Robust self-nonself discrimination requires negative T cell selection on non-random peptides

Negative selection on well-chosen peptides allows T cells to distinguish self from highly self-similar foreign.

Inge M. N. Wortel,[1] Can Keşmir,[2] Rob J. de Boer,[2]
Judith N. Mandl,[3] Johannes Textor[1,2*]

[1]Department of Tumor Immunology, Radboud Institute of Molecular Life Sciences, Nijmegen, the Netherlands

[2]Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, the Netherlands

[3]Department of Physiology, McGill University, Montreal, Canada

[*]To whom correspondence should be addressed; E-mail: Johannes.Textor@radboudumc.nl.

Self-foreign discrimination by the immune system was long thought to arise because negative selection in the thymus silences self-reactive T cells. Yet recent data show that this silencing is remarkably incomplete. Here we ask how a repertoire containing many self-reactive cells can nevertheless discriminate self from foreign. We address this question using realistic-scale computational models of the T cell repertoire. Our models show that when foreign peptides differ systematically from self, moderate T cell cross-reactivity skews the post-selection repertoire towards foreign recognition. When no such systematic differences exist, self-foreign discrimination is only possible if peptide presentation in the thymus minimizes the co-occurrence of similar, redundant self peptides. These results imply that negative selection needs to be based on non-random self peptides to allow robust self-foreign discrimination for both self-similar and -dissimilar pathogens.

To eliminate pathogens without damaging healthy cells, the immune system must discriminate between self and foreign (nonself). The innate arm of the immune system is able to do so with a limited number of germline-encoded receptors that recognize pathogen-associated molecular patterns. By contrast, the adaptive arm of the immune system, which is found in all jawed vertebrates and is mediated by T and B lymphocytes, uses a vastly diverse repertoire of receptors to generate specific protective responses against any pathogen it encounters [1, 2]. For example, humans have a repertoire of at least 10^7 different T cells [3], each expressing one or two of the $>10^{15}$ unique receptor sequences that can arise from the stochastic recombination of V(D)J gene segments and addition of non-templated nucleotides [4, 5]. These T cell receptors (TCRs) recognize short foreign peptides presented on major histocompatibility complex (MHC) molecules on the surface of infected or cancerous cells.

However, the random TCR generation process inevitably also produces TCRs that recognize self peptides presented by healthy cells. It was long thought that the majority of these self-reactive receptors are effectively eliminated during T cell development in the thymus through a process termed negative selection [6], but recent studies have shown that this process is nowhere near as complete as it was thought to be [7, 8, 9]. In fact, given that T cells may only encounter an estimated 10^3-10^5 different peptides during negative selection – a small fraction of all MHC-binding self peptides – it is not trivial how negative selection can achieve self-foreign discrimination at all [10, 11, 12].

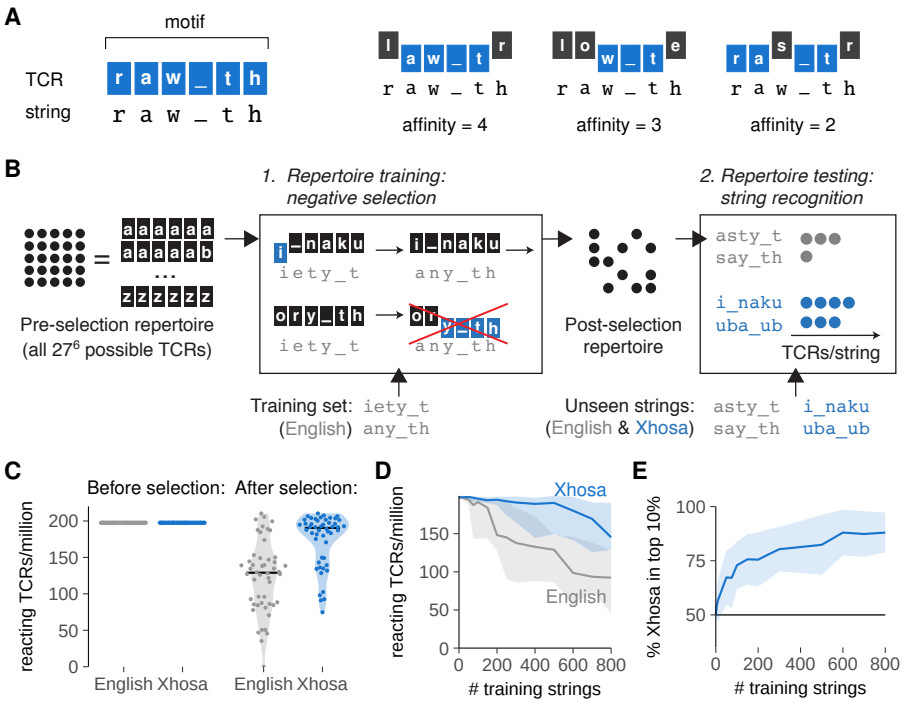

Figure 1: Negative selection on a subset of the whole "self" can achieve self-foreign discrimination. (a) Our model of string recognition represents TCRs by a *binding motif* – the string they bind to most strongly (top). Their affinity for any given string equals the maximum number of adjacent positions where the binding motif matches the string (bottom). (b) Simulating negative selection *in silico*: (1) TCRs in the unbiased pre-selection repertoire (with all possible $27^6 \approx 400$ million TCR motifs of 6 characters [a-z and _]) are deleted if their affinity for any of the *training strings* exceeds the functional response threshold t. (2) Unseen English and Xhosa strings are exposed to the post-selection repertoire to find the number of remaining TCRs reacting to them (that is, TCRs with affinity $\geq t$). (c) Reacting TCRs per million of unseen English and Xhosa strings, before and after negative selection on 500 English strings. Horizontal lines indicate medians. (d) Median and interquartile range of English- and Xhosa-reactivity after negative selection on English strings. (e) Percentage of Xhosa strings among the 10% of strings with the most reacting TCRs after negative selection on English strings (mean±standard deviation, SD, of 30 simulations). No discrimination should result in equal amounts (50%) of English and Xhosa strings in this top 10%. Throughout this figure, we tested 50 English and 50 Xhosa strings using an affinity threshold $t = 3$ for negative selection.

Results

An artificial immune system discriminates self from foreign after negative selection

To investigate how incomplete negative selection can still foster effective self-foreign discrimination, we devised an "artificial immune system" (AIS) [13]. Our AIS is an algorithmic model of a T cell repertoire [14], similar to how an artificial neural network (ANN) is an algorithmic model of the central nervous system. Because it was important to consider T cell repertoires of realistic scale and complexity, we exploited data compression techniques that allow building AISs containing billions of TCRs [15].

Like ANNs, AISs are not only used for *in silico* modelling of the biological system, but also as general-purpose classification algorithms. We took advantage of this property by first using a well-interpretable classification problem outside of immunology to investigate how a TCR repertoire could discriminate a foreign peptide from a self peptide it has not encountered during selection. Specifically, we built an AIS that distinguishes English from other languages based on short strings (letter sequences) of text. This artificial problem mimics the task of self-foreign discrimination because in both cases, classes (languages or proteomes) are to be distinguished based on a limited amount of information (short strings or peptides). A useful property of the language problem is that it can take on a range of difficulties, as very dissimilar languages such as English and the South-African language Xhosa are much easier to distinguish than related languages such as modern and medieval English.

Our model is based on an existing AIS [14, 15, 16] that represents each TCR as a binding motif, and defines a TCR's affinity for a string as the maximum number of adjacent positions where this motif matches the string

(Fig. 1A) [17]. A TCR is defined to *react* to all strings for which it has an affinity of at least some threshold t, which represents a functional response threshold rather than a mere binding threshold. Crucially, reaction does not require a perfect match between the string and TCR motif. Thus, our TCRs are *cross-reactive* and react to multiple, related peptides.

To test how well TCR repertoires could discriminate between two very dissimilar languages (English and Xhosa) after incomplete negative selection, we started with an unbiased pre-selection repertoire with equal numbers of TCRs reacting to English and Xhosa, and then performed *in silico* negative selection on an English *training set* by deleting all TCRs reacting to any of the (<1000) training strings (Fig. 1B, using a threshold t = 3 leading to intermediate cross-reactivity). Although this negative selection did not completely abrogate TCR reactivity towards English strings outside of the training set, it still biased the post-selection repertoire to contain more TCRs reacting to Xhosa than to English (Fig. 1C,D).

Given that peptides to which many TCRs react tend to elicit stronger immune responses [18], it is important that these most frequently recognized peptides are predominantly foreign. The 10% most frequently recognized strings in our simulation were indeed predominantly Xhosa strings (Fig. 1E). The affinity distribution of these TCR interactions was shifted towards higher affinities for Xhosa, but only very slightly (Fig S1A). For sake of simplicity, we therefore focus only on the number of reacting TCRs throughout this paper, rather than considering different affinities separately.

Discrimination success relies on moderate cross-reactivity and sequence dissimilarity

These results confirm that our AIS can easily distinguish English from Xhosa even after incomplete negative selection. To investigate in more detail under which conditions this discrimination arises, we analyzed which TCRs were deleted during negative selection on English strings (Fig. 2). TCRs reacting to "unseen" English strings (those absent from the training set TCRs were exposed to during negative selection) had a reduced survival compared to TCRs reacting to Xhosa strings (Fig. 2A). Because TCRs are only deleted when they react to at least one string in the training set, this implies that strings eliciting reactions from the same TCRs tend to represent the same language. To visualize this, we created graphs in which each node represents a string, and two nodes become connected neighbors when at least 5 TCRs per million pre-selection TCRs react to both of them (Fig. 2B). Indeed, neighbor strings are largely from the same language (Fig. 2B, left), which is quantified by the *concordance*, the average proportion of neighbors from the same language. To show that the high concordance (0.81) of English and Xhosa strings represents intrinsic differences between English and Xhosa strings, we randomly divided English strings into two groups and constructed a similar graph, which as expected has a concordance of only 0.5 (Fig. 2B, right). This confirms that our TCRs can only discriminate between two sets of strings that are intrinsically different.

Our results indicate two key requirements for achieving self-foreign discrimination through negative selection on an incomplete subset of self: an appropriate level of TCR *cross-reactivity* towards multiple, related strings, and sufficient *dissimilarity* between self- and foreign.

To illustrate the importance of cross-reactivity, we set the affinity threshold in our model to t = 6, so that each TCR was maximally specific and only reacted to the one string matching its binding motif perfectly (i.e., no cross-reactivity). The corresponding graph contains no neighbors at all (Fig. 2C, left) and has a concordance of 0.5 (Fig. 2D,E). Consequently, maximal TCR specificity abolishes self-foreign discrimination in our model (Fig. 2E) because without cross-reactivity, negative selection cannot delete TCRs for strings that are not part of the training set – it therefore deletes very few TCRs (Fig. S1B). However, very low specificity (t = 1) is equally problematic as it results in a graph where any two strings are neighbors irrespective of language (Fig. 2C, right), which leads to low concordance even between dissimilar languages (Fig. 2D,E), poor self-foreign discrimination (Fig. 2E), and often even deletion of the entire repertoire (Fig. S1B). Only intermediate specificities allow TCRs to preferentially react to either English or Xhosa strings (Fig. 2C, middle). This results in both a high concordance (Fig. 2D,E) and a preference for Xhosa-reactivity in the post-selection repertoire (Fig. 2E).

As shown in Fig. 2B, even an optimal level of cross-reactivity will not result in a high concordance unless the languages are intrinsically different. The accomplished level of self-foreign discrimination therefore depends directly on the similarity between self-

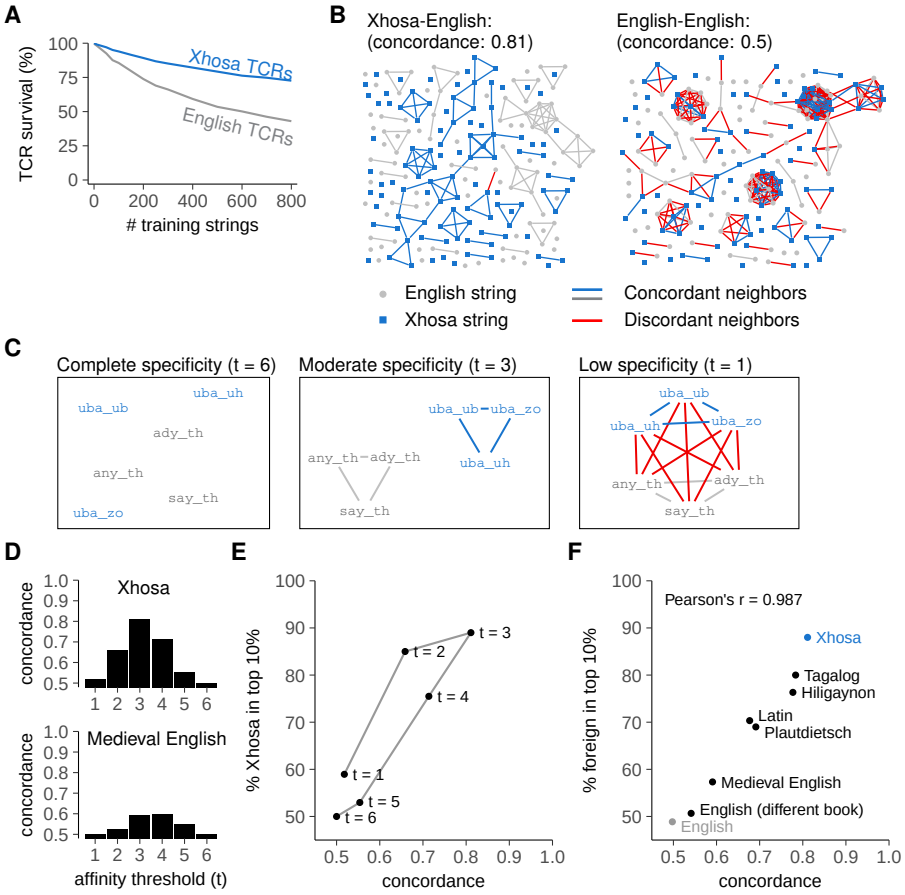

Figure 2: Discrimination requires moderate TCR cross-reactivity and dissimilar self- and foreign strings. (a) Mean percentage of surviving TCRs reacting to English and Xhosa strings after negative selection (using threshold t = 3). (b) String similarity visualized in a graph where nodes (strings) are neighbors (connected by edges) if at least 5/million pre-selection TCRs react to both. (c) Cross-reactivity increases the number of edges between example English and Xhosa strings (demonstrated here for a few examples). Edges between strings from different languages are shown in red. (d) Concordance in the English-Xhosa and English-Medieval English graphs for different thresholds t. (e) Concordance and discrimination between English and Xhosa for different thresholds t. Negative selection was performed on 800 English strings. (f) Language concordance versus enrichment of foreign strings among the top 10% most frequently recognized strings after negative selection (t = 3, selection on 800 English strings). Pearson's correlation coefficient r = 0.977, with 95% confidence interval [0.890, 0.995]. The control "English" compares two sets of English strings from the same book that was used for training (Moby Dick), whereas "English (different book)" compares unseen English strings from the training book to those from the Bible.

and foreign sequences. Indeed, when we repeated our analysis for a number of other languages with varying similarity to English, we found a linear correlation between concordance and the acquired level of discrimination (Fig. 2F). This was a property of the tested languages rather than the specific texts chosen, as our model could not discriminate between English strings from different books (Fig. 2F).

Sequence similarity hampers discrimination between self- and foreign peptides

These results on natural languages suggest that TCR cross-reactivity and sequence dissimilarity should also be important for self-foreign discrimination in the immune system. We therefore applied our AIS model to self-foreign discrimination by $CD8^+$ T cells, which recognize peptides bound to the MHC class I (MHC-I) complex with a typical length of nine amino acids (AAs). The six residues at positions 3-8 are thought to be most relevant for TCR binding [19]. Accordingly, we modified our TCR model to accommodate 6-mer peptide sequences rather than six-letter strings (Fig. 3A). Setting the affinity threshold to an intermediate value of t = 4 in this model allowed each TCR to react to roughly one in every 55,000 peptides (Fig. S2A) – a cross-reactivity level that reasonably matches an experimental estimate of one in 30,000 [20]. Furthermore, at this level of cross-reactivity, peptides elicited reactions from 0 to 20 TCRs per million in our simulated

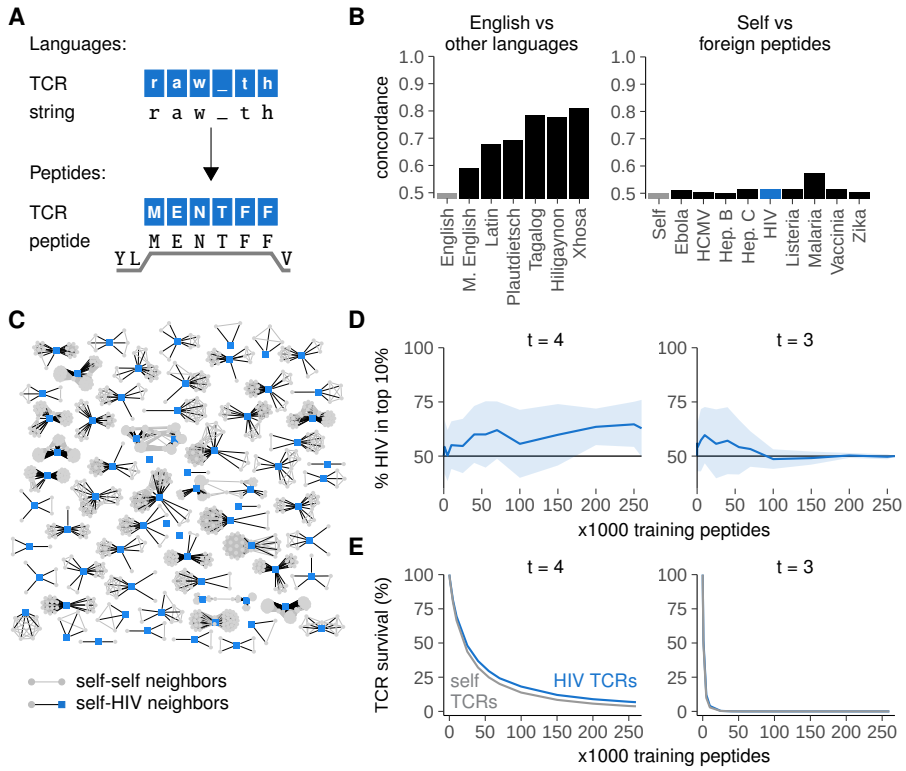

Figure 3: High similarity between self- and foreign peptides hampers their discrimination by the immune system. (a) TCR binding to peptides on MHC-I (HLA-A2:01) focuses on the 6 residues at positions 3-8 and resembles the TCR-string model as in Fig. 1A. (b) Concordance for English versus other languages (left) compared to that for self versus foreign peptides (right). (c) Graph of HIV peptides and their neighbors. Edges connect peptides that have at least 5/million pre-selection TCRs in common. (d) Percentage of HIV-peptides among the 10% most frequently recognized peptides after negative selection (mean±SD of 30 simulations). (e) Mean percentage surviving TCRs for self and HIV peptides after negative selection.

repertoires (Fig. S2B), in line with experimental data [21, 22, 23, 24]. These results suggest that the cross-reactivity level of TCRs roughly matches that of our model at $t = 4$, well within the "moderate" range allowing discrimination between dissimilar strings (Fig. 2D,E).

To examine whether self- and foreign peptides are dissimilar enough to allow self-foreign discrimination, we first predicted MHC-I-binding peptides from the human proteome [25] and used the residues 3-8 as MHC-bound self peptides in our model. To obtain foreign sequences, we predicted MHC binders for a variety of pathogens associated with T cell immunity: the malaria parasite, the bacterium *Listeria monocytogenes*, and the viruses ebola, hepatitis B, hepatitis C, human cytomegalovirus (HCMV), human immunodeficiency virus (HIV), and vaccinia (Table S1).

Graphs of self versus foreign peptides had strikingly low concordances (Fig. 3B)[17], barely exceeding the control concordance observed between two random, different sets of self peptides ("Self", negative control), and lower than the concordance we had observed between modern and medieval English. This was a property of the sequences themselves rather than the chosen threshold t (Fig. S3A). In a graph of all HIV peptides and their neighbors, the majority of HIV peptides had many self neighbors whereas none of them had HIV neighbors (Fig. 3C) – indicating that most HIV peptides are more similar to peptides from the human proteome than to other HIV peptides.

This high similarity between self- and foreign peptides suggests that achieving self-foreign discrimination via negative selection is difficult. Indeed, although the realistic cross-reactivity at $t = 4$ allowed some discrimination between self- and HIV peptides as shown by a small enrichment of HIV among most frequently recognized peptides (Fig. 3D, left), this effect came nowhere close to that observed for languages (Fig. 1E), even with very large numbers of training self peptides. Consistent with this observation, the survival of self-reactive TCRs was only slightly lower than that of HIV-reactive TCRs (Fig. 3E, left). These results were not specific for HIV peptides, as we obtained similarly low levels of self-foreign discrimination for all other

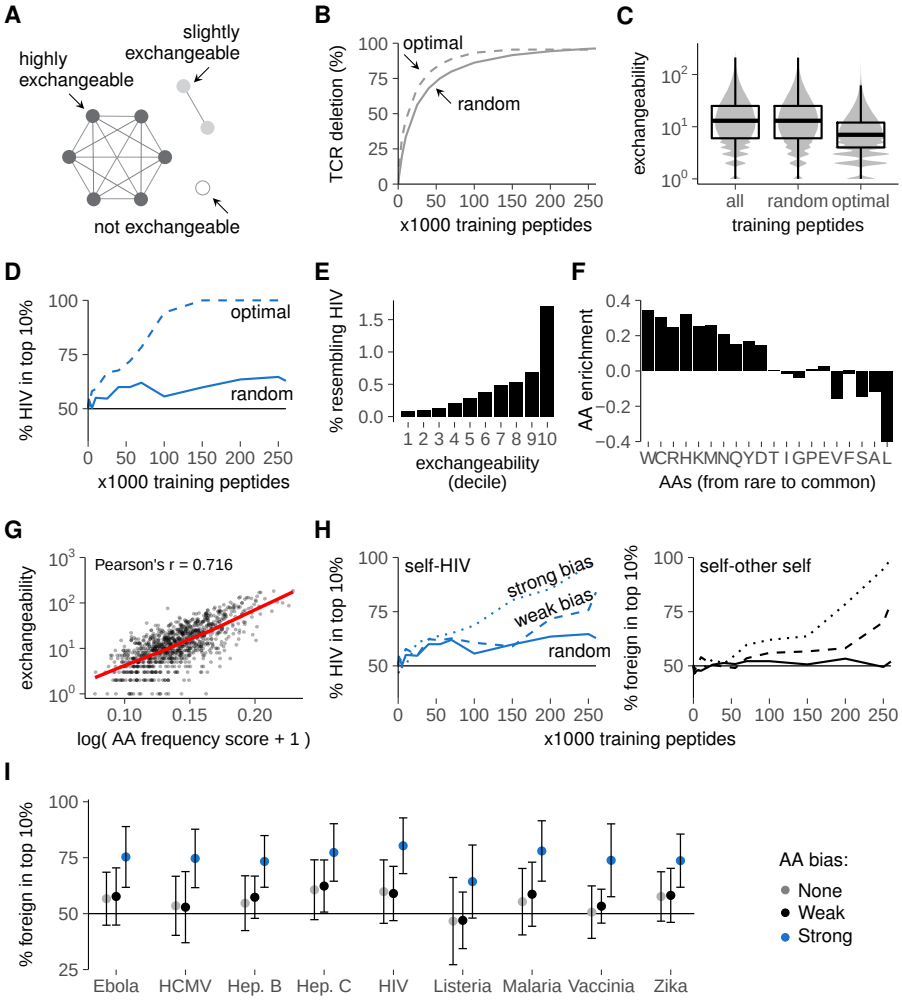

Figure 4: Improved self representation during negative selection allows self-foreign discrimination.
(a) Self peptides from large clusters delete the same TCRs as their neighbors and are thus exchangeable during negative selection, whereas peptides from small clusters are not. (b) Mean percentage of self-reactive TCRs deleted by optimal versus random training sets of self peptides during negative selection. (c) Peptide exchangeability distribution in the full set of all self peptides compared to that in random and optimal subsets of 100,000 peptides. Exchangeability is defined as the number of self neighbors + 1. (d) Self-HIV discrimination after selection on random or optimal training sets. (e) Percentage of self peptides with HIV neighbor(s) plotted against exchangeability (self peptides were divided into 10 equal-number deciles from low to high exchangeability). (f) AA enrichment in optimal training set. Enrichment is the log of the observed frequency divided by the frequency among all self peptides. Negative values indicate depletion. (g) Exchangeability versus peptide AA frequency score in a random sample of 1000 self peptides (frequency score is low for peptides with many rare AAs, [17]). Pearson's correlation coefficient r = 0.716, with 95% confidence interval [0.684, 0.745]. (h) Discrimination after negative selection on self peptides chosen with a (weak/strong) bias for rare AAs. Plots show self-HIV discrimination (left), and self-other self discrimination (right, where a random sample of self was assigned the label "foreign" before selection on training sets from the remaining "self" peptides). (e) Self-foreign discrimination for different pathogens after negative selection on 150,000 self peptides chosen randomly or with AA bias. Negative selection in panels b, d, h, and i was performed with t = 4, and results were plotted as mean±SD of 30 simulations.

pathogens tested (Fig. S3B). Self-HIV discrimination was even worse for t = 3 and rapidly disappeared completely as TCR survival diminished for large training sets (Fig. 3D,E, right), confirming that self-foreign discrimination becomes more difficult when TCRs are too cross-reactive.

Selection on non-random peptides greatly improves self-foreign discrimination

Thus, although incomplete negative selection can achieve self-foreign discrimination in principle, achieving sufficient discrimination is very difficult in practice because self- and foreign peptides can be extremely similar and therefore can be recognized by the same TCRs. Clearly, the immune system must overcome this problem in order to balance the removal of self-reactivity with the

preservation of foreign recognition. It has previously been suggested that thymic selection should occur on a non-random set of self peptides to achieve self-foreign discrimination [12]. We therefore used our model to investigate what an "optimal" set of self peptides would look like, and how much this might improve self-foreign discrimination.

As a starting point, we based the optimization of the training set on the peptide cluster structure as observed in Fig 3C. The large clusters in this graph contain many similar self peptides, which can delete the same TCRs during negative selection (Fig 4A). Exchanging one such peptide for one of its neighbors during selection thus has little effect on the post-selection repertoire – and presenting both has little added value. By contrast, self peptides in smaller clusters are far less *exchangeable* (Fig 4A): their TCRs cannot be removed as easily by other peptides. Thus, negative selection on randomly chosen training sets is inefficient: these sets often contain several exchangeable peptides that delete the same TCRs, while simultaneously missing many non-exchangeable peptides and allowing the corresponding self-reactive TCRs to escape. We therefore used combinatorial optimization techniques [17] to compute peptide combinations that deleted as many different self-reactive TCRs as possible ("optimal" training sets, Fig 4B). As expected, these optimal training sets contained fewer exchangeable peptides (Fig 4C, where exchangeability equals the number of self neighbors plus one).

We then tested whether these training sets optimized for inducing *tolerance* could also establish self-foreign *discrimination*. This is not guaranteed, as the latter requires not only the removal of self-reactive TCRs, but also the preservation of foreign-reactivity. Nevertheless, our optimal training sets substantially improved self-foreign discrimination (Fig 4D). This seems to be a consequence of the enrichment for low exchangeability peptides (Fig 4C), which are less likely to delete HIV-reactive TCRs (Fig 4E). Importantly, this discrimination still required appropriate TCR cross-reactivity and was absent at $t = 3$ (Fig. S4B). From these results, we conclude that negative selection on a representative set of self peptides can alleviate the problem of self-foreign similarity, but only when TCRs are sufficiently specific.

Obviously, our optimal training sets are artificial, and biological negative selection cannot calculate which self peptides should be present in the thymus. We therefore investigated how a representative set of self peptides might reasonably be obtained during real negative selection. Analysis of our optimal training sets revealed an enrichment for rare AAs compared to the total set of self peptides (Fig. 4F). Interestingly, peptides with many rare AAs were typically less exchangeable (Fig 4G). This finding suggests that training sets enriched for rare AAs – similar to our optimal sets – contain fewer exchangeable peptides, and might thus result in better self-foreign discrimination.

To test this hypothesis, we again generated training sets of different sizes, but this time picked our training peptides with a probability that depended on the AA composition of each peptide [17]. These probabilities introduced either a weak or a strong bias for self peptides with rare AAs, mimicking the AA enrichment pattern observed in our optimal training sets. This AA bias substantially improved self-foreign discrimination after negative selection, for HIV (Fig. 4H, left) and all other pathogens tested (Fig. 4I, S4A). Interestingly, this strategy also worked when we first set aside a random sample of other self peptides as "foreign" before selecting training sets from the remaining "self" peptides. In this scenario, biased training sets still yielded substantial self-"foreign" discrimination, whereas random sets did not (Fig 4H, right). This result demonstrates that negative selection on non-random training peptides facilitates self-foreign discrimination – even in the extreme case where no inherent difference between self and foreign peptides exists.

Discussion

Our AIS model explains how negative selection on an incomplete set of self peptides can nonetheless bias a T cell repertoire towards foreign recognition. We demonstrate that a non-random subset of self peptides enriched for rare AAs can balance the removal of self-reactive TCRs with the preservation of foreign-reactive receptors. Importantly, this strategy works even when self and foreign peptides are not inherently different. In fact, for the pathogens we considered, the similarity to self was so high that it is hard to conceive how any self-foreign discrimination could be achieved through negative selection on random peptides. By contrast, a "smart" peptide presentation strategy could still ensure that the peptides best recognized by the immune system are predominantly foreign – even in this difficult

scenario. This notion reconciles textbook negative selection theory with recent observations that T cells see only a fraction of all self peptides during thymic selection, and that even healthy individuals have many self-reactive T cells [7].

Our finding that non-random peptide presentation is a prerequisite for efficient self-foreign discrimination raises the question how the thymus might obtain a preference for presenting low-exchangeability peptides. Although it remains unclear exactly which and how many peptides a T cell sees during selection, the importance of the thymic peptidome in shaping the TCR repertoire is evident from the existence of specialized antigen presenting cells, transcription factors such as AIRE, and even special proteasomes controlling thymic peptide presentation [26]. We suggest that the biased presentation of low-exchangeability peptides required for self-foreign discrimination might arise from special binding preferences of thymic antigen presentation proteins. As has already been shown for the thymoproteasome during thymic positive selection [27, 28], such binding preferences can enrich for specific subsets of self peptides and thereby impact the ability of a TCR repertoire to recognize self and foreign. While a bias for specific AAs such as described in this paper would be one way to enrich for low-exchangeability peptides, we do not exclude that other binding preferences could have a similar impact on self-foreign discrimination.

Importantly, our imperfect selection accomplishes self-foreign discrimination by also reducing the recognition of peptides the T cell repertoire has not seen during selection. This capability of the T cell repertoire to generalize beyond given examples is a fundamental property of learning systems [29], and allows the repertoire to perform a cognitive task: learning to distinguish self from foreign. Even though this learning process mechanistically differs from learning by the central nervous system, its high-level outcome is remarkably similar, and shares many properties with "slow learning" systems as described in psychology and neuroscience [30].

References and Notes

[1] Cooper MD and Alder MN. The Evolution of Adaptive Immune Systems. *Cell*, 124(4):815–822, 2006. doi:10.1016/j.cell.2006.02.001.

[2] Flajnik MF and Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):nrg2703, 2009. doi:10.1038/nrg2703.

[3] Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, Olshen RA, Weyand CM, Boyd SD, and Goronzy JJ. Diversity and clonal selection in the human t-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(36):13139–13144, 2014. doi:10.1073/pnas.1409155111.

[4] Davis MM and Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):334395a0, 1988. doi:10.1038/334395a0.

[5] Zarnitsyna V, Evavold B, Schoettle L, Blattman J, and Antia R. Estimating the Diversity, Completeness, and Cross-Reactivity of the T Cell Repertoire. *Frontiers in Immunology*, 4, 2013. doi:10.3389/fimmu.2013.00485.

[6] Silverstein AM. Autoimmunity *versus horror autotoxicus*: The struggle for recognition. *Nature Immunology*, 2(4):ni0401_279, 2001. doi:10.1038/86280.

[7] Yu W, Jiang N, Ebert PR, Kidd B, Müller S, Lund P, Juang J, Adachi K, Tse T, Birnbaum M, Newell E, Wilson D, Grotenbreg G, Valitutti S, Quake S, and Davis M. Clonal Deletion Prunes but Does Not Eliminate Self-Specific CD8+ T Lymphocytes. *Immunity*, 42(5):929–941, 2015. doi:10.1016/j.immuni.2015.05.001.

[8] Legoux FP, Lim JB, Cauley AW, Dikiy S, Ertelt J, Mariani TJ, Sparwasser T, Way SS, and Moon JJ. CD4+ T Cell Tolerance to Tissue-Restricted Self Antigens Is Mediated by Antigen-Specific Regulatory T Cells Rather Than Deletion. *Immunity*, 43 (5):896–908, 2015. doi:10.1016/j.immuni.2015.10.011.

[9] Davis M. Not-So-Negative Selection. *Immunity*, 43(5):833–835, 2015. doi:10.1016/j.immuni.2015.11.002.

[10] Detours V, Mehr R, and Perelson AS. Deriving Quantitative Constraints on T Cell Selection from Data on the Mature T Cell Repertoire. *The Journal of Immunology*, 164(1):121–128, 2000. doi:10.4049/jimmunol.164.1.121.

[11] Müller V and Bonhoeffer S. Quantitative constraints on the scope of negative selection. *Trends in Immunology*, 24(3):132–135, 2003. doi:10.1016/S1471-4906(03)00028-0.

[12] Vrisekoop N, Monteiro J, Mandl J, and Germain R. Revisiting Thymic Positive Selection and the Mature T Cell Repertoire for Antigen. *Immunity*, 41(2):181–190, 2014. doi:10.1016/j.immuni.2014.07.007.

[13] Castro Ld and Timmis J. *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer Science & Business Media, 2002.

[14] Forrest S, Hofmeyr SA, and Somayaji A. Computer Immunology. *Commun. ACM*, 40(10):88–96, 1997. doi:10.1145/262793.262811.

[15] Elberfeld M and Textor J. Negative selection algorithms on strings with efficient training and linear-time classification. *Theoretical Computer Science*, 412(6):534–542, 2011. doi:10.1016/j.tcs.2010.09.022.

[16] Percus JK, Percus OE, and Perelson AS. Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self-nonself discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, 90(5):1691–1695, 1993. doi:10.1073/pnas.90.5.1691.

[17] Materials and methods are available in the supplementary materials.

[18] Jenkins MK and Moon JJ. The Role of Naive T Cell Precursor Frequency and Recruitment in Dictating Immune Response Magnitude. *The Journal of Immunology*, 188(9):4135–4140, 2012. doi:10.4049/jimmunol.1102661.

[19] Frankild S, Boer RJd, Lund O, Nielsen M, and Kesmir C. Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for Holes in the T Cell Repertoire. *PLOS ONE*, 3(3):e1831, 2008. doi:10.1371/journal.pone.0001831.

[20] 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 JR, McMichael A, and Yewdell JW. Quantitating T Cell Cross-Reactivity for Unrelated Peptide Antigens. *The Journal of Immunology*, 183(7):4337–4345, 2009. doi:10.4049/jimmunol.0901607.

[21] Blattman JN, Antia R, Sourdive DJD, Wang X, Kaech SM, Murali-Krishna K, Altman JD, and Ahmed R. Estimating the Precursor Frequency of Naive Antigen-specific CD8 T Cells. *Journal of Experimental Medicine*, 195(5):657–664, 2002. doi:10.1084/jem.20001021.

[22] Alanio C, Lemaitre F, Law HKW, Hasan M, and Albert ML. Enumeration of human antigenspecific naive CD8+ T cells reveals conserved precursor frequencies. *Blood*, 115(18):3718–3725, 2010. doi:10.1182/blood-2009-10-251124.

[23] Legoux F, Debeaupuis E, Echasserieau K, Salle HDL, Saulquin X, and Bonneville M. Impact of TCR Reactivity and HLA Phenotype on Naive CD8 T Cell Frequency in Humans. *The Journal of Immunology*, 184(12):6731–6738, 2010. doi:10.4049/jimmunol.1000295.

[24] Schmidt J, Neumann-Haefelin C, Altay T, Gostick E, Price DA, Lohmann V, Blum HE, and Thimme R. Immunodominance of HLA-A2-Restricted Hepatitis C Virus-Specific CD8+ T Cell Responses Is Linked to Naïve-Precursor Frequency. *Journal of Virology*, 85(10):5232–5236, 2011. doi:10.1128/JVI.00093-11.

[25] Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, and Nielsen M. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, 61(1):1, 2009. doi:10.1007/s00251-008-0341-z.

[26] Klein L, Kyewski B, Allen PM, and Hogquist KA. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology*, 14(6):nri3667, 2014. doi:10.1038/nri3667.

[27] Nitta T, Murata S, Sasaki K, Fujii H, Ripen AM, Ishimaru N, Koyasu S, Tanaka K, and Takahama Y. Thymoproteasome Shapes Immunocompetent Repertoire of CD8+ T Cells. *Immunity*, 32 (1):29–40, 2010. doi:10.1016/j.immuni.2009.10.009.

[28] Sasaki K, Takada K, Ohte Y, Kondo H, Sorimachi H, Tanaka K, Takahama Y, and Murata S. Thymoproteasomes produce unique peptide motifs for positive selection of $CD8^+$ T cells. *Nature Communications*, 6:ncomms8484, 2015. doi:10.1038/ncomms8484.

[29] Gold EM. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967. doi:10.1016/S0019-9958(67)91165-5.

[30] McClelland JL, McNaughton BL, and O'Reilly RC. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995. doi:10.1037/0033-295X.102.3.419.

Acknowledgments

IW and JT performed experiments and analysed/interpreted data. All authors contributed to experimental design and formulation of theory. IW and JT wrote the manuscript; CK, RdB, and JM critically revised the manuscript.

IW was supported by a Radboudumc PhD grant. JT was supported by a Young Investigator Grant (10620) from KWF. CK and JT were supported by an NWO-ALW grant (823.02.014), and CK was supported by the EU HORIZON2020 program (APERIM project).

The authors declare no conflict of interest.

Supplementary materials

Materials and Methods

Figs. S1 to S4

Table S1