

1 **TCR-peptide contact profile determines immunogenicity in**  
2 **pathogen/tumor-derived MHC-I epitopes.**

3

4 **Authors:** Masato Ogishi<sup>1,2\*</sup> and Hiroshi Yotsuyanagi<sup>2</sup>

5 **Affiliations:**

6 <sup>1</sup> National Center for Global Health and Medicine Hospital, Tokyo, 162-8655, Japan

7 <sup>2</sup> Department of Infectious Diseases and Applied Immunology, Research Hospital, The  
8 Institute of Medical Science, The University of Tokyo, Tokyo, 108-0071, Japan

9

10 \* To whom correspondence should be addressed: Masato Ogishi

11 ([oogishi-tky@umin.ac.jp](mailto:oogishi-tky@umin.ac.jp))

12 **One Sentence Summary:** Accurate epitope prediction was achieved via machine  
13 learning by incorporating TCR-peptide contact profiles.

14 **Keywords:** Epitope prediction, T cell receptor repertoire, amino acid contact potential,  
15 machine learning

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

17

## 18 **Abstract**

19           Computational methodologies to predict epitopes for cytotoxic T lymphocytes  
20 (CTLs) will galvanize vaccine research and pave the way toward targeted  
21 immunotherapy of infections and cancer. However, the classification of immunogenic  
22 epitopes and non-immunogenic major histocompatibility complex (MHC) class I ligands  
23 *in silico* has yet to attain sufficient accuracy. Here, we demonstrated highly accurate  
24 epitope prediction by a machine learning-based classifier incorporating T cell receptor  
25 (TCR)-peptide contact profiles, with an accuracy of 0.77 and an area under the curve of  
26 0.84 in hold-out validation. Predictive accuracy was retained for five major human  
27 leucocyte antigen supertypes. Successful prediction using independent datasets of viral  
28 epitopes and tumor neoepitopes was demonstrated. Collectively, this is the first study  
29 demonstrating accurate and generalizable CTL immunogenicity prediction from the  
30 TCR-peptide axis. The R package *Repitope* was implemented to maximize code  
31 reusability. Prospective validation in vaccination and/or cancer immunotherapy cohorts is  
32 warranted.

33

## 34 **Introduction**

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

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

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

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

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

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

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

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

106 predictive accuracy. Successful prediction was demonstrated using independent  
107 epitope/MHCL datasets of viral and tumor origin. Moreover, using a mutational  
108 landscape dataset obtained from checkpoint inhibitor trials, a correlation between  
109 predicted neoepitope burden and clinical outcome was shown. Overall, this is the first  
110 study demonstrating a highly accurate and generalizable epitope prediction by  
111 integrating the TCR-peptide axis. The codes for rTPCP and mrTPCP analysis were  
112 compiled into the R package *Repitope* (<https://github.com/masato-ogishi/Repitope/>).  
113 Prospective validation of this tool in independent cohorts of vaccination and cancer  
114 immunotherapy is necessary to evaluate its possible clinical applicability.  
115

## 116 **Results**

### 117 **Preparation of pooled human TCR repertoire dataset**

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

### 128 **Immunogenicity prediction from repertoire-wide TCR-peptide contact profiles** 129 **parametrized using amino acid contact potentials**

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



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

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

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

### 165 **Improved immunogenicity prediction by incorporating position-specific contact** 166 **profiles**

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

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

183       The amino acid compositions of MHCLs are restricted by the HLA to which they  
184 are coupled. Since our mrTPCP framework is not dependent on HLA information, it  
185 might be useful for pan-specific immunogenicity prediction. To test this hypothesis,  
186 4738 unique nonapeptides in the Chowell dataset were stratified based on their

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

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

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

205 promising methodology for accurate epitope prediction.

## 206 **Immunogenicity prediction using independent datasets**

207 Any pattern learned from one dataset is not always extendable to other datasets  
208 constructed in different contexts. Therefore, we tested the performance of our  
209 immunogenicity prediction model by utilizing independent datasets adopted from  
210 previous publications(4, 10, 24, 29–32), after removing peptides overlapping with those  
211 in the Chowell dataset. As expected, randomly selected 10,000 MHCLs retrieved from  
212 the Immune Epitope Database (IEDB) were predicted as either immunogenic or  
213 non-immunogenic in an approximately 1:1 ratio, with a uniform distribution of  
214 predicted probabilities (Fig. 5A and Table 1). In contrast, epitope datasets of viral and  
215 tumor origin were significantly enriched with peptides predicted as epitopes ( $p < 0.01$  by  
216 Wilcoxon's rank sum test in comparison with randomly selected MHCLs from IEDB). It  
217 is notable that 16 out of 22 (73%) well-defined neoepitopes and 22 out of 35 (63%) best  
218 neoepitopes reported by Stronen et al. had probabilities of  $> 0.80$  (Fig. 5A). With the  
219 probability threshold of 0.80, our model also effectively classified the epitope/MHCL  
220 dataset from various pathogens originally reported by Calis et al.(21) (accuracy, 0.71;  
221 95% CI, 0.67 to 0.75; AUC, 0.77; 95% CI, 0.71 to 0.82) (Fig. 5B).

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

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

240 Finally, we compared estimated neoepitope burden across 21 tumor types in The

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

## 252 **R package implementation of immunogenicity prediction framework**

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

258

## 259 **Discussion**

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

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



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

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

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

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

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

331 infection.

332 Immune checkpoint inhibitors achieved revolutionary success in some types of

333 tumor including melanoma and non-small cell lung carcinoma (NSCLC)(38–40).

334 However, relatively few explanations have been proposed about the reason why these two

335 types of tumor are the most sensitive to checkpoint blockade. To address this, we

336 explored the entire TCGA cancer genome dataset(34). As anticipated, skin cutaneous

337 melanoma and NSCLC were most enriched with predicted MHCLs or neoepitopes (Fig.

338 7A). The ratio of predicted neoepitope to predicted MHCL significantly varied by

339 respective gene analyzed (Fig. 7B). Focusing on the high-yield genes such as

340 mitochondrial enzymes and olfactory receptors may expedite the development of

341 pan-cancer targeted immunotherapy.

342 Similarly to previous studies on immunogenicity prediction, this study has several

343 limitations. First, this is a retrospective observational study; no prospective identification

344 of novel epitopes is demonstrated. Thus, prospective validation is indispensable before

345 this model can be clinically applied. Second, the process of quantitative parametrization

346 of TCR-peptide interactions could further be optimized, as our window-based pairwise

347 interaction model might oversimplify the biophysicochemical nature of the

348 TCR-peptide-MHC interactions. In particular, the hypothesis that either a 4-mer or

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

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

367 datasets and key scripts are bundled. Disease-specific, prospective cohort studies could

368 be conducted to evaluate clinical usefulness in the future.

369

## 370 **Materials and Methods**

### 371 **Study design**

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

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

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

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

### 382 **Computational analysis**

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

## 388 **Preparation of pooled human TCR repertoire dataset**

389 TCR repertoire sequencing was carried out as previously described(43), except the  
390 primers being used. Briefly, total RNA from CD8<sup>+</sup> T cells collected from donated  
391 peripheral blood of German origin was purchased (Miltenyi Biotec) and used as the  
392 source of a pooled TCR repertoire. Primers for human TCR-V $\beta$  were adopted from  
393 previously published work(44). All primers were synthesized by Life Technologies, and  
394 a template-switching oligo (TSO) containing 5' terminal unique molecular identifier  
395 (UMI) and 3' terminal guanidine locked nucleic acid (LNA) was synthesized by  
396 Exiqon(45). The sequences of the oligonucleotides utilized in this study are summarized  
397 in table S4. Reverse transcription with template-switching and semi-nested step-out PCR  
398 were performed using SMARTScribe (Clontech) and KAPA2G Fast Multiplex PCR  
399 master mix (Kapa Biosystems)(46). Amplified PCR products were gel-excised, repaired,  
400 and re-purified. Paired-end libraries were prepared, and paired-end sequencing of 2 x 300  
401 bp was performed using MiSeq (Illumina). UMI-guided de-multiplexing was performed  
402 using MiGEC software in order to reduce the effect of PCR amplification bias(47). CDR3  
403 regions were identified using IMGT/HighV-QUEST(48)  
404 (<https://www.imgt.org/HighV-QUEST/>).

## 405 **Epitope/ligand dataset for training/testing immunogenicity prediction model**



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

#### 410   **Machine learning for immunogenicity prediction**

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

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

#### 431 **Epitope/ligand datasets for external validation**

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

442 Chowell dataset.

443 To independently assess the generalizability of the trained immunogenicity  
444 prediction model, epitope datasets were collected from the following sources: (i)  
445 hepatitis C virus (HCV) CTL epitopes  
446 ([https://hcv.lanl.gov/content/immuno/tables/ctl\\_summary.html](https://hcv.lanl.gov/content/immuno/tables/ctl_summary.html)), (ii) human  
447 immunodeficiency virus type-I (HIV) CTL epitopes  
448 ([https://www.hiv.lanl.gov/content/immunology/tables/ctl\\_summary.html](https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html)), (iii)  
449 well-established tumor neoepitopes from multiple publications(4, 10, 29–32), (iv)  
450 neoepitopes predicted to be the most stable MHC binders identified in the study by  
451 Stronen et al.(24), and (v) all neoepitopes identified in the study by Stronen et al. For  
452 comparison, we also obtained a human MHCL dataset from IEDB (<http://www.iedb.org/>).  
453 Note that the MHCL data lacks T cell assay annotation, and thus the true ratio of 'epitopes'  
454 to 'MHCLs' in the definitions discussed in this study is unknown. Moreover, we retrieved  
455 the epitope/MHCL dataset originally reported by Calis et al(21). This dataset is suitable  
456 for assessing the specificity of our immunogenicity prediction model, because it contains  
457 experimentally validated non-immunogenic MHCLs, mostly originated from dengue  
458 virus. Peptide sequences containing alphabetical characters other than those representing  
459 20 authentic amino acid residues were removed. Any peptide contained in the Chowell

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

#### 462 **Correlation with clinical outcomes in checkpoint inhibitor trials**

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

#### 468 **Neoepitope burden across TCGA tumor types**

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

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

478 **Statistical analysis**

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

484

## 485 **Supplementary Materials**

486 Supplementary Materials and Methods

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

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

489 Fig. S3. Exploration of AACP scales most important in the immunogenicity prediction

490 models trained using rTPCP variables.

491 Fig. S4. Variable importance analysis of the immunogenicity prediction model trained

492 using mrTPCP variables.

493 Table S1. AAIndex AACP scales used in this study.

494 Table S2. Performance evaluation of SVM classifiers with rTPCP variables from all

495 AAIndex AACP scales.

496 Table S3. Performance evaluation of SVM classifiers with rTPCP variables from

497 MIYS990106 AACP scale.

498 Table S4. Oligonucleotides used in this study.

499 Data file S1. HCV CTL epitopes.

500 Data file S2. HIV CTL epitopes.

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

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

- 503 Data file S5. All neoepitopes reported by Stronen et al.
- 504 Data file S6. Pathogen-derived epitopes reported by Calis et al.
- 505 Data file S7. Tumor neoepitopes and clinical data reported by Rizvi et al.
- 506 Data file S8. Tumor neoepitopes and clinical data reported by van Allen et al.
- 507 Data file S9. Tumor neoepitope candidates identified from TCGA datasets.
- 508

## 509 **References**

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



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

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

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

- 581 of melanoma neoantigen-specific T cells. *Science (80-. )*. **348**, 803–808 (2015).
- 582 30. E. Tran *et al.*, Immunogenicity of somatic mutations in human gastrointestinal  
583 cancers. *Science (80-. )*. **350**, 1387–1390 (2015).
- 584 31. L. Kong *et al.*, Structural basis of hepatitis C virus neutralization by broadly  
585 neutralizing antibody HCV1. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 9499–9504  
586 (2012).
- 587 32. P. F. Robbins *et al.*, Mining exomic sequencing data to identify mutated antigens  
588 recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19**, 747–  
589 752 (2013).
- 590 33. B. Diedenhofen, J. Musch, Cocor: A comprehensive solution for the statistical  
591 comparison of correlations. *PLoS One*. **10** (2015),  
592 doi:10.1371/journal.pone.0121945.
- 593 34. K. Chang *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat.*  
594 *Genet.* **45**, 1113–1120 (2013).
- 595 35. L. Wojciech *et al.*, The same self-peptide selects conventional and regulatory  
596 CD4<sup>+</sup> T cells with identical antigen receptors. *Nat. Commun.* **5**, 5061 (2014).
- 597 36. R. Eil *et al.*, Ionic immune suppression within the tumour microenvironment  
598 limits T cell effector function. *Nature*. **537**, 539–543 (2016).

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

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

635

636 **Acknowledgments:** We thank Dr. Takahashi of Leavanest, Inc. for technical assistance in

637 MiSeq sequencing. We are also grateful to Dr. Couture-Cossette and Dr. Ueno for

638 thoughtful advices. **Funding:** This work was supported by a Grant-in-Aid for Scientific

639 Research (No. 30251234) from the Ministry of Education, Culture, Sports, Science and

640 Technology, Japan, and by a Grant-in-Aid (No. HIV 28) from the Ministry of Health,

641 Labor and Welfare, Japan. **Author contributions:** M.O. and H.Y. designed the study;

642 M.O. performed computational analyses, prepared the figures and tables, and drafted the

643 manuscript; M.O. and H.Y. wrote the manuscript. All authors critically discussed the

644 results and their implications, and reviewed and approved the final version of the

645 manuscript. **Competing interests:** The authors declare no competing financial interests.

646 **Data and materials availability:** The TCR sequence dataset was deposited in the DDBJ

647 Sequence Read Archive (DRA) under accession number DRA005827. The datasets used

648 in this study, the trained immunogenicity prediction model, and key codes were bundled

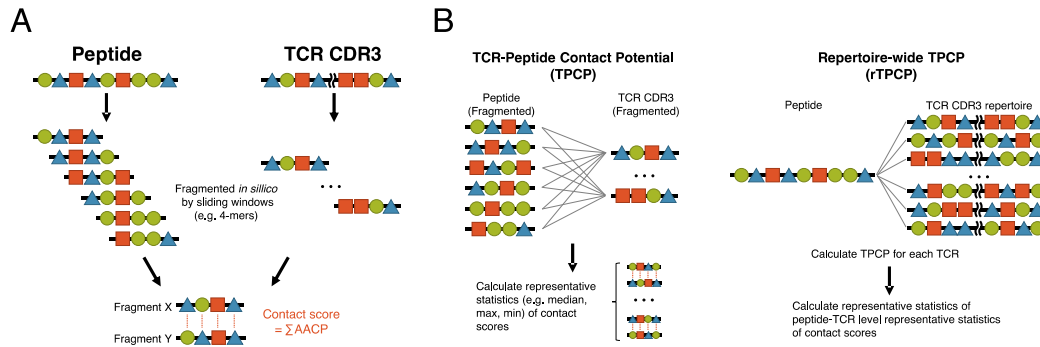
649 and publicly distributed as the R package *Repitope* on GitHub

650 (<https://github.com/masato-ogishi/Repitope/>).

651

## 652 Figures

653



654 **Figure 1. Schematic diagram of repertoire-wide TCR-peptide contact profile**

655 **(rTPCP).** Sequence-based modeling of TCR-peptide interactions is proposed. The

656 interaction is restricted to a "window" of a fixed size, on the basis of the hypothesis that

657 not all residues in the MHC-loaded peptide and TCR CDR3 are necessarily involved in

658 the interactions. The energetic stability of the interactions is approximated as the summed

659 amino acid pairwise contact potential (AACP). (A) TCR-peptide contact profile (TPCP)

660 at a single TCR level. Both a peptide and a TCR CDR3 sequence are "windowed" in a

661 sliding manner. Each fragment is paired, and the summed contact potential is calculated.

662 TPCP is expressed as a set of representative statistics (e.g. median, maximum, minimum)

663 of the set of inter-fragment contact potentials. (B) Repertoire-wide TPCP (rTPCP).

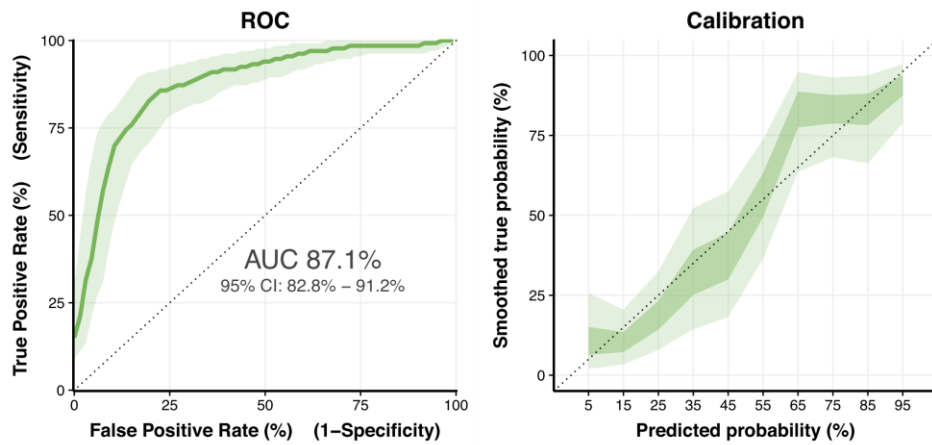


664 TPCPs were calculated against multiple TCR CDR3 sequences, and rTPCP is expressed

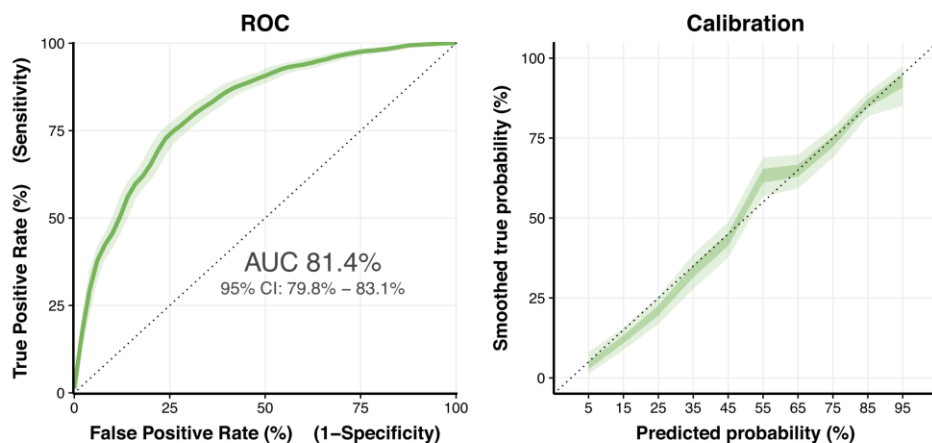
665 as a set of representative statistics of TPCPs.

666

A

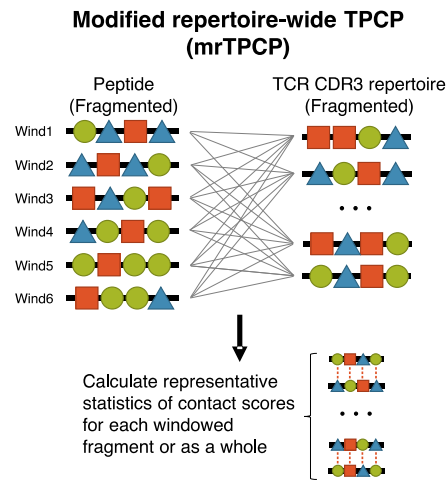


B

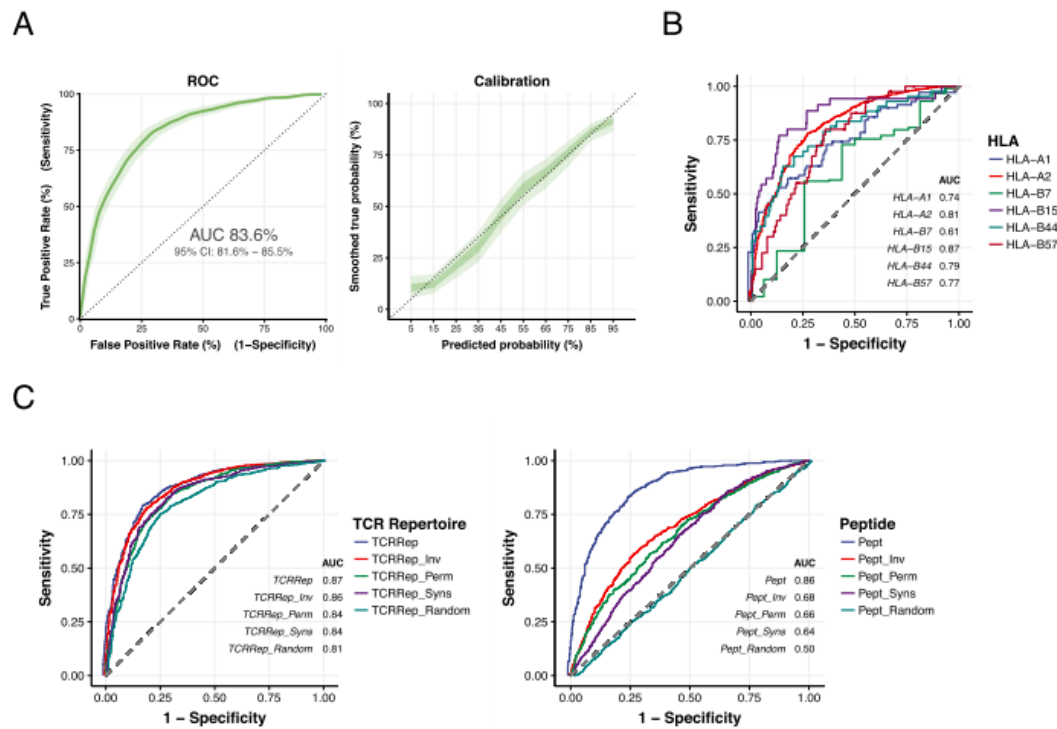


667 **Figure 2. Immunogenicity prediction through rTPCP.** (A) Representative ROC and  
668 calibration plots of the SVM classifier trained using rTPCP variables from all AACP  
669 scales. A total of 900 HLA-A2-restricted peptides (450 epitopes and 450 MHCLs) were  
670 randomly selected from the Chowell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

673 SVM classifier trained using rTPCP variables from AAIndex MIYS990106. A full set of  
674 peptides in the Chowell dataset were used for model training and validation. Graphics  
675 were generated using the *classifierplots* package in R.  
676

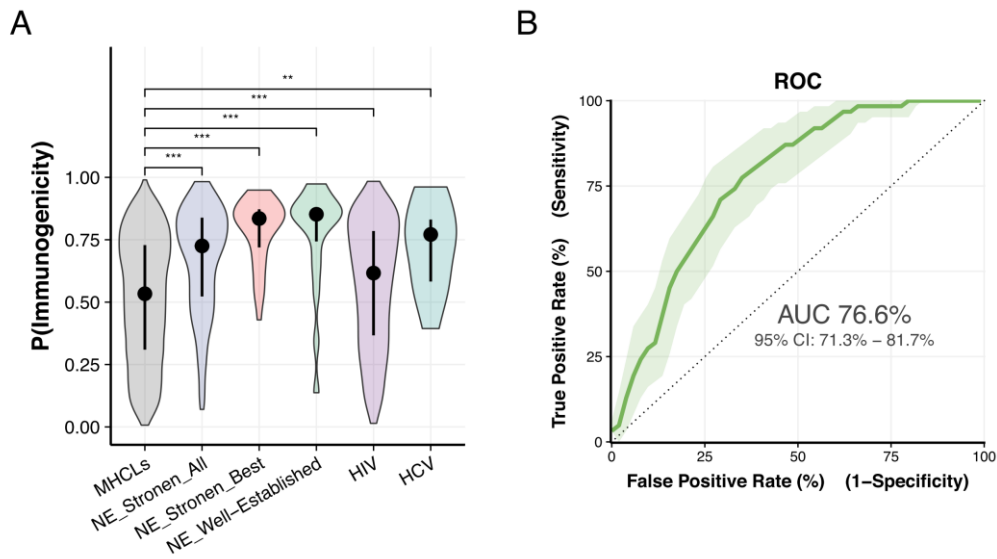


677 **Figure 3. Schematic diagram of modified rTPCP (mrTPCP).** All TCR sequences in  
678 the reference repertoire are fragmented according to the fixed window size. A  
679 position-specific peptide-derived fragment was matched against a set of TCR-derived  
680 fragments. Representative statistics were calculated both in a position-specific and  
681 position-blind (i.e., pooled) manner. Owing to the position-dependent nature of the  
682 analysis, only nonapeptides (=9-mers) were considered in the subsequent analysis.  
683



684 **Figure 4. Improved immunogenicity prediction using mrTPCP.** (A) ROC and  
685 calibration plots of the SVM classifier trained using mrTPCP variables derived from  
686 MIYS990106. See the legend of Fig. 2B and method sections for further details. (B)  
687 HLA-stratified ROC analysis. The entire Chowell dataset was sorted according to their  
688 HLA restriction, and six most data-rich HLA supertypes were selected for visualization.  
689 (C) Sequence manipulation analysis. Either the input peptide sequences or the reference  
690 TCR repertoire sequences were manipulated, and mrTPCP variables were calculated. The  
691 authentically trained SVM classifier was applied. Inv, inversion of the sequence; Perm,  
692 permutation of the sequence; Syns, randomly synthesized sequences with relative amino

693 acid frequencies retained. For peptides, amino acid frequencies of immunogenic and  
694 non-immunogenic peptides were separately considered; Random, completely random  
695 sequences. (B-C) AUC was calculated and graphics were generated using *pROC* and  
696 *plotROC* packages in R, respectively.  
697



698 **Figure 5. Immunogenicity prediction of independent datasets of viral epitopes and**

699 **tumor neopeptides.** Peptide data were collected from various sources. Any peptide

700 overlapping with those in the Chowell dataset was excluded. The probability of

701 immunogenicity was estimated by applying the mrTPCP-based SVM classifier (Fig. 4A).

702 (A) A metadataset of viral epitopes and tumor neopeptides(4, 10, 24, 29–32). Dots and

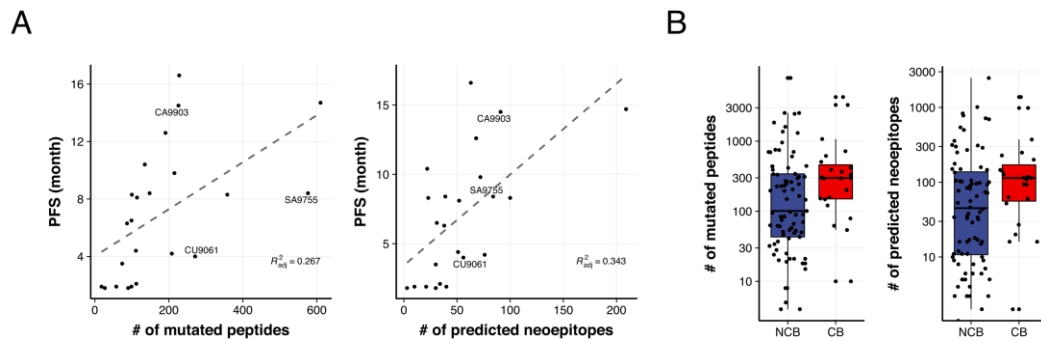
703 bars represent the median and interquartile range, respectively. \*\*\*:  $p < 0.001$ , \*\*:  $p$

704  $< 0.01$ .  $P$  values were determined using Wilcoxon's rank sum test. NE, neopeptides. (B)

705 Epitope/ligand data originally reported by Calis et al.(21). The probability threshold was

706 set to be 0.80.

707



708 **Figure 6. Correlation between predicted neopeptide burden and clinical outcome in**

709 **checkpoint inhibitor trials.** The mrTPCP-based SVM classifier (Fig. 4A) was applied to

710 external datasets of tumor mutational landscapes obtained from checkpoint inhibitor

711 trials. We set the threshold of immunogenicity to be above 0.80, on the basis of the

712 observation that most of the well-established tumor neopeptides exhibited probabilities of

713 higher than 0.80 (Fig. 5A) (A) Progression-free survival (PFS) correlated with mutational

714 burden/predicted neopeptide burden in non-small cell lung carcinoma (NSCLC) patients

715 treated with pembrolizumab(3). The three patients were labeled, for which scaled fitting

716 residuals decreased by more than 1 when the predicted neopeptide burden was used as a

717 correlate. Adjusted correlation coefficient ( $R^2_{adj}$ ) was calculated using the *stat\_poly\_eq*

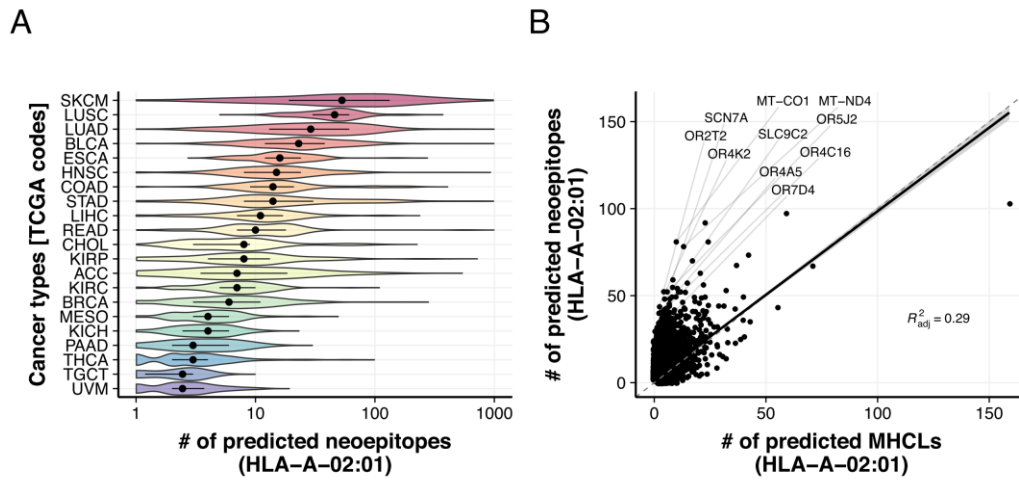
718 package.(B) Clinical benefit (CB) was associated with heavier mutational

719 burden/predicted neopeptide burden in melanoma patients treated with ipilimumab(5).

720 CB was defined as in the original paper(5). NCB, no clinical benefit.

721





722 **Figure 7. Neopeptide burdens in TCGA datasets.** Mutation data were retrieved from  
723 all advanced stage tumors in The Cancer Genome Atlas (TCGA)(34). For the 105959  
724 HLA-A-02:01-restricted nonapeptides predicted to be stable MHC binders,  
725 immunogenicity prediction was carried out. (A) Predicted neopeptide burden was  
726 visualized for the 21 tumor types registered in TCGA. SKCM, Skin Cutaneous  
727 Melanoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma. For a  
728 complete set of abbreviation used in TCGA, visit the NCI Genomic Data Commons  
729 (<https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations>).  
730 (B) The correlation between predicted MHCL burden and neopeptide burden per gene.  
731 HUGO symbols were depicted for genes enriched with neopeptides. Enrichment was  
732 defined as fitting residuals being larger than 10 for the purpose of tidy visualization.  
733 Adjusted correlation coefficient was calculated using the *stat\_poly\_eq* package.  
734

## 735 Tables

736

### 737 Table 1. Prediction results on datasets independent from training/validation data.

738 Immunogenicity was predicted using the mrTPCP-based SVM classifier (Fig. 4A). Note

739 that any peptide contained in the Chowell dataset was excluded. Unadjusted  $p$  values

740 were calculated against the IEDB MHCL data as the negative control using the *prop.test*

741 function implemented in R with continuity correction. N.A., not applicable.

742

Dataset	Predicted Epitope	$P$ value	Data Source
HCV CTL epitopes [n = 11]	10 (90.9%)	< 0.01	HCV immunology database ( <a href="https://hcv.lanl.gov/content/immunology/tables/ctl_summary.html">https://hcv.lanl.gov/content/immunology/tables/ctl_summary.html</a> )
HIV CTL epitopes [n = 867]	545 (62.9%)	< 0.001	HIV molecular immunology database ( <a href="https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html">https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html</a> )
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 ( <a href="http://www.iedb.org/">http://www.iedb.org/</a> )

743