arrow shows the observed proportion of predicted risk-associated epitopes in candidate proteins. *P*-value is one-sided.

Table S1. HLA-DRB1 and HLA-DQ variants represented in our dataset.

DRB1 (N = 35)	DQ (N = 39)
HLA-DRB1-0101	HLA-DQA10501-DQB10201
HLA-DRB1-0102	HLA-DQA10201-DQB10402
HLA-DRB1-0103	HLA-DQA10501-DQB10301
HLA-DRB1-0301	HLA-DQA10101-DQB10501
HLA-DRB1-0401	HLA-DQA10301-DQB10301
HLA-DRB1-0402	HLA-DQA10301-DQB10303
HLA-DRB1-0403	HLA-DQA10103-DQB10603
HLA-DRB1-0404	HLA-DQA10102-DQB10603
HLA-DRB1-0405	HLA-DQA10401-DQB10402
HLA-DRB1-0406	HLA-DQA10301-DQB10302
HLA-DRB1-0407	HLA-DQA10201-DQB10202
HLA-DRB1-0408	HLA-DQA10102-DQB10602
HLA-DRB1-0701	HLA-DQA10101-DQB10503
HLA-DRB1-0801	HLA-DQA10201-DQB10303
HLA-DRB1-0802	HLA-DQA10102-DQB10504
HLA-DRB1-0803	HLA-DQA10102-DQB10609
HLA-DRB1-0804	HLA-DQA10103-DQB10601
HLA-DRB1-0901	HLA-DQA10102-DQB10604
HLA-DRB1-1001	HLA-DQA10101-DQB10602
HLA-DRB1-1101	HLA-DQA10501-DQB10302
HLA-DRB1-1102	HLA-DQA10102-DQB10502
HLA-DRB1-1103	HLA-DQA10601-DQB10301
HLA-DRB1-1104	HLA-DQA10102-DQB10501
HLA-DRB1-1201	HLA-DQA10102-DQB10201
HLA-DRB1-1301	HLA-DQA10201-DQB10201
HLA-DRB1-1302	HLA-DQA10301-DQB10201
HLA-DRB1-1303	HLA-DQA10301-DQB10304
HLA-DRB1-1401	HLA-DQA10401-DQB10301
HLA-DRB1-1402	HLA-DQA10102-DQB10601
HLA-DRB1-1404	HLA-DQA10501-DQB10202
HLA-DRB1-1501	HLA-DQA10101-DQB10502
HLA-DRB1-1502	HLA-DQA10103-DQB10602
HLA-DRB1-1503	HLA-DQA10301-DQB10402
HLA-DRB1-1601	HLA-DQA10201-DQB10301
HLA-DRB1-1602	HLA-DQA10301-DQB10603
	HLA-DQA10103-DQB10503
	HLA-DQA10301-DQB10602
	HLA-DQA10201-DQB10302
	HLA-DQA10301-DQB10502

Table S2. Predicted T1D-associated epitopes (N = 233) bound by DRB1 variants.

Epitope	Protein	Odds ratio	P-value
GNYKLTASAPGYLAI	CBPE	2.07	9.9e-166
NYKLTASAPGYLAIT	CBPE	0.83	2.9e-06
ETIVNLIHSTRIHIM	CBPE	0.14	0.0e+00
TIVNLIHSTRIHIMP	CBPE	0.14	0.0e+00
IVNLIHSTRIHIMPS	CBPE	0.22	7.2e-292
WGGGLLMSRKHRHKL	DCE1	0.2	2.3e-150
AAWGGGLLMSRKHRH	DCE1	0.18	3.0e-30
LLMSRKHRHKLNGIE	DCE1	0.18	3.0e-30
FKFWLMWKAKGTVGF	DCE1	0.36	2.3e-20
VNFFRMVISNPAATH	DCE2	0.53	5.8e-125
KVNFFRMVISNPAAT	DCE2	0.58	4.9e-100
FFRMVISNPAATHQD	DCE2	0.53	1.7e-124
GDKVNFFRMVISNPA	DCE2	1.27	6.5e-23
FRMVISNPAATHQDI	DCE2	2.53	5.5e-261
FVPFLVSATAGTTVY	DCE2	0.83	2.9e-06
KGFVPFLVSATAGTT	DCE2	0.84	3.4e-06
ISNMYAMMIARFKMF	DCE2	0.41	1.1e-93
SNMYAMMIARFKMFP	DCE2	0.35	1.2e-100
AISNMYAMMIARFKM	DCE2	0.41	1.1e-93
NMYAMMIARFKMFPE	DCE2	0.44	9.6e-61
QKGFVPFLVSATAGT	DCE2	0.53	6.7e-06
VLLEYVTLKKMREII	DCE2	0.39	2.1e-58
SRLSKVAPVIKARMM	DCE2	0.26	3.7e-60
GGGLLMSRKHKWKLS	DCE2	0.2	2.3e-150
MYAMMIARFKMFPEV	DCE2	0.37	2.3e-24
KQKGFVPFLVSATAG	DCE2	0.2	3.3e-09
AWGGGLLMSRKHKWK	DCE2	0.2	3.7e-150
AAWGGGLLMSRKHKW	DCE2	0.18	3.0e-30
FKLWLMWRAKGTTGF	DCE2	0.36	2.3e-20
GGAISNMYAMMIARF	DCE2	0.33	1.0e-16
LPRLIAFTSEHSHFS	DCE2	0.03	6.6e-164
LSFRLLCALTSLTIL	G6PC2	2.37	9.1e-233
TLSFRLLCALTSLTI	G6PC2	2.52	4.3e-261
SFRLLCALTSLTILQ	G6PC2	1.24	3.1e-12
VLSFCKSASIPLTVV	G6PC2	0.44	2.8e-154
YTLSFRLLCALTSLT	G6PC2	1.25	1.2e-12
LSFCKSASIPLTVVA	G6PC2	0.34	1.9e-150
PDWIHIDTTPFAGLV	G6PC2	2.91	2.2e-274
DWIHIDTTPFAGLVR	G6PC2	3.23	0.0e+00
FRLLCALTSLTILQL	G6PC2	1.61	1.4e-23
YYTFLNFMSNVGDPR	G6PC2	5.07	4.2e-40
RAYYTFLNFMSNVGD	G6PC2	2.92	4.5e-32
AYYTFLNFMSNVGDP	G6PC2	5.31	6.7e-41
YRAYYTFLNFMSNVG	G6PC2	2.91	6.8e-32
QLYHFLQIPTHEEHL	G6PC2	5.31	6.7e-41
NYTLSFRLLCALTSL	G6PC2	2.03	9.2e-18
LQLYHFLQIPTHEEH	G6PC2	5.31	6.7e-41
LFYVLSFCKSASIPL	G6PC2	0.34	3.7e-137
CFQFNQTVGTKMIWV	G6PC2	0.34	3.7e-137
NETIVNLIHSTRIHI	CBPE	0.25	1.7e-243
APGYLAITKKVAVPY	CBPE	0.48	4.0e-34
GYLAITKKVAVPYSP	CBPE	0.48	4.0e-34
SAPGYLAITKKVAVP	CBPE	0.47	2.9e-34
ASAPGYLAITKKVAV	CBPE	0.16	6.4e-58
AIFQSLARAYSSFNP	CBPE	0.26	2.6e-53
QFNQTVGTKMIWVAV	G6PC2	0.28	2.7e-153
YSVHMLMKQSGKKSQ	G6PC2	0.17	1.9e-88
PYSVHMLMKQSGKKS	G6PC2	0.16	1.3e-83
NNYTLSFRLLCALTS	G6PC2	0.2	3.3e-09
VWYVMVTAALSHTVC	G6PC2	0.2	3.3e-09
VLNIDLLWSVPIAKK	G6PC2	0.06	1.2e-224
MDFLHRNGVLIIQHL	G6PC2	0.75	6.9e-06
TVVAFIPYSVHMLMK	G6PC2	0.04	1.3e-176
PLTVVAFIPYSVHML	G6PC2	0.03	6.6e-164
VVAFIPYSVHMLMKQ	G6PC2	0.03	6.6e-164
SRVFIATHFPHQVIL	G6PC2	0.03	6.6e-164
RVDFSLAGALNAGFK	GFAP	1.81	1.8e-118
TRVDFSLAGALNAGF	GFAP	0.83	2.9e-06
NVKLALDIEIATYRK	GFAP	3.23	0.0e+00
QDLLLLDVAPLSLGL	HS71A	4.43	0.0e+00
VQDLLLLDVAPLSLG	HS71A	4.68	0.0e+00
NVQDLLLLDVAPLSL	HS71A	5.18	0.0e+00
DLLLLDVAPLSLGLE	HS71A	4.51	0.0e+00
LLLLDVAPLSLGLET	HS71A	2.9	1.8e-273
NVLRIINEPTAAAIA	HS71A	3	5.5e-294
LNVLRIINEPTAAAI	HS71A	2.95	5.9e-287
GLNVLRIINEPTAAA	HS71A	1.74	4.4e-31
VLRIINEPTAAAIAY	HS71A	0.41	5.3e-16
AGGVMTALIKRNSTI	HS71A	0.39	2.0e-58
IDFYTSITRARFEEL	HS71A	0.16	6.4e-58
AGVIAGLNVLRIINE	HS71A	0.16	1.0e-57
GVIAGLNVLRIINEP	HS71A	0.33	1.0e-16
SVALNHLKATPIESH	IAPP	0.84	3.6e-05
ANFLVHSSNNFGAIL	IAPP	2.7	1.6e-222
FLIVLSVALNHLKAT	IAPP	0.39	4.1e-36
NFLVHSSNNFGAILS	IAPP	0.13	7.2e-26
RLANFLVHSSNNFGA	IAPP	0.13	7.2e-26
IVLSVALNHLKATPI	IAPP	0.16	2.8e-57
PYEFTTLKSLQDPMK	ICA69	5.31	6.7e-41
ALWMRLLPLLALLAL	INS	1.29	2.1e-09

LWMRLLPLLALLALW	INS	1.3	1.2e-09
DDAIFQSLARAYSSF	CBPE	0.16	6.4e-58
VNLIHSTRIHIMPSL	CBPE	0.25	2.4e-60
PTVFRQMRPVSRVLA	CH60	0.59	8.8e-87
TVFRQMRPVSRVLAP	CH60	0.59	8.8e-87
LPTVFRQMRPVSRVL	CH60	0.59	1.5e-86
VFRQMRPVSRVLAPH	CH60	0.34	3.9e-171
RLPTVFRQMRPVSRV	CH60	0.47	7.6e-70
MRLLPLLALLALWGP	INS	0.33	1.0e-16
FLFNKYISRPRRVEL	PDX1	0.29	9.1e-40
HIKIWFQNRRMKWKK	PDX1	0.18	3.0e-30
IKIWFQNRRMKWKKE	PDX1	0.34	2.1e-05
LFNKYISRPRRVELA	PDX1	0.18	3.0e-30
EFLFNKYISRPRRVE	PDX1	0.42	1.2e-14
KEFLFNKYISRPRRV	PDX1	0.42	1.2e-14
ELAVMLNLTERHIKI	PDX1	0.05	1.9e-32
LAVMLNLTERHIKIW	PDX1	0.05	5.5e-31
NVKMALDIEIATYRK	PERI	3.23	0.0e+00
PGAFSYSSSSRFSSS	PERI	0.34	3.7e-137
RHHLMAALSAYAAQR	PTPR2	1.25	8.2e-20
DRHHLMAALSAYAAQ	PTPR2	2.41	3.9e-239
VDRHHLMAALSAYAA	PTPR2	1.16	1.7e-06
HHLMAALSAYAAQRP	PTPR2	1.16	1.8e-06
LLLLLLLPPRVLPAA	PTPR2	0.4	7.1e-246
SILTYVAHTSALTYP	PTPR2	1.16	1.1e-09
LLLLLLLLPPRVLPA	PTPR2	0.4	7.1e-246
ILTYVAHTSALTYPP	PTPR2	1.17	3.2e-11
ESILTYVAHTSALTY	PTPR2	0.67	5.9e-63
MDFYRYEVSPVALQR	PTPR2	2.06	2.1e-168
LLLLLLPPRVLPAAP	PTPR2	0.45	2.6e-173
GVDRHHLMAALSAYA	PTPR2	0.8	5.3e-08
LLLLLLLLLPPRVLP	PTPR2	0.45	2.6e-173
AMDFYRYEVSPVALQ	PTPR2	2.06	2.1e-168
HLMAALSAYAAQRPP	PTPR2	0.82	7.0e-07
PLLLLLLLLLPPRVL	PTPR2	0.14	0.0e+00
LLLLLPPRVLPAAPS	PTPR2	0.43	4.5e-18
SPVALQRLRVALQKL	PTPR2	0.45	1.7e-72
PVALQRLRVALQKLS	PTPR2	0.45	4.3e-72
VSPVALQRLRVALQK	PTPR2	0.45	3.3e-72
VALQRLRVALQKLSG	PTPR2	0.62	1.6e-23
ARGYIVTDRDPLRPE	PTPR2	3.23	0.0e+00
GRRLVEDVARLLQVP	PTPR2	3.23	0.0e+00
GTYVLIDMVLNKMAK	PTPR2	3.23	0.0e+00
PAMDFYRYEVSPVAL	PTPR2	2.74	4.1e-246
FYRYEVSPVALQRLR	PTPR2	2.9	1.9e-260
AVTFKVSANVQNVTT	PTPR2	2.9	6.0e-261
VTFKVSANVQNVTTE	PTPR2	2.66	3.4e-217
GPAVTFKVSANVQNV	PTPR2	2.78	1.2e-232
VPAMDFYRYEVSPVA	PTPR2	3.46	0.0e+00
GCVVIVMLTPLAENG	PTPR2	1.79	9.3e-33
TFKVSANVQNVTTED	PTPR2	0.1	2.5e-20
EVSPVALQRLRVALQ	PTPR2	0.27	6.9e-77
KEQFEFALTAVAEEV	PTPR2	0.2	3.3e-09
PAYIATQGPLPATVA	PTPR2	0.2	3.3e-09
LPLLLLLLLLLPPRV	PTPR2	0.33	1.0e-16
MPAYIATQGPLSHTI	PTPRN	1.93	2.0e-139
PAYIATQGPLSHTIA	PTPRN	2.09	1.4e-165
RMPAYIATQGPLSHT	PTPRN	2.07	4.7e-164
AYIATQGPLSHTIAD	PTPRN	0.8	5.3e-08
YGYIVTDQKPLSLAA	PTPRN	2.69	5.2e-243
EYGYIVTDQKPLSLA	PTPRN	3.23	0.0e+00
FRQMRPVSRVLAPHL	CH60	0.38	1.3e-112
EVIVTKDDAMLLKGK	CH60	3.23	0.0e+00
IGIEIIKRTLKIPAM	CH60	0.32	1.8e-124
GIEIIKRTLKIPAMT	CH60	0.32	3.5e-124
KIGIEIIKRTLKIPA	CH60	0.32	1.8e-124
GTYILIDMVLNRMAK	PTPRN	2.66	3.5e-240
TGTYILIDMVLNRMA	PTPRN	3.22	0.0e+00
TYILIDMVLNRMAKG	PTPRN	2.66	3.5e-240
YILIDMVLNRMAKGV	PTPRN	2.66	3.5e-240
RTGTYILIDMVLNRM	PTPRN	3.23	0.0e+00
SPIIEHDPRMPAYIA	PTPRN	3.23	0.0e+00
LLCLLLLSSRPGGCS	PTPRN	1.94	7.9e-45
QKPLSLAAGVKLLEI	PTPRN	0.34	3.7e-137
EAPALFSRTASKGIF	PTPRN	0.28	2.7e-153
KDQFEFALTAVAEEV	PTPRN	0.2	3.3e-09
PRMPAYIATQGPLSH	PTPRN	0.2	3.3e-09
ILIDMVLNRMAKGVK	PTPRN	0.18	1.4e-102
LIDMVLNRMAKGVKE	PTPRN	0.25	2.4e-60
LLGIIKGMKVELSTV	RT31	0.43	1.7e-214
DLLGIIKGMKVELST	RT31	0.38	2.5e-215
KDLLGIIKGMKVELS	RT31	0.35	4.3e-261
VSTFLPLRPLSRHPL	RT31	0.61	2.6e-24
PRVSTFLPLRPLSRH	RT31	0.3	9.8e-63
STFLPLRPLSRHPLS	RT31	0.52	1.5e-33
RRPLKSLEATLGRLR	RT31	1.57	2.5e-23
GIIKGMKVELSTVNV	RT31	1.56	6.8e-23
TFLPLRPLSRHPLSS	RT31	0.52	1.5e-33
KKDLLGIIKGMKVEL	RT31	0.1	0.0e+00
KRRPLKSLEATLGRL	RT31	1.6	1.6e-24
FSNIISDMKVARSAT	RT31	2.69	9.3e-244
SFSNIISDMKVARSA	RT31	3.23	0.0e+00
NIISDMKVARSATAR	RT31	3.22	0.0e+00
AAAIMLLTVRHGTVR	RT31	0.67	1.4e-34

IIKGMKVELSTVNVR	RT31	1.79	9.3e-33
AAIMLLTVRHGTVRY	RT31	0.55	5.1e-77
SAAAIMLLTVRHGTV	RT31	1.62	7.2e-30
AIMLLTVRHGTVRYR	RT31	0.33	7.9e-106
IMLLTVRHGTVRYRS	RT31	0.26	2.3e-60
RHGTVRYRSSALLAR	RT31	0.03	6.2e-164
FSILLMEGVPKSLNY	ZNT8	1.69	3.5e-33
DFSILLMEGVPKSLN	ZNT8	1.7	1.1e-33
KDFSILLMEGVPKSL	ZNT8	1.7	1.1e-33
RTYLVNDKAAKMYAF	ZNT8	2.99	6.2e-285
KELILAVDGVLSVHS	ZNT8	3.16	9.3e-306
ELILAVDGVLSVHSL	ZNT8	3.23	0.0e+00
VKELILAVDGVLSVH	ZNT8	3.17	1.3e-306
HIWSLTMNQVILSAH	ZNT8	0.23	5.5e-24
SILLMEGVPKSLNYS	ZNT8	1.74	4.4e-31
LHIWSLTMNQVILSA	ZNT8	0.13	3.3e-62
RLTFGWHRAEILGAL	ZNT8	0.34	3.7e-137
VDGVLSVHSLHIWSL	ZNT8	0.14	0.0e+00
SLHIWSLTMNQVILS	ZNT8	0.09	6.8e-05
GVLSVHSLHIWSLTM	ZNT8	0.33	1.0e-16
RQMRPVSRVLAPHLT	CH60	0.3	1.6e-134
IEIIKRTLKIPAMTI	CH60	0.36	5.3e-93
QKIGIEIIKRTLKIP	CH60	0.23	5.5e-90
FETLRGDERILSILR	CMGA	3.23	0.0e+00
HKEVQANASVRAAFV	ZNT8	0.75	6.9e-06
ISALIIYFKPEYKIA	ZNT8	0.03	6.6e-164
AVDGVLSVHSLHIWS	ZNT8	0.03	6.2e-164
ERILSILRHQNLLKE	CMGA	0.27	2.3e-186
GDERILSILRHQNLL	CMGA	0.31	4.3e-116
SSMKLSFRARAYGFR	CMGA	0.32	3.5e-124
RGDERILSILRHQNL	CMGA	0.78	1.7e-04
DSSMKLSFRARAYGF	CMGA	0.32	1.8e-124
NRDSSMKLSFRARAY	CMGA	0.36	8.5e-98
ILSILRHQNLLKELQ	CMGA	0.33	1.0e-16
ANFFRMVISNPAATQ	DCE1	0.53	5.8e-125
KANFFRMVISNPAAT	DCE1	0.58	4.9e-100
DKANFFRMVISNPAA	DCE1	0.58	1.2e-99
FFRMVISNPAATQSD	DCE1	0.53	1.7e-124
GDKANFFRMVISNPA	DCE1	1.47	7.6e-56
FRMVISNPAATQSDI	DCE1	2.53	5.5e-261
ISNMYSIMAARYKYF	DCE1	0.34	0.0e+00
AISNMYSIMAARYKY	DCE1	0.37	8.4e-305
SNMYSIMAARYKYFP	DCE1	0.26	0.0e+00
GAISNMYSIMAARYK	DCE1	0.49	5.0e-161
NMYSIMAARYKYFPE	DCE1	0.53	6.8e-59
SLEQILVDCRDTLKY	DCE1	3.23	0.0e+00
KLVLFTSEQSHYSIK	DCE1	3.18	2.6e-275
GGAISNMYSIMAARY	DCE1	0.39	9.2e-161
PGGAISNMYSIMAAR	DCE1	1.79	9.3e-33

Table S3. Predicted T1D-associated epitopes (N = 322) bound by DQ variants.

Epitope	Protein	Odds ratio	P-value
MAGRGGSALLALCGA	CBPE	0.18	3.9e-130
TKAVIHWIMDIPFVL	CBPE	0.67	3.5e-33
ETKAVIHWIMDIPFV	CBPE	0.63	1.0e-35
AVIHWIMDIPFVLSA	CBPE	0.67	3.5e-33
KANFFRMVISNPAAT	DCE1	0.24	1.1e-60
GIFSPGGAISNMYSI	DCE1	0.33	1.3e-163
GDGIFSPGGAISNMY	DCE1	0.43	3.9e-130
IFSPGGAISNMYSIM	DCE1	0.18	3.9e-130
YSIKKAGAALGFGTD	DCE1	0.18	3.9e-130
ANTNMFTYEIAPVFV	DCE1	0.79	5.1e-36
EQTVQFLLEVVDILL	DCE1	0.75	2.8e-13
FYVNATAGTTVYGAF	DCE1	0.19	0.0e+00
YVNATAGTTVYGAFD	DCE1	0.19	0.0e+00
PFYVNATAGTTVYGA	DCE1	0.19	0.0e+00
VNATAGTTVYGAFDP	DCE1	0.11	0.0e+00
NATAGTTVYGAFDPI	DCE1	0.14	7.1e-300
PGGAISNMYSIMAAR	DCE1	0.08	3.0e-287
NFFRMVISNPAATQS	DCE1	0.24	8.3e-59
ATAGTTVYGAFDPIQ	DCE1	0.03	7.9e-164
PAATQSDIDFLIEEI	DCE1	0.42	7.9e-70
TAGTTVYGAFDPIQE	DCE1	0.03	9.0e-157
EHTNVCFWYIPQSLR	DCE1	0.06	5.2e-33
TANTNMFTYEIAPVF	DCE1	0.23	1.2e-32
THQDIDFLIEEIERL	DCE2	1.96	4.2e-139
NMFTYEIAPVFVLLE	DCE2	1.79	3.3e-115
HQDIDFLIEEIERLG	DCE2	1.96	2.9e-139
TNMFTYEIAPVFVLL	DCE2	1.85	1.1e-125
MFTYEIAPVFVLLEY	DCE2	2.8	5.4e-284
QDIDFLIEEIERLGQ	DCE2	1.96	2.9e-139
KVNFFRMVISNPAAT	DCE2	0.24	1.1e-60
AISNMYAMMIARFKM	DCE2	0.03	2.0e-163
GIFSPGGAISNMYAM	DCE2	0.33	1.3e-163
RPTLAFLQDVMNILL	DCE2	0.76	2.8e-13
ERPTLAFLQDVMNIL	DCE2	0.63	4.3e-36
PTLAFLQDVMNILLQ	DCE2	0.75	2.8e-13
TLAFLQDVMNILLQY	DCE2	0.75	2.8e-13
FLVSATAGTTVYGAF	DCE2	0.19	0.0e+00

PFLVSATAGTTVYGA	DCE2	0.19	0.0e+00
LVSATAGTTVYGAFD	DCE2	0.19	0.0e+00
VSATAGTTVYGAFDP	DCE2	0.19	0.0e+00
VPFLVSATAGTTVYG	DCE2	0.36	2.1e-109
SATAGTTVYGAFDPL	DCE2	0.14	7.1e-300
FVPFLVSATAGTTVY	DCE2	0.29	8.8e-61
VNFFRMVISNPAATH	DCE2	0.24	1.0e-59
NFFRMVISNPAATHQ	DCE2	0.24	8.3e-59
ATAGTTVYGAFDPLL	DCE2	0.03	7.4e-164
VNYAFLHATDLLPAC	DCE2	0.17	2.0e-235
ATHQDIDFLIEEIER	DCE2	0.42	7.9e-70
PHKMMGVPLQCSALL	DCE2	0.03	3.5e-157
KMMGVPLQCSALLVR	DCE2	0.03	7.0e-157
WNPHKMMGVPLQCSA	DCE2	0.03	9.0e-157
TAGTTVYGAFDPLLA	DCE2	0.03	7.0e-157
GGAISNMYAMMIARF	DCE2	0.17	6.9e-248
QHTNVCFWYIPPSLR	DCE2	0.06	5.2e-33
GTTVYGAFDPLLAVA	DCE2	0.05	1.7e-32
GFVPFLVSATAGTTV	DCE2	0.05	1.4e-57
LRVLNIDLLWSVPIA	G6PC2	0.78	2.5e-12
LFGLGFAINSEMFLL	G6PC2	2.49	3.1e-258
VLFGLGFAINSEMFL	G6PC2	2.62	2.4e-280
LGVIGGMLVAEAFEH	G6PC2	1.78	7.2e-105
ILGVIGGMLVAEAFE	G6PC2	1.72	3.1e-98
GVIGGMLVAEAFEHT	G6PC2	1.88	1.4e-121
VILGVIGGMLVAEAF	G6PC2	0.23	8.6e-234
VIGGMLVAEAFEHTP	G6PC2	0.16	3.8e-227
GVLFGLGFAINSEMF	G6PC2	4.2	0.0e+00
FGLGFAINSEMFLLS	G6PC2	0.71	9.4e-13
CVWYVMVTAALSHTV	G6PC2	0.05	1.4e-57
PETKAVIHWIMDIPF	CBPE	0.76	1.2e-12
VIHWIMDIPFVLSAN	CBPE	0.76	4.6e-13
YSPAAGVDFELESFS	CBPE	3.35	2.5e-05
KPLVIIAEDVDGEAL	CH60	1.96	2.9e-139
RVDFSLAGALNAGFK	GFAP	0.6	9.4e-33
TRVDFSLAGALNAGF	GFAP	0.61	1.4e-29
EAASYQEALARLEEE	GFAP	6.14	0.0e+00
EAVAYGAAVQAAILM	HS71A	0.69	1.9e-50
DEAVAYGAAVQAAIL	HS71A	0.69	5.5e-50
PDEAVAYGAAVQAAI	HS71A	0.7	8.3e-48
NPDEAVAYGAAVQAA	HS71A	0.73	3.3e-38
AYGAAVQAAILMGDK	HS71A	0.17	0.0e+00
LGLETAGGVMTALIK	HS71A	0.17	0.0e+00
SLGLETAGGVMTALI	HS71A	0.17	0.0e+00
GLETAGGVMTALIKR	HS71A	0.17	0.0e+00
LETAGGVMTALIKRN	HS71A	0.27	3.2e-245
LSLGLETAGGVMTAL	HS71A	0.3	4.2e-231
IINEPTAAAIAYGLD	HS71A	0.7	8.3e-48
RIINEPTAAAIAYGL	HS71A	0.7	8.3e-48
INEPTAAAIAYGLDR	HS71A	0.7	8.3e-48
NEPTAAAIAYGLDRT	HS71A	0.72	9.7e-39
INPDEAVAYGAAVQA	HS71A	0.38	4.2e-150
LRIINEPTAAAIAYG	HS71A	0.18	0.0e+00
NVQDLLLLDVAPLSL	HS71A	0.74	9.2e-19
VLRIINEPTAAAIAY	HS71A	0.2	0.0e+00
PLSLGLETAGGVMTA	HS71A	0.52	2.4e-41
PTAAAIAYGLDRTGK	HS71A	0.13	1.5e-300
YPVTNAVITVPAYFN	HS71A	0.14	2.6e-136
GYPVTNAVITVPAYF	HS71A	0.14	1.9e-135
PVTNAVITVPAYFND	HS71A	0.15	4.8e-131
SSSTQASLEIDSLFE	HS71A	0.39	1.7e-76
STQASLEIDSLFEGI	HS71A	0.39	1.7e-76
TAAAIAYGLDRTGKG	HS71A	0.03	9.0e-157
YGAAVQAAILMGDKS	HS71A	0.03	9.0e-157
CQEVISWLDANTLAE	HS71A	0.05	1.7e-32
AKRTLSSSTQASLEI	HS71A	0.05	1.2e-57
RAKRTLSSSTQASLE	HS71A	0.05	1.4e-57
LGYPVTNAVITVPAY	HS71A	0.05	1.2e-57
VQDLLLLDVAPLSLG	HS71A	0.74	2.1e-06
HSSNNFGAILSSTNV	IAPP	0.03	5.9e-157
LTAWFSLFADLDPLS	ICA69	1.47	7.5e-54
TAWFSLFADLDPLSN	ICA69	1.46	2.2e-52
DLTAWFSLFADLDPL	ICA69	5.71	0.0e+00
AWFSLFADLDPLSNP	ICA69	1.46	2.2e-52
WFSLFADLDPLSNPD	ICA69	1.46	2.2e-52
SDLTAWFSLFADLDP	ICA69	1.48	7.6e-56
DEHVVASDADLDAKL	ICA69	1.96	1.1e-138
EHVVASDADLDAKLE	ICA69	3.27	0.0e+00
ASDLTAWFSLFADLD	ICA69	1.48	7.6e-56
LSEIFNASSLEEGEF	ICA69	5.11	0.0e+00
KEDEHVVASDADLDA	ICA69	3.24	0.0e+00
EKTSHTMAAIHESFK	ICA69	0.11	7.2e-227
WEKTSHTMAAIHESF	ICA69	0.22	1.4e-213
KTSHTMAAIHESFKG	ICA69	0.03	1.2e-158
AASDLTAWFSLFADL	ICA69	0.76	4.8e-13
SEIFNASSLEEGEFS	ICA69	5.4	0.0e+00
TSHTMAAIHESFKGY	ICA69	0.03	6.6e-157
CGSHLVEALYLVCGE	INS	6.14	0.0e+00
LALLALWGPDPAAAF	INS	0.05	1.0e-32
LLALLALWGPDPAAA	INS	0.05	1.7e-32
PPGGAVPPAAPVAAR	PDX1	0.33	3.6e-165
PGGAVPPAAPVAARE	PDX1	0.33	3.6e-165
PPPPGGAVPPAAPVA	PDX1	0.33	1.3e-163
WKGQWAGGAYAAEPE	PDX1	0.34	1.3e-161

PPPPPGGAVPPAAPV	PDX1	0.18	3.9e-130
RGGGTAVGGGGVAEP	PDX1	0.18	3.9e-130
GGAVPPAAPVAAREG	PDX1	0.18	3.9e-130
KGQWAGGAYAAEPEE	PDX1	0.34	1.3e-161
HAWKGQWAGGAYAAE	PDX1	0.34	2.1e-161
GQWAGGAYAAEPEEN	PDX1	6.15	0.0e+00
LERKIESLMDEIEFL	PERI	1.94	1.7e-136
ELERKIESLMDEIEF	PERI	1.96	1.1e-138
RRGVMLAVDAVIAEL	CH60	1.77	9.9e-108
RGVMLAVDAVIAELK	CH60	2.63	8.4e-251
HRKPLVIIAEDVDGE	CH60	3.24	0.0e+00
VIIAEDVDGEALSTL	CH60	3.24	0.0e+00
GTTTATVLARSIAKE	CH60	0.07	0.0e+00
DGTTTATVLARSIAK	CH60	0.07	0.0e+00
TTTATVLARSIAKEG	CH60	0.03	8.7e-225
GDGTTTATVLARSIA	CH60	0.09	8.5e-305
RLDFSMAEALNQEFL	PERI	5.18	0.0e+00
LDFSMAEALNQEFLA	PERI	5.18	0.0e+00
ERLDFSMAEALNQEF	PERI	5.18	0.0e+00
SRLLGSASPSSSVRL	PERI	0.26	3.0e-247
RLLGSASPSSSVRLG	PERI	0.33	3.5e-167
LLGSASPSSSVRLGS	PERI	0.18	2.4e-133
FALEAGGYQAGAARL	PERI	0.34	2.1e-161
ALEAGGYQAGAARLE	PERI	0.34	2.1e-161
FRSPRAGAGALLRLP	PERI	0.18	3.9e-130
QFALEAGGYQAGAAR	PERI	0.53	2.1e-38
SERLDFSMAEALNQE	PERI	5.39	0.0e+00
KEQFEFALTAVAEEV	PTPR2	5.29	0.0e+00
EQFEFALTAVAEEVN	PTPR2	5.3	0.0e+00
QFEFALTAVAEEVNA	PTPR2	2.3	0.0e+00
FEFALTAVAEEVNAI	PTPR2	2.3	0.0e+00
EFALTAVAEEVNAIL	PTPR2	2.3	0.0e+00
TKEQFEFALTAVAEE	PTPR2	5.34	0.0e+00
FALTAVAEEVNAILK	PTPR2	2.28	0.0e+00
DDYTQYVMDQELADL	PTPR2	1.96	2.9e-139
LLQVPSSAFADVEVL	PTPR2	5.19	0.0e+00
RHHLMAALSAYAAQR	PTPR2	0.19	0.0e+00
ESILTYVAHTSALTY	PTPR2	0.15	2.8e-132
HHLMAALSAYAAQRP	PTPR2	0.22	0.0e+00
SILTYVAHTSALTYP	PTPR2	0.06	9.7e-60
FSESILTYVAHTSAL	PTPR2	0.24	7.9e-60
DRHHLMAALSAYAAQ	PTPR2	0.22	0.0e+00
VDRHHLMAALSAYAA	PTPR2	0.22	0.0e+00
HLMAALSAYAAQRPP	PTPR2	0.18	0.0e+00
LMAALSAYAAQRPPA	PTPR2	0.18	0.0e+00
GVDRHHLMAALSAYA	PTPR2	0.11	0.0e+00
MAALSAYAAQRPPAP	PTPR2	0.08	4.1e-294
RYSREGGAALANALR	PTPR2	0.37	2.8e-152
RRYSREGGAALANAL	PTPR2	0.37	4.0e-152
YSREGGAALANALRR	PTPR2	0.37	2.8e-152
SREGGAALANALRRH	PTPR2	0.33	3.6e-165
ERRYSREGGAALANA	PTPR2	0.34	1.3e-161
GVARGSPGRAALGES	PTPR2	0.18	3.9e-130
TGHMILSYMEDHLKN	PTPR2	0.63	1.0e-35
STGHMILSYMEDHLK	PTPR2	0.63	8.5e-36
ISTGHMILSYMEDHL	PTPR2	0.76	1.2e-12
HMILSYMEDHLKNKN	PTPR2	0.76	1.2e-12
QTKEQFEFALTAVAE	PTPR2	5.18	0.0e+00
LQVPSSAFADVEVLG	PTPR2	2.76	8.2e-284
VPSSAFADVEVLGPA	PTPR2	5.4	0.0e+00
RLLQVPSSAFADVEV	PTPR2	6.14	0.0e+00
QFHFLSWYDRGVPSS	PTPR2	0.06	4.5e-33
TQFHFLSWYDRGVPS	PTPR2	0.24	1.0e-37
TKFIALTLVSLACIL	PTPR2	0.75	9.4e-06
FIALTLVSLACILGV	PTPR2	0.23	1.3e-14
KDQFEFALTAVAEEV	PTPRN	5.29	0.0e+00
DQFEFALTAVAEEVN	PTPRN	5.3	0.0e+00
SKDQFEFALTAVAEE	PTPRN	5.34	0.0e+00
VGKGGAGASSSLSPL	PTPRN	0.33	3.5e-167
PVGKGGAGASSSLSP	PTPRN	0.33	1.6e-163
PPVGKGGAGASSSLS	PTPRN	0.33	1.3e-163
QPPVGKGGAGASSSL	PTPRN	0.18	3.9e-130
GKGGAGASSSLSPLQ	PTPRN	0.18	4.7e-130
QDIPTGSAPAAQHRL	PTPRN	0.34	2.4e-161
LQDIPTGSAPAAQHR	PTPRN	0.34	2.1e-161
DIPTGSAPAAQHRLP	PTPRN	0.34	2.1e-161
KGGAGASSSLSPLQA	PTPRN	0.18	3.9e-130
IPTGSAPAAQHRLPQ	PTPRN	0.33	1.3e-163
TGHMILAYMEDHLRN	PTPRN	0.63	8.5e-36
ISTGHMILAYMEDHL	PTPRN	0.63	6.4e-36
VGQREEAAAVLPQTA	PTPRN	5.17	0.0e+00
RSKDQFEFALTAVAE	PTPRN	5.18	0.0e+00
GSFINISVVGPALTF	PTPRN	0.15	5.8e-131
VALAGVAGLLVALAV	PTPRN	0.21	3.4e-19
ALAGVAGLLVALAVA	PTPRN	0.04	2.4e-197
KDMAIATGGAVFGEE	CH60	0.33	3.6e-165
LKDMAIATGGAVFGE	CH60	0.33	3.6e-165
QLKDMAIATGGAVFG	CH60	0.33	2.3e-167
AIATGGAVFGEEGLT	CH60	0.19	2.6e-128
KISSIQSIVPALEIA	CH60	3.19	0.0e+00
KKISSIQSIVPALEI	CH60	3.49	0.0e+00
ISSIQSIVPALEIAN	CH60	5.21	0.0e+00
SSIQSIVPALEIANA	CH60	5.22	0.0e+00
HVHMSSGSFINISVV	PTPRN	0.03	7.0e-157

NMDISTGHMILAYME	PTPRN	0.03	7.0e-157
FINISVVGPALTFRI	PTPRN	0.05	1.4e-57
LAGVAGLLVALAVAL	PTPRN	0.23	1.3e-14
LVALAGVAGLLVALA	PTPRN	0.23	7.0e-15
MKVARSATARVRSRP	RT31	0.23	6.0e-07
DMKVARSATARVRSR	RT31	0.23	6.0e-07
ISDMKVARSATARVR	RT31	0.21	6.7e-08
SPELVAAASAVADSL	RT31	0.69	5.7e-49
ELVAAASAVADSLPF	RT31	0.69	4.3e-49
PELVAAASAVADSLP	RT31	0.69	5.7e-49
LSPELVAAASAVADS	RT31	0.73	4.1e-38
LVAAASAVADSLPFD	RT31	0.82	1.4e-15
PLSPELVAAASAVAD	RT31	1.32	2.1e-29
PLSSGSPETSAAAIM	RT31	0.33	3.5e-167
LSSGSPETSAAAIML	RT31	0.37	4.0e-152
EPLSPELVAAASAVA	RT31	0.43	9.9e-94
VAAASAVADSLPFDK	RT31	2.47	1.6e-232
SGSPETSAAAIMLLT	RT31	0.2	0.0e+00
GSPETSAAAIMLLTV	RT31	0.2	0.0e+00
SSGSPETSAAAIMLL	RT31	0.2	0.0e+00
SPETSAAAIMLLTVR	RT31	0.1	0.0e+00
PETSAAAIMLLTVRH	RT31	0.03	3.6e-157
ETSAAAIMLLTVRHG	RT31	0.03	7.0e-157
NIISDMKVARSATAR	RT31	0.2	5.6e-07
AAKMYAFTLESVELQ	ZNT8	5.7	0.0e+00
AKMYAFTLESVELQQ	ZNT8	5.7	0.0e+00
KMYAFTLESVELQQK	ZNT8	1.94	1.1e-139
KAAKMYAFTLESVEL	ZNT8	5.56	0.0e+00
MYAFTLESVELQQKP	ZNT8	1.94	4.0e-135
LSKSFTMHSLTIQME	ZNT8	5.32	0.0e+00
VQANASVRAAFVHAL	ZNT8	0.16	0.0e+00
QANASVRAAFVHALG	ZNT8	0.16	0.0e+00
NQVILSAHVATAASR	ZNT8	0.16	0.0e+00
SKSFTMHSLTIQMES	ZNT8	1.5	2.2e-55
ANASVRAAFVHALGD	ZNT8	0.16	0.0e+00
VILSAHVATAASRDS	ZNT8	0.16	0.0e+00
NASVRAAFVHALGDL	ZNT8	0.19	0.0e+00
KSFTMHSLTIQMESP	ZNT8	1.86	1.5e-116
EVQANASVRAAFVHA	ZNT8	0.19	0.0e+00
TMNQVILSAHVATAA	ZNT8	0.18	0.0e+00
ILSAHVATAASRDSQ	ZNT8	0.27	5.4e-308
KEVQANASVRAAFVH	ZNT8	0.19	0.0e+00
ASVRAAFVHALGDLF	ZNT8	0.16	1.4e-298
SFTMHSLTIQMESPV	ZNT8	1.86	7.1e-117
HKEVQANASVRAAFV	ZNT8	0.19	0.0e+00
EVVGGHIAGSLAVVT	ZNT8	0.3	9.3e-235
AEVVGGHIAGSLAVV	ZNT8	0.33	9.0e-171
IAEVVGGHIAGSLAV	ZNT8	0.33	3.9e-165
VVGGHIAGSLAVVTD	ZNT8	0.3	1.1e-234
MIAEVVGGHIAGSLA	ZNT8	0.33	5.6e-165
VGGHIAGSLAVVTDA	ZNT8	0.17	0.0e+00
GGHIAGSLAVVTDAA	ZNT8	0.16	0.0e+00
FMIAEVVGGHIAGSL	ZNT8	0.18	3.9e-130
MIIVSSCAVAANIVL	ZNT8	0.18	0.0e+00
IIVSSCAVAANIVLT	ZNT8	0.18	0.0e+00
GHIAGSLAVVTDAAH	ZNT8	0.17	0.0e+00
TVMIIVSSCAVAANI	ZNT8	0.18	0.0e+00
IVSSCAVAANIVLTV	ZNT8	0.22	0.0e+00
LSAHVATAASRDSQV	ZNT8	0.42	4.0e-97
ATVMIIVSSCAVAAN	ZNT8	0.35	3.3e-127
HIAGSLAVVTDAAHL	ZNT8	0.12	3.0e-232
LTMNQVILSAHVATA	ZNT8	0.15	1.1e-136
QATVMIIVSSCAVAA	ZNT8	0.31	5.2e-131
FTMHSLTIQMESPVD	ZNT8	1.87	1.6e-118
IAGSLAVVTDAAHLL	ZNT8	0.13	1.5e-05
IWSLTMNQVILSAHV	ZNT8	0.7	1.3e-08
LHIWSLTMNQVILSA	ZNT8	0.74	2.6e-06
SLHIWSLTMNQVILS	ZNT8	0.74	2.6e-06
HIWSLTMNQVILSAH	ZNT8	0.74	2.6e-06
DLFQSISVLISALII	ZNT8	0.75	9.4e-06
LFQSISVLISALIIY	ZNT8	0.75	9.4e-06
SSCAVAANIVLTVVL	ZNT8	0.75	9.4e-06
FQSISVLISALIIYF	ZNT8	0.23	1.3e-14
HSLHIWSLTMNQVIL	ZNT8	0.74	3.5e-06
LGDLFQSISVLISAL	ZNT8	0.23	1.3e-14
IRRGVMLAVDAVIAE	CH60	0.7	1.0e-08
AGDGTTTATVLARSI	CH60	0.03	1.5e-223
GVMLAVDAVIAELKK	CH60	0.74	2.1e-06
DQELESLSAIEAELE	CMGA	5.28	0.0e+00
LESLSAIEAELEKVA	CMGA	5.18	0.0e+00
EDQELESLSAIEAEL	CMGA	5.28	0.0e+00
ESLSAIEAELEKVAH	CMGA	5.18	0.0e+00
PEDQELESLSAIEAE	CMGA	5.19	0.0e+00
SLSAIEAELEKVAHQ	CMGA	0.39	1.7e-76
TQSDIDFLIEEIERL	DCE1	1.96	4.2e-139
NMFTYEIAPVFVLME	DCE1	1.79	3.3e-115
QSDIDFLIEEIERLG	DCE1	1.96	2.9e-139
TNMFTYEIAPVFVLM	DCE1	1.93	2.4e-142
NTNMFTYEIAPVFVL	DCE1	1.39	7.6e-141
MFTYEIAPVFVLMEQ	DCE1	3.08	1.3e-307
AATQSDIDFLIEEIE	DCE1	1.96	2.9e-139
SDIDFLIEEIERLGQ	DCE1	1.96	2.9e-139
ISNMYSIMAARYKYF	DCE1	0.08	6.9e-296
AISNMYSIMAARYKY	DCE1	0.08	4.9e-296

GAISNMYSIMAARYK	DCE1	0.08	4.9e-296
SNMYSIMAARYKYFP	DCE1	0.08	1.9e-289
GGAISNMYSIMAARY	DCE1	0.08	4.9e-296

Table S4. Predicted T1D-associated epitopes (N = 500) bound by trans-encoded DQ variants.

Epitope	Protein	Odds ratio	P-value
NYKLTASAPGYLAIT	CBPE	0.15	3.7e-216
YKLTASAPGYLAITK	CBPE	0.04	9.3e-42
GNYKLTASAPGYLAI	CBPE	0.13	2.2e-102
VNLIHSTRIHIMPSL	CBPE	0.28	1.0e-89
NLIHSTRIHIMPSLN	CBPE	0.45	3.9e-30
IHSTRIHIMPSLNPD	CBPE	0.13	8.3e-09
KAVIHWIMDIPFVLS	CBPE	0.62	4.0e-66
TKAVIHWIMDIPFVL	CBPE	0.62	1.0e-66
GGSALLALCGALAAC	CBPE	0.13	4.8e-36
FFRMVISNPAATQSD	DCE1	0.05	1.8e-17
PGGAISNMYSIMAAR	DCE1	0.23	1.8e-123
ATAGTTVYGAFDPIQ	DCE1	0.02	2.7e-36
SPGGAISNMYSIMAA	DCE1	0.01	4.9e-15
IFKFWLMWKAKGTVG	DCE1	1.53	4.9e-06
DIFKFWLMWKAKGTV	DCE1	1.68	5.8e-08
NMYSIMAARYKYFPE	DCE1	1.53	4.9e-06
KFWLMWKAKGTVGFE	DCE1	1.69	4.4e-08
FWLMWKAKGTVGFEN	DCE1	1.85	8.2e-09
EQTVQFLLEVVDILL	DCE1	0.5	7.9e-49
TVQFLLEVVDILLNY	DCE1	0.49	6.8e-42
GIFSPGGAISNMYSI	DCE1	0.56	3.8e-36
DGIFSPGGAISNMYS	DCE1	0.56	2.4e-35
GDGIFSPGGAISNMY	DCE1	0.58	6.7e-41
IFSPGGAISNMYSIM	DCE1	0.33	7.0e-42
YSIKKAGAALGFGTD	DCE1	0.33	1.3e-42
MASSTPSSSATSSNA	DCE1	0.26	7.3e-33
TAGTTVYGAFDPIQE	DCE1	0.05	2.2e-14
PEHTNVCFWYIPQSL	DCE1	0.27	2.8e-05
FLVSATAGTTVYGAF	DCE2	0.53	1.1e-131
LVSATAGTTVYGAFD	DCE2	0.54	6.3e-124
PFLVSATAGTTVYGA	DCE2	0.58	6.5e-103
VSATAGTTVYGAFDP	DCE2	0.56	1.7e-113
FVPFLVSATAGTTVY	DCE2	1.31	9.5e-19
THQDIDFLIEEIERL	DCE2	2.37	1.9e-223
HQDIDFLIEEIERLG	DCE2	2.36	2.1e-220
ATHQDIDFLIEEIER	DCE2	3.51	6.9e-252
NMFTYEIAPVFVLLE	DCE2	1.63	5.3e-99
QDIDFLIEEIERLGQ	DCE2	2.48	2.2e-241
VPFLVSATAGTTVYG	DCE2	0.27	9.5e-184
KVNFFRMVISNPAAT	DCE2	0.69	7.7e-22
VNFFRMVISNPAATH	DCE2	0.6	3.1e-38
NFFRMVISNPAATHQ	DCE2	0.03	6.9e-23
NVCFWYIPPSLRTLE	DCE2	0.22	2.3e-24
HTNVCFWYIPPSLRT	DCE2	0.18	1.1e-55
GTTVYGAFDPLLAVA	DCE2	0.25	6.2e-32
FFRMVISNPAATHQD	DCE2	0.04	2.8e-16
ATAGTTVYGAFDPLL	DCE2	0.06	4.9e-130
GGAISNMYAMMIARF	DCE2	0.32	8.4e-136
GAISNMYAMMIARFK	DCE2	0.29	2.1e-136
PHKMMGVPLQCSALL	DCE2	0.02	5.8e-37
AISNMYAMMIARFKM	DCE2	0.41	4.1e-99
VNYAFLHATDLLPAC	DCE2	0.06	2.0e-87
TAGTTVYGAFDPLLA	DCE2	0.02	3.0e-40
ISNMYAMMIARFKMF	DCE2	0.52	3.6e-37
DVNYAFLHATDLLPA	DCE2	0.07	3.5e-68
PGGAISNMYAMMIAR	DCE2	0.19	1.3e-41
NYAFLHATDLLPACD	DCE2	0.02	4.3e-28
KMMGVPLQCSALLVR	DCE2	0.02	2.7e-30
FTYEIAPVFVLLEYV	DCE2	0.63	3.3e-18
GFVPFLVSATAGTTV	DCE2	0.28	1.3e-44
RPTLAFLQDVMNILL	DCE2	0.36	2.0e-116
VFKLWLMWRAKGTTG	DCE2	1.54	3.8e-06
DVFKLWLMWRAKGTT	DCE2	1.69	4.4e-08
LWLMWRAKGTTGFEA	DCE2	1.85	8.2e-09
AGTTVYGAFDPLLAV	DCE2	0.2	3.8e-23
PTLAFLQDVMNILLQ	DCE2	0.46	3.6e-65
ERPTLAFLQDVMNIL	DCE2	0.43	8.9e-78
TLAFLQDVMNILLQY	DCE2	0.46	3.6e-65
LAFLQDVMNILLQYV	DCE2	0.3	2.9e-61
GIFSPGGAISNMYAM	DCE2	0.56	3.8e-36
WNPHKMMGVPLQCSA	DCE2	0.05	2.2e-14
QHTNVCFWYIPPSLR	DCE2	0.38	5.4e-11
TTVYGAFDPLLAVAD	DCE2	0.12	2.8e-11
VMNILLQYVVKSFDR	DCE2	0.02	1.1e-07
LGVIGGMLVAEAFEH	G6PC2	1.57	1.3e-70
ILGVIGGMLVAEAFE	G6PC2	1.56	1.2e-60
VILGVIGGMLVAEAF	G6PC2	1.22	7.8e-14
GVIGGMLVAEAFEHT	G6PC2	2.03	3.9e-135
VIGGMLVAEAFEHTP	G6PC2	1.95	2.0e-90
QVILGVIGGMLVAEA	G6PC2	4.7	4.0e-306
VLFGLGFAINSEMFL	G6PC2	1.15	2.4e-09

GVLFGLGFAINSEMF	G6PC2	1.4	9.8e-41
LNIDLLWSVPIAKKW	G6PC2	0.63	2.7e-29
FGLGFAINSEMFLLS	G6PC2	0.85	1.7e-09
LRVLNIDLLWSVPIA	G6PC2	0.8	6.3e-08
VLNIDLLWSVPIAKK	G6PC2	0.75	5.4e-10
LLRVLNIDLLWSVPI	G6PC2	0.16	1.2e-48
VAFIPYSVHMLMKQS	G6PC2	1.53	4.9e-06
AFIPYSVHMLMKQSG	G6PC2	1.69	4.4e-08
CVWYVMVTAALSHTV	G6PC2	0.08	8.3e-46
WYVMVTAALSHTVCG	G6PC2	0.02	1.8e-18
SIPLTVVAFIPYSVH	G6PC2	0.55	5.8e-20
LGVLFGLGFAINSEM	G6PC2	0.03	8.3e-10
NIDLLWSVPIAKKWC	G6PC2	0.02	3.3e-08
RGGSALLALCGALAA	CBPE	0.13	4.8e-36
AVIHWIMDIPFVLSA	CBPE	0.63	3.2e-65
ETIVNLIHSTRIHIM	CBPE	1.53	4.9e-06
AIFQSLARAYSSFNP	CBPE	0.53	3.7e-19
PETKAVIHWIMDIPF	CBPE	0.55	2.1e-10
MAGRGGSALLALCGA	CBPE	0.33	1.3e-41
YSPAAGVDFELESFS	CBPE	6.77	0.0e+00
DDAIFQSLARAYSSF	CBPE	0.4	1.1e-25
IFQSLARAYSSFNPA	CBPE	0.08	3.8e-43
RVDFSLAGALNAGFK	GFAP	0.88	5.3e-06
PTRVDFSLAGALNAG	GFAP	0.44	3.9e-24
DFSLAGALNAGFKET	GFAP	0.48	5.5e-14
RRRITSAARRSYVSS	GFAP	1.68	5.8e-08
ERRRITSAARRSYVS	GFAP	1.69	4.4e-08
LLNVKLALDIEIATY	GFAP	0.33	2.4e-43
LPTRVDFSLAGALNA	GFAP	0.59	1.1e-08
EAASYQEALARLEEE	GFAP	2.27	2.3e-67
EAVAYGAAVQAAILM	HS71A	0.65	5.7e-67
DEAVAYGAAVQAAIL	HS71A	0.64	4.6e-72
PDEAVAYGAAVQAAI	HS71A	0.65	8.1e-65
AVAYGAAVQAAILMG	HS71A	0.65	4.1e-68
INEPTAAAIAYGLDR	HS71A	0.71	2.9e-45
IINEPTAAAIAYGLD	HS71A	0.7	2.6e-46
VAYGAAVQAAILMGD	HS71A	0.7	6.2e-48
RIINEPTAAAIAYGL	HS71A	0.71	4.5e-45
NPDEAVAYGAAVQAA	HS71A	0.75	2.1e-30
NEPTAAAIAYGLDRT	HS71A	0.7	1.0e-45
AYGAAVQAAILMGDK	HS71A	0.52	2.2e-127
EPTAAAIAYGLDRTG	HS71A	0.73	6.7e-42
LRIINEPTAAAIAYG	HS71A	0.49	3.7e-173
SLGLETAGGVMTALI	HS71A	0.57	2.5e-106
LGLETAGGVMTALIK	HS71A	0.57	2.6e-103
INPDEAVAYGAAVQA	HS71A	1.12	2.6e-05
VLRIINEPTAAAIAY	HS71A	0.9	7.1e-05
PTAAAIAYGLDRTGK	HS71A	1.49	4.0e-34
GLETAGGVMTALIKR	HS71A	0.61	1.9e-81
SSSTQASLEIDSLFE	HS71A	0.33	2.3e-35
STQASLEIDSLFEGI	HS71A	0.33	1.2e-34
LETAGGVMTALIKRN	HS71A	0.26	4.5e-286
IAEAYLGYPVTNAVI	HS71A	0.08	1.2e-09
AKRTLSSSTQASLEI	HS71A	0.56	5.1e-20
LSLGLETAGGVMTAL	HS71A	0.34	4.0e-192
ETAGGVMTALIKRNS	HS71A	0.31	1.6e-05
QDLLLLDVAPLSLGL	HS71A	0.33	1.7e-125
VQDLLLLDVAPLSLG	HS71A	0.32	1.6e-130
NVQDLLLLDVAPLSL	HS71A	0.46	3.2e-126
DLLLLDVAPLSLGLE	HS71A	0.21	2.3e-136
ENVQDLLLLDVAPLS	HS71A	0.21	2.7e-135
LLLLDVAPLSLGLET	HS71A	0.13	4.8e-36
AGVIAGLNVLRIINE	HS71A	0.13	4.8e-36
IDFYTSITRARFEEL	HS71A	1.53	4.9e-06
FEGIDFYTSITRARF	HS71A	1.68	5.8e-08
YPVTNAVITVPAYFN	HS71A	0.39	3.6e-66
GYPVTNAVITVPAYF	HS71A	0.37	1.6e-65
KRTLSSSTQASLEID	HS71A	0.68	3.0e-09
RAKRTLSSSTQASLE	HS71A	0.13	7.8e-15
PVTNAVITVPAYFND	HS71A	0.7	3.8e-08
LGYPVTNAVITVPAY	HS71A	0.75	9.0e-06
SENVQDLLLLDVAPL	HS71A	0.23	1.4e-17
FDVSILTIDDGIFEV	HS71A	0.23	1.4e-17
PLSLGLETAGGVMTA	HS71A	0.54	1.2e-29
TAAAIAYGLDRTGKG	HS71A	0.05	2.2e-14
YGAAVQAAILMGDKS	HS71A	0.05	2.2e-14
LFEGIDFYTSITRAR	HS71A	2.85	1.5e-12
TQRLANFLVHSSNNF	IAPP	0.2	2.8e-24
HSSNNFGAILSSTNV	IAPP	0.19	2.3e-94
QRLANFLVHSSNNFG	IAPP	0.39	2.2e-07
RLANFLVHSSNNFGA	IAPP	0.4	3.8e-07
LTAWFSLFADLDPLS	ICA69	1.99	1.6e-163
DLTAWFSLFADLDPL	ICA69	2.99	0.0e+00
TAWFSLFADLDPLSN	ICA69	1.98	1.4e-162
AWFSLFADLDPLSNP	ICA69	2.16	2.4e-169
SDLTAWFSLFADLDP	ICA69	2.58	1.4e-262
WFSLFADLDPLSNPD	ICA69	1.78	7.0e-108
LSEIFNASSLEEGEF	ICA69	1.56	1.6e-76
ASDLTAWFSLFADLD	ICA69	2.45	5.1e-243
DEHVVASDADLDAKL	ICA69	2.11	1.9e-173
EHVVASDADLDAKLE	ICA69	3.78	0.0e+00
KEDEHVVASDADLDA	ICA69	1.48	1.0e-30
EKTSHTMAAIHESFK	ICA69	0.29	7.7e-209
KTSHTMAAIHESFKG	ICA69	0.13	7.1e-164

TSHTMAAIHESFKGY	ICA69	0.05	2.5e-83
AASDLTAWFSLFADL	ICA69	0.61	4.4e-08
FWEKTSHTMAAIHES	ICA69	0.08	3.1e-43
SEIFNASSLEEGEFS	ICA69	2.38	1.1e-78
LLSEIFNASSLEEGE	ICA69	2.23	5.4e-66
LLLSEIFNASSLEEG	ICA69	2.21	4.0e-65
EIFNASSLEEGEFSK	ICA69	2.24	3.1e-66
LALLALWGPDPAAAF	INS	0.57	2.5e-19
ALLALWGPDPAAAFV	INS	0.23	4.6e-22
LLALLALWGPDPAAA	INS	0.56	4.5e-19
ALWMRLLPLLALLAL	INS	0.11	2.0e-46
LWMRLLPLLALLALW	INS	0.11	2.0e-46
CGSHLVEALYLVCGE	INS	2.27	2.3e-67
RTRTAYTRAQLLELE	PDX1	0.42	5.5e-06
WKGQWAGGAYAAEPE	PDX1	0.55	4.4e-38
GQWAGGAYAAEPEEN	PDX1	2.14	9.3e-63
KGQWAGGAYAAEPEE	PDX1	0.55	4.4e-38
PPGGAVPPAAPVAAR	PDX1	0.49	1.2e-69
PPPGGAVPPAAPVAA	PDX1	0.47	1.4e-66
AWKGQWAGGAYAAEP	PDX1	0.55	7.1e-39
PGGAVPPAAPVAARE	PDX1	0.47	3.7e-66
LFNKYISRPRRVELA	PDX1	1.85	8.2e-09
PPPPGGAVPPAAPVA	PDX1	0.56	2.4e-35
PPPPPGGAVPPAAPV	PDX1	0.33	1.3e-41
RGGGTAVGGGGVAEP	PDX1	0.34	6.7e-41
GGAVPPAAPVAAREG	PDX1	0.33	1.3e-41
GGTAVGGGGVAEPEQ	PDX1	0.34	6.7e-41
HAWKGQWAGGAYAAE	PDX1	0.57	4.5e-35
NKRTRTAYTRAQLLE	PDX1	2.13	2.7e-10
LERKIESLMDEIEFL	PERI	1.94	4.0e-135
ELERKIESLMDEIEF	PERI	1.96	1.1e-138
KIESLMDEIEFLKKL	PERI	1.96	8.8e-139
RLDFSMAEALNQEFL	PERI	3.34	0.0e+00
LDFSMAEALNQEFLA	PERI	3.1	0.0e+00
ERLDFSMAEALNQEF	PERI	3.1	0.0e+00
KISSIQSIVPALEIA	CH60	2.08	2.6e-173
KKISSIQSIVPALEI	CH60	2.36	1.0e-219
ISSIQSIVPALEIAN	CH60	2.41	1.1e-229
SSIQSIVPALEIANA	CH60	3.03	0.0e+00
KPLVIIAEDVDGEAL	CH60	2.47	2.5e-240
PLVIIAEDVDGEALS	CH60	2.43	5.7e-232
RRGVMLAVDAVIAEL	CH60	1.6	3.4e-67
LVIIAEDVDGEALST	CH60	2.44	5.0e-233
RGVMLAVDAVIAELK	CH60	0.89	3.5e-06
HRKPLVIIAEDVDGE	CH60	1.46	1.9e-27
VIIAEDVDGEALSTL	CH60	1.46	1.9e-27
DGTTTATVLARSIAK	CH60	0.26	3.2e-207
SRLLGSASPSSSVRL	PERI	0.34	1.2e-161
RLLGSASPSSSVRLG	PERI	0.46	2.9e-81
SSSRLLGSASPSSSV	PERI	0.42	1.4e-104
ALEAGGYQAGAARLE	PERI	0.56	7.6e-36
SERLDFSMAEALNQE	PERI	1.39	1.0e-25
LLNVKMALDIEIATY	PERI	0.28	2.7e-69
LLGSASPSSSVRLGS	PERI	0.29	1.9e-75
FALEAGGYQAGAARL	PERI	0.56	7.0e-36
FRSPRAGAGALLRLP	PERI	0.33	7.0e-42
SFRSPRAGAGALLRL	PERI	0.33	1.3e-42
FSSSRLLGSASPSSS	PERI	0.39	1.4e-31
PSERLDFSMAEALNQ	PERI	0.33	2.4e-43
ELLNVKMALDIEIAT	PERI	0.33	2.4e-43
QFALEAGGYQAGAAR	PERI	0.74	8.5e-07
KEQFEFALTAVAEEV	PTPR2	4.5	0.0e+00
RHHLMAALSAYAAQR	PTPR2	0.42	2.6e-237
QFEFALTAVAEEVNA	PTPR2	2.11	0.0e+00
EQFEFALTAVAEEVN	PTPR2	4.45	0.0e+00
DRHHLMAALSAYAAQ	PTPR2	0.43	2.3e-219
VDRHHLMAALSAYAA	PTPR2	0.48	6.9e-180
HHLMAALSAYAAQRP	PTPR2	0.41	4.1e-238
FEFALTAVAEEVNAI	PTPR2	2.21	0.0e+00
HLMAALSAYAAQRPP	PTPR2	0.43	2.0e-229
EFALTAVAEEVNAIL	PTPR2	2.24	0.0e+00
TKEQFEFALTAVAEE	PTPR2	4.09	0.0e+00
GVDRHHLMAALSAYA	PTPR2	0.78	2.6e-18
LMAALSAYAAQRPPA	PTPR2	0.5	1.0e-156
FALTAVAEEVNAILK	PTPR2	1.73	0.0e+00
LLQVPSSAFADVEVL	PTPR2	3.51	0.0e+00
DDYTQYVMDQELADL	PTPR2	2.31	1.4e-212
DYTQYVMDQELADLP	PTPR2	2.44	5.0e-233
LQVPSSAFADVEVLG	PTPR2	1.6	5.8e-51
QVPSSAFADVEVLGP	PTPR2	1.68	6.2e-55
MAALSAYAAQRPPAP	PTPR2	0.22	6.2e-243
ESILTYVAHTSALTY	PTPR2	0.66	1.4e-27
SESILTYVAHTSALT	PTPR2	0.35	6.9e-107
SILTYVAHTSALTYP	PTPR2	0.47	4.8e-60
FSESILTYVAHTSAL	PTPR2	0.24	5.7e-151
AMDFYRYEVSPVALQ	PTPR2	1.22	1.2e-05
MDFYRYEVSPVALQR	PTPR2	1.49	7.8e-13
DFYRYEVSPVALQRL	PTPR2	1.63	1.9e-17
PAMDFYRYEVSPVAL	PTPR2	1.64	5.6e-19
QTKEQFEFALTAVAE	PTPR2	1.77	3.6e-57
TQFHFLSWYDRGVPS	PTPR2	0.45	1.6e-10
VTQFHFLSWYDRGVP	PTPR2	0.52	2.0e-07
STGHMILSYMEDHLK	PTPR2	0.6	1.6e-10
VPAMDFYRYEVSPVA	PTPR2	1.97	4.1e-30

YSREGGAALANALRR	PTPR2	0.35	6.3e-178
RYSREGGAALANALR	PTPR2	0.35	6.3e-178
SGVDRHHLMAALSAY	PTPR2	0.09	4.6e-49
RRYSREGGAALANAL	PTPR2	0.37	1.5e-158
SREGGAALANALRRH	PTPR2	0.35	5.8e-149
ERRYSREGGAALANA	PTPR2	0.55	3.5e-38
LLLLLLLLPPRVLPA	PTPR2	0.35	2.8e-70
LLLLLLLLLPPRVLP	PTPR2	0.1	1.0e-48
SGCVVIVMLTPLAEN	PTPR2	0.13	4.8e-36
RFSESILTYVAHTSA	PTPR2	0.59	3.9e-13
TGHMILSYMEDHLKN	PTPR2	0.55	1.9e-12
ISTGHMILSYMEDHL	PTPR2	0.55	2.1e-10
HMILSYMEDHLKNKN	PTPR2	0.55	2.1e-10
GVARGSPGRAALGES	PTPR2	0.33	1.3e-41
TKFIALTLVSLACIL	PTPR2	0.01	6.4e-06
STKFIALTLVSLACI	PTPR2	0.01	9.4e-06
RLLQVPSSAFADVEV	PTPR2	2.43	4.1e-80
VPSSAFADVEVLGPA	PTPR2	2.39	5.1e-79
FYRYEVSPVALQRLR	PTPR2	0.11	3.3e-10
PLLLLLLLLLPPRVL	PTPR2	0.02	1.1e-07
LLLLLLPPRVLPAAP	PTPR2	0.02	3.3e-08
LLLLLPPRVLPAAPS	PTPR2	0.02	1.1e-07
DRFSESILTYVAHTS	PTPR2	0.45	1.8e-05
PSSAFADVEVLGPAV	PTPR2	4.99	3.4e-17
ALTAVAEEVNAILKA	PTPR2	2.23	3.4e-17
KDQFEFALTAVAEEV	PTPRN	4.5	0.0e+00
DQFEFALTAVAEEVN	PTPRN	4.51	0.0e+00
SKDQFEFALTAVAEE	PTPRN	4.09	0.0e+00
AGASSSLSPLQAELL	PTPRN	5.36	0.0e+00
NMDISTGHMILAYME	PTPRN	0.06	7.3e-67
HVHMSSGSFINISVV	PTPRN	0.19	2.6e-93
MDISTGHMILAYMED	PTPRN	0.08	3.7e-41
VHMSSGSFINISVVG	PTPRN	0.02	1.5e-30
DISTGHMILAYMEDH	PTPRN	0.01	4.9e-15
EHVHMSSGSFINISV	PTPRN	0.01	3.5e-15
HMSSGSFINISVVGP	PTPRN	0.01	4.9e-15
RSKDQFEFALTAVAE	PTPRN	1.83	1.2e-61
HMILAYMEDHLRNRD	PTPRN	0.6	1.6e-10
VALAGVAGLLVALAV	PTPRN	0.11	2.5e-78
LTLVALAGVAGLLVA	PTPRN	0.3	3.1e-08
SFINISVVGPALTFR	PTPRN	0.44	9.8e-35
FINISVVGPALTFRI	PTPRN	0.32	1.4e-67
GSFINISVVGPALTF	PTPRN	0.42	4.7e-39
GTYILIDMVLNRMAK	PTPRN	0.11	2.0e-46
PMRSVLLTLVALAGV	PTPRN	0.13	4.8e-36
SPMRSVLLTLVALAG	PTPRN	0.13	4.8e-36
GASSSLSPLQAELLP	PTPRN	1.56	3.3e-23
ASSSLSPLQAELLPP	PTPRN	1.46	3.1e-12
VGKGGAGASSSLSPL	PTPRN	0.48	8.8e-63
GTTTATVLARSIAKE	CH60	0.2	2.5e-194
GDGTTTATVLARSIA	CH60	0.28	3.4e-203
IRRGVMLAVDAVIAE	CH60	0.62	3.2e-51
GVMLAVDAVIAELKK	CH60	0.53	3.9e-60
LQGVDLLADAVAVTM	CH60	0.6	7.2e-23
KDMAIATGGAVFGEE	CH60	0.34	2.2e-142
MAIATGGAVFGEEGL	CH60	0.47	5.2e-69
LKDMAIATGGAVFGE	CH60	0.36	1.0e-144
AIATGGAVFGEEGLT	CH60	0.33	2.9e-44
TALLDAAGVASLLTT	CH60	0.28	1.9e-06
ALLDAAGVASLLTTA	CH60	0.31	7.1e-05
RTALLDAAGVASLLT	CH60	0.26	2.0e-08
PVGKGGAGASSSLSP	PTPRN	0.48	1.1e-62
PPVGKGGAGASSSLS	PTPRN	0.56	2.4e-35
QPPVGKGGAGASSSL	PTPRN	0.33	1.3e-41
GKGGAGASSSLSPLQ	PTPRN	0.29	3.5e-74
QDIPTGSAPAAQHRL	PTPRN	0.48	5.8e-62
LQDIPTGSAPAAQHR	PTPRN	0.57	4.5e-35
DIPTGSAPAAQHRLP	PTPRN	0.57	9.7e-35
KGGAGASSSLSPLQA	PTPRN	0.33	1.3e-41
IPTGSAPAAQHRLPQ	PTPRN	0.56	2.4e-35
ALAGVAGLLVALAVA	PTPRN	0.07	4.9e-47
VGQREEAAAVLPQTA	PTPRN	1.62	1.5e-41
QREEAAAVLPQTAHS	PTPRN	0.32	5.0e-17
REEAAAVLPQTAHST	PTPRN	0.36	8.0e-13
FHFLSWPAEGTPAST	PTPRN	2.85	1.5e-12
LAAVLAGYGVELRQL	PTPRN	0.49	2.9e-05
AAVLAGYGVELRQLT	PTPRN	0.11	3.3e-10
YILIDMVLNRMAKGV	PTPRN	0.02	1.1e-07
ELVAAASAVADSLPF	RT31	0.72	2.2e-38
SPELVAAASAVADSL	RT31	0.79	2.1e-22
PELVAAASAVADSLP	RT31	0.8	3.7e-19
LSPELVAAASAVADS	RT31	0.79	3.6e-26
PLSPELVAAASAVAD	RT31	0.88	5.8e-09
VAAASAVADSLPFDK	RT31	1.67	1.1e-91
GSPETSAAAIMLLTV	RT31	0.88	2.8e-07
EPLSPELVAAASAVA	RT31	0.87	1.9e-08
SSGSPETSAAAIMLL	RT31	1.31	2.5e-22
ISDMKVARSATARVR	RT31	0.18	1.4e-170
SDMKVARSATARVRS	RT31	0.21	1.8e-141
DMKVARSATARVRSR	RT31	0.21	1.8e-141
MKVARSATARVRSRP	RT31	0.22	4.6e-123
IISDMKVARSATARV	RT31	0.2	5.0e-143
PETSAAAIMLLTVRH	RT31	0.34	1.9e-150
ETSAAAIMLLTVRHG	RT31	0.03	1.0e-51

LLGIIKGMKVELSTV	RT31	0.08	8.3e-70
AAASAVADSLPFDKQ	RT31	0.24	1.2e-08
LSSGSPETSAAAIML	RT31	0.48	1.2e-72
PRVSTFLPLRPLSRH	RT31	1.53	4.9e-06
RVSTFLPLRPLSRHP	RT31	1.69	4.4e-08
GTVRYRSSALLARTK	RT31	1.53	4.9e-06
NIISDMKVARSATAR	RT31	0.43	1.9e-10
TVRYRSSALLARTKN	RT31	1.54	3.8e-06
KVARSATARVRSRPE	RT31	1.68	5.8e-08
HGTVRYRSSALLART	RT31	1.54	3.8e-06
RHGTVRYRSSALLAR	RT31	1.69	4.4e-08
SNIISDMKVARSATA	RT31	0.43	1.7e-10
PLSSGSPETSAAAIM	RT31	0.48	8.8e-63
IEPLSPELVAAASAV	RT31	0.36	5.7e-13
VARSATARVRSRPEL	RT31	2.8	2.6e-12
MFPRVSTFLPLRPLS	RT31	1.85	8.0e-08
VQANASVRAAFVHAL	ZNT8	0.42	7.4e-227
MIIVSSCAVAANIVL	ZNT8	0.44	1.5e-242
IIVSSCAVAANIVLT	ZNT8	0.46	2.4e-208
VMIIVSSCAVAANIV	ZNT8	0.44	2.5e-243
GGHIAGSLAVVTDAA	ZNT8	0.61	2.6e-78
QANASVRAAFVHALG	ZNT8	0.44	9.9e-207
VGGHIAGSLAVVTDA	ZNT8	0.66	6.8e-56
ANASVRAAFVHALGD	ZNT8	0.44	6.4e-205
NASVRAAFVHALGDL	ZNT8	0.58	5.5e-90
GHIAGSLAVVTDAAH	ZNT8	0.75	1.0e-27
VVGGHIAGSLAVVTD	ZNT8	0.8	9.0e-18
EVQANASVRAAFVHA	ZNT8	0.55	8.0e-118
IVSSCAVAANIVLTV	ZNT8	0.66	6.0e-61
NQVILSAHVATAASR	ZNT8	0.51	2.1e-155
TVMIIVSSCAVAANI	ZNT8	0.47	8.7e-205
HIAGSLAVVTDAAHL	ZNT8	1.35	9.0e-25
ASVRAAFVHALGDLF	ZNT8	1.33	1.7e-22
EVVGGHIAGSLAVVT	ZNT8	0.72	1.5e-37
QVILSAHVATAASRD	ZNT8	0.53	2.1e-141
VILSAHVATAASRDS	ZNT8	0.57	4.5e-114
MNQVILSAHVATAAS	ZNT8	0.51	1.9e-155
KEVQANASVRAAFVH	ZNT8	0.58	1.1e-102
HKEVQANASVRAAFV	ZNT8	0.56	1.1e-112
AAKMYAFTLESVELQ	ZNT8	2.49	1.3e-261
AKMYAFTLESVELQQ	ZNT8	2.64	3.5e-304
KMYAFTLESVELQQK	ZNT8	2.09	1.1e-164
KAAKMYAFTLESVEL	ZNT8	1.32	2.5e-39
ILSAHVATAASRDSQ	ZNT8	0.29	0.0e+00
TMNQVILSAHVATAA	ZNT8	0.28	0.0e+00
AEVVGGHIAGSLAVV	ZNT8	0.3	1.6e-241
IAEVVGGHIAGSLAV	ZNT8	0.31	4.4e-199
MIAEVVGGHIAGSLA	ZNT8	0.35	3.8e-121
LTFGWHRAEILGALL	ZNT8	1.42	4.2e-14
KRLTFGWHRAEILGA	ZNT8	1.52	2.1e-18
SVRAAFVHALGDLFQ	ZNT8	0.82	2.1e-06
TFGWHRAEILGALLS	ZNT8	0.47	2.0e-30
SFTMHSLTIQMESPV	ZNT8	0.7	3.6e-43
SKSFTMHSLTIQMES	ZNT8	0.66	2.2e-58
KSFTMHSLTIQMESP	ZNT8	0.69	4.0e-42
FTMHSLTIQMESPVD	ZNT8	0.85	5.4e-09
LSKSFTMHSLTIQME	ZNT8	1.2	9.8e-14
LHIWSLTMNQVILSA	ZNT8	0.67	3.4e-30
SLHIWSLTMNQVILS	ZNT8	0.65	1.1e-35
ATVMIIVSSCAVAAN	ZNT8	0.29	1.2e-281
LSAHVATAASRDSQV	ZNT8	0.29	3.9e-155
QATVMIIVSSCAVAA	ZNT8	0.25	2.1e-126
VSSCAVAANIVLTVV	ZNT8	0.27	1.4e-75
IAGSLAVVTDAAHLL	ZNT8	0.4	1.9e-23
HIWSLTMNQVILSAH	ZNT8	0.18	2.5e-88
IWSLTMNQVILSAHV	ZNT8	0.12	1.7e-70
KELILAVDGVLSVHS	ZNT8	0.23	1.5e-121
VKELILAVDGVLSVH	ZNT8	0.23	1.8e-122
GVKELILAVDGVLSV	ZNT8	0.23	2.0e-121
WSLTMNQVILSAHVA	ZNT8	0.1	8.7e-49
ELILAVDGVLSVHSL	ZNT8	0.21	2.4e-35
HSLHIWSLTMNQVIL	ZNT8	0.71	4.5e-18
SLTMNQVILSAHVAT	ZNT8	0.13	4.8e-36
KALSKSFTMHSLTIQ	ZNT8	1.53	4.9e-06
LTMNQVILSAHVATA	ZNT8	0.15	4.9e-114
FMIAEVVGGHIAGSL	ZNT8	0.33	7.0e-42
VRAAFVHALGDLFQS	ZNT8	1.55	2.0e-19
QLKDMAIATGGAVFG	CH60	0.46	7.0e-70
LLDAAGVASLLTTAE	CH60	0.22	7.9e-09
VRTALLDAAGVASLL	CH60	0.26	4.1e-07
EIRRGVMLAVDAVIA	CH60	0.29	1.4e-140
GCALLRCIPALDSLT	CH60	0.11	5.2e-42
RALMLQGVDLLADAV	CH60	0.36	1.8e-131
ARALMLQGVDLLADA	CH60	0.34	3.1e-137
VMLAVDAVIAELKKQ	CH60	0.38	5.7e-79
ALMLQGVDLLADAVA	CH60	0.39	7.9e-105
VSRVLAPHLTRAYAK	CH60	1.68	5.8e-08
PVSRVLAPHLTRAYA	CH60	1.69	4.4e-08
NHKEVQANASVRAAF	ZNT8	0.08	3.8e-43
DLFQSISVLISALII	ZNT8	0.02	8.3e-12
GDLFQSISVLISALI	ZNT8	0.15	2.1e-25
SSCAVAANIVLTVVL	ZNT8	0.01	3.1e-09
LFQSISVLISALIIY	ZNT8	0.01	3.1e-09
SCAVAANIVLTVVLH	ZNT8	0.02	4.1e-05

LGDLFQSISVLISAL	ZNT8	0.01	1.4e-05
SVHSLHIWSLTMNQV	ZNT8	0.15	3.8e-27
SRVLAPHLTRAYAKD	CH60	1.69	4.4e-08
PTVFRQMRPVSRVLA	CH60	1.64	1.5e-06
TTTATVLARSIAKEG	CH60	0.14	8.3e-131
LMLQGVDLLADAVAV	CH60	0.54	3.5e-46
AGDGTTTATVLARSI	CH60	0.21	1.9e-142
DARALMLQGVDLLAD	CH60	0.28	2.7e-69
TTATVLARSIAKEGF	CH60	0.12	5.8e-44
MLQGVDLLADAVAVT	CH60	0.33	2.4e-43
ASLLTTAEVVVTEIP	CH60	1.81	9.5e-06
DQELESLSAIEAELE	CMGA	5.07	0.0e+00
QELESLSAIEAELEK	CMGA	5.07	0.0e+00
ELESLSAIEAELEKV	CMGA	5.07	0.0e+00
EDQELESLSAIEAEL	CMGA	5.13	0.0e+00
LESLSAIEAELEKVA	CMGA	3.3	0.0e+00
ESLSAIEAELEKVAH	CMGA	3.25	0.0e+00
PEDQELESLSAIEAE	CMGA	3.5	0.0e+00
SLSAIEAELEKVAHQ	CMGA	0.33	2.3e-35
SMKLSFRARAYGFRG	CMGA	1.53	4.9e-06
NRDSSMKLSFRARAY	CMGA	1.53	4.9e-06
KLSFRARAYGFRGPG	CMGA	1.69	4.4e-08
FYVNATAGTTVYGAF	DCE1	0.54	3.2e-126
YVNATAGTTVYGAFD	DCE1	0.55	8.7e-123
VNATAGTTVYGAFDP	DCE1	0.63	9.1e-70
PFYVNATAGTTVYGA	DCE1	0.59	4.4e-99
TQSDIDFLIEEIERL	DCE1	2.37	1.9e-223
QSDIDFLIEEIERLG	DCE1	2.36	2.1e-220
ATQSDIDFLIEEIER	DCE1	2.47	2.5e-240
AATQSDIDFLIEEIE	DCE1	2.36	2.7e-220
NMFTYEIAPVFVLME	DCE1	1.63	9.2e-99
SDIDFLIEEIERLGQ	DCE1	2.48	2.2e-241
PAATQSDIDFLIEEI	DCE1	3.51	6.9e-252
AISNMYSIMAARYKY	DCE1	0.31	0.0e+00
ISNMYSIMAARYKYF	DCE1	0.33	4.4e-277
GAISNMYSIMAARYK	DCE1	0.31	0.0e+00
GGAISNMYSIMAARY	DCE1	0.47	2.3e-196
SNMYSIMAARYKYFP	DCE1	0.21	2.5e-276
KANFFRMVISNPAAT	DCE1	0.69	2.9e-22
ANFFRMVISNPAATQ	DCE1	0.37	6.6e-97
DKANFFRMVISNPAA	DCE1	0.54	1.1e-35
NFFRMVISNPAATQS	DCE1	0.39	4.6e-58
ANTNMFTYEIAPVFV	DCE1	0.85	6.0e-23

Table S5. Predicted T1D-associated deamidated epitopes (N = 20) bound by DRB1 variants.

Deamidated-epitope	Protein	Odds ratio	P-value
ANFFRMVISNPAATE	DCE1	0.53	5.8e-125
FFRMVISNPAATESD	DCE1	0.58	1.2e-99
DERILSILRHENLLK	CMGA	0.47	2.0e-33
EKGFVPFLVSATAGT	DCE2	0.2	3.3e-09
KDEFEFALTAVAEEV	PTPRN	0.2	3.3e-09
KEEFEFALTAVAEEV	PTPR2	0.2	3.3e-09
DFLHRNGVLIIQHLE	G6PC2	0.75	6.9e-06
FRMVISNPAATESDI	DCE1	2.43	3.7e-242
SLEEILVDCRDTLKY	DCE1	3.23	0.0e+00
EDLLLLDVAPLSLGL	HS71A	4.68	0.0e+00
VEDLLLLDVAPLSLG	HS71A	5.18	0.0e+00
NVEDLLLLDVAPLSL	HS71A	5.18	0.0e+00
YGYIVTDEKPLSLAA	PTPRN	3.23	0.0e+00
AEEYGYIVTDEKPLS	PTPRN	3.23	0.0e+00
GRRLVEDVARLLEVP	PTPR2	3.23	0.0e+00
TLRGDERILSILRHE	CMGA	3.23	0.0e+00
PYEFTTLKSLEDPMK	ICA69	5.31	6.7e-41
RLDFSMAEALNEEFL	PERI	5.31	6.7e-41
IWSLTMNEVILSAHV	ZNT8	0.1	2.5e-20
EKPLSLAAGVKLLEI	PTPRN	0.34	3.7e-137

Table S6. Predicted T1D-associated deamidated epitopes (N = 94) bound by DQ variants.

Deamidated-epitope	Protein	Odds ratio	P-value
TESDIDFLIEEIERL	DCE1	1.96	1.6e-138
ESDIDFLIEEIERLG	DCE1	1.96	2.9e-139
MFTYEIAPVFVLMEE	DCE1	1.94	1.1e-139
FTYEIAPVFVLMEEI	DCE1	5.18	0.0e+00
EEQTVEFLLEVVDIL	DCE1	1.96	1.1e-138
AATESDIDFLIEEIE	DCE1	1.96	2.9e-139
INPDEAVAYGAAVEA	HS71A	3.24	0.0e+00
QDIPTGSAPAAEHRL	PTPRN	0.61	1.8e-30
LQDIPTGSAPAAEHR	PTPRN	0.54	7.5e-37
EFALEAGGYQAGAAR	PERI	0.53	2.1e-38
RSKDEFEFALTAVAE	PTPRN	2.74	2.2e-283
QTKEEFEFALTAVAE	PTPR2	2.77	3.6e-288
NFFRMVISNPAATES	DCE1	0.24	8.3e-59
EATVMIIVSSCAVAA	ZNT8	0.31	5.2e-131
LTMNEVILSAHVATA	ZNT8	0.24	2.1e-59
ISSIESIVPALEIAN	CH60	2.71	1.5e-276
SSIESIVPALEIANA	CH60	5.22	0.0e+00
SIESIVPALEIANAH	CH60	5.4	0.0e+00
EKKISSIESIVPALE	CH60	5.4	0.0e+00
AGASSSLSPLEAELL	PTPRN	5.4	0.0e+00
GASSSLSPLEAELLP	PTPRN	5.4	0.0e+00
DDLTEYVISQEMERI	PTPRN	0.53	6.5e-52
WHDDLTEYVISQEME	PTPRN	0.53	6.5e-52
LEVPSSAFADVEVLG	PTPR2	2.76	8.2e-284
RLLEVPSSAFADVEV	PTPR2	5.4	0.0e+00
YAFTLESVELQEKPV	ZNT8	2.78	1.1e-288
KEEEEEMAVVPEGLF	CMGA	5.4	0.0e+00
EAASYEEALARLEEE	GFAP	5.4	0.0e+00
SDGLELEVQPSEEEA	PTPR2	6.14	0.0e+00
EEEMAVVPEGLFRGG	CMGA	6.14	0.0e+00
QQKEEEEEMAVVPEG	CMGA	6.14	0.0e+00
AASYEEALARLEEEG	GFAP	6.14	0.0e+00
GEEQTVEFLLEVVDI	DCE1	0.42	7.9e-70
IEKRIEEIIEELDVT	CH60	0.42	7.9e-70
SSTEASLEIDSLFEG	HS71A	0.39	1.7e-76
TEASLEIDSLFEGID	HS71A	0.39	1.7e-76
DLTEYVISQEMERIP	PTPRN	0.39	1.7e-76
TWEDDYTEYVMDQEL	PTPR2	0.42	7.9e-70
CEEVISWLDANTLAE	HS71A	0.05	1.7e-32
EFHFLSWYDRGVPSS	PTPR2	0.24	1.0e-37
VTEFHFLSWYDRGVP	PTPR2	0.06	4.5e-33
LHIWSLTMNEVILSA	ZNT8	0.74	2.1e-06
SLHIWSLTMNEVILS	ZNT8	0.74	2.1e-06
HIWSLTMNEVILSAH	ZNT8	0.74	2.1e-06
IWSLTMNEVILSAHV	ZNT8	0.74	2.1e-06
HSLHIWSLTMNEVIL	ZNT8	0.75	5.6e-06
EQTVEFLLEVVDILL	DCE1	2.52	3.2e-04
AFLEDVMNILLEYVV	DCE2	2.52	3.2e-04
EKRIEEIIEELDVTT	CH60	3.35	2.5e-05
ERPTLAFLEDVMNIL	DCE2	1.46	2.0e-52
KISSIESIVPALEIA	CH60	5.28	0.0e+00
KKISSIESIVPALEI	CH60	5.07	0.0e+00
QIEKRIEEIIEELDV	CH60	1.96	2.9e-139
SSSTEASLEIDSLFE	HS71A	5.06	0.0e+00
STEASLEIDSLFEGI	HS71A	1.96	2.9e-139
EAVAYGAAVEAAILM	HS71A	2.01	3.2e-112
DEAVAYGAAVEAAIL	HS71A	2.01	1.0e-112
KDEFEFALTAVAEEV	PTPRN	5.29	0.0e+00
DEFEFALTAVAEEVN	PTPRN	5.3	0.0e+00
EFEFALTAVAEEVNA	PTPRN	2.3	0.0e+00
SKDEFEFALTAVAEE	PTPRN	5.34	0.0e+00
KEEFEFALTAVAEEV	PTPR2	5.29	0.0e+00
EEFEFALTAVAEEVN	PTPR2	5.3	0.0e+00
TKEEFEFALTAVAEE	PTPR2	5.34	0.0e+00
DDYTEYVMDQELADL	PTPR2	1.96	2.9e-139
LLEVPSSAFADVEVL	PTPR2	5.19	0.0e+00
AKMYAFTLESVELQE	ZNT8	5.7	0.0e+00
KMYAFTLESVELQEK	ZNT8	5.18	0.0e+00
MYAFTLESVELQEKP	ZNT8	5.18	0.0e+00
RLDFSMAEALNEEFL	PERI	5.18	0.0e+00
LDFSMAEALNEEFLA	PERI	5.18	0.0e+00
ERLDFSMAEALNEEF	PERI	5.19	0.0e+00
DFSMAEALNEEFLAT	PERI	5.18	0.0e+00
FSMAEALNEEFLATR	PERI	5.19	0.0e+00
PSERLDFSMAEALNE	PERI	5.18	0.0e+00
LAACGWLLGAEAEEP	CBPE	6.52	0.0e+00
ACGWLLGAEAEEPGA	CBPE	6.52	0.0e+00
ANFFRMVISNPAATE	DCE1	0.24	1.1e-60
NEVILSAHVATAASR	ZNT8	0.16	0.0e+00
EVILSAHVATAASRD	ZNT8	0.16	0.0e+00
TMNEVILSAHVATAA	ZNT8	0.18	0.0e+00
AVAYGAAVEAAILMG	HS71A	0.7	3.2e-47
PDEAVAYGAAVEAAI	HS71A	0.85	8.6e-11
VAYGAAVEAAILMGD	HS71A	0.86	2.4e-09
NPDEAVAYGAAVEAA	HS71A	1.58	4.2e-72
AYGAAVEAAILMGDK	HS71A	0.2	0.0e+00
GKGGAGASSSLSPLE	PTPRN	0.18	4.7e-130
KGGAGASSSLSPLEA	PTPRN	0.18	3.9e-130

RPTLAFLEDVMNILL	DCE2	0.65	1.7e-32
PTLAFLEDVMNILLE	DCE2	0.5	2.9e-108
TLAFLEDVMNILLEY	DCE2	0.57	7.6e-73
LAFLEDVMNILLEYV	DCE2	0.78	4.0e-11
GERPTLAFLEDVMNI	DCE2	0.57	2.0e-72
NVEDLLLLDVAPLSL	HS71A	0.74	9.2e-19

Table S7. Predicted T1D-associated deamidated epitopes (N = 124) bound by trans-encoded DQ variants.

Deamidated-epitope	Protein	Odds ratio	P-value
KISSIESIVPALEIA	CH60	3.83	0.0e+00
KKISSIESIVPALEI	CH60	2.78	0.0e+00
ISSIESIVPALEIAN	CH60	2.35	2.2e-195
SSIESIVPALEIANA	CH60	4.21	0.0e+00
EAVAYGAAVEAAILM	HS71A	0.68	7.5e-47
DEAVAYGAAVEAAIL	HS71A	0.85	3.3e-10
PDEAVAYGAAVEAAI	HS71A	0.76	9.0e-28
AVAYGAAVEAAILMG	HS71A	0.72	3.8e-37
NFFRMVISNPAATES	DCE1	0.37	1.5e-56
FFRMVISNPAATESD	DCE1	0.04	2.8e-16
WEMVWESGCVVIVML	PTPR2	1.48	1.2e-04
FWEMVWESGCVVIVM	PTPR2	1.86	1.2e-08
SLHIWSLTMNEVILS	ZNT8	0.79	1.6e-24
HSLHIWSLTMNEVIL	ZNT8	0.68	2.8e-28
LHIWSLTMNEVILSA	ZNT8	0.79	2.3e-24
HIWSLTMNEVILSAH	ZNT8	0.56	1.2e-61
TEFHFLSWYDRGVPS	PTPR2	0.45	1.6e-10
VTEFHFLSWYDRGVP	PTPR2	0.52	2.0e-07
EATVMIIVSSCAVAA	ZNT8	0.25	2.1e-126
QDIPTGSAPAAEHRL	PTPRN	0.65	3.0e-15
EDLLLLDVAPLSLGL	HS71A	0.33	5.3e-127
VEDLLLLDVAPLSLG	HS71A	0.33	5.3e-127
NVEDLLLLDVAPLSL	HS71A	0.49	5.4e-106
ENVEDLLLLDVAPLS	HS71A	0.31	3.0e-134
IWSLTMNEVILSAHV	ZNT8	0.23	1.8e-121
WSLTMNEVILSAHVA	ZNT8	0.13	4.8e-36
RPTLAFLEDVMNILL	DCE2	2.08	1.6e-115
LAFLEDVMNILLEYV	DCE2	2.76	2.2e-207
AFLEDVMNILLEYVV	DCE2	3.28	4.7e-260
SENVEDLLLLDVAPL	HS71A	0.3	2.9e-61
AGASSSLSPLEAELL	PTPRN	1.96	4.8e-89
GASSSLSPLEAELLP	PTPRN	2.07	2.6e-93
LTMNEVILSAHVATA	ZNT8	0.38	1.5e-20
TVEFLLEVVDILLNY	DCE1	0.55	1.0e-32
VEFLLEVVDILLNYV	DCE1	0.49	6.8e-42
EQTVEFLLEVVDILL	DCE1	3	1.2e-218
LSHTIADFWEMVWES	PTPRN	0.74	1.6e-05
GKGGAGASSSLSPLE	PTPRN	0.29	3.5e-74
KGGAGASSSLSPLEA	PTPRN	0.33	1.3e-41
EKRIEEIIEELDVTT	CH60	6.77	0.0e+00
EFLLEVVDILLNYVR	DCE1	0.33	2.4e-43
FLEDVMNILLEYVVK	DCE2	0.33	2.4e-43
CEEVISWLDANTLAE	HS71A	0.05	3.0e-05
KCEEVISWLDANTLA	HS71A	0.06	1.2e-04
SIESIVPALEIANAH	CH60	1.74	1.3e-59
EKKISSIESIVPALE	CH60	1.81	4.5e-61
GAGASSSLSPLEAEL	PTPRN	1.85	3.4e-64
ASSSLSPLEAELLPP	PTPRN	2.22	8.1e-106
KEEEEEMAVVPEGLF	CMGA	1.94	2.8e-64
LQDIPTGSAPAAEHR	PTPRN	0.75	1.2e-06
DIPTGSAPAAEHRLP	PTPRN	0.75	1.2e-06
EFALEAGGYQAGAAR	PERI	0.74	8.5e-07
RLLEVPSSAFADVEV	PTPR2	2.66	1.1e-122
AKEKGFVPFLVSATA	DCE2	2.85	1.5e-12
SDGLELEVQPSEEEA	PTPR2	2.27	2.3e-67
QKEEEEEMAVVPEGL	CMGA	2.39	6.3e-79
EEEMAVVPEGLFRGG	CMGA	2.27	2.3e-67
QQKEEEEEMAVVPEG	CMGA	2.27	2.3e-67
EAASYEEALARLEEE	GFAP	2.24	3.1e-66
AASYEEALARLEEEG	GFAP	2.27	2.3e-67
SVHSLHIWSLTMNEV	ZNT8	0.15	3.8e-27
EMVWESGCVVIVMLT	PTPR2	0.4	2.4e-04
VAYGAAVEAAILMGD	HS71A	0.75	2.4e-29
INPDEAVAYGAAVEA	HS71A	2.06	2.4e-169
AYGAAVEAAILMGDK	HS71A	0.89	3.0e-05
KDEFEFALTAVAEEV	PTPRN	4.64	0.0e+00
EFEFALTAVAEEVNA	PTPRN	2.11	0.0e+00
DEFEFALTAVAEEVN	PTPRN	4.51	0.0e+00
SKDEFEFALTAVAEE	PTPRN	4.09	0.0e+00
KEEFEFALTAVAEEV	PTPR2	4.5	0.0e+00
SSSLSPLEAELLPPL	PTPRN	4.99	3.4e-17
EEFEFALTAVAEEVN	PTPR2	4.45	0.0e+00
TKEEFEFALTAVAEE	PTPR2	5.17	0.0e+00
EVILGVIGGMLVAEA	G6PC2	4.7	4.0e-306
NEVILSAHVATAASR	ZNT8	0.45	2.7e-203
EVILSAHVATAASRD	ZNT8	0.53	2.1e-141
MNEVILSAHVATAAS	ZNT8	0.52	3.6e-146
TESDIDFLIEEIERL	DCE1	2.07	6.1e-161

AATESDIDFLIEEIE	DCE1	2.36	2.7e-220
ATESDIDFLIEEIER	DCE1	2.36	2.7e-220
ESDIDFLIEEIERLG	DCE1	2.36	2.1e-220
PAATESDIDFLIEEI	DCE1	2.47	2.5e-240
EEQTVEFLLEVVDIL	DCE1	2.05	4.7e-158
FTYEIAPVFVLMEEI	DCE1	3.51	0.0e+00
GEEQTVEFLLEVVDI	DCE1	3.51	6.9e-252
MFTYEIAPVFVLMEE	DCE1	1.88	2.1e-153
TLAFLEDVMNILLEY	DCE2	2.06	2.6e-114
PTLAFLEDVMNILLE	DCE2	1.64	2.7e-58
ERPTLAFLEDVMNIL	DCE2	1.77	1.2e-106
GERPTLAFLEDVMNI	DCE2	2.16	8.6e-127
IEKRIEEIIEELDVT	CH60	3.51	6.9e-252
QIEKRIEEIIEELDV	CH60	2.44	5.0e-233
SSSTEASLEIDSLFE	HS71A	3.01	0.0e+00
STEASLEIDSLFEGI	HS71A	2.47	2.5e-240
SSTEASLEIDSLFEG	HS71A	0.48	2.2e-25
TEASLEIDSLFEGID	HS71A	0.33	1.2e-34
LSSSTEASLEIDSLF	HS71A	0.66	1.2e-11
DDLTEYVISQEMERI	PTPRN	3.45	3.0e-247
WHDDLTEYVISQEME	PTPRN	3.47	5.5e-249
RSKDEFEFALTAVAE	PTPRN	1.31	7.5e-18
DLTEYVISQEMERIP	PTPRN	0.33	1.2e-34
DDYTEYVMDQELADL	PTPR2	2.36	2.7e-220
WEDDYTEYVMDQELA	PTPR2	2.44	5.0e-233
EDDYTEYVMDQELAD	PTPR2	2.31	1.4e-212
LLEVPSSAFADVEVL	PTPR2	3.47	0.0e+00
TWEDDYTEYVMDQEL	PTPR2	3.51	6.9e-252
QTKEEFEFALTAVAE	PTPR2	3.23	0.0e+00
LEVPSSAFADVEVLG	PTPR2	1.6	2.7e-50
EVPSSAFADVEVLGP	PTPR2	1.68	6.2e-55
AKMYAFTLESVELQE	ZNT8	2.73	1.5e-298
KMYAFTLESVELQEK	ZNT8	3.05	0.0e+00
MYAFTLESVELQEKP	ZNT8	2.46	9.1e-266
YAFTLESVELQEKPV	ZNT8	3.33	0.0e+00
RLDFSMAEALNEEFL	PERI	3.21	0.0e+00
LDFSMAEALNEEFLA	PERI	3.36	0.0e+00
ERLDFSMAEALNEEF	PERI	3.52	0.0e+00
DFSMAEALNEEFLAT	PERI	3.56	0.0e+00
SERLDFSMAEALNEE	PERI	3.46	0.0e+00
FSMAEALNEEFLATR	PERI	3	0.0e+00
PSERLDFSMAEALNE	PERI	3.19	0.0e+00
LAACGWLLGAEAEEP	CBPE	1.8	1.3e-93
ACGWLLGAEAEEPGA	CBPE	1.8	1.3e-93
TMNEVILSAHVATAA	ZNT8	0.28	0.0e+00
ANFFRMVISNPAATE	DCE1	0.37	6.6e-97

Table S8. Pathogens (N = 16) tested for molecular mimicry with predicted T1D-assocaited epitopes. All protein sequences were taken from UniProt database.

Pathogens
BovineINS
CoxsackievirusB1
CoxsackievirusB4
CoxsackievirusB5 (strain Peterborough)
Cytomegalovirus (strain Merlin)
EMCV (strain Rueckert)
Epstein-Barr (strain B95-8)
LCMV (strain Armstrong)
LDV
Mumps (strain Miyahara vaccine)
P. tuberculosis (strain ATCC BAA-968)
Parechovirus2 (strain Willianson)
ParvovirusB19 (strain HV)
Poliovirus (strain Mahoney)
RotavirusC (isolate RVC UK Briston 1989)
Rubella (strain Therien)

Conclusion

Pathogens and autoimmunity are considered as two major antagonistic selection forces to operate on classical HLA genes. Previous research on the HLA's association with infectious and autoimmune diseases have suggested a significant implication of the peptide-repertoires of HLA alleles in the outcome of these diseases (*65*, *66*, *79*, *187*, *188*). In order to understand which and how the properties of peptide-repertoire of HLA alleles modulate their specific association with these diseases and by that the selection on HLA genes themselves, I predicted and characterized the peptide-repertoire of several common and rare HLA alleles in the context of one infectious disease (HIV-1/AIDS; Chapters 1 and 2) and one autoimmune disease (Type 1 Diabetes; Chapters 3 and 4).

In case of HIV-1 infection, which is one of the most-studied infectious diseases in humans (*189*), the peptide-repertoire of HLA alleles illustrated multiple aspects of HLA-mediated immune defense against this viral pathogen. Generally, HLA-binding of a broader array of viral peptides, whether it results from the combination of different alleles due to HLA heterozygosity or individual alleles, conferred better viral control. This quantitative advantage could potentially be due to simultaneous T-cell targeting of multiple peptides, which might generate a strong cumulative immune response, and at the same time make it difficult for the virus to concurrently evolve escape mutations at multiple sites without significant replicative fitness cost. Pereyra *et al.* 2014 have shown that the viral control can also be achieved by T-cell targeting of specific peptides (*122*). HLA-binding of a broader array of viral peptides in an individual can make it more likely that specific immunodominant peptides are presented on the cell surface. It is also possible that an allele can alone present such peptides and therefore carrying it might be sufficient to control the disease, irrespective of HLA zygosity. This is supported by the observed superior viral control conferred by specific alleles compared to general heterozygote advantage.

In addition, PepWAS predicted HIV-1-associated epitopes showed that individual peptides can have differential association with the disease, even when bound by the same allele. The motifs revealed by clustering HIV-1-associated epitopes indicated that an allele might confer protection by binding those peptides which are not bound by other alleles, as a pathogen is more likely to evolve first at those peptides which are bound by multiple alleles in a population. This suggests that, in a population, the alleles whose peptide-repertoire are very different from each other would be beneficial, as proposed by

Divergent allele advantage hypothesis (*38*, *41*, *100*, *102*). This would not only allow a broader immune response and make it difficult for the pathogen to adapt to immune defense in an individual, but also ensure that no single pathogen can adapt to the whole population. Altogether, the insights from this work suggest that multiple aspects of HLA-bound peptide-repertoire, such as how many peptides and which peptides, in an individual can modulate the immune response, and a single pathogen can select for both HLA heterozygosity as well as individual HLA alleles.

On the other side, in case of Type 1 Diabetes (T1D), taken as an example of autoimmune conditions, HLA heterozygosity was observed to confer T1D risk. The comparison of the individual-specific peptide-repertoire between HLA heterozygous and homozygous individuals suggested that HLA heterozygote disadvantage can be mediated by both a large repertoire of HLA-bound self-peptides and higher chances of carrying specific risk alleles in an individual. In order to elucidate the functional basis of allele-specific risk for T1D, I compared the peptide-repertoire of T1D risk and protective alleles, which did not support the size of peptide-repertoire as a significant determinant of an allele's effect on T1D. Rather, the findings indicate low peptide-binding affinity of risk alleles as a determinant of their effect on the disease. It has been suggested that weak interaction of immature T-cells with self-peptide-HLA complex is required for their survival and maturation in thymus (*23*), which is thought to be influenced by the binding affinity between peptides and HLA molecules (*164*, *165*). It has also been proposed that negative selection of T-cells in thymus depends on the specificity of peptide-MHC complexes (*47*). These arguments lead to the speculation that the HLA alleles with high peptide-binding affinity might form stable peptide-HLA complexes, making it more likely that autoreactive T-cells are removed in thymus, whereas the HLA alleles with low peptide-binding affinity might form instable peptide-HLA complexes that do not allow efficient removal of autoreactive T-cells. The autoreactive T-cells that survived in thymus might get activated by non-genetic triggers in the periphery (*23*). The findings in this thesis shed light on two possible non-genetic triggers of autoimmunity; (1) neo-peptides generated by post-translational modification due to cellular stress and (2) the infection with pathogen, like *P.* tuberculosis, containing self-mimicking peptides. Due to sequence homology of neo-peptides and pathogen-derived peptides with the native human self-peptides, the immune response initiated by these triggers might spread to self-peptides in the genetically susceptible individual (*183*, *190*), possibly leading to overt autoimmune

diseases. Notably, the molecular mimicry between T1D-associated self-peptides and pathogen-derived peptides, both derived from Heat shock proteins which is present in a diverse taxa of organisms (*191*, *192*), suggests that occurrence of autoimmunity might be a side-effect of the selection for HLA alleles that bind antigens shared by multiple pathogens.

Considering that the binding affinity between a peptide and an HLA molecule might influence the strength of interaction of peptide-HLA complexes with T-cells (*164*, *165*), HLA alleles with high peptide-binding affinity could be perceived to be useful against pathogens because they allow stable synapse between peptide-HLA complex and T-cells (*165*). However, the evolutionary advantage of HLA alleles with low peptide-binding affinity has not been elucidated. Such alleles are present at nominal frequency in human populations, indicating that they have been selectively maintained (*28*). As discussed above, HLA alleles with low peptide-binding affinity might not allow efficient removal of T-cells in thymus, resulting in an overall large repertoire of matured T-cells (including autoreactive T-cells) in periphery. This might bring the advantage against a broader array of pathogens, though at the risk of autoimmunity.

Together, these findings shed light on different factors that can possibly contribute to HLA heterozygote disadvantage as well as allele-specific risk in T1D. It is important to note that there might be additional factors contributing to the specific association of HLA genotypes or alleles with autoimmune diseases that I did not investigate in this thesis, but have been suggested by others, e.g. additional risk of autoimmune diseases (including T1D) conferred by the interaction between two different alleles of the same HLA gene as shown by Lenz *et al.* 2015 (*58*) or the risk of Behçet's disease conferred by epistasis between HLA and non-HLA genes as shown by Kirino *et al.* 2013 (*193*).

Overall, the findings in this thesis suggest that the specific association of HLA genotypes or individual HLA alleles with infectious and autoimmune diseases could be mediated by both quantitative as well as qualitative features of HLA-bound peptides in an individual. The findings also suggest that different variables, such as the number of bound peptides, peptide binding-affinity, peptide-motifs etc., that possibly modulate HLA-mediated immune response might not be mutually exclusive. For example, an effective immune attack to a pathogen might be achieved by HLA-binding of a few immunodominant peptides with high affinity or a large number of peptides with both high and low affinities. Moreover, the extent to which these modulatory variables play a role in HLA-mediated

immune response might depend upon specific condition, e.g. whether the pathogen causes acute or chronic infection (*194*). Altogether, the insights from this thesis shed light on different mechanisms at the molecular level through which the pathogen and autoimmunity-mediated antagonistic selection might operate on classical HLA genes, leading to the persistence of a very diverse variety of HLA alleles in a population without any one being fixed.

Perspectives

In this thesis, I used large datasets on HIV and Type 1 Diabetes (T1D), yet not a large number of disease-associated alleles were represented in the datasets. As the decreasing cost of genome sequencing is facilitating the generation of much bigger datasets, e.g. Biobanks, the characterization of peptide-properties of rare disease-associated alleles would not only help to substantiate the findings from this thesis, but also provide a deeper understanding of the antagonistic selection that could operate on the classical HLA genes. Notably, the alleles of HLA genes, which I analyzed the peptide-repertoires of, were not shared between HIV-1 and T1D. So, the approaches and the findings in this thesis might provide motivation for the future studies focused on individual HLA alleles associated across different diseases. In addition, the PepWAS is applicable to any HLA-mediated diseases. The identification of the HLA-bound peptides associated with other diseases might provide the functional basis for the HLA's association with other diseases as well as aid in designing effective vaccine programs.

References

1. W. F. Zimmerman, Evolution . By Nicholas H. Barton, Derek E. G. Briggs, Jonathan A. Eisen, David Goldstein, and , Nipam H. Patel. Cold Spring Harbor (New York): Cold Spring Harbor Laboratory Press. $100.00. xiv + 833 p.; ill.; index. 978-0-87969-684-9. 2007. *Q. Rev. Biol.* **83**, 204–205 (2008).
2. M. Fumagalli *et al.*, Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* **7** (2011), doi:10.1371/journal.pgen.1002355.
3. M. Sironi, R. Cagliani, D. Forni, M. Clerici, Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**, 224–236 (2015).
4. J. Charles A Janeway, P. Travers, M. Walport, M. J. Shlomchik, *Immunobiology: The Immune System in Health and Disease. 5th edition.* (2001).
5. K. Hoebe, E. Janssen, B. Beutler, The interface between innate and adaptive immunity. *Nat Immunol*. **5**, 971–974 (2004).
6. Y. Kawashima *et al.*, Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*. **458**, 641–645 (2009).
7. A. J. Leslie *et al.*, HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.* **10**, 282–289 (2004).
8. X. Didelot, A. S. Walker, T. E. Peto, D. W. Crook, D. J. Wilson, Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150–162 (2016).
9. G. Sorci, A. P. Møller, T. Boulinier, Genetics of host-parasite interactions. *Trends Ecol. Evol.* **12**, 196–200 (1997).
10. M. . W. J. . D. E. . C. B. . & L. B. R. Woolhouse, Biological and biomedical implications of the co-ecolution of pathogens and their hosts. *Nat. Genet.* . **32**, 569–577 (2002).
11. I. Bartha *et al.*, A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife*. **2013**, 1–16 (2013).
12. N. C. Elde, H. S. Malik, The evolutionary conundrum of pathogen mimicry. *Nat. Rev. Microbiol.* **7**, 787–797 (2009).
13. M. Milinski, The Major Histocompatibility Complex, Sexual Selection, and Mate Choice. *Annu. Rev. Ecol. Evol. Syst.* **37**, 159–186 (2006).
14. E. K. Karlsson, D. P. Kwiatkowski, P. C. Sabeti, Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15**, 379–393 (2014).
15. M. E. J. Woolhouse, J. P. Webster, E. Domingo, B. Charlesworth, B. R. Levin, Biological and biomedical implicatios of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32**, 569–577 (2002).
16. O. Kaltz, J. A. Shykoff, Local adaptation in host – parasite systems. *Heredity (Edinb).* **81**, 361–370 (1998).
17. J. Lighten *et al.*, Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nat. Commun.* **8**, 1–10 (2017).
18. G. W. Litman, J. P. Rast, S. D. Fugmann, The origins of vertebrate adaptive immunity. *Nat. Rev. Immunol.* **10**, 543–553 (2010).
19. P. A. Gorer, The detection of a hereditary antigenic difference in the blood of mice by means of human group a serum. *J. Genet.* (1936), doi:10.1007/BF02982499.
20. Klein, *Natural History of the Major Histocompatibility Complex* (Wiley, 1986).
21. R. Horton *et al.*, Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
22. D. J. Penn, Major Histocompatibility. *Encycl. Life Sci.*, 1–7 (2002).
23. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
24. N. A. Danke, D. M. Koelle, C. Yee, S. Beheray, W. W. Kwok, Autoreactive T Cells in Healthy Individuals. *J. Immunol.* **172**, 5967–5972 (2004).
25. J. Robinson *et al.*, The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).
26. M. Yeager, A. L. Hughes, Evolution of the mammalian MHC: Natural selection,

recombination, and convergent evolution. *Immunol. Rev.* **167**, 45–58 (1999).
27. J. T. Martinsohn, A. B. Sousa, L. A. Guethlein, J. C. Howard, The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics*. **50**, 168–200 (1999).
28. F. F. González-Galarza *et al.*, 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 (2015).
29. O. D. Solberg *et al.*, Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Hum. Immunol.* **69**, 443–464 (2008).
30. P. W. Hedrick, G. Thomson, Evidence for balancing selection at HLA. *Genetics*. **104**, 449–456 (1983).
31. V. Apanius, D. Penn, P. R. Slev, L. R. Ruff, W. K. Potts, The Nature of Selection on the Major Histocompatibility Complex. *Crit. Rev. Immunol.* (1997), doi:10.1615/CritRevImmunol.v17.i2.40.
32. S. B. Piertney, M. K. Oliver, The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb).* **96**, 7–21 (2006).
33. F. Prugnolle *et al.*, Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* **15**, 1022–1027 (2005).
34. N. Qutob *et al.*, Signatures of historical demography and pathogen richness on MHC class i genes. *Immunogenetics*. **64**, 165–175 (2012).
35. E. M. Leffler *et al.*, Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science (80-.).* **340**, 1578–1582 (2013).
36. A. L. Hughes, M. Yeager, Natural Selection At Major Histocompatibility Complex Loci of Vertebrates. *Annu. Rev. Genet.* **32**, 415–435 (1998).
37. T. L. Lenz, Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution (N. Y).* **65**, 2380–2390 (2011).
38. F. Pierini, T. L. Lenz, Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection. *Mol. Biol. Evol.*, 1–14 (2018).
39. A. L. Hughes, M. Nei, Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*. **335**, 167–170 (1988).
40. W. K. Potts, E. K. Wakeland, Evolution of diversity at the major histocompatibility complex. *Trends Ecol. Evol.* **5**, 181–187 (1990).
41. E. Wakeland *et al.*, Ancestral polymorphisms of MHC class II genes: Divergent allele advantage. *Immunol. Res.* **9**, 115–122 (1990).
42. T. L. Lenz, Adaptive value of novel MHC immune gene variants. *Proc. Natl. Acad. Sci.* **115**, 201722600 (2018).
43. R. W. Slade, H. I. McCallum, Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*. **132**, 861–4 (1992).
44. J. Robinson *et al.*, Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLOS Genet.* **13**, e1006862 (2017).
45. H. G. Rammensee, T. Friede, S. Stevanović, MHC ligands and peptide motifs: first listing. *Immunogenetics*. **41**, 178–228 (1995).
46. H. G. Rammensee, Chemistry of peptides associated with MHC class I and class II molecules. *Curr. Opin. Immunol.* **7**, 85–96 (1995).
47. B. Woelfing, A. Traulsen, M. Milinski, T. Boehm, Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 117–128 (2009).
48. M. a Nowak, K. Tarczy-Hornoch, J. M. Austyn, The optimal number of major histocompatibility complex molecules in an individual. *Proc. Natl. Acad. Sci.* **89**, 10896–10899 (1992).
49. K. M. Wegner, M. Kalbe, J. Kurtz, T. B. H. Reusch, M. Milinski, Parasite selection for immunogenetic optimality. *Science (80-.).* **301**, 1343 (2003).
50. J. Fellay *et al.*, A whole-genome association study of major determinants for host control of HIV-1. *Science (80-.).* **317**, 944–947 (2007).

51. F.-R. Zhang *et al.*, Genomewide Association Study of Leprosy. *N. Engl. J. Med.* **361**, 2609–2618 (2009).
52. Y. Kamatani *et al.*, A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat. Genet.* **41**, 591–595 (2009).
53. G. Sveinbjornsson *et al.*, HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet.* **48**, 318–322 (2016).
54. J. C. Barrett *et al.*, Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
55. S. Raychaudhuri *et al.*, Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
56. R. C. Betz *et al.*, Genome-wide meta-analysis in alopecia areata resolves HLA associations and reveals two new susceptibility loci. *Nat. Commun.* **6**, 1–8 (2015).
57. J. Trowsdale, J. C. Knight, Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323 (2013).
58. T. L. Lenz *et al.*, Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).
59. T. Shiina, K. Hosomichi, H. Inoko, J. K. Kulski, The HLA genomic loci map: Expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).
60. C. Tian *et al.*, Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8** (2017), doi:10.1038/s41467-017-00257-5.
61. T. L. Lenz, V. Spirin, D. M. Jordan, S. R. Sunyaev, Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. *Mol. Biol. Evol.* **33**, 2555–2564 (2016).
62. J. H. Karnes *et al.*, Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* **9**, 1–14 (2017).
63. J. L. Mobley, Is rheumatoid arthritis a consequence of natural selection for enhanced tuberculosis resistance? *Med. Hypotheses*. **62**, 839–843 (2004).
64. R. Cagliani *et al.*, Crohn's disease loci are common targets of protozoa-driven selection. *Mol. Biol. Evol.* **30**, 1077–1087 (2013).
65. P. J. McLaren *et al.*, 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 (2015).
66. X. Hu *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).
67. R. Apps *et al.*, Influence of HLA-C Expression Level on HIV Control. *Science (80-.).* **340**, 87–91 (2013).
68. B. Krause-Kyora *et al.*, Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-03857-x.
69. V. Matzaraki, V. Kumar, C. Wijmenga, A. Zhernakova, The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18** (2017), doi:10.1186/s13059-017-1207-1.
70. G. Sorci, S. Cornet, B. Faivre, Immune Evasion, Immunopathology and the Regulation of the Immune System. *Pathogens*. **2**, 71–91 (2013).
71. M. Kalbe *et al.*, Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proc. R. Soc. B Biol. Sci.* **276**, 925–934 (2009).
72. A. Kloch, W. Babik, A. Bajer, E. Siński, J. Radwan, Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole Myodes glareolus. *Mol. Ecol.* **19**, 255–265 (2010).
73. J. Kurtz *et al.*, Major histocompatibility complex diversity influences parasite resistance and innate immunity in sticklebacks. *Proc. R. Soc. B Biol. Sci.* **271**, 197–204 (2004).
74. L. Råberg *et al.*, On the adaptive significance of stress-induced immunosuppression On the adaptive signi ® cance of stress-induced immunosuppression. *Society*, 1637–1641 (1998).
75. M. Carrington *et al.*, HLA and HIV-1: heterozygote advantage and B*35-Cw*04

disadvantage. (1999).
76. M. R. Thursz, H. C. Thomas, B. M. Greenwood, A. V. Hill, Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nat. Genet.* **17**, 11–12 (1997).
77. P. Hraber, C. Kuiken, K. Yusim, Evidence for human leukocyte antigen heterozygote advantage against hepatitis C virus infection. *Hepatology*. **46**, 1713–1721 (2007).
78. C. A. Dendrou, J. Petersen, J. Rossjohn, L. Fugger, HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
79. N. A. Patsopoulos *et al.*, Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLoS Genet.* **9** (2013), doi:10.1371/journal.pgen.1003926.
80. J. G. Abelin *et al.*, Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*. **46**, 315–326 (2017).
81. J. Neefjes, M. L. Jongsma, P. Paul, O. Bakke, Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol*. **11**, 823–36 (2011).
82. X. Rao, I. Hoof, A. I. C. A. Fontaine Costa, D. Van Baarle, C. Keşmir, HLA class i allele promiscuity revisited. *Immunogenetics*. **63**, 691–701 (2011).
83. J. Neefjes, H. Ovaa, A peptide's perspective on antigen presentation to the immune system. *Nat. Chem. Biol.* **9**, 769–775 (2013).
84. M. Nei, A. L. Hughes, Polymorphism and evolution of the major histocompatibility complex loci in mammals. *Evol. Mol. Lev.* (1991), pp. 222–247.
85. S. Sommer, The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* **2**, 1–18 (2005).
86. L. G. Spurgin, D. S. Richardson, How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. B Biol. Sci.* **277**, 979–988 (2010).
87. C. Eizaguirre, T. L. Lenz, Major histocompatability complex polymorphism: Dynamics and consequences of parasite-mediated local adaptation in fishes. *J. Fish Biol.* **77** (2010), pp. 2023–2047.
88. T. L. Lenz, Adaptive value of novel MHC immune gene variants. *Proc. Natl. Acad. Sci.*, 201722600 (2018).
89. W. E. Stutz, D. I. Bolnick, Natural selection on MHC IIβ in parapatric lake and stream stickleback: Balancing, divergent, both or neither? *Mol. Ecol.* **26**, 4772–4786 (2017).
90. M. J. Ejsmond, J. Radwan, Red Queen Processes Drive Positive Selection on Major Histocompatibility Complex (MHC) Genes. *PLoS Comput. Biol.* **11**, 1–14 (2015).
91. P. C. Doherty, R. M. Zinkernagel, Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, 50–52 (1975).
92. D. J. Penn, K. Damjanovich, W. K. Potts, MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11260–11264 (2002).
93. J. M. Blackwell, S. E. Jamieson, D. Burgner, HLA and Infectious Diseases. *Clin. Microbiol. Rev.* **22**, 370–385 (2009).
94. J. Trowsdale, The MHC, disease and selection. *Immunol. Lett.* **137**, 1–8 (2011).
95. P. E. Chappell *et al.*, Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife*. **4**, 1–22 (2015).
96. S. Takeshima *et al.*, 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 (2008).
97. S. L. O. Connor *et al.*, MHC Heterozygote Advantage in Simian Immunodeficiency Virus – Infected Mauritian Cynomolgus Macaques. **2** (2010).
98. M. L. Evans, B. D. Neff, Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (Oncorhynchus tshawytscha). *Mol. Ecol.* **18**, 4716–4729 (2009).
99. A. K. Niskanen *et al.*, Balancing selection and heterozygote advantage in major histocompatibility complex loci of the bottlenecked Finnish wolf population. *Mol. Ecol.* **23**, 875–889 (2014).
100. T. L. Lenz, B. Mueller, F. Trillmich, J. B. W. Wolf, 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** (2013), doi:10.1098/rspb.2013.0714.
101. C. Landry, D. Garant, P. Duchesne, L. Bernatchez, "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 (2001).
102. T. L. Lenz, K. Wells, M. Pfeiffer, S. Sommer, 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 (2009).
103. A. D. Richman, L. G. Herrera, D. Nash, MHC class II beta sequence diversity in the deer mouse (Peromyscus maniculatus): Implications for models of balancing selection. *Mol. Ecol.* **10**, 2765–2773 (2001).
104. B. D. Neff, S. R. Garner, J. W. Heath, D. D. Heath, The MHC and non-random mating in a captive population of Chinook salmon. *Heredity (Edinb).* **101**, 175–185 (2008).
105. D. Meyer, V. R. C. Aguiar, B. D. Bitarello, D. Y. C. Brandt, K. Nunes, A genomic perspective on HLA evolution. *Immunogenetics*. **70**, 5–27 (2018).
106. M. Carrington, S. J. O'Brien, The Influence of *HLA* Genotype on AIDS. *Annu. Rev. Med.* **54**, 535–551 (2003).
107. J. W. Mellors *et al.*, Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science (80-.).* **272**, 1167–1170 (1996).
108. A. Stenzel *et al.*, Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum. Genet.* **114**, 377–385 (2004).
109. A. Blomhoff *et al.*, 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 (2006).
110. T. Ahmad *et al.*, Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* **12**, 647–656 (2003).
111. P. Parham, P. J. Norman, L. Abi-Rached, L. A. Guethlein, 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 (2012).
112. S. Buhler, J. M. Nunes, A. Sanchez-Mazas, HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*. **68**, 401–416 (2016).
113. C. Bronke *et al.*, HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes. *Aids*. **27**, 899–905 (2013).
114. J. R. Bailey, T. M. Williams, R. F. Siliciano, J. N. Blankson, Maintenance of viral suppression in HIV-1-infected HLA-B*57+ elite suppressors despite CTL escape mutations. *J. Exp. Med.* **203**, 1357–69 (2006).
115. S. A. Migueles *et al.*, 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 (2000).
116. C. W. Pohlmeyer, R. W. Buckheit, R. F. Siliciano, J. N. Blankson, CD8+T cells from HLA-B*57 elite suppressors effectively suppress replication of HIV-1 escape mutants. *Retrovirology*. **10**, 27–32 (2013).
117. P. Parham, MHC class I molecules and KIRS in human history, health and survival. *Nat. Rev. Immunol.* **5**, 201–214 (2005).
118. M. Colonna, G. Borsellino, M. Falco, G. B. Ferrara, J. L. Strominger, 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 (1993).
119. S. Rajagopalan, E. O. Long, Understanding how combinations of HLA and KIR genes influence disease. *J. Exp. Med.* **201**, 1025–1029 (2005).
120. D. Zipeto, A. Beretta, HLA-C and HIV-1 : friends or foes ?, 1–9 (2012).
121. C. Körner *et al.*, HIV-1-Mediated Downmodulation of HLA-C Impacts Target Cell Recognition and Antiviral Activity of NK Cells. *Cell Host Microbe*. **22**, 111–119.e4 (2017).
122. F. Pereyra *et al.*, HIV Control Is Mediated in Part by CD8+ T-Cell Targeting of Specific Epitopes. *J. Virol.* **88**, 12937–12948 (2014).
123. The International HapMap 3 Consortium, Integrating common and rare genetic variation

in diverse human populations. *Nature*. **467**, 52–58 (2010).
124. B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G. R. Abecasis, Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
125. O. Delaneau, J.-F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*. **10**, 5–6 (2013).
126. B. N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. **5**, e1000529 (2009).
127. T. Jacks *et al.*, Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*. **331**, 280–283 (1988).
128. K. Falk, O. Rötzschke, S. Stevanović, G. Jung, H. G. Rammensee, Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*. **351**, 290–296 (1991).
129. I. A. York *et al.*, The Er aminopeptidase ERAP I enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat. Immunol.* **3**, 1177–1184 (2002).
130. V. Jurtz *et al.*, NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).
131. B. L. Aken *et al.*, The Ensembl gene annotation system. *Database (Oxford).* **2016**, 1–19 (2016).
132. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
133. R. Grantham, Amino Acid Difference Formula to Help Explain Protein Evolution Amino Acid Difference Formula to Help Explain Protein Evolution. *Science (80-.).* **185**, 862–864 (1974).
134. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
135. R. Gouveia-Oliveira, P. W. Sackett, A. G. Pedersen, MaxAlign: Maximizing usable data in an alignment. *BMC Bioinformatics*. **8**, 1–8 (2007).
136. S. Guindon, O. Gascuel, A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52**, 696–704 (2003).
137. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*. **20**, 289–290 (2004).
138. H. Wickham, M. H. Wickham, The ggplot package (2007).
139. J. Neefjes, M. L. M. Jongsma, P. Paul, O. Bakke, Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
140. F. M. Key, J. C. Teixeira, C. de Filippo, A. M. Andrés, Advantageous diversity maintained by balancing selection in humans. *Curr. Opin. Genet. Dev.* **29**, 45–51 (2014).
141. L. B. Barreiro, L. Quintana-Murci, From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).
142. S. L. O'Connor *et al.*, MHC heterozygote advantage in simian immunodeficiency virus-infected Mauritian cynomolgus macaques. *Sci. Transl. Med.* **2** (2010), doi:10.1126/scitranslmed.3000524.
143. S. M. Weenink, M. R. Christie, Autoantibodies in Diabetes. *Autoantibodies Autoimmun. Mol. Mech. Heal. Dis.* **54**, 321–349 (2006).
144. V. Lampasona, D. Liberati, Islet Autoantibodies. *Curr. Diab. Rep.* **16** (2016), doi:10.1007/s11892-016-0738-2.
145. S. C. Kent *et al.*, Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope. *Nature*. **435**, 224–228 (2005).
146. B. Stadinski, J. Kappler, G. S. Eisenbarth, Molecular Targeting of Islet Autoantigens. *Immunity*. **32**, 446–456 (2010).
147. D. Daniel, R. G. Gill, N. Schloot, D. Wegmann, Epitope specificity, cytokine production profile and diabetogenic activity of insulin-specific T cell clones isolated from NOD mice. *Eur. J. Immunol.* (1995), doi:10.1002/eji.1830250430.

148. B. Roep *et al.*, Antigen Targets of Type 1 Diabetes. *Cold Spring Harb Perspect Med Perspect*. **2**, 1–14 (2012).
149. M. Manczinger, L. Kemény, Peptide presentation by HLA-DQ molecules is associated with the development of immune tolerance. *PeerJ*. **6**, e5118 (2018).
150. A. Katsarou *et al.*, Type 1 diabetes mellitus. *Nat. Rev. Dis. Prim.* **3**, 1–18 (2017).
151. H. Miyadera, J. Ohashi, Å. Lernmark, T. Kitamura, K. Tokunaga, Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA. *J. Clin. Invest.* **125**, 275–291 (2015).
152. J. Cárdenas-Roldán, A. Rojas-Villarraga, J. M. Anaya, How do autoimmune diseases cluster in families? A systematic review and meta-analysis. *BMC Med.* **11**, 73 (2013).
153. P. Cruz-Tapias, J. Castiblanco, J.-M. Anaya, HLA association with autoimmune diseases. *Autoimmun. From Bench to Bedside.*, 271–284 (2013).
154. K. K. Jensen *et al.*, Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, 394–406 (2018).
155. M. L. Marré, E. A. James, J. D. Piganelli, β cell ER stress and the implications for immunogenicity in type 1 diabetes. *Front. Cell Dev. Biol.* **3**, 1–14 (2015).
156. S. M. Anderton, Post-translational modifications of self antigens: Implications for autoimmunity. *Curr. Opin. Immunol.* **16**, 753–758 (2004).
157. S. G. Fonseca, J. Gromada, F. Urano, Endoplasmic reticulum stress and pancreatic β-cell death. *Trends Endocrinol. Metab.* **22**, 266–274 (2011).
158. R. J. McLaughlin *et al.*, Human islets and dendritic cells generate post-translationally modified islet autoantigens. *Clin. Exp. Immunol.* **185**, 133–140 (2016).
159. M. Van Lummel *et al.*, Posttranslational modification of HLA-DQ binding islet autoantigens in type 1 diabetes. *Diabetes*. **63**, 237–247 (2014).
160. J. W. McGinty *et al.*, Recognition of posttranslationally modified GAD65 epitopes in subjects with type 1 diabetes. *Diabetes*. **63**, 3033–3040 (2014).
161. L. W. Vader *et al.*, Specificity of tissue transglutaminase explains cereal toxicity in celiac disease. *J. Exp. Med.* **195**, 643–9 (2002).
162. S. Tollefsen *et al.*, HLA-DQ2 and-DQ8 signatures of gluten T cell epitopes in celiac disease. *J. Clin. Invest.* **116**, 2226–2236 (2006).
163. B. H. Johansen *et al.*, Both α and β chain polymorphisms determine the specificity of the disease-associated HLA-DQ2 molecules, with β chain residues being most influential. *Immunogenetics*. **45**, 142–150 (1996).
164. N. A. Bowerman, L. A. Colf, K. C. Garcia, D. M. Kranz, Different strategies adopted by Kb and Ld to generate T cell specificity directed against their respective bound peptides. *J. Biol. Chem.* **284**, 32551–32561 (2009).
165. B. Engels *et al.*, Relapse or eradication of cancer is predicted by peptide-major histocompatibility complex affinity. *Cancer Cell*. **23**, 516–526 (2013).
166. M. Knip, O. Simell, Environmental triggers of type 1 diabetes. *Cold Spring Harb. Perspect. Biol.* **3**, 1–15 (2011).
167. U. Christen, C. Bender, M. G. Von Herrath, Infection as a cause of type 1 diabetes? *Curr. Opin. Rheumatol.* **24**, 417–423 (2012).
168. K. T. Coppieters, T. Boettler, M. von Herrath, Virus Infections in Type 1 Diabetes. *Cold Spring Harb. Perspect. Med.* **2**, a007682–a007682 (2012).
169. R. S. Fujinami, M. G. Von Herrath, U. Christen, J. L. Whitton, Molecular Mimicry , Bystander Activation , or Viral Persistence : Infections and Autoimmune Disease Molecular Mimicry , Bystander Activation , or Viral Persistence : Infections and Autoimmune Disease. *Clin. Microbiol. Rev.* **19**, 80–94 (2006).
170. H. A. Doyle, M. J. Mamula, Post-translational protein modifications in antigen recognition and autoimmunity. *Trends Immunol.* **22**, 443–449 (2001).
171. X. Jia *et al.*, Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One*. **8** (2013), doi:10.1371/journal.pone.0064683.
172. A. Bateman *et al.*, UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
173. I. D. Federation, *Diabetes Atlas 2015* (2015).

174. T. P. Di Lorenzo, M. Peakman, B. O. Roep, Translational mini-review series on type 1 diabetes: Systematic analysis of T cell epitopes in autoimmune diabetes. *Clin. Exp. Immunol.* **148**, 1–16 (2007).
175. M. A. Degli-Esposti *et al.*, Ancestral haplotypes: conserved population MHC haplotypes. *Hum. Immunol.* **34**, 242–252 (1992).
176. C. A. Alper *et al.*, The Haplotype Structure of the Human Major Histocompatibility Complex. *Hum. Immunol.* **67**, 73–84 (2006).
177. M. van Lummel *et al.*, Discovery of a Selective Islet Peptidome Presented by the Highest-Risk HLA-DQ8 trans Molecule. *Diabetes*. **65**, 732–741 (2016).
178. H. A. Doyle, M. J. Mamula, Autoantigenesis: The evolution of protein modifications in autoimmune disease. *Curr. Opin. Immunol.* **24**, 112–118 (2012).
179. M. A. Atkinson *et al.*, Cellular Immunity to a determinant common to glutamate decarboxylase and Coxsackie virus in insulin dependent diabetes. *J. Clin. Invest.* **94**, 2125–2129 (1994).
180. M. Mansilla, X. Montalban, Heat Shock Protein 70: Roles in Multiple Sclerosis. *Mol. Med.* **18**, 1 (2012).
181. M. J. Muraro *et al.*, A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **3**, 385–394 (2016).
182. A. Volchuk, D. Ron, The endoplasmic reticulum stress response in the pancreatic β-cell. *Diabetes, Obes. Metab.* **12**, 48–57 (2010).
183. C. L. Vanderlugt, S. D. Miller, Epitope spreading in immune-mediated diseases: Implications for immunotherapy. *Nat. Rev. Immunol.* **2**, 85–95 (2002).
184. S. Masala *et al.*, Antibodies recognizing mycobacterium avium paratuberculosis epitopes cross-react with the beta-cell antigen znt8 in sardinian type 1 diabetic patients. *PLoS One*. **6**, 4–13 (2011).
185. H.-S. Jun, J.-W. Yoon, A new look at viruses in type 1 diabetes. *Diabetes. Metab. Res. Rev.* **19**, 8–31 (2003).
186. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
187. C. Hammer *et al.*, Amino acid variation in HLA class II proteins is a major determinant of humoral response to common viruses. *Am. J. Hum. Genet.* **97**, 738–743 (2015).
188. Y. Okada *et al.*, Fine mapping major histocompatibility complex associations in psoriasis and its clinical subtypes. *Am. J. Hum. Genet.* **95**, 162–172 (2014).
189. F. Barré-Sinoussi, A. L. Ross, J. F. Delfraissy, Past, present and future: 30 years of HIV research. *Nat. Rev. Microbiol.* **11**, 877–883 (2013).
190. M. Knip, O. Simell, Enviornmental Triggers of Type 1 Diabetes. *Cold Spring Harb. Perspect. Med.* **2**, a007690 (2012).
191. C. Hunt, R. I. Morimoto, Conserved features of eukaryotic hsp70 genes revealed by comparison with the nucleotide sequence of human hsp70. *Proc. Natl. Acad. Sci.* **82**, 6455–6459 (1985).
192. S. Lindquist, E. A. Craig, The Heat-Shock Proteins. *Annu. Rev. Genet.* **22**, 631–677 (1988).
193. Y. Kirino *et al.*, Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nat. Genet.* **45**, 202–207 (2013).
194. P. Chappell *et al.*, Expression levels of mhc class i molecules are inversely correlated with promiscuity of peptide binding. *Elife*. **2015**, 1–22 (2015).

Annex

Chapter 1

Dynamic allelic expression during T cell activation uncovers *cis* regulatory effects in HLA-DQB1 and autoimmune disease genes

Maria Gutierrez-Arcelus[1,2,3]*, Yuriy Baglaenko[1,2,3]*, Susan Hannes[1,2,3], Jatin Arora[4], Nikola Teslovich[1,2,3], Yang Luo[1,2,3], Kamil Slowikowski[1,2,3], Tiffany Amariuta[1,2,3], Harm-Jan Westra[1,2,3], Deepak A. Rao[3], Joerg Ermann[3], Anna H. Jonsson[3], Tonu Esko[5], Michael B. Brenner[3], Soumya Raychaudhuri[1,2,3,6]

1.Division of Genetics and Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
2. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.
3. Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, BTM, 60 Fenwood Road, Boston, MA 02115, USA.
4. Max Planck Institute for Evolutionary Biology, Plön, Germany, 24306.
5. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
6. Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK.
*Equal contribution

Introduction

Understanding the mechanisms by which genetic factors contribute to autoimmune disease is key for designing interventions (1). Most of the hundreds of susceptibility variants for autoimmune disease are of small effects, non-coding and enriched in regulatory elements specific for CD4+ T cells (2, 3). Hence, it is hypothesized that causal alleles affect gene regulation in CD4+ T cells. Even within the region contributing most to disease risk, the Major Histocompatibility (MHC) locus, amino acid changes in Human Leukocyte Antigen (HLA) genes do not fully explain all the risk at that locus (4), and regulatory variation could also play a role. Studies quantifying transcriptomes in hundreds of genotyped individuals have identified genetic variants affecting gene expression (expression quantitative trait loci, eQTLs) in T cells (5–10) . However, only a small proportion of fine-mapped autoimmune risk alleles have been reported to have an effect on gene expression in these studies (11), which are limited to T cells mostly in resting state and sometimes at one or two stimulated states. Recent studies have identified causal alleles from two loci that exert their regulatory function only under

certain T cell stimulated states (12, 13). Hence, ascertaining the genetic regulatory effects across multiple time points during CD4+ T cell activation may be key to elucidate autoimmune mechanisms. Since allelic expression is largely driven by cis genetic regulatory variation (14), it can be exploited to define regulatory effects in multiple cell states without the need of a large sample size (15).

Results

Here, we looked for cell-state dependent *cis* regulatory effects by identifying allele specific expression that changes with time upon stimulation of $CD4^+$ memory T cells. We isolated $CD4^+$ memory T cells from peripheral blood of 24 genotyped individuals of European ancestry, for which imputation was performed (fig. S1). We stimulated cells with anti-CD3/CD28 beads and performed high input and deep RNA sequencing at up to eight time points, specifically at: 0, 2, 4, 8, 12, 24, 48 and 72 hours (**Fig. 1A**). We performed full time course replicates in two of our individuals. Principal component analysis on gene expression levels shows the 200 samples separate well by time point (**Fig. 1B,** fig. S2). We clustered genes by expression profile across time and identified expected early, intermediate and late activation clusters, as well as pulsed and sustained repression clusters (fig. S3). We quantified allele specific expression at heterozygous sites of each individual, at each of the ascertained time points (fig. S4). Requiring 20 minimum reads per site with both alleles seen in at least four time points per individual, we queried a total of 225,924 events, representing 38,890 unique SNPs in 8,322 genes and some in transcribed intergenic regions (3%).

In order to identify allele specific expression and its time dependency, we used a logistic regression framework, where each read gets assigned a 1 for reference allele or 0 for alternative allele. We first tested a conservative nested approach to identify high confidence dynamic allele-specific expression (ASE) events in two pilot individuals for which we performed full time course replicates. First, for each heterozygote site in an individual, we merged counts from all time points and identified 1,484 sites with evidence of ASE (intercept $P < 2.8e-06$, Bonferroni threshold). Next, for those sites we looked for cases where ASE changes with time by fitting a polynomial model with time point as an ordinal variable. We noticed an inflation of low P-values, perhaps driven by the over-dispersed nature of allelic counts (fig. S5A). To account for this, we treated sample as a random effect and compared it to a mixed effect model with sample as random effect and a polynomial term for time (fig. S5B). This yielded 64 dynamic ASE calls ($P < 3.7e-03$, <5%

FDR) for which we assessed replication in an independent time course experiment for each individual. Dynamic ASE sites had a strong enrichment of low P-values in their replicate (~70% of cases with P < 0.05), and a high correlation of effects (beta) for time (Spearman rho: 0.92 and 0.86) and time squared (Spearman rho: 0.51 and 0.68; fig S6). Having shown good replication, we applied the same approach for all 24 individuals. From 207,519 events tested for the whole cohort, 15,268 had evidence of ASE (P < 2.4e-07). Of these, 561 had significant dynamic ASE events (P < 3.2e-03, <5% FDR), spanning 356 unique SNPs in 186 genes and 7 intergenic sites (**Fig. 1C**). Comparing with previous eQTL studies in T cells, we found that 32-60% of our dynamic ASE genes have been reported to have an eQTL in resting T cells in previous studies (*5, 7, 9*), with an enrichment for T cell specific eQTLs (fig. S8), which indicates we are able to capture known *cis* regulatory genetic effects.

We observed a great amount of significant dynamic ASE events in the MHC locus (**Fig. 1C**), with 15 events for the gene HLA-DQB1. This gene is part of the HLA class II genes that code for antigen presenting proteins, and is the major risk gene for Type 1 Diabetes and Celiac Disease (*4, 16*). Besides being expressed in antigen presenting cells, HLA class II proteins are also observed in activated T cells, although their function in that cell type remains to be established (*17*). **Fig. 1D** shows an example of an individual with total expression level of DQB1 that slightly increases after stimulation, then gets down-regulated at 8 hours and gets re-activated at 48 and 72 hours (left panel). If we look at the allelic read counts in a heterozygous SNP in exon 6 of the HLA-DQB1 gene (middle panel) and the proportion of read counts coming from the reference allele (right panel), we can observe that the reference allele is more abundant in the early time points, whereas at 48 and 72 hours there is an allelic switch where the alternative allele strongly dominates (P = 2.3e-05).

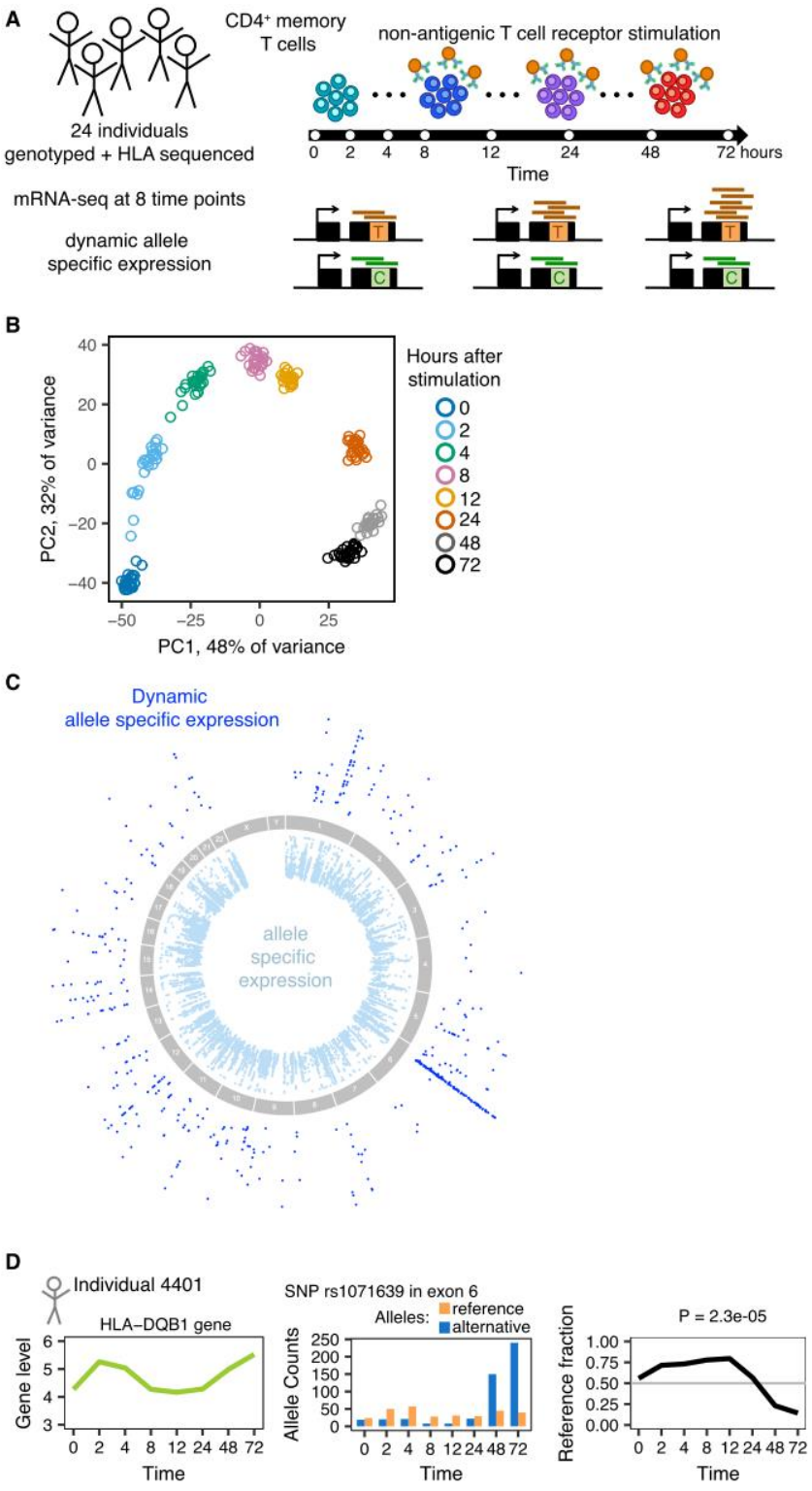


Fig. 1 Dynamic allele specific expression during T cell activation. (A) Study design. (B) Principal Component Analysis on top 1000 most variable genes. Shown are PC1 and PC2 scores for the 200 samples colored by time point. (C) Plot showing position across the genome of allele specific expression (ASE) events (light blue), and dynamic ASE events (dark blue). Y-axis for dynamic ASE events if FDR. (D) Example of dynamic ASE event for a SNP in HLA-DQB1. For each time point, we show the level of expression of HLA-DQB1 (left), allele counts for the SNP (middle) and fraction of reads with the reference allele (right).

Since HLA genes are highly polymorphic and this can create mapping bias, we developed a computational pipeline to accurately quantify allele specific expression in HLA genes. We performed next generation targeted sequencing for HLA-DQB1, HLA-DRB1, HLA-A, HLA-B and HLA-C, and high-resolution PCR for HLA-DQA1, for all 24 individuals, revealing the classical alleles for each HLA gene. We then built an HLA personalized genome for each individual (**Fig. 2A**), where we added to the reference genome the cDNA allelic sequences for each HLA gene, and masked the corresponding exonic coordinates in the reference genome. We then mapped the reads to this HLA personalized genome and quantified the number of uniquely mapped read to each HLA allele. Given our initial observations in a pilot study, these 24 individuals were purposely pre-selected to be heterozygous for HLA-DQB1 with at least 20 mismatches between the two alleles present at each individual. Among the 48 HLA-DQB1 ~780bp sequences, there are 14 HLA-DQB1 4-digit classical alleles. As shown in **Fig. 2B-C**, allelic counts for the whole HLA-DQB1 allele recapitulate those patterns observed for the individual SNP shown in **Fig. 1D**. However, this was less often the case for SNPs found in middle exons that are more polymorphic (show this in fig. S9?). Overall, our strategy allowed for accurate allele specific quantification over >20 SNPs in four HLA-DQB1 exons which is more robust than for individual SNPs.

Using our mixed effects logistic regression framework on the HLA-DQB1 allelic counts, we determined that 15 out of 24 individuals have significant dynamic ASE for HLA-DQB1 ($P < 0.002$, Bonferroni threshold). **Fig. 2C** depicts profiles of allelic expression levels (normalized by library size) for three individuals, with their respective allelic fraction patterns over time. Clustering of these allelic profiles revealed three main groups (**Fig. 2D**, shown also with PCA on fig. S10A-B). HLA-DQB1 4-digit alleles cluster together more than expected by chance (permutation $P < 0.001$, fig S10C), suggesting *cis* regulatory variants often track with HLA-DQB1 classical allele haplotypes. We named these three HLA-DQB1 *cis* regulatory programs based on their expression dynamics: Late Spike, Low Constant and Fluctuating (**Fig. 2D**).

We then asked whether the clear up-regulation of mRNA levels in the Late Spike *cis* regulatory program compared to the other two programs translated to the protein level. To do this, we recruited 5 homozygous individuals for HLA-DQB1 4-digit alleles pertaining to the Late Spike *cis* regulatory program (one DQB1*05:03 and four DQB1*05:01 individuals), and 5 homozygous individuals for HLA-DQB1 4-digit alleles present in the Low Constant or Fluctuating *cis* regulatory programs (one DQB1*0302, one DQB1*0201, one DQB1*0202 and two DQB1*0301 individuals). As before, we isolated

peripheral CD4+ memory T cells and stimulated them with anti-CD3/CD28 beads. We quantified protein levels with flow cytometry for HLA-DQ (the protein complex involving HLA-DQB1 and HLA-DQA1), at 0, 1, 3 and 7 days after stimulation. As shown in **Fig. 2E**, protein levels of HLA-DQ (Median Fluorescence Intensity, MFI) in HLA-DQ+ cells recapitulate HLA-DQB1 mRNA levels. Interestingly, the difference in HLA-DQ levels between the 2 groups becomes statistically significant only at 7 days after stimulation, indicating that protein levels take longer to take off. Other metrics, such as percent HLA-DQ+ cells, or fold change in HLA-DQ+ with respect to 0 hours, presented similar patterns across time (fig. S11). Overall, these results show that the mRNA dynamics of the Late Spike *cis* regulatory program for HLA-DQB1 translate to the protein levels, raising the possibility of differences in CD4+ memory T cell function among individuals.

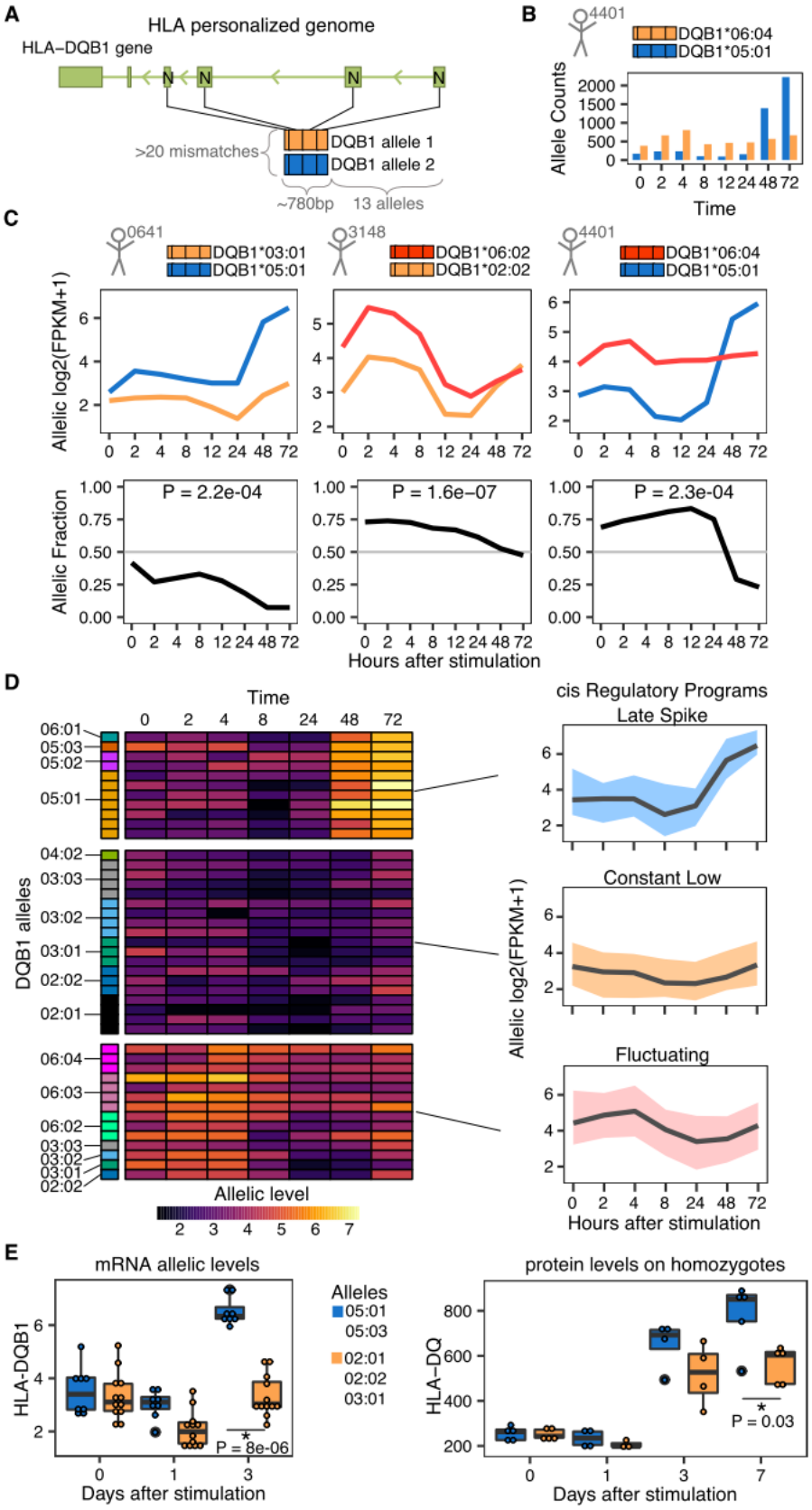


Fig. 2 HLA-DQB1 dynamic ASE at mRNA and protein levels. (A) Scheme of HLA personalized pipeline and study design for capturing highly divergent HLA-DQB1 alleles within each individual. (B) HLA-DQB1 allele counts for individual 4401 along the time course. (C) Normalized allelic expression for HLA-DQB1 (top), and allelic fraction (bottom) for three individuals. (D) Heatplot shows normalized allelic expression levels

(log2(FPKM+1)) for each of the 48 HLA-DQB1 alleles in hour cohort. Allelic profiles were clustered into three *cis* Regulatory Programs, for which average expression profile is shown on the right with black line, and total expression area occupied by all alleles in that cluster shown with the colored ribbon. (E) Left panel shows normalized allelic expression levels (log2(FPKM+1)). Right panel shows protein levels (median fluorescence intensity of HLA-DQ+ CD4+ memory T cells) for 5 homozygote individuals for alleles within the Late-Spike regulatory program (blue) and 5 homozygote individuals for alleles in other programs (yellow).

Next, we sought to fine-map the genetic variant driving the Late Spike *cis* regulatory program with genetic and epigenetic tools. To do this, we used 2,198 fully sequenced Estonian genomes from which we called SNP genotypes and HLA 4-digit alleles. We then looked for SNPs that were in tight linkage disequilibrium (LD) with the Late Spike HLA-DQB1 4-digit alleles (i.e. 05:01, 05:02, 05:03, 06:01), that together represent what we hereon call the Late Spike haplotype. Specifically, we calculated Pearson correlation coefficient between SNP genotypes and Late Spike haplotype dosage (0, 1 or 2), for 27,210 SNPs with 5% minor allele frequency (MAF) that are within 1Mb of HLA-DQB1 transcription start site (TSS). As expected, most of the SNPs with the highest correlation coefficient (0.98-0.99) are within HLA-DQB1 gene (91%), and the correlation decays with distance to the gene (**Fig. 3A**). There were six intergenic SNPs with r=0.98. One SNP, rs71542466, is 34bp upstream from the TSS of HLA-DQB1. There are 4 intergenic SNPs further away from the TSS (206-430bp), and one SNP 52bp downstream of the end of the gene. An eQTL analysis on our limited set of 24 individuals confirmed these 6 intergenic SNPs to be within the top 5 most significant P values (P = 3.8e-08) explaining 76% of HLA-DQB1 expression variance at 72 hours (**Fig. 3B**, fig. S12). In order to look for open chromatin regions, we performed ATAC-seq in $CD4^+$ memory T cells stimulated for 72 hours (as described previously) from one donor, in replicate. Out of the three visible peaks in the HLA-DQB1region, the highest peak is located at the promoter and overlaps the SNP rs71542466 (**Fig. 3C**). The other four promoter-proximal SNPs are outside of the peak region, and the end-of-gene SNP does not overlap any open chromatin peak. Looking at published data, the promoter SNP rs71542466 also overlaps ChIP-seq chromatin marks in primary memory T cells and PMA Iionomycin stimulated Th17 cells that typically mark promoters and enhancers (*18*). Interestingly, the promoter SNP overlaps binding sites for both the transcription factor RFX5 (*19*) (in LCLs) and the co-activator CIITA (*20*) (B-cells), which are major regulators of HLA class II genes. Together, these results indicate that the

promoter SNP rs71542466 is the most likely variant driving the Late Spike *cis* regulatory program.

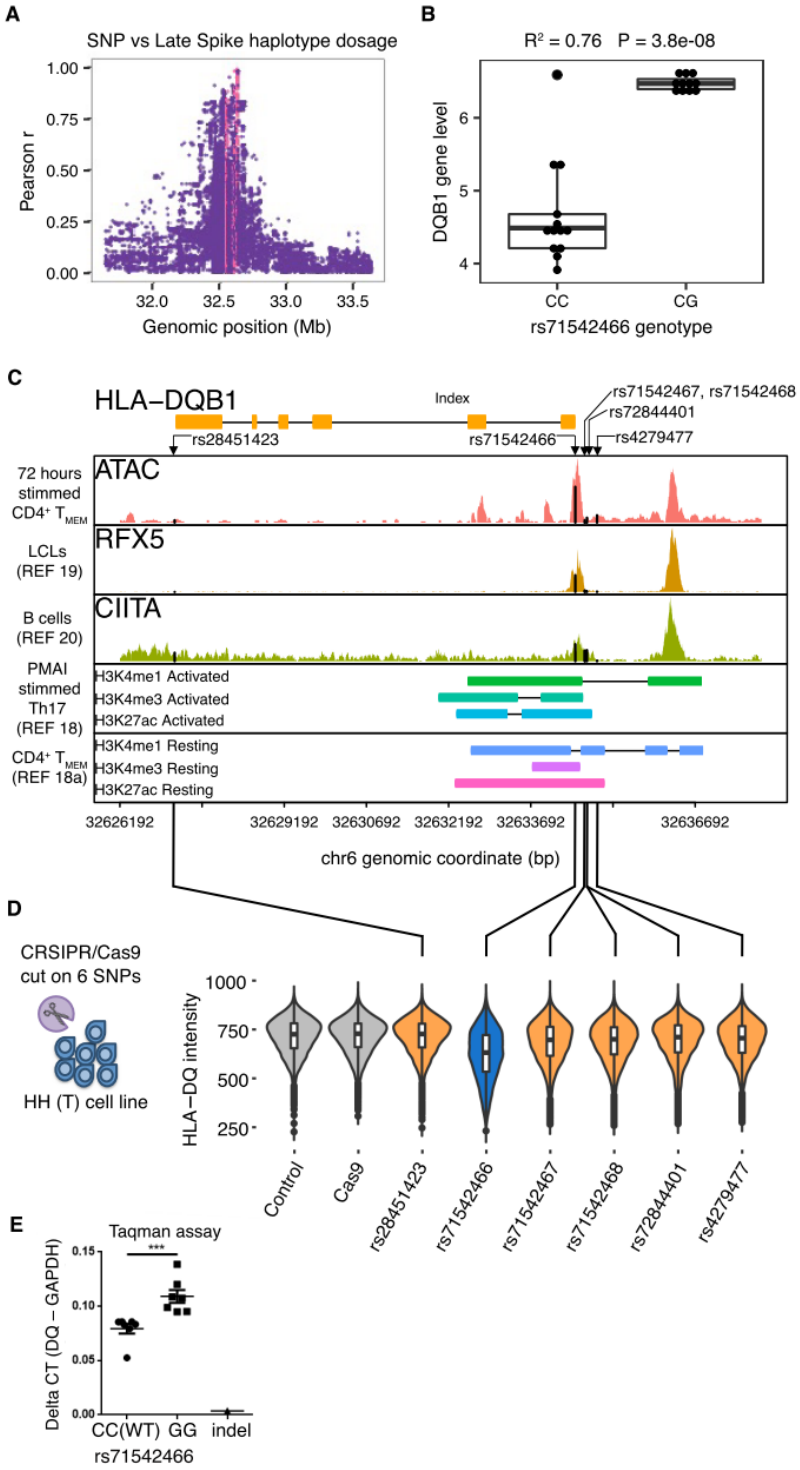


Fig. 3 Fine-mapping and validation of causal variant for Late-Spike *cis* regulatory program. (A) Pearson correlation coefficient between SNP genotypes and Late-Spike allele dosage. Orange vertical lines indicate location of HLA-DRB1, HLA-DQB1 and HLA-DQA1 genes. (B) DQB1 gene expression levels (log2(FPKM+1)) for individuals separated by their rs71542466 genotype. (C) Location of 6 fine-mapped non-coding SNPs around

HLA-DQB1. Tracks showing open chromatin regions (ATAC-seq) or regions bound by RFXA, CIITA or markers by histone modifications (ChIP-seq). (D) CRISPR/Cas9 cuts at or near six fine-mapped SNPs in HH T cell lines. Shown is HLA-DQ median fluorescence intensity. (E) Relative HLA-DQB1 expression levels measured with Taqman assay, performed for 8 wilt type (WT, CC genotype), and 8 SNP edited (GG genotype) expanded clones for rs71542466, as well as a cell line clone with an indel at the same target position.

We hypothesized that among the 6 intergenic SNPs with strongest correlation with the Late Spike haplotype, the top candidate promoter SNP rs71542466 is the causative one. In order to test this, we used CRISPR/Cas9 editing in the MHC class II expressing T cell line HH. First, we Sanger sequenced ~2kb regions at the start and end of HLA-DQB1 gene, covering the 6 SNPs of interest, and determined their genotypes. We designed probes to cut at or as close as possible to the 6 SNPs to assess the regulatory potential of each region in a heterogeneous mixture of cells with varying insertions and deletions. We observed that the only cut that caused a significant decrease in expression of HLA-DQ was that designed for the top candidate promoter SNP rs71542466 (**Fig. 3D**), indicating that this SNP is in an important regulatory region, whereas the other SNPs are not. As a next step, we performed targeted base-editing providing HDR donor ssODN templates while cutting using CRISPR/Cas9 RNPs. Asymmetric ssDNA donors of 120nt, as previously described, were used to maximize efficiency. After editing, we single cell sorted and expanded clones. Sanger sequencing was performed after outgrowth to find clones with the exact desired SNP substitution. In the case for rs71542466, the cell lines HH was homozygous for the reference allele (WT), which is expected to have low expression. Hence, clones with the alternative allele (non-WT) were expected to have higher expression of HLA-DQB1. Among 192 sequenced independent clones, we found 8 that are homozygous for the alternative allele. All of them showed higher HLA-DQB1 expression compared to the wild type (reference allele, **Fig. 3E**, fig S13). And indels at the same region showed a drastic decrease in HLA-DQB1 expression, as expected. Overall, these results show that the promoter SNP rs71542466 causes a change in HLA-DQB1 expression in the expected direction in a T cell line, indicating it is the causal SNP for the Late Spike *cis* regulatory program.

We then analyzed the dynamic ASE events in the rest of the genome. First, we asked whether for a given SNP, its allelic imbalance (reference fraction distance from 0.5) is associated with expression levels of its gene across time. We found there is a bimodal distribution of spearman rank correlation coefficients between allelic imbalance and gene

expression levels (**Fig. 4A**). This indicates that as gene expression increases across time, the allelic imbalance tends to change either negatively or positively, as shown in the examples for genes F5 and CTSL, respectively (**Fig. 4B**).

We found that multiple of our dynamic ASE genes are in non-MHC autoimmune disease loci. For example, gene NDUFAF1 is in a locus associated with Ulcerative Colitis and Cohn's Disease. In T cell activation, NDUFAF1 is necessary for expression and secretion of IL-2 and IL-4 (*21*). In our time course, NDUFAF1 is a late activation gene, with expression peaking at 24 hours. The SNP rs3204853 in the NDUFAF1 gene shows both of its alleles are expressed at similar levels initially, and the reference allele starts gradually dominating at 8 hours and peaking at 72 hours (dynamic ASE P = 3.2e-04, **Fig. 4C**). Another example is the PLEK gene that is in a locus associated with Celiac Disease and Multiple Sclerosis. PLEK codes for Plekstrin, the platelet and leukocyte C kinase substrate, which has been implicated in exocytosis (*22*). In our time course, PLEK is up-regulated early at 2 hours with a slow decay after that. The SNP rs1063479 starts with the reference allele being slightly more expressed than the alternative, and makes a drastic allelic switch at 2 hours with the alternative allele strongly dominating for the rest of the time course (dynamic ASE P = 1.9e-03, **Fig. 4D**).

We evaluated whether dynamic ASE genes are significantly enriched in autoimmune disease loci. To do this, we compared the number of dynamic ASE genes within risk loci of each of 13 autoimmune diseases with those found within 1000 null sets of loci matched for number of genes per locus. We found that dynamic ASE genes are significantly enriched for risk loci for type 1 diabetes (OR=6.6), inflammatory bowel disease (OR=5.2), rheumatoid arthritis (OR=4.1), ankylosing spondylitis (OR=5.9), vitiligo (OR=5.7), primary biliary cirrhosis (OR=4.5), multiple sclerosis (OR=3.9), juvenile idiopathic arthritis (OR=3.3) and psoriasis (OR=2.2; all P <= 0.001). Whereas they are not enriched for non-immune mediated diseases such as schizophrenia (OR=0.8), coronary heart disease (OR=0) and type 2 diabetes (OR=0.9) (**Fig. 4E**). Similar enrichments have been previously reported for T cell eQTLs (5, 7). To see whether autoimmune disease genes are particularly enriched for dynamic regulatory variation compared to other types of regulatory effects where less conditions are captured, we called ASE events that changed between 0 hours and 24 hours using our logistic regression framework with time as continuous variable, we called ASE events significant at 0 hours only, and we used published eQTLs for resting CD4+ naïve T cells from the BLUEPRINT consortium. We found that our dynamic ASE genes, spanning up to 8 different cellular states, are the most

enriched for autoimmune disease genes, followed by condition specific regulatory effects between 0 and 24 hours, regulatory effects at 0 hours only in our CD4+ memory T cells, and finally eQTLs found in resting CD4+ naïve T cells (**Fig. 4F**). These results show autoimmune disease *cis* regulatory variation is highly cell-state dependent and highlight the importance on ascertaining allele dynamics in multiple cellular states in order to understand the mechanisms of disease.

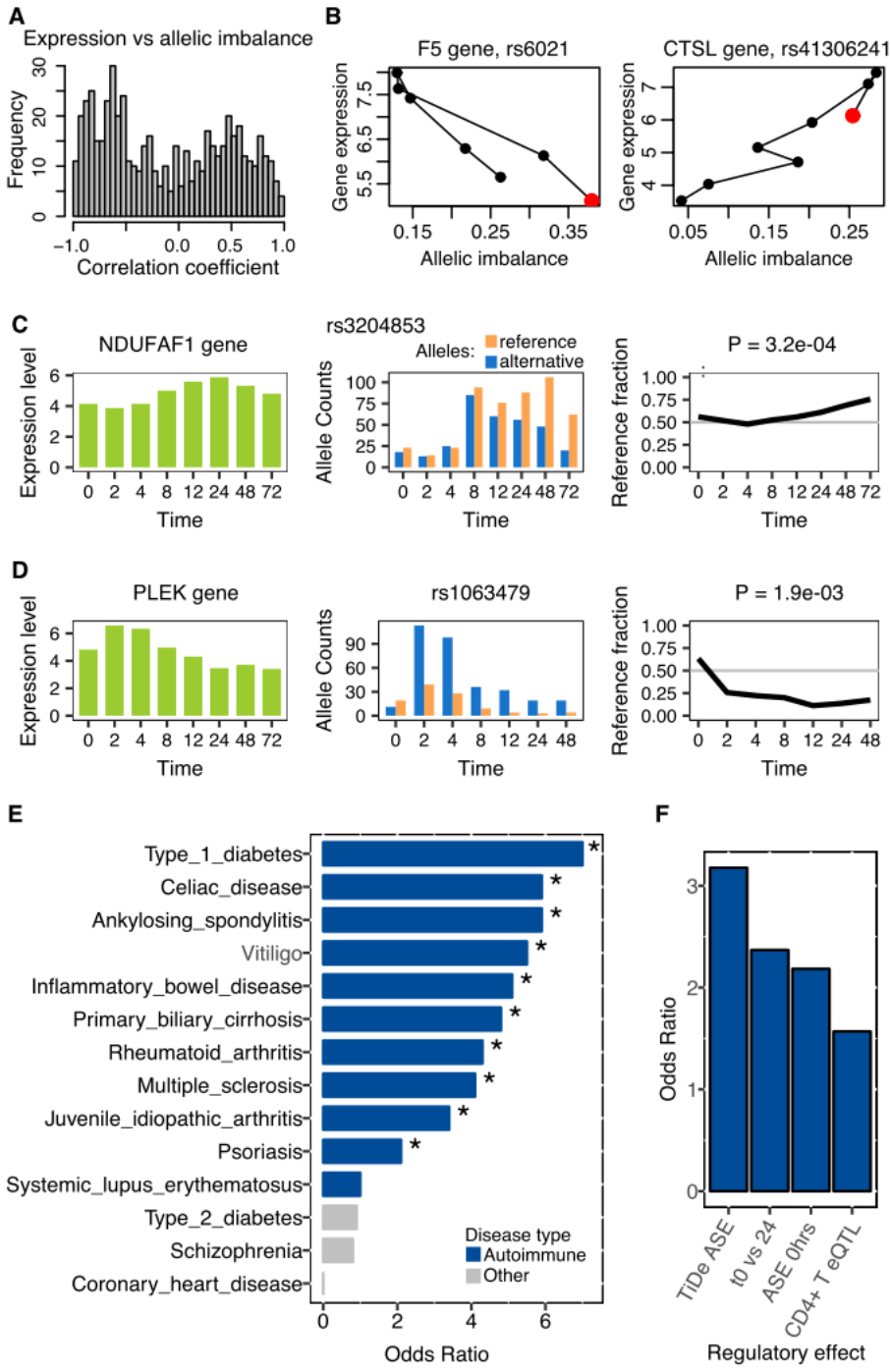


Fig. 4 Non-MHC dynamic ASE genes are enriched in autoimmune disease loci. (A) Spearman correlation coefficient between expression levels (log2(tpm+1)) and allelic imbalance (distance to 0.5 allelic fraction). (B) Gene expression levels by allelic imbalance trajectories for SNPs in two genes. Red dot indicates start of trajectory (0-hour time point). (C-D) Dynamic ASE examples for two autoimmune disease genes. (E) Enrichment of dynamic ASE genes in risk loci for autoimmune diseases (blue) and other control diseases where no enrichment is expected (gray). (F) Enrichment of genes identified by different genetic regulatory effect approaches in autoimmune diseases.

Discussion

Overall, these results can explain why investigators have found limited shared genetic effects between autoimmune GWAS and eQTL variants in three immune cell types in resting state (11), despite extensive evidence indicating autoimmune disease variants affect regulatory elements in immune cell types (2, 3, 23). Some studies have highlighted the importance of looking at stimulated state and particular sub-populations of cells in order to capture those "missing eQTLs" (12, 13, 24). Our study supports that view, and proposes to go even further and study multiple stimulation states in specific cell populations to understand the dynamics of *cis* regulatory variation that contributes to autoimmunity. Furthermore, our study shows that a cost-effective way to study genetic regulatory effects in multiple states is by leveraging the power of allele specific expression, as has been used by other groups (15).

Thanks to this high-resolution time course approach, we discovered new dramatic and condition-dependent *cis* regulatory variation for a major autoimmune disease gene: HLA-DQB1. This raises the question of whether, and to what extent, genetic regulatory variation of HLA genes could participate in disease susceptibility or disease penetrance (25). For most autoimmune diseases, the MHC region is the major contributing locus to disease risk. For example, in T1D the MHC region contributes ~30% of disease liability, compared to 9% by the rest of the susceptibility loci combined (4). We and others have shown that specific amino acid changes in HLA-DQB1 and HLA-DRB1 explain the majority, albeit not all, of the risk assigned to the MHC region in Type 1 Diabetes and Rheumatoid Arthritis (4, 26). However, the genetics in this region are complex, and haven't been fully elucidated. Here, the promoter regulatory SNP causal for up-regulation of HLA-DQB1 during late activation is associated with T1D risk (OR = 2, P=5e-139), even after controlling for the HLA-DQB1 major risk amino acid at position 57 (OR=1.7, P=2e-11), with the lowly expressed allele more common in cases than controls. Detailed analyses, on large sample sizes, will be needed to disentangle the regulatory effects from the strong

amino acid effects. While amino acid changes causing differential antigen display may be the primary autoimmune mechanism at the HLA locus, our data underscore the possibility that the quantity of antigen displayed may also play a role.

We recognize that the role of HLA class II genes is not well established in T cells yet, hence it is harder to speculate how variation in expression there could have an impact in disease or immune function. However, recent studies have uncovered rheumatoid arthritis associated effector memory and helper T cell subsets that express HLA class II (27, 28). Early studies on function of HLA class II in human T cells suggest they are capable of presenting antigen (29–33) and HLA class II could potentially play a role in anergy (17). While more studies are needed to define the function of HLA class II genes in human T cells, we believe variation in expression could lead to functional differences in CD4+ memory T cells among individuals.

The three dynamic HLA-DQB1 *cis* regulatory programs identified in our study in CD4+ memory T cells, together with evidence from other studies reporting independent eQTL variants for HLA genes (5, 9, 34–36) (Table S2), indicates there is substantial genetic variation affecting expression of HLA genes in cell type and condition dependent manner. Besides the Late Spike regulatory SNP, we also mapped an eQTL at 0 hours (fig S14), that is not correlated with the Late Spike SNP. Furthermore, even if we proved causality for the Late Spike regulatory SNP, we cannot discard the possibility of additional completely linked variants inside the gene that could also contribute to expression variation. Other studies with HLA-specialized pipelines have identified genetic regulatory variation in class II genes in LCLs (37, 38). And general eQTL studies have also reported independent regulatory variation for HLA-DQB1 in monocytes, macrophages and T cells (Table S2). This pervasive genetic regulatory variation is possibly mirroring the high level of polymorphisms at the protein coding level of HLA genes.

All in all, our study underscores the need to (1) perform HLA-specialized analyses to characterize the regulatory landscape of HLA genes and evaluate its impact in disease, and (2) study the high-resolution dynamics of genetic *cis* regulatory variation in other immune subsets and stimulation conditions for a better understanding of autoimmune disease mechanisms.

Methods

Healthy controls were recruited from the GAP registry and selected based on imputed HLA-DQB1 alleles. Peripheral blood was collected and PBMCs were extracted and frozen.

When all samples were ready to be assayed, memory CD4 T cells were isolated, rested overnight and stimulated the next day with anti-CD3/CD28 dynabeads. At each time point, cells were lysed and RNA was extracted. TruSeq RNA-seq libraries were prepared for 200 samples from 24 individuals, and sequenced at high depth. Targeted sequencing was performed for HLA class I and class II genes from each individual. ATAC-seq was performed in duplicate on CD4+ memory T cells stimulated for 72 hours for one individual. CRISPR/Cas9 experiments were performed on HH cutaneous T cell lines. Protein levels were measured with flow cytometry.

References

1. M. Gutierrez-Arcelus, S. S. Rich, S. Raychaudhuri, Autoimmune diseases connecting risk alleles with molecular traits of the immune system. Nat. Rev. Genet. 17, 160–174 (2016).
2. G. Trynka et al., Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. 45, 124–130 (2013).
3. K. K.-H. Farh et al., Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 518, 337–343 (2015).
4. X. Hu 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).
5. L. Chen et al., Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 167, 1398–1414.e24 (2016).
6. X. Hu et al., Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells. PLoS Genet. 10, e1004404 (2014).
7. T. Raj et al., Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 344, 519–523 (2014).
8. A. S. Dimas et al., Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 325, 1246–1250 (2009).
9. M. Gutierrez-Arcelus et al., Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife. 2, e00523 (2013).
10. C. J. Ye et al., Intersection of population variation and autoimmunity genetics in human T cell activation. Science. 345, 1254665 (2014).
11. S. Chun et al., Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat. Genet. 49, 600 (2017).
12. D. R. Simeonov et al., Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature. 549, 111–115 (2017).
13. H.-J. Westra et al., Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. Nat. Genet. 50, 1366–1374 (2018).
14. A. Buil et al., Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat. Genet. 47, 88–91 (2015).
15. N. Hauff, X. Zhou, X. Wen, R. Pique-Regi, F. Luca, High-throughput allele-specific expression across 250 environmental conditions. Genome (2016).
16. L. M. Sollid et al., Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J. Exp. Med. 169, 345–350 (1989).

17. T. M. Holling, E. Schooten, P. J. van Den Elsen, Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men. Hum. Immunol. 65, 282–290 (2004).
18. Roadmap Epigenomics Consortium et al., Integrative analysis of 111 reference human epigenomes. Nature. 518, 317–330 (2015).
19. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature. 489, 57–74 (2012).
20. D. Wong et al., Genomic mapping of the MHC transactivator CIITA using an integrated ChIP-seq and genetical genomics approach. Genome Biol. 15, 494 (2014).
21. M. M. Kaminski et al., Mitochondrial reactive oxygen species control T cell activation by regulating IL-2 and IL-4 expression: mechanism of ciprofloxacin-mediated immunosuppression. J. Immunol. 184, 4827–4841 (2010).
22. L. Lian et al., Loss of pleckstrin defines a novel pathway for PKC-mediated exocytosis. Blood. 113, 3577–3584 (2009).
23. H. K. Finucane et al., Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
24. M. R. Mumbach et al., Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat. Genet. 49, 1602–1612 (2017).
25. S. E. Castel et al., Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat. Genet. 50, 1327–1334 (2018).
26. S. Raychaudhuri et al., Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat. Genet. 44, 291–296 (2012).
27. D. A. Rao et al., Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. Nature. 542, 110–114 (2017).
28. C. Y. Fonseka et al., Mixed-effects association of single cells identifies an expanded effector CD4+ T cell subset in rheumatoid arthritis. Sci. Transl. Med. 10 (2018), doi:10.1126/scitranslmed.aaq0305.
29. M. Brandes, K. Willimann, B. Moser, Professional antigen-presentation function by human gammadelta T Cells. Science. 309, 264–268 (2005).
30. W. J. Pichler, T. Wyss-Coray, T cells as antigen-presenting cells. Immunol. Today. 15, 312–315 (1994).
31. A. Lanzavecchia, E. Roosnek, T. Gregory, P. Berman, S. Abrignani, T cells can present antigens such as HIV gp120 targeted to their own surface molecules. Nature. 334, 530–532 (1988).
32. C. R. Hewitt, M. Feldmann, Human T cell clones present antigen. J. Immunol. 143, 762–769 (1989).
33. J. M. LaSalle, K. Ota, D. A. Hafler, Presentation of autoantigen by human T cells. J. Immunol. 147, 774–780 (1991).
34. GTEx Consortium et al., Genetic effects on gene expression across human tissues. Nature. 550, 204–213 (2017).
35. Y. Nédélec et al., Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell. 167, 657–669.e21 (2016).
36. T. Lappalainen et al., Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 501, 506–511 (2013).
37. R. C. Pelikan et al., Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. Nat. Commun. 9, 2905 (2018).
38. D. Meyer, V. R. C Aguiar, B. D. Bitarello, D. Y. C Brandt, K. Nunes, A genomic perspective on HLA evolution. Immunogenetics (2017), doi:10.1007/s00251-017-1017-3.

Email regarding the acceptance of Chapter 2 in PNAS.

PNAS MS# 2018-12548RR Decision Notification

December 7, 2018 8:15 PM

From: journalstaff@pnascentral.org
To: arora@evolbio.mpg.de
Reply To: pnas@nas.edu

December 7, 2018

Title: "HIV Peptidome-Wide Association Study Reveals Patient-Specific Epitope Repertoires Associated with HIV Control"
Tracking #: 2018-12548RR
Authors: Arora et al.

Dear Dr. Arora,

We are pleased to inform you that the PNAS Editorial Board has given final approval of your article for publication. David Ho, the Editor who conducted the initial review of your manuscript [MS# 2018-12548RR], will also be informed of the decision.

Papers "in press" at PNAS are under embargo and not for public release before 3:00 PM Eastern Time, the Monday before publication. Authors may talk with the press about their work prior to the embargo but should coordinate this with the PNAS News Office or their institution's press office so that reporters are aware of PNAS policy and understand that papers are embargoed until the week of publication. If you plan to present your embargoed paper at a conference prior to publication, please contact the PNAS News Office immediately at 202-334-1310, or PNASnews@nas.edu.

Sincerely yours,
PNAS Editorial Office

Acknowledgements

It was in IMPRS week in June 2015, I met Tobias Lenz. During the selection process, I decided to give the preference to the supervisor over the projects. I was then provided with the opportunity by Tobias to work in his group. Now, when I look at my journey with him so far, I feel very happy about every aspect of it. He is an incredible supervisor and a great mentor. Although no words can entirely acknowledge my regards for him, I would like to thank him particularly for putting his trust in me.

I would also like to acknowledge the kind support from Derk and his awesome IT team as well as Britta, Kerstin, other members of the administration and Iben.

Special thanks to my group members (Ana, Federica, Marc, Malavi, Onur), colleagues and friends (Alice, Dominik, Ela, Ezgi, Loukas, Maryam-Moji, Maria, Roman, Sina, Zahra), my collaborators (Soumya, Maria, Yang in Boston and Shyam, Mohammad in Singapore) and my friends who are out-of-Germany (Maria Kondili, Kaushik, Vibhor, Brijveen mam, Pande, Lather, Khurana, Rishi, Anannya, Bhavuk, Naussad, Nirmala, Nooshin, Zahid). Also, special thanks to Aditi and Bilal, particularly for accompanying me until Bahnhof on the nights I worked until late.

And, I am grateful to all who ever believed in me…

Curriculum vitae

Personal Data
Name: Jatin Arora
Date and place of birth: 06.12.1988 in Ambala City, India
Nationality: Indian
Address of residence: Prinzen Str. 7, Plön, 24306 Germany

Education
Since 2016: Doctoral studies at Max Planck Institute for Evolutionary Biology, Plön
2018: Research trainee at Genome Institute of Singapore, Singapore
2017: Research trainee at Harvard Medical School, Boston, USA
2014-2015: Researcher at Center of Molecular Medicine, Vienna Austria
2012-2014: Master of Science, University Pierre and Marie Curie, Paris, France
2014: Master thesis, Okinawa Institute of Science and Technology, Japan
2006-2010: Bachelors of Technology, YMCA Institute of Engineering, Faridabad, India
2005-2006: High school, S.A. Jain Senior Model School, Ambala City, India

Professional experience
2010-2012: Software engineer for GE Healthcare at Birlasoft Ltd., India

Publications
J Arora, PJ McLaren, N Chaturvedi, M Carrington, J Fellay, TL Lenz (In press) HIV peptidome-wide association study reveals patient-specific epitope repertoires associated with HIV control. ***PNAS***

AS Mikheyev, MMY Tin, **J Arora**, TD Seeley (2015) Museum samples reveal rapid evolution by wild honeybees exposed to a novel parasite. ***Nature communications*** 6

PP Singh, **J Arora**, H Isambert (2015) Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. ***PLoS Computational Biology*** 11 (7), e100439

Awards
2018: EMBO Short Term Fellowship for research visit to Singapore
2017: IMPRS research grant for visit to Harvard Medical School, USA
2015: IMPRS fellowship for doctoral studies at Max Planck Institute in Plön, Germany
2015: CCHD research fellowship in Vienna, Austria
2014: OIST fellowship for master thesis in Okinawa, Japan
2012, 2013: Charpak scholarships by the Government of France to do masters in Paris

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 (Jatin Arora)