Immunogenicity landscape of cytotoxic T lymphocyte epitopes in sequence space.

Authors: Masato Ogishi\textsuperscript{1*} and Hiroshi Yotsuyanagi\textsuperscript{1}

Affiliations:

\textsuperscript{1} Department of Infectious Diseases and Applied Immunology, Research Hospital, The Institute of Medical Science, The University of Tokyo, Tokyo, 108-0071, Japan

* To whom correspondence should be addressed: Masato Ogishi (oogishi-tky@umin.ac.jp)

Keywords: Epitope immunogenicity prediction, human leucocyte antigen class I, T cell receptor, escape, neoepitope
Abstract

Cytotoxic T lymphocytes (CTLs) recognize peptides known as epitopes presented onto MHC-I molecules to trigger an immune response. However, factors demarcating epitopes from non-immunogenic MHC binders remain poorly understood. Here, using 21,146 human MHC-I-loaded peptide sequences, we demonstrated an accurate immunogenicity prediction through a sequence-based approximation of the contact potential profiles (CPPs) of peptides against pooled TCR repertoires in conjunction with MHC binding information, achieving the area under the curve (AUC) of 79%. Predictive features provided insights into molecular scanning by TCR repertoire. Our framework also worked for primate and mouse peptide datasets. The quantitative immunogenicity scores delineated the landscapes of viral escaping and epitope formation through epitope mapping in sequence space and predicted overall survivals of cancer patients in checkpoint blockade cohorts. Overall, our analysis provides insights into the high-dimensional architectures of CTL activation and expedites translational efforts toward precision immunotherapy in infectious diseases and cancer.
Introduction

Adaptive immunity is triggered by antigens. Antigens are processed into peptide fragments by proteasomes, transported into the endoplasmic reticulum, and coupled to major histocompatibility complex [MHC; also called the human leukocyte antigen (HLA in humans)] molecules on the surface of antigen-presenting cells (APCs). Naïve T cells interact with MHC-loaded peptides via their receptor (T cell receptor, TCR), and successful recognition activates them to initiate subsequent immunological orchestration.

The term “epitope” means an immunogenic MHC-loaded peptide. However, strong MHC binding does not ensure immunogenicity, and vice versa. Although antigen processing and MHC binding have been successfully predicted, prediction of immunogenicity remains a challenging task to date.

Why not all MHC binders subject to TCR recognition serve as epitopes? Since immunogenicity is a product of the amplitude of MHC-loaded peptides and the likelihood of recognition by T cells, not only the interactions between the peptide and MHC molecule but also the recognition of the peptide-MHC complex by TCRs are the critical determinants of immunogenicity. In this context, we hypothesized that the intermolecular contact potential profiles (CPPs) between MHC-loaded peptides and the host TCR repertoire would serve as predictive features of immunogenicity. Because
of the extreme diversity of TCR repertoire, neither in vitro assay nor structural modeling is a feasible approach. Instead, we hereby established an alternative, sequence-based framework which we termed Repitope, approximating the CPPs using amino acid pairwise contact potential (AACP) scales. We made three assumptions. First, only a small fraction of TCRs in the host TCR repertoire recognize a given MHC-loaded peptide. Second, CDR3 loops, which exhibit the highest degree of genetic variability, are sufficient in modeling the TCR-peptide interactions. Finally, only a small part of a CDR3 loop interacts with a small part of a given peptide. With those assumptions in mind, molecular scanning by the host TCR repertoire can be viewed as a parallelized pairwise sequence alignment problem between peptide sequences and CDR3 fragment sequences, with alignment scores calculated from AACP-derived custom substitution matrix as surrogate measurements of inter-fragment contact potentials.

With the Repitope framework, we demonstrated an accurate and quantitative prediction of immunogenicity of cytotoxic T lymphocyte (CTL) epitopes presented on human HLA class I molecules. Incorporation of binding to representative HLA supertypes improved the overall predictive accuracy, indicating the complementary roles of HLA restriction and TCR recognition. The most predictive features provided mechanistic insights into CTL immunity. Mapping adjacent peptides in sequence space
enabled prediction of bidirectional immune transitions including escaping and epitope formation. *In silico* mutagenesis followed by neighbor network clustering analysis enabled identification of transition-prone peptides after single amino acid substitutions. Finally, we demonstrated stratification of long-surviving patients treated with checkpoint inhibitors in melanoma and lung carcinoma cohorts from the accumulation of changes in immunogenicity between wild-type and mutated peptides. Collectively, our findings theoretically bolster our understanding of CTL immunity and galvanize the rational design of precision immunotherapy. We distributed datasets and essential functions as the R package Repitope ([https://github.com/masato-ogishi/Repitope/](https://github.com/masato-ogishi/Repitope/)) for ensuring reproducibility and warranting prospective validations in independent cohorts.
Results

Immunogenicity prediction from TCR-peptide contact potential profiling

We first collected a set of 21,146 peptides of lengths 8 to 11 amino acids with known HLA restrictions and T cell functional assay results from the Immune Epitope Database (IEDB) and other literature8,9,16–21, yielding 6,942 immunogenic epitopes and 14,204 non-immunogenic MHC binders. Experimentally verified HLA genotype and serotype information was integrated as HLA restriction (Figure 1a and Supplementary Table 1 and Methods), against which prediction by NetMHC 4.06 showed stronger binding than non-restricting HLAs (Figure 1b). Meanwhile, epitopes showed significantly stronger binding than mere MHC binders (Figure 1c). Tumor neoepitopes showed at least comparable binding compared to virus-derived epitopes (Figure 1d).
Figure 1. MHC binding prediction. (a) The distribution of restricting HLA supertypes in our metadataset. (b) The distributions of predicted rank percentiles. Binding was predicted using NetMHC 4.0 with default settings. Red, corresponding HLAs. Grey, all of the twelve HLA representatives [A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62]. (c) The distributions of predicted rank percentiles between epitopes and non-immunogenic MHC binders. (d) The distributions of predicted rank percentiles in epitopes and MHC binders of various sources. NE, tumor neoepitope. IAV, influenza A virus. DGV, dengue virus. HBV, hepatitis B virus. HCV, hepatitis C virus. HIV, human immunodeficiency virus type 1. Neoepitope and viral data were obtained from TANTIGEN and IEDB database, respectively. In the right panel, tumor neoepitopes were depicted the same as the left panel for comparative purposes. Grey dashed line indicates the threshold of strong binders (0.5%). P values were calculated by Kruskal-Wallis test followed by Wilcoxon rank sum test with false discovery rate (FDR)-adjustment. **, p < 0.01; ***, p < 0.001; ****, p < 0.0001; ns, not significant.

As expected, previously reported tools for immunogenicity prediction showed only marginal predictive capabilities (Supplementary Figure 1). To achieve accurate immunogenicity prediction, we hereby propose the concept of TCR-peptide contact potential profiling (CPP) (Figure 2a). We extracted and pooled human TCR β CDR3 sequences from datasets deposited in the Sequence Read Archive using MiXCR
software\textsuperscript{22}. CPP features were calculated against the fragment libraries generated from pooled CDR3 sequences, with the fragment size ranging from 3 to 8. The depth of fragment library was set 10,000 from preliminary experiments (Supplementary Figure 2). Iterative feature selection reduced the number of features down to 62 (Supplementary Figure 3). Heatmap analysis showed that immunogenic peptides were clustered based on these features (Figure 2b).
Figure 2. TCR-peptide contact potential profiling (CPP). (a) The concept of CPP. (b) Heatmap clustering analysis. (c) Feature importance analysis. Calculations were iterated five times with different random seeds. Bars represent mean and sd. (d-e) Residue/motif enrichment analysis. The volcano plots show the odds ratios and FDR-adjusted p values by chi-squared test from MHC binders to epitopes for each of the specified (d) residues and (e) motifs. (f) Distributions of the most predictive CPP feature.

The predictive features are likely to reflect the underlying mechanisms of molecular scanning of MHC-loaded peptides by the host TCR repertoire. Feature importance analysis suggested that CPP features derived from the fragment library of
length 3 played vital roles (Figure 2c). Also notably, the existence of residues K and Y was indicated as essential features (Supplementary Figure 3). Indeed, both residues were more enriched to MHC binders (Figure 2d). In contrast, RYQ and other trimer sequence motifs were enriched toward epitopes (Figure 2e). Given that nonamer peptides bound to MHC class II are required to share only five residues to bind the same TCR, and that MHC-I presents peptides shorter than MHC-II, it is likely that trimer motifs are critical for at least the initial phase of molecular scanning by TCRs.

Moreover, the AACP scale KESO98010124, which represents the transfer energy derived from interfacial regions of protein-protein complexes and thus a plausible metric of TCR-peptide interactions, consisted of the most predictive feature (Figure 2f and Supplementary Figure 3).

By integrating physicochemical peptide descriptors, CPP features, and HLA restrictions, we established a framework for immunogenicity prediction, which we termed Repitope. Incorporation of either annotation-based HLA restrictions or binding prediction results against twelve HLA representatives by NetMHC 4.06 improved predictive performance (Supplementary Figure 4). Although annotation-based predictions slightly outperformed those utilizing predicted binding categories, we adopted the latter as “immunogenicity scores” for subsequent analyses (Figure 3a, b).
This makes the entire framework dependent solely on input peptide sequences. The classification of epitopes and MHC binders via immunogenicity score yields sensitivity and specificity of 70% and 73%, respectively, with the best score threshold being 0.35. The overall accuracy, kappa, and area under the curve (AUC) were 72%, 41%, and 79%, respectively (Table 1). The high predictive performance was retained across peptide lengths and HLA supertypes (Supplementary Figure 5). Reduction of the number of CPP features down to 14 still yielded AUC of 76% (Supplementary Figure 6).

However, sequence-level peptide clustering led to slightly lower AUCs (Supplementary Figure 7). Furthermore, we found that either our framework or the prediction models constructed on the human peptide dataset successfully predicted immunogenicity for the peptide datasets of other species (Supplementary Figure 8). Particularly, immunogenicity scores calculated on the primate peptides using human models without HLA information yielded AUC of 70%. This finding implies the existence of shared regulatory mechanisms of CTL immunity between human and other primates. Predictive performances of various algorithms and conditions tested were summarized in Table 1.

In contrast to HLA binding, tumor neoepitopes exhibited significantly lower immunogenicity scores compared to epitopes of various viral origins (Figure 1c and.
Figure 3c), with some exceptions such as gp100\textsubscript{280-288} (score = 0.85) and NY-ESO-1\textsubscript{157-165} (score = 0.70). This is consistent with previous observations that neoepitopes dissimilar to self and similar to pathogen-derived epitopes are more likely to be immunogenic\textsuperscript{25}.

Collectively, the aforementioned observations strongly suggest that the Repitope framework recapitulates the high-dimensional architectures of CTL immunity to enable \textit{in silico} prediction of epitope immunogenicity.
Figure 3. Immunogenicity prediction. (a-b) Immunogenicity scores. Probabilities were estimated from the most predictive 62 CPP features and predicted binding categories against twelve HLA representatives by NetMHC 4.0 (see Supplementary Figure 3 and Methods). (a) Evaluation of immunogenicity scores. (b) Distributions of immunogenicity scores. Left, averaged scores. Right, coefficients of variance of scores between iterations. (c) The distributions of scores in epitopes and MHC binders of various sources. Grey dashed line indicates the median immunogenicity score in the tumor neoepitope dataset. See the legend of Figure 1d for data sources. P values were calculated by Kruskal-Wallis test followed by Wilcoxon rank sum test with FDR adjustment; *, p < 0.05; ***, p < 0.001; ****, p < 0.0001; ns, not significant.
Table 1. Summary of immunogenicity prediction results. Bolded was the prediction adopted as immunogenicity score. HLA, whether annotated HLA binding information or predicted binding by NetMHC 4.0 was used for machine learning. NFeat, The number of CPP features (excluding HLA-related features) used. NPept, the number of peptides for machine learning. NEpi, the number of epitopes in the dataset. ROC-AUC, the area under the receiver-operating curve. PRC-AUC, the area under the precision-recall curve. Thr, the best threshold determined from ROC.

| Experiment ID | Data       | HLA | NFeat | NPept | NEpi | ROC-AUC | PRC-AUC | Thr  | Acc  | Sen  | Spe  | PPV  | NPV  | Prec | Recall |
|---------------|------------|-----|-------|-------|------|---------|---------|------|------|------|------|------|------|------|--------|
| IEDH          | Human      | n.a. | 4,093 | 4,631 | 0.59 | 0.49    | 0.94    | 0.91 | 0.70 | 0.72 | 0.80 | 0.72 | 0.80 | 0.72 |
| POPISK        | Human      | n.a. | 4,114 | 2,112 | 0.68 | 0.68    | 0.60    | 0.62 | 0.45 | 0.57 | 0.70 | 0.60 | 0.70 | 0.60 |
| PLAQU        | Human      | n.a. | 14,093| 4,631 | 0.59 | 0.49    | 0.94    | 0.91 | 0.70 | 0.72 | 0.80 | 0.72 | 0.80 | 0.72 |

Immunogenicity score ratios of adjacent peptide pairs predict immune transitions

Single amino acid substitutions can drastically affect the immunogenicity of the MHC-bound peptide. Delineating the landscape of “immune transitions,” i.e., escaping and epitope formation, is the critical step toward understanding the evolutionary trajectories of microbes and cancer. To this end, we first characterized the 6,175 adjacent peptide pairs identified from our epitope metadataset, of which 940 were transitional pairs based on their annotations. As expected, the ratios of their immunogenicity scores predicted immune transitions, with AUC and the best threshold of 74% and 1.26, respectively (Figure 4a, b). Substitutions at anchor positions 2 and 9...
in nonamers significantly affected the immunogenicity of the mutated counterparts (Figure 4c). Moreover, mutating residues K and Y, residues enriched in non-immunogenic MHC binders, significantly augmented immunogenicity (Figure 2d and Figure 4d). Next, we characterized transitional peptide pairs of the same viral origins (Figure 4e, f). Ratios of immunogenicity scores of transitional peptide pairs were significantly higher than those of non-transitional pairs in all of the five viruses analyzed (Figure 4e). The combined analysis yielded AUC of 79% (Figure 4f). These findings strongly suggested that the viral adaptation trajectory can be predicted from the dynamic changes in immunogenicity scores.
Figure 4. Immune transitions. (a-b) Prediction of immune transitions from the ratios of immunogenicity scores. (c-d) The distributions of immunogenicity score ratios. Gray dashed line indicates the overall median. (e-f) Immune transitions in different viruses. See the legend of Figure 1d for data sources. In (c-e), P values were calculated by Kruskal-Wallis test followed by Wilcoxon rank sum test with FDR adjustment; **, p < 0.01; ****, p < 0.0001. NT, non-transitional; T, transitional.

**Immune transitional landscapes delineated by epitope mapping in sequence space**

Strong immunogenicity does not necessarily mean robust immunogenicity because of potential immune evasion. The discrepancy of these two aspects was exemplified by the neighbor network clustering analysis on an epitope RRYQKSTEL and 132 neighbors (Figure 5a). This epitope, whose immunogenicity score was 0.994, was chosen because it has the largest number of neighbors in our metadataset. Indeed, there were 32 non-immunogenic neighbors, forming two clusters. The Cluster 3 consisted of P9 mutants, in which both annotations and predictions fluctuated. In
contrast, the Clusters 2, a cluster of P8 mutants, consisted of non-immunogenic peptides with only one exception. These observations indicated that, while P8 mutation deprives the epitope of immunogenicity, P9 mutation undermines the robustness of its immunogenicity, possibly by affecting its binding to restricting HLAs.

To gain more comprehensive insights into the immunogenicity landscapes in epitope sequence space, we next conducted in silico mutagenesis analysis; we computationally generated all possible neighbors only single amino acid distant from original viral peptides (Figure 4e, f). While mean scores were higher in epitopes, coefficients of variance were higher in MHC binders (Figure 5b, c). When focusing on the central cluster in neighbor networks, similar but more enhanced trends were observed (Figure 5d, e). As for nonamers, P1- and P8-mutating clusters had higher average scores, whereas P2- and P9-mutated clusters had higher score variances, potentially because of disrupted interactions with restricting HLAs (Figure 5f, g). We next classified peptides based on their annotated immunogenicity and existence of their adjacent but transitional counterparts. The differences in immunogenicity scores, per-cluster mean scores, and intracluster score variances between transitional and non-transitional peptides were statistically significant (Figure 5h-j). Most notably, non-immunogenic but next to immunogenic peptides exhibited higher scores and lower
intracluster score variances than robustly non-immunogenic ones. These observations indicate that increased intracluster variance, as well as decreased immunogenicity score, represents the hallmark of escaping mutations. Our findings may help to choose less escape-prone epitopes for the development of robustly effective vaccines.

Figure 5. Mapping immune transitional landscapes in sequence space. (a) An example neighbor network. An epitope RRYQKSTEL was chosen as representative. Edges represent single amino acid substitution, weighted by inverted score ratios. Colors indicate predicted immunogenicity scores, and shapes indicate annotated immunogenicity (circle, immunogenic; square, non-immunogenic). A walktrap algorithm was adopted for clustering. Labels indicate consensus sequences for each cluster. (b-c) The mean score and coefficient of variance for the peptide of interest and all in silico mutated neighbors. (d-e) The mean score and coefficient of variance for the peptide of interest and neighbors residing in the same cluster. (f-g) Mean scores and variance per cluster stratified by the mutated positions in nonamers. (h-i) Mean scores and variance per cluster stratified by the immunogenicity-transition patterns. Peptides that have at least one transitional neighbors were categorized as transitional. In (b-i), P values were calculated by Wilcoxon rank sum test with FDR adjustment; *, p < 0.05; **, p < 0.01; ***, p < 0.001; ****, p < 0.0001. P, immunogenic; N, non-immunogenic; NT, non-transitional; T, transitional.

Accumulation of transitional neoepitopes predicts patient survival in checkpoint blockade cohorts.
Tumor neoepitope burden defined by intracellular processing and MHC binding has been shown to correlate with response and patient survival in various checkpoint blockade cohorts\textsuperscript{26–28}. However, neoepitopes that elicit the most prominent and sustained CTL response often remain elusive. Here, we demonstrated that accumulation of transitional neoepitopes successfully identified patients responding favorably to checkpoint inhibitors in melanoma and lung carcinoma cohorts (Figure 6a; \( P \)-value = 6.1x10\(^{-5} \) in the log-rank test). Unlike the previously proposed neoantigen fitness model\textsuperscript{11}, this immune transition model predicted overall survivals without any model-specific parameter; predictions can be made from the linear sums of the score ratios. Moreover, our model successfully reduced the total number of neoepitopes without compromising predictive efficacy; the median numbers of neoepitopes were 39 vs. 4 in high- and low-burden groups in our model, considerably small numbers compared to 285 vs. 34.5 according to the original criteria (Figure 6b, c). These findings place tumor neoepitope biology in a more generalized context of CTL immunity. Larger-scale cancer mutanome profiling study is warranted to validate and further sophisticate the prediction of neoepitope candidates, thus galvanizing research in precision oncoimmunology.
Figure 6. Prediction of survivals of patients treated with checkpoint inhibitors. (a) A Kaplan-Mayer curve. The dataset compiled by Lukzsa et al.\textsuperscript{11} derived from two melanoma cohorts\textsuperscript{27,28} and one lung carcinoma cohort\textsuperscript{26} was used. Patients whose sums of score ratios exceed 100 were classified as high neoepitope burden. (b-c) Neoepitope burdens defined by (b) the original criteria\textsuperscript{11} and (c) our criteria.
Discussion

The ultimate goal of immunogenicity prediction is to predict which peptide best triggers sustained immunological response in vivo. In contrast to the complexity of the regulatory mechanisms in vivo including regulatory T cells, CTL exhaustion mediated through chronic immune checkpoint signals, and the immunosuppressive microenvironment engendered by solid tumors, T cell activation is primarily governed by the MHC-peptide-TCR interactions, with some assistance from coreceptors such as CD8. Considering this asymmetric complexity, T cell activation assay in vitro inevitably yields some false positives in vivo. That being said, however, eliminating candidates least likely to be immunogenic in silico would significantly expedite the research of precision immunotherapy.

Our analysis demonstrates highly accurate immunogenicity predictions in silico. Our framework is novel in that it utilizes TCR fragment library as a peptide homology evaluation function from the viewpoint of TCR repertoire, in contrast to previous studies that define homology from sequence-level pairwise alignment scores. TCR-peptide contact potential profiling recapitulates the context-dependent effects of single amino acid substitutions on immunogenicity. In contrast to high-throughput in vitro T cell assays that only provide estimates of immunogenicity of the peptides tested, our
framework can extrapolate the high-dimensional rules of immunogenicity learned from more than 20,000 peptide sequences to delineate the immunogenic landscapes in sequence space. Finally, our framework is advantageous in that it does not depend on the \textit{a priori} knowledge of HLA restrictions, thus enabling \textit{in silico} screening solely from genomic data such as tumor mutation datasets.

Escaping is one of the most prominent issues in developing effective vaccines\textsuperscript{34,35}, and neoepitope formation is a promising target towards tumor-specific therapeutic vaccines\textsuperscript{36,37}. These phenomena are two sides of the same coin in sequence space.

However, the bidirectional effects of single amino acid substitutions in terms of immunogenicity are likely highly context-dependent and remain barely understood. Even a single mutation could affect various aspects of CTL immunity, including proteolytic processing, transportation to endoplasmic reticulum, stability and structural orientation in MHC presentation, and propensity for TCR recognition\textsuperscript{1,9,38}. We found that dynamic changes in immunogenicity scores are predictive of viral escaping and response to checkpoint blockade therapy in cancer patients. Moreover, we characterized the peculiarity of peptides adjacent to transitional boundaries defined from neighbor network architectures in sequence space. To our knowledge, this work is the first attempt to systematically explain the effects of single amino acid substitutions in the
context of bidirectional immune transitions.

Our study has several caveats and limitations. First, the presumed correlation between the sequence-based contact potential profiles and the real TCR affinity/avidity profiles is not experimentally validated. Although we only utilized TCRβ CDR3 sequences from unsorted T cell sources, this point could further be explored in the future. Moreover, since the majority of the epitope data utilized were obtained from microbial sources, the application of the Repitope framework for autoimmunity and oncoimmunology may require further validations and optimizations. Finally, it remains an open question whether similar strategies could reveal underlying rules governing the antigen recognition by helper T cells and regulatory T cells, which if exist would undoubtedly engender a more unified view of T cell biology.

In conclusion, our analyses establish a framework for accurate epitope immunogenicity prediction by introducing sequence-based approximation of TCR-peptide contact potential profiles and provide insights into CTL immunogenicity landscapes. Simultaneous prediction of immunogenicity and immune transitional potentials represents a promising strategy toward more effective and less escape-prone precision immunotherapy against infectious diseases and cancer.
Methods

MHC-I epitope/ligand dataset

Human data. HLA-I-restricted peptide sequences with known immunogenicity for CTL were collected from public databases [Immune Epitope Database (IEDB, as of Mar 30, 2018)\textsuperscript{15}, the best-characterized CTL epitopes from Los Alamos National Laboratory (LANL) HIV Sequence Database\textsuperscript{18}, LANL HCV Sequence Database\textsuperscript{17}, EPIMHC\textsuperscript{16}, MHCBN\textsuperscript{20}, and TANTIGEN\textsuperscript{21}] and previous publications\textsuperscript{8,9,19}. For the peptides from IEDB database, only those with the evidence listed in Supplementary Table 2 were considered valid. Peptides presented on non-human MHC molecules were discarded, whereas those presented on HLA class I molecules expressed in non-human hosts (e.g., transgenic mice) were included. Either peptide sequences whose lengths were other than 8, 9, 10, and 11, or those containing non-standard letters were discarded. In this manner, a total of 21,146 unique peptides were identified. A total of 1,872 (8.9\%) peptides that had conflicting immunogenicity annotations were identified. In this study, peptides that were annotated as immunogenic in at least one dataset were considered immunogenic. Consequently, 6,942 (32.8\%) unique epitopes and 14,204 (67.2\%) MHC binders were identified.

Other species. CTL epitopes and MHC binders loaded on species-specific MHC
molecules were retrieved from the IEDB database (as of Mar 30, 2018). Species studied were mice and primates (bonobo, chimpanzee, gorilla, marmoset, and rhesus macaque).

**Contact potential profiling (CPP)**

*Concept.* Immunogenic peptides presented on the MHC class I molecules must be recognized by the TCRs of CD8\(^+\) CTLs with strong affinity to trigger subsequent immunological cascades\(^{14}\). The entire TCR repertoires do not necessarily recognize those peptides; instead, a quite limited set of epitope-specific TCRs would be sufficient. Moreover, given the flexible conformational changes upon binding between TCR and peptide-MHC complex, it is plausible that only a restricted set of residue pairs between the peptide and the CDR3 loop initially act as a “seed” of molecular scanning that may endow the complex the time for conformational changes to further enhance the affinity and trigger CTL activation\(^{14,39,40}\). In this view, the interacting surfaces could be approximated as a pair of linear sequences, and by deconvoluting the TCR repertoire into a set of fragments of a defined amino acid length, the process of molecular scanning could be interpreted as a pairwise sequence alignment problem. One caution is that, although an ordinary sequence alignment algorithm involves a similarity matrix which gives higher values to biochemically similar residue pairs, higher scores must be given to more strongly interacting residue pairs rather than more biochemically similar
residue pairs in order to interpret the alignment as the reflection of intermolecular interaction. For this purpose, amino acid pairwise contact potentials (AACPs) were adopted from the AAIndex database\textsuperscript{41} (http://www.genome.jp/aaindex/AAindex/list_of_potentials) to generate custom substitution matrices. The Smith-Waterman local alignment algorithm maximizes the alignment score of a given TCR-derived sequence fragment against the whole sequence of a given peptide, and its alignment score can be interpreted as a correlate of the intermolecular contact potential. Consequently, the interactions between a given peptide and a given TCR repertoire can be profiled by summarizing the distribution of alignment scores.

\textit{Fragment library}. We hypothesized that human public TCR clonotypes are biased in favor of “publicly” immunogenic peptides rather than non-immunogenic ones, and therefore could be utilized as a probe for immunogenicity. We focused on the CDR3 loops because they are primarily responsible for the interactions with MHC-loaded peptides, whereas MHC $\alpha$-helices are typically contacted with more conserved CDR1 and CDR2 loops\textsuperscript{1,42,43}. We restricted our analysis to human TCR$\beta$ (TRB) repertoire. Due to the frequent lack of annotations of T cell sorting (e.g., CD8$^+$ T cells), we included TCR repertoires from all kinds of T cells. TCR repertoire datasets were collected from
NCBI Sequence Read Archive (SRA) and a previous study led by Britanova et al.\textsuperscript{44}.

SRA was searched with the following terms ‘(T Cell Receptor \(\beta\)) AND human.’ The following BioProjects were included: PRJNA389805, PRJNA329041, PRJNA298417, PRJNA273698, PRJNA258001, PRJNA229070, PRJNA79707, and PRJNA79435.

Projects in which healthy donors were not involved (e.g., cancer, autoimmune diseases) were omitted. Datasets derived from HIV-infected subjects in the BioProject PRJNA258001 were discarded. Meanwhile, PRJNA316572, involving the unique molecular identifier strategy, were not included in order to maintain consistency in the way of analysis. Fastq files were obtained using fastq-dump script with the following options: --gzip --skip-technical --readids --read-filter pass --dumpbase --split-files --clip --accession [SRA run number]. By using MiXCR software\textsuperscript{22}, a total of 23,006,555 TRB CDR3 clonotypes were identified. Subsequently, a total of 191,326 clonotypes observed in at least 22 out of 220 (=10\%) different datasets were retained as public clonotypes.

Finally, the pooled CDR3 sequences were fragmented by a sliding window strategy to generate a fragment library. Sliding windows of amino acid length 3 to 8 were tested. Moreover, since CDR3 loops could interact with peptides in either a forward parallel or antiparallel fashion, reversed CDR3 sequences were simultaneously fragmented and pooled. Resultant fragment pools were compressed to a defined library depth (e.g.,...
10,000) by random down-sampling while preserving the original proportion of each fragment.

Pairwise sequence alignment. The `pairwiseAlignment` function implemented in the `Biostrings` package in `Bioconductor` was utilized. This function seeks an optimal alignment which maximizes the overall alignment score defined as a sum of pairwise scores. Alignment type was set “global-local” to obtain an optimal alignment of a CDR3 fragment with consecutive subsequences of the peptide. Gaps were not allowed. A set of substitution matrices were generated from a set of thirty-five AACP scales retrieved from the AAIndex database. The scales utilized in this study are summarized in Supplementary Table 3. AACP matrices were rescaled so that numerical indicators resided within the range between 0 and 1 for comparative purposes. For analytical purposes, we also prepared inverted AACP matrices (rescaled as well) to mimic the weakest interactions as well as the strongest interactions. A distribution of alignment scores was summarized by calculating representative statistics. Following statistics were utilized: mean, standard deviation, median, 10% trimmed mean, median absolute deviation, skew, kurtosis, standard error, interquartile range, 10% quantile, and, 90% quantile. Predictive features were generated by combining the fragment length, the AACP scale, and the type of statistics. In this manner, a total of 4,620 CPP features
were generated for each of the peptides.

**Peptide descriptors**

Apart from CPP features, sequence-based physicochemical features were also calculated. Each peptide sequence was converted into a set of consecutive fragments of a defined amino acid length, and peptide descriptors were calculated against each of the fragments using functions in the *Peptide* package. Following functions were utilized:

- *aIndex*, *blosumIndices*, *boman*, *charge*, *crucianProperties*, *fasgaiVectors*, *hmoment*,
- *hydrophobicity*, *instaIndex*, *kideraFactors*, *mswhimScores*, *pI*, *protFP*, *vhseScales*, and,
- *zScales*. The distributions of the values were summarized as described in the paragraph of CPP. Moreover, a set of categorical features indicating whether the peptide of interest is free from a specific amino acid residue (e.g., tyrosine) were included. Finally, the length of the peptide was included as a feature as well. In this manner, combined with CPP features, a total of 6,081 features were generated for each of the peptides.

**HLA restriction**

*Annotation*. Experimental annotations on binding to HLA class I molecules were retrieved from IEDB and other databases. HLA genotypes were converted to supertypes according to the rules from Sidney *et al.*[^46] Serotypes were treated *per se*. In this work, supertypes and serotypes were not strictly distinguished and treated as “HLA restriction”
in a combined fashion. Annotations of HLA restriction against twelve HLA representatives [A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62] were summarized into a matrix of twelve columns indicating whether the peptide in question was bound to the specified HLA type.

**Prediction.** Binding prediction against twelve HLA representatives was performed using NetMHC 4.0\(^6\). Either predicted rank percentiles or their categorical conversions were summarized in a matrix format as described above. The thresholds of rank percentiles for strong and weak binders were 0.5% and 2%, respectively.

**Immunogenicity prediction**

**Clustering.** Overrepresentation of peptides sharing high sequence homology could lead to biases in the subsequent analysis. Peptides clustering using IEDB epitope clustering tool\(^16\) followed by random down-sampling was performed to reduce the bias and rule out the possibility of overfitting. A homology threshold of either 60%, 80%, or 100% (i.e., no clustering) was compared.

**Preprocessing and feature selection.** An iterative machine learning strategy was employed to find out a restrictive set of predictive features for immunogenicity uniformly applicable to the entire epitope/ligand dataset. First, the dataset was partitioned into training and testing subdatasets in a ratio of 4:1. Second, the training
subdataset was centered and rescaled using the `preProcess` function implemented in the `caret` package. Third, highly correlating features were eliminated using the `findCorrelation` function in `caret`. The threshold of correlational coefficients was set to be 0.75. Finally, features were selected based on the importance values calculated using the `randomForestSRC::rfsrc` method implemented in the `mlr` package. The 100 most important features were retained unless otherwise stated. The entire feature selection process was iterated five times with different random seeds. Moreover, considering the possibility that the initial order of features matters in the feature selection process, the same five iterations were performed using the inversely reordered feature set. The testing subdataset was not utilized for feature selection purposes.

**Machine learning.** Preliminary experiments suggested that the optimal algorithm is the extremely randomized trees (ERT) with `mtry` and `numRandomCuts` being 35 and 2, respectively. The class imbalance was taken into consideration by inverse weighting. ERT model was trained using the training subdataset. The testing subdataset was preprocessed using the processing function defined by the training subdataset and was subjected to predictions.

**Immunogenicity score.** Peptides were divided into five chunks of approximately same sizes. ERT model training and immunogenicity prediction were performed in a
leave-one-chunk-out fashion with a predefined set of the most predictive features. HLA binding data from either experimental annotations or predictions by NetMHC 4.0 was incorporated as features when indicated. The entire prediction process (data splitting, preprocessing, and machine learning) was iterated five times with different random seeds. The estimated probability values were averaged for each of the peptides. We termed the averaged probability of immunogenicity as “immunogenicity score.” We extrapolated this framework to an external dataset by applying the twenty-five ERT models to the entire external dataset.

**Immune transition**

*Concept.* Single amino acid substitutions have multifaceted effects on the immunogenicity of epitopes and MHC binders. Loss of immunogenicity is called escape or immune evasion. Also, the term immunoediting is used particularly in the field of cancer\(^49\). In this work, we unify the terminology as “escape.” For the sake of simplicity, we do not take other mechanisms of evasion such as impaired intracellular antigen processing, somatic loss of HLA heterozygocity\(^50,51\), and checkpoint-mediated CTL exhaustion and apoptosis\(^52\). In contrast, there is no appropriate terminology for the acquisition of immunogenicity. In this work, we refer to this phenomenon as “epitope/neoeptope formation.” Escaping and epitope formation can be understood as
two sides of the same coin in sequence space. Therefore, we defined a more unified concept, “immune transition,” in this work, which is defined as a change in immunogenicity between a pair of MHC-presented peptides with just one edit distance.

By definition, the immune transition is bidirectional.

**Neighbor network analysis.** A graph network was constructed from a set of peptides, where pairs of peptides with one edit distance were considered edges. The low-to-high score ratios were considered edge distances, and edges were weighted by the inversed score ratios. Network clustering was performed using the `cluster_walktrap` function implemented in the `igraph` package\(^53\). The consensus sequence per cluster was generated using the ClustalW algorithm implemented in the `msa` package\(^54\) in Bioconductor\(^45\).

**In silico mutagenesis analysis.** To delineate the immunogenicity landscape in a broader context, we computationally expand neighbor networks by simulating the immunogenicity scores of computationally single amino acid substituted peptides. Simulated peptides with indels were not generated. Note that we ignore the possibility that the artificially introduced mutation could potentially affect factors outside the MHC-peptide-TCR axis such as altered proteolytic degradation. Instead, our primary attention is on exploring the effects of those mutations on MHC presentation and TCR
Viral escaping analysis

Viral epitopes and MHC binders were retrieved from the IEDB database (as of Mar 30, 2018). Five viruses, namely, influenza A virus (IAV), dengue virus (DGV), hepatitis B virus (HBV), hepatitis C virus (HCV), and human immunodeficiency virus type I (HIV) were analyzed. In virus-specific escaping analyses, only peptide pairs derived from the same viral species with one edit distance were considered transitional.

Tumor neoepitope analysis

The compiled dataset of predicted tumor neoepitopes and clinical annotations by Luksza et al.\textsuperscript{11} was utilized. These data were derived from two melanoma cohorts and one lung carcinoma cohort treated with anti-PD-1 and anti-PD-1 checkpoint inhibitors, respectively\textsuperscript{26–28}. The score ratios of mutant peptides relative to their wild-type counterparts were calculated and summed up per patient. The best threshold minimizing the $P$-value by the log-rank test was explored using all cohorts combined. The optimal threshold was 100. Therefore, tumors whose sums of score ratios exceeded this threshold were classified as “high neoepitope burden.”

Computational analysis

Main computational analyses were conducted using R ver. 3.4.4
The latest versions of R packages were consistently used. The key datasets and essential in-house functions were bundled as the R package Repitope, which is publicly distributed in GitHub (https://github.com/masato-ogishi/Repitope/). Other scripts are available upon request.

Statistics

All statistical analysis is exploratory; no predetermined experimental protocols were applied before initiating the entire project. All data inclusion and exclusion criteria were described in the above sections. No accounting for missing data values is applicable. Non-parametric hypothesis testing was employed unless otherwise stated. P-values were adjusted for multiple comparisons on the basis of false discovery rate (FDR) unless otherwise stated. All statistical analyses were conducted in R.
Acknowledgments

We thank Dr. Couture-Cossette and Dr. Ueno for the valuable advice.

**Funding:** This work is not supported by any external funding sources.

**Author contributions:** M.O. designed the study, performed computational analyses, and drafted the manuscript; M.O. and H.Y. wrote the manuscript.

**Competing interests:** The authors declare no competing interests.

**Data availability:** Both the datasets compiled and the custom functions are available as the R package *Repitope* on GitHub (https://github.com/masato-ogishi/Repitope/). Other analytical scripts are available at the GitHub repository and also available upon request.
References

1. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–66 (2006).

2. Sette, A. *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* **153**, 5586–92 (1994).

3. Schmidt, J. *et al.* In silico and cell-based analyses reveal strong divergence between prediction and observation of T-cell-recognized tumor antigen T-cell epitopes. *J. Biol. Chem.* **292**, 11840–11849 (2017).

4. Duan, F. *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* **211**, 2231–2248 (2014).

5. Stranzl, T., Larsen, M. V., Lundegaard, C. & Nielsen, M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* **62**, 357–368 (2010).

6. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).
7. Editorial. The problem with neoantigen prediction. *Nat. Biotechnol.* **35**, 97–97 (2017).

8. Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, (2013).

9. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci.* **112**, E1754–E1762 (2015).

10. Zhang, W. *et al.* Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. *PLoS One* **10**, e0128194 (2015).

11. Luksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).

12. Balachandran, V. P. *et al.* Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **551**, 512–516 (2017).

13. Rasmussen, M. *et al.* Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J. Immunol.* **197**, 1517–1524 (2016).

14. Tian, S., Maile, R., Collins, E. J. & Frelinger, J. A. CD8+ T Cell Activation Is
Governed by TCR-Peptide/MHC Affinity, Not Dissociation Rate. *J. Immunol.* 179, 2952–2960 (2007).

15. Vita, R. *et al.* The Immune Epitope Database 2.0. *Nucleic Acids Res.* 38, (2009).

16. Reche, P. A., Zhang, H., Glutting, J. P. & Reinherz, E. L. EPIMHC: A curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* 21, 2140–2141 (2005).

17. Kuiken, C., Yusim, K., Boykin, L. & Richardson, R. The Los Alamos hepatitis C sequence database. *Bioinformatics* 21, 379–384 (2005).

18. Llano, A., Williams, A., Olvera, A., Silva-Arrieta, S. & Brander, C. Best-Characterized HIV-1 CTL Epitopes: The 2013 Update. *HIV Mol. Immunol.* 2013 3–25 (2013).

19. Tung, C. W. *et al.* POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* 12, 446 (2011).

20. Lata, S., Bhasin, M. & Raghava, G. P. S. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res. Notes* 2, 61 (2009).

21. Olsen, L. R. *et al.* TANTIGEN: a comprehensive database of tumor T cell antigens. *Cancer Immunol. Immunother.* 66, 731–735 (2017).

22. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity
profiling. *Nat. Methods* **12**, 380–381 (2015).

23. Nelson, R. W. *et al.* T Cell Receptor Cross-Reactivity between Similar Foreign and Self Peptides Influences Naive Cell Population Size and Autoimmunity. *Immunity* **42**, 95–107 (2015).

24. Keskin, O., Bahar, I., Badretdinov, A. Y., Ptitsyn, O. B. & Jernigan, R. L. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.* **7**, 2578–2586 (1998).

25. Bjerregaard, A. M. *et al.* An analysis of natural T cell responses to predicted tumor neoepitopes. *Front. Immunol.* **8**, 1–9 (2017).

26. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (80-. ).* **348**, 124–128 (2015).

27. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* 2189–2199 (2014). doi:10.1056/NEJMo1406498

28. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (80-. ).* **350**, 207–211 (2015).

29. Wojciech, L. *et al.* The same self-peptide selects conventional and regulatory CD4+ T cells with identical antigen receptors. *Nat. Commun.* **5**, 5061 (2014).

30. Attanasio, J. & Wherry, E. J. Costimulatory and Coinhibitory Receptor Pathways
in Infectious Disease. *Immunity* **44**, 1052–1068 (2016).

31. Eil, R. *et al.* Ionic immune suppression within the tumour microenvironment limits T cell effector function. *Nature* **537**, 539–543 (2016).

32. Chang, C. H. *et al.* Metabolic Competition in the Tumor Microenvironment Is a Driver of Cancer Progression. *Cell* **162**, 1229–1241 (2015).

33. Luescher, I. F. *et al.* CD8 modulation of T-cell antigen receptor–ligand interactions on living cytotoxic T lymphocytes. *Nature* **373**, 353–356 (1995).

34. Goulder, P. J. R. & Watkins, D. I. HIV and SIV CTL escape: implications for vaccine design. *Nat. Rev. Immunol.* **4**, 630–640 (2004).

35. Bowen, D. G. & Walker, C. M. Mutational escape from CD8+ T cell immunity: HCV evolution, from chimpanzees to man. *J. Exp. Med.* **201**, 1709–14 (2005).

36. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* (2017). doi:10.1038/nature22991

37. Hu, Z., Ott, P. A. & Wu, C. J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.* **18**, 168–182 (2017).

38. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
39. Boniface, J. J., Reich, Z., Lyons, D. S. & Davis, M. M. Thermodynamics of T cell receptor binding to peptide-MHC: evidence for a general mechanism of molecular scanning. *Proc Natl Acad Sci U S A* **96**, 11446–11451 (1999).

40. Zhong, S. *et al.* T-cell receptor affinity and avidity defines antitumor response and autoimmunity in T-cell immunotherapy. *Proc. Natl. Acad. Sci.* **110**, 6973–6978 (2013).

41. Kawashima, S. *et al.* AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, 202–205 (2008).

42. Garcia, K. C., Teyton, L. & Wilson, I. A. Structural basis of T cell recognition. *Annu. Rev. Immunol.* **17**, 369–97 (1999).

43. Cole, D. K. *et al.* T-cell Receptor (TCR)-peptide specificity overrides affinity-enhancing TCR-major histocompatibility complex interactions. *J. Biol. Chem.* **289**, 628–638 (2014).

44. Britanova, O. V. *et al.* Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *J. Immunol.* **192**, 2689–2698 (2014).

45. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
Sidney, J., Peters, B., Frahm, N., Brander, C. & Sette, A. HLA class I supertypes: A revised and updated classification. *BMC Immunol.* **9**, 1–15 (2008).

Kuhn, M. Building Predictive Models in *R* Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

Bischl, B. *et al.* mlr: Machine Learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016).

Efremova, M. *et al.* Targeting immune checkpoints potentiates immunoediting and changes the dynamics of tumor evolution. *Nat. Commun.* **9**, (2018).

McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **1259**–1271 (2017). doi:10.1016/j.cell.2017.10.001

Chowell, D. *et al.* Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science (80-. ).* **4572**, eaao4572 (2017).

Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).

Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Sy*, 1695 (2006).

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, btv494 (2015).
55. R Core Team. R: A Language and Environment for Statistical Computing. (2016).

doi:ISBN 3-900051-07-0
Supplementary Figure 1. Immunogenicity prediction using previously published tools. Immunogenicity was predicted using (a) Immune Epitope Database (IEDB) class I prediction tool, (b) POPISK, and (c) PAAQD. (b-c) POPISK and PAAQD analyses were restricted to HLA-A2-restricted nonamers and nonamers of any HLA restriction, respectively.
**Supplementary Figure 2.** Increased fragment library depths reduced variability in CPP feature calculation. The result of epitope RRYYQKSTEL using AACP scale KESO980101 was plotted as an example. Calculations were iterated five times with different random seeds. A depth of 10,000 was sufficient to calculate representative statistics with minimal variability between iterations. SD, standard deviation; Med, median; TrM, 10% trimmed mean; MAD, median absolute deviation; Skew, skewness; Kurt, kurtosis; SE, standard errors of mean; IQR, interquartile range; Q10, 10% quantile; Q90, 90% quantile.
Supplementary Figure 3. Iterative feature selection. Feature importance was calculated by constructing random forest models using 80% of the dataset, and 100 the most important features were retained. Calculations were iterated five times with different random seeds. (a-b) Shared features in five iterations. In the phase of importance calculation, random forest models were trained with a set of features whose names were sorted in (a) an ascending order, and (b) a descending order, respectively. (c) Shared features between ascending vs. descending consensus feature sets. (d) The distributions of importance values for the 62 the most predictive features. See Methods for the nomenclature of the features.
Supplementary Figure 4. The effects of HLA binding information for immunogenicity prediction.
(a) Prediction without HLA information. (b) Prediction with annotation-based HLA restrictions.
(c-d) Predictions with either (c) rank percentiles or (d) binding categories predicted by NetMHC 4.0 against twelve HLA representatives. Note that (d) depicts the same probabilities as in Figure 3a.
Supplementary Figure 5. Predictive performance of immunogenicity scores stratified by (a) peptide lengths, or (b) restricting HLA supertypes.
Supplementary Figure 6. The effects of the number of features for immunogenicity prediction. Feature importance was calculated, and the most predictive features were selected as described in Supplementary Figure 3 and Methods. The numbers of features retained in each iteration were (a) 100, (b) 75, (c) 50, and (d) 25. The numbers of features retained after feature selection steps and thus used for model training were indicated on top of each plot. Note that (a) is the same as Figure 3a.
Supplementary Figure 7. The effects of sequence-level peptide homology for immunogenicity prediction. Peptides were clustered using the IEDB clustering tool. (a) No clustering (same as Figure 3a). (b-c) Clustering with the similarity threshold of (b) 80% and (c) 60%, respectively.
Supplementary Figure 8. Immunogenicity prediction for peptides restricted to the MHC molecules of other species. Peptide information was retrieved from the IEDB database. MHC information was not utilized for model construction. Peptides presented on (a-b) primate and (c-d) mouse MHCs were subject to the immunogenicity prediction. In (a) and (c), dataset-specific models were constructed, whereas in (b) and (d), the models constructed from the human peptide dataset (N=21,146) were extrapolated.
Supplementary Tables

Supplementary Table 1. The list of HLA genotype-supertype conversion rules derived from Sidney et al.\textsuperscript{46} (A CSV file)

Supplementary Table 2. The list of allowed evidence of immunogenicity for the IEDB dataset. (A CSV file)

Supplementary Table 3. The list of amino acid pairwise contact potential (AACP) scales utilized in this study. (A CSV file)