How the thymus designs antigen-specific and self-tolerant T cell receptor sequences

Andrej Košmrlj*, Abhishek K. Jha†, Eric S. Huseby‡, Mehran Kardar*§, and Arup K. Chakraborty†§¶‖

Departments of *Physics, †Chemical Engineering, ¶Chemistry, and ‖Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; and ‡Department of Pathology, University of Massachusetts, Worcester, MA 01655

Communicated by Herman N. Eisen, Massachusetts Institute of Technology, Cambridge, MA, August 15, 2008 (received for review June 2, 2008)

T lymphocytes (T cells) orchestrate adaptive immune responses that clear pathogens from infected hosts. T cells recognize short peptides (p) derived from antigenic proteins bound to protein products of the MHC genes. Recognition occurs when T cell receptor (TCR) proteins expressed on T cells bind sufficiently strongly to antigen-derived pMHC complexes on the surface of antigen-presenting cells. A diverse repertoire of self-pMHC-tolerant TCR sequences is shaped during development of T cells in the thymus by processes called positive and negative selection. Combining computational models and analysis of experimental data, we parsed the contributions of positive and negative selection to the design of TCR sequences that recognize antigenic peptides with specificity, yet also exhibit cross-reactivity. A dominant role for negative selection in mediating antigen specificity of mature T cells and a molecular mechanism for TCR recognition of antigen are described.

statistical mechanics | T cell antigen specificity | thymic selection

Because T cell receptor (TCR) genes undergo stochastic somatic rearrangement, most T cells express a distinct TCR, thereby enabling the T cell population to recognize many different antigenic short peptide (p)MHC complexes. TCR recognition of pMHC is both specific and degenerate. It is specific, because if a TCR recognizes a particular pMHC complex, most mutations to the peptide amino acids abrogate recognition (1, 2). It is degenerate because a given TCR can interact productively with several antigenic peptides (3). pMHC complexes where the peptide is derived from the cell's own proteins are also displayed on antigen-presenting cell (APC) surfaces. TCRs are self-tolerant because they bind weakly to these "self"-pMHC complexes, thereby avoiding frequent autoimmune responses.

The diverse, specific/degenerate, and self-tolerant T cell repertoire is designed during T cell development in the thymus (4–8). Immature T cells (thymocytes) interact with a variety of self-pMHC molecules expressed on the surface of thymic epithelial cells as well as hematopoietically derived macrophages and dendritic cells. Thymocytes expressing a TCR that binds with high affinity to any self-pMHC molecule are deleted in the thymus (a process called negative selection). However, a thymocyte's TCR must also bind sufficiently strongly to at least one type of self pMHC complex to receive survival signals and emigrate from the thymus (a process called positive selection).

Signaling events, gene transcription programs, and cell migration during T cell development in the thymus have been studied extensively (4–14). Despite important advances, how interactions with self-pMHC complexes in the thymus shape the peptide-binding properties of selected TCR amino acid sequences such that mature T cells exhibit their special properties is poorly understood.

Recent experiments carried out by Huseby *et al.* (1, 2) provided important clues in this regard. These experiments determined differences in how T cells interact with foreign (antigenic) pMHC depending on whether they developed in conventional mice that display a diverse array of self-pMHC complexes in the thymus or if they develop in mice that were engineered to express only one type of peptide in the thymus. For T cells that develop in conventional mice, T cell recognition of antigenic pMHC was found to be sensitive to most mutations of the antigenic peptide's amino acids. In contrast, T cells selected in mice with only one type of peptide in the thymus were much more peptide-degenerate, with some T cells being tolerant to most mutations of antigenic peptide amino acids.

We reasoned that a detailed understanding of the origin of these experimental results may shed light on the broader question of how the thymus designs diverse self-tolerant TCR sequences that mediate specific/degenerate antigen recognition. Toward this end, we studied a computational model of thymic selection. Our main conclusions can be summarized as follows. Avoiding negative selection against diverse peptides in the thymus imposes strong constraints on the amino acid composition of the peptide contact residues of selected TCRs. Specifically, TCR peptide contact residues are greatly enriched in amino acids that bind weakly to all other amino acids, a result consistent with our analysis of available crystal structures of TCR–pMHC complexes. We show that such TCRs recognize antigenic peptides via multiple modest interactions, each of which contributes a significant fraction of the binding affinity required for recognition. Therefore, mutations to most peptide amino acids abrogate recognition, thus conferring specificity. Positive selection is important for many properties, such as MHC restriction, but not antigen specificity. Our results, and a model for TCR recognition of antigen that emerges from it, illuminate how thymic selection meets the apparently conflicting demands of antigen specificity, cross-reactivity, and self-tolerance.

Model Development

To describe the interactions between TCRs and pMHC complexes, we represent them as strings of sites (Fig. 1*A*). Each site on a TCR can interact with the corresponding site on a pMHC molecule. Such "string models" for studying TCR–pMHC interactions have been used to study various issues, including thymic selection (12, 14, 15), and employed simplified representations of amino acids (e.g., a string of numbers, bits, etc.). From the standpoint of our work, the most pertinent result revealed by these past studies are calculations showing that negative selection reduces TCR cross-reactivity. The mechanistic reasons underlying this numerical result or how it relates to amino acid sequences of selected TCRs were not described. Our goal was to elucidate how the diversity of endogenous peptides bound to host MHC proteins encountered in the thymus determines the amino acid sequences of peptide contact residues on selected TCRs and how such TCRs are antigen specific while also being cross-reactive and self-tolerant.

The specific features of our model were chosen to address these issues and to relate our results closely to known experimental data

Author contributions: A.K., A.K.J., E.S.H., M.K., and A.K.C. designed research; A.K. and A.K.J. performed research; A.K., A.K.J., M.K., and A.K.C. contributed new reagents/analytic tools; A.K., A.K.J., E.S.H., M.K., and A.K.C. analyzed data; and A.K., A.K.J., E.S.H., M.K., and A.K.C. wrote the paper.

The authors declare no conflict of interest.

§To whom correspondence may be addressed. E-mail: arupc@mit.edu or kardar@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0808081105/DCSupplemental.

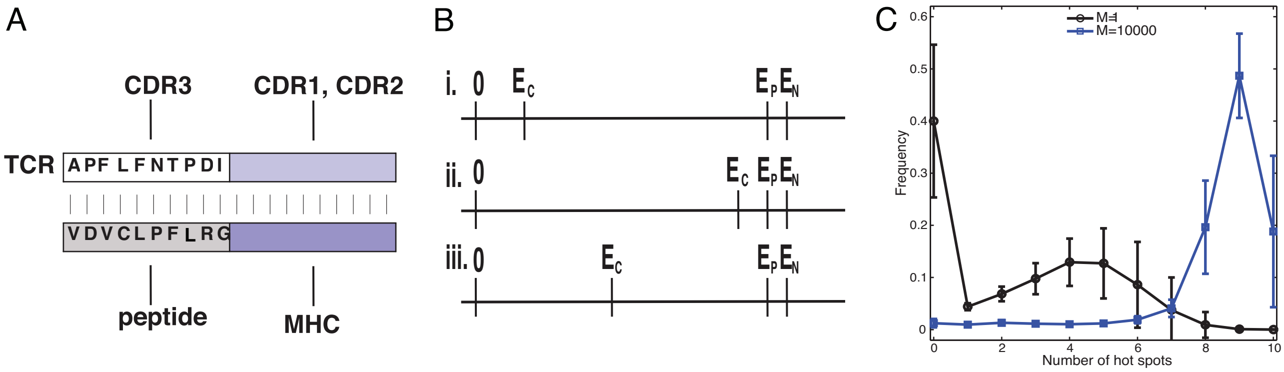

Fig. 1. A simple model recapitulates differences in specificities of T cells selected in mice with one or many types of peptides in the thymus (1, 2). (*A*). Schematic description of the model. The interactions between CDR1 and CDR2 regions of the TCR and conserved residues on the MHC are described by a TCR-dependent energy equal to E_c. Amino acids on the peptide (and variant MHC residues) as well as the corresponding contact residues on the CDR3 loops of the TCR are treated explicitly, and their interactions are described in *Model Development* (Eq. **1**). (*B*) Cartoon representation of the three regimes of values of TCR-MHC interactions (E_c). In these regimes the TCR–MHC interactions are (*i*) weak, (*ii*) strong, and (*iii*) moderate compared with the threshold for negative selection, E_N. (*C*) Selection against many peptides in the thymus results in a larger number of hot spots characterizing antigen recognition. The frequencies of occurrence of one, two, three, etc., hot spots (defined in *Results*) on MHC-bound antigenic peptide moieties recognized by selected TCRs. For TCRs that develop in a thymus with many types of self-peptides (blue curve, M = 10,000 peptides) many sites on the antigenic peptide moiety are hot spots. For TCRs that develop in a thymus with only one type of self-pMHC complex (black curve, M = 1 peptide) there are far fewer hot spots, indicating less specific (more degenerate or cross-reactive) recognition.

such as that of Huseby *et al.* (1, 2). Because Huseby *et al.* used transgenic mice that expressed a single type of MHC, we divided the string of sites on the pMHC molecule into a conserved part representing the MHC and a variable part representing the peptides. One could also view the variable sites more generally as representative of the peptides and the variable residues of the MHC. The CDR1 and CDR2 loops of the TCR mostly contact MHC residues, whereas the CDR3 loop primarily contacts the peptide residues. We partitioned the TCR interaction sites in to two parts: a region representing the CDR1 and CDR2 loops and a part that mimics the CDR3 loop. Because the CDR3 loops are hypervariable, the amino acids of the peptide contact residues of the CDR3 region are explicitly considered, whereas those of the less variable CDR1 and CDR2 regions are not (Fig. 1*A*). For ease of reference, the CDR3 sites are called, "variable." These variable sites represent only those CDR3 amino acids that contact peptide amino acids (or variable MHC residues). Thus, we do not explicitly treat the conformation of the CDR3 loop, which would be necessary if the entire sequence of CDR3 amino acids was considered. Similarly, because peptides bound to MHC are short, peptide conformation is not an important variable. Although we vary the peptide length (data not shown), most results we present are for peptides that are 10 aa long.

We generate panels of TCR and self pMHC molecules on the computer by picking amino acids for the peptides and peptide contact residues on the CDR3 loops of the TCR according to the probabilities with which amino acids appear in the human (or mouse) proteome (16) (Table S1). Antigenic peptides are generated using the frequency of occurrence of amino acids in *Listeria monocytogenes*, a common bacterial pathogen (17). To assess the effects of thymic selection as well as antigen recognition, we evaluate the energy of interaction between TCR-pMHC pairs. The interaction energy between the CDR1 and CDR2 regions of TCRs and the MHC is given a value equal to E_c (and it is varied to describe different TCRs). The total interaction energy equals the sum of E_c and the value obtained by aligning the TCR and pMHC amino acids that are treated explicitly and adding the pairwise interactions between corresponding amino acids. For a given TCR–pMHC pair, the total interaction energy is

$$E = E_c + \sum_{i=1}^{N} J(l_i, j_i), \qquad [1]$$

where E_c is defined above, and J (l_i, j_i) is the interaction energy between the ith amino acids on the variable part of the TCR (l_i) and the peptide (j_i), respectively, and N is the length of the variable regions. The matrix J encodes the values of interaction energies between specific types of amino acids. For most results presented, J was taken to be the parameterized potential due to Miyazawa and Jernigan (MJ matrix) which has been used fruitfully to study proteins (18, 19). However, we also used other potentials (*vide infra*), including ones where the interaction between a pair of juxtaposed amino acids depends on the neighboring residues, to show that our qualitative results and mechanistic insights are independent of this choice [supporting information (SI) Fig. S1]. We express energy values in units of the thermal energy, k_BT, where k_B is Boltzmann's constant, and T is absolute temperature. At 37°C, the thermal energy equals 0.6 Kcal/mole. We emphasize that the purpose of our study is not to compute specific values of energies but to use them to obtain qualitative mechanistic insights.

Recent experiments show that negative selection occurs when the TCR–pMHC interaction affinity exceeds a sharply defined threshold (9). Because affinity correlates directly with the free energy (or energy) gained upon binding, in our model, if the interaction energy between a TCR and self-pMHC is more attractive than (exceeds) a threshold value, E_N, this TCR is negatively selected. It is possible that the off-rate characterizing TCR–pMHC binding, rather than affinity, determines ligand potency, and, indeed, ligands that induce positive and negative selection are separated by a sharp boundary in off-rate as well. Off-rate correlates with the free-energy barrier associated with dissociation of the TCR–pMHC complex. For a related set of reactions, this barrier and the binding energy scale similarly (20) (*Linear Free-Energy Relationships* in *SI Text*) and so use of the interaction energy should correlate with trends in off-rate as well. The ability of a pMHC ligand to stimulate positive selection does not go to zero abruptly (9). In our model, if the interaction energy between a particular TCR–pMHC pair exceeds a threshold value, E_p, the TCR is positively selected. Replacing the soft threshold associated with positive selection with a sharp boundary does not affect qualitative results (Fig. S2) because we find that the characteristics of peptide binding residues on selected T cells are largely shaped by negative selection. The effects of varying E_p and E_N over wide ranges are described in the context of our results.

Results

Selection Against Many Endogenous pMHC Molecules Is Required for Antigen-Specific TCR Sequences. We first tested whether our computational model could recapitulate the experimental observation (1, 2) that T cell recognition of an antigenic peptide is sensitive to mutations at many peptide sites for T cells selected against many endogenous thymic peptides, whereas very few sites on the antigenic peptide are important for recognition for T cells selected in mice that express one type of peptide in the thymus.

For a specific choice of the interaction energy between the CDR1/CDR2 region of the TCR and the MHC (E_c), a panel of one million sequences of TCR peptide contact residues was generated by choosing different amino acids for the variable region according to the frequency with which they appear in the human proteome (results for mouse are in Fig. S3). In the case where there was only one type of self pMHC complex in the thymus, the interaction energy between each TCR and a MHC-bound peptide moiety representative of the human proteome was computed by using the MJ interaction energy matrix and Eq. **1**. Only those TCRs that have interaction energies lying between the positive- and negative-selection thresholds (E_p and E_N) were selected. The selected T cells were then challenged with many antigenic peptides characteristic of *L. monocytogenes* (17). A TCR was considered to recognize an antigenic pMHC if the interaction energy exceeded the negative-selection threshold, E_N. In this way, panels of selected T cells that recognize different antigens were generated. Each amino acid on the antigenic peptides was then mutated to the 19 other possibilities, and recognition by the reactive TCRs was again assessed. If more than half the mutations at a particular amino acid site led to abrogation of recognition for an originally reactive T cell, the site was labeled a "hot spot." This procedure was repeated 1,000 times with a different panel of preselection TCRs, and choices for the peptide in the thymus and antigenic peptides to obtain statistics on the number of hot spots characterizing interactions between a typical antigenic peptide and selected TCRs.

For many types of peptides in the thymus, we generated a panel of 10,000 self-peptides using amino acid frequencies characteristic of the human proteome (16). The results we obtain are qualitatively robust if at least 100 types of pMHC complexes are in the thymus (Fig. S4). Pathologically large numbers of peptides in the thymus result in deletion of all thymocytes. Interaction energies of the panel of TCRs with self-pMHCs were calculated. A TCR was positively selected if it interacted with at least one such pMHC with an energy that exceeded the positive-selection threshold (E_p). To avoid negative selection, a TCR must not interact with any self-pMHC with an energy that exceeds the negative selection threshold (E_N). Hot spots characterizing antigen recognition were determined in the manner described above.

Although the interaction energy between the CDR1 and CDR2 regions of the TCR and MHC (E_c) varies continuously as residues on the CDR1 and CDR2 regions change, TCRs can be grouped into three classes based on the relative values of E_c and the negative-selection threshold, E_N (Fig. 1*B*): (*i*) TCR–MHC interactions are very weak (E_c and E_N are separated by a large value); (*ii*) TCR–MHC interactions are very strong (E_c and E_N are separated by a small value); (*iii*) TCR–MHC interactions are moderate in scale (E_c and E_N are separated by a moderate value). Based on recent experimental data (9), for results reported, the difference between E_p and E_N is taken to be relatively small ($5\,k_BT$). For completeness, we consider cases where this gap is large, and the qualitative results are unchanged (see *Results for Cases Where the Gap Between the Positive and Negative Selection Thresholds Is Large, and TCR-MHC Interactions Are Weak* in *SI Text*).

Very few preselection TCRs with CDR1 and CDR2 loops that interact very weakly with conserved MHC (case *i*) are positively selected if E_p and E_N are relatively close (Table S2). In effect, they are not MHC restricted. These TCRs are irrelevant for our studies of how thymic selection shapes antigen-specific peripheral T cells. TCRs with CDR1 and CDR2 loops that interact very strongly with MHC (E_c close to or greater than E_N, case *ii*) are negatively selected with very high probability (Table S2) and so are not relevant for our studies of understanding the origin of how thymic selection results in antigen specificity in the periphery. Not surprisingly then, our studies focus on TCRs with values of E_c that correspond to moderate interactions between the CDR1/CDR2 loops and MHC (case *iii*). These TCRs are positively selected with high probability and must avoid negative selection to emerge into the periphery (Table S2).

Fig. 1*C* shows the frequency of hot spots resulting from our calculations when the conserved TCR–MHC interactions are moderate in scale ($E_N - E_c$ taken to be $40\,k_BT$ for the results). For TCR selected against many types of peptides, a large fraction of the antigenic peptide's amino acids are hot spots. In contrast, when TCR are selected against one type of peptide in the thymus, very few antigenic peptide amino acids are hot spots. This mirrors previous experimental observations (2). We also find that for moderate TCR–MHC interactions, the ability of T cells to mature when only one type of peptide is present in the thymus is limited by positive selection, whereas T cell survival is limited by negative selection when there are many types of peptides in the thymus (Fig. S5). Because our computational model recapitulates known experimental data (Fig. 2 and refs. 1 and 2), we used the model to obtain insights into the mechanistic origins of antigen specificity.

Frustration During Negative Selection Strongly Constraints Selected TCR Sequences. For a TCR to emerge from the thymus when only one type of pMHC complex is present therein, the binding energy of the TCR for this pMHC must lie in the interval between E_N and E_p. Because the interaction energy between the TCR's peptide contact residues and the peptide's amino acids is a sum over individual contact energies (Eq. **1**), many sequences of peptide contact residues on the TCR can satisfy this criterion. A type of selected sequence that occurs with high probability is one where a small number of TCR residues make strong contacts with the corresponding peptide amino acids, and all of the others make irrelevant (i.e., weak) contacts (Fig. 2*A*). A TCR with such a sequence of peptide contact residues on the TCR would almost certainly be negatively selected when many types of peptides are present in the thymus. This is because it will likely encounter another peptide in the thymus that can differ by only a single amino acid, leading to an additional significant interaction and a total energy that exceeds E_N.

Thus, surviving negative selection presents a frustrating situation because a TCR that avoids negative selection with one peptide in the thymus could be negatively selected by another peptide. Positive selection does not present this problem because, once a TCR receives survival signals by binding a single peptide more strongly than E_p, interactions with other peptides are only relevant for negative selection. The frustration associated with subsequently avoiding negative selection by all these diverse pMHCs is the dominant constraint determining peripheral TCR sequences.

To explore how this frustration influences the character of the peptide contact residues of selected TCRs, we developed an analytical approximation (*Methods*) that suggested that the peptide contact residues on selected TCRs are greatly enriched in amino acids that bind weakly to other amino acids.

Negative Selection Against Many Peptides Results in TCR Sequences with Peptide Contact Residues Enriched in Weakly Interacting Amino Acids. To test this suggestion, we first examined the amino acid compositions of the peptide contact residues of the selected TCRs obtained from our computer simulations. When there are many types of peptides in the thymus, peptide contact residues of selected TCRs are enriched in amino acids that interact weakly with other amino acids, whereas strongly interacting amino acids are attenu-

CHEMISTRY

IMMUNOLOGY

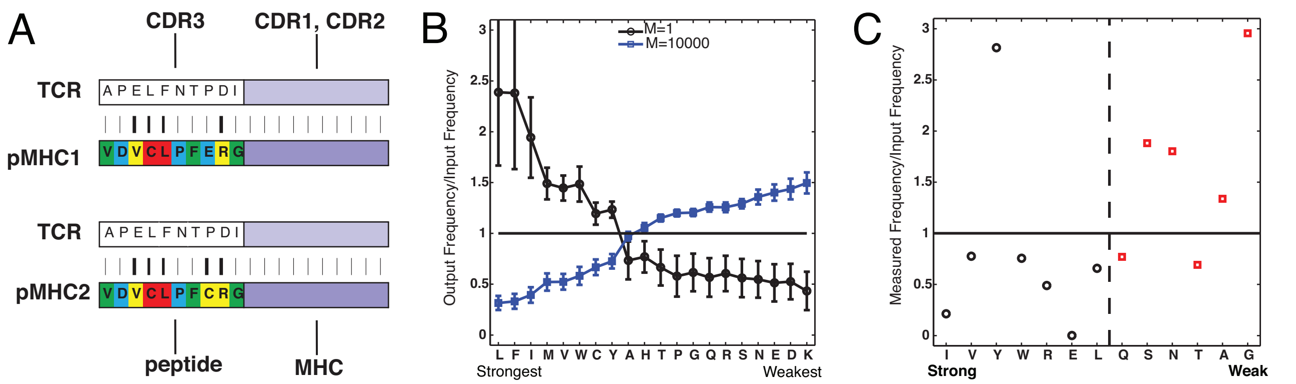

Fig. 2. Consequences of avoiding negative selection on the composition of peptide contact residues of selected TCRs. (*A*) Schematic description of frustration due to negative selection. The thickness of the bars (or color of peptide amino acids: strong, red; moderate, yellow; weak, blue; very weak, green) is proportional to the interaction energy between TCR and pMHC residues. When developing in a thymus with only one type of endogenous peptide, a TCR that results in a few strong interactions and several weak or moderate interactions with this peptide can survive selection. This is because the total interaction energy falls between the positive- and negative-selection thresholds. The sequence of TCR peptide contact residues shown, that survives selection against one type of peptide in the thymus, would likely be negatively selected when there are many types of peptides in the thymus. For example, a peptide that differs by one amino acid from the first one (shown as a change from E to C) may lead to an additional moderate interaction energy that is sufficient to increase the total interaction energy past the negative selection threshold. (*B* and *C*) Selection against many types of peptides in the thymus results in selected TCRs with peptide contact residues with an enhanced frequency of amino acids that interact weakly with all other amino acids. The ordinate is the ratio of the frequencies of occurrence of an amino acid in the peptide contact residues of selected TCRs to preselection TCRs. (*B*) For the computational results, the abscissa is a list of amino acids ordered according to the maximum energy (as per the MJ interaction potential) with which it interacts with all other amino acids. The qualitative results are robust to changes in potential (Fig. S1 *B*, *D*, and *F*). (*C*) The ordinate was obtained by analyzing the 18 available crystal structures of TCR-pMHC (I) complexes as described in the text. Amino acids were classified as strongly interacting (IVYWREL) or weakly interacting (QSNTAG) following ref. 23.

ated (Fig. 2*B*). The opposite is true when T cell selection is mediated by a single peptide species in the thymus, with preferential selection of TCR that contain strongly interacting amino acids. In Fig. 2*B*, amino acids were ordered according to the maximum value of the strength with which each amino acid interacts with all others. The nature of the MJ interaction potential is such that this order also reflects the ordering obtained by considering the average value of the interaction energy of an amino acid with all others. The qualitative results shown in Fig. 2*B* are robust to changes in the interaction potential (Fig. S1). Using different potentials only changes the identities of the amino acids that interact weakly or strongly or the criterion used to define interaction strength. For example, if a potential is such that the order of amino acids obtained by using the average interaction energies with other amino acids is quite different from that obtained by considering the largest interaction energies, the qualitative results in Fig. 2*B* are obtained if we use the latter quantities to order amino acids.

Do experimental data support our conclusion that frustration due to negative selection skews the mature T cell repertoire to TCRs composed of peptide contact residues enriched in amino acids that bind weakly to other amino acids? We analyzed the 18 available crystal structures of TCR bound to class I pMHC complexes to obtain the frequency with which different amino acids are represented at residues of the TCR that contact the peptide (21). All TCR moieties that contact peptide amino acids were considered, and two methods were used to identify these contact residues. One was to define a contact as a position where a water molecule does not fit in the gap between a TCR residue and a peptide amino acid. In the other method, residues in contact have their C_α atoms within 6.5 Å of each other. The qualitative results are the same for both methods (Fig. S6), and in Fig. 2*C*, we show results using the second criterion.

Whereas the qualitative computational results (Fig. 2*B*) are independent of interaction potential, to compare the experimental data with this prediction, we need to know whether a particular amino acid is "weak" or "strong" in reality. We have used two different prescriptions to order the amino acids according to the strength of their interactions with other amino acids. One is to use the MJ matrix, but the order thus obtained has been criticized because it overemphasizes hydrophobic interactions and considers interactions between charged amino acids to be weak (22). Data obtained by examining the stability of thermophiles are proposed to be better suited for analyzing the strength of interactions between amino acids (23), and posit that the strongly interacting amino acids are IVYWREL, and the weakly interacting ones are QNSTAG (23).

Fig. 2*C* shows results where amino acids are divided into two classes (weak and strong) according to this prescription. The data obtained from crystal structures are in qualitative agreement with the theoretical prediction in that weakly interacting amino acids are enriched on peptide contact residues of the TCR, and strongly interacting amino acids are attenuated. Using the MJ matrix leads to similar results (Fig. S6), except that charged amino acids (R, E, K), which are "weak" according to the MJ matrix, are additional outliers. Tyrosine is considered to be a strongly interacting amino acid by either approach, but is well represented in the TCR–peptide contact residues. This may be because a germ-line-encoded tyrosine interacts with a conserved MHC residue that is close to the peptide amino acids (24, 25), and so it may interact ubiquitously with peptide amino acids.

Our results suggest that negative selection against many types of thymic peptides results in mature TCRs with peptide contact residues that interact weakly with other amino acids. How does this influence their antigen specificity?

Antigen Specificity Is the Result of TCR Residues Binding Peptides via Multiple Moderate Interactions. In our model, the interaction energy between an antigenic peptide and residues of a TCR that recognizes it is the sum of 10 numbers, with each number being the interaction energy between an amino acid on the peptide and the corresponding TCR contact site (Fig. 1*A*). We computed the values of these site–site interaction energies using all our TCR–antigenic peptide pairs. In Fig. 3, we compare the frequency with which each value of these interaction energies occurs for three cases: preselection TCRs, TCRs that developed in a thymus with many types of pMHC, and TCRs that developed in a thymus with one type of pMHC.

Our results indicate that, compared with the preselection TCRs, antigen recognition by TCRs selected against many types of pMHC

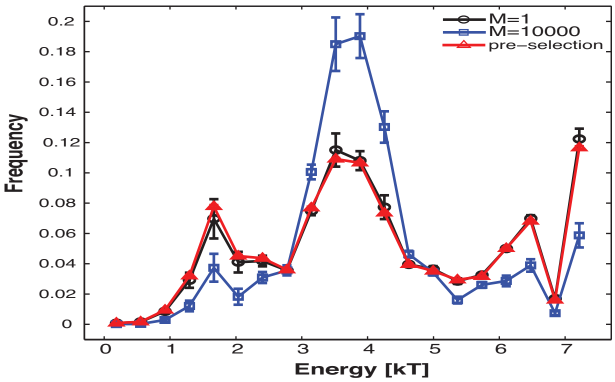

Fig. 3. Distribution of amino acid–amino acid contact energies (in units of kT, described in text) characterizing interactions between selected reactive TCRs and antigenic peptides suggest the basis for specificity. The distribution of interaction energies between individual amino acids on peptide contact residues on the TCR and antigenic peptides are shown. The distribution for TCRs that develop in a thymus with many endogenous peptides (blue curve) is very different from that for preselection TCRs (red curve). The distribution of contact energies is not significantly altered for TCRs that develop in a thymus with only one type of peptide (black curve) compared with preselection TCRs.

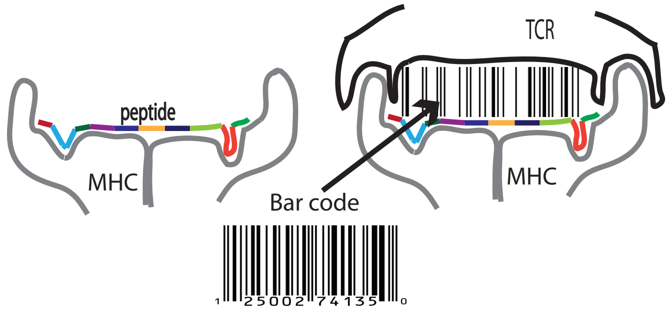

Fig. 4. A bar code scanning model for specificity of TCR recognition of antigenic peptides. The thickness of the lines in the cartoon is proportional to the strength of TCR–peptide interactions.

complexes is mediated by fewer strong and weak amino acid–amino acid interactions, resulting in a pronounced enhancement of moderate interactions. This result is consistent with experimental observations of Savage and Davis (26). This focusing on moderate interactions is because negative selection constrains mature TCR peptide contact residues to be composed of weakly interacting amino acids (Fig. 2). The weakly interacting amino acids on the TCR bind to strongly interacting amino acids on antigenic peptides (Fig. S7) resulting in multiple moderate scale interactions that add up to a total binding energy that is large enough for recognition. Because antigen recognition is mediated by multiple interactions of moderate value, each contact makes a significant contribution to the total interaction energy necessary for recognition. Therefore, disrupting most interactions by mutating peptide amino acids results in abrogation of recognition. This is the origin of antigen specificity. This prediction is consistent with measurements reported for the B3K 506 TCR, which was selected against many types of pMHC complexes in the thymus and recognizes the 3K–IA^b pMHC (1). Many mutations of the antigenic peptide correspond to moderate $\Delta\Delta G$ values, and each contributes significantly to recognition.

When there is one type of pMHC complex in the thymus, the peptide-binding residues of selected TCRs are not subject to the important constraints of avoiding negative selection against many types of peptides, and moderate amino acid–amino acid interactions do not dominate (Fig. 3). Strongly interacting amino acids are represented more than in the preselection repertoire (Fig. 2*B*), resulting in a small enhancement of strong interactions between amino acids (Fig. 3). These strong interactions make dominant contributions to the total interaction energy required for antigen recognition (see also Fig. 2*A*). Thus, mutating the antigenic peptide amino acids that contact strongly interacting amino acid residues on the TCR should abrogate recognition, but mutations at most other sites should have little impact. This is reflected in the experimental data reported by Huseby *et al.* (1). For one example, consider the YAe62.8 TCR, which is selected against a single type of peptide in the thymus and recognizes variants of the 3K-IA^b antigenic peptide. Most mutations to the antigenic peptide result in small changes in $\Delta\Delta G$, but one mutation results in a large change. This one major peptide contact dominates the interaction energy with the others being irrelevant, and this is the origin of enhanced cross-reactivity.

TCRs that survive negative selection against many types of peptides are quite diverse because many sequences are consistent with the constraint that peptide contact residues are predominantly composed of amino acids that interact weakly with all others.

Discussion

Although important clues were provided by the experimental data reported by Huseby *et al.* (1, 2), a mechanistic understanding of how thymic selection designs TCR sequences that are simultaneously antigen specific, cross-reactive, diverse, and self-tolerant remained unclear. Our computational studies shed light on these issues.

If a TCR receives survival signals from a self-pMHC complex, it is positively selected. Interactions with the other peptides expressed in the thymus are then only relevant for negative selection. Positive selection ensures MHC restriction, enables weak binding of TCRs to self pMHC, and influences the fraction of T cells that survive thymic selection. Thus, it mediates important properties. However, antigen specificity appears to be determined by the requirement that positively selected T cells must survive negative selection.

TCR sequences must simultaneously avoid being negatively selected by many endogenous MHC-bound peptides, and this imposes strong constraints on the nature of the peptide contact residues of selected TCRs. We find that this is why, in mature T cells, these residues are enriched in amino acids that interact weakly with other amino acids (referred to as "weak" amino acids). For a selected TCR to recognize an antigenic peptide in the periphery, it must bind to it with an affinity that exceeds a threshold. This can occur only if the peptide is composed of amino acids that are among the strongest binders of the corresponding weak amino acids of the TCR's peptide contact residues (Fig. S7), resulting in a number of moderate scale interactions that sum up to exceed the threshold affinity required for recognition. Because each moderate interaction contributes a substantial fraction of the overall affinity, disrupting most of them (via mutations) abrogates recognition. Thus, antigen specificity emerges because TCR residues that contact the peptide are enriched in amino acids that interact weakly with other amino acids. It is worth remarking that weakly binding amino acids are not always the mediators of recognition; TCR selected against one type of peptide do not exhibit this behavior (1, 2), and the EGFR receptors-binding sites are cysteine rich (27).

Because the amino acids treated explicitly in our model include variable MHC residues, our results are also consistent with data showing that TCR selected against many peptides are also MHC specific. We note in passing that we have also studied the alloreactivity of selected TCRs (data not shown). Our findings suggest that the relative importance of the peptide (compared with the MHC) in mediating alloreactive responses depends on how different the allo- and endogenous MHCs are vis-à-vis their interaction energies with the CDR1 and CDR2 loops of a particular TCR (E_c in our model); the greater this difference, the less important the peptide.

Our results suggest a model for specificity of TCR–antigenic pMHC recognition that is different from Fisher's "lock and key" metaphor for the specificity with which an enzyme binds its substrate. It also appears to be different from that applicable to specificity of antibody–antigen interactions where shape complementarity and multiple weak interactions are inextricably coupled (28). Shape complementarity is important for TCR recognition of antigen in two ways (Fig. 4). First, it plays a key role in peptide binding to the MHC groove, and hence influences antigen presentation. Secondly, shape complementarity is possibly important in mediating interactions of the TCR with MHC moieties, which results in orienting the TCR in a way that juxtaposes its peptide contact residues with the peptide. Indeed, it has been suggested that if the peptide has a conformation that is not relatively flat, it disrupts TCR–MHC interactions, thereby preventing positive selection (29). But, these TCR–MHC interactions required for positive selection and binding of peripheral TCRs to MHC in the proper orientation do not confer peptide specificity.

Once properly oriented, a TCR scans the relatively flat conformation of the short peptide, and recognizes the epitope if a number of peptide amino acids correspond to strong binders for the weak peptide contact residues of this TCR. For reasons described above, recognition is specific because each resulting interaction is moderate. Shape complementarity seems to be decoupled from the origin of specificity. TCR recognition of antigen is analogous to scanning a flat "bar code" for the appropriate number of moderately thick lines. In this metaphor, the moderately thick lines represent moderate interactions mediated by peptide amino acids that are strong binders for the weak amino acids that comprise the TCR's peptide contact residues. This bar-code model also makes vivid why specificity and cross-reactivity can coexist. For example, consider a situation where any three of four contacts with the peptide amino acids need to be of moderate scale for recognition; i.e., three of the four lines need to be moderately thick. If a particular peptide satisfies this criterion (say, lines 1, 3, and 4 are moderately thick), mutations at any one of these sites will abrogate recognition (specificity). But another peptide that leads to lines 1, 2, and 3 being moderately thick will also be recognized by this TCR (cross-reactivity). One might say that TCRs scan a bar code and recognize statistical patterns—ones that have a sufficient number of moderately thick lines.

We hope that the results we have reported will motivate experimental and computational studies that will ultimately elucidate how one of nature's intriguing designers (the thymus) works and how its aberrant regulation can contribute to autoimmune disease. An important question unresolved by our studies is how variability in expression levels of different types of endogenous peptides in the thymus influences the T cell repertoire.

Methods

How Negative Selection Against Many Peptides Constrains Selected TCR Sequences. The probability (P) that a TCR characterized by a sequence of peptide contact residues, $\vec{l} = \{l_1, l_2, l_3, \ldots\}$, is not negatively selected can be written as:

$$P(\vec{l}) = \prod_{j=1}^{M} [1 - \theta(E(\vec{l}, \vec{j}) - E_{\mathrm{N}})] p(\vec{j}), \qquad [2]$$

where M is the number of peptides in the thymus, $E(\vec{l}, \vec{j})$ is the absolute value of the interaction energy between the TCR and a peptide composed of a sequence of amino acids, $\vec{j} = \{j_1, j_2, j_3, \ldots\}$, which occurs with probability $p(\vec{j})$. The step function, θ, represents the negative selection threshold. Approximations (described in *Probability that a TCR Will Escape Negative Selection* in *SI Text*) allowed us to rewrite Eq. **2** as:

$$P(\vec{l}) \propto \exp\left[-M \prod_{i=1}^{10} \sum_{k=1}^{20} h_{ik}\right] = \exp[-M\{(h_{11} + h_{12} + h_{13} + \ldots) \cdot (h_{21} + h_{22} + \ldots)\ldots(h_{10,1} + h_{10,2} + \ldots + h_{10,20})\}], \qquad [3]$$

$$h_{ik} = \exp\left[b\left(J(l_i, k) - \frac{E_N}{10}\right)\right] p(k),$$

where b is a positive constant.

Eq. **3** suggests that if any of the quantities, h_{ik}, becomes large, the probability of survival of that TCR becomes small, and that h_{ik} becomes large if the TCR's peptide contact residues interact strongly with its corresponding peptide amino acid. Thus, TCR with a high probability of survival must be composed of peptide contact residues that bind weakly to other amino acids.

ACKNOWLEDGMENTS. We thank Profs. Herman Eisen and Eugene Shakhnovich for fruitful discussions and comments. This work was supported by National Institutes of Health (NIH) Grant 1-PO1-AI071195-01 and a NIH Director's Pioneer award (to A.K.C.). E.S.H. was supported in part by Beckman Young Investigator and Searle Scholar awards.

1. Huseby ES, *et al.* (2006) Interface-disrupting amino acids establish specificity between T cell receptors and complexes of major histocompatibility complex and peptide. *Nat Immunol* 7:1191–1199.
2. Huseby ES, *et al.* (2005) How the T cell repertoire becomes peptide and MHC specific. *Cell* 122:247–260.
3. Unanue ER (1984) Antigen-presenting function of the macrophage. *Annu Rev Immunol* 2:395–428.
4. von Boehmer H, *et al.* (2003) Thymic selection revisited: How essential is it? *Immunol Rev* 191:62–78.
5. Werlen G, Hausmann B, Naeher D, Palmer E (2003) Signaling life and death in the thymus: Timing is everything. *Science* 299:1859–1863.
6. Siggs OM, Makaroff LE, Liston A (2006) The why and how of thymocyte negative selection. *Curr Opin Immunol* 18:175–183.
7. Hogquist KA, Baldwin TA, Jameson SC (2005) Central tolerance: Learning self-control in the thymus. *Nat Rev Immunol* 5:772–782.
8. Jameson SC, Hogquist KA, Bevan MJ (1995) Positive selection of thymocytes. *Annu Rev Immunol* 13:93–126.
9. Daniels MA, *et al.* (2006) Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling. *Nature* 444:724–729.
10. Bousso P, Bhakta NR, Lewis RS, Robey E (2002) Dynamics of thymocyte–stromal cell interactions visualized by two-photon microscopy. *Science* 296:1876–1880.
11. Borghans JAM, Noest AJ, De Boer RJ (2003) Thymic selection does not limit the individual MHC diversity. *Eur J Immunol* 33:3353–3358.
12. Detours V, Mehr R, Perelson AS (1999) A quantitative theory of affinity-driven T cell repertoire selection. *J Theor Biol* 200:389–403.
13. Scherer A, Noest A, de Boer RJ (2004) Activation-threshold tuning in an affinity model for the T-cell repertoire. *Proc R Soc London Ser B* 271:609–616.
14. Detours V, Perelson AS (1999) Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection. *Proc Natl Acad Sci USA* 96:5153–5158.
15. Chao DL, Davenport MP, Forrest S, Perelson AS (2005) The effects of thymic selection on the range of T cell cross-reactivity. *Eur J Immunol* 35:3452–3459.
16. Flicek P, *et al.* (2008) Ensembl 2008. *Nucleic Acids Res* 36:D707–D714.
17. Moszer I, Glaser P, Danchin A (1995) Subtilist—A relational database for the *Bacillus subtilis* Genome. *Microbiol UK* 141:261–268.
18. Li H, Tang C, Wingreen NS (1997) Nature of driving force for protein folding: A result from analyzing the statistical potential. *Phys Rev Lett* 79:765–768.
19. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644.
20. Edwards JO (1954) Correlation of relative rates and equilibria with a double basicity scale. *J Am Chem Soc* 76:1540–1547.
21. Kaas Q, Ruiz M, Lefranc MP (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res* 32:D208–D210.
22. Sorenson JM, Head-Gordon T (1999) The importance of hydration for the kinetics and thermodynamics of protein folding: Simplified lattice models. *Biophys J* 76:A109–A109.
23. Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 3:62–72.
24. Feng D, *et al.* (2007) Structural evidence for a germline-encoded T cell receptor—Major histocompatibility complex interaction 'codon'. *Nat Immunol* 8:975–983.
25. Dai S, *et al.* (2008) Crossreactive T cells spotlight the germline rules for alpha beta T cell–receptor interactions with MHC molecules. *Immunity* 28:324–334.
26. Savage PA, Davis MM (2001) A kinetic window constricts the T cell receptor repertoire in the thymus. *Immunity* 14:243–252.
27. Abe Y, *et al.* (1998) Disulfide bond structure of human epidermal growth factor receptor. *J Biol Chem* 273:11150–11157.
28. Perelson AS, Oster GF (1979) Theoretical studies of clonal selection—Minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* 81:645–670.
29. Schumacher TNM, Ploegh HL (1994) Are MHC-bound peptides a nuisance for positive selection. *Immunity* 1:721–723.

Supporting Information

Košmrlj *et al.* 10.1073/pnas.0808081105

SI Text

Linear Free-Energy Relationships. The relationship between the change in free energy at equilibrium (related to affinity) and the free-energy barrier for the reaction to occur (related to off-rate) for a set of related reactions has been studied extensively. Reactions are considered related if the change from one reaction to another is a change in some moieties that does not change the class of reactions (e.g., reactions of amines ($-RNH_2$) with an acid and varying R groups). For related reactions, the free-energy surfaces usually do not intersect. As such, if the equilibrium free-energy change is larger for one reaction compared with another, then so is the free-energy barrier. Thus, the reaction with the higher affinity will also have a lower off-rate. These relationships are called linear free-energy relationships (1, 2).

Results for Cases Where the Gap Between the Positive and Negative Selection Thresholds Is Large, and TCR-MHC Interactions Are Weak. If TCR–MHC interactions are weak and E_p and E_N were separated by a large gap, regardless of the number of peptides in the thymus, almost all preselection TCRs characterized by weak TCR–MHC interactions (E_c) would be positively selected, and almost none would be negatively selected (Table S2). This contradicts the fact that very few T cells are positively selected (3–8). Our calculations also show that, for this situation, TCRs selected against 1 or 10,000 types of pMHC in the thymus display many hot spots vis-à-vis recognition of antigenic peptides (Fig. S8), a result contradicting observations (9, 10). The origin of this result is that, in this case, positive selection determines TCR sequences. Positive selection requires only that a TCR interact with any one pMHC molecule with energy greater than E_p, making selection against one or many pMHC complexes have similar consequences. For these reasons, we do not consider this situation.

Probability that a TCR Will Escape Negative Selection. The probability (P) that a TCR characterized by a sequence of peptide contact residues composed of a set of amino acids, $\{l_1, l_2, l_3, \ldots\}$, denoted by $\vec{l}$, is not negatively selected can be written as:

$$P(\vec{l}) = \prod_{j=1}^{M} [1 - \theta(E(\vec{l},\vec{j}) - E_N)]p(\vec{j}), \qquad [1]$$

where M is the number of peptides in the thymus, $E(\vec{l},\vec{j})$ is the interaction energy between the TCR and a peptide composed of a sequence of amino acids, denoted by $\vec{j}$. The absolute values of this interaction energy and E_N are used in Eq. **1**. The step function, θ, is used to represent the sharply defined negative selection threshold, and $p(\vec{j})$ is the probability of finding a peptide characterized by the amino acid sequence $\vec{j}$ in the thymus. Because the probability P that a particular TCR escapes the negative-selection process is the product of the probabilities to escape M encountered peptides, we can alternatively write:

$$P(\vec{l}) = \exp\left\{\sum_{j=1}^{M}[\ln p(\vec{j}) + \ln(1 - \theta(E(\vec{l},\vec{j}) - E_N))]\right\}$$
$$\approx \exp\{M\langle \ln p(\vec{j})\rangle + M\langle \ln[1 - \theta(E(\vec{l},\vec{j}) - E_N)]\rangle\}. \qquad [2]$$

The approximation rests on the reasonable assumption that the sum of logarithms of the individual escape probabilities is a self-averaging quantity and should be valid in the limit of large M. The first factor in the exponent is related to the entropy of the probability distribution of finding peptides in the thymus and is independent of TCR sequence $\vec{l}$; the second factor restricts the choice of sequence of the peptides that escape negative selection, i.e.:

$$P(\vec{l}) \propto \exp\{M\langle \ln[1 - \theta(E(\vec{l},\vec{j}) - E_N)]\rangle\}. \qquad [3]$$

It is hard to evaluate averages by using step function, but we can approximate the step function with the following smooth function

$$1 - \theta(\Delta E) \approx \exp[-e^{b\Delta E}], \qquad [4]$$

where b is a positive constant. Note that when ΔE is negative, $e^{b\Delta E}$ is ≈ 0, whose exponential is roughly unity, whereas if ΔE is positive, $e^{b\Delta E}$ is a large positive number, whose exponential is ≈ 0. How sharply the change from 0 to 1 occurs as ΔE changes from negative to positive can be controlled by changing the constant b, and a sharp cutoff is obtained for $b \rightarrow \infty$.

With this approximation, and noting that ΔE is the sum of N contributions, where N is peptide length, we find:

$$\langle \ln[1 - \theta(E(\vec{l},\vec{j}) - E_N)]\rangle \approx -\langle e^{\sum_{i=1}^{N} b[J(l_i,j_i) - E_N/N]}\rangle$$
$$= -\prod_{i=1}^{N}\left\langle \exp\left[b\left(J(l_i, j_i) - \frac{E_N}{N}\right)\right]\right\rangle = -\prod_{i=1}^{N}\sum_{j=1}^{20} h_{ij}, \qquad [5]$$

where

$$h_{ij} = p_j \exp\left[b\left(J(l_i, j) - \frac{E_N}{N}\right)\right], \qquad [6]$$

and p_j is the frequency with which amino acid j occurs. We were able to take the averaging operation inside the product, by assuming that the sites are independent. The expression for the probability that a particular TCR sequence escapes negative selection then takes the form

$$P(\vec{l}) \propto \exp\left\{-M\prod_{i=1}^{N}\sum_{j=1}^{20} h_{ij}\right\}. \qquad [7]$$

1. Swain CG, Scott CB (1953) Quantitative Correlation of Relative Rates. Comparison of Hydroxide Ion with Other Nucleophilic Reagents toward Alkyl Halides, Esters, Epoxides and Acyl Halides. *J Am Chem Soc* 75:141–147.
2. Edwards JO (1954) Correlation of Relative Rates and Equilibria with a Double Basicity Scale. *J Am Chem Soc* 76:1540–1547.
3. Detours V, Perelson AS (1999) Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection. *Proc Natl Acad Sci USA* 96:5153–5158.
4. vanMeerwijk JPM, *et al.* (1997) Quantitative impact of thymic clonal deletion on the T cell repertoire. *J Exp Med* 185:377–383.
5. Egerton M, Scollay R, Shortman K (1990) Kinetics of mature T cell development in the thymus. *Proc Natl Acad Sci USA* 87:2579–2582.
6. Scollay RG, Butcher EC, Weissman IL (1980) Thymus-cell migration quantitative aspects of cellular traffic from the thymus to the periphery in mice. *Eur J Immunol* 10:210–218.
7. Shortman K, Vremec D, Egerton M (1991) The kinetics of T-cell antigen receptor expression by subgroups of Cd4+8+ thymocytes—Delineation of Cd4+8+32+ thymocytes as post-selection intermediates leading to mature T-cells. *J Exp Med* 173:323–332.

8. Merkenschlager M, *et al.* (1997) How many thymocytes audition for selection? *J Exp Med* 186:1149–1158.
9. Huseby ES, *et al.* (2006) Interface-disrupting amino acids establish specificity between T cell receptors and complexes of major histocompatibility complex and peptide. *Nat Immunol* 7:1191–9.
10. Huseby ES, *et al.* (2005) How the T cell repertoire becomes peptide and MHC specific. *Cell* 122:247–260.
11. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644.
12. Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 3:62–72.

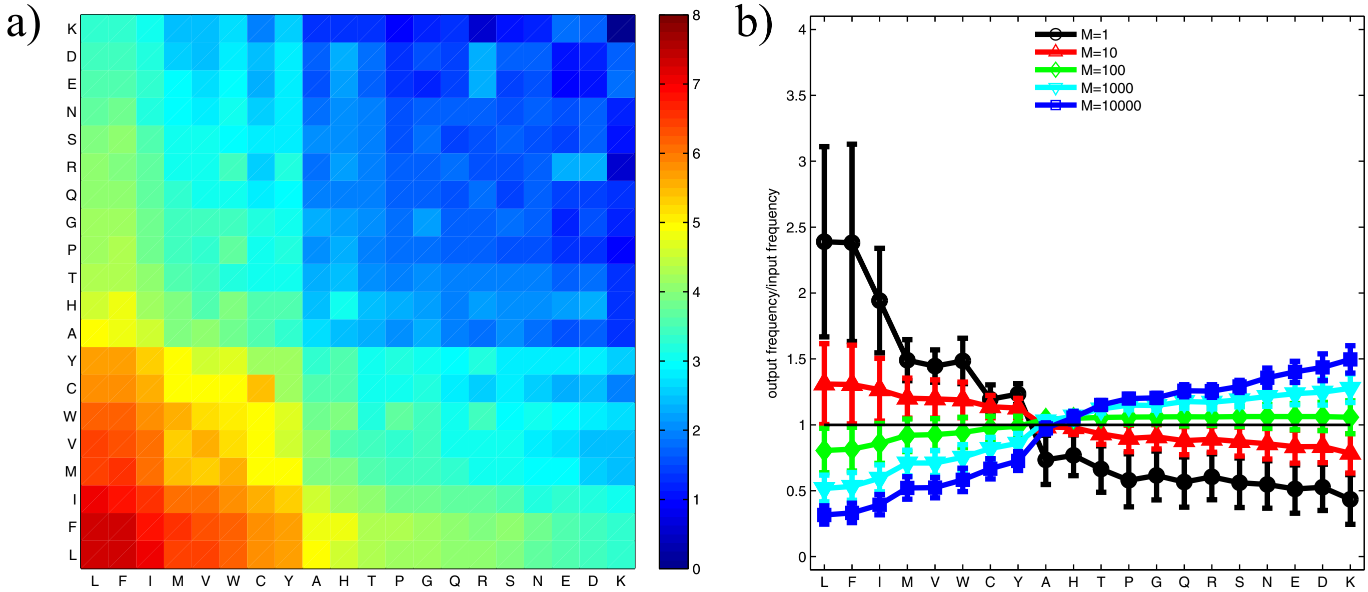

Fig. S1. Results for random statistical potential between amino acids. (*a–f*) In all calculations reported in the main text, the MJ matrix (11) was used to determine the interaction energy between peptide contact residues of the TCR and peptide amino acids. Here, we explore what happens if we use semi-random symmetric matrices with the same values of mean and variance as the MJ matrix and controlled differences between the largest values in each row (column). As shown in *a* there is a clear gradation of interaction energies (color scale in k_BT units) in the MJ matrix, from the strong (lower left, red color) to weak (upper right, dark blue color), enabling a clear ordering of the amino acids. For the MJ matrix, the order of amino acids obtained by using the average interaction energy with other amino acids or that obtained by using the strongest interaction with other amino acids is quite similar. Therefore, the computational results are unchanged from that shown in Fig. 2*B* if results using the MJ matrix are graphed with the amino acids ordered according to their average interaction with other amino acids (*b*). For random matrices (e.g., *c* and *e*), the average value of an amino acid's interaction energies with other amino acids and the strongest interaction of this amino acid with all others are not correlated. Our analytical calculation (*Probability That a TCR Will Escape Negative Selection* in *SI Text*) shows that ordering amino acids according to their strongest interaction with other amino acids is appropriate when there are many types of peptides in the thymus. Therefore, we use this criterion in *b* (MJ matrix) and *d* and *f* (different random potentials). Results for the random potentials are qualitatively similar to that for the MJ matrix when this criterion is used. We varied the random potential by varying the difference between the maximum interaction energies characterizing the strongest and weakest amino acids (L and K). If this difference is the same as that for the MJ matrix (4 k_BT), the results look like those shown in *b*. When we make this difference smaller (e.g., 2 k_BT as in *c*), there is no clear trend of amino acid composition when TCR develop in a thymus with a small number of types of peptides (*d*). Importantly, for many types of peptides in the thymus, the qualitative trends obtained for the MJ matrix are recovered. This is also true for even smaller differences between the strongest and weakest amino acids (e.g., 0.6 k_BT) in *e* and *f*. For random potentials, there are more "bumps" in the distribution, but these disappear if an even larger number of endogenous peptides are displayed in the thymus. For nonsymmetric interaction matrices, statistical properties of selected TCRs are also similar to that we have reported, and the order of amino acids is determined by the strongest interactions with other amino acids (data not shown). (*g–i*) More complex interactions between peptide contact residues of TCRs and peptide amino acids are used to check the robustness of our results. The qualitative features of the post-thymic selection TCR repertoire are robust to more complex interactions between peptide contact residues of the TCR and peptide amino acids. We show: the number of hot spots (*g*), the amino acid composition of selected TCRs (*h*), and the distribution of contact energies (*i*) between selected TCRs and antigenic pMHC for the following more complex potential, which includes interactions with "nearest neighbor" amino acids

$$E = E_c + \sum_{i=1}^{N}\left[J(l_i, j_i) + \frac{1}{2}\{J(l_i, j_{i+1}) + J(l_i, j_{i-1})\}\right].$$

$J(l_i, j_i)$ is the interaction energy between the *i*th amino acids on the variable part of the TCR (l_i) and the peptide (j_i), respectively, and N is the length of the variable regions. In fact, the statistical properties of the TCR repertoire (*g–i*) remain unchanged for any bilinear combination

$$E = E_c + \sum_{\alpha=1}^{N}\sum_{\beta=1}^{N} C_{\alpha\beta} J(l_\alpha, j_\beta).$$

$(E_N - E_c = 75\ k_BT,\ E_N - E_p = 5\ k_BT)$.

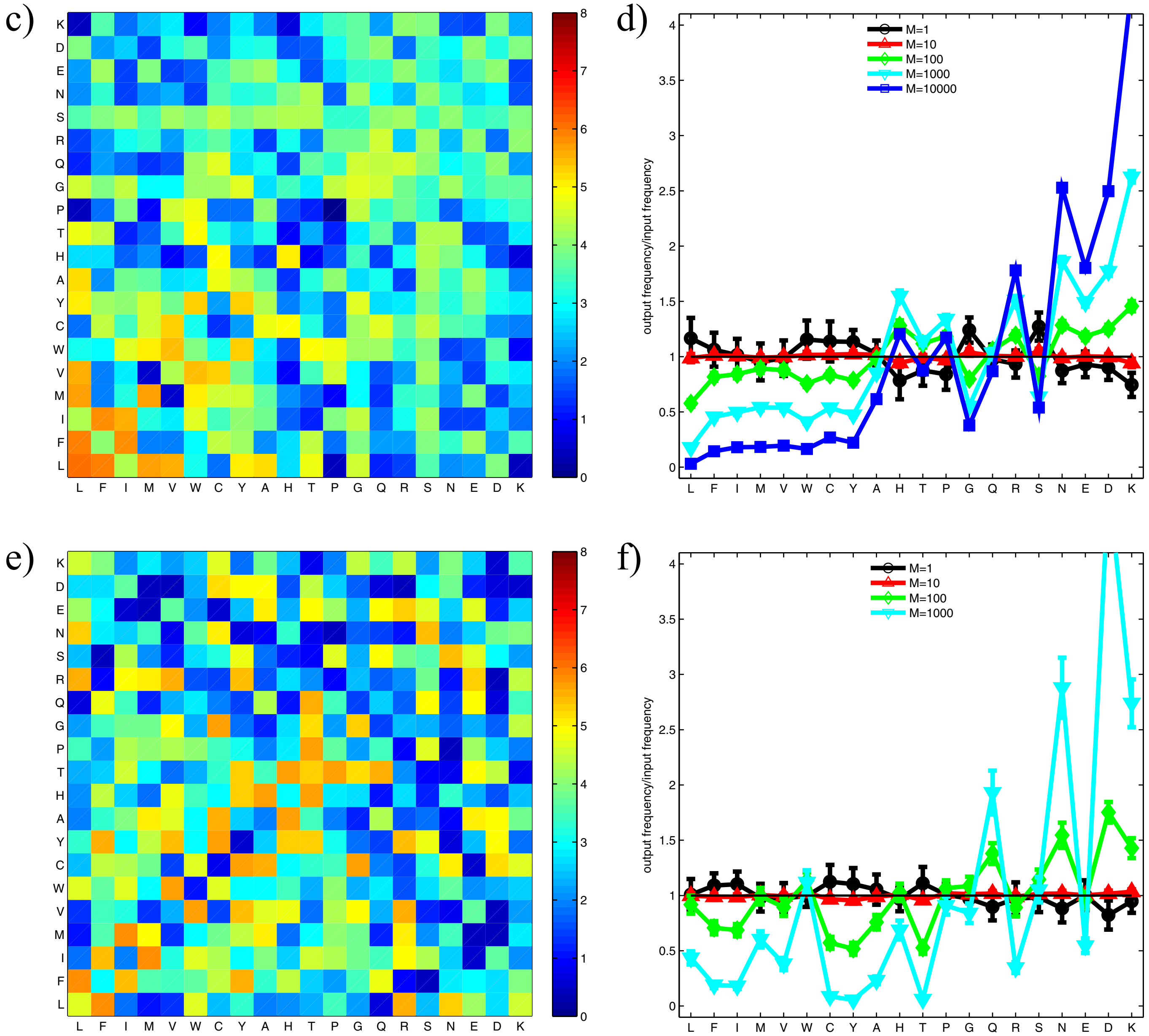

Fig. S1. *Continued.*

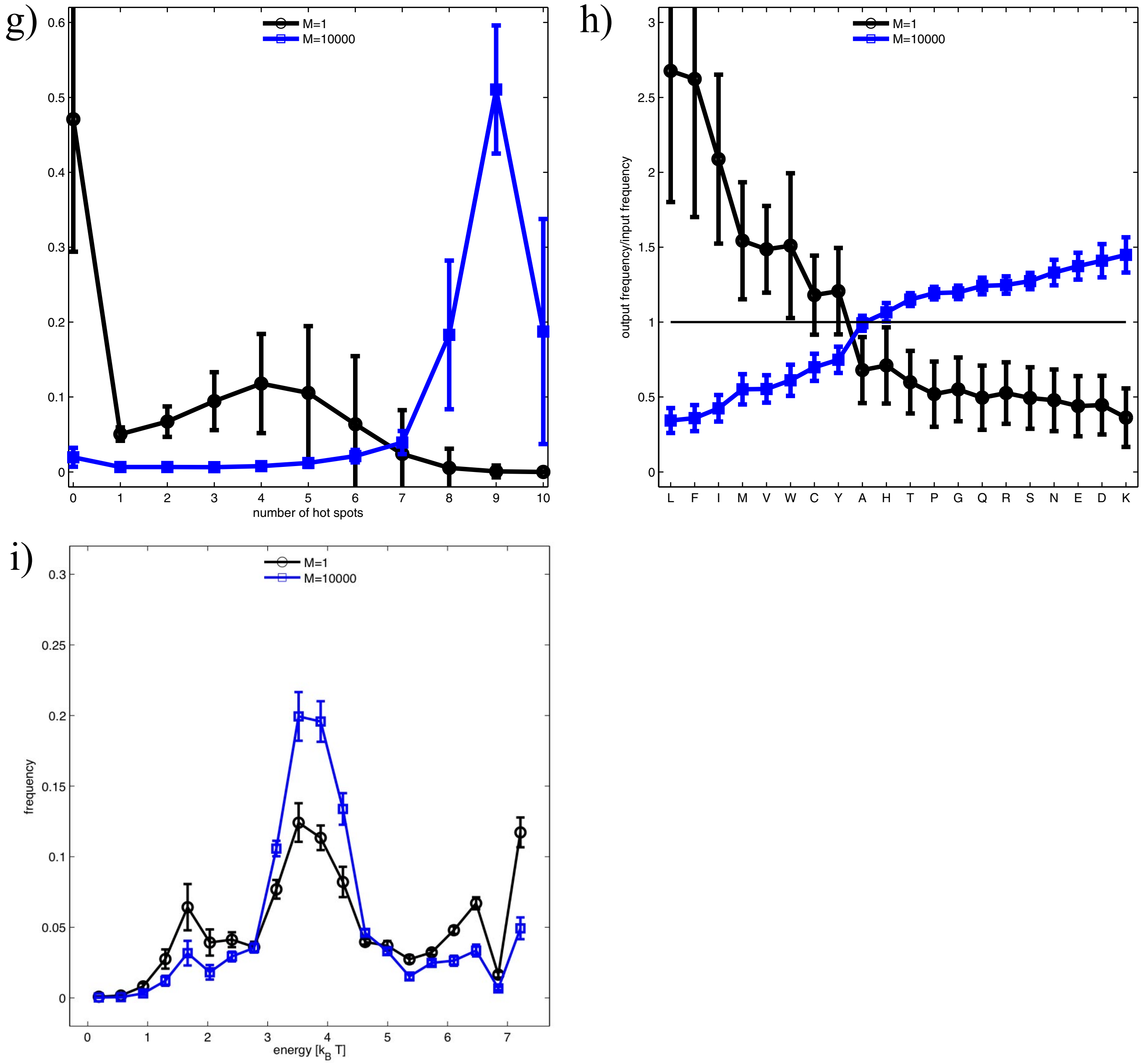

Fig. S1. *Continued.*

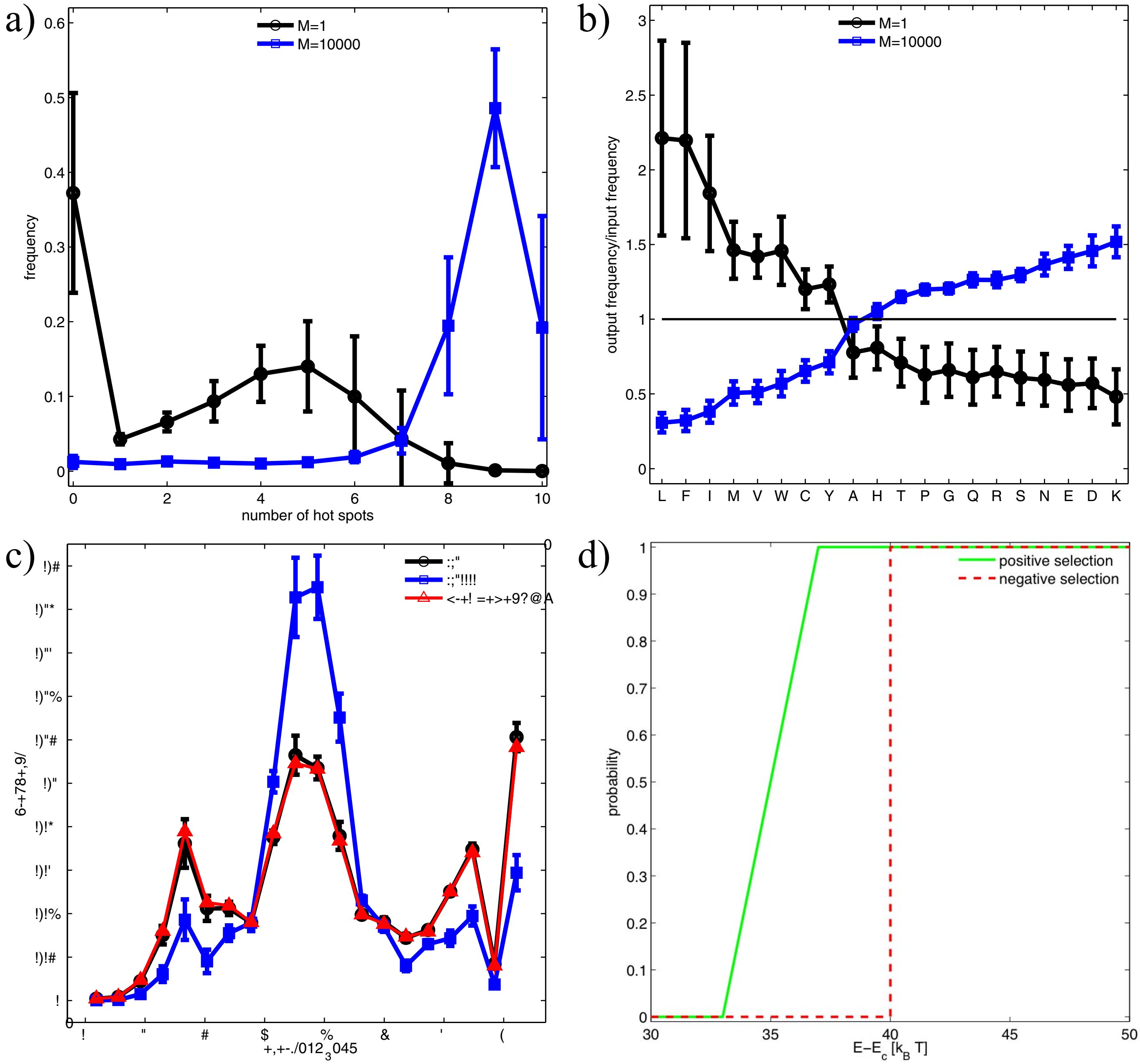

Fig. S2. Soft threshold for positive selection. (*a–c*) The qualitative features of the post-thymic selection TCR repertoire are robust to the nature of threshold for positive selection. We show the number of hot spots (*a*), the amino acid composition of selected TCRs (*b*), and the distribution of contact energies between selected TCRs and antigenic pMHC (*c*) with a soft threshold for positive selection. (*d*) The interaction energy dependence of selection probability [positive selection (green curve) and negative selection (red curve)] for a given TCR when it interacts with self-peptide during thymic selection in our model is shown. The statistical properties of the TCR repertoire (*a–c*) remain unchanged upon introduction of a soft threshold for positive selection. ($E_N - E_c = 40\ k_BT$, $E_N - E_p = 5\ k_BT$).

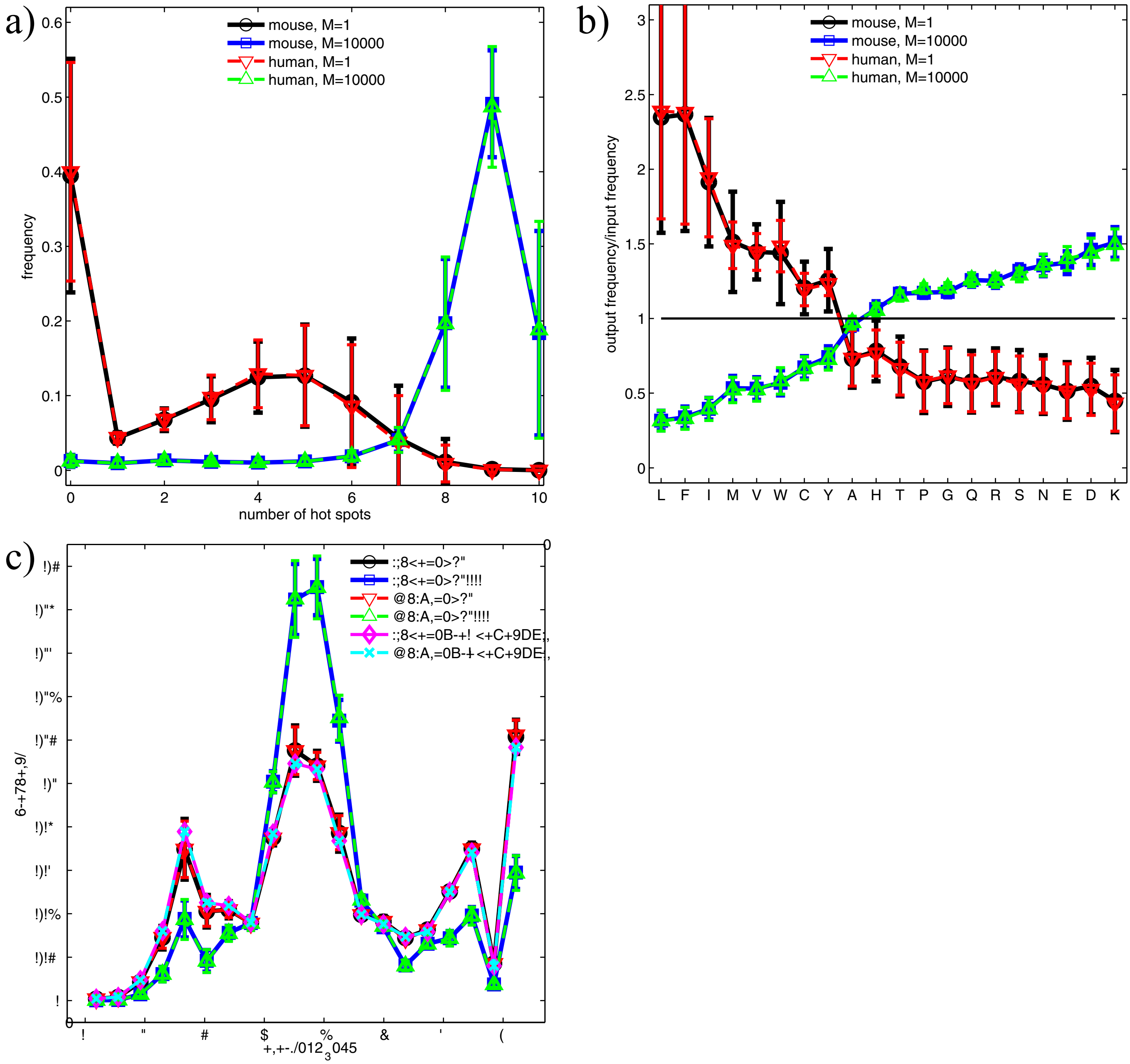

Fig. S3. Thymic selection using amino acid frequencies from mouse proteome. Distribution of hot spots (*a*), amino acid composition of selected TCRs (*b*), and distribution of contact energies between selected TCRs and antigenic pMHC (*c*) are similar whether using amino acid frequencies from mouse or human proteome to generate TCRs and self-peptides. ($E_N - E_c = 40\ k_BT$, $E_N - E_p = 5\ k_BT$).

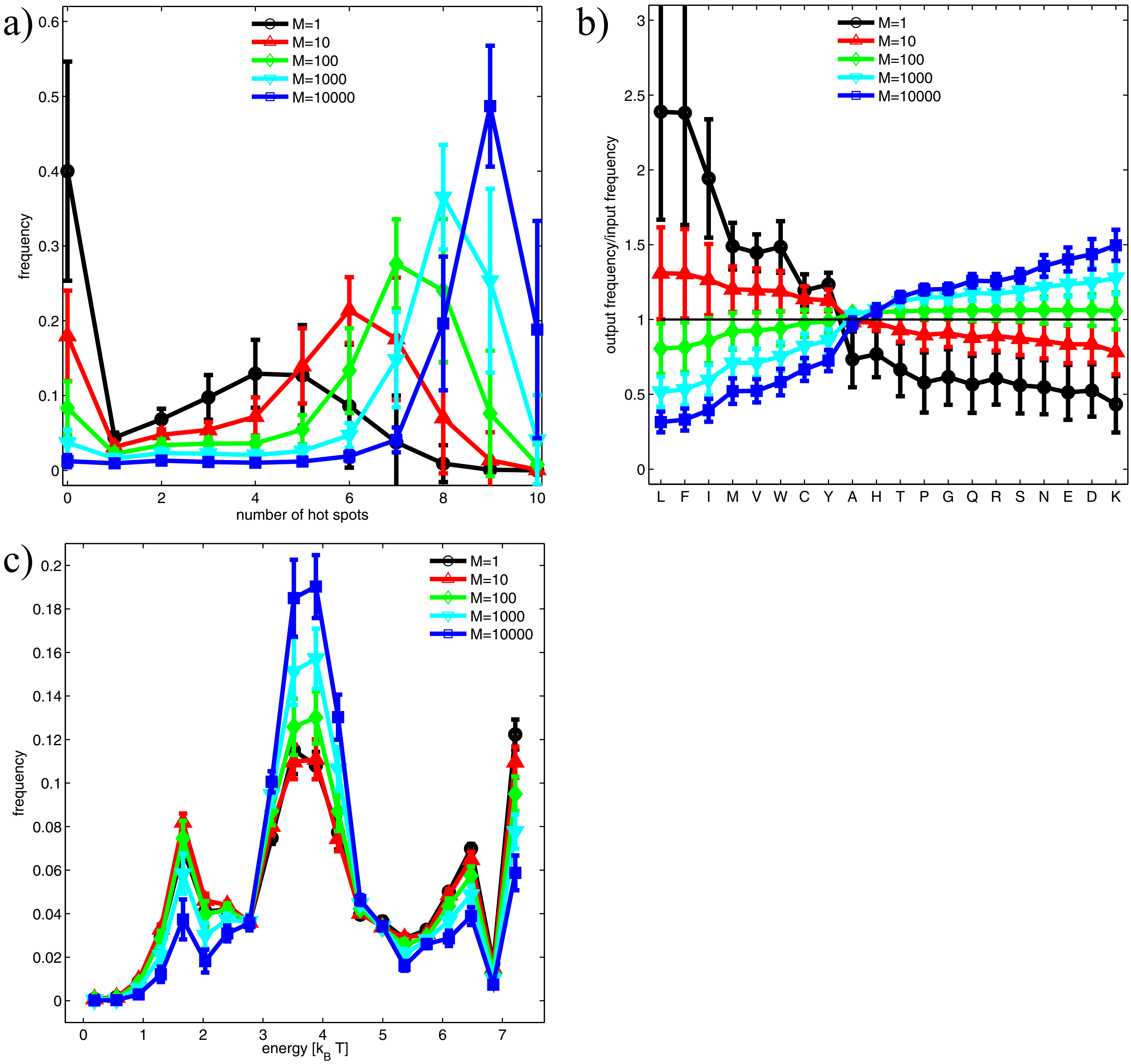

Fig. S4. Results of thymic selection as a function of the number of self-peptides M in the thymus. (*a*) The number of hot spots increase with number of self-peptides presented in the thymus. (*b*) The dependence of the amino acid distribution of selected TCR sequences as a function of the number of self peptides in the thymus. (*c*) The distribution of contact energies between selected reactive TCRs and antigenic peptides. Increasing the number of self-peptides in the thymus results in more moderate contacts and less weak and strong contacts. These results (particularly *b*) show that as long as there are $>$100 types of endogenous pMHC in the thymus, the qualitative results reported in the main text would be obtained. ($E_N - E_c = 40\ k_BT$, $E_N - E_p = 5\ k_BT$).

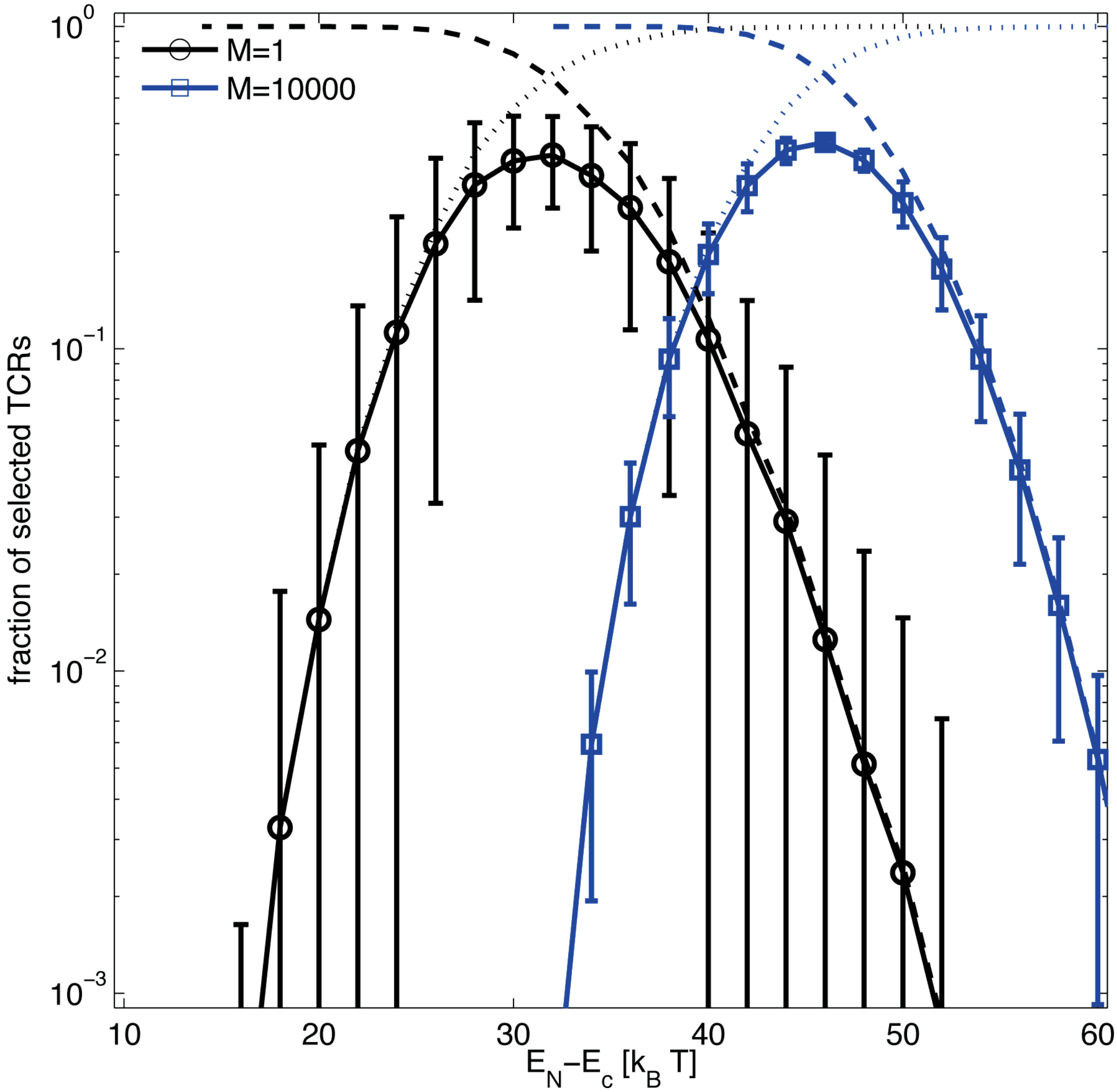

Fig. S5. TCR selection probabilities. Fraction of selected TCRs against one self-peptide (black curve) and many types of self-peptides (blue curve, $M = 10{,}000$) as a function of the threshold for negative selection $E_N - E_c$, whereas the gap between thresholds for negative and positive selection is kept constant at $E_N - E_p = 5\ k_BT$. At small values of $E_N - E_c$ negative selection is dominant—dotted lines show fraction of TCRs that are not negatively selected. At large values of $E_N - E_c$ positive selection is dominan—broken lines show fraction of TCRs that are positively selected.

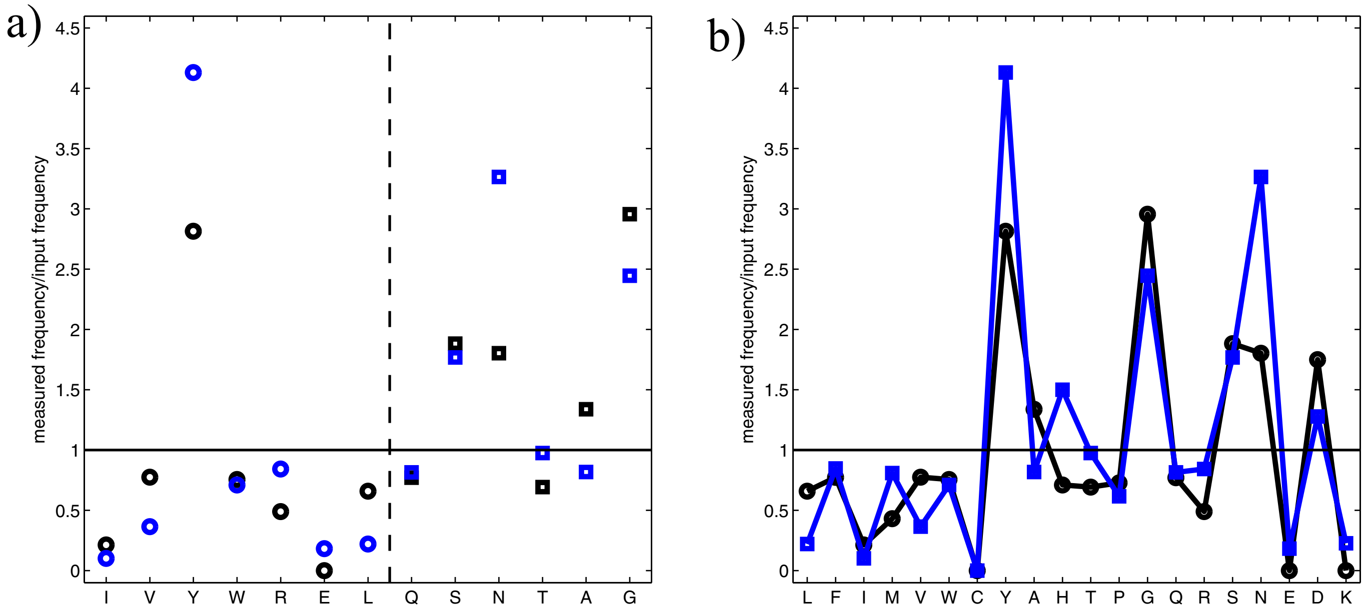

Fig. S6. Frequency distribution of amino acids in TCR that are in contact with peptide. The ratio of amino acid frequencies derived from the list of amino acids of TCRs in contact with peptides calculated from 18 available crystal structures of TCR–pMHC(I) complexes with respect to the amino acid frequencies from human proteome are presented in these graphs. The residues are said to be in contact with each other if the Cα–Cα distance is $<$ 6.5 Å (black points and Fig. 4*B* of main text). In a separate analysis, any two residues are defined to be in contact if a water molecule cannot fit between them (blue points). The dominance of weakly interacting amino acids is robust to the definition of contact between residues. (*a*) The abscissa is divided into the two types of strong amino acids (IVYWREL) and weak amino acids (QSNTAG) according to ref. 12. (*b*) The amino acids on abscissa are ordered from strongest (L) to weakest (K) according to the strongest interaction with another amino acid using the MJ matrix. This ordering presents charged amino acids (REDK) as weak. In contrast, according to ref. 12, amino acids R and E are strong, and amino acids D and K are not weak.

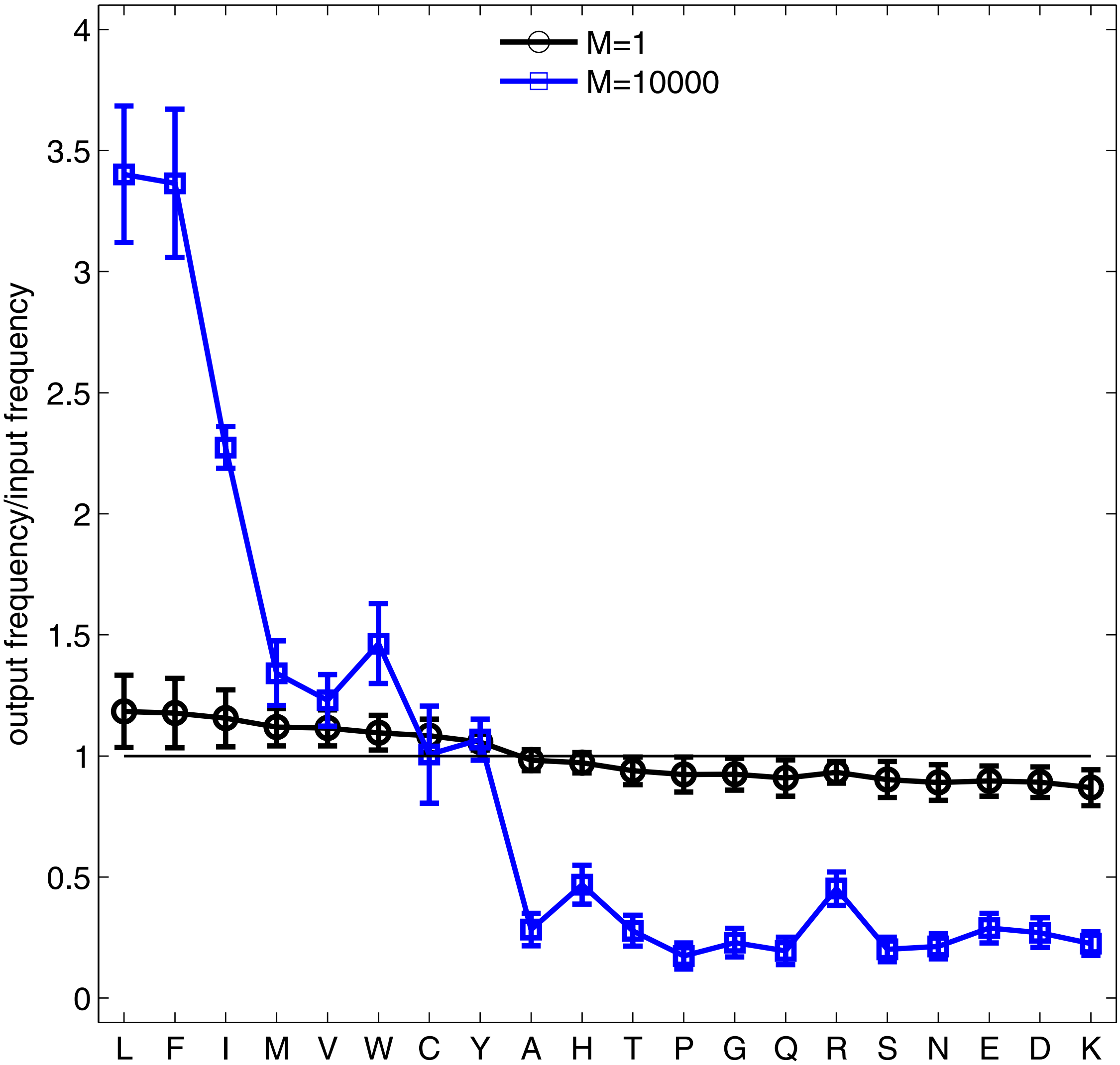

Fig. S7. Amino acid frequencies of recognized antigenic peptides. Depicted is the ratio of amino acid frequencies of reactive antigenic peptides, defined as those that are recognized by at least one of the selected TCRs with respect to amino acid frequencies of all antigenic peptides (*Listeria monocytogenes*). The black curve depicts the results for TCRs selected against one self-peptide, whereas the blue curve corresponds to selection against many self-peptides (M = 10,000). For TCRs selected against many self-peptides, the reactive antigens are composed of more strong amino acids. The amino acids on the abscissa are ordered from strongest (L) to weakest (K) according to the strongest interaction with another amino acid in the MJ matrix. ($E_N - E_c = 40\ k_BT$, $E_N - E_p = 5\ k_BT$).

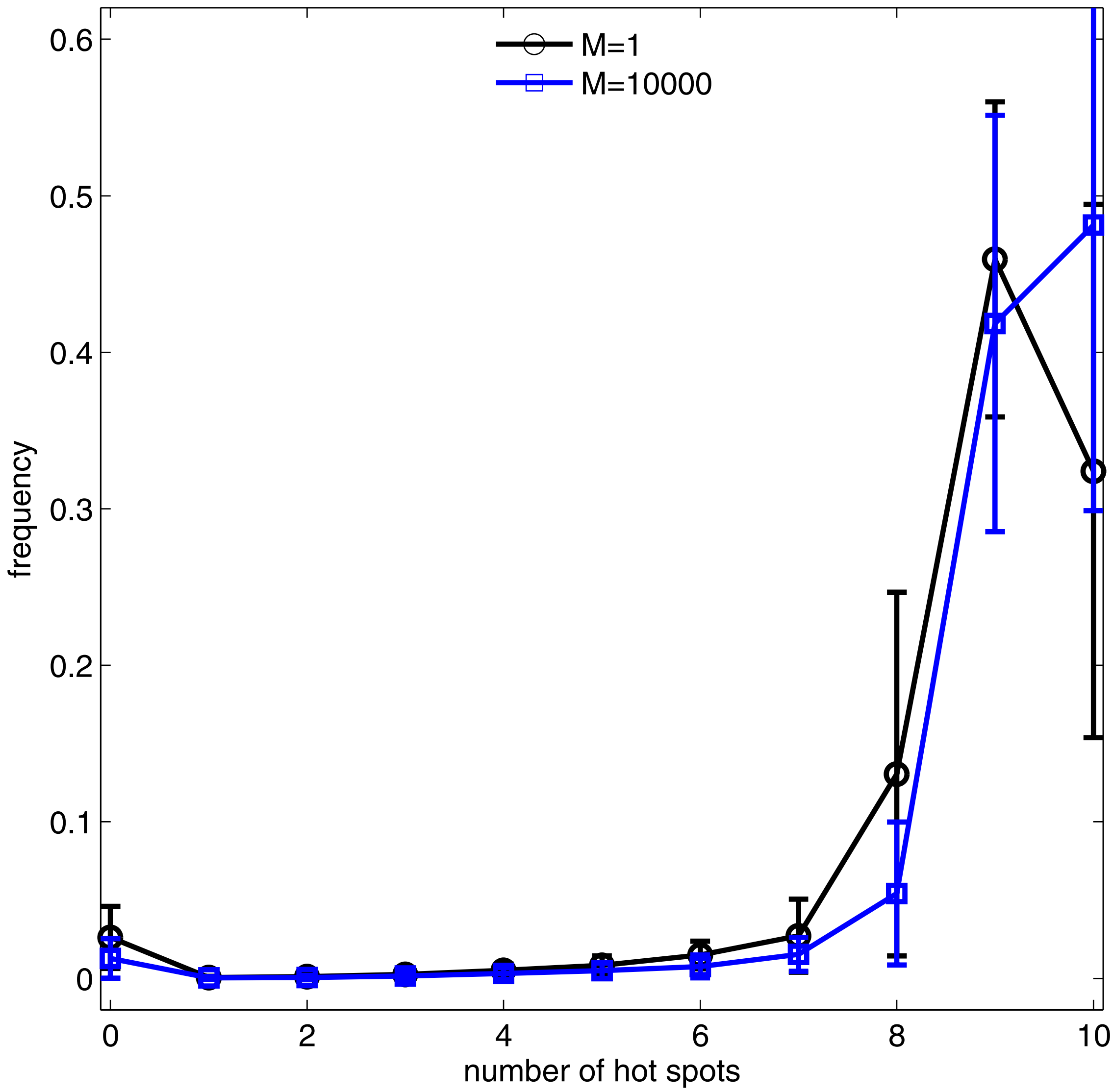

Fig. S8. Distribution of hot spots for small value of E_c (weak TCR–MHC interactions) and large gap, $E_N - E_p$. When interactions between TCRs and MHC are weak ($E_N - E_c = 60\ k_BT$) and the gap between negative and positive selection thresholds ($E_N - E_p = 30\ k_BT$) is large, the distribution of the number of hot spots shows a peak at large values for TCRs selected in thymus both against one self-peptide (black curve) and against many self-peptides (blue curve, $M = 10{,}000$).

Table S1. Amino acid frequencies of *Homo sapiens*, mouse and *Listeria monocytogenes* proteomes

	Homo sapiens	*Mus musculus* (house mouse)	*Listeria monocytogenes*
A	0.0692	0.0681	0.0774
C	0.0225	0.0228	0.0061
D	0.0476	0.0481	0.0544
E	0.0718	0.0700	0.0744
F	0.0359	0.0369	0.0453
G	0.0658	0.0641	0.0667
H	0.0261	0.0263	0.0178
I	0.0434	0.0439	0.0784
K	0.0576	0.0576	0.0716
L	0.0985	0.0993	0.0951
M	0.0215	0.0221	0.0275
N	0.0360	0.0358	0.0462
P	0.0636	0.0619	0.0347
Q	0.0481	0.0479	0.0346
R	0.0568	0.0563	0.0365
S	0.0836	0.0850	0.0580
T	0.0536	0.0541	0.0611
V	0.0598	0.0609	0.0704
W	0.0123	0.0120	0.0093
Y	0.0263	0.0269	0.0345

Table S2. TCR selection probabilities

Weak TCR–MHC interactions (small value of E_c, $E_N - E_c > 55\ k_BT$)		Strong TCR–MHC interactions (large value of E_c, $E_N - E_c < 35\ k_BT$)	
Small gap between selection thresholds ($E_N - E_p \leq 5\ k_BT$)	Large gap between selection thresholds ($E_N - E_p > 20\ k_BT$)	Small gap between selection thresholds ($E_N - E_p \leq 5\ k_BT$)	Large gap between selection thresholds ($E_N - E_p > 20\ k_BT$)
Very few TCRs are positively selected in thymus, e.g. ≈0.02% are negatively selected and ≈0.5% positively selected at $E_N - E_c$=60 k_BT, $E_N - E_p = 5\ k_BT$	Almost all TCRs are positively selected and very few TCRs are negatively selected in thymus, e.g. ≈0.02% are negatively selected and ≈100% positively selected at $E_N - E_c = 60\ k_BT$, $E_N - E_p = 30\ k_BT$	Almost all TCRs are negatively selected in thymus, e.g. ≈100% are negatively selected at $E_N - E_c = 30\ k_BT$	Almost all TCRs are negatively selected in thymus, e.g. ≈100% are negatively selected at $E_N - E_c$ =30 k_BT

Fraction of selected TCRs for different values of parameters E_c (TCR–MHC interaction energy), E_N (threshold for negative selection) and E_p (threshold for positive selection) for $M = 10{,}000$ types of endogenous peptides in thymus.