Immunogenetics
DOI 10.1007/s00251-014-0811-4

ORIGINAL PAPER

Viral CD8 T cell epitope nucleotide composition shows evidence of short- and long-term evolutionary strategies

Yaakov Maman · Uri Hershberg · Yoram Louzoun

Received: 10 July 2014 / Accepted: 26 October 2014

Abstract Viral epitopes have a distinct codon usage that reflects their dual role in infection and immunity. On the one hand, epitopes are part of proteins important to viral function; on the other hand, they are targets of the immune response. Studies of selection are most commonly based on changes of amino acid and seen through the accumulation of non-synonymous mutations. An independent measure of selection is the codon usage and underlying changeability of the nucleotide sequences. We here use multiple tools and a large-scale analysis of viral genomes to demonstrate that viral epitopes have a distinct codon usage and that this codon usage reflects distinct short- and long-term types of selection during viral evolution. We show that $CD8^+$ T cell epitopes are encoded by codons more distant from stop codons and more changeable than codons outside epitopes. This biased codon usage reflects the viral population toggling back and forth from a wild-type sequence to an escape mode, which enable them to avoid immune detection when needed, and go back to the functionally favorable form when the threat is removed (i.e., in a new host).

Keywords Codon bias · CTL epitopes · Selection

Electronic supplementary material The online version of this article (doi:10.1007/s00251-014-0811-4) contains supplementary material, which is available to authorized users.

Y. Maman
Department of Immunobiology, Yale University School of Medicine, 300 Cedar Street, Box 208011, New Haven, CT 06520-8011, USA

U. Hershberg
Department of Biomedical Engineering, Drexel University, Philadelphia, PA 19104, USA

Y. Louzoun (✉)
Department of Mathematics and Gonda Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel
e-mail: louzouy@math.biu.ac.il

Introduction

Evolution and the selection processes that underlie it are linked directly to the ability of different species to function in the environment. Classically, this is taken to mean selection of the function of proteins and thus, selection is measured by changes at the amino acid level. However, it has been shown that the codon usage can have functional consequences (Bulmer 1991; Duret 2002; Hershberg and Petrov 2008, 2009) including DNA stability and tRNA usage in bacteria (Hershberg and Petrov 2009; Kanaya et al. 1999; Yamao et al. 1991), and the generation of viable immune repertoire variability in B cells (Hershberg and Shlomchik 2006). Immune-induced selection for codon bias was suggested by several studies. Plotkin et al. (Plotkin and Dushoff 2003) show that in influenza virus, the codons of hemagglutinin are biased toward non-synonymous point mutations and that this bias is more significant at codons in regions involved in antibody combination. Kijak et al. (2004) show similar pattern in HIV, where a bias is observed toward codons that have a direct path to point mutation. We here present results showing that the codon usage in different regions of a given gene can be differentially affected by the inter-host and intra-host evolution and can serve as a marker for such an evolution.. Advantageous mutations occurring in populations undergoing selection in a fixed environment will eventually be fixed in the populations, and no resulting polymorphism will be observed (Kimura 1962). However, in a continuously changing environment, we do not expect mutation fixation, since the increased fitness induced by a mutation may be of limited advantage. In such a case, selection may actually favor mutations leading to a higher flexibility (Hershberg and Shlomchik 2006).

Viral populations must maintain their diversity in order to survive in this host as proposed by the quasispecies model (Elena et al. 2008; Wilke et al. 2001). This model proposes

Published online: 08 November 2014

that selection works by maximizing the averaged replication rate of a cloud of similar, but not identical genotypes rather than fix one genotype with the highest replication rate, since the latter is more sensitive to deleterious mutations.

Another selective force that maintains diversity in viral population is viral escape from the cellular immune response, mediated by $CD8^+$ T lymphocytes (CTL). CTL epitope presentation is deadly for viruses, since it leads to their rapid elimination (McMichael et al. 1983). The presentation of a CTL epitope requires well-defined motifs (Sherman 2006) that can be changed by a small number of mutations (Yokomaku et al. 2004). Viral populations avoid the immune response by accumulating escape mutations in these precise positions (Lichterfeld et al. 2005). Viruses typically have a very high mutation rate, which is of the order of one mutation per viral life cycle (Coffin 1995; Sanjuan et al. 2010). One of the most common escape mechanisms is to specifically mutate epitopes that are presented by an MHC-I molecule from a given HLA alleles (Maman et al. 2011; Vider-Shalit et al. 2007, 2009).

The HLA locus is the most polymorphic locus in the human genome. In the class I locus HLA-A, HLA-B, and HLA-C have over 1000 known alleles each (HLA-B has almost 2000 alleles) (http://www.ncbi.nlm.nih.gov/projects/gv/mhc/ihwg.cgi). Each allele has a distinct binding groove and a distinct epitope repertoire (Borghans et al. 2004). Thus, it would be in principle impossible to escape from all alleles. However, each host expresses at most six MHC class-I alleles. Thus, during infection within a single host, the virus only has to accumulate escape mutations preventing the presentation of epitopes in these alleles and could in principle rapidly eliminate all these CTL epitopes. When the virus moves to another host, it has to adapt to the alleles of the new host. The environment the virus has to deal with is thus constantly changing.

Another factor limiting the viral escape is the fitness cost induced by some mutations, which may be stronger than the advantage induced by the escape from the immune response (Seibert et al. 1995). Such a fitness cost is often the result of suboptimal assembly or function of a protein. While the advantage induced by the escape from the immune response is host-specific, the cost is often not.

Thus, the evolution of many viruses that escape from the immune response can be analyzed at two time scales: The longer time scale involves the optimization of viral functionality, the ability to infect, and proliferate in a host, as well as selection of traits attributed to common features of the host immune system (such as proteasomal cleavage, TAP binding (Schmid et al. 2008), and perhaps super-antigens (Torres et al. 2001)). The shorter time scale is the intra-host evolution, which is often driven by the need to escape the host-specific immune system (Maman et al. 2011; Vider-Shalit et al. 2007, 2009). While the longer time scale evolution might result in quite directed selection for mutations, the shorter time scale evolution can result in back and forth mutations, toggling between the more functional viral sequences that serve as CTL epitopes and those that are less functional, but harder for the immune system to detect. This type of selection is called toggling selection (Fig. S1) and has been suggested to occur in HIV (Delport et al. 2008).

These contradicting effects can have a signature in the codon usage and the frequency of mutations between codons. A genetic position where toggling between multiple states is advantageous when moving from one host to the other has to maximize its ability to safely change amino acid composition by mutations. This implies a selection of specific codon usage that follows two types of constraints: (1) Codons will be as far as possible from the stop codons. This allows the sequence to mutate back and forth in its codon neighborhood, without destroying the full length protein (see Fig. S1 for example). (2) Codons must have a high changeability. The changeability of codons is typically defined as the fraction of its first neighbors in the genetic code that do not share the same amino acid (Hershberg and Shlomchik 2006). Due to the nature of the genetic code and the variation in the number of codons that encode each amino acid, the codons encoding a specific amino acid tend to have very similar if not identical changeability levels. The clearest example of a non-uniform distribution of changeability can be found in the sixfold redundant amino acids (arginine, serine, and leucine). Therefore, we expect to see the clearest indication of codon bias in these three amino acids. We here show that epitopes indeed exhibit codon bias that fits these two restrictions and allows them to more efficiently undergo toggling mutations. Using epitope prediction algorithms (Vider-Shalit and Louzoun 2011) and phylogenetic analysis (Maman et al. 2011), we systematically show a signature for both short- and long-term evolution in the codon usage and dynamics of viral CTL epitopes.

Methods

Viral sequences

In the epitopes/non-epitopes codon bias analysis, we used sequences from ten human viruses (Table S1). In the phylogenetic tree analysis, we used nucleotide sequences of viral genes from three viruses, on which enough data was available to perform a phylogenetic analysis: HBV (Core, Pol, Surface, and X); HIV (env, gag, nef, pol, rev, tat, and vpu), and HPV (E2, E6, E7, L1, and L2). All sequences were taken from NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore).

Epitope prediction

Epitopes were computed by sequentially using three algorithms: a proteasomal cleavage algorithm (Ginodi et al. 2008), a TAP binding algorithm developed by Peters et al. (2003), and the MLVO (Multi Label Vector Optimization)-MHC binding (Vider-Shalit and Louzoun 2011) algorithm (Fig. 1). The algorithms' quality was systematically validated versus epitope databases and was found to induce low FP and FN error rates (Louzoun et al. 2006). A different threshold was defined for each stage as well as for each HLA allele. Epitopes were defined as the supra-threshold nine-mers (along with their flanking regions).

Phylogenetic trees

The DNA sequences of different viral proteins were aligned using Muscle (version 3.6) for each protein in the datasets analyzed (Edgar 2004). Phylogenic trees were then produced from the aligned sequences using the neighbor-joining method of the Phylip bioinformatics tool package (version 3.69) (http://evolution.genetics.washington.edu/phylip/getme.htm). For each group of sequences, a genetically distant "outgroup" sequence from a non-human virus homologue was added to properly position the root of the tree and the ancestral sequences. To avoid ambiguous nucleotides in internal nodes, when both child sequences had a gap in a certain locus, the parental nucleotide was changed to a gap as well. If one of the child sequences had a non-ambiguous nucleotide, the parental nucleotide was changed accordingly.

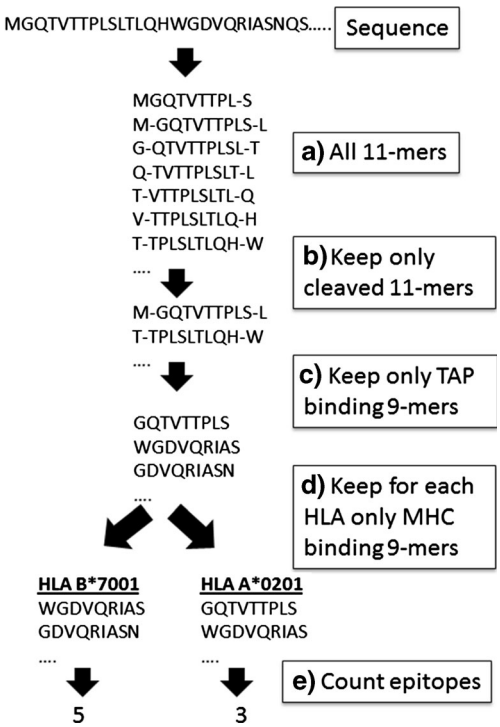

Fig. 1 Epitope repertoire analysis. Each protein is divided into all nine-mers and the appropriate flanking positions (*a*). For each 11-mer (a nine-mer and the C and N flanking positions), a cleavage score is computed (*b*). We compute for all peptides with a positive cleavage score a TAP binding score and choose only supra-threshold peptides (*c*). The MHC binding score of all TAP binding and cleaved nine-mers is computed (*d*). Nine-mers passing all these stages are defined as epitopes. We then compute the number of epitopes per protein per HLA allele (*e*)

Evolutionary model for immune-induced codon bias

To track immune-induced selection on codon usage, a distinction has to be made between sequence positions that encode for epitopes and those that do not. For this purpose, we first defined the "epitope score," E, of the ith position in the kth leaf: $E_{i,k} = \sum_{a \in A} a_j M(k)_{i,j}$, where a_j is the frequency of the jth HLA allele in the Caucasian population and $M(k)_{i,j}$ is a binary variable that denotes whether the ith position of sequence of the kth leaf is a part of an epitope presented by the jth HLA allele.

For a whole phylogenetic tree consisting of K leaves, the calculated score for the ith position is given by the following: $S_i = \frac{\sum_{k=1}^{K} E_{i,k}}{K}$.

Finally, the 15 % highest scored positions were defined as epitopes and all the other positions as non-epitopes (Fig. S2).

Having this classification for each sequence position, we tracked the mutations that occurred on each codon along the tree lineage. We defined "toggling mutation" in a lineage as a replacement mutation from A to B, and back to A.

Statistical analysis

A two-way t test with unknown and unequal variance was used for the comparison of toggling mutations in epitopes and non-epitopes along phylogenetic trees. A paired t test was used for HIV versus SIV protein conservation. A chi-squared test was used for the analysis of the codon bias and mutation rates. Correlations between codon usage and their changeability were computed using a Pearson correlation. Mean square error (MSE) was calculated to evaluate the differences in toggling/replacement ratio between regions of proteasomal cleavage size and either epitope or non-epitope regions.

Before applying a t test, we tested in each case that the distribution does not deviate from a normal distribution using a Kolomogorov-Smirnov test. In no case did we observe a deviation from normal distribution.

Results

Epitopes' codon usage is biased toward codons that are changeable and far from stop codon

In previous studies, we have shown that viruses undergo selection to remove their CTL epitopes (Maman et al. 2011; Vider-Shalit et al. 2007, 2009). This was shown at the amino acid level by comparing the epitope density in human viruses to the epitope density in their non-human counterparts, as well as through direct measurements of epitope removal. In all tested pathogens, we found that some epitopes are removed, while some are maintained. The epitopes that are not removed may remain, since they either pose no danger to the virus or since the viral fitness would be reduced following mutations in the epitope. In some epitopes, the fitness cost is low enough to allow for these mutations to occur in hosts where the epitope is presented, while high enough to lead to a mutation back to the original sequence in hosts with different HLA haplotypes where the epitope is not presented. We here show a signature of such toggling mutations on the codon bias.

Positions where toggling mutations took place are expected to have a biased codon usage for two reasons: (1) greater changeability (the changeability of a codon is defined as the fraction of its neighbors that encode for different amino acids (among all viable neighbors)) and (2) greater distance from stop codon, as shall be further discussed. To test for a codon usage difference between epitopes and non-epitopes, we used the MLVO epitope prediction algorithm (Vider-Shalit and Louzoun 2011), combined with algorithms for pre-processing (proteasomal cleavage (Ginodi et al. 2008) and TAP binding (Peters et al. 2003)) to define positions inside and outside of epitopes for each HLA allele for a large number of viral proteins. We then checked the frequency of each codon in viral epitopes and outside such epitopes (Fig. 1, see Methods for details). The frequency was normalized to 1 for each amino acid to avoid any bias due to amino acid selection.

As expected from the considerations above, epitopes were observed to have a distinct codon usage, as can be seen from the ratio of the codon frequency within and outside epitopes (the "relative epitope codon usage") (Fig. 2).

Not all amino acids and not all codons are equally likely to be influenced by changeability or by the distance from a stop codon. Methionine and tryptophan, for instance, are encoded by a single codon. For nine amino acids, the mutation distance from stop is uniform among their codons: 1 away from stop for tyrosine, cysteine, glutamine, glutamic acid, and lysine, and 2 away from stop for aspartic acid, asparagine, histidine, and phenylalanine. No significant difference was detected between codon frequency in each of these amino acids. Among the other nine amino acids, where codons have different distance from stop codon, the codons that are 1 away from stop codons are indeed significantly less abundant in epitopes than in non-epitopes (t test; $p<0.05$) (Fig. 3a).

Most of these codons are found in the sixfold redundant amino acids serine, arginine, and leucine. In terms of changeability, the sixfold redundant amino acids (arginine, serine, and leucine) can be clearly divided into two subgroups: A fourfold redundant subgroup and a twofold redundant subgroup. These amino acids show a clear distinction in their changeability, with the fourfold redundant codons being less changeable than the twofold redundant codons. This is especially clear for the codons encoding serine as its twofold redundant subgroup is more than a mutation away from its fourfold redundant subgroup. As expected, in serine, arginine, and leucine, the correlation between inherent changeability of codons and their relative epitope codon usage is positive

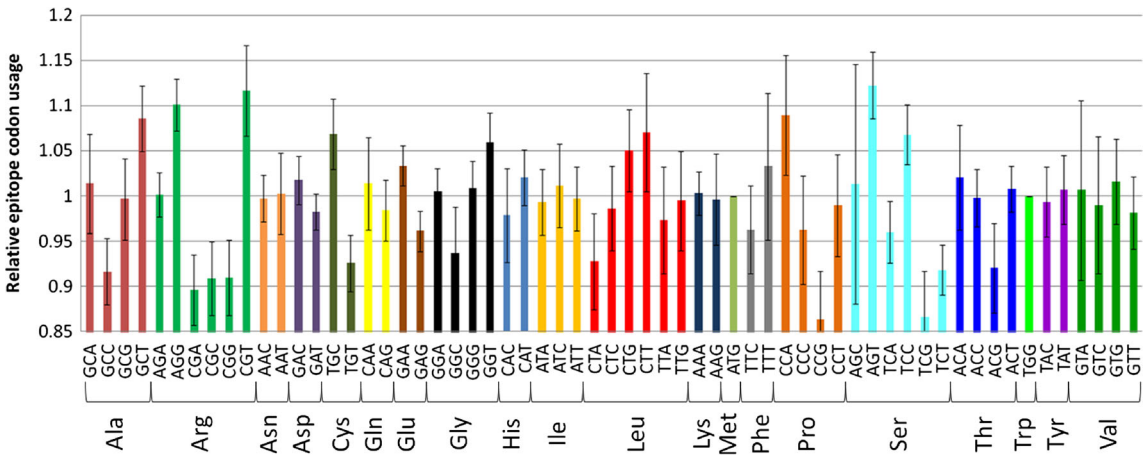

Fig. 2 Codon bias in epitopes. The relative epitope codon usage of 61 coding codons was calculated as the ratio of the usage in epitopes to usage in non-epitopes for each HLA allele, weighted by the frequency of the corresponding allele in the population. The different amino acids are represented in *different colors*. Values below 1 are underrepresented in epitopes, while codons with value above 1 are overrepresented in epitopes (color figure online)

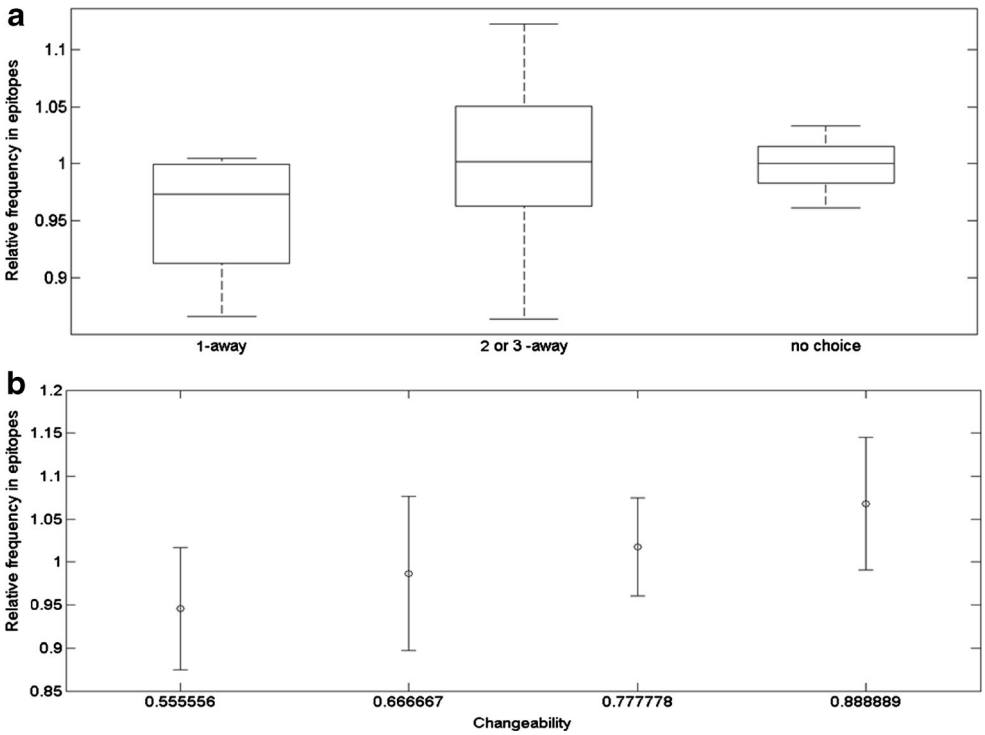

Fig. 3 Characteristics of codons in epitopes. Relative epitope codon usage was calculated with respect to the codons' distance from stop codon (**a**) and their changeability (**b**). Epitopes are more abundant in codons that are farther from stop codon (*t* test, $p<0.05$) and more changeable (Pearson, $R=0.45$, $p=0.06$)

(Fig. 3b; $R=0.45$, $p=0.06$). One can thus conclude that the codon bias in epitopes is correlated with both changeability and distance from stop. However, looking at both of these results in combination, it would appear that while both characteristics are important, distance from stop codon is more so. As can be seen in Fig. 2, in leucine, the twofold redundant subgroup (TTA/G) is not overrepresented, since it is a single mutation away from stop. Similarly, the AGA codon from the twofold subgroup or arginine is also not overrepresented in epitopes, in comparison to AGG, that is, both changeable and far from stop and is overrepresented in epitopes. In serine, the twofold subgroup codons (AGT/C) both have more than one mutation from stop and they are both overrepresented.

Epitope lineage exhibits abundant toggling mutations and indications of long-term evolution

In order to test if epitopes do toggle more than other regions in the same genes, we constructed phylogenetic trees for a large number of genes from human viruses. We used all genes, where we could find at least 100 sequences per gene. For each tree, we divided the gene positions into epitopes and non-epitopes (Fig. S2, see Methods for details). We then defined a toggling mutation as a replacement mutation from A to B that is followed by replacement mutation from B to A in the same path from the root of the tree to a leaf. Overall, the fraction of toggling (out of all replacement mutations) is significantly higher in epitopes than in other regions (Fig. 4a, $p<0.05$). This is also true at the single protein level for most proteins tested. The interesting exception to this rule are tat and rev. These proteins have been reported by us and others to lose their epitopes very fast in HIV evolution and to have maintained a very limited number of epitopes. An in-depth analysis of the accumulation of escape mutations in tat and rev can be found in the study of Vider-Shalit et al. (2009).

Validation using published epitopes

To confirm our computational results, we have repeated the above analysis with published experimental epitopes. We analyzed HIV-I clad B sequences, where a large enough number of published epitopes are available (http://iedb.org/; http://www.hiv.lanl.gov/); (Vita et al. 2009). Figure S3 shows the toggling mutations as a fraction from all replacement mutations in published epitopes from HIV proteins and in non-epitope sequences. Four out of six proteins detected show higher toggling fraction in epitopes compared to non-epitopes ($p<1.e{-}8$). Of the remaining two, in env, the toggling fraction is slightly higher in non-epitopes, although to a lesser extent ($p<1.e{-}4$), and in pol, there is no difference. The simplest explanation for the toggling selection outside of epitopes in env is that this is a protein expressed on the surface of the virus that needs to avoid an antibody response, as was indeed reported (Wei et al. 2003). However, the detailed analysis of

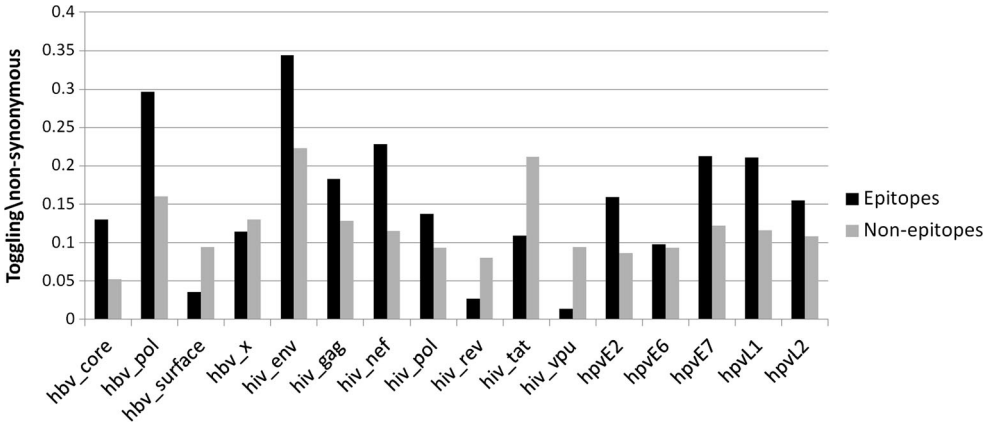

Fig. 4 Toggling mutations in human viruses' lineages. Toggling (back and forth non-synonymous mutation) fraction from all non-synonymous mutations was calculated for viral genes in epitopes (*black bars*) and non-epitopes (*grey bars*), using phylogenetic analysis. Toggling fraction was higher in epitopes for most genes in predicted epitopes (*t* test, $p<0.05$)

the antibody response is beyond the scope of the current analysis.

Toggling mutations occur at changeable codons

We have demonstrated above the preferential codon usage of changeable codons in epitopes. We hypothesize that this occurs because a population of viruses with more changeable codons would be more able to toggle back and forth from the amino acid which these codons encode. This phenomenon is predominantly observable in arginine, serine, and leucine. As above, these three amino acids are each encoded by six codons. The six codons are divided into twofold and fourfold redundant subgroups. However, looking merely at the codon does not give us a direct measure of this dynamic. To this end, we used a phylogenetic analysis (see Methods), to track toggling mutation from and to codon encode for arginine, serine, and leucine. Specifically, we measured all the toggling "cycles" in these three sixfold redundant amino acids with a cycle being a set of mutations that start with an amino acid and eventually return to it. We compared the fraction of cycles within and between the changeable (twofold redundant) and stable (fourfold redundant) codon groups of each amino acid in epitope and non-epitope regions (see Methods). If an immune-induced selection drives this changeable codon usage, we would expect to see enrichment in toggling mutations around changeable codons in epitope regions compared to non-epitope regions.

Indeed, when epitopes and non-epitopes are compared, toggling mutations in epitopes occur within the changeable group of codons significantly more than in non-epitopes. For example, in arginine, the stable group contain CGA/C/T/G, and the changeable twofold redundant group contains AGA/G. In this case, about one half of the toggling in arginine are within the changeable group in comparison with non-epitopes where AGA/G are used about one third of the times (chi-squared test; $p<0.05$) (Fig. 5). Similarly, mutations within the stable (fourfold redundant) group are found less in epitopes than in non-epitopes. However, the mutations between these groups are similar in epitopes and in non-epitopes.

We here infer patterns of codon usage in the population that allow more rapid changes. It is only at the level of population that toggling codons are stable. By following the phylogenetic pattern exhibited by these viruses, we see that changing between unstable positions can indeed result in their prevalence in the population despite their changeability in each specific individual.

Not surprisingly, outside of epitopes, where bias is not influenced by a selective advantage in maintaining changeable codons toggling occurs mainly from the four-codon less changeable group or between the changeable and stable groups.

Proteasomal cleavage sites are conserved among hosts and therefore do not exhibit toggling evolution

While features such as HLA binding are host-specific and are therefore expected to lead to a high level of toggling mutations, selection against proteasomal cleavage is not host-specific. The proteasome is highly conserved among humans and among mammals in general. Therefore, proteasomal cleavage sites are not expected to pass toggling mutations. We have compared the fraction of toggling mutations in cleavage sites and compared it to mutations in epitopes. Mutations in cleavage sites can remove epitopes, but they are not diverse among hosts, while mutations in the epitope body are host-specific. As expected, cleavage sites exhibit toggling mutation frequencies more similar to the ones of non-epitopes than the ones of epitope (Fig. 6) (MSE(cleavage, epitopes)=8.6e−3, MSE(cleavage, non-epitopes)=1.4e−3).

Epitopes remain where they have no choice

We have shown above multiple indicators of the heightened frequency of toggling mutations and the resulting effect on the codon usage. These all point to the same conclusion that epitopes remain in regions that are important for viral fitness.

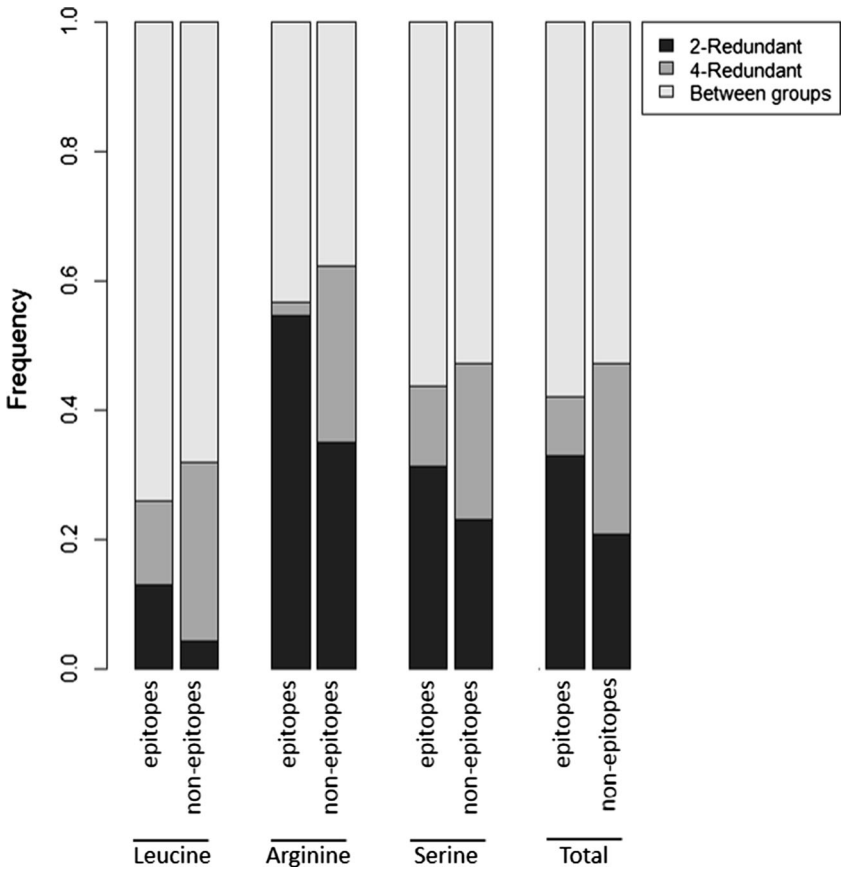

Fig. 5 Codons' changeability in toggling mutations in epitopes and non-epitopes. Toggling mutations either from leucine, arginine, and serine to themselves were analyzed. For all these amino acids, toggling mutations in epitopes occurred within changeable group of codons (twofold redundant) to a greater extent than in non-epitopes (chi-squared test, $p<0.05$). *Each circle* represents a different amino acid, and *each color* represents a different type of toggling mutations. All mutations start and end within the same amino acid

Thus, the positions where epitopes remain should be highly conserved. However, the extent of their conservation may not be the same in all epitope positions.

In order to test this conclusion, let us separate the viral sequences into positions with a high mutation fitness costs and those with a low fitness cost. We expect epitopes to be present more in the conserved set of positions than the second one. Thus, we expect epitope positions to be more conserved than non-epitope positions. In order to test that this is indeed the case, we computed the fraction of synonymous and non-synonymous mutations in published epitope positions in HIV proteins (HIV-EP) and non-epitope positions (HIV-

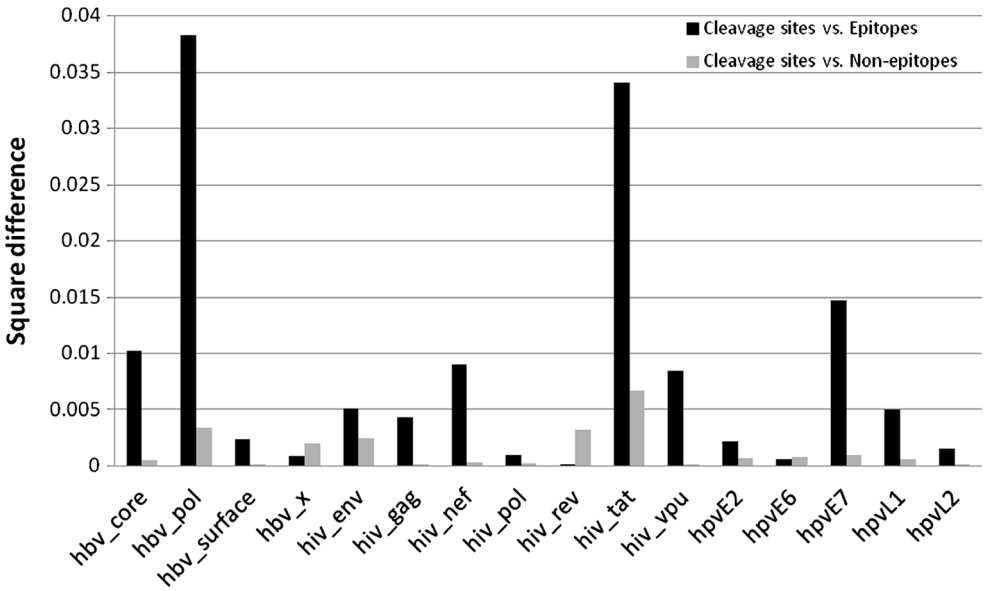

Fig. 6 Toggling mutations in cleavage sites. Toggling/replacement ratio was calculated in proteasome and cleavage sites, and compared to both epitopes and non-epitopes. Toggling signature in cleavage site is more similar to the one of non-epitopes. *Each column* represents the square difference between the frequencies in the different groups. The mean square difference (MSE) is much larger between cleavage sites and epitopes than between cleavage sites and non-epitopes (MSE(cleavage, epitopes)=8.6e−3, MSE(cleavage, non-epitopes)=1.4e−3)

NEP). Indeed, the fraction of non-synonymous mutations is much lower in HIV-EP than in HIV-NEP in all tested proteins (Fig. 7a; t test, $p<0.006$). If these regions are conserved since they are functionally important, we expect the same regions to be also conserved in the same positions in the SIV sequences. We have thus aligned SIV sequences to the HIV sequences and tested the fraction of non-synonymous mutation in HIV-EP and HIV-NEP on the SIV sequences, and indeed, these positions are conserved in all tested proteins (Fig. 7b; t test, $p<0.0002$).

Still, a difference is expected between HIV and SIV. Since the published HIV epitope positions are based on human HLA molecules, we expect some of these positions not to be epitopes in SIV (since monkeys and humans have different MHC binding motifs). Let us focus on these specific positions. Such positions are conserved in both species, but in humans, they may toggle back and forth between the conserved sequence and the escape sequence inducing non-synonymous mutations. This toggling is not expected to occur in the SIV, since these positions are not epitopes in the primate immune system. We thus expect the fraction of non-synonymous mutations to be even lower in HIV-EP positions in the SIV than in the HIV sequences, but no difference in the HIV-NEP positions, as is indeed the case (paired t test, $p<0.05$ in HIV-EP; no significant difference in HIV-NEP) (Fig. 7).

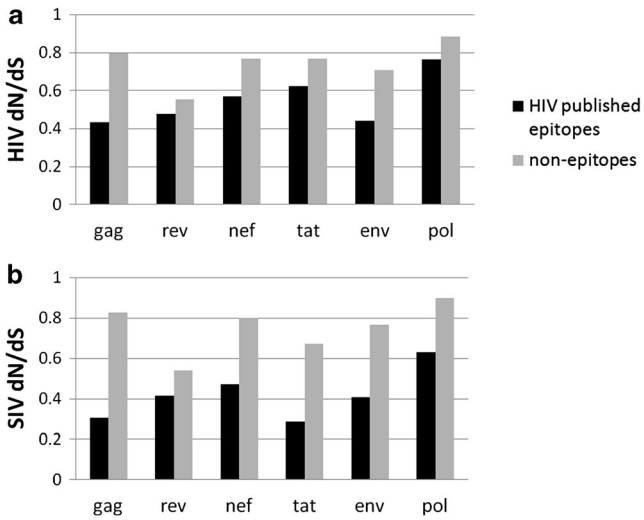

Fig. 7 Ratio of non-synonymous to synonymous mutations (dN/dS) in published HIV epitopes identified in both SIV and HIV sequences. Epitope positions were determined by appearance of published epitopes in the consensus sequence (allowing 1 mismatch for pol, rev, tat, env, and gag and 3 mismatches for nef). dN/dS ratio was calculated for epitope and non-epitope positions for each gene in HIV (**a**) and SIV (**b**). In both cases, epitopes have lower dN/dS ratio than non-epitopes (paired t test, $p<0.006$ and $p<0.002$ for HIV and SIV, respectively), with significantly greater differences in SIV (paired t test, $p<0.04$)

Discussion

We have shown here that viral epitopes remain where the functional price of their removal is too large and that this leads to more pronounced toggling selection and a more changeable codon usage in these regions.

Multiple previous studies have demonstrated that viruses are selected to evade immune detection by CTLs, such as MHC level downregulation or interference with the epitope cleavage process (Maman et al. 2011; Schmid et al. 2008). One of the main escape mechanisms is the removal of CTL epitopes through point mutations (Maman et al. 2011; Vider-Shalit et al. 2007, 2009). Few mutations are required to remove epitopes (Yokomaku et al. 2004), and the viral mutation rate is very high (Sanjuan et al. 2010). Thus, one would assume that most pathogenic viruses should be able to remove all epitopes in the most frequent alleles. Despite this, viral proteins still contain epitopes that can be presented on MHC-I molecules in alleles highly frequent in the population, some of which have been shown to induce a CTL response.

One possible explanation, pointed to by recent studies, is that some viral proteins, such as late expressed or lower expressed proteins, are under weaker selection to remove epitopes (Maman et al. 2011; Vider-Shalit et al. 2007, 2009). However, this cannot explain all the found epitopes, since viral proteins that are under a strong selection against immune recognition also still present some epitopes (Maman et al. 2011).

We have shown here that epitopes remain in genomic regions where removing them would lead to a high fitness cost. This was previously proposed at the level of specific epitopes (Seibert et al. 1995). We have shown here that this is actually a general feature of the remaining epitopes in most viral proteins. The balance between fitness cost and the need to escape the immune response has multiple expected effects at the nucleotide level that are not observed at the amino acid level. These effects are indeed observed in viral CTL epitopes:

1. Comparison of HIV and SIV sequences reveals that epitopes remain in conserved regions: regions that consist of HIV epitopes are more conserved in SIV.
2. Within the population of the human viruses, epitopes exhibit high level of toggling mutations. These back and forth mutations occur in response to the changing environment induced by the high HLA polymorphism. This polymorphism challenges the virus to escape from HLA recognition in one host and go back to the wild-type form as soon as it is transmitted to another host with different haplotype.
3. These toggling mutations bias the viral epitope codon usage toward codons more distant from stop codons than other parts of the viral sequences. Since epitopes toggle

back and forth, viruses with a high probability of mutating to a stop codon will be removed.

4. Similarly, codons in epitopes are more changeable than other parts of the viral sequences. The sequences that successfully toggle are those sequences that can efficiently move from one amino acid and back. This is most clearly exemplified in the codon usage of serine, arginine, and leucine. In these three sixfold redundant amino acids, epitope codon usage is biased both toward the use of their less stable twofold redundant block, and to toggling back in forth within it rather than passing through their stable fourfold redundant codons.

It is important to note that there is a built-in limitation in this kind of toggling escape. It presupposes some variability in the response where the escape mutant has a disadvantage, for instance, in a host with a more permissive HLA profile. This kind of indirect evolution involved with toggling mutations, accounts for short-term, host-specific evolution. Other aspects of immune pressure may be more rigid. For example, in the case of proteasomal cleavage, since there is practically no polymorphism in the proteasome, we would expect only direct evolution and no toggling. Indeed, we found that regions identified as proteasomal cleavage sites in different viruses are actually similar to regions outside epitopes in their lack of toggling selection.

These results taken together show an interesting case of evolution with a backup solution. Viruses evolve in real time to maximize their fitness in a host, and this leads to a population of viruses that maintains a possibility to easily mutate to avoid being detected by the immune response when needed and where immune variability allows.

References

Borghans JA, Beltman JB, De Boer RJ (2004) MHC polymorphism under host-pathogen coevolution. Immunogenetics 55:732–739

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Coffin JM (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267:483–489

Delport W, Scheffler K, Seoighe C (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog 4:e1000242

Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640–649

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5:113

Elena S, Agudelo-Romero P, Carrasco P, Codoner F, Martin S, Torres-Barcelo C, Sanjuán R (2008) Experimental evolution of plant RNA viruses. Heredity 100:478–483

Ginodi I, Vider-Shalit T, Tsaban L, Louzoun Y (2008) Precise score for the prediction of peptides cleaved by the proteasome. Bioinformatics 24:477–483

Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287–299

Hershberg R, Petrov DA (2009) General rules for optimal codon choice. PLoS Genet 5:e1000556

Hershberg U, Shlomchik MJ (2006) Differences in potential for amino acid change after mutation reveals distinct strategies for kappa and lambda light-chain variation. Proc Natl Acad Sci U S A 103:15963–15968

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143–155

Kijak GH, Currier JR, Tovanabutra S, Cox JH, Michael NL, Wegner SA, Birx DL, McCutchan FE (2004) Lost in translation: implications of HIV-1 codon usage for immune escape and drug resistance. AIDS Rev 6:54–60

Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719

Lichterfeld M, Yu XG, Le Gall S, Altfeld M (2005) Immunodominance of HIV-1-specific CD8(+) T-cell responses in acute HIV-1 infection: at the crossroads of viral and host genetics. Trends Immunol 26:166–171

Louzoun Y, Vider T, Weigert M (2006) T-cell epitope repertoire as predicted from human and viral genomes. Mol Immunol 43:559–569

Maman Y, Blancher A, Benichou J, Yablonka A, Efroni S, Louzoun Y (2011) Immune-induced evolutionary selection focused on a single reading frame in overlapping hepatitis B virus proteins. J Virol 85:4558–4566

McMichael AJ, Gotch FM, Noble GR, Beare PA (1983) Cytotoxic T-cell immunity to influenza. N Engl J Med 309:13–17

Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 171:1741–1749

Plotkin JB, Dushoff J (2003) Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci 100:7152–7157

Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R (2010) Viral mutation rates. J Virol 84:9733–9748

Schmid BV, Kesmir C, de Boer RJ (2008) The specificity and polymorphism of the MHC class I prevents the global adaptation of HIV-1 to the monomorphic proteasome and TAP. PLoS One 3:e3525

Seibert SA, Howell CY, Hughes MK, Hughes AL (1995) Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 12:803–813

Sherman LA (2006) To each (MHC molecule) its own (binding motif). J Immunol 177:2739–2740

Torres BA, Kominsky S, Perrin GQ, Hobeika AC, Johnson HM (2001) Superantigens: the good, the bad, and the ugly. Exp Biol Med (Maywood) 226:164–176

Vider-Shalit T, Louzoun Y (2011) MHC-I prediction using a combination of T cell epitopes and MHC-I binding peptides. J Immunol Methods 374:43–46

Vider-Shalit T, Fishbain V, Raffaeli S, Louzoun Y (2007) Phase-dependent immune evasion of herpesviruses. J Virol 81:9536–9545

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

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2009) The immune epitope database 2.0. Nucleic Acids Res: doi:10.1093/nar/gkp1004

Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD, Shaw GM (2003) Antibody neutralization and escape by HIV-1. Nature 422:307–312

Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412:331–333

Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S (1991) Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res 19:6119–6122

Yokomaku Y, Miura H, Tomiyama H, Kawana-Tachikawa A, Takiguchi M, Kojima A, Nagai Y, Iwamoto A, Matsuda Z, Ariyoshi K (2004) Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. J Virol 78:1324–1332