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 $\to \infty$.

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.

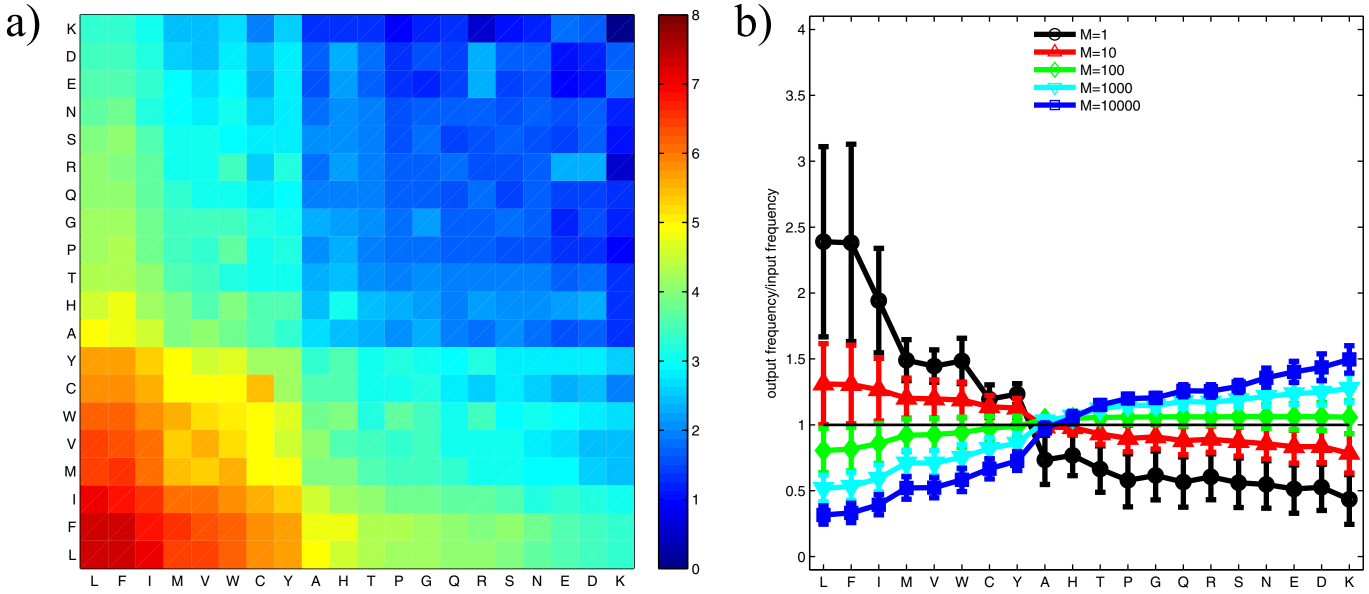

Fig. S1. Results for random statistical potential between amino acids. (*a–f*) In all calculations reported in the main text, the MJ matrix (11) was used to determine the interaction energy between peptide contact residues of the TCR and peptide amino acids. Here, we explore what happens if we use semi-random 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

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

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

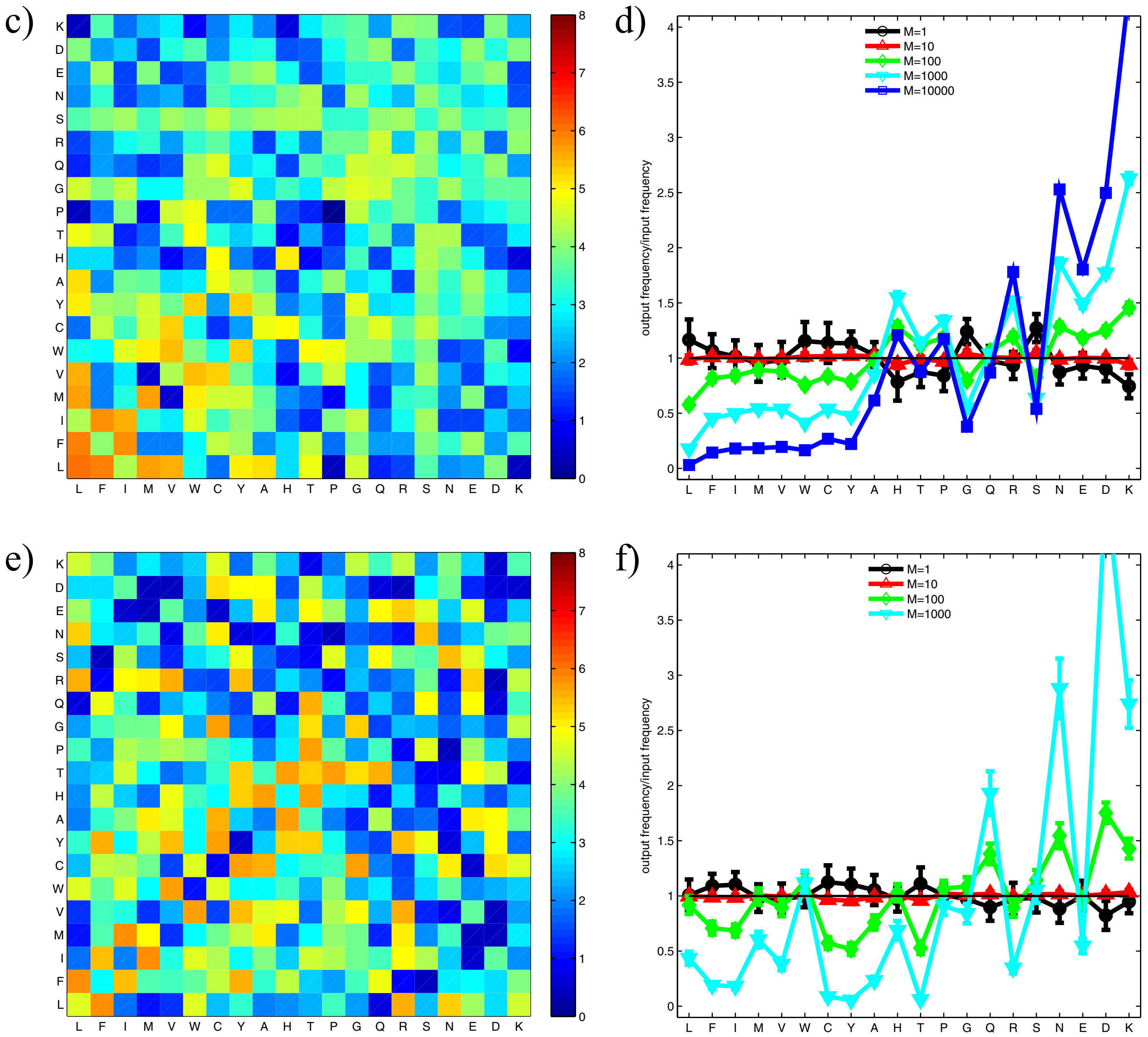

Fig. S1. *Continued.*

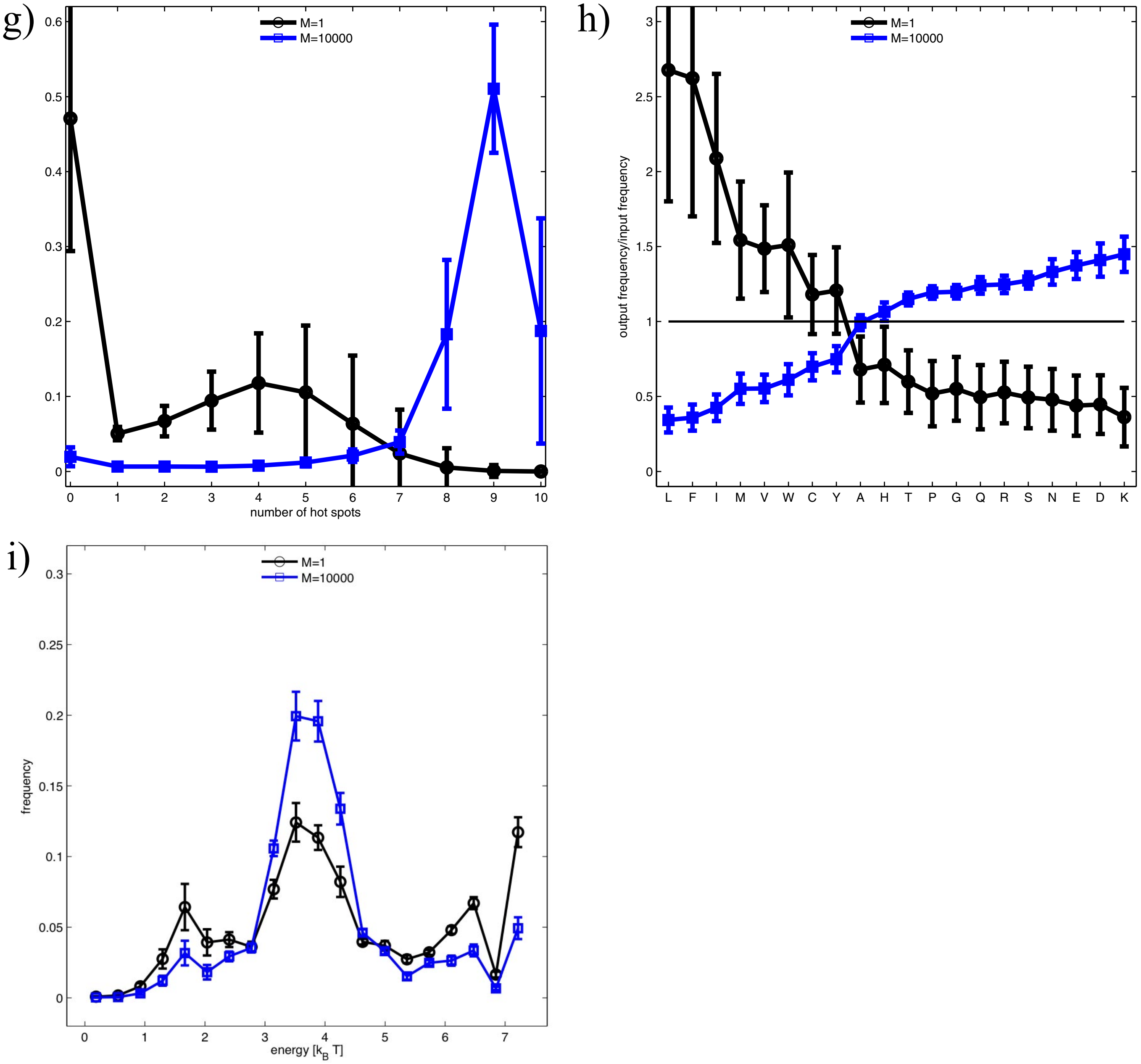

Fig. S1. *Continued.*

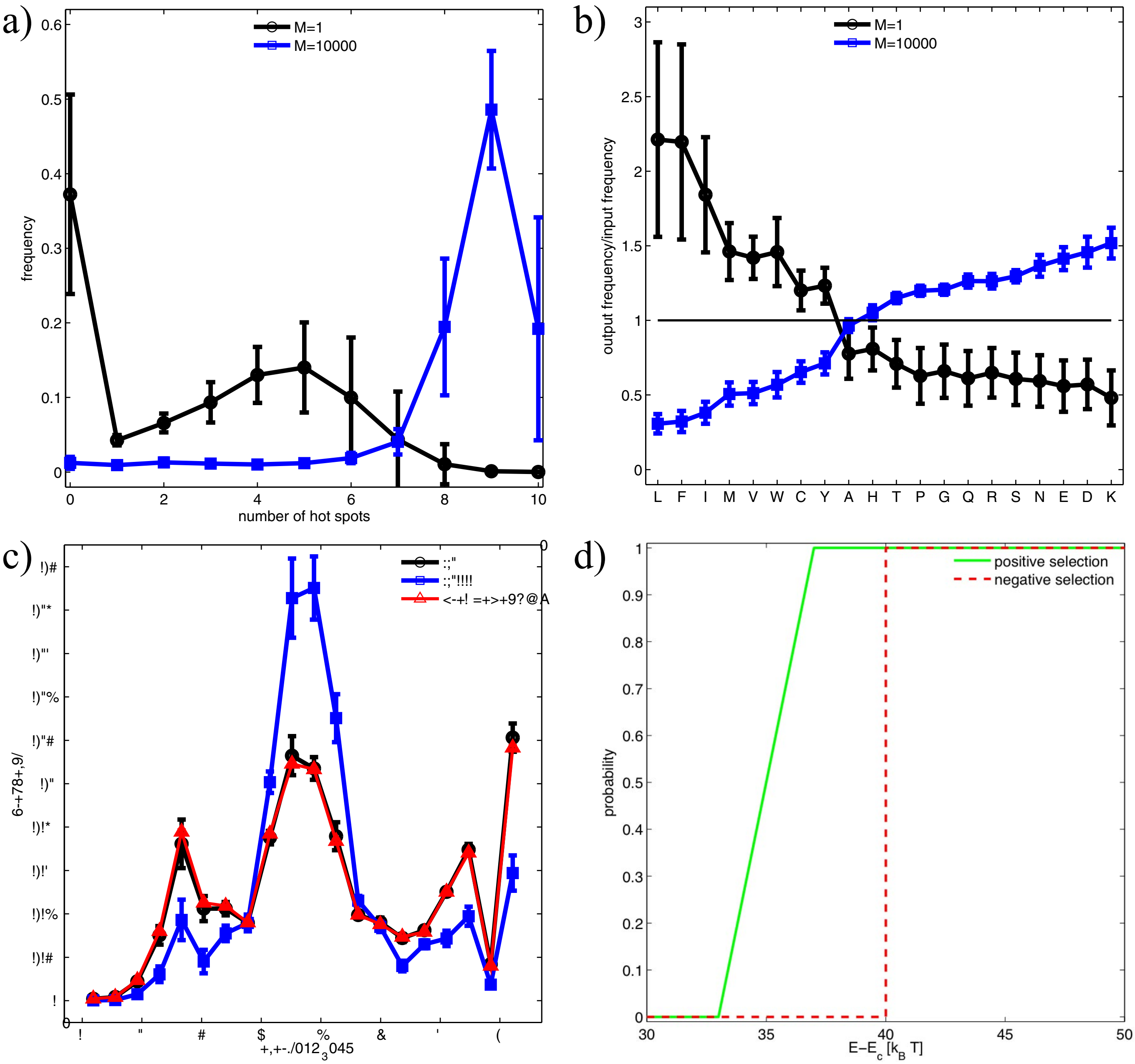

Fig. S2. Soft threshold for positive selection. (*a*–*c*) The qualitative features of the post-thymic selection TCR repertoire are robust to the nature of threshold for positive selection. We show the number of hot spots (*a*), the amino acid composition of selected TCRs (*b*), and the distribution of 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. ($E_{\mathrm{N}} - E_{\mathrm{c}} = 40\ k_BT$, $E_N - E_{\mathrm{p}} = 5\ k_{\mathrm{B}}T$).

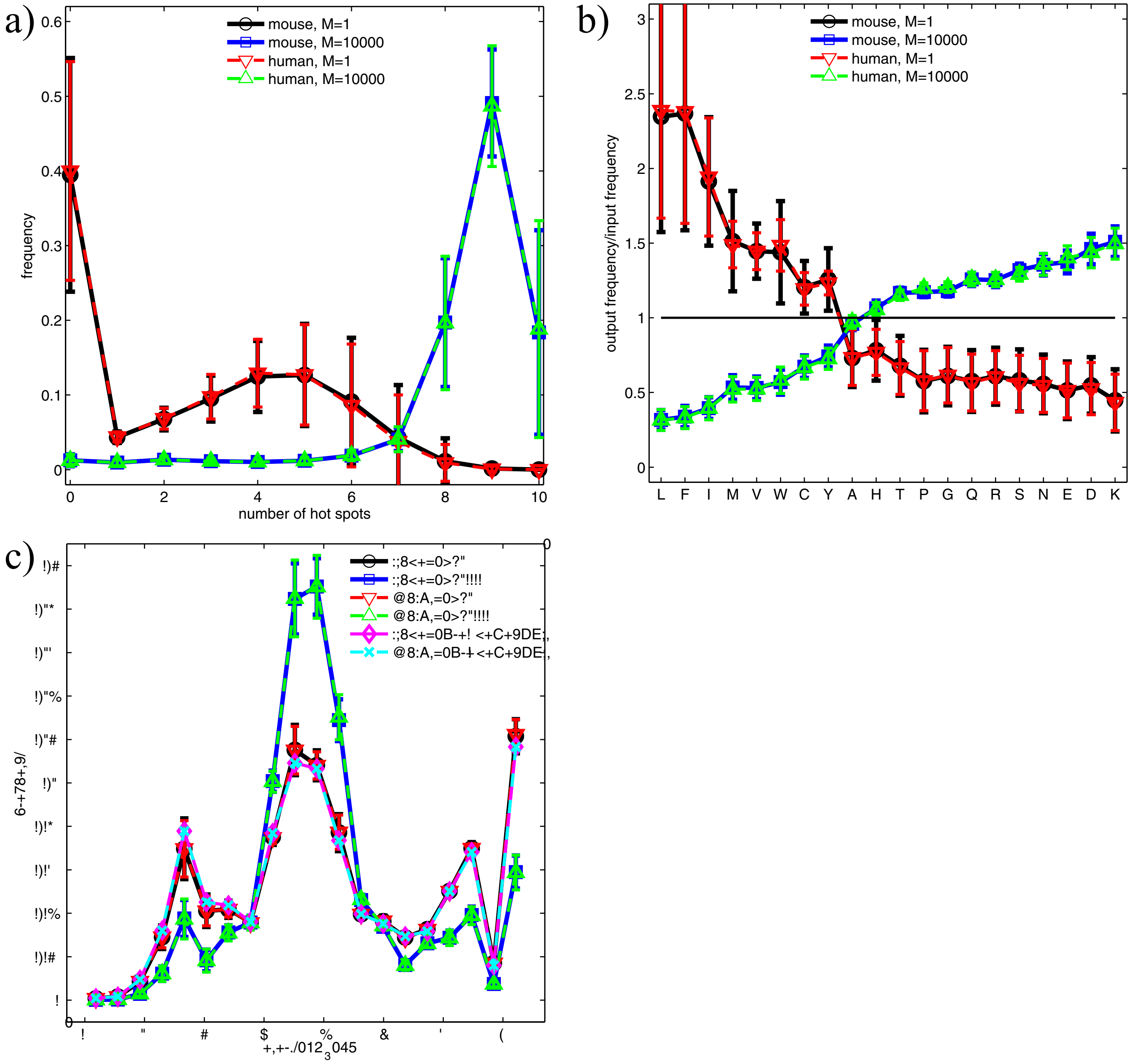

Fig. S3. Thymic selection using amino acid frequencies from mouse proteome. Distribution of hot spots (*a*), amino acid composition of selected TCRs (*b*), and distribution of contact energies between selected TCRs and antigenic pMHC (*c*) are similar whether using amino acid frequencies from mouse or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

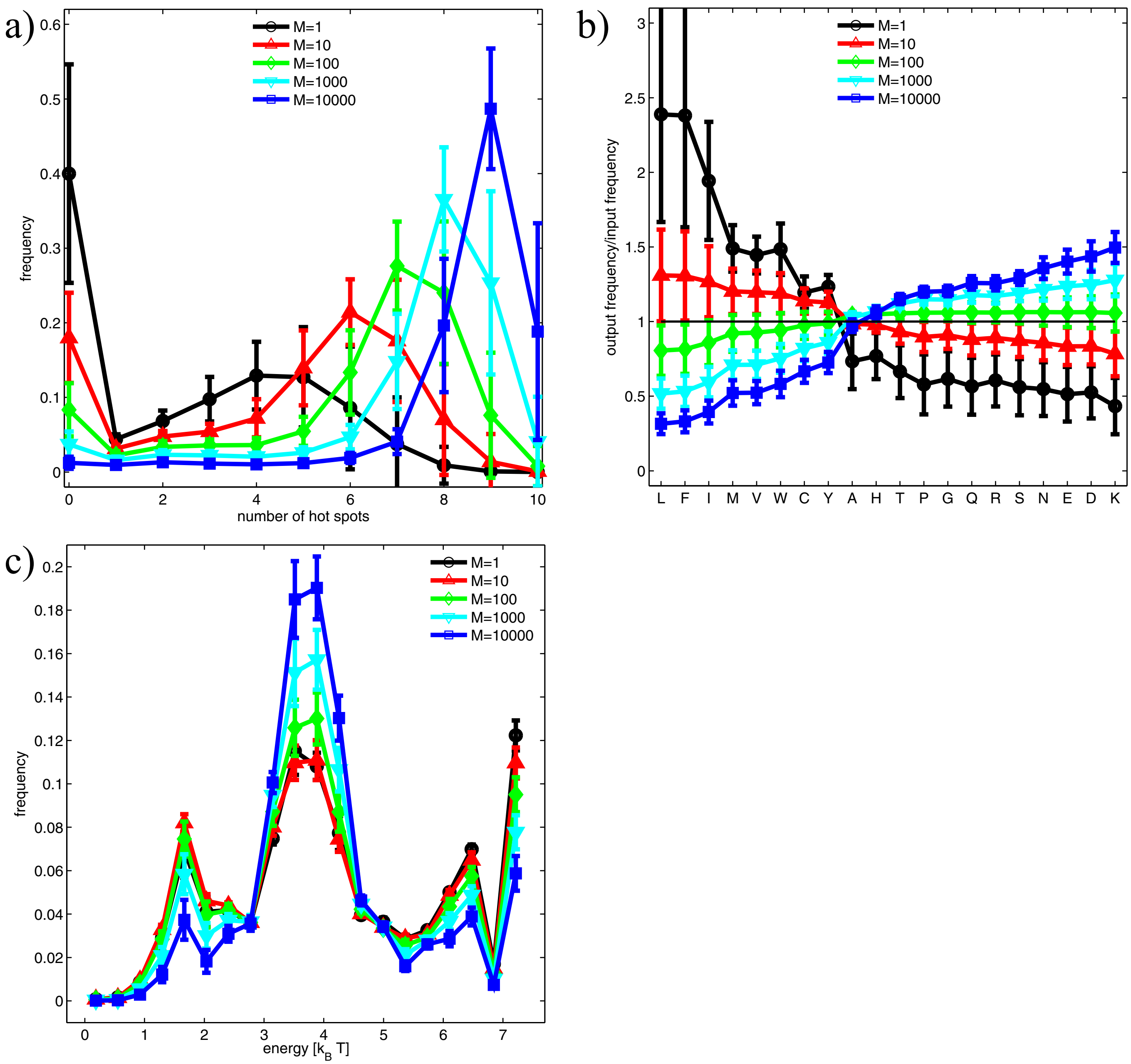

Fig. S4. Results of thymic selection as a function of the number of self-peptides *M* in the thymus. (*a*) The number of hot spots increase with number of self-peptides presented in the thymus. (*b*) The dependence of the amino acid distribution of selected TCR sequences as a function of the number of 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. ($E_N - E_c = 40\ k_BT$, $E_N - E_p = 5\ k_BT$).

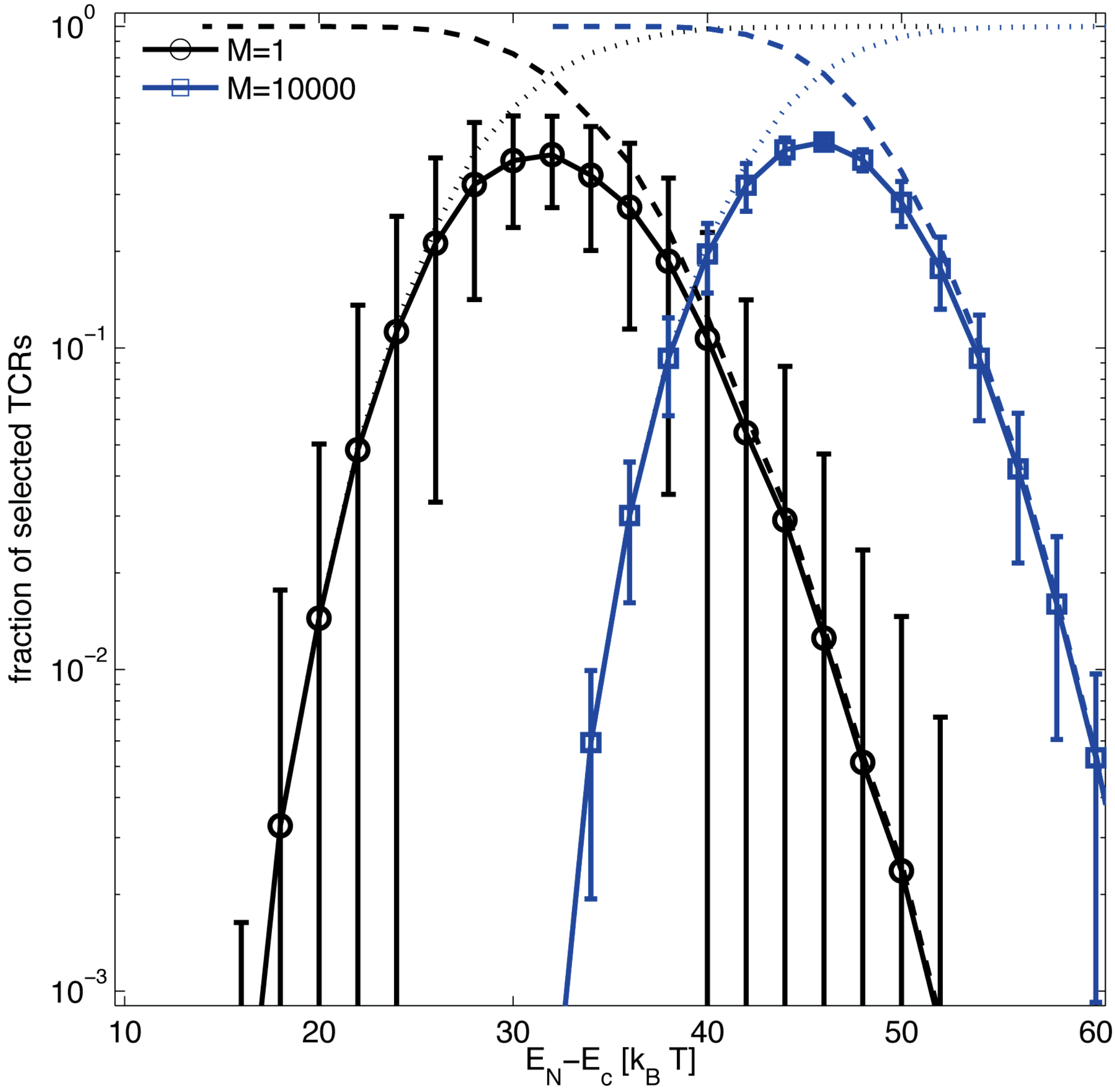

Fig. S5. TCR selection probabilities. Fraction of selected TCRs against one self-peptide (black curve) and many types of self-peptides (blue curve, $M = 10{,}000$) as a function of the threshold for negative selection $E_N - E_c$, whereas the gap between thresholds for negative and positive selection 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.

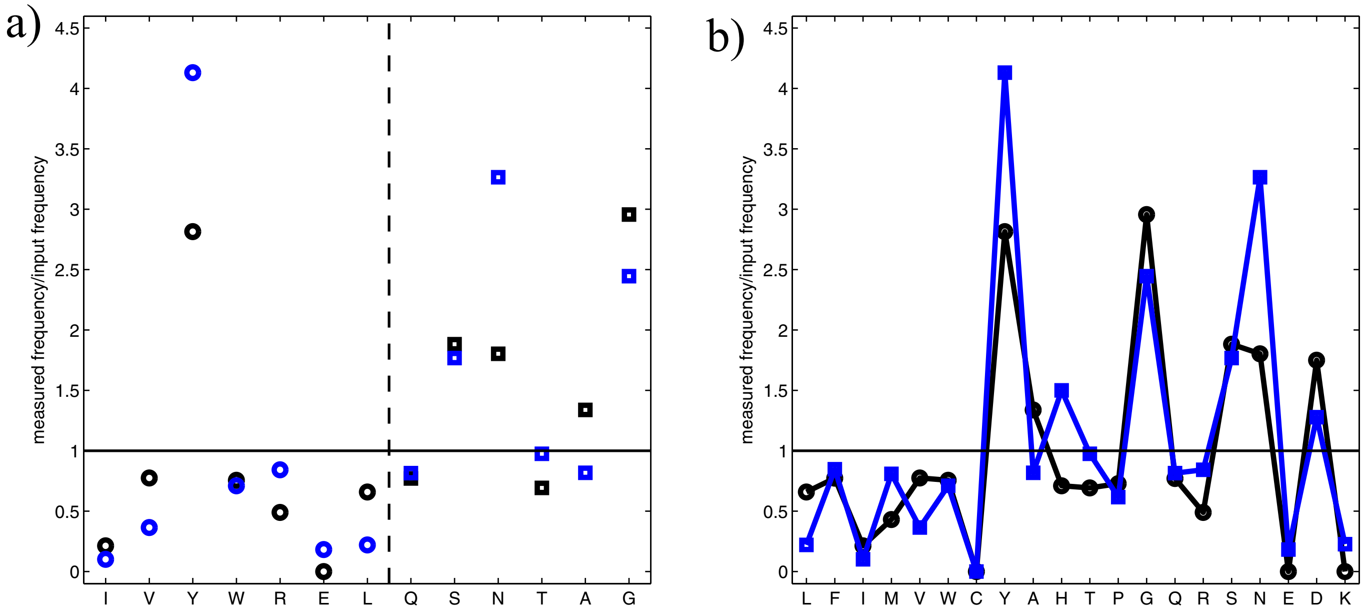

Fig. S6. Frequency distribution of amino acids in TCR that are in contact with peptide. The ratio of amino acid frequencies derived from the list of amino acids of TCRs in contact with peptides calculated from 18 available crystal structures of TCR–pMHC(I) complexes with respect to the amino acid frequencies 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.

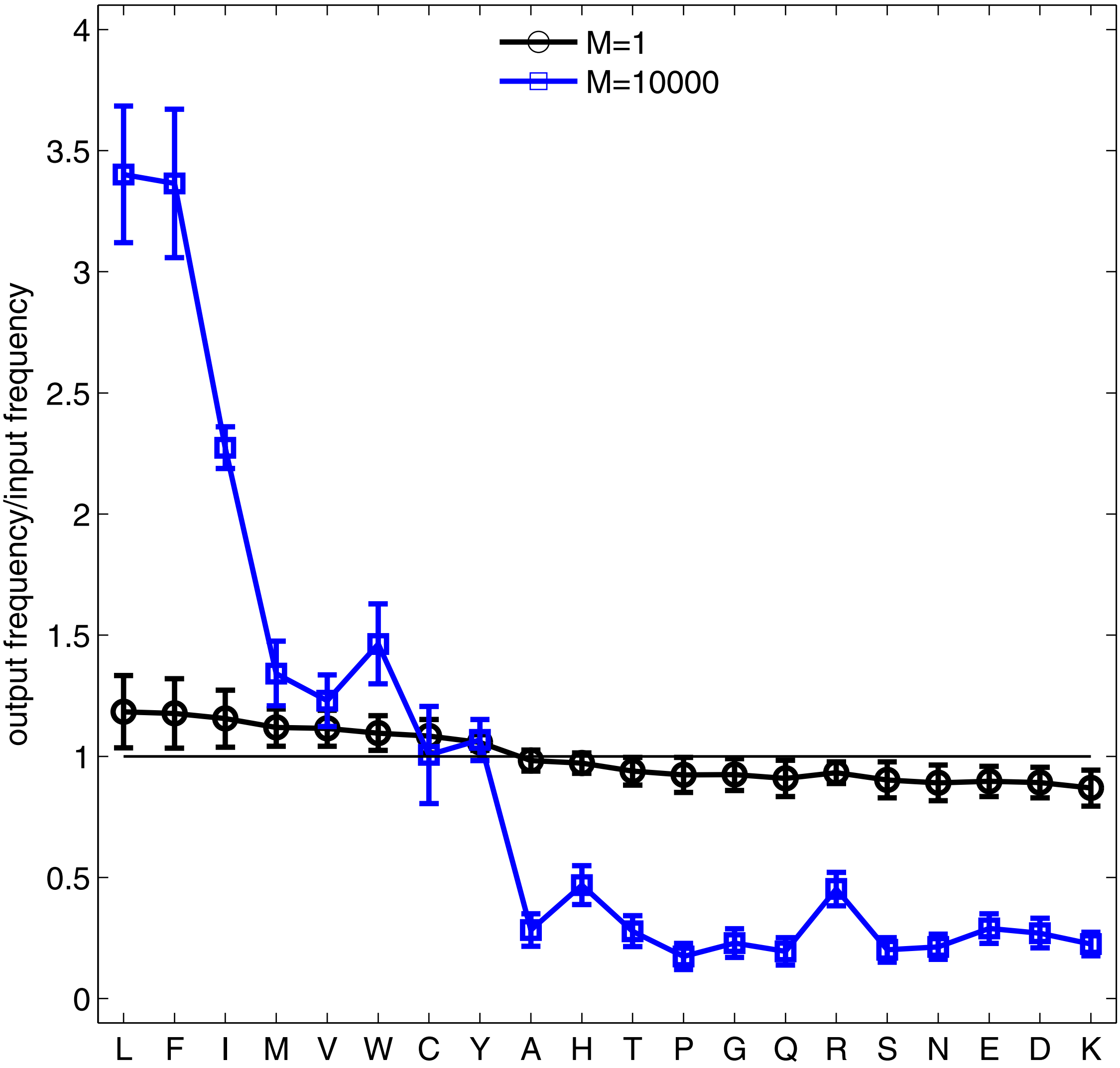

Fig. S7. Amino acid frequencies of recognized antigenic peptides. Depicted is the ratio of amino acid frequencies of reactive antigenic peptides, defined as those that are recognized by at least one of the selected TCRs with respect to amino acid frequencies of all antigenic peptides (*Listeria monocytogenes*). 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. ($E_{\mathrm{N}} - E_{\mathrm{c}} = 40\ k_BT$, $E_N - E_{\mathrm{p}} = 5\ k_{\mathrm{B}}T$).

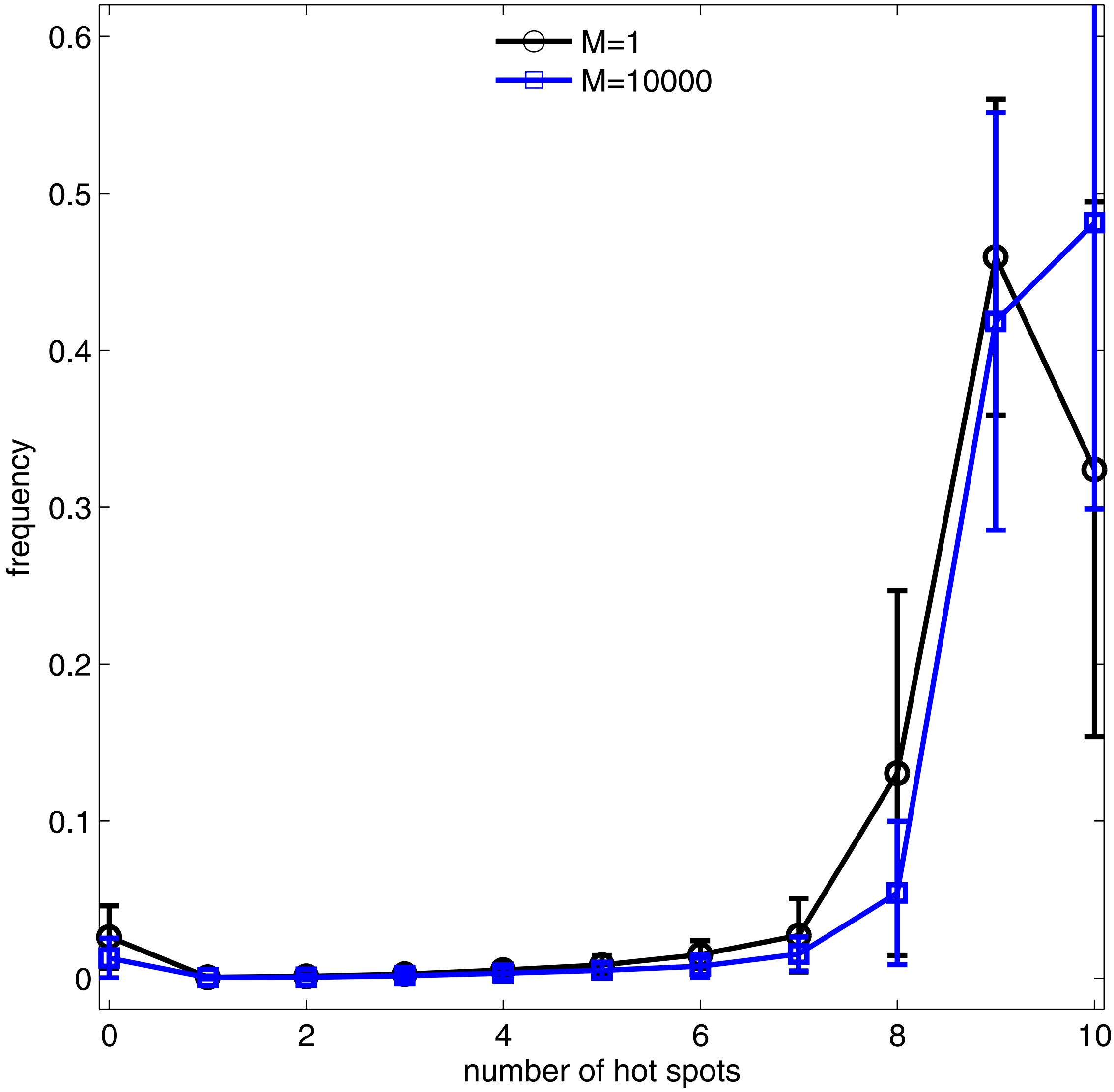

Fig. S8. Distribution of hot spots for small value of E_c (weak TCR–MHC interactions) and large gap, $E_N - E_p$. When interactions between TCRs and MHC are weak ($E_N - E_c = 60\ k_BT$) and the gap between negative and positive selection thresholds ($E_N - E_p = 30\ k_BT$) is large, the distribution 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.