Bound to be an Epitope:
Determinants of the T-cell response to MHC-I presented peptides

Jorg Calis

Cover A metaphor for the immune system and the conduct of science. Inspired by artwork from Ursus Wehrli. Technical assistance by Bas Boukens is gratefully acknowledged.
Print Proefschriftmaken.nl || Uitgeverij BOXPress
ISBN 978-90-8891-460-7

Bound to be an Epitope:
Determinants of the T-cell response to MHC-I presented peptides

Epitopen aan het oppervlakte: Factoren die de T-cel reactie op MHC-I gepresenteerde peptiden bepalen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 3 oktober 2012 des middags te 2.30 uur

door

Jorg Justin Aimé Calis

geboren op 10 maart 1983 te Hilversum

Promotor: Prof. dr. R.J. De Boer
Co-promotor: Dr. C. Keşmir

Contents

Chapter 1

General Introduction and Overview

1.1 Preface

The immune system has to cope with a variety of threats such as viral or bacterial infections, and tumor formation. On the one hand, these threats can be located extracellular, such that the immune system can deal with it directly. For instance, bacteria that grow in a wound on the skin can be targeted by antibodies or the Complement system. On the other hand, a threat can be intracellular, for instance if a cell is infected with a virus. The immune system deals with such threats in an indirect manner, using the membrane glycoprotein MHC.
Peptides from degraded proteins can be presented at the cell surface on MHC molecules, thereby giving a representation of intracellular processes. A healthy cell would present peptides derived from human (self) proteins, an infected cell should also present nonself peptides derived from proteins that are part of the infection. Two important classes of MHC molecules can be distinguished: first, the MHC class I molecules that are expressed by all nucleated cells and that present peptides derived from endogenous proteins. Second, the MHC class II molecules, that present peptides from proteins that are degraded in endolysosomal compartments, MHC-II is only expressed by so-called professional antigen presenting cells that are part of the immune system. In this thesis we focus on the MHC-I presentation system.
What now happens if for instance an influenza virus infects an epithelial cell? Many (nonself) peptides derived from the viral proteins can be presented on MHC-I molecules of the epithelial cell, among the (self) peptides derived from

house-keeping or epithelial-specific proteins. T-cells from the immune system are selected to interact with specific peptide-MHC complexes (pMHCs), some of them have a specific interaction with an influenza-derived pMHC that enables the mounting of a specific immune response. The few pMHCs that are used by the specific T-cells as targets of the immune response are called epitopes. It is not known why certain pMHCs become epitopes and others are not, and this is the central question in this thesis: "Which pMHCs become epitopes?".
The central question can be broken up into three sub-questions that we want to explore in this introduction:

1. Which peptides are MHC-I presented?
2. Which pMHCs can be recognized by T-cells, i.e. are immunogenic?
3. Which immunogenic pMHCs are used in the immune response, i.e. are epitopes?

1.2 The MHC-I presentation pathway

The question which peptides are presented on MHC-I molecules has been addressed for many years using various techniques. As a result, many components of the cellular machinery that enable MHC-I presentation have been elucidated. Moreover, polymorphisms in the MHC-I presentation pathway have been studied intensively, some of which were shown to result in the presentation of different peptides by different cells, individuals and species. The whole set of peptides that is presented on the MHC-I molecules of a cell has been referred to as the MHC-I ligandome. When we ask which peptides are MHC-I presented we ultimately like to know the MHC-I ligandome dependent on the state of a certain cell or tissue. An important contribution to the study of MHC-I ligandomes has been made in the area of bioinformatics, where successful predictors of the MHC-I ligandome were developed.

1.2.1 Components of the MHC-I presentation pathway

Even though additional components might be found in the future, a major pathway enabling the presentation of peptides from intracellular proteins on MHC-I molecules has been elucidated (Figure 1.1). Three functions can be distinguished that are necessary for MHC-I presentation. First, the production of peptides by degradation of proteins and subsequent trimming of peptide fragments. Second, the transportation of peptides to the ER where peptides are loaded on MHC-I molecules. Third, the loading of peptides on MHC-I molecules and assessment of their suitability for MHC-I presentation. Let us go through these functions one-by-one.

Most peptides are formed by degradation of cellular proteins in a protein complex called the proteasome. The concentration of proteasomes is highest in the nucleus, but many are also found in the cytosol, in total $\sim 10^6$ proteasomes are present in a single cell [1]. The high abundance of proteasomes reflects the high turnover of proteins in a cell, about 3% of all proteins are degraded each hour [1]. Peptides can be seen as a side-product of this dynamical process, or as an intermediate to full degradation. Further degradation of peptides is performed by aminopeptidases such as ERAP and TOP [2], that can cleave of a few amino acids from the N-terminus of a peptide. The peptidase activity can trim peptides to a length that might be suitable for MHC-I binding, or destroy potential MHC-I ligands [2]. Carboxypeptidase activity from TPPII, Nardylisin and ACE has been shown to affect the generation of MHC-I ligands as well [3–6]. The peptidase activity is located in the cytosol and the Endoplasmatic Reticulum (ER) [7].
The Transporter Associated with Antigen Presentation (TAP) enables the transport of peptides from the cytosol to the ER, where peptides can be loaded on MHC-I molecules. TAP preferably transports peptides of 7-16 amino acids (7-16mers) [8, 9], in concordance with the length preference of MHC-I molecules (8-11 amino acids) [10]. TAP is one of many ABC-transporters that have very diverse functions in cross-membrane transport, but its preference to transport 7-16mer peptides as well as its location in the MHC-region on the genome seem to imply its important role in MHC-I presentation.
Besides its role in peptide transport, TAP is also involved in peptide loading on MHC-I by forming a complex in the ER that is called the Peptide Loading Complex (PLC). Empty MHC-I molecules are in the PLC. The MHC-I molecule is composed of two subunits that are referred to as the heavy and light chain, the heavy chain is variable for different MHC-I molecules and encoded by a polymorphic and polygenic gene family, the light chain is constant and is called β2-microglobulin (β2m). Other molecules in the PLC are tapasin, calreticulin and ERp57, these chaperones stabilize MHC-I when it is not bound to a peptide [11]. The PLC is thought to ensure that peptide-MHC complexes (pMHCs) with an unstable or low-affinity binding dissociate, such that these pMHCs are not transported to the cell surface. Once a peptide is loaded onto an MHC-I molecule, the pMHC can be exported to the cell surface, where it is presented to the immune system.

1.2.2 Efficiencies in the MHC-I presentation pathway

The chance that a peptide from a certain protein will be presented on an MHC-I molecule at the cell surface is small. First, the peptide must be produced by a proteasomal cleavage at the C-terminus, while not being destroyed by internal cleavages. The chance that the C-terminus of a certain peptide is produced has been estimated to be 1 in $\sim$2-4 based on in vitro digestion studies [12–14] and theoretical work [15, 16]. Before a peptide can bind to MHC-I or is transported to the ER, it should not be destroyed by aminopeptidases. Peptides are degraded

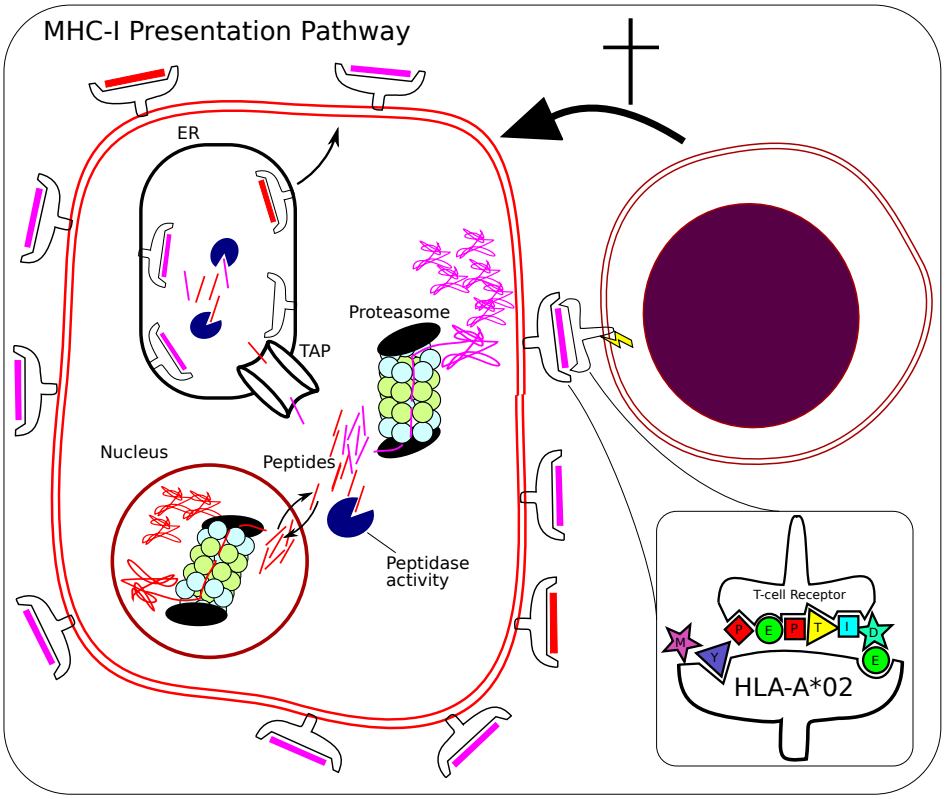


Figure 1.1. The MHC-I presentation pathway. An overview of the MHC-I presentation pathway in which proteins are degraded by proteasomes into peptides, and where peptides are transported by TAP to the ER, where they can bind to MHC-I molecules that go to the cell surface. An MHC-I presented peptide (pMHC) interacts with the T-cell receptor of a T-cell, this interaction triggers the effector T-cell to kill the MHC-I presenting cell.

rapidly, Reits et al. [7] showed that the half-life of a peptide in the cytosol is in the order of seconds, as peptides in the cytosol are unbound and move through the cell by diffusion, unprotected from peptidases [7]. Third, a peptide must find TAP to be transported to the ER, but only 1 in $\sim$3 peptides has a motif that is required by TAP for translocation to the ER [17, 18]. Finally, a peptide in the ER must find an MHC-I molecule that is capable of loading the peptide. MHC-I molecules have a very specific binding motif, only 1 in $\sim$40 peptides will be able to interact with large enough affinity [18]. When we consider these dynamics (summarized in Table 1.1) it is evident that not every degraded protein can be represented with at least one peptide on an MHC-I molecule. This becomes even more evident when we realize that a cell degrades $\sim 2 \times 10^6$ proteins per minute [1], but has

only ∼100,000 MHC-I molecules [19]. If every protein would be represented by a single pMHC, it could be on the cell surface for maximally ∼3 seconds. This would contradict measurements from Parker et al. that showed that the half-life of many pMHCs lies in a range of 5-500 minutes [20, 21].

Effect	Correct peptides	Reference
Proteasomal cleavage	1 in ∼2-4	[12–16]
TAP transport	1 in ∼3	[17, 18]
Peptidase destruction	<1%	[7]
MHC-I binding	1 in ∼40	[18]

Table 1.1. Efficiencies in the MHC-I presentation pathway. Here we summarize different processes that underlie the MHC-I presentation of a peptide (column 1), and the chance that a peptide will be able to pass this process (column 2). These effects are further discussed in Section 1.2.2.

1.2.3 Diversity in the MHC-I presentation pathway

Diversity is a hall-mark of the immune system and the threats it has to cope with [22]. Which peptides are presented on the MHC-I molecules differs per cell, tissue, individual and species. In a cell, different states will lead to the expression and degradation of different proteins that serve as substrates for MHC-I ligands. In tissues, besides the altered expression of proteins, diversity within the MHC-I presentation pathway plays an important role. For instance, other proteasome-types or aminopeptidases can be expressed in different cell-types, with a marked influence on the MHC-I presented peptides [11, 23]. When individuals and species are compared, other variants of MHC-I molecules, TAP molecules and proteasomes are found with different peptide interaction motifs that can have a dramatic effect on the MHC-I ligandome [24–26].
The MHC-I molecule is the most diverse component of the MHC-I presentation pathway, it is encoded by a gene-system that is polygenic (i.e. duplicates of the MHC-I gene) and polymorphic. In the human genome, three genes encode the most important classical MHC-I molecules, they are called HLA-A, HLA-B and HLA-C. All three are polymorphic, 1884, 2490 and 1384 alleles of HLA-A, -B and -C have been described so far (May 2012), respectively [27, 28] (Table 1.2). The diversity among the encoded MHC-I molecules is greatest in the binding groove that determines the peptide binding motif [29]. As a result of this diversity, many MHC-I molecules exist with different peptide-binding specificities [30]. For instance, HLA-A*0201 prefers to bind 9mer peptides with a Leucine at position 2 and a Valine at position 9, whereas HLA-A*2402 binds 9mer peptides with a Tyrosine at position 2 and a Phenylalanine at position 9 [30] (Figure 1.2).
The extreme diversity of MHC-I molecules can best be explained by an antagonistic selection model based on the co-evolution with pathogens [31]. In this

model, pathogens adapt to common MHC-I molecules in a population and are more pathogenic for hosts that express these MHC-I molecules. Vice versa, hosts with rare MHC-I molecules have an advantage, as the pathogen is not likely to be adapted to the rare MHC-I molecule and therefore less pathogenic in these hosts. Recently, Kubinak et al. [32] provided experimental proof for the key-assumption in this model: they studied the pathogenicity of Friend virus complex after it passages (by infection) through mice with different MHC-I backgrounds. They showed that the virus adapts to the MHC-I background of the host through which it passaged, and that it was much more pathogenic in a next host with the same MHC-I background than in a non-MHC-I-matched host [32]. As a result of the evolutionary process in which rare MHC-I molecules confer advantage, the HLA system became so diverse that the chance that two persons have exactly the same HLA-A and HLA-B background is only 0.001% (calculated based on European HLA-A/B haplotype frequencies [33]).
Schmid et al. [34] investigated how different parts of the MHC-I presentation pathway (i.e. proteasomal degradation, TAP transport and MHC-I binding) become polymorphic by modeling the co-evolution of this pathway with pathogens. They showed that polymorphism does not evolve at every step of the MHC-I presentation pathway, as the different steps in many hosts would than be incompatible, and not enough peptides would be presentable [34]. In agreement with the predictions from this modeling work, human TAP is functionally invariant and not more than a handful of proteasome-variants are known. In addition, the proteasome does not vary in the population, instead variants are tissue-specific. Best studied is the so-called immunoproteasome variant that is expressed in immune cells, and in cells that are exposed to IFN-γ [35]. Recently, the thymoproteasome was discovered, that is only expressed in cortical thymic epithelial cells. Even though TAP and proteasome motifs are very similar within species, there can be differences among species, this has been clearly shown for TAP that was suggested to co-evolve with the different MHC-I preferences [36, 37].

Gene	Alleles (n)	Proteins (n)
HLA-A	1884	1365
HLA-B	2490	1898
HLA-C	1384	1006

Table 1.2. HLA class I diversity. For different HLA class I loci (column 1), the number of different nucleotide (column 2) and protein (column 3) sequences is shown.

1.2.4 Motifs in the MHC-I presentation pathway

Even though the peptide-binding preferences of different human MHC-I (HLA-I) molecules are very different, there are also similarities. First, most HLA-I molecules prefer to present peptides with a similar length, often 9mers. Second, in

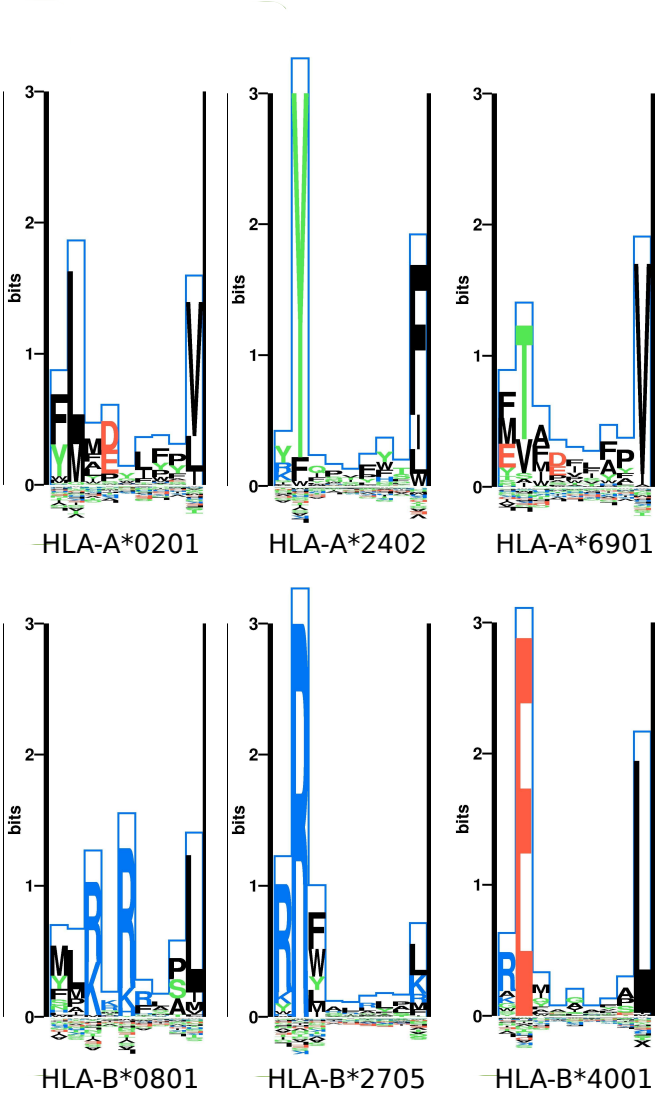


Figure 1.2. MHC-I peptide binding motifs can be very different. Here we show the binding motif of six different MHC-I molecules, derived from the motifviewer website [30]. The height of the blue box that surrounds the amino acids letters at every position displays the information content that predicted MHC-I ligands have at that position, specific amino acids are shown in this box relative to their frequency in the predicted set of MHC-I ligands [38].

most HLA-I molecules two positions are important for peptide binding, these so-called anchor-positions are predominantly positions P2 and P9 of the 9mer peptide. However, exceptions are possible, HLA-B*4402 has been shown to present many 10mers [24], and HLA-B*0801 uses position P5 as an anchor position [30]. At the anchor positions, a wide range of preferences for different amino acid residues can be observed. For instance, the acidic Glutamic acid residue is preferred by HLA-B*4001 at position P2, whereas HLA-B*2705 prefers the basic Arginine at this position [30] (Figure 1.2). Despite the diverse specificities, MHC-I molecules with relatively similar peptide-binding motifs have been grouped in so-called supertypes [39]. Sette et al. suggested that whereas MHC-I frequencies in a population are variable, the frequencies of supertypes in different populations are relatively stable [39], which suggests that certain binding motifs are kept during evolution. Other groups defined supertypes based on a clustering of binding motifs [40]. The peptide binding positions of MHC-I molecules can evolve independently, this poses a difficulty when one tries to classify MHC-I molecules in supertypes. For instance, one can group all MHC-I molecules together that bind peptides with a basic residue at position P9, but these can have diverse binding motifs at position P2, and some of them might be very similar to other MHC-I molecules that bind non-basic residues at P9.

Just as MHC-I molecules have a certain binding motif, proteasomes and TAP also require their substrates to fit to a certain motif. The proteasomal cleavage motif is most specific at the amino-terminal residue of the cleavage site, i.e. position P1' [41]. If an hydrophobic, basic or acidic residue is present at position P1', the site is preferably cleaved by the constitutive proteasomes, the immunoproteasomes preferably cleaves at sites where the P1' position is an hydrophobic or basic, but not an acidic amino acid [12, 42, 43]. In addition to the preferences at position P1', other positions determine if it is a proteasomal cleavage site [41]. TAP requires substrates to have a length of 7 to 16 amino acids [8, 9], and transports peptides with certain residues at the three N-terminal residues and at the C-terminus. A negative effect on the transportation by TAP has been shown for Proline residues at the N-terminal positions, a positive effect of aromatic, hydrophobic or basic charged residues has been shown at the C-terminus [36]. The specificities of proteasomes and TAP are much weaker than that of MHC-I molecules: while MHC-I molecules are expected to bind only 1 in $\sim$40 9mer peptides [18], 1 in 2-4 sites can be cleaved by the proteasome [12–16], and 1 in 3 peptides can be transported by TAP [17, 18]. Moreover, proteasome cleavage is a stochastic event, identical proteins can be degraded into different fragments [12–14].

The preferences of non-polymorphic components in the MHC-I pathway are expected to become similar in order to increase the number of MHC-I presentable peptides. This is observed for TAP and the immunoproteasome, that respectively transport and generate peptides with hydrophobic residues at the C-terminus [36, 42]. In addition, immunoproteasomes have been shown to generate more MHC-I ligands than constitutive proteasomes, and to evolve faster to keep up with the fast evolution of MHC-I binding preferences [43]. As the presentation of pathogen-derived peptides on MHC-I molecules is crucial for an effective im-

mune response, a preference for pathogen-derived peptides might be expected. For HLA-A molecules as well as the immunoproteasome, it has been shown that they have a larger predicted affinity for pathogen-derived peptides [15, 44].

1.2.5 Predicting what is MHC-I presented

The elucidation of MHC binding motifs sparked the idea that peptide-MHC binding might be predictable. At first, simple predictions were made based on simple rules, for instance, a potential binder of HLA-A*0201 should have a Leucine at position P2 [45]. Later, predictors were developed in an iterative cycle of measuring affinities of different peptides, training a predictor on the outcome, predicting the affinity of peptides and test those for which the prediction was inconclusive [46–48]. These predictors were based on a quantitative prediction of the binding affinity and relied on neural networks [46]. In addition, they were MHC-I molecule specific as they were trained on affinity measurements of peptides on a single MHC-I molecule, thus a new predictor had to be developed for every MHC-I molecule. Recently, a next generation predictor has been developed that combines the measurements of peptides on multiple MHC-I molecules. This predictor has been trained to predict the peptide binding affinity of an MHC-I molecule given the sequence of the MHC-I molecule, therefore it can predict the affinity for MHC-I molecules for which no or scarce peptide-binding measurements are available [49]. Current MHC-I binding predictions are so precise that the difference with actual measurements is as large as the difference between measurements from different labs [50].
Even though proteasome cleavage and TAP transport have been studied to a lesser extend than MHC-I binding, predictors have been developed based on the studies of these steps in the MHC-I presentation pathway. Proteasome cleavage sites have been determined by studying in vitro proteasome digestions [12–14] and MHC-I ligands that are the result of in vivo proteolysis. Based on these measurements, different proteolysis predictors have been developed [16, 41, 51]. To predict TAP transportability, a TAP transport predictor has been developed based on TAP-peptide binding affinity measurements [52], these affinities were shown to correlate with the chance that a certain peptide is transported by TAP [19].
As a result of the efforts described above, we can now accurately predict the binding affinity of peptides for different MHC-I molecules, but this is not the same as predicting which peptides are MHC-I presented. Peptides need to be made by the proteasome and aminopeptidases, and need to be transported to the ER by TAP. Even though there are predictors for these processes, how to combine them into a model to predict MHC-I ligands is unknown. Moreover, pMHCs need to be stable as they are transported to the cell surface and when they are presented at the cell surface. Finally, the abundance of MHC-I ligand precursors can affect which peptides are ultimately presented [53]. It is unknown how to include these issues in ligandome predictions.

1.3 T-cells - which pMHCs are immunogenic?

The peptides presented on MHC-I molecules enable T-cells of the immune system to screen intracellular processes. T-cells recognize pMHCs with their T-cell receptor (TCR) that is specific for a small fraction of the pMHCs, and a co-receptor that recognizes an invariable part of the MHC molecule. The co-receptor for MHC-I molecules is CD8, thus $CD8^+$ T-cells can interact with MHC-I presented peptides, CD4 acts as a co-receptor for MHC-II molecules. Only 1 in $\sim$100,000 T-cells have a T-cell receptor that can interact with a specific pMHC [54–57], but many pMHCs can be recognized by a host because there is a huge repertoire of TCRs [58]. The extend to which a certain pMHC is recognized by T-cells is referred to (in this Introduction) as its immunogenicity. Immunogenic pMHCs can stimulate T-cells to proliferate and become effector cells. These effector cells can kill other cells that express the pMHC. Therefore, if we want to identify which pMHCs are used as epitopes it is essential to be able to classify which pMHCs are immunogenic.

1.3.1 The T-cell receptor

The T-cell receptor (TCR) enables T-cells to specifically recognize pMHCs. The T-cells with a TCR that is composed of an α- and a β-chain play a dominant role in the adaptive immune response [59]. Other T-cells have a TCR with γ/δ-chains, we will not discuss these T-cells in this thesis. The TCR-chains form loops that interact with both the MHC-I molecule and the presented peptide [60]; the amino acid sequences in these loops determine the specificity of the TCR [61]. The genetic loci that encode the TCR are recombined during T-cell development, in a recombination process that involves three gene segments for the β-chain (V_β, D_β and J_β) and two gene segments for the α-chain (V_α and J_α) [62]. For the α-chain $\sim$70 V_α and 61 J_α segments are available, for the β-chain 52 V_β, 2 D_β and 13 J_β segments are available, these can form already 5.77 million combinations [62]. Moreover, where segments are joined random nucleotides can be inserted, deleted or substituted, as a result of which the estimated number of possible TCR sequences increases to $>10^{15}$ [58].
Both the α- and the β-chain make three loops that contact the pMHC, these are also referred to as the Complementary Determining Regions (CDRs). The three CDRs differ in their variability, and how well they interact with the peptide or MHC part of the pMHC. CDR1 and CDR2 from both the α and β-chain are germline encoded by the V-segment, and mostly in contact with the MHC [60]. The α-chain CDR1/2 contact the area that is at the N-terminal side of the presented peptide, the β-chain CDR1/2 contact the C-terminal side [60]. CDR3 from both the α- and β-chain are encoded by the sequence that is formed when joining V and J (via D in case of the β-segment), thus the CDR3 sequence is not germline

encoded and is extremely variable [60]. The CDR3s are in close contact with the presented peptide, but interactions between CDR3 and MHC also take place [60].

1.3.2 The T-cell receptor repertoire

Even though $>10^{15}$ TCRs can be made in theory [58], in practice some TCR sequences are more likely to form or be functional than others. First, the recombined α- and β-chains should be able to form a functional TCR, T-cells die if their TCR does not pass quality-control mechanisms that ensure a stable TCR [63]. Second, T-cells undergo two selection processes in the thymus, called positive and negative selection (Figure 1.3). During positive selection, T-cells are selected that have a strong enough interaction with any of the hosts MHC-I molecules [63]. During negative selection, T-cells are deleted that have a too strong interaction with any of the hosts pMHCs [63]. Finally, some T-cells might be deleted in the periphery if they either recognize a peripheral pMHC [64, 65], of if they are unable to interact with any of the peripheral pMHCs [66, 67].
Positive selection takes place in the cortex of the thymus with pMHCs expressed on cortical thymic epithelial cells (cTECs) [68–70]. In this phase, the T-cells change from $CD4^+CD8^+$ (double-positive), to $CD4^-CD8^+$ or $CD4^+CD8^-$ (single-positive) dependent on their recognition of MHC-I or MHC-II molecules, respectively [71]. Interestingly, cTECs express a special proteasome-type called the thymoproteasome and a special peptidase called Cathepsin L [72, 73]. As a result of these specific proteases, cTECs are expected to present a different peptide-repertoire on their MHC-I and MHC-II molecules than other cells [74]. The importance of cTEC-specific pMHCs is observed in mice that do not express the thymoproteasome, the number of $CD8^+$ T-cells in these mice is $\sim$80% lower, despite normal MHC-I expression in the cTECs [74]. Similarly, in Cathepsin L knockout-mice $CD4^+$ T-cell numbers are affected [73].
Medullary thymic epithelial cells (mTECs) are involved in negative selection, a T-cell that interacts too strong with any of the pMHCs on an mTEC will be deleted. This process is thought to ensure that autoreactive T-cells are deleted that might otherwise cause auto-immunity after thymic education. To present pMHCs that can be encountered outside the thymus, mTECs express peripheral proteins using a transcription factor called AIRE [75]. On the one hand, the importance of AIRE in negative selection has been shown using AIRE knock-out mice; these mice develop auto-immune diseases in tissues such as the retina and the stomach [76]. On the other hand, not every peripheral protein is expressed by mTECs, and AIRE knock-out mice do not develop auto-immune diseases in all peripheral tissues [76]. The existence of T-cell mediated auto-immune diseases [77] such as diabetes type 1 [78–80] confirms that not all autoreactive T-cells are deleted from the repertoire. Possibly, other peripheral regulatory processes prevent auto-immune responses in some tissues.
An estimated 95% of T-cells do not survive thymic selection [81, 82]. These

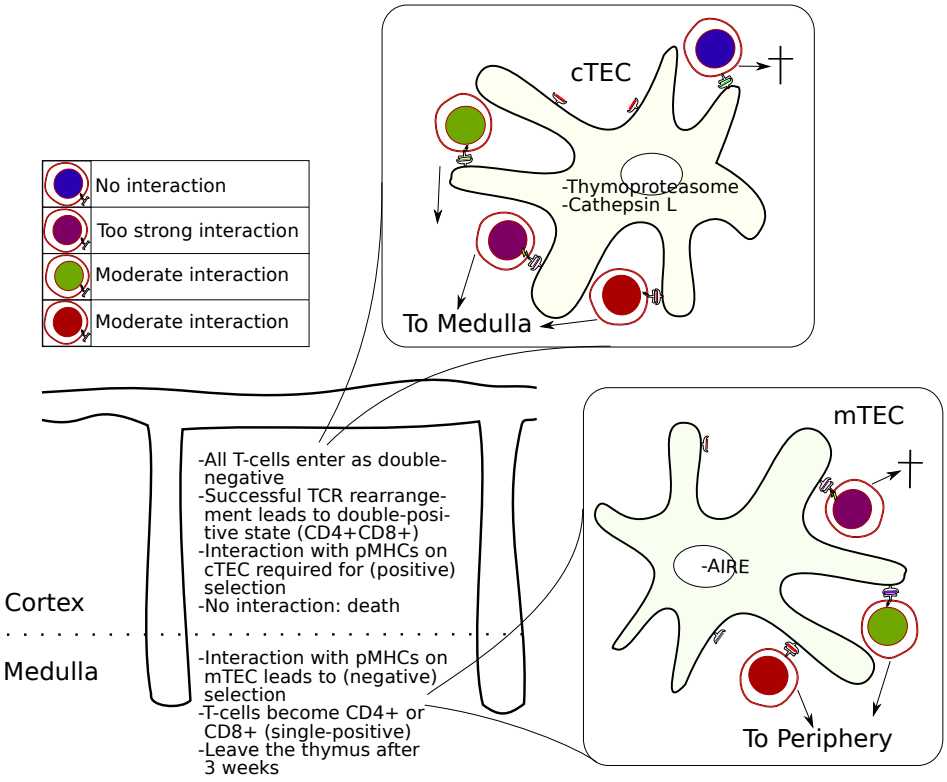


Figure 1.3. Thymic selection of T-cells. T-cells enter the thymus via the cortex, where they are selected on their ability to recognize MHC complexes, this process is called positive selection. Selected T-cells go to the medulla, where they are deleted if they strongly interact with a self pMHC, this process is called negative selection. The faith of four T-cells with different pMHC interaction strengths in the cortex and the medulla upon interaction with thymic epithelial cells is illustrated.

deaths might reflect a high chance of not forming a functional TCR, as well as a consequence of positive and negative selection, ∼70-80% [83] and ∼50% [84] of the T-cells have been estimated not to survive positive and negative selection, respectively. Thus, based on these estimates, both positive and negative selection are expected to have a large influence on the T-cell receptor repertoire, and therefore the repertoire should vary between hosts with different MHC backgrounds. However, ∼10-20% of the TCRs are shared in the naive T-cell pool of different hosts [85, 86], or in the T-cell response to specific pMHCs [87], these TCRs are so-called public TCRs. Moreover, the frequencies of V_β-segments or $V_\beta J_\beta$ and $V_\alpha J_\alpha$-segment combinations are hardly affected by MHC-I background differences [88–90]. Even more striking, most V_β-frequencies are not very different when $CD8^+$ T-cells are compared with $CD4^+$ T-cells [89, 91]. This is striking,

as V-segments form a large part of the TCR that interacts with MHC [60], and segment-usage has been shown to affect pMHC specificity [92], thus it seems difficult to reconcile the observations that segment-usage is largely invariable with a strong influence of positive and negative selection on the T-cell repertoire. A possible explanation could be found in the work of Gebe et al., they showed that some autoreactive T-cells are not lost from the repertoire because they adopt a regulatory function or down-regulate TCR-expression [93]. In line with this hypothesis, regulatory Foxp3$^+$ T-cells and conventional Foxp3$^-$ T-cells have distinct TCR repertoires, even though the overlap between these TCR repertoires is larger when most frequent TCRs are analyzed [94, 95].
Recent advances in high-throughput sequencing made it possible to directly sequence parts of the TCR β-chain repertoire of a host. In such experiments, T-cells are isolated and the variable part of the β-chain is amplified and sequenced. Robins et al. showed that a healthy human has at least 1×10^6 distinct β-chains in both the naive CD8$^+$ and the naive CD4$^+$ T-cell receptor repertoires [96]. However, α-chains have not been sequenced in a high-throughput fashion, as they are too long and complex, therefore α-chain diversity has been measured indirectly by extrapolation on the diversity measured for a few $V_\alpha J_\alpha$ combinations, resulting in an estimated α-chain diversity of 0.5×10^6 sequences [97]. Arstila et al. showed that 2.5% of the α-chain repertoire that was $V_\alpha 12^+$, still contained two-thirds of all β-chains, therefore is was extrapolated that the actual TCR repertoire (α- and β-chain combined) is at least ($\frac{1}{0.025} \times \frac{2}{3} \approx$) 25-fold larger than the β-chain repertoire [97]. Thus, current estimates are that the human T-cell receptor repertoire is composed of at least 2.5×10^7 TCRs [96, 97]. This estimate should be considered a lower-bound estimate, as different β-chains can still bind to different α-chain variants that are $V_\alpha 12^+$ [98]. Only direct TCR sequencing on single T-cells of both the α- and β–chain can eventually tell how large the TCR repertoire in an individual is. Excitingly, such techniques are now under development [99].

1.3.3 The functional T-cell repertoire

The level of recognition of different pMHCs by the T-cell repertoire is referred to as the functional T-cell repertoire [100]. The functional repertoire can be very different from the T-cell receptor repertoire as a single TCR can recognize multiple pMHCs and multiple TCRs can recognize the same pMHC. The number of T-cells that recognize a pMHC has been referred to as the precursor frequency, as these cells are the potential precursors of a T-cell clone that could form an immune response to the pMHC. Especially interesting are the precursor frequencies of naive T-cells, as they reflect the immunogenicity of pMHCs in a primary immune response, i.e. when the immune system is confronted with a pathogen for the first time.
As most naive precursor frequencies are very low, they are hard to determine. Some initial estimates were based on a titration with known amounts of labeled

pMHC-specific T-cells and measurements of the fraction of labeled cells upon antigen stimulation [54]. As this method is laborious and not sensitive enough to detect very low precursor frequencies, a new method was developed by Moon et al. that relies on the enrichment of specific T-cells using soluble tetramers and magnetic beads [101]. In addition, assays to measure multiple T-cell frequencies at once using tetramers and multi-color coding have been developed [102, 103]. The measurements with new techniques were in line with previous results, and showed that the naive precursor frequency of most pMHCs is about 1 in ∼300,000 [101]. However, the fraction of naive T-cells that recognize a specific pMHC was shown to vary substantially for different pMHCs, from 1 in $\sim 10^5$ to 1 in $\sim 10^6$ for naive CD4$^+$ T-cells, from 1 in $\sim 10^4$ to 1 in $\sim 10^5$ for naive CD8$^+$ T-cells [101, 104]. Despite the large variation among pMHCs, the variation among hosts with similar MHC background was unexpectedly small [104]. Legoux et al. took this a step further and showed that precursor frequencies for certain pMHCs are ∼10-fold higher, only when hosts that express the presenting MHC molecule are compared with hosts that do not have this MHC molecule [105]. Even when CD4$^+$ and CD8$^+$ T-cells were compared from hosts that were HLA-A2$^-$, the fraction that recognized HLA-A2 presented peptides was only ∼2-fold higher in the CD8$^+$ T-cell repertoire, when the effect of CD8-MHC-I binding was cancelled [105]. Thus, on the one hand the stable precursor frequencies in different hosts suggest that the functional T-cell repertoire is robust for the randomicity of T-cell receptor generation. On the other hand, there seems to be an influence of positive selection as precursor frequencies for peptides on MHC-I molecules of the host are ∼10-fold higher, but this effect seems to be very MHC-I specific.
Not enough measurements of precursor frequencies for foreign and self pMHCs in hosts with different MHC-I backgrounds are available to describe the effect of negative selection on the functional repertoire. However, an indirect effect was described by Huseby et al. [106] who measured the cross-reactivity of T-cells in hosts that could use only a single pMHC for negative selection. The cross-reactivity was determined by testing for pMHC-specific T-cell clones how many positions of the presented peptide were important for T-cell recognition [106]. They showed that about half of the T-cells surviving negative selection on a single pMHC, were more cross-reactive than the T-cells that underwent normal negative selection [106], in agreement with previous estimates that about half of the T-cells are deleted during negative selection [84]. If more pMHCs are used during negative selection, cross-reactive T-cells have a larger chance to recognize any of the self-pMHCs to be deleted from the repertoire, later modeling work on the functional T-cell repertoire confirmed this effect [106, 107]. Taken together, both positive and negative selection have an effect on the functional T-cell repertoire.
Peptide-immunization studies are used to directly test whether the T-cell repertoire can make an immune response to a certain pMHC. In such a study, peptides with a large enough affinity for the hosts MHC-I molecule are used to try to elicit an immune response. After the immunization, the immune response can be measured as the number of T-cells that give an effector response when confronted with the pMHC. Multiple peptide-immunization studies have shown that ∼50% of the

pMHCs are non-immunogenic [18, 108]. Possibly, precursor frequencies for these pMHC are too low to elicit a functional T-cell response.

1.3.4 Predicting immunogenicity

As the T-cell repertoire has an important influence on the development and outcome of immune responses, it would be of interest to model and predict how the T-cell repertoire develops. An important model of the generation of a T-cell receptor repertoire was made by Venturi et al. [87]. In their model, the chances of different V(D)J-recombinations and mutations at segment joining sites are taken into account to enable the prediction of an artificial T-cell receptor repertoire [87]. The model correctly predicted that TCRs that are shared by multiple hosts were more likely to be formed than other TCRs that were not shared [87, 109]. The limited number of N-additions at segment recombination sites is an important feature in this model, that not many N-additions take place was derived from an analysis of TCR sequences, and this was later also shown in other studies [85, 110]. Thus, also at the site where TCR-segments join, that is the most variable part of the TCR-gene, there is a tendency to encode most of the TCR-sequence with germline sequences.
To model the functional T-cell repertoire is more difficult given the limited number of precursor frequency measurements. However, a theoretical estimate on the possible range of precursor frequencies has been made by Borghans et al. [111], who showed that a T-cell must be specific enough to prevent large self/nonself overlaps, but not too specific to be able to recognize most foreign peptides. Given a thymic expression of 80% of the peripheral proteins, Borghans et al. showed that immune responses would only be possible for precursor frequencies that lie in a range from 1 in 10^6 to 10^{10} [111].
Borghans et al. did not take the heterogeneity of precursor frequencies for different pMHCs into account, for which at that time no estimates were available. However, recent advances in the measurement of precursor frequencies made it possible to show that such heterogeneity exists [101, 104]. The immunogenicity of specific pMHCs differs dramatically, and this has an effect on the chance that a pMHC is used as an epitope [56, 104]. At the moment, not enough precursor frequency measurements are available to associate certain features of a pMHC with its immunogenicity. However, as the immunogenicity has an effect on the mounting of an immune response, the presence of an immune response or the absence thereof can be used to try to predict immunogenicity. Tung et al. were the first to use this idea to make a predictor of immunogenicity for specific peptides on HLA-A*0201 called POPISK [112]. The difficulty when studying immunogenicity based on immune responses is in assembling a set of peptides that are non-epitopes because they are non-immunogenic, and not because they are not MHC-I presentable, processed or expressed.

1.4 Immune responses - which pMHCs become epitopes?

The outcome of an infection varies between hosts. For instance, some HIV-1 infected patients are symptom-free for more than a decade, whereas others get AIDS within two years. For many diseases, the disease outcome was shown to depend on variations in the MHC-I genes [77, 113–115], implying a different effectiveness of immune responses dependent on which pMHCs are presented to and used by the immune system. To understand these variations in effectiveness, we first have to know which pMHCs are used in an immune response, i.e. are epitopes. Only then, the quantity and quality of the responses can be investigated to see which responses confer protection.

1.4.1 Determinants of a T-cell response

For a pMHC to become an epitope, it needs to be presented (Section 1.2) and be immunogenic (Section 1.3). In addition, there are other determinants: First, responses to self pMHCs are avoided and inhibited by negative selection [75] and regulatory mechanisms [116]. Avoiding auto-immune responses can affect the response to foreign pMHCs as well, especially if the foreign pMHC is identical or similar to self [117, 118]. Negative selection lowers the number of possible T-cells that can react to a pMHC, i.e. the immunogenicity of certain pMHCs. In addition, specific immune responses can be prevented by regulatory T-cells (Treg), that have been suggested to prevent the stimulation of naive T-cells by indirectly influencing the dendritic cells (DCs) that are involved in this stimulation [116]. Other regulatory mechanisms play a role, for instance, activated T-cells are not allowed to exert their cytotoxic functions in every tissue, and their entry in peripheral tissues can be regulated by $CD4^+$ T-cells and endothelial cells [65, 119]. Second, multiple immune responses can outcompete each other due to limited survival factors such as IL-2 and IL-12 [120, 121]. This competition takes place between T-cells that recognize different peptides on the same HLA molecule [18], but also between T-cells that recognize the peptides presented on different HLA molecules, i.e. HLA-A, HLA-B or HLA-C [122, 123]. It has been observed that the responses to HLA-B presented peptides often dominate the response to HLA-A presented peptides [124–126]. A good explanation for this phenomenon is lacking, but might be found in an increased immunogenicity of peptides presented on HLA-B. Third, the timing of peptide presentation seems to be important; for multiple pathogens the peptides of early expressed proteins were shown to be prominent immune targets [127, 128]. The targeting of early proteins might be due to an increased presentation on MHC-I, as the peptides are available for presentation early on and the presentation of late proteins can be hindered by MHC-I downregulation by the pathogen [129, 130].

1.4.2 Predicting immune responses

Even though for various HLA molecules significant associations with diseases have been reported [77, 113–115], how this relates to specific T-cell responses is unknown. The study of the effect of specific T-cell responses is obstructed by an incomplete picture of all T-cell epitopes. Only when all epitopes are known, can we start to study their effectiveness, while correcting for the effects of co-occurring responses. The prediction of epitopes could help to gather a complete picture of the T-cell epitopes in an immune response.
Even though T-cell immune responses can be affected by factors that vary in different hosts, such as the peptides presented on other HLA molecules, the T-cell repertoire, thymic selection, regulatory responses and competition, the outcome is highly reproducible [126, 131]. For instance, in $\sim$50% of the HIV-1 patients that express HLA-A*0201, a response towards the SLYNTVATL peptide is observed [126, 131], this response was not observed in patients with another HLA background [131]. Similarly, a response to the EBV-derived FLRGRAYGL peptide was observed in $\sim$90% of the patients with an HLA-B*0801 background [126]. Thus, predicting epitopes seems to be possible, if one wants to attempt this, predictions of MHC-I ligands (section 1.2.5) and immunogenicity (section 1.3.4) would make a good start. Next, predictions on self/nonself overlaps might be possible if one could make a good prediction of the self MHC-I ligandome. It has been shown that the overlap with self helps to explain why certain pMHCs are used in an immune response [117, 118]. Frankild et al. proposed a method to predict the degree of self-similarity that might be used to discriminate epitopes from non-epitopes [118]. Textor et al. proposed another elegant prediction of self-similarity, based on the simulation of negative selection on the T-cell repertoire [132]. Finally, a comparison of different pathogens might help to predict which pathogen-derived proteins serve as a source for T-cell epitopes. We already mentioned the increased chance of early-expressed proteins to serve as antigens. Possibly, other characteristics might predict if peptides from a protein are immune targeted, for instance location of expression, hydrophobicity, or protein stability.

1.5 Overview

We know that peptides should bind to MHC to be recognized by T-cells, but that not every binding peptide is *bound to be an epitope*. In this thesis, we address this problem by asking the question that was introduced in the preface: “Which pMHCs become epitopes?”. We broke up this question in three sub-questions:
1. Which peptides are MHC-I presented? We have studied how some HLA molecules can consistently present more pathogen-derived peptides than self peptides (**Chapter 2**). Hereby, we unraveled an important and general feature of MHC-I peptide binding preferences, namely a preference to present peptides from

pathogens with a low G+C content. Furthermore, we address the cellular proteolysis that is important for the production of MHC-I ligands. We show which predictors should be used to simulate this step, and investigate how proteolysis, TAP transport and MHC-I binding predictors should be combined for an optimal MHC-I ligandome prediction (**Chapter 3**).
2. Which pMHCs are recognized by T-cells, i.e. are immunogenic? We bring together and analyze a large set of pMHCs that have been shown to be immunogenic or non-immunogenic (**Chapter 4**). This analysis enables us to derive determinants of immunogenicity, such as the contribution of different amino acid residues to immunogenicity, and the importance of different positions in the presented peptide for T-cell recognition.
3. Which immunogenic pMHCs are used in an immune response, i.e. are epitopes? We investigate how self-tolerance shapes eventual immune responses, by estimating the self/nonself overlap on different HLA molecules (**Chapter 5**). Our best estimate is that $\sim$30% of the foreign pMHCs overlaps with a self-derived pMHC, and we show how this affects the chance that a pMHC will be an epitope. In addition, we provide a supplement (Addendum) to Chapter 5 to discuss how large self/nonself overlaps influence the establishment of a successfull immune response. Furthermore, we show how the similarity to self affects the immune response to HLA class II presented peptides, and thereby the outcome of an immune response to mismatched MHC-I molecules in kidney transplantation (**Chapter 6**).

Chapter 2

MHC class I Molecules Exploit the Low G+C Content of Pathogenic Genomes for Increased Presentation

JORG J.A. CALIS[*,†], GABINO F. SANCHEZ-PEREZ[*] AND CAN KEŞMIR[*,†] (2010)

[*]Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, The Netherlands
[†]Academic Biomedical Centre, Utrecht University, Utrecht, The Netherlands

European Journal of Immunology, **40**: 2699–2709

Abstract

Distinguishing self from non-self and pathogenic from non-pathogenic is a fundamental challenge to the immune system. Several pathogen associated molecular patterns (PAMPs) are used for this purpose by the innate immune systems. At the adaptive branch of immunity, however, the role of PAMPs is largely unknown. By investigating the presentation of large sets of viruses and bacteria on HLA molecules, we analyze if and how MHC molecules prefer pathogen-derived peptides. The fraction of MHC binders in different organisms are found to vary up to 8 fold. We find that this variation is largely due to strong (dis)preferences for G+C content, which are reflected in amino acid frequencies. Interestingly, a significant majority of HLA-A, but not HLA-B, has a preference for low G+C contents, which seems to be a universal signature for pathogenicity, i.e., a form of PAMP. Finally, we find the same G+C preferences in Chimpanzee and Rhesus Macaque MHC class I molecules. These results demonstrate that despite the fast evolution of MHC alleles, their extreme polymorphism and diversity in peptide-binding preferences, MHC molecules can acquire a preference for presenting PAMPs.

2.1 Introduction

MHC class I (MHC-I) presentation of peptides is crucial for the T-cell response. Peptide-MHC-I complexes (pMHC) are generated in a pathway that mainly involves the cytosolic degradation of proteins by the proteasome, translocation of peptides to the endoplasmatic reticulum by TAP, N-terminal trimming of peptides by aminopeptidases, binding of $8 - 11$ amino acid long peptides to MHC-I molecules and transport and degradation of the pMHC [133–137]. T-cells recognizing specific pMHCs, representative for an infection or tumor, will become activated and proliferate to seek and destroy other cells presenting the same pMHC.

Self- and pathogen-derived antigens are presented on MHC-I molecules simultaneously. Given the enormous turn-over of self-proteins (about 10^6 peptides are generated by the proteasome every second [138]) and limited number of MHC-I molecules ($\sim 10^5$), only a small fraction of pathogen-derived peptides will be presented on the cell surface [139]. Still, the presentation of pathogen-derived peptides is required to elicit an effective immune response. Another requirement is that pathogen-derived pMHC are different from self-derived pMHC, since most self-reactive T-cells are deleted during negative selection [106]. To fulfill these requirements it would help if MHC-I molecules have a preference for pathogen-derived peptides.

Due to the diversity of pathogens that need to be presented on HLA molecules and evolutionary pressures such as the heterozygous advantage and the rare allele advantage, HLA-A/B are the most polymorph genes in the human population [31, 140, 141]. The polymorphic sites are mostly located in the parts that encode the peptide binding groove [141, 142], as a result different HLA molecules prefer to bind different peptides. Most HLA-A/B molecules use the second and ninth position of a peptide as binding anchors, therefore the specificity at these positions is highest. As an example, HLA-A0201 has a preference for peptides with a Leucine at the second and Valine at the ninth position, whereas HLA-B5101 preferably binds peptides with an Proline and Isoleucine at those positions, respectively.

The elucidation of MHC molecule binding motifs and the development of MHC-I pathway predictors enabled the study of MHC preferences.For instance, we recently showed that, despite the diversity in peptide preferences, HLA-A molecules have a shared preference for pathogen-derived peptides [44]. On a small scale, MHC-I predictors have been used to successfully predict new HIV-1 epitopes [143]. On a large scale, the preference for human versus non-human viruses has been investigated [144] and reported to be lower in human viruses. In another study of HLA preferences we showed that HIV-1 gains as many epitopes as it loses during its last 30 years of evolution [145]. Furthermore, the preference for self-derived peptides was studied by Almani et al. and they reported predicted epitopes in the human proteome to contain more non-synonymous single nucle-

otide polymorphisms [146] than non-epitopes.

None of the above studies came up with a mechanism explaining the observed preferences of HLA-A/B molecules, despite the diversity of peptide binding preferences. To get more insight in such a mechanism we investigated the presentation of large sets of viral and bacterial proteomes by simulating key-processes of the MHC-I presentation pathway, the proteasomal degradation, TAP translocation and MHC-I binding [50, 143, 147, 148]. HLA molecules display a large variation in the predicted fraction of presented peptides in these organisms. We here show that up to 95% of this variation can be explained by genomic G+C content differences. Also, using naturally occurring experimentally verified epitopes, we could demonstrate these G+C preferences. Most interestingly, a majority of HLA-A, but not HLA-B molecules, targets G+C low species. Since pathogenicity is associated with a low G+C content [149], these results suggest that HLA-A alleles have been under selection to encode binding motifs that preferentially present G+C low, and therefore most likely pathogenic, species. Finally, we observe the same G+C negative preference in non-human primate MHC-I molecules. Taken together, we believe that we discovered a novel feature of HLA class I presentation that provides a mechanism to explain previous observations on HLA preferences.

2.2 Results

2.2.1 HLA molecules are responsive to G+C content differences

HLA alleles encode HLA molecules with a large variety of binding preferences [10]. Though random mutations generate new MHC molecules with different peptide binding groove, some binding preferences might be selected. For instance, if an HLA molecule can present a large repertoire of pathogen-derived peptides it could confer an advantage to its host and grow in the population. Indeed, we showed that HLA-A prefer pathogen-derived peptides over human peptides [44]. To further investigate mechanisms behind this preference, we predicted the presentation of a large set of viruses (1223 non-redundant viral proteomes, see Methods) using high quality, validated MHC-peptide binding predictors. Unsimilar HLA-A/B molecules for which two predictors were available were selected, this unfortunately excluded HLA-C (explained in detail in Methods). The 11 selected HLA-A molecules and 10 HLA-B molecules are expressed in 88% and 52% of the caucasian population, respectively. The fractions of presented peptides of viruses on HLA-A/B molecules vary greatly, up to 60-fold (results not shown). A large part of the variation is due to a small proteome size: the maximal variation within the 100 smallest viruses is 25-fold, while the variation within the largest 100 viruses is maximally 8-fold (see Supplementary Figure 2.S1). To circumvent

this source of variation, the analysis was repeated with bacterial species, which have larger proteomes than viruses. Though not as extreme as with viruses, in a data set of 174 bacteria the fractions of presented peptides can vary largely, for some HLA molecules up to 6-fold.

A group of 15 bacteria, (containing among others Bacillus cereus, that causes food poisoning and pneumonia) showed very high predicted fractions of presented peptides on most HLA molecules. This specific subset could be described by their low G+C contents, which suggested that variations in fractions of presented peptides could be explained by variations in G+C contents. To test this hypothesis, a possible correlation between G+C content and fraction of presented peptides was examined. In Figure 2.1 an example is shown for 6 HLA molecules, every dot represents the fraction of presented peptides and the G+C content of a bacterium (all 21 HLA molecules are shown in Supplementary Figure 2.S2). Clear G+C preferences can be observed, e.g. HLA-A*1101 has a preference for bacteria with a low G+C content, whereas HLA-B*0702 favours the presentation of G+C high bacteria. To formalize these observations an HLA molecule was coined G+C responsive upon observation of a significant ($p<0.001$) correlation between G+C content and fraction of presented peptides within a set of 174 bacteria. In addition, correlation thresholds to assess G+C negativity, neutrality or positivity were chosen such that at least 70% of the variation in fraction of presented peptides would be explained by variations in G+C content. A majority of the HLA molecules (14 out of 21) was responsive to G+C content: 10 G+C negative and 4 G+C positive (see Table 2.1). This result was consistent for different MHC binding prediction methods, different parameter settings and not dependent on peptide processing predictions. A large determinant of peptide binding preferences are the anchor positions, G+C preferences based on only these two positions overlap significantly with those determined using all positions (Spearman rank test: correlation=0.83, $p<.001$).

To test whether the observed high fraction of G+C responsive molecules was to be expected, random HLA molecules were made, by shuffling the binding motifs of original HLA molecules (see Methods), and checked for G+C responsiveness. Random HLA molecules have a specificity similar to real HLA, e.g. on anchor positions only one or two amino acids are allowed to bind. However, since the binding specificities are randomly distributed over the 20 amino acids there is no bias for specific (e.g. G+C informative) amino acids. Contrary to our expectations, 66% of the random HLA molecules is G+C responsive. Apparently, the observed 67% of G+C responsive HLA molecules (14/21) was to be expected (Permutation test, $p=0.57$).

In the random-HLA-model G+C positive and negative molecules were equally distributed, 32.1% and 33.8 % respectively. In contrast, the majority of real HLA-A molecules (9 out of 11) is G+C negative (See Table 2.1). Compared to the distribution of random HLA molecule G+C responsiveness, this fraction of G+C

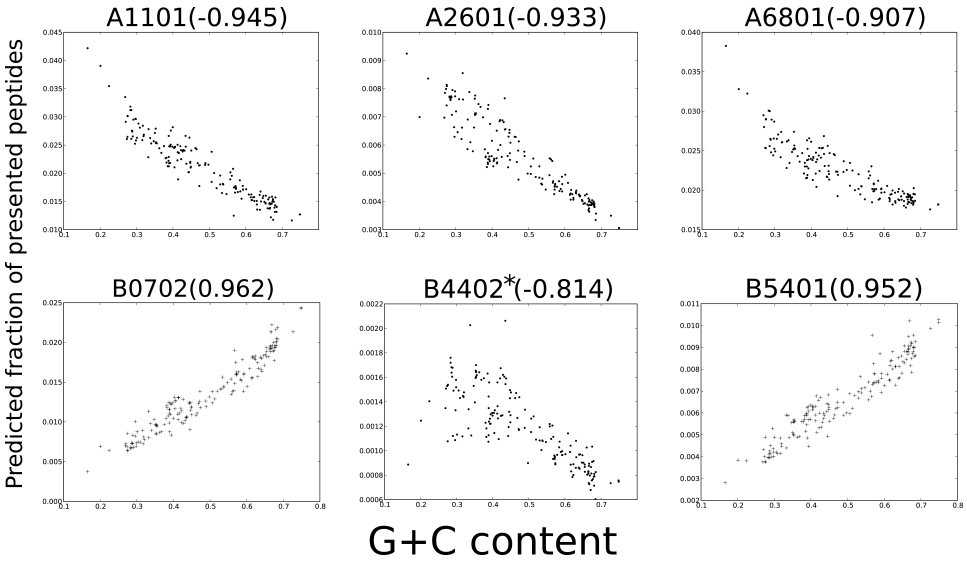


Figure 2.1. HLA molecules have strong G+C preferences. G+C content and fractions of presented peptides are shown for six representative HLA molecules. Every data point gives the predicted fraction of presented peptides (y-axis) and G+C content (x-axis) of one of the 174 bacteria. Correlations between G+C content and the predicted fractions of presented peptides are significant (Spearman rank test: $p<0.001$) in all six cases. The Spearman rank correlation coefficient for each HLA is given in brackets. Data points are presented as dots in case of a G+C-negative HLA molecule (A1101, A2601, A6801 and B4402), or as plus signs for G+C-positive HLA molecules (B0702 and B5401). HLA molecules indicated with an asterisk (*) showed inconsistent correlation values for different methods or parameter settings.

negative HLA-A molecules is significantly high (Permutation test: $p=0.0017$). Similarly, there is a significant underrepresentation (0 of 11) of G+C positive HLA-A molecules (Permutation test: $p=0.014$, see Table 2.1). For HLA-B, the number of G+C positive and G+C negative molecules was not significantly different from the random HLA model (Permutation test: $p=0.41$, see Table 2.1). Taken together, these results suggest that while all HLA molecules are sensitive to G+C content, only HLA-A molecules have a strong preference to present G+C low organisms.

2.2.2 Pathogens have a low G+C content and different amino acid usage

Although G+C content is related to genome size and pathogenicity [149, 150], causality remains elusive and it is unclear what selection pressures determine the

GC+ molecules	GC- molecules		GC uncorrelated molecules
	A0101(-0.89)	A2601(-0.93)	A0201(-0.31)
	A0301(-0.95)	A2902(-0.93)	A3301(0.79)*
	A1101(-0.95)	A3001(-0.89)	
	A2301(-0.94)	A6801(-0.91)	
	A2402(-0.94)		
B0702(0.96)	B1801(-0.73)		B1501(-0.59)*
B2705(0.90)			B4402(-0.81)*
B3501(0.86)			B5101(0.83)*
B5401(0.95)			B5701(0.45)*
			B5801(0.45)

Table 2.1. HLA molecules and their G+C preferences. Spearman-rank Correlation values between G+C contents and fractions of presented peptides of the bacteria set for all HLA molecules. HLA-molecules indicated with an asterisk (*) showed inconsistent correlation values for different MHC-thresholds or MHC prediction method.

G+C content of organisms. We tested whether in our bacterial and viral data set pathogenicity relates to G+C content, as this would provide an explanation for the G+C low preference of HLA-A molecules. Here we focus only on pathogens presented by the MHC-I pathway, viruses and intracellular bacteria. In our data sets the G+C content of pathogenic intracellular bacteria and viruses is significantly lower than that of non-pathogenic bacteria (Ranksums test: $p< 0.01$, see Figure 2.2). Similarly, the G+C content of pathogenic intracellular bacteria and viruses is significantly lower than that of human coding sequences (Ranksums test: $p< 0.01$, see Figure 2.2). Thus, pathogens tend to have a low G+C content which may have been exploited by HLA class I molecules to preferably present pathogens.

How can HLA class I molecules detect G+C content differences? Because HLA molecules bind short protein fragments, this is only possible if genomic G+C content differences relate to differences at the proteomic level. Differences in G+C content have been shown to bias amino acid usage in bacteria [151]. High G+C contents result in an increased frequency of the amino acids G, A, R and P and low G+C contents result in an increased usage of F, I, N, K and Y [151–153]. This trend is also visible in our data set: In Figure 2.3 we depict relative amino acid frequencies for all bacteria in our data set, sorted on their G+C content. The amino acids G, A, R and P are positively correlated and amino acids F, I, N, K and Y are negatively correlated to G+C content. Also, the G+C positive and negative amino acids cluster together as their frequency profiles are highly similar (highlighted in red and green in Figure 2.3). So, genomic G+C content

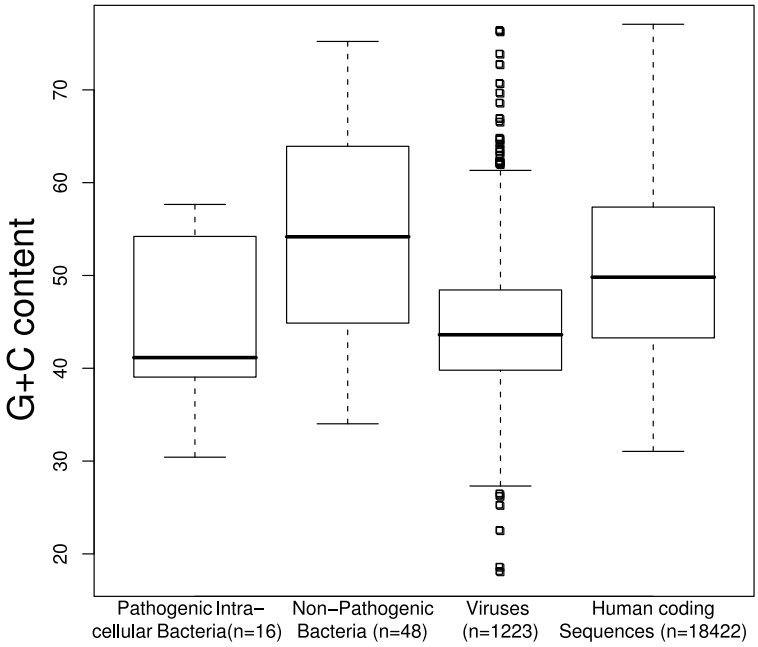


Figure 2.2. Pathogenic organisms have a low G+C content. G+C contents of pathogenic intracellular bacteria, nonpathogenic bacteria, viruses and human coding sequences are shown. Pathogenic intracellular bacteria and viruses have lower G+C contents than nonpathogenic bacteria (rank sums-test; $p=0.004$ and $p<0.001$, respectively) and human coding sequences (rank sums-test; $p=0.006$ and $p<0.001$, respectively). The pathogenic groups, pathogenic intracellular bacteria and viruses, are not different from each other (rank sums test; $p=0.60$). The median (and average) G+C contents of pathogenic intracellular bacteria, nonpathogenic bacteria, viruses and human coding sequences are 40.8%(44.1%), 53.9%(53.4%), 43.3%(44.0%) and 49.5%(50.1%), respectively. The box-and-whisker-plot shows the median of the data set as a thick black line, the box is formed by the first and third quartile of the data set, the upper (or lower) whisker is the minimum (or maximum) of either the third quartile plus (or the first quartile minus) 1.5 times the interquartile range or the maximal (or minimal) value in the data set. Data points outside the plot are shown as boxes.

differences translate into proteomic differences that can be sampled by HLA class I molecules.

Next, we analyze experimentally verified epitopes for G+C positive (GARP) and G+C negative (FINKY) amino acid frequencies to validate the predicted G+C preferences as reported in Table 2.1. If the predicted G+C preferences are correct, an increased frequency of GARP or FINKY should be observed in naturally occurring epitopes. A database perfectly suited for this test is the SYFPEITHI database [45], that almost exclusively contains naturally occurring epitopes. For 9 G+C negat-

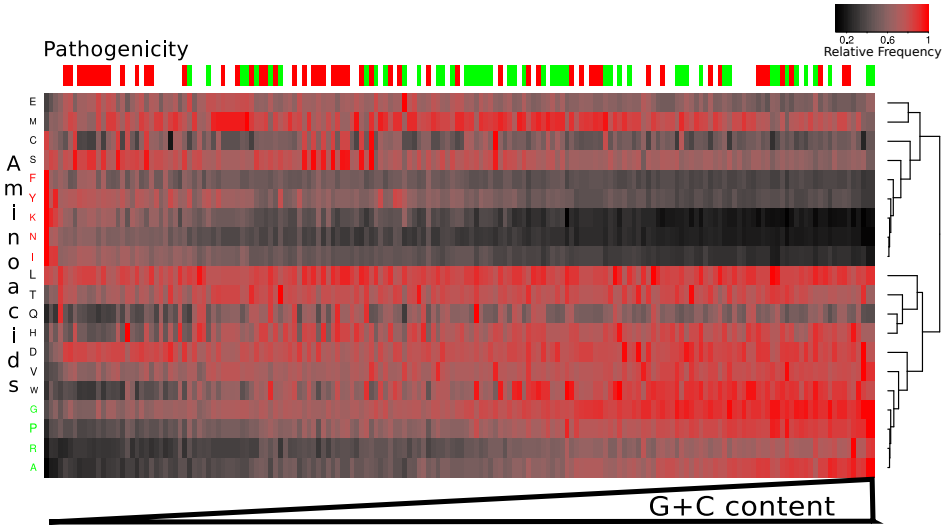


Figure 2.3. G+C content and amino acid frequencies correlate. All bacteria in our data set are sorted based on the G+C content, highest G+C contents on the right, lowest G+C contents on the left. Above the heat map, pathogenic bacteria are indicated red, nonpathogenic bacteria green and unknown bacteria white. Per amino acid (in rows) the relative frequency of that amino acid in each bacterial proteome (in columns) is plotted. The relative frequency is the amino acid frequency in a certain bacterium divided by the maximum of the frequencies of that amino acid in all bacteria. The colors in the heat map correspond to the relative frequency as shown in the upper right panel. Amino acids are clustered according to their frequency profiles, the G+C-negative amino acids F, I, N, K and Y (red) and the G+C-positive amino acids G, A, R and P (green) form separated clusters.

ive, 4 G+C positive and 3 G+C neutral HLA class I molecules that have at least 10 epitopes in SYFPEITHI, the frequency of GARP and FINKY in their epitope sets was determined. As expected, the G+C negative HLA class I molecules have epitopes with a higher FINKY content than the neutral (Ranksums test: $p=0.013$) or positive (Ranksums test: $p=0.031$) HLA class I molecules. Similarly, the epitopes of the G+C positive HLA molecules have a higher GARP content than the neutral (Ranksums test: $p=0.034$) or negative (Ranksums test: $p=0.021$) HLA class I molecules. So, the predicted G+C preferences are confirmed by the naturally occurring experimentally verified epitopes found in SYFPEITHI.

On the whole, a low G+C content is a property of organisms with a pathogenic lifestyle [149, 150] and differences in G+C content are reflected in amino acid usage [151]. These properties appear to have shaped the binding motifs of the common present day HLA-A, but not HLA-B, molecules to enhance the presentation of pathogens.

2.2.3 HLA-A negativity is not caused by a lack of diversity or rigid G+C preferences

The preferred presentation of pathogens provides a tantalizing explanation for the observed G+C negativity of HLA-A molecules. However, this preference does not need to be a direct result of natural selection. First of all, the G+C negativity of HLA-A can be due to a lack of divergence: if the common ancestor of all HLA-A alleles encoded a G+C negative HLA molecule and divergence of their binding motifs during evolution would be minimal, present day HLA-A molecules would still be G+C negative. Such a lack of divergence would be visible when HLA-A molecules were compared with HLA-B molecules, which show large variations in G+C preferences. However, phylogenetic distances among HLA-A alleles were not different from the distances among HLA-B alleles (Ranksums test: $p=0.89$). Also at the protein level, HLA-A molecules are as different as HLA-B molecules (Ranksums test: $p=0.089$). Even when divergence was determined on peptide binding sites only (see Methods), that are known to be most variable, no difference between HLA-A and HLA-B could be observed (Ranksums test: $p=0.24$). Therefore, a lack of divergence does not seem to explain the shared G+C negativity of HLA-A molecules, since a similar amount of diversity can result in mixed G+C preferences in the case of HLA-B.

Second, a restrained flexibility of HLA-A binding groove preferences might be the reason for G+C negativity. If HLA molecules were unable to alter their binding preferences, despite divergence at the genomic and proteomic level as described above, this would lead to a shared G+C preference. A phylogeny of all HLA-A alleles was made to investigate the flexibility of G+C preferences during HLA evolution. The maximum likelihood phylogeny of HLA-A alleles, Figure 2.4, shows that the two non-G+C negative encoding HLA-A alleles (i.e. A*0201 and A*3301) do not cluster together, and thus suggests that G+C preferences were able to change at least twice independently during HLA-A evolution. To further emphasize the flexibility of HLA preferences during evolution, binding motifs of the encoded HLA molecules are displayed in Figure 2.4. The diversity of binding motifs shows that HLA-A molecules are not likely to have rigid binding preferences.

Finally, since G+C content and amino acid usage are linked (Figure 2.3), it might be that HLA-A molecules evolved G+C negativity by all binding the same G+C informative amino acid, e.g. the G+C negative isoleucine (I). To test this explanation, the strategies behind a preferred presentation of G+C low bacteria were examined by investigating G+C responsiveness upon *knocking out*, i.e. setting the binding preference to 0.0, any of the 20 amino acids for all HLA-A molecules (see Methods). In Supplementary Figure 2.S3 all 20 *knock out* results per HLA are presented. Only 8 out of 20 amino acids had an effect on the G+C preferences of any of the HLA alleles. Setting the binding preferences for lysine to zero had the largest effect on HLA-A G+C preferences, (highlighted in Supplementary Figure 2.S3): only a modest fraction of 3 (of the 11) HLA-A alleles changed their

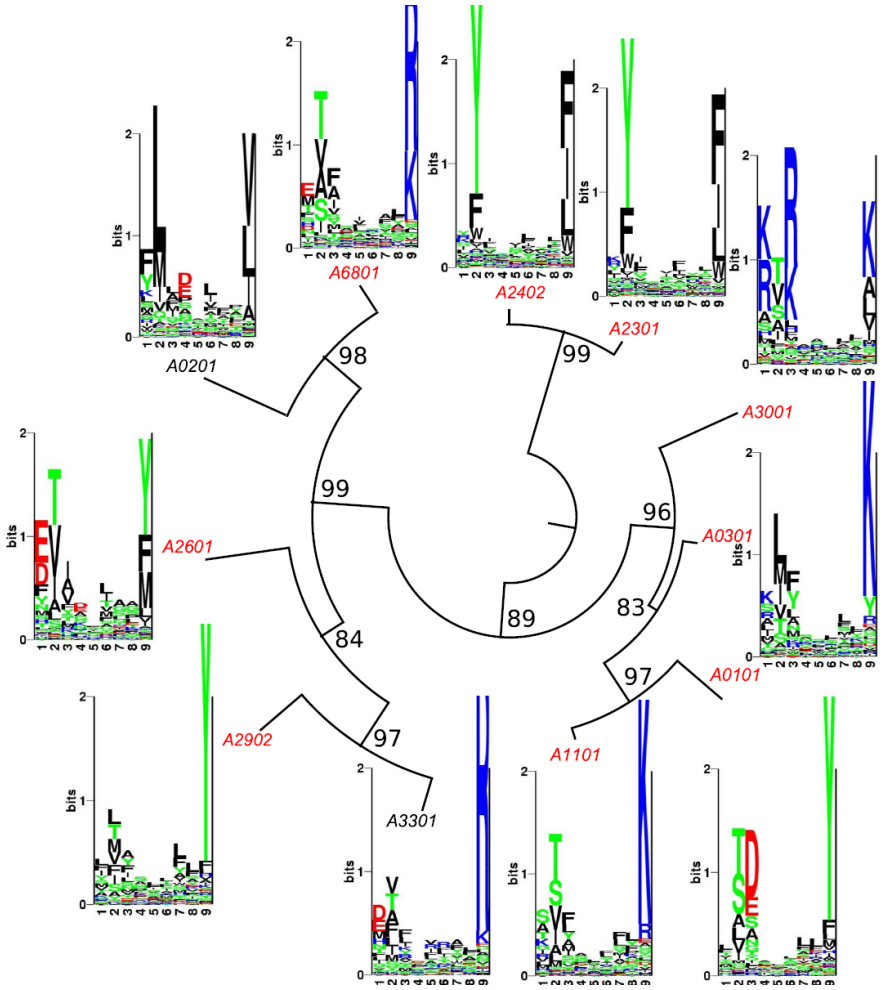


Figure 2.4. HLA-A phylogeny and binding motifs. The phylogeny analysis of HLA-A alleles is based on ClustalW alignment and maximum-likelihood clustering. The numbers shown in the phylogeny indicate how often a clade in the phylogeny was observed in 100 bootstrap analyses. Red and black HLA names are indicating G+C-negative and G+C-neutral binding preferences, respectively. For every HLA, the binding motif obtained from the MHC motif viewer website [30] is shown to indicate the encoded peptide-binding motif.

G+C preference. This result indicates that G+C negative preferences are not obtained by the preferred binding of a single G+C informative amino acid. Instead, as is suggested by the diversity of binding motifs (Figure 2.4), a diverse set of strategies have evolved by HLA-A molecules to become G+C negative.

In summary, G+C negativity among HLA-A alleles is neither due to a lack of divergence, nor is it caused by constrains on G+C preferences during HLA evolution. Therefore, we conclude that an optimization to present pathogen-derived peptides provides the best explanation for the shared G+C negativity of HLA-A alleles.

2.2.4 G+C responsiveness of non-human MHC class I molecules

If the selection of G+C negative MHC class I molecules, and thereby an enhanced presentation of pathogens, is an important aspect of MHC class I evolution, subsets of MHC molecules from other species would also be G+C negative. This hypothesis was tested by investigating the G+C preference of Chimpanzee MHC-I molecules (Patr molecules). Chimpanzees were chosen because their MHC region is quite similar to that of humans. Moreover, well investigated Chimpanzee MHC-I binding predictors are available [49].

The G+C preference of 47 Patr molecules was investigated on our bacteria data set. Similar to HLA molecules, a majority of Patr molecules, 26 of 47, was found to be G+C responsive (see Supplementary Table 2.S4). Moreover, a majority of Patr-A molecules, i.e. 15 out of 18 (Permutation test: $p<0.001$), are G+C negative. Patr-B molecules, like HLA-B molecules, show mixed G+C preferences and the fraction of G+C negative Patr-B molecules is not significantly different than expected (Permutation test: $p=0.91$). These results suggest that not only in humans but also in Chimpanzees there has been a selection pressure to obtain G+C negative peptide binding preferences in MHC molecules encoded by the A-locus.

A more distant primate for which the MHC-I G+C preferences can be investigated is the Rhesus Macaque. However, compared to the HLA locus, the Rhesus Macaque MHC-I (Mamu) locus is organized quite differently, i.e. the Mamu-A and -B genes duplicated at least once and twice, respectively [154]. Possibly, as a result of these duplications, G+C preferences of Mamu molecules seem to be different: not Mamu-A molecules, but Mamu-B molecules are mostly G+C negative (see Supplementary Table 2.S5). Unfortunately, becasuse they are rather diverged from HLA molecules where the majority of peptide binding data is available, Mamu binding predictions are less accurate. 23 of the 40 Mamu molecules are G+C neutral. Still, 11 Out of 23 Mamu-B molecules are G+C negative (Permutation test: $p=0.117$), while none of them can be classified as G+C positive (Permutation test: $p<0.001$). To conclude, in both humans and Chimpanzees, and to some extent in Rhesus macaques, we find one MHC class I gene that seems to be dedicated to presenting G+C low pathogens.

2.3 Discussion

It is well established that the innate immune system uses conserved microbial components for danger signaling and self/non-self discrimination. A good example is the recognition of LPS by TLR 4 [155]. We here find that also the adaptive immune system uses a pathogen specific signature, i.e. a low G+C content, to enhance the presentation of pathogenic organisms and increase self/non-self discrimination. A significant majority of HLA-A molecules has G+C low preferences, despite a large variety in HLA-A binding motifs. Also, Chimpanzees and Rhesus Macaques have a G+C negative MHC class I locus, demonstrating that G+C preferences are an important factor in MHC class I evolution across species. Opposite to humans and Chimpanzees, in Rhesus Macaques the B-locus instead of the A-locus is G+C negative. Humans and Rhesus Macaques diverged approximately 27 Million years (Myr) ago, whereas Chimpanzees diverged about 5.4 Myr ago from humans [156]. It will be interesting to know if the common ancestor of all Catarrhine primates had a G+C negative A- or B-locus. This question could be solved if more high-quality MHC-I predictors for Gorilla, Orangutan or Gibbon would be developed.

We previously observed that HLA-A, but not HLA-B molecules favor the presentation of bacteria and viruses over self [44]. This observation can now be explained by differences in G+C contents (see Figure 2.2) and different G+C preferences. HLA-B molecules, though at least as diverse as HLA-A, do not show any sign of selection for G+C negativity. Still, HLA-B molecules are as functional as their HLA-A counterparts in antigen presentation and for some pathogens they elicit dominant immune responses more frequently than HLA-A molecules [124–126]. If HLA-B molecules are important in the presentation of (G+C low) pathogens, one expects them to become G+C negative as well. We propose that one G+C negative HLA locus (HLA-A) might be sufficient to capture the general feature of A+T richness. Consequently, the other HLA locus (HLA-B) can evolve with specific pathogens. That HLA-B and not HLA-A alleles evolve via recombination might help HLA-B to perform pathogen-specific adaptations faster [141, 157, 158]. The higher polymorphism of HLA-B alleles (1249 HLA-B versus 853 HLA-A alleles known in the HLA/IMGT-database (October 2009) [159] and more rapid selection of new HLA molecules in the B-locus [141, 157, 158, 160] are in agreement with this view. Also, pathogen specific evolution of HLA-B molecules might explain the observed immunodominance of HLA-B molecules in fast evolving pathogens such as HIV-1 [124].

Although having a low G+C content is a general feature of pathogenic life [149], several G+C rich pathogens exist. These pathogens can be presented by G+C positive HLA-B molecules. Furthermore, G+C rich pathogens are enriched in unmethylated CpG-motifs, that can be recognized by the immune system using TLR9 [161, 162]. So, G+C positive HLA class I molecules and molecules like

TLR9 might complement the G+C negative HLA-A molecules to cover a possible hole in the immune responses to G+C high pathogens.

Recently, Vider-Shalit et al. claimed that human herpes viruses have fewer CTL epitopes than their non-human family members [144]. They analyzed different viruses by a presentation score called "Size of Immune Repertoire (SIR) score", that is claimed to express how well a sequence is presented population-wide. Since we here demonstrate that G+C content is of major influence on how well a pathogen is presented, we tested whether the SIR scores reported by Vider-Shalit et al. [144] correlate with the G+C content of different viruses. For 16 herpes viruses, the G+C content could be determined and a very good correlation with the SIR scores was observed (Spearman rank test: correlation=-0.81; $p<0.001$). Possibly, the claimed difference between human and non-human herpes viruses is due to G+C content differences in these viruses.

Eventhough it was well established that a low G+C content is a general feature of pathogenic life, to our knowledge, this is the first study demonstrating the selection of HLA molecules based on G+C preferences. To conclude, we believe that the results reported here contribute significantly to understanding the nature of MHC-I presentation preferences and reveal an important aspect of MHC-I evolution: namely that a subset of MHC molecules have been selected to capture a general feature of pathogenic life, a low G+C content.

2.4 Methods

2.4.1 Genomics, proteomics and pathogenicity data

The fractions of presented peptides of the human proteome, 300 bacterial proteomes and 2165 viral proteomes were analyzed. All proteomes and genomes were downloaded via http://www.ebi.ac.uk, the human proteome and genome in May 2008, the bacterial and viral proteomes and genomes in October 2008. G+C contents were calculated from the genomic data sets. The number of bacterial proteomes was reduced to 174 by randomly picking one strain per species. Viral proteomes with a similarity higher than 80% were considered to be non-unique and were excluded, resulting in a selection of 1223 unique viral proteomes.
The 174 bacteria were sorted into three groups based on pathogenicity, i.e. "pathogenic", "non-pathogenic" and "unknown". Pathogenicity data was derived from the NCBI prokaryotic genome project table (attributes "Pathogenic in:" and "Habitat", http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Bacteria were defined "pathogenic" if they are pathogenic to humans. Bacteria were defined "non-pathogenic" if they are not pathogenic to plants or animals. To be confident on the non-pathogenic bacteria, bacteria with a known pathogen in their

genus or host-associated habitat were excluded from the non-pathogenic group and defined “unknown”. Also, bacteria pathogenic to non-humans were defined “unknown”. Following these criteria, 53 bacteria were defined “pathogenic”, 48 bacteria were defined “non-pathogenic” and 73 bacteria were defined “unknown”. 16 Intracellular bacteria were selected from the pathogenic group based on the organism description in the NCBI genome project.

2.4.2 MHC-I presentation predictions

The number of epitopes present in a protein or proteome can be predicted by using tools for the three key-processes of MHC-I presentation. Out of several prediction methods, we used the ones that are proven to be the best in large benchmark studies [50, 147], and we checked the reliability of our results by testing alternative methods and parameter settings. Since this study focuses on differences between MHC molecules, the most crucial predictions are for MHC-peptide binding affinities. Based on the benchmark study of Peters et al. [50], we used the best performing MHC binding predictor, NetMHC-3.0 [46, 47], which is an allele-specific neural network based predictor trained with large sets of experimental peptide-MHC binding data. All binding predictions were checked for consistency by an alternative method, a Scoring Matrix Method (SMM)-based MHC-binding prediction tool [163]. Unless mentioned otherwise we obtained similar results with both methods.
To assess whether a peptide binds to an MHC molecule depends on the choice of binding threshold, and recently it has been discussed extensively [164]. If one assumes that all MHC molecules use a fixed threshold, the threshold of 500 and 5000 nM threshold can be used [46, 47] to define binding peptides. If, on the other hand, one assumes that all MHC molecules present the same number of peptides, e.g. 2% of a proteome, a scaled threshold can be used. Since we investigate MHC preferences and want to ensure absolute consistency of our results with respect to methods and parameters we choose to use the fixed thresholds of 500 nM and 5000 nM and a scaled MHC allele specific threshold defined as the threshold that corresponds to a 2% fraction of presented peptides of the human proteome. Results presented are derived using a 500 nM threshold and they are similar for all parameter settings, unless mentioned otherwise.
For Chimpanzee and Rhesus Macaque MHC-I binding predictions NetMHCpan-2.0 [49, 165] was used, a neural network based predictor that uses MHC-I binding groove polymorphisms to predict peptide-MHC binding affinities. Unfortunately, no alternative MHC-I binding prediction methods for these non-human primate MHC-I molecules are available on a large scale to check the predictions made using NetMHCpan-2.0. Consistency of the predictions was checked for the fixed MHC binding thresholds 500 nM and 5000 nM, and the results were similar unless mentioned otherwise.
Processing of the peptides, proteasomal degradation and TAP transport, is pre-

dicted by NetChop Cterm3.0 [16, 51]. To test whether the results depend on the peptide processing predictions, tests were repeated without peptide processing. The results remain similar, unless mentioned otherwise. Finally, the "fraction of presented peptides" is defined as the number of peptides predicted to be presented, divided by the number of all possible 9mers in the protein or proteome.

2.4.3 MHC-I allele selection

NetMHC-3.0 provides neural network predictors for 43 HLA molecules, covering 28 HLA serotypes [46, 47]. To reduce similarity in our analysis we choose to use only the most frequent molecule per serotype (e.g. A*0201 is chosen to represent the A2 serotype and B*2705 for the B27 serotype). To be able to check for consistency of our predictions we excluded those HLA molecules for which either SMM predictions [163] were not available, or the predictions of both methods showed inconsistent qualities (Spearman rank test: correlation <0.7 or $p>0.001$). This resulted in the selection of 11 HLA-A and 10 HLA-B molecules (Table 2.1). Unfortunately, NetMHC-3.0 does not provide neural network predictors for HLA-C molecules, furthermore NetMHCpan-2.0 has been described to have bad HLA-C predictors [49]. Therefore, HLA-C molecules were not included in this study.
NetMHCpan-2.0 provides predictors for 70 Patr molecules and 72 Mamu molecules [49]. To reduce similarity in our analysis we used only unique 2-digit molecules, resulting in the selection of 47 Patr molecules (Supplementary Table 2.S4) and 57 Mamu molecules. NetMHCpan-2.0 predictors are based on similarity to well-characterized MHC-I molecules with large peptide binding data-sets, the pMHC binding affinity for MHC molecules more distant to the training-set is underestimated [49]. To prevent using these low-quality predictors we excluded MHC-I predictors that predicted a fraction of presented peptides of zero in any of the 174 bacteria below a 500 nM threshold, thus selecting 40 Mamu molecules (Supplementary Table 2.S5).

2.4.4 Random HLA model

Predictors for random HLA molecules were generated using SMM-matrices [163]. An SMM matrix consists of nine position vectors corresponding to the nine positions of a 9mer. Every position vector consists of likelihood estimates for 20 amino acids. Random HLA molecules were made by combining position vectors from nine different randomly chosen HLA molecules, from the HLA-A/B set described above and in Table 2.1. Subsequent randomization of the amino acid binding-values, e.g. Lysine gets the preference of Arginine, ensured random aminoacid preferences. Importantly, because position vectors in the random HLA molecules maintain their original positions, general HLA characteristics, such as having two anchor positions, are preserved in the random HLA molecules.

2.4.5 HLA phylogeny and distances

Genomic and protein sequences of HLA molecules were downloaded from the IMGT/HLA database [159] in January 2009. Using ClustalW a multiple sequence alignment was made from the genomic sequences. From this alignment, using the dnadist program from the Phylip-3.68 package, pairwise distances were calculated using the Jukes-Cantor distance measure. A maximum likelihood tree was estimated, also on the genomic sequences, by RAxML_HPC version 7.0.4 [166] using the GTRMIX model and bootstrap support from 200 replicates (Figure 2.4). MHC binding motifs complementing the phylogeny are obtained from the MHC motif viewer website, www.cbs.dtu.dk/biotools/MHCMotifViewer/ [30]. From a ClustalW multiple sequence alignment of protein sequences, pairwise distances were calculated using the Dayhoff PAM matrix measure in the protdist program provided by the Phylip-3.68 package. HLA peptide binding sites are defined as amino acids within maximally 5 Å of the binding peptide in determined pMHC structures [165]. A list of all peptide binding sites is provided in supplementary file S6. From a ClustalW multiple sequence alignment of binding sites only, pairwise distances were calculated using the Dayhoff PAM matrix measure in the protdist program.

2.4.6 Amino acid knock out analysis

The contribution of single amino acids on the G+C responsiveness of HLA molecules was tested by setting the binding preference of a specific amino acids to 0.0 in the SMM matrices [163] in each of the 9 position vectors. These revised SMMs were used to predict fractions of presented peptides and assess G+C responsiveness on the bacteria set. We refer to this analysis as an amino acid *knock out* analysis.

2.5 Acknowledgments

We thank Rob de Boer for valuable comments on the manuscript and discussion on this research project, Ronald Bontrop for discussion on the manuscript and Boris Schmid, Xiangyu Rao and Ilka Hoof for technical support. This study was financially supported by the University of Utrecht and the Academic Biomedical Centre. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

2.6 Supporting Information

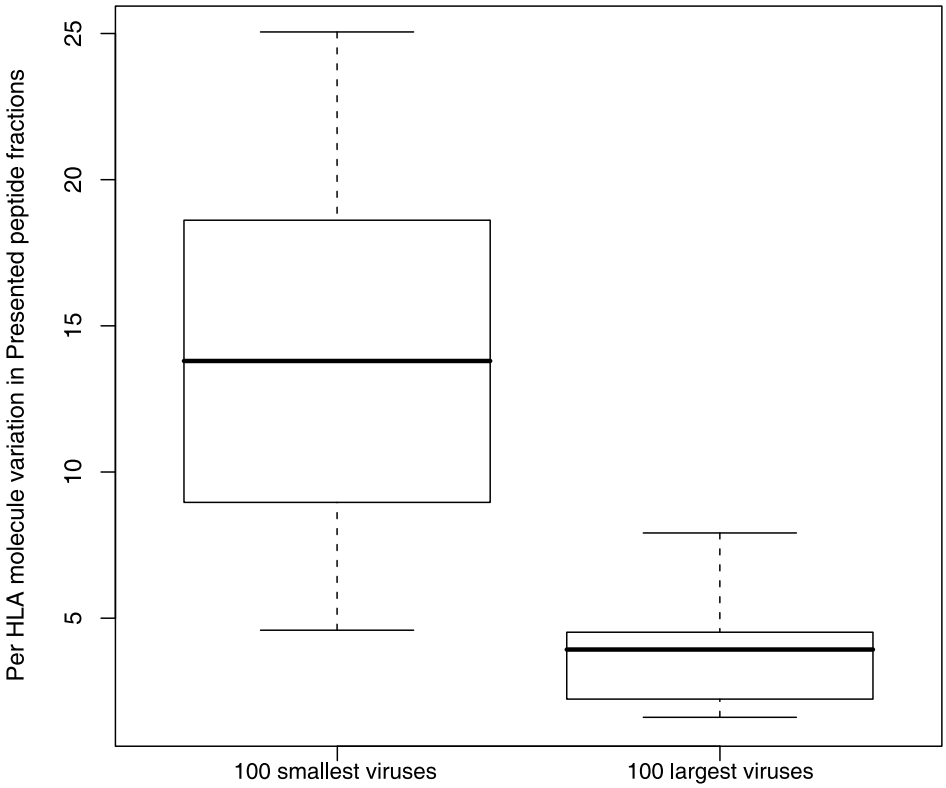


Figure 2.S1. Vatiation in the predicted fractions of presented peptides for small and large viruses. For the 100 smallest and largest viruses the variation in predicted fractions of presented peptides was determined per HLA molecule. These variations for all HLA molecules in our study (n=21; see Table 2.1) are presented in the boxplots.

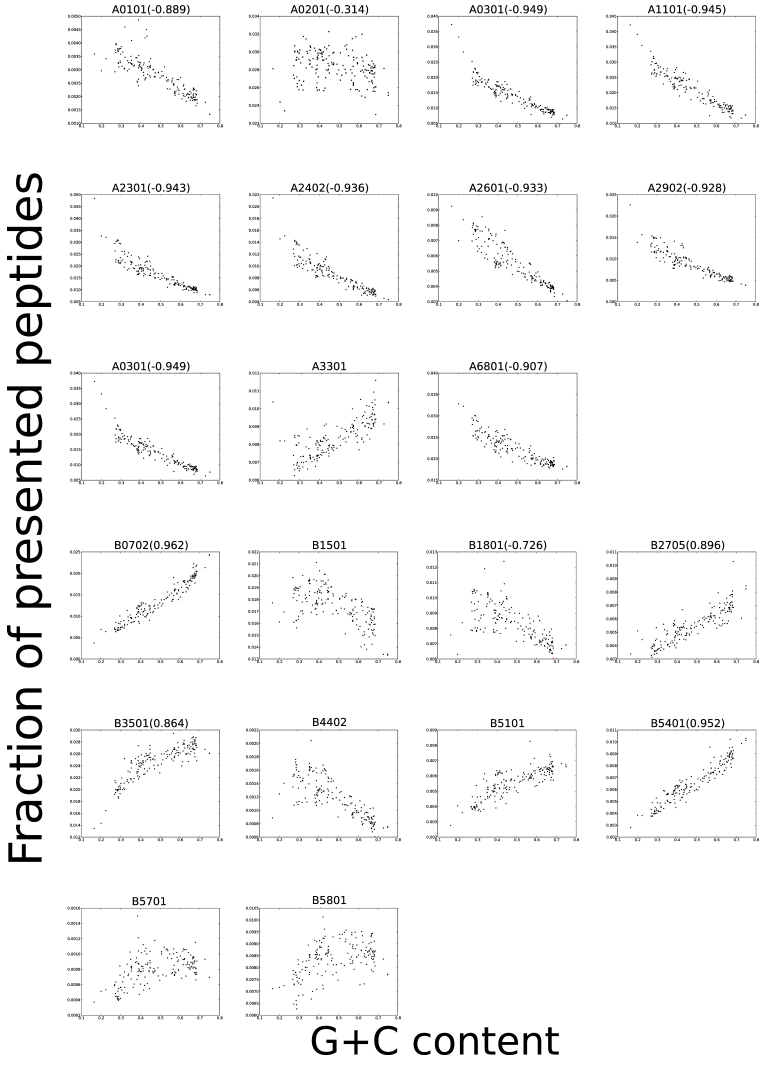


Figure 2.S2. G+C content versus the predicted fraction of presented peptides for all HLA molecules included in this study. In brackets is the Spearman rank test correlation value for G+C content versus the fraction of presented peptides given. Every data point depicts the predicted fraction of presented peptides and G+C content of one of the 174 bacteria and is colored red in case of a G+C negative HLA molecule, black in case of a G+C neutral HLA molecule, or green if the HLA molecule is G+C positive. In case of inconsistent correlation values for different parameter settings of methods, the correlation value is not presented.

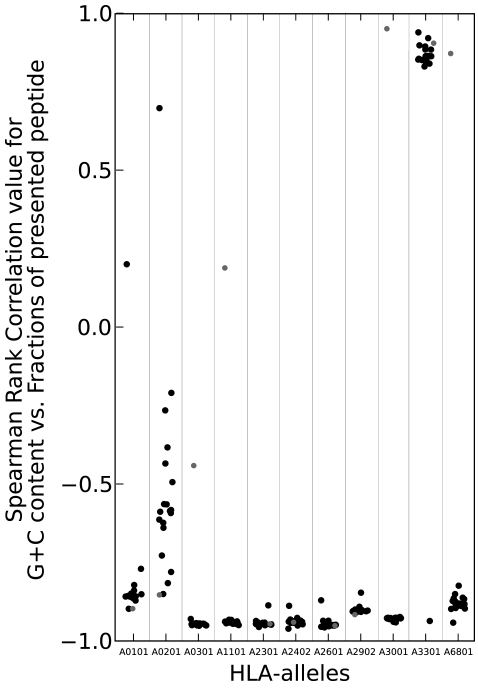


Figure 2.S3. Spearman-rank correlation values between G+C contents and predicted fractions of presented peptides of the bacteria set in an amino acid knock out analysis (see Methods) are shown. The analysis is performed to get more insight in the impact of single amino acids on the G+C preference of HLA-A molecules. Knocking out lysine (K) had the largest effect on HLA-A G+C preferences, therefore this analysis is highlighted in grey. HLA-A*3001 and HLA-A*6801 clearly rely on Lysine for their G+C negativity, upon knocking out Lysine they become G+C positive. For HLA-A*1101 and HLA-A*0301 the effect is smaller, both are classified as G+C neutral when knocking out Lysine, though for HLA-A*0301 we still observe a negative correlation value of -0.44. Five G+C negative HLA-A molecules rely on other amino acids for their G+C preference, here knocking out Lysine had no effect.Spearman-rank correlation values between G+C contents and predicted fractions of presented peptides of the bacteria set in an amino acid knock out analysis (see Methods) are shown. The analysis is performed to get more insight in the impact of single amino acids on the G+C preference of HLA-A molecules. Knocking out lysine (K) had the largest effect on HLA-A G+C preferences, therefore this analysis is highlighted in grey. HLA-A*3001 and HLA-A*6801 clearly rely on Lysine for their G+C negativity, upon knocking out Lysine they become G+C positive. For HLA-A*1101 and HLA-A*0301 the effect is smaller, both are classified as G+C neutral when knocking out Lysine, though for HLA-A*0301 we still observe a negative correlation value of -0.44. Five G+C negative HLA-A molecules rely on other amino acids for their G+C preference, here knocking out Lysine had no effect.

GC+ alleles	GC- alleles		uncorrelated alleles	
	Patr-A0101(-0.78)	Patr-A1101(-0.87)	Patr-A0201(-0.44)*	
	Patr-A0301(-0.95)	Patr-A1201(-0.89)	Patr-A0501(-0.38)*	
	Patr-A0401(-0.72)	Patr-A1301(-0.88)	Patr-A1601(0.02)*	
	Patr-A0601(-0.88)	Patr-A1401(-0.88)		
	Patr-A0701(-0.88)	Patr-A1501(-0.90)		
	Patr-A0801(-0.80)	Patr-A1701(-0.88)		
	Patr-A0901(-0.85)	Patr-A1801(-0.91)		
	Patr-A1001(-0.90)			
Patr-B1101(0.94)	Patr-B0301(-0.96)	Patr-B2301(-0.93)	Patr-B0101(-0.78)*	Patr-B1401(-0.05)*
Patr-B1301(0.95)	Patr-B0701(-0.95)	Patr-B2401(-0.83)	Patr-B0201(0.62)*	Patr-B1601(-0.40)*
Patr-B1701(0.73)	Patr-B1801(-0.92)	Patr-B2701(-0.80)	Patr-B0401(-0.50)*	Patr-B1901(0.02)*
Patr-B2801(0.95)	Patr-B2001(-0.95)		Patr-B0501(0.62)*	Patr-B2101(0.35)*
			Patr-B0601(-0.81)*	Patr-B2201(0.59)*
			Patr-B0801(-0.03)*	Patr-B2501(-0.76)*
			Patr-B0901(-0.72)*	Patr-B2601(-0.41)*
			Patr-B1001(-0.78)*	Patr-B2901(-0.23)*
			Patr-B1202(0.58)*	Patr-B3001(-0.58)*

Table 2.S4. Spearman-rank Correlation values between G+C contents and fractions of presented peptides of the bacteria set for all Chimp MHC-I alleles (Patr). HLA-alleles indicated with an asterix (*) showed inconsistent correlation values for different MHC-thresholds or MHC prediction method.

GC+ alleles	GC- alleles		uncorrelated alleles	
Mamu-A03(0.93)			Mamu-A01(0.58)*	Mamu-A1602(0.12)*
Mamu-A04(0.75)			Mamu-A02(-0.16)*	Mamu-A19(0.28)*
Mamu-A06(0.72)			Mamu-A0505(0.12)*	Mamu-A25(-0.18)*
Mamu-A21(0.92)			Mamu-A07(0.43)*	Mamu-A26(0.26)*
Mamu-A23(0.88)			Mamu-A11(-0.69)*	Mamu-A28(-0.30)*
Mamu-A24(0.82)			Mamu-A1305(0.62)*	
	Mamu-B01(-0.96)	Mamu-B48(-0.92)	Mamu-B03(0.23)*	Mamu-B36(-0.20)*
	Mamu-B12(-0.90)	Mamu-B49(-0.71)	Mamu-B05(-0.42)*	Mamu-B44(-0.39)*
	Mamu-B19(-0.90)	Mamu-B5002(-0.91)	Mamu-B08(-0.58)*	Mamu-B45(0.23)*
	Mamu-B20(-0.88)	Mamu-B61(-0.91)	Mamu-B17(0.67)*	Mamu-B47(-0.48)*
	Mamu-B37(-0.91)	Mamu-B69(-0.90)	Mamu-B27(-0.23)*	Mamu-B57(0.09)*
	Mamu-B43(-0.74)		Mamu-B28(-0.44)*	Mamu-B66(-0.68)*

Table 2.S5. Spearman-rank Correlation values between G+C contents and fractions of presented peptides of the bacteria set for all Rhesus macaque MHC-I alleles (Mamu). HLA-alleles indicated with an asterix (*) showed inconsistent correlation values for different MHC-thresholds or MHC prediction method.

Chapter 3

A comparison of antigen processing predictors and their impact on MHC-I ligand predictions

JORG J.A. CALIS*, PETER REININK*, CHRISTIN KELLER†, PETER M. KLOETZEL† AND CAN KEŞMIR* (2012)

*Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, The Netherlands †Institut für Biochemie-Charité, Medical Faculty of the Humboldt University Berlin, Berlin, Germany

In preparation

Abstract

Proteasomes are the main actors in cellular proteolysis. In proteolysis, proteins are digested into peptide fragments. Some of these fragments can be transported to the ER by TAP, where they become potential ligands for MHC-I presentation. Based on *in vitro* proteasome digestion and MHC-I ligand data, different proteolysis predictors have been developed. In this study, we compared how well proteolysis predictors capture basic proteasome activity or complete cellular protolysis. As expected, the *in vitro* proteasome digestion patterns were best captured by methods that are trained on *in vitro* proteasome cleavage data (ProteaSMM and NetChop 20S), whereas cellular proteolysis is best predicted by a method trained on MHC-I ligand data (NetChop Cterm). These results suggest that a substantial fraction of the MHC-I ligands are generated by other proteases. Further, we investigated the optimal way of combining a proteolysis, TAP transport and MHC-I binding predictor, to obtain the best MHC-I ligandome prediction model. The best MHC-I ligand predictions were obtained by an additive model making use of NetChop-3.0 Cterm and MHC-I binding predictions, that were normalized such that every MHC-I molecule has the same specificity. Thus, the TAP predictor was dispensable for optimal predictions, and proteasome specificity alone is not sufficient to predict the proteolysis that underlies the MHC-I ligandome.

3.1 Introduction

The proteasome degrades intracellular proteins, marked for degradation by the ubiquitination pathway [167]. Protein degradation, i.e. proteolysis, is important to efficiently remove miss-folded proteins, and to regulate cellular processes such as the cell-cycle or the production of MHC-I ligands [168–171]. Peptide fragments that result from proteolysis are rapidly degraded by cytosolic aminopeptidases [7]. Few fragments escape this degradation and are transported to the Endoplasmatic Reticulum (ER) by the Transporter associated with Antigen Processing (TAP), where they can bind to MHC-I molecules [11].

Most cells express the constitutive proteasome, which is a barrel shaped multi-subunit protein complex, composed of two α- and two β-rings, where each ring contains seven subunits. In the β ring of the constitutive proteasome, three proteins are present that have proteolytic capacity, β1, β2 and β5 [172]. Under influence of Interferon-γ (IFNγ), these subunits can be substituted by $\beta1_i$, $\beta2_i$ and $\beta5_i$, respectively, to form the so-called immunoproteasome [35]. Whereas the constitutive proteasome has a preference to cleave hydrophobic, acidic and basic amino acids, the immunoproteasome is more specific and prefers to cleave after hydrophobic and basic amino acids only [12, 42, 43]. Other proteasome types can be formed by a combination of constitutive and immunoproteasomal subunits [173], or with the $\beta5_t$ subunit that is only expressed in cortical thymic epithelial cells [74]. These different proteasome types partially overlap in their cleavage preferences [74, 173, 174], which has an influence on the repertoire of MHC-I presented peptides [25].

The cleavage specificity of different types of proteasomes can be determined using *in vitro* digestion experiments or MHC-I-ligand elutions. In an *in vitro* digestion experiment, a protein is incubated with proteasomes. The peptide fragments that are formed during the digestion can be detected using mass spectrometry, and cleavage sites can be inferred from these fragments [12–14]. The cleavage sites of only three proteins, i.e. β-casein, enolase and prion protein, have been determined in such *in vitro* assays [12–14]. *In vivo* proteolytic activity is measured indirectly via the analysis of digestion fragments. Even though most of the fragments are degraded by aminopeptidases within seconds [7], some are TAP-transported to the ER, bound to MHC-I molecules, and subsequently presented at the cell surface. Such MHC-I presented peptides can be eluted from a cell and also be identified by mass spectrometry. The C-terminus of an MHC-I presented peptide is expected to be generated by protease activity and to reflect an *in vivo* cleavage site in the protein from which the MHC-I ligand was derived [171]. However, as many cleavage sites will result in fragments that do not become MHC-I ligands, only a small subset of all cleavage sites can be detected. In addition, other peptidases, e.g. ACE, TPPII and Nardilysin [3–5], can influence the C-terminus of MHC-I ligands. Therefore the MHC-I-ligand data is more likely to reflect the

activity of all cellular proteases, rather than the activity of just the proteasomes or one proteasome-type.

To predict cellular proteolyis, different predictors have been developed based on different data sets [16, 41, 51, 175–179]. Most predictors, e.g. NetChop 20S [16, 51], FragPredict [175, 176], ProteaSMM [41], PAProC [178, 179] and PepCleave [177], have been trained on the *in vitro* proteasome digestion data from B-casein and enolase [12, 13]. The so-called enhanced versions of ProteaSMM and NetChop 20S are trained on the *in vitro* proteasome digestion data from B-casein, enolase and the prion-protein [12–14]. Unlike the other predictors, NetChop Cterm is trained on *in vivo* MHC-I ligand data [16, 51]. Besides the different data sets that were used for training the methods, different computational techniques were used to construct the predictors. For instance, ProteaSMM models the cleavage pattern with a Stabilized Matrix Method (SMM) using six amino acids Cterminal and four amino acids N-terminal of a potential cleavage site, and NetChop is based on a neural network that uses nine amino acids Cterminal and eight amino acids N-terminal of a potential cleavage site. Tenzer et al. [41] bench-marked FragPredict, PAProC, NetChop-2.0 and ProteaSMM on proteasome digestion and MHC-I ligand data, and showed that ProteaSMM best predicted *in vitro* proteasome digestion cleavage patterns, whereas NetChop-2.0 Cterm best predicted the cleavage patterns based on MHC-I ligands. Tenzer et al. argued that the increased performance of NetChop-2.0 Cterm on the MHC-I ligand data was due to a recognition of TAP-transportable peptides. After this study, NetChop was updated to version 3.0 [51] and a new method, PepCleave, was developed [177]. Unfortunately, PepCleave cannot be compared to the other predictors as it predicts fragments and not cleavages [177]. Therefore, we have chosen to compare ProteaSMM and the newest version of NetChop on new *in vitro* proteasome digestion data sets, and a benchmark set of MHC-I ligands to see how well the "proteasome only" predictions compare with the cellular proteolysis that underlies the processing of MHC-I ligand precursors.

Proteasome predictors have been used in combination with TAP transport and MHC-I binding predictors to perform MHC-I ligand predictions. How to combine the different predictors has been the subject of much discussion: for instance an almost full day of discussions in the EpiRep2011 meeting at the Bar-Ilan University were centered around the question “How to best predict MHC-I ligands?”. Different groups employ different strategies for MHC-I pathway predictions [26, 49, 164, 180–182], which can be distinguished along three main questions. First, are all three predictors needed? If predictions for MHC-I binding are more precise than for TAP transport or proteolysis, it could be better to only use the MHC-I binding predictor. Second, should an MHC-I ligand prediction model work like a serie of filters, or should the predictions of each step be added? If a peptide with low TAP affinity can be presented because it is a very good MHC-I binder, one should use an additive model in which the TAP transport, proteasome cleavage and MHC-I binding can compensate for each other. If not, one better

uses a filter-model in which an MHC-I ligand is required to have good scores for TAP transport, proteolysis and MHC-I binding. Third, should MHC-I binding scores be scaled per HLA molecule? The predicted specificities differ for different HLA molecules. For instance, a much larger fraction of peptides are predicted to bind with an affinity of 500 nM or stronger to HLA-A*0201 than to other HLA molecules (not shown). We do not know whether these differences reflect true differences in the binding specificities of MHC-I molecules. If such differences do not reflect true differences in specificity, one should scale the predicted binding scores per HLA molecule. We evaluated these three questions and show that the best MHC-I pathway predictions are obtained with the additive model that combines proteolysis probability with scaled MHC-I binding prediction scores.

3.2 Results

3.2.1 Predicting *in vitro* cleavage patterns

To test which proteasome predictors can predict best the proteolytic activity of proteasomes *in vitro,* we used a recently generated independent data set. This data set was based on *in vitro* digestions of 17-30 amino acids long HIV-1 peptides, from which the digestion products were analyzed using mass spectrometry to determine cleavage and non-cleavage sites (see Methods). Digestions were performed with either constitutive or immunoproteasomes. Of 240 possible cleavage sites that could be used to test the predictors (see Methods), 99 (41%) were used by the constitutive proteasomes, and 99 by the immunoproteasomes, 68 (28%) of these sites were shared between the two proteasome-types.

The prediction performance of ProteaSMM and NetChop-3.0 was analyzed using Receiver Operator Characteristic (ROC)-curves, where the number of correct and false predictions is plotted for every possible prediction threshold [183]. The area under a ROC-curve (AUC) is a performance measure of the predictor, and it is widely used because it is threshold independent [183]. For each predictor (and different versions of the predictors) the AUCs were determined on both constitutive and immunoproteasomal cleavage patterns obtained from the *in vitro* digestions (Table 3.1). In general, the methods perform better in predicting the immunoproteasomal cleavage pattern. This could be explained by immunoproteasomes being more specific; they prefer to cleave after hydrophobic and basic amino acids whereas constitutive proteasomes also cleave after acidic amino acids [12, 42, 43]. Possibly, a more specific cleavage pattern is easier to predict. The immunoproteasomal cleavage pattern was best predicted by proteaSMM-immuno and proteaSMM-constitutive (ROC-comparison test: $p<0.001$; Table 3.1), and the constitutive cleavage pattern was best captured by proteaSMM-constitutive and NetChop-3.0 20S (ROC-comparison test: $p<0.001$; Table 3.1). Surprisingly,

Predictor name	Constitutive cleavage prediction (AUC)	Immunoproteasomal cleavage prediction (AUC)
NetChop-3.0 Cterm	0.671	0.731
NetChop-3.0 20s	**0.713**	0.748
ProteaSMM Immuno	0.685	**0.803**
ProteaSMM Immuno E.	0.650	0.767
ProteaSMM Constitutive	**0.702**	**0.792**
ProteaSMM Constitutive E.	0.656	0.776

Table 3.1. Predictor performances on *in vitro* proteasome cleavage pattern predictions. The prediction performance was determined for constitutive and immunoproteasomal cleavage patterns (second and third column, respectively), as AUC of ROC-curves. Best performing predictors that are not significantly different in their performance (ROC-comparison test: $p>0.001$), but that are significantly better than the other predictors (ROC-comparison test: $p<0.001$) are indicated in boldface. The Enhanced versions of ProteaSMM are indicated with E.

the enhanced ProteaSMM versions did not perform better, even though they are trained on extra data from proteasomally digested prion protein [14]. Possibly, the digestion of this protein is very different from the other digestions that were used to train and test the predictors. NetChop-3.0 20S is also trained on prion data, but the effect of this data on NetChop-3.0 20S performance cannot be estimated as no version of this method is available that is not trained on prion data. In summary, the methods that have been trained on *in vitro* proteasome digestion data (proteaSMMs and NetChop-3.0 20S) outperformed the method that has been trained on *in vivo* MHC-I ligand data (NetChop-3.0 Cterm), which agrees with previous observations [41, 184] and the expectation that methods trained on *in vitro* data can best predict the basic proteasome cleavage patterns.

3.2.2 Predicting *in vivo* cleavage patterns

Cellular proteolysis can be very different from proteasomal activity, as other peptidases (e.g. ACE, TPPII or Nardilysin [3–6]) contribute to the *in vivo* proteolysis. Therefore, we compared which proteasome predictors best predict cellular proteolysis using MHC-I ligand data. To this end, we inferred cleavage sites from non-redundant MHC-I ligands, that have been identified after 2004, when NetChop Cterm was last updated ($n=832$; see Figure 3.1 and Methods). To determine non-cleavage sites is much more challenging. We have generated a non-cleavage data set using two alternative methods. First, for every cleavage site, 100 non-cleavage sites were made by shuffling an area of 19 amino acids around the cleavage site (the area used by NetChop for predictions plus one N-terminal and one C-terminal extension, see Figure 3.2). Second, all other sites in the source pro-

teins of the MHC-I ligands that we used, were taken as non-cleavage sites (Figure 3.1 and Methods). The predictors were assessed for their capacity to discriminate cleavage sites from non-cleavage sites, by comparing AUC values. Not surprisingly, NetChop-3.0 Cterm most accurately captured the *in vivo* cleavage pattern using both non-cleavage sets (ROC-comparison test: $p<0.001$; Figures 3.5), as this method has been trained on MHC-I ligand data. Nevertheless, this indicates that cellular proteolysis differs from proteasomal activity *in vitro*.

As we infer cellular proteolysis from MHC-I ligand data, the superior performance of NetChop might be due to a biased recognition of peptides with a high TAP affinity [41]. To exclude this effect, the performance of the different proteasome predictors was tested in combination with the TAP transport predictor [52]. In an AUC-analysis one can test the predictive performance of a single set of scores, however we now want to test the performance of a combination of two scores, i.e. proteasome cleavage and TAP transport scores. Therefore, we developed a new method, in which first for every TAP binding threshold, the performance of the cleavage predictor was measured as the AUC. Next, an integration over all AUCs was combined in a score called Volume Under the Plane (VUP; see Methods). Based on VUP scores, NetChop-3.0 Cterm still outperformed the other proteasome predictors (in both non-cleavage definitions $p<0.001$; Figure 3.5), indicating that its higher performance is not due to having a biased recognition of TAP ligands. Tenzer et al. chose a simpler approach to measure the prediction performance of combined TAP-transport and proteasome cleavage scores, by summing both scores into a single score [41]. We repeated this analysis with both non-cleavage site definitions. In every case, NetChop-3.0 Cterm outperformed the other predictors, even when the TAP transport and proteasome cleavage scores were differently weighted prior to summation (Figures 3.3 and S3.S1). Taken together, we show that NetChop Cterm best predicts cellular proteolysis. This result suggests that not only the proteasome but also other cellular proteases contribute substantially to MHC-I ligand production.

We wanted to understand how NetChop-3.0 Cterm predicts cellular proteolysis better than the other predictors. Therefore, we decided to examine for each predictor, which of the cleavage sites were given a low prediction score. For every predictor, cleavage sites with a bottom-5% prediction score were selected for this analysis. A striking difference between NetChop-3.0 Cterm and the other predictors was observed at position P1' of these cleavage sites (i.e. the C-terminus of the MHC-I ligand; Figure 3.2). Whereas the amino acids at position P1' were equally distributed for NetChop-3.0 Cterm, a Lysine was found in at least 50% of the cleavage sites with a low prediction score for the other predictors (Figure 3.4). In other words, the predictors based on *in vitro* proteasomal cleavage data fail to capture the *in vivo* cleavage after Lysine residues. This fits with the described proteolytic preferences of TPPII and Nardilysin [3, 4], and the suggested role of these proteases in the generation of MHC-I ligands for HLA-A*03 and HLA-A*11 [23, 172, 186]. In addition, other proteases such as ACE have been shown to

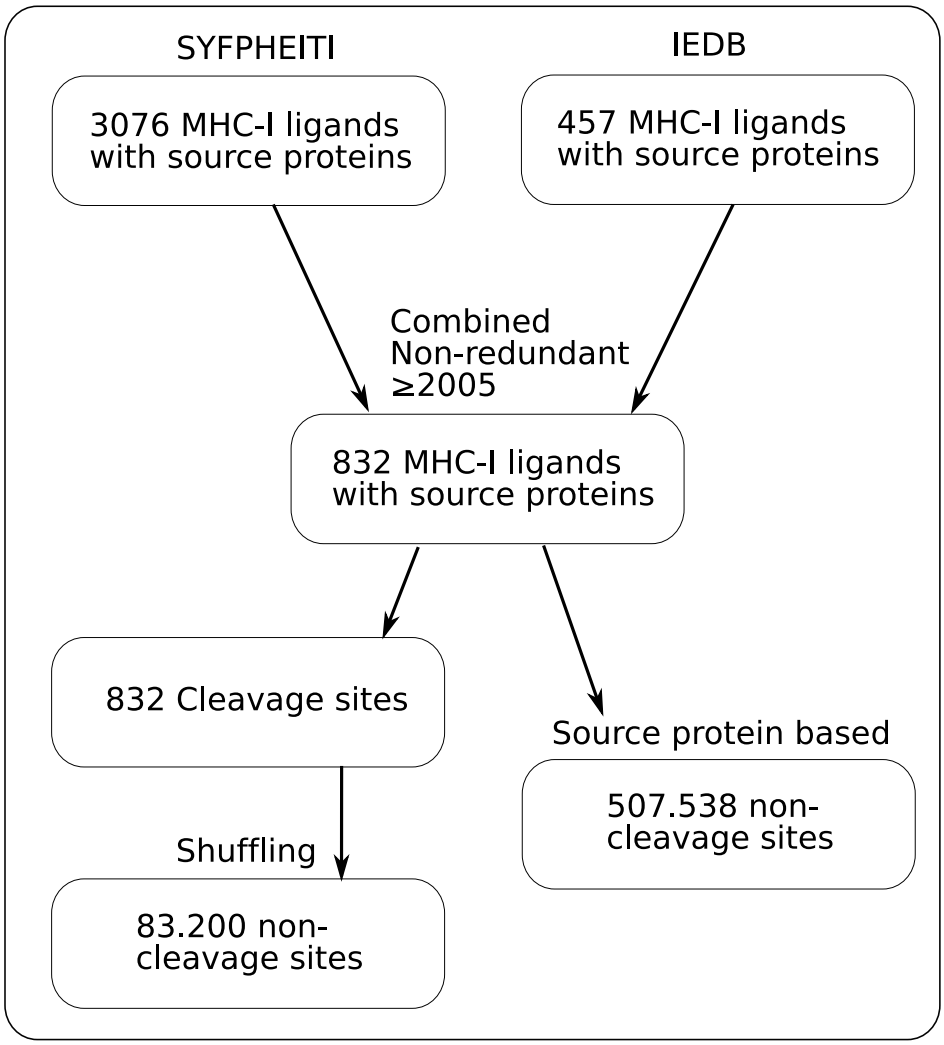


Figure 3.1. Constructing the MHC-I ligand data set. MHC-I ligands and source proteins, that were obtained in elution studies, were derived from the SYFPHEITI database [45] and the IEDB database [185]. The data sets were combined and non-redundant ligands that were not published before 2005 were selected. Every MHC-I ligand in its source protein represents a cleavage site, non-cleavage sites were derived by shuffling an area of 19 amino acids around the cleavage site (Figure 3.2), or by defining all other sites in the source protein as a non-cleavage site.

influence the generation of MHC-I ligands [5, 6], their proteolytic activity should as well be captured by NetChop-3.0 Cterm. Taken together, these results sug-

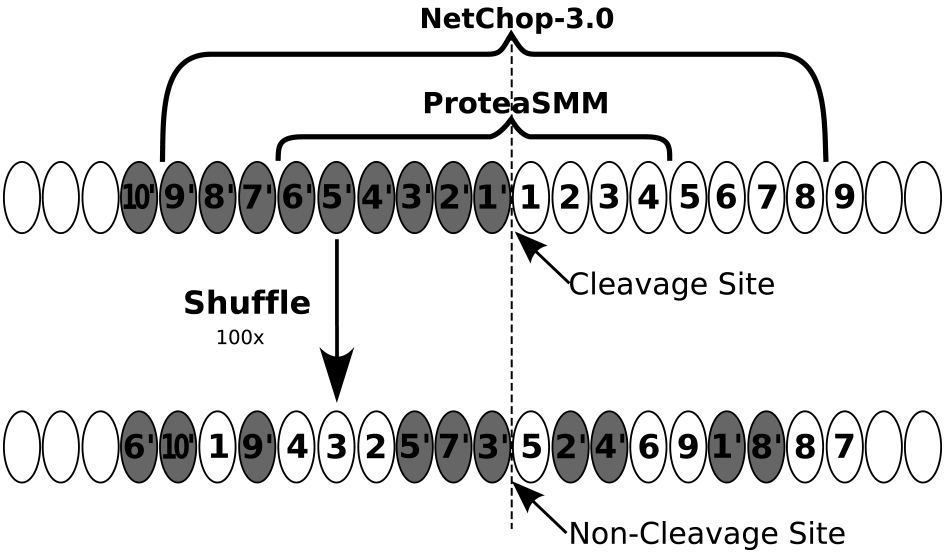


Figure 3.2. Constructing the non-cleavage sites data set. The C-terminus of an MHC-I ligand (in orange) is defined as a cleavage site. An area of 19 amino acids (from P10' to P9) around a cleavage site was shuffled and the middle position was assigned as a non-cleavage site. For every cleavage site, 100 non-cleavage sites were constructed. The positions that are used by NetChop-3.0 (P9' to P8) and ProteaSMM (P6' to P4) for predicting cleavage probabilities are indicated above the figure.

gest that NetChop-3.0 Cterm incorporates the activity of different proteases, and therefore is the best predictor of cellular proteolysis.

3.2.3 MHC-I ligandome predictions

Even though cellular proteolysis, TAP transport and MHC-I binding are predictable, it remains an open question how to combine these predictions into a model for predicting MHC-I ligands. To search for the best MHC-I pathway prediction method, we made use of the 9mer ligands in our data set, as MHC-I binding predictions are most reliable for peptides of this length. Within the source proteins of these ligands, all other 9mer peptides were taken as non-ligands. For all ligands and non-ligands, proteolysis, TAP transport and MHC-I binding prediction scores were determined, to be used as inputs for MHC-I ligandome prediction models. Prediction performances were assessed using the Matthews correlation coefficient (MCC) [187], as ligand and non-ligand data sets are of very different sizes (n=418 and n=262.806, respectively). To compare the performance of different prediction models, a random half of the data was used to determine op-

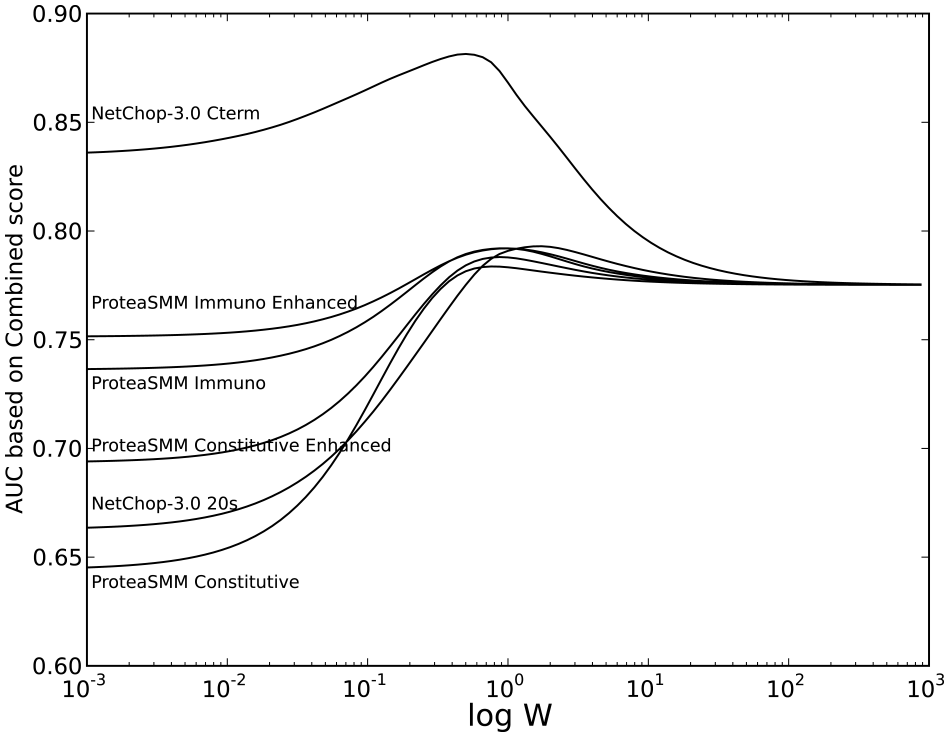


Figure 3.3. Predicting cellular proteolysis in combination with the TAP transport predictor. For the different proteasome cleavage predictors, the proteasome cleavage prediction score was added to the TAP transport prediction score (as proposed by Tenzer et al. [41]). Prediction performance was measured as the AUC of an ROC-curve (Y-axis), using the shuffled sequences as non-cleavage sites (see Figure 3.2 and Methods). When combining the scores, the weight of the TAP transport score was changed by the factor W (on the X-axis). The combined score (C), based on the TAP transport (T) and proteolysis (P) score would than be $C = W * T + P$. As a result, the proteasome cleavage or the TAP transport predictor had a larger influence in the combined score, if W is smaller or larger, respectively.

timal parameters (see Methods) and the other half was used to assess the actual performance. The procedure of defining optimal parameters and performance assessment was repeated 100 times.

First, we investigated whether a prediction model should include predictors for every step in the MHC-I presentation pathway (i.e. proteolysis, TAP transport and/or MHC-I binding). In every model, excluding MHC-I binding or proteolysis predictions led to a worse performance, whereas TAP transport predictions were most often dispensable (Table 3.2). Second, we investigate whether a peptide with a sub-optimal score in one process (i.e. proteolysis, TAP transport of

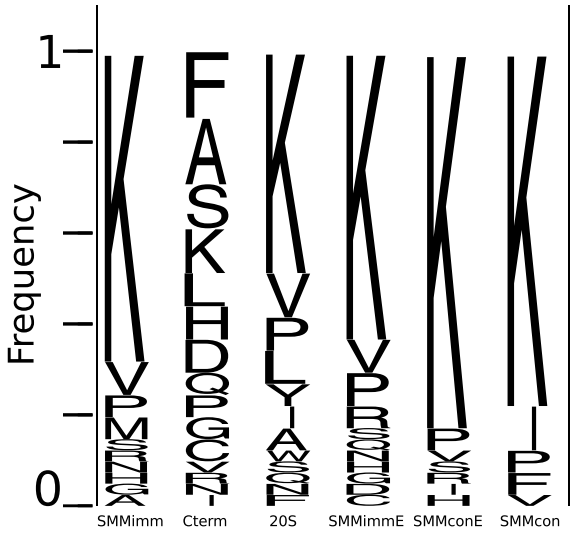
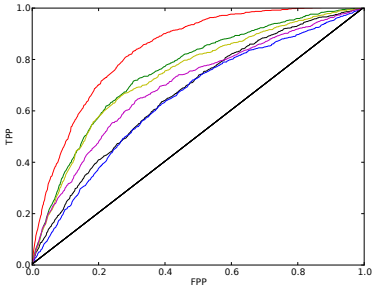
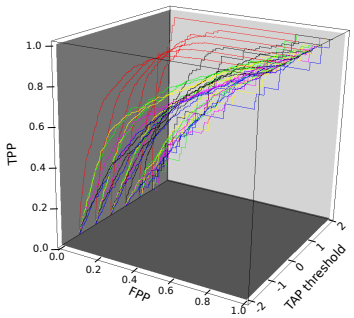


Figure 3.4. Proteolytic activity after Lysine residues is only predicted by NetChop-3.0 Cterm. For every proteasome cleavage predictor, 5% of the true cleavage sites with the lowest prediction scores were determined. The amino acid profile at P1' (i.e. the C-terminus of the presented MHC-I ligand) of these cleavage sites with a low prediction score was analyzed. The height of the letters represents their frequency at position P1' of the peptides with low proteolysis prediction scores, these are shown for every proteolysis predictor.

MHC-I binding) can be presented if it has a very good score in other steps. If such compensations are possible, the prediction scores for cleavage, TAP transport and MHC-I binding should be additive; we refer to this model as the additive model. Alternatively, one might claim that an MHC-I ligand should have a minimum level of proteolysis, TAP transportability and MHC-I binding; we refer to this as the filter-model. The additive model always outperformed the filter model when proteolysis and MHC binding predictions were included (Rank-sums test: $p<0.001$; Table 3.2)). Third, we investigate whether MHC-I binding predictions should be normalized by applying a procedure known as rescaling. When models using rescaled MHC-I binding prediction scores were compared with the same model using unscaled binding scores, the better description of the MHC-I ligandome was always obtained with the model with rescaled scores (Rank-sums test: $p<0.001$). Summarizing, a most optimal prediction of the MHC-I ligandome was made with a model that combined proteolysis predictions and rescaled MHC-I binding prediction scores in an additive way.





Predictor	AUC[a]	AUC[b]	VUP[a]	VUP[b]
Netchop-3.0 Cterm	0.835	0.844	0.895	0.746
Netchop-3.0 20s	0.663	0.698	0.711	0.638
ProteaSMM Immuno	0.736	0.761	0.680	0.689
ProteaSMM Immuno enhanced	0.751	0.773	0.676	0.683
ProteaSMM Constitutive	0.644	0.671	0.629	0.640
ProteaSMM Constitutive enhanced	0.693	0.710	0.646	0.658

Figure 3.5. Predicting cellular proteolysis. Proteasome cleavage predictors were tested as a stand-alone predictor, or in combination with a TAP predictor, and performance was assessed using AUC and VUP, respectively (see Methods). The performance was tested using either non-cleavage data sets that were derived by shuffling cleavage sites (a) or by taking other sites from source protein as non-cleavage sites (b) (see Figure 3.1 and Methods). Examples of the AUC and the VUP analyses are shown in the upper part, AUC and VUP scores are given in the lower part. In all analyses, NetChop-3.0 Cterm reached the highest score (ROC-comparison test: $p<0.001$). NetChop-3.0 Cterm is shown in red, NetChop-3.0 20s in black, ProteaSMM Immuno in yellow, ProteaSMM Immuno Enhanced in green, ProteaSMM Constitutive in blue, and ProteaSMM Constitutive Enhanced in magenta.

3.3 Discussion

In this study we analyzed how well different methods can predict the cleavage patterns of basic proteasome activity *in vitro* and cellular proteolysis. *In vitro* cleavage patterns of constitutive and immunoproteasomes were shown to be best captured by method trained on proteasome digestion data, i.e. ProteaSMM and NetChop-3.0 20S (Table 3.1). In contrast, cellular proteolysis was best predicted by the method that is trained on MHC-I ligand data, NetChop-3.0 Cterm (Figure

Prediction Scores			Filter model		Additive model	
PRT	TAP	MHC	scaled MHC scores	unscaled	scaled MHC scores	unscaled
x	x	x	0.32	0.28	**0.34**	0.32
	x	x	0.31	0.27	0.31	0.30
x		x	0.32	0.28	**0.33**	0.29
x	x		0.06	NA	0.07	NA
		x	0.31	0.27	0.31	0.27
	x		0.04	NA	0.04	NA
x			0.06	NA	0.06	NA

Table 3.2. Comparing different MHC-I pathway prediction models. Prediction models were constructed using proteolysis scores from NetChop-3.0 Cterm and/or TAP transport and/or MHC-I binding prediction scores (indicated in first three columns as PRT, TAP and MHC, respectively). MHC-I binding scores were either scaled per MHC-I molecule (column 4 and 6) or left unscaled (column 5 and 7), and prediction scores were combined in a single score (additive model, column 4-5), or used as a threshold (filter model, column 6-7). For every prediction model, the average Matthew's correlation coefficient (MCC) of 100 performance tests (see Methods) is shown. Best predictions were done using an additive model that combined proteolysis prediction scores and scaled MHC-I binding prediction scores (indicated in boldface). NA: no MHC-I binding predictions, scaling/unscaling not applicable here.

3.5). Further, we showed that the better result was not due to an embedded recognition of TAP transportable peptides (Figures 3.5 and 3.3). There are two main explanations for our findings: First, the proteolytic activity of proteasomes *in vitro* might be different than their activity in a cell. This difference might result from the mix of proteasome subtypes that are present in a cell [173], or from interactions with other molecules such as PA28 or the 19S cap regulatory particle [13, 188]. Second, other proteases such as TPPII, ACE or Nardilysin might contribute to the proteolytic activity in a cell [3–6]. The best example of cellular proteolytic activity that is not observed *in vitro*, is the cleavage after Lysine-residues. This activity is required to generate ligands for HLA-A*03 and HLA-A*11, that bind peptides with a Lysine at the C-terminus [23, 172, 186]. We show that only NetChop-3.0 Cterm captures this hallmark of cellular proteolysis (Figure 3.4). In addition, other more subtle non-proteasomal cleavage patterns might be captured by NetChop-3.0 Cterm.

From our analysis on models to predict MHC-I ligands, we conclude that the addition of a TAP transport predictor does not increase the overall performance. One might suggest that this is due to embedded TAP affinity preferences in the NetChop-3.0 Cterm proteolysis predictor. However, NetChop does not consider amino acids N-terminal from the potential 9mer MHC-I ligand, that are known to affect TAP transport [36, 52], and was trained on MHC-I ligands of differ-

ent lengths. Alternatively, TAP transport is probably the least specific step in the MHC-I presentation pathway, and therefore dispensable in an MHC-I ligandome prediction. Next, we conclude that an additive prediction model outperforms a filter model, which suggests that inefficiencies in one step of the MHC-I presentation pathway can be compensated by others. Finally, we show that scaled MHC-I binding predictions led to a better description of the MHC-I ligandome, which indicates that predicted specificities of different HLA molecules are not reflecting true biological variation.

MacNamara et al. [164] also adressed the question if scaling of MHC-I binding predictions increases the prediction of MHC-I ligands, but concluded that unscaled predictions lead to better predictions. We think that the different definitions of non-ligands underlie these opposite results. Whereas we define non-ligands as all other 9mers in the source protein of a ligand, restricted *only* by the MHC-I molecule that presents this ligand, MacNamara et al. consider all other 9mers as non-ligands for *all* MHC-I molecules that are part of their study. If we take over the definition from MacNamara et al. we indeed reproduce their result, i.e. that unscaled rather than scaled binding predictions lead to improved pathway predictions (results not shown). We have two reasons to believe our definition provides a more realistic, though still not fully correct, set of non-ligands to evaluate the performance of MHC-I ligand prediction models. First, the ligand/non-ligand ratios vary enormously for different MHC-I molecules in the definition of MacNamara et al., from 1 in 2100 to 1 in 44000. Second, more peptides are falsely assigned as non-ligands (i.e. more False Negatives) as not all MHC-I ligand source proteins are tested on every MHC-I molecule. In addition, our result that rescaled MHC-I binding predictions improve MHC-I ligand predictions, agrees with work from Stranzl et al., who analyzed the performance of an MHC-I ligand predictors on a different benchmark set [180].

The development of proteasome predictors serves two goals. First, to understand the preferences and biochemical processes that underly cellular proteolysis. Second, to predict and understand how this process influences the MHC-I ligandome. With respect to the first goal, we found profound differences between proteasomal proteolysis *in vitro* and cellular proteolysis, this suggests a strong activity of non-proteasomal proteases. Additionally, we show how proteolysis predictions should be combined with TAP transport and MHC-I binding predictions for an optimal prediction of the MHC-I ligandome. The right combination of predictors was shown to significantly improve MHC-I ligand predictions. These findings can serve as a guideline for future MHC-I ligand prediction studies, meanwhile they also demonstrate the need for improved predictions of proteolysis and other processes affecting MHC-I presentation.

3.4 Methods

3.4.1 Data collection

Proteasomal *in vitro* cleavage patterns were derived from a digestion of HIV-1 peptides with constitutive or immuno-proteasomes, as explained in [189]. 16 Peptides from the HIV-1 proteins GAG and TAT, with a length of 17 to 30 amino acids were degraded. After 0, 1, 2, 4, 8 and 24 hours of degradation, peptide fragments were analyzed using mass spectrometry (as in [189]). To avoid analyzing secondary cleavage products, peptide fragments found after 4 hours of degradation were used to infer cleavage sites. Of 368 possible cleavage sites 150 were cleaved by the immunoproteasome and 148 were cleaved by the constitutive proteasome, 103 cleavage sites were found in both sets (Suplementary Table S3.S2). The ProteaSMM proteasome cleavage predictors require six amino acids N-terminal and four amino acids C-terminal of a possible cleavage site. Therefore, cleavage predictions cannot be made at the beginning and end of a peptide sequence. As a result of this limitation, only 240 (of the 368) sites could be used to compare the different proteasome predictions. Of these 240 sites, 99 were cleaved by the immunoproteasome and 99 were cleaved by the constitutive proteasome, 68 cleavage sites were found in both sets.
In vivo cleavage sites were derived from MHC-I ligand data. Ligands that were identified in MHC-I elution studies were downloaded from the SYFPHEITI database [45] and the IEDB database [185]. Source proteins of the MHC-I ligands were downloaded from the NCBI via links that were provided by the SYFPHEITI and IEDB databases. The C-terminal residue of an MHC-I ligand was regarded as position P1' of a cleavage site (Figure 3.2). In total 3076 MHC-I ligands with their source protein were derived from the SYFPHEITI database and 457 MHC-I ligands with their source protein were derived from the IEDB database. Identical peptides, or peptides that were either a C- or N-terminal extension of each other, were regarded as redundant. In addition, the ligands and their corresponding source proteins that were published before 2005, or which were redundant/identical to an MHC-I ligand published before 2005 were excluded, because they could have been used for training of NetChop-3.0 Cterm. This filtering resulted in 832 MHC-ligands and their source proteins, of which every MHC-I ligand corresponds to a peptide fragment that is apparently generated by *in vivo* proteolytic activity (Figure 3.1).
Detecting *in vivo* non-cleavage sites based on the absence of a certain peptide fragment, is not possible, as many other reasons might underlie the absence of an MHC-I ligand, e.g. further degradation of the fragment or low affinity to MHC-I molecules. Therefore, non-cleavage sites were generated in two ways: First, by we shuffled an area of 19 amino acids around the cleavage site (the longest flanking region used by a proteasome predictor method plus one extra amino acid on each side, as indicated in Figure 3.2). After shuffling, the middle position, pre-

viously corresponding to the cleavage site, was now assigned as a non-cleavage site. For every cleavage site, 100 non-cleavage sites were generated, i.e. in total 83.200 non-cleavages sites were created this way (Figure 3.1). The advantage of this method is that the amino acid frequencies in the sequences that contain cleavage and non-cleavage sites remain the same. Second, all sites in the source proteins of an MHC-I ligand that were not assigned as a cleavage site were taken as non-cleavage sites (N=507.538, Figure 3.1). Results are presented with the first definition but the analysis was made for both, unless mentioned otherwise.

3.4.2 MHC-I pathway predictions

Proteasome cleavage, TAP transport and MHC-I binding predictions were performed as indicated by the developers of these prediction methods [41, 48, 51, 52, 147]. Because the TAP and MHC-I binding scores range on a logarithmic scale, they were log-transformed, in order to be able to combine them with proteolysis scores that range between 0 and 1. In addition, the log-transformed scores MHC scores were multiplied with -1 to indicate that higher scores relate to stronger binding (i.e. low IC50). Throughout this paper the negative log-transformed scores were used for all analyses, but untransformed scores are shown when presented in figures.
A set of random peptides can have on average stronger predicted binding affinity for one HLA molecule than for another HLA molecules, due to biases in the training data. A normalization procedure was performed (often referred to as scaling) to minimize such biases. To this end, 100,000 random 9mer peptides were generated with equal amino acid frequencies, and their binding affinities on the different HLA molecules were predicted with NetMHC-3.2 [48, 51, 147]. To derive the scaled binding affinity of a new peptide on a HLA molecule, we determined how it ranked among the random peptides on the same HLA molecule. Finally, we assigned the scaled affinity of this peptide as the binding affinity of a random peptide on HLA-A*0201 with the same rank. For instance, on HLA-A*0201 the random peptide with rank 1000 is LMPRYLTGL with a predicted binding affinity of 37 nM. The MSYWKARAM peptide on HLA-B*0801 had a predicted binding affinity of 428 nM, and also ranks at position 1000 among the random peptide affinities on HLA-B*0801. Therefore, the scaled binding affinity of this latter peptide becomes 37 nM. In this way, all HLA molecules have the same affinity distribution on the random peptides and we do not need to assume anything on their specificity. In addition, we repeated all analyses with a previously used rescaling method where binding affinities are normalized by the affinity at which 1% of the random peptides bind (as described in [164]), with which we could reproduce all results reported in this paper (not shown).

3.4.3 Performance measures

Different proteasome predictors were assessed for their performance in discriminating cleavage from non-cleavage sites. First, the performance of the proteasome predictors was tested using Receiver Operator Characteristic (ROC) curves [183]. In a ROC curve, true positive proportions (TPP) and false positive predictions (FPP) are plotted on the y- and x-axis, respectively, for every possible threshold. The area under the ROC curve (AUC) is a measure of the predictors performance. If a predictor performes well, the TPPs increase faster than the FPP, and the AUC becomes larger than 0.5; the maximal AUC is 1.0.
The AUC can only be determined on a single set of prediction scores. However, we aimed to compare the prediction performance of the proteasome predictors in combination with the TAP transport predictor. Therefore, we developed an alternative performance measure, where for every TAP transport prediction value, we determined the AUC based on the cleavage and non-cleavage sites that exceeded the TAP transport value (T^{TAP}). If less than 25 cleavage sites or non-cleavage sites exceeded the TAP threshold, it was discarded. A score was derived by integrating over all the AUCs with respect to the TAP threshold values and subsequent normalisation by the range of TAP thresholds (equation 3.1). The resulting score ranges between 0 and 1, a random predictor would score 0.5 and a perfect predictor would score 1, similar to the scores that are obtained in an AUC analysis. This score reflects the predictive performance of the proteolysis predictor for different data sets which have been selected over a range of possible TAP values, we call this performance measure Volume under the Plane (VUP).

$$VUP = \frac{\sum_{i=1}^{n}(T_{i-1}^{TAP} - T_i^{TAP}) \times AUC_{T_{i-1}^{TAP}} + \frac{(T_{i-1}^{TAP} - T_i^{TAP}) \times (AUC_{T_{i-1}^{TAP}} - AUC_{T_i^{TAP}})}{2}}{Max(T^{TAP}) - Min(T^{TAP})} \tag{3.1}$$

3.4.4 MHC-I ligand prediction models

Two different MHC-I ligand prediction models were assessed for their capacity to discriminate MHC-I ligands from non-ligands, the additive and the filter model. In both cases, a random half of the data was used to determine optimal parameters, and the other half was used to assess the actual performance. The performance was assessed using Matthews correlation coefficient (MCC) [187]. In the filter model, a peptide is required to have a certain proteolysis, TAP transport and MHC-I binding prediction score to be defined as a ligand. To find the optimal parameter settings in this model, all combinations of possible threshold values were investigated. In the additive model, the prediction scores for proteolysis, TAP transport and MHC-I binding are summed according to some weight, and the summed score of a peptide is required to exceed a certain threshold for that peptide to be defined as an MHC-I ligand. Thus, to determine optimal parameter

settings in the additive model, weights for the different prediction steps and a threshold to define ligands from non-ligands based on the summed score had to be found. An optimal weight for each prediction step was determined using a binomial linear model in R [190], and summed scores were determined for ligands and non-ligands. Every summed score was assessed for its capacity to discriminate ligands from non-ligands, and the score at which this was most efficient was defined as the threshold. For both models, the derivation of optimal prediction settings on one half of the data and performance assessment based on the other half was repeated 100 times.

3.4.5 Statistics

Statistical tests were performed using the stats-package from the scipy-module in Python. The difference between AUC/VUP performance measures was determined by deriving AUCs/VUPs on 50 new datasets, that were generated by bootstrapping the original data set. The derived AUCs/VUPs were compared using a paired two-tailed t-test, p-values less than 0.001 were considered significant (as in [41]). We refer to this test as the ROC-comparison test.

3.5 Acknowledgments

We thank Berend Snel, Ilka Hoof, Hanneke van Deutekom and Xiangyu Rao for discussion on this research project and technical support.

3.6 Supporting Information

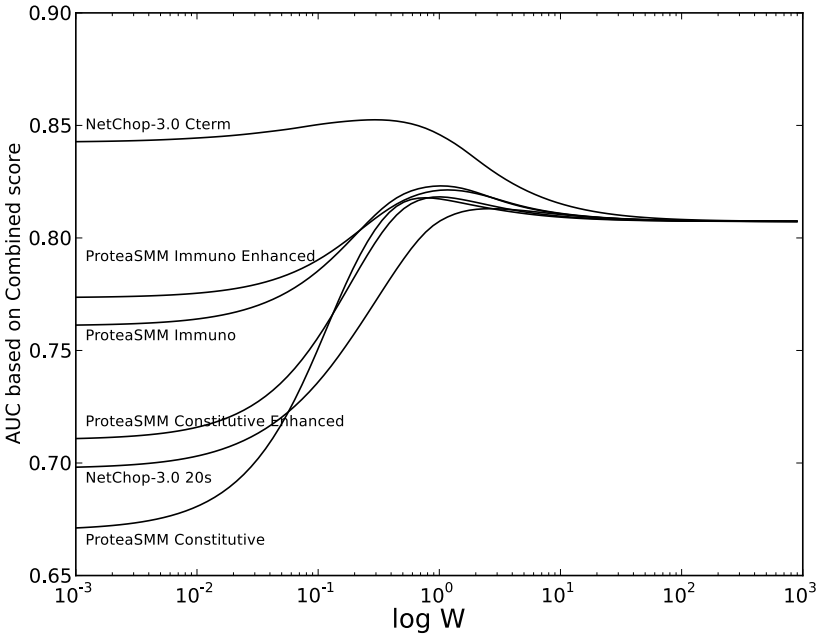


Figure 3.S1. Predicting cellular proteolysis in combination with the TAP transport predictor, with non-cleavage sites defined as all other sites in a source protein. For the different proteasome cleavage predictors, the proteasome cleavage prediction score was added to the TAP transport prediction score (as proposed by Tenzer et al. [41]). Prediction performance was measured as the AUC of an ROC-curve (Y-axis), using all other sites in a source proteins as non-cleavage sites (see Methods). When combining the scores, the weight of the TAP transport score was changed by the factor W (on the X-axis). The combined score (C), based on the TAP transport (T) and proteolysis (P) score would than be $C = W * T + P$. As a result, the proteasome cleavage or the TAP transport predictor had a larger influence in the combined score, if W is smaller or larger, respectively.

Proteasome type	name	cleavage sites
I	tat2	CFHCQVC*FI*TK*GLGISY*G*RKK*RR
C	tat2	CFHCQVC*FITK*GLGISY*GRKK*RR
I	p_seq3	RLIY*A*T*R*QL*Q*R*F*A*V*N*PGL*LI*T
C	p_seq3	RLIYA*T*R*QL*Q*R*F*A*V*N*PGL*LI*T
I	tat1	MEPVD*PRLEPWKH PG*SQPKTA*C TN C*Y*C*K
C	tat1	MEPVD*PRLEPWKH*PG*SQPKTA*C*TN*C*Y*C*K
I	nat12mod	RWL*L*L*GL*NPLV*G*GGR*L*Y*SPTSI*L*G
C	nat12mod	RWLLLGLNPLV*GGGRLYSPTSILG
I	nat12	KRWIILGL*NKIVRMYSPVSIL*D
C	nat12	K*R*WIIL*G*L*NK*I*V*R*M*Y*SPV*S*I*L*D
I	p_seq2	YVL*F*L*T*K*GL*SI*SY*L*GKK
C	p_seq2	Y*V*L*F*L*T*K*GL*S*I*SYL*GKK
I	p24	ALSEGATPQD*LNTML*NTV*GGHQA*AMQML
C	p24	ALSE*GATP*Q*DLNTM*L*NTVGGHQA*AMQML
I	p_seq1	FVIH*R*L*EPWL*HPG*SQHI*TA*S*TN
C	p_seq1	FVIH*R*L*E*PWL*H*PGSQ*H*I*T*A*STN
I	nat10mod	R*FII*PXF*T*A*L*SGGRR*A*L*L*Y*GATPY*AI*G
C	nat10mod	R*FII*PXFT*A*L*SGGRR*A*L*L*Y*GA*TPYA*I*G
I	p_seq4	YAIP*Q*A*L*N*T*L*L*N*TV*GGHQAA
C	p_seq4	YAIP*Q*A*L*N*T*L*L*N*TV*GGHQAA
I	p17	YKLK*HI*VW*A*S*RELER*F*AVNPGL*L*E*V*TS*E*GC
C	p17	YKLKHIVWASREL*ERFAVNPGLLEVTSEGC
I	nat13mod	R*AL*GPA*A*TL*QTPWTA*SL*GV*G
C	nat13mod	R*AL*G*PA*A*T*L*Q*T*P*WTA*SLG*VG
I	nat13	KALGPA*A*TL*EEMM*T*A*CQGVGGPGH
C	nat13	KALGPAATL*EEMM*T*A*CQGVGGPGH
I	nat11mod	RAIPIPA*GTL*L*SGGGR*AIYK*R*W*AI*L*G
C	nat11mod	R*A*I*PIP*A*GT*LL*SGGGR*A*IY*K*R*W*A*I*L*G
I	nat11	NNPP*IPVG*E*I*Y*K*R*W*II*L*G*L*N*KI*V
C	nat11	NNPPIPVG*EIY*K*R*W*IIL*G*L*NKI*V
I	nat10	PEVIPMF*S*AL*SE*GATPQ*D*L*NTML*NTVGGH
C	nat10	PEVIPMFS*ALSE*GATPQ*D*LNTML*NTVGGH

Table 3.S2. *In vitro* proteasome digestion cleavage patterns. From the HIV-1 proteins GAG and TAT, 16 peptides with a length of 17 to 30 amino acids were digested *in vitro* by immuno and constitutive proteasomes. Peptide fragments from the digest were analyzed using mass spectrometry (see [189] and Methods), cleavage sites were inferred from the fragments. For every peptide, the inferred constitutive (C) and immunoproteasomal (I) cleavages are shown. The cleavages are indicated by an asterisk (*), e.g. LINDA*C presents a cleavage between the A and the C.

Chapter 4

Properties of MHC class I presented peptides that enhance immunogenicity

Jorg J.A. Calis*, Matt Maybeno†, Jason A. Greenbaum†, Daniela Weiskopf†, Aruna D. De Silva‡†, Alessandro Sette†,Can Keşmir*,Bjoern Peters† (2012)

*Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, The Netherlands †Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, California, USA ‡ Genetech Research Institute, Colombo, Sri Lanka

Submitted

Abstract

T-cells have to recognize peptides presented on MHC molecules to be activated and elicit their effector functions. Several studies demonstrate that some peptides are more immunogenic than others and therefore more likely to be T-cell epitopes. We set out to determine which properties cause such differences in immunogenicity. To this end, we collected and analyzed a large set of data describing the immunogenicity of peptides presented on various MHC-I molecules. Two main conclusions could be drawn from this analysis: First, we showed that positions P4-6 of a presented peptide are more important for immunogenicity. Second, some amino acids, especially those with large and aromatic side chains, are associated with immunogenicity. This information was combined into a simple model that was used to demonstrate that immunogenicity is, to a certain extent, predictable. The immunogenicity prediction model (made available at *http://tools-int-01.liai.org/immunogenicity/immuno_tool.html*) was shown to be applicable to epitope discovery studies. After the past successful elucidation of different steps in the MHC-I presentation pathway, the identification of variables that influence immunogenicity will be an important next step in the investigation of T-cell epitopes and our understanding of cellular immune responses.

4.1 Introduction

Peptides presented on MHC class I (MHC-I) molecules at the cell-surface are screened by $CD8^+$ T-cells to detect aberrancies, such as an infection. The strength of the interaction between the peptide-MHC complexes (pMHC) and T-cell receptors (TCRs), depends both on the MHC-I molecule and the presented peptide. A specific pMHC will be recognized by an estimated average of one in 100,000 naive T-cells [54–57], but this precursor frequency differs for different pMHCs [56, 104, 191]. In the context of an infection, recognized pMHCs can stimulate T-cells to proliferate into an effector T-cell population that finds and kills infected cells presenting this pMHC. Such a pMHC, that is the target of a specific T-cell immune response, is called an epitope.

In past years, many efforts have been put in determining which peptides are presented on MHC-I molecules. For numerous peptide-MHC combinations the binding affinity has been measured [45, 50], and this data enabled the development of highly accurate MHC-I binding predictors [21, 45, 49, 147, 163, 165, 192, 193]. Furthermore, the generation of MHC ligands by the proteasome and other proteases cleavage and TAP transport have been studied extensively [7, 12–14, 17, 36, 194], data from these studies were used to construct successful processing-predictors [16, 41, 52]. Thanks to this progress, for a pathogen such as HIV-1 it is now possible to predict reliably which peptides will be presented on a certain MHC-I molecule, and test subsequently if these predicted pMHCs are epitopes [143].

Despite high accuracy predictions of which pMHCs are formed upon infection, what distinguishes epitopes from non-epitopes is still an open question, several factors have been described though. First, the abundance of a pMHC plays a role in immune targeting [195–197], which can be affected by peptide-MHC binding affinity [198] and stability [199], the abundance of the precursor protein [196, 197, 200] and the efficiency of MHC ligand processing [196, 197, 201, 202]. Second, an epitope should be immunogenic: In this paper, we will refer to T-cell recognized and unrecognized pMHCs as immunogenic and non-immunogenic pMHCs. A peptide-immunization experiment, in which the ability of a peptide to elicit a pMHC specific immune response is determined, provides a perfect platform to directly measure immunogenicity, as there are no other factors like the right processing of the peptide and expression of a source protein that can affect the generation of a T-cell response. Peptide-immunization experiments show that about half of the pMHCs are immunogenic [18, 108]. Some pMHCs are more likely to be an epitope because they are recognized by more T-cells [56, 104], i.e. such pMHCs have a higher immunogenicity. Third, the pMHCs derived from certain proteins that are expressed early in infection are more likely to evoke a response [127, 128]. Fourth, even if an immunogenic peptide is presented under the right conditions, a response might be blocked by regu-

latory processes if a (nonself) pMHC is too similar to a self pMHC [93, 117, 118]. We recently estimated that about one-third of the nonself pMHCs is too similar to self [203]. Finally, an immune response might be outcompeted by other T-cell responses in a competition for limited survival factors, a phenomenon called competitive exclusion [120, 121].

The identification of epitopes is key to the study and understanding of cellular immune responses, and is of great importance in vaccine development. Therefore, we studied an important step that influences whether a pMHC can be an epitope: immunogenicity (the recognition by T-cells). We generated a set of immunogenic and non-immunogenic pMHCs, and compared the amino acid frequencies in both sets. This analysis showed that T-cells have a preference for certain amino acids, especially aromatic and large residues. Moreover, the middle part of the presented peptide (P4-P6) was shown to be most important for immunogenicity. These findings were combined in a simple enrichment model, to estimate the immunogenicity of a pMHC. The performance of this predictor was tested in cross-validation and independent new data sets, and shown to predict immunogenicity (AUC=0.65). We used this predictor to examine a possible adaptation of the immune system to recognize pathogen-derived peptides, and showed that a preference for these peptides exists. Finally, our immunogenicity predictions were shown to assist the prediction of T-cell epitopes recognized in humans and mice.

4.2 Results

4.2.1 Classifying immunogenic pMHCs

To investigate the peptide preferences in T-cell recognition, one needs well defined sets of immunogenic and non-immunogenic pMHCs. Therefore, strict parameters were set to classify only those pMHCs for which immunogenicity or the absence thereof was strongly shown upon infection or vaccination. The classification of immunogenic pMHCs from positive immune responses upon infection or vaccination is relatively straight-forward. In contrast, the classification of non-immunogenic (i.e. unrecognized by T-cells) pMHCs upon natural infections is difficult, as many other factors could cause the lack of an immune response besides non-immunogenicity (see Introduction). Therefore, for the classification of non-immunogenic pMHC, we required a peptide-immunization study in combination with a high predicted peptide-MHC-I binding affinity, to ensure that MHC-I presentation of the assayed peptide to T-cells was feasible. However, this strict definition excluded humans as a host for the identification of non-immunogenic pMHCs, since peptide-immunization studies have rarely been conducted in humans. Even though immunogenic pMHCs could be derived from humans, we

decided to collect only data from mice, to avoid any bias caused by disparate sampling from different hosts. In addition, we compared only peptides presented on MHC-I molecules from the same species (H-2 or HLA, i.e. from HLA-transgenic mice), of the same length (9mers only), and a redundancy reduction method was applied to avoid oversampling effects (see Methods for a detailed description on the data collection and classification process).

Four sources of data were used, the Immune Epitope Database (IEDB) [185], and three immunogenicity studies in mice, from Assarsson et al. on Vaccinia-derived peptides in HLA-A*02-transgenic mice [18], from Kotturi et al on Arenavirus-derived peptides in HLA-A*1101-transgenic mice [108] and an unpublished data set on Coxiella Burnetti-derived peptides in (non-transgenic) C57BL/6 mice (see methods). 600 Immunogenic and 181 non-immunogenic non-redundant pMHCs with 9mers that fulfilled our strict criteria, were selected for further characterization (see Figure 4.1). Thus, a relatively large set of immunogenic and non-immunogenic pMHCs could be made to compare what properties determine the difference in immunogenicity.

4.2.2 Amino acid properties of immunogenic pMHCs

The immunogenic and non-immunogenic pMHCs, classified above, can be compared to see what properties associate with immunogenicity. We hypothesize that certain amino acids are more likely to interact with TCRs, and therefore increase the immunogenicity of a pMHC. Conversely, some amino acids could abolish TCR interactions. To test this hypothesis, per amino acid the association with immunogenicity was tested, and a comparison with background amino acid frequencies was made. To prevent any bias that might rise due to the binding motif of MHC-I molecules, residues at positions with an influence on the binding affinity were excluded from the analysis (see Methods). In addition, all peptides in our data set were required to have a predicted binding affinity stronger than 500 nM (see Methods). As most classified peptides were HLA restricted (Figure 4.1), and because of interest for the human immune system, we decided to restrict the analysis to these pMHCs. The positive association of the large, aromatic and non-polar Phenylalanine (permutation test, explained in detail in Methods: $p<0.01$) and the negative association of the small and polar Serine (permutation test: $p<0.001$) were most prominent (Figure 4.2). In addition, significant associations with immunogenicity were observed for Isoleucine, Lysine, Methionine and Tryptophan (permutation test: $p<0.05$; False discovery rate (FDR) for multiple testing determined as in [204]: $q<0.05$). Thus, in line with the hypothesis that T-cells have a preference for certain amino acids, some amino acids seem to (significantly) increase or decrease the immunogenicity of a pMHC.

We wanted to now if the observed associations might be the result of some underlying preference for certain amino acid characteristics. Therefore, the enrichment

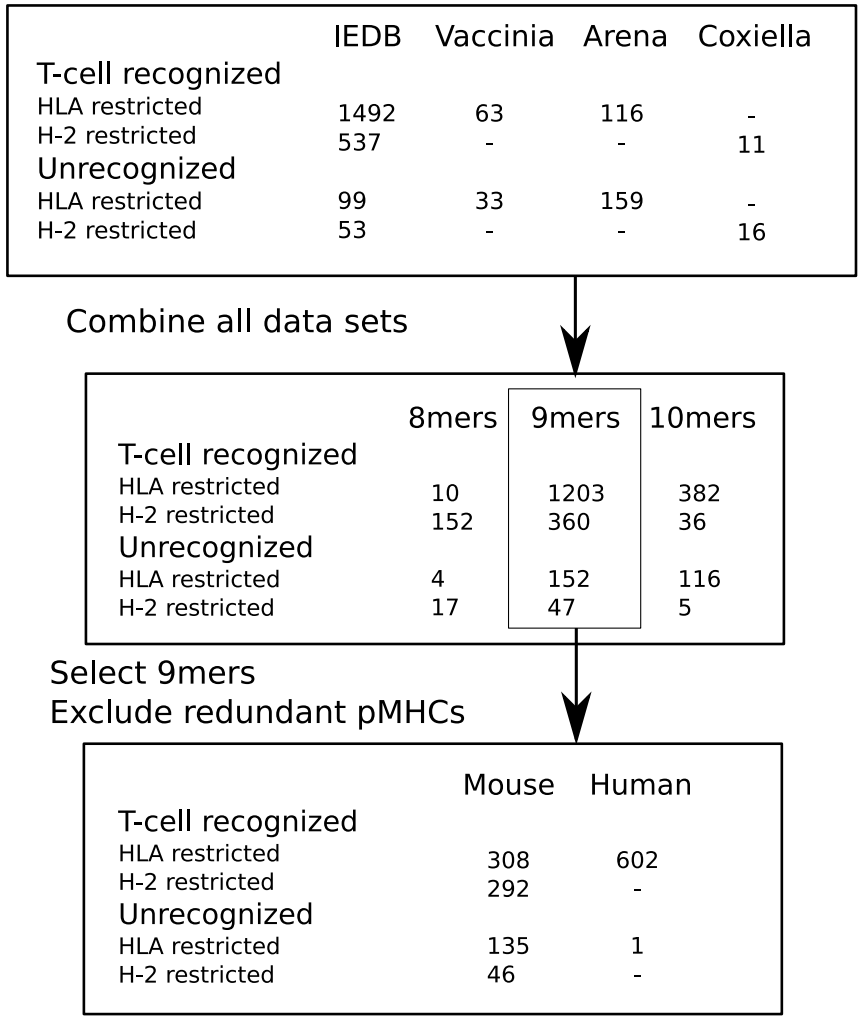


Figure 4.1. Data acquisition and handling oversight. Data was collected from four different sources (see Methods). The first panel shows how many pMHCs were derived from each data set and their respective MHC restrictions and immunogenicity status. Data from all sets was combined, and for different peptide lengths the number of pMHCs is shown in the second panel. The number of non-redundant 9mers with respect to the host in which the data was obtained is shown in the third panel.

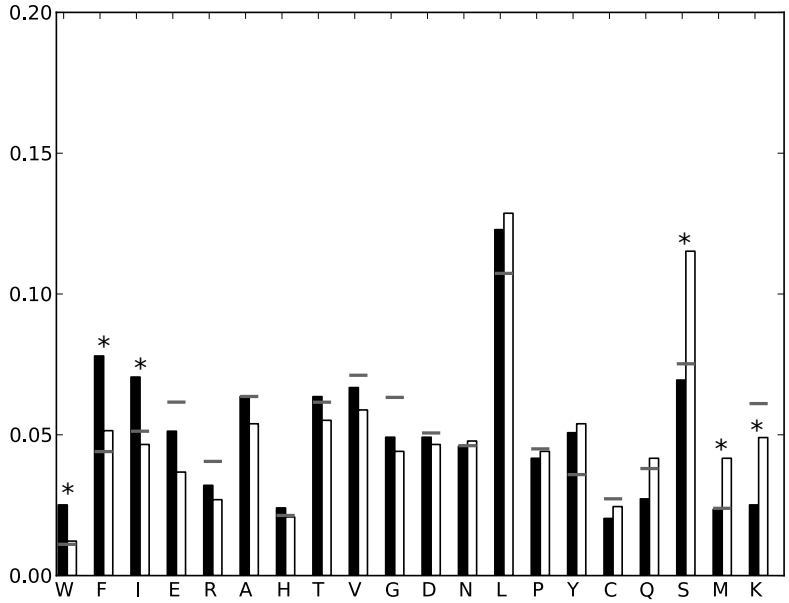


Figure 4.2. T-cell preferences for different amino acids in HLA class I presented peptides. The fraction of an amino acid in immunogenic (left bar, filled) and non-immunogenic (right bar, unfilled) peptides presented on HLA class I molecules is shown. Significantly different distributions are indicated with a star (Permutation test, see Methods: $p<0.05$; False discovery rate (FDR) for multiple testing determined as in [204]: $q<0.05$). The background frequency for each amino acid in the protein sequences that were a source of the immunogenic or non-immunogenic peptides is shown by a grey line.

of every amino acid in immunogenic vs non-immunogenic peptides was determined, and the enrichments were compared to physicochemical and biochemical properties described in the AAindex database [205]. In the AAindex database, near-identical properties are defined by their strong correlation (Spearman-rank test: absolute correlation coefficient > 0.8). None of the amino acid properties described in AAindex were near-identical to our enrichments (supplementary Table S4.S3). Thus, T-cell preferences do not seem to follow a known amino acid property, possibly a combination of properties are prefered that contribute to a better interaction with the T-cell receptors repertoire. To try to unravel this combination, an analysis of amino acids grouped according to broad characteristics such as size, polarity, charge and aromaticity was performed (see Methods). For groups of amino acids with opposite characters, e.g. small and large amino acids, the number of residues in immunogenic versus non-immunogenic peptides

were compared. This analysis showed that large, non-polar and aromatic residues were overrepresented in immunogenic peptides presented on HLA (Fisher's test: $p<0.02$; see Table 4.1), for acidic residues a trend for overrepresentation was observed ($p=0.06$). Unfortunately, it is difficult to unravel if size, polarity and/or aromaticity was most important for immunogenicity, because amino acids share combinations of such characteristics.

Amino acid feature	Total AA count		Enrichment in immunogenic peptides	Fisher's exact test (p-value)
	immunogenic	non-imm.		
large	384	132	1.28	0.014
small	653	304	0.94	
polar	600	307	0.86	0.0046
non-polar	916	358	1.12	
aromatic	326	111	1.29	0.012
non-aromatic	1522	699	0.95	
acidic	185	67	1.21	0.06
basic	147	78	0.83	
charged	332	145	1.00	1.00
non-charged	1516	665	1.00	

Table 4.1. Amino acid properties of immunogenic peptides presented on HLA class I molecules. Sets of amino acids were counted in immunogenic and non-immunogenic (non-imm.) peptides based on size, polarity, aromaticity, acidity and charge (see Methods). The enrichment of amino acid counts in the immunogenic peptides (immunogenic/non-immunogenic) was first determined in every specific group. Second, this enrichment was corrected for the overal larger number of peptides that were classified as immunogenic, this corrected enrichment is shown in the fourth column. The association of opposite sets with immunogenicity was compared using Fisher's exact test.

Our results might be biased by the large set of HLA-A*0201 presented peptides. Therefore, we excluded all HLA-A*0201 presented peptides and repeated our analysis. First, the amino acid profile was shown to be very similar, for every amino acid that was significantly associated with immunogenicity based on all pMHCs (F,I,K,M,S and W in Figure 4.2, indicated by stars), the same trend (i.e. over- or underrepresentation) was observed for the non-HLA-A*0201 presented peptides (Supplementary Figure S4.S1). In addition, more large and aromatic residues were observed in the immunogenic pMHCs of the non-HLA-A*0201 presented peptides, though not significant due to the small number of non-immunogenic pMHCs ($n=75$, respectively). Moreover, all these results were repeated in an analysis based on only HLA-A*0201 presented peptides (Supplementary Figure S4.S1). Thus, our result that T-cell preferences for certain peptides exist is not biased by a large set of HLA-A*0201 presented peptides, and was robust to either excluding or selecting these pMHCs.

4.2.3 TCR-pMHC interactions

The data set of immunogenic and non-immunogenic pMHCs enabled us to investigate another aspect of immunogenicity: the importance of different positions in the presented peptide. Structural studies, as well as immunogenicity studies of specific T-cell clones with altered peptide ligands, suggest that some positions in a presented peptide, especially positions 4-6, are in close contact with the TCR [60, 203, 206] and important for specific T-cell responses [118, 207–211]. If a certain position has a large effect on T-cell recognition, the amino acid profile at that position is expected to be different for immunogenic (i.e. T-cell recognized) compared to non-immunogenic (i.e. T-cell unrecognized) pMHCs. This difference was determined per position, using only non-anchor positions to avoid any effect of HLA binding (see Methods), thus excluding positions P1, P2 and P9 from the analysis. The difference between the amino acid profiles of immunogenic and non-immunogenic pMHCs was measured using Kullback-Leibler's measure of divergence, that calculates how well one profile can be described using the other profile, the divergence is larger if the profiles are more different from each other. In line with previous studies, the largest difference between immunogenic and non-immunogenic pMHCs was observed at positions 4, 5 and 6 (Fisher's test: $p<0.01$; Table 4.2). Thus, this analysis supports the hypothesis that these positions are most important for immunogenicity.

Position	Kullback-Leibler divergence
1	NA (anchor)
2	NA (anchor)
3	0.10
4	0.31 **
5	0.30 **
6	0.29 **
7	0.26 *
8	0.18
9	NA (anchor)

Table 4.2. Position dependent differences in immunogenic peptides. For peptides presented on HLA class I molecules in HLA transgenic mice that were either known to be immunogenic or non-immunogenic (see Methods), amino acids were counted per position. The 20 counts for immunogenic and non-immunogenic pMHCs were compared per position using the Kullback-Leibler divergence. A Fisher's test in R was done, with asymptotic chi-squared probabilities if the "Cochran conditions" (no cell has count zero, at least 80% of the cells have 5 or more counts) were satisfied [190, 212], to determine if the distributions were significantly different (* $p<0.05$; ** $p<0.01$).

4.2.4 Predicting immunogenicity

The associations of certain amino acids with immunogenicity, and the importance of different positions, can be combined into a model to predict the immunogenicity of a pMHC. Such a model based on the enrichment of non-anchor amino acids in immunogenic 9mer peptides presented on HLA, and weighted by the importance of different positions measured as KL divergence (Table 4.2) was made (see Methods). The performance of this immunogenicity model was tested in a 3-fold cross-validation experiment, where two-thirds of the data were used for building the model and one-third for testing. The immunogenicity model could distinguish immunogenic from non-immunogenic peptides on HLA class I molecules with a significant accuracy, on average 66% of the immunogenic pMHCs got a positive score, i.e. predicted to be recognized, compared to 44% of the non-immunogenic pMHCs (Ranksums test: $p<0.001$; AUC=0.65; Figure 4.3). Thus, it was possible to make a model to predict the immunogenicity of pMHCs, based on amino acid enrichment- and position importance-scores that were described in the previous two chapters. That such a model was possible agrees with our previous observation that distinct preferences underlie immunogenicity. The immunogenicity model (see Supporting Table S4.S2), based on all HLA presented 9mer peptides, is further used throughout the paper.

The observed T-cell preferences could be the result of neutral evolution, where random mutations have led to certain V-D-J-segments that encode for T-cell receptors with a certain preference. Alternatively, V-D-J-segments might have evolved to encode for TCRs with a preference for pathogen-derived peptides, similar to what is observed for HLA-A molecules [213]. Our immunogenicity model (Supporting Table S4.S2) enables the investigation of such scenarios. For 13 HLA-A and 15 HLA-B molecules, binding ligands were predicted for a large set of viruses and the human proteome (data selection and ligand predictions were previously described in [203]). Next, for each HLA molecule, the predicted viral and human pMHCs were compared by assessing the fraction of ligands with a positive score in the immunogenicity model. For 27 of the 28 HLA molecules, the fraction of viral ligands with a positive score was higher than the fraction of human ligands with a positive score (sign-test: $p<0.001$, see Figure 4.4). The enriched immunogenicity of viral ligands was largest for HLA-A*3001, HLA-B*0702 and HLA-B*4501, where the fraction of viral ligands with a positive score was 11% higher compared to the fraction of human ligands. Only for HLA-A*2301 were there more human ligands with a positive score in the immunogenicity model. Thus, viral MHC-I ligands were predicted to be more immunogenic than human ligands. This result suggests that T-cell preferences have been selected to favour the recognition of foreign peptides.

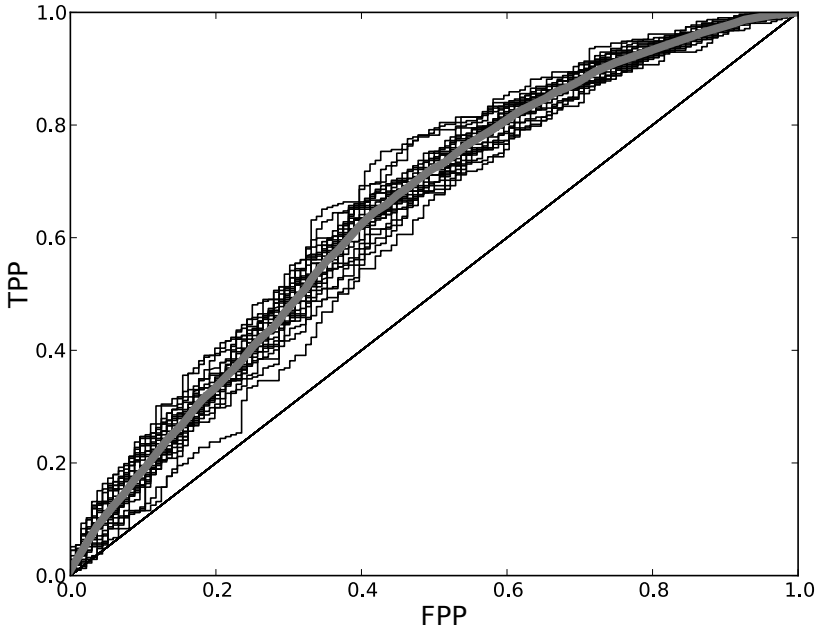


Figure 4.3. Cross-validation of the immunogenicity model. Two-thirds of the data were used for making the immunogenicity model (see methods) and one-third for cross-validation. The average ROC (thick grey line) of 25 of such cross-validations (thin lines) are plotted. The average AUC was 0.65.

4.2.5 Predicting epitopes in mice

Recently, Weiskopf et al. analyzed the immune targeting of a large number of Dengue-derived peptides presented on HLA-A*0101, HLA-A*0201, HLA-A*1101 and HLA-B*0702, upon infection of HLA-transgenic mice with Dengue virus [214]. 22 non-redundant 9mer epitopes and 110 non-redundant 9mer non-epitopes with a high predicted binding affinity ($<$500nM) were reported in this study [214]. Because this novel data set had no similarity to the pMHCs that were used to build the immunogenicity model, it presented an opportunity to test the epitope identification performance of our immunogenicity model. Predicting epitopes is a more challenging task than predicting immunogenicity: while epitopes are expected to be immunogenic and therefore score high in our immunogenicity model, some non-epitopes may well be immunogenic in immunization experiments, but lack immune targeting due to other factors such as a lack of processing or expression of the peptide during infection. Surpassing our expectations, the immunogenicity model scored the epitopes much higher than

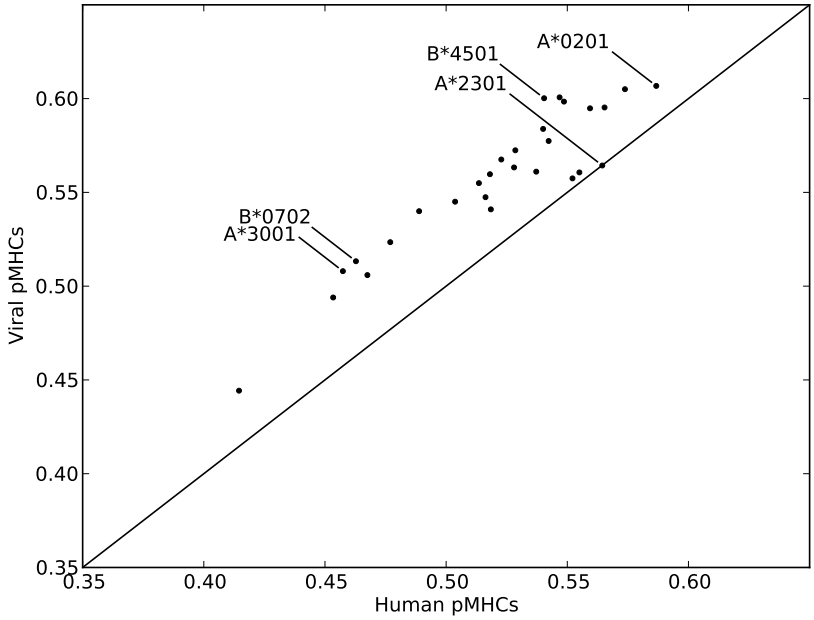


Figure 4.4. Viral pMHCs are better recognized by T-cells. For 13 HLA-A and 15 HLA-B molecules, the fraction of viral pMHCs (y-axis) and human pMHCs (x-axis) with a positive immunogenicity score is shown. The diagonal denotes the line y=x, HLA molecules with a larger fraction of positively scoring viral pMHCs fall above this line, which was the case for 27 of the 28 HLA molecules (sign-test: $p<0.001$). The three HLA molecules on which the viral ligands were most immunogenic, the one HLA molecule on which human ligands were most immunogenic and HLA-A*0201 are indicated in the figure.

the non-epitopes (Ranksums test: $p<0.01$; see Figure 4.5). Besides validating our prediction model, this analysis provided an example of how one could apply the immunogenicity model to enrich for epitopes by exluding non-immunogenic pMHCs. If 38% of the peptides would not be tested because the immunogenicity model gave them a negative score, still 86% of the epitopes would be identified. In a large study where many peptides have to be tested this means a significant fraction of the work can be saved when using the immunogenicity model. Taken together, in an independent data set of Dengue peptides presented on different HLA class I molecules in infected mice, we could validate the immunogenicity model and show that it could assist the identification of epitopes.

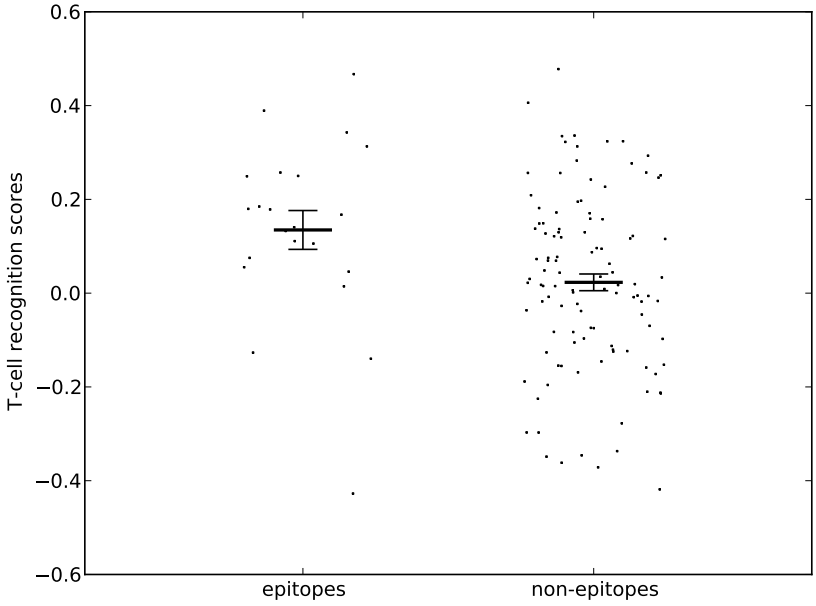


Figure 4.5. Predicting Dengue-derived epitopes with the immunogenicity model in mice. For non-redundant epitopes (n=22) and non-epitopes (n=110) identified by Weiskopf et al. [214], the immunogenicity scores were determined. Average and variation of the average are shown as thick lines with error bars, individual scores are shown as dots. The epitopes had a significantly higher immunogenicity score than the non-epitopes (Ranksums test: $p<0.01$; AUC=0.69).

4.2.6 Predicting epitopes in humans

Next, we tested whether we could extrapolate our immunogenicity predictions to HLA presented peptides in humans. As mentioned before, few peptide immunization studies are performed in humans, therefore very few pMHCs (n=1) could be classified as non-immunogenic in humans. However, human immunogenic pMHCs could be identified (Figure 4.1), and the amino acid profile of these pMHCs was compared to the amino acid profile of murine immunogenic and non-immunogenic pMHCs (Figure 4.1). The human immunogenic pMHCs were more similar to the immunogenic pMHCs in HLA-transgenic mice (Kullback-Leibler divergence = 0.024), than to the non-immunogenic pMHCs in HLA-transgenic mice (Kullback-Leibler divergence = 0.069). Thus, immunogenic pMHCs have a similar amino acid profile in mice and men, suggesting that T-cell preferences between these two species are comparable.

To test if our immunogenicity model can assist the identification of epitopes in humans, we made use of a recent large epitope discovery study that was recently conducted in Dengue seropositive donors by Weiskopf et al. (Weiskopf et al., manuscript in preparation). In this study, T-cell responses were measured in Dengue seropositive donors, to predicted MHC ligands on the HLA molecules of those donors. Ethics approval was granted for the dengue virus large scale epitope discovery study from the LIAI IRB and the Ethical Review Committee at Medical Faculty, University of Colombo, Sri Lanka. In total, 42 non-redundant 9mer epitopes and 477 non-redundant 9mer non-epitopes were derived from this study (see Methods for selection and redundancy reduction criteria), for which immunogenicity scores were determined and compared. Similar to our result based on murine data (Figure 4.5), the human epitopes scored much higher than the non-epitopes (Ranksums test: $p=0.014$; see Figure 4.6). Thus, also in the human immune system could we validate the immunogenicity model. This finding confirms that studies in HLA-transgenic mice provide usefull data to understand human T-cell recognition, in agreement with other studies that compared the immune responses in HLA-transgenic mice and men [215]. In addition, we could show its usefullness in a human epitope discovery project, if 43% of the peptides with a negative score would be excluded from the experiment, still 71% of the epitopes would be identified. Taken together, our immunogenicity model can predict the immunogenicity of pMHCs in both mice and men, and assist the identification of epitopes in these hosts.

4.3 Discussion

Immunogenicity (i.e. T-cell recognition) is an important factor that determines if a pMHC can be targeted in an immune response. We showed that certain pMHCs are more likely to be immunogenic because they have certain amino acid residues. More precisely, the presence of large and aromatic residues seemed to be associated with immunogenicity. In addition, positions 4-6 of the presented peptide were shown to have a large effect on immunogenicity. We combined these findings into a simple model and demonstrated that the model can predict the immunogenicity of pMHCs in independent data sets.

We benchmarked our immunogenicity prediction model on epitope and non-epitope data sets that were derived in mice and men. As expected given their immunogenicity, most epitopes obtained high scores in our model. Conversely, some non-epitopes do not elicit an immune response because they are non-immunogenic, indeed some of the non-epitopes scored much lower in our immunogenicity prediction model. The prediction of non-immungenicity will be usefull in future large-scale epitope discovery studies, as it might shorten the list of potential peptides that have to be tested without finding less epitopes. In both data sets that were used to test the immunogenicity prediction model, we showed

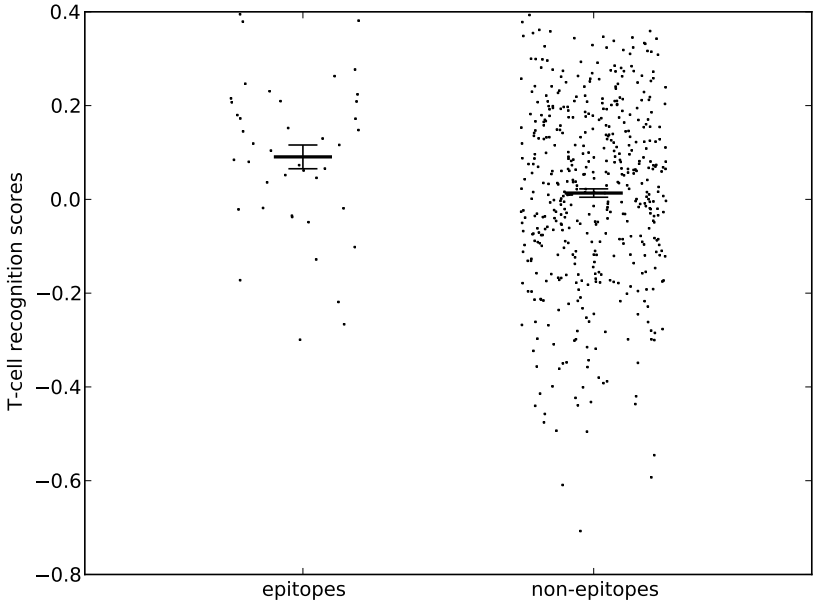


Figure 4.6. Predicting Dengue-derived epitopes with the immunogenicity model in humans. For non-redundant epitopes (n=42) and non-epitopes (n=477) identified by Weiskopf et al. (manuscript in preparation), the immunogenicity scores were determined. Average and variation of the average are shown as thick lines with error bars, individual scores are shown as dots. The epitopes had a significantly higher immunogenicity score than the non-epitopes (Ranksums test: p=0.014; AUC=0.61).

that ∼40% of the peptides can be discarded, while losing only 15-30% of the epitopes. Even though such an enrichment of 25-40% of the number of epitopes in the positive predicted peptides seems like a small improvement, the effect can be large in studies where patient-derived samples or other resources are limited.

Previously, other groups have studied the importance of different positions in an MHC-I presented peptide using two distinct approaches. First, specific T-cell clones have been assessed for the recognition of variant peptides [118, 207–211, 216], most T-cell clones in such studies lost the recognition of peptides that were substituted at positions between the anchors (P3-8). In well-studied systems, such as the T4 T-cell clone recognizing the SLFNTVATL peptide on HLA-A2, recognition of position P5 was most specific, followed by a high specificity at the flanking positions P4 and P6 [118, 208]. Second, the study of TCR-pMHC structures contributed to the understanding of immunogenicity. In such structures, the number of interactions between the TCR and different positions of the

MHC-I presented peptide have been evaluated [60, 203, 206], and more interactions were observed with the positions P4-P8. The results from both approaches seem to agree: that positions P4-8 and of those especially positions P4-6 are most important for immunogenicity. Now, we present a third line of evidence that supports these observations: the amino acids at different positions of the MHC-I presented peptide in immunogenic and non-immunogenic pMHCs differ most at positions P4-P6, suggesting that these positions are most important determinants of immunogenicity.

Based on the comparison of immunogenic and non-immunogenic pMHCs we derived a simple model to predict immunogenicity. We call our immunogenicity model simple because it does not account for non-linear influences on immunogenicity, or position-specific amino acid enrichment scores. Position-specific scores (i.e. 20 scores per position) seem to present an opportunity for further improvement, as different preferences seem to occur at different positions, e.g. a preference at position 6 for non-charged and non-polar residues (Fisher's exact test: $p<0.05$; data not shown). Unfortunately, the current data sets are too small to incorporate position specific preferences into the immunogenicity model without running the risk of overfitting. We believe our simple model provides a proof-of-principle that immunogenicity is predictable, and that more complex and possibly more accurate predictors can be made if more data, especially non-immunogenic pMHCs, is available.

The group of Ho et al. have pioneered the field of immunogenicity predictors, and recently published a method for immunogenicity prediction called POPISK [112]. POPISK aims to predict the immunogenicity of HLA-A*0201 presented peptides and reports a high accuracy in cross-validation (AUC=0.74, see [112]). POPISK is different from our predictor in three important ways. First, it is trained on all peptide positions of HLA-A*0201 presented peptides, whereas we exclude positions that influence the binding affinity such as the anchor positions P2 and P9. Second, non-immunogenic pMHCs in the IMMA2 data set that was used to train POPISK were not defined based on negative results in a peptide-immunization experiment, therefore other explanations for the absence of an immune response besides non-immunogenicity cannot be excluded. Third, POPISK is a rather complex model using support vector machines and string kernels. A complex model runs the risk to be overtrained, especially on a limited data set, which will not be noticed in cross-validation if redundant peptides are not excluded from the data sets, as is the case for the IMMA2 data set that was used to build POPISK [112]. Possibly due to such differences, POPISK is not able to score the Dengue-derived epitopes that were recently published by Weiskopf et al. [214] higher than the non-epitopes from this study, neither based on all pMHCs (1-sided t-test: $p=0.28$; not shown), nor on the HLA-A*0201 presented pMHCs (1-sided t-test: $p=0.39$; not shown). A model like POPISK might perform better if it is trained on more high quality data. For now, we think that the available data only permits the construction of simple proof-of-principle immunogenicity predictors, and the study

of basic features of immunogenicity, that we studied and describe here.

The TCR repertoire can be influenced by the hosts genetics, e.g. the HLA-background of a host [91, 105, 106], or the likelyhood of certain VDJ-recombinations [85, 88, 90, 105]. Even though the T-cell pool might vary in every individual as a result of such influences, we found that T-cells have a preference for certain amino acids (see Supporting Table S4.S2, the immunogenicity model). That preferences are similar among hosts agrees with the observation from Alanio et al. that T-cell precursor frequencies for the same pMHC are similar in different hosts, whereas precursor frequencies for different pMHCs vary substantially [191]. Furthermore, we showed that these preferences resulted in a better recognition of pathogen-derived pMHCs (Figure 4.4). The observed preferences might be the result of natural selection for the increased immunogenicity of pathogen-derived pMHCs, additional to the widely suggested selection for TCR-genes that interact with conserved MHC-I motifs [60, 61, 217, 218].

We focused in this paper on MHC-I presented peptides and showed a preference for large, aromatic residues. This fits with a previous study by Alexander et al. who tested the immunogenicity of so-called PADRE peptides, that are presented on most MHC class II molecules but that differ in T-cell recognition sites [219]. Interestingly, they showed that PADRE peptides with large residues are very immunogenic. Thus, T-cell preferences for peptides presented either on MHC class I or II molecules seem to overlap. We have also performed an analysis of amino acid preferences for H-2 restricted pMHC complexes on a limited dataset, and found a somewhat different pattern of preferred amino acids (data not shown). Such a difference might be expected given the altered peptide binding preferences of H-2 molecules, that present short 8mer peptides and use more and different auxiliary anchor positions than HLA class I molecules [220]. More experimental data and further studies are necessary to analyze if this difference is significant, and if so, if it is due to structural, evolutionary, or other differences between the immunogenicity of peptides that are presented on HLA class I or H-2 molecules. Similarly, preferences for peptides might be different as they are presented on different HLA molecules, but we currently lack the data to investigate the influence of MHC-I restriction on immunogenicity.

The identification of all pMHCs that are epitopes would be prerequisite to a complete understanding of the cellular immune response. That understanding would help the study of host-pathogen interactions, for instance how pathogens try to escape from immune recognition by mutating the epitopes that are under pressure of the immune system [221, 222]. In addition, the identification of epitopes will help the development of better vaccines, that effectively elicit protective immune responses. In past years, investigations of the MHC-I presentation pathway led to the development of highly accurate predictors that can predict which pMHCs are formed upon infection. However, we do not know which pMHCs are used by the immune system to mount a T-cell response. Previously, we and others showed

that self-similarity plays an important role in excluding some pMHCs as potential epitopes [117, 118, 203], and we estimated that at least one-third of the foreign pMHCs would be ignored to prevent otherwise autogenic responses [203]. Now, we add another piece to the epitope-puzzle, and show that the important role that immunogenicity plays in determining which pMHCs are epitopes is predictable. A combination model that integrates predictions from the MHC-I presentation pathway and immunogenicity will allow us to more accurately predict epitopes, and will be used in the future to assist large-scale epitope discovery projects.

4.4 Methods

4.4.1 Generation of data sets

The aim of this study is to compare immunogenic and non-immunogenic peptides on MHC class I molecules. These peptides were obtained from data sets from Assarsson et al [18], Kotturi et al. [108] and an unpublished data set on Coxiella Burnetti-derived peptides as well as the IEDB [185]. Only 8-10mer peptides were selected, for which reliable MHC-I binding predictions are possible. Using NetMHC-3.2 [147], a binding affinity predictor that was shown to perform best in a large benchmark study [50], binding affinities were predicted for all pMHC. Only pMHC with a high predicted binding affinity were selected ($<$500nM).
The data set by Assarsson et al of vaccinia-derived peptides presented in an HLA-A*02 transgenic mouse model, has been classified by the authors into “dominant", “subdominant", “cryptic" and “negatives" [18]. Peptides classified as immunogenic induced a positive response in the peptide-immunization experiment performed in this study (categories “dominant", “subdominant" or “cryptic"; n=63; see Figure 4.1), while the non-immunogenic peptides were the ones that did not induce an immune response in this experiment (category “negative"; n=33; see Figure 4.1).
Data described by Kotturi et al. was kindly provided by the authors. Kotturi et al. studied the immunogenicity of peptides presented on HLA-A*1101 that are derived from Arenaviruses [108]. In an HLA-transgenic mice, T-cell recognition upon peptide-immunization was measured. If a significantly high T-cell response was elicited (t-test: $p<0.05$; SFC$>$20 per million; stimulation index$>$2.0) in at least two independent measurements (detailed in [108]), a peptide was classified as immunogenic (n=116, see Figure 4.1). All other peptides were classified as non-immunogenic (n=159, see Figure 4.1).
A previously unpublished set of peptides derived from Coxiella burnetti proteins was tested for immunogenicity in wild type Bl/6 mice. The immunization protocol and criteria for positivity were the same as for the Kotturi data set [108]. 11 Immunogenic and 16 non-immunogenic pMHCs were derived from this experiment.

A large data set was derived from the IEDB, where all T-cell response experiments (i.e. peptide-immunizations, but also vaccination and infection experiments) with MHC class I presented peptides in mice or humans were downloaded (www.iedb.org [185]). All entries from HLA-A*1101-transgenic mice were excluded, to rule out any bias resulting from the incompatibility of the HLA-A*1101 binding motif and the preferences of murine TAP [223]. This requirement was alleviated for the data from the Kotturi study as we know that in this peptide-immunization study there was no need for peptides to be TAP transported. If no restricting MHC molecule was reported it was estimated from the reported mouse strain MHC background, if multiple MHC class I molecules were possible the molecule with highest predicted binding affinity was selected as the restricting MHC-I molecule. Immunogenic pMHCs were selected based on a reported positive T-cell response, and the absence of restimulation in vitro. Non-immunogenic peptides were selected based on a reported negative T-cell response and the absence of any reported positive T-cell response. In addition, as with the other data sets, non-immunogenic pMHCs were required to be identified in a peptide-immunization experiment: the antigen-epitope relation had to be “epitope" and the first in vivo immunogen had to be “peptide from protein". This resulted in the identification of 2029 immunogenic and 152 non-immunogenic pMHCs (see Figure 4.1).

4.4.2 Non-redundancy selection

The data in databases such as the IEDB is biased towards pMHCs that are well-studied. For instance, for the SIINFEKL peptide we find 358 entries in the IEDB, and 22 entries of single amino acid mutants. To eliminate such cases in our dataset, a redundancy reduction based on source protein mapping was applied. First, for all peptides in our datasets that were identified as immunogenic or non-immunogenic following the above requirements (see Figure 4.1), source proteins were downloaded via the sequence information provided in the IEDB. In addition, for the Vaccinia-, Coxiella- and Arenavirus-derived pMHCs source proteomes were downloaded via EBI/EMBL in July 2011. Next, all peptides were mapped to all source proteins using BLASTP 2.2.18 [224], and a mapping was considdered successful if more than 75% of the peptide identities matched. Two peptides were defined as redundant if more than half of their residues map to the same positions in any of the source proteins. In addition, two peptides were defined redundant if both could not be mapped. Redundant peptides were filtered out, wherein we prioritized the selection of 9mer peptides and the selection of prominent pMHC with more entries in the IEDB. If redundant pMHCs with equal priority remained, the selection of one of them was based on chance, this was the case for 11 of the 193 non-immunogenic pMHCs and 28 of the 636 immunogenic pMHCs. As a result, selected pMHC sets can vary slightly. A single non-redundant pMHC set was selected and used for the presented analysis, but every result was tested

and repeated in ten (of ten) non-redundancy selections. When selecting non-redundant Dengue-derived peptides from the HLA-transgenic mouse experiments of Weiskopf et al. [214], the selection of epitopes with a high T-cell response and non-epitopes with a strong binding affinity was prioritized. Selected epitopes (n=22) and non-epitopes (n=110) did not differ significantly in their predicted binding affinities.

4.4.3 Selecting Dengue-derived epitopes in humans

Weiskopf et al. tested the immune responses to Dengue-derived peptides in Dengue seropositive donors (manuscript in preparation). For every donor, the HLA background was determined, and peptides predicted to be presented on these HLA molecules were tested. We defined pMHCs with a positive immune response in any of the donors as epitopes, a pMHC that never evoked an immune response and that was not redundant with an epitope was defined as a non-epitope. Only 9mer peptides from the epitope and non-epitope sets were selected, and non-redundant pMHCs were selected from both sets. In addition, as 5 of the 229 donors contributed to 50% of all detected immune responses, we selected per donor 5 epitopes with a highest immune response, to prevent a bias that might have been caused due to the very broad T-cell response in these donors. Selected epitopes (n=42) and non-epitopes (n=477) did not differ significantly in their predicted binding affinities.

4.4.4 The immunogenicity model

The immunogenicity model is build based on the enrichment of amino acids in immunogenic versus non-immunogenic peptides and the importance scores of different positions of the MHC-I presented peptide (Table 4.2). Only non-anchor positions were used, these positions were defined for each MHC-I molecule as the six positions with least impact on the binding affinity. The impact on binding affinity was determined for each position by calculating the selectivity of the MHC-I binding predictor NetMHC-3.2 at that position (as explained in [203]). Per amino acid, the enrichment is calculated as the ratio between the fraction of that amino acid in the immunogenic versus non-immunogenic data sets. For instance, if 2.5% of the residues in immunogenic and 1.5% of the residues in non-immunogenic peptides are a Tryptophan, the enrichment in immunogenic peptides is 1.7-fold, and the natural logarithm of this enrichment is 0.54, we call this a log enrichment score. To predict the immunogenicity of a new pMHC, per non-anchor residue of the presented peptide the immunogenicity score is found and weighted according to the importance of that position, this importance was measured as the Kullback-Leibler divergence (see Table 4.2). The scores of all (non-anchor) residues are summed, the larger this score, the more the pMHC is

like the immunogenic peptides and therefore expected to be immunogenic. The scores of amino acids at anchor residues are masked, i.e. not used to derive the immunogenicity score.
Given a peptide ligand, L, presented on an HLA molecule, H, the immunogenicity score, S, is calculated as follows:

$$S(H,L) = \sum_{p=1}^{9} E_{A(L,p)} \times I_p \times M(H,p) \tag{4.1}$$

Where for every position p in the ligand L, the enrichment E for the amino acid at that position A(L,p) times the importance of that position I_p times the eventual masking of that position on that HLA because it is an anchor position, M(H,p) that equals 1 (no masking) or 0 (masking), is summed (eq. 4.1).
The immunogenicity model was tested in a 3-fold cross-validation experiment, where a random two-thirds of the data is used to calculate the enrichments. These enrichments, together with the position importance weights (Table 4.2) were then used to construct the model as described above, and the other one-third was used to test the performance of this model. 25 Cross-validations were performed and the model was shown to be able to discriminate immunogenic from non-immunogenic pMHCs in all cases. Our final immunogenicity model, that is used throughout this paper is based on all HLA class I presented peptides found in HLA-transgenic mice. The exclusion of redundant peptides is an important step in the collection of this data. However, if two redundant peptides are equally fit to be selected, the selection of a single non-redundant peptide will be a matter of chance. As a result of this, the selected set of peptides and the immunogenicity model based on these peptides can vary slightly. For the validations on Dengue-derived epitope data from humans and mice, a final immunogenicity model was constructed by repeating the non-redundancy selection and model building 100 times, and taking average log enrichment scores per amino acid from these 100 models. The final enrichment scores, position importance weights and explanations on constructing the immunogenicity model are shown in Supporting Table S4.S2.

4.4.5 Amino acid properties

Different groups of amino acids were assembled based on shared characteristics. These groups were used to test if certain characteristics associate with immunogenic or non-immunogenic peptides. Small amino acids were grouped with a size of less than 120 Da (A,G,P,S,T,V), and large amino acids with a size of more than 150 Da (F,H,R,W,Y). Definitions of the other groups, based on conventional views: Polar amino acids (C,G,N,Q,S,T,Y), non-polar amino acids (all amino acids that are not polar and not charged), aromatic amino acids (F,H,W,Y), non-aromatic amino acids (all amino acids that are not aromatic), charged amino acids

(D,E,H,K,R), non-charged amino acids (all amino acids that are not charged), acidic amino acids (D,E) and basic amino acids (H,K,R).

4.4.6 Statistics

Statistical tests were performed using the stats-package from the scipy-module in Python. To assess the significance of the association of a certain amino acid with immunogenicity, a permutation test was performed. For each amino acid, the frequency in non-anchor positions of immunogenic and non-immunogenic peptides, and the background frequency in source proteins was determined (data used for Figure 4.2). Per permutation, based on the background frequency and the total number of immunogenic and non-immunogenic amino acids, a random sample of immunogenic and non-immunogenic amino acids was generated and the frequency was determined for each amino acid. In every permutation, the difference between simulated immunogenic and non-immunogenic frequencies was determined for each amino acid, this difference was compared with the difference in the real peptides. 10000 Permutations were performed and per amino acid the fraction of permutations in which the difference was larger or equal than the real difference determined the chance of finding our result by chance, i.e. the p-value.

4.5 Acknowledgments

We thank Rob de Boer and Johannes Textor for valuable discussion on this research project. We thank the National Blood Center, Ministry of Health, Colombo, Sri Lanka for providing buffy coat samples used in the dengue virus large scale epitope discovery study. This study was financially supported by the University of Utrecht, and the National Institutes of Health contracts HHSN272201200010C and HHSN272200900042C.

4.6 Supporting Information

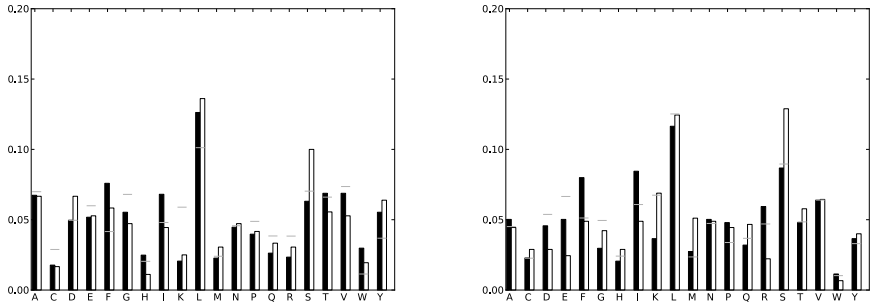


Figure 4.S1. T-cell preferences for different amino acids in HLA-A*0201 presented peptides (left panel) or peptides presented on other HLA molecules (right panel). The fraction of an amino acid in immunogenic (left bar, filled) and non-immunogenic (right bar, unfilled) peptides restricted on these HLA molecule sets is shown. The background frequency for each amino acid in the protein sequences that were a source of the immunogenic or non-immunogenic peptides is shown by a grey line.

Formula: $S(H,L) = \sum_{p=1}^{9} E_{A(L,p)} \times I_p \times M(H,p)$

Amino Acid	log enrichment score (E)
A	0.127
C	-0.175
D	0.072
E	0.325
F	0.380
G	0.110
H	0.105
I	0.432
K	-0.700
L	-0.036
M	-0.570
N	-0.021
P	-0.036
Q	-0.376
R	0.168
S	-0.537
T	0.126
V	0.134
W	0.719
Y	-0.012

Position	Importance (I)
1	NA
2	NA
3	0.10
4	0.31
5	0.30
6	0.29
7	0.26
8	0.18
9	NA

Example: SLFNTVATL (L) on HLA-A*0201 (H)

-masked positions (M) for HLA-A*0201: 1,2 and 9

-score position 3, F: 0.10 * 0.380 = 0.038
-score position 4, N: 0.31 * -0.021=-0.007
-score position 5, T: 0.30 * 0.126 = 0.038
-score position 6, V: 0.29 * 0.134 = 0.039
-score position 7, A: 0.26 * 0.127 = 0.033
-score position 8, T: 0.18 * 0.126 = 0.023

-Final score, S, (sum of all position scores): 0.164

Figure 4.S2. The immunogenicity model. The immunogenicity score, S, is derived by summing the log-enrichment scores of amino acids that are found at non-masked positions, weighted by the importance of that position (see formula and Methods). The final log-enrichment scores for all amino acids are given in the left table, importance scores for the different positions are shown in the right table (also shown in table 4.2). An example to calculate the score for HLA-A*0201:SLFNTVATL is given.

AAindex	description	corr.	p-value	q-value
CHOP780213	Frequency of the 2nd residue in turn	-0.52	0.020	0.59
PRAM820101	Intercept in regression analysis	-0.52	0.020	0.59
RICJ880114	Relative preference value at C1	-0.51	0.021	0.59
TANS770104	Normalized frequency of chain reversal R	-0.49	0.028	0.59
RICJ880103	Relative preference value at N-cap	-0.46	0.043	0.59
OOBM770103	Long range non-bonded energy per atom	-0.45	0.044	0.59
VINM940104	Normalized flexibility parameters	-0.45	0.045	0.59
MEIH800101	Average reduced distance for C-α	-0.45	0.047	0.59
VASM830102	Relative population of conformational state C	-0.45	0.048	0.59
CHOP780210	Normalized frequency of N-terminal non β region	-0.45	0.048	0.59
MEEJ810102	Retention coefficient in NaH2PO4	0.44	0.050	0.59
MEEJ810101	Retention coefficient in NaClO4	0.45	0.047	0.59
NAKH900113	Ratio of average and computed composition	0.46	0.039	0.59
LEVM780102	Normalized frequency of β-sheet, with weights	0.48	0.031	0.59
PRAM900103	Relative frequency in β-sheet	0.48	0.031	0.59
OOBM850103	Optimized transfer energy parameter	0.49	0.029	0.59
PALJ810112	Normalized frequency of β-sheet in α/β class	0.51	0.021	0.59
PONP800107	Accessibility reduction ratio	0.56	0.011	0.59
WILM950102	Hydrophobicity coefficient in RP-HPLC	0.67	0.001	0.49

Table 4.S3. Amino acid characteristics that correlate with our enrichment values (Figure 4.S2). For all amino acid indices that are described in the AAindex-database [205], the Spearman Rank correlation (corr.) with enrichment scores in immunogenic pMHCs was determined. All significant ($p<0.05$) correlations are reported, q-values are given in the fifth column, they give the estimated False Discovery Rate (see [204]) which was very high in all cases >0.4 due to the large number of tests ($n=505$). The "hydrophobicity coefficient in RP-HPLC"-index showed the best correlation, but is not the only measure of hydrophobicity, all other indices with the term "hydrophobic" or "hydrophobicity" in their description ($n=35$) were not significantly correlated with our enrichment scores ($p>0.1$, not shown).

Chapter 5

Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire

JORG J.A. CALIS*, ROB J. DE BOER* AND CAN KEŞMIR* (2012)

*Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, The Netherlands

PLoS Computational Biology, **8**(3):e1002412

Abstract

The cellular immune system screens peptides presented by host cells on MHC molecules to assess if the cells are infected. In this study we examined whether the presented peptides contain enough information for a proper self/nonself assessment by comparing the presented human (self) and bacterial or viral (nonself) peptides on a large number of MHC molecules. For all MHC molecules tested, only a small fraction of the presented nonself peptides from 174 species of bacteria and 1000 viral proteomes ($\sim$0.2%) is shown to be *identical* to a presented self peptide. Next, we use available data on T-cell receptor-peptide-MHC interactions to estimate how well T-cells distinguish between similar peptides. The recognition of a peptide-MHC by the T-cell receptor is flexible, and as a result, about one-third of the presented nonself peptides is expected to be *indistinguishable* (by T-cells) from presented self peptides. This suggests that T-cells are expected to remain tolerant for a large fraction of the presented nonself peptides, which provides an explanation for the "holes in the T-cell repertoire" that are found for a large fraction of foreign epitopes. Additionally, this overlap with self increases the need for efficient self tolerance, as many self-similar nonself peptides could initiate an autoimmune response. Degenerate recognition of peptide-MHC-I complexes by T-cells thus creates large and potentially dangerous overlaps between self and nonself.

5.1 Introduction

The recognition of peptide-MHC-I complexes (pMHC) by the T-cell receptor (TCR) is required for effector T-cells to kill an infected cell. Although some MHC-I molecules have a preference to present pathogen-derived peptides [213], pMHC are formed with both self and nonself peptides. Therefore, to allow $CD8^+$ T-cells of the cellular immune system to discriminate self from nonself, presented nonself peptides should be different from presented self peptides. What would happen if a nonself peptide is so similar to a self peptide that it is recognized by the same T-cell (we will call such peptides "overlapping peptides")? Firstly, an effector T-cell response to an overlapping peptide, could cause T-cell mediated autoimmune disease, such as type 1 diabetes [78–80] or multiple sclerosis [225, 226]. Secondly, to avoid autoimmunity, T-cells recognizing self-pMHCs are tolerized during negative selection [106]. Due to this self tolerance, overlapping nonself peptides should fail to elicit a T-cell response, and this may limit the number of pathogen-derived peptides that are available for an immune response and hence the chance to control a pathogen [117, 118]. Assarsson et al. showed that $\sim 50\%$ of the MHC-I presented vaccinia derived peptides are not recognized by T-cells [18]. Similarly, for HIV-1-derived peptides predicted to be presented on the well-studied HLA-A*0201 molecule, only $\sim 50\%$ has been reported to elicit a T-cell response [118]. Taken together, these studies suggest large "holes" in the T-cell repertoire [117, 227], which could be caused by overlaps with self pMHCs.

We have previously shown that on HLA-A2 molecules only a minute fraction (0.26%) of the presented nonself peptides are identical to presented self peptides [15]. Such a small overlap can not cause the large holes in the T-cell repertoire. However, at that time there was too little data available on T-cell recognition of pMHCs, to study its impact on the self/nonself overlap. It is well established that T-cells are cross-reactive and can recognize similar, and sometimes even unrelated, peptides presented on the same MHC molecule [228]. The principles of TCR-pMHC interactions that allow for this flexibility are not fully understood. CTL recognition-studies using peptide libraries with altered peptide ligands [118, 207–211] and pMHC-TCR structures [60, 206] allow some inferences to be made. The middle (P4-P6) part of the peptide forms the core of the interaction [60, 118, 206–211], where the majority of amino acid substitutions (with exception of those with very similar amino acids) tend to perturb pMHC recognition. Other positions in the peptide, although not in direct contact with the TCR, can still be important for the TCR-pMHC interaction if they affect the configuration of the P4-P6 residues [210], or MHC-binding [229]. In most cases, the N-terminal position (P1) of the peptide is unimportant for the TCR-pMHC interaction [60, 118, 206–210].

Given these new insights, we here extend our previous investigations on self/nonself overlaps by including the T-cell recognition of pMHCs. In addition, we analyze

the self/nonself overlap of peptides presented on several HLA-A and HLA-B molecules, to estimate the degree of variance among different MHC-I molecules. Using high-quality predictors of the MHC-I presentation pathway [16, 50, 51, 147], we show that presented peptides derived from nonself are in almost all cases ($> 99.7\%$) distinct from presented self peptides, for all common MHC molecules. This result is in agreement with our original observation that most peptides with a length of nine amino acids (9mers) of unrelated species are unique [15]. However, the cross-reactivity of T-cell recognition is shown to increase the self/nonself overlap between sufficiently similar peptides to about one-third. Our results suggest an explanation for the observed holes in the T-cell repertoire during an infection, and we show that our self/nonself overlap estimates can be used to distinguish immunogenic from non-immunogenic pMHCs. Moreover, the estimates of self/nonself overlap demonstrate that the risk of autoimmunity due to molecular mimicry with pathogens is nonnegligible.

5.2 Results

5.2.1 Self/nonself overlaps based on peptides

MHC class I molecules shape $CD8^+$ T-cell responses via the presentation of peptides derived from intracellular proteins. These peptides are short: most MHC-I molecules prefer to bind peptides of 9 amino acids (9mers). To investigate how similar self and nonself peptides are, the human and a large number of nonself proteomes (data selection is detailed in Methods) were cut into fragments of various lengths (1-20 amino acids long) and peptides that occur both in self and nonself proteomes were identified (i.e. without considering MHC-I presentation). The fraction of foreign peptides that are also present in the human proteome defines the "overlap", i.e. the chance that a randomly chosen nonself peptide is identical to a self peptide. For small peptides shorter than five amino acids, the overlap is 100%, since almost every 5mer is present in the human proteome (see Figure 5.1). For longer peptides the overlap decreases rapidly, and at a length of 9 amino acids the average overlap is only 0.20% for viruses (between 0-0.5% for 95% of all viruses) and 0.19% for bacteria (0.1-0.4% for 95% of all bacteria). These results are in excellent agreement with our previous estimates based on a much smaller set of nonself proteomes [15]. To conclude, 9mers contain enough information to discriminate self from nonself, i.e. the chance that a nonself 9mer overlaps with a self 9mer is only 0.2%.

Surprisingly, the overlaps do not decrease much further for peptides longer than 9mers (see Figure 5.1). To characterize these overlapping sequences further, for each human protein we counted the number of viruses or bacteria that has at least one overlapping 9mer peptide. The proteins where this number was larger

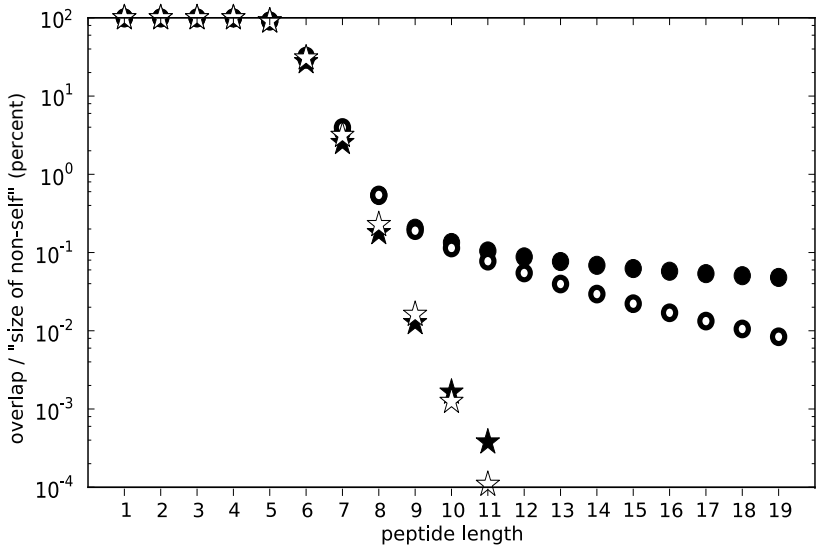


Figure 5.1. Viral and bacterial self/nonself overlaps for peptides of different lengths. The chance that a bacterial or viral peptide overlaps with a peptide in the human proteome is shown as open and closed circles for bacteria and viruses, respectively. Stars indicate the self/nonself overlaps with shuffled bacterial (open stars) or viral (closed stars) proteins. For all peptides of 5 amino acids or longer, the overlap of unshuffled viruses and bacteria is significantly smaller than the shuffled (representing the expected) overlap (Ranksums test: $p<0.05$).

than expected ($p<0.01$, see Methods) were analyzed by a functional annotation cluster analysis [230, 231]. This analysis showed that bacterial 9mers tend to overlap with human proteins of mitochondrial origin, which is in line with the bacterial origin of mitochondria [232]. In addition, proteins involved in metabolic processes that might be common to bacteria and humans had more overlapping 9mers (see Supporting Table 5.S1). For viruses, the overlap is largest with nuclear proteins and transcription factors that are possibly acquired via horizontal gene transfer to modulate host cellular processes (see Supporting Table 5.S1). In order to test the effects of homologous sequences or convergent evolution on self/nonself overlaps, sequences were shuffled before examining the overlap to break up any overlap that might be the result of these effects. Indeed, this shows that a far majority of the overlaps were due to these homologous sequences as the overlaps in shuffled sequences are much lower than the actual overlaps (Figure 5.1, in stars).

5.2.2 Self/nonself overlaps based on peptide-MHC-I complexes

Only peptides that are presented on an MHC-I molecule, i.e. about 1-3% of all 9mers [18], can be recognized by T-cells. Due to the binding preferences of different MHC-I molecules, the self/nonself overlap of MHC-I presented peptides can be different per MHC-I molecule and does not need to be the same as the overlap based on all 9mers. For instance, we recently showed that certain MHC-I molecules have a preference for pathogen-specific peptides [213]; such a preference should decrease the self/nonself overlap for that MHC-I molecule. To estimate the self/nonself overlap of MHC-I presented peptides, an *in silico* approach was undertaken using state-of-the-art MHC-I pathway predictors [16, 50, 51, 147] (see Methods).

For a large set of common human MHC-I molecules (13 HLA-A molecules and 15 HLA-B molecules, see Methods for selection criteria), the presented peptides in the human proteome and a large set of nonself proteomes were predicted. To define presented peptides we made use of the well studied HLA-A*0201 molecule. For this molecule an IC50 value of 500nM is often taken as threshold to seperate the binders from non-binders. Applying this threshold to all self peptides we find that HLA-A*0201 has a specificity of 2.3%, i.e. 2.3% of the tested peptides would be binders. For other HLA molecules we determined "scaled" binding thresholds, so that they have the same specificity as HLA-A*0201, i.e. they present 2.3% of all self peptides. Next, the overlap between presented self and nonself peptides was enumerated per MHC-I molecule, by comparing for each HLA molecule, self and nonself peptides presented on that HLA molecule. On average, only 0.15% of the MHC-I presented nonself peptides is identical to a presented self peptide (see Figure 5.2A, left). The average overlap of MHC-I presented peptides is somewhat smaller than the overlap of all 9mers in the proteome (0.2%, see Figure 5.1), which is in agreement with the fact that many MHC-I molecules have a slight preference for pathogen-derived peptides [213]. The maximal overlap of 0.33%, which is still very low, was found for peptides presented by HLA-B*5401. These results demonstrate that for all common human MHC-I molecules, only a minute fraction of the presented nonself peptides is identical to a presented self peptide. By using scaled binding thresholds, we take the conservative assumption that different HLA molecules have similar specificities, this does not have to be so. The self/nonself overlaps were also calculated by using a fixed binding threshold of 500 nM, which leads to different specificities for different HLA molecules. The self/nonself overlap determined for peptides presented on different HLA molecules remain as low as in the case of using scaled thresholds (see Figure 5.2A, right).

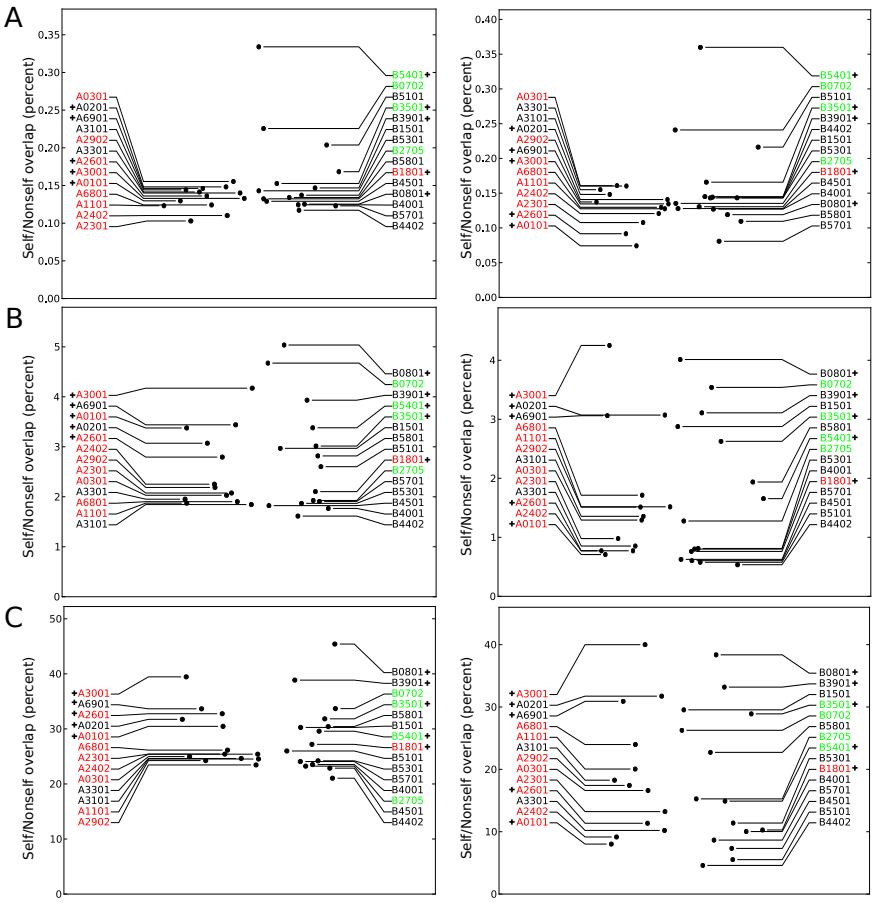


Figure 5.2. Self/nonself overlaps of peptides presented on different HLA molecules. In A, the exact overlap of the complete peptide (positions 1-9). In B, the exact overlap of the middle positions of the peptide (positions 3-8) that are assumed to be in contact with the TCR. In C, the degenerate overlap of positions 3-8, i.e. a cross-reactive T-cell overlap. In all cases, the left and right figures show the self/nonself overlaps determined using a scaled or fixed MHC binding threshold, respectively (see Methods). HLA molecules that have been described to have a GCpositive, GC-negative or GC-neutral preference [213] are colored green, red and black, respectively. HLA molecules with additional anchors (see Methods) are indicated with a plus-sign.

5.2.3 Self/nonself overlaps based on T-cell recognition

So far, we only considered identical self and nonself peptides as overlaps. However, also non-identical MHC-I presented peptides can be recognized by the same

T-cell [228]. This cross-reactivity is partly due to the fact that not all the residues on a presented peptide are accessible for the TCR. For example, most MHC-I molecules have two binding pockets that bind positions 2 and 9 (i.e. anchor-residues) of the presented peptide. These anchor-residues are hidden in the binding pocket of an MHC-I molecule, and are not exposed to the TCR [233]. Recently, we analyzed the T-cell recognition of the HIV-1 derived SLFNTVATL peptide presented on HLA-A*02 and suggested that not only the anchor-residues (P2 and P9), but also the first position (P1) of the presented peptide, hardly affects T-cell recognition [118]. Furthermore, at the remaining six middle positions (P3-8), some amino acid substitutions did not perturb T-cell recognition, especially those between amino acids with similar physical-chemical properties. TCR recognition was most stringent at the fifth position (P5), where only a Threonine-to-Serine substitution did not affect recognition [118].

To see if other TCR-pMHC contacts follow the same interaction-"rules", all non-redundant TCR-pMHC-I structures found in the PDB-database (www.pdb.org [234]) encompassing a 9mer (n=9, see Methods for selection criteria) were studied. In agreement with Frankild et al. [118], the majority of interactions in these structures involved the middle positions of the presented peptide (Figure 5.3). Several other reports on TCR-pMHC structures, and on different T-cell clones, confirm the degeneracy at the first position, and confirm that substitutions among similar amino acids are allowed in other positions [60, 206–211]. Our structural analysis suggests that the third position has less contacts with the TCR than the other middle positions (Figure 5.3). However, Tynan et al. [210] show examples in which position 3 is important for T-cell recognition. Therefore, we conservatively assume that the third position is as important for T-cell recognition as the other middle positions (P4-8).

Given these data, we studied how much of presented nonself can be discriminated from presented self by T-cells. First, the self/nonself overlaps were determined on those positions recognized by T-cells, i.e. the middle positions (P3-8) of MHC-I presented peptides. The self/nonself overlap of these 6mer fragments is on average 18 times higher than the overlap based on all positions (i.e., 2.7% for scaled thresholds and 1.7% for fixed thresholds see Figure 5.2B). This increase in the overlaps is mainly due to excluding the first position: if only both anchor positions are discarded, the overlap determined on the non-anchor positions (P1 and P3-8) remains low (i.e. 0.4% on average, see Table 5.1 and Supporting Figure 5.S2). Similarly, if only one of the anchor positions and position P1 are discarded, the overlap is much higher (Supporting Table 5.S3). We showed previously that highly specific anchor-positions of MHC molecules do not have to be exposed to the TCR to contribute to self/nonself discrimination because T-cells are MHC restricted [15]. For instance, HLA-A*0101 has a very specific preference for Tyrosine at the second anchor position (P9), and even if an HLA-A*0101 restricted T-cell is not interacting with this amino acid, all presented peptides it can possibly respond to must have a Tyrosine at position 9.

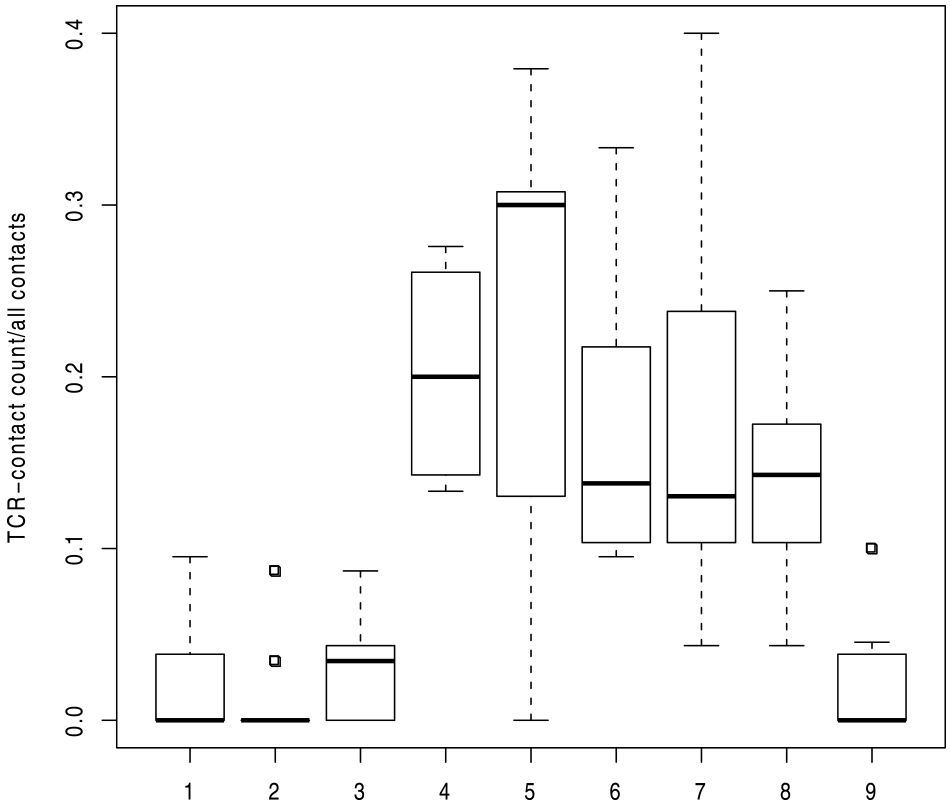


Figure 5.3. TCR interactions per peptide position. TCR contacts for 9 pMHC-TCR structures that have a 9mer (see Methods for details on selection and analysis criteria) were determined per position of the peptide. Per position the fraction of TCR-contacts relative to the total number of peptide- TCR contacts in a structure is shown. Positions 4-8 all have a significantly higher number of interactions than positions 1-3 and 9 have (Ranksums test: $p<0.005$).

Next, overall self/nonself overlaps were estimated with a novel model of degenerate T-cell binding. As above, T-cells were assumed to bind to the middle positions (P3-8) of the MHC-I presented peptides only. In addition, the degeneracy was modeled by considering two peptides as overlapping if they have mismatches in maximally two regions. We allow one mismatch at the N-terminal side of the fifth position (P1-4) and one at the C-terminal side of that position (P6-9) (see Methods). Moreover, only mismatches between amino acids having similar peptide-protein interaction properties were allowed, as such conservative substitutions have been shown to have a limited influence on T-cell recognition [118, 210, 211, 228]. The similarity between amino acids was derived from the PMBEC amino acid substitution matrix, that is based on peptide-MHC interac-

		Recognized peptide positions		
	Self percentage	P1-9 (complete)	P1 and P3-8 (non-anchor)	P3-8 (middle)
Exact	100	0.15%*	0.41%	2.7%*
	50	0.09%	0.25%	1.6%
Degenerate	100	0.7%	5.2%	29%*

Table 5.1. Summary of all the average self/nonself overlaps obtained using peptides predicted to be presented on HLA molecules. Overlaps were determined using all positions of the peptide (P1-9), the nonanchor positions (P1 and P3-8) or the middle positions between the anchors (P3-8). Further, overlaps were determined as exact, i.e. every position should be identical, or as degenerate, i.e. with 1 or 2 substitutions being allowed to mimic T-cell recognition (see Methods). Finally, overlaps with 100% or (a randomly chosen) 50% of the human proteome are shown. Self/nonself overlaps indicated with a star (*) are shown per HLA molecule in Figure 5.2.

tions and therefore specifically tailored to estimate the influence of amino acid substitutions on peptide-protein interactions [235]. We refer to this new overlap as the "degenerate" overlap. The degenerate self/nonself overlap is much higher than the identical overlaps of P3-8, on average 29% (see Figure 5.2C, left). These results can be ascribed to the degenerate nature of T-cell recognition: when using an alternative model of TCR recognition described by Frankild et al., the "peptide similarity score"-method (see Methods) [118], similarly high self/nonself overlaps were observed (results not shown). The self/nonself overlaps based on middle positions of the presented peptide (P3-8), determined using fixed binding thresholds were very similar to the overlap based on scaled thresholds (see Figures 5.2C, right), though more varied and somewhat lower. This is a result of the differences in the specificities of HLA molecules. The specificity determines the fraction of presented self and nonself peptides, which in turn influences the chance of finding a self/nonself overlap. One can explain this intuitively as the following: if an MHC molecule is very specific, it presents a small set of self peptides. For every presented nonself peptide, the chance of having an overlap with self would then become smaller. Therefore, there is a strong correlation between binding specificity and self/nonself overlaps (see Supporting Figure 5.S4). Furthermore, we tested the robustness of our results for various methods of peptide binding predictions, measures of amino acid similarity, and assumptions on T-cell recognition (summarized in Supporting Table 5.S3). In all cases did degenerate T-cell recognition lead to a high self/nonself overlap of $\sim 20 - 40\%$.

Despite the high overlaps, our assumptions on the degenerate T-cell recognition can be considered conservative. For example, position 3 of the presented peptide tends to have few interactions with the TCR (see Figure 5.3) and our model should probably allow more mismatches at this position. Furthermore, many peptides with more than two substitutions at the middle positions (P3-8) have been

shown to be cross-reactive [118]. If we assume that only a fraction of the self proteins provides a source of presented peptides, our estimates on self/nonself overlap decrease proportionally (see Table 5.1 and Supporting Table 5.S3). Cole et al. [229] recently showed that in some cases, the anchor residues are involved in T-cell recognition. This observation might be more of an exception rather than the general mode of T-cell recognition, as in most cases T-cell recognition has been described to be less specific and not influenced by the anchor residues [118, 206, 210, 228, 233]. Recent estimates on T-cell crossreactivity confirm that our model remains conservative. Ishizuka et al. tested the T-cell recognition of 30.000 unrelated MHC-I presented peptides using human and murine T-cell clones, and found a single cross-reactive response, which suggested a cross-reactivity level of 3.3×10^{-5} (1/30000) [57]. Typical T-cell precursor frequencies in a mouse are 1/100000 [54–56], i.e. on average 1 in a 100.000 T-cells are expected to recognize a particular pMHC, and 1 in a 100.000 pMHCs are expected to be recognized by a single T-cell clone. In other words, precursor frequency and cross-reactivity are similar concepts reflecting the specificity of a T-cell [100]. In our degenerate T-cell recognition model, single T-cells recognize only one in 2.7 million (3.7×10^{-7}) pMHCs (see Methods). Since this is much more specific than the experimental estimates, we think that our degenerate self/nonself overlap of about one-third is conservative and underestimates the actual overlap

5.2.4 Consequences of a high self/nonself overlap

Although these estimates on cross-reactive overlaps remain relatively crude, our results show that the degenerate recognition of MHC-I presented peptides by T-cells has a profound effect on self/nonself discrimination. This reconfirms that deletion of self reactive T-cells is important, as many of them would be activated during an infection and induce an autoimmune response. As a consequence, we estimate that about a third ($\sim 20-40\%$) of the foreign pMHCs is expected not to trigger an immune response. To test this prediction, the self/nonself overlap of HIV-1 derived peptides presented on HLA-A*0201 was studied to see if our model can account for the observed poor immunogenicity of these peptides. The presentation of, and T-cell responses to, HIV-1 derived peptides presented on HLA-A*0201 has been the subject of extensive investigations. Because it is such an intensively studied system, the lack of a reported T-cell response for one of the predicted pMHCs can be used as a reasonable indication for the lack of immunogenicity of that pMHC [118]. One explanation for the lack of immunogenicity is an overlap of the epitope with a self pMHC, and hence the self tolerance of the corresponding T-cell clone. We tested this by comparing overlaps of immunogenic and non-immunogenic HIV-1 pMHCs with self (see Methods). Only 4 of the 33 immunogenic pMHC (12%) were found to overlap with self according to our degenerate T-cell recognition model using the PMBEC similarity matrix. A significantly higher fraction of non-immunogenic pMHC, i.e. 18 of 54 (33%), over-

lapped with self (Chi-square test: p=0.027) (see Table 5.2), which is comparable to the overlaps reported by Frankild, using a different model for self-similarity but the same pMHCs[118]. We extended the analysis of self/nonself overlaps to vaccinia-derived peptides presented in HLA-A*02-transgenic mice for which Assarsson et al. [18] have determined the immunogenicity (see Methods). The overlap between (murine) self and immunogenic peptides is again lower than the self overlap of non-immunogenic peptides, although not significant due to the small number of data points (see Table 5.2).

	Immunogenic		Non-Immunogenic		Chi^2-test
	overlap	no overlap	overlap	no overlap	(p-value)
HIV-1 peptides on HLA-A*0201	4	29	18	36	0.027
vaccinia peptides on HLA-A*0201	3	15	8	18	0.29
HIV-1 peptides on non-HLA-A*0201 molecules	0	9	4	9	0.066
HLA-A*0201 pMHC from the IEDB	54	143	230	362	0.0038

Table 5.2. The self/nonself overlap of immunogenic versus non-immunogenic pMHCs. For immunogenic or non-immunogenic HIV-1 peptides presented on HLA-A*0201 determined by Frankild et al. [118], for immunogenic and non-immunogenic vaccinia derived peptides determined by Assarsson et al. [18], for immunogenic and non-immunogenic HIV-1 peptides on non-HLA-A*0201 determined by Perez et al. [131] and for immunogenic and non-immunogenic pMHCs sampled from the IEDB on HLA-A*0201 (see Methods for selection criteria applied to all four data sets), the presence of a self/nonself overlap was determined with the degenerate T-cell recognition model. For all sets of peptides, the immunogenic peptides have less overlaps with self, the significance of this association was tested using a Chi-square test, the p-value is reported in the last column.

These results are also valid for other HLA molecules: using data provided by Perez et al. [131] on non-HLA-A*0201 presented HIV-1 peptides we found the same trend, that immunogenic peptides have less self/nonself overlaps than their non-immungenic counterparts (see Table 5.2, and Methods). Finally, we analysed immunogenic/non-immunogenic pMHCs derived from the IEDB [185] that were presented on the same HLA molecule (see Methods for selection criteria). The number of immunogenic and non-immunogenic pMHCs was large enough only for HLA-A*0201, and therefore the self/nonself overlaps of these sets were compared. Again, we found significantly less self overlaps among immunogenic peptides than non-immunogenic ones (Chi-square test: $p < 0.01$; see Table 5.2). These results on the HLA-A*0201 presented HIV-1 and IEDB peptides are robust to the model assumptions: In all alternative overlap models described in Supporting Table 5.S3, the number of overlaps with self was smaller for immunogenic pMHCs than for non-immunogenic pMHCs. This difference was always significant

for the large set of IEDB peptides, for the smaller set of HIV-1 peptides a significant difference was not always observed (data not shown). Thus, in various data sets and model assumptions we find a correlation between pMHCs being immunogenic and their overlap with self, but these correlations only become significant for HLA-A*0201 where there is enough data. Summarizing, high self/nonself overlaps can explain the observed large "holes" in the T-cell repertoire [117, 227], and play an important role in determining the immunogenicity of foreign pMHCs.

5.3 Discussion

Previously, we have shown that the few epitopes sampled from a pathogens proteome are likely to be unique and are not expected to be present in the host (human) proteome [15]. Here, we extend this study by investigating a much larger set of nonself proteomes and a larger set of common HLA molecules. From this analysis we conclude that the pMHC of all common HLA-A and HLA-B molecules carry enough information for self/nonself discrimination, as a small minority (0.1% to 0.3%) of nonself derived peptides is expected to be identical to presented self-peptides. However, if the degenerate T-cell recognition of pMHCs is taken into account, the results change drastically. The cross-reactive recognition by T-cells results in a much higher self/nonself overlap of $\sim 20-40\%$ that is robust to various assumptions on degenerate T-cell recognition (see Supporting Table 5.S3), i.e. in the “eyes” of a T-cell, about a third of the epitopes is expected to be similar to a self peptide presented on the same MHC-I molecule. Such a large overlap is expected to have a strong effect on the immunogenicity of pathogen-derived epitopes.

One might intuitively think that the high self/nonself overlap estimates are in disagreement with the exquisite specifity of T-cell recognition. However, in our "degenerate" model of the middle positions (P3-8) with maximally 2 conservative mismatches, an individual T-cell recognizes only one in 2.7 million pMHCs. This level of specificity is much higher than experimental measurements of about one in 100.000 [54–57]. Therefore, we think that our current self/nonself overlap estimates are conservative.

Could longer peptides be a solution for the high self/nonself overlaps caused by degenerate T-cell recognition? Given that T-cells cannot use all the information that is present in an MHC-I presented 9mer, we do not expect that the presentation of longer peptides would make much difference. Even though a longer peptide would contain more information, if that is not detected by the T-cells it would not improve self/nonself discrimination. Alternatively, MHC binding could be more specific at for instance position 1, thus preserving self/nonself information as now happens at the anchor positions. The disadvantage of more specific binding motifs would be the reduced presentation of foreign peptides and more

opportunities for a virus to escape MHC presentation.

Another consequence of a high self/nonself overlap could be high risk of autoimmunity. The identification of self antigens targeted in autoimmune diseases remains an enormous challenge, and our method of identifying overlapping peptides could possible help to narrow the search for these auto antigens. This requires a thorough understanding of the pathogens that might trigger a particular autoimmune disease and the corresponding HLA risk factors. Unfortunately, only for few autoimmune diseases sufficient data is available to extract such associations. For instance, Epstein Barr virus and HLA-B*4402 are associated with multiple sclerosis [236, 237], and HTLV-1 and HLA-B*5401 are associated with HAM/TSP [113]. We are currently searching the overlaps between the presented peptides of these viruses and the human self peptides presented on these HLA molecules for potential CTL targets in these autoimmune diseases (work in progress).

The predicted self/nonself overlap varies between HLA molecules (see Figure 5.2), and two factors explain most of this variation. First, some HLA molecules have a preference for peptides derived from organisms with a low G+C content [213], which seems to be a universal signature for pathogenicity [149]. HLA molecules with such a preference for presenting nonself (e.g. HLA-A*2301) have a lower self/nonself overlap than other HLA molecules, because they present peptides that are less likely to occur in the human proteome. Second, the usage of additional (auxiliary or atypical) anchors at positions that also interact with the TCR increases the chance that presented peptides overlap according to our model. For example, HLA-B*0801 with atypical anchors at the third and fifth position will present more peptides that overlap at position three and five, and has the highest estimated self/nonself overlap (see Figure 5.2C). Indeed, a strong correlation between the use of additional anchors (see Methods) and self/nonself overlaps is found (Spearman Rank test: correlation=0.88, $p < 0.001$, not shown). Possibly, peptides presented on HLA-B*0801 have more specific TCR-interactions at the conventional anchor positions (P2 and P9) than in our T-cell recognition model, leading to an overestimate of the self/nonself overlap for this HLA molecule and others with atypical anchors. If the degenerate self/nonself overlap is not based on the middle positions of the presented peptide (P3-8), but on an HLA molecule specific choice of the six least specific positions (see Methods), the overlaps are however very comparable to an overlap based on the middle positions (see Supporting Table 5.S3).

Our estimates on self/nonself overlaps can explain why MHC-I restricted cellular immune responses to a pathogen are more narrow than the (predicted) number of pMHCs for that organism [18, 118]. We show that about one-third of the nonself pMHC should not elicit T-cell responses because they overlap with a self pMHC, i.e. this explains the large "holes" found in the T-cell repertoire [117, 118, 227]. We validated this prediction by comparing the over-

laps of immunogenic and non-immunogenic pMHC from HIV-1, vaccinia or the IEDB, and showed that the number of self overlaps is significantly higher for non-immunogenic pMHC than for immunogenic pMHC. Still, a fraction of the immunogenic pMHCs were predicted to be overlapping with self, possibly because not all self-proteins induce tolerance or because regulatory processes are overridden during some viral infections causing autoimmunity [77]. In addition, an improved understanding of the rules of T-cell recognition could result in an even better distinction between overlapping/non-overlapping, and non-immunogenic/immunogenic pMHCs. This would be important in vaccine design and the understanding of immunogenicity in cellular immune responses.

5.4 Methods

5.4.1 Proteome data collection

Human, murine, viral and bacterial proteomes were downloaded via http://www.ebi.ac.uk, the human proteome in May 2008, bacterial and viral proteomes in October 2008 and the Mouse proteome in January 2011. Only human and mouse proteins that have been shown at the protein or transcript level were included in the "self" data set. Redundant bacterial proteomes were removed by selecting only one strain per species, which resulted in 174 species of bacteria. 1000 non-redundant viral proteomes were selected with a maximum similarity of 80%. The similarity between viruses was determined as the number of exact matches in an all-to-all alignment of proteome sequences using BLASTP 2.2.18 relative to the smallest virus. Human viruses were selected based on the reported host information in the downloaded proteome, or on the term 'human' in their species name (e.g. Human Immunodeficiency Virus). A list of all bacteria and viruses used in this study is available upon request.

5.4.2 MHC-I presentation predictions

The peptides presented on a certain MHC-I molecule can be predicted by simulating three key-processes of MHC-I presentation, i.e. proteasomal cleavage, TAP transport and peptide-MHC-I binding. The combination of proteasomal cleavage and TAP-transport determines which peptides reach the ER to potentially bind MHC-I. This process was predicted using NetChop Cterm3.0 [16, 51]. Peptide-MHC-I binding was predicted using NetMHC-3.2, an improved version of NetMHC-3.0, that was shown to perform best in a large benchmark study of Peters et al. [50, 147]. The fraction of nonself peptides that overlap with a self peptide presented on an MHC-I molecule depends on the number of self peptides

that is predicted to bind to this MHC-I molecule. Because we want to compare the self/nonself overlap of different MHC-I molecules, we have chosen to exclude the variance in the number of presented self peptides by using scaled thresholds, i.e., the number of self peptides predicted to bind to each MHC molecules is scaled to be similar. Unfortunately, this procedure will eliminate the variation as a result of possible differences in specificity among MHC molecules. For each MHC molecule the threshold was set such that the presented fraction of self was similar to that on HLA-A*0201 with a 500 nM threshold (2.3%) [46, 47]. This results in on average 250.492 self pMHCs, 3.750.428 bacterial and 196.265 viral pMHCs, per HLA molecule. Alternatively, we repeated the analysis with a fixed threshold of 500 nM (see supporting figure 2 and supporting Table 5.S3). In order to exclude HLA molecules with too similar binding motifs from our analysis, we selected the most frequent HLA molecule available in NetMHC-3.2 at two digit resolution. This resulted in a set of 13 HLA-A and 15 HLA-B molecules.
All results were checked for consistency with two other MHC-I binding prediction methods, NetMHCpan-2 [49] and a Stabilized Matrix Method (SMM)-based MHC-binding prediction tool [163], for HLA-A*0101, HLA-A*0201, HLA-A*0301, HLA-B*0702, HLA-B*0801 and HLA-B*3501. Note that for the HLA molecules that we have included in our analysis the average AUC for NetMHC and NetMHCpan predictions is 0.809 and 0.812, respectively [238]. As expected, similar results were obtained with NetMHCpan, but also when using SMMs (supporting Table 5.S3).

5.4.3 Self/nonself overlap estimations

Per MHC-I molecule, the set of presented 9mers derived from viral or bacterial (nonself) proteomes and that from the human (self) proteome were compared to see how much these sets overlap. In the self/nonself overlap determination for vaccinia-derived pMHC from Assarsson et al. [18], the Mouse proteome was used as self. Overlaps were determined in different ways. First a “complete overlap” was determined as the exact match of all positions of the 9mer (positions 1-9, as in Figure 5.2A). Second, a “middle positions 6mer overlap” was defined as an exact match of the amino acids at positions 3-8 (as in Figure 5.2B). Third, the “non-anchor 7mer overlap” was determined as the exact match of the amino acids at position 1 and 3-8 (as in Supporting Figure 5.S2). Finally, a “degenerate overlap” was determined by allowing two amino acid mismatches. Amino acid mismatches were not allowed at the most specifically recognized position 5. Moreover, we reasoned that two amino acid substitutions closeby would be more likely to abolish T-cell recognition. Therefore, only a single mismatch was allowed at the positions N-terminal from position 5 (P1-P4) and at the positions C-terminal (P6-P9) from position 5. Finally, only mismatches between amino acids with similar peptide-protein interaction properties were allowed. Following Kim et al., amino acids were considered similar if their absolute covariance was

greater than 0.05 in the PMBEC matrix [235]. The PMBEC matrix is based on measured binding affinities between peptides libraries and MHC-I molecules, and was shown to capture similarity features common to substitution matrices such as BLOSUM50, and outperform other matrices when used as a Bayesian prior in MHC-I binding predictor training [235]. Furthermore, repeating our analysis using a positive score in the BLOSUM62 or BLOSUM50 matrix to identify allowed mismatches, similar results were found (Supporting Table 5.S3). The self/nonself overlap is the chance a nonself pMHC overlaps with self, and was calculated by dividing the total number of overlaps in all nonself proteomes by the total number of pMHCs in all nonself proteomes. The self/nonself overlap was determined for bacteria and viruses separately, and the average of these two self/nonself overlaps is presented throughout the paper.
Additionally, self/nonself overlaps were estimated using the "peptide similarity score"-method described in detail by Frankild et al. [118]. In this method the similarity between two peptides is determined using the BLOSUM35 amino acid substitution matrix and all positions of the compared peptides. The similarity score is subsequently scaled to the minimal and maximal similarity scores for the reference peptide, in order to normalize for the intrinsic similarity that a certain peptide has to all other peptides. If for instance the BLOSUM35 similarity score between peptide A and peptide B is 3, and the minimum and maximum possible similarities for any peptide with peptide A are 1 and 11, respectively, the peptide similarity score is $(3-1)/(11-1) = 0.2$ (see [118] for a full description of the method). Frankild et al. showed that a self similarity score of 0.85 tends to separate too self-similar, and hence non-immunogenic, from immunogenic HIV-epitopes [118]. This analysis and an analysis of cross-reactive peptides from literature was used for verification of this method [118]. We used the same threshold when determining overlaps with this "peptide similarity score"-method, i.e. nonself peptides with a similarity score exceeding 0.85 with a self peptide are considered as overlapping.

5.4.4 Cross-reactivity

The cross-reactivity in our degenerate overlap model of T-cell recognition (described above) was determined in order to compare it with experimentally determined levels. For every possible 9mer peptide, the number of variants at the T-cell recognized middle positions (P3-8) was determined that would be recognized by the same T-cell in our degenerate overlap model. In other words, for every combination of amino acids at P3-8 we performed an exhaustive search to determine how many other combinations would also be recognized. On average, 24 of such combinations were found. Thus, given the number of possible variants at positions P3-8 (20^6), the cross-reactivity in our model is $24/(20^6)$, which is 1 in 2.7 million or 3.8×10^{-7}.

5.4.5 Immunogenic/non-immunogenic pMHCs

Four sets of pMHCs were obtained for which the immunogenicity had been determined previously. The first set of HIV-1 derived peptides presented on HLA-A02 was determined by Frankild et al. [118], who predicted which HIV-1 peptides were presented on HLA-A02 and then defined the ones as immunogenice if there was at least one report of a T-cell response in a patient in the Los Alamos Database. Because HIV-1 responses for the most frequent HLA-A*02 molecule are studied extensively, we defined all other peptides as non-immunogenic. Thus, 33 immunogenic and 54 non-immunogenic HIV-1 derived peptides were defined using this strategy. The second set is derived from Assarsson et al. [18], who tested the immunogenicity of vaccinia derived peptides in a humanized mouse-system expressing HLA-A*02. We classified the 9mers shown to be naturally processed and immunogenic (termed “Dominant” and “Subdominant”) as immunogenic peptides, and non-immunogenic peptides (termed “Negative”) were classified as such. This resulted in the selection of 18 immunogenic and 26 non-immunogenic vaccinia derived peptides. The third data set is derived from Perez et al [131], who measured the T-cell response in HIV-1 patients to a set of HIV-1 peptides. The patients were HLA class I genotyped [131]. We only considered responses to 9mer peptides with a predicted binding affinity of less than 500nM, to only one of the patients HLA-A and HLA-B molecules. Binding predictions were done with NetMHCpan-2 [49]. The virus in every patient was sequenced by Perez et al. [131], and we excluded all T-cell responses in which the peptide that was used for testing the T-cell response was not encoded by the viral genome. Only peptides presented on HLA molecules other than HLA-A*0201 were selected since HLA-A*0201 presented HIV-1 peptides were already compared in the data set derived from Frankild et al [118]. Peptide-HLA combinations with only negative T-cell responses measured by Perez et al. were classified as non-immunogenic (n=13), all other peptide-HLA combinations were classified as immunogenic (n=9). The fourth data set was derived from the IEDB [185], by downloading all entries that describe a T-cell response assay to a 9mer peptide presented on one of the HLA molecules in our test set, performed in a human subject upon infection. Only peptide-HLA combinations in which the predicted binding affinity was less than 500 nM were considered. Furthermore, we required that the assayed T-cells were not restimulated in vitro, and that the peptide was used in the T-cell response assay. Peptide-HLA combinations were classified as immunogenic if a "Positive(-High)" or "Positive-Low" T-cell response was measured, and classified as non-immunogenic if the T-cell response was always reported to be "negative". We were able to classify more than 20 immunogenic and 20 non-immunogenic peptides only for HLA-A*0201 (i.e. 197 immunogenic and 592 non-immunogenic peptides).

5.4.6 Additional anchor selectivity

For all HLA molecules, we predicted the binding of 1.000.000 random peptides with equal amino acid frequencies using NetMHC-3.2 and the thresholds described above. The Shannon entropy was determined per position on the predicted binders, per HLA molecule, and used as a measure of selectivity. Based on this selectivity, the six least specific positions were determined for each HLA molecule to use in the "allele specific" analysis of degenerate self/nonself overlaps (Supporting Table 5.S3). Additional anchor selectivity was calculated as the sum of the entropy at the non-anchor positions (P1 and P3-8), per HLA molecule. An HLA molecule was defined to have additional anchors if the additional anchor selectivity was larger than 25% of the sum of entropy at all positions (P1-9) for an HLA molecule.

5.4.7 Analyzing TCR-pMHC structures

Structures of HLA-I-9mer-TCR-complexes were downloaded in August 2011 from the PDB-database (www.pdb.org [234]). After redundancy reduction we selected nine structures for further analysis: 1AO7, 1BD2, 1LP9, 1MI5, 2ESV, 3GSN, 3KPR, 3O4L and 2F53 [110, 239–246]. The selected structures consist of HLA-A*02 (n=6), HLA-B*08, HLA-B*44 and HLA-E molecules. Per peptide position the number of TCR contacts was determined as the number of TCR amino acids within a 5.0 Å distance. For each structure, we determined per peptide position the fraction of TCR contacts relative to all peptide-TCR contacts in that structure. Boxplots of these fractions are shown in Figure 5.3.

5.4.8 Statistics

Statistical tests were performed using the stats-package from the scipy-module in Python. A Permutation test was also done in Python, using the shuffle function in the random-package from the numpy-module, to identify human proteins that have more than expected peptides that overlap with viruses or bacteria. The permutation test was performed as follows: per human protein, we counted the number of viruses or bacteria that overlap with a 9mer peptide in this protein. These counts were normalized by the length of the protein, i.e. the number of overlapping viruses or bacteria was divided by the protein length. In 1000 permutations, per human protein a number of overlapping viruses or bacteria was drawn based on the expected fraction of overlaps and given the protein length. If the actual number of overlaps was higher than the number in all 1000 permutations, the human protein was selected as a protein with a significantly high number of viral or bacterial overlaps.

A similar analysis was performed to identify proteins with more than expected HLA-B*5401 ligands. First, per protein the number of HLA-B*5401 binding peptides was predicted as described above. Next, this prediction was compared in 1000 permutations where a number of binding peptides was drawn based on the specificity of HLA-B*5401 (i.e. 2.3% as described above). If the actual number of binding peptides was higher than the number in all 1000 permutations, the protein was selected as a protein with a significantly high number HLA-B*5401 ligands.

5.5 Acknowledgments

We thank Johannes Textor for valuable comments on the manuscript and discussion on this research project, and Hanneke van Deutekom, Xiangyu Rao and Ilka Hoof for discussion and technical support.

5.6 Supporting Information

Human proteins overlapping with bacteria (n=2055; 1726 found in DAVID)		
Cluster	Description	Number of proteins
1	nucleotide binding	525
2	mitochondrion	276
3	ATPase, AAA+ type, core	95
4	ATPase activity	137
5	membrane-enclosed lumen	281
6	NAD(P)-binding domain	72
7	monosaccharide metabolic process	62
8	magnesium ion binding	100
9	fatty acid metabolic process	54
10	mitochondrial part	160
Human proteins overlapping with viruses (n=722; 590 found in DAVID)		
Cluster	Description	Number of proteins
1	nucleus	299
2	transcription factor activity	116
3	membrane-enclosed lumen	112
4	regulation of transcription from RNA polymerase II promoter	85
5	RNA binding	58
6	pattern specification process	35
7	negative regulation of macromolecule metabolic process	51
8	non-membrane-bounded organelle	95
9	transcription factor binding	48
10	compositionally biased region: Arg/Ser-rich (RS domain)	10

Table 5.S1. Human proteins that overlap with more than expected bacteria and viruses. Human proteins with peptides that cause them to overlap with a significantly large number of viruses or bacteria at the 9mer level were analyzed using the online annotation analyzer DAVID [230, 231]. For the 10 most enriched non-redundant annotation clusters, the category encompassing most proteins is shown. All categories were significantly enriched ($p<10^{-4}$).

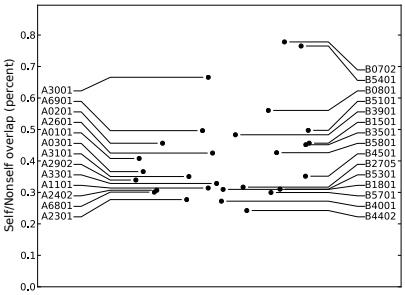


Figure 5.S2. Self/nonself overlaps of different HLA molecules. Overlaps were determined as an exact overlap of the non-anchor positions (P1 and P3-8) of HLA presented peptides.

		Recognized peptide positions					
	Self	P1-9 (complete)	P1 and P3-8 (non-anchor)	P3-8 (middle)	allele specific	P3-9	P2-8
Exact	100%	0.15%	0.41%	2.7%	2.5%	0.92%	0.8%
	50%	0.085%	0.24%	1.6%	1.5%	0.54%	0.48%
Degenerate (PMBEC)	100%	0.71%	5.2%	29%	28%	12%	10%
	50%	0.44%	3.2%	19%	19%	7.8%	6.6%
Degenerate (Blosum 62)	100%	0.86%	6.8%	36%	35%	16%	13%
	50%	0.52%	4.1%	25%	24%	10%	8.5%
Degenerate (Blosum 50)	100%	0.98%	7.9%	40%	40%	18%	15%
	50%	0.59%	4.9%	28%	28%	12%	9.8%
NetMHC-pan[1,3] (PMBEC)	100%	0.85%	7%	36%	35%	16%	13%
	50%	0.52%	4.3%	26%	25%	10%	8.4%
SMM[2,3] (PMBEC)	100%	0.89%	7.9%	38%	39%	16%	14%
	50%	0.55%	4.9%	27%	28%	10%	8.9%
fixed[4] (PMBEC)	100%	0.63%	3.5%	18%	18%	8.9%	7.4%
	50%	0.4%	2.1%	12%	12%	5.6%	4.6%
Fully degene-rate[5] (PMBEC)	100%	0.94%	7.1%	35%	35%	17%	14%
	50%	0.57%	4.3%	24%	24%	11%	8.7%

Table 5.S3. Degenerate T-cell recognition leads to high self/nonself overlaps under various conditions. The self/nonself overlap was determined for the HLA molecules in our set (see Methods) and the average of the set is shown per cell. In the six columns on the right, the positions are shown on which the overlap is based, in the "allele specific" case the 6 least specific positions (see Methods) were selected for every HLA molecule, to allow for a-typical anchors in other positions. Overlaps were determined as "exact", i.e. every position should be identical, or as degenerate (all other columns), i.e. with 1 or 2 substitutions being allowed to mimic the degeneracy of T-cell recognition (see Methods). The matrix that was used for determining amino acid similarity is shown in brackets. Overlaps with 100% or (a randomly chosen) 50% of the human proteome are shown in different rows. [1]NetMHCpan-2 predictions (see Methods). [2]SMM binding predictions (see Methods). [3]The analysis was done only for HLA-A*0101, HLA-A*0201, HLA-A*0301, HLA-B*0702, HLA-B*0801 and HLA-B*3501. [4]Using a fixed binding threshold of 500nM instead of a scaled threshold. [5]Amino acid substitutions were allowed next to each other.

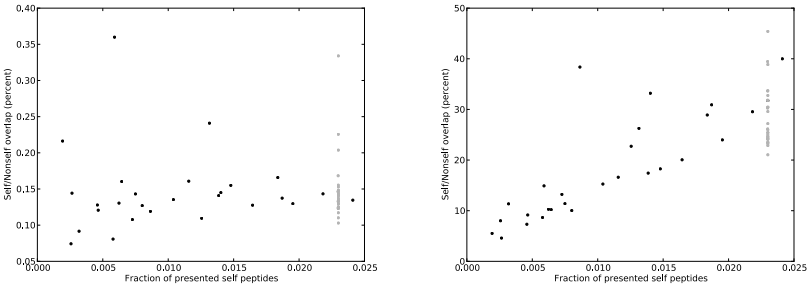


Figure 5.S4. The self/nonself overlap of identical and non-identical overlaps versus the binding specificity. The precise overlap of all peptide positions (P1-9, left figure, y-axis), and the degenerate overlap of the T-cell recognized middle positions (P3-8, right figure, y-axis), as well as the the fraction of presented self peptides (both figures, x-axis) for each HLA molecule. The overlap and binding fraction were determined for every HLA molecule using scaled (in grey) and fixed (in black) binding thresholds. As discussed in the main text, a larger number of presented self peptides will lead to a larger chance of finding a self/nonself overlap. However, this does not hold if the self and nonself peptides are required to be identical to overlap (left figure), in which case the binding affinities of the self and nonself peptide are the same, and the chance of having an overlap with self depends solely on the presence of that peptide in the self proteome. Since the overlap is based on presented nonself peptides, if the self peptide is present it must be presented given the identical binding affinities. The correlations of overlap versus binding specificity illustrate this difference between identical and non-identical overlaps, data points obtained under the fixed threshold (in black) were used in a Spearman Rank test. In the left figure where an identical overlap is shown the correlation is absent (correlation=0.25, p=0.20), in the right figure a strong correlation between the overlap and binding specificity is observed (correlation=0.89, $p<0.001$).

I

Addendum to Chapter 5: Quantifying the constraints that a large self/nonself overlap puts on T-cell responses

JORG J.A. CALIS*, ROB J. DE BOER* AND CAN KEŞMIR* (2012)

*Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, The Netherlands

Addendum to Chapter 5

Abstract

We recently estimated that due to degenerate T-cell recognition, about 30% of the foreign peptides presented on an MHC-I molecule (pMHCs) is overlapping with a self pMHC. Here, the effect of this overlap on the T-cell response to a pathogen was studied by modeling the chance of mounting an immune response in the absence of an auto-immune response. We show that to prevent auto-immunity, the efficiency of self tolerance induction needs to be near complete, self tolerance should be elicited for at least 97% of the self pMHCs. Thus, the self/nonself overlap puts a large constraint on the possibility of errors in the self tolerance induction process.

I.1 Introduction

Pathogen-derived peptides presented on MHC-I molecules (pMHCs) can elicit T-cell immune responses. These immune responses are very specific, as T-cell receptors (TCRs) specifically recognize only 1 in ∼200,000 pMHCs [54–57]. Despite the high level of specificity that characterizes T-cells, we recently showed that ∼30% of the foreign pMHCs resembles a self pMHC, we refer to such self-resembling foreign pMHCs as self-overlapping, and we refer to the fraction of self-overlapping foreign pMHCs as the self/nonself overlap [203].

Previously, Borghans et al. studied the impact of T-cell specificity, by modeling the cellular immune response towards pathogens and the chance to develop autoimmune responses [111]. Inspired by this work, we draw a new model that includes our recent estimate that the self/nonself overlap is ∼30% [203]. To prevent an autoimmune response, T-cells should not respond to self-overlapping pMHCs by a process called self tolerance. We show that self/nonself overlaps of 30% dramaticly increase the risk of auto-immunity, unless self tolerance mechanisms are near-perfect. We quantify the level of self tolerance that is required to be able to elicit a protective immune response, and show that self tolerance should be elicited for at least 97% of the self pMHCs.

In the model we present here, immune responses are only evoked by pathogen-derived pMHCs. If these pathogen-derived pMHCs are self-overlapping, an auto-immune response is induced. The auto-immunity that might be caused by stimulating naive T-cells directly with self pMHCs is not taken into account. In other words, we study the constraints that the large self/nonself overlap puts on the immune system when a pathogen-specific immune response is elicited, but overall constraints on self tolerance can be even more strict.

I.2 Modeling independent effector T-cells

We aim to model the chance of a protective immune response (P_p) to a pathogen with n pathogen-derived pMHCs. We define a protective response as one in which at least one of the pathogen-derived pMHCs elicits an immune response (P_r), while an auto-immune response due to the immune targeting of a self-overlapping pMHC (P_a) should not be elicited. Here, we only model the auto-immunity that is caused by a cross-reactive immune response to a pathogen-derived pMHC, not the auto-immunity that might be caused by a direct immune response to a self pMHC. Inspired by the modeling work from Borghans et al. [111], we first determine per pathogen-derived pMHC the chance of an immune response (P_r) and an auto-

immune response (P_a):

$$P_r = P_i \times (1 - P_t \times P_o) \tag{I.1}$$

$$P_a = P_i \times P_o \times (1 - P_t) \tag{I.2}$$

Where the chance of an immune response is calculated as the chance that the pMHC is immunogenic (i.e. recognizing T-cells exist; P_i), and is not prevented by an overlap with a self pMHC (P_o) for which self tolerance (P_t) exists ($1 - P_t \times P_o$). The chance of an auto-immune response is calculated as the chance that the pMHC is immunogenic (P_i), and overlapping with a self pMHC (P_o) for which no self tolerance exists ($1 - P_t$). Based on the per epitope chances of an immune response (P_r, eq. I.1) and an auto-immune response (P_a, eq. I.2), the following equation was derived to calculate P_p:

$$P_p = (1 - P_a)^n - (1 - P_r)^n \tag{I.3}$$

Where the chance of a protective immune response to a pathogen with n pathogen-derived pMHCs (P_p, eq. I.3) is modeled as the chance of no auto-immune response ($(1 - P_a)^n$) minus the chance of no immune response ($(1 - P_r)^n$), these chances can be subtracted as the auto-immune responses are a subset of all immune responses.

Based on our recent estimate that the self/nonself overlap is 30% [203], we fix $P_o = 0.3$. To estimate the chance that a pMHC is immunogenic (P_i), we rewrite eq. I.1 as:

$$P_i = \frac{P_r}{1 - P_t \times P_o} \tag{I.4}$$

Assarsson et al. reported that ∼50% of the foreign pMHCs elicit an immune response ($P_r = 0.5$) [18]. If we make the assumption that self tolerance is close to perfect ($P_t \approx 1.0$) and that the self/nonself overlap is 30% ($P_o = 0.3$), we can derive P_i by filling in eq. I.4: $P_i = \frac{0.5}{1-0.3} = 0.7$.

Next, we analyzed the chance of a protective immune response (P_p) for a wide range of possible foreign pMHCs ($n = [1..100]$) and all possible degrees of self tolerance efficiency ($0 \leq P_t \leq 1$; Figure I.1). As expected, if self tolerance is highly efficient, the chance of an auto-immune response is reduced and a protective immune response is more likely. As both the chance of an immune response and the chance of auto-immunity increase with the number of pathogen-derived pMHCs, there is an intermediate number of pMHCs at which the chance of a protective immune response is highest and the requirements for self tolerance are smallest. At this optimal number of foreign pMHCs (n=6.3, see Figure I.1), the chance of a protective immune response is fairly large ($P_p > 0.95$), when the fraction of self pMHCs for which self tolerance is induced is higher than 97% (Figure I.1, shown in blue). Thus, at least 97% of the self pMHCs should be used in the mechanisms that induce self tolerance when pathogens with ∼6 pMHCs are considered, for

other pathogens with another number of pMHCs the requirements on self tolerance for a protective immune response are even more strict.

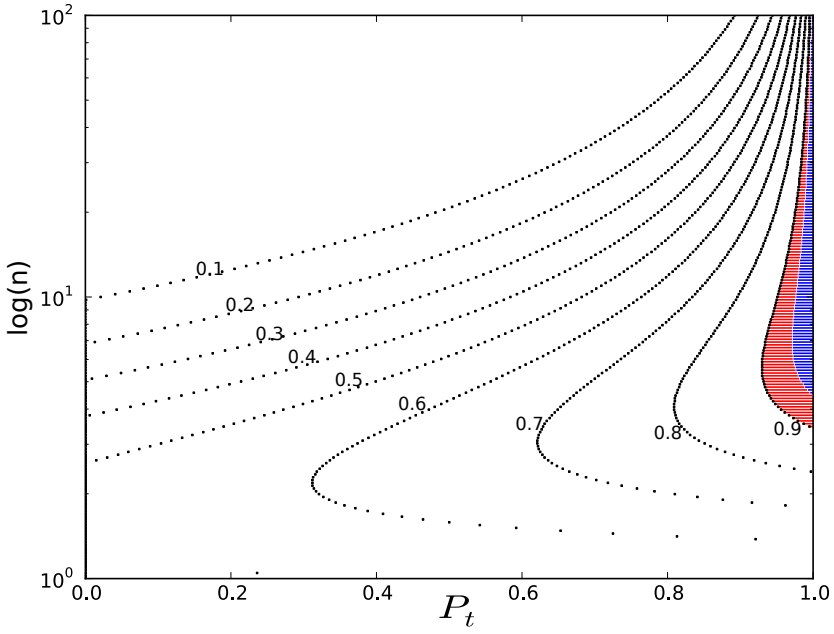


Figure I.1. Limits to a protective immune response. The chance of a protective immune response (P_p) is shown for a range of possible foreign pMHCs ($n = [1..100]$) and all possible efficiencies of self tolerance ($0 \leq P_t \leq 1$). Blue coloring corresponds to n, P_t-combinations where $P_p > 0.95$, in red are n, P_t-combinations where $P_p > 0.90$, dotted lines represent 0.1-intervals of P_p.

I.3 Discussion

Here we present an analysis of the CD8$^+$ T-cell immune response, where we implement the recently estimated large self/nonself overlap of ∼30% [203]. We show that in such a model, the host is at a high risk of auto-immunity, unless self tolerance control is extremely efficient.

Several mechanisms exist to enable self tolerance, three of these induce self tolerance before an effector T-cell immune response is set up. First, T-cells undergo apoptosis if they recognize self pMHCs that are presented during negative selec-

tion on medullary thymic epithelial cells (mTECs) [63]. Second, T-cells can be deleted in the periphery if they recognize self pMHCs [64, 65]. Third, regulatory T-cells (T_{reg}s) can inhibit the development of specific immune responses, by influencing Dendritic Cells (DCs) that stimulate naive T-cells in the lymph node (LN) to become effector T-cells [116, 247]. In our model, the combined effect of these three mechanisms is modeled as the fraction of self pMHCs for which self tolerance exists (P_t). As the immune system cannot a priori know which self pMHCs are overlapping with foreign pMHCs, we think that our analysis shows that at least 97% of the self pMHCs should be used in the self tolerance mechanisms.

Certain HLA backgrounds or pathogen infections have been associated with specific auto-immune diseases [77]. Even though the auto-antigen is often unknown, these associations suggest that certain pathogens trigger an immune response that becomes self-reactive. A well-studied auto-immune disease is type 1 diabetes (T1D); where $CD8^+$ T-cells play an important role [80], and HLA class I molecules such as HLA-B*39 and HLA-B*18 have been reported to be associated with T1D [78]. Myelopathy (HAM/TSP) is a well-described example of an auto-immune disease that is triggered by Human T-lymphotropic virus 1 [113]. That viral infections can trigger autoimmunity was shown by Oldstone et al., whom studied a transgenic mouse model in which a viral protein was stably expressed in Langerhans cells, thus the viral protein mimicked a self-antigen [248]. Upon infection with a virus that expressed the same protein, an autoimmune response was elicited that caused type 1 diabetes [248]. The auto-immune diseases and the model from Oldstone et al. show that self-reactive T-cells are part of the repertoire, despite the self tolerance induction efficiency of more than 97% that we require for a protective immune response in our model.

In our model, effector T-cells are independent, i.e. once activated they could respond to every cell that expresses a pMHC they recognize, even if it is a self pMHC that is expressed in a non-infected tissue. However, effector T-cells have been described to be not that independent, as they require help to be recruited to non-lymphoid tissues [65]. For instance, Nakanishi et al. showed that effector $CD8^+$ T-cells require the help from $CD4^+$ T-cells to enter reproductive organ tissue and elicit their effector function [119]. The $CD4^+$ T-cell help is dependent on cross-presentation by local tissue DCs, in addition cross-presentation by local endothelial cells seems to help the tissue entry of effector T-cells [65]. Access to other tissues is even more restricted, e.g. T-cells can be killed if they enter the central nervous system [249]. Such restrictions or regulation of effector T-cells might be a fourth mechanism to enable self tolerance, though they require a model in which effector T-cells are no longer independent.

The chance of an auto-immune response increases with the number of foreign pMHCs, as more immune responses are set up that have the potential to become an auto-immune response. However, the competition between T-cells can limit the number of responses, even if more foreign pMHCs are presented [120, 121].

Nevertheless, the immune targeting of twenty or more pMHCs is observed in HIV-1 patients [131]. This might be attributed to viral escapes from immune responses, and subsequent possibility for new immune responses. Such an effect would be especially dangerous in case of a chronic infection, where every new immune response to an escape variant can cause auto-immunity. In addition, the immune system is faces multiple pathogens from which many different pMHCs are presented that all require an adequate immune response.

We have fixed the chance of a pMHC to be immunogenic (P_i) at 0.7, based on the observation that 50% of the foreign peptides can elicit an immune response in a peptide immunization experiment [18], and the assumption that 30% of all peptides is self-overlapping and therefore non-immunogenic ($P_o = 0.3$ and $P_t = 1.0$). However, it has been shown that some autoreactive T-cells are not deleted, but adopt a regulatory function or down regulate TCR expression [93, 116]. If such autoreactive T-cells could be forced into an immune response by the peptide immunization protocol, not all self-overlapping foreign pMHCs would be non-immunogenic (i.e. $P_t < 1$). In the most extreme case (i.e. $P_t = 0$), this would mean $P_i = \frac{0.5}{1-0} = 0.5$ (see eq. I.4). When we fix P_i at 0.5 instead of 0.7, the parameter regimes in which good protection is possible ($P_p > 0.95$) were similar (not shown).

One could suggest that self tolerance should be perfect as self pMHCs are presented by DCs when they stimulate naive T-cells. In contrast, DCs might exclusively present pathogen-derived pMHCs, special antigen storage depots in DCs could facilitate such exclusivity [250]. Here, we only study the response towards pathogen-derived pMHCs, for which the presentation on DCs and stimulation of naive T-cells is certain. We show that the response to these pMHCs requires a high efficiency of self tolerance to prevent auto-immunity, as many foreign pMHCs are self-overlapping.

Chapter 6

De novo development of donor-specific HLA IgG antibodies after kidney transplantation is facilitated by donor HLA-derived T-helper epitopes

HENNY G. OTTEN[†,*], JORG J.A. CALIS[‡,*], CAN KEŞMIR[‡], ARJAN D. VAN ZUILEN[¶], AND ERIC SPIERINGS[†] (2012)

[†]Department of Immunology, University Medical Center Utrecht, Utrecht, The Netherlands [‡]Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, The Netherlands [¶]Department of Nephrology, University Medical Center Utrecht, Utrecht, The Netherlands
[*]Contributed equally

Submitted

Abstract

Background: The de novo development of donor-specific IgG HLA antibodies may be dependent on the presence of T-helper epitopes within the recognized HLA antigens.
Methods: The correlation between antibody production against mismatched donor human leukocyte antigens (HLA) class I and the number of predicted T-helper epitopes in the respective HLA class-I mismatches was investigated. To this end, we analyzed a cohort of 21 nonimmunized individuals that received and rejected a renal transplant with HLA-mismatches which were all immunogenic according to HLAMatchmaker.
Results: Mismatched HLA class I antigens for which donor-specific antibodies were detectable after transplantation, contained a larger number of T-helper epitopes than those against which no antibodies were raised. Most T-helper epitopes (more than 60%) were not part of an eplet.
Conclusions: Our data suggest that cross presentation of donor-derived HLA by recipient HLA class-II molecules may be an important mechanism in IgM-to-IgG isotype switching of donorspecific HLA antibodies.

6.1 Introduction

Matching for human leukocyte antigens (HLA) significantly improves the outcome of kidney transplantation (reviewed in [251]). However, as a result of the high level of polymorphism of the various HLA loci and the limited number of donors, HLA mismatches between donor and recipient exist in approximately 85% of cadaveric kidney transplantations (Eurotransplant database; http://www.eurotransplant.org, accessed April 24, 2012). Evidently, these HLA mismatches frequently lead to production of HLA-specific antibodies, which shorten graft survival [252].

In order to prevent antibody formation against HLA, the optimal kidney grafts are either HLA identical to the recipient, or express acceptable HLA mismatches which do not induce antibody formation. To a certain extent, these acceptable mismatches can be identified with the HLAMatchmaker algorithm, which was initially based upon linear epitopes consisting of 3 adjacent amino acids, and currently upon eplets defined by the quaternary structure of the HLA protein [253, 254]. Using the HLA typing of the recipient, this structural algorithm splits the HLA molecules into conformational sets of amino-acid residues on alloantibody-accessible sites of HLA molecules. As long as all donor HLA eplets are covered by recipient HLA eplets, no antibody responses are to be expected.

HLAMatchmaker identifies HLA-antigens unlikely to be recognized on a subsequent kidney graft [255]. The likelihood of developing HLA antibodies in combination with graft rejection seems to be strongly correlated with the number of triplet mismatches between the HLA alleles of the donor and the recipient [256]. Furthermore, retrospective clinical analyses using HLAMatchmaker showed that HLA-A,-B mismatched kidneys with zero to two triplet mismatches, display similar graft survival rates as the zero HLA-A,-B antigen mismatches [257, 258]. This improved graft survival rate as compared to the high-triplet mismatch group was observed in both sensitized and nonsensitized recipients [257], suggesting that HLAMatchmaker could be used as a prognostic tool for matching renal transplants. The latter issue could, however, not be confirmed in an independent cohort, possibly due to the lack of high-resolution DNA typing of HLA alleles [259]. Therefore, the use of HLAMatchmaker as a prognostic tool for matching kidney transplants is still a matter of debate.

The induction of HLA antibodies depends on the HLA phenotype of both the donor and the recipient [260, 261]. Consequently, the immunogenicity of an HLA mismatch must be considered in the context of the HLA phenotype of the recipient, implying that a single donor-HLA mismatch can have different effects in recipients with different HLA types. This concept is supported by studies on the influence of acceptable single HLA mismatches on graft survival [262] and by data on Bw4-specific antibodies [263, 264]. Subsequent analyses on the antibody production

after kidney transplantation indicated that selection of HLA class-I mismatches of the donor in the context of the HLA-DR phenotype of the responder might reduce the incidence of humoral graft rejection and minimize the sensitization grade of retransplant candidates [261]. The correlation between antibody formation against specific HLA class-I antigens and the presence of a particular HLA class II molecule suggest a role for indirect presentation of donor-derived HLA peptides by HLA class-II molecules on the antigen-presenting cells of the donor. This indirect presentation, may lead to Thelper-2-cell responses, which are required for IgM to IgG isotype switching of the B-cell response, leading to the production of donor-specific antibodies (DSA) of the IgG isotype [265]. Although HLAMatchmaker predicts which HLA-antigens can potentially induce HLA antibody formation, it does not predict T-cell reactivity towards allogeneic HLA [266].

Binding of peptides to HLA molecules is predictable. For many HLA molecules, binding affinities of different peptides have been determined and used to train and test HLA-binding predictors. Nowadays, these predictors display a good performance; the differences between predicted binding affinities and experimental measurements have been shown to be as small as the differences in measurements between different laboratories [50]. Predictability is particularly high for HLA class-I molecules, as these molecules have a more strict preference for nine amino acid long peptides (9-mers) and require specific amino acids as anchor residues at clearly defined anchor positions [10]. For HLA class II molecules predictability is lower, as peptides of different length can bind using different positions as anchor residues [267]. Therefore, it is difficult to determine how a peptide aligns to the HLA class II-binding groove and which amino-acid residues in the peptide are preferred as anchors. To solve this problem, Nielsen et al. used a socalled core-predictor to estimate how a peptide positions in the class II binding groove [268]. The core-predictor enabled the development of an accurate HLA class-II predictor, called NetMHCII [269].

To investigate the role of donor HLA-derived T-helper epitopes in the de novo development of DSA, we used NetMHCII to identify allogeneic HLA class I-derived, HLA class II-presented epitopes. We subsequently investigated whether the inability of the recipient's HLA-DRB1 molecule to present specific non-self HLA class-I epitopes explained the lack of antibody production in a cohort of 21 non-immunized individuals who received and lost their transplant and developed HLA class-I to some but not all HLA-mismatches.

6.2 Results

6.2.1 Overview of specificities

A total of 22 recipients matched the inclusion criteria (table 6.1). In sera of 21 of them, donorspecific antibodies could be detected. A total of 38 immunogenic (18 HLA-A and 20 HLA-B) and 11 non-immunogenic alleles HLA alleles (3 HLA-A and 8 HLA-B) could be identified. These numbers were equally distributed (Chi-square test, data not shown).

6.2.2 Immunogenic alleles contain more T-helper ligands

For all mismatched alleles we predicted the number of non-self peptides that should be able to bind to the HLA-DRB1 allele of the recipient. As shown in figure 6.1A, the immunogenic group contains a higher number of DR-presentable T-helper ligands as compared to the nonimmunogenic group ($p<0.01$ in the Mann-Whitney U test). The mean values were 3.0 and 1.2, respectively. These differences were not observed when the mismatched donor-derived HLA alleles were analyzed against a scrambled recipient DRB1 background (figure 6.1B), indicating that the DRB1 background of the specific recipient plays a crucial role in the analyses.

6.2.3 Immunogenic alleles have increased numbers of triplets and eplets

The immunogenic group and the non-immunogenic group were compared for their number of triplets and eplets as determined by HLAmatchmaker. Both the number of triplets (figure 6.2A) and the number of eplets (figure 6.2B) are significantly higher in the immunogenic group than in the non-immunogenic group (triplets: $p<0.005$; eplets: $p<0.0005$ in Mann-Whitney U tests).

6.2.4 Eplets do not co-localize with DRB1-presented T-helper ligands

Based upon the shared biological origin of the eplets and the DRB1-presented T-helper ligands a correlation between the number of eplets and the number of T-helper ligands is to be expected. To address this issue, we plotted the number of eplets against the number of T-helper ligands and performed correlation analyses.

nr.	Age of recipient [years]	Gender recipient/ donor	time till eccomy [months]	HLA class I mismatch	HLA-DRB1 typing of recipient	Immunogenic alleles	Non-immunogenic alleles
1	26	M/F	143	A*32:01, B*15:17	DRB1*03:01, -	A*32:01, B*15:17	-
2	51	M/M	46	A*03:01, B*18:01	DRB1*03:01, -	A*03:01	B*18:01
3	48	M/F	0	A*02:01, B*41:01	DRB1*04:01, DRB1*07:01	A*02:01, B*41:01	-
4	54	M/M	0	A*01:01, A*24:02, B*39:06	DRB1*03:01, DRB1*15:01	A*01:01, A*24:02	B*39:06
5	47	M/M	10	A*03:01, B*07:02	DRB1*07:01, DRB1*13:01	B*07:02	A*03:01
6	29	M/M	2	A*02:01, B*57:01, B*49:01#	DRB1*07:01, DRB1*15:01	A*02:01, B*49:01	B*57:01
7	32	M/M	0	A*01:01, B*08:01	DRB1*04:01, DRB1*15:01	A*01:01, B*08:01	-
8	54	M/F	0	B*44:02	DRB1*04:01, DRB1*11:01	B*44:02	-
9	43	M/F	0	B*40:01	DRB1*01:01, DRB1*04:01	B*40:01	-
10	69	M/M	0	A*24:02, B*18:01, B:35:03	DRB1*04:01, DRB1*07:01,	A*24:02, B*18:01	B:35:03
11	53	M/F	0	A*03:01, B*07:02, B*15:01#	DRB1*15:01, DRB1*03:01	A*03:01, B*07:02, B*15:01	-
12	66	M/M	0	A*24:02, B*44:02, B*39:06	DRB1*04:01, DRB1*11:01	-	A*24:02, B*44:02, B*39:06
13	35	M/M	0	A*24:02, B*39:06, B*51:01	DRB1*04:01, -	A*24:02, B*39:06, B*51:01	-
14	47	M/M	101	A*02:01, B*27:05	DRB1*09:01, DRB1*15:01	A*02:01, B*27:05	-
15	59	F/M	4	A*02:01, B*51:01	DRB1*12:01, -	A*02:01	B*51:01
16	54	F/F	3	A*01:01, B*07:02	DRB1*03:01, DRB1*15:01	A*01:01, B*07:02	-
17	27	F/M	0	B*51:01	DRB1*03:01, -	B*51:01	-
18	53	F/M	0	A*01:01, A*0201, B*08:01, B44&	DRB1*08:01, DRB1*11:01	A*01:01, A*0201, B*08:01, B44	-
19	45	F/M	0	A*02:01, B*07:02	DRB1*07:01, DRB113:01#	A*02:01, B*07:02	-
20	30	F/F	0	A*32:01	DRB1*03:01,	A*32:01	-
21	17	F/F	38	B*57:01	DRB1*03:01, DRB1*13:01#	B*57:01	-

Table 6.1. Description of the transplant study group matching our inclusion criteria, extracted from a cohort of 869 kidney transplant pairs. (#) No DNA available for high resolution typing. Extrapolated from the serological broad typing. (&) No DNA available for high resolution typing, number of eplets/peptides calculated as average of B*44:02 and B*44:03.

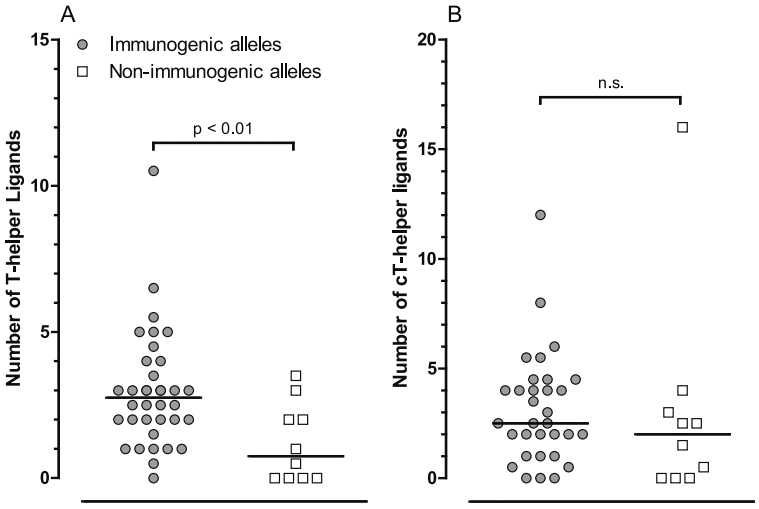


Figure 6.1. Comparison of the number of T-helper ligands in immunogenic (solid dots) and non-immunogenic (open boxes) HLA class I alleles. A) Higher number of DR-presentable T-helper ligands were observed in the immunogenic alleles as compared to the non-immunogenic alleles. B) No differences were observed when the mismatched donor-derived HLA alleles were analyzed against a scrambled recipient DRB1 background. The reported p-values are derived from the Mann-Whitney U tests.

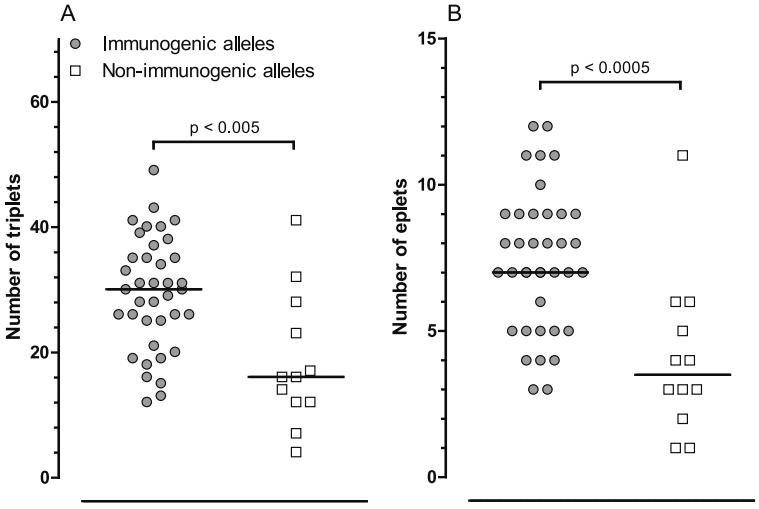


Figure 6.2. Comparison of the number of HLAMatchmaker triplets (A) and eplets (B) in immunogenic (solid dots) and non-immunogenic (open boxes) HLA class I alleles. The reported p-values are derived from the Mann- Whitney U tests.

Correlations were observed both for the non-immunogenic alleles (R^2=0.83; significance of the slope: p<0.0005) and the immunogenic alleles (R^2=0.14; significance of the slope: p<0.05; figure 6.3A).

We subsequently analyzed the topographic location of the T-helper ligands in our study cohort, indicated by the position of the involved amino acids. The location of these T-helper ligands was compared to the location of the eplets. The polymorphic amino acids of T-helper ligands are highly overrepresented in the β-plated sheet and in the alpha-3 domain of the HLA protein (figure 3B), whereas the eplets are located on surface residues of the HLA protein, accessible to antibodies. These data were confirmed by simulation experiments on virtual transplant pairs (figure 6.3C), showing that a significant number of polymorphic amino acids (62%) can be part of an HLA class II-presented T-helper ligands, while not being part of an eplet.

6.3 Discussion

The HLA-DR phenotype of the responder may play a determinative role in the immunogenicity of HLA antigens [261, 263, 264]. The production of Bw4-specific antibodies strongly correlates with the presence of either the HLA-DR1 or HLA-DR3 phenotype in the responder. *In vitro*, a Bw4- derived peptide binds strongly to DRB1*01- and DRB1*03-expressing cells, while the corresponding Bw6 peptide does not. Similarly, HLA-DRB1*15:01 shows an enrichment in the production of HLA-A2 antibodies in HLA-A2-mismatched transplant pairs [263]. Based upon this concept, it has been proposed to consider HLA class-I mismatches of the donor in the context of the HLA-DR phenotype of the responder in order to improve the outcome of kidney transplantation. So far, this concept has only been evaluated without taking the HLA class-I background of the recipient into account. Moreover, given the large number of HLA alleles, enormous transplantation cohorts would be required to define all combinations that are at risk.

In the present study we applied a computational approach for HLA binding with subtraction of the self-HLA to explain donor-specific HLA antibodies. We demonstrate that immunogenic donor-derived HLA class-I alleles, defined as alleles towards which DSA are detectable, contain a higher number of epitopes that can be presented by HLA class-II molecules from the recipient. Evidently, the level of homology between donor's and recipient's HLA class-I alleles may affect the number of T-helper ligands; the lower the homology the higher the chance to find non-self epitopes. To support the hypothesis that the recipient's HLA-DR rather than the level of homology was explaining our observations, we counted the number of T-helper ligands using a random HLA-DR background. As immunogenic and non-immunogenic alleles showed a similar number of peptides in the context of the random HLA-DR background (figure 6.1B), we conclude that the recipient-

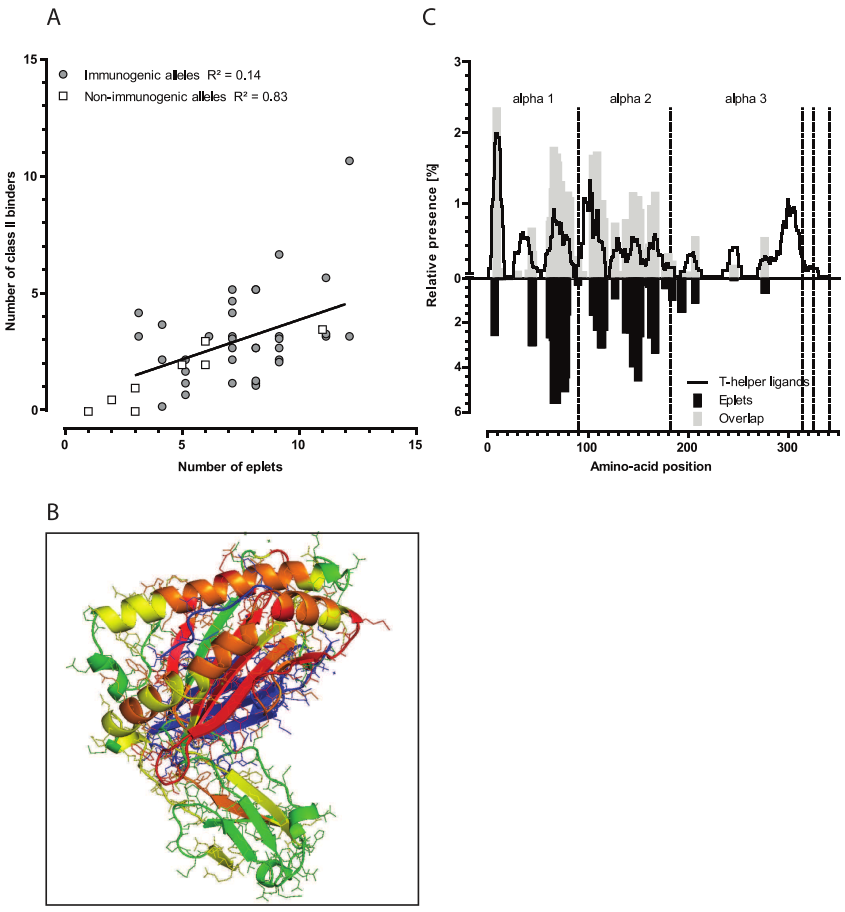


Figure 6.3. Identification of eplets and HLA class II-presented T-helper ligands as two separate entities. A) Correlation between the number of eplets and the number of T-helper ligands. Immunogenic alleles have been depicted as grey dots; non-immunogenic alleles as open boxes. The resulting regression curves have been depicted by solid lines and dotted lines respectively. The regression coefficient is based upon combined analysis of the two groups. Overlapping data have been shifted 0.1 unit for visualization purposes only. B) Location of the HLA class II-presented T-helper ligands on the HLA class-I molecule, as observed in the studied kidney transplant cohort. Colors indicate the relative presence of an amino acid in immunogenic HLA class I antigen; (Green = 0, yellow = 1 to 4, orange = 5 to 9, and red $>$ 9). The nonpolymorphic beta-2m molecule has been depicted in blue. C) Location of the HLA class II presented T-helper ligands versus eplet-related residues in the HLA class-I molecule, as observed in a virtual transplantation cohort of 10000 simulated transplants. The eplet-related residues were defined as polymorphic residues present within 3.0 Angstrom eplet patch.

specific HLA-DR background is essential in predicting the chance of developing DSA after transplantation.

Two factors may improve the outcome of our analyses; the quality of the HLA class II-binding prediction and the resolution of HLA typing of recipient and donor. In NetMHCII, the HLA binding motives are well-defined for 9 HLA-DR antigens including 11 different HLA-DRB1 alleles. Thus, for a number of HLA-DRB1 alleles the peptide binding characteristics have not been determined. Given the large diversity in HLA-DRB1 alleles, characterization of each individual HLA class-II binding motif via peptide-screening binding assays is not feasible. Therefore, an alternative algorithm, NetMHCIIpan, has been developed via a computational approach [270]. NetMHCIIpan can define binding motifs on the basis of the primary amino-acid sequence, providing information for alleles for which limited experimental binding data have been reported [270]. For our data set, NetMHCIIpan would only provide a better prediction for the HLA-DRB1*13:01 allele. As such, analyses with NetMHCIIpan did not enhance the performance of our models (data not shown).

In the present retrospective study, low resolution typing data could not be extrapolated to high resolution HLA typing for 3 donors and 2 recipients. In these cases, the results from all likely options were averaged (in case of donor typing) or both subtracted (in case of recipient typing). This approach may have lead to an incorrect assignment and subsequently to an underestimation of the effect the number of T-helper ligands on the induction of specific antibodies. Thus, the effect of T-helper ligands on the production of anti-HLA IgG antibodies may be stronger than currently reported. Extended analyses on high-resolution typed recipient-donor are required to estimate the magnitude of the effect in more detail.

Both the T-helper ligands described in this study and the eplets as determined by HLAMatchmaker are based upon the same phenomenon; mismatched amino acids in the HLA alleles of recipient and donor. As such, these two parameters cannot be fully dissected. However, although immunogenic alleles show higher numbers of eplets/triplets than the non-immunogenic alleles, various aspects of our analyses indicate that T-helper ligands act, at least partly, independently from the number of eplets/triplets. First, we show that the actual HLA-DRB1 background is essential; when using a scrambled HLA-DRB1 background, no correlation was found (figure 6.1B). Second, while there is a strong correlation between the number of T-helper ligands and the number of eplets when analyzing the non-immunogenic group, this correlation is much weaker in the immunogenic group (figure 6.3A). Third, we demonstrate the physical location of amino acids that are included in potential T-helper ligands are differently distributed when compared to the location of eplet-involved amino acids (figure 6.3B-C); the alpha-3 domain and the N-terminal part of the alpha-1 domain seem to be enriched for T-helper ligands, while they rarely result in eplets. Taken together, we conclude that these two

parameters are complementary to each other while predicting the chance of DSA development.

In summary, we show that the *de novo* development of donor-specific HLA IgG antibodies correlates with the number of non-self donor-derived HLA epitopes that can be presented by recipient-HLA class II molecules and with the number of HLAMatchmaker eplets in the mismatched HLA-allele of the donor. Topographic analyses and scrambling of the HLA-DRB1 background suggest that these two phenomena result from two in part independent entities. Therefore, cross presentation of donor-derived HLA by recipient HLA class-II molecules may be an important mechanism in IgM-to-IgG isotype switching of donor-specific HLA antibodies. Large clinical studies are required to evaluate the effect of this mechanism on graft survival.

6.4 Methods

6.4.1 Transplant recipients

We analyzed the entire cohort of 869 kidney transplants that were performed between 1990 and 2008 in the University Medical Center Utrecht, Utrecht, The Netherlands. From this cohort, we selected recipients that lost their kidney graft and had no pre-transplant alloimmunizing event, i.e. no pregnancy, blood transfusions, or previous organ or stem-cell transplantation. Recipient pairs that were fully matched for the HLA-A and HLA-B antigens were excluded, as they were not informative for this study purpose. One pair was excluded because no binding algorithm was available for the recipient's HLA class-II alleles. These selection criteria resulted in 21 analyzable recipient-donor pairs. For all donor-recipients combinations T-cell cross-match assays were performed using the basic NIH technique on unseparated peripheral blood mononuclear cells in the presence of dithiothreitol before transplantation. All cross-match results in were negative.

6.4.2 Samples

Serum samples were obtained at two time points. First, pre-transplant sera used for crossmatching was analyzed. Second, post-transplant sera were used which were obtained three months after transplantectomy. The reason for the latter time point is that at that time immune suppression was absent and antibody analysis was no longer influenced by any antibody filtering effect of the donor kidney. All sera of the recipients were obtained for purposes of regular panelreactive HLA antibody (PRA) screening.

6.4.3 HLA typing

For each recipient, two independently collected samples were typed with different methods; one sample was typed serologically, using the conventional complement-dependent cytotoxicity (CDC) procedure using commercial typing trays (Biotest, Dreieich, Germany) and one sample was typed molecularly at intermediate resolution for the HLA class-I and -II alleles based upon the PCR-SSO technique in combination with Luminex using commercial reagents and following the instructions of the manufacturer (OneLambda Inc., Canoga Park, CA, USA). For donor typing, only one sample was available locally to perform both serological and molecular typing, following the identical procedure as for recipient typing. In all cases, donor typing in our center confirmed the HLA typing provided by the donor center. An additional high-resolution typing was performed from all recipients and donors of whom DNA was still available. From the remaining 5 individuals, all typing results were converted to the most likely high resolution typing based upon the reported HLA frequencies within the observed NMDP multiple allele codes [33]. In one case, this approach led to multiple options with a frequency of more than 10%; a B44 could be converted into either a B*44:02 or a B*44:03 (table 6.1). For this pair, data were analyzed for both possibilities, averaging the results.

6.4.4 HLA antibody screening and characterization

Tests were performed to determine the presence or absence of HLA antibodies to HLA-A and -B using the Labscreen Single Antigen kits (OneLambda Inc.) following the standard manufacturer's guidelines. Beads were analyzed on a Luminex 200 flow cytometer (Luminex Inc., Austin, TX, USA). Results with an MFI of $>$ 4000 were scored as positive.

6.4.5 HLA class II-binding predictions

For all mismatched HLA class-I molecules, the number of allogeneic T-helper ligands was examined to explain a potential antibody response to the epitope-containing HLA class-I allele. Allogeneic T-helper ligands were defined as recipient HLA class-II binding epitopes within the mismatched donor-derived HLA class-I molecule, that were not covered by any of the other HLA class-I alleles of the recipient. HLA class I-derived T-helper ligands were predicted using the HLA class-II binding predictor NetMHCII [268]. This predictor is based upon the SMM-align predictor [269] to predict how a potential ligand aligns to the binding groove of an HLA class-II molecule, and subsequently predicts how well the aligned ligand is expected to bind. If the predicted binding affinity was higher

than 1000 nM [271], the nine amino acids that aligned to the binding groove were defined as an HLA class-II epitope.

6.4.6 Matchmaker analyses

HLAMatchmaker eplets were assigned to the HLA alleles based on HLAMatchmaker version 2.1 (http://www.HLAMatchmaker.net) [272]. Only the HLA-A, and -B loci were included in these analyses. The number of mismatched eplets was determined as the number of donor eplets that were absent in the recipient's HLA-A and -B locus.

6.4.7 Location of T-helper ligands and eplets

Different polymorphic residues within the HLA molecule may contribute to the different types of mismatches, i.e. as determined by either the eplet- or the T-helper-ligand method. To identify the polymorphic residues that were involved in eplets and/or T-helper ligands, we analyzed the data obtained from our study cohort and a cohort of randomly generated virtual recipient-donor pairs. For the latter cohort, we first randomly generated a virtual population reflecting the HLA A/B/C/DR-haplotype frequencies in Caucasians. These frequencies were obtained from previous studies [33]. To simulate a recipient-donor combination matching our allocation protocol, at least one HLA-A or -B, and one HLA-DR match between recipient and donor was required. Within 10000 random recipient-donor combinations that fitted these requirements, the mismatched HLA class-I alleles were assessed using the eplet-method and the Th-ligand-method as described above. Relative frequency plots were constructed based upon the location of eplets and Th-ligands within the HLA molecule.
To measure the overlap in position usage between the Th-ligand-method and the eplet-method, the usage of each amino acid by either methods was determined at those positions that are variable among MHC-I molecules. For both methods, the usage counts were normalized such that they sum up to 100%. The overlap between the counts was determined as the overlap between these normalized usage counts.

6.4.8 Statistical analyses

All mismatched alleles from the donor were separated into a group for which DSA were detected (immunogenic group) and a group for which no DSA could be demonstrated (non-immunogenic group). Between these two groups, the number of non-self HLA class-I epitopes presented by HLA class II and/or the number of

HLAMatchmaker triplets and eplet were compared using the Mann-Whitney U test (GraphPad Prism 5.03, GraphPad Software, Inc., La Jolla, CA).

Chapter 7

Summarizing Discussion

In this thesis we approach the question of which pMHCs become epitopes from three different perspectives. First, we investigate which peptides are more likely to be MHC-I presented than others. We show that HLA-A molecules present more pathogen-derived peptides compared to self-derived peptides, by making use of the fact that pathogens have a low genomic G+C content, and therefore an enhanced frequency of the amino acids F, I, N, K and Y (Chapter 2). Furthermore, we investigate the optimal prediction of the MHC-I ligandome using proteolysis, TAP transport and MHC-I binding predictors (Chapter 3). Second, we show that certain positions and amino acid content of MHC-I presented peptides are associated with immunogenicity. The observed correlates with immunogenicity were summarized in a model that could be applied to successfully predict the immunogenicity of novel pMHCs in large-scale epitope discovery projects (Chapter 4). Third, we studied how the similarity to self influences the formation of an immune response to foreign pMHCs. We show that $\sim$30% of the MHC-I presented foreign peptides is unlikely to be an epitope, as they resemble a self-peptide on the same MHC-I molecule (Chapter 5). Furthermore, we model the constraints that this self/nonself overlap puts on the capacity of the immune system to elicit immune responses without causing auto-immune responses (Addendum to Chapter 5). Finally, we investigate how self-overlaps affect host-versus-graft immune responses after kidney transplantation. We show that a humoral immune response to mismatched HLA class I molecules depends on the presentation of non-overlapping class-I-derived peptides on HLA class II (Chapter 6). Here, we will provide a more detailed discussion for each of the chapters, and conclude by demonstrating the current capacity of epitope prediction tools.

7.1 Peptide binding preferences

In Chapter 2, we have shown that pathogen-derived peptides bind better to HLA-A molecules than self-peptides. The binding of amino acid residues that associate with a low G+C content underlies this preference; we refer to such a binding preference as a G+C-negative preference. Why pathogens have a low G+C content is unknown, but the subsequent influence on their amino acid frequencies is clear. The low G+C content leads to an increased number of F, I, N, K and Y, and decreased number of G, A, R and P amino acids. Due to the G+C negative preference, pathogen-derived peptides are better presented to the immune system than peptides from the human proteome. For instance, *Plasmodium falciparum* that causes malaria has a G+C content of $<$20%, which results in a 63% enhanced presentation compared to human proteins for the G+C-negative HLA-A*0301 molecule.
The MHC-I preference for pathogen-derived peptides seems to be an important feature for the MHC-I molecules in different species. We found a dominant G+C-negative preference in the Patr-A locus of Chimpanzees, as well as in the Mamu-B locus of Macaques (Chapter 2). We therefore propose that MHC-I molecules that are G+C-negative have an increased chance to be selected and become more frequent. The HLA-B locus evolves faster than HLA-A, probably due to frequent recombination events [141, 157, 158, 160], and possibly a stronger selection for alleles that encode rare or different binding motifs. We think that the increased selection for rarity explains why there is no fixation of G+C-negative preferences in the HLA-B locus. In a commentary on our publication of Chapter 2, Levasseur and Pontarotti suggested an opposite model, in which the G+C preferences reflect an ancient preference of MHC-I molecules [273]. However, we showed that G+C-negative preferences of HLA-A molecules are conferred by many different binding motifs (Figure 2.4). Therefore, one can conclude that similar G+C preferences are not likely to be a signature that is remaining from the common ancestry of MHC-I molecules.
We believe that the G+C-negative preference of HLA-A molecules explains several related findings in the literature. For example, Vider-Shalit et al. [274] calculated a so-called "Size of Immune Repertoire" (SIR)-score to express how well a virus is presented by the human MHC-I molecules, and claimed that herpesviruses adapt to the human HLA system, because some human herpes-viruses were presented worse than non-human herpes-viruses. However, the reported SIR-scores correlate perfectly with the G+C content of these herpes-viruses and provide a better explanation for this observation (Spearman rank test: ρ=-0.81; p$<$0.001). Hertz et al. [275] suggested that there is a selection pressure on the MHC-I system to present conserved sequences from self and nonself proteomes. As G+C content is strongly correlated with recombination rates [276], we believe that G+C-preferences provide the mechanism that explains the suggested selection pressure. Hertz et al. calculated for several viruses and HLA molecules the correlation between the conservation of a sequence and how well it is MHC-I

presented; and refer to this correlation as the "efficiency score" [275]. As with the SIR-scores, the efficiency scores correlated perfectly with the G+C content of double-stranded DNA viruses and negative-sense single-stranded RNA viruses (Spearman rank test: $\rho \geq 0.8$; $p \leq 0.001$). Moreover, many efficiency scores were greater for HLA-A molecules, in line with the G+C negative preference that we describe for this locus. Recently, Granados et al. [277] claimed that microRNA (miRNA) targeted genes express proteins that are better presented on MHC-I molecules. As miRNAs are known to target mRNAs with a high A+U content [278], we expect that also this result can be explained in terms of G+C preferences of MHC-I molecules.

7.2 Predicting the MHC-I ligandome

Not only peptide-binding preferences determine which peptides are MHC-I presented; precursor proteins need to be expressed and processed into peptides, and those peptides need to be transported to the ER and bind to MHC-I, followed by transportation to the cell surface. In Chapter 3 we study how to combine predictors of different steps in the MHC-I pathway into a prediction of the MHC-I ligandome, and show that optimal predictions are made when the predicted chance of proteolysis is added to a normalized MHC-I-peptide binding prediction (Chapter 3).

The proposed optimal MHC-I ligand prediction method is closest to the NetCTL and NetCTLpan methods, in which MHC-I binding, proteolysis and TAP transport predictions are differentially weighted and summed into a final score [147, 180]. The suggested weightings for MHC-I binding, proteolysis and TAP transport are 0.15, 0.75 and 0.05, respectively [147], which is close to our finding that TAP transport scores are dispensable. In Chapters 2 and 4, we require peptides to exceed a certain MHC-I binding, proteolysis and TAP transport score to be predicted as an MHC-I ligand, we refer to this method as the filter method. In Chapters 2, we do not expect that using the additive method would change our results, as we compared the preferences of different HLA class I molecules and this preference is the main factor that determines which MHC-I ligands are presented. In Chapter 4, MHC-I ligand predictions were used to compare the overlap of pathogen- and self-derived peptides when presented on the same HLA molecule. Here, the specific prediction of a self-overlap for individual pathogen-derived pMHCs might improve when a better MHC-I ligand prediction model is used. However, our main conclusion that the self/nonself overlap is large, depends on degenerate T-cell recognition, and should not be affected by the type of MHC-I ligand prediction model.

The binding affinity of a certain peptide on an MHC-I molecule is measured in a competition experiment with a reporter peptide, or in a renaturation assay where denaturated HLA molecules refold in the presence of a peptide ligand [50]. As

the affinities of the reporter peptides for different MHC-I molecules vary or the capacity of different MHC-I molecules to renaturate, the variation among measured binding affinities for different MHC-I molecules is unavoidable. This variation can be removed by a normalization method (scaling), assuming that every MHC-I molecule binds the same fraction of binders. We show in Chapter 3 that MHC-I ligandome predictions significantly improve if MHC-I binding predictions are scaled. This finding does not show that all MHC-I molecules have the same specificity; it only demonstrates that possible biological variation in specificity is not properly predicted by current MHC-I binding predictors. Future investigations into the binding specificity for different MHC-I molecules might provide a method to take such biological variation into account and improve MHC-I ligandome predictions. Until that time, we show that the best strategy is to assume all MHC-I molecules have the same specificity.
Which proteins are used as precursors for MHC-I ligands has an important effect on MHC-I ligandome predictions. In our group, Hoof et al. showed that the abundance of a precursor protein influences to a large extent its chance of being sampled by the MHC-I presentation pathway [53]. In addition, the abundance of precursor proteins provided better information on the chance of being sampled than gene expression data [53]. Nevertheless, both types of data as well as other properties like the protein length were shown to be informative for the MHC-I precursor predictions [53]. The analysis of Hoof et al. [53] was based on data from MHC-I elution studies that directly assess the MHC-I ligandome, however such data sets are unfortunately still sparse. More peptide elution efforts are needed to assess the MHC-I ligandome in different tissues, with different MHC-I backgrounds and in different (e.g. infected or uninfected) states.

7.3 Predicting immunogenicity

T-cell precursor frequencies vary substantially for different pMHCs [101, 104]. Kotturi et al. showed that these frequencies associate with the chance of mounting an immune response [56]. In Chapter 4 we investigate the factors that determine the chances of a pMHC to be recognized by T-cells, i.e. its immunogenicity. We show that certain amino acids are more abundant in immunogenic than in non-immunogenic HLA class I presented peptides, probably because more TCRs can interact with these amino acid residues. In addition, we describe the importance of different positions of the presented peptide; the amino acids at the middle positions of the presented peptide (P4-P6) are most important for immunogenicity, in line with the close contact of these positions with the TCR. These findings were summarized in a model that can predict the immunogenicity of pMHCs.
For our study, we defined non-immunogenic pMHCs as the ones that give a negative response in a peptide-immunization experiment. Unfortunately, not many peptide-immunization experiments have been performed, which limits the size of

our data set and the strength of our analysis. In the future, we might be able to derive a better picture of the determinants of immunogenicity, if more data becomes available. A refined model to predict immunogenicity could be made MHC-I molecule specific, as TCR-pMHC interactions can differ per MHC-I molecule. Similarly, different parts of the TCR interact with different positions of an MHC-I presented peptide, which could lead to different preferred amino acid interactions at different positions of the presented peptide. Different position specific amino acid associations with immunogenicity could be the result from such differential preferences, that could be used to improve immunogenicity predictions. Finally, our model could be improved if it was trained on quantitative measures of immunogenicity, just as MHC-I binding predictors improved from the training on quantitative binding measurements. Precursor frequency measurements can provide a quantitative measure of immunogenicity, and can be realized in humans on a large scale due to the recent development of sensitive assays [101–103].

7.4 Self/nonself overlaps

An important and predictable factor that determines if a pMHC generates an immune response is its similarity to self [117, 118]. In Chapter 5 we estimated the fraction of foreign peptides that overlaps with self, for different MHC-I molecules. We show that the overlap is large: $\sim$30% of all foreign pMHC resembles self. This estimate is based on a model of T-cell-pMHC interactions. On the one hand, our model is strict compared to previous estimates of T-cell specificity [54], so the actual self/nonself overlap might be higher. On the other hand, our prediction of self-pMHCs might be too large as we predict all proteins in the human proteome to be possible MHC-I ligand precursors. Thus, the predicted self/nonself overlaps should be considered as a best estimate and an indication that the influence of self-overlaps on T-cell immune responses is non-negligible.
Next, we analyzed how the large self-overlap of foreign pMHCs influences the chances of mounting immune responses and auto-immune responses to a pathogen in an Addendum to Chapter 5. Using a simple probabilistic model, in which naive T-cells are stimulated by pathogen-derived pMHCs to seek and kill infected cells independently, we show that the risk of auto-immunity becomes very high, unless induction of self tolerance is extremely efficient ($>$97%). I think that this required level of self tolerance control is too high to be achieved by mechanisms that act prior or during the naive-to-effector transition of T-cells. However, after this transition, when effector T-cells are performing their effector function, there is a possibility of requiring extra information and thereby increase the specificity of an immune response and reduce the chance of auto-immunity. For instance, the regulation of effector T-cell entry into infected tissues [119] might be a mechanism to increase the specificity of an immune response, as well as other mechanisms that can involve the integration of signals from other cells of the immune

system, e.g. NK cells, to allow for a more specific cytotoxic signal.
In Chapter 6 we tested if not only $CD8^+$, but also $CD4^+$ T-cell responses are affected by self-overlaps. In this case, the $CD4^+$ T-cell immune response was measured indirectly as an antibody immune response to mismatched HLA class I molecules after a kidney transplantation. The overlap with self was determined as the number of donor (non-self) derived MHC-II ligands, that were not present in the recipients background (i.e. that were not self-overlapping). Only when the overlap with self (i.e. the recipient) was considered, could we predict which mismatches were immune targeted. In this work we determined overlaps based on the so-called core (9mer) peptide of an HLA class II presented 15mer that was predicted to be aligned to the MHC-II binding groove. When more data on the T-cell receptor interaction with MHC class II presented peptides would be available, we could improve the self-overlap predictions for peptides presented on MHC-II molecules to $CD4^+$ T-cells. Such an improved analysis could be applied to future host-donor transplant matching to minimize the chance of transplant rejection.

7.5 Predicting Epitopes

We have investigated and discussed three processes that influence which peptide-MHC-I complexes become epitopes, i.e. MHC-I presentation of peptides, T-cell recognition of pMHCs and self-tolerance induction. For each process did we provide a new method to predict which peptides or pMHCs are likely candidates, or did we provide suggestions on how to optimally make such a prediction. Now, we would like to demonstrate how the combination of these predictions, helps us to predict epitopes. In other words, we want to quantify in how many cases a peptide "bound to be an epitope" is also an epitope.
We decided to test our epitope predictions on a well-studied immune response, the T-cell response to HIV-1-derived 9mer peptides presented on the most common HLA molecule, HLA-A*0201. Reported immune responses to HIV-1 are made accessible via the Los Alamos HIV Immunology Database. However, not all these reports were reproduced, therefore we limited our test set to 33 9mer peptides that were shown to elicit an HLA-A*0201 restricted T-cell response in at least three reports (data provided by Marit van Buuren). In addition, we included four other well known 9mer peptides from the so-called "A-list" of HIV-1 derived pMHCs that were unequivocally shown to be HLA-A*0201 restricted epitopes in many HIV patients. We compared the list of epitopes (n=37) with all other 9mer peptides in the HIV-1 proteome that were not reported as HLA-A*0201 restricted epitopes (n=3003), that we refer to in this section as non-epitopes.
We wanted to test how many of the peptides that are bound to be epitopes are actually epitopes. To quantify this, we focused on the top-1% candidates (n=30) and determined how many of those were actual epitopes based on several predictions discussed in this thesis. We first evaluated how well MHC-I binding predic-

tions would help to enrich epitopes, thus the 30 peptides with highest predicted binding affinities were selected, of which 12 were epitopes. In Chapter 3, we showed that combining MHC-I binding predictions with proteolysis predictions results in a better MHC-I ligand prediction model than MHC-I binding alone. If we apply the combined prediction model to the HIV-1 peptides, 15 of the 30 selected peptides are epitope. Out of these epitopes 13 have no self-overlap (Chapter 5), while only 8 have a positive immunogenicity score (Chapter 4). How to combine self-overlap and immunogenicity predictors with MHC-I ligand predictions can be investigated and improved when more epitope/non-epitope data sets for different pathogens are available, analogous to our study of the combination of predictions for different steps in the MHC-I presentation pathway (Chapter 3). In addition, self-overlap and immunogenicity predictors are expected to improve in the future if more data on self-pMHCs and immunogenicity of pMHCs will be available. This small analysis illustrates how close we are to answering our initial question, $\sim$50% of the peptides that were "bound to be an epitope" according to our MHC-I ligand and self-overlap models have been confirmed to be epitopes. For the other predicted epitopes, future studies can still confirm that they are immune targeted. To conclude, the ultimate goal of correct epitope predictions seems to be in reach, and we explored exciting steps to lead us there.

Bibliography

[1] Princiotta, M. F., Finzi, D., Qian, S.-B., Gibbs, J., Schuchmann, S., Buttgereit, F., Bennink, J. R. and Yewdell, J. W. 2003. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity* 18(3):343. (Cited on pages 3 and 4.)

[2] York, I. A., Mo, A. X. Y., Lemerise, K., Zeng, W., Shen, Y., Abraham, C. R., Saric, T., Goldberg, A. L. and Rock, K. L. 2003. The cytosolic endopeptidase, thimet oligopeptidase, destroys antigenic peptides and limits the extent of MHC class I antigen presentation. *Immunity* 18(3):429. (Cited on page 3.)

[3] Geier, E., Pfeifer, G., Wilm, M., Lucchiari-Hartz, M., Baumeister, W., Eichmann, K. and Niedermann, G. 1999. A giant protease with potential to substitute for some functions of the proteasome. *Science* 283(5404):978. (Cited on pages 3, 43, 46, 47, and 53.)

[4] Kessler, J. H., Khan, S., Seifert, U., Gall, S. L., Chow, K. M., Paschen, A., Bres-Vloemans, S. A., de Ru, A., van Montfoort, N., Franken, K. L. M. C., Benckhuijsen, W. E., Brooks, J. M., van Hall, T., Ray, K., Mulder, A., Doxiadis, I. I. N., van Swieten, P. F., Overkleeft, H. S., Prat, A., Tomkinson, B., Neefjes, J., Kloetzel, P. M., Rodgers, D. W., Hersh, L. B., Drijfhout, J. W., van Veelen, P. A., Ossendorp, F. and Melief, C. J. M. 2011. Antigen processing by nardilysin and thimet oligopeptidase generates cytotoxic T cell epitopes. *Nat Immunol* 12(1):45. (Cited on pages 3, 43, 46, 47, and 53.)

[5] Shen, X. Z., Lukacher, A. E., Billet, S., Williams, I. R. and Bernstein, K. E. 2008. Expression of angiotensin-converting enzyme changes major histocompatibility complex class I peptide presentation by modifying C termini of peptide precursors. *J Biol Chem* 283(15):9957. (Cited on pages 3, 43, 46, 48, and 53.)

[6] Shen, X. Z., Billet, S., Lin, C., Okwan-Duodu, D., Chen, X., Lukacher, A. E. and Bernstein, K. E. 2011. The carboxypeptidase ACE shapes the MHC class I peptide repertoire. *Nat Immunol* 12(11):1078. (Cited on pages 3, 46, 48, and 53.)

[7] Reits, E., Griekspoor, A., Neijssen, J., Groothuis, T., Jalink, K., van Veelen, P., Janssen, H., Calafat, J., Drijfhout, J. W. and Neefjes, J. 2003. Peptide diffusion, protection, and degradation in nuclear and cytoplasmic compartments before antigen presentation by MHC class I. *Immunity* 18(1):97. (Cited on pages 3, 4, 5, 43, and 63.)

[8] Momburg, F., Roelse, J., Hammerling, G. J. and Neefjes, J. J. 1994. Peptide size selection by the major histocompatibility complex-encoded peptide transporter. *J Exp Med* 179(5):1613. (Cited on pages 3 and 8.)

[9] van Endert, P. M., Tampe, R., Meyer, T. H., Tisch, R., Bach, J. F. and McDevitt, H. O. 1994. A sequential model for peptide binding and transport by the transporters associated with antigen processing. *Immunity* 1(6):491. (Cited on pages 3 and 8.)

[10] Falk, K., Rotzschke, O., Stevanovic, S., Jung, G. and Rammensee, H. G. 1991. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351(6324):290. (Cited on pages 3, 22, and 122.)

[11] Neefjes, J., Jongsma, M. L. M., Paul, P. and Bakke, O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 11(12):823. (Cited on pages 3, 5, and 43.)

[12] Toes, R. E., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T. P., Muller, J., Schonfisch, B., Schmid, C., Fehling, H. J., Stevanovic, S., Rammensee, H. G. and Schild, H. 2001. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J Exp Med* 194(1):1. (Cited on pages 3, 5, 8, 9, 43, 44, 45, and 63.)

[13] Emmerich, N. P., Nussbaum, A. K., Stevanovic, S., Priemer, M., Toes, R. E., Rammensee, H. G. and Schild, H. 2000. The human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate. *J Biol Chem* 275(28):21140. (Cited on pages 3, 5, 8, 9, 43, 44, 53, and 63.)

[14] Tenzer, S., Stoltze, L., Schonfisch, B., Dengjel, J., Muller, M., Stevanovic, S., Rammensee, H.-G. and Schild, H. 2004. Quantitative analysis of prion-protein degradation by constitutive and immuno-20S proteasomes indicates differences correlated with disease susceptibility. *J Immunol* 172(2):1083. (Cited on pages 3, 5, 8, 9, 43, 44, 46, and 63.)

[15] Burroughs, N. J., De Boer, R. J. and Kesmir, C. 2004. Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics*. 56(5):311. (Cited on pages 3, 5, 8, 9, 89, 90, 94, and 99.)

[16] Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V. and Brunak, S. 2002. Prediction of proteasome cleavage motifs by neural networks. *Protein. Eng.* 15(4):287. (Cited on pages 3, 5, 8, 9, 34, 44, 63, 90, 92, and 101.)

[17] Uebel, S., Kraas, W., Kienle, S., Wiesmuller, K. H., Jung, G. and Tampe, R. 1997. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. *Proc Natl Acad Sci U S A* 94(17):8976. (Cited on pages 4, 5, 8, and 63.)

[18] Assarsson, E., Sidney, J., Oseroff, C., Pasquetto, V., Bui, H.-H., Frahm, N., Brander, C., Peters, B., Grey, H. and Sette, A. 2007. A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *J Immunol* 178(12):7890. (Cited on pages 4, 5, 8, 15, 16, 63, 65, 78, 89, 92, 98, 100, 102, 104, 114, and 117.)

[19] Fruci, D., Lauvau, G., Saveanu, L., Amicosante, M., Butler, R. H., Polack, A., Ginhoux, F., Lemonnier, F., Firat, H. and van Endert, P. M. 2003. Quantifying recruitment of cytosolic peptides for HLA class I presentation: impact of TAP transport. *J Immunol* 170(6):2977. (Cited on pages 5 and 9.)

[20] Motozono, C., Yanaka, S., Tsumoto, K., Takiguchi, M. and Ueno, T. 2009. Impact of intrinsic cooperative thermodynamics of peptide-MHC complexes on antiviral activity of HIV-specific CTL. *J Immunol* 182(9):5528. (Cited on page 5.)

[21] Parker, K. C., Bednarek, M. A. and Coligan, J. E. 1994. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* 152(1):163. (Cited on pages 5 and 63.)

[22] Borghans, J. A. M. 2000. Diversity in the Immune Systematic. Ph.D. thesis, Utrecht University. (Cited on page 5.)

[23] Kloetzel, P. M. 2004. Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII. *Nat Immunol* 5(7):661. (Cited on pages 5, 47, and 53.)

[24] Hillen, N., Mester, G., Lemmel, C., Weinzierl, A. O., Muller, M., Wernet, D., Hennenlotter, J., Stenzl, A., Rammensee, H.-G. and Stevanovic, S. 2008. Essential differences in ligand presentation and T cell epitope recognition among HLA molecules of the HLA-B44 supertype. *Eur J Immunol* 38(11):2993. (Cited on pages 5 and 8.)

[25] Kincaid, E. Z., Che, J. W., York, I., Escobar, H., Reyes-Vargas, E., Delgado, J. C., Welsh, R. M., Karow, M. L., Murphy, A. J., Valenzuela, D. M., Yancopoulos, G. D. and Rock, K. L. 2012. Mice completely lacking immunoproteasomes show major changes in antigen presentation. *Nat Immunol* 13(2):129. (Cited on pages 5 and 43.)

[26] van Deutekom, H. W. M., Hoof, I., Bontrop, R. E. and Kesmir, C. 2011. A comparative analysis of viral peptides presented by contemporary human and chimpanzee MHC class I molecules. *J Immunol* 187(11):5995. (Cited on pages 5 and 44.)

[27] Robinson, J., Malik, A., Parham, P., Bodmer, J. G. and Marsh, S. G. 2000.

IMGT/HLA database–a sequence database for the human major histocompatibility complex. *Tissue Antigens* 55(3):280. (Cited on page 5.)

[28] Robinson, J., Mistry, K., McWilliam, H., Lopez, R., Parham, P. and Marsh, S. G. E. 2011. The IMGT/HLA database. *Nucleic Acids Res* 39(Database issue):D1171. (Cited on page 5.)

[29] Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. 1987. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 329(6139):512. (Cited on page 5.)

[30] Rapin, N., Hoof, I., Lund, O. and Nielsen, M. 2008. MHC motif viewer. *Immunogenetics* 60(12):759. (Cited on pages 5, 7, 8, 29, and 35.)

[31] Borghans, J. A. M., Beltman, J. B. and de Boer, R. J. 2004. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* 55(11):732. (Cited on pages 5 and 21.)

[32] Kubinak, J. L., Ruff, J. S., Hyzer, C. W., Slev, P. R. and Potts, W. K. 2012. Experimental viral evolution to specific host MHC genotypes reveals fitness and virulence trade-offs in alternative MHC types. *Proc Natl Acad Sci U S A* 109(9):3422. (Cited on page 6.)

[33] Maiers, M., Gragert, L. and Klitz, W. 2007. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol* 68(9):779. (Cited on pages 6, 130, and 131.)

[34] Schmid, B. V. 2009. Limits of viral adaptation to the antigen presentation pathway. Ph.D. thesis, Utrecht University. (Cited on page 6.)

[35] Aki, M., Shimbara, N., Takashina, M., Akiyama, K., Kagawa, S., Tamura, T., Tanahashi, N., Yoshimura, T., Tanaka, K. and Ichihara, A. 1994. Interferon-gamma induces different subunit organizations and functional diversity of proteasomes. *J Biochem* 115(2):257. (Cited on pages 6 and 43.)

[36] Gubler, B., Daniel, S., Armandola, E. A., Hammer, J., Caillat-Zucman, S. and van Endert, P. M. 1998. Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. *Mol Immunol* 35(8):427. (Cited on pages 6, 8, 53, and 63.)

[37] Walker, B. A., Hunt, L. G., Sowa, A. K., Skjodt, K., Gobel, T. W., Lehner, P. J. and Kaufman, J. 2011. The dominantly expressed class I molecule of the chicken MHC is explained by coevolution with the polymorphic peptide transporter (TAP) genes. *Proc Natl Acad Sci U S A* 108(20):8396. (Cited on page 6.)

[38] Schneider, T. D. and Stephens, R. M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18(20):6097. (Cited on page 7.)

[39] Sette, A. and Sidney, J. 1999. Nine major HLA class I supertypes account

for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50(3-4):201. (Cited on page 8.)

[40] Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S. and Brunak, S. 2004. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics.* 55(12):797. (Cited on page 8.)

[41] Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H. and Holzhutter, H.-G. 2005. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci* 62(9):1025. (Cited on pages 8, 9, 44, 46, 47, 50, 56, 58, 59, and 63.)

[42] Gaczynska, M., Rock, K. L. and Goldberg, A. L. 1993. Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes. *Nature* 365(6443):264. (Cited on pages 8, 43, and 45.)

[43] Kesmir, C., Van Noort, V., De Boer, R. J. and Hogeweg, P. 2003. Bioinformatic analysis of functional differences between the immunoproteasome and the constitutive proteasome. *Immunogenetics.* 55(7):437. (Cited on pages 8, 43, and 45.)

[44] Rao, X., Costa, A. I., Van Baarle, D. and Kesmir, C. 2009. A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant $CD8^+$ T cell responses. *J. Immunol.* 182(3):1526. (Cited on pages 9, 21, 22, and 31.)

[45] Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. and Stevanovic, S. 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50(3-4):213. (Cited on pages 9, 26, 48, 55, and 63.)

[46] Buus, S., Lauemoller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A. and Brunak, S. 2003. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens.* 62(5):378. (Cited on pages 9, 33, 34, and 102.)

[47] Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S. L., Lamberth, K., Buus, S., Brunak, S. and Lund, O. 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 12(5):1007. (Cited on pages 9, 33, 34, and 102.)

[48] Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O. and Nielsen, M. 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res* 36(Web Server issue):W509. (Cited on pages 9 and 56.)

[49] Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S. and Nielsen, M. 2009. NetMHCpan, a method for MHC class I binding

prediction beyond humans. *Immunogenetics* 61(1):1. (Cited on pages 9, 30, 33, 34, 44, 63, 102, and 104.)

[50] Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., Wilson, S. S., Sidney, J., Lund, O., Buus, S. and Sette, A. 2006. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol* 2(6):e65. (Cited on pages 9, 22, 33, 63, 78, 90, 92, 101, 122, and 135.)

[51] Nielsen, M., Lundegaard, C., Lund, O. and Kesmir, C. 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics.* 57(1-2):33. (Cited on pages 9, 34, 44, 56, 90, 92, and 101.)

[52] Peters, B., Bulik, S., Tampe, R., Endert, P. M. V. and Holzhutter, H.-G. 2003. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J Immunol* 171(4):1741. (Cited on pages 9, 47, 53, 56, and 63.)

[53] Hoof, I., van Baarle, D., Hildebrand, W. H. and Kesmir, C. 2012. Proteome Sampling by the HLA Class I Antigen Processing Pathway. *PLoS Comput Biol* 8(5):e1002517. (Cited on pages 9 and 136.)

[54] Blattman, J. N., Antia, R., Sourdive, D. J. D., Wang, X., Kaech, S. M., Murali-Krishna, K., Altman, J. D. and Ahmed, R. 2002. Estimating the precursor frequency of naive antigen-specific CD8 T cells. *J Exp Med* 195(5):657. (Cited on pages 10, 14, 63, 97, 99, 113, and 137.)

[55] Hataye, J., Moon, J. J., Khoruts, A., Reilly, C. and Jenkins, M. K. 2006. Naive and memory CD4+ T cell survival controlled by clonal abundance. *Science* 312(5770):114. (Cited on pages 10, 63, 97, 99, and 113.)

[56] Kotturi, M. F., Scott, I., Wolfe, T., Peters, B., Sidney, J., Cheroutre, H., von Herrath, M. G., Buchmeier, M. J., Grey, H. and Sette, A. 2008. Naive precursor frequencies and MHC binding rather than the degree of epitope diversity shape CD8+ T cell immunodominance. *J Immunol* 181(3):2124. (Cited on pages 10, 15, 63, 97, 99, 113, and 136.)

[57] Ishizuka, J., Grebe, K., Shenderov, E., Peters, B., Chen, Q., Peng, Y., Wang, L., Dong, T., Pasquetto, V., Oseroff, C., Sidney, J., Hickman, H., Cerundolo, V., Sette, A., Bennink, J. R., McMichael, A. and Yewdell, J. W. 2009. Quantitating T cell cross-reactivity for unrelated peptide antigens. *J Immunol* 183(7):4337. (Cited on pages 10, 63, 97, 99, and 113.)

[58] Davis, M. M. and Bjorkman, P. J. 1988. T-cell antigen receptor genes and T-cell recognition. *Nature* 334(6181):395. (Cited on pages 10 and 11.)

[59] Carpenter, A. C. and Bosselut, R. 2010. Decision checkpoints in the thymus. *Nat Immunol* 11(8):666. (Cited on page 10.)

[60] Rudolph, M. G., Stanfield, R. L. and Wilson, I. A. 2006. How TCRs bind

MHCs, peptides, and coreceptors. *Annu Rev Immunol* 24:419. (Cited on pages 10, 11, 13, 69, 76, 77, 89, and 94.)

[61] Garcia, K. C., Adams, J. J., Feng, D. and Ely, L. K. 2009. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol* 10(2):143. (Cited on pages 10 and 77.)

[62] Janeway, C. A., Immunobiology: the immune system in health and disease. Garland Science Publishing, 2005, 6th edn. (Cited on page 10.)

[63] Morris, G. P. and Allen, P. M. 2012. How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nat Immunol* 13(2):121. (Cited on pages 11 and 116.)

[64] Kurts, C., Kosaka, H., Carbone, F. R., Miller, J. F. and Heath, W. R. 1997. Class I-restricted cross-presentation of exogenous self-antigens leads to deletion of autoreactive CD8(+) T cells. *J Exp Med* 186(2):239. (Cited on pages 11 and 116.)

[65] Kurts, C., Robinson, B. W. S. and Knolle, P. A. 2010. Cross-priming in health and disease. *Nat Rev Immunol* 10(6):403. (Cited on pages 11, 16, and 116.)

[66] Tanchot, C., Lemonnier, F. A., Perarnau, B., Freitas, A. A. and Rocha, B. 1997. Differential requirements for survival and proliferation of CD8 naive or memory T cells. *Science* 276(5321):2057. (Cited on page 11.)

[67] Freitas, A. A. and Rocha, B. 1999. Peripheral T cell survival. *Curr Opin Immunol* 11(2):152. (Cited on page 11.)

[68] Hogquist, K. A., Jameson, S. C., Heath, W. R., Howard, J. L., Bevan, M. J. and Carbone, F. R. 1994. T cell receptor antagonist peptides induce positive selection. *Cell* 76(1):17. (Cited on page 11.)

[69] Lo, W.-L., Felix, N. J., Walters, J. J., Rohrs, H., Gross, M. L. and Allen, P. M. 2009. An endogenous peptide positively selects and augments the activation and survival of peripheral CD4+ T cells. *Nat Immunol* 10(11):1155. (Cited on page 11.)

[70] Ebert, P. J. R., Jiang, S., Xie, J., Li, Q.-J. and Davis, M. M. 2009. An endogenous positively selecting peptide enhances mature T cell responses and becomes an autoantigen in the absence of microRNA miR-181a. *Nat Immunol* 10(11):1162. (Cited on page 11.)

[71] Teh, H. S., Kisielow, P., Scott, B., Kishi, H., Uematsu, Y., Bluthmann, H. and von Boehmer, H. 1988. Thymic major histocompatibility complex antigens and the alpha beta T-cell receptor determine the CD4/CD8 phenotype of T cells. *Nature* 335(6187):229. (Cited on page 11.)

[72] Murata, S., Takahama, Y. and Tanaka, K. 2008. Thymoproteasome: probable role in generating positively selecting peptides. *Curr Opin Immunol* 20(2):192. (Cited on page 11.)

[73] Honey, K., Nakagawa, T., Peters, C. and Rudensky, A. 2002. Cathepsin L regulates CD4+ T cell selection independently of its effect on invariant chain: a role in the generation of positively selecting peptide ligands. *J Exp Med* 195(10):1349. (Cited on page 11.)

[74] Murata, S., Sasaki, K., Kishimoto, T., Niwa, S.-I., Hayashi, H., Takahama, Y. and Tanaka, K. 2007. Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science* 316(5829):1349. (Cited on pages 11 and 43.)

[75] Anderson, M. S. and Su, M. A. 2011. Aire and T cell development. *Curr Opin Immunol* 23(2):198. (Cited on pages 11 and 16.)

[76] Anderson, M. S., Venanzi, E. S., Klein, L., Chen, Z., Berzins, S. P., Turley, S. J., von Boehmer, H., Bronson, R., Dierich, A., Benoist, C. and Mathis, D. 2002. Projection of an immunological self shadow within the thymus by the aire protein. *Science* 298(5597):1395. (Cited on page 11.)

[77] Gough, S. C. L. and Simmonds, M. J. 2007. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genomics* 8(7):453. (Cited on pages 11, 16, 17, 101, and 116.)

[78] Howson, J. M. M., Walker, N. M., Clayton, D., Todd, J. A. and Consortium, T. . D. G. 2009. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes Obes Metab* 11 Suppl 1:31. (Cited on pages 11, 89, and 116.)

[79] Pociot, F. and McDermott, M. F. 2002. Genetics of type 1 diabetes mellitus. *Genes Immun* 3(5):235. (Cited on pages 11 and 89.)

[80] Toma, A., Haddouk, S., Briand, J.-P., Camoin, L., Gahery, H., Connan, F., Dubois-Laforgue, D., Caillat-Zucman, S., Guillet, J.-G., Carel, J.-C., Muller, S., Choppin, J. and Boitard, C. 2005. Recognition of a subregion of human proinsulin by class I-restricted T cells in type 1 diabetic patients. *Proc Natl Acad Sci U S A* 102(30):10581. (Cited on pages 11, 89, and 116.)

[81] Shortman, K., Vremec, D. and Egerton, M. 1991. The kinetics of T cell antigen receptor expression by subgroups of CD4+8+ thymocytes: delineation of CD4+8+3(2+) thymocytes as post-selection intermediates leading to mature T cells. *J Exp Med* 173(2):323. (Cited on page 11.)

[82] Egerton, M., Scollay, R. and Shortman, K. 1990. Kinetics of mature T-cell development in the thymus. *Proc Natl Acad Sci U S A* 87(7):2579. (Cited on page 11.)

[83] Merkenschlager, M., Graf, D., Lovatt, M., Bommhardt, U., Zamoyska, R. and Fisher, A. G. 1997. How many thymocytes audition for selection? *J Exp Med* 186(7):1149. (Cited on page 12.)

[84] van Meerwijk, J. P., Marguerat, S., Lees, R. K., Germain, R. N., Fowlkes,

B. J. and MacDonald, H. R. 1997. Quantitative impact of thymic clonal deletion on the T cell repertoire. *J Exp Med* 185(3):377. (Cited on pages 12 and 14.)

[85] Quigley, M. F., Greenaway, H. Y., Venturi, V., Lindsay, R., Quinn, K. M., Seder, R. A., Douek, D. C., Davenport, M. P. and Price, D. A. 2010. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci U S A* 107(45):19414. (Cited on pages 12, 15, and 77.)

[86] Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J. R. and Holt, R. A. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* 21(5):790. (Cited on page 12.)

[87] Venturi, V., Kedzierska, K., Price, D. A., Doherty, P. C., Douek, D. C., Turner, S. J. and Davenport, M. P. 2006. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci U S A* 103(49):18691. (Cited on pages 12 and 15.)

[88] Pham, H.-P., Manuel, M., Petit, N., Klatzmann, D., Cohen-Kaminsky, S., Six, A. and Marodon, G. 2011. Half of the T-cell repertoire combinatorial diversity is genetically determined in humans and humanized mice. *Eur J Immunol* . (Cited on pages 12 and 77.)

[89] Melenhorst, J. J., Lay, M. D. H., Price, D. A., Adams, S. D., Zeilah, J., Sosa, E., Hensel, N. F., Follmann, D., Douek, D. C., Davenport, M. P. and Barrett, A. J. 2008. Contribution of TCR-beta locus and HLA to the shape of the mature human Vbeta repertoire. *J Immunol* 180(10):6484. (Cited on page 12.)

[90] Fuschiotti, P., Pasqual, N., Hierle, V., Borel, E., London, J., Marche, P. N. and Jouvin-Marche, E. 2007. Analysis of the TCR alpha-chain rearrangement profile in human T lymphocytes. *Mol Immunol* 44(13):3380. (Cited on pages 12 and 77.)

[91] Houston, E. G. and Fink, P. J. 2009. MHC drives TCR repertoire shaping, but not maturation, in recent thymic emigrants. *J Immunol* 183(11):7244. (Cited on pages 12 and 77.)

[92] Balamurugan, A., Ng, H. L. and Yang, O. O. 2010. Rapid T cell receptor delineation reveals clonal expansion limitation of the magnitude of the HIV-1-specific CD8+ T cell response. *J Immunol* 185(10):5935. (Cited on page 13.)

[93] Gebe, J. A., Yue, B. B., Unrath, K. A., Falk, B. A. and Nepom, G. T. 2009. Restricted autoantigen recognition associated with deletional and adaptive regulatory mechanisms. *J Immunol* 183(1):59. (Cited on pages 13, 64, and 117.)

[94] Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R. M., Hudson, J. R., Davis, R. W. and Han, J. 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci U S A* 107(4):1518. (Cited on page 13.)

[95] Nguyen, P., Liu, W., Ma, J., Manirarora, J. N., Liu, X., Cheng, C. and Geiger, T. L. 2010. Discrete TCR repertoires and CDR3 features distinguish effector and Foxp3+ regulatory T lymphocytes in myelin oligodendrocyte glycoprotein-induced experimental allergic encephalomyelitis. *J Immunol* 185(7):3895. (Cited on page 13.)

[96] Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H. and Carlson, C. S. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114(19):4099. (Cited on page 13.)

[97] Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J. and Kourilsky, P. 1999. A direct estimate of the human alphabeta T cell receptor diversity. *Science* 286(5441):958. (Cited on page 13.)

[98] Kesmir, C., Borghans, J. A. and de Boer, R. J. 2000. Diversity of Human alpha beta T Cell Receptors. *Science* 288(5469):1135. (Cited on page 13.)

[99] Dash, P., McClaren, J. L., Oguin, T. H., Rothwell, W., Todd, B., Morris, M. Y., Becksfort, J., Reynolds, C., Brown, S. A., Doherty, P. C. and Thomas, P. G. 2011. Paired analysis of TCRa and TCRb chains at the single-cell level in mice. *J Clin Invest* 121(1):288. (Cited on page 13.)

[100] Borghans, J. A. and de Boer, R. J. 1998. Crossreactivity of the T-cell receptor. *Immunol Today* 19(9):428. (Cited on pages 13 and 97.)

[101] Moon, J. J., Chu, H. H., Pepper, M., McSorley, S. J., Jameson, S. C., Kedl, R. M. and Jenkins, M. K. 2007. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* 27(2):203. (Cited on pages 14, 15, 136, and 137.)

[102] Hadrup, S. R., Bakker, A. H., Shu, C. J., Andersen, R. S., van Veluw, J., Hombrink, P., Castermans, E., Straten, P. T., Blank, C., Haanen, J. B., Heemskerk, M. H. and Schumacher, T. N. 2009. Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. *Nat Methods* 6(7):520. (Cited on pages 14 and 137.)

[103] Hombrink, P., Hadrup, S. R., Bakker, A., Kester, M. G. D., Falkenburg, J. H. F., von dem Borne, P. A., Schumacher, T. N. M. and Heemskerk, M. H. M. 2011. High-throughput identification of potential minor histocompatibility antigens by MHC tetramer-based screening: feasibility and limitations. *PLoS One* 6(8):e22523. (Cited on pages 14 and 137.)

[104] Obar, J. J., Khanna, K. M. and Lefrancois, L. 2008. Endogenous naive CD8+ T cell precursor frequency regulates primary and memory responses

to infection. *Immunity* 28(6):859. (Cited on pages 14, 15, 63, and 136.)

[105] Legoux, F., Debeaupuis, E., Echasserieau, K., Salle, H. D. L., Saulquin, X. and Bonneville, M. 2010. Impact of TCR reactivity and HLA phenotype on naive CD8 T cell frequency in humans. *J Immunol* 184(12):6731. (Cited on pages 14 and 77.)

[106] Huseby, E. S., White, J., Crawford, F., Vass, T., Becker, D., Pinilla, C., Marrack, P. and Kappler, J. W. 2005. How the T cell repertoire becomes peptide and MHC specific. *Cell* 122(2):247. (Cited on pages 14, 21, 77, and 89.)

[107] Kosmrlj, A., Jha, A. K., Huseby, E. S., Kardar, M. and Chakraborty, A. K. 2008. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci U S A* 105(43):16671. (Cited on page 14.)

[108] Kotturi, M. F., Botten, J., Sidney, J., Bui, H.-H., Giancola, L., Maybeno, M., Babin, J., Oseroff, C., Pasquetto, V., Greenbaum, J. A., Peters, B., Ting, J., Do, D., Vang, L., Alexander, J., Grey, H., Buchmeier, M. J. and Sette, A. 2009. A multivalent and cross-protective vaccine strategy against arenaviruses associated with human disease. *PLoS Pathog* 5(12):e1000695. (Cited on pages 15, 63, 65, and 78.)

[109] Venturi, V., Chin, H. Y., Asher, T. E., Ladell, K., Scheinberg, P., Bornstein, E., van Bockel, D., Kelleher, A. D., Douek, D. C., Price, D. A. and Davenport, M. P. 2008. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. *J Immunol* 181(11):7853. (Cited on page 15.)

[110] Miles, J. J., Bulek, A. M., Cole, D. K., Gostick, E., Schauenburg, A. J. A., Dolton, G., Venturi, V., Davenport, M. P., Tan, M. P., Burrows, S. R., Wooldridge, L., Price, D. A., Rizkallah, P. J. and Sewell, A. K. 2010. Genetic and structural basis for selection of a ubiquitous T cell receptor deployed in Epstein-Barr virus infection. *PLoS Pathog* 6(11):e1001198. (Cited on pages 15 and 105.)

[111] Borghans, J. A., Noest, A. J. and de Boer, R. J. 1999. How specific should immunological memory be? *J Immunol* 163(2):569. (Cited on pages 15 and 113.)

[112] Tung, C.-W., Ziehm, M., Kamper, A., Kohlbacher, O. and Ho, S.-Y. 2011. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* 12(1):446. (Cited on pages 15 and 76.)

[113] Jeffery, K. J., Siddiqui, A. A., Bunce, M., Lloyd, A. L., Vine, A. M., Witkover, A. D., Izumo, S., Usuku, K., Welsh, K. I., Osame, M. and Bangham, C. R. 2000. The influence of HLA class I alleles and heterozygosity on the outcome of human T cell lymphotropic virus type I infection. *J Immunol* 165(12):7278. (Cited on pages 16, 17, 100, and 116.)

[114] Kuniholm, M. H., Kovacs, A., Gao, X., Xue, X., Marti, D., Thio, C. L., Peters, M. G., Terrault, N. A., Greenblatt, R. M., Goedert, J. J., Cohen, M. H.,

Minkoff, H., Gange, S. J., Anastos, K., Fazzari, M., Harris, T. G., Young, M. A., Strickler, H. D. and Carrington, M. 2010. Specific human leukocyte antigen class I and II alleles associated with hepatitis C virus viremia. *Hepatology* 51(5):1514. (Cited on pages 16 and 17.)

[115] Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., de Bakker, P. I. W., Walker, B. D., Ripke, S., Brumme, C. J., Pulit, S. L., Carrington, M., Kadie, C. M., Carlson, J. M., Heckerman, D., Graham, R. R., Plenge, R. M., Deeks, S. G., Gianniny, L., Crawford, G., Sullivan, J., Gonzalez, E., Davies, L., Camargo, A., Moore, J. M., Beattie, N., Gupta, S., Crenshaw, A. and Others. 2010. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 330(6010):1551. (Cited on pages 16 and 17.)

[116] Shevach, E. M. 2011. Biological functions of regulatory T cells. *Adv Immunol* 112:137. (Cited on pages 16, 116, and 117.)

[117] Rolland, M., Nickle, D. C., Deng, W., Frahm, N., Brander, C., Learn, G. H., Heckerman, D., Jojic, N., Jojic, V., Walker, B. D. and Mullins, J. I. 2007. Recognition of HIV-1 peptides by host CTL is related to HIV-1 similarity to human proteins. *PLoS One* 2(9):e823. (Cited on pages 16, 17, 64, 78, 89, 99, 100, and 137.)

[118] Frankild, S., De Boer, R. J., Lund, O., Nielsen, M. and Kesmir, C. 2008. Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLoS. ONE.* 3(3):e1831. (Cited on pages 16, 17, 64, 69, 75, 78, 89, 94, 95, 96, 97, 98, 100, 103, 104, and 137.)

[119] Nakanishi, Y., Lu, B., Gerard, C. and Iwasaki, A. 2009. CD8(+) T lymphocyte mobilization to virus-infected tissue requires CD4(+) T-cell help. *Nature* 462(7272):510. (Cited on pages 16, 116, and 137.)

[120] de Boer, R. J. and Perelson, A. S. 1994. T cell repertoires and competitive exclusion. *J Theor Biol* 169(4):375. (Cited on pages 16, 64, and 116.)

[121] Zhang, N. and Bevan, M. J. 2011. CD8(+) T cells: foot soldiers of the immune system. *Immunity* 35(2):161. (Cited on pages 16, 64, and 116.)

[122] Boon, A. C. M., de Mutsert, G., Graus, Y. M. F., Fouchier, R. A. M., Sintnicolaas, K., Osterhaus, A. D. M. E. and Rimmelzwaan, G. F. 2002. The magnitude and specificity of influenza A virus-specific cytotoxic T-lymphocyte responses in humans is related to HLA-A and -B phenotype. *J Virol* 76(2):582. (Cited on page 16.)

[123] Lacey, S. F., Villacres, M. C., Rosa, C. L., Wang, Z., Longmate, J., Martinez, J., Brewer, J. C., Mekhoubad, S., Maas, R., Leedom, J. M., Forman, S. J., Zaia, J. A. and Diamond, D. J. 2003. Relative dominance of HLA-B*07 restricted CD8+ T-lymphocyte immune responses to human cytomegalovirus pp65 in persons sharing HLA-A*02 and HLA-B*07 alleles. *Hum Immunol* 64(4):440. (Cited on page 16.)

[124] Kiepiela, P., Leslie, A. J., Honeyborne, I., Ramduth, D., Thobakgale, C.,

Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K. J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M. M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L. D., Szinger, J., Day, C., Klenerman, P., Mullins, J., Korber, B., Coovadia, H. M., Walker, B. D. and Goulder, P. J. R. 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 432(7018):769. (Cited on pages 16 and 31.)

[125] Lewinsohn, D. A., Winata, E., Swarbrick, G. M., Tanner, K. E., Cook, M. S., Null, M. D., Cansler, M. E., Sette, A., Sidney, J. and Lewinsohn, D. M. 2007. Immunodominant tuberculosis CD8 antigens preferentially restricted by HLA-B. *PLoS Pathog* 3(9):1240. (Cited on pages 16 and 31.)

[126] Bihl, F., Frahm, N., Giammarino, L. D., Sidney, J., John, M., Yusim, K., Woodberry, T., Sango, K., Hewitt, H. S., Henry, L., Linde, C. H., Chisholm, J. V., Zaman, T. M., Pae, E., Mallal, S., Walker, B. D., Sette, A., Korber, B. T., Heckerman, D. and Brander, C. 2006. Impact of HLA-B alleles, epitope binding affinity, functional avidity, and viral coinfection on the immunodominance of virus-specific CTL responses. *J Immunol* 176(7):4094. (Cited on pages 16, 17, and 31.)

[127] Oseroff, C., Kos, F., Bui, H.-H., Peters, B., Pasquetto, V., Glenn, J., Palmore, T., Sidney, J., Tscharke, D. C., Bennink, J. R., Southwood, S., Grey, H. M., Yewdell, J. W. and Sette, A. 2005. HLA class I-restricted responses to vaccinia recognize a broad array of proteins mainly involved in virulence and viral gene regulation. *Proc Natl Acad Sci U S A* 102(39):13980. (Cited on pages 16 and 63.)

[128] Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H.-H., Grey, H. and Sette, A. 2006. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat Biotechnol* 24(7):817. (Cited on pages 16 and 63.)

[129] Althaus, C. L. and de Boer, R. J. 2011. Implications of CTL-mediated killing of HIV-infected cells during the non-productive stage of infection. *PLoS One* 6(2):e16468. (Cited on page 16.)

[130] Hansen, T. H. and Bouvier, M. 2009. MHC class I antigen presentation: learning from viral evasion strategies. *Nat Rev Immunol* 9(7):503. (Cited on page 16.)

[131] Perez, C. L., Larsen, M. V., Gustafsson, R., Norstrom, M. M., Atlas, A., Nixon, D. F., Nielsen, M., Lund, O. and Karlsson, A. C. 2008. Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. *J Immunol* 180(7):5092. (Cited on pages 17, 98, 104, and 117.)

[132] Textor, J. 2011. Search and Learning in the Immune System: Models of Immune Surveillance and Negative Selection. Ph.D. thesis, University of Lubeck. (Cited on page 17.)

[133] Vyas, J. M., der Veen, A. G. V. and Ploegh, H. L. 2008. The known unknowns of antigen processing and presentation. *Nat Rev Immunol* 8(8):607. (Cited on page 21.)

[134] Groothuis, T. A. M., Griekspoor, A. C., Neijssen, J. J., Herberts, C. A. and Neefjes, J. J. 2005. MHC class I alleles and their exploration of the antigen-processing machinery. *Immunol Rev* 207:60. (Cited on page 21.)

[135] Craiu, A., Akopian, T., Goldberg, A. and Rock, K. L. 1997. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc Natl Acad Sci U S A* 94(20):10850. (Cited on page 21.)

[136] Falk, K., Rotzschke, O. and Rammensee, H. G. 1990. Cellular peptide composition governed by major histocompatibility complex class I molecules. *Nature* 348(6298):248. (Cited on page 21.)

[137] Paulsson, K. M. 2004. Evolutionary and functional perspectives of the major histocompatibility complex class I antigen-processing machinery. *Cell Mol Life Sci* 61(19-20):2446. (Cited on page 21.)

[138] Yewdell, J. W. 2001. Not such a dismal science: the economics of protein synthesis, folding, degradation and antigen processing. *Trends Cell Biol* 11(7):294. (Cited on page 21.)

[139] Yewdell, J. W., Reits, E. and Neefjes, J. 2003. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat Rev Immunol* 3(12):952. (Cited on page 21.)

[140] Prugnolle, F., Manica, A., Charpentier, M., Guesgan, J. F., Guernier, V. and Balloux, F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol* 15(11):1022. (Cited on page 21.)

[141] Parham, P. and Ohta, T. 1996. Population biology of antigen presentation by MHC class I molecules. *Science* 272(5258):67. (Cited on pages 21, 31, and 134.)

[142] Hughes, A. L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335(6186):167. (Cited on page 21.)

[143] Schellens, I. M., Kesmir, C., Miedema, F., Van Baarle, D. and Borghans, J. A. 2008. An unanticipated lack of consensus cytotoxic T lymphocyte epitopes in HIV-1 databases: the contribution of prediction programs. *AIDS* 22(1):33. (Cited on pages 21, 22, and 63.)

[144] Vider-Shalit, T., Sarid, R., Maman, K., Tsaban, L., Levi, R. and Louzoun, Y. 2009. Viruses selectively mutate their CD8+ T-cell epitopes–a large-scale immunomic analysis. *Bioinformatics* 25(12):i39. (Cited on pages 21 and 32.)

[145] Schmid, B., Kesmir, C. and De Boer, R. J. 2008. The specificity and poly-

morphism of the MHC class I prevents the global adaptation of HIV-1 to the monomorphic proteasome and TAP. *PLoS. ONE.* 3(10):e3525. (Cited on page 21.)

[146] Almani, M., Raffaeli, S., Vider-Shalit, T., Tsaban, L., Fishbain, V. and Louzoun, Y. 2009. Human self-protein CD8+ T-cell epitopes are both positively and negatively selected. *Eur J Immunol* 39(4):1056. (Cited on page 22.)

[147] Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O. and Nielsen, M. 2007. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* 8:424. (Cited on pages 22, 33, 56, 63, 78, 90, 92, 101, and 135.)

[148] Lundegaard, C., Nielsen, M. and Lund, O. 2006. The validity of predicted T-cell epitopes. *Trends Biotechnol* 24(12):537. (Cited on page 22.)

[149] Rocha, E. P. C. and Danchin, A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet* 18(6):291. (Cited on pages 22, 24, 27, 31, and 100.)

[150] Bentley, S. D. and Parkhill, J. 2004. Comparative genomic structure of prokaryotes. *Annu Rev Genet* 38:771. (Cited on pages 24 and 27.)

[151] Singer, G. A. and Hickey, D. A. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* 17(11):1581. (Cited on pages 25 and 27.)

[152] Muto, A. and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* 84(1):166. (Cited on page 25.)

[153] Knight, R. D., Freeland, S. J. and Landweber, L. F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* 2(4):RESEARCH0010. (Cited on page 25.)

[154] Otting, N., Heijmans, C. M. C., Noort, R. C., de Groot, N. G., Doxiadis, G. G. M., van Rood, J. J., Watkins, D. I. and Bontrop, R. E. 2005. Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc Natl Acad Sci U S A* 102(5):1626. (Cited on page 30.)

[155] Nurnberger, T., Brunner, F., Kemmerling, B. and Piater, L. 2004. Innate immunity in plants and animals: striking similarities and obvious differences. *Immunol Rev* 198:249. (Cited on page 31.)

[156] Fukami-Kobayashi, K., Shiina, T., Anzai, T., Sano, K., Yamazaki, M., Inoko, H. and Tateno, Y. 2005. Genomic evolution of MHC class I region in primates. *Proc Natl Acad Sci U S A* 102(26):9230. (Cited on page 31.)

[157] Watkins, D. I., McAdam, S. N., Liu, X., Strang, C. R., Milford, E. L., Levine, C. G., Garber, T. L., Dogon, A. L., Lord, C. I. and Ghim, S. H. 1992. New

recombinant HLA-B alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. *Nature* 357(6376):329. (Cited on pages 31 and 134.)

[158] McAdam, S. N., Boyson, J. E., Liu, X., Garber, T. L., Hughes, A. L., Bontrop, R. E. and Watkins, D. I. 1994. A uniquely high level of recombination at the HLA-B locus. *Proc Natl Acad Sci U S A* 91(13):5893. (Cited on pages 31 and 134.)

[159] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. and Duroux, P. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* 37(Database issue):D1006. (Cited on pages 31 and 35.)

[160] Belich, M. P., Madrigal, J. A., Hildebrand, W. H., Zemmour, J., Williams, R. C., Luz, R., Petzl-Erler, M. L. and Parham, P. 1992. Unusual HLA-B alleles in two tribes of Brazilian Indians. *Nature* 357(6376):326. (Cited on pages 31 and 134.)

[161] Bauer, S., Pigisch, S., Hangel, D., Kaufmann, A. and Hamm, S. 2008. Recognition of nucleic acid and nucleic acid analogs by Toll-like receptors 7, 8 and 9. *Immunobiology* 213(3-4):315. (Cited on page 31.)

[162] Krieg, A. M., Yi, A. K., Matson, S., Waldschmidt, T. J., Bishop, G. A., Teasdale, R., Koretzky, G. A. and Klinman, D. M. 1995. CpG motifs in bacterial DNA trigger direct B-cell activation. *Nature* 374(6522):546. (Cited on page 31.)

[163] Peters, B., Tong, W., Sidney, J., Sette, A. and Weng, Z. 2003. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 19(14):1765. (Cited on pages 33, 34, 35, 63, and 102.)

[164] MacNamara, A., Kadolsky, U., Bangham, C. R. M. and Asquith, B. 2009. T-cell epitope prediction: rescaling can mask biological variation between MHC molecules. *PLoS Comput Biol* 5(3):e1000327. (Cited on pages 33, 44, 54, and 56.)

[165] Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O. and Buus, S. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* 2(8):e796. (Cited on pages 33, 35, and 63.)

[166] Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688. (Cited on page 35.)

[167] Hershko, A. and Ciechanover, A. 1992. The ubiquitin system for protein degradation. *Annu Rev Biochem* 61:761. (Cited on page 43.)

[168] Goldberg, A. L. 2003. Protein degradation and protection against misfolded or damaged proteins. *Nature* 426(6968):895. (Cited on page 43.)

[169] Seifert, U., Bialy, L. P., Ebstein, F., Bech-Otschir, D., Voigt, A., Schroter, F., Prozorovski, T., Lange, N., Steffen, J., Rieger, M., Kuckelkorn, U., Aktas, O., Kloetzel, P.-M. and Kruger, E. 2010. Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell* 142(4):613. (Cited on page 43.)

[170] van Leuken, R., Clijsters, L. and Wolthuis, R. 2008. To cell cycle, swing the APC/C. *Biochim Biophys Acta* 1786(1):49. (Cited on page 43.)

[171] Kloetzel, P. M. 2001. Antigen processing by the proteasome. *Nat Rev Mol Cell Biol* 2(3):179. (Cited on page 43.)

[172] Kloetzel, P. M. and Ossendorp, F. 2004. Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. *Curr Opin Immunol* 16(1):76. (Cited on pages 43, 47, and 53.)

[173] Guillaume, B., Chapiro, J., Stroobant, V., Colau, D., Holle, B. V., Parvizi, G., Bousquet-Dubouch, M.-P., Theate, I., Parmentier, N. and den Eynde, B. J. V. 2010. Two abundant proteasome subtypes that uniquely process some antigens presented by HLA class I molecules. *Proc Natl Acad Sci U S A* 107(43):18599. (Cited on pages 43 and 53.)

[174] Florea, B. I., Verdoes, M., Li, N., van der Linden, W. A., Geurink, P. P., van den Elst, H., Hofmann, T., de Ru, A., van Veelen, P. A., Tanaka, K., Sasaki, K., Murata, S., den Dulk, H., Brouwer, J., Ossendorp, F. A., Kisselev, A. F. and Overkleeft, H. S. 2010. Activity-based profiling reveals reactivity of the murine thymoproteasome-specific subunit beta5t. *Chem Biol* 17(8):795. (Cited on page 43.)

[175] Holzhutter, H. G., Frommel, C. and Kloetzel, P. M. 1999. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J Mol Biol* 286(4):1251. (Cited on page 44.)

[176] Holzhutter, H. G. and Kloetzel, P. M. 2000. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. *Biophys J* 79(3):1196. (Cited on page 44.)

[177] Ginodi, I., Vider-Shalit, T., Tsaban, L. and Louzoun, Y. 2008. Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics* 24(4):477. (Cited on page 44.)

[178] Kuttler, C., Nussbaum, A. K., Dick, T. P., Rammensee, H. G., Schild, H. and Hadeler, K. P. 2000. An algorithm for the prediction of proteasomal cleavages. *J Mol Biol* 298(3):417. (Cited on page 44.)

[179] Nussbaum, A. K., Kuttler, C., Hadeler, K. P., Rammensee, H. G. and Schild, H. 2001. PAProC: a prediction algorithm for proteasomal cleavages avail-

able on the WWW. *Immunogenetics* 53(2):87. (Cited on page 44.)

[180] Stranzl, T., Larsen, M. V., Lundegaard, C. and Nielsen, M. 2010. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62(6):357. (Cited on pages 44, 54, and 135.)

[181] Vider-Shalit, T., Almani, M., Sarid, R. and Louzoun, Y. 2009. The HIV hide and seek game: an immunogenomic analysis of the HIV epitope repertoire. *AIDS* 23(11):1311. (Cited on page 44.)

[182] Kim, Y., Sette, A. and Peters, B. 2011. Applications for T-cell epitope queries and tools in the Immune Epitope Database and Analysis Resource. *J Immunol Methods* 374(1-2):62. (Cited on page 44.)

[183] Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857):1285. (Cited on pages 45 and 57.)

[184] Saxova, P., Buus, S., Brunak, S. and Kesmir, C. 2003. Predicting proteasomal cleavage sites: a comparison of available methods. *Int. Immunol.* 15(7):781. (Cited on page 46.)

[185] Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A. and Peters, B. 2010. The immune epitope database 2.0. *Nucleic Acids Res* 38(Database issue):D854. (Cited on pages 48, 55, 65, 78, 79, 98, and 104.)

[186] Seifert, U., Maranon, C., Shmueli, A., Desoutter, J.-F., Wesoloski, L., Janek, K., Henklein, P., Diescher, S., Andrieu, M., de la Salle, H., Weinschenk, T., Schild, H., Laderach, D., Galy, A., Haas, G., Kloetzel, P.-M., Reiss, Y. and Hosmalin, A. 2003. An essential role for tripeptidyl peptidase in the generation of an MHC class I epitope. *Nat Immunol* 4(4):375. (Cited on pages 47 and 53.)

[187] Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405(2):442. (Cited on pages 49 and 57.)

[188] de Graaf, N., van Helden, M. J. G., Textoris-Taube, K., Chiba, T., Topham, D. J., Kloetzel, P.-M., Zaiss, D. M. W. and Sijts, A. J. A. M. 2011. PA28 and the proteasome immunosubunits play a central and independent role in the production of MHC class I-binding peptides in vivo. *Eur J Immunol* 41(4):926. (Cited on page 53.)

[189] Peters, B., Janek, K., Kuckelkorn, U. and Holzhutter, H.-G. 2002. Assessment of proteasomal cleavage probabilities from kinetic analysis of time-dependent product formation. *J Mol Biol* 318(3):847. (Cited on pages 55 and 60.)

[190] R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. (Cited on pages 58 and 69.)

[191] Alanio, C., Lemaitre, F., Law, H. K. W., Hasan, M. and Albert, M. L. 2010. Enumeration of human antigen-specific naive CD8+ T cells reveals conserved precursor frequencies. *Blood* 115(18):3718. (Cited on pages 63 and 77.)

[192] Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S. and Lund, O. 2004. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 20(9):1388. (Cited on page 63.)

[193] Schuler, M. M., Nastke, M.-D. and Stevanovikc, S. 2007. SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol* 409:75. (Cited on page 63.)

[194] Sijts, E. J. A. M. and Kloetzel, P. M. 2011. The role of the proteasome in the generation of MHC class I ligands and immune responses. *Cell Mol Life Sci* 68(9):1491. (Cited on page 63.)

[195] Gallimore, A., Dumrese, T., Hengartner, H., Zinkernagel, R. M. and Rammensee, H. G. 1998. Protective immunity does not correlate with the hierarchy of virus-specific cytotoxic T cell responses to naturally processed peptides. *J Exp Med* 187(10):1647. (Cited on page 63.)

[196] Pang, K. C., Sanders, M. T., Monaco, J. J., Doherty, P. C., Turner, S. J. and Chen, W. 2006. Immunoproteasome subunit deficiencies impact differentially on two immunodominant influenza virus-specific CD8+ T cell responses. *J Immunol* 177(11):7680. (Cited on page 63.)

[197] Tenzer, S., Wee, E., Burgevin, A., Stewart-Jones, G., Friis, L., Lamberth, K., hao Chang, C., Harndahl, M., Weimershaus, M., Gerstoft, J., Akkad, N., Klenerman, P., Fugger, L., Jones, E. Y., McMichael, A. J., Buus, S., Schild, H., van Endert, P. and Iversen, A. K. N. 2009. Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance. *Nat Immunol* 10(6):636. (Cited on page 63.)

[198] Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., Oseroff, C., Yuan, L., Ruppert, J., Sidney, J., del Guercio, M. F., Southwood, S., Kubo, R. T., Chesnut, R. W., Grey, H. M. and Chisari, F. V. 1994. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J Immunol* 153(12):5586. (Cited on page 63.)

[199] Lazarski, C. A., Chaves, F. A., Jenks, S. A., Wu, S., Richards, K. A., Weaver, J. M. and Sant, A. J. 2005. The kinetic stability of MHC class II:peptide complexes is a key parameter that dictates immunodominance. *Immunity* 23(1):29. (Cited on page 63.)

[200] Gruta, N. L. L., Kedzierska, K., Pang, K., Webby, R., Davenport, M., Chen, W., Turner, S. J. and Doherty, P. C. 2006. A virus-specific CD8+ T cell immunodominance hierarchy determined by antigen dose and precursor

frequencies. *Proc Natl Acad Sci U S A* 103(4):994. (Cited on page 63.)

[201] Chen, W., Norbury, C. C., Cho, Y., Yewdell, J. W. and Bennink, J. R. 2001. Immunoproteasomes shape immunodominance hierarchies of antiviral CD8(+) T cells at the levels of T cell repertoire and presentation of viral antigens. *J Exp Med* 193(11):1319. (Cited on page 63.)

[202] Crowe, S. R., Turner, S. J., Miller, S. C., Roberts, A. D., Rappolo, R. A., Doherty, P. C., Ely, K. H. and Woodland, D. L. 2003. Differential antigen presentation regulates the changing patterns of CD8+ T cell immunodominance in primary and secondary influenza virus infections. *J Exp Med* 198(3):399. (Cited on page 63.)

[203] Calis, J. J. A., de Boer, R. J. and Kesmir, C. 2012. Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLoS Comput Biol* 8(3):e1002412. (Cited on pages 64, 69, 70, 76, 78, 80, 113, 114, and 115.)

[204] Storey, J. D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100(16):9440. (Cited on pages 65, 67, and 85.)

[205] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. and Kanehisa, M. 2008. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36(Database issue):D202. (Cited on pages 67 and 85.)

[206] Wucherpfennig, K. W., Call, M. J., Deng, L. and Mariuzza, R. 2009. Structural alterations in peptide-MHC recognition by self-reactive T cell receptors. *Curr Opin Immunol* 21(6):590. (Cited on pages 69, 76, 89, 94, and 97.)

[207] Hausmann, S., Biddison, W. E., Smith, K. J., Ding, Y. H., Garboczi, D. N., Utz, U., Wiley, D. C. and Wucherpfennig, K. W. 1999. Peptide recognition by two HLA-A2/Tax11-19-specific T cell clones in relationship to their MHC/peptide/TCR crystal structures. *J Immunol* 162(9):5389. (Cited on pages 69, 75, 89, and 94.)

[208] Lee, J. K., Stewart-Jones, G., Dong, T., Harlos, K., Gleria, K. D., Dorrell, L., Douek, D. C., van der Merwe, P. A., Jones, E. Y. and McMichael, A. J. 2004. T cell cross-reactivity and conformational changes during TCR engagement. *J Exp Med* 200(11):1455. (Cited on pages 69, 75, 89, and 94.)

[209] Boggiano, C., Moya, R., Pinilla, C., Bihl, F., Brander, C., Sidney, J., Sette, A. and Blondelle, S. E. 2005. Discovery and characterization of highly immunogenic and broadly recognized mimics of the HIV-1 CTL epitope Gag77-85. *Eur J Immunol* 35(5):1428. (Cited on pages 69, 75, 89, and 94.)

[210] Tynan, F. E., Elhassen, D., Purcell, A. W., Burrows, J. M., Borg, N. A., Miles, J. J., Williamson, N. A., Green, K. J., Tellam, J., Kjer-Nielsen, L., McCluskey, J., Rossjohn, J. and Burrows, S. R. 2005. The immunogenicity of a viral

cytotoxic T cell epitope is controlled by its MHC-bound conformation. *J Exp Med* 202(9):1249. (Cited on pages 69, 75, 89, 94, 95, and 97.)

[211] Hoof, I., Perez, C. L., Buggert, M., Gustafsson, R. K. L., Nielsen, M., Lund, O. and Karlsson, A. C. 2010. Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity. *J Immunol* 184(9):5383. (Cited on pages 69, 75, 89, 94, and 95.)

[212] Clarkson, D. B., Fan, Y.-a. and Joe, H. 1993. A remark on algorithm 643: FEXACT: an algorithm for performing Fisher's exact test in r x c contingency tables. *ACM Trans. Math. Softw.* 19(4):484. (Cited on page 69.)

[213] Calis, J. J. A., Sanchez-Perez, G. F. and Kesmir, C. 2010. MHC class I molecules exploit the low G+C content of pathogen genomes for enhanced presentation. *Eur J Immunol* 40(10):2699. (Cited on pages 70, 89, 92, 93, and 100.)

[214] Weiskopf, D., Yauch, L. E., Angelo, M. A., John, D. V., Greenbaum, J. A., Sidney, J., Kolla, R. V., Silva, A. D. D., de Silva, A. M., Grey, H., Peters, B., Shresta, S. and Sette, A. 2011. Insights into HLA-restricted T cell responses in a novel mouse model of dengue virus infection point toward new implications for vaccine design. *J Immunol* 187(8):4268. (Cited on pages 71, 73, 76, and 80.)

[215] Kotturi, M. F., Assarsson, E., Peters, B., Grey, H., Oseroff, C., Pasquetto, V. and Sette, A. 2009. Of mice and humans: how good are HLA transgenic mice as a model of human immune responses? *Immunome Res* 5:3. (Cited on page 74.)

[216] Kessels, H. W. H. G., de Visser, K. E., Tirion, F. H., Coccoris, M., Kruisbeek, A. M. and Schumacher, T. N. M. 2004. The impact of self-tolerance on the polyclonal CD8+ T cell repertoire. *J Immunol* 172(4):2324. (Cited on page 75.)

[217] Scott-Browne, J. P., White, J., Kappler, J. W., Gapin, L. and Marrack, P. 2009. Germline-encoded amino acids in the alphabeta T-cell receptor control thymic selection. *Nature* 458(7241):1043. (Cited on page 77.)

[218] Li, L.-P., Lampert, J. C., Chen, X., Leitao, C., Popovic, J., Muller, W. and Blankenstein, T. 2010. Transgenic mice with a diverse human T cell antigen receptor repertoire. *Nat Med* 16(9):1029. (Cited on page 77.)

[219] Alexander, J., Sidney, J., Southwood, S., Ruppert, J., Oseroff, C., Maewal, A., Snoke, K., Serra, H. M., Kubo, R. T. and Sette, A. 1994. Development of high potency universal DR-restricted helper epitopes by modification of high affinity DR-blocking peptides. *Immunity* 1(9):751. (Cited on page 77.)

[220] Sette, A., Sidney, J., Livingston, B. D., Dzuris, J. L., Crimi, C., Walker, C. M., Southwood, S., Collins, E. J. and Hughes, A. L. 2003. Class I molecules with similar peptide-binding specificities are the result of both common

ancestry and convergent evolution. *Immunogenetics* 54(12):830. (Cited on page 77.)

[221] Basta, S. and Bennink, J. R. 2003. A survival game of hide and seek: cytomegaloviruses and MHC class I antigen presentation pathways. *Viral Immunol* 16(3):231. (Cited on page 77.)

[222] Fischer, W., Ganusov, V. V., Giorgi, E. E., Hraber, P. T., Keele, B. F., Leitner, T., Han, C. S., Gleasner, C. D., Green, L., Lo, C.-C., Nag, A., Wallstrom, T. C., Wang, S., McMichael, A. J., Haynes, B. F., Hahn, B. H., Perelson, A. S., Borrow, P., Shaw, G. M., Bhattacharya, T. and Korber, B. T. 2010. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* 5(8):e12303. (Cited on page 77.)

[223] Braud, V. M., McMichael, A. J. and Cerundolo, V. 1998. Differential processing of influenza nucleoprotein in human and mouse cells. *Eur J Immunol* 28(2):625. (Cited on page 79.)

[224] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389. (Cited on page 79.)

[225] Brynedal, B., Duvefelt, K., Jonasdottir, G., Roos, I. M., Akesson, E., Palmgren, J. and Hillert, J. 2007. HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. *PLoS One* 2(7):e664. (Cited on page 89.)

[226] McDole, J., Johnson, A. J. and Pirko, I. 2006. The role of CD8+ T-cells in lesion formation and axonal dysfunction in multiple sclerosis. *Neurol Res* 28(3):256. (Cited on page 89.)

[227] Wolfl, M., Rutebemberwa, A., Mosbruger, T., Mao, Q., mei Li, H., Netski, D., Ray, S. C., Pardoll, D., Sidney, J., Sette, A., Allen, T., Kuntzen, T., Kavanagh, D. G., Kuball, J., Greenberg, P. D. and Cox, A. L. 2008. Hepatitis C virus immune escape via exploitation of a hole in the T cell repertoire. *J Immunol* 181(9):6435. (Cited on pages 89, 99, and 100.)

[228] Welsh, R. M., Che, J. W., Brehm, M. A. and Selin, L. K. 2010. Heterologous immunity between viruses. *Immunol Rev* 235(1):244. (Cited on pages 89, 94, 95, and 97.)

[229] Cole, D. K., Edwards, E. S. J., Wynn, K. K., Clement, M., Miles, J. J., Ladell, K., Ekeruche, J., Gostick, E., Adams, K. J., Skowera, A., Peakman, M., Wooldridge, L., Price, D. A. and Sewell, A. K. 2010. Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. *J Immunol* 185(4):2600. (Cited on pages 89 and 97.)

[230] Huang, D. W., Sherman, B. T. and Lempicki, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.

Nat Protoc 4(1):44. (Cited on pages 91 and 107.)

[231] Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4(5):P3. (Cited on pages 91 and 107.)

[232] Embley, T. M. and Martin, W. 2006. Eukaryotic evolution, changes and challenges. *Nature* 440(7084):623. (Cited on page 91.)

[233] Bakker, A. B., van der Burg, S. H., Huijbens, R. J., Drijfhout, J. W., Melief, C. J., Adema, G. J. and Figdor, C. G. 1997. Analogues of CTL epitopes with improved MHC class-I binding capacity elicit anti-melanoma CTL recognizing the wild-type epitope. *Int J Cancer* 70(3):302. (Cited on pages 94 and 97.)

[234] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res* 28(1):235. (Cited on pages 94 and 105.)

[235] Kim, Y., Sidney, J., Pinilla, C., Sette, A. and Peters, B. 2009. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* 10:394. (Cited on pages 96 and 103.)

[236] Banwell, B., Krupp, L., Kennedy, J., Tellier, R., Tenembaum, S., Ness, J., Belman, A., Boiko, A., Bykova, O., Waubant, E., Mah, J. K., Stoian, C., Kremenchutzky, M., Bardini, M. R., Ruggieri, M., Rensel, M., Hahn, J., Weinstock-Guttman, B., Yeh, E. A., Farrell, K., Freedman, M., Iivanainen, M., Sevon, M., Bhan, V., Dilenge, M.-E., Stephens, D. and Bar-Or, A. 2007. Clinical features and viral serologies in children with multiple sclerosis: a multinational observational study. *Lancet Neurol* 6(9):773. (Cited on page 100.)

[237] MHC, I., Network, A. G., Rioux, J. D., Goyette, P., Vyse, T. J., Hammarstrom, L., Fernando, M. M. A., Green, T., Jager, P. L. D., Foisy, S., Wang, J., de Bakker, P. I. W., Leslie, S., McVean, G., Padyukov, L., Alfredsson, L., Annese, V., Hafler, D. A., Pan-Hammarstrom, Q., Matell, R., Sawcer, S. J., Compston, A. D., Cree, B. A. C., Mirel, D. B., Daly, M. J., Behrens, T. W., Klareskog, L., Gregersen, P. K., Oksenberg, J. R. and Hauser, S. L. 2009. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc Natl Acad Sci U S A* 106(44):18680. (Cited on page 100.)

[238] Karosiene, E., Lundegaard, C., Lund, O. and Nielsen, M. 2011. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* . (Cited on page 102.)

[239] Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. and Wiley, D. C. 1996. Structure of the complex between human T-cell receptor, viral

peptide and HLA-A2. *Nature* 384(6605):134. (Cited on page 105.)

[240] Ding, Y. H., Smith, K. J., Garboczi, D. N., Utz, U., Biddison, W. E. and Wiley, D. C. 1998. Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. *Immunity* 8(4):403. (Cited on page 105.)

[241] Buslepp, J., Wang, H., Biddison, W. E., Appella, E. and Collins, E. J. 2003. A correlation between TCR Valpha docking on MHC and CD8 dependence: implications for T cell selection. *Immunity* 19(4):595. (Cited on page 105.)

[242] Kjer-Nielsen, L., Clements, C. S., Purcell, A. W., Brooks, A. G., Whisstock, J. C., Burrows, S. R., McCluskey, J. and Rossjohn, J. 2003. A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity. *Immunity* 18(1):53. (Cited on page 105.)

[243] Hoare, H. L., Sullivan, L. C., Pietra, G., Clements, C. S., Lee, E. J., Ely, L. K., Beddoe, T., Falco, M., Kjer-Nielsen, L., Reid, H. H., McCluskey, J., Moretta, L., Rossjohn, J. and Brooks, A. G. 2006. Structural basis for a major histocompatibility complex class Ib-restricted T cell response. *Nat Immunol* 7(3):256. (Cited on page 105.)

[244] Dunn, S. M., Rizkallah, P. J., Baston, E., Mahon, T., Cameron, B., Moysey, R., Gao, F., Sami, M., Boulter, J., Li, Y. and Jakobsen, B. K. 2006. Directed evolution of human T cell receptor CDR2 residues by phage display dramatically enhances affinity for cognate peptide-MHC without increasing apparent cross-reactivity. *Protein Sci* 15(4):710. (Cited on page 105.)

[245] Gras, S., Saulquin, X., Reiser, J.-B., Debeaupuis, E., Echasserieau, K., Kissenpfennig, A., Legoux, F., Chouquet, A., Gorrec, M. L., Machillot, P., Neveu, B., Thielens, N., Malissen, B., Bonneville, M. and Housset, D. 2009. Structural bases for the affinity-driven selection of a public TCR against a dominant human cytomegalovirus epitope. *J Immunol* 183(1):430. (Cited on page 105.)

[246] Macdonald, W. A., Chen, Z., Gras, S., Archbold, J. K., Tynan, F. E., Clements, C. S., Bharadwaj, M., Kjer-Nielsen, L., Saunders, P. M., Wilce, M. C. J., Crawford, F., Stadinsky, B., Jackson, D., Brooks, A. G., Purcell, A. W., Kappler, J. W., Burrows, S. R., Rossjohn, J. and McCluskey, J. 2009. T cell allorecognition via molecular mimicry. *Immunity* 31(6):897. (Cited on page 105.)

[247] Robb, R. J., Lineburg, K. E., Kuns, R. D., Wilson, Y. A., Raffelt, N. C., Olver, S. D., Varelias, A., Alexander, K. A., Teal, B. E., Sparwasser, T., Hammerling, G. J., Markey, K. A., Koyama, M., Clouston, A. D., Engwerda, C. R., Hill, G. R. and Macdonald, K. P. A. 2012. Identification and expansion of highly suppressive CD8+FoxP3+ regulatory T cells after experimental allogeneic bone marrow transplantation. *Blood* . (Cited on page 116.)

[248] Oldstone, M. B., Nerenberg, M., Southern, P., Price, J. and Lewicki,

H. 1991. Virus infection triggers insulin-dependent diabetes mellitus in a transgenic model: role of anti-self (virus) immune response. *Cell* 65(2):319. (Cited on page 116.)

[249] Gold, R., Hartung, H. P. and Lassmann, H. 1997. T-cell apoptosis in autoimmune diseases: termination of inflammation in the nervous system and other sites with specialized immune-defense mechanisms. *Trends Neurosci* 20(9):399. (Cited on page 116.)

[250] van Montfoort, N., Camps, M. G., Khan, S., Filippov, D. V., Weterings, J. J., Griffith, J. M., Geuze, H. J., van Hall, T., Verbeek, J. S., Melief, C. J. and Ossendorp, F. 2009. Antigen storage compartments in mature dendritic cells facilitate prolonged cytotoxic T lymphocyte cross-priming capacity. *Proc Natl Acad Sci U S A* 106(16):6730. (Cited on page 117.)

[251] Opelz, G., Wujciak, T., Dohler, B., Scherer, S. and Mytilineos, J. 1999. HLA compatibility and organ transplant survival. Collaborative Transplant Study. *Rev Immunogenet* 1(3):334. (Cited on page 121.)

[252] Terasaki, P. I. and Cai, J. 2008. Human leukocyte antigen antibodies and chronic rejection: from association to causation. *Transplantation* 86(3):377. (Cited on page 121.)

[253] Duquesnoy, R. J. 2002. HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. I. Description of the algorithm. *Hum Immunol* 63(5):339. (Cited on page 121.)

[254] Duquesnoy, R. J. 2006. A structurally based approach to determine HLA compatibility at the humoral immune level. *Hum Immunol* 67(11):847. (Cited on page 121.)

[255] Duquesnoy, R. J. 2011. Antibody-reactive epitope determination with HLAMatchmaker and its clinical applications. *Tissue Antigens* 77(6):525. (Cited on page 121.)

[256] Dankers, M. K. A., Witvliet, M. D., Roelen, D. L., de Lange, P., Korfage, N., Persijn, G. G., Duquesnoy, R., Doxiadis, I. I. N. and Claas, F. H. J. 2004. The number of amino acid triplet differences between patient and donor is predictive for the antibody reactivity against mismatched human leukocyte antigens. *Transplantation* 77(8):1236. (Cited on page 121.)

[257] Duquesnoy, R. J., Takemoto, S., de Lange, P., Doxiadis, I. I. N., Schreuder, G. M. T., Persijn, G. G. and Claas, F. H. J. 2003. HLAmatchmaker: a molecularly based algorithm for histocompatibility determination. III. Effect of matching at the HLA-A,B amino acid triplet level on kidney transplant survival. *Transplantation* 75(6):884. (Cited on page 121.)

[258] Haririan, A., Fagoaga, O., Daneshvar, H., Morawski, K., Sillix, D. H., El-Amm, J. M., West, M. S., Garnick, J., Migdal, S. D., Gruber, S. A. and Nehlsen-Cannarella, S. 2006. Predictive value of human leucocyte antigen epitope matching using HLAMatchmaker for graft outcomes in a predomin-

antly African-American renal transplant cohort. *Clin Transplant* 20(2):226. (Cited on page 121.)

[259] Laux, G., Mytilineos, J. and Opelz, G. 2004. Critical evaluation of the amino acid triplet-epitope matching concept in cadaver kidney transplantation. *Transplantation* 77(6):902. (Cited on page 121.)

[260] Dankers, M. K. A., Roelen, D. L., Meer-Prins, E. M. W. V. D., Lange, P. D., Korfage, N., Smits, J. M. A., Persijn, G. G., Welsh, K. I., Doxiadis, I. I. N. and Claas, F. H. J. 2003. Differential immunogenicity of HLA mismatches: HLA-A2 versus HLA-A28. *Transplantation* 75(3):418. (Cited on page 121.)

[261] Dankers, M. K. A., Roelen, D. L., Nagelkerke, N. J. D., de Lange, P., Persijn, G. G., Doxiadis, I. I. N. and Claas, F. H. J. 2004. The HLA-DR phenotype of the responder is predictive of humoral response against HLA class I antigens. *Hum Immunol* 65(1):13. (Cited on pages 121, 122, and 126.)

[262] Maruya, E., Takemoto, S. and Terasaki, P. I. 1993. HLA matching: identification of permissible HLA mismatches. *Clin Transpl* :511. (Cited on page 121.)

[263] Fuller, T. C. and Fuller, A. 1999. The humoral immune response against an HLA class I allodeterminant correlates with the HLA-DR phenotype of the responder. *Transplantation* 68(2):173. (Cited on pages 121 and 126.)

[264] Papassavas, A. C., Barnardo, M. C. N. M., Bunce, M. and Welsh, K. I. 2002. Is there MHC Class II restriction of the response to MHC Class I in transplant patients? *Transplantation* 73(4):642. (Cited on pages 121 and 126.)

[265] Suciu-Foca, N., Liu, Z., Harris, P. E., Reed, E. F., Cohen, D. J., Benstein, J. A., Benvenisty, A. I., Mancini, D., Michler, R. E. and Rose, E. A. 1995. Indirect recognition of native HLA alloantigens and B-cell help. *Transplant Proc* 27(1):455. (Cited on page 122.)

[266] Claas, F. H. J. 2002. Predictive parameters for in vivo alloreactivity. *Transpl Immunol* 10(2-3):137. (Cited on page 122.)

[267] Nielsen, M., Lund, O., Buus, S. and Lundegaard, C. 2010. MHC class II epitope predictive algorithms. *Immunology* 130(3):319. (Cited on page 122.)

[268] Nielsen, M., Lundegaard, C. and Lund, O. 2007. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 8:238. (Cited on pages 122 and 130.)

[269] Nielsen, M. and Lund, O. 2009. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 10:296. (Cited on pages 122 and 130.)

[270] Nielsen, M., Justesen, S., Lund, O., Lundegaard, C. and Buus, S. 2010. NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res* 6:9. (Cited on page 128.)

[271] Southwood, S., Sidney, J., Kondo, A., del Guercio, M. F., Appella, E., Hoffman, S., Kubo, R. T., Chesnut, R. W., Grey, H. M. and Sette, A. 1998. Several common HLA-DR types share largely overlapping peptide binding repertoires. *J Immunol* 160(7):3363. (Cited on page 131.)

[272] Duquesnoy, R. J. and Marrari, M. 2009. Correlations between Terasaki's HLA class I epitopes and HLAMatchmaker-defined eplets on HLA-A, -B and -C antigens. *Tissue Antigens* 74(2):117. (Cited on page 131.)

[273] Levasseur, A. and Pontarotti, P. 2010. Was the ancestral MHC involved in innate immunity? *Eur J Immunol* 40(10):2682. (Cited on page 134.)

[274] Vider-Shalit, T., Fishbain, V., Raffaeli, S. and Louzoun, Y. 2007. Phase-dependent immune evasion of herpesviruses. *J Virol* 81(17):9536. (Cited on page 134.)

[275] Hertz, T., Nolan, D., James, I., John, M., Gaudieri, S., Phillips, E., Huang, J. C., Riadi, G., Mallal, S. and Jojic, N. 2011. Mapping the landscape of host-pathogen coevolution: HLA class I binding and its relationship with evolutionary conservation in human and viral proteins. *J Virol* 85(3):1310. (Cited on pages 134 and 135.)

[276] Katzman, S., Capra, J. A., Haussler, D. and Pollard, K. S. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol* 3:614. (Cited on page 134.)

[277] Granados, D. P., Yahyaoui, W., Laumont, C. M., Daouda, T., Muratore-Schroeder, T. L., Cote, C., Laverdure, J.-P., Lemieux, S., Thibault, P. and Perreault, C. 2012. MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements. *Blood* . (Cited on page 135.)

[278] von Roretz, C., Marco, S. D., Mazroui, R. and Gallouzi, I.-E. 2011. Turnover of AU-rich-containing mRNAs during stress: a matter of survival. *Wiley Interdiscip Rev RNA* 2(3):336. (Cited on page 135.)

Samenvatting

Het lichaam wordt elke dag blootgesteld aan virussen, bacteriën en andere gevaren, het immuunsysteem moet hier op een adequate manier op reageren. Sommige gevaren kunnen op een directe manier worden aangepakt omdat ze zich manifesteren tussen de lichaamseigen cellen, bacteriën in een snijwond zijn hiervan een voorbeeld. Andere gevaren, zoals virale infecties, spelen zich in de lichaamscel af (intracellulair). Om intracellulaire gevaren te kunnen detecteren maakt het immuunsysteem gebruik van MHC-I moleculen waarop eiwitfragmenten (peptiden) gepresenteerd worden. In elke cel in ons lichaam worden eiwitten gemaakt en ook weer afgebroken. Peptiden zijn een bijproduct van de afbraak van eiwitten. Sommige peptiden kunnen binden aan MHC-I moleculen en vormen een peptide-MHC-I-complex (pMHC), elke cel presenteert deze complexen op het celoppervlakte. De pMHCs geven zo een beeld van de processen die zich in de cel afspelen.
Als een cel door een virus is geïnfecteerd, worden er ook virale peptiden geproduceerd, die als pMHC op het celoppervlakte gepresenteerd worden. T-cellen zijn gespecialiseerde cellen van het immuunsysteem die een interactie aangaan met specifieke pMHCs. Ze kunnen een specifiek viraal peptide herkennen en die herkenning gebruiken om geïnfecteerde cellen op te sporen en op te ruimen. Wanneer een pMHC gebruikt wordt voor een specifieke immuunrespons noemen we het een epitoop. Niet alle virale peptiden die op MHC-I moleculen worden gepresenteerd zijn epitopen. In deze thesis beschrijven we ons onderzoek naar de verschillende factoren die bepalen welke peptiden epitopen zijn, aan de hand van drie vragen:

1. Welke peptiden worden gepresenteerd op MHC-I moleculen?

In ons onderzoek bouwen we voort op eerder uitgevoerde studies aan het MHC-I presentatie systeem. Op basis van deze studies zijn voorspellers gemaakt die

kunnen voorspellen hoe eiwitten worden afgebroken in peptiden, welke peptiden beschikbaar zijn voor MHC-I binding en hoe sterk de interactie tussen verschillende MHC-I moleculen en peptiden is. In hoofdstuk 2 beschrijven we ons onderzoek naar de peptide voorkeuren van verschillende MHC-I moleculen. Sommige MHC-I moleculen zijn beter in het binden van peptiden die afkomstig zijn van pathogenen dan in het binden van humane peptiden. Wij laten zien dat dit komt doordat pathogenen een lagere G+C fractie hebben in hun genomen. De lagere G+C fractie leidt tot een veranderd gebruik van bepaalde aminozuren, waarvoor de MHC-I moleculen een voorkeur hebben. In hoofdstuk 3 beschrijven we ons onderzoek naar de optimale voorspelling van MHC-I gepresenteerde peptiden. We laten zien hoe de verschillende voorspellers voor eiwitafbraak, peptide transport en MHC-I-peptide binding het beste kunnen worden gecombineerd om te voorspellen welke peptiden worden gepresenteerd.

2. Welke pMHCs worden herkend door T-cellen?

Als er geen T-cellen zijn die een bepaald pMHC herkennen, dan zal het niet als epitoop kunnen worden gebruikt. De mate waarin een pMHC wordt herkend noemen we de immunogeniciteit van een pMHC. In hoofstuk 4 beschrijven we ons onderzoek naar de factoren die de immunogeniciteit van pMHCs beïnvloeden. Voor ons onderzoek selecteren we uit verschillende bronnen een set pMHCs die wel herkend wordt door T-cellen en een set waarbij dit niet het geval is. Wanneer we de twee sets met elkaar vergelijken zien we dat bepaalde aminozuren leiden tot een betere herkenning, en dat bepaalde posities in de gepresenteerde peptiden belangrijker zijn voor de immunogeniciteit. We brengen deze resultaten samen in een model waarmee we de immunogeniciteit van nieuwe pMHCs kunnen voorspellen.

3. Welke herkende pMHCs worden gebruikt in de immuunrespons?

Niet elk viraal peptide dat wordt gepresenteerd op MHC-I en waarvoor er herkennende T-cellen zijn, is een epitoop. Een factor die hierbij van invloed is, is de mate waarin een viraal peptide lijkt op de lichaamseigen peptiden. Als het virale peptide gelijk is aan, of sterk lijkt op een lichaamseigen peptide, dan zou een immuunrespons hiertegen ook gericht kunnen worden op gezonde cellen die het lichaamseigen peptide presenteren. Hierdoor zou een auto-immuunziekte kunnen worden veroorzaakt. Lichaamsvreemde peptiden die teveel lijken op lichaamseigen peptiden noemen we zelf-gelijkend. We hebben onderzocht hoe groot de invloed van zelf-gelijkenis is op het opzetten van immuunresponsen, deze stu-

die staat beschreven in hoofdstuk 5. Uit ons onderzoek blijkt dat zelf-gelijkenis een grote invloed heeft, aangezien ∼30% van de lichaamsvreemde MHC-I gepresenteerde peptiden veel lijkt op een lichaamseigen peptide dat op MHC-I gepresenteerd kan worden. In een addendum bij hoofdstuk 5 beschrijven we in welke mate de grote zelf-gelijkenis druk uitoefent op het immuunsysteem, om T-cellen die lichaamseigen peptiden herkennen te onderdrukken. Meer dan 95% van deze zelf-herkennende T-cellen moet verwijderd worden om auto-immuniteit te voorkomen. Of zelf-gelijkenis ook een belangrijke rol speelt bij het opzetten van andere immuunresponsen bestuderen we tenslotte voor niertransplantaties, waarbij immuunreacties ongewenst zijn omdat ze tot afstoting kunnen leiden. In hoofdstuk 6 laten we zien dat de kans op een immuunrespons tegen een lichaamsvreemd MHC-I molecuul in de donornier groter is als er meer peptiden in zitten die niet zelf-gelijkend zijn.

Conclusies

Het correct voorspellen van epitopen is belangrijk voor een beter begrip van de T-cel immuunrespons en daarmee voor een beter begrip van het immuunsysteem. Wanneer we dit beter begrijpen kunnen we betere vaccins ontwikkelen, betere transplantaties uitvoeren, en beter begrijpen hoe pathogenen de gezondheid van individuën en populaties beïnvloeden. Op basis van de onderzoekingen die hier beschreven zijn en vroegere bevindingen, kunnen we goed voorspellen welke peptiden epitopen zijn, ∼50% van de voorspelde epitopen waren correct. In de toekomst zal er meer data beschikbaar komen die de immunogeniciteit van verschillende pMHCs beschrijft. Tevens kunnen er meer pMHCs worden ontdekt in zogenaamde peptide-elutie studies, op basis waarvan de eiwitten die als peptide op MHC-I worden gepresenteerd beter kunnen worden bestudeerd. Wij laten zien dat met deze data het voorspellen van epitopen verder verbeterd kan worden.

Curriculum vitæ

Jorg Justin Aimé Calis was born 10^{th} of March 1983 in Hilversum, the Netherlands. From 1995 to 2001 he attended Alberdingk Thijm College in Hilversum, where he obtained his VWO diploma. He studied biomedical sciences in Utrecht, and followed the Master's programme Cancer Genomics and Developmental Biology. His first internship in this master was at the Functional Genomics group in the Hubrecht Laboratory for Developmental Biology in Utrecht, supervised by dr. M. Tijsterman and prof. dr. R.H.A. Plasterk, studying transposon silencing in *C. elegans*. His second internship was at the Theoretical Biology and Bioinformatics group of Utrecht University, supervised by dr. V.V. Ganusov and prof. dr. Rob J. de Boer, studying the immune response to SIV/HIV. He received his Master of Science degree in February 2008. Subsequently, he started his PhD research on MHC-I presentation and epitope predictions in the Theoretical Biology and Bioinformatics group of Utrecht University, supervised by dr. C. Keşmir and prof. dr. Rob J. de Boer. During a three month period starting June 2011, Jorg conducted his research at the La Jolla Institute for Allergy and Immunology research in the Vaccine Discovery group of dr. B. Peters and prof. dr. A. Sette. The results of his PhD research are described in this thesis.

List of Publications

Jorg J. A. Calis, Matt Maybeno, Jason A. Greenbaum, Daniela Weiskopf, Aruna D. De Silva, Alessandro Sette, Can Keşmir & Bjoern Peters Properties of MHC class I presented peptides that enhance immunogenicity. **Submitted** (2012)

Henny G. Otten*, **Jorg J. A. Calis***, **Can Keşmir, Arjan D. van Zuilen & Eric Spierings** De novo development of donor-specific HLA IgG antibodies after kidney transplantation is facilitated by donor HLA derived T-helper epitopes. **Submitted** (2012)
*Contributed equally

Jorg J. A. Calis, Rob J. de Boer & Can Keşmir Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLoS Comput Biol* 8(3):e1002412 (2012)

Jorg J. A. Calis, Gabino F. Sanchez-Perez & Can Keşmir MHC class I molecules exploit the low G+C content of pathogen genomes for enhanced presentation. *Eur J Immunol* 40: 2699–2709 (2010)
• Received commentary: **Anthony Levasseur & Pierre Pontarotti** Was the ancestral MHC involved in innate immunity? *Eur J Immunol* 40: 2682–2685 (2010)

Dankwoord

Wanneer je een roos uit elkaar haalt en de losse onderdelen onderzoekt zul je een beter begrip krijgen van de bouw. Hoe alle blaadjes, meeldraden en andere onderdelen samenkomen blijkt dan een ingewikkeld systeem te zijn dat interessante vragen opwerpt. Vaak leidt zo'n beter begrip tot een grotere waardering. Hoe alle hulp die ik van verschillende mensen tijdens mijn promotie heb gekregen samenkomt is niet in één beeld te vangen. Daarom heb ik besloten om de verschillende onderdelen van deze hulp te beschrijven, opdat dit tot een groter begrip en waardering leidt: jullie verdienen niets minder!

Can, het is een groot plezier om met je te werken en van je te leren. Ik had altijd weer zin om aan de slag te gaan met de nieuwe ideeën die ik opdeed tijdens onze meetings. Vaak kwam ik met een gecompliceerd verhaal aan wat jij kon samenvatten in één figuur, van jouw "denken in figuren" heb ik veel geleerd. Niet alleen op wetenschappelijk, maar ook op persoonlijk vlak stuur je je groep fantastisch aan. Ik denk dat je mij ook op positieve wijze hebt bijgestuurd waardoor ik beter en professioneler heb leren samenwerken.

Rob, als Can en ik het niet meer wisten konden we altijd nog naar jou toe, daarbij heb ik veel van je kunnen leren. Ondanks je drukke agenda was er altijd wel ergens een gaatje te vinden, en kon op een rustige manier het onderzoek besproken worden. Dat je dit voor vele projecten doet, en daarnaast ook het TBB-bootje veilig door reorganisatiestormen weet te loodsen vind ik bewonderenswaardig.

I want to thank the members of my reading committee for investing their valuable time and effort in reading my thesis: Ole Lund, David Price, Ton Schumacher, Willem van Eden and Ronald Bontrop.

During my PhD I had the pleasure of working together and discussing my work with people from outside our group. I gratefully acknowledge Bjoern Peters and Alex Sette from the LIAI in La Jolla, US, for hosting me for a very productive and interesting 3-month collaboration in California. Thanks to all the Viro-Immunomeeting members: especially José Borghans, Kiki Tesselaar, Maaike Ressing and Debbie van Baarle. Jeroen en alle anderen van de Bontrop-groep, bedankt voor goede discussies en fun tijdens de EFI-meetings.

Jop en Wytse, leuk dat jullie mijn paranimfen willen zijn. Mooi om te zien hoe verschillend jullie zijn, hetgeen waarschijnlijk dé basis van onze 11-jarige vriend-

schap is. Ik zat net te denken hoe jullie je talenten (Wytse's geheugen en Jop's schrijversschap) het beste zouden kunnen combineren: waarschijnlijk door mijn biografie te schrijven. Maar het zou ook een wetenschappelijk onderzoek naar de gezondheideffecten van Ethiopische koffie kunnen zijn. Een van de weinige punten waarop jullie niet verschillen betreft de voorkeur voor speciaalbier, waarin ik gelukkig ook kan delen. Misschien dat we dat gedrieën nog eens kunnen vieren in (Café) België.
Je staat er niet elke dag bij stil, maar eigenlijk is de TBB groep een perfecte groep om onderzoek te doen. Een goede mix van discussie over en interesse in elkaars onderzoek, maar ook de gezelligheid daarnaast maakt dat ik elke dag met veel plezier richting Utrecht trok. Allereerst wil ik hier Paulien en Ben voor bedanken, zonder wie de groep nooit had bestaan, ook bedankt voor de bioinformatica cursus die een unieke blik op biologie en evolutie geeft, en natuurlijk bedankt voor de groepsdiners op de Maliebaan. Van de vaste krachten bij de TBB wil ik Jan Kees bedanken voor het prachtige computersysteem waar we op mogen werken, Berend, Kirsten en Ronald bedankt voor jullie interesse en hulp.
Thanks for my direct colleagues in Can's group. Rao, never a dull moment when you are around, thanks for all your fresh views (sometimes misunderstandings) on Dutch society, and contributions to my research. Hanneke, bedankt voor al je bijdragen aan mijn onderzoek, en voor je goede verzorging van onze groep middels de verspreiding van zelf-gebakte cakejes en zelf-geteelde rucola. Ilka, numerous times I turned my chair to ask for your help, and you could always help me further. Your experience and helpfulness were great to have around, and have helped me a lot. I hope you keep coming to Utrecht every now and than, as it accelerates and improves our work tremendously, and adds a lot of fun as well.
Het is interessant om vijf jaar als student en aio te werken en een grote groep mensen te zien komen en gaan. Van de oude garde (zij die zijn gegaan) wil ik de (Midden-Oosten) kamergenoten Henk-Jan en Boris bedanken voor jullie (programmeer-)hulp, (verre en later dichtbijzijnde) koffietripjes naar het zuiden, en mooie discussies. Already some time ago, but nevertheless very influential was Vitaly, Vitaly thanks for supervising me as a student and for many discussions on immunology and how we can contribute to it. Many thanks also to Jos and Gabino from the bioinformatics group for practical help and discussion on bioinformatical problems. Folkert, je bent ondertussen ook lid van de oude garde. Ik vond het erg leuk om bijna elke dag koffie met je te drinken, volgens mij ben je mijn sterkere koffie steeds beter gaan waarderen.
Ondertussen is er ook een nieuwe garde (zij die er nog zijn). First of all, I'd like to thank all current members of the Immunology-room, Hanneke, Sai, Paola, Leila, Ioana, Bram (in order of appearance) for a great time, nice discussions and for taking me to the Gutenberg. Also other members of the Gutenberg-posse I'd like to thank for a daily escape from the nest, thanks to Adrian, Lidija, Michael, Renske, Johannes, Alessia, Eelco, Chris, Thomas, Like, Sandro, Daniel, Rutger (basically everybody, in order of willingness to go to the Gutenberg). Thomas, bedankt voor je interesse en gezelligheid sinds we in de studentenkamer naast elkaar zaten. Daniel, thanks for many lunches and coffee-breaks, it is always fun

chatting with you. Sai and Paola, thanks for spicing up the immunology-group as well as our groupdiners.
Joost, Klaartje, Levien en Ana, ik ga jullie niet in een hokje stoppen, maar wel bedanken voor jullie aanwezigheid tijdens mijn aio-tijd. Tenslotte wil ik ook Peter en Leon bedanken die een studenten project met mij hebben willen doen. Ik heb veel geleerd van het begeleiden van studenten, ik hoop dat jullie ook wat van mij konden leren.

Niet alleen collega's, maar ook andere mensen hebben indirect een invloed gehad op dit boekje. Al zullen ze niet direct bij het onderzoek betrokken zijn, ik weet zeker dat het resultaat anders was geweest zonder jullie aanwezigheid. Pap en mam, bedankt voor jullie rotsvaste vertrouwen, en voor het rustig beantwoorden van mijn nimmer aflatende vragenstroom de afgelopen 29 jaar. Merel, ik bewonder je eigenheid, inzichten en doorzettingsvermogen. Mikel, als goede vriend of (bijna-)familie, maakt eigenlijk niet uit, in beide gevallen valt je (droge) humor in goede aarde. Vele avonden en tripjes werden in wisselende samenstelling gevierd met: Tobias, gelukkig wordt je geloof in je zelf overtroffen door een goede portie zelfspot, je anti-pathie voor gezondheidsbevorderende producten is fenomenabel. Mannen (Jop, Wytse, Pleun, Bas, Bart, Joep, Tijs, Kees, Marcel), van "een serietje 6-8-10" tot een huttentocht in Zwitserland, eigenlijk maakt het niet uit wat we doen want de verhalen blijven sterk en de gesprekken zijn goed. Het bestuur met aanhang, waarmee de lol maar ook de agenda-drukte nu bijna groter is dan 8 jaar (!) geleden, tx voor alle etentjes. De semi-vrienden van VV: bedankt voor veel gezelligheid, feest en advies, in het bijzonder maar niet exclusief van Tom, Michiel, Sebas en beide Michels. Dan zijn er nog wat mensen die iets meer dan minder via Linda kwamen aanwaaien: Liset mijn collega-forens, gelukkig zien we elkaar vaker met een glas wijn in de hand dan in de trein; De Spaanse Pepers; Myrte, Eelke, Kim, Mieke en Sietske; Via het NKI, Chris en Sietske (ik hoop op een Californische wijnproeverij), Rik, Dalila en Marieke (opdat de (rook)dekentjes nog vaak (over/)in het Rembrandtpark worden uitgerold). Benjamin and Ruben (not related), even though we don't often meet, I can be certain it will be fun and interesting whenever/wherever that happens.

Tenslotte en meest belangrijk wil ik natuurlijk Linda bedanken. Lieve Linda, niets fascineert me meer dan bij jou te zijn. Op fantastische wijze breng je leven in de brouwerij, en zet je me aan tot actie. Ik kijk uit naar het verdere verloop van ons avontuur, met evenveel liefde, geluk en plezier als de afgelopen jaren.