Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself

Balázs Koncz

PhD Thesis

Szeged

2021

UNIVERSITY OF SZEGED, ALBERT SZENT-GYÖRGYI MEDICAL SCHOOL
DEPARTMENT OF DERMATOLOGY AND ALLERGOLOGY
DOCTORAL SCHOOL OF CLINICAL MEDICINE

Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself

Balázs Koncz

PhD Thesis

Supervisor:
Dr. Máté Manczinger

Szeged

2021

PUBLICATIONS

Scientific paper related to the thesis

I. Balázs Koncz, Gergő M. Balogh, Benjamin T. Papp, Leó Asztalos, Lajos Kemény, Máté Manczinger: Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself. *Proceedings of the National Academy of Sciences* (2021). DOI: 10.1073/pnas.2100542118. IF: 11.205, SCImago Journal Rank: Q1 / D1.

Publications not directly related to the thesis

II. Máté Manczinger, Balázs Koncz, Gergő Mihály Balogh, Benjamin Tamás Papp, Leó Asztalos, Lajos Kemény, Balázs Papp & Csaba Pál. Negative trade-off between neoantigen repertoire breadth and the specificity of HLA-I molecules shapes antitumor immunity. *Nature Cancer* (2021). DOI: 10.1038/s43018-021-00226-4.

III. Párniczky A, Lantos T, Tóth EM, Szakács Z, Gódi S, Hágendorn R, Illés D, Koncz B, Márta K, Mikó A, Mosztbacher D, Németh BC, Pécsi D, Szabó A, Szücs Á, Varjú P, Szentesi A, Darvasi E, Erőss B, Izbéki F, Gajdán L, Halász A, Vincze Á, Szabó I, Pár G, Bajor J, Sarlós P, Czimmer J, Hamvas J, Takács T, Szepes Z, Czakó L, Varga M, Novák J, Bod B, Szepes A, Sümegi J, Papp M, Góg C, Török I, Huang W, Xia Q, Xue P, Li W, Chen W, Shirinskaya NV, Poluektov VL, Shirinskaya AV, Hegyi PJ, Bátovský M, Rodriguez-Oballe JA, Salas IM, Lopez-Diaz J, Dominguez-Munoz JE, Molero X, Pando E, Ruiz-Rebollo ML, Burgueño-Gómez B, Chang YT, Chang MC, Sud A, Moore D, Sutton R, Gougol A, Papachristou GI, Susak YM, Tiuliukin IO, Gomes AP, Oliveira MJ, Aparício DJ, Tantau M, Kurti F, Kovacheva-Slavova M, Stecher SS, Mayerle J, Poropat G, Das K, Marino MV, Capurso G, Małecka-Panas E, Zatorski H, Gasiorowska A, Fabisiak N, Ceranowicz P, Kuśnierz-Cabala B, Carvalho JR, Fernandes SR, Chang JH, Choi EK, Han J, Bertilsson S, Jumaa H, Sandblom G, Kacar S, Baltatzis M, Varabei AV, Yeshy V, Chooklin S, Kozachenko A, Veligotsky N, Hegyi P; Hungarian Pancreatic Study Group. Antibiotic therapy in acute pancreatitis: From global overuse to evidence based recommendations. *Pancreatology* (2019). DOI: 10.1016/j.pan.2019.04.003. IF: 3.996, SCImago Journal Rank: Q1 / D3.

IV. Balázs Koncz, Erika Darvasi, Dalma Erdősi, Andrea Szentesi, Katalin Márta, Bálint Erőss, Dániel Pécsi, Zoltán Gyöngyi, János Girán, Nelli Farkas, Maria Papp, Eszter Fehér, Zsuzsanna Vitális, Tamás Janka, Áron Vincze, Ferenc Izbéki, Veronika Dunás-Varga, László Gajdán, Imola Török, Sándor Károly, Judit Antal, Noémi Zádori, Markus M Lerch, John Neoptolemos, Miklós Sahin-Tóth, Ole H Petersen, and Péter Hegyi. LIFEStyle, Prevention and Risk of Acute PaNcreatitis (LIFESPAN): Protocol of a Multicentre and Multinational Observational Case-Control Study. *BMJ Open.* (2020). DOI: 10.1136/bmjopen-2019-029660. IF: 2.692, SCImago Journal Rank: Q1 / D2.

Cumulative impact factor: 17,893

TABLE OF CONTENTS

ABBREVIATIONS

9-mer	nine amino acids long peptide
10-mer	ten amino acid long peptide
AIC	Akaike information criterion
ANOVA	analysis of variance
AUC	area under the curve
BLAST	basic local alignment search tool
BLOSUM62	blocks substitution matrix 62
cTEC	cortical thymic epithelial cell
CTL	cytotoxic T cell, cytotoxic T lymphocyte
DS1	dataset 1
DS2	dataset 2
HLA	human leukocyte antigen
HLA-I	HLA class I (molecule)
HUGO	human genome organisation
IEDB	immune epitope database
IQR	interquartile range
IP	immunoproteasomal
MCA	multiple correspondence analysis
MIRA	multiplex identification of T-cell receptor antigen specificity
np-TCEMs	referring to TCEMs for which we expect to find specific positively selected T cells with lower probability
OR	odds ratio
pHLA	peptide HLA molecule complex
ROC curve	receiver operating characteristic curve
RPKM	reads per kilobase million
TAP	transporter associated with antigen processing
TCEM	T cell exposed motif
TCR	T cell receptor
TP	thymoproteasomal

1. INTRODUCTION

1.1. Basis of adaptive immune recognition

In the second half of the 20th century, the self-nonself theory was the general view in immunology, which suggests that the immune system's primary goal is to discriminate between self and nonself (1, 2). Therefore, an immune response is triggered against foreign entities, but not against the organism's own materials (3).

In the 1990s, *Matzinger* outlined the danger theory, which suggests the immune system has a principal role in the danger detection and protection against harmful agents (4). It claims that self-elements can trigger an immune response if they are dangerous (e.g., cellular stress); and nonself constituents can be tolerated if they are not dangerous (e.g., commensal bacteria) (4). Immune responses are triggered by 'danger/alarm signals', which are released by the organism's own cells (4).

In sum, if the immune system could effectively recognize cells that contain mutated proteins or intracellular pathogens, then there is a higher chance for it to destroy them.

The immune system is typically divided into two subsystems - innate and adaptive – which operate in various but coordinated ways. Innate immunity is ready for action from the very beginning of an infection. After it has detected the presence of pathogens, it enhances the gene expression and protein synthesis of molecules associated with immune response. Innate immunity alone is unable to defeat most infections. Therefore, the adaptive immune system joins the battle and responds to the intruder in a very specific and more effective way. Antigens play a key role in the adaptive immune response. These molecules are originated from pathogens or are produced by human cells and can trigger an immune response. The adaptive immune response is specific for antigens and provides long-lasting immunity against pathogens (5).

The T cell activation is the result of innate immune processes in cells (phagocytosis, increasing of the expression of certain cytokines) and the formation of the immunological synapse (6). The latter structure consists of the peptide-human leukocyte antigen (pHLA) complex, T cell receptor (TCR), adhesion molecules, and checkpoint receptors (6). In the following, I present the most important properties of human leukocyte antigen (HLA) molecules.

HLA molecules have a major role in peptide presentation during the adaptive immune response. These molecules are encoded by the human leukocyte antigen super-locus, which represents a genomic region at the chromosomal position 6p21 (7). HLA genes are extremely polymorphic: more than 20,000 HLA alleles are registered currently (8). Generally, HLA molecules are classified into two classes: HLA-I and HLA-II. Classical HLA-I molecules (HLA-A, HLA-B, and HLA-C) are expressed in almost all cells, while HLA-II molecules (HLA-DP, HLA-DQ, and HLA-DR) are mainly found in B cells, myeloid dendritic cells and monocytes (9). Normal and mutated self-peptides and the ones that are originated from intracellular pathogens are presented by HLA-I molecules. At the same time peptides of extracellular proteins and extracellular pathogens appear on the cell surface bound to HLA-II molecules.

In my thesis, I am going to focus on HLA-I-presented peptides therefore I explain the intrinsic antigen presentation pathway in detail. Self or pathogen-associated proteins are cleaved to 8-11 (most often 9) amino acids long fragments by proteasomes (10) and aminopeptidases (11) in the cytosol. These peptide fragments are translocated to the endoplasmic reticulum via the transporter associated with antigen processing (TAP) molecule (12). HLA-I molecules bind peptides within the peptide loading complex (13). The resulting pHLA-I complex is transported to the cell surface, where it is anchored to the cell membrane (5). It is important to emphasize that the HLA-I presentation of peptides is highly dependent on the expression of the encoding gene (14).

Adaptive immune recognition is dependent on the presence of antigen-specific T cells in the T cell repertoire (15). If antigen-specific cytotoxic T cells (CTL) can be found in the repertoire, the immunological synapse can be formed. In a healthy state, the immune response is dependent on the peptides presented on the cell surface: tolerance is developed to cells presenting only self-peptides, and the immune system eliminates cells presenting foreign or dangerous peptides (infected cells, tumor cells). If there are no specific T cells in the T cell repertoire, the antigen-specific immune response is lacking.

1.2. T cell receptors recognize T cell exposed motifs of peptide sequences

In the peptide-HLA-TCR complex, some amino acids of the peptide have a critical role in HLA binding (anchor residues) and are partially hidden from the TCR, while others are mainly responsible for TCR binding (16, 17).

A computational study described that in a nine amino acids long (9-mer) peptide positions 4–8 have a significantly higher number of interactions with TCRs than positions 1–3 and 9 (17). In my analyses, I will focus on the *T cell exposed motif* (TCEM) which can be found between positions 4 to 8 of 9-mers as defined by *Bremel* and *Homan* (*Figure 1*) (18).

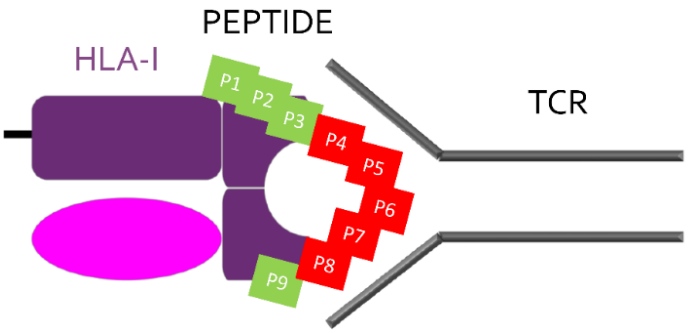


Figure 1. Schematic diagram of the interaction between peptide, HLA-I, and TCR. *T-cell exposed amino acids of a 9-mer are colored red. Amino acids having only a minor role in TCR binding are colored green.*

1.3. The development of the T cell repertoire

The T cell repertoire is formed during fetal development, and it is already established at birth (19). Positive and negative selection processes are key to establish a functionally competent and self-tolerant T cell repertoire. These processes occur in a discrete microenvironment, in the cortex and the medulla of the thymus (15, 20). Although TCRs are generated by a quasi-random process of somatic recombination, the mature T cell repertoire – the subset of all possible TCRs – is far from random, whereas the aim is to develop a T cell population that is effective in fighting pathogens (21). Selection processes are dependent on the presented self-peptides by thymic antigen-presenting cells.

1.4. Positive selection and thymoproteasomal cleavage in cortical thymic epithelial cells

T cell precursors (thymocytes) evolve to double-positive T cells expressing CD4 and CD8 coreceptors as a result of certain intracellular stimuli (e.g. RAG1, RAG2, TCRβ then TCRα expression) (22). During positive selection, these cells become CTLs if they recognize self-pHLA-I complexes on cortical thymic epithelial cells (cTEC) (22). This encounter represents an essential signal for thymocyte survival and differentiation. T cells incapable of binding pHLA-I complexes *die by neglect* (15, 21, 23). It is important to emphasize that the TCR ligand pool is composed of self-pHLAs (15, 21, 23–25).

The unique protein degradation machinery of cTECs was described in the previous two decades (23). cTECs exclusively express a particular proteinase complex, called thymoproteasome (26), which produces unique peptide motifs for the positive selection of CTLs (25, 27). As thymoproteasomes contain β5t subunit in contrast to the constitutive proteasome and immunoproteasome (26), and the pocket of β5t is mostly composed of hydrophilic amino acids, these are less potent in generating peptides with hydrophobic C termini (25, 27). Peptides generated by thymoproteasomes exhibit low affinity to TCRs and induce large fractions of CD8+ cells (25). A recent study published the amino acid preferences of the thymoproteasome around the cleavage site (25). Experiments confirmed that thymoproteasomes are essential for the positive selection of an adequate number of CTLs (26, 28). To note, cathepsin-L and the thymus-specific serine protease in the cTECs are responsible for generating ligands of CD4+ T cell positive selection (15).

1.5. Negative selection

Positively selected T cells migrate into the medulla where they interact with antigen-presenting cells (medullary thymic epithelial cells, dendritic cells) (22). During negative selection (also known as clonal deletion) T cells expressing TCRs that bind self-pHLA ligands with high affinity die by apoptosis (15, 22, 29). It is critical because mature T cells that bind self-pHLAs with high affinity could trigger an autoimmune response. Negative selection mainly takes place in the medulla, but the process already begins around cortical dendritic cells (15). Medullary thymic epithelial cells express many tissue-specific genes generating qualitatively and quantitatively different peptide pools compared to cTECs (30, 31). It is reported that after

negative selection the estimated number of different TCRs is < 10^8 in the human naive T cell pool, which is much less than the count of different peptide motifs (32).

A relevant fraction of autoreactive T cells is transformed to regulatory T cells, which fulfill a key role in the maintenance of tolerance to self-antigens and in preventing autoimmune diseases (33, 34).

1.6. Peripheral tolerance

Although central tolerance is highly efficient, the control of self-reactivity is not yet perfect. The essence of the peripheral tolerance processes is to provide additional control over self-reactivity: self-reactive T cells become functionally unresponsive (anergic) or are deleted after binding self-antigens outside of the thymus (35).

1.7. The binding strength of pHLA-TCR affects the fate of T cells

During positive selection, the recognition of self-pHLA complexes eventuates in survival, while negative selection – mediated also by self-pHLA complexes - can be a death verdict. The classical affinity model could dissolve this paradox. The strength of the interaction between the TCR and self-pHLA complexes is substantial in the further fate of the T cell (23). At least weak interaction is required to survive positive selection, but strong binding results in apoptosis of the cell during negative selection (15, 34, 36) *(Figure 2)*. Consequently, TCRs interacting with weak to medium affinity with self-pHLA-I complexes are the most likely to survive.

On the other hand, recent studies reported a direct positive relationship between the strength of self and foreign pHLA binding by TCRs. Consequently, T cells with higher self-reactivity predominate the acute immune response to pathogens (21, 37). In other words, self-pHLA complexes make it possible that T cells in the repertoire function well enough against pathogens.

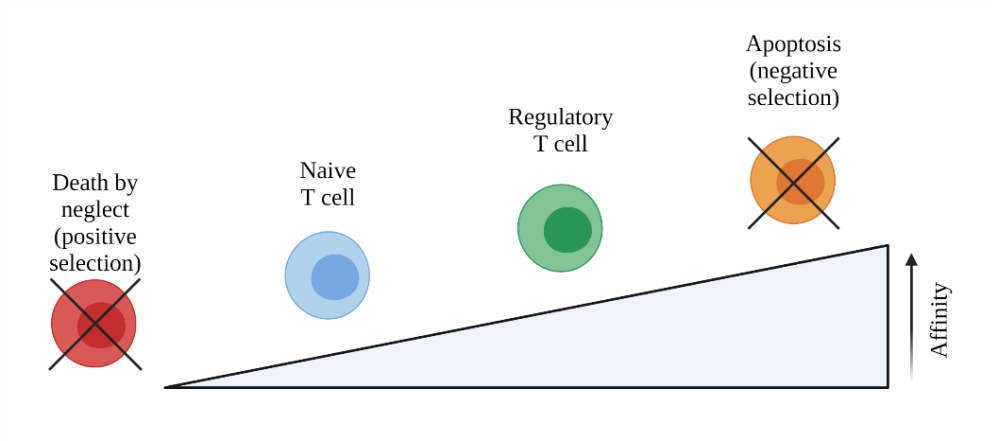


Figure 2. The affinity model of thymocyte selection. *Weak interactions are required to protect thymocytes from death during positive selection. Strong interactions cause negative selection by apoptosis. (Figure based on Klein et al. 2014 Box 1. (15))*

1.8. Immunogenicity of nonself peptides and cross-reactivity of T cells

How can T cells differentiate between self and nonself peptides, dangerous and harmless signs? T-cell responses to a given peptide are influenced by several factors. One of the most important factors is the similarity to self-antigens or commensal antigens. Numerous studies showed that peptides similar to self-antigens have lower immunogenicity, presumably as a result of the negative selection of T cells (38, 39). TCRs are typically able to bind not just one particular peptide but a set of peptides (cross-reactivity of TCRs) (40), which consists of closely related sequences. As T cells are cross-reactive, a significant fraction of nonself peptides is indistinguishable from presented self-peptides and T cells are missing from the repertoire or they are tolerant to these peptides (39). Note, the size of the peptide set (polispecificity) varies between TCRs (21, 41, 42). Moreover, cross-reactivity is not independent of the development of the T cell repertoire: a study showed that thymic negative selection against fewer self-peptides resulted in a more cross-reactive T cell repertoire (43).

At the same time, it is widely accepted that nonself peptides highly dissimilar to human proteins are more immunogenic (44, 45). A study suggested that the immune recognition of peptides is less likely if they are conserved in the commensal microbiome (38). This observation can be explained as a result of peripheral tolerance. Another study showed that microbiota peptide similarity can either enhance or reduce the immunogenicity of peptides (46).

2. AIMS - HYPOTHESIS - RESEARCH QUESTIONS

It has been suggested that as a result of T cell positive selection, both the CD4+ and CD8+ T cell repertoires are skewed to greater self-reactivity, and T cells that bind self-peptides stronger also bind the foreign agonist peptides more effectively (21, 37, 47). In other words, self-peptides mediating positive selection can be considered as a 'test-set' selecting T cells that recognize foreign peptides with higher effectivity. But is there any negative consequence of this mechanism?

Our hypothesis suggests a fundamental side-effect of T cell positive selection on the recognition of nonself peptides.

> As sequences of self-proteins mediate positive selection, a large fraction of nonself peptides is not recognized by the immune system even if T cells are cross-reactive.

As T cell positive selection is mediated by TCEMs of self-peptides, the hypothesis predicts that it is less likely to detect specific T cells in the repertoire for TCEMs that are 1) extremely rare or missing from human proteins, 2) not expressed in cTECs, or 3) not presented on the surface of cTECs.

Additionally, the hypothesis raises several questions. Is it possible that a peptide is overly different from self-proteins and consequently it is not recognized by T cells? Could T cell cross-reactivity compensate for this side-effect of T cell positive selection? Does this phenomenon have any effect on the susceptibility to infections? In my thesis, I aim to confirm the predictions of the hypothesis and answer the questions that arose.

3. METHODS

3.1. In vitro datasets

Our hypothesis was tested on two independent in vitro datasets. HLA-I bound peptides were collected from the Immune Epitope Database (IEDB) (48) which contains the details of nearly 400,000 T cell activation assays. The database is strictly curated and has a standardized decision algorithm to determine whether a given assay is positive or not (49). The final datasets were compiled using strict selection criteria *(Figure 3)* detailed below.

After the first selection steps, the allele-peptide pairs met the following criteria: alleles were determined with high resolution (both the allele group and the specific HLA protein are known (50)), peptides were linear, 9 or 10 amino acids long and contained only the 20 standard amino acids.

HLA binding is the prerequisite of T cell activation. Therefore, the HLA binding of peptides was confirmed by two alternative approaches:

(1) The binding between peptides and alleles was determined using the state-of-the-art bioinformatics tool, *NetMHCpan* (51). This software predicts the binding of peptides to numerous HLA molecules utilizing artificial neural networks. The general guidelines were used to select bound peptides: either the binding affinity had to be lower than 500 nM, or the binding rank percentile had to be lower than 2% (**dataset 1**, left branch on *Figure 3*).

(2) To avoid the bias associated with the computational prediction of HLA binding (52, 53), HLA binding assays that were also collected in IEDB were matched with allele-peptide pairs of activation assays. Pairs that were found in at least two binding assays were retained and the fraction of positive binding assays was more than 60% (**dataset 2**, right branch on *Figure 3*).

Previous works have suggested that the overrepresentation of highly similar sequences due to collection bias in the IEDB could influence the analysis results (54, 55). Consequently, in the case of dataset 2, a highly diverse peptide set was created using a previously established iterative method, which excluded similar peptide sequences from peptides (56). Briefly, the k-tuple distance between all peptide sequences was determined in each iteration using *Clustal Omega* (57). Peptide pair(s) with the lowest distance values were determined and the peptide having the lowest mean distance from all other sequences was excluded. We repeated these

iterations until only peptides with at least 0.5 k-tuple distance from all other sequences remained in the dataset (58). This distance value corresponds to a maximum 50% overlap between sequences. As the result of this step, the filtered peptide set covered the sequence space more homogeneously.

In the case of both datasets, allele-peptide pairs were kept if they occurred in at least two activation assays. Sequences occurring in the human reference proteome (59) were excluded.

Peptides were classified into two groups: in both datasets, allele-peptide pairs with solely negative T cell assays were defined as **nonimmunogenic** and the ones with more positive than negative T cell assays as **immunogenic**. Allele-peptide pairs that did not belong to any of these categories were dropped. Peptide sequences tested for multiple alleles, but with the opposite T cell activation results were also excluded. Finally, to avoid any overlap between the two datasets, peptides found in both datasets were only kept in the second one. After filtering, the number of peptides in datasets 1 and 2 were 3,380 and 635, respectively (*Table 1*).

Table 1. The number of immunogenic and nonimmunogenic peptides in datasets 1 and 2.

	Dataset 1	**Dataset 2**
Immunogenic peptides	1093	360
Nonimmunogenic peptides	2287	275

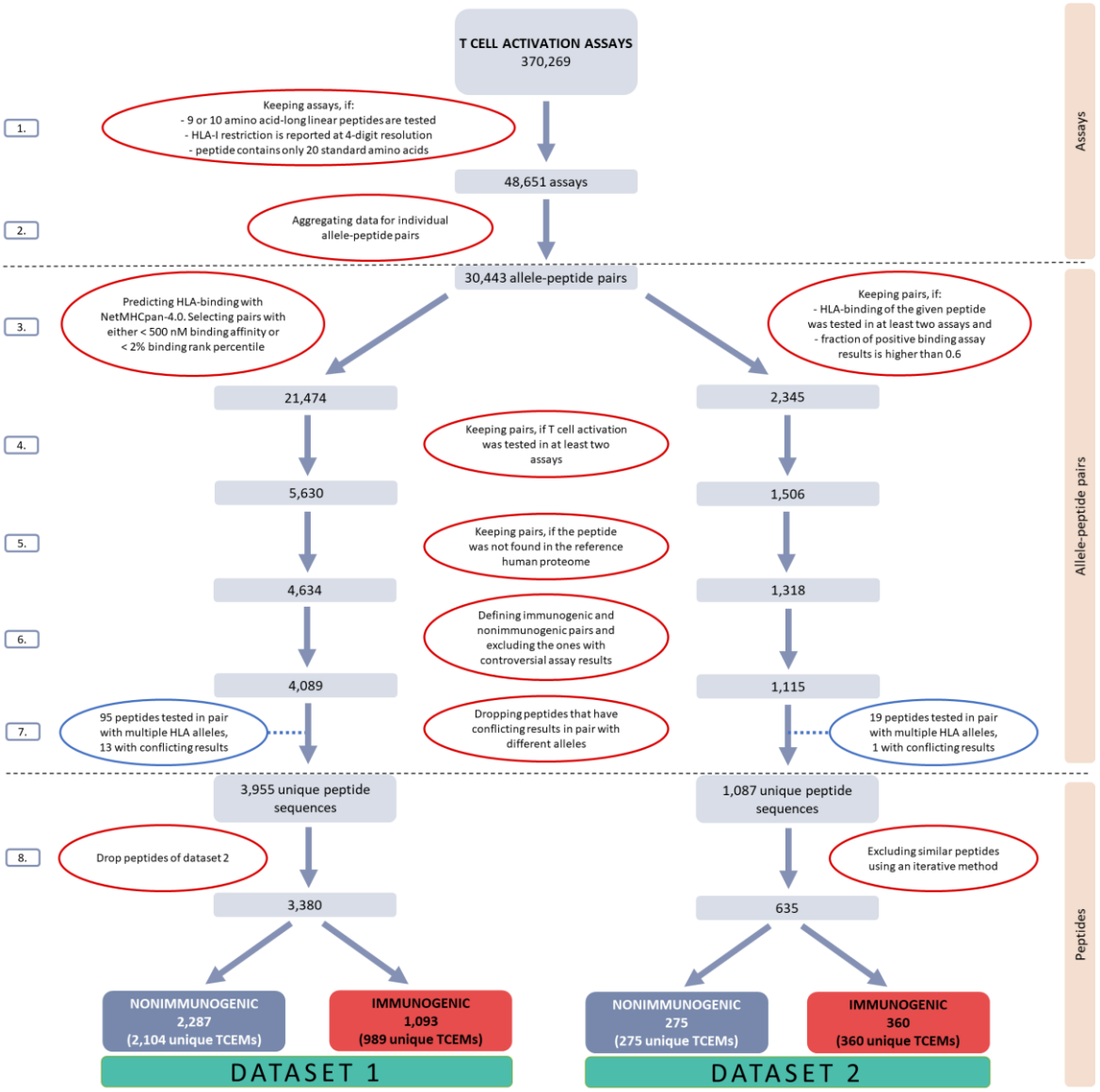


Figure 3. The assembly of peptide sets used throughout our study. *Ellipses indicate filtering criteria at each step. T cell activation data on peptide-HLA pairs were collected from the IEDB and filtered (steps 1 and 2). The HLA binding of peptides was confirmed with computational prediction in the first dataset and with HLA binding assay data in the second dataset (step 3). Allele-peptide pairs, whose binding was not confirmed were discarded. Moreover, allele-peptide pairs were excluded, if they were tested in only one T cell assay (step 4) and/or the peptide was found in the reference human proteome (step 5). Next, the allele-peptide pairs were classified into immunogenic and nonimmunogenic groups, and the pairs having controversial assay results were excluded (step 6). In both datasets, peptides having conflicting results with different alleles were also excluded (step 7). The remaining peptides in dataset 2 were filtered for nonsimilar sequences (step 8, right branch). To avoid overlap, peptides found in both datasets were kept only in dataset 2 (step 8, left branch).*

3.2. TCEM frequency

For each peptide, TCEM was defined as the amino acid sequence from positions 4 through 8 for 9-mers (18) and the amino acid sequence between positions 5 and 9 for 10-mers. The latter definition is based on the fact, that numerous common HLA alleles, like A*02:01, prefer certain amino acids at position 4 (see binding logos in a recently published immunopeptidomics study (60)). Since there are 20 standard amino acids, 20^5 = 3,200,000 different five amino acid-long sequences (pentamers) are possible. **TCEM frequency** in the human proteome was determined for each pentamer as follows. Proteins in the human reference proteome were decomposed into overlapping 9-mers. Sequences between the 4^{th} and the 8^{th} amino acids were determined and the number of occurrences was counted for each possible pentamer. Pentamers that contain selenocysteine were excluded.

3.3. TCEM expression

A recently published study reported the gene expression of human thymic cortical epithelial cells from infants (61). Raw data were downloaded. Columns of the count matrix were scaled using the *calcNormFactors* function in the *edgeR R* library. Next, RPKM values were calculated using the *rpkm* function of *edgeR* and exon length data of the *GenomicFeatures R* library. The median RPKM value in cTEC samples was determined for each gene. Matching of ENSEMBL gene IDs (used in the expression dataset) with UniProt IDs was unsatisfactory, as 40% of UniProt protein IDs in the dataset did not have corresponding ENSEMBL gene IDs in the downloaded expression set. Consequently, ENSEMBL gene IDs and UniProt IDs were first converted to HUGO IDs using the *org.Hs.eg.db R* library and protein information in the UniProt database, respectively. Next, the proteins and genes were matched using HUGO IDs. With this approach, the expression of encoding genes could be determined for more than 90% of proteins. To assign an expression value to a TCEM, the proteins containing a given TCEM were collected. The median expression of genes encoding these proteins was calculated to approximate the chance for a given TCEM being expressed in cTECs (**TCEM expression**). If a given TCEM was found multiple times in the same protein, the expression of the encoding gene was included the same number of times in the calculation. TCEMs encoded by housekeeping genes were determined using data from a recent study (62).

3.4. Thymoproteasomal cleavage score

A previous study reported the amino acid prevalence around the cleavage site of the thymo- and immunoproteasome (25). Briefly, the authors carried out thymo- and immunoproteasomal digestion of three proteins and determined the fraction of amino acids found on five positions towards the C and N-termini around the cleavage site. They also provided amino acid frequencies in the three proteins, which would be found if they were randomly cleaved. We developed a score, which estimated the probability of cleavage between two amino acids at a given amino acid environment. We first normalized amino acid prevalence values at each position around the cleavage site by dividing them with their prevalence in the substrates yielding amino acid preference scores: $c_{i,j}$ referring to the score of amino acid j at position i (between -5 and 5) around the cleavage site (*Figure 4 A*). Next, we determined the probability of proteasomal cleavage (*C*) at each site of the human proteome by calculating the median of $c_{i,j}$ values at positions around the cleavage site (*Figure 4 B*). We approximated the probability of peptide formation upon proteasomal cleavage by implementing cleavage scores. Specifically, for each 9-mer peptide in the human proteome, we determined the probability of peptide formation upon proteasomal cleavage by calculating the mean of *C* values before the N-terminal (C_N) and the C-terminal (C_C) of the 9-mer yielding $\bar{C}$ (*Figure 4 C*). Then for each pentamer, we calculated the median of $\bar{C}$ values of all 9-mers that contain the given pentamer in the TCEM region yielding the **thymoproteasomal cleavage score** (*Figure 4 D*).

We calculated the immunoproteasomal cleavage score using the same approach and considering the amino acid preference of the immunoproteasome. This score served as a control in the analysis.

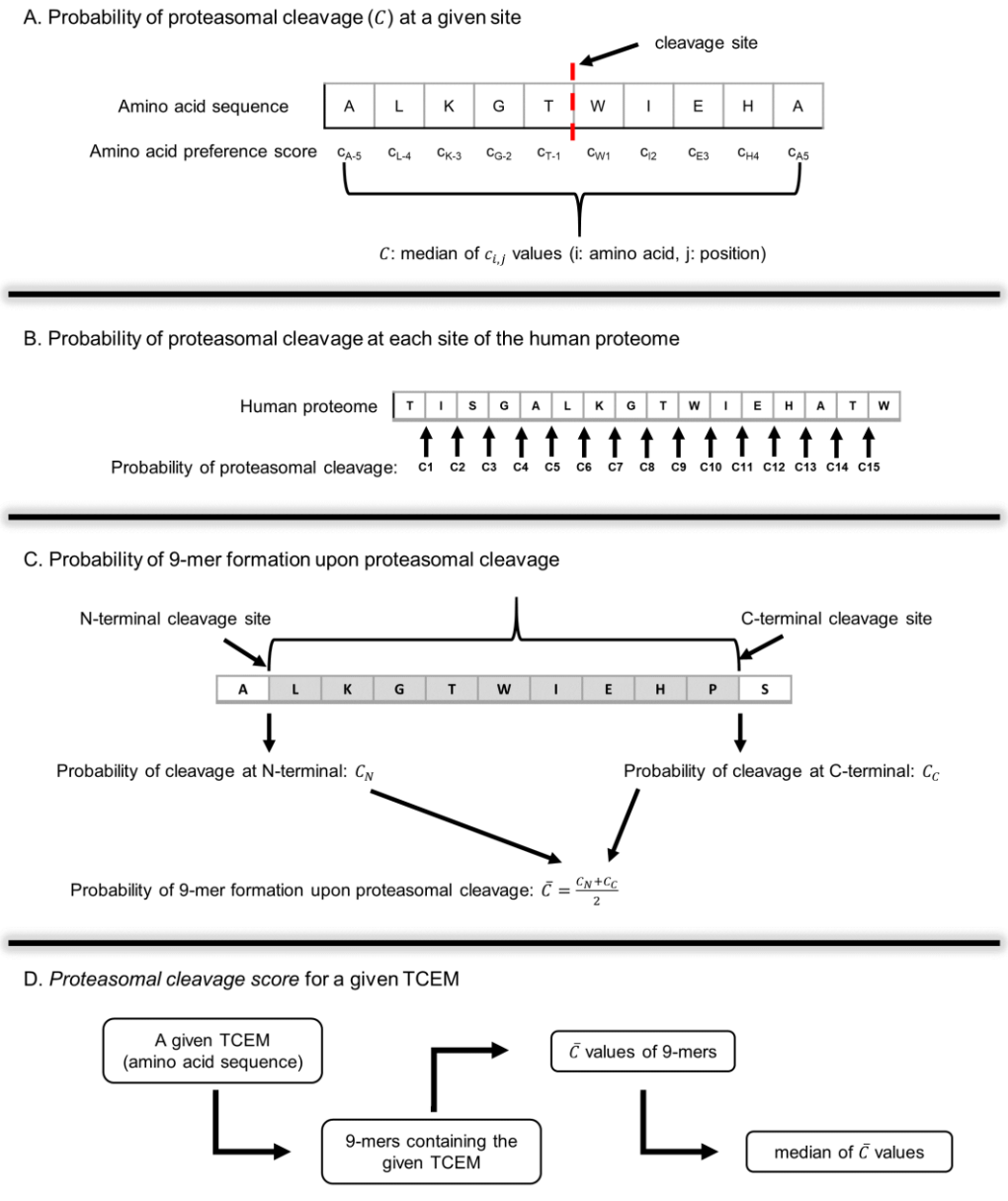


Figure 4. Calculation of proteasomal cleavage score. A) *Amino acid preference values around the proteasomal cleavage site were calculated using data from a previous study (25). The probability of cleavage at a given site of a protein sequence was estimated by calculating the median of preference values associated with the amino acids that were found at the five positions towards the C and N-termini.* ***B)*** *The probability of proteasomal cleavage (C) was calculated at each site of the human proteome.* ***C)*** *For each 9-mer in the human proteome, the C values were averaged before the N- and after the C-terminal amino acids to estimate the probability of peptide formation upon proteasomal cleavage ($\bar{C}$).* ***D)*** *Finally, for each TCEM, the median of $\bar{C}$ values associated with the peptides that include the given TCEM were calculated.*

3.5. Sequence similarity score

Sequence similarity score was calculated using an established method (46). The similarity between a given peptide and the most similar sequence in the human reference proteome was estimated using the BLOSUM62 substitution matrix. First, the most similar peptides in the human proteome for a given peptide were found with *BLAST* software, and then the similarity score was calculated between each pair using an established formula (38).

3.6. SARS-CoV-2 specific T cells in the repertoire

Multiplex Identification of T-cell Receptor Antigen Specificity (MIRA) data on 27 healthy individuals were acquired from the website of Adaptive Biotechnologies (63). The authors co-cultured naive CD8+ T cells of healthy donors with dendritic cells, loaded with a pool of examined peptides of SARS-CoV-2. Then, they used the MIRA technology to identify antigen-specific T cells. The MIRA technology combines conventional T cell assays with immune repertoire sequencing to identify a large number of antigen-specific T cells in the repertoire simultaneously (64).

First, HLA-I allele-peptide pairs were determined, for which carrying of a given allele could potentially be associated with the prevalence of specific naive CD8+ T cells in the repertoire. Specifically, the binding of each examined peptide by the HLA alleles carried by any individuals was predicted using *NetMHCpan* software (51). The prevalence of peptide-specific T cells was associated with carrying a given HLA allele, if the predicted values suggested strong binding (i.e., affinity was lower than 50 nM and rank percentile was under 0.5%) and peptide-specific T cells were found in at least two individuals carrying the given allele. These rigorous criteria for HLA binding were used to decrease false positive hits. For each patient, the expected peptides with specific T cells in the repertoire were determined considering the previously specified peptide-allele pairs and the HLA genotype of the individual. Participants with specific T cells found for at least 20 SARS-CoV-2 peptides were included. This filtering step yielded data on 22 individuals for further analysis.

3.7. The level of T cell cross-reactivity

Data on the binding strength of T cells were reported in two studies (65, 66). Each study examined the shift in peptide-binding by TCRs when sequentially changing amino acids at each peptide position. The analysis was narrowed down to the TCEM sequence, and the BLOSUM62 similarity was determined between the TCEM of the original and the modified peptides as described previously (38). ROC curves and ROC AUC values were determined using the *ROCR R* library. The optimal cutoff was estimated by implementing an established cost-benefit method (67). In the case of the NY-ESO-1 epitope and TCR C^{259}, the level of TCR binding strength was determined, under which T cell activation is negligible. To identify this value, T cell activation data of the sequentially modified NY-ESO-1 epitopes (reported in the same study) were used (65). Lower than 10% of original TCR binding strength was selected as insufficient binding because the median level of T cell activation by peptides under this cutoff was only 7.9% of the original peptide's T cell activating ability.

3.8. Determining bound peptides and np-TCEMs in the proteomes of intracellular pathogens

The reference proteomes of 50 well-known intracellular pathogens were downloaded from the UniProt database (59). First, the TCEMs of each 9-mer in the proteome of each pathogen and their prevalence in the human proteome, expression in cTECs, and the probability of proteasomal cleavage were determined as previously described. np-TCEMs were defined as the ones found less than 4 times in the human proteome or have low expression in cTECs or low probability of thymoproteasomal cleavage. Then, the binding of each 9-mer to common HLA alleles was predicted with *NetMHCpan* (51). HLA-A and B alleles listed in a reference set with maximal population coverage were used (68). As the list did not include data for HLA-C, the first four-digit allele from each two-digit HLA-C allele class was selected. To decrease the prevalence of false-positive binding results, 9-mers bound strongly were identified (i.e., rank percentile value was under 0.5% and the binding affinity value was under 50 nM). To alleles, for which the prediction could not detect any bound peptides in the proteome of the pathogen, *N* peptides with the lowest predicted binding affinity values were assigned, in which *N* refers to the median number of peptides bound by other alleles at the same loci. For each allele-species pair, the fraction of np-TCEMs in bound peptides was calculated.

3.9. HLA association data

To identify HLA allele associations with infectious diseases, a literature mining was carried out. We focused on meta-analyses to collect highly reliable HLA associations. We searched PubMed with the "hla infection meta analysis" and "hla association meta analysis" keywords. HLA association meta-analysis studies were found for hepatitis B (69), hepatitis C (70), dengue virus (68), and human papillomavirus (71). In the case of hepatitis B, C, and HPV studies, significant associations between HLA allele groups and infections or response to treatment were selected. In the case of the meta-analysis for dengue infection, P-values were not determined by the authors. They ranked associations of different HLA allele groups with the infection along 17 studies and considered the allele with the best rank as protective ones. We followed the method of the authors but only considered those allele groups that were included in at least 75% of studies to increase the reliability of the analysis. After calculating the rank percentile of odds ratio (OR) values associated with the allele groups in each study, the mean rank percentile of each allele group was calculated. The group with the lowest rank percentile was associated with protection and the group with the highest rank percentile was associated with susceptibility. As the results of all studies were published for allele groups or serotypes and not individual alleles, the fraction of np-TCEMs in presented peptides was calculated as follows. For serotypes, the mean values of alleles belonging to the given serotype were calculated. In the case of allele groups (i.e., associations published for two digits resolution), the mean values of alleles were calculated, which are marked as common in the Common and Well-Documented Alleles Catalog (72).

3.10. Statistical analysis and visualization

For statistical analyses, *R* (version 3.6.3) was used in the *RStudio* (version 1.2.5033) environment. The *ggplot2, ggpubr, grid, gridExtra, ggsci, scales, ComplexHeatmap, ggrepel,* and *png R* libraries were used for visualization. Smooth curves on plots were fitted with cubic smoothing spline method (73). *Figure 14 A and Figure 16* were created with BioRender.com.

A typical boxplot in figures includes a horizontal line within the box (median data value), a box (encompasses half of the data values; lower limit and upper limit of the box indicates the 1st and 3rd quartile, respectively), vertical lines indicate 1st quartile - 1.5 x IQR and 3rd quartile + 1.5 x IQR.

4. RESULTS

4.1. Analysing the effect of T cell positive selection on peptide immunogenicity

The three predictions of the hypothesis were examined in parallel on two nonoverlapping in vitro T cell activation datasets. Dataset 1 contained a high number of peptides, while dataset 2 ensured that the findings are not confounded by computational prediction or the presence of similar sequences. In dataset 2, the diversity of peptides was high and showed a similar distribution in immunogenic and nonimmunogenic groups *(Figure 5)*.

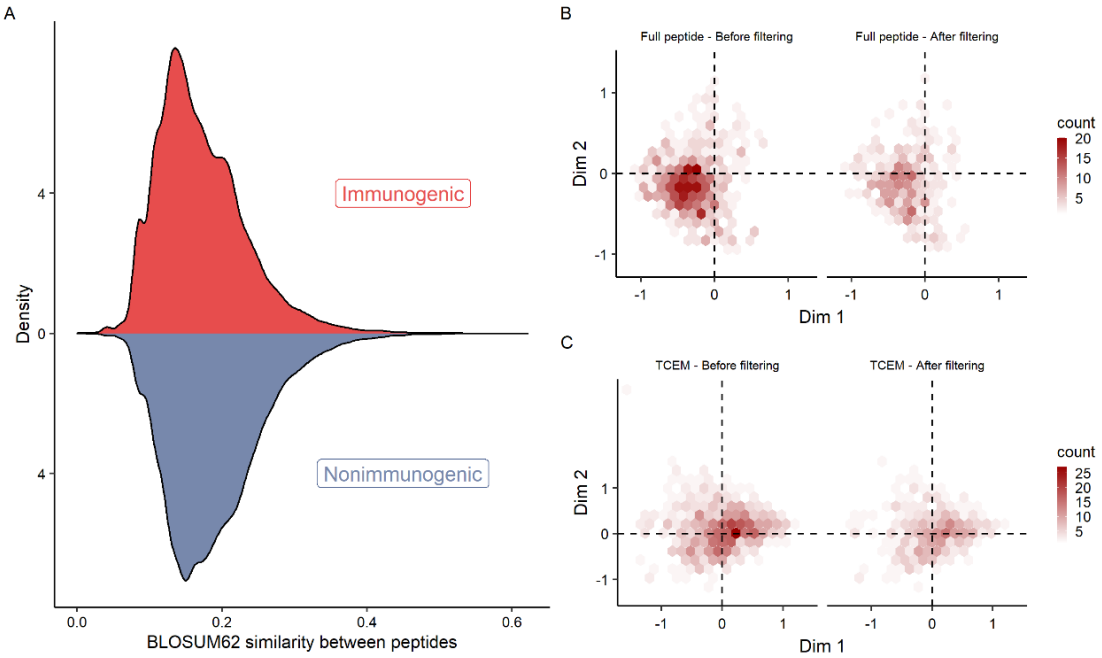


Figure 5. Diversity of peptide sequences. A) *Plots indicate the density of BLOSUM62 similarity values (n = 64,620 and 37,675 for immunogenic and nonimmunogenic sequences, respectively) between all pairs of immunogenic and nonimmunogenic peptides in dataset 2. Both groups contain highly diverse and dissimilar sequences. BLOSUM62 similarity values between peptide pairs were calculated with the protr R library.* ***B-C)*** *Peptide sequences of dataset 2 cover the sequence space more homogeneously after excluding similar sequences (n = 853 and 525 before and after filtering). Multiple correspondence analysis (MCA) was carried out on 9 amino acid-long peptide sequences as follows. Each position of the peptide (B) or its TCEM region (C) was treated as a categorical variable having the 20 amino acids as possible categories. The distribution of peptides (B) or TCEMs (C) in sequence space is shown on MCA biplots. The number of sequences in equal-size hexagons is shown color-coded. The distribution of sequences was less heterogeneous after similarity reduction (B and C).*

To confirm our predictions, the following attributes of peptides of datasets 1 and 2 were determined (*Methods*):

i. TCEM frequency in human proteins.

ii. Median expression of genes in cTECs that encode proteins containing a given TCEM.

iii. Thymoproteasomal and immunoproteasomal cleavage score of TCEMs.

4.1.1. TCEM frequency in the human proteome and peptide immunogenicity

The distribution of TCEM frequency in the human proteome had a long right tail: a considerable fraction of motifs was rarely or not found in the human proteome, but some reached very high frequencies *(Figure 6).*

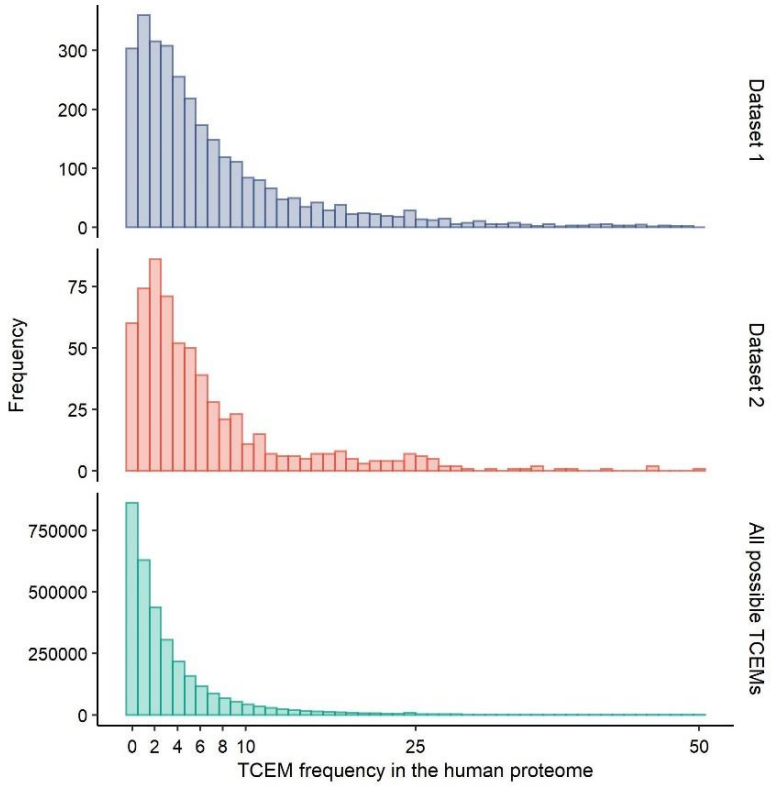


Figure 6. The distribution of TCEM frequency in the human proteome. *The histograms indicate the number of times TCEMs were found in the human proteome (n = 3,194,577, 3,031 and 630 for all possible TCEMs and motifs in datasets 1 and 2, respectively). Note, that only TCEM sequences occurring less than 51 times in the human proteome are shown on the plot for visualization purposes.*

Our hypothesis predicted the following: if a TCEM is very rare in the human proteome, specific T cells will unlikely survive the positive selection around cTECs, and they will be potentially missing from the T cell repertoire. Consequently, motifs very rarely or not found in the human proteome are less likely to be immunogenic. Indeed, the TCEM frequency of immunogenic and nonimmunogenic peptides was significantly different in both datasets: immunogenic peptides contained more frequent TCEMs *(Figure 7 A)*. Accordingly, if peptides were classified into two groups based on their TCEM frequencies (cutoff = 4), immunogenic peptides were more likely found in the group of relatively frequent TCEMs *(Figure 7 B)*. The result suggests that an appropriate occurrence of TCEMs in human proteins is needed for immunogenicity.

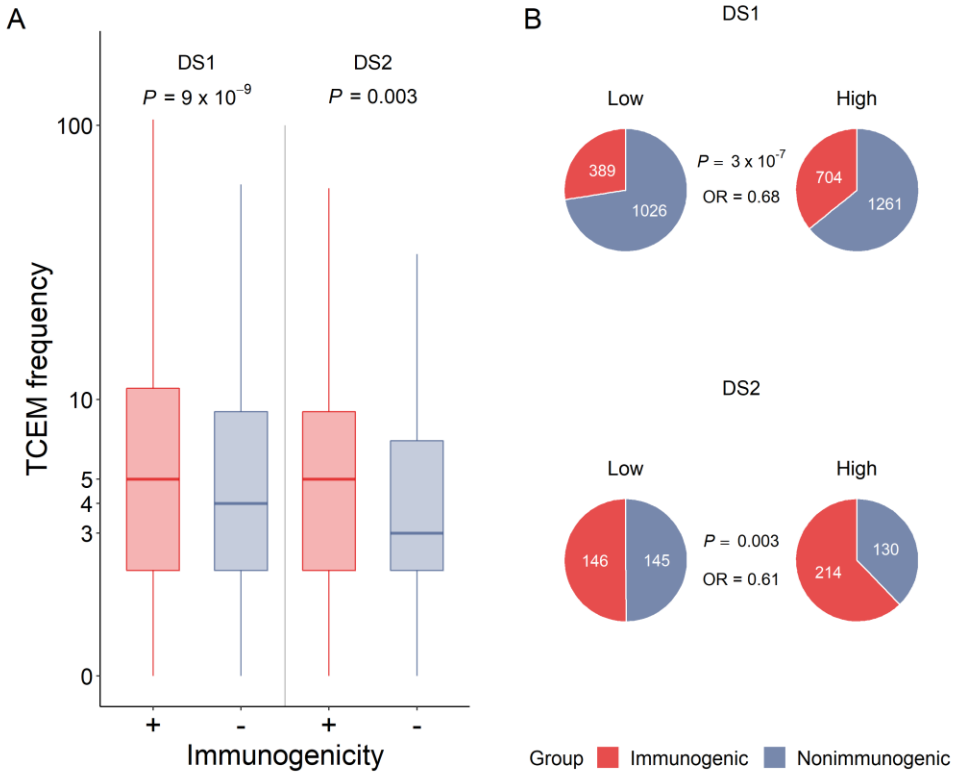


Figure 7. Peptide immunogenicity is influenced by TCEM frequency in human proteins. A) *The plot indicates the number of times immunogenic (+, n = 1093 and 360 in datasets 1 and 2, respectively) and nonimmunogenic (-, n = 2287 and 275 in datasets 1 and 2, respectively) TCEMs found in human proteins. In both datasets, TCEMs of immunogenic peptides were found more times in human proteins than TCEMs of nonimmunogenic ones. Outliers are not shown for visualization purposes. The P-values of two-sided Wilcoxon's rank-sum tests are indicated.* ***B)*** *TCEMs found more than 3 times in the human proteome (group 'High' on the plot) were more likely to activate T cells. The ORs and P-values of two-sided Fisher's exact tests are shown. DS1: dataset 1, DS2: dataset 2.*

4.1.2. TCEMs not expressed in cTECs are less immunogenic

In the subsequent analyses, the focus was on motifs that occurred at least once in the human proteome. It was reported that the HLA-I presentation of peptides is highly dependent on the expression of the encoding gene (14).

Our hypothesis predicts that TCEMs encoded by genes having low or undetectable expression in cTECs cannot mediate the positive selection of specific T cells. At the same time, the immune response is not expected to TCEMs that are encoded by abundantly expressed housekeeping genes, because the response to these TCEMs may be blocked by central or peripheral immune tolerance mechanisms (74, 75).

To examine the potentially biphasic relationship between TCEM expression and T cell activation, the probability for a TCEM of being immunogenic was plotted as a function of its expression using lowess smoothing *(Figure 8 A)*. The distribution density of TCEM expression in the immunogenic and nonimmunogenic peptide groups was also examined separately *(Figure 8 B)*. In line with our expectation, TCEMs having either low or high expression in cTECs were similarly less likely to activate T cells than the ones in the medium expression group *(Figure 8 A-B)*.

Indeed, if the peptides were classified into low (bottom 15 percentile), medium (15-75 percentiles), and high (upper 75 percentile) expression groups, immunogenic peptides were less likely found in both low and high expression groups *(Figure 8 C)*. These results suggest missing T cell responses to TCEMs not expressed in cTECs, at the site of positive selection.

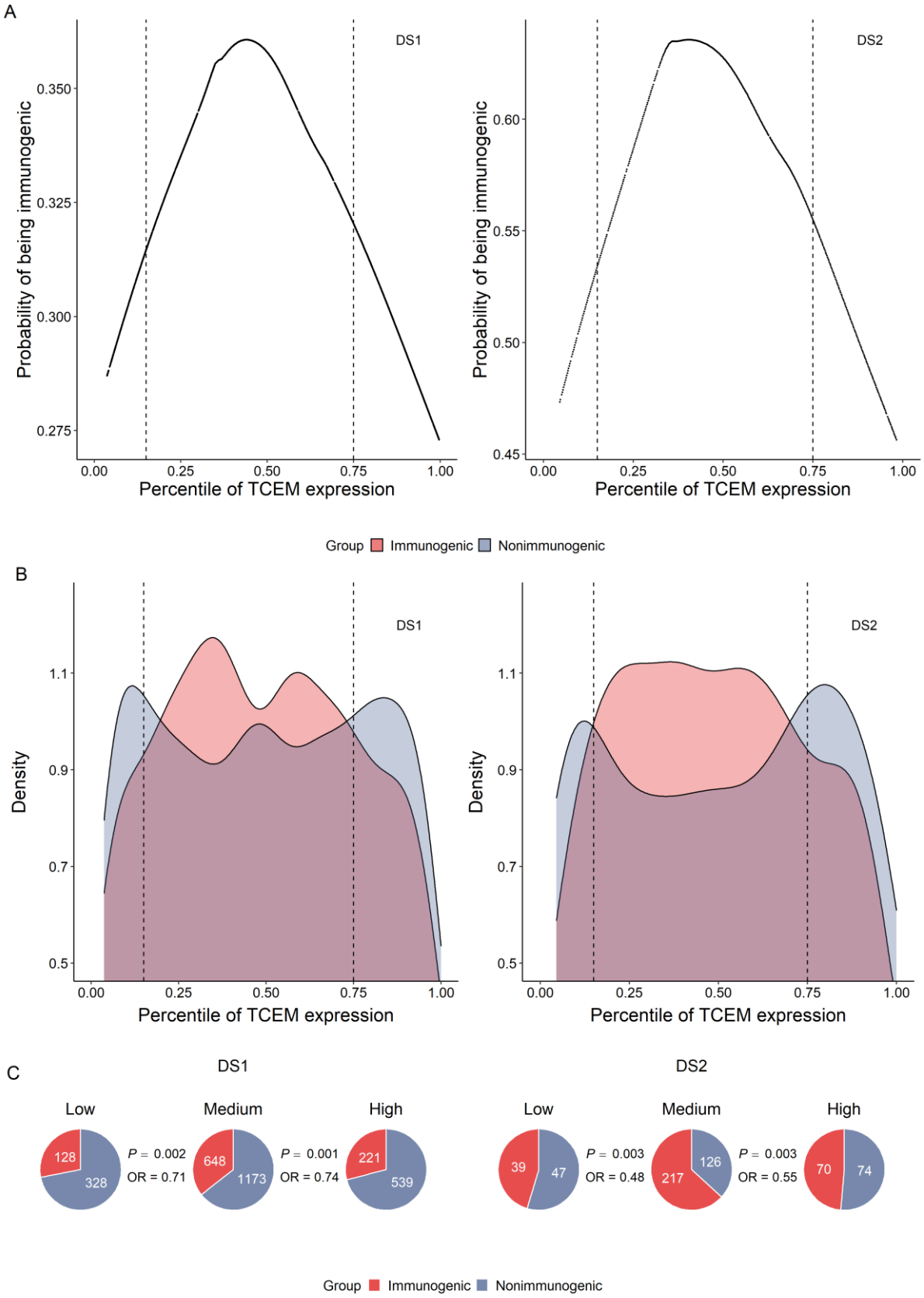

A
DS1
DS2
Probability of being immunogenic
Percentile of TCEM expression
Group Immunogenic Nonimmunogenic
B
Density
C
Low
Medium
High
P = 0.002
OR = 0.71
P = 0.001
OR = 0.74
P = 0.003
OR = 0.48
P = 0.003
OR = 0.55
128
328
648
1173
221
539
39
47
217
126
70
74

Figure 8. Peptide immunogenicity is influenced by TCEM expression in cTECs. A) *The plots show the probability of a TCEM being immunogenic as the function of its expression in cTECs. The curves were fitted using lowess regression.* ***B)*** *The plots indicate the probability density of the expression of immunogenic (n = 997 and 326 for datasets 1 and 2, respectively) and nonimmunogenic (n = 2040 and 247 for datasets 1 and 2, respectively) TCEMs. For visualization purposes, gene expression values were transformed by calculating their percentile rank. Vertical dashed lines indicate cutoff values used for OR calculation in panel C.* ***C)*** *Peptides were classified based on their TCEM's expression in cTECs into "Low" (bottom 15 percentile), "Medium" (15-75 percentile), and "High" (upper 75 percentile) groups. Immunogenic peptides were less likely found in both "Low" and "High" expression groups. The ORs and P-values of two-sided Fisher's exact tests are shown. DS1: dataset 1, DS2: dataset 2.*

Next, the association between the prevalence of TCEM-encoding housekeeping genes and gene expression in cTECs was examined. As expected, TCEMs in the high expression group were more likely to be found in proteins encoded by housekeeping genes *(Figure 9)*.

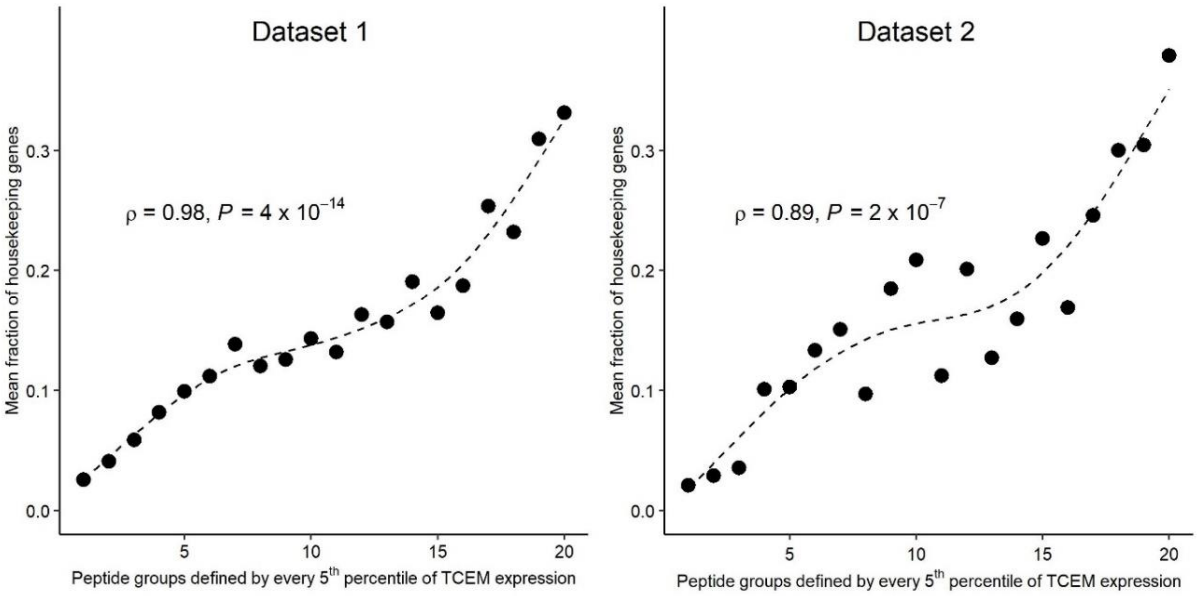


Figure 9. The prevalence of TCEM-encoding housekeeping genes in different TCEM expression groups. *Peptides were classified into twenty groups with increasing TCEM expression in cTECs. Highly expressed TCEMs in cTECs are more likely to be encoded by housekeeping genes. For each TCEM, the genes encoding their sequence were collected. Then, the relative fraction of housekeeping genes was determined among them. The mean of these values is indicated for each TCEM expression group. The dashed lines indicate smooth curve fitted using cubic smoothing spline method in R (Methods).*

4.1.3. TCEMs that are unlikely to be presented on cTECs are less immunogenic

In cTECs, a specific proteasome called the thymoproteasome generates most peptides from intracellular proteins for T cell positive selection (25). If a peptide is generated with a low probability after thymoproteasomal cleavage, it has a little chance to be presented on the cell surface in complex with HLA molecules even if it has a high expression in the cell. Consequently, lower immunogenicity was expected for TCEMs that are less likely to be generated after thymoproteasomal cleavage. At the same time, no effect of immunoproteasomal cleavage was expected on immunogenicity, because it had only minor importance in cTECs (25). The so-called immuno- and thymoproteasomal cleavage scores were calculated to approximate the chance of TCEM formation in the cell after proteasomal cleavage. In line with expectations, TCEMs of immunogenic peptides were more likely to be generated by thymoproteasomal cleavage than nonimmunogenic ones *(Figure 10 A)*, while immunoproteasomal cleavage did not affect immunogenicity *(Figure 10 B)*. Accordingly, immunogenic peptides were more likely found in the group containing TCEMs with high thymoproteasomal cleavage score (i.e., score higher than the 25^{th} percentile, *Figure 10 C*), while this is not the case for immunoproteasomal cleavage *(Figure 10 D)*.

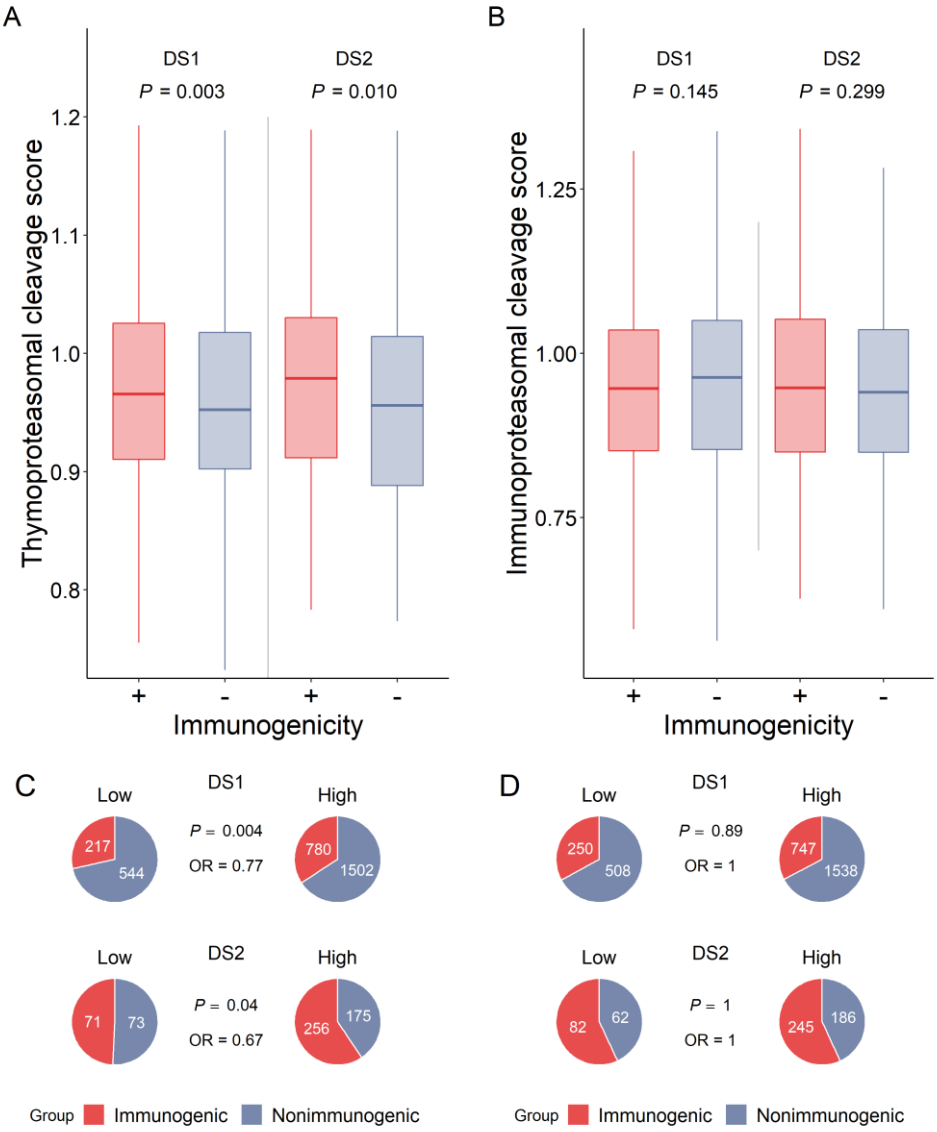


Figure 10. Peptide immunogenicity is influenced by TCEM presentation on cTECs. A-B) *The scores representing the likelihood of TCEM formation after thymoproteasomal (A) and immunoproteasomal (B) cleavage are presented on the vertical axes. TCEMs of immunogenic peptides were more likely to be generated and presented after thymoproteasomal, but not immunoproteasomal cleavage (n = 997 and 327 for immunogenic and 2,046 and 248 for nonimmunogenic TCEMs in datasets 1 and 2, respectively). Outliers are not shown for visualization purposes. The P-values of two-sided Wilcoxon's rank-sum tests are indicated.* ***C-D)*** *Peptides were classified based on their thymoproteasomal/immunoproteasomal cleavage score into a "Low" and a "High" group (cutoff = 25th percentile). Immunogenic peptides were less likely found in groups with a low thymoproteasomal cleavage score (C). Immunoproteasomal cleavage did not affect immunogenicity (D). The ORs and P-values of two-sided Fisher's exact tests are shown. DS1: dataset 1, DS2: dataset 2.*

4.1.4. The robustness of results

Three lines of evidence were reported suggesting that the self-mediated positive selection of T cells results in a defective T cell repertoire with implications on the recognition of nonself peptides. First, TCEMs that are very rare or not found in human proteins are less likely to be immunogenic *(Figure 7)*. Second, the scarce expression of TCEMs in cTECs is also associated with lower immunogenicity *(Figure 8)*. Third, TCEMs that are improbably generated by the cTEC-specific thymoproteasome are less likely to be immunogenic *(Figure 10)*.

These effects on immunogenicity are held in multivariate logistic regression models indicating that they are not confounded by and independent of each other *(Figure 11, Table 2)*.

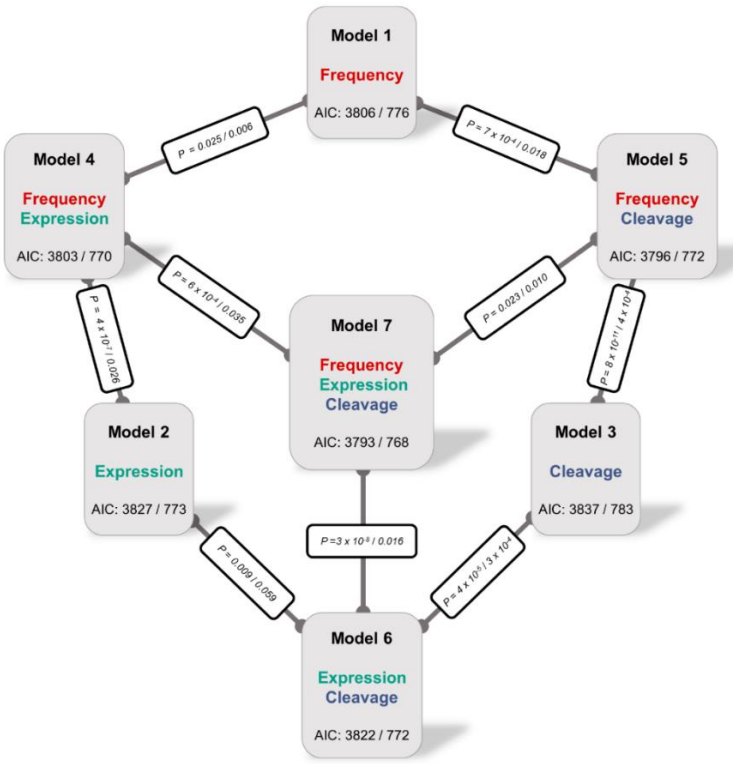


Figure 11. The effects of TCEM frequency, expression and thymoproteasomal cleavage score are not confounded by and independent of each other. *Univariate, bivariate and trivariate logistic regression models were constructed to examine the effect of each variable on T cell activation. Models were compared with ANOVA and the two-sided P-values for datasets 1 (left) and 2 (right) are shown on arrows. P-values lower than 0.05 indicate that the more complex model fits better than the simpler ones (See also Table 2). Akaike information criterion (AIC) values are shown for datasets 1 (left) and 2 (right). All bivariate models fitted significantly better than univariate ones. Additionally, the trivariate model fitted significantly better than the bivariate ones. For detailed data on models, see Table 2.*

Table 2. A detailed description of logistic regression models in Figure 11. *The coefficients and the two-sided P-values of Z statistics are indicated for each independent variable. AIC: Akaike information criterion.*

Dataset	Model	TCEM frequency coefficient	TCEM frequency P-value	TCEM expression coefficient	TCEM expression P-value	TCEM thymoprot. cleavage score coefficient	TCEM thymoprot. cleavage score P-value	AIC
Dataset 1	model 1 (univariate)	2.31	2×10^{-9}					3806
	model 2 (univariate)			0.32	8×10^{-5}			3827
	model 3 (univariate)					1.02	0.017	3837
	model 4 (bivariate)	0.20	4×10^{-7}	0.19	0.025			3803
	model 5 (bivariate)	0.25	1×10^{-10}			1.50	7×10^{-4}	3796
	model 6 (bivariate)			0.33	5×10^{-5}	1.13	0.009	3822
	model 7 (trivariate)	0.23	4×10^{-8}	0.19	0.023	1.51	6×10^{-4}	3793
Dataset 2	model 1 (univariate)	0.29	9×10^{-4}					776
	model 2 (univariate)			0.65	2×10^{-4}			773
	model 3 (univariate)					1.96	0.035	783
	model 4 (bivariate)	0.21	0.027	0.51	0.006			770
	model 5 (bivariate)	0.31	5×10^{-4}			2.18	0.020	772
	model 6 (bivariate)			0.63	3×10^{-4}	1.76	0.061	772
	model 7 (trivariate)	0.23	0.017	0.47	0.010	1.97	0.037	768

Additionally, the effect of these attributes was additive: rare TCEMs having low expression in cTECs and low thymoproteasomal cleavage score were less likely to be immunogenic than TCEMs having only one or two of these attributes (*Table 3*).

Table 3. The effect of TCEM attributes on immunogenicity is additive. *In both datasets, rare TCEMs having low expression in cTECs and low thymoproteasomal cleavage score are associated with much lower immunogenicity than TCEMs explained by one or two of the three attributes. OR: Odds ratio (immunogenic vs. nonimmunogenic) in the examined TCEM group. Two-sided P-values of Fisher's exact tests are shown.*

Dataset	TCEM frequency < 4	TCEM expression < 15%	TCEM thymoproteasomal cleavage score < 25%	OR	P-value
Dataset 1	•			0.68	3×10^{-7}
		•		0.71	0.002
			•	0.77	0.004
	•	•		0.44	2×10^{-6}
	•		•	0.64	0.003
		•	•	0.50	0.002
	•	•	•	0.31	**0.002**
Dataset 2	•			0.61	0.003
		•		0.48	0.003
			•	0.67	0.041
	•	•		0.43	0.009
	•		•	0.47	0.007
		•	•	0.33	0.018
	•	•	•	0.26	**0.027**

Next, we aimed to exclude the possibility that our findings may be confounded by a single amino acid with a peculiar effect on immunogenicity. To test it, the prevalence of the twenty amino acids in immunogenic and nonimmunogenic TCEMs was examined. The most significant difference was found for tyrosine and phenylalanine enriched in nonimmunogenic motifs, and glycine and alanine enriched in immunogenic ones (*Table 4, 3rd column*). This is in line with expectation as the former amino acids are rare, while the latter ones are commonly found in human proteins (*Table 4, 2nd column*). Surprisingly, tryptophan, the rarest amino acid was more common in immunogenic TCEMs, which can be explained by its major role in peptide immunogenicity (76–78). Reassuringly, this phenomenon did not affect our results: the main analysis was iteratively repeated by excluding TCEMs containing certain amino acids, and all findings remained significant (*Table 4*).

Table 4. The results remain when TCEMs containing certain amino acids are excluded from the analysis. *In the 3rd, 7th, 8th, 9th and 10th columns, OR values and two-sided P-values of Fisher's exact tests are indicated. In the 6th column. two-sided P-values of Wilcoxon's rank-sum tests are shown.*

Full name and one letter code of amino acids		Prevalence in human proteins*	Fisher's exact test OR (P) - Amino acid prevalence in IMM vs. NIMM TCEMs	Median TCEM freq. in IMM	Median TCEM freq. in NIMM	Wilcoxon's rank-sum test P - TCEM freq. in IMM vs. NIMM (*Figure 7 A*)	Fisher's exact test OR (P) - low vs. medium TCEM expression (*Figure 8 C*)	Fisher's exact test OR (P) - high vs. medium TCEM expression (*Figure 8 C*)	Fisher's exact test OR (P) - low vs. high thymoprot. cleavage score (*Figure 10 C*)	Fisher's exact test OR (P) - low vs. high immunoprot. cleavage score (*Figure 10 D*)
Tyrosine	Y	0.027	0.46 (4×10^{-28})	6	5	8×10^{-4}	0.75 (0.011)	0.73 (6×10^{-4})	0.70 (5×10^{-5})	1.08 (0.400)
Alanine	A	0.070	1.69 (5×10^{-19})	5	4	0.001	0.66 (9×10^{-4})	0.81 (0.036)	0.79 (0.017)	0.88 (0.181)
Glycine	G	0.066	1.77 (2×10^{-18})	5	4	9×10^{-6}	0.63 (1×10^{-4})	0.75 (0.003)	0.74 (0.001)	0.92 (0.373)
Phenylalanine	F	0.036	0.63 (5×10^{-15})	6	5	1×10^{-6}	0.68 (0.001)	0.73 (0.001)	0.81 (0.024)	1.03 (0.778)
Isoleucine	I	0.043	0.65 (1×10^{-14})	6	4	7×10^{-9}	0.67 (0.001)	0.66 (4×10^{-5})	0.82 (0.040)	1.01 (0.961)
Glutamic acid	E	0.071	1.61 (2×10^{-10})	5	4	1×10^{-7}	0.72 (0.004)	0.69 (9×10^{-5})	0.74 (9×10^{-4})	1.08 (0.374)
Proline	P	0.063	1.52 (6×10^{-10})	5	4	1×10^{-7}	0.66 (4×10^{-4})	0.71 (2×10^{-4})	0.79 (0.010)	1.06 (0.555)
Tryptophan	W	0.012	1.95 (6×10^{-9})	6	4	1×10^{-14}	0.64 (2×10^{-5})	0.70 (5×10^{-5})	0.74 (5×10^{-4})	1.06 (0.458)
Serine	S	0.083	0.71 (6×10^{-9})	4	3	6×10^{-9}	0.66 (8×10^{-4})	0.67 (6×10^{-5})	0.83 (0.055)	1.05 (0.597)
Lysine	K	0.057	0.64 (4×10^{-8})	5	4	4×10^{-8}	0.65 (1×10^{-4})	0.73 (6×10^{-4})	0.79 (0.008)	0.99 (0.896)
Valine	V	0.060	1.25 (2×10^{-4})	5	4	2×10^{-6}	0.70 (0.003)	0.76 (0.006)	0.82 (0.045)	1.04 (0.703)
Asparagine	N	0.036	0.82 (0.004)	5	4	2×10^{-7}	0.63 (6×10^{-5})	0.75 (0.002)	0.71 (2×10^{-4})	1.00 (1.000)
Threonine	T	0.054	1.19 (0.004)	5	4	2×10^{-11}	0.77 (0.029)	0.64 (1×10^{-5})	0.65 (2×10^{-5})	1.02 (0.886)
Glutamine	Q	0.048	1.28 (0.013)	5	4	3×10^{-8}	0.69 (5×10^{-4})	0.72 (2×10^{-4})	0.78 (0.005)	1.05 (0.584)
Aspartic acid	D	0.047	1.18 (0.031)	5	4	2×10^{-6}	0.76 (0.0153)	0.74 (0.001)	0.77 (0.004)	1.03 (0.756)
Arginine	R	0.056	1.15 (0.065)	5	4	2×10^{-8}	0.64 (7×10^{-5})	0.75 (0.002)	0.72 (2×10^{-4})	1.00 (1.000)
Methionine	M	0.021	0.85 (0.091)	6	5	3×10^{-8}	0.72 (0.002)	0.75 (9×10^{-4})	0.75 (8×10^{-4})	1.04 (0.614)
Histidine	H	0.026	0.89 (0.276)	5	4	4×10^{-9}	0.65 (8×10^{-5})	0.74 (7×10^{-4})	0.74 (3×10^{-4})	1.09 (0.316)
Cysteine	C	0.023	0.91 (0.359)	5	4	2×10^{-9}	0.68 (2×10^{-4})	0.70 (4×10^{-5})	0.72 (1×10^{-4})	1.05 (0.589)
Leucine	L	0.100	1.01 (0.892)	3	3	1×10^{-5}	0.53 (2×10^{-5})	0.77 (0.025)	0.76 (0.018)	1.05 (0.695)

IMM: immunogenic, NIMM: nonimmunogenic; *Selenocysteine (U): 3.17×10^{-6}

4.2. The frequency, expression, and presentation of TCEMs determine the prevalence of specific naïve CD8+ T cells in the repertoire

To confirm the previous findings, the predictions of the primary hypothesis were directly demonstrated on T cell repertoires of healthy individuals. Based on the hypothesis, it is less likely to detect a given naïve T cell in the repertoire that is specific for infrequent TCEMs in human proteins, for TCEMs not expressed in cTECs, or for TCEMs not presented on the surface of cTECs. Recently published data were utilized to demonstrate the absence of such T cells in the repertoire of healthy individuals (63). The dataset characterized in the Methods section is exceptional because the peptide sequence specificity of a large number of TCRs was determined using the MIRA technology. The peptides were grouped based on the prevalence, expression, and proteasomal cleavage scores of their TCEMs as described previously. For each individual and in each peptide group, the fraction of HLA-presented peptides recognized by at least one TCR in the repertoire was determined (*Methods*). Specific naive CD8+ T cells were less likely to be present for rare than nonrare TCEMs in the repertoire of healthy individuals *(Figure 12 A)*. Similarly, it was less likely to observe specific T cells for TCEMs having either negligible or overly high expression in cTECs *(Figure 12 B)*. Moreover, TCEMs with low thymoproteasomal cleavage scores were less likely to be associated with the presence of specific T cells in the repertoire *(Figure 12 C)*, while the immunoproteasomal cleavage score did not show this relationship *(Figure 12 D)*. In sum, these findings on T cell repertoire data confirmed the ones on in vitro T cell activation data.

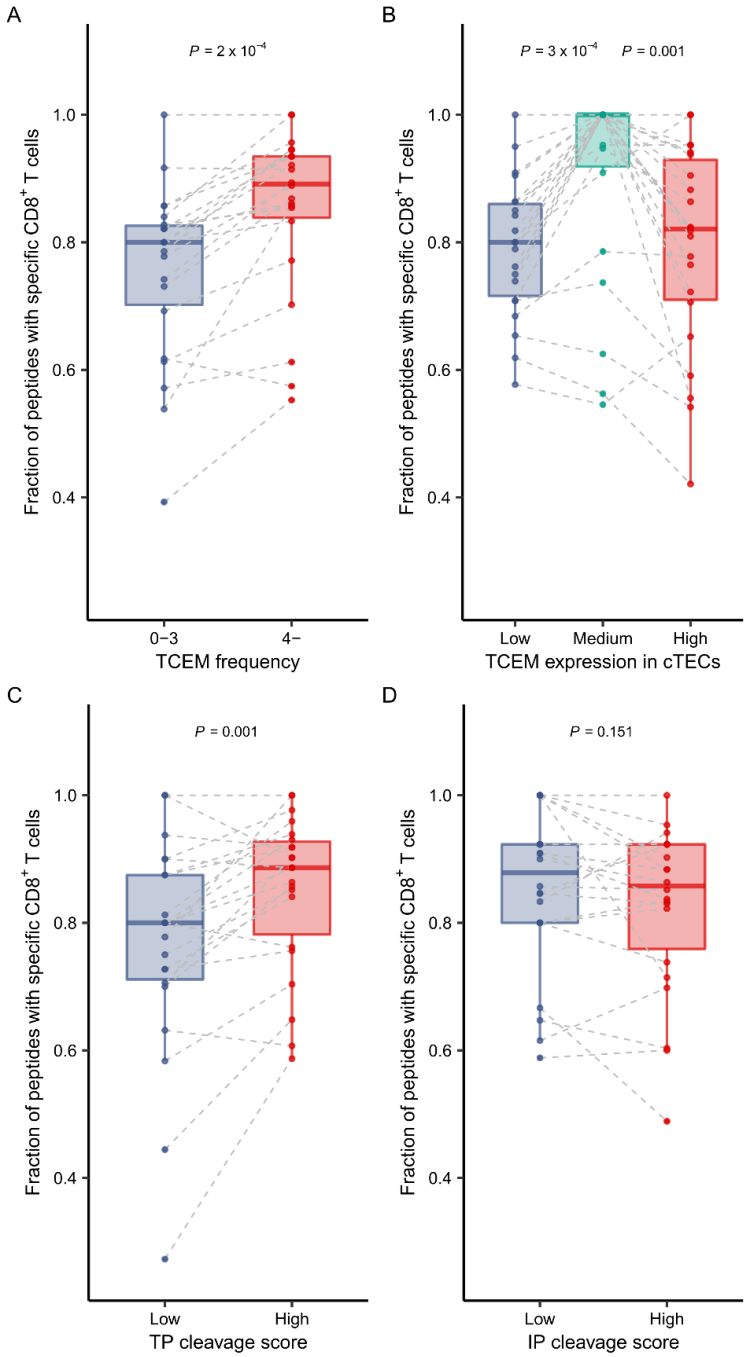


Figure 12. Specific naive CD8+ T cells were less likely to be present for TCEMs found rarely in human proteins (A), having low expression in cTECs (B), or low thymoproteasomal cleavage score (C). *The vertical axes represent the fraction of peptides, for which specific T cells were detected. Point pairs (or triplets on panel B) indicate values belonging to a given individual (n = 22). Two-sided P-values of paired Wilcoxon's rank-sum tests are shown. TCEMs were stratified into expression groups based on tertiles, and into thymoproteasomal (TP) or immunoproteasomal (IP) cleavage score groups based on the 1st quartile.*

4.3. Decreased immunogenicity of overly dissimilar peptides to human proteins

The leading hypothesis predicted a rather provocative relationship: in contrast with expectation, overly dissimilar peptides are not recognized by the immune system, because self-peptides mediate the positive selection of specific T cells. Put differently, it is less likely to find TCEMs of highly dissimilar peptides in the human proteome and, thus, specific positively selected T cells are potentially absent from the repertoire. To examine this assumption, for each peptide of the datasets used in our study its most similar counterpart in the human proteome was determined using established methods (46, 79) (*Methods*). As expected, peptides with exceptionally rare TCEMs (occurring zero to three times) in the human proteome had lower similarity than other peptides *(Figure 13 A)*. Accordingly, overly dissimilar peptides of datasets 1 and 2 were less likely to be immunogenic just like highly similar ones *(Figure 13 B)*. To corroborate these results, the self-similarity of SARS-CoV-2 peptides was analyzed. Reassuringly, naive CD8+ T cells specific for highly dissimilar peptides were found in the repertoire of fewer individuals *(Figure 13 C)*.

We conclude that while a given level of peptide dissimilarity to human proteins is essential for self-nonself discrimination, overly dissimilar peptides are less likely to be recognized by the immune system because specific T cells are not present in the repertoire.

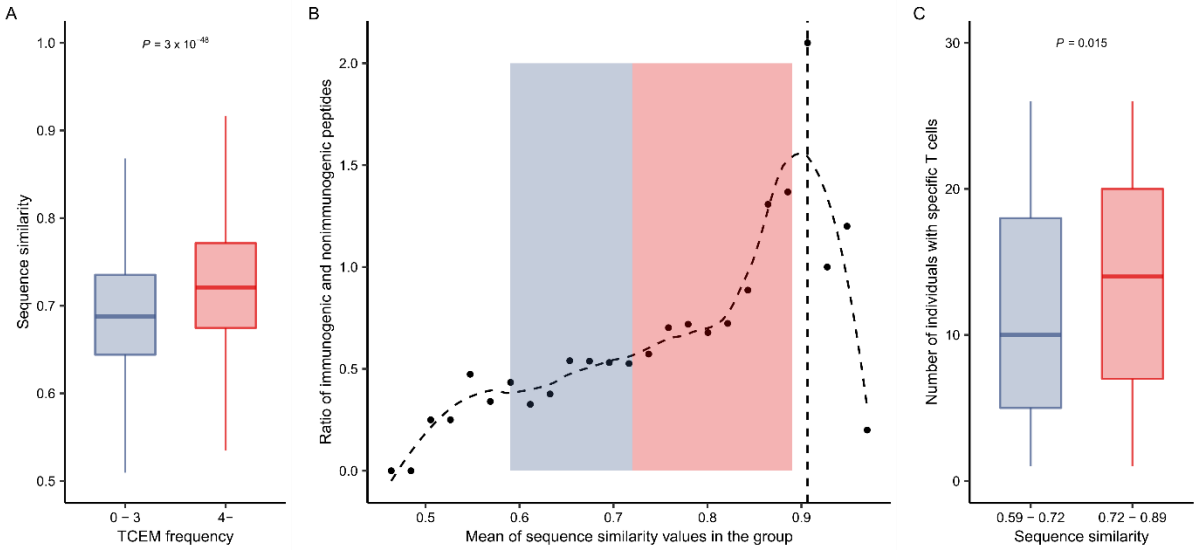


Figure 13. Overly dissimilar peptides to human proteins are less immunogenic. A) *Peptides in datasets 1 and 2 with TCEMs found less than four times in human proteins are less similar to the closest hit in the human proteome. (n = 1,706 and 2,309 in the 0-3 and 4- TCEM frequency groups, respectively) Outliers are not shown for visualization purposes.* ***B)*** *Peptides in datasets 1 and 2 were pooled and stratified into twenty-five groups based on similarity. In each group, the ratio between immunogenic and nonimmunogenic peptides was calculated. Groups are shown in increasing order of similarity. The horizontal axis indicates the mean similarity in the given group. The vertical dashed line indicates the group having the highest fraction of immunogenic peptides. The curve was fitted with a cubic smoothing spline method in R (Methods). The background shading represents the similarity ranges of peptide groups on panel C.* ***C)*** *T cells specific for overly dissimilar peptides were found in the repertoire of fewer individuals. Peptides were stratified into sequence similarity groups based on the median value (n = 149 and 147 in the lower and higher similarity groups, respectively). The similarity ranges are also indicated on panel B with background colors. Note, that the dataset of SARS-CoV-2 peptides did not include peptides that are highly similar to human proteins. On panels A and C, P-values of two-sided Wilcoxon's rank-sum tests are indicated.*

4.4. Cross-reactivity is not able to compensate for the side-effect of self-mediated positive selection of T cells

Our results suggest that the mechanism of positive selection results in a defective T cell repertoire. Is the cross-reactivity of TCRs able to compensate for these defects? To answer this question, two groups of TCEMs were first created *(Figure 14 A)*. The first group consisted of motifs in datasets 1 and 2, for which it is the least likely to find specific positively selected T cells in the repertoire based on the previous findings on T cell activation data (n = 43, these TCEMs were nonimmunogenic, found less than 4 times in the human proteome, had low expression in cTECs and low thymoproteasomal cleavage score). The second group consisted of all possible TCEM sequences (n = 323,470), for which the presence of specific T cells in the repertoire is likely: they were found more than 3 times in the human proteome, expressed in cTECs, and had normal thymoproteasomal cleavage scores. For each TCEM in the first set, its BLOSUM62 similarity to every TCEM in the second set was calculated to explain their proximity in sequence space *(Figure 14 A)*.

Next, the level of TCR cross-reactivity was estimated in the sequence space of TCEMs. Empirical data was downloaded from a recent study. The authors measured the binding strength of the well-known NY-ESO-1 epitope to TCR C^{259}, when sequentially replacing every amino acid in different positions of the epitope (65). We determined the BLOSUM62 similarity between the TCEM of the original and the modified peptide sequences and we found a strong positive correlation between the similarity of the original to the modified TCEM and the peptide binding strength to TCR C^{259} *(Figure 14 B)*.

Then a TCEM similarity cutoff value was determined, under which the binding to the TCR is too weak to induce T cell activation *(Figure 14 A)*. The relationship between the TCR binding strength and T cell activation was examined. Less than 10% of original binding strength as insufficient binding was fixed because the ability of peptides to activate T cells was negligible below this cutoff (*Methods*). The similarity between the modified and the original TCEM was able to accurately predict whether the peptide will be bound by the TCR with strength above this level *(Figure 14 C)*. An established "cost-benefit" method (67) was then used to determine the optimal TCEM similarity cutoff for binding (value = 0.61, *Methods*). Reassuringly, a very

similar cutoff value was obtained, when using data of an independent study on the A6 TCR and its target epitope, the Tax peptide of HTLV-1 (66) (*Methods*).

Are T cells that are specific for TCEMs in the second group able to bind TCEMs in the first group? We found that only an insignificant minority (ranging from 0.0006% to 0.043% for TCEMs in the first group, median = 0.015%) of similarity values reached the previously determined cutoff values of T cell cross-reactivity *(Figure 14 D)*. This result suggests that T cells in the repertoire (specific for TCEMs in the second group) are unlikely to recognize TCEMs, whose recognition is negatively affected by self-mediated positive selection (i.e., TCEMs in the first group). Although the result is indicative, it is important to highlight that cross-reactivity was inferred based on the data for two TCRs, and the results need future validation using data of more TCRs.

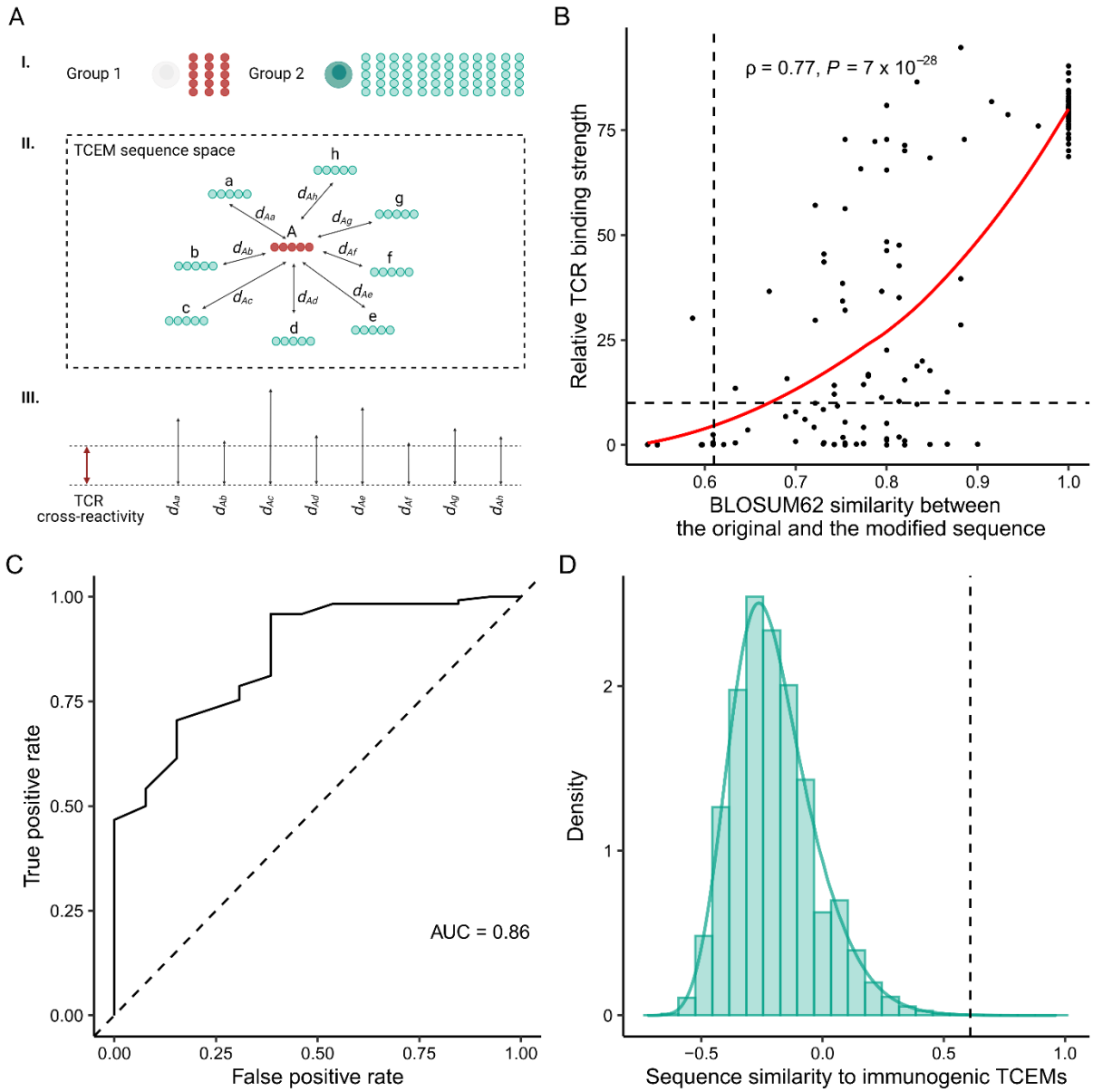


Figure 14. TCR cross-reactivity is unlikely to compensate for the defects in the T cell repertoire. A) *Schematic diagram of the analysis. To determine whether T cell cross-reactivity can bridge defects in the repertoire, two groups of TCEMs were created (I). The first group consisted of motifs, for which it is the least likely to find specific positively selected T cells in the repertoire based on our results (marked with red color on the sketch). The second group consisted of all TCEMs, for which it is likely to find specific T cells in the repertoire (marked with green color on the sketch). The pairwise similarity was calculated between the members of the two TCEM groups (II). The higher the similarity between two TCEMs, the closer they reside in the sequence space resulting in smaller distance (d) values. Next, the level of T cell cross-reactivity was estimated in TCEM sequence space (III). Cross-reactivity was defined as the lowest similarity between a given TCEM sequence and the TCR's cognate TCEM sequence that*

is needed for a reasonable TCR binding strength and T cell activation. Finally, the number of cases was determined when members of the first and second groups are close enough to be recognized by the same TCR. ***B)*** *The amino acids of the NY-ESO-1 epitope were sequentially changed and the binding strength to TCR* C^{259} *was measured in a previous study (65). The relative binding strength of the modified (n = 135) and the original peptide to TCR* C^{259} *is shown as a function of the BLOSUM62 similarity between their TCEM sequences. The horizontal line indicates 10% of the original binding value, which was considered as a cutoff for improbable binding (Methods). Spearman's rho and the P-value of a two-sided correlation test are indicated. The red line indicates a smooth curve fitted using a cubic smoothing spline method in R (Methods).* ***C)*** *The ROC curve demonstrates the accuracy of BLOSUM62 similarity in classifying peptides into binding and nonbinding (i.e., lower than 10% of original binding) groups. AUC: area under the curve* ***D)*** *The density of all similarity values (n = 13,909,210) between TCEMs in group 1 and group 2. Vertical lines on panels B and D represent the optimal cutoff (0.61) for classification.*

4.5. Positive selection of T cells and susceptibility to infections

The adaptive recognition of pathogen-associated peptide sequences is essential for the initiation of an effective immune response. Presented results suggest that many such sequences are potentially nonimmunogenic because specific T cells are not observed in the CD8+ T cell repertoire. We aimed to determine the frequency of these peptides in proteins of intracellular pathogens. To this end, reference proteomes of 50 familiar intracellular pathogens were used. In the proteome of each species, the prevalence of TCEMs was determined, that are either rare or not found in human proteins and/or not or lowly expressed in cTECs and/or unlikely to be presented after thymoproteasomal cleavage (called np-TCEMs hereafter, referring to TCEMs for which we expect to find specific positively selected T cells with lower probability) (*Methods*). The frequency of these np-TCEMs ranged from 58% to 71% in different species. *(Figure 15 A).*

This high fraction of np-TCEMs could hinder immune recognition, especially when only a few peptides of the pathogen are presented because either the proteome of the pathogen is small and/or the HLA allele has a narrow binding repertoire. To this end, the binding of all 9-mer peptides found in the proteome of pathogens was predicted to the most common HLA-I alleles (*Methods*). For each allele-species pair, the fraction of np-TCEMs was calculated in the presented peptides and the result was visualized on a heatmap *(Figure 15 B).* As expected, the fraction of presented peptides with np-TCEMs was extremely variable between HLA alleles when the pathogens had small proteomes *(Figure 15 C).* This group of pathogens was dominated by viruses, like Human parvovirus B19, Hepatitis viruses, Human papillomavirus, etc. On the contrary, HLA alleles presented a similar fraction of np-TCEMs from large proteomes of protozoal and bacterial species.

We expected an HLA-dependent effect of np-TCEMs on disease risk. To this end, HLA association meta-analysis data were collected (*Methods*). Allele groups with positive or negative associations were selected and the fraction of np-TCEMs presented by alleles in each group from all peptides of the causative pathogens were calculated. Allele groups associated with infections or treatment failure dominantly presented np-TCEMs in contrast with protective allele groups *(Figure 15 D).*

These results suggest that the proposed side-effect of T cell positive selection influences the adaptive immune recognition of intracellular pathogens.

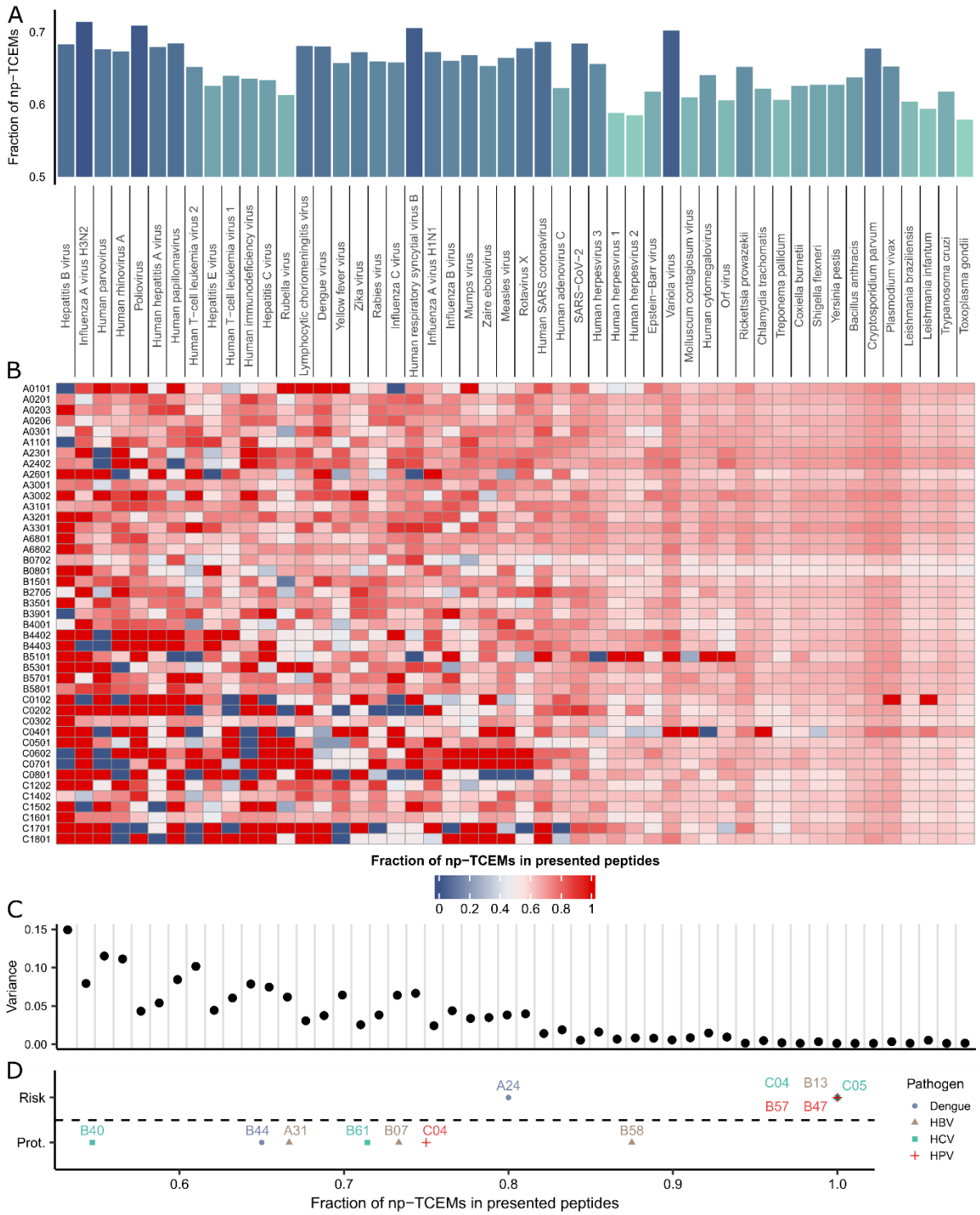


Figure 15. The effect of self-mediated positive selection on the recognition of pathogens. A) *The prevalence of np-TCEMs in the proteome of different pathogens (n = 50). np-TCEMs were defined as*

being found less than 4 times in the human proteome or having low expression in cTECs or low thymoproteasomal cleavage score. Pathogens are ordered by increasing proteome size. ***B)*** *The heatmap shows the prevalence of np-TCEMs in peptides of intracellular pathogens that are presented by common HLA alleles.* ***C)*** *The plot shows the variance of presented np-TCEMs by different HLA alleles. The variance decreases with increasing proteome size of pathogens (Spearman's rho: -0.93, two-sided correlation test* $P = 9.13 \times 10^{-23}$*).* ***D)*** *The fraction of np-TCEMs in peptides that are presented by risk (n = 6) and protective (n = 7) HLA allele groups. Group-specific values were calculated by averaging values for common alleles in each group (Methods). In contrast with protective allele groups, predisposing ones present mainly peptides with np-TCEMs in their sequence (two-sided Wilcoxon's rank-sum test P = 0.004). HBV: Hepatitis B virus; HCV: Hepatitis C virus; HPV: Human papillomavirus.*

5. DISCUSSION

5.1. The relationship between T cell positive selection and the nonresponsiveness to nonself peptides

The prevalence of specific T cells in the repertoire is essential for adaptive immune recognition of HLA-presented peptides. It has been suggested that during positive selection, self-peptides on the surface of cTECs can be considered as a test set for thymocytes: cells that recognize these peptides survive, potentially recognize nonself peptides more effectively and, consequently, dominate the immune response to the foreign antigens (21, 37, 47). Our results suggest that the nonresponsiveness to many nonself peptides can also be explained by the mechanism of T cell positive selection because it is mediated by self-peptides. Put differently, self-mediated positive selection has a negative trade-off on the recognition of foreign peptides. Importantly, the T cell cross-reactivity cannot compensate for this side effect set up by T cell positive selection *(Figure 14)*.

5.2. The TCEM region is crucial in the recognition of peptide sequences

In our research, three lines of evidence were presented supporting the leading hypothesis on two reliable and nonoverlapping peptide sets *(Figures 7, 8 and 10)*. We focused on the TCEM region of peptides. Although amino acids in other positions may also contact with TCR, extensive literature information supports that the amino acids between positions 4 and 8 are the main contacting residues with TCRs. In an extensive and highly cited review, authors examined molecular structures of many pHLA-TCR complexes and reported that mainly amino acids at positions 4 through 8 are in contact with TCRs (17). Importantly, other studies also suggested that self-nonself discrimination is governed by these short motifs (18, 80–82). Moreover, our analysis on TCR cross-reactivity also supported these findings: the TCEM sequences of the modified NY-ESO-1 peptides alone were able to determine the binding of the peptide to the TCR C^{259} *(Figure 14 B)*. While the TCEM region is essential in TCR binding, it has been at issue how such short peptides can make it possible for the immune system to differentiate between self and nonself peptides (81, 83). Namely, human peptides contain around 75% of all possible pentamer sequences (81) (73.1% in our analysis) that largely overlap with the ones found in commensal and pathogenic bacteria (81). Our findings suggest that the overlap

between self and nonself motifs is far from being disadvantageous. On the contrary, it is crucial for the positive selection of T cells that are specific for foreign peptides. In other words, the overlap between motifs makes it possible to recognize nonself.

5.3. Are there holes in the T cell repertoire?

Our hypothesis was supported with direct evidence by examining naïve CD8+ T cell repertoires of healthy individuals. The results suggest that self-mediated positive selection has a negative effect on the prevalence of SARS-CoV-2 peptide-specific T cells in the repertoire *(Figure 12)*. Importantly, it has already been suggested that "holes" in the T cell repertoire hinder the recognition of certain pathogens (84–87) and the studies explained the presence of such holes by central tolerance (84, 87). On the other hand, *Yu et al.* suggested that there are no significant holes in the repertoire because clonal deletion affects only the most self-reactive T cells (88). Consequently, every possible HLA-presented peptide could be recognized by T cells, but many T cells are anergic due to immune tolerance mechanisms. However, our results suggest there are gaps in the T cell repertoire caused by positive selection.

5.4. T cell positive selection affects the recognition of peptides of pathogens

The fraction of those peptides in pathogens whose recognition could be affected by self-mediated positive selection to some extent was also estimated. A significant proportion of peptides - varying between 58% and 71% in various species – fell into this category *(Figure 15 A)*. If we also consider that around one-third of nonself peptides are indistinguishable from self-ones due to high similarity (39), it is not surprising that at least 50% of HLA-A*02:01-presented vaccinia and HIV sequences were reported to be nonimmunogenic in previous studies (39, 84, 89). At the same time, it depends on the specificity of HLA alleles, which peptides are presented to the T cells. Accordingly, the results showed that HLA alleles, which predominantly present peptides of HBV, HCV, HPV, and dengue virus without specific positively selected T cells, are associated with infections or worse response to therapy *(Figure 15 D)*. To note, a similar mechanism could also explain variable responses to vaccines that deserve further investigation.

5.5. Overly dissimilar peptides are potentially not recognized by the immune system

Finally, presented results do not support a conventional interpretation of self-nonself discrimination, which suggests that the more dissimilar a peptide to self, the more likely it is to be immunogenic (44, 45, 90, 91). Our results showed that the more dissimilar a peptide to human proteins, the less likely it is to find its TCEM in the human proteome *(Figure 13 A)*. Consequently, specific positively selected T cells are potentially absent from the repertoire above a level of dissimilarity *(Figure 13 B and C)*. In sum, although a certain level of dissimilarity is essential for the discrimination of self and nonself, overly dissimilar peptides are potentially unrecognized by the immune system *(Figure 16)*. While these results indicate the importance of this blind spot in the immune response to infections, it is a question to be clarified in future works, whether mutated cancer peptides can also reach this level of dissimilarity. In the same way, evaluating our hypothesis on HLA-II presented peptides and CD4+ T cells is also an important area of subsequent research.

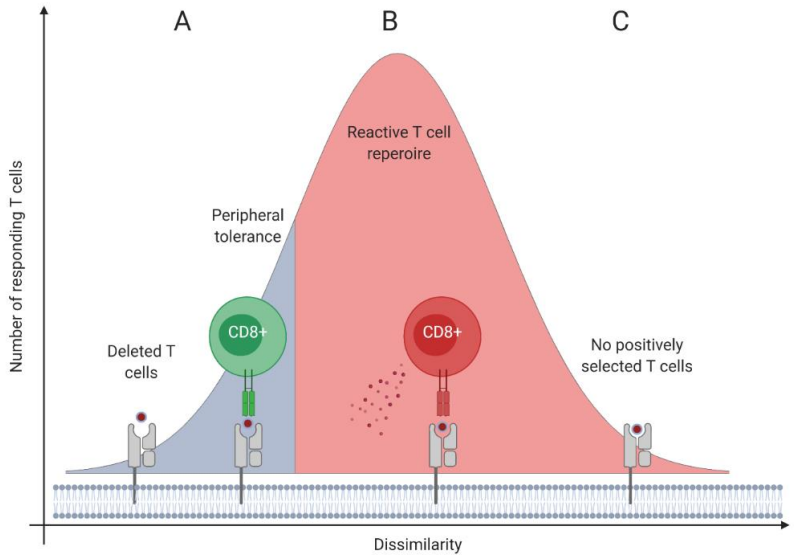


Figure 16. The blindness of immune recognition for peptides that are overly dissimilar to human proteins. A) *Immune system tolerates peptides that are similar to self-proteins. T cells recognizing these peptides are either deleted in the thymus or unresponsive due to peripheral tolerance mechanisms (38).* ***B)*** *Peptides with a certain level of dissimilarity to human proteins are recognized as nonself resulting in T cell activation and immune-mediated destruction of cells.* ***C)*** *Peptides that are overly dissimilar to human proteins are not recognized by the immune system, because specific positively selected T cells are absent from the repertoire.*

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor, Máté Manczinger, for his constant support and supervision of my work. He helped to learn the R programming language and to search and to manage online databases. He established the Computational Immunological Research Group where I could participate in several projects (*E. coli* and IBD, HLA promiscuity in TCGA and immunotherapy cohorts, mutational signatures, SARS-CoV-2 evolution, etc.). He always gave me much advice to proceed and motivated me. He constantly monitored the progress of the work therefore the processes could not be stranded. I would like to thank him for his support during the preparation of the manuscript and thesis.

I would like to express my special thanks to Professor Lajos Kemény for the opportunity to perform my studies at the Department of Dermatology and Allergology, University of Szeged.

I am grateful to my colleagues Gergő Mihály Balogh, Benjamin Tamás Papp, Leó Asztalos and Anna Tácia Fülöp for helping and supporting me during the years. They gave me much advice in programming and helped to optimize the presentations at conferences. I am grateful for their help and comments on my work.

I also want to thank Csaba Pál and Balázs Papp for allowing me to participate in their group seminars and giving advice on our projects.

I would like to thank Prof. Dr. Péter Hegyi for the opportunity to do my research in his laboratory in the first year of my training and to all members of the Hungarian Pancreatic Study Group, especially Andrea Szentesi for their help and support.

My special thanks go to my wife, Barbara, for her patience, constant support, encouragement, and to my daughters (Orsi and Flóra) and my whole family for their love.

REFERENCES

1. F. M. Burnet, The Production of Antibodies. A Review and a Theoretical Discussion. *The Production of Antibodies. A Review and a Theoretical Discussion.* (1941) (February 9, 2021).

2. A. I. Tauber, The immune self: theory or metaphor? *Immunology Today* **15**, 134–136 (1994).

3. T. Pradeu, E. L. Cooper, The danger theory: 20 years later. *Front Immunol* **3** (2012).

4. P. Matzinger, Tolerance, danger, and the extended family. *Annu Rev Immunol* **12**, 991–1045 (1994).

5. P. Parham, *The Immune System*, 4th Ed. (Garland Science, 2014).

6. M. L. Dustin, The Immunological Synapse. *Cancer Immunol Res* **2**, 1023–1033 (2014).

7. T. Shiina, K. Hosomichi, H. Inoko, J. K. Kulski, The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of Human Genetics* **54**, 15–39 (2009).

8. J. Robinson, *et al.*, The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* **43**, D423–D431 (2015).

9. S. Boegel, *et al.*, HLA and proteasome expression body map. *BMC Med Genomics* **11** (2018).

10. K. L. Rock, A. L. Goldberg, Degradation of cell proteins and the generation of MHC class I-presented peptides. *Annu Rev Immunol* **17**, 739–779 (1999).

11. L. Stoltze, *et al.*, Two new proteases in the MHC class I processing pathway. *Nat Immunol* **1**, 413–418 (2000).

12. J. C. Shepherd, *et al.*, TAP1-dependent peptide translocation in vitro is ATP dependent and peptide selective. *Cell* **74**, 577–584 (1993).

13. A. Blees, *et al.*, Structure of the human MHC-I peptide-loading complex. *Nature* **551**, 525–528 (2017).

14. H. Pearson, *et al.*, MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest* **126**, 4690–4701 (2016).

15. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology* **14**, 377–391 (2014).

16. K. Falk, O. Rötzschke, S. Stevanović, G. Jung, H.-G. Rammensee, Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **351**, 290–296 (1991).

17. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How Tcrs Bind Mhcs, Peptides, and Coreceptors. *Annual Review of Immunology* **24**, 419–466 (2006).

18. R. D. Bremel, E. J. Homan, Frequency Patterns of T-Cell Exposed Amino Acid Motifs in Immunoglobulin Heavy Chain Peptides Presented by MHCs. *Front. Immunol.* **5** (2014).

19. B. F. Haynes, M. L. Markert, G. D. Sempowski, D. D. Patel, L. P. Hale, The Role of the Thymus in Immune Reconstitution in Aging, Bone Marrow Transplantation, and HIV-1 Infection. *Annual Review of Immunology* **18**, 529–560 (2000).

20. J. F. A. P. Miller, The golden anniversary of the thymus. *Nature Reviews Immunology* **11**, 489–495 (2011).

21. N. Vrisekoop, J. P. Monteiro, J. N. Mandl, R. N. Germain, Revisiting Thymic Positive Selection and the Mature T Cell Repertoire for Antigen. *Immunity* **41**, 181–190 (2014).

22. H. Takaba, H. Takayanagi, The Mechanisms of T Cell Selection in the Thymus. *Trends in Immunology* **38**, 805–816 (2017).

23. K. Takada, K. Kondo, Y. Takahama, Generation of Peptides That Promote Positive Selection in the Thymus. *The Journal of Immunology* **198**, 2215–2222 (2017).

24. T. K. Starr, S. C. Jameson, K. A. Hogquist, Positive and Negative Selection of T Cells. *Annual Review of Immunology* **21**, 139–176 (2003).

25. K. Sasaki, *et al.*, Thymoproteasomes produce unique peptide motifs for positive selection of CD8 + T cells. *Nature Communications* **6**, 7484 (2015).

26. S. Murata, *et al.*, Regulation of CD8+ T Cell Development by Thymus-Specific Proteasomes. *Science* **316**, 1349–1353 (2007).

27. Y. Xing, S. C. Jameson, K. A. Hogquist, Thymoproteasome subunit-β5T generates peptide-MHC complexes specialized for positive selection. *PNAS* **110**, 6979–6984 (2013).

28. T. Nitta, *et al.*, Thymoproteasome Shapes Immunocompetent Repertoire of CD8+ T Cells. *Immunity* **32**, 29–40 (2010).

29. E. Palmer, Negative selection — clearing out the bad apples from the T-cell repertoire. *Nature Reviews Immunology* **3**, 383–391 (2003).

30. B. Kyewski, J. Derbinski, Self-representation in the thymus: an extended view. *Nature Reviews Immunology* **4**, 688–698 (2004).

31. A. M. Gallegos, M. J. Bevan, Central Tolerance to Tissue-specific Antigens Mediated by Direct and Indirect Antigen Presentation. *J Exp Med* **200**, 1039–1049 (2004).

32. T. P. Arstila, *et al.*, A Direct Estimate of the Human αβ T Cell Receptor Diversity. *Science* **286**, 958–961 (1999).

33. K. Wing, S. Sakaguchi, Regulatory T cells exert checks and balances on self tolerance and autoimmunity. *Nature Immunology* **11**, 7–13 (2010).

34. I. Bains, H. M. van Santen, B. Seddon, A. J. Yates, Models of Self-Peptide Sampling by Developing T Cells Identify Candidate Mechanisms of Thymic Selection. *PLOS Computational Biology* **9**, e1003102 (2013).

35. Y. Xing, K. A. Hogquist, T-Cell Tolerance: Central and Peripheral. *Cold Spring Harb Perspect Biol* **4**, a006957 (2012).

36. A. E. Moran, K. A. Hogquist, T-cell receptor affinity in thymic development. *Immunology* **135**, 261–267 (2012).

37. J. N. Mandl, J. P. Monteiro, N. Vrisekoop, R. N. Germain, T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity* **38**, 263–274 (2013).

38. A. Bresciani, *et al.*, T-cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome. *Immunology* **148**, 34–39 (2016).

39. J. J. A. Calis, R. J. de Boer, C. Keşmir, Degenerate T-cell Recognition of Peptides on MHC Molecules Creates Large Holes in the T-cell Repertoire. *PLOS Computational Biology* **8**, e1002412 (2012).

40. D. Mason, A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol Today* **19**, 395–404 (1998).

41. P. Kraj, R. Pacholczyk, L. Ignatowicz, αβTCRs Differ in the Degree of Their Specificity for the Positively Selecting MHC/Peptide Ligand. *The Journal of Immunology* **166**, 2251–2259 (2001).

42. M. E. Birnbaum, *et al.*, Deconstructing the Peptide-MHC Specificity of T Cell Recognition. *Cell* **157**, 1073–1087 (2014).

43. A. Kosmrlj, *et al.*, Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* **465**, 350–354 (2010).

44. P. Lydyard, A. Whelan, M. Fanger, *BIOS Instant notes in immunology* (Taylor & Francis, 2011).

45. S. K. Mohanty, K. S. Leela, *Textbook of immunology* (JP Medical Ltd, 2013).

46. S. C. Pro, *et al.*, Microbiota epitope similarity either dampens or enhances the immunogenicity of disease-associated antigenic epitopes. *PLOS ONE* **13**, e0196551 (2018).

47. R. B. Fulton, *et al.*, The TCR's sensitivity to self peptide-MHC dictates the ability of naive CD8(+) T cells to respond to foreign antigens. *Nat Immunol* **16**, 107–117 (2015).

48. R. Vita, *et al.*, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* **47**, D339–D343 (2019).

49. R. Vita, B. Peters, A. Sette, The curation guidelines of the immune epitope database and analysis resource. *Cytometry Part A* **73A**, 1066–1070 (2008).

50. S. G. E. Marsh, *et al.*, Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* **75**, 291–455 (2010).

51. V. Jurtz, *et al.*, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

52. S. Paul, *et al.*, HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. *J.I.* **191**, 5831–5839 (2013).

53. M. Nielsen, M. Andreatta, NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* **8**, 33 (2016).

54. M. Collatz, *et al.*, EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics* **37**, 448–455 (2021).

55. S. E. C. Caoili, Expressing Redundancy among Linear-Epitope Sequence Data Based on Residue-Level Physicochemical Similarity in the Context of Antigenic Cross-Reaction. *Advances in Bioinformatics* **2016**, e1276594 (2016).

56. M. Manczinger, *et al.*, Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLOS Biology* **17**, e3000131 (2019).

57. F. Sievers, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011).

58. K. Yang, L. Zhang, Performance comparison between k -tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Research* **36**, e33–e33 (2008).

59. The UniProt Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2019).

60. S. Sarkizova, *et al.*, A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology* **38**, 199–209 (2020).

61. A. J. Coles, *et al.*, Keratinocyte growth factor impairs human thymic recovery from lymphopenia. *JCI Insight* **4** (2019).

62. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends in Genetics* **29**, 569–574 (2013).

63. T. M. Snyder, *et al.*, "Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels" (Infectious Diseases (except HIV/AIDS), 2020) https:/doi.org/10.1101/2020.07.31.20165647 (September 30, 2020).

64. M. Klinger, *et al.*, Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PLOS ONE* **10**, e0141561 (2015).

65. A. R. Karapetyan, *et al.*, TCR Fingerprinting and Off-Target Peptide Identification. *Front. Immunol.* **10** (2019).

66. R. S. Gejman, *et al.*, Identification of the Targets of T-cell Receptor Therapeutic Agents and Cells by Use of a High-Throughput Genetic Platform. *Cancer Immunol Res* **8**, 672–684 (2020).

67. C. E. Metz, Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298 (1978).

68. D. Weiskopf, *et al.*, Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *PNAS* **110**, E2046–E2053 (2013).

69. V. Seshasubramanian, G. Soundararajan, P. Ramasamy, Human leukocyte antigen A, B and Hepatitis B infection outcome: A meta-analysis. *Infect Genet Evol* **66**, 392–398 (2018).

70. E. Gauthiez, *et al.*, A systematic review and meta-analysis of HCV clearance. *Liver International* **37**, 1431–1445 (2017).

71. M. Bhaskaran, G. ArunKumar, A meta-analysis of association of Human Leukocyte Antigens A, B, C, DR and DQ with Human Papillomavirus 16 infection. *Infect Genet Evol* **68**, 194–202 (2019).

72. S. J. Mack, *et al.*, Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* **81**, 194–203 (2013).

73. C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, C. De Boor, *A practical guide to splines* (springer-verlag New York, 1978).

74. D. Malhotra, *et al.*, Tolerance is established in polyclonal CD4+ T cells by distinct mechanisms, according to self-peptide expression patterns. *Nat Immunol* **17**, 187–195 (2016).

75. D. Von Bubnoff, *et al.*, Antigen-presenting cells and tolerance induction. *Allergy* **57**, 2–8 (2002).

76. M. Zeeshan, K. Tyagi, Y. D. Sharma, CD4+ T Cell Response Correlates with Naturally Acquired Antibodies against Plasmodium vivax Tryptophan-Rich Antigens. *Infection and Immunity* **83**, 2018–2029 (2015).

77. J. Schmidt, *et al.*, Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine* **2**, 100194 (2021).

78. T. P. Riley, *et al.*, Structure Based Prediction of Neoantigen Immunogenicity. *Front. Immunol.* **10** (2019).

79. L. P. Richman, R. H. Vonderheide, A. J. Rech, Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Systems* **9**, 375-382.e4 (2019).

80. J. J. A. Calis, *et al.*, Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLOS Computational Biology* **9**, e1003266 (2013).

81. R. D. Bremel, E. J. Homan, Extensive T-Cell Epitope Repertoire Sharing among Human Proteome, Gastrointestinal Microbiome, and Pathogenic Bacteria: Implications for the Definition of Self. *Front. Immunol.* **6** (2015).

82. M. J. Reddehase, J. B. Rothbard, U. H. Koszinowski, A pentapeptide as minimal antigenic determinant for MHC class I-restricted T lymphocytes. *Nature* **337**, 651–653 (1989).

83. N. J. Burroughs, R. J. de Boer, C. Keşmir, Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics* **56**, 311–320 (2004).

84. S. Frankild, R. J. de Boer, O. Lund, M. Nielsen, C. Kesmir, Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for "Holes" in the T Cell Repertoire. *PLoS ONE* **3**, e1831 (2008).

85. M. Wölfl, *et al.*, Hepatitis C Virus Immune Escape via Exploitation of a Hole in the T Cell Repertoire. *The Journal of Immunology* **181**, 6435–6446 (2008).

86. E. J. Yager, *et al.*, Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *Journal of Experimental Medicine* **205**, 711–723 (2008).

87. D. Vidović, P. Matzinger, Unresponsiveness to a foreign antigen can be caused by self-tolerance. *Nature* **336**, 222–225 (1988).

88. W. Yu, *et al.*, Clonal Deletion Prunes but Does Not Eliminate Self-Specific αβ CD8(+) T Lymphocytes. *Immunity* **42**, 929–941 (2015).

89. M. Rolland, *et al.*, Recognition of HIV-1 peptides by host CTL is related to HIV-1 similarity to human proteins. *PLoS ONE* **2**, e823 (2007).

90. M. Yarmarkovich, J. M. Warrington, A. Farrel, J. M. Maris, Identification of SARS-CoV-2 Vaccine Epitopes Predicted to Induce Long-Term Population-Scale Immunity. *Cell Reports Medicine* **1**, 100036 (2020).

91. M. Ogishi, H. Yotsuyanagi, The landscape of T cell epitope immunogenicity in sequence space. *bioRxiv*, 155317 (2018).

I.

Q: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself

Balázs Koncz[a], Gergő M. Balogh[a], Benjamin T. Papp[a,b], Leó Asztalos[a,b], Lajos Kemény[a,c,d], and Máté Manczinger[a,c,d,e,1]

[a]Department of Dermatology and Allergology, University of Szeged, Hungary; [b]Szeged Scientists Academy, Hungary; [c]MTA-SZTE Dermatological Research Group, University of Szeged, Hungary; [d]HCEMM-USZ Skin Research Group, Szeged, Hungary; and [e]Biological Research Centre, Institute of Biochemistry, Synthetic and Systems Biology Unit, Eötvös Loránd Research Network, Szeged, Hungary

Edited by Philippa Marrack, National Jewish Health, Denver, CO, and approved July 19, 2021 (received for review January 11, 2021)

Adaptive immune recognition is mediated by the binding of peptide–human leukocyte antigen complexes by T cells. Positive selection of T cells in the thymus is a fundamental step in the generation of a responding T cell repertoire: only those T cells survive that recognize human peptides presented on the surface of cortical thymic epithelial cells. We propose that while this step is essential for optimal immune function, the process results in a defective T cell repertoire because it is mediated by self-peptides. To test our hypothesis, we focused on amino acid motifs of peptides in contact with T cell receptors. We found that motifs rarely or not found in the human proteome are unlikely to be recognized by the immune system just like the ones that are not expressed in cortical thymic epithelial cells or not presented on their surface. Peptides carrying such motifs were especially dissimilar to human proteins. Importantly, we present our main findings on two independent T cell activation datasets and directly demonstrate the absence of naïve T cells in the repertoire of healthy individuals. We also show that T cell cross-reactivity is unable to compensate for the absence of positively selected T cells. Additionally, we show that the proposed mechanism could influence the risk for different infectious diseases. In sum, our results suggest a side effect of T cell positive selection, which could explain the nonresponsiveness to many nonself peptides and could improve the understanding of adaptive immune recognition.

adaptive immune recognition | T cell repertoire | infectious diseases | positive selection

The human immune system has to differentiate between self and nonself. The prerequisite of adaptive immune recognition is the formation of the immunological synapse (1). This structure is made up of human leukocyte antigen (HLA) molecules presenting short peptide sequences to T cells (1). T cell receptors (TCRs) recognize T cell exposed motifs (TCEMs) of peptide sequences (2–5). These are short, usually five amino acid–long motifs in contact with the CDR3 region of TCRs and are not involved in anchoring the peptides to HLA molecules (2–5).

Adaptive immune recognition is dependent on the presence of peptide-specific T cells in the T cell repertoire (6). The T cell repertoire is shaped by positive and negative selection steps in the thymus (6). Positive selection takes place around cortical thymic epithelial cells (cTECs) (6). cTECs present a special set of peptides on the cell surface produced by the thymoproteasome and cathepsin L (6–8). Recognition of these cTEC-specific peptides by T cell precursors (called thymocytes) is essential for the formation of a functioning T cell repertoire. Nonetheless, these peptides are cleavage products of human proteins (6–9). Thymocytes recognizing HLA-bound self-peptides survive, while others die by neglect (7, 9). Positively selected T cells then go through negative selection: T cells binding self-peptide–HLA complexes with high affinity are deleted from the repertoire, referred to as central tolerance (6).

The positive selection of T cells is an essential step in the formation of a responsive T cell repertoire. It has been suggested that both the CD4+ and CD8+ T cell repertoires are skewed to greater self-reactivity and that T cells that bind self-peptides stronger also bind the foreign agonist peptides more effectively (9–11). In other words, self-peptides mediating positive selection can be considered as a "test-set" selecting T cells that recognize foreign peptides with higher effectivity. However, is there any negative consequence of this mechanism?

We propose a fundamental side effect of T cell positive selection on the recognition of nonself peptides: as positive selection is mediated by self-peptides, a large fraction of nonself peptides is not recognized by the immune system even if T cells are cross-reactive. To test our hypothesis, we focused on the TCEMs of HLA class I (HLA-I) restricted peptides. As T cell positive selection is mediated by TCEMs of self-peptides, we expected that it is less likely to find specific T cells in the repertoire for TCEMs that are 1) very rare or missing from human proteins, 2) not expressed in, or 3) not presented on the surface of cTECs. Accordingly, we expected that peptides carrying such motifs are less immunogenic. We demonstrate the predictions of our hypothesis on two nonoverlapping T cell activation datasets and provide more direct evidence by examining naive CD8+ T cell repertoires of healthy individuals. Although it is widely accepted that nonself peptides that are highly dissimilar to human proteins are more immunogenic (12–16), we found that the dominantly nonimmunogenic peptides having rare TCEMs were

Significance

It is well established that peptides that are dissimilar to human proteins are more immunogenic. However, the immune system is still unable to recognize a large fraction of highly dissimilar peptides found in a wide variety of pathogens. We propose that this phenomenon could be explained by the mechanism of T cell positive selection. During this process, only those cells survive that recognize human peptides on the surface of thymic epithelial cells. As self-peptides mediate positive selection, the immune system is unable to recognize many nonself peptides, most of which are highly dissimilar to human peptides.

Author contributions: B.K., L.K., and M.M. designed research; B.K., G.M.B., B.T.P., L.A., and M.M. performed research; M.M. contributed new reagents/analytic tools; B.K., G.M.B., B.T.P., L.A., and M.M. analyzed data; B.K. and M.M. wrote the paper; L.K. provided supervision and funding acquisition; and M.M. provided supervision and project administration.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: manczinger.mate@med.u-szeged.hu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2100542118/-/DCSupplemental.

more dissimilar to human proteins than immunogenic ones. Such peptides dominated the proteome of many intracellular pathogens, and their presentation by HLA-I molecules was associated with an increased risk of infectious diseases. Our results suggest that the self-mediated positive selection of T cells generates a "blind spot" in adaptive immune recognition with implications on the susceptibility to infectious diseases.

Results

Dataset Assembly. To test the predictions of our hypothesis, we focused on the immunogenicity of peptides that are presented by HLA-I molecules, and thus, the lack of T cell response cannot be explained by missing antigen presentation. We collected T cell activation data for nonhuman peptides from the Immune Epitope Database (IEDB) (17) and assembled two nonoverlapping datasets using different criteria (*SI Appendix*, Fig. S1 and Dataset S1).

In the first dataset, we predicted the binding of each peptide to the reported HLA allele and excluded the ones whose HLA-binding was not confirmed by prediction. To note, this approach has already been used by previous studies focusing on peptide immunogenicity (5). In the case of the second dataset, we aimed to control for certain confounding factors that could bias the analysis. First, the computational prediction of HLA-binding can be inaccurate especially for certain HLA alleles (18, 19). Second, previous works have suggested that the overrepresentation of highly similar sequences due to collection bias in the IEDB could influence the analysis results (20, 21). Consequently, we kept allele–peptide pairs if the binding of the peptide to the reported HLA allele was also verified empirically and excluded similar sequences using an iterative method (*Methods* and *SI Appendix*, Fig. S1). Importantly, we also excluded peptides having controversial assay results or discordant results for different HLA alleles from both datasets. We defined peptides with exclusively negative assay results as nonimmunogenic and peptides with dominantly positive assay results as immunogenic (for detailed curation and filtering steps, refer to *Methods* and *SI Appendix*, Fig. S1).

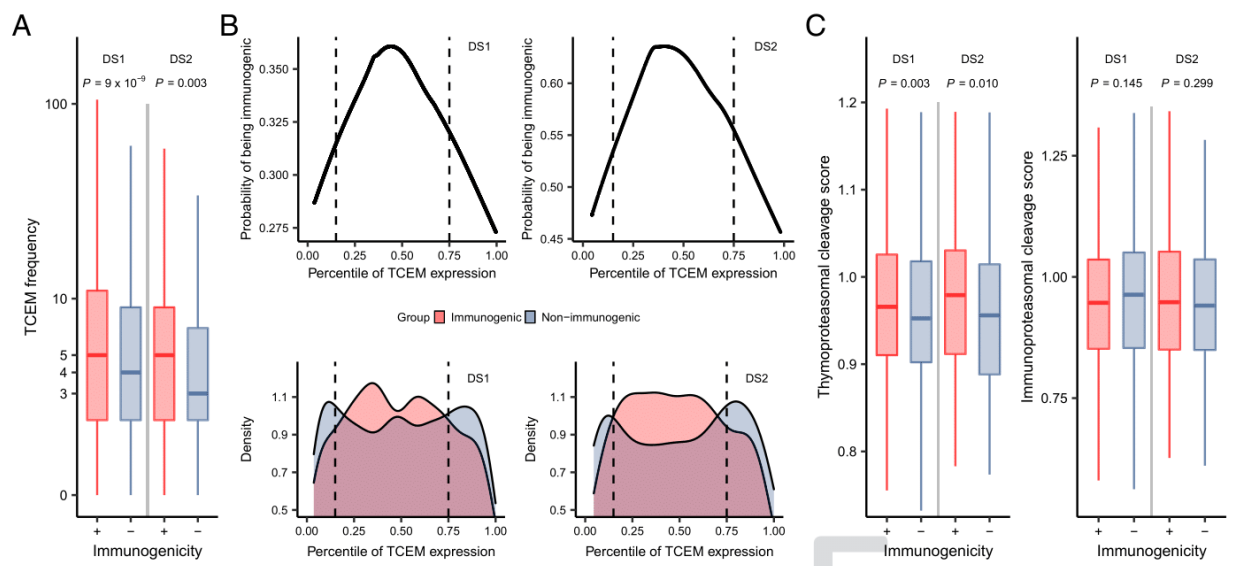
The number of immunogenic and nonimmunogenic peptides was 1,093 and 2,287 in the first dataset and 360 and 275 in the second one. Peptides in the second dataset had high diversity and covered the sequence space more homogeneously after excluding similar sequences (*SI Appendix*, Fig. S2). We analyzed the two nonoverlapping datasets in parallel to present our findings on a large number of peptides (dataset 1) and to ensure that they are not confounded by computational prediction or the presence of similar sequences (dataset 2).

TCEMs Occurring Very Rarely or Missing from Human Proteins Are Less Likely to Be Immunogenic. The positive selection of T cells is mediated by peptide sequences found in the human proteome. Certain amino acids of presented peptides are buried in the binding pockets of HLA molecules, and only five amino acids are in contact with TCRs (2–5). These sequence motifs mediate the recognition of presented peptides (2, 22, 23) and, consequently, the positive selection of T cells in the thymus. We expected that motifs very rarely or not found in the human proteome are less likely to be immunogenic because specific T cells are potentially missing from the repertoire as their precursors did not survive positive selection. Similarly to previous studies (2, 4, 22), we defined TCEMs as five amino acid–long sequences between the anchoring positions of presented peptides (*Methods*). We then determined their frequency in the reference human proteome. We aimed to use TCEM frequency in human proteins as a proxy of their presentation on the cell surface. The prevalence of TCEMs in human proteins showed a long-tailed distribution: a large fraction of motifs was rarely or not found in the human proteome, but many still reached high frequencies (*SI Appendix*, Fig. S3). Next, we collected data of immunopeptidomics studies (Dataset S2) and found that the frequency of TCEMs in human proteins can accurately predict their occurrence in HLA-I–bound peptides on the cell surface (*SI Appendix*, Fig. S4). Importantly, the analysis suggested that TCEMs found less than four times in human proteins are unlikely to be presented on the cell surface (*SI Appendix*, Fig. S4).

In line with expectation, nonimmunogenic peptides contained TCEMs that are very uncommon or not found in the human proteome (Fig. 1*A*). Accordingly, motifs occurring less than four times in the human proteome were less likely to be immunogenic than others. (Fig. 1*D*). The result suggests that TCEMs need to occur in sufficient numbers in human proteins to be recognized by the immune system. Otherwise, specific T cells are potentially absent from the repertoire as their precursors have not survived positive selection.

TCEMs That Are Not Expressed in cTECs Are Less Likely to Be Immunogenic. In the subsequent analyses, we focused on motifs occurring at least once in the human proteome. It was reported that the HLA-I presentation of peptides is highly dependent on the expression of the encoding gene (25). We assumed that TCEMs encoded by genes having low or undetectable expression in cTECs cannot mediate the positive selection of specific T cells. At the same time, we did not expect an immune response to TCEMs encoded by abundantly expressed housekeeping genes, because the response to these TCEMs may be blocked by central or peripheral immune tolerance (26, 27). In sum, we expected a bimodal relationship between the expression of TCEM-encoding genes and immunogenicity. We downloaded gene expression data of human cTECs from a recent study (28). For each TCEM, we determined the proteins containing its sequence. We then calculated the median expression of genes encoding these proteins to approximate the chance for a given TCEM of being expressed in cTECs. To examine the potentially bimodal relationship between TCEM expression and T cell activation, we plotted the probability for a TCEM of being immunogenic as a function of its expression using lowess smoothing (Fig. 1*B*). We also examined the distribution density of TCEM expression in the immunogenic and nonimmunogenic peptide groups separately (Fig. 1*B*). We found that in line with expectation, TCEMs having either low or high expression in cTECs are similarly less likely to activate T cells than the ones in the medium expression group (Fig. 1 *B* and *D*). These results suggest absent T cell responses to TCEMs that are not expressed at the site of T cell positive selection. As expected, TCEMs in the high expression group were more likely to be found in proteins encoded by housekeeping genes (*SI Appendix*, Fig. S5).

TCEMs That Are Not Presented on cTECs after Proteasomal Cleavage Are Less Likely to Be Immunogenic. Even if a given TCEM is expressed in cTECs, proper proteasomal cleavage is essential for its presentation on the cell surface by HLA molecules. Proteasomal cleavage is special in cTECs (6, 8, 29). Thymoproteasomes are exclusively expressed in these cells and are responsible for the generation of peptides that mediate the positive selection of T cells. In contrast with constitutive and immunoproteasomes, thymoproteasomes have a reduced ability to cleave peptide bonds after hydrophobic amino acids (i.e., they have lower chymotrypsin-like activity) (8). A previous study reported the amino acid preference of thymo- and immunoproteasomes around their cleavage sites (8). Using the presented data, we approximated the probability of thymo- and immunoproteasomal cleavage at each position of the reference human proteome. We then calculated a score associated with the chance of a given TCEM being generated after thymoproteasomal cleavage and,



D

	TCEM frequency < 4	TCEM expression < 15%	TCEM expression > 75%	Thymoproteasomal clevage score < 25%	Immunoproteasomal clevage score < 25%
Dataset 1 OR (P)	0.68 (3.2 x 10^{-7})	0.71 (0.002)	0.74 (0.001)	0.77 (0.004)	1.01 (0.89)
Dataset 2 OR (P)	0.61 (0.003)	0.48 (0.003)	0.55 (0.003)	0.67 (0.041)	1.00 (1.00)

Fig. 1. Peptide immunogenicity is influenced by TCEM frequency in human proteins (*A*), TCEM expression in cTECs (*B*), and TCEM presentation on cTECs (*C*). (*A*) The plot indicates the number of times immunogenic (+, $n = 1{,}093$ and 360 in datasets S1 and S2, respectively) and nonimmunogenic (–, $n = 2{,}287$ and 275 in datasets 1 and 2, respectively) TCEMs found in human proteins. In both datasets, TCEMs of immunogenic peptides were found more times in human proteins than TCEMs of nonimmunogenic ones. Outliers are not shown for visualization purposes. (*B*) The upper plots show the probability of a TCEM being immunogenic as the function of its expression in cTECs. The curves were fitted using lowess regression (24). The lower plots indicate the probability density of the expression of immunogenic ($n = 997$ and 326 for datasets 1 and 2, respectively) and nonimmunogenic ($n = 2{,}040$ and 247 for datasets 1 and 2, respectively) TCEMs. For visualization purposes, gene expression values were transformed by calculating their percentile rank. Vertical dashed lines indicate cutoff values used for OR calculation in *D*. (*C*) The likelihood of TCEM formation after thymoproteasomal (*Left*) and immunoproteasomal (*Right*) cleavage is shown. TCEMs of immunogenic peptides were more likely to be generated and presented after thymoproteasomal but not immunoproteasomal cleavage. $n = 997$ and 327 for immunogenic and 2,046 and 248 for nonimmunogenic TCEMs in datasets 1 and 2, respectively. Outliers are not shown for visualization purposes. (*D*) Peptides were classified based on their TCEM's frequency in human proteins, expression in cTECs, and thymo- or immunoproteasomal cleavage scores. TCEMs found rarely in the human proteome, having low expression in cTECs or low thymoproteasomal cleavage score, were less likely to be immunogenic. *P* values of two-sided Fisher's exact tests are shown. In panels *A* and *C*, the *P* values of two-sided Wilcoxon's rank-sum tests are indicated. On *A* and *C*, the bottom and top of boxes indicate the first and third quartile, horizontal lines indicate median, and vertical lines indicate first quartile − 1.5*IQR and third quartile + 1.5*IQR. DS1: dataset 1, DS2: dataset 2.

thus, presented on the surface of cTECs (*Methods* and *SI Appendix*, Fig. S6). We expected lower immunogenicity for TCEMs that are less likely to be presented on the surface of cTECs after thymoproteasomal cleavage. At the same time, we expected no effect of immunoproteasomal cleavage on immunogenicity because the immunoproteasome has only minor importance in cTECs (8). In line with expectation, TCEMs of immunogenic peptides were more likely to be generated by thymoproteasomal cleavage than nonimmunogenic ones, while immunoproteasomal cleavage did not affect immunogenicity (Fig. 1*C*). Accordingly, TCEMs that are unlikely to be presented on cTECs were less immunogenic (Fig. 1*D*).

The Robustness of Results. We reported three lines of evidence suggesting that the positive selection of T cells results in a defective T cell repertoire with implications on the recognition of nonself peptides. First, TCEMs that are very rare or not found in the human proteins are less likely to be immunogenic (Fig. 1 *A* and *D*). Second, the scarce expression of TCEMs in cTECs is also associated with lower immunogenicity (Fig. 1 *B* and *D*). Third, TCEMs that are improbably generated by the cTEC-specific thymoproteasome are less likely to be immunogenic (Fig. 1 *C* and *D*).

These effects on immunogenicity held in multivariate logistic regression models, indicating that they are not confounded by and independent of each other (*SI Appendix*, Fig. S7 and Table S1). Similarly, the effect of these attributes was additive: rare TCEMs having low expression in cTECs and low thymoproteasomal cleavage score were less likely to be immunogenic than TCEMs having only one or two of these attributes (*SI Appendix*, Table S2).

Next, we tested whether our findings are confounded by a single amino acid with a peculiar effect on immunogenicity. First, we examined the prevalence of the 20 amino acids in immunogenic and nonimmunogenic TCEMs. The most significant difference was found for tyrosine and phenylalanine enriched in

nonimmunogenic motifs and glycine and alanine enriched in immunogenic ones (*SI Appendix*, Table S3). This is in line with expectation as the former amino acids are rarely while the latter ones are commonly found in human proteins (*SI Appendix*, Table S3). Surprisingly, tryptophan, the rarest amino acid, was more common in immunogenic TCEMs, which can be explained by its major role in peptide immunogenicity (30–32). Reassuringly, this phenomenon had no effect on our results as all findings remained significant when we iteratively repeated the analysis by excluding TCEMs containing certain amino acids (*SI Appendix*, Table S3). The hydrophobicity of TCR contact residues is reported to influence peptide immunogenicity (33) and could confound our results. The effect of all TCEM attributes on immunogenicity remained significant when controlling for hydrophobicity in logistic regression models (*SI Appendix*, Fig. S8).

Finally, it is reported that peptides bind to certain HLA variants with secondary anchors in their TCEM region (34). We determined these HLA variants using data from a recent immunopeptidomics study (35) (*Methods*). All of our results held after excluding peptides that bind to all reported HLA variants with secondary anchors at the TCEM region (n = 69 and 16 in datasets 1 and 2, respectively; *SI Appendix*, Fig. S9 and Dataset S1).

The Frequency, Expression, and Presentation of TCEMs Determine the Prevalence of Specific Naïve CD8+ T Cells in the Repertoire. To confirm our previous findings, we aimed to directly demonstrate the predictions of our hypothesis. Specifically, we expected that it is less likely to find a given naïve T cell in the repertoire that is specific for infrequent TCEMs in human proteins, for TCEMs not expressed in cTECs, or for TCEMs not presented on the surface of cTECs. We used recently published data on the peptide-specificity of naive CD8+ T cells in the repertoire of healthy individuals (*Methods*) (36). The authors reported the prevalence of naïve CD8+ T cells specific for any of the examined 296 nine amino acid–long (9-mer) SARS-CoV-2 peptides in the repertoire of 27 individuals. In the case of each individual, we focused on sequences that were bound by at least one of their HLA-I alleles (*Methods*), because we expected to find specific T cells for HLA-presented peptides only. We grouped the peptides based on the prevalence, expression, and proteasomal cleavage scores of their TCEMs as described previously for T cell activation datasets. For each individual and in each peptide group, we determined the fraction of HLA-presented peptides that are recognized by at least one TCR in the repertoire. Specific naive CD8+ T cells were less likely to be present for rare than nonrare TCEMs in the repertoire of healthy individuals (Fig. 2*A*). Similarly, it was less likely to find specific T cells for TCEMs having either negligible or overly high expression in cTECs (Fig. 2*B*). Moreover, TCEMs with low thymoproteasomal cleavage scores were less likely to be associated with the presence of specific T cells in the repertoire (Fig. 2*C*), while the immunoproteasomal cleavage score did not show this relationship (Fig. 2*D*). In sum, our findings for T cell repertoires of healthy individuals confirmed the presented results on the T cell activation datasets.

Decreased Immunogenicity of Overly Dissimilar Peptides to Human Proteins. Our hypothesis predicted a rather provocative relationship: in contrast with expectation, overly dissimilar peptides are not recognized by the immune system, because self-peptides mediate the positive selection of specific T cells. To test this prediction, for each peptide in the T cell activation datasets, we determined its most similar counterpart in the human proteome using the BLAST software as in previous studies (37, 38) (*Methods*). As expected, peptides with very rare TCEMs (occurring zero to three times) in the human proteome had lower similarity to human proteins than other peptides (Fig. 3*A*). Accordingly, overly dissimilar peptides of datasets 1 and 2 were less likely to be immunogenic just like highly similar ones (Fig. 3*B*). Importantly, we found the same relationship when analyzing dissimilarity values of an independent study published recently (37) (*SI Appendix*, Fig. S10). To corroborate these results, we analyzed the self-similarity of SARS-CoV-2 peptides. Reassuringly, we found naive CD8+ T cells that are specific for highly dissimilar peptides in the repertoire of fewer individuals (Fig. 3*C*).

We conclude that while a given level of peptide dissimilarity to human proteins is essential for self-nonself discrimination, overly dissimilar peptides are less likely to be recognized by the immune system, because specific T cells are not present in the repertoire.

Cross-Reactivity Is Not Able to Compensate for the Side Effect of Self-Mediated Positive Selection of T Cells. We propose that the mechanism of positive selection results in a defective T cell repertoire. Is the cross-reactivity of TCRs able to compensate for these defects? To answer this question, we first created two groups of TCEMs (Fig. 4*A*). The first group consisted of motifs in datasets 1 and 2, for which we assumed that it is the least likely to find specific positively selected T cells in the repertoire (n = 43, they were nonimmunogenic, found less than four times in the human proteome, and had low expression in cTECs and low thymoproteasomal cleavage score; *SI Appendix*, Table S2). The second group consisted of all possible TCEM sequences (n = 323,470), for which the presence of specific T cells in the repertoire is likely: they were found more than three times in the human proteome, expressed in cTECs, and had normal thymoproteasomal cleavage score. For each TCEM in the first set, we calculated its BLOSUM62 similarity to every TCEM in the second set to explain their proximity in sequence space (Fig. 4*A*).

Next, we estimated the level of TCR cross-reactivity in the sequence space of TCEMs. We downloaded empirical data from a recent study. The authors measured the binding strength of the well-known NY-ESO-1 epitope to TCR C^{259}, when sequentially replacing every amino acid in different positions of the epitope (23). We determined the BLOSUM62 similarity between the TCEM of the original and the modified peptide sequences and found a strong positive correlation between the similarity of the original to the modified TCEM and the peptide binding strength to TCR C^{259} (Fig. 4*B*).

We then aimed to determine a TCEM similarity cutoff value, under which the binding to the TCR is too weak to induce T cell activation (Fig. 4*A*). To this end, we examined the relationship between the TCR binding strength and the activation of T cells which was reported in the same study. We fixed less than 10% of the original binding strength as insufficient binding because the ability of peptides to activate T cells was negligible below this cutoff (*Methods*). We found that the similarity between the modified and the original TCEM can accurately predict whether the peptide will be bound by the TCR strong enough to cause T cell activation (Fig. 4*C*). We then used an established "cost–benefit" method (39) to determine the optimal TCEM similarity cutoff for binding (*Methods*). We considered this value as the lowest similarity between two TCEM sequences that can be bridged by the examined TCR. Reassuringly, we got a very similar cutoff value, when using data of an independent study on the A6 TCR and its target epitope, the Tax peptide of HTLV-1 (40) (*SI Appendix*, Fig. S11 and *Methods*). As the result of this analysis, we had an estimate on the magnitude of cross-reactivity of a given TCR in sequence space.

We then determined whether T cells that are specific for TCEMs in the second group are able to bind TCEMs in the first group (Fig. 4*D*). We found that only an insignificant minority (ranging from 0.0006 to 0.043% for TCEMs in the first group,

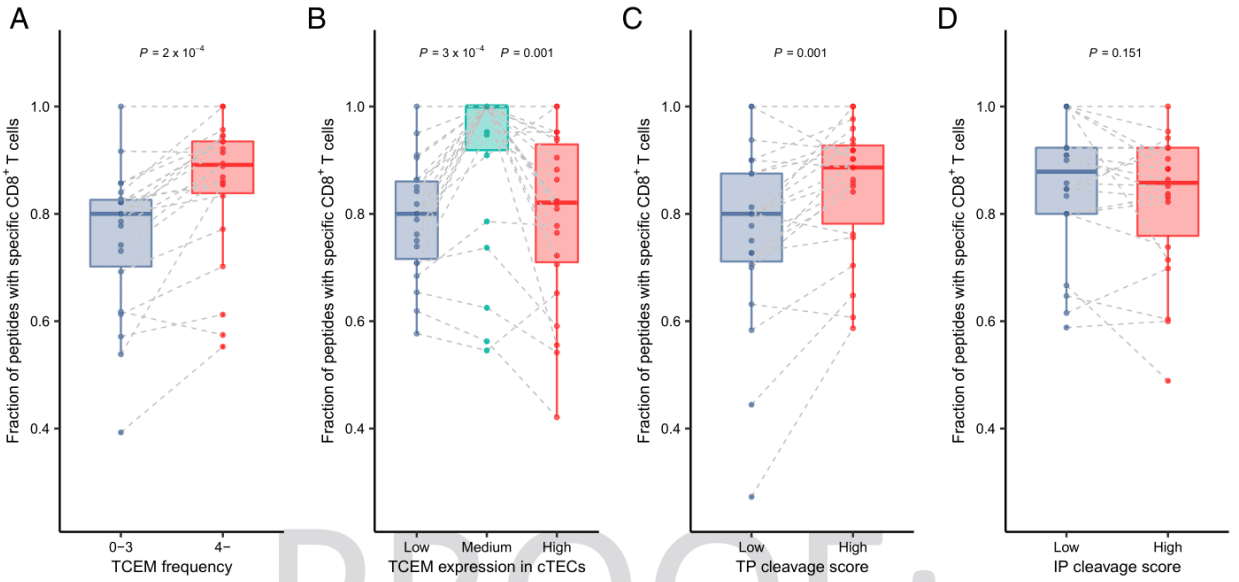


Fig. 2. Specific naive CD8+ T cells were less likely to be present for TCEMs found rarely in human proteins (*A*), having low expression in cTECs (*B*) or low thymoproteasomal cleavage score (*C*). The vertical axes represent the fraction of peptides, for which specific T cells were detected. Point pairs (or triplets on *B*) indicate values belonging to a given individual ($n = 22$). Two-sided *P* values of paired Wilcoxon's rank-sum tests are shown. TCEMs were stratified into expression groups based on tertiles and into thymoproteasomal (TP) or immunoproteasomal (IP) cleavage score groups based on the first quartile. The bottom and top of boxes indicate the first and third quartile, horizontal lines indicate median, and vertical lines indicate first quartile - 1.5*IQR and third quartile + 1.5*IQR.

median = 0.015%) of similarity values reached the previously determined cutoff values of T cell cross-reactivity (Fig. 4*D*). This result suggests that T cells in the repertoire (specific for TCEMs in the second group) are not likely to recognize TCEMs, whose recognition is negatively affected by self-mediated positive selection (i.e., TCEMs in the first group). Although the result is indicative, it is important to highlight that we inferred cross-reactivity based on the data for two TCRs, and the results need future validation using data of more TCRs.

Positive Selection of T Cells and Susceptibility to Infections. The adaptive immune recognition of pathogen-associated peptides is essential for the initiation of an effective immune response. Our results suggest that many such peptides are potentially non-immunogenic, because specific T cells are not found in the repertoire of CD8+ T cells. We aimed to determine the frequency of peptides in proteins of intracellular pathogens that could be affected by the side effect of T cell positive selection to some extent. To this end, we downloaded reference proteomes of 50 common intracellular pathogens. In the proteome of each species, we determined the prevalence of TCEMs that are either rare or not found in human proteins, not or lowly expressed in cTECs, or unlikely to be presented after thymoproteasomal cleavage (we call them np-TCEMs hereafter, referring to TCEMs for which we expect to find specific positively selected T cells with lower probability). We found that the frequency of np-TCEMs is ranging from 58 to 71% in different species. (Fig. 5*A* and Dataset S3).

The high prevalence of np-TCEMs could hinder immune recognition, especially when only a few peptides of the pathogen are presented to T cells. This might be the case when either the proteome of the given pathogen is small and/or the given HLA allele has a narrow binding repertoire. To this end, we predicted the binding of all 9-mer peptides to common HLA-I alleles (*Methods*). For each alleles–pecies pair, we calculated the fraction of np-TCEMs in the presented peptides and visualized the results on a heatmap (Fig. 5*B*). As expected, the fraction of presented peptides with np-TCEMs was highly variable between HLA alleles when the pathogens had small proteomes (Fig. 5*C*). This group of pathogens was dominated by viruses, like Human parvovirus B19, Hepatitis viruses, Human papillomavirus, etc. On the contrary, HLA alleles presented a similar fraction of np-TCEMs from the large proteomes of protozoal and bacterial species (Fig. 5*C*).

We expected that the fraction of HLA-presented np-TCEMs could influence disease risk. To this end, we carried out literature mining to find HLA association meta-analysis data (*Methods*). We found such data for chronic hepatitis B (41), human papillomavirus (42), and dengue virus (43) infection, and Hepatitis C viral persistence after IFN-alpha therapy (44). We selected allele groups with positive or negative associations and determined the prevalence of np-TCEMs in HLA-bound peptides of the causative pathogens. In contrast with protective alleles, risk HLA allele groups bound peptides with dominantly np-TCEMs (Fig. 5*D*).

These results suggest that the proposed side effect of T cell positive selection influences the adaptive immune recognition of intracellular pathogens.

Discussion

The prevalence of specific T cells in the repertoire is essential for adaptive immune recognition of HLA-presented peptides. It has been suggested that during positive selection, self-peptides on the surface of cTECs can be considered as a test set for thymocytes: cells that recognize these peptides survive potentially recognize nonself peptides more effectively and consequently dominate the immune response to the foreign antigens (9–11). We propose that the nonresponsiveness to many nonself peptides can also be explained by the mechanism of T cell positive selection because it is mediated by self-peptides. In other words,

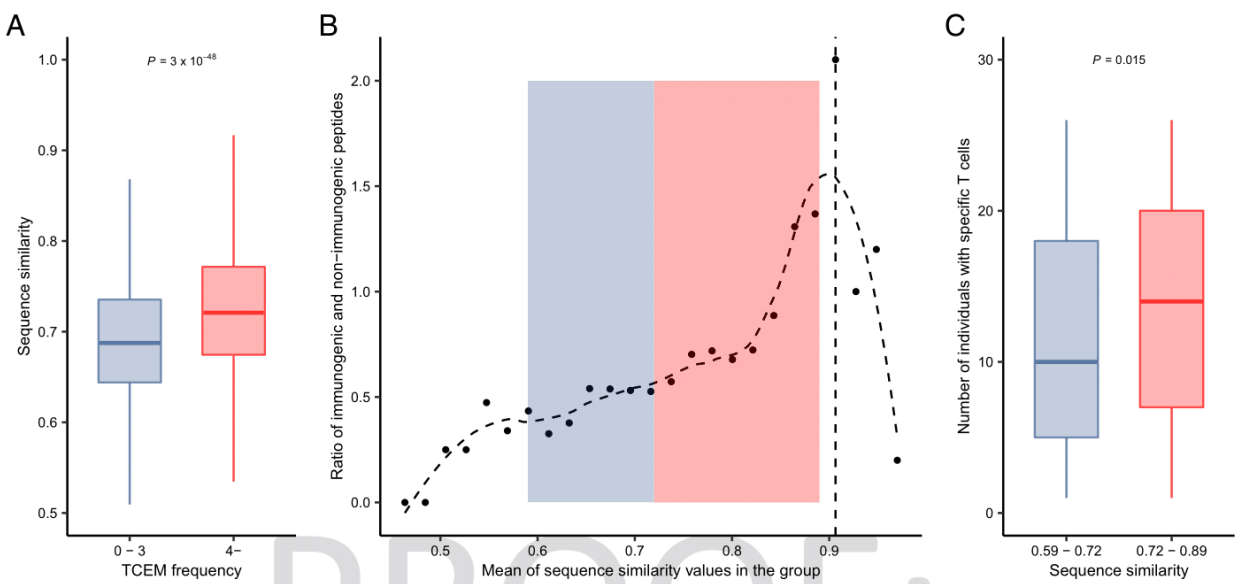


Fig. 3. Overly dissimilar peptides to human proteins are less immunogenic. (*A*) Peptides in datasets 1 and 2 with TCEMs found less than four times in human proteins are less similar to the closest hit in the human proteome. (n = 1,706 and 2,309 in the 0- through 3- and 4-TCEM frequency groups, respectively) Outliers are not shown for visualization purposes. (*B*) Peptides in datasets 1 and 2 were pooled and stratified into 25 groups based on similarity. In each group, the ratio between immunogenic and nonimmunogenic peptides was calculated. Groups are shown in increasing order of similarity. The horizontal axis indicates the mean similarity in the given group. The vertical dashed line indicates the group having the highest fraction of immunogenic peptides. The curve was fitted with a cubic smoothing spline method in R (*Methods*). Background shading represents the similarity ranges of peptide groups on *C*. (*C*) T cells specific for overly dissimilar peptides were found in the repertoire of fewer individuals. Peptides were stratified into sequence similarity groups based on the median value (n = 149 and 147 in the lower and higher similarity groups, respectively). The similarity ranges are also indicated on *B* with background colors. Note that the dataset of SARS-CoV-2 peptides did not include peptides that are highly similar to human proteins. On *A* and *C*, *P* values of two-sided Wilcoxon's rank-sum tests are indicated. The bottom and top of boxes indicate the first and third quartile, horizontal lines indicate median, and vertical lines indicate first quartile − 1.5*IQR and third quartile + 1.5*IQR.

self-mediated positive selection has a negative trade-off on the recognition of foreign peptides. Importantly, our results suggest that T cell cross-reactivity is unable to compensate for this negative consequence of positive selection (Fig. 4).

We presented three lines of evidence supporting our hypothesis on two reliable and nonoverlapping peptide sets (Fig. 1). We focused on the five amino acid–long TCEM region of peptides, because numerous studies suggested that self–nonself discrimination is governed by these short motifs (2–5). Importantly, our analysis on TCR cross-reactivity supports these findings: the TCEM sequences of the modified NY-ESO-1 peptides alone were able to determine the binding of the peptide to the TCR C^{259} (Fig. 4*B*). At the same time, it has been at issue how such short peptides can make it possible for the immune system to differentiate between self and nonself peptides (2, 45). Namely, human peptides contain around 75% of all possible pentamer sequences (2) (73.1% in our analysis) that largely overlap with the ones found in commensal and pathogenic bacteria (2). Our findings suggest that the overlap between self and nonself motifs is far from being disadvantageous; in fact, it is crucial for the positive selection of T cells that are specific for foreign peptides. In other words, the overlap between motifs makes it possible to recognize nonself.

It is an important issue to be clarified whether our findings are affected by regulatory T cell (Treg) activation. Namely, the positivity of T cell assays in our in vitro datasets could reflect the activation of Treg cells to some extent, which could explain the positivity of assays for self-similar peptides. However, peptide immunogenicity in our datasets is supported by dominantly IFN-gamma ELISpot assays (Dataset S1). Although a small subset of induced Treg cells is able to produce IFN-gamma (46), clear positivity of IFN-gamma ELISpot assays is predominantly associated with inflammatory but not tolerogenic responses (38). Consequently, it is not likely that our results are confounded by Treg cell activation.

We aimed to support our hypothesis with direct evidence by examining naïve CD8+ T cell repertoires of healthy individuals (Fig. 2). The results suggest that self-mediated positive selection has a negative effect on the prevalence of SARS-CoV-2, peptide-specific T cells in the repertoire (Fig. 2). Importantly, it has already been suggested that "holes" in the T cell repertoire hinder the recognition of certain pathogens (47–50), and the studies explained the presence of such holes by central tolerance (47, 50). On the other hand, Yu et al. suggested that there are no significant holes in the repertoire, because clonal deletion affects only the most self-reactive T cells (51). Consequently, every possible HLA-presented peptide could be recognized by T cells, but many T cells are anergic due to immune tolerance mechanisms. However, our results suggest that the self-dependent positive selection of T cells causes gaps in the immune recognition of nonself peptides potentially through a biased T cell repertoire.

We also estimated the fraction of peptides in pathogens whose recognition could be affected by self-mediated positive selection to some extent. A significant proportion of peptides—varying between 58% and 71% in different species—fell into this category (Fig. 5*A*). If we also consider that around one-third of nonself peptides are indistinguishable from self-ones due to high similarity (52), it is not surprising that at least 50% of HLA-A*02:01–presented vaccinia and HIV sequences were reported to be nonimmunogenic in previous studies (47, 52, 53). At the same time, which peptides are presented to the T cells depends on the specificity of HLA alleles. We showed that HLA alleles that predominantly present peptides whose recognition is

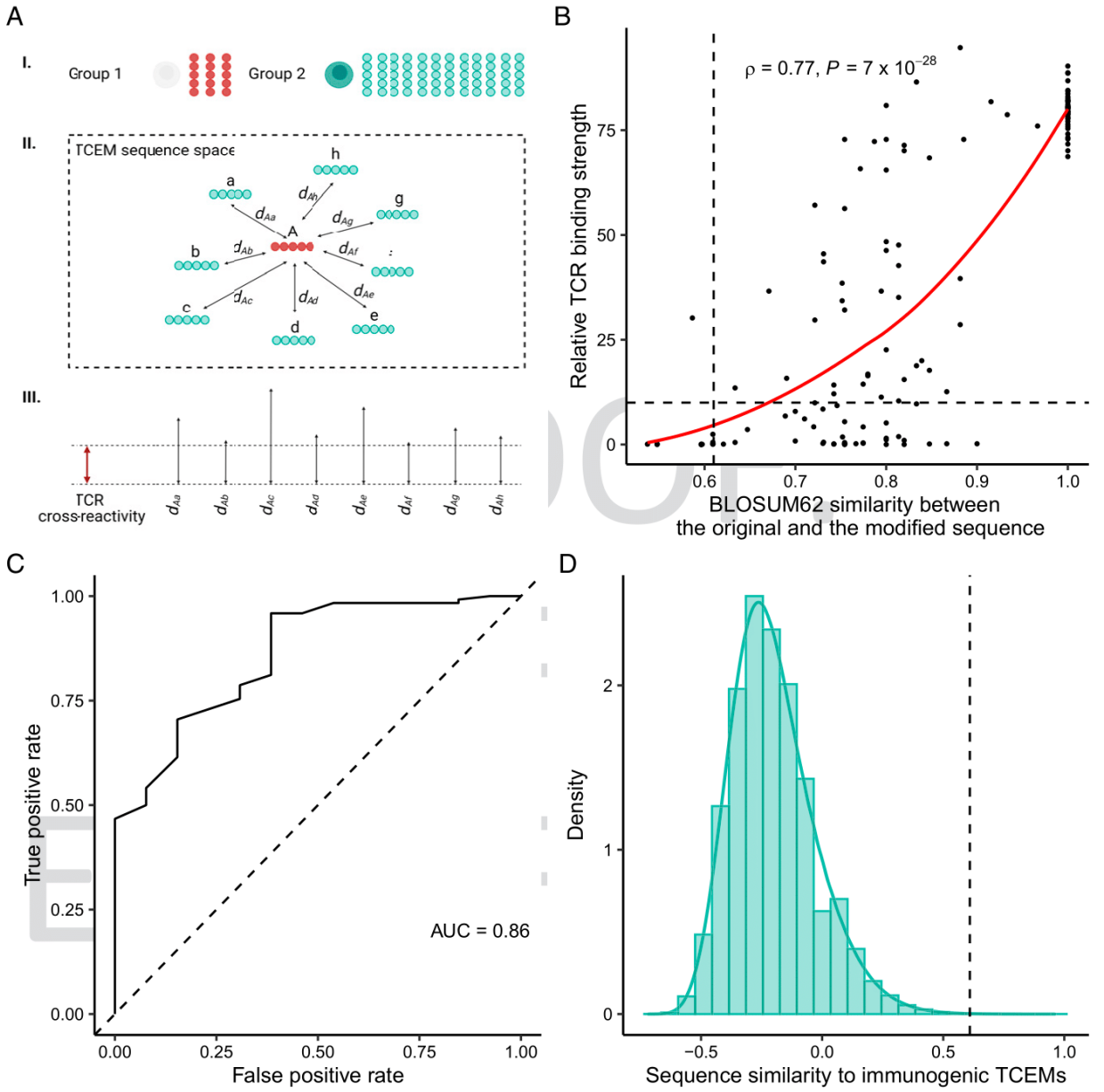


Fig. 4. TCR cross-reactivity is not likely to compensate for the defects in the T cell repertoire. (*A*) Schematic diagram of the analysis. To determine whether T cell cross-reactivity is able to bridge defects in the repertoire, we created two groups of TCEMs (I). The first group consisted of motifs, for which we assumed that it is the least likely to find specific positively selected T cells in the repertoire (marked with red color on the figure). The second group consisted of all TCEMs, for which we expected to find specific T cells in the repertoire (marked with green color on the figure). We calculated the pairwise similarity between the members of the two TCEM groups (II). The higher the similarity between two TCEMs, the closer they reside in the sequence space resulting in smaller distance (*d*) values. Next, we estimated the level of T cell cross-reactivity in TCEM sequence space (III). We defined cross-reactivity as the lowest similarity between a given TCEM sequence and the TCR's cognate TCEM sequence that is needed for a reasonable TCR binding strength and T cell activation. Finally, we determined the number of cases when members of the first and second groups are close enough to be recognized by the same TCR. (*B*) The amino acids of the NY-ESO-1 epitope were sequentially changed, and the binding strength to TCR C^{259} was measured in a previous study (23). The relative binding strength of the modified ($n = 135$) and the original peptide to TCR C^{259} is shown as a function of BLOSUM62 similarity between their TCEM sequences. The horizontal line indicates 10% of the original binding value, which was considered as a cutoff for improbable binding (*Methods*). Spearman's rho and the *P* value of a two-sided correlation test are indicated. The red line indicates a smooth curve fitted using a cubic smoothing spline method in R (*Methods*). (*C*) The ROC curve demonstrates the accuracy of BLOSUM62 similarity in classifying peptides into binding and nonbinding (i.e., lower than 10% of original binding) groups. AUC: area under the curve. (*D*) The density of all similarity values between TCEMs in group 1 and group 2 ($n = 323{,}470$). Vertical lines on *A* and *C* represent the optimal cutoff (0.61) for classification.

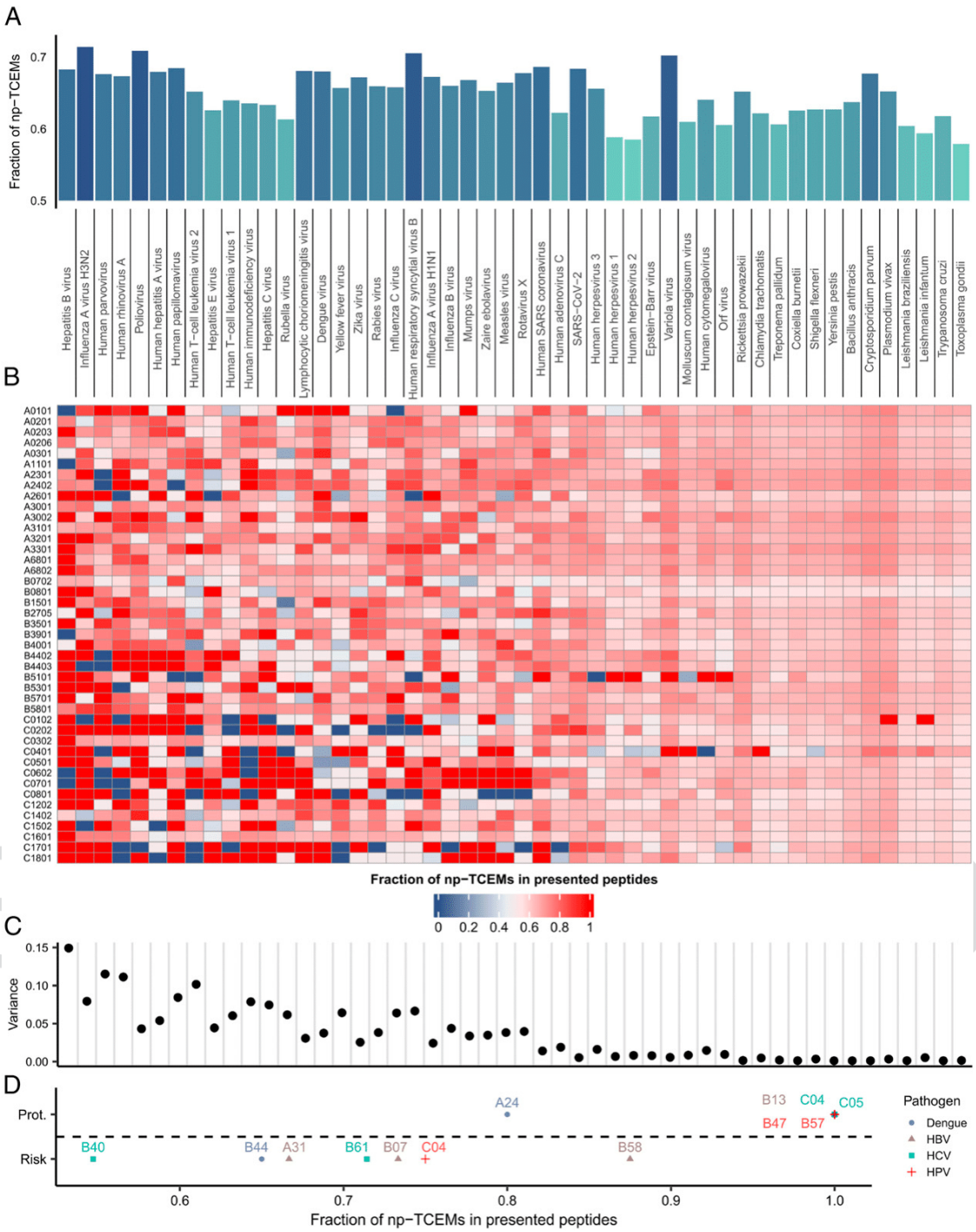


Fig. 5. The effect of self-mediated positive selection on the recognition of pathogens. (*A*) The prevalence of np-TCEMs in the proteome of different pathogens (n = 50). np-TCEMs were defined as being found less than four times in the human proteome or having low expression in cTECs or low thymoproteasomal cleavage score. Pathogens are ordered by increasing proteome size. (*B*) The heatmap shows the prevalence of np-TCEMs in peptides of intracellular pathogens that are presented by common HLA alleles. (*C*) The plot shows the variance of presented np-TCEMs by different HLA alleles. The variance decreases with increasing proteome size of pathogens (Spearman's rho: −0.93, two-sided correlation test $P = 9.13*10^{-23}$). (*D*) The fraction of np-TCEMs in peptides that are presented by risk (n = 6) and protective (n = 7) HLA allele groups. Group-specific values were calculated by averaging values for common alleles in each group (*Methods*). In contrast with protective allele groups, predisposing ones present mainly peptides with np-TCEMs in their sequence (two-sided Wilcoxon's rank-sum test $P = 0.004$). HBV: hepatitis B virus; HCV: hepatitis C virus; and HPV: human papillomavirus. For full pathogen names, refer to Dataset S3.

potentially hindered by self-mediated positive selection are associated with risk for certain infections (Fig. 5*D*). To note, a similar mechanism could also explain variable responses to vaccines which need further investigation.

Finally, our results do not support a common interpretation of self-nonself discrimination, which suggests that the more dissimilar a peptide to self, the more likely it is to be immunogenic (12–15). We showed that the more dissimilar a peptide to human proteins, the less likely it is to find its TCEM in the human proteome (Fig. 3*A*). Consequently, specific positively selected T cells are potentially absent from the repertoire above a level of dissimilarity (Fig. 3 *B* and *C* and *SI Appendix*, Fig. S10). We conclude that although a certain level of dissimilarity is essential for the discrimination of self and nonself, overly dissimilar peptides are potentially not recognized by the immune system (Fig. 6). While our results indicate the importance of this blind spot in the immune response to infections, it is a question to be clarified in future works, whether mutated cancer peptides can also reach this level of dissimilarity. Similarly, testing our hypothesis on HLA-II presented peptides and CD4+ T cells is also an important area of future research.

Methods

Collecting and Filtering Peptide Immunogenicity Data. We collected HLA binding and T cell activation data from the IEDB (17). The IEDB contains experimental data on T and B cell epitopes. Data on MHC binding and T cell specificity are continuously collected from the literature or directly submitted by researchers working in the field. The database is strictly curated and has a standardized decision algorithm to determine whether a given assay is positive or not (55, 56). The authors of the database always refer to experts when they are facing novel assays or immunological content (56). Consequently, the positivity of an assay always means that the interaction between the adaptive immune receptor and the peptide is highly probable (55, 56). We downloaded raw T cell assay results from the website (as of February 3, 2020). We selected nine and 10 amino acid–long linear nonhuman peptides containing only the 20 standard amino acids and tested for HLA-I alleles genotyped with at least 4-digit resolution (*SI Appendix*, Fig. S1). It is important to note that the inclusion of human peptides in our analysis could severely confound our results, because 1) specific positively selected T cells are more likely to be found for these peptides and 2) they are dominantly nonimmunogenic due to central and peripheral tolerance mechanisms (16).

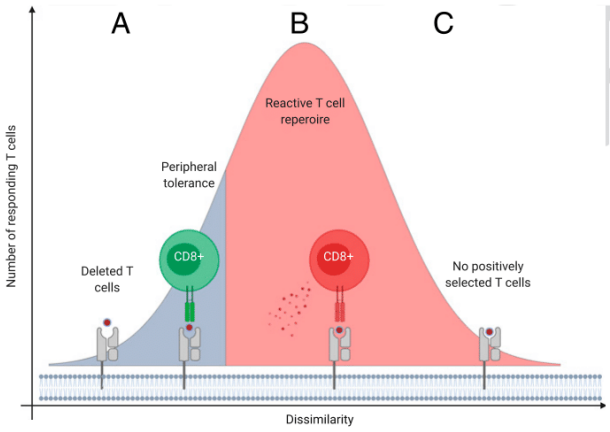


Fig. 6. The blindness of immune recognition for peptides that are overly dissimilar to human proteins. (*A*) The immune system tolerates peptides that are similar to self-proteins. T cells recognizing these peptides are either deleted in the thymus or unresponsive due to peripheral tolerance mechanisms (54). (*B*) Peptides with a certain level of dissimilarity to human proteins are recognized as nonself resulting in T cell activation and immune-mediated destruction of cells (Fig. 3*B* and *SI Appendix*, Fig. S10). (*C*) Peptides that are overly dissimilar to human proteins are not recognized by the immune system, because specific positively selected T cells are absent from the repertoire (Fig. 3*B* and *SI Appendix*, Fig. S10).

Consequently, the presence of human peptides in our datasets could obscure the effect of T cell–positive selection on peptide immunogenicity. Next, we created two independent datasets. The first dataset was created using established methods (5). Specifically, we collected HLA allele–peptide pairs, in which the binding of the peptide by the HLA allele was supported by the prediction results of the NetMHCpan-4.0 algorithm (57) (either the binding affinity was lower than 500 nM, or the binding rank percentile was lower than 2%) (*SI Appendix*, Fig. S1). We considered it particularly important to confirm HLA-binding of peptides in a unified way using an accurate algorithm, because especially in older studies, HLA restriction of peptides was determined with less accurate computational methods.

To avoid the inaccuracy of computational prediction (18, 19), we created the second dataset using raw MHC binding assay data downloaded from the IEDB (17) (as of February 3, 2020). We collected allele–peptide pairs that were found in both MHC binding and T cell assay data at least twice. We considered a given peptide sequence as being bound by the given HLA allele if more than 60% of binding assays were positive. We also aimed to exclude similar sequences from the second dataset. To this end, we used a previously established iterative method yielding peptides with high sequence diversity (58). Briefly, the k-tuple distance between all peptide sequences was determined in each iteration using Clustal Omega 1.2 (59). Peptide pair(s) with the lowest distance values were determined, and the peptide having the lowest mean distance from all other sequences was excluded. We repeated these iterations until only peptides with at least 0.5 k-tuple distance from all other sequences remained in the dataset (60). This distance value corresponds to a maximum 50% overlap between sequences.

In both datasets, we defined allele–peptide pairs with solely negative T cell assays as nonimmunogenic and the ones with more positive than negative T cell assays as immunogenic. Allele–peptide pairs not meeting these criteria were excluded. Peptide sequences tested for multiple alleles but were also excluded with the opposite T cell activation results (*SI Appendix*, Fig. S1). To avoid any overlap between the two datasets, peptides found in both were only kept in the second one. Using the results of a recent large immunopeptidomics study (35), we identified HLA alleles, to which peptides bind with secondary anchor residues in the TCEM region. We considered a peptide position as an anchoring residue to a given HLA allele if the amino acid entropy at the position of allele-bound peptides was lower than 0.8.

Calculating TCEM Frequency, Expression, and the Probability of Thymo- and Immunoproteasomal Cleavage. According to previous studies (2–5), we defined TCEMs of nine amino acid–long peptides as amino acids from position 4 through 8. For 10 amino acid–long sequences, we defined TCEMs as amino acids from positions 5 through 9 according to sequence logos published in a recent immunopeptidomics study (35). We determined the frequency of TCEM sequences in human proteins as follows. We downloaded the reference human proteome from the UniProt database (61) (Proteome ID: UP000005640; only reviewed sequences are included, downloaded on January, 27th 2020). We decomposed each protein in the proteome to overlapping 9-mers, and for each 9-mer, we determined its TCEM sequence (2–5). We then quantified the incidence of every possible TCEM sequence ($n = 20^5$) in the human proteome.

To calculate the expression of TCEMs in cTECs, we first downloaded gene expression data of cTECs reported in a recent study (unadjusted counts file under GEO accession GSE127209) (28). We scaled columns of the count matrix using the calcNormFactors function in the edgeR R library (62, 63). Next, we calculated RPKM values using the rpkm function of edgeR and exon length data of the GenomicFeatures R library (64). Then, for each gene, we determined the median RPKM value in cTEC samples. We matched ENSEMBL gene IDs (used in the expression dataset) with UniProt IDs as follows. Direct conversion between IDs was unsatisfactory, as 40% of UniProt protein IDs in the dataset did not have a corresponding ENSEMBL gene ID in the downloaded expression set. Consequently, we first converted ENSEMBL gene IDs and UniProt IDs to HUGO IDs using the org.Hs.eg.db R library and protein information in the UniProt database, respectively. Next, we matched proteins with genes using HUGO IDs. With this approach, we were able to determine the expression of encoding genes for more than 90% of proteins. For each TCEM, we first determined genes encoding proteins that include the given TCEM in their sequence and calculated their median expression. If a given TCEM was found multiple times in the same protein, we included the expression of the encoding gene the same number of times in the calculation. We identified TCEMs encoded by housekeeping genes by using an established list of these genes (65). The relationship between expression and immunogenicity was plotted using lowess regression (24) implemented with the Hmisc R library. The probability density of expression values was plotted

after determining the smoothed kernel density estimate by the ggplot2 R library.

The probability of thymo- and immunoproteasomal cleavage was determined using amino acid prevalence data around the cleavage site provided by a previous study (8). Briefly, the authors carried out thymo- and immunoproteasomal digestion of three proteins and determined the fraction of amino acids found at the five positions toward the C and N termini around the cleavage site. They also provided amino acid frequencies in the three proteins that would be found if the proteins were randomly cleaved. We first normalized amino acid prevalence values at each position around the cleavage site by dividing them with their prevalence in the substrates yielding amino acid preference scores: $c_{i,j}$ referring to the score of amino acid j at position i around the cleavage site ($i \epsilon \{-5; -4; -3; -2; -1; 1; 2; 3; 4; 5\}$) (*SI Appendix*, Fig. S6*A*). Next, we determined the probability of proteasomal cleavage (C) at each site of the human proteome by calculating the median of $c_{i,j}$ values at positions around the cleavage site (*SI Appendix*, Fig. S6*B*). Then, for each 9-mer peptide in the human proteome, we determined the probability of peptide formation upon proteasomal cleavage by calculating the mean of C values ($\overline{C}$) at the two sites before the N-terminal and after the C-terminal amino acids (*SI Appendix*, Fig. S6*C*). Finally, for each TCEM, we calculated the median of $\overline{C}$ values associated with peptides containing the sequence of the given TCEM (*SI Appendix*, Fig. S6*D*). Note that the presented method was carried out for thymo- and immunoproteasomal cleavage separately.

For each peptide sequence in the study, the most similar sequence in the human proteome was identified using BLAST 2.9.0 (66) with the option of ungapped alignment. The BLOSUM62 similarity between the sequence pairs was calculated as described previously (54).

Analyzing SARS-CoV-2 TCR Sequencing Dataset. Data on antigen-specific naïve CD8+ T cells of 27 healthy individuals were downloaded from the website of Adaptive Biotechnologies (36). The authors cocultured naive CD8+ T cells of healthy donors with dendritic cells, loaded with a pool of all examined peptides of SARS-CoV-2. Then, they used the MIRA technology to identify antigen-specific T cells. The MIRA technology combines conventional T cell assays with immune repertoire sequencing to identify a large number of antigen-specific T cells in the repertoire simultaneously (67). We first determined HLA-I allele–peptide pairs, to which it applies that carrying the given HLA-I allele could potentially be associated with the prevalence of specific naive CD8+ T cells in the repertoire. Specifically, we predicted the binding of each examined peptide by all HLA alleles carried by any individuals using NetMHCpan-4.0. We associated the prevalence of T cells with carrying a given HLA allele if the predicted binding affinity and rank percentile values suggested strong binding (i.e., affinity was lower than 50 nM, and rank percentile was under 0.5%) and peptide-specific T cells were found in at least two individuals carrying the given allele. We used these strict criteria for HLA-binding to decrease false positive hits. For each individual, we determined the peptides, for which we expected to find specific T cells in the repertoire by considering the previously specified peptide–allele pairs and the HLA genotype of the individual. We included a given individual in the analysis only if we found specific T cells in the repertoire for at least 20 SARS-CoV-2 peptides. This filtering step yielded data on 22 individuals for further analysis.

Approximating the Level of T Cell Cross-Reactivity. We acquired data on T cell binding strength reported by two studies (23, 40). Both studies examined the shift in TCR binding strength when sequentially changing amino acids at each epitope position. We narrowed down our analysis to the TCEM sequence and determined the BLOSUM62 similarity between the TCEM of the original and the modified peptides as described previously (54). We determined ROC curves and ROC AUC values using the ROCR R library (68). The optimal cutoff was determined using an established cost–benefit method (39) implemented by the OptimalCutpoints R library (69). In the case of the NY-ESO-1 epitope and TCR C^{259}, we were able to determine the level of TCR binding strength under which T cell activation is negligible. We used T cell activation data of the sequentially modified NY-ESO-1 peptides provided by the same study (23). We defined lower than 10% of the original epitope's TCR binding strength as insufficient binding, because under this value, the median level of T cell activation was only 7.9% of the original.

Determining TCEMs in Intracellular Pathogens and Analyzing HLA Association Data. We downloaded the reference proteomes of 50 well-known intracellular pathogens from the UniProt database (61) (on November 18, 2019, SARS-CoV-2 proteome was downloaded on March 26, 2020; Dataset S3). First, we determined the TCEMs of each 9-mer in the proteome of each pathogen. Next, we calculated their prevalence in the human proteome, expression in cTECs, and probability of proteasomal cleavage as explained before. We defined np-TCEMs as the ones found less than four times in the human proteome or have low expression in cTECs or low probability of thymoproteasomal cleavage. We used cutoff values determined for Fig. 2 to define low expression and low probability of thymoproteasomal cleavage. We then predicted the binding of each 9-mer to common HLA alleles with the NetMHCpan-4.0 algorithm (57). We used HLA-A and B alleles listed in a reference set with maximal population coverage found on the IEDB web page (43). As the list did not include data for HLA-C, we selected the first four-digit alleles from all two-digit HLA-C allele classes. To decrease the prevalence of false-positive binding results, we considered a 9-mer to be bound by a given allele if the predicted binding rank percentile value was under 0.5% and the predicted binding affinity value was under 50 nM. To those alleles, for which the prediction could not detect any bound peptides in the proteome of the pathogen, we assigned N peptides having the lowest predicted binding affinity values, in which N refers to the median number of peptides bound by other alleles at the same HLA locus. For each allele–species pair, we calculated the fraction of np-TCEMs in bound peptides and visualized the results on a heatmap.

To identify HLA allele associations with infectious diseases, we carried out literature mining. We aimed to involve only reliable HLA association data in the analysis, so we focused on meta-analyses. We searched PubMed with the "hla infection meta analysis" and "hla association meta analysis" keywords on August 27, 2020. We found HLA association meta-analysis studies for hepatitis B (41), hepatitis C (44), dengue virus (43), and human papillomavirus (42). In the case of hepatitis B, C, and HPV studies, we selected significant associations between HLA allele groups and either risk for infections or response to treatment. In the case of dengue virus infection, P values were not determined by the authors of the study. They ranked associations of different HLA allele groups with the infection along 17 studies and considered the allele with the best rank as protective ones. We followed the method of the authors and carried out the same analysis but considered only those allele groups that were included in at least 75% of studies to increase the reliability of the analysis. In each study, we calculated the rank percentile of odds ratio values associated with carrying different allele groups. Next, we calculated the average of study-specific rank percentile values for each allele group. We associated the group with the lowest-rank percentile with protection and the group with the highest-rank percentile with susceptibility. As the examined meta-analysis studies were carried out for allele groups or serotypes and not individual alleles, we aggregated the allele-specific values of np-TCEM presentation as follows. For serotypes, we averaged the values specific for the alleles that belong to the given serotype. In the case of allele groups (i.e., HLA type is defined at two-digit resolution), we averaged the values specific for the alleles that are classified as common in the Common and Well-Documented Alleles Catalog (70).

Statistical Analysis and Visualization. We used R version 3.6.3 (71) in RStudio version 1.2.5033 environment for statistical analyses. We used the ggplot2 (72), ggpubr, grid, gridExtra, ggsci, scales, png, ComplexHeatmap (73), and ggrepel R libraries for visualization. Smooth curves on plots were fitted with cubic smoothing spline method (74). Figs. 4*A* and 6 were created with BioRender.com.

Data Availability. Code data have been deposited in GitHub (https://github.com/KBalazs1987/ImObsRecognition). All other study data are included in the article and/or supporting information. Previously published data were used for this work (https://www.iedb.org/ https://doi.org/10.1101/2020.07.31.20165647).

ACKNOWLEDGMENTS. We are grateful to Bence Hegyi, Attila Bebes, and Roy Bitkover for their useful comments on a previous version of the manuscript. We thank Ron S. Gejman for providing raw data on minigene depletion binding experiments. M.M. was supported by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences. M.M. was supported by the ÚNKP-20-5, A.L. was supported by the ÚNKP-20-2, B.G.M. was supported by the ÚNKP-19-3, KB was supported by the ÚNKP-20-4, and B.T.P. was supported by the ÚNKP-20-3—New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. This research work was conducted with the support of the Szeged Scientists Academy under the sponsorship of the Hungarian Ministry of Innovation and Technology (FEIF/433-4/

2020-ITM_SZERZ). The project has received funding from the EU's Horizon 2020 research and innovation program under grant agreement 739593.

1. M. L. Dustin, The immunological synapse. *Cancer Immunol. Res.* **2**, 1023–1033 (2014).
2. R. D. Bremel, E. J. Homan, Extensive T-cell epitope repertoire sharing among human proteome, gastrointestinal microbiome, and pathogenic bacteria: Implications for the definition of self. *Front. Immunol.* **6**, 538 (2015).
3. M. J. Reddehase, J. B. Rothbard, U. H. Koszinowski, A pentapeptide as minimal antigenic determinant for MHC class I-restricted T lymphocytes. *Nature* **337**, 651–653 (1989).
4. R. D. Bremel, E. J. Homan, Frequency patterns of T-cell exposed amino acid motifs in immunoglobulin heavy chain peptides presented by MHCs. *Front. Immunol.* **5**, 541 (2014).
5. J. J. A. Calis *et al*., Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).
6. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
7. K. Takada, K. Kondo, Y. Takahama, Generation of peptides that promote positive selection in the thymus. *J. Immunol.* **198**, 2215–2222 (2017).
8. K. Sasaki *et al*., Thymoproteasomes produce unique peptide motifs for positive selection of CD8(+) T cells. *Nat. Commun.* **6**, 7484 (2015).
9. N. Vrisekoop, J. P. Monteiro, J. N. Mandl, R. N. Germain, Revisiting thymic positive selection and the mature T cell repertoire for antigen. *Immunity* **41**, 181–190 (2014).
10. J. N. Mandl, J. P. Monteiro, N. Vrisekoop, R. N. Germain, T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity* **38**, 263–274 (2013).
11. R. B. Fulton *et al*., The TCR's sensitivity to self peptide-MHC dictates the ability of naive CD8(+) T cells to respond to foreign antigens. *Nat. Immunol.* **16**, 107–117 (2015).
12. M. Yarmarkovich, J. M. Warrington, A. Farrel, J. M. Maris, Identification of SARS-CoV-2 vaccine epitopes predicted to induce long-term population-scale immunity. *Cell Rep Med* **1**, 100036 (2020).
13. M. Ogishi, H. Yotsuyanagi, The landscape of T cell epitope immunogenicity in sequence space. *bioRxiv* [Preprint] 155317 (2018). https://doi.org/10.1101/155317. Q:27
14. P. Lydyard, A. Whelan, M. Fanger, *BIOS Instant Notes in Immunology* (Taylor & Francis, 2011).
15. S. K. Mohanty, S. K. Mohanty, K. S. Leela, *Textbook of Immunology* (JP Medical Ltd, 2013).
16. Y. Xing, K. A. Hogquist, T-cell tolerance: Central and peripheral. *Cold Spring Harb. Perspect. Biol.* **4**, a006957 (2012).
17. R. Vita *et al*., The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
18. S. Paul, *et al*., HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* **191**, 5831–5839 (2013).
19. M. Nielsen, M. Andreatta, NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33 (2016).
20. M. Collatz *et al*., EpiDope: A deep neural network for linear B-cell epitope prediction. *Bioinformatics* **37**, 448–455 (2021).
21. S. E. C. Caoili, Expressing redundancy among linear-epitope sequence data based on residue-level physicochemical similarity in the context of antigenic cross-reaction. *Adv. Bioinforma.* **2016**, 1276594 (2016).
22. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
23. A. R. Karapetyan *et al*., TCR fingerprinting and off-target peptide identification. *Front. Immunol.* **10**, 2501 (2019).
24. R. Andersen, Nonparametric methods for modeling nonlinearity in regression analysis. *Annu. Rev. Sociol.* **35**, 67–85 (2009).
25. H. Pearson *et al*., MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126**, 4690–4701 (2016).
26. D. Malhotra *et al*., Tolerance is established in polyclonal CD4(+) T cells by distinct mechanisms, according to self-peptide expression patterns. *Nat. Immunol.* **17**, 187–195 (2016).
27. D. von Bubnoff *et al*., Antigen-presenting cells and tolerance induction. *Allergy* **57**, 2–8 (2002).
28. A. J. Coles *et al*., Keratinocyte growth factor impairs human thymic recovery from lymphopenia. *JCI Insight* **5**, e125377 (2019).
29. S. Murata *et al*., Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science* **316**, 1349–1353 (2007).
30. M. Zeeshan, K. Tyagi, Y. D. Sharma, CD4+ T cell response correlates with naturally acquired antibodies against Plasmodium vivax tryptophan-rich antigens. *Infect. Immun.* **83**, 2018–2029 (2015).
31. J. Schmidt *et al*., Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep. Med.* **2**, 100194 (2021).
32. T. P. Riley *et al*., Structure based prediction of neoantigen immunogenicity. *Front. Immunol.* **10**, 2047 (2019).
33. D. Chowell *et al*., TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1754–E1762 (2015).
34. J. Sidney *et al*., Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res.* **4**, 2 (2008).
35. S. Sarkizova *et al*., A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
36. T. M. Snyder *et al*., Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels (infectious diseases (except HIV/AIDS). *medRxiv* [Preprint] (2020) https://doi.org/10.1101/2020.07.31.20165647 (Accessed 30 September 2020).
37. L. P. Richman, R. H. Vonderheide, A. J. Rech, Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst.* **9**, 375–382.e4 (2019).
38. S. Carrasco Pro *et al*., Microbiota epitope similarity either dampens or enhances the immunogenicity of disease-associated antigenic epitopes. *PLoS One* **13**, e0196551 (2018).
39. C. E. Metz, Basic principles of ROC analysis. *Semin. Nucl. Med.* **8**, 283–298 (1978).
40. R. S. Gejman *et al*., Identification of the targets of T-cell receptor therapeutic agents and cells by use of a high-throughput genetic platform. *Cancer Immunol. Res.* **8**, 672–684 (2020).
41. V. Seshasubramanian, G. Soundararajan, P. Ramasamy, Human leukocyte antigen A, B and Hepatitis B infection outcome: A meta-analysis. *Infect. Genet. Evol.* **66**, 392–398 (2018).
42. M. Bhaskaran, G. ArunKumar, A meta-analysis of association of human leukocyte antigens A, B, C, DR and DQ with human papillomavirus 16 infection. *Infect. Genet. Evol.* **68**, 194–202 (2019).
43. D. Weiskopf *et al*., Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E2046–E2053 (2013).
44. E. Gauthiez *et al*.; Swiss Hepatitis C Cohort Study, A systematic review and meta-analysis of HCV clearance. *Liver Int.* **37**, 1431–1445 (2017).
45. N. J. Burroughs, R. J. de Boer, C. Keşmir, Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics* **56**, 311–320 (2004).
46. V. Daniel, H. Wang, M. Sadeghi, G. Opelz, Interferon-gamma producing regulatory T cells as a diagnostic and therapeutic tool in organ transplantation. *Int. Rev. Immunol.* **33**, 195–211 (2014).
47. S. Frankild, R. J. de Boer, O. Lund, M. Nielsen, C. Kesmir, Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLoS One* **3**, e1831 (2008).
48. M. Wölfl *et al*., Hepatitis C virus immune escape via exploitation of a hole in the T cell repertoire. *J. Immunol.* **181**, 6435–6446 (2008).
49. E. J. Yager *et al*., Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *J. Exp. Med.* **205**, 711–723 (2008).
50. D. Vidović, P. Matzinger, Unresponsiveness to a foreign antigen can be caused by self-tolerance. *Nature* **336**, 222–225 (1988).
51. W. Yu *et al*., Clonal deletion prunes but does not eliminate self-specific αβ CD8(+) T lymphocytes. *Immunity* **42**, 929–941 (2015).
52. J. J. A. Calis, R. J. de Boer, C. Keşmir, Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLOS Comput. Biol.* **8**, e1002412 (2012).
53. M. Rolland *et al*., Recognition of HIV-1 peptides by host CTL is related to HIV-1 similarity to human proteins. *PLoS One* **2**, e823 (2007).
54. A. Bresciani *et al*., T-cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome. *Immunology* **148**, 34–39 (2016).
55. R. Vita, B. Peters, A. Sette, The curation guidelines of the immune epitope database and analysis resource. *Cytometry A* **73**, 1066–1070 (2008).
56. W. Fleri *et al*., The immune epitope database: How data are entered and retrieved. *J. Immunol. Res.* **2017**, 5974574 (2017).
57. V. Jurtz *et al*., NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368 (2017).
58. M. Manczinger *et al*., Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLoS Biol.* **17**, e3000131 (2019).
59. F. Sievers *et al*., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
60. K. Yang, L. Zhang, Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Res.* **36**, e33 (2008).
61. UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

62. D. J. McCarthy, Y. Chen, G. K. Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
63. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
64. M. Lawrence *et al*., Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
65. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
66. C. Camacho *et al*., BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
67. M. Klinger *et al*., Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One* **10**, e0141561 (2015).
68. T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: Visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
69. M. López-Ratón, M. X. Rodríguez-Álvarez, C. C. Suárez, F. G. Sampedro, OptimalCutpoints: An *R* package for selecting optimal cutpoints in diagnostic tests. *J. Stat. Soft.* **61** (2014). Q:28
70. S. J. Mack *et al*., Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* **81**, 194–203 (2013).
71. R. Ihaka, R. Gentleman, R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
72. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
73. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
74. C. de Boor, *A Practical Guide to Splines* (Springer-Verlag New York, 1978).