The landscapes of T cell epitope immunogenicity in sequence space

Authors: M. Ogishi,[1]* H. Yotsuyanagi[1]

Affiliations:

[1]Division of Infectious Diseases and Applied Immunology, The Institute of Medical Sciences Research Hospital, The University of Tokyo, Tokyo, 108-0071, Japan.

*Correspondence to: M. Ogishi (oogishi-tky@umin.ac.jp)

Abstract: Epitopes bound to major histocompatibility complex (MHC) triggers T cell immunity. However, determinants of epitope immunogenicity remain poorly understood. Here, we describe sequence-based framework mimicking contacts between T cell receptors (TCRs) and peptides. Machine-learning compressed the simulated contact profiles of 21,162 and 31,693 human MHC-I and MHC-II-restricted peptides against public TCR repertoires into a linear coordinate system termed "immunogenicity score." Simulated contact profiles against individual TCRs showed remarkable consistency with observed affinities and contact footprints. Furthermore, comprehensive analysis untangled position- and residue-specific impact of single amino acid change on immunogenicity. Finally, patients with longer survival after checkpoint blockade immunotherapies exhibited higher-immunogenicity neoepitope profiles in addition to higher neoepitope burden. Overall, our framework quantitatively formulates T cell epitope immunogenicity by integrating structural and physicochemical parameters.

One Sentence Summary: Prediction of MHC-I and MHC-II-restricted T cell epitope immunogenicity from sequence-based TCR-peptide contact potential profiling.

Main Text: T cell epitopes bound to major histocompatibility complex [MHC; also called the human leukocyte antigen (HLA) in humans] molecules, when recognized by T cell receptor (TCR), activate T cells to initiate subsequent immunological orchestration (*1–3*). MHC class I (MHC-I) molecules typically present 8- to 11-aa peptides generated through proteasomal cleavage of intracellular proteins to activate $CD8^+$ cytotoxic T lymphocytes (CTLs), whereas MHC class II (MHC-II) molecules with an open-ended binding groove accommodate peptides of more variable length derived from endocytosed proteins to activate $CD4^+$ T helper (Th) cells (*4*). In contrast to the success in predicting intracellular processing and MHC binding (*5–7*), however, the observation that only a small subset of MHC binders trigger functional T cell immunity hinders expeditious exploration of potent T cell epitopes essential for the development of effective vaccines and targeted immunotherapies (*8–10*). By definition, functional T cell activation necessitates TCR recognition of peptide-MHC (pMHC) complexes (*11*, *12*). Previous studies have associated hydrophobicity at TCR contact residues and sequence-level similarity to known pathogen-derived epitopes with immunogenicity (*13*, *14*). However, there remains no unifying explanation linking physicochemical and structural constraints of TCR-peptide interactions and T cell immunity. In this context, we show that a computational framework mimicking the molecular scanning of pMHC complexes by host TCR repertoire, which we term TCR-peptide contact potential profiling (CPP), enables accurate prediction of T cell epitope immunogenicity.

The concept of CPP stands on the hypothesis that public TCR repertoire is biased toward pathogen-derived epitopes through evolutional purifying selection and postnatal negative selection and therefore could be used as a non-linear evaluation function to dichotomize "publicly" immunogenic peptides and those least likely immunogenic (Fig. 1A and Method)

(*15–17*). We postulate our fragment-based interaction model based on the poor surface complementarity, the relatively low affinities (K_D = 0.1 ~ >500 μM) compared to other immunoglobulin-like molecules including antibodies, and the substantial conformational changes induced upon recognition of pMHC complexes by the complementarity-determining region (CDR) loops (*18–20*). We focused on TCR β chain CDR3 (CDR3β) loop because this region has the highest genetic variability and is primarily responsible for the interactions with peptides presented onto the MHC grooves, whereas more conserved CDR1 and CDR2 loops typically interact with MHC α-helices (*2*, *3*, *20*, *21*).

To gain a proof-of-concept, we first compiled epitope metadatasets comprising 21,162 8- to 11-aa HLA-I-restricted and 31,693 11- to 30-aa HLA-II-restricted peptide sequences from various sources (Data S1) (*13*, *22–29*). We defined positivity as the existence of at least one positive functional T cell assay result. A total of 6,957 and 16,642 HLA-I and HLA-II-restricted epitopes met the criteria. Next, we extracted 23,006,555 unique CDR3β sequences from TCR repertoire datasets derived from healthy individuals using MiXCR software (*30*). We obtained 191,326 unique CDR3β sequences identified in at least 22 out of 220 (10%) different datasets to construct fragment libraries for the CPP feature computation. Importance-guided feature selections covering CPP-specific features and other peptide descriptors yielded the 32 and 27 most predictive features for MHC-I and MHC-II predictions, respectively (Data S2). We applied machine-learning techniques to compress these features in conjunction with predicted percentile ranks to representative HLA alleles by NetMHC4.0 (*6*) and NetMHCIIpan3.2 (*7*) into a unidimensional scale. More specifically, we termed the averaged probability estimates of immunogenicity from 5 times repeated 5-fold cross-validations (CVs) as “immunogenicity scores.” Probabilistic estimates were reasonably consistent across CVs; the normalized standard

deviations distributed with medians and interquartile ranges (IQRs) of 9.3% (6.4% to 13.6%) and 6.5% (4.3% to 9.4%) in MHC-I and MHC-II, respectively (fig. S1). The resultant scores showed high predictive performances (Figs. 1B and 1C). The optimal thresholds determined from the receiver operating characteristic (ROC) curves were 0.36 and 0.51 for MHC-I and MHC-II, respectively. We found consistent dichotomization of epitopes and binders across various subsets of the datasets (Figs. 1D-G). Predictions remained accurate even when the numbers of features were curtailed down to seven and twelve for MHC-I and MHC-II, respectively. Meanwhile, removal of peptides based on sequence-level homology modestly undermined the predictive performances, especially in MHC-I (fig. S2). Contrastingly, the currently available MHC-I immunogenicity prediction tool (*22*) showed only marginal predictive power, with the area under the ROC curve (AUC) of 54% on our metadataset (fig. S3). There is no tool for MHC-II to compare.

We next decomposed the CPP-derived features into a set of parameters, namely, fragment lengths, summarizing statistics, and AAIndex amino acid pairwise contact potential scales (*31*), to investigate their predictive contributions and potential biological relevance. Notably, features derived from short fragments (*i.e.*, 3- and 4-aa) and the longest fragment (*i.e.*, 8-aa and 11-aa in MHC-I and MHC-II, respectively) exhibited higher cumulative importance values (Fig. 2A and fig. S4). Meanwhile, features reflecting kurtosis and skewness exhibited higher importance, indicating that distinct distributions of contact potentials are the hallmark of immunogenicity (Fig. 2B). Interestingly, the inverted version of BETM990101 (termed BETM990101inv) (*32*) and other six AAIndex scales of the highest importance for both MHC-I and MHC-II (Fig. 2C). To explore the biological implications of these parameters, we analyzed 82 TCR-peptide complexes with experimentally determined affinities collected from literature (Data S3) through

a modified framework which we termed single-TCR contact potential profiling (sCPP). We found that BETM990101inv and other six AAIndex scales important for immunogenicity prediction also generated sCPP features strongly correlating with affinities (Figs. 3A-B and table S1). Sequence permutation experiments significantly undermined the observed correlations (Fig. 3C). Structural analysis using PRODIGY (*33*) suggested that around 50% and more than 80% of the intermolecular contacts are limited within the 3-aa and 5-aa ranges, respectively, in both peptides and CDR3β regions (Fig. 3D), a finding supportive of the previous observation that 3-aa and 4-aa fragments are indispensable in predicting immunogenicity (Fig. 2A). Moreover, we found strong correlations between structure-guided and predicted contact footprints, *i.e.*, the density distributions of TCR contacts by peptide positions (Figs. 3E-F). Surprisingly, BETM990101inv was again among the best-correlating AAIndex scales. The observed correlations depended significantly upon the peptide sequences, and less evidently, the CDR3β regions. Collectively, these observations strongly suggest that our CPP framework recapitulates the high-dimensional architectures of TCR recognition of pMHC complexes and subsequent activation of CTL and Th cells.

Single amino acid mutations could modulate peptide immunogenicity bidirectionally. Mutation-driven acquisition and loss of immunogenicity are in principle the two sides of the same coin in sequence space (Fig. 4A). To systematically investigate the mutational impacts, we screened our epitope datasets to identify 6,128 and 3,413 MHC-I and MHC-II-restricted peptides that have their neighbors, *i.e.*, counterparts with single amino acid differences (Data S4). We termed the discordance of annotated immunogenicity between neighbors as “immune transition.” Peptides are defined as “transitional” if there is at least one immune transition (Fig. 3A). In this manner, we found 660 and 700 transitional MHC-I epitopes and binders, and 500 and 476

transitional MHC-II epitopes and binders, respectively (Data S5). Next, using the same feature sets and the same machine-learning framework for immunogenicity prediction, we constructed classifiers predicting immune transition status. The averaged probability estimates, which we termed “immune transition scores,” predicted immune transition status with AUCs of 92% and 94% in MHC-I and MHC-II, respectively (Fig. 4B). Notably, the estimation of immunogenicity was near-perfect for non-transitional peptides (AUC: 92% and 95% in MHC-I and MHC-II, respectively), but not comparably accurate for transitional ones (Fig. 4C). Consistently, when extrapolated to the entire peptide datasets (N=21,162 and N=31,693 for MHC-I and MHC-II, respectively), we found a monotonous decrease in the predictive accuracy regarding immunogenicity across the five quintiles of immune transition scores (fig. S5). To further characterize the immunogenicity landscape of transitional peptides, we next examined the properties of neighbor networks, where nodes and edges represent peptides and single-aa mutations with edge directions and weights defined from their changes in immunogenicity scores. We showed a representative neighbor network in Figure 4D. Notably, only epitopes in binder-enriched clusters were assigned low immunogenicity scores. Therefore, we computationally generated all possible single-aa mutants from transitional peptides and computed the averaged immunogenicity scores. As expected, averaged immunogenicity scores accurately predicted immunogenicity (Fig. 4E). Removal of mutants residing in other clusters even magnified the differences. Stratification of peptides by their transitional status, in conjunction with the simulation of single-aa mutations for transitional peptides, achieved extremely high predictive performances (Figs. 4F-G). Moreover, when focusing on neighboring peptide pairs, transitional changes were associated with more substantial score discrepancies (Fig. 5A). Mutations at anchor residues and indels led to more substantial impact than those at

non-anchor positions and substitutions, respectively (Figs. 5B-E). Heatmap clustering analysis revealed residue-specific effects on predicted immunogenicity that were partially interpretable by residue-level physicochemical properties [e.g., hydrophobic/aliphatic (V, I, L, M), negative (D and E), and aromatic (F and W)], with some exceptions [e.g., positive (K and R) in the MHC-I prediction]. Collectively, our comprehensive analysis delineates the landscape of bidirectional effects of single-aa mutations on immunogenicity in a position-dependent context.

Checkpoint blockade therapy augmenting neoepitope-targeting T cell immunity has fundamentally transformed the current oncologic treatment landscape (*34–36*). Although the neoepitope quantity correlates with clinical benefits, the quality of neoepitope triggering active tumor rejection and serving better outcome has remained barely understood (*37–39*). Indeed, the majority of predicted neoepitope candidates fail to elicit T cell responses (*10*). In this context, we tested the feasibility of applying our framework to assess the immunogenicity of tumor neoepitopes. Neoepitopes identified from the TANTIGEN database (*26*) had scores comparable to or even higher than other MHC-I-restricted epitopes (Figs. 6A-C). Furthermore, our model assigned high scores to neoepitopes from established neoantigens including PMEL/gp100, CTG1B, and MAR1/MART1 (Fig. 6D). Lastly, although not a neoepitope, the score of the famous tumor-associated antigen NY-ESO-1 (SLLMWITQV) was 0.82. Encouraged by these observations, we next computed immunogenicity scores for a set of neoepitope candidates extracted from mutational profiling datasets of three checkpoint blockade cohorts (Data S6) (*37–39*). We constructed a univariate Cox proportional-hazards model based on the log-transformed per-patient neoepitope burden to dichotomize patients into good and poor responders (figs. S6A-E). The optimal threshold for neoepitope burden was 90. Patients with better prognosis were more likely to have higher diversity in immunogenicity scores as measured by kurtosis (figs.

S6F-G). Moreover, we observed a higher proportion of high-immunogenicity neoepitopes in good responders (Figs. 6E-F), with a score threshold of 0.55, indicating that quality, as well as quantity, may be associated with survival. Overall, our analysis places tumor neoepitope biology in a more generalized context of CTL immunity and enables quantitative comparison of epitope quality between diseases.

Our results indicate that TCR-peptide contact potential profiling *in silico* can explain the TCR affinities and contact sites and predict immunogenicity of both MHC-I and MHC-II-restricted T cell epitopes by utilizing a restricted set of amino acid pairwise contact potential scales. Our framework utilizes public TCR repertoire as an evaluation function to assess the non-linear similarity between MHC-presented peptides and map them into a linear coordinate system termed immunogenicity scores. With this coordinate system, we can not only predict the immunogenicity of a given peptide but also quantitatively discuss the distance and dynamics of immunological states that are highly heterogeneous between specimens and individuals. These findings warrant prospective evaluation of the system in various clinical contexts including vaccination and cancer immunotherapy cohorts.

References and Notes:

1. A. Grakoui *et al.*, The immunological synapse: a molecular machine controlling T cell activation. *Science*. **285**, 221–7 (1999).

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

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

4. J. Neefjes, M. L. M. Jongsma, P. Paul, O. Bakke, Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).

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

6. M. Andreatta, M. Nielsen, Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*. **32**, 511–517 (2016).

7. K. K. Jensen *et al.*, Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. **154**, 394–406 (2018).

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

9. E. W. Newell *et al.*, Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat. Biotechnol.* **31**, 623–629 (2013).

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

11. S. Tian, R. Maile, E. J. Collins, J. A. Frelinger, CD8+ T cell activation is governed by TCR-peptide/MHC affinity, not dissociation rate. *J. Immunol.* **179**, 2952–60 (2007).

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

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

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

15. J. J. Miles, D. C. Douek, D. A. Price, Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol. Cell Biol.* **89**, 375–387 (2011).

16. A. Madi *et al.*, T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *Elife*. **6**, 1–17 (2017).

17. H. S. Robins *et al.*, Overlap and Effective Size of the Human CD8+ T Cell Receptor Repertoire. *Sci. Transl. Med.* **2**, 47ra64 (2010).

18. K. C. Garcia *et al.*, Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science*. **279**, 1166–72 (1998).

19. J. J. Boniface, Z. Reich, D. S. Lyons, M. M. Davis, Thermodynamics of T cell receptor binding to peptide-MHC: evidence for a general mechanism of molecular scanning. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11446–11451 (1999).

20. J. S. Bridgeman, A. K. Sewell, J. J. Miles, D. A. Price, D. K. Cole, Structural and biophysical determinants of αβ T-cell antigen recognition. *Immunology*. **135**, 9–18 (2012).

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

22. J. J. A. Calis *et al.*, Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol*. **9**, e1003266 (2013).

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

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

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

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

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

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

29. W. Fleri *et al.*, The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol.* **8**, 1–16 (2017).

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

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

32. M. R. Betancourt, D. Thirumalai, Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **8**, 361–369 (1999).

33. A. Vangone, A. M. J. J. Bonvin, Contacts-based prediction of binding affinity in protein–protein complexes. *Elife*. **4**, 1–15 (2015).

34. M. M. Gubin *et al.*, Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*. **515**, 577–581 (2014).

35. E. M. E. Verdegaal *et al.*, Neoantigen landscape dynamics during human melanoma–T cell interactions. *Nature*. **536**, 91–95 (2016).

36. S. J. Antonia *et al.*, Durvalumab after Chemoradiotherapy in Stage III Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **377**, 1919–1929 (2017).

37. N. A. Rizvi *et al.*, Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. **348**, 124–8 (2015).

38. A. Snyder *et al.*, Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

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

40. R Core Team, R: A Language and Environment for Statistical Computing (2018).

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

42. R. C. Gentleman *et al.*, Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

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

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

45. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).

46. T. Borrman *et al.*, ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins Struct. Funct. Bioinforma.* **85**, 908–916 (2017).

47. L. C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, A. Vangone, PRODIGY: A web server for predicting the binding affinity of protein-protein complexes. *Bioinformatics*. **32**, 3676–3678 (2016).

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

49. U. Bodenhofer, E. Bonatesta, C. Horejš-Kainrath, S. Hochreiter, msa: an R package for multiple sequence alignment. *Bioinformatics*. **31**, btv494 (2015).

50. N. McGranahan *et al.*, Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*, 1259–1271 (2017).

51. D. Chowell *et al.*, Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*. **359**, 582–587 (2018).

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

53. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. **32**, 2847–2849 (2016).

54. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*. **57**, 289–300 (1995).

55. E. Ledell, M. Petersen, M. Van Der Laan, Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.* **9**, 1583–1607 (2015).

Acknowledgments: We thank Dr. R. Ueno for helpful advice. **Funding:** This work is not supported by any external funding sources. **Author contributions:** M.O. conceived the idea, designed the study, performed computational analyses, and drafted the manuscript; M.O. and H.Y. wrote the manuscript. **Competing interests:** The authors declare no competing interests.

Data availability: Compiled datasets and essential in-house functions are available as the R package *Repitope* on GitHub (https://github.com/masato-ogishi/Repitope/). Other scripts are available upon request.

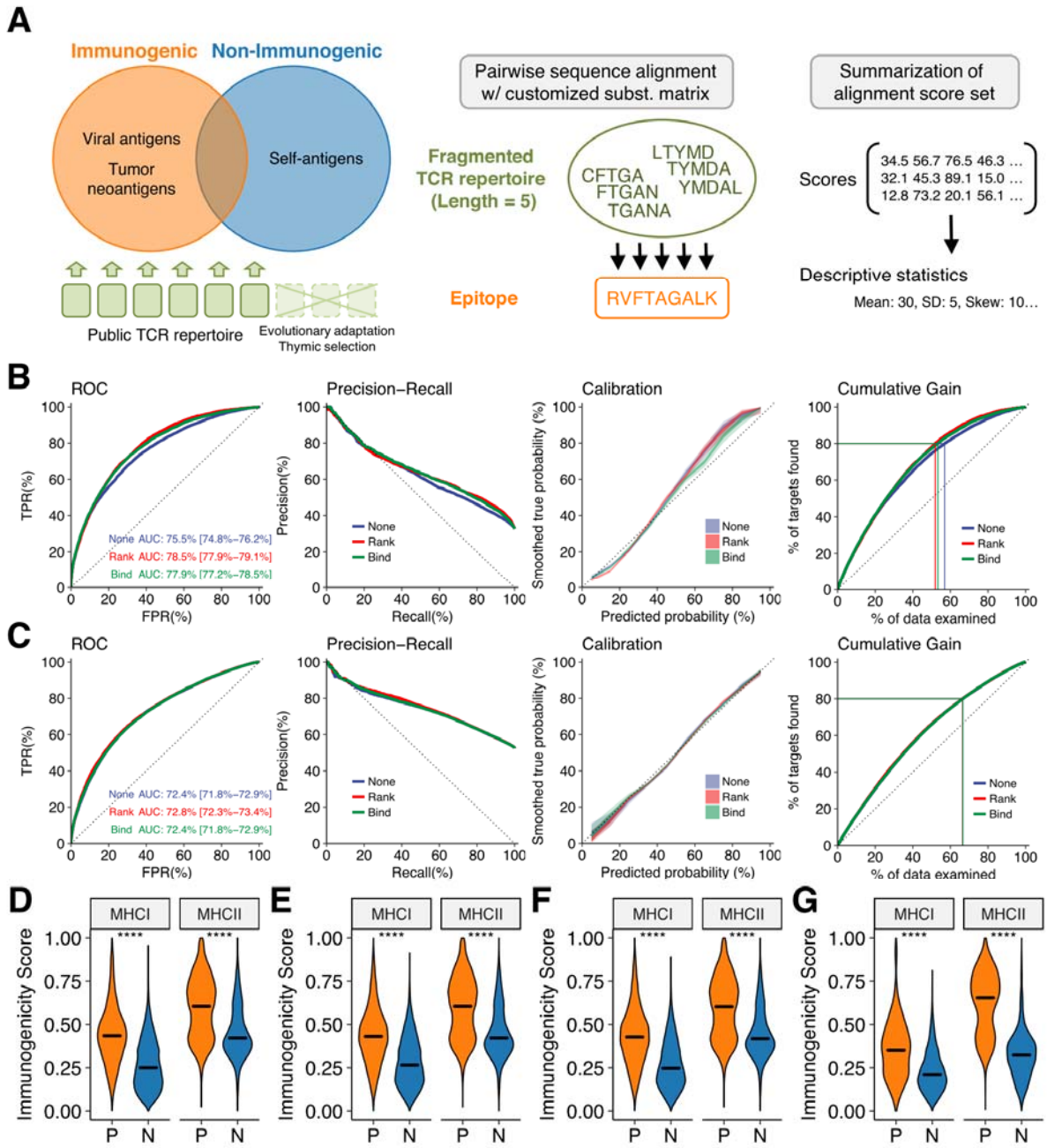

Fig. 1. Immunogenicity prediction from TCR-peptide contact potential profiling. (**A**) We formulate TCR-peptide contact potential profiles to identify "publicly" immunogenic epitopes by exploiting public TCR repertoire inherently biased toward common pathogen antigens but away from self-antigens due to genome-level evolution and postnatal thymic selection. (**B-C**)

Immunogenicity prediction for (**B**) MHC-I and (**C**) MHC-II-restricted peptides. Using the 32 and 27 most predictive features for MHC-I and MHC-II, respectively (Data S2 and fig. S4), averaged probabilistic estimates of immunogenicity were computed through five times repeated 5-fold cross-validations (CVs), either without (blue) or with predicted rank percentiles (red; termed "immunogenicity scores") or binder categories (green) to representative HLA alleles (*6*, *7*). (**D-G**) Distributions of immunogenicity scores. P, positive (immunogenic); N, negative (non-immunogenic). (**D**) All peptides (N=21,162 and 31,693 for MHC-I and MHC-II, respectively). (**E-F**) IEDB-derived peptides with functional T cell assay annotations (**E**) in any host (N=13,058 and 31,693) or (**F**) in human (N=11,084 and 30,188). MHC tetramer binding was not considered functional evidence. (**G**) IEDB-derived peptides annotated from direct *ex vivo* assays using autologous effector and antigen-presenting cells in human (N=6,048 and 2,023). Bars represent medians. *P*-values were calculated by Wilcoxon rank sum test. ****, $P < 0.0001$.

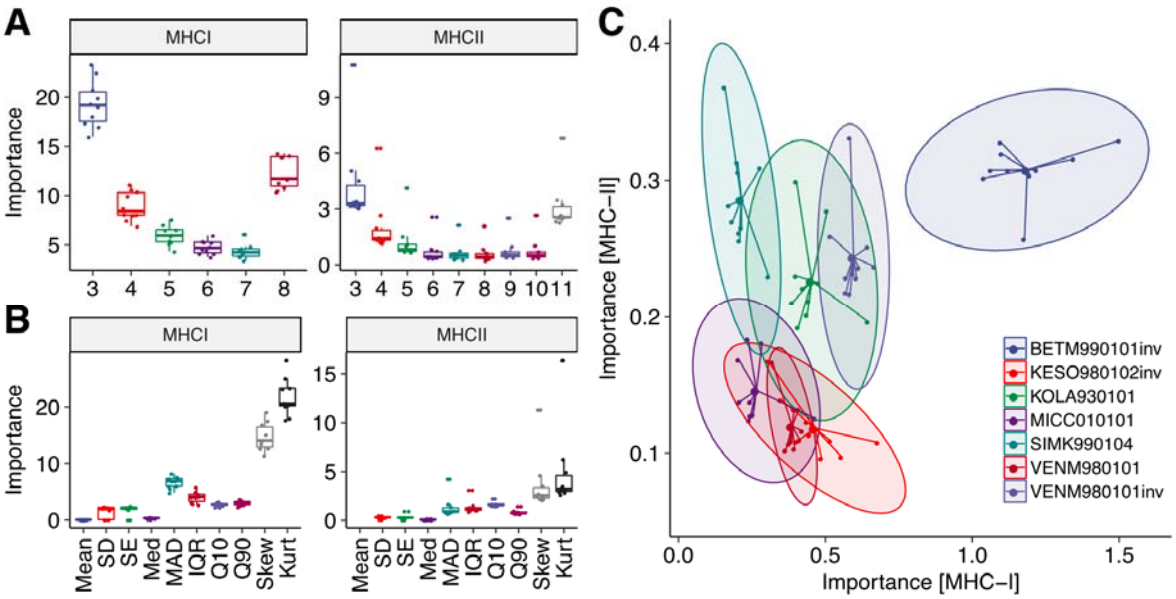

Fig. 2. Contact potential profiling parameters consisting of the features most predictive of immunogenicity. Feature importance estimates from ten iterations during the feature selection steps were retrieved. (**A**-**B**) Importance of all CPP features aggregated by (**A**) fragment length and (**B**) descriptive statistics. (**C**) Importance of the most predictive 21 and 26 CPP features for MHC-I and MHC-II, respectively, aggregated by their source AAIndex scales. Only the seven most predictive AAIndex scales commonly utilized in MHC-I and MHC-II immunogenicity predictions are shown.

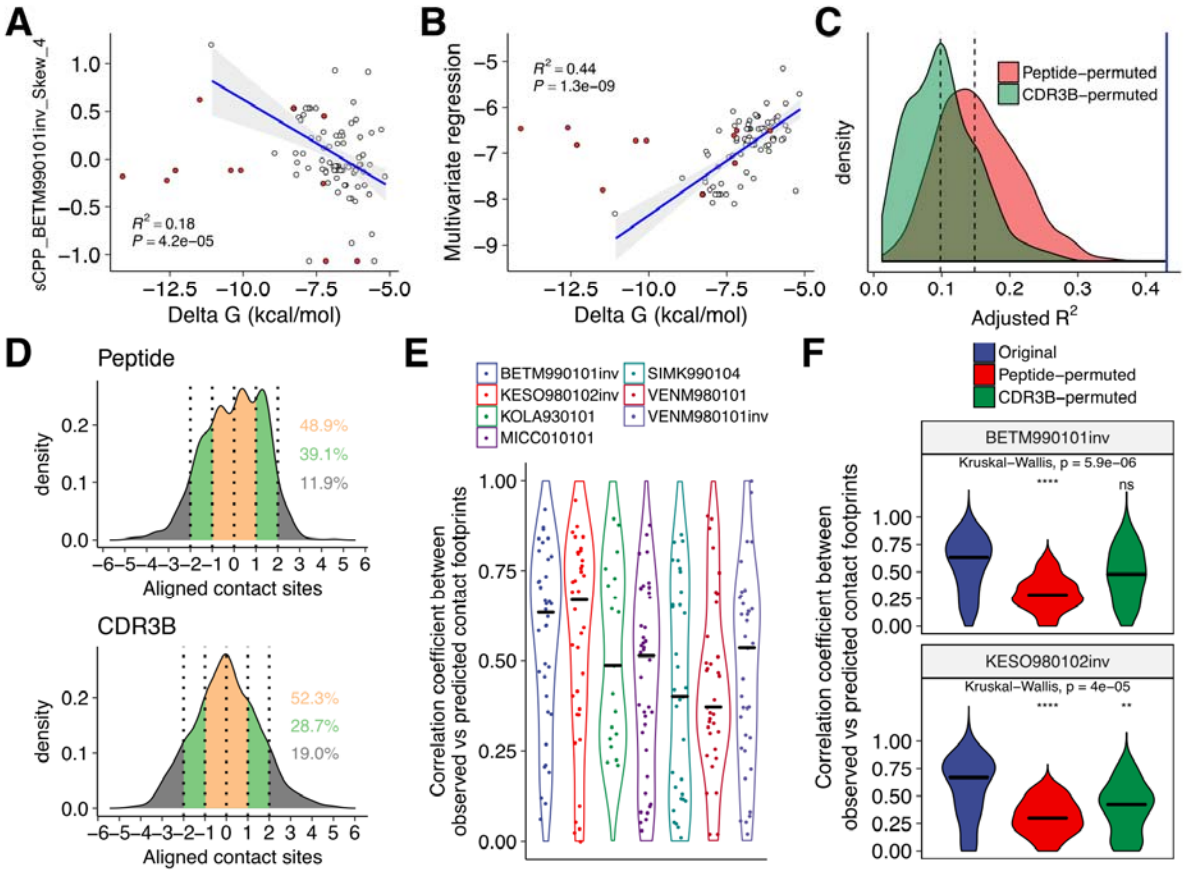

Fig. 3. Concordance of experimentally determined and simulated TCR-peptide interactions. (**A-C**) Correlations between affinity and features derived from single-TCR contact potential profiling (sCPP) in 82 wildtype (WT) TCR-peptide complexes. Thirteen mutant (MT) TCRs were excluded from correlation analysis and only shown for comparison (red points). Adjusted squared Pearson's correlation coefficients are presented. (**A**) The best-correlating BETM990101inv-derived feature. (**B**) Multivariate regression with features derived from each of the AAIndex scales predictive of immunogenicity (Fig. 2C). (**C**) Permutation experiments with 1,000 iterations. Multivariate regression was performed as in (**B**) using sCPP features computed from either randomly permuted peptide sequences or CDR3B sequence fragments. The blue line represents the non-permuted regression result. Dashed lines represent medians. (**D**) Center-

aligned contact footprints, *i.e.*, density distributions, determined from 51 unique peptide-WT TCR structures using the PRODIGY server (*33*). Inset numbers indicate the proportions of contacts falling into the color-specified ranges. (**E-F**) Concordance of contact footprints predicted from 4-aa TCR fragment library to observed footprints. In (**F**), permutation experiments were conducted as in (**C**). Bars represent medians. *P*-values were calculated by Kruskal-Wallis test followed by Wilcoxon rank sum test with false-discovery rate (FDR) adjustment. ns, not significant; **, $P < 0.01$; ****, $P < 0.0001$.

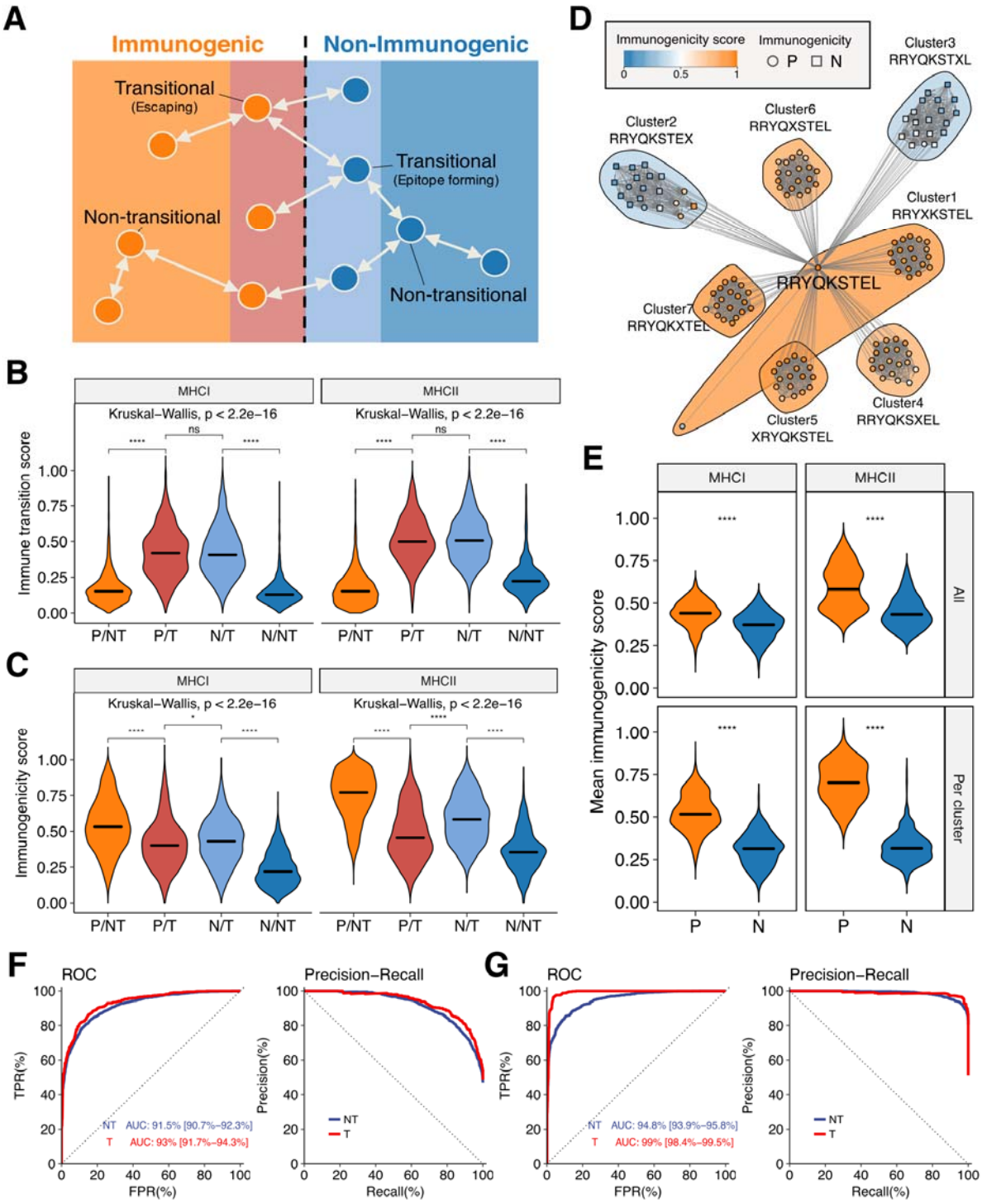

Fig. 4. Accurate immunogenicity prediction by integrating adjacent single amino acid mutant information in sequence space. (**A**) Schematic of the concept of immune transition in

sequence space. We identified 6,128 and 3,413 MHC-I and MHC-II-restricted peptides with neighbors, *i.e.*, single-aa mutated counterparts. We termed the change in immunogenicity between neighbors as "immune transition." Peptides are defined as "transitional" if there is at least one immune transition in our epitope metadataset. (**B-C**) Distributions of (**B**) immune transition scores and (**C**) immunogenicity scores stratified by transitional patterns. Probabilistic estimates of immune transition status, immune transition scores, were computed similarly to immunogenicity scores using the same feature sets. (**D**) A representative neighbor network of an HLA-I-restricted epitope RRYQKSTEL and 132 neighbors of which 32 are non-immunogenic (Method). (**E**) *In silico* mutagenesis. Distributions of averaged immunogenicity scores of simulated single amino acid mutants derived from each of the transitional peptides are shown. (**F-G**) Predictive performance of immunogenicity scores for (**F**) MHC-I and (**G**) MHC-II-restricted peptides stratified by transitional status. P, positive (immunogenic); N, negative (non-immunogenic); T, transitional; NT, non-transitional. Bars represent medians. FDR-adjusted *P*-values were calculated by Kruskal-Wallis test followed by Wilcoxon rank sum test. ns, not significant; *, $P < 0.05$; ****, $P < 0.0001$.

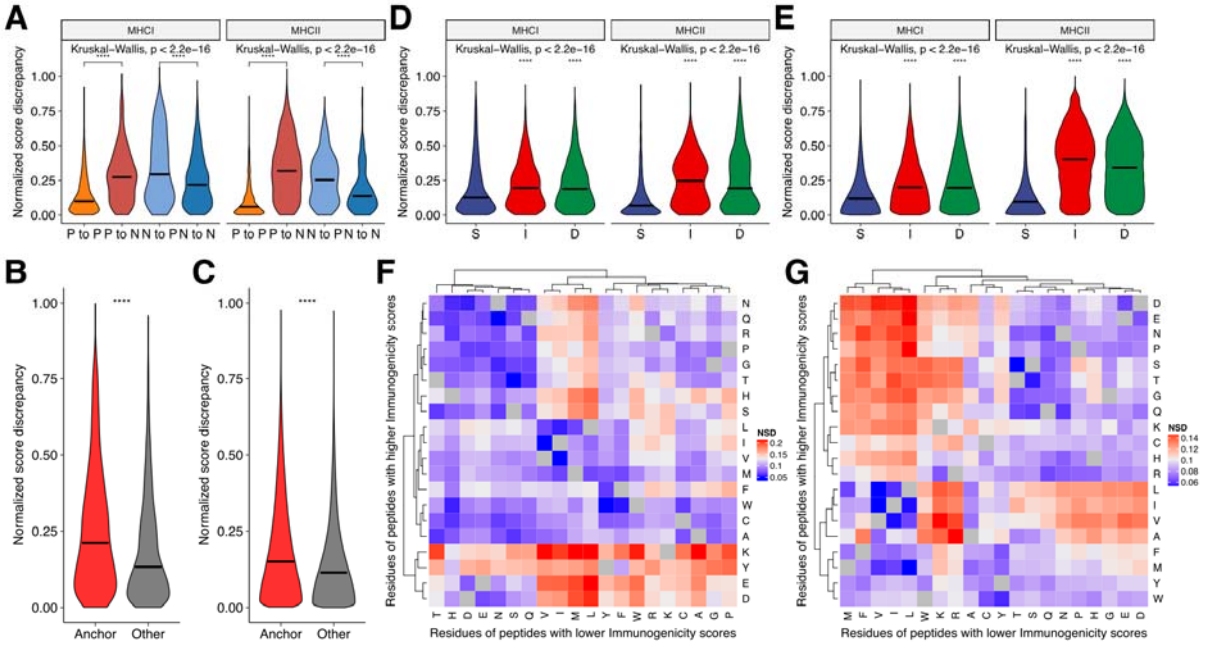

Fig. 5. Effects of single mutations on immunogenicity. We identified 6,174 and 3,079 single-mutated peptide pairs in our datasets, of which 940 and 535 were transitional, for MHC-I and MHC-II, respectively. We also computationally generated 2,580,603 and 3,100,488 single-mutated peptide pairs from 232,723 and 292,437 all possible single-aa mutants of transitional MHC-I- and MHC-II-restricted peptides, respectively. (**A**) Normalized absolute score discrepancies (NSDs) in single-mutated peptide pairs stratified by transitional patterns. P, positive (immunogenic); N, negative (non-immunogenic). (**B-C**) NSDs in (**B**) observed and (**C**) simulated MHC-I-restricted peptide pairs stratified by mutated positions. Anchor residues were assumed to be at position 2 and the C-terminus. (**D-E**) NSDs in (**D**) observed and (**E**) simulated peptide pairs stratified by mutation types. S, substitution; I, insertion; D, deletion. (**F-G**) Median NSDs in simulated (**F**) MHC-I-restricted and (**G**) MHC-II-restricted peptide pairs stratified by mutated residues. Bars represent medians. FDR-adjusted *P*-values were calculated by Kruskal-Wallis test followed by Wilcoxon rank sum test. ****, $P < 0.0001$.

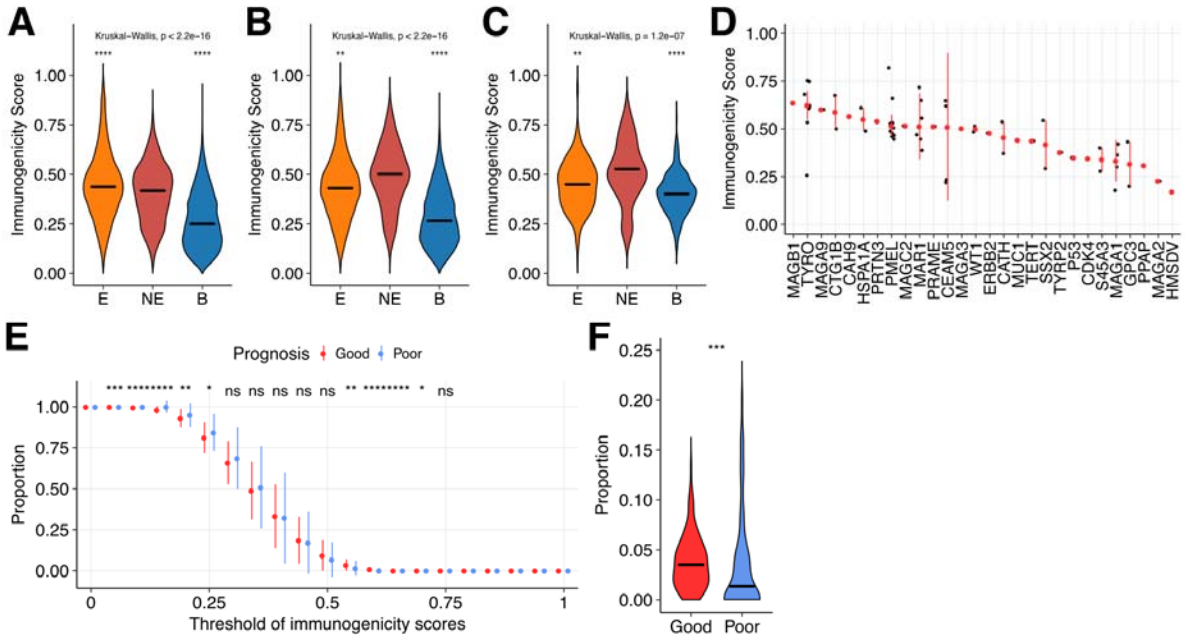

Fig. 6. Prediction of neoepitope immunogenicity landscapes. (**A**-**C**) Distributions of MHC-I immunogenicity scores of (**A**) all peptides (N=21,162), (**B**) IEDB-derived peptides annotated from functional T cell assays (N=13,058), and (**C**) IEDB-derived peptides derived from human antigens annotated from direct *ex vivo* assays using autologous effector/antigen-presenting cells in human (N=503). NE, neoepitopes identified from the TANTIGEN database(*26*); E, other epitopes; B, non-immunogenic MHC binders. Bars represent medians. (**D**) Immunogenicity scores of 66 neoepitopes derived from 28 proteins. Red knots and bars represent medians and interquartile ranges. (**E**-**F**) Proportions of high-immunogenicity neoepitopes among patients with good (red) or poor (blue) response to checkpoint blockade immunotherapies. Tumor mutation datasets from three checkpoint blockade trials were combined to construct a univariate Cox proportional-hazards model using per-patient neoepitope burden predicted based on binding to autologous HLAs. Patients were stratified into either good or poor responder based on the estimated relative risks. (**E**) Proportions of above-threshold neoepitopes per patient. Knots and

bars represent medians and interquartile ranges. (**F**) Proportions of neoepitopes with immunogenicity scores of 0.55 or higher. FDR-adjusted *P*-values were calculated by Wilcoxon rank sum test. ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

Supplementary Materials：

Materials and Methods
Figs. S1-S6
Table S1
Data S1-S7