Epitope immunogenicity prediction through repertoire-wide TCR-peptide contact profiles

Authors: Masato Ogishi[1, 2*] and Hiroshi Yotsuyanagi[2]

Affiliations:

[1] National Center for Global Health and Medicine Hospital, Tokyo, 162-8655, Japan

[2] 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 prediction, T cell receptor repertoire, amino acid contact potential, machine learning

Conflict of Interest (COI): The authors declare no conflicts of interest.

Abstract

Computational methodologies to predict epitopes for cytotoxic T lymphocytes (CTLs) will galvanize vaccine research and pave the way toward targeted immunotherapy of infections and cancer. However, the classification of immunogenic epitopes and non-immunogenic major histocompatibility complex (MHC) class I ligands *in silico* remains difficult. Here, we defined a novel framework quantifying the interactions between a given peptide and T cell receptor (TCR) repertoire. Using 4738 peptide sequences and a pooled TCR repertoire, an epitope classifier with unprecedented accuracy in hold-out validation was constructed. The classifier was applicable to multiple human leucocyte antigen supertypes. The classifier was further validated independently using pathogen epitope datasets and tumor neoepitope datasets. A panel of neoepitope-rich genes were identified using The Cancer Genome Atlas (TCGA) datasets. The R package *Repitope* was implemented to maximize code reusability. This is the first study demonstrating *in silico* CTL epitope prediction with clinically meaningful robustness, thus prospective validation is warranted.

Introduction

The adaptive immune system is driven by antigen recognition. The capability of triggering immune responses is termed 'immunogenicity'. Antigens are processed into fragments of peptides by proteasomes, and coupled to major histocompatibility complex [MHC; also called the human leucocyte antigen (HLA) in humans] molecules on the surface of antigen-presenting cells (APCs). Antigenic peptides presented by MHC-bearing cells are called MHC ligands. Naïve T cells interact with the MHC ligands (MHCLs) via their receptor (T cell receptor, TCR), and successful recognition activates them to initiate subsequent immunological orchestration(*1*). Immunogenic MHCLs are termed 'epitopes'. Conversely, being MHCLs does not ensure immunogenicity(*2*).

Acquired immunity plays an indispensable role in rejecting both pathogens and tumors. Accumulating evidence is shedding light on mutation-derived epitopes, or neoepitopes, as the targets of anticancer T cell immunity. First, the efficacy of immune checkpoint inhibitors correlates with tumor mutational burden(*3–6*). Second, mismatch-repair deficiency, which increases the overall genomic instability and tumor mutational burden, has been shown to predict a better outcome in patients receiving checkpoint blockade therapy(*7*), which eventually led to the FDA approval of the first pan-cancer efficacy biomarker(*8*). Third, the presence of neoepitope-specific T cells in

patients has been established(*9–11*). Finally, even outside the context of immunotherapy, a heavier mutational burden has been shown to predict a longer overall survival through the meta-analysis of genomic sequencing datasets from studies of six tumor types(*12*). Collectively, these observations led to the unprecedented progress of precision immunotherapy initiatives in oncoimmunology. However, personalized anticancer immunotherapy is still at a nascent stage, in important part owing to the lack of a fast and scalable methodology to screen potent neoepitopes. The test-one-by-one strategy is not feasible given the heavy mutational burden observed in most types of cancer, and, albeit extensively studied, immunoinformatics has achieved only minimal success to date in predicting potent neoepitopes from their genomic profiles(*13*).

Two types of computational tools have been explored for applications in epitope prediction. The first type predicts properties in the processes involved in antigen presentation, including antigen processing, peptide transport, and the affinity and stability of the MHC-peptide complex (*14–17*). A limitation of this approach is the high false discovery rate in terms of immunogenicity (i.e., only a small fraction of the peptides predicted and subsequently shown to bind to MHC are actually recognized by T cells and elicit effective responses). The second category of tool used for epitope discovery is aimed at the direct prediction of immunogenicity. Several immunoinformatic tools have

69 been proposed for the prediction of MHC class I (MHC-I) epitopes, which activate
70 cytotoxic T lymphocytes (CTLs) (*18–23*). However, none of them has demonstrated
71 adequate predictive performance in validation datasets, or has been successfully applied
72 to real-world datasets such as tumor neoepitope sequences obtained from immune
73 checkpoint inhibitor clinical trials.

74 Epitopes, by definition, are recognized by T cells via TCRs. However, in contrast to
75 the MHC-peptide axis, the TCR-peptide axis has yet to be thoroughly examined in the
76 context of immunogenicity. Exceptional research led by Chowell et al. demonstrated that
77 immunogenicity prediction was improved by incorporating hydrophobicity at TCR
78 contact residues(*23*). However, their model focuses on the biochemical properties of
79 MHC-I-loaded peptides but does not specifically address TCR-peptide interactions
80 themselves. On the other hand, a groundbreaking study conducted by Strønen et al. shed
81 light on the TCR-dependent nature of peptide immunogenicity in the context of
82 oncoimmunology(*24*). They showed that naïve T cell repertoires of healthy blood donors
83 were able to trigger effective immune responses against a variety of neoepitopes isolated
84 from cancer patients treated with immune checkpoint inhibitors. Many of the targeted
85 neoepitopes were overlooked by autologous tumor-infiltrating lymphocytes *in vivo*.
86 Furthermore, patient-derived T cells transformed with the appropriate donor TCR

87 successfully invigorated anti-neoepitope immunity. Their results suggest that even
88 MHCLs non-immunogenic to autologous TCRs can serve as epitopes if recognized by
89 appropriate TCRs.

90 We started the whole project aiming at unveiling the enigma of the
91 immunogenicity on MHC-I-loaded peptides on the basis of the following hypothesis:
92 are peptides stably interacting with the host TCR repertoire more likely to be
93 immunogenic? If this is the case, prediction of peptide immunogenicity may be
94 significantly improved by incorporating the TCR-peptide axis. Given that human TCR
95 repertoires are evolutionarily optimized so as to effectively combat pathogens and
96 cancers, we utilized a pooled human TCR repertoire sequenced from the commercial
97 RNA of peripheral blood $CD8^+$ T cells for reference. We defined repertoire-wide
98 TCR-peptide contact profiles (rTPCP) using amino acid pairwise contact potential
99 (AACP) scales to quantitatively parametrize TCR-peptide interactions to classify
100 epitopes and MHCLs through a machine learning (ML) approach. Our initial model
101 achieved unprecedented accuracy in hold-out validation. When the rTPCP definition
102 was modified to incorporate position-specific effects (mrTPCP), comparable accuracy
103 was achieved with just one AACP scale. Prediction was not biased for at least five HLA
104 supertypes. Permutation of peptide sequences, but not TCR sequences, undermined

105 predictive accuracy. Successful prediction was demonstrated using independent
106 epitope/MHCL datasets of viral and tumor origin. Moreover, using a mutational
107 landscape dataset obtained from checkpoint inhibitor trials, a correlation between
108 predicted neoepitope burden and clinical outcome was shown. Overall, this is the first
109 study demonstrating a highly accurate and generalizable epitope prediction by
110 integrating the TCR-peptide axis. The codes for rTPCP and mrTPCP analysis were
111 compiled into the R package *Repitope* (https://github.com/masato-ogishi/Repitope/).
112 Prospective validation of this tool in independent cohorts of vaccination and cancer
113 immunotherapy is necessary to evaluate its possible clinical applicability.
114

115 # Results

116 **Preparation of pooled human TCR repertoire dataset**

117 First, we screened public databases such as Sequence Read Archive, but failed to
118 find a suitable human TCR sequence dataset. Therefore, we generated an in-house TCR
119 repertoire data by sequencing the variable regions of TCR β chains (TCR-Vβ) from
120 commercially available pooled human peripheral $CD8^+$ T cells. Among the three
121 complementarity-determining regions (CDRs), we focused on CDR3, because it has the
122 largest diversity among CDR regions, and CDR1 and CDR2 are primarily involved in the
123 recognition of MHC, not the ligand presented(*1*). Rarefaction analysis estimated the total
124 CDR3 clonotype diversity be approximately 1500, out of which 872 unique clonotypes
125 were identified (fig. S1). No apparent bias in CDR3 length or Variable (V) and Junction
126 (J) segment usage was observed (fig. S2).

127 **Immunogenicity prediction from repertoire-wide TCR-peptide contact profiles**
128 **parametrized using amino acid contact potentials**

129 Immunogenicity prediction model necessitates quantitative parametrization of the
130 likelihood that a given peptide stably interacts with a given set of TCRs. Although
131 molecular dynamics simulation would be the most accurate method, it is not appropriate

132 because of its high demand for computational power; our goal is to construct a "portable"
133 prediction framework that can be run on ordinary desktop computers. To simplify the
134 framework, we adopted a sequence-based prediction strategy using AACPs listed in the
135 AAIndex database(*25*) (http://www.genome.jp/aaindex/AAindex/list_of_potentials) as
136 the measurement of energetic stability, or the decrease in free energy, of TCR-peptide
137 interaction. We hereby propose the concept of rTPCP, where a given peptide contacts
138 with all TCRs in a given repertoire with varying contact potentials (Fig. 1; see
139 Supplementary Materials and Methods for details). Using the rTPCP variables, we
140 attempted ML-based classification of MHCL peptides into immunogenic (functional
141 epitope) and non-immunogenic subsets. We utilized the peptide dataset compiled by
142 Chowell et al., which contains 7582 distinct human peptides (*23*). Preliminary analyses
143 suggested support vector machine (SVM) as the most accurate and balanced algorithm.
144 As an initial attempt, we focused on 450 epitopes and 450 ligands restricted on human
145 leucocyte antigen A2 (HLA-A2). We retrieved 35 AAIndex AACP scales (table S1) to
146 calculate rTPCP variables. To our surprise, the resultant SVM-based immunogenicity
147 prediction model achieved an unprecedentedly high predictive performance in the
148 hold-out validation dataset [accuracy, 0.81; 95% confidence interval (CI), 0.76 to 0.86;
149 receiver operating characteristic (ROC) area under the curve (AUC), 0.87; 95% CI, 0.83

150 to 0.91] (Fig. 2A). Four additional iterations using different random seeds yielded
151 comparable results (table S2).

152 One caveat of ML-based prediction is over-parametrization, which may lead to
153 model instability and limited generalizability, as is the case for our model (2520 rTPCP
154 variables against 900 HLA-A2-restricted peptides). Variable importance analysis
155 revealed that the AAIndex AACP scale MIYS990106, which represents inter-residue
156 pairwise contact energies(*26*), yielded the most consistently important variables (fig.
157 S3). Therefore, we retrained the SVM-based classifier solely using the
158 MIYS990106-derived rTPCP variables. This time, we included a full set of
159 epitopes/MHCLs in the dataset to maximize overall data size, resulting in a matrix with
160 72 variables for 7575 distinct peptides. The model achieved considerably high
161 performance despite the relatively small number of variables (accuracy, 0.75; 95% CI,
162 0.73 to 0.76; AUC, 0.81; 95% CI, 0.80 to 0.83) (Fig. 2B). Four additional iterations with
163 different random seeds yielded comparable results (table S3).

164 **Improved immunogenicity prediction by incorporating position-specific contact**
165 **profiles**

166 MHC-loaded peptides interact with TCRs at specific positions(*1*). The effects of
167 position-specific interactions may counterbalance each other in TPCP. To test this

168 hypothesis, we modified the rTPCP definition to incorporate position-specific
169 interactions (mrTPCP; schematically depicted in Figure 3). In this iteration, every TCR
170 was fragmented and pooled to generate a TCR fragment repertoire, and representative
171 statistics were calculated on a set of AACPs (see Supplementary Materials and Methods
172 for details). Because of the position-specific nature of the analysis, we limited the
173 subsequent analysis to 4738 unique nonapeptides in the Chowell dataset. The
174 SVM-based classifier trained from 187 mrTPCP variables outperformed our previous
175 rTPCP-based classifier (accuracy, 0.77; 95% CI, 0.75 to 0.79; AUC, 0.84; 95% CI, 0.82
176 to 0.86), with statistical significance [$p = 0.048$, according to the *roc.test* function
177 implemented in the *pROC* package(*27*)] (Fig. 4A). For comparison, the same dataset
178 was used to test three previously published immunogenicity prediction tools with
179 publicly available source code or web implementation. However, none of the tested
180 tools achieved similarly meaningful prediction; the accuracies were 0.56, 0.59, and 0.57
181 for POPISK(*19*), PAAQD(*20*), and EpitopePrediction(*28*), respectively.

182 The amino acid compositions of MHCLs are restricted by the HLA to which they
183 are coupled. Since our mrTPCP framework is not dependent on HLA information, it
184 might be useful for pan-specific immunogenicity prediction. To test this hypothesis,
185 4738 unique nonapeptides in the Chowell dataset were stratified based on their

186 corresponding HLA supertypes, and ROC analysis was conducted (Fig. 4B). The trained
187 classifier worked with no significant decrease in accuracy for at least five major HLA
188 supertypes (HLA-A1, A2, B15, B44, and B57) for which a sufficient amount of peptide
189 data was available.

190 Previous studies suggest that position-specific amino acid usage biases in
191 MHC-coupled peptides affect their immunogenicity(*21*, *23*). In our model, windows 1
192 and 2 seemed to be of higher importance, but no exceptionally important window was
193 identified (fig. S4). To further evaluate these position-dependent characteristics, we next
194 conducted sequence manipulation analysis; mrTPCP variables were calculated for
195 manipulated peptide sequences or using manipulated reference TCR repertoire
196 sequences. The classifier trained from authentic TCRs and peptides was then applied to
197 perform ROC analysis. Manipulation of TCRs led to a minimal decrease in AUC,
198 whereas manipulation of peptides led to a significant decrease in AUC (Fig. 4C).
199 Difference in amino acid compositions between epitopes and MHCLs was only of
200 partial predictive significance, indicating that position-specific or sequence-specific
201 features are the major determinants of immunogenicity.

202 Collectively, these observations suggest that the mrTPCP framework effectively
203 mimics the biological mechanisms of CTL immunogenicity, thereby providing a

204 promising methodology for accurate epitope prediction.

205 **Immunogenicity prediction using independent datasets**

206 Any pattern learned from one dataset is not always extendable to other datasets
207 constructed in different contexts. Therefore, we tested the performance of our
208 immunogenicity prediction model by utilizing independent datasets adopted from
209 previous publications(*4*, *10*, *24*, *29–32*), after removing peptides overlapping with those
210 in the Chowell dataset. As expected, randomly selected 10,000 MHCLs retrieved from
211 the Immune Epitope Database (IEDB) were predicted as either immunogenic or
212 non-immunogenic in an approximately 1:1 ratio, with a uniform distribution of
213 predicted probabilities (Fig. 5A and Table 1). In contrast, epitope datasets of viral and
214 tumor origin were significantly enriched with peptides predicted as epitopes ($p < 0.01$ by
215 Wilcoxon's rank sum test in comparison with randomly selected MHCLs from IEDB). It
216 is notable that 16 out of 22 (73%) well-defined neoepitopes and 22 out of 35 (63%) best
217 neoepitopes reported by Stronen et al. had probabilities of > 0.80 (Fig. 5A). With the
218 probability threshold of 0.80, our model also effectively classified the epitope/MHCL
219 dataset from various pathogens originally reported by Calis et al.(*21*) (accuracy, 0.71;
220 95% CI, 0.67 to 0.75; AUC, 0.77; 95% CI, 0.71 to 0.82) (Fig. 5B).

221 Encouraged by these observations, we next explored the possibility that our

222 immunogenicity prediction model improves the correlation between neoepitope burden
223 and clinical outcomes in checkpoint inhibitor trials. First, we adopted clinical and
224 mutational data from non-small cell lung carcinoma (NSCLC) patients treated with
225 pembrolizumab (n = 23)(*3*). We observed a slightly improved correlation between
226 neoepitope burden and progression-free survival (PFS) ($R = 0.55$, $p = 0.007$), compared
227 with the correlation between originally reported mutated peptide burden and PFS ($R =$
228 0.61, $p = 0.002$), although the improvement is not statistically significant as determined
229 by the methods implemented in the *cocor* package(*33*) (Fig. 6A). The PFSs of three
230 patients, namely, CA9903, CU9061, and SA9755, were better predicted (Fig. 6A). Next,
231 we analyzed clinical and mutational data from melanoma patients treated with
232 ipilimumab (n = 110)(*5*). Clinical benefit (CB) was defined as originally reported(*5*).
233 There were significant differences in both mutational burden and predicted neoepitope
234 burden between patients with and without CB (Fig. 6B). Overall, our results showed
235 that neoepitope burden predicted through the mrTPCP framework retains at least
236 comparable usefulness as a biomarker as compared with conventional mutational
237 burden, with greatly reduced number of neoepitope candidates, enabling more focused
238 approach in view of precision immunotherapy.
239 Finally, we compared estimated neoepitope burden across 21 tumor types in The

240 Cancer Genome Atlas (TCGA)(*34*). HLA-A-02:01 was chosen for subsequent analysis
241 as an example. Using the *EpitopePrediction* package(*28*), a total of 108,730 9-mer
242 MHCLs, of which 105,959 were unique, were identified. Immunogenicity prediction
243 was performed as described, with the probability threshold of 0.80. A total of 69,587
244 (64%) mutated peptides were predicted as neoepitopes. Skin cutaneous melanoma
245 (SKCM), lung squamous cell carcinoma (LUSC), and lung adenocarcinoma (LUAD)
246 were the three most MHCL-enriched, and neoepitope-enriched types of cancer (Fig. 7A).
247 There was a significant gene-by-gene variation of the ratio of neoepitope burden to the
248 MHCL burden (Fig. 7B). Mitochondrial enzymes (MT-CO1 and MT-ND4) and
249 olfactory receptors (OR2T2, OR4A5, OR4C16, OR4K2, OR5J2, and OR7D4) were the
250 genes that were particularly high-yield in terms of neoepitopes.

251 **R package implementation of immunogenicity prediction framework**

252 We implemented the R package *Repitope* to maximize code reusability. *Repitope*
253 contains datasets used in this study, functions to calculate rTPCP and mrTPCP variables
254 for user-provided peptide datasets and reference TCR repertoire data, and the mrTPCP
255 SVM classifier developed in this study. Source codes are deposited for public use at
256 GitHub (https://github.com/masato-ogishi/Repitope/).

257

Discussion

In this work, the accurate classification of epitopes and non-immunogenic MHC-I ligands was achieved by introducing the concept of repertoire-wide TCR-peptide contact profiles. Considering that current concepts of CTL epitope prediction are mostly focused on the peptide-MHC axis, it is of interest that our immunogenicity prediction model incorporating the TCR-peptide axis showed improved predictive capability over previous models.

We decided to use the dataset previously compiled by Chowell et al. for the following reasons. First, we eschewed compiling peptide datasets from scrach to avoid potential selection bias. Second, the dataset contains a sufficiently large number of human peptide data from IEDB, the largest and least biased data source available. Finally, the mutual exclusiveness of epitopes and MHCLs included in the dataset is ideal for model training and evaluation; the immunogenic CTL epitopes included were defined by T cell assays, and non-immunogenic MHCLs included were proven by MHC ligand elution assays, with any potentially immunogenic eluted ligand associated with autoimmunity or cancer being excluded. Consequently, the SVM classifier developed in this work successfully classified epitopes and MHCLs with unprecedented accuracy (Figs. 2 and 4). Moreover, our model also improved upon previous ones in that it employs smaller

276 number of variables(*19*, *22*) (fig. S3). Generally, ML classifiers using smaller numbers of
277 variables are preferable, since over-parametrization frequently causes ML algorithms to
278 "cheat", or to find variables distributed unevenly between the two classes in question just
279 because of stochastic fluctuation (with no generalizability for external data). Our
280 mrTPCP model employs only one AACP scale for parametrization, which resulted in 187
281 mrTPCP variables. This is a fairly small size when considering the number of input
282 peptides.

283 Our mrTPCP framework has two notable features: independence from HLA
284 specificity, and dependence on a reference TCR repertoire. First, pan-specific
285 immunogenicity prediction may be feasible, as it does not depend on HLA information.
286 We showed that our model worked with minimal performance reduction for at least five
287 major HLA supertypes (HLA-A1, A2, B15, B44, and B57), for which sufficient amount
288 of peptide data was available (Fig. 4B). This point could further be explored with more
289 immunogenicity data for various HLA alleles in the future. Second, the framework
290 requires reference TCR repertoire. The model discussed in this study relies on the pooled
291 TCR repertoire of German origin, which could be a source of bias. However,
292 immunogenicity could still be predicted with a minimal decrease in AUC, even when
293 using completely random sequences instead of TCR repertoire (Fig. 4C). Conversely,

294 manipulation of input peptide sequences resulted in a significant decrease in predictive
295 accuracy (Fig. 4C). These observations suggested that the mrTPCP framework is
296 primarily dependent on the inherent features of epitope sequence but not the reference
297 repertoire. Interestingly, peptide sequence permutation and randomization with relative
298 amino acid compositions retained led to moderately decreased AUC (0.68 and 0.64,
299 respectively), whereas completely random peptide sequences could not be classified
300 (AUC = 0.50). This is consistent with the previous research of Calis et al., in which an
301 AUC of 0.65 was obtained when residue-specific properties but not sequence-specific
302 properties were taken into consideration(*21*). Collectively, both amino acid composition
303 and sequence-specific features recapitulated by the mrTPCP framework are important in
304 determining peptide immunogenicity.

305 The regulatory mechanisms of CTL activation are asymmetric, and it is this
306 asymmetry that makes the construction of immunogenicity prediction models
307 complicated. The activation part is relatively simple; stable and strong interactions in the
308 TCR-peptide-MHC complex are the main driving force of T cell activation(*1*). In contrast,
309 there are several immunomodulatory systems outside the TCR-peptide-MHC axis
310 affecting the T cell activation process *in vivo*, including regulatory T cells (Tregs)(*35*),
311 CTL exhaustion mediated by chronic immune checkpoint signals, and the

312 immunosuppressive microenvironment engendered by solid tumors(*36*, *37*). Considering
313 this asymmetry, immunogenicity prediction models based solely on peptide sequence
314 may in principle yield some false positives. Therefore, our results should be recognized as
315 preliminary, warranting prospective validation to evaluate their clinical applicability.
316 That being said, however, eliminating candidates least likely to be immunogenic *in silico*
317 should greatly expedite research in targeted immunotherapy, and the findings in our
318 present study are indeed encouraging; epitopes of viral and tumor origin not included in
319 the training/testing dataset were successfully predicted with high sensitivity, whereas
320 predicted probabilities of MHCLs randomly retrieved from IEDB distributed almost
321 uniformly from 0 to 1 (Fig. 5 and Table 1). It is reasonable to assume a distribution of
322 levels of immunogenicity in the dataset of randomly selected MHCLs without T cell
323 assay-based annotation. Furthermore, we showed that the usefulness of neoepitope
324 burden as a biomarker for clinical outcome was not affected, or even slightly improved,
325 when candidate mutations were filtered using our prediction model (Fig. 6). One caveat to
326 be mentioned is its relatively low sensitivity in predicting HIV epitopes. In addition to the
327 "general" rules learned from the Chowell dataset which contains epitopes from various
328 sources, some additional rules may be critical for HIV-specific CTL immunity and could
329 be machine-learned with more data obtained specifically in the context of chronic HIV

330 infection.

331 Immune checkpoint inhibitors achieved revolutionary success in some types of
332 tumor including melanoma and non-small cell lung carcinoma (NSCLC)(*38–40*).
333 However, relatively few explanations have been proposed about the reason why these two
334 types of tumor are the most sensitive to checkpoint blockade. To address this, we
335 explored the entire TCGA cancer genome dataset(*34*). As anticipated, skin cutaneous
336 melanoma and NSCLC were most enriched with predicted MHCLs or neoepitopes (Fig.
337 7A). The ratio of predicted neoepitope to predicted MHCL significantly varied by
338 respective gene analyzed (Fig. 7B). Focusing on the high-yield genes such as
339 mitochondrial enzymes and olfactory receptors may expedite the development of
340 pan-cancer targeted immunotherapy.

341 Similarly to previous studies on immunogenicity prediction, this study has several
342 limitations. First, this is a retrospective observational study; no prospective identification
343 of novel epitopes is demonstrated. Thus, prospective validation is indispensable before
344 this model can be clinically applied. Second, the process of quantitative parametrization
345 of TCR-peptide interactions could further be optimized, as our window-based pairwise
346 interaction model might oversimplify the biophysicochemical nature of the
347 TCR-peptide-MHC interactions. In particular, the hypothesis that either a 4-mer or

348 5-mer window size is sufficient for recapitulating TCR-peptide interactions is not
349 experimentally verified, necessitating further exploration. Moreover, we limited our
350 modeling to TCR-Vβ, omitting TCR-Vα; this point could further be explored. Despite
351 these caveats, however, both the proposed framework of mrTPCP recapitulating the
352 biology of TCR-peptide interactions and the demonstrated robustness of
353 immunogenicity prediction represent noticeable progress toward fully unveiling the
354 mechanisms underlying CTL immunity, paving the way toward precision
355 immunotherapy against pathogens and cancer.

356 In conclusion, accurate epitope prediction was achieved through a machine
357 learning approach by incorporating TCR-peptide interactions parametrized using an
358 optimal amino acid pairwise contact potential scale. Unbiased prediction was
359 demonstrated for peptides coupled to multiple major HLA supertypes. The framework
360 was primarily reliant on the sequence-dependent features of the peptides, and only
361 minimally affected by the perturbation of the reference TCR repertoire. The resultant
362 classifier worked well for independent viral epitopes and tumor neoepitopes. These
363 findings not only provide valuable insights into the mechanisms underlying CTL
364 immunity, but could also bolster the ongoing precision immunotherapy initiatives. Code
365 reusability was maximized by publicly distributing the R package *Repitope*, in which

datasets and key scripts are bundled. Disease-specific, prospective cohort studies could be conducted to evaluate clinical usefulness in the future.

Materials and Methods

Study design

Research objectives. The purpose of this study was to construct a sequence-based epitope prediction model by incorporating TCR-peptide contact profiles.

Design. This is a retrospective, observational study. The entire analysis is exploratory; no predetermined experimental protocol was applied *a priori*.

Data collection. Peptide sequences accompanied by annotations on immunogenicity and other clinical profiles (if applicable) were manually retrieved from public database and previously published articles by the authors.

Data size. The optimal sizes of the epitope and MHCL datasets are unknown, since we hereby proposed a novel framework. Therefore, no statistical estimation was performed to predetermine sample size.

Computational analysis

All in-house computational analyses were conducted using R ver. 3.4.0 (https://www.r-project.org/) (*42*). The latest versions of R packages were consistently used. Key datasets and scripts were bundled as the R package *Repitope*, and publicly distributed in GitHub (https://github.com/masato-ogishi/Repitope/). Other scripts are available upon request.

387 **Preparation of pooled human TCR repertoire dataset**

388 TCR repertoire sequencing was carried out as previously described(*43*), except the
389 primers being used. Briefly, total RNA from $CD8^+$ T cells collected from donated
390 peripheral blood of German origin was purchased (Miltenyi Biotec) and used as the
391 source of a pooled TCR repertoire. Primers for human TCR-Vβ were adopted from
392 previously published work(*44*). All primers were synthesized by Life Technologies, and
393 a template-switching oligo (TSO) containing 5′ terminal unique molecular identifier
394 (UMI) and 3′ terminal guanidine locked nucleic acid (LNA) was synthesized by
395 Exiqon(*45*). The sequences of the oligonucleotides utilized in this study are summarized
396 in table S4. Reverse transcription with template-switching and semi-nested step-out PCR
397 were performed using SMARTScribe (Clontech) and KAPA2G Fast Multiplex PCR
398 master mix (Kapa Biosystems)(*46*). Amplified PCR products were gel-excised, repaired,
399 and re-purified. Paired-end libraries were prepared, and paired-end sequencing of 2 x 300
400 bp was performed using MiSeq (Illumina). UMI-guided de-multiplexing was performed
401 using MiGEC software in order to reduce the effect of PCR amplification bias(*47*). CDR3
402 regions were identified using IMGT/HighV-QUEST(*48*)
403 (https://www.imgt.org/HighV-QUEST/).

404 **Epitope/ligand dataset for training/testing immunogenicity prediction model**

405 The dataset primarily utilized in this study is originated from the research led by
406 Chowell et al(*23*). Any peptide derived from a mouse experiment was removed to create
407 a human-specific immunogenicity dataset. No additional data filtering was performed to
408 avoid deliberate peptide selection.

409 **Machine learning for immunogenicity prediction**

410 Machine learning (ML) procedures were streamlined using the *caret* package in
411 R(*49*). The hold-out validation strategy was adopted; the input dataset was randomly split
412 into training and testing subdatasets in a ratio of 2:1. The training subdataset was
413 preprocessed (i.e., centered and scaled) using the *preProcess* function in *caret*. Ten-fold
414 cross-validations (CVs) were repeated ten times to train classifiers. Testing subdataset
415 was preprocessed using the preprocessing model generated from the training subdataset,
416 and immunogenicity was predicted. Unless otherwise noted, the performance metric in
417 each testing subdataset was reported. As any ML algorithm is designed to self-optimize
418 through CVs, the performance metric obtained in the process of CVs is an optimized
419 value for the input dataset. Our true interest is the performance of the trained classifier
420 when applied to an external dataset not involved in either model training or optimization.
421 Preliminary assessment suggested that the support vector machine (SVM) was the best
422 algorithm. SVM has a long history of providing state-of-the-art, well-generalizable

423 predictions in various biological contexts(*50*). Four SVM methods, namely, *svmLinear*,
424 *svmPoly*, *svmRadial*, and *svmRadialSigma*, were tested. We chose *svmPoly* as the best
425 algorithm on the basis of various factors including accuracy, AUC, balance between
426 sensitivity and specificity, and the smoothness of the calibration curve. Accuracy was
427 calculated using the *confusionMatrix* function implemented in *caret*, and AUC was
428 calculated using either the *classifierplots* function in the *classifierplots* package, or the
429 *roc* and *auc* functions implemented in the *pROC* package(*27*).

430 **Epitope/ligand datasets for external validation**

431 The hold-out validation strategy is by itself not sufficient for evaluating the
432 generalizability of the ML classifier for external datasets. The ML algorithm, after all,
433 mines hidden patterns applicable across the training dataset. When the hold-out
434 validation strategy is adopted, training and testing subdatasets derived from a single
435 data source lie in a single context, and consequently, patterns learned from the training
436 subdataset is highly likely applicable to the testing subdataset. Therefore, the trained
437 classifier should be tested and validated with other external datasets constructed in
438 different contexts. In this study, the trained classifier may be biased, since
439 autoimmunity- and cancer-associated immunogenic peptides were excluded from the
440 epitope data, and pathogen-derived MHCLs were excluded from the MHCL data in the

441 Chowell dataset.

442 To independently assess the generalizability of the trained immunogenicity
443 prediction model, epitope datasets were collected from the following sources: (i)
444 hepatitis C virus (HCV) CTL epitopes
445 (https://hcv.lanl.gov/content/immuno/tables/ctl_summary.html), (ii) human
446 immunodeficiency virus type-I (HIV) CTL epitopes
447 (https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html), (iii)
448 well-established tumor neoepitopes from multiple publications(*4*, *10*, *29–32*), (iv)
449 neoepitopes predicted to be the most stable MHC binders identified in the study by
450 Stronen et al.(*24*), and (v) all neoepitopes identified in the study by Stronen et al. For
451 comparison, we also obtained a human MHCL dataset from IEDB (http://www.iedb.org/).
452 Note that the MHCL data lacks T cell assay annotation, and thus the true ratio of 'epitopes'
453 to 'MHCLs' in the definitions discussed in this study is unknown. Moreover, we retrieved
454 the epitope/MHCL dataset originally reported by Calis et al(*21*). This dataset is suitable
455 for assessing the specificity of our immunogenicity prediction model, because it contains
456 experimentally validated non-immunogenic MHCLs, mostly originated from dengue
457 virus. Peptide sequences containing alphabetical characters other than those representing
458 20 authentic amino acid residues were removed. Any peptide contained in the Chowell

459 dataset was excluded. Datasets are available as supplementary data files (Data files
460 S1-S6).

461 **Correlation with clinical outcomes in checkpoint inhibitor trials**

462 Correlation between mutational landscapes and clinical outcome has been shown
463 in various tumor types in checkpoint inhibitor trials(*3–5*) To test the predictive
464 usefulness of neoepitope burden predicted through the proposed framework, we
465 re-analyzed mutational datasets from two studies(*3*, *5*). Datasets are available as
466 supplementary data files (Data files S7 and S8).

467 **Neoepitope burden across TCGA tumor types**

468 To assess the difference in neoepitope burden across tumor types, we analyzed
469 genomic datasets derived from The Cancer Genome Atlas project(*34*). For all advanced
470 stage (Stage III and Stage IV) tumors, mutation annotation format (MAF) files were
471 downloaded from the National Cancer Institute (NCI) Genomic Data Commons (GDC)
472 Data Portal (https://portal.gdc.cancer.gov/). Mutation data were parsed, and
473 nonapeptides harboring mutation were reconstructed by referencing the UniProt human
474 proteome (ID: UP000005640). HLA-A-02:01 is used as a representative allele, and the
475 binding stability of all nonapeptides were estimated using the *EpitopePrediction*

476 package(*28*). The dataset is available as a supplementary data file (Data file S9).

477 **Statistical analysis**

478 No variable distribution was assumed *a priori*, and data were presented as median
479 and interquartile range, unless otherwise stated. *P* values were reported unadjusted unless
480 otherwise stated. No accounting for missing data values is applicable. All statistical
481 analysis is exploratory; no predetermined experimental protocols were applied before
482 initiating the entire project. All statistical analyses were conducted in R.

483

Supplementary Materials

Supplementary Materials and Methods

Fig. S1. Clonotype rarefaction analysis in the reference TCR repertoire.

Fig. S2. Composition of the reference TCR repertoire..

Fig. S3. Exploration of AACP scales most important in the immunogenicity prediction models trained using rTPCP variables.

Fig. S4. Variable importance analysis of the immunogenicity prediction model trained using mrTPCP variables.

Table S1. AAIndex AACP scales used in this study.

Table S2. Performance evaluation of SVM classifiers with rTPCP variables from all AAIndex AACP scales.

Table S3. Performance evaluation of SVM classifiers with rTPCP variables from MIYS990106 AACP scale.

Table S4. Oligonucleotides used in this study.

Data file S1. HCV CTL epitopes.

Data file S2. HIV CTL epitopes.

Data file S3. Well-established tumor neoepitopes from multiple publications.

Data file S4. Best-predicted tumor neoepitopes reported by Stronen et al.

502 Data file S5. All neoepitopes reported by Stronen et al.

503 Data file S6. Pathogen-derived epitopes reported by Calis et al.

504 Data file S7. Tumor neoepitopes and clinical data reported by Rizvi et al.

505 Data file S8. Tumor neoepitopes and clinical data reported by van Allen et al.

506 Data file S9. Tumor neoepitope candidates identified from TCGA datasets.

507

508 # References

509 1. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How TCRs bind MHCs, peptides,
510 and coreceptors. *Annu. Rev. Immunol.* **24**, 419–66 (2006).
511 2. G. T. Gibney, L. M. Weiner, M. B. Atkins, Predictive biomarkers for checkpoint
512 inhibitor-based immunotherapy. *Lancet Oncol.* **17**, e542–e551 (2016).
513 3. N. A. Rizvi *et al.*, Mutational landscape determines sensitivity to PD-1 blockade
514 in non-small cell lung cancer. *Science (80-.).* **348**, 124–128 (2015).
515 4. A. Snyder *et al.*, Genetic Basis for Clinical Response to CTLA-4 Blockade in
516 Melanoma. *N. Engl. J. Med.*, 2189–2199 (2014).
517 5. E. M. Van Allen *et al.*, Genomic correlates of response to CTLA-4 blockade in
518 metastatic melanoma. *Science (80-.).* **350**, 207–211 (2015).
519 6. L. M. Colli *et al.*, Burden of nonsynonymous mutations among TCGA cancers
520 and candidate immune checkpoint inhibitor responses. *Cancer Res.* **76**, 3767–
521 3772 (2016).
522 7. A. D. Skora *et al.*, PD-1 Blockade in Tumors with Mismatch-Repair Deficiency.
523 *N. Engl. J. Med.*, 2509–2520 (2017).
524 8. K. Garber, Oncologists await historic first: a pan-tumor predictive marker, for
525 immunotherapy. *Nat. Biotechnol.* **35**, 297–298 (2017).

526 9. M. M. Gubin *et al.*, Checkpoint blockade cancer immunotherapy targets
527 tumour-specific mutant antigens. *Nature*. **515**, 577–581 (2014).
528 10. N. van Rooij *et al.*, Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell
529 Reactivity in an Ipilimumab-Responsive Melanoma. *J. Clin. Oncol.* **31**, e439–
530 e442 (2013).
531 11. A. Gros *et al.*, Prospective identification of neoantigen-specific lymphocytes in
532 the peripheral blood of melanoma patients. *Nat. Med.* (2016),
533 doi:10.1038/nm.4051.
534 12. S. D. Brown *et al.*, Neo-antigens predicted by tumor genome meta-analysis
535 correlate with increased patient survival. *Genome Res.* **24**, 743–750 (2014).
536 13. Editorial, The problem with neoantigen prediction. *Nat. Biotechnol.* **35**, 97–97
537 (2017).
538 14. M. Nielsen, C. Lundegaard, O. Lund, C. Keşmir, The role of the proteasome in
539 generating cytotoxic T-cell epitopes: insights obtained from improved predictions
540 of proteasomal cleavage. *Immunogenetics*. **57**, 33–41 (2005).
541 15. T. Stranzl, M. V. Larsen, C. Lundegaard, M. Nielsen, NetCTLpan: pan-specific
542 MHC class I pathway epitope predictions. *Immunogenetics*. **62**, 357–368 (2010).
543 16. M. Nielsen, M. Andreatta, NetMHCpan-3.0; improved prediction of binding to

544 MHC class I molecules integrating information from multiple receptor and
545 peptide length datasets. *Genome Med.* **8**, 33 (2016).
546 17. M. Rasmussen *et al.*, Pan-Specific Prediction of Peptide–MHC Class I Complex
547 Stability, a Correlate of T Cell Immunogenicity. *J. Immunol.* **197**, 1517–1524
548 (2016).
549 18. C. W. Tung, S. Y. Ho, POPI: Predicting immunogenicity of MHC class I binding
550 peptides by mining informative physicochemical properties. *Bioinformatics*. **23**,
551 942–949 (2007).
552 19. C. W. Tung, M. Ziehm, A. Kamper, O. Kohlbacher, S. Y. Ho, POPISK: T-cell
553 reactivity prediction using support vector machines and string kernels. *BMC*
554 *Bioinformatics*. **12**, 446 (2011).
555 20. T. Saethang *et al.*, PAAQD: Predicting immunogenicity of MHC class I binding
556 peptides using amino acid pairwise contact potentials and quantum topological
557 molecular similarity descriptors. *J. Immunol. Methods*. **387**, 293–302 (2013).
558 21. J. J. A. Calis *et al.*, Properties of MHC Class I Presented Peptides That Enhance
559 Immunogenicity. *PLoS Comput. Biol.* **9** (2013),
560 doi:10.1371/journal.pcbi.1003266.
561 22. W. Zhang *et al.*, Accurate prediction of immunogenic T-cell epitopes from

562 epitope sequences using the genetic algorithm-based ensemble learning. *PLoS*
563 *One*. **10**, e0128194 (2015).
564 23. D. Chowell *et al.*, TCR contact residue hydrophobicity is a hallmark of
565 immunogenic CD8 + T cell epitopes. *Proc. Natl. Acad. Sci.* **112**, E1754–E1762
566 (2015).
567 24. E. Strønen *et al.*, Targeting of cancer neoantigens with donor-derived T cell
568 receptor repertoires. *Science (80-.).* **2288**, 1–11 (2016).
569 25. S. Kawashima *et al.*, AAindex: Amino acid index database, progress report 2008.
570 *Nucleic Acids Res.* **36**, 202–205 (2008).
571 26. S. Miyazawa, R. L. Jernigan, Self-consistent estimation of inter-residue protein
572 contact energies based on an equilibrium mixture approximation of residues.
573 *Proteins*. **34**, 49–68 (1999).
574 27. X. Robin *et al.*, pROC: an open-source package for R and S+ to analyze and
575 compare ROC curves. *BMC Bioinformatics*. **12**, 77 (2011).
576 28. Y. Kim, J. Sidney, C. Pinilla, A. Sette, B. Peters, Derivation of an amino acid
577 similarity matrix for peptide:MHC binding and its application as a Bayesian prior.
578 *BMC Bioinformatics*. **10**, 394 (2009).
579 29. B. M. Carreno *et al.*, A dendritic cell vaccine increases the breadth and diversity

580 of melanoma neoantigen-specific T cells. *Science (80-.).* **348**, 803–808 (2015).

581 30. E. Tran *et al.*, Immunogenicity of somatic mutations in human gastrointestinal

582 cancers. *Science (80-.).* **350**, 1387–1390 (2015).

583 31. L. Kong *et al.*, Structural basis of hepatitis C virus neutralization by broadly

584 neutralizing antibody HCV1. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 9499–9504

585 (2012).

586 32. P. F. Robbins *et al.*, Mining exomic sequencing data to identify mutated antigens

587 recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19**, 747–

588 752 (2013).

589 33. B. Diedenhofen, J. Musch, Cocor: A comprehensive solution for the statistical

590 comparison of correlations. *PLoS One*. **10** (2015),

591 doi:10.1371/journal.pone.0121945.

592 34. K. Chang *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat.*

593 *Genet.* **45**, 1113–1120 (2013).

594 35. L. Wojciech *et al.*, The same self-peptide selects conventional and regulatory

595 CD4+ T cells with identical antigen receptors. *Nat. Commun.* **5**, 5061 (2014).

596 36. R. Eil *et al.*, Ionic immune suppression within the tumour microenvironment

597 limits T cell effector function. *Nature*. **537**, 539–543 (2016).

598 37. C. H. Chang *et al.*, Metabolic Competition in the Tumor Microenvironment Is a
599 Driver of Cancer Progression. *Cell*. **162**, 1229–1241 (2015).
600 38. J. D. Wolchok *et al.*, Nivolumab plus ipilimumab in advanced melanoma. *N.*
601 *Engl. J. Med.* **369**, 122–33 (2013).
602 39. M. A. Postow *et al.*, Nivolumab and Ipilimumab versus Ipilimumab in Untreated
603 Melanoma. *N. Engl. J. Med.* **372**, 2006–2017 (2015).
604 40. H. Borghaei *et al.*, Nivolumab versus Docetaxel in Advanced Nonsquamous
605 Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **373**, 1627–39 (2015).
606 41. M. Yadav *et al.*, Predicting immunogenic tumour mutations by combining mass
607 spectrometry and exome sequencing. *Nature*. **515**, 572–6 (2014).
608 42. R Core Team, R: A Language and Environment for Statistical Computing
609 (2016), , doi:ISBN 3-900051-07-0.
610 43. M. Ogishi, H. Yotsuyanagi, K. Moriya, K. Koike, Delineation of autoantibody
611 repertoire through differential proteogenomics in hepatitis C virus-induced
612 cryoglobulinemia. *Sci. Rep.* **6**, 29532 (2016).
613 44. J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, R. A. Holt, Profiling the
614 T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome*
615 *Res.* **19**, 1817–24 (2009).

616 45. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in
617 single cells. *Nat. Methods*. **10**, 1096–1098 (2013).
618 46. M. Matz *et al.*, Amplification of cDNA ends based on template-switching effect
619 and step-out PCR. *Nucleic Acids Res.* **27**, 1558–1560 (1999).
620 47. M. Shugay *et al.*, Towards error-free profiling of immune repertoires. *Nat.*
621 *Methods*. **11**, 653–655 (2014).
622 48. S. Li *et al.*, IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype
623 diversity and next generation repertoire immunoprofiling. *Nat. Commun.* **4**, 2333
624 (2013).
625 49. M. Kuhn, Building Predictive Models in *R* Using the caret Package. *J. Stat. Softw.*
626 **28**, 1–26 (2008).
627 50. W. S. Noble, Support vector machine applications in computational biology.
628 *Kernel Methods Comput. Biol.*, 71–92 (2004).
629 51. T. C. Hsieh, K. H. Ma, A. Chao, iNEXT: an R package for rarefaction and
630 extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–
631 1456 (2016).
632 52. D. V. Bagaev *et al.*, VDJviz: a versatile browser for immunogenomics data. *BMC*
633 *Genomics*. **17**, 453 (2016).

Acknowledgments: We thank Dr. Takahashi of Leavanest, Inc. for technical assistance in MiSeq sequencing. We are also grateful to Dr. Couture-Cossette and Dr. Ueno for thoughtful advices. **Funding**: This work was supported by a Grant-in-Aid for Scientific Research (No. 30251234) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and by a Grant-in-Aid (No. HIV 28) from the Ministry of Health, Labor and Welfare, Japan. **Author contributions**: M.O. and H.Y. designed the study; M.O. performed computational analyses, prepared the figures and tables, and drafted the manuscript; M.O. and H.Y. wrote the manuscript. All authors critically discussed the results and their implications, and reviewed and approved the final version of the manuscript. **Competing interests**: The authors declare no competing financial interests. **Data and materials availability**: The TCR sequence dataset was deposited in the DDBJ Sequence Read Archive (DRA) under accession number DRA005827. The datasets used in this study, the trained immunogenicity prediction model, and key codes were bundled and publicly distributed as the R package *Repitope* on GitHub (https://github.com/masato-ogishi/Repitope/).

Figures

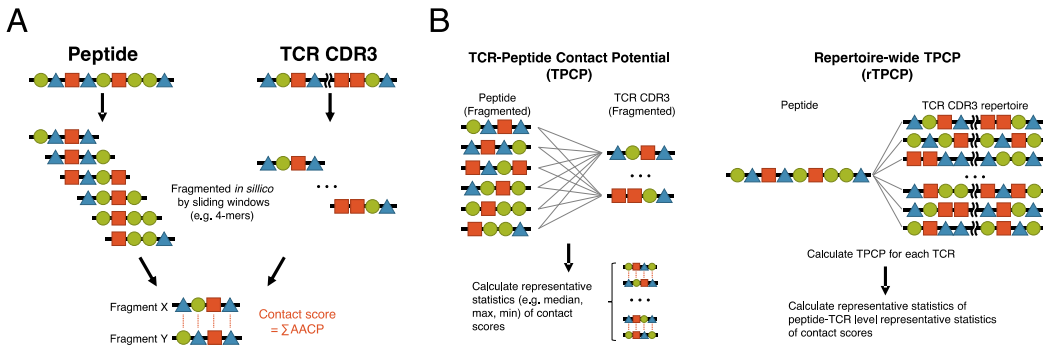


Figure 1. Schematic diagram of repertoire-wide TCR-peptide contact profile (rTPCP). Sequence-based modeling of TCR-peptide interactions is proposed. The interaction is restricted to a "window" of a fixed size, on the basis of the hypothesis that not all residues in the MHC-loaded peptide and TCR CDR3 are necessarily involved in the interactions. The energetic stability of the interactions is approximated as the summed amino acid pairwise contact potential (AACP). (A) TCR-peptide contact profile (TPCP) at a single TCR level. Both a peptide and a TCR CDR3 sequence are "windowed" in a sliding manner. Each fragment is paired, and the summed contact potential is calculated. TPCP is expressed as a set of representative statistics (e.g. median, maximum, minimum) of the set of inter-fragment contact potentials. (B) Repertoire-wide TPCP (rTPCP).

663 TPCPs were calculated against multiple TCR CDR3 sequences, and rTPCP is expressed

664 as a set of representative statistics of TPCPs.

665

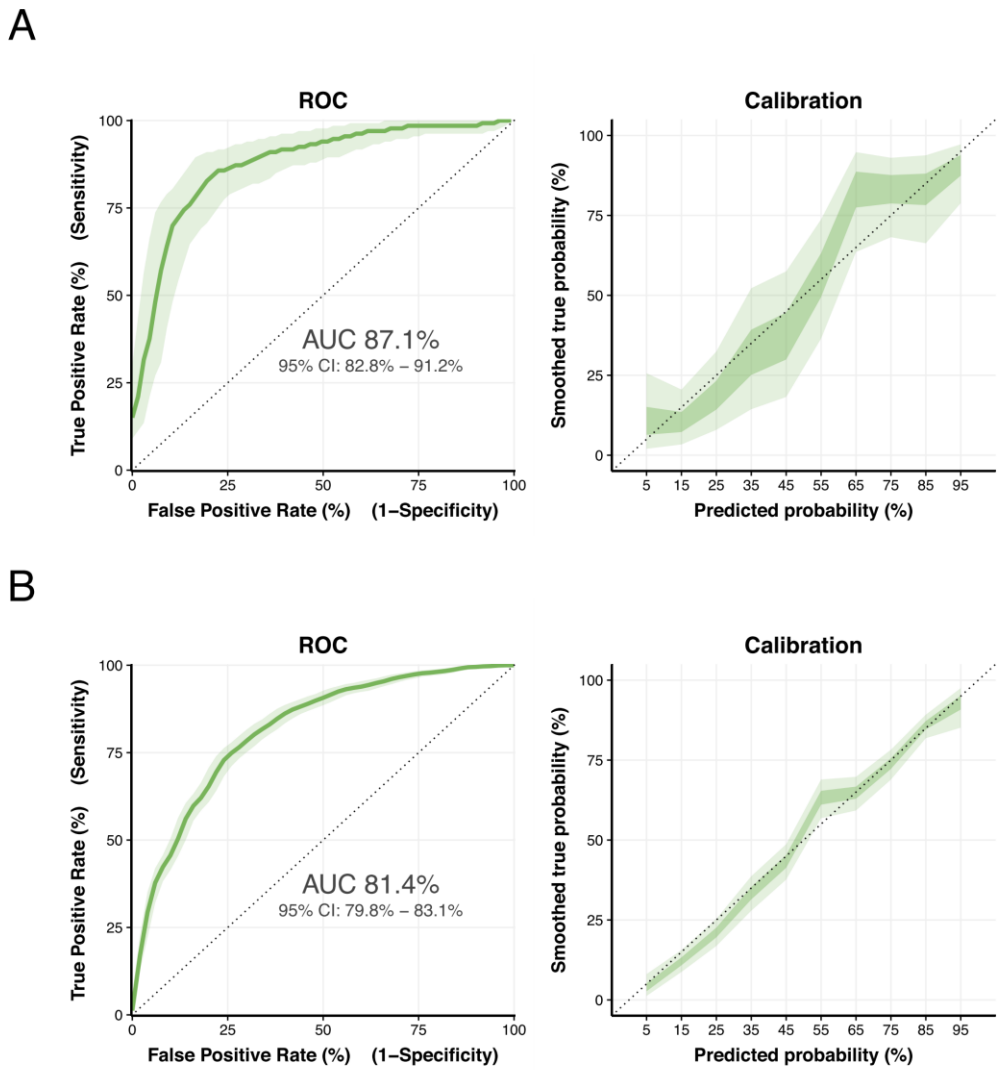
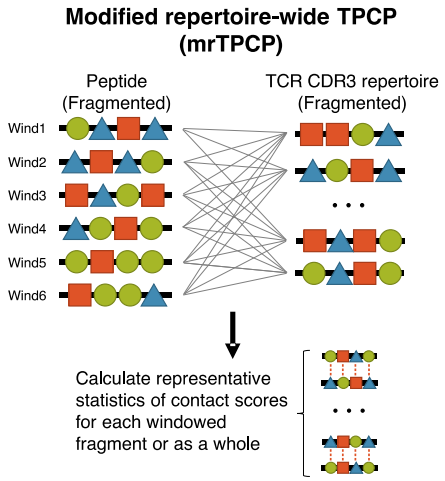


Figure 2. Immunogenicity prediction through rTPCP. (A) Representative ROC and calibration plots of the SVM classifier trained using rTPCP variables from all AACP scales. A total of 900 HLA-A2-restricted peptides (450 epitopes and 450 MHCLs) were randomly selected from the Chowell dataset, and split into training and testing subdatasets. The performance in the hold-out testing subdataset is shown. The AACP scales used are listed in table S1. (B) Representative ROC and calibration plots of the

SVM classifier trained using rTPCP variables from AAIndex MIYS990106. A full set of peptides in the Chowell dataset were used for model training and validation. Graphics were generated using the *classifierplots* package in R.



676 **Figure 3. Schematic diagram of modified rTPCP (mrTPCP).** All TCR sequences in

677 the reference repertoire are fragmented according to the fixed window size. A

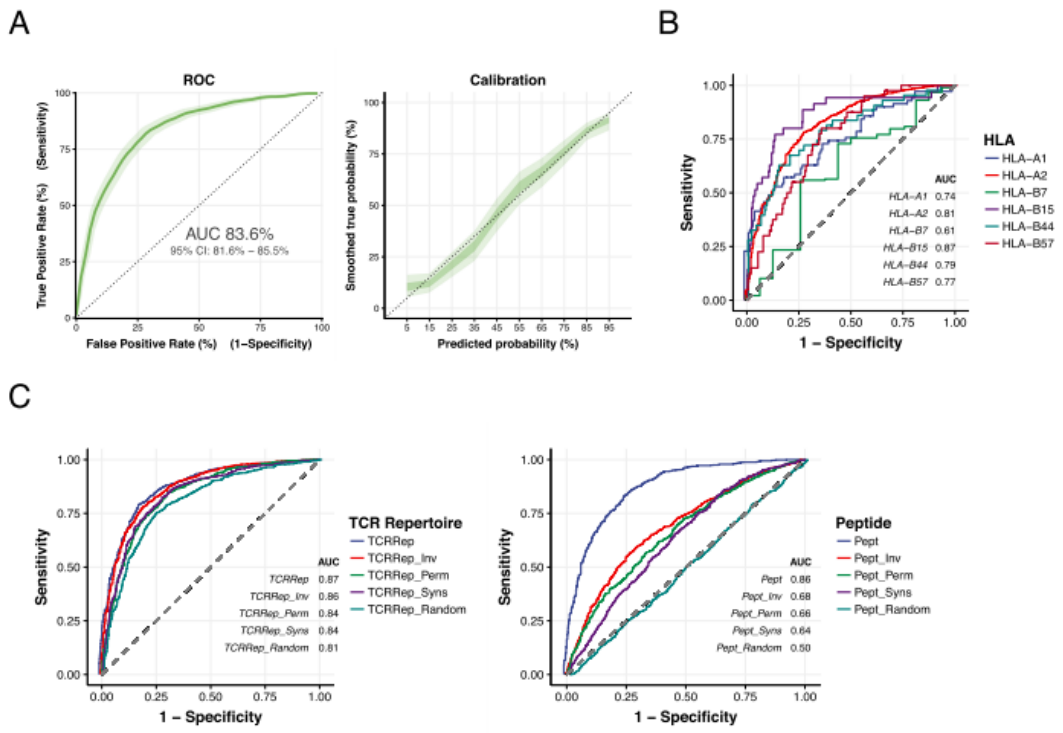
678 position-specific peptide-derived fragment was matched against a set of TCR-derived

679 fragments. Representative statistics were calculated both in a position-specific and

680 position-blind (i.e., pooled) manner. Owing to the position-dependent nature of the

681 analysis, only nonapeptides (=9-mers) were considered in the subsequent analysis.

682



683 **Figure 4. Improved immunogenicity prediction using mrTPCP.** (A) ROC and

684 calibration plots of the SVM classifier trained using mrTPCP variables derived from

685 MIYS990106. See the legend of Fig. 2B and method sections for further details. (B)

686 HLA-stratified ROC analysis. The entire Chowell dataset was sorted according to their

687 HLA restriction, and six most data-rich HLA supertypes were selected for visualization.

688 (C) Sequence manipulation analysis. Either the input peptide sequences or the reference

689 TCR repertoire sequences were manipulated, and mrTPCP variables were calculated. The

690 authentically trained SVM classifier was applied. Inv, inversion of the sequence; Perm,

691 permutation of the sequence; Syns, randomly synthesized sequences with relative amino

692 acid frequencies retained. For peptides, amino acid frequencies of immunogenic and

693 non-immunogenic peptides were separately considered; Random, completely random

694 sequences. (B-C) AUC was calculated and graphics were generated using *pROC* and

695 *plotROC* packages in R, respectively.

696

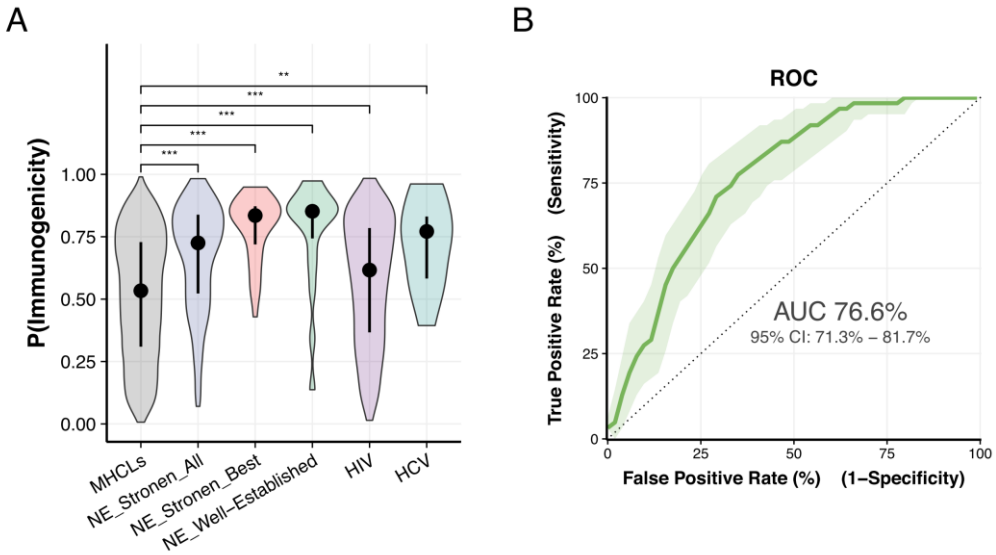


Figure 5. Immunogenicity prediction of independent datasets of viral epitopes and tumor neoepitopes. Peptide data were collected from various sources. Any peptide overlapping with those in the Chowell dataset was excluded. The probability of immunogenicity was estimated by applying the mrTPCP-based SVM classifier (Fig. 4A). (A) A metadataset of viral epitopes and tumor neoepitopes(*4*, *10*, *24*, *29–32*). Dots and bars represent the median and interquartile range, respectively. ***: $p < 0.001$, **: $p < 0.01$. *P* values were determined using Wilcoxon's rank sum test. NE, neoepitopes. (B) Epitope/ligand data originally reported by Calis et al.(*21*). The probability threshold was set to be 0.80.

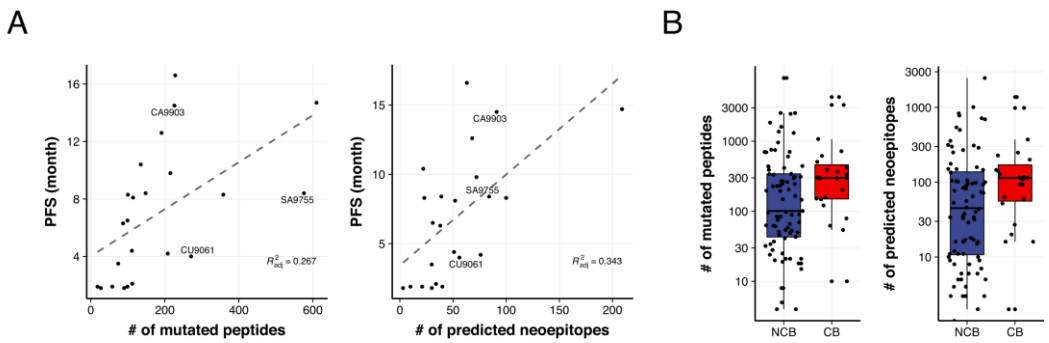
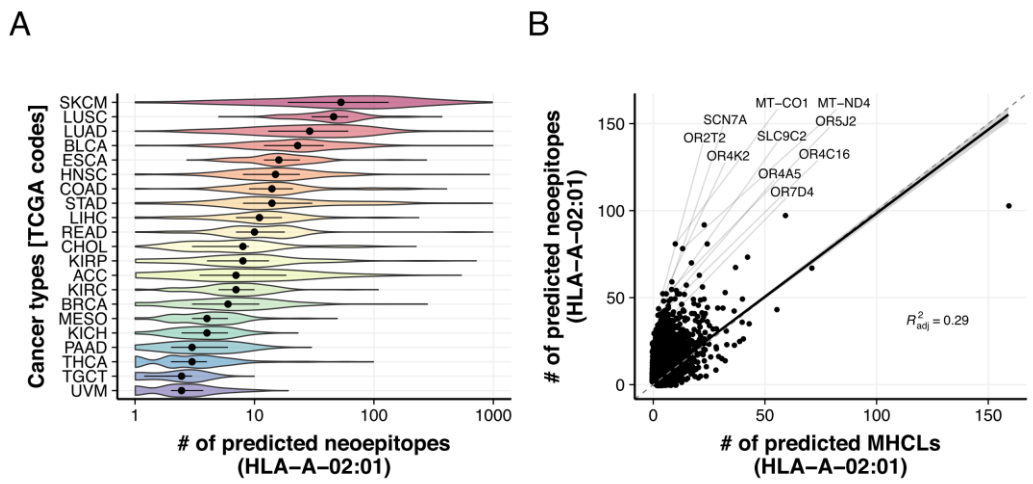


Figure 6. Correlation between predicted neoepitope burden and clinical outcome in checkpoint inhibitor trials. The mrTPCP-based SVM classifier (Fig. 4A) was applied to external datasets of tumor mutational landscapes obtained from checkpoint inhibitor trials. We set the threshold of immunogenicity to be above 0.80, on the basis of the observation that most of the well-established tumor neoepitopes exhibited probabilities of higher than 0.80 (Fig. 5A) (A) Progression-free survival (PFS) correlated with mutational burden/predicted neoepitope burden in non-small cell lung carcinoma (NSCLC) patients treated with pembrolizumab(*3*). The three patients were labeled, for which scaled fitting residuals decreased by more than 1 when the predicted neoepitope burden was used as a correlate. Adjusted correlation coefficient ($R^2{}_{adj}$) was calculated using the *stat_poly_eq* package.(B) Clinical benefit (CB) was associated with heavier mutational burden/predicted neoepitope burden in melanoma patients treated with ipilimumab(*5*). CB was defined as in the original paper(*5*). NCB, no clinical benefit.



721 **Figure 7. Neoepitope burdens in TCGA datasets.** Mutation data were retrieved from

722 all advanced stage tumors in The Cancer Genome Atlas (TCGA)(*34*). For the 105959

723 HLA-A-02:01-restricted nonapeptides predicted to be stable MHC binders,

724 immunogenicity prediction was carried out. (A) Predicted neoepitope burden was

725 visualized for the 21 tumor types registered in TCGA. SKCM, Skin Cutaneous

726 Melanoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma. For a

727 complete set of abbreviation used in TCGA, visit the NCI Genomic Data Commons

728 (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations).

729 (B) The correlation between predicted MHCL burden and neoepitope burden per gene.

730 HUGO symbols were depicted for genes enriched with neoepitopes. Enrichment was

731 defined as fitting residuals being larger than 10 for the purpose of tidy visualization.

732 Adjusted correlation coefficient was calculated using the *stat_poly_eq* package.

733

Tables

Table 1. Prediction results on datasets independent from training/validation data. Immunogenicity was predicted using the mrTPCP-based SVM classifier (Fig. 4A). Note that any peptide contained in the Chowell dataset was excluded. Unadjusted *p* values were calculated against the IEDB MHCL data as the negative control using the *prop.test* function implemented in R with continuity correction. N.A., not applicable.

Dataset	**Predicted Epitope**	***P* value**	**Data Source**
HCV CTL epitopes [n = 11]	10 (90.9%)	< 0.01	HCV immunology database (https://hcv.lanl.gov/content/immuno/tables/ctl_summary.html)
HIV CTL epitopes [n = 867]	545 (62.9%)	< 0.001	HIV molecular immunology database (https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html)
Well-established tumor neoepitopes [n = 20]	18 (90.0%)	< 0.001	(*4*, *10*, *29–32*)
Tumor neoepitopes (Best) [n = 35]	34 (97.1%)	< 0.001	(*24*)
Tumor neoepitopes (All) [n = 492]	382 (77.6%)	< 0.001	(*24*)
IEDB MHC-I ligands [10,000 randomly selected]	5412 (54.1%)	N.A.	IEDB (http://www.iedb.org/)

Supplementary Materials

Supplementary Materials and Methods

Quantitative parametrization of repertoire-wide TCR-peptide contact profiles

Contact potential scales. The AAIndex database is a curated source of numerical scales representing various biophysicochemical properties of amino acid pairwise contact potentials (AACPs)(*25*). Thirty-five AACP scales were retrieved from the database (http://www.genome.jp/aaindex/AAindex/list_of_potentials) on the basis of the relevance to amino acid residue contact energy (table S1).

Inter-peptide contact potential. Consider peptide X and peptide Y. The length of the peptides X and Y is designated as l. The inter-peptide contact potential between X and Y, or $cp(X,Y)$, is defined as

$$cp(X,Y) = \sum_{i=1}^{l} aacp(\mathrm{R}_{X,i}, \mathrm{R}_{Y,i})$$

where $\mathrm{R}_{X,i}$ and $\mathrm{R}_{Y,i}$ are the residues at position i $(1 \leq i \leq l)$, and $aacp(\mathrm{x},\mathrm{y})$ is a function of residue x and y that returns the AACP of the pair of x and y.

TCR-peptide contact profile (TPCP). For simplicity, only the interactions between the TCR β chain CDR3 region and the MHC-loaded peptide were considered. Suppose X

denotes an MHC-loaded peptide of length l_X, and Y denotes a CDR3 sequence of length l_Y. The window size w, representing the interacting interface between the peptide and CDR3, is assumed to be either 4 or 5. Owing to the loop structure of TCR CDR3, not every residue is likely involved in the interaction with the MHC-loaded peptide(*1*). In this setting, the sets of fragments F_X and F_Y, and the set of contact potentials of fragment pairs $CP_{X,Y}$ will be expressed as

$$F_X = \{ substring(X, i, i + w - 1) \mid 1 \leq i \leq l_X - w + 1 \}$$

$$F_Y = \{ substring(Y, i, i + w - 1) \mid 1 \leq i \leq l_Y - w + 1 \}$$

$$CP_{X,Y} = \{ cp(f_X, f_Y) \mid f_X \in F_X, f_Y \in F_Y \}$$

where $substring(s, i, j)$ is a function that returns the substring of sequence s from position i to position j, and f denotes a fragment belonging to the specified set of fragments F. TCR-peptide contact profile (TPCP) is defined as a set of representative statistics calculated from the set $CP_{X,Y}$. Specifically, the following statistics were considered: mean, median, maximum (max), minimum (min), second max, second min, the difference of max and second max (delta max), the difference of min and second min (delta min), and the difference of max and min (range).

Repertoire-wide TCR-peptide contact profile (rTPCP). Consider a set of TCR CDR3

sequences R. The set of TPCP statistics $TPCP_{X,R}$ is defined as

$$TPCP_{X,R} = \{ tpcp(X, Y_R) | Y_R \in R \}$$

where $tpcp(x, y)$ is a function that returns TPCP statistics of a peptide x and a TCR CDR3 y. Repertoire-wide TPCP (rTPCP) is defined as a set of representative statistics calculated from the set $TPCP_{X,R}$. Specifically, the following statistics were considered: mean, median, maximum, and minimum.

Modified repertoire-wide TCR-peptide contact profile (mrTPCP). Hereafter, only nonapeptides were utilized because of the position-specific nature of the analysis. Consider the peptide X and the TCR CDR3 repertoire R. Both are fragmented with a fixed window size w, as described above. The sets of fragments F_X and F_R, and the set of contact potentials $CP_{X,R}$ will be expressed as

$$F_X = \{ substring(X, i, i + w - 1) \mid 1 \leq i \leq 10 - w \}$$

$$F_R = \{ substring(Y, i, i + w - 1) \mid 1 \leq i \leq l_Y - w + 1, Y \in R \}$$

$$CP_{X,R} = \{ cp(f_X, f_R) \mid f_X \in F_X, f_R \in F_R \}$$

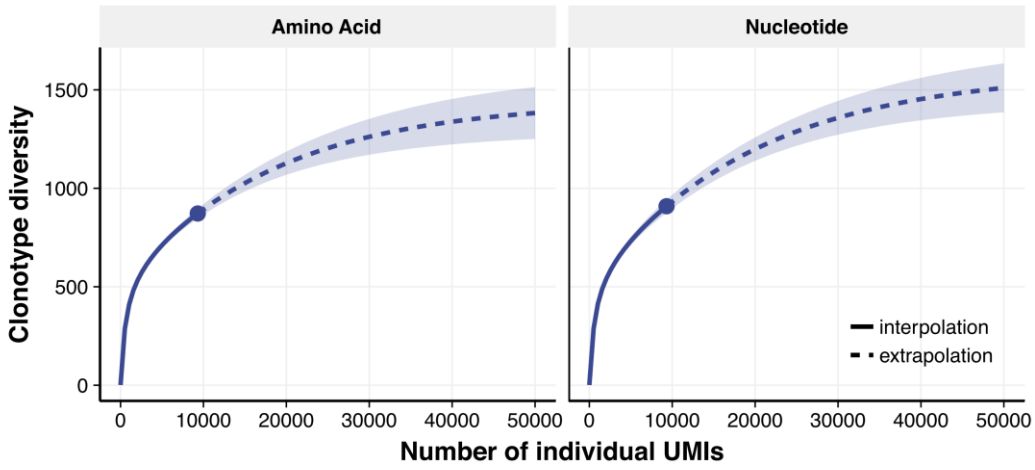
Two types of summarization strategy were adopted: “by-TCR-fragment” and “by-window”.

I. By-TCR-fragment: First, for each TCR fragment $f \in F_R$, $TPCP_{X,f}$ was calculated. Three representative statistics, namely, median, minimum, and maximum, were adopted to summarize the intra-peptide distribution of contact potentials. Then, representative statistics were calculated against $TPCP_X = \{TPCP_{X,f} \mid f \in F_R\}$. To more precisely capture the characteristics of the contact potential distribution, deciles, minimum, and maximum were calculated.

II. By-window: Representative statistics were calculated against a window-specific set of contact potentials $CP_{X,R,i} = \{ cp(substring(X, i, i + w - 1), f_R) \mid f_R \in F_R\}$, where i is an integer of $1 \leq i \leq 10 - w$. As representative statistics, deciles, minimum, and maximum were adopted.

Modified repertoire-wide TCR-peptide contact profile (mrTPCP) is defined as a set of "by-TCR-fragment" and "by-window" representative statistics.

808 **Supplementary Figures**



809 Fig. S1. Clonotype rarefaction analysis of reference TCR repertoire. TCR-Vβ sequences

810 were de-multiplexed using MiGEC(*47*) on the basis of unique molecular indexing

811 (UMI). De-multiplexed sequences were analyzed using IMGT/HighV-QUEST(*48*)

812 (https://www.imgt.org/HighV-QUEST/). Results were parsed into clonotype count tables.

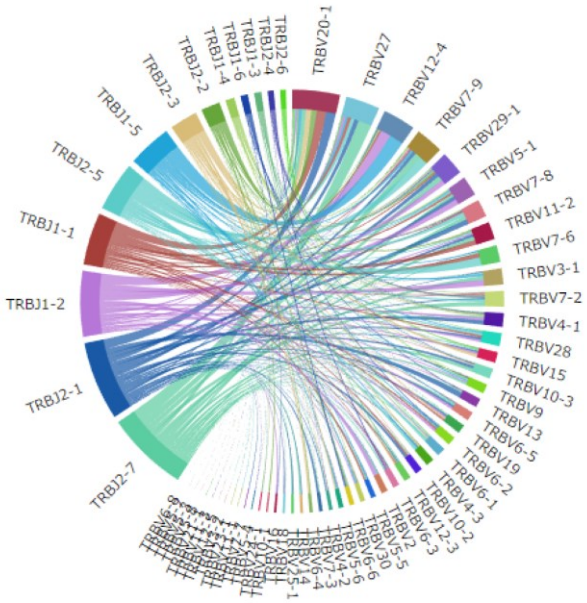
813 Clonotype is defined by the equity of the CDR3 sequence (either nucleotide or amino

814 acid). Rarefaction analysis was conducted using the *iNEXT* package(*51*).

815

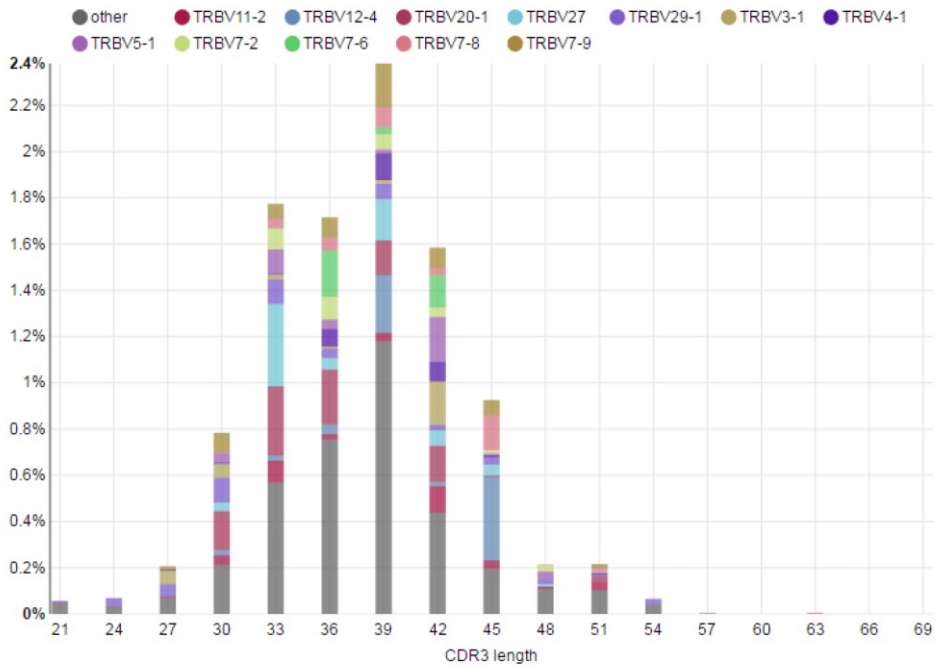
A

CD8+ V-J Usage



B

CD8+ V Spectratype



816 Fig. S2. Composition of the reference TCR repertoire. The Variable (V) and Junction (J)

817 segment usage in the CDR3 region was analyzed using VDJviz web tool(*52*)

818 (https://vdjviz.cdr3.net/). (A) VJ segment usage analysis. (B) V spectratype analysis.

819

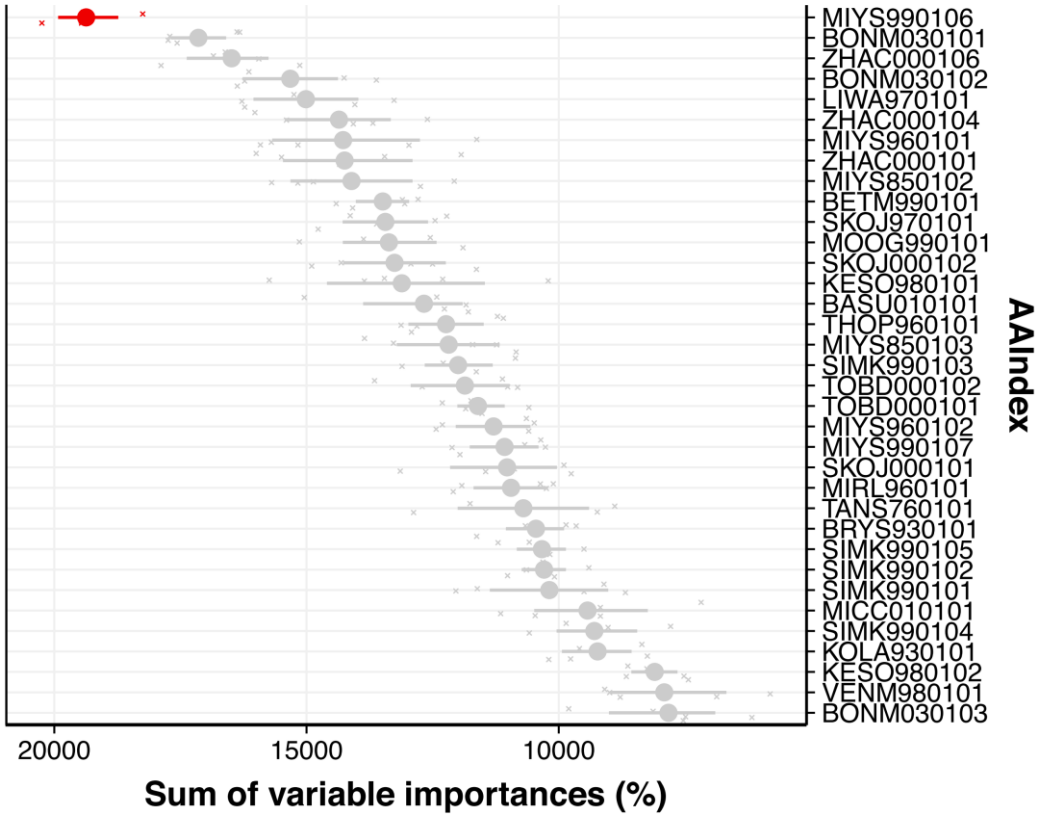


Fig. S3. Exploration of AACP scales most important in the immunogenicity prediction models trained using rTPCP variables. HLA-A2 restricted peptides were selected from the Chowell dataset. Four support vector machine (SVM) algorithms implemented in the *caret* package(*49*) were utilized to train immunogenicity prediction models based on the 2520 rTPCP variables from a total of 35 AAIndex AACP scales calculated against 2121 peptides. Classifier training was iterated using five different random seeds. Variable importance was estimated against a total of 20 classifiers (five iterations for each of the four algorithms) trained using the *varImp* function implemented in *caret*. The importance values were grouped according to their deriving AAIndex, summed, and visualized.

829

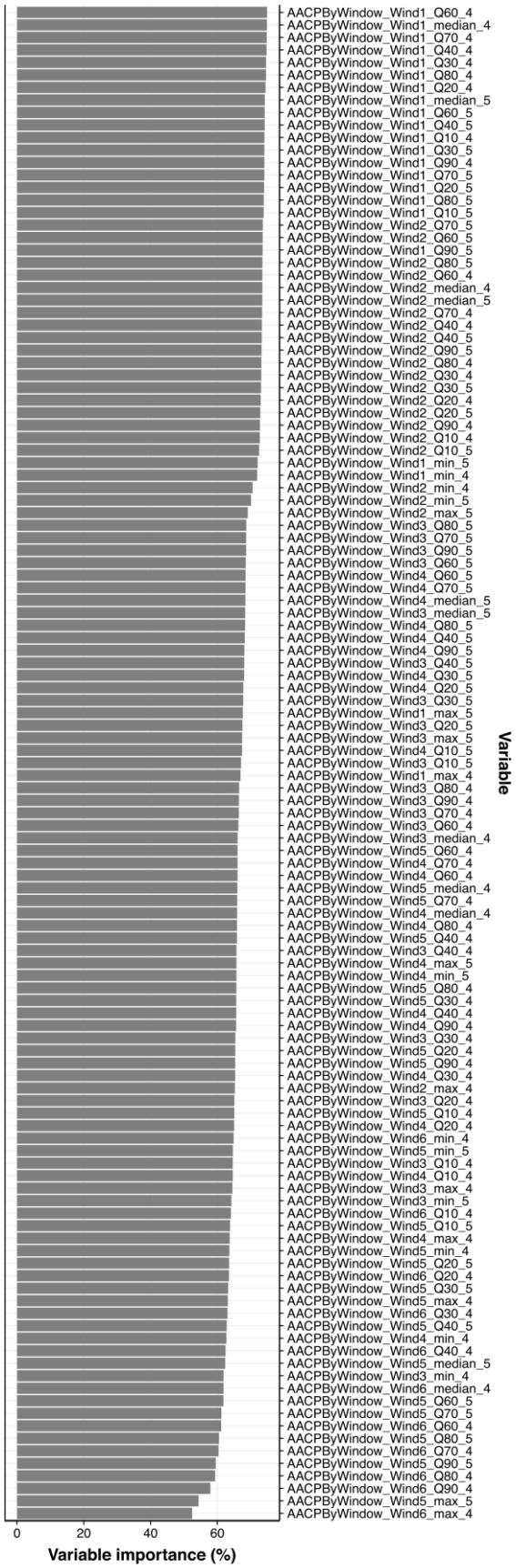

AACPByWindow_Wind1_Q60_4
AACPByWindow_Wind1_median_4
AACPByWindow_Wind1_Q70_4
AACPByWindow_Wind1_Q40_4
AACPByWindow_Wind1_Q30_4
AACPByWindow_Wind1_Q80_4
AACPByWindow_Wind1_Q20_4
AACPByWindow_Wind1_median_5
AACPByWindow_Wind1_Q60_5
AACPByWindow_Wind1_Q40_5
AACPByWindow_Wind1_Q10_4
AACPByWindow_Wind1_Q30_5
AACPByWindow_Wind1_Q90_4
AACPByWindow_Wind1_Q70_5
AACPByWindow_Wind1_Q20_5
AACPByWindow_Wind1_Q80_5
AACPByWindow_Wind1_Q10_5
AACPByWindow_Wind2_Q70_5
AACPByWindow_Wind2_Q60_5
AACPByWindow_Wind1_Q90_5
AACPByWindow_Wind2_Q80_5
AACPByWindow_Wind2_Q60_4
AACPByWindow_Wind2_median_4
AACPByWindow_Wind2_median_5
AACPByWindow_Wind2_Q70_4
AACPByWindow_Wind2_Q40_4
AACPByWindow_Wind2_Q40_5
AACPByWindow_Wind2_Q90_5
AACPByWindow_Wind2_Q80_4
AACPByWindow_Wind2_Q30_4
AACPByWindow_Wind2_Q30_5
AACPByWindow_Wind2_Q20_4
AACPByWindow_Wind2_Q20_5
AACPByWindow_Wind2_Q90_4
AACPByWindow_Wind2_Q10_4
AACPByWindow_Wind2_Q10_5
AACPByWindow_Wind1_min_5
AACPByWindow_Wind1_min_4
AACPByWindow_Wind2_min_4
AACPByWindow_Wind2_min_5
AACPByWindow_Wind2_max_5
AACPByWindow_Wind3_Q80_5
AACPByWindow_Wind3_Q70_5
AACPByWindow_Wind3_Q90_5
AACPByWindow_Wind3_Q60_5
AACPByWindow_Wind4_Q60_5
AACPByWindow_Wind4_Q70_5
AACPByWindow_Wind4_median_5
AACPByWindow_Wind3_median_5
AACPByWindow_Wind4_Q80_5
AACPByWindow_Wind4_Q40_5
AACPByWindow_Wind4_Q90_5
AACPByWindow_Wind3_Q40_5
AACPByWindow_Wind4_Q30_5
AACPByWindow_Wind4_Q20_5
AACPByWindow_Wind3_Q30_5
AACPByWindow_Wind1_max_5
AACPByWindow_Wind3_Q20_5
AACPByWindow_Wind3_max_5
AACPByWindow_Wind4_Q10_5
AACPByWindow_Wind3_Q10_5
AACPByWindow_Wind1_max_4
AACPByWindow_Wind3_Q80_4
AACPByWindow_Wind3_Q90_4
AACPByWindow_Wind3_Q70_4
AACPByWindow_Wind3_Q60_4
AACPByWindow_Wind3_median_4
AACPByWindow_Wind5_Q60_4
AACPByWindow_Wind4_Q70_4
AACPByWindow_Wind4_Q60_4
AACPByWindow_Wind5_median_4
AACPByWindow_Wind5_Q70_4
AACPByWindow_Wind4_median_4
AACPByWindow_Wind4_Q80_4
AACPByWindow_Wind5_Q40_4
AACPByWindow_Wind3_Q40_4
AACPByWindow_Wind4_max_5
AACPByWindow_Wind4_min_5
AACPByWindow_Wind5_Q80_4
AACPByWindow_Wind5_Q30_4
AACPByWindow_Wind4_Q40_4
AACPByWindow_Wind4_Q90_4
AACPByWindow_Wind3_Q30_4
AACPByWindow_Wind5_Q20_4
AACPByWindow_Wind5_Q90_4
AACPByWindow_Wind4_Q30_4
AACPByWindow_Wind2_max_4
AACPByWindow_Wind3_Q20_4
AACPByWindow_Wind5_Q10_4
AACPByWindow_Wind4_Q20_4
AACPByWindow_Wind6_min_4
AACPByWindow_Wind5_min_5
AACPByWindow_Wind3_Q10_4
AACPByWindow_Wind4_Q10_4
AACPByWindow_Wind3_max_4
AACPByWindow_Wind3_min_5
AACPByWindow_Wind6_Q10_4
AACPByWindow_Wind5_Q10_5
AACPByWindow_Wind4_max_4
AACPByWindow_Wind5_min_4
AACPByWindow_Wind5_Q20_5
AACPByWindow_Wind6_Q20_4
AACPByWindow_Wind5_Q30_5
AACPByWindow_Wind5_max_4
AACPByWindow_Wind6_Q30_4
AACPByWindow_Wind5_Q40_5
AACPByWindow_Wind4_min_4
AACPByWindow_Wind6_Q40_4
AACPByWindow_Wind5_median_5
AACPByWindow_Wind3_min_4
AACPByWindow_Wind6_median_4
AACPByWindow_Wind5_Q60_5
AACPByWindow_Wind5_Q70_5
AACPByWindow_Wind6_Q60_4
AACPByWindow_Wind5_Q80_5
AACPByWindow_Wind6_Q70_4
AACPByWindow_Wind5_Q90_5
AACPByWindow_Wind6_Q80_4
AACPByWindow_Wind6_Q90_4
AACPByWindow_Wind5_max_5
AACPByWindow_Wind6_max_4
Variable
0
20
40
60
Variable importance (%)

Fig. S4. Position-specific variable importance analysis of the immunogenicity prediction model trained using mrTPCP variables. All human immunogenicity data in the Chowell dataset were utilized. Variable importance was estimated using the *filterVarImp* function implemented in *caret* using the 187 mrTPCP variables derived from the AAIndex AACP scale MIYS990106. Window-specific variable importance values were selected and visualized.

Supplementary Tables

Table S1. AAIndex AACP scales used in this study.

AAIndexID	Description
TANS760101	Statistical contact potential derived from 25 x-ray protein structures
BRYS930101	Distance-dependent statistical potential (only energies of contacts within 0-5 Angstrooms are included)
THOP960101	Mixed quasichemical and optimization-based protein contact potential
MIRL960101	Statistical potential derived by the maximization of the harmonic mean of Z scores
VENM980101	Statistical potential derived by the maximization of the perceptron criterion
BASU010101	Optimization-based potential derived by the modified perceptron criterion
MIYS850102	Quasichemical energy of transfer of amino acids from water to the protein environment
MIYS850103	Quasichemical energy of interactions in an average buried environment
MIYS960101	Quasichemical energy of transfer of amino acids from water to the protein environment
MIYS960102	Quasichemical energy of interactions in an average buried environment
MIYS990106	Quasichemical energy of transfer of amino acids from water to the protein environment
MIYS990107	Quasichemical energy of interactions in an average buried environment
LIWA970101	Modified version of the Miyazawa-Jernigan transfer energy
KESO980101	Quasichemical transfer energy derived from interfacial regions of protein-protein complexes

KESO980102	Quasichemical energy in an average protein environment derived from interfacial regions of protein-protein complexes
MOOG990101	Quasichemical potential derived from interfacial regions of protein-protein complexes
BETM990101	Modified version of the Miyazawa-Jernigan transfer energy
TOBD000101	Optimization-derived potential obtained for small set of decoys
TOBD000102	Optimization-derived potential obtained for large set of decoys
KOLA930101	Statistical potential derived by the quasichemical approximation
SKOJ970101	Statistical potential derived by the quasichemical approximation
SKOJ000101	Statistical quasichemical potential with the partially composition-corrected pair scale
SKOJ000102	Statistical quasichemical potential with the composition-corrected pair scale
BONM030101	Quasichemical statistical potential for the antiparallel orientation of interacting side groups
BONM030102	Quasichemical statistical potential for the intermediate orientation of interacting side groups
BONM030103	Quasichemical statistical potential for the parallel orientation of interacting side groups
MICC010101	Optimization-derived potential
SIMK990101	Distance-dependent statistical potential (contacts within 0-5 Angstrooms)
SIMK990102	Distance-dependent statistical potential (contacts within 5-7.5 Angstrooms)
SIMK990103	Distance-dependent statistical potential (contacts within 7.5-10 Angstrooms)
SIMK990104	Distance-dependent statistical potential (contacts within 10-12 Angstrooms)

SIMK990105	Distance-dependent statistical potential (contacts longer than 12 Angstrooms)
ZHAC000101	Environment-dependent residue contact energies (rows = helix, cols = helix)
ZHAC000104	Environment-dependent residue contact energies (rows = strand, cols = strand)
ZHAC000106	Environment-dependent residue contact energies (rows = coil, cols = coil)

Table S2. Performance evaluation of SVM classifiers with rTPCP variables from all AAIndex AACP scales.

Algorithm	**Sensitivity**	**Specificity**	**Accuracy [95% CI]**	**Seed**
SVM-Linear	70.70%	78.20%	74.4% [68.8%-79.6%]	12345
SVM-Poly	75.90%	79.70%	77.8% [72.3%-82.7%]	12345
SVM-Radial	75.90%	74.40%	75.2% [69.5%-80.3%]	12345
SVM-RadialSigma	72.90%	75.90%	74.4% [68.8%-79.6%]	12345
SVM-Linear	69.90%	78.90%	74.4% [68.8%-79.6%]	23451
SVM-Poly	69.90%	82.70%	76.3% [70.7%-81.3%]	23451
SVM-Radial	68.40%	77.40%	72.9% [67.2%-78.2%]	23451
SVM-RadialSigma	73.70%	76.70%	75.2% [69.5%-80.3%]	23451
SVM-Linear	82.70%	74.40%	78.6% [73.1%-83.3%]	34512
SVM-Poly	81.20%	81.20%	81.2% [76.0%-85.7%]	34512
SVM-Radial	78.20%	82.70%	80.5% [75.2%-85.0%]	34512
SVM-RadialSigma	78.90%	82.00%	80.5% [75.2%-85.0%]	34512
SVM-Linear	70.70%	70.70%	70.7% [64.8%-76.1%]	45123
SVM-Poly	77.40%	71.40%	74.4% [68.8%-79.6%]	45123
SVM-Radial	70.70%	74.40%	72.6% [66.8%-77.8%]	45123
SVM-RadialSigma	72.20%	73.70%	72.9% [67.2%-78.2%]	45123
SVM-Linear	76.70%	68.40%	72.6% [66.8%-77.8%]	51234
SVM-Poly	75.90%	77.40%	76.7% [71.1%-81.6%]	51234
SVM-Radial	77.40%	69.90%	73.7% [68.0%-78.9%]	51234
SVM-RadialSigma	77.40%	70.70%	74.1% [68.4%-79.2%]	51234

Table S3. Performance evaluation of SVM classifiers with rTPCP variables from MIYS990106 AACP scale.

Algorithm	**Sensitivity**	**Specificity**	**Accuracy [95% CI]**	**Seed**
SVM-Poly	72.90%	71.80%	72.3% [70.6%-74.1%]	12345
SVM-Poly	73.10%	75.20%	74.2% [72.5%-75.9%]	23451
SVM-Poly	73.50%	75.60%	74.6% [72.9%-76.3%]	34512
SVM-Poly	72.20%	72.40%	72.3% [70.5%-74.0%]	45000
SVM-Poly	72.30%	76.30%	74.4% [72.6%-76.1%]	51234

849 Table S4. Oligonucleotides used in this study.

Name	Sequence	Description
SmartTSO‡	ACAGCAGGTCAGTCAAGCAGTATTTDHBVtDHBVtDHBVtDHBVTTTrGrG+G	Template-switching oligo
TCRSeq-RT	CACGTGGTCGGGGWAGAAGC	Reverse-transcription primer specific to TCR-Vβ region with adapter for 3'-RACE
TCRSeq-Fwd-HC	CGGATAACAATTTCACACAGGGCACAGCAGGTCAGTCAAGCAGTA	Heal-carrier oligo for step-out suppression PCR
TCRSeq-Fwd-HS	CGGATAACAATTTCACACAGGGC	Heal-specific forward primer for step-out suppression PCR
TCRSeq-Fwd-HS-N1†	NCGGATAACAATTTCACACAGGGC	
TCRSeq-Fwd-HS-N2†	NNCGGATAACAATTTCACACAGGGC	
TCRSeq-Fwd-HS-N3†	NNNCGGATAACAATTTCACACAGGGC	
TCRSeq-Rev1	CTCTGCTTCTGATGGCTCAAAC	Reverse primer for 1st round PCR
TCRSeq-Rev2	CGGGTGGGAACACSTTKTTCAGGT	Reverse primer for 2nd round PCR
TCRSeq-Rev2-N1†	NCGGGTGGGAACACSTTKTTCAGGT	
TCRSeq-Rev2-N2†	NNCGGGTGGGAACACSTTKTTCAGGT	
TCRSeq-Rev2-N3†	NNNCGGGTGGGAACACSTTKTTCAGGT	

850

851 †: Either N, NN, or NNN was attached to the primers in order to prevent signal saturation

852 through the base calling process in MiSeq, as previously described(*43*).

853 ‡: 3’ terminus consists of guanosine (rG) and guanidine locked nucleotide acid (+G), as

854 previously described(*45*).

855

Supplementary Data Files

Data file S1. HCV CTL epitopes.

Data file S2. HIV CTL epitopes.

Data file S3. Well-established tumor neoepitopes from multiple publications.

Data file S4. Best-predicted tumor neoepitopes reported by Stronen et al.

Data file S5. All neoepitopes reported by Stronen et al.

Data file S6. Pathogen-derived epitopes reported by Calis et al.

Data file S7. Tumor neoepitopes and clinical data reported by Rizvi et al.

Data file S8. Tumor neoepitopes and clinical data reported by van Allen et al.

Data file S9. Tumor neoepitope candidates identified from TCGA datasets.