European Journal of Immunology

MHC class I molecules exploit the low G+C content of pathogen genomes for enhanced presentation

Jorg J. A. Calis, Gabino F. Sanchez-Perez and Can Keşmir

Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, The Netherlands

Distinguishing self from nonself and pathogenic from nonpathogenic is a fundamental challenge to the immune system but whether adaptive immune systems use pathogen-specific signatures to achieve this is largely unknown. By investigating the presentation of large sets of viruses and bacteria on MHC class I molecules, we analyze whether MHC-I molecules have a preference for pathogen-derived peptides. The fraction of potential MHC-I binders in different organisms can vary up to eight-fold. We find that this variation can be largely explained by G+C content differences of the organisms, which are reflected in amino acid frequencies. A significant majority of HLA-A, but not HLA-B, molecules has a preference for peptides derived from organisms with a low G+C content. Interestingly, a low G+C content seems to be a universal signature for pathogenicity. Finally, we find the same preferences in chimpanzee and rhesus macaque MHC-I molecules. These results demonstrate that despite the fast evolution of MHC-I alleles and their extreme polymorphism and diversity in peptide-binding preferences, MHC-I molecules can acquire a preference to exploit pathogen-specific signatures.

Key words: G+C content · HLA alleles · MHC-I presentation · Self/nonself discrimination

See accompanying article by Levasseur and Pontarotti

Supporting Information available online

Introduction

MHC class I (MHC-I) presentation of peptides is crucial for the cytotoxic T-cell response. The process of MHC-I presentation 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 to form peptide–MHC-I complexes (pMHC) and pMHC-transport to and from the cell surface [1–5]. T cells recognizing foreign pMHC will become activated and proliferate to seek and destroy other cells presenting the same pMHC. Thus, the presentation of pathogen-derived peptides on MHC-I molecules is required to elicit an effective immune response. Self- and pathogen-derived peptides compete for presentation on the same MHC-I molecules. Given the enormous turnover of self-proteins (about 10^6 peptides are generated by the proteasome every second [6]) and limited number of MHC-I molecules ($\sim 10^5$), only a small fraction of all pathogen-derived peptides will be presented on the cell surface [7]. Another requirement for an effective immune response is that pathogen-derived pMHC are different from self-derived pMHC, since most self-reactive T-cells are tolerized during negative selection [8]. MHC-I molecules that prefer to bind pathogen-specific peptides will be able to fulfill these

Correspondence: Jorg J. A. Calis
e-mail: j.j.a.calis@uu.nl

requirements best and provide a selective advantage. However, such a preference to enhance the MHC-I presentation of pathogen-specific peptides has thus far not been described.

HLA molecules are encoded by the most polymorphic genes in the human population [9–11], and the more polymorphic sites are located in the parts that encode the peptide-binding groove [10, 12]. As a result, different HLA molecules bind different peptides. Most HLA-A/B molecules use the second and ninth position of a peptide as binding anchors. At these anchor positions, the specificity is highest. For example, HLA-A*0201 has a preference for peptides with a leucine at the second and valine at the ninth position, whereas HLA-B*5101 preferably binds peptides with a proline and isoleucine at those positions, respectively.

In the last decade, large sets of peptide–MHC-I-binding affinities have been measured and made available *via* databases such as SYFPEITI [13] or IEDB [14]. These data enabled the elucidation of peptide-binding rules for specific MHC-I molecules and the development of MHC-I pathway predictors. These predictors have been successful in identifying new HIV-1 epitopes for specific MHC-I molecules [15] and are used to investigate the preferences of MHC-I molecules. For instance, MHC-I molecules have been reported to preferably present self-peptides containing non-synonymous single nucleotide polymorphisms [16]. Also, human viruses have been reported to be presented less on HLA molecules as compared with nonhuman viruses [17]. Finally, we recently showed that HLA-A molecules share a preference for binding peptides derived from pathogens [18]; however, none of the described studies has come up with a mechanism explaining the observed preferences of MHC-I molecules. Due to the diversity of peptide-binding preferences, such a mechanism is hard to find.

In the search for a mechanism explaining the observed shared MHC-I preferences, we investigated the MHC-I presentation of large sets of viral and bacterial proteomes. The predicted fraction of MHC-I-presented peptides in these proteomes was determined using predictors for the key processes of the MHC-I presentation pathway: proteasomal degradation, TAP translocation and MHC-I binding [15, 19–21]. We find that MHC-I molecules display a large variation in the predicted fraction of presented peptides for different proteomes, and that up to 95% of this variation can be explained by differences in the genomic G+C content. Most interestingly, a majority of HLA-A, but not HLA-B, molecules prefers to present the peptides from species with a low G+C content. Since pathogenicity is associated with a low G+C content [22], these results suggest that HLA-A alleles have been selected to encode binding motifs that preferentially present pathogen-specific peptides. Analyzing experimentally verified epitopes, we confirm the preference of HLA-A molecules to present peptides enriched in amino acids encoded by a G+C low genome. Finally, we observe the same preference for low G+C contents in nonhuman primate MHC-I molecules. Taken together, we propose a novel feature of MHC class I presentation that provides a mechanism to explain how MHC-I enhances the presentation of pathogens.

Results

MHC class I molecules are responsive to differences in G+C content

Previously, we and others described that HLA molecules have preferences for peptides that are derived from pathogens [18]. To investigate the mechanisms behind these observations, we predicted for a large set of viruses (1223 nonredundant viral proteomes, see *Materials and methods* section) how well they are presented on different HLA class I molecules. We limited our analysis to HLA class I molecules for which two high-quality MHC-peptide-binding predictors were available, and therefore we had to exclude HLA-C molecules (explained in detail in *Materials and methods* section). The eleven HLA-A molecules and ten HLA-B molecules that fulfill the criteria are listed in Table 1

Table 1. HLA molecules and their G+C preferences[a)]

G+C-*positive* molecules	G+C-*negative* molecules		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)

[a)] Spearman rank correlation coefficients between G+C content and the predicted fractions of presented peptides are reported in brackets for each HLA molecule. For G+C-positive and G+C-negative HLA molecules, the correlation for different methods and parameter settings is significant (p<0.001) and consistent. HLA molecules indicated with an asterisk (*) showed inconsistent correlation values for different methods or parameter settings.

and are expressed in 88% and 52% of the Caucasian population, respectively. The fraction of presented peptides is the number of nonamer (9mer) peptides that is predicted to be presented on an HLA-A/B molecule relative to all 9mer in that virus (explained in detail in *Materials and methods* section). This fraction was predicted *per* virus and is a measure of how much a virus is preferably presented by an MHC-I molecule. The predicted fractions of presented peptides of viruses on HLA-A/B molecules vary greatly, up to 60-fold (results not shown). However, a large part of this variation is due to the small proteome of viruses: the variation within the 100 smallest viruses is 25-fold, whereas the variation within the largest 100 viruses is maximally eight-fold (Supporting Information Fig. S1). To circumvent this trivial source of variation, the analysis was repeated with bacterial species, which have larger proteomes than viruses. In a data set of 174 bacteria, the predicted fractions of presented peptides still vary largely, for some HLA molecules up to six-fold. For example, HLA-A*0101 has a strong preference for *Candidatus carsonella*, but not for *Anaeromyxobacter dehalogenans*, with predicted fractions of presented peptides of 3.7% and 0.8%, respectively.

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. Interestingly, this specific subset stands out by their low G+C contents, suggesting that part of the variation in predicted fractions of presented peptides is due to G+C contents. To test this hypothesis, the correlation between G+C content of a genome and its predicted fraction of presented peptides was examined. In Fig. 1, we show this correlation for six representative HLA molecules, every dot shows the predicted fraction of presented peptides and the G+C content of a bacterium (all tested HLA molecules ($n = 21$) are shown in Supporting Information Fig. S2). Clear G+C preferences can be observed, *e.g.* HLA-A*1101 has a preference for bacteria with a low G+C content, these are presented almost four times better than bacteria with a high G+C content. Conversely, HLA-B*0702 favors the presentation of bacteria with a high G+C content. To formalize these observations, an HLA molecule was termed G+C responsive if two requirements were fulfilled. First, a significant (Spearman rank test: $p<0.001$) correlation between G+C content and the predicted fraction of presented peptides within the set of 174 bacteria should exist. Second, the absolute value of the correlation coefficient should be high, *i.e.* >0.7. HLA molecules that do not fulfill both requirements are called G+C neutral, whereas G+C-responsive HLA molecules are called G+C positive (correlation coefficient>0.7) or G+C negative (correlation coefficient<-0.7). The majority of the HLA molecules studied here (14 out of 21) were G+C responsive: ten G+C negative and four G+C positive (Table 1). This result was consistent for

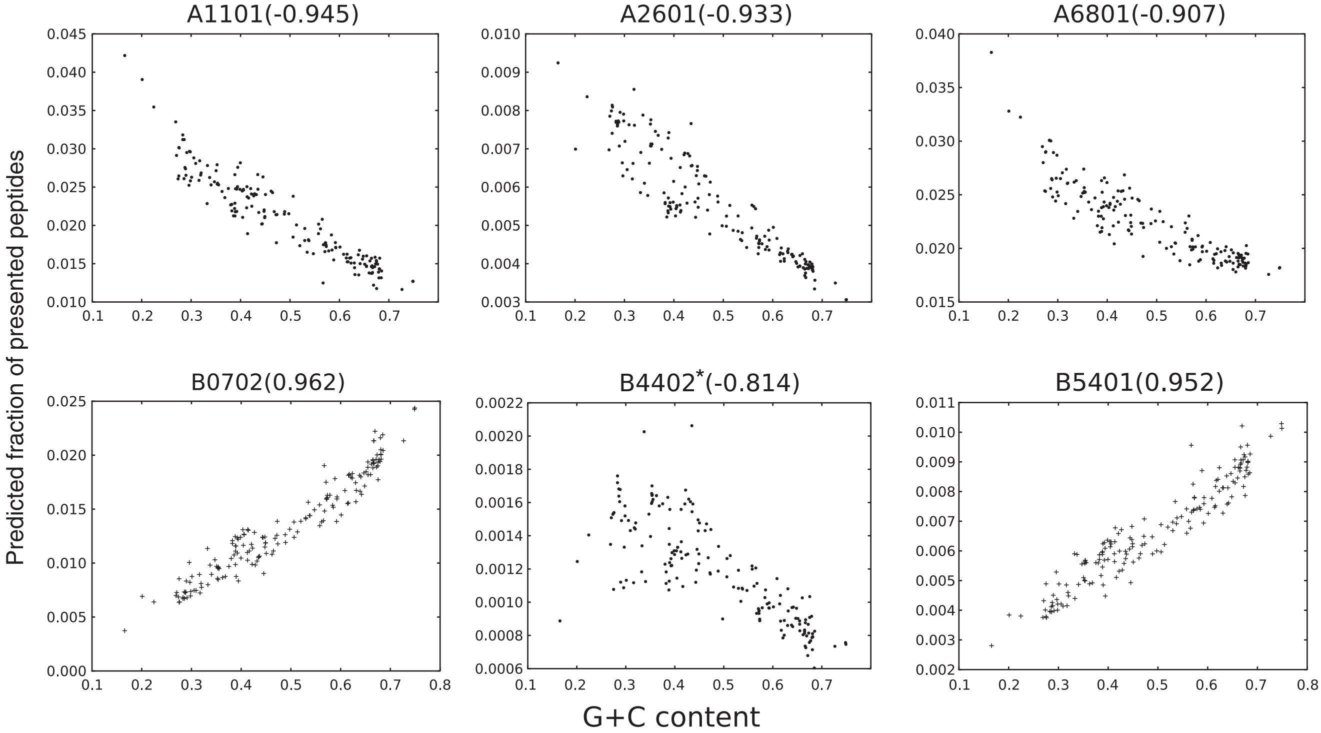

Figure 1. HLA molecules have strong G+C preferences. G+C content and fractions of presented peptides are shown for six representative HLA molecules. Every data point gives the predicted fraction of presented peptides (*y*-axis) and G+C content (*x*-axis) of one of the 174 bacteria. Correlations between 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.

different MHC-binding prediction methods, different parameter settings and was not dependent on proteasome-cleavage and TAP-transport predictions. A large determinant of peptide-binding preferences are the anchor positions; G+C preferences predicted using only these two positions overlap significantly with those determined using all positions (Spearman rank test: correlation coefficient = 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 *Materials and methods* section), and checked for G+C responsiveness. The random HLA molecules had similar binding specificities as real HLA molecules, *i.e.* on anchor positions only few amino acids are permitted. However, since the permitted amino acids were randomly distributed over the 20 amino acids, there is no bias for specific (*e.g.* G+C informative) amino acids. Of the random HLA molecules, 66% were G+C responsive, which is not different from the observed ratio of G+C-responsive HLA molecules (14/21) (Permutation test: $p = 0.57$). G+C-positive and G+C-negative molecules occur with equal frequency in the random-HLA-model, 32.1% and 33.8%, respectively. On the contrary, the majority of real HLA-A molecules (nine out of eleven) are G+C negative (Table 1), which compared with random HLA is significantly high (Permutation test: $p = 0.0017$). Similarly, there is a significant under-representation (zero of eleven) of G+C-positive HLA-A molecules (Permutation test: $p = 0.014$, see Table 1). The number of G+C-positive and G+C-negative HLA-B molecules is not significantly different from the random HLA model (Permutation test: $p = 0.41$, see Table 1). Taken together, these results suggest that while most HLA molecules are G+C responsive, only HLA-A molecules have a strong preference to present the peptides derived from G+C low organisms.

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

Although G+C content is related to genome size and pathogenicity [22, 23], causality between these three features remains elusive and it is unclear what selection pressures determine the G+C content of organisms. We tested whether in our bacterial and viral data set pathogenicity is related to G+C content, as this would provide an evolutionary explanation for the G+C low preference of HLA-A molecules. Here, we focus only on pathogens presented by the MHC-I pathway, *i.e.* viruses and intracellular bacteria (see *Materials and methods* section for the selection criteria). In our data sets, the G+C content of pathogenic intracellular bacteria and viruses is significantly lower than that of nonpathogenic bacteria (rank sums test: p<0.01, see Fig. 2). Similarly, the G+C content of pathogenic intracellular bacteria and viruses is significantly lower than that of human coding sequences (rank sums test: p<0.01, see Fig. 2). Thus, pathogens tend to have a low G+C content which may have been exploited by HLA class I molecules for enhanced presentation.

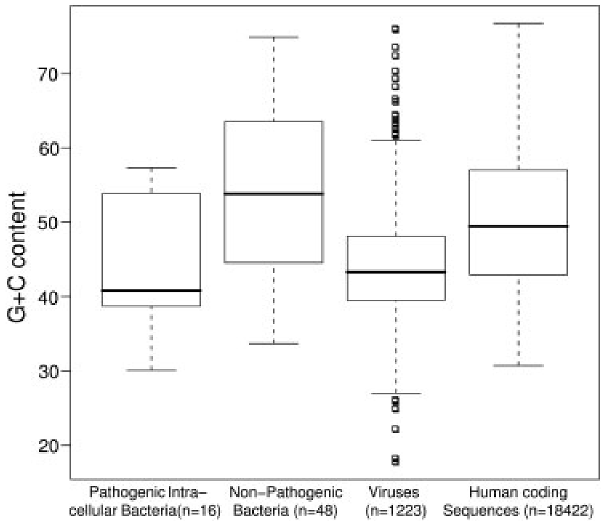

Figure 2. Pathogenic organisms have a low G+C content. G+C contents of pathogenic intracellular bacteria, nonpathogenic bacteria, viruses and human coding sequences are shown. Pathogenic intracellular bacteria and viruses have lower G+C contents than nonpathogenic bacteria (rank sums-test; $p = 0.004$ 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.

How can HLA class I molecules be G+C responsive? Because HLA molecules bind short protein fragments, this is only possible if differences in the genomic G+C content are reflected in differences at the proteomic level. G+C contents have been shown to bias amino acid usage in bacteria [24]. High G+C contents result in an increased frequency of the amino acids G, A, R and P, whereas low G+C contents result in an increased usage of F, I, N, K and Y [24–26]. This trend is also visible in our data set: in Fig. 3, we show relative amino acid frequencies for all bacteria in our data set, sorted by 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. Additionally, when clustering the amino acids on their frequencies in the bacteria, the G+C-positive and G+C-negative amino acids form separate clusters as their frequency profiles are highly distinct (highlighted in red and green in Fig. 3). Thus, genomic G+C content differences translate into proteomic differences, which can explain why HLA class I molecules have G+C preferences.

Next, to validate the predicted G+C preferences summarized in Table 1, we analyzed the G+C-positive (GARP) and G+C-

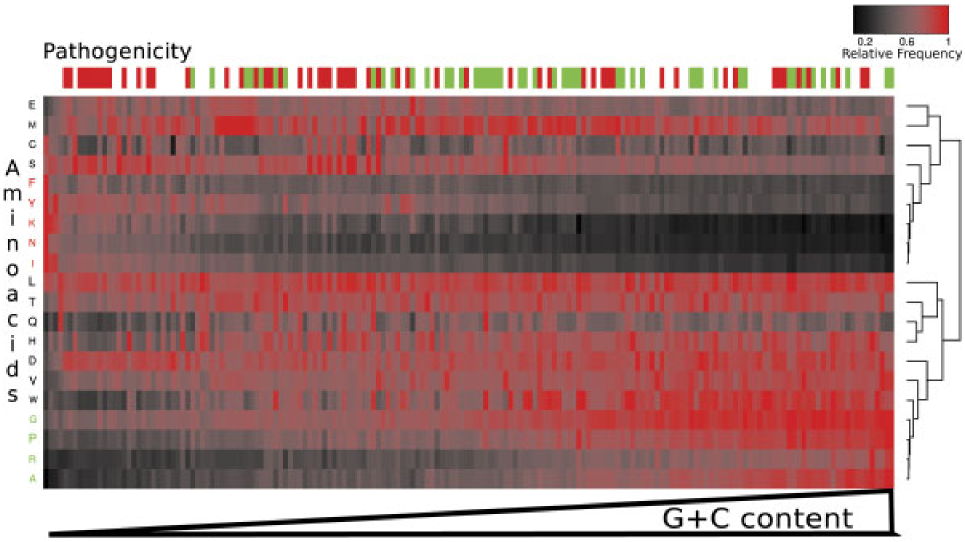

Figure 3. G+C content and amino acid frequencies correlate. All bacteria in our data set are sorted based on the G+C content, highest G+C contents on the right, lowest G+C contents on the left. Above the heat map, pathogenic bacteria are indicated red, nonpathogenic bacteria green and unknown bacteria 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.

negative (FINKY) amino acid frequencies of experimentally verified HLA-A/B presented peptides. A database perfectly suited for this test is the SYFPEITHI database [13] that almost exclusively contains naturally occurring HLA ligands. The frequency of GARP and FINKY in the peptide set was determined for 16 HLA molecules that have at least ten presented peptides in SYFPEITHI. As expected, the G+C-negative HLA class I molecules ($n = 9$) present peptides with a higher FINKY content than the G+C neutral ($n = 3$) (rank sums test: $p = 0.013$) and G+C-positive ($n = 4$) (rank sums test: $p = 0.031$) HLA class I molecules. Similarly, the peptides presented by the G+C-positive HLA molecules have a higher GARP content than the G+C neutral (rank sums test: $p = 0.034$) and G+C-negative (rank sums test: $p = 0.021$) HLA class I molecules. Taken together, these results, based on the naturally occurring experimentally verified MHC-I-presented peptides, confirm the predicted G+C preferences.

On the whole, a low G+C content is a property of organisms with a pathogenic lifestyle [22, 23] and differences in G+C content are reflected in amino acid usage [24]. These properties appear to have shaped the repertoire of binding motifs of the common present day HLA-A, but not HLA-B, molecules, to preferably present pathogen-derived peptides.

The shared G+C negativity of HLA-A molecules is not caused by a lack of diversity

The preferred presentation of pathogens provides a tantalizing explanation for the shared G+C negativity of HLA-A molecules. However, other hypotheses might also explain this observation. First of all, the G+C negativity of HLA-A can be due to a lack of diversity: if the common ancestor of all HLA-A alleles encoded a G+C-negative HLA-A molecule and divergence from this ancestor would be minimal, present day HLA-A molecules would still be G+C negative. If this were true, the diversity of HLA-A molecules would be expected to be much smaller than that of HLA-B molecules, which show large variations in their G+C preferences. The diversity among HLA molecules can be measured from the nucleotide sequences that encode these molecules, or from the protein sequences using standard sequence analysis techniques (see *Methods* section). Comparing the HLA-A and HLA-B molecules included in our analysis, we did not find a difference in the encoding nucleotide sequence variation (rank sums test: $p = 0.89$). Also at the protein level, HLA-A molecules are as diverse as HLA-B (rank sums test: $p = 0.089$). Even the most variable part of the HLA molecule, the peptide-binding sites (see *Materials and methods* section) did not show a difference in diversity between HLA-A and HLA-B molecules (rank sums test: $p = 0.24$). Therefore, a lack of diversity does not explain the shared G+C negativity of HLA-A molecules: a similar amount of diversity can result in mixed G+C preferences in the case of HLA-B.

Second, a restrained structural flexibility of the HLA-A-binding groove might be the reason for G+C negativity. In other words, HLA-A molecules might have structurally similar binding grooves, despite diversity at the genomic and proteomic level. As a result, binding preferences would be similar and this could explain the shared G+C preference. A maximum likelihood phylogeny of all HLA-A alleles was constructed to test this scenario. The two HLA-A molecules that do not prefer G+C low organisms (*i.e.* A*0201 and A*3301) do not cluster together (Fig. 4, shown in black), which suggests that G+C preferences

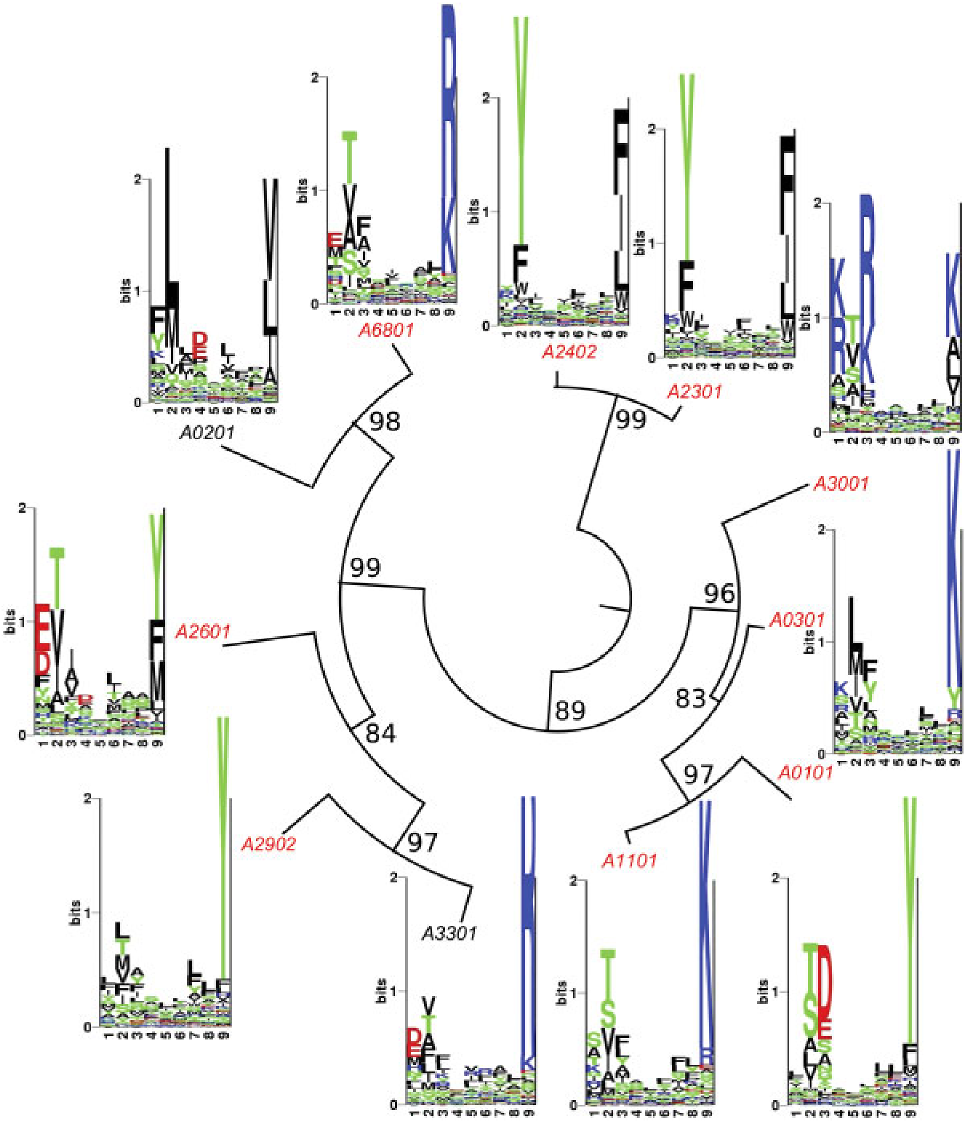

Figure 4. HLA-A phylogeny and binding motifs. The phylogeny analysis of HLA-A alleles is based on ClustalW alignment and maximum-likelihood clustering. The numbers shown in the phylogeny indicate how often a clade in the phylogeny was observed in 100 bootstrap analyses. Red and black HLA names are indicating G+C-negative and G+C-neutral binding preferences, respectively. For every HLA, the binding motif obtained from the MHC motif viewer website [48] is shown to indicate the encoded peptide-binding motif.

changed at least twice independently during HLA-A evolution. To further demonstrate the substantial structural flexibility of HLA-A-binding preferences during evolution, we display the binding motifs of the encoded HLA-A molecules in the maximum likelihood phylogeny in Fig. 4. A binding motif, here presented as a sequence logo, shows *per* position which amino acids are preferred by an HLA molecule. For instance, HLA-A*0101 prefers a peptide with Tyrosine (Y) at its ninth position, whereas HLA-A*1101 prefers Lysine (K) at that position (Fig. 4). These different binding preferences, however, result in a similar G+C-negative preference (Table 1). Further, we performed an amino acid knockout analysis (see *Materials and methods* section) to see if shared binding preferences can explain G+C negativity. This analysis showed that no single amino acid preference can explain the shared G+C negativity of the HLA-A molecules (Supporting Information Fig. S3). Taken together, these results suggest that the binding preferences of HLA-A molecules are dynamic during evolution. Therefore, structurally restrained binding grooves do not offer an explanation for the shared G+C negativity.

G+C responsiveness of nonhuman MHC class I molecules

G+C negativity among HLA-A molecules is neither due to a lack of diversity, nor is it caused by constraints on G+C preferences during HLA evolution. Therefore, an evolutionary selection pressure to present peptides that are derived from pathogens provides the best explanation for the shared G+C negativity of HLA-A 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 expected to have evolved G+C negativity. This hypothesis was tested by investigating the G+C preference of Chimpanzee MHC-I molecules (Patr molecules). Due to the similarity shared with the MHC region of humans, high-quality binding predictors are available for Patr molecules [27].

The G+C preference of all Patr molecules was investigated using the bacteria data set. Similar to HLA molecules, a majority of Patr molecules, 26 of 47, was found to be G+C responsive (Supporting Information Table S1). Moreover, a majority of Patr-A molecules, *i.e.* 15 out of 18, are G+C negative (compared with random HLA molecules, permutation test: $p<0.001$). Patr-B molecules, such as HLA-B molecules, show mixed G+C preferences and the fraction of G+C-negative Patr-B molecules is not significantly different from random HLA molecules (Permutation test: $p = 0.91$). These results suggest that also in Chimpanzees there has been an evolutionary pressure to select and/or keep G+C-negative peptide-binding preferences in MHC molecules encoded by the A-locus.

Finally, the MHC-I G+C preferences of a more distant primate, the Rhesus Macaque, were investigated. Compared with the HLA locus, the Rhesus Macaque MHC-I (Mamu) locus is organized quite differently, *i.e.* the Mamu-A and -B genes are duplicated at least once and twice, respectively [28]. Possibly, as a result of these duplications, G+C preferences of Mamu molecules seem to be different: Mamu-B, but not Mamu-A, molecules are mostly G+C negative (Supporting Information Table S2). Unfortunately, because they are more distant from HLA molecules, for which the majority of peptide-binding data are available, Mamu-binding predictions are less accurate. This probably leads to an underestimate of the actual specificity and G+C preference: a significantly large set of Mamu molecules (23 of 40) are G+C neutral (compared with random HLA molecules, permutation test: $p = 0.002$). Still, 11 out of 23 Mamu-B molecules are G+C negative (Permutation test: $p = 0.117$), whereas 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, one MHC class I gene seems to be dedicated to present peptides derived from G+C low pathogens.

Discussion

It is well established that the innate immune system uses conserved microbial components for danger signaling and self/nonself discrimination. A good example is the recognition of LPS by TLR-4 [29]. We here find that the adaptive immune system also uses a pathogen-specific signature, *i.e.* a low G+C content, to enhance the MHC-I presentation of pathogenic organisms. A significant majority of HLA-A molecules have G+C low preferences, despite a large variety in HLA-A-binding motifs. Chimpanzees and rhesus macaques also 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. In contrast to humans and chimpanzees, in rhesus macaques it is the B-locus instead of the A-locus that is G+C negative. Humans and rhesus macaques diverged approximately 27 Million years ago, whereas chimpanzees diverged only about 5.4 Million years ago from humans [30]. 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 addressed if more high-quality MHC-I predictors for gorilla, orangutan or gibbon were developed.

We previously observed that HLA-A, but not HLA-B molecules, favor the presentation of bacteria and viruses over self [18]. This observation can now be explained by differences in G+C content between self and pathogens (Fig. 2) and the shared G+C preferences of HLA-A molecules (Table 1). HLA-B molecules lack a common G+C preference. Still, HLA-B molecules are as functional as their HLA-A counterparts in antigen presentation, and for some pathogens they are more frequently involved in dominant immune responses than HLA-A molecules [31–33]. If HLA-B molecules were important in the presentation of (G+C low) pathogens, one would expect them to also evolve a G+C-negative preference. However, we propose that one G+C-negative HLA locus (HLA-A) is sufficient to capture the general feature of G+C low pathogens. Consequently, the other HLA locus (HLA-B) can coevolve with specific pathogens. HLA-B, and not HLA-A, alleles have been reported to evolve *via* recombination: this might help fast adaptation to specific pathogens [10, 34, 35]. The higher polymorphism of HLA-B alleles (1605 HLA-B *versus* 1001 HLA-A alleles known in the HLA/IMGT-database in May 2010 [36]), and the more rapid selection of new HLA molecules in the B-locus [10, 34, 35, 37] are in agreement with a fast adaptation of HLA-B to specific pathogens. Pathogen-specific evolution of HLA-B molecules might explain the observed immunodominance of HLA-B molecules in rapidly evolving pathogens such as HIV-1 [31].

Although having a low G+C content is a general feature of pathogenic life [22], 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 [38, 39]. G+C-positive HLA class I molecules and molecules like TLR9 could therefore 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.* reported that human herpes viruses have fewer CTL epitopes than their nonhuman family members [17]. Vider-Shalit *et al.* 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 by the human population. 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.* [17] correlate with the G+C content of different viruses. For 16 herpes viruses, we could determine the G+C content and a very good correlation with the reported SIR scores (Spearman rank test: correlation $=-0.81$; $p<0.001$) was observed. Thus, differences in G+C content provide an alternative explanation for the reported SIR scores.

Although it was well established that a low G+C content is a general feature of pathogenic life, this is the first study, to the best of our knowledge, demonstrating that HLA molecules evolved to employ this pathogen-specific signature. This result contributes to understanding the nature of MHC-I presentation preferences and reveals an important aspect of MHC-I evolution: namely that a subset of MHC molecules has been selected to capture a general feature of pathogenic life, a low G+C content.

Materials and methods

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. Only the human proteins with evidence at the protein or transcript level were included in this study. 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 redundant and were excluded, resulting in a selection of 1223 nonredundant viral proteomes.

The 174 bacteria were sorted into three groups based on pathogenicity, *i.e.* "pathogenic", "nonpathogenic" and "unknown". Pathogenicity data were 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 "nonpathogenic" if they are not pathogenic to plants or animals. To be confident on the nonpathogenic bacteria, bacteria with a known pathogen in their genus or host-associated habitat were excluded from the nonpathogenic group and defined as "unknown". Following these criteria, 53 bacteria were defined as "pathogenic", 48 bacteria were defined as "nonpathogenic" and 73 bacteria were defined as "unknown". In total, 16 Intracellular pathogenic bacteria (used in Fig. 2) were selected from the pathogenic group based on the organism description in the NCBI genome project. All 174 bacteria and their classification are listed in Supporting Information Table S3.

MHC-I presentation predictions

The epitopes present in a protein or proteome can be predicted by using tools for the three key processes of MHC-I presentation. 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.* [19, 20], we used the best-performing MHC-binding predictor, NetMHC-3.0 [40, 41], 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 [42]. Unless mentioned otherwise, we obtained similar results with both methods.

Once the peptide–MHC predictions are made, one needs to define a threshold that separates predicted binders from nonbinders. Recently, the choice of thresholds has been discussed extensively [43]. One possibility is to assume a fixed threshold holds for all MHC molecules, *e.g.* 500nM or 5000 nM [40, 41]. Alternatively, an MHC-specific threshold can be used if one assumes that all MHC molecules present the same number of peptides, *e.g.* 2% of a proteome. In order to check for consistency of our results, we chose to use both the fixed thresholds of 500 nM and 5000 nM and an MHC-specific threshold defined as the threshold that corresponds to the top 2% of predicted peptides from the human proteome. Results presented in the text were derived using a 500 nM threshold, and they are similar for all thresholds, unless mentioned otherwise.

For Chimpanzee and Mamu, we used MHC-I-binding predictions from NetMHCpan-2.0 [27, 44], a neural network-based predictor that uses MHC-I-binding groove polymorphisms to predict peptide–MHC-binding affinities. Unfortunately, no alternative predictors for these nonhuman primate MHC-I molecules are available. Consistency of the predictions was checked only for the fixed MHC-binding thresholds 500nM and 5000 nM, and the results were similar unless mentioned otherwise.

Processing of the peptides, proteasomal degradation and TAP transport, was predicted by NetChop Cterm3.0 [45, 46]. To test whether the results depend on the peptide-processing predictions, the analysis was repeated without peptide processing. In no case, did this affect the reported results. Finally, the "fraction of presented peptides" is defined as the number of peptides predicted to be presented, divided by the number of all possible 9mer in the protein or proteome.

MHC-I allele selection

NetMHC-3.0 provides neural network predictors for 43 HLA molecules, covering 28 HLA serotypes [40, 41]. To obtain a nonredundant HLA set we chose 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 SMM predictions [42] were not available. Moreover, HLA molecules where the predictions of both methods were inconsistent (Spearman rank test: correlation <0.7 or $p>0.001$) were excluded. This resulted in the selection of eleven

HLA-A and ten HLA-B molecules (Table 1). HLA-C molecules were not included in this study as NetMHC-3.0 does not provide neural network predictors for these molecules and NetMHCpan-2.0 is reported to be poor for HLA-C molecules [27].

NetMHCpan-2.0 provides predictors for 70 Patr molecules and 72 Mamu molecules [27]. Again, to obtain a nonredundant HLA set, we used only unique 2-digit molecules, resulting in the selection of 47 Patr molecules (Supporting Information Table S1) and 57 Mamu molecules. NetMHCpan-2.0 predictors are based on similarity to well-characterized MHC-I molecules with large peptide-binding data sets. Consequently, the pMHC-binding affinity for MHC molecules more distant to the training set, *e.g.* Mamu molecules, is underestimated [27]. To prevent using these low-quality predictors, we excluded Mamu predictors that predicted zero binders in any of the 174 bacteria below a 500 nM threshold. The final set included 40 Mamu molecules (Supporting Information Table S2).

Random HLA model

Predictors for random HLA molecules were generated using SMM-matrices [42]. 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 1. Subsequent randomization of the amino acid binding-values, *e.g.* Lysine gets the preference of Arginine, ensured random amino acid 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.

HLA phylogeny and distances

Genomic and protein sequences of HLA molecules were downloaded from the IMGT/HLA database [36] in January 2009. Multiple sequence alignments were made using ClustalW. 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 [47] using the GTRMIX model and bootstrap support from 200 replicates (Fig. 4). MHC-binding motifs complementing the phylogeny are obtained from the MHC motif viewer website, www.cbs.dtu.dk/biotools/MHCMotifViewer/ [48]. Pairwise distances of protein sequences 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 known pMHC structures [44]. A list of all peptide-binding sites is provided in Supporting Information Table S4.

Amino acid *knockout* 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 acid to 0.0 in the SMM matrix [42] in each of the nine position vectors. These revised SMM 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 *knockout* analysis.

Acknowledgements: The authors thank Rob de Boer for valuable comments on the manuscript and discussion on this research project, Ronald Bontrop and Liset Westera 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.

Conflict of interest: The authors declare no financial or commercial conflict of interest.

References

1 **Vyas, J. M., der Veen, A. G. V. and Ploegh, H. L.,** The known unknowns of antigen processing and presentation. *Nat. Rev. Immunol.* 2008. **8**: 607–618.

2 **Groothuis, T. A. M., Griekspoor, A. C., Neijssen, J. J., Herberts, C. A. and Neefjes, J. J.,** MHC class I alleles and their exploration of the antigen-processing machinery. *Immunol. Rev.* 2005. **207**: 60–76.

3 **Craiu, A., Akopian, T., Goldberg, A. and Rock, K. L.,** Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc. Natl. Acad. Sci. USA* 1997. **94**: 10850–10855.

4 **Falk, K., Rötzschke, O. and Rammensee, H. G.,** Cellular peptide composition governed by major histocompatibility complex class I molecules. *Nature* 1990. **348**: 248–251.

5 **Paulsson, K. M.,** Evolutionary and functional perspectives of the major histocompatibility complex class I antigen-processing machinery. *Cell. Mol. Life Sci.* 2004. **61**: 2446–2460.

6 **Yewdell, J. W.,** Not such a dismal science: the economics of protein synthesis, folding, degradation and antigen processing. *Trends Cell. Biol.* 2001. **11**: 294–297.

7 **Yewdell, J. W., Reits, E. and Neefjes, J.,** Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat. Rev. Immunol.* 2003. **3**: 952–961.

8 **Huseby, E. S., White, J., Crawford, F., Vass, T., Becker, D., Pinilla, C., Marrack, P. and Kappler, J. W.,** How the T cell repertoire becomes peptide and MHC specific. *Cell* 2005. **122**: 247–260.

9 **Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V. and Balloux, F.,** Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* 2005. **15**: 1022–1027.

10 **Parham, P. and Ohta, T.,** Population biology of antigen presentation by MHC class I molecules. *Science* 1996. **272**: 67–74.

11 **Spurgin, L. G. and Richardson, D. S.,** How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. Biol. Sci.* 2010. **277**: 979–988.

12 **Hughes, A. L. and Nei, M.,** Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 1988. **335**: 167–170.

13 **Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. and Stevanović, S.,** SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999. **50**: 213–219.

14 **Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R. *et al.*,** The immune epitope database 2.0. *Nucleic Acids Res.* 2010. **38**: D854–D862.

15 **Schellens, I. M., Keşmir, C., Miedema, F., Van Baarle, D. and Borghans, J. A.,** An unanticipated lack of consensus cytotoxic T lymphocyte epitopes in HIV-1 databases: the contribution of prediction programs. *AIDS* 2008. **22**: 33–37.

16 **Almani, M., Raffaeli, S., Vider-Shalit, T., Tsaban, L., Fishbain, V. and Louzoun, Y.,** Human self-protein CD8$^+$ T-cell epitopes are both positively and negatively selected. *Eur. J. Immunol.* 2009. **39**: 1056–1065.

17 **Vider-Shalit, T., Sarid, R., Maman, K., Tsaban, L., Levi, R. and Louzoun, Y.,** Viruses selectively mutate their CD8$^+$ T-cell epitopes-a large-scale immunomic analysis. *Bioinformatics* 2009. **25**: i39–i44.

18 **Rao, X., Costa, A. I., Van Baarle, D. and Keşmir, C.,** A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant CD8$^+$ T cell responses. *J. Immunol.* 2009. **182**: 1526–1532.

19 **Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., Basch, D. *et al.*,** A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* 2006. **2**: e65.

20 **Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O. and Nielsen, M.,** Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *Biomed. Chromatogr. Bioinform.* 2007. **8**: 424.

21 **Lundegaard, C., Nielsen, M. and Lund, O.,** The validity of predicted T-cell epitopes. *Trends Biotechnol.* 2006. **24**: 537–538.

22 **Rocha, E. P. C. and Danchin, A.,** Base composition bias might result from competition for metabolic resources. *Trends Genet.* 2002. **18**: 291–294.

23 **Bentley, S. D. and Parkhill, J.,** Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* 2004. **38**: 771–792.

24 **Singer, G. A. and Hickey, D. A.,** Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* 2000. **17**: 1581–1588.

25 **Muto, A. and Osawa, S.,** The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA* 1987. **84**: 166–169.

26 **Knight, R. D., Freeland, S. J. and Landweber, L. F.,** 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.* 2001. **2**: RESEARCH0010.

27 **Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S. and Nielsen, M.,** NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 2009. **61**: 1–13.

28 **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.,** Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc. Natl. Acad. Sci. USA* 2005. **102**: 1626–1631.

29 **Nürnberger, T., Brunner, F., Kemmerling, B. and Piater, L.,** Innate immunity in plants and animals: striking similarities and obvious differences. *Immunol. Rev.* 2004. **198**: 249–266.

30 **Fukami-Kobayashi, K., Shiina, T., Anzai, T., Sano, K., Yamazaki, M., Inoko, H. and Tateno, Y.,** Genomic evolution of MHC class I region in primates. *Proc. Natl. Acad. Sci. USA* 2005. **102**: 9230–9234.

31 **Kiepiela, P., Leslie, A. J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P. *et al.*,** Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 2004. **432**: 769–775.

32 **Bihl, F., Frahm, N., Giammarino, L. D., Sidney, J., John, M., Yusim, K., Woodberry, T. *et al.*,** Impact of HLA-B alleles, epitope binding affinity, functional avidity, and viral coinfection on the immunodominance of virus-specific CTL responses. *J. Immunol.* 2006. **176**: 4094–4101.

33 **Lewinsohn, D. A., Winata, E., Swarbrick, G. M., Tanner, K. E., Cook, M. S., Null, M. D., Cansler, M. E. *et al.*,** Immunodominant tuberculosis CD8 antigens preferentially restricted by HLA-B. *PLoS Pathog.* 2007. **3**: 1240–1249.

34 **Watkins, D. I., McAdam, S. N., Liu, X., Strang, C. R., Milford, E. L., Levine, C. G., Garber, T. L. *et al.*,** New recombinant HLA-B alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. *Nature* 1992, **357**: 329–333, doi:10.1038/357329a0.

35 **McAdam, S. N., Boyson, J. E., Liu, X., Garber, T. L., Hughes, A. L., Bontrop, R. E. and Watkins, D. I.,** A uniquely high level of recombination at the HLA-B locus. *Proc. Natl. Acad. Sci. USA* 1994. **91**: 5893–5897.

36 **Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y. *et al.*,** IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 2009. **37**: D1006–D1012.

37 **Belich, M. P., Madrigal, J. A., Hildebrand, W. H., Zemmour, J., Williams, R. C., Luz, R., Petzl-Erler, M. L. and Parham, P.,** Unusual HLA-B alleles in two tribes of Brazilian Indians. *Nature* 1992. **357**: 326–329.

38 **Bauer, S., Pigisch, S., Hangel, D., Kaufmann, A. and Hamm, S.,** Recognition of nucleic acid and nucleic acid analogs by Toll-like receptors 7, 8 and 9. *Immunobiology* 2008. **213**: 315–328.

39 **Krieg, A. M., Yi, A. K., Matson, S., Waldschmidt, T. J., Bishop, G. A., Teasdale, R., Koretzky, G. A. and Klinman, D. M.,** CpG motifs in bacterial DNA trigger direct B-cell activation. *Nature* 1995. **374**: 546–549.

40 **Buus, S., Lauemoller, S. L., Worning, P., Keşmir, C., Frimurer, T., Corbet, S., Fomsgaard, A. *et al.*,** Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens* 2003. **62**: 378–384.

41 **Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S. and Lund, O.,** Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 2003. **12**: 1007–1017.

42 **Peters, B., Tong, W., Sidney, J., Sette, A. and Weng, Z.,** Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 2003. **19**: 1765–1772.

43 **MacNamara, A., Kadolsky, U., Bangham, C. R. M. and Asquith, B.,** T-cell epitope prediction: rescaling can mask biological variation between MHC molecules. *PLoS Comput. Biol.* 2009. **5**: e1000327.

44 **Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G. *et al.*,** NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* 2007. **2**: e796.

45 **Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V. and Brunak, S.,** Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 2002. **15**: 287–296.

46 **Nielsen, M., Lundegaard, C., Lund, O. and Keşmir, C.,** The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 2005. **57**: 33–41.

47 **Stamatakis, A.,** RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006. **22**: 2688–2690.

48 **Rapin, N., Hoof, I., Lund, O. and Nielsen, M.,** MHC motif viewer. *Immunogenetics* 2008. **60**: 759–765.

Abbreviations: **Mamu:** Rhesus Macaque MHC-I · **Patr:** Chimpanzee MHC-I · **pMHC:** peptide–MHC-I complex · **9mer:** nonamer · **SIR:** Size of Immune Repertoire · **SMM:** Scoring Matrix Method

Full correspondence: Jorg J. A. Calis, Padualaan 8, 3584 CH Utrecht, The Netherlands
Fax: +31-0-30-251-3655
e-mail: j.j.a.calis@uu.nl

See accompanying Commentary:
http://dx.doi/10.1002/eji.201040856

Received: 20/1/2010
Revised: 18/6/2010
Accepted: 21/7/2010
Accepted article online: 30/7/2010