Published in final edited form as:

Nat Mach Intell. 2021 October ; 3(10): 864–875. doi:10.1038/s42256-021-00383-2.

Deep learning-based prediction of the T cell receptor-antigen binding specificity

Tianshi Lu[1,+], Ze Zhang[1,+], James Zhu[1], Yunguan Wang, Ph.D.[1], Peixin Jiang[2], Xue Xiao, Ph.D.[1], Chantale Bernatchez, Ph.D.[3], John V. Heymach, M.D., Ph.D.[2], Don L. Gibbons, M.D., Ph.D.[2], Jun Wang, Ph.D.[4], Lin Xu, Ph.D.[1], Alexandre Reuben, Ph.D.[2,*], Tao Wang, Ph.D.[1,5,*]

[1]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.

[2]Department of Thoracic/Head & Neck Medical Oncology, MD Anderson Cancer Center, Houston, TX USA, 77030.

[3]Department of Melanoma Medical Oncology, MD Anderson Cancer Center, Houston, TX USA, 77030.

[4]Department of Pathology, New York University Grossman School of Medicine, New York, NY 10016.

[5]Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.

Abstract

Neoantigens play a key role in the recognition of tumor cells by T cells. However, only a small proportion of neoantigens truly elicit T cell responses, and fewer clues exist as to which neoantigens are recognized by which T cell receptors (TCRs). We built a transfer learning-based model, named pMHC-TCR binding prediction network (pMTnet), to predict TCR-binding specificities of neoantigens, and T cell antigens in general, presented by class I major histocompatibility complexes (pMHCs). pMTnet was comprehensively validated by a series of

[*]Corresponding authors: (1) Tao Wang, Ph.D., Quantitative Biomedical Research Center, Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX, 75390, USA; Tao.Wang@UTSouthwestern.edu (lead contact). (2) Alexandre Reuben, Ph.D., Department of Thoracic/Head & Neck Medical Oncology, MD Anderson Cancer Center, Houston, TX, 77030, USA; AReuben@mdanderson.org.

[+]Co-first authors

AUTHOR CONTRIBUTIONS

T.L. created the TCR/pMHC pairing prediction model and carried out the primary data analyses. Z.Z. created the TCR embedding algorithm. J.Z. carried out structural analyses. P.J., C.B., J.H., D.G., and A.R. contributed the in-house validation data. A.R., Y.W., X.X., and L.X. provided input on the study design and reviewed the manuscript. T.W. supervised the whole study.

Supplementary Information The details of data used for the training (Fig. 1) and validation (Fig. 2) of pMTnet and the detailed explanation of the NIES score.

Code availability

The pMTnet software is available on GitHub at https://github.com/tianshilu/pMTnet[73]. Pipeline for HERV expression detection is available on GitHub at https://github.com/jcao89757/HERVranger[74]. QBRC mutation calling pipeline is available on GitHub at https://github.com/tianshilu/QBRC-Somatic-Pipeline[75]. QBRC neoantigen calling pipeline is available on GitHub at https://github.com/tianshilu/QBRC-Neoantigen-Pipeline[76].

COMPETING INTERESTS

We are applying for formal intellectual property protection on the pMTnet model and software. Alexandre Reuben serves on the Scientific Advisory Board and has received honoraria from Adaptive Biotechnologies.

analyses, and showed advance over previous work by a large margin. By applying pMTnet in human tumor genomics data, we discovered that neoantigens were generally more immunogenic than self-antigens, but HERV-E, a special type of self-antigen that is re-activated in kidney cancer, is more immunogenic than neoantigens. We further discovered that patients with more clonally expanded T cells exhibiting better affinity against truncal, rather than subclonal, neoantigens, had more favorable prognosis and treatment response to immunotherapy, in melanoma and lung cancer but not in kidney cancer. Predicting TCR-neoantigen/antigen pairs is one of the most daunting challenges in modern immunology. However, we achieved an accurate prediction of the pairing only using the TCR sequence (CDR3β), antigen sequence, and class I MHC allele, and our work revealed unique insights into the interactions of TCRs and pMHCs in human tumors using pMTnet as a discovery tool.

Keywords

TCR; pMHC; binding; prediction; neoantigen

INTRODUCTION

Neoantigens are short peptides presented by MHC proteins on the surface of tumor cells, which are transcribed and translated from somatically mutated genes. Neoantigens serve as targets for cytotoxic T cells *via* their interactions with T cell receptors (TCRs) and are therefore key players in immunoediting[1]. Immunotherapies, while having transformed cancer patient care, benefit only a small subset of patients[2-5]. Importantly, these immunotherapies have highlighted the role of neoantigens in checkpoint inhibitor-induced immune responses[6]. Therefore, an accurate and comprehensive characterization of the interactions between neoantigens/antigens and TCRs is central to understanding cancer progression, prognosis, and responsiveness to immunotherapy.

One of the most fundamental and unsolved questions regarding neoantigens and antigen biology in general is the lack of understanding of why not all neoantigens elicit T cell responses (immunogenic)[78], notwithstanding that they are expressed and presented on the cell surface. Even less is known about the TCR binding specificity to immunogenic neoantigens presented by MHC molecules (pMHCs). The ability to link pMHCs to TCR sequences is essential for monitoring the interactions between the immune system and tumors, and critical for enhancing the design or implementation of various immunotherapies. For example, selection of neoantigen vaccine candidates could be informed by pre-existence of compatible TCRs in the patient’s circulation. Accordingly, a number of experimental approaches, such as tetramer analysis[9], TetTCR-seq[10] and T-scan[11], have been developed to detect pairing of TCRs and pMHCs. However, these methods are time-consuming, technically challenging, and costly. Furthermore, each technique has its caveats. Ito *el al*[12] examined multiple studies involving such techniques and found their validation rates to be as low as 1%. However, this is likely an underestimation due to many factors, including the rarity of matching TCRs in the patient’s sampled T cell repertoire. These deficiencies call for the development of state-of-the-art bioinformatics algorithms to predict TCR binding specificity of neoantigens, which will significantly reduce the time and cost of identifying the pairings, and will greatly complement experimental approaches.

In this work, we employed transfer learning, a newer branch of deep learning, to train a model, named pMTnet, that can predict the TCR binding specificity of class I pMHCs. We systematically validated pMTnet using a large number of independent validation data and demonstrated the advance of our model over previous works. We applied pMTnet in human tumor sequencing data and made a series of novel observations regarding the sources of immunogenicity, prognosis and treatment response to immunotherapies. Overall, pMTnet addressed the long-standing TCR-pMHC pairing prediction problem, revealed biological insights on the genome-wide scale, and could serve as a basis for constructing biomarkers for predicting immunotherapeutic response.

RESULTS

Deep learning TCR-antigen binding specificity

Conceptually, we employed a staged approach of dividing the goal of learning the TCR-binding specificity of antigens (pMHCs) into three steps, to lower the difficulty level of the prediction task. First, we trained a numeric embedding of pMHCs (class I only) using Long short-term memory (LSTM) network so the protein sequences of antigens and MHCs could be represented numerically. Second, we trained an embedding of TCR sequences using stacked auto-encoders, which again encoded text strings of TCR sequences numerically. These two steps create numeric vectors that are manageable for mathematical operations and set the stage for the final pairing prediction. At the final stage, we created a deep neural network on top of these two embeddings to combine the knowledge from TCRs, antigenic peptide sequences and MHC alleles in a biologically meaningful way. We employed fine-tuning to finalize the prediction model for the pairing between TCRs and pMHCs.

To numerically embed TCRs, we focused on the CDR3 regions of TCRβ chains, which is the key determinant of specificity in antigen recognition[13]. We first encoded amino acid symbols using Atchley factors[14], which use five numbers to comprehensively represent the physicochemical nature of each amino acid. We then built a stacked autoencoder (Fig. 1a) to learn a small numeric embedding of TCRs with the “Atchley” version of TCRs as input from 243,747 unique human TCRβ CDR3 sequences. Details of these data are shown in Supplementary Information. Auto-encoders are capable of capturing key features of complex input through an unsupervised decompose-reconstruction process and embed the captured features of the input in the form of a short numeric vector. Although we only used CDR3β sequences, these CDR3s were composed of V, D and J genes allowing their identities to be indirectly infused into the embedding. We validated this auto-encoder by comparing the input TCRs and reconstructed TCRs. Our analyses show that CDR3s can be reconstructed *via* CDR3 embeddings in a highly faithful manner (Fig. 1b), demonstrating the successful training of this auto-encoder. More examples are shown in Extended Data Figure 1a,b. The Pearson correlations between the original TCR CDR3 Atchley matrices and the reconstructed matrices were generally larger than 0.95 (Extended Data Figure 1c). The validity of this TCR auto-encoder was also supported by our recent publication[15] where we built a statistical model called Tessa, on top of this auto-encoder, which successfully interpreted the functional significance of TCR repertoire from single cell sequencing data.

For embedding of pMHCs, we first re-implemented the netMHCpan model using a deep LSTM neural network (Fig. 1c). This was done so that the internal layers of the netMHCpan model were available for integration with the other parts of our model. The input of this model is the MHC sequence (class I only) and the antigen protein sequence. The output, at this stage of training, is whether the antigens bind to the MHC molecule or not. Although this output layer is dedicated to predicting antigen and MHC binding, the layers prior to it should contain important information regarding the overall structure of the pMHC complex. The same data used for training netMHCpan were used to re-train our model, which consist of 172,422 measurements of peptide-MHC binding affinity covering 130 types of class I MHC from humans. The Pearson Correlation of the predicted binding probability and true binding strength in the independent testing dataset reached 0.781 (Fig. 1d), which is comparable with the Pearson Correlation of 0.76 from the original netMHCpan publication[16]. For the next stage of learning the pairing between TCR and pMHCs, we extracted the immediate layer (a numeric vector) before the final output layer, as the numeric embedding of pMHCs.

Finally, we leveraged the trained numeric vector encodings of TCRs and pMHCs for learning the pairing between them. We constructed a fully connected deep learning network based on the output of these two sub-models, leading to a final layer with a single neuron for predicting the pairing (Fig. 1e). Based on this integrated model, we innovatively employed a differential learning schema, where this model is fed a true binding pair of TCR and pMHC and another negative pair with the same pMHC in each training cycle. We collected a total of 32,607 pairs of binding TCR-pMHCs from a series of peer-reviewed publications[10,17-23] (N=13,388), and four Chromium Single Cell Immune Profiling Solution datasets (N=19,219). The details of these data are shown in Supplementary Information. Some databases provided quality metrics, which we used to filter the records to keep only pairs with high confidence. For example, in the VDJdb data, we only included records with vdj.score>0, as is also done in TCRGP[24]. Duplicated records were removed. We created 10 times more negative pairs, by random mismatching TCR and pMHC of these 32,607 pairs. The training was performed for 150 epochs (Fig. 1f). We named the final model, pMTnet for pMHC-TCR binding prediction network. Following our differential training, the prediction output was also generated in a comparative manner. pMTnet outputs a continuous variable between 0 and 1, reflecting the percentile rank of the predicted binding strength between the TCR and the pMHC, with respect to a pool of 10,000 randomly sampled TCRs (as a background distribution) against the same pMHC. We use a smaller rank to denote a stronger binding, similar to netMHCpan. Importantly, as we always bundle antigen and MHC together and let the model focus on discerning binding or non-binding TCRs, all validations are specific for distinguishing TCR binding specificity, rather than antigen-MHC binding or the overall immunogenicity.

pMTnet predicts TCR-pMHC pairing in independent experimental data

We performed a series of validation analyses with a large number of known TCR-pMHC binding pairs collected from independent studies. First, we collected 619 experimentally validated TCR-pMHC binding pairs (Supplementary Information). Compared with the training cohort, which is mainly constructed from bulk export from databases like VDJdb and high throughput experiments, the binding pairs that comprise the test cohort have mostly been subjected to stringent interrogation by the original reports on an individual basis. In this and all following validation analyses, TCR-pMHC pairs that appeared in the training dataset were removed, so the testing sets were completely independent of the training set. 10 times negative pairs were generated by random mismatching. We used two metrics, Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) and Precision-Recall (PR). Strikingly, the AUC of ROC reached 0.827 in this cohort and AUC of PR reached 0.565 (Fig. 2a). To test whether pMTnet truly "learned" the features that determine binding, or is simply "remembering" pairing cases, we looked at the prediction performance for TCRs with different degrees of similarity to the training TCRs (Fig. 2b, left group). For calculating "similarity", we calculated the minimum of each testing TCR's Euclidean Distances to all the training TCRs based on the TCR embeddings (representative examples shown in Supplementary Information). The AUCs of ROC and PR are shown for the subset of the testing TCRs with minimum distances over each cutoff, and the performance of pMTnet is relatively robust with respect to increasing levels of TCR dissimilarities. For pMHC, we performed the same analyses, and made similar observations (Fig. 2b, right group).

We also compared the performance of pMTnet with other software developed to predict TCR/epitope pairing, including netTCR[25], TCRex[26], and TCRGP[24]. Unlike pMTnet, all three softwares were limited by the type of epitopes/MHCs/TCRs that can be used for prediction. For example, netTCR only accommodates for the HLA-A:0201 allele, epitopes shorter than 10 amino acids, and CDR3s shorter than 10 amino acids. When tested on the same epitopes/MHCs/TCRs that satisfy the criterion of these three software, pMTnet demonstrates a large margin of improvement over each one (Supplementary Information). We also validated pMTnet on additional high quality pairing data from VDJdb and Gee *et al*[27] that are not used during the training (Supplementary Information), and showed that the AUROC of pMTnet achieved >0.8 on them, and out-performed competing software.

Next, we validated the predicted binding between TCRs and pMHCs *via* the expected impact of the binding on the T cells, *i.e.*, T cells with higher pMHC affinity should be more clonally expanded. The 10x Genomics Chromium Single Cell Immune Profiling platform generates single cell 5' libraries and V(D)J enriched libraries in combination with highly multiplexed pMHC multimer reagents. The antigen specificity between the TCR of one T cell and each tested pMHC is profiled by counting the number of barcodes sequenced for that particular pMHC in this cell. We examined four single-cell datasets, which profiled the antigen specificities of 44 pMHCs for CD8+ T cells from four healthy donors. For each TCR clone, we recorded the pMHC with the strongest predicted binding strength, by pMTnet, among all 44 pMHCs. Interestingly, we found the clone sizes and predicted ranks for T cell clonotypes were negatively correlated with statistical significance achieved (Fig. 2c).

In other words, T cells with TCRs whose predicted pMHC binding strengths were stronger were also much more expanded than others without a strong binding partner. This is more clearly demonstrated by the odds ratios (Supplementary Information) testing enrichment of the expanded T cell clonotypes with high affinity binding antigens. Conversely, we observed some TCRs with small clone sizes having small predicted binding ranks to pMHC, which was likely caused by the stochastic nature of binding between TCRs and pMHC, and possibly the constantly incoming new clones whose expansion had not happened yet.

We further analyzed whether pMTnet is capable of distinguishing the impact of the fine details of peptide sequences on TCR binding specificity. We acquired 186 pMHC-TCR pairs from Liu *et al*[28], Cole *et al*[29], and Tran *et al*[30]. In Liu *et al*, LPEP peptide analogs with single amino acid substitutions were tested for specificity towards three distinct TCRs with different CDR3βs. Out of all 94 analogs, 36 were determined to be stronger binders (<100pM of peptide needed to induce cytotoxic lysis by T cell) with the others deemed weaker binders. In Cole's study, alanine-substituted MART-1 peptides were tested for the affinity to TCR MEL5 and ILA1. 15 out of 70 peptides had interactions with TCRs (KD value<500mM). In Tran's study, 11 out of all 22 analog peptides activated T cells validated by IFN-γ ELISPOT. pMTnet generated predictions for each peptide analog (in complex with MHC) and the stronger binding analogs were indeed predicted to have stronger binding strength than their analogs (Fig. 2d, AUC=0.726).

We further validated pMTnet in prospective experimental data. We performed bulk TCR-sequencing and HLA allele typing for one donor seropositive for prior Influenza, EBV and HCMV infections. The experiments were performed in the blood and the *in vitro* expanded T cells from this donor's lung tumor. We analyzed the bulk TCR-sequencing data and predicted the binding between TCRs and four viral pMHCs, including Influenza M (GILGFVFTL), Influenza A (FMYSDFHFI), EBV BMLF1 (GLCTLVAML), and HCMV pp65 (NLVPMVATV). We found that TCRs predicted to have stronger binding (smaller ranks) to any of these peptides exhibited higher clonal proportions than the other TCRs (Fig. 3a), in both the blood (left panel) and *in vitro* expanded T cells (right panel). We calculated the odds ratios for the enrichment of highly expanded TCRs with stronger predicted binding, where a higher odds ratio referred to a higher positive enrichment. We observed a stronger enrichment in both the blood and expanded T cells, while we performed permutations of the predicted binding ranks and observed much smaller odds ratios (Fig. 3b). Then we treated the expanded T cells with each of the viral peptides and performed scRNA-seq with paired TCR-seq, and we also performed vehicle treatment. We identified TCRs captured in each of the treatment groups and the vehicle treatment group, and used pMTnet to predict the binding of the TCRs to each peptide. We selected the top TCRs (predicted rank<2% by pMTnet) from each experiment, and first examined the gene expression of the T cells of these top binding TCR clonotypes. By comparing T cells with predicted top binding TCRs and the other T cells, we observed differentially expressed genes enriched in pathways essential for T cell proliferation, migration, survival, and cytotoxicity (results for GLCTLVAML shown in Fig. 3c as an example). We also calculated the clonal sizes of these top TCR clonotypes, and found that the majority of these TCR clonotypes exhibited larger clonal fractions in the treatment group than the vehicle group (Fig. 3d, clonal size ratio >1).

Structural analyses support predicted TCR-pMHC interactions

We performed *in silico* mutational analyses to look for structural evidence for the CDR3 residues whose mutations led to dramatic changes in the predicted binding between TCR and pMHCs. For each CDR3 residue, we mutated its numeric embedding to a vector of all 0s (“0-setting”). This is similar to but different from the alanine scanning technique in biophysics studies[31]. We first performed residue-wise mutations for all the TCRs of the 619 testing cohort, and recorded the differences in the predicted binding ranks (rank difference) between the wild type TCRs and the mutated TCRs. We divided each TCR CDR3 into six segments of equal lengths (Fig. 4a), and as expected, residues in the middle segments of CDR3s, which bulge out and are in closer contact with pMHCs, were more likely to induce larger changes in predicted binding affinity, when compared with the outer segments (T-test P-value between the third or fourth segment and any other segment is <0.00001). Furthermore, we extracted a total of 13 TCR-pMHC pairs from the IEDB cohort (Supplementary Information), with 3D crystal structures available in Protein Data Bank (PDB) and whose predicted binding affinity rank was less than 2%. According to the structures, we grouped CDR3 residues by whether or not they formed any direct contact with pMHCs residues within 4Å. We found that the contact residues were more likely to induce larger changes in predicted pMHC binding strength than non-contact residues (Fig. 4b, P value=0.036). We also performed *in silico* alanine scanning and found a similar trend (Fig. 4c). The alanine scan was not as significant as for the “0-setting” scan, which could be attributed to the fact that, in the alanine scan, all alanines are presumed to have no effect after mutation (alanine->alanine). However, replacing one alanine with other residues with large side chains could affect the overall structural integrity of the protein complex, which may actually lead to a change in binding affinity. In Fig. 4a-c, we showed the absolute changes in rank percentiles (change to either stronger or weaker binding). But examination of the direction of the changes in rank percentiles showed that the *in silico* mutations mainly resulted in weaker binding.

In Fig. 4de, we showed an example TCR-pMHC structure with the PDB id of 5hhm, generated by Valkenburg *et al*[32]. Overall, we found that R98 and S99 had the biggest differences in predicted ranks after the “0-setting” scan (Fig. 4d, upper panel) and alanine scan (Fig. 4d, lower panel), which were the residues located in the middle of the CDR3 and had the most contacts with pMHC. The other two amino acids with relatively high rank changes could be explained by their crucial role in formation and stabilization of the CDR3 loop. We observed that S95 formed intra-chain contacts with the small loop formed by Q103 and the side chains of E102 and Y104.

Characterizing the TCR-pMHC interactions in human tumors

To further validate pMTnet and demonstrate the value of pMTnet as a knowledge discovery tool, we characterized the TCR and pMHC interactions in several of the most immunogenic tumor types, where the tumor antigen presentation machinery is more likely to be active[33]. We analyzed the genomics data of The Cancer Genome Atlas (TCGA) and the in-house Renal Cell Carcinoma (RCC) data from our prior publication[33]. TCGA patients included lung adenocarcinoma patients (LUAD)[34], lung squamous cell carcinoma patients (LUSC)[35], clear cell renal cell carcinoma patients (KIRC)[36] and melanoma patients (SKCM)[37].

We investigated several classes of antigens that could affect T cell populations in the tumor microenvironment. The first class of antigens that could affect T cell retention and expansion is tumor neoantigens. The other class of antigens is tumor self-antigens (also referred to as tumor associated antigens, TAAs), such as CAIX[38]. In kidney cancer, in particular, Cherkasova *et al* discovered the re-activation of a special class of self-antigens, HERV-E retrovirus, which encodes several immunogenic peptides that have been experimentally validated[39]. The rest T cell infiltration may be explained by prior virus infection, or may simply be bystanders. The field has been debating for a long time which of these factors is most potent in shaping the landscape of the T cell repertoire in tumors. To answer this question, we identified candidate neoantigens and self-antigens from the genomic data (Materials and Methods). For RCCs, we profiled the expression of this very well characterized HERV-E found by Cherkasova *et al* (Materials and Methods). This pipeline also examined all other HERVs, but Extended Data Figure 2 shows that the tumor-over-normal expression ratios of the other top HERVs were much smaller than that of HERV-E, and reports are lacking regarding whether these HERVs encode immunogenic peptides like this HERV-E does. In each patient sample, we assigned each TCR to one of the antigens (neoantigen and self-antigens) with the lowest predicted binding ranking, and also satisfying the criterion that this binding rank has to be lower than each one of a series of cutoffs between 0.00% and 2% (otherwise, this TCR will be unassigned).

For each patient sample, we calculated the percentages of neoantigens or self-antigens predicted to bind at least one TCR (defined as immunogenic antigen) for each class of antigens. Fig. 5a shows the total and immunogenic antigen numbers for one example RCC patient. Then for all patients of all cancer types, we calculated the proportion of immunogenic antigens for neoantigen, self-antigen (excluding HERV-E), and HERV-E (kidney cancer only) for each patient, and averaged them across all patients. We observed that neoantigens were generally more immunogenic than self-antigens (higher proportions of neoantigens are predicted to bind TCRs) (Fig. 5b). This is fitting because neoantigens, unlike self-antigens, are mutated peptides that have not been encountered by T cells during the developmental process. However, we observed that HERV-E antigens were more likely to be immunogenic than both neoantigens and the other self-antigens in RCCs, confirming prior reports on the importance of HERV-E in inducing immune responses in kidney cancers[39].

Next we examined the impact of TCR-pMHC interactions on the clonal expansion of T cells. For each patient, we compared the clonal fractions of TCRs (#specific TCR clonotype/#all TCRs) that were predicted to be binding to any of the neoantigens and self-antigens, and also the clonal fractions of the other non-binding T cells. In an example patient (Fig. 5c), we showed the average clonal fraction of TCRs that can bind or that cannot bind to any antigen in this patient (1% binding rank cutoff). This patient's binding T cells had a higher average clonal fraction than non-binding T cells. For each of the four cancer types, we calculated the number of patients with binding T cells having a higher average clone fraction, divided by the number of patients with non-binding T cells having a higher average clone fraction. Strikingly, we observed that more and more patients demonstrated clonal expansion of their antigen-targeting T cells compared to other T cells (Fig. 5d), with smaller and smaller rank percentile cutoffs (stronger affinity) to define antigen-TCR pairing. Consistent with Fig. 2c and Fig. 3, this result also shows that more immunogenic tumor antigens induce stronger T cell clonal expansion in human tumors.

Finally, we tested the TCR binding affinity of neoantigens generated by missense mutations and frameshift mutations. Frameshift mutations usually generate epitopes that are completely new and not similar to any epitope from the normal human proteome, while missense neoantigens differ from the normal epitopes by one mismatch. Therefore, frameshift neoantigens are likely more potent in inducing strongly reactive T cells/TCRs. Indeed, the neoantigens generated by frameshift mutations exhibited significantly stronger binding to TCRs (average rank=0.81%) than neoantigens generated by missense mutations (average rank=0.92%) (P=8.1E-9).

TCR-neoantigen interactions impact tumor progression and immunotherapy treatment response

We evaluated whether the TCR-pMHC interactions profiled by pMTnet are physiologically important. We focused on tumor neoantigens, as they are associated with somatic mutations, which can be directly linked to tumor clone fitness. In a given tumor, some neoantigens may bind TCRs of T cells that are more clonally expanded while others may bind T cells that are less expanded. Conversely, some neoantigens arise from truncal mutations (higher variant allele frequency), while others arise from subclonal mutations. When truncal neoantigens are bound by clonally-expanded TCRs, the distribution of neoantigens and T cells may favor the elimination of tumor cells, which could be beneficial for prognosis and immunotherapy treatment response[42,43]. To quantitatively measure this effect, we developed a neoantigen immunogenicity effectiveness score (NIES), which is based on the product of the variant allele frequency (VAF) of the neoantigen's corresponding mutation and the clonal fraction of the TCRs that bind the same neoantigen (details in Supplementary Information). Proper normalizations were carried out to remove the confounding effect of tumor purity and total T cell infiltration. The higher the NIES score, the more expanded TCRs are specific against truncal neoantigens, which is generally favorable for clinical outcomes[42,43].

We examined the association between NIES and prognosis in the LUAD, LUSC, SKCM, and RCC cohorts. We first focused on patients with high levels of total T cell infiltration. We speculated that the neoantigen-T cell axis is more likely to be functionally active when there is sufficient T cell infiltration. Interestingly, in lung cancer and melanoma patients, higher NIES scores were associated with better survival (Fig. 6a, LUAD, P=1.74E-3; Fig. 6b, LUSC, P=0.0238; Fig. 6c, SKCM, P=6.65E-4). In comparison, NIES was not prognostic in kidney cancer (Fig. 6d). For all four cohorts, overall survival of patients with low T cell infiltration was unrelated to the levels of NIES, further supporting our hypothesis. Interestingly, the difference between kidney cancer and other cancer types may have reflected the unique features of kidney cancers such as low mutational load and HERV-E reactivation. We next combined lung cancer and melanoma patients with high T cell infiltration, and the survival analysis of this integrated cohort revealed that patients with higher NIES had a better overall prognosis (P=1.12E-6, Fig. 6e). Multivariate analysis was performed adjusting for disease type, stage, gender, age, and TCR repertoire diversity[44] in the combined cohort and the association between survival rate and NIES still held

(P<0.001, Fig. 6f). Analyses shown in Fig. 6a-f were carried out using a binding ranking cutoff of 1%. Using a series of different cutoffs, we obtained similar results (Fig. 6g). As a benchmark, we dichotomized patients by the median neoantigen load, T cell infiltration, or TCR diversity and performed the same analyses. We observed that NIES was much more strongly prognostic than other candidate biomarkers (Fig. 6g).

Similarly, we evaluated the implication of TCR-neoantigen interaction efficiency for treatment response prediction. We analyzed a total of 139 melanoma patients on immune checkpoint inhibitor treatment from Liu *et al*[45], Van Allen *et al*[46] and Hugo *et al*[47]. Patients were divided into two groups based on the median NIES and we demonstrated that patients with higher NIES had better overall survival (Extended Data Figure 3a, P=2.44E-3, binding affinity cutoff at 1%). We also analyzed a cohort of anti-PD-L1-treated metastatic gastric cancer patients, of which more than one-third were found to harbor high mutation loads[48]. For this cohort, survival information was unavailable so we analyzed RECIST responses, and found that patients with better responses had higher NIES (Extended Data Figure 3b, P=9.9E-3). Results of other binding rank cutoffs are shown in Extended Data Figure 4ab. In comparison, we analyzed a cohort of ccRCC patients on anti-PD1/anti-PD-L1 from Miao *et al*[49]. As expected, no significant association was observed between NIES and survival rate (Extended Data Figure 4c). NIES was then benchmarked against total neoantigen load, T cell infiltration, and TCR repertoire diversity and demonstrated an advantage over these three other biomarkers (Extended Data Figure 5, 1% cutoff). To systematically assess the significance of these comparisons, we leveraged the bootstrap technique and confirmed that the advances were statistically significant (Extended Data Figure 3c).

DISCUSSION

Our work enabled prediction of the TCR-binding specificity of class I pMHCs, just given the TCR sequence, (neo)antigen sequence, and MHC type, which has not been achieved before to our knowledge. This is enabled by several innovative algorithmic designs, including transfer learning to take advantage of a large amount of related TCR and pMHC data without pairing information, and the differential training paradigm that allows pMTnet to focus on differentiating binding *vs.* non-binding TCRs. Although TCRs directly interact with the epitopes, MHC proteins restrict the spatial locations of the anchor positions of the epitopes, which further limits the possible conformations of the epitopes and influences their interactions with TCRs. This led us to incorporate MHC protein sequences in pMTnet. In our work, we showed that pMTnet significantly outperforms competing software such as netTCR. Other methods such as GLIPH[18] and TCRdist[50] were developed to group TCR sequences profiled in a given sample into clusters, with each cluster of TCRs assumed to be specific to a single epitope. However, such methods still cannot pinpoint the exact sequence of neoantigens or antigens without prior knowledge.

Furthermore, a suite of genome-wide analyses was now enabled by pMTnet, which has revealed interesting biological discoveries. Our work provided a large scale and unbiased estimate of the immunogenicity potential of neoantigens and self-antigens (including HERV-E). Recently, Gee *et al* carried out yeast-display screening in two HLA-A*02:01 homozygous patients with colorectal adenocarcinoma and identified four TCRs and their peptide targets[27]. Surprisingly, three of the four receptors recognized unmutated self-antigens. Consistent with the observations of Gee *et al* in a limited number of patients, we confirmed in several large cohorts that self-antigens do have immunogenic potential, though neoantigens are still more likely to be immunogenic. But HERVs, a special class of self antigens in kidney cancer, seems to be more immunogenic than neoantigens.

Our work demonstrated the potential of leveraging pMTnet to enhance the care of cancer patients, such as generating prognostic tools and predictive tools for immunotherapy treatment response. Yost *el al*[51] discovered dramatic clonal replacement of T cells in cancer patients after anti-PD-1 therapy. pMTnet could make it more feasible, both in terms of time and cost, to closely monitor the patients' TCR repertoire after immunotherapy treatment, and to achieve the most informative treatment decisions in real-time. pMTnet could also be used for designing TCR-T or neoantigen vaccine therapies, where pMTnet can generate a narrowed down list of candidate TCRs or neoantigens for engineering. We showed that NIES is prognostic and predictive for checkpoint inhibitor treatment, though not in kidney cancer, perhaps due to the re-activation of HERVs and its low mutational load.

One caveat of the current study is the potential problem caused by the biased representation of certain epitopes and their clonally expanded pairing TCRs in our training dataset. Admittedly, our training dataset collection has many common epitopes such as those well studied ones from CMV. In the future, we expect more training TCR-pMHC pairing data to be accumulated by the field, especially given the advent of high-throughput technologies such as T-scan and 10X Immune Profiling. These data will more accurately represent the whole space of possible epitopes for training pMTnet, and will be powerful for helping move the field forward.

Overall, we proved that the pairing between TCRs and pMHCs, just given the TCR, the antigen, and the MHC sequences, is "machine learnable", which sets a foundation for future studies based on our work. We expect pMTnet to propel tumor immunogenomics research and also to enhance the design and implementation of immunotherapy in the modern era of personalized medicine.

METHODS

Embedding TCR CDR3β sequences

We encoded the TCR CDR3β sequences by the "Atchley factor"[14], which represents each amino acid with 5 numeric values. These 5 values can comprehensively characterize the biochemical properties of each amino acid. The resulting numeric matrix has the number of rows being the number of Atchley factors and the number of columns being 80. Then the "Atchley matrices" of TCR sequences were fed into a stacked auto-encoder, which is a powerful algorithm capable of learning sophisticated signals in an unsupervised manner. Atchley matrices of TCR sequences are input into a 2D convolutional layer with 30 5x2 kernels and activated with the 'SELU' function, followed by a batch normalization layer and a 2D average pooling layer with 4x1 kernels. The pooling layer is followed by another 2D convolutional layer with 20 4x2 kernels, and the same batch normalization layer and a 2D average pooling layer as previously described. After pooling, the matrices are converted into a flattened layer, followed by a 30-neuron dense layer activated with the 'SELU' function, and a dropout layer with a dropout rate 0.01, and another 30-neuron dense layer activated with the 'SELU' function, which is the 'bottleneck' layer of the auto-encoder model. Layers before the bottleneck layer are reversed to create the decoder part of the model. The input of the encoder and output of decoder are exactly the same – the Atchley matrices. The training process instructed the auto-encoder to reconstruct the input data and capture their inherent structure using a simple numeric vector. After training is finished, the smallest fully connected layer in the middle of the auto-encoder (bottleneck) forms a 30-neuron numeric vector embedding of the original CDR3s. The TCR CDR3β sequences were padded to 80 amino acids long for several reasons. We leave room here for potentially adding CDR3 of the α chains in the future. CDR1 and CDR2 may also be added. This could be convenient as the structure of the auto-encoder does not need to be changed, or just needs to be minimally changed, even when the other CDRs are added.

Embedding pMHCs

The embedding of pMHCs mostly follows the netMHCpan algorithm. The netMHCpan algorithm uses a pseudo sequence method to encode the MHC proteins[52]. The pseudo-sequences consist of amino acids in contact with the peptide and only 34 polymorphic residues were included. Then the BLOSUM50 matrix is used to encode these 34 residues. On the other hand, the (neo)antigens were also encoded by the BLOSUM50 matrix as in netMHCpan. We constructed a deep learning model with the HLA pseudo sequence and the antigen sequence as the input. Here, we used the MHC sequence rather than type as the input, so the use can be extended to unknown MHC types not seen in the training cohort. The major difference of our implementation from the original netMHCpan model is that, instead of simple feed-forward neural networks, we used a Long short-term memory (LSTM) layer with the output size of 16 on top of the antigen input, and an LSTM layer with the output size of 16 on top of the MHC input. We found this change to seem to have increased the speed of reaching model convergence on our hand. The LSTM outputs for antigen and MHC are concatenated to form a 32-dimensional vector in the same layer. This layer is followed by a dense layer with 60 neurons activated by "tanh" and a-single-neuron dense layer as the last output layer. We trained this network with the exact data that were used to train the netMHCpan model. After training is completed, we extracted the immediate 60-dimensional fully connected layer before the single-neuron output layer (again a short numeric vector), as the embedding of pMHCs.

Learning TCR binding specificity of pMHCs

We employed transfer learning to leverage the trained numeric encodings of TCRs and pMHCs. These pre-trained models were fixed and incorporated into the final prediction model as early layers (save parameters needed for training). The two encodings both yield the final output layers in the form of numeric vectors. We concatenated the two numerical vectors into a single layer, added a dense layer with 300 neurons activated by "RELU", a dropout layer with dropout rate of 0.2, a dense layer with 200 neurons activated by "RELU", a dense layer with 100 neurons activated by "RELU", and the last layer with a single neuron with tanh activation. Mathematically, the output prediction for a given pMHC, p^*, towards a given TCR, T^*, can be written as $f(p^*,T^*)$. For the training process, known interactions between pMHCs and TCRs were treated as positive data. And we randomly mismatched these TCRs and pMHCs to create 10 times more negative data.

Differential loss function

Rather than directly learning the positive and negative labels of the training data, we developed a novel differential training method to instruct pMTnet to distinguish binding TCRs from non-binding TCRs through comparison. To implement this, we created two duplicates of the above-described networks, always sharing weights throughout the training process. During one training step, one positive (known interaction) training point (p,T^+) is fed into the first network, and a negative training point (p,T^-) is fed into the second network. A loss function of

$$Loss = Relu(f(p,T^-) - f(p,T^+)) + 0.03[f^2(p,T^-) + f^2(p,T^+)]$$

is defined. In other words, the learning process focuses on the same pMHC each time and tries to identify the TCRs that truly bind to it, out of other TCRs. The second item in the loss function serves the purpose of a regularization term to reduce overfitting and also to push the output of the network to be closer to 0. This helps make sure the model parameters stay in a dynamic range where gradients are neither too small nor too large.

In accordance with this differential training method, the output of pMTnet is also not the direct output of the deep learning network. In fact, for each pMHC, $p*$, we sample 10,000 TCR sequences randomly from our databases to form a background distribution, $\{T^b\}$. We will calculate the percentile of $f(p*,T*)$ in the whole distribution of $\{f(p*,T^b)\}$, where $T*$ is the TCR of interest. The larger this value, the stronger we predict the binding is between $p*$ and $T*$. In line with how netMHCpan generates the ranked prediction of the binding strength between antigens and HLA proteins (percentile_rank), we also inverted this rank. Therefore, in our final output, a smaller rank between a pMHC and a TCR refers to a stronger binding prediction between them.

Defining self-antigens

To detect self-antigens in tumor samples, we focused on genes that are lowly expressed (<0.01 RPKM) in all normal tissue types according to the GTEx project (https://www.gtexportal.org/home/datasets), and are expressed at >1RPKM in each tumor sample. There are a total of 52 normal tissue types collected by GTEx. We translated the protein sequences from such genes and also used the netMHCpan to detect 8-11 mers that will bind to the same patient's class I HLA alleles.

To detect HERV expression levels in bulk RNA sequencing data from patient samples, we built a pipeline named 'HERVranger', available at: https://github.com/jcao89757/HERVranger. Description is provided at this link regarding how it detects HERV expression from genomic sequencing data.

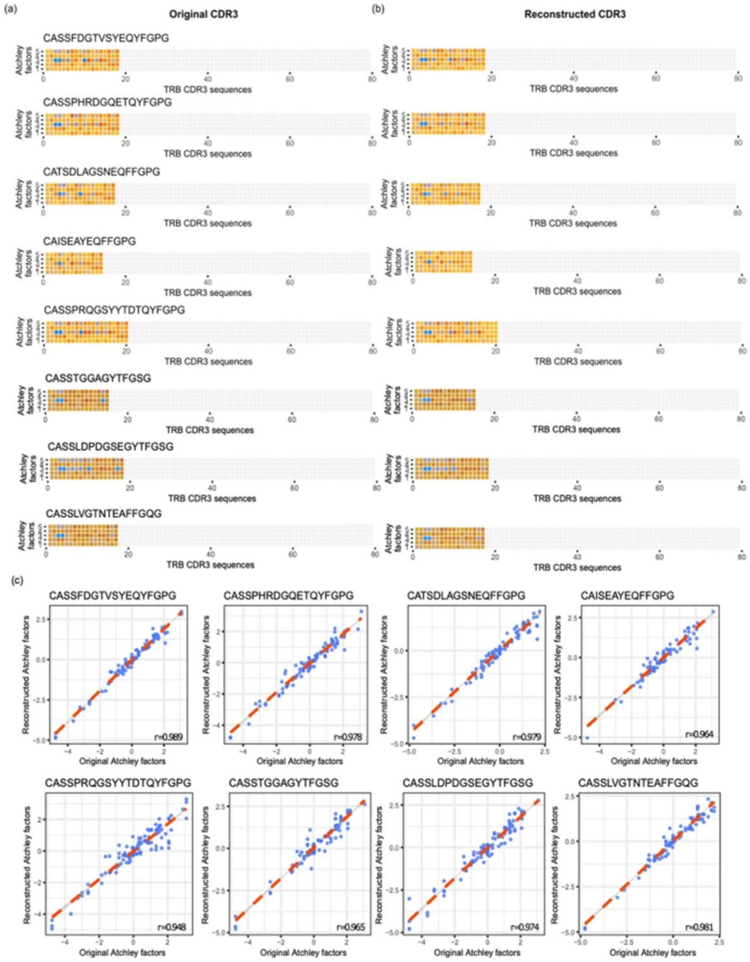
Generation of in-house validation data

The donor of this study was enrolled in the MDACC Immunogenomic Profiling of Non-Small Cell Lung Cancer (ICON) project. The ICON study was approved by MDACC’s institutional review board, and patients were consented before enrollment. This donor expressed HLA-A*02. We extracted the peripheral blood and tumor infiltrating T cell specimens from this donor. Bulk sequencing of the CDR3 regions of human TCRβ chains was performed using the immunoSEQ Assay (Adaptive Biotechnologies, Seattle, WA) following the manufacturer's protocol. The expanded T cells (IL2 plus anti-CD3 treatment) from the tumor were seeded on 96-well plates at the density of 100,000 cells per well. The T cells were co-cultured with 10ug/ml individual HLA-matched viral peptides, including Influenza M (GILGFVFTL), Influenza A (FMYSDFHFI), EBV BMLF1 (GLCTLVAML), and HCMV pp65 (NLVPMVATV) (GenScript, Piscataway, NJ, USA) or AIM-V medium alone overnight. The T cells were harvested and submitted for library preparation of 10X Single Cell 5’ Gene Expression and Immune Profiling at MedGenome (Foster City, CA, USA). Sequencing was conducted with an Illumina NovaSeq6000 with 150-bp paired-end reads (Illumina, San Diego, CA, USA) following the manufacturer’s protocol. The 10X sequencing data was analyzed using the 10X CellRanger software.

Statistical analyses

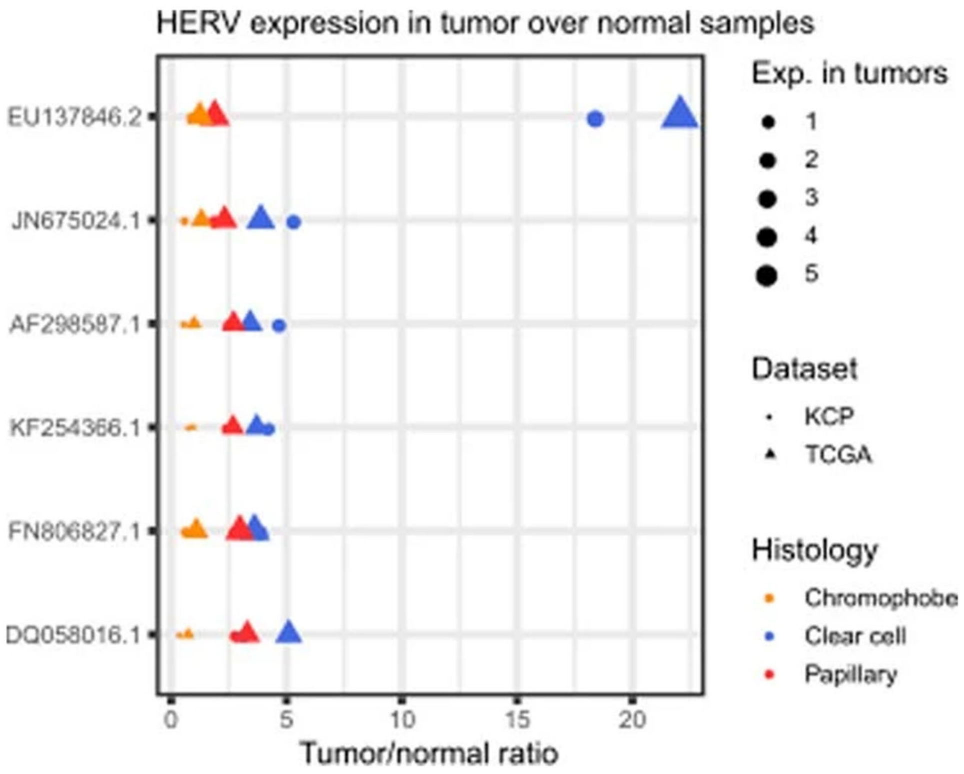
All computations and statistical analyses were carried out in the R computing environment. All P values are two-way unless otherwise noted. The AUC of ROC and Precision-Recall was calculated by the ‘*RPPROC*’ R package. The GOrilla webserver was used to detect enriched gene ontology pathways from single-cell expression analysis[71]. For analysis of neoantigens in the tumor, we used the QBRC mutation calling and neoantigen calling pipelines[43]. For correlating NIES with survival, the NIES scores were split on the median within each patient cohort, and we employed the log-rank test for evaluating whether patients with higher NIES scores had better survival. For the metastatic gastric cancer cohort, we employed the ordinal Jonckheere test for investigating whether there is an overall trend of patients with better responses (CR->PR->SD->PD) having higher NIES scores. The same criterion was applied for neoantigen load, T cell infiltration, and TCR diversity. T cell infiltration was profiled by our recently published eTME gene signature[33] using the ssGSEA method[72]. For model comparison, 5,000 bootstrap resamples of the original cohorts were generated, and each resample was used to evaluate the performance of the NIES scores (or neoantigen load or T cell infiltration or TCR diversity). The P values of 5,000 bootstraps of each approach were compared using the two-sided Wilcoxon signed-rank test.

Extended Data



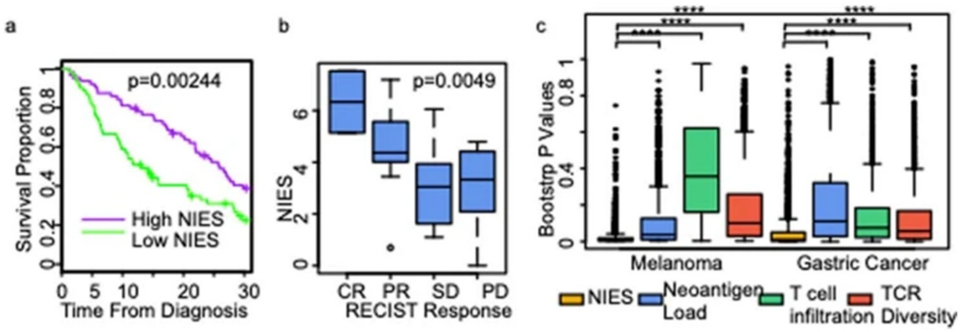
Extended Data Figure 1.

More examples showing the successful embedding of TCRs by the auto-encoder. (a) Heatmaps of the original TCR CDR3β sequences, embedded by the “Atchley factors” and all padded with zeros to the length of 80 amino acids. (b) Heatmaps of the re-constructed TCR CDR3β sequences for the same TCRs. (c) Scatterplots showing the consistency between ‘Atchley factor’ values of the original and re-constructed TCRs. Blue points represent tiles in the heatmaps in (a) and (b). The red dashed lines are for y=x.



Extended Data Figure 2.

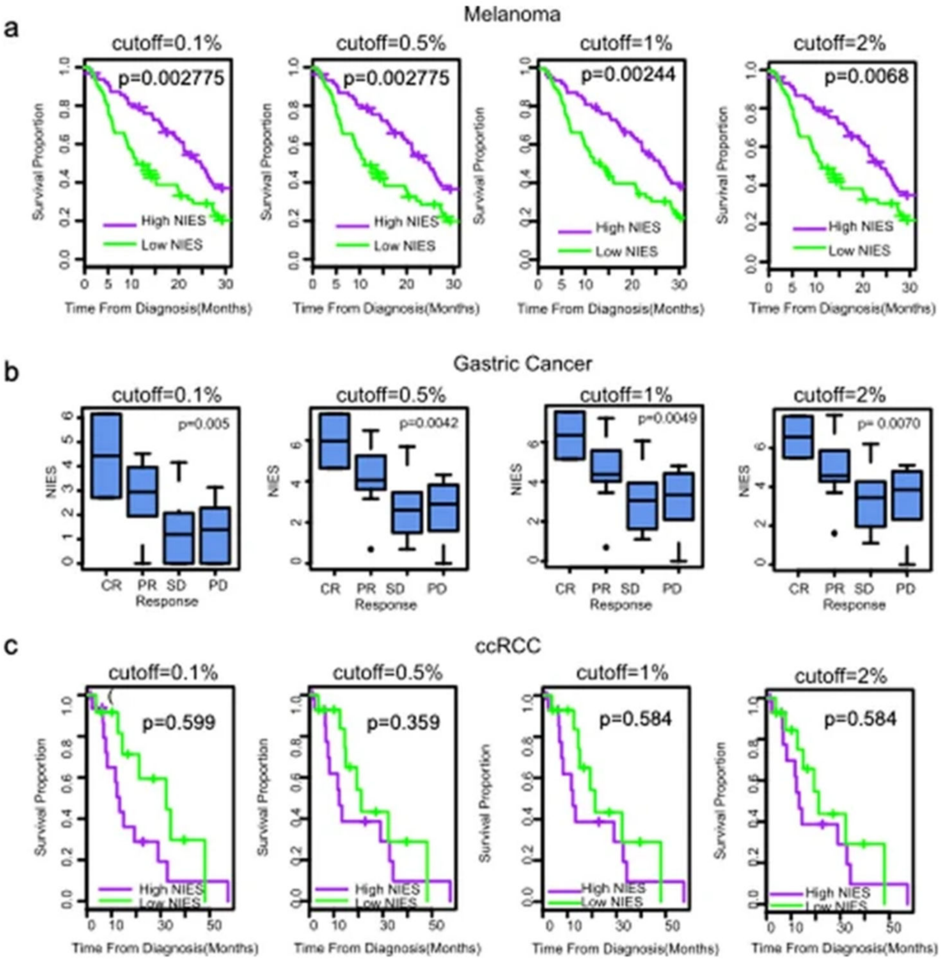
Differential analysis of the expression levels of HERVs between tumor samples and normal samples in different RCC cancer types and data cohorts. In addition to EU137846.2 (the known HERV-E), the HERVs whose tumor-over-normal expression ratio is >3 in any of the type/cohort, and whose normal tissue expression is <3 are also shown. There are five such HERVs.



Extended Data Figure 3.

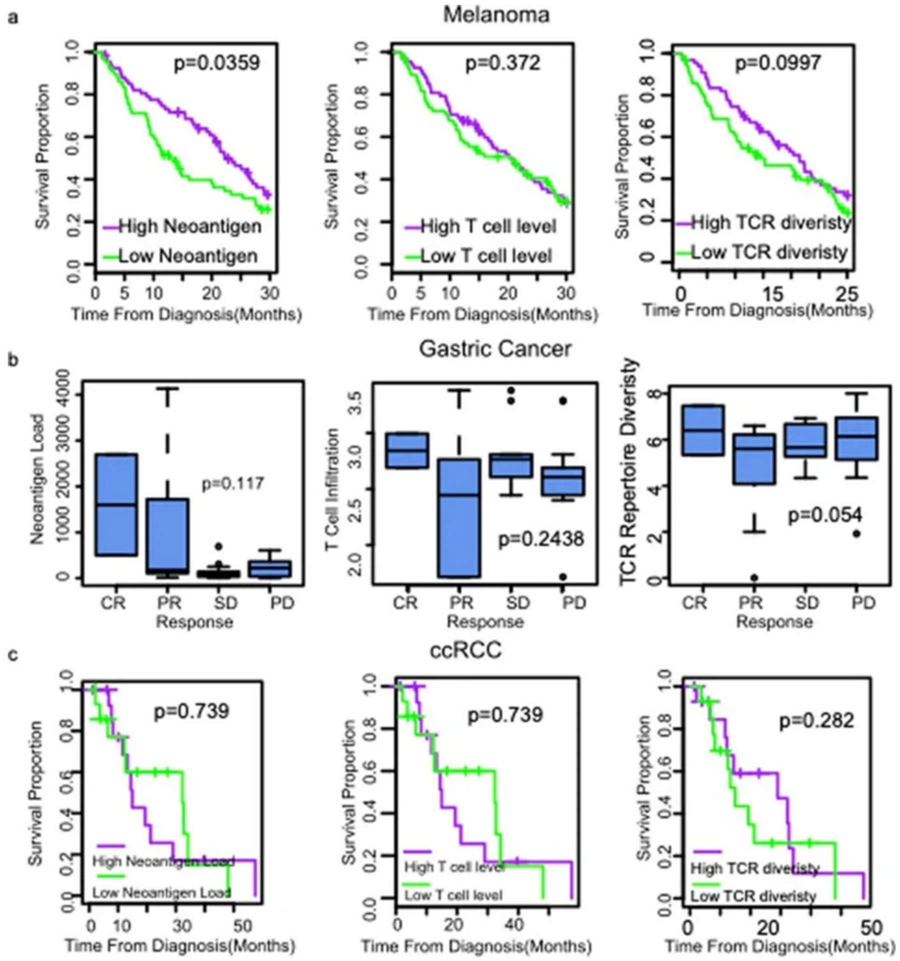
Efficiencies of TCR-neoantigen interactions impact response to immunotherapies. (a) Association between NIES and overall survival of melanoma patients on immunotherapies. The patients were split by the median of NIES in each cohort and then combined. The P-value for the log-rank test is shown. (b) Association between NIES and the response of metastatic gastric cancer patients. The overall survival or progression-free survival data are not made available from the original publication, so we used the RECIST response variables.

Complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). There are 40 gastric cancer patients. An ordinal Jonckheere test is employed to investigate whether patients with better response to immunotherapies also have higher NIES scores. In this test, all categories are compared together to investigate whether an overall trend exists across all categories. (c) Boxplots of bootstrap P values evaluating the robustness of comparison between NIES, neoantigen load, T cell infiltration level, and TCR diversity. One P-value is generated from one bootstrap resample of each cohort, and the two-sided Wilcoxon signed-rank test was carried out for the bootstrap P values to assess whether differences are significant between different biomarkers. NS: P>0.01, *: P=0.01-0.05, **: P=0.001-0.01, ***: P=0.0001-0.001, ****:P<0.0001. For boxplots in (b) and (c), box boundaries represent interquartile ranges, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median.



Extended Data Figure 4.

Association of NIES with treatment response of (a) melanoma, (b) metastatic gastric cancer, and (c) kidney cancer patients on checkpoint-inhibitor treatment. There are 33 kidney cancer patients from the Miao cohort. The same analyses as in Extended Data Figure 3 were carried out, except that the binding affinity cutoffs for assigning TCRs to neoantigens were varied at several possible values.



Extended Data Figure 5.

Association of neoantigen load, T cell infiltration level, and TCR repertoire diversity with treatment response of (a) melanoma, (b) metastatic gastric cancer, and (c) kidney cancer patients on checkpoint-inhibitor treatment. The same analyses as in Extended Data Figure 3 were carried out for these biomarkers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 10/01/19. We acknowledge Drs. Daifei Liu, Bo Li and Jared Ostmeyer from UT Southwestern for their helpful advice on our project. We acknowledge the authors of the phs000452.v3.p1 and phs001493.v1.p1 datasets, as well as the funding agencies that supported these studies and dbGaP that supported the archiving of these datasets.

Funding

This study was supported by the National Institutes of Health (NIH) [CCSG 5P30CA142543/TW, and R01CA258584/TW], Cancer Prevention Research Institute of Texas [CPRIT RP190208/TW], University of Texas MD Anderson Cancer Center [Lung Cancer Moon Shot/AR], the University Cancer Foundation at the University of Texas MD Anderson Cancer Center [Institutional Research Grant/AR], the Waun Ki Hong Lung Cancer Research Fund [AR], Exon 20 Group [AR], and Rexanna's Foundation for Fighting Lung Cancer [AR].

Data availability

The details of data used for the training and validation of pMTnet, including sample size and role in the machine learning process, are shown in Supplementary Information. The training and testing datasets are shared on our github repository: https://github.com/tianshilu/pMTnet. The processed TCR-seq and scRNA-seq data generated from the in-house patient donor are archived on the https://github.com/tianshilu/pMTnet link as well. The raw scRNA-seq plus TCR-seq data have been archived on NIH GEO with the accession number of GSE173165.

For the NIES analyses, the public patient sequencing datasets are from TCGA, Liu *et al*[45], Van Allen *et al*[46] and Hugo *et al*[47]. The raw RNA-Seq and exome-seq data of the in-house IL2 cohort patients can be downloaded from the European Genome Phenome Archive with accession number EGAS00001003605 through controlled access.

References

1. Dunn GP, Old LJ & Schreiber RD The three Es of cancer immunoediting. Annu. Rev. Immunol 22, 329–360 (2004). [PubMed: 15032581]
2. Ascierto PA & Marincola FM 2015: The Year of Anti-PD-1/PD-L1s Against Melanoma and Beyond. EBioMedicine 2, 92–93 (2015). [PubMed: 26137543]
3. Anagnostou V et al. Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. Cancer Discov. 7, 264–276 (2017). [PubMed: 28031159]
4. Reck M et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. N. Engl. J. Med 375, 1823–1833 (2016). [PubMed: 27718847]
5. Rizvi NA et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348, 124–128 (2015). [PubMed: 25765070]
6. Schumacher TN & Schreiber RD Neoantigens in cancer immunotherapy. Science 348, 69–74 (2015). [PubMed: 25838375]
7. Linette GP & Carreno BM Neoantigen vaccines pass the immunogenicity test. Trends Mol. Med 23, 869–871 (2017). [PubMed: 28867556]
8. Verdegaal EME et al. Neoantigen landscape dynamics during human melanoma-T cell interactions. Nature 536, 91–95 (2016). [PubMed: 27350335]
9. Altman JD et al. Phenotypic analysis of antigen-specific T lymphocytes. Science 274, 94–96 (1996). [PubMed: 8810254]
10. Zhang S-Q et al. High-throughput determination of the antigen specificities of T cell receptors in single cells. Nat. Biotechnol (2018) doi:10.1038/nbt.4282.
11. Kula T et al. T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. Cell 178, 1016–1028.e13 (2019). [PubMed: 31398327]
12. Ito A et al. Cancer neoantigens: A promising source of immunogens for cancer immunotherapy. J. Clin. Cell. Immunol 06, (2015).
13. Hou X et al. Analysis of the Repertoire Features of TCR Beta Chain CDR3 in Human by High-Throughput Sequencing. Cell. Physiol. Biochem 39, 651–667 (2016). [PubMed: 27442436]
14. Atchley WR, Zhao J, Fernandes AD & Drüke T Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102, 6395–6400 (2005). [PubMed: 15851683]

15. Zhang Z, Xiong D, Wang X, Liu H & Wang T Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat. Methods 18, 92–99 (2021). [PubMed: 33408405]
16. Nielsen M & Andreatta M NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. 8, 33 (2016). [PubMed: 27029192]
17. Tickotsky N, Sagiv T, Prilusky J, Shifrut E & Friedman N McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 33, 2924–2929 (2017). [PubMed: 28481982]
18. Glanville J et al. Identifying specificity groups in the T cell receptor repertoire. Nature 547, 94–98 (2017). [PubMed: 28636589]
19. Huth A, Liang X, Krebs S, Blum H & Moosmann A Antigen-Specific TCR Signatures of Cytomegalovirus Infection. J. Immunol 202, 979–990 (2019). [PubMed: 30587531]
20. Chen G et al. Sequence and structural analyses reveal distinct and highly diverse human CD8+ TCR repertoires to immunodominant viral antigens. Cell Rep. 19, 569–583 (2017). [PubMed: 28423320]
21. Joglekar AV et al. T cell antigen discovery via signaling and antigen-presenting bifunctional receptors. Nat. Methods 16, 191–198 (2019). [PubMed: 30700902]
22. Bagaev DV et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res. 48, D1057–D1062 (2020). [PubMed: 31588507]
23. Zhang W et al. PIRD: pan immune repertoire database. Bioinformatics 36, 897–903 (2020). [PubMed: 31373607]
24. Jokinen E, Heinonen M, Huuhtanen J, Mustjoki S & Lähdesmäki H TCRGP: Determining epitope specificity of T cell receptors. BioRxiv (2019) doi:10.1101/542332.
25. Jurtz VI et al. NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. BioRxiv (2018) doi:10.1101/433706.
26. Gielis S et al. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. Front. Immunol 10, 2820 (2019). [PubMed: 31849987]
27. Gee MH et al. Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. Cell 172, 549–563.e16 (2018). [PubMed: 29275860]
28. Liu YC et al. Highly divergent T-cell receptor binding modes underlie specific recognition of a bulged viral peptide bound to a human leukocyte antigen class I molecule. J. Biol. Chem 288, 15442–15454 (2013). [PubMed: 23569211]
29. Cole DK et al. T-cell receptor (TCR)-peptide specificity overrides affinity-enhancing TCR-major histocompatibility complex interactions. J. Biol. Chem 289, 628–638 (2014). [PubMed: 24196962]
30. Tran E et al. Immunogenicity of somatic mutations in human gastrointestinal cancers. Science 350, 1387–1390 (2015). [PubMed: 26516200]
31. Weiss GA, Watanabe CK, Zhong A, Goddard A & Sidhu SS Rapid mapping of protein functional epitopes by combinatorial alanine scanning. Proc Natl Acad Sci USA 97, 8950–8954 (2000). [PubMed: 10908667]
32. Valkenburg SA et al. Molecular basis for universal HLA-A*0201-restricted CD8+ T-cell immunity against influenza viruses. Proc Natl Acad Sci USA 113, 4440–4445 (2016). [PubMed: 27036003]
33. Wang T et al. An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. Cancer Discov. 8, 1142–1155 (2018). [PubMed: 29884728]
34. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550 (2014). [PubMed: 25079552]
35. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489, 519–525 (2012). [PubMed: 22960745]
36. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499, 43–49 (2013). [PubMed: 23792563]
37. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. Cell 161, 1681–1696 (2015). [PubMed: 26091043]

38. Lo AS-Y, Xu C, Murakami A & Marasco WA Regression of established renal cell carcinoma in nude mice using lentivirus-transduced human T cells expressing a human anti-CAIX chimeric antigen receptor. Mol. Ther. Oncolytics 1, 14003 (2014). [PubMed: 27119093]
39. Cherkasova E et al. Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer. Cancer Res. 76, 2177–2185 (2016). [PubMed: 26862115]
40. Bolotin DA et al. Antigen receptor repertoire profiling from RNA-seq data. Nat. Biotechnol 35, 908–911 (2017). [PubMed: 29020005]
41. Scanlan MJ, Gure AO, Jungbluth AA, Old LJ & Chen Y-T Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol. Rev 188, 22–32 (2002). [PubMed: 12445278]
42. Reuben A et al. TCR Repertoire Intratumor Heterogeneity in Localized Lung Adenocarcinomas: An Association with Predicted Neoantigen Heterogeneity and Postsurgical Recurrence. Cancer Discov. 7, 1088–1097 (2017). [PubMed: 28733428]
43. Lu T et al. Tumor neoantigenicity assessment with CSiN score incorporates clonality and immunogenicity to predict immunotherapy outcomes. Sci. Immunol 5, (2020).
44. Simnica D et al. T cell receptor next-generation sequencing reveals cancer-associated repertoire metrics and reconstitution after chemotherapy in patients with hematological and solid tumors. Oncoimmunology 8, e1644110 (2019). [PubMed: 31646093]
45. Liu D et al. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. Nat. Med 25, 1916–1927 (2019). [PubMed: 31792460]
46. Van Allen EM et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science 350, 207–211 (2015). [PubMed: 26359337]
47. Hugo W et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell 165, 35–44 (2016). [PubMed: 26997480]
48. Kim ST et al. Comprehensive molecular characterization of clinical responses to PD-1 inhibition in metastatic gastric cancer. Nat. Med 24, 1449–1458 (2018). [PubMed: 30013197]
49. Miao D et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. Science 359, 801–806 (2018). [PubMed: 29301960]
50. Dash P et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 547, 89–93 (2017). [PubMed: 28636592]
51. Yost KE et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat. Med 25, 1251–1259 (2019). [PubMed: 31359002]
52. Nielsen M et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS ONE 2, e796 (2007). [PubMed: 17726526]
53. Gao W, Mahajan SP, Sulam J & Gray JJ Deep learning in protein structural modeling and design. Patterns (N Y) 1, 100142 (2020). [PubMed: 33336200]
54. Liu J & Gong X Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. BMC Bioinformatics 20, 609 (2019). [PubMed: 31775612]
55. Guo Y, Li W, Wang B, Liu H & Zhou D DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. BMC Bioinformatics 20, 341 (2019). [PubMed: 31208331]
56. Li H & Durbin R Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009). [PubMed: 19451168]
57. DePristo MA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet 43, 491–498 (2011). [PubMed: 21478889]
58. McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010). [PubMed: 20644199]
59. Van der Auwera GA et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 11, 11.10.1–11.10.33 (2013).
60. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol 31, 213–219 (2013). [PubMed: 23396013]

61. Koboldt DC et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012). [PubMed: 22300766]

62. Chiang C et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods 12, 966–968 (2015). [PubMed: 26258291]

63. Saunders CT et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012). [PubMed: 22581179]

64. Wilm A et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40, 11189–11201 (2012). [PubMed: 23066108]

65. Wang K, Li M & Hakonarson H ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010). [PubMed: 20601685]

66. Liu C et al. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Res. 41, e142 (2013). [PubMed: 23748956]

67. Jurtz V et al. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J. Immunol 199, 3360–3368 (2017). [PubMed: 28978689]

68. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]

69. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]

70. Bolotin DA et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods 12, 380–381 (2015). [PubMed: 25924071]

71. Eden E, Navon R, Steinfeld I, Lipson D & Yakhini Z GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48 (2009). [PubMed: 19192299]

72. Barbie DA et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462, 108–112 (2009). [PubMed: 19847166]

73. Lu T Code for ‘Deep Learning-based prediction of T cell receptor-antigen binding specificity ’. Zenodo 10.5281/zenodo.4670312 (2021).

74. Lu T Code for ‘Deep Learning-based prediction of T cell receptor-antigen binding specificity ’. Zenodo 10.5281/zenodo.4681560 (2021).

75. Lu T Code for ‘Deep Learning-based prediction of T cell receptor-antigen binding specificity ’. Zenodo 10.5281/zenodo.4670314 (2021).

76. Lu T Code for ‘Deep Learning-based prediction of T cell receptor-antigen binding specificity ’. Zenodo 10.5281/zenodo.4670320 (2021).

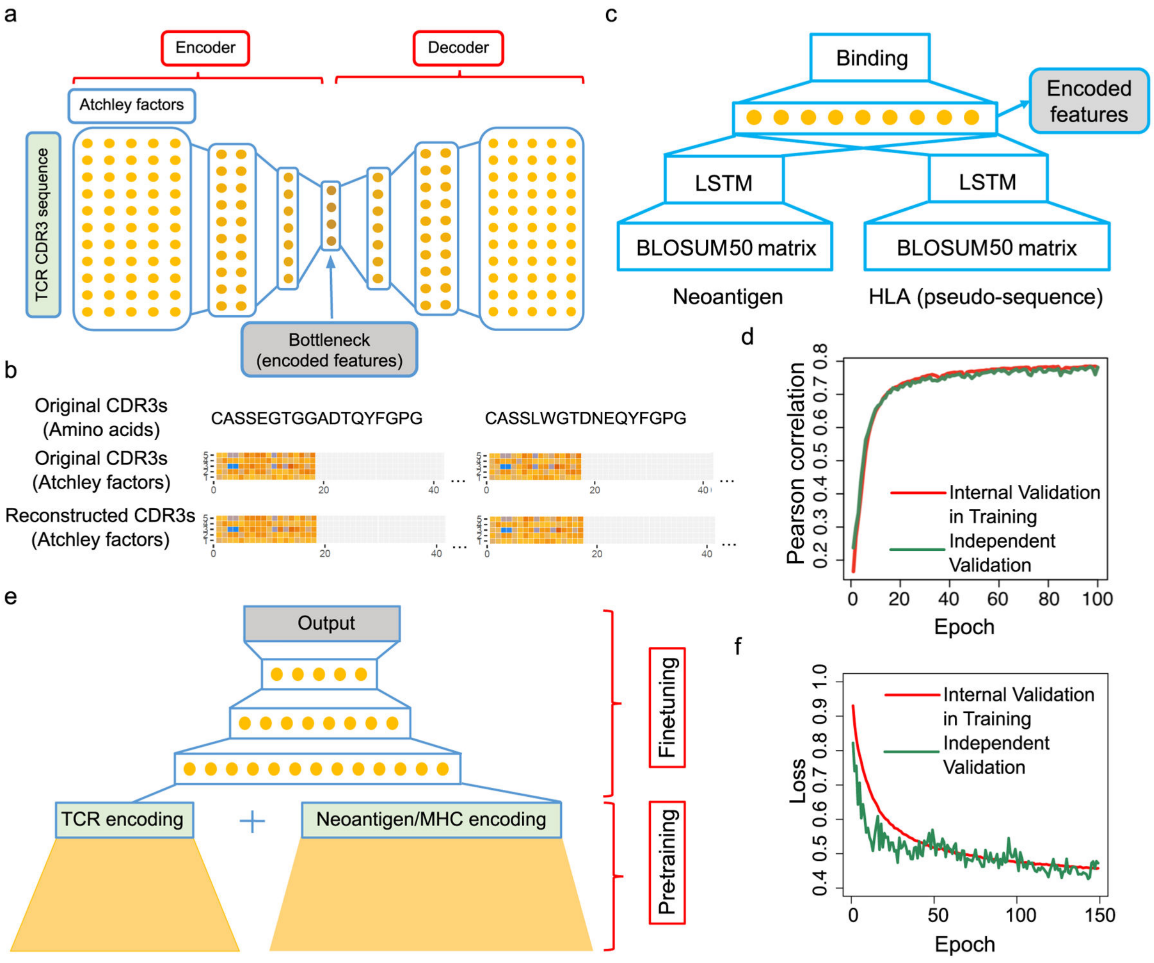


Fig. 1.

Deep learning the TCR binding specificity of neoantigens. (a) The structure of the stacked auto-encoder for learning TCR embeddings. (b) Original TCRs and reconstructed TCRs are almost the same. Original TCRs (amino acid sequences), Atchley factor-encoded TCRs (Atchley matrices of numbers), reconstructed TCRs (in the form of reconstructed Atchley matrices), and reconstructed TCR sequences (amino acid symbols determined by means of closest Euclidean distance) are shown. (c) The structure of the re-implemented netMHCpan model. (d) Validation of the predicted binding between (neo)antigens and MHC proteins generated by the pMHC embedding model, by the experimentally obtained data. The increase in the Pearson Correlation over training cycles (epochs) is shown. (e) Structure of the final pMTnet model. (f) The loss function of pMTnet over training time, in the units of epochs. The performances on both the internal validation subset that is split within the training cohort (red) and the independent validation cohort (green) are shown.

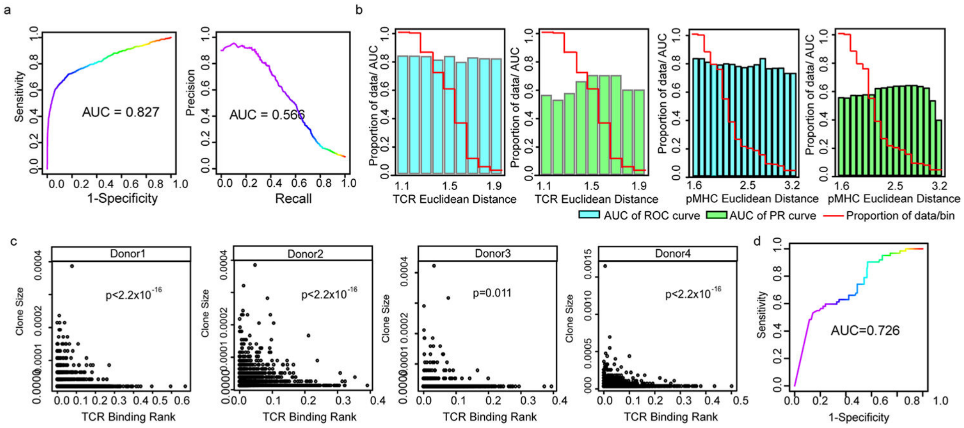


Fig. 2.

Validation of pMTnet. (a) AUCs of Receiver operating characteristic (ROC) and precision-recall (PR) of the predicted binding ranks (smaller ranks refer to stronger binding) were shown for the 619 experimentally validated TCR-pMHC binding pairs and 10 times more randomly shuffled negative pairs. (b) AUCs of ROC and PR for different cutoffs of euclidean distances of the 30-dimension PCs for embeddings were shown, where the cutoffs were used for subsetting TCRs (left group) and pMHCs (right group) of the 619 testing cohort. The AUCs were shown in light pink and green. The proportions of the selected TCRs and pMHCs out of the total 619 testing cohort, chosen by these cutoffs, were shown in blue. (c) The expansion of TCR clonotype is associated with their binding strength to pMHCs in the 10x Genomics Chromium Single Cell Immune Profiling datasets. The portion of this 10X Genomics dataset that was used in the validation phase is totally independent of the portion used in the training phase (see Supplementary Information for details). Y-axis shows the percentage of each clonotype in the whole pool of TCRs. The P values were calculated by the Spearman correlation test. (d) Peptide analogs that were experimentally validated as having stronger affinity towards the target TCR are predicted as having stronger affinity by pMTnet. An ROC plot was shown correlating the predictions (continuous variable) against the ground truth (binary variable). The Liu study dataset was shown.

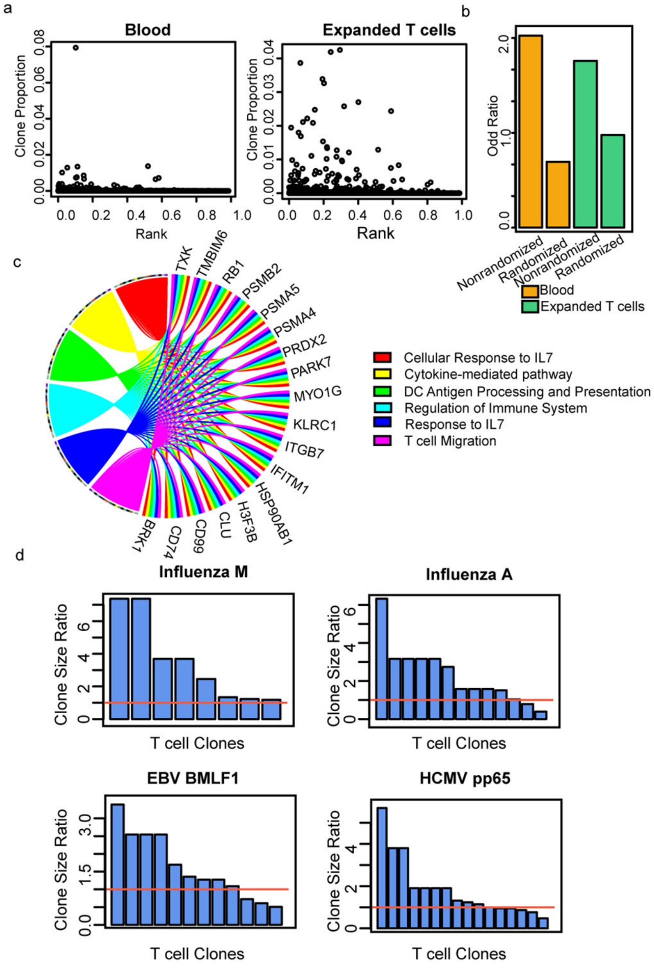


Fig. 3.

Prospective validation of pMTnet predictions. (a) TCR CDR3s predicted to have smaller binding ranks have higher clonal sizes. Blood cells: left panel and *in vitro* expanded T cells: right panel. X-axis shows the minimum of the binding ranks to any of the four viral pMHCs. Y-axis shows the clonal proportions of each TCR CDR3 clonatype in each sample. (b) Odds ratios for enrichment of highly expanded T cells with smaller binding rank for blood/expanded-T cells. We extracted the #CDR3s with clonal proportions>0.1% and with predicted rank<2% (HB); #CDR3s with clonal proportions<0.1% and predicted rank>2% (Ls); #CDR3 with clonal proportions>0.1% and predicted rank>2% (LB); #CDR3 with clonal proportions<0.1% and predicted rank<2% (Hs). Odds ratios are calculated as (HB *Ls)/(LB *Hs). Permutation of predicted ranks were performed, and the odds ratios were calculated again for control purposes. (c) Genes differentially expressed in T cells with predicted binding to viral pMHC (EBV BMLF1 as an example, rank cutoff=0.1) and T cells without binding are enriched in pathways essential for T cell functions. Right part of the circos plot shows differentially expressed genes and they are enriched in the corresponding pathways with the same colors on the left. (d) Ratios of clonal proportions in the viral pMHC treatment group *vs.* the vehicle treatment group. The red horizontal line (ratio=1) indicates no change.

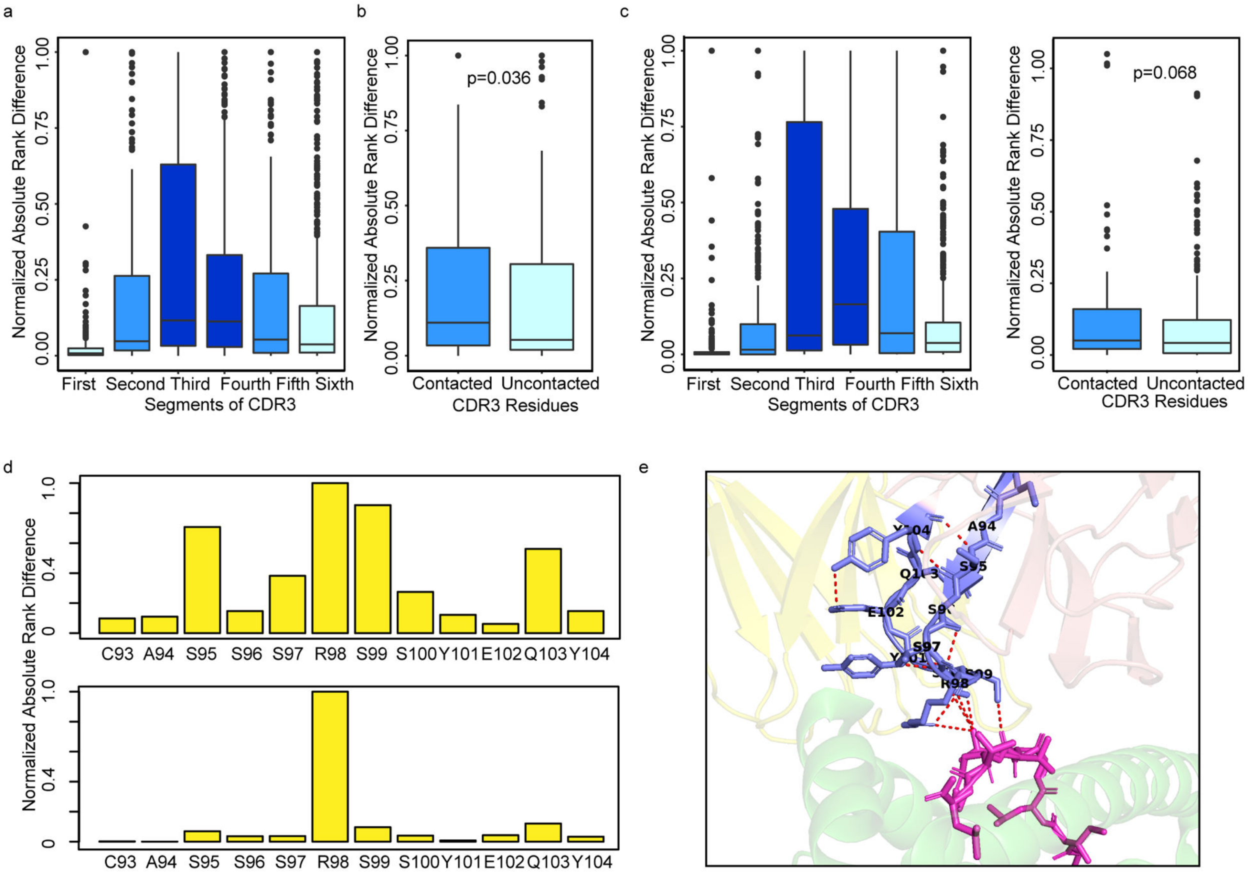
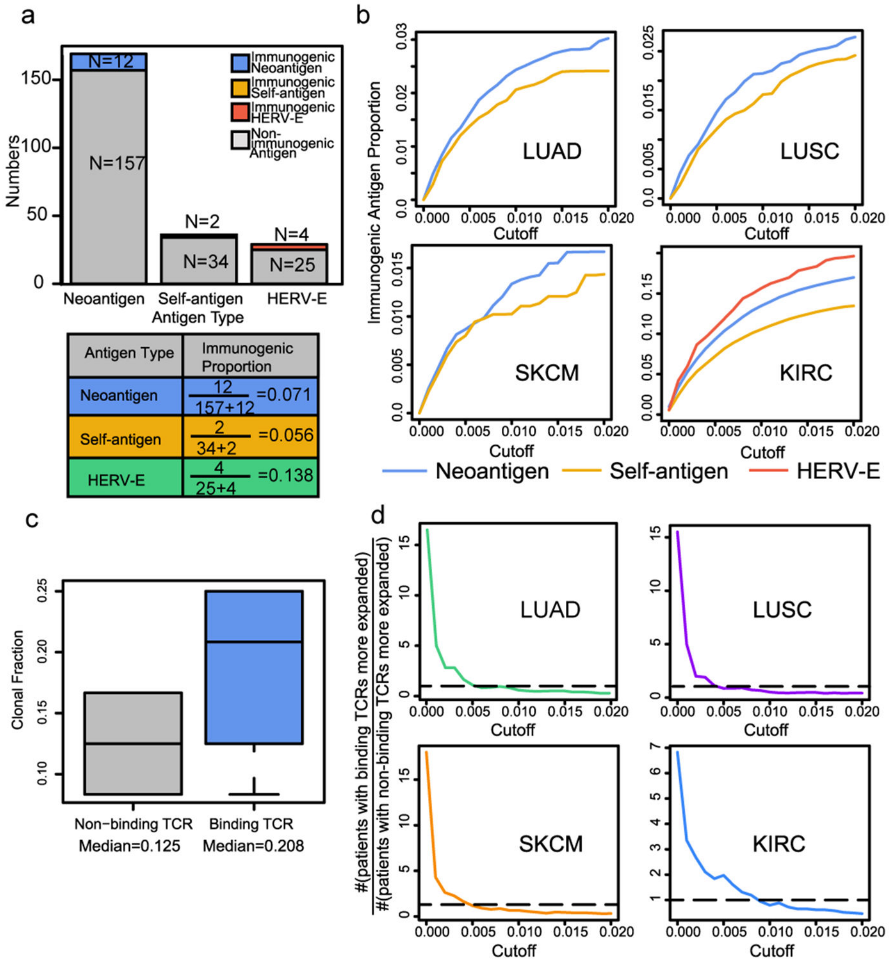


Fig. 4.

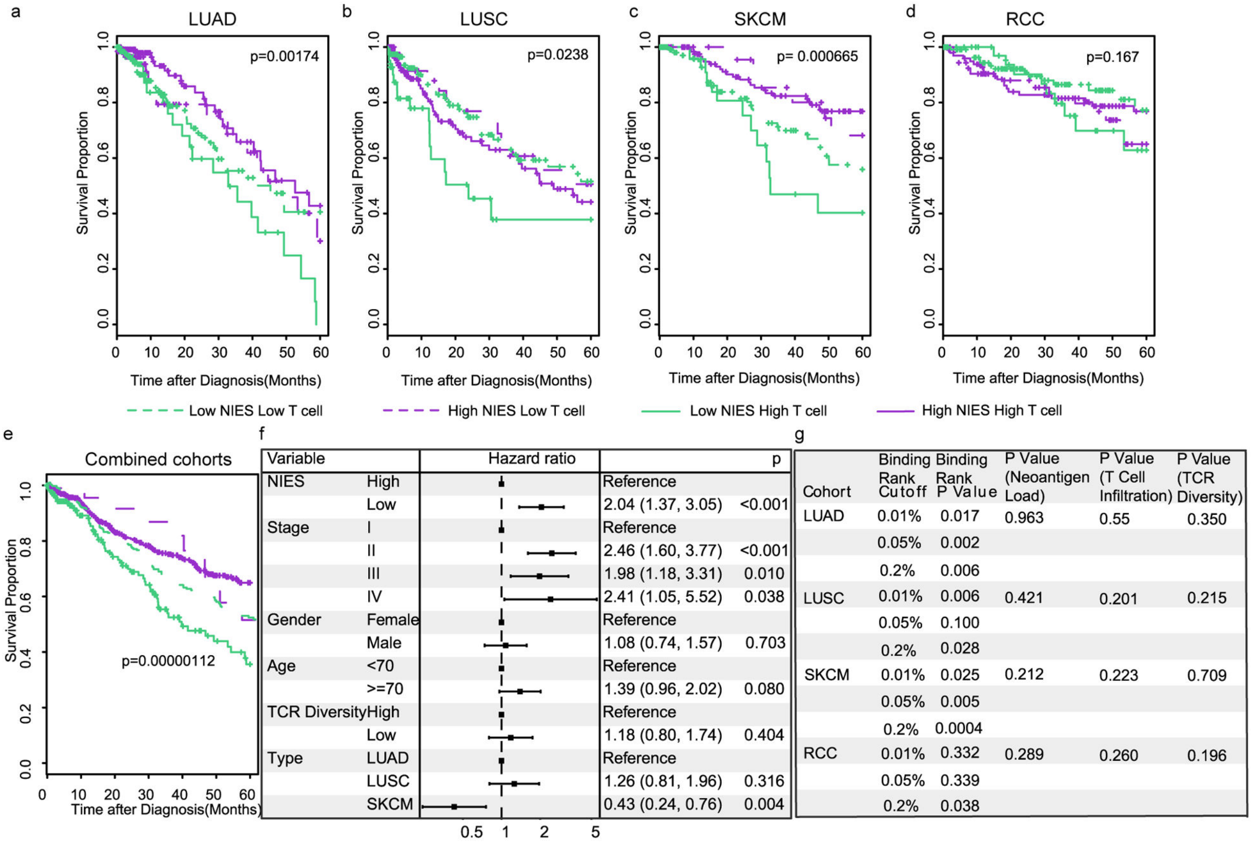
Structural analyses support the predicted TCR-pMHC interactions. (a) Residues in the middle segments of CDR3s are more likely to induce larger changes in predicted binding affinity. We divided each TCR CDR3 into six segments of equal lengths, and plotted the normalized changes in predicted binding ranks of residues in each segment of all CDR3s investigated. The absolute value of rank changes for each amino acid of a peptide are normalized by the maximal absolute value of rank changes for that peptide. (b) Residues with direct contacts are more likely to induce larger changes in the predicted pMHC binding strength than non-contacted residues. According to the 3D crystal structures, the CDR3 residues were grouped by whether or not they formed any direct contacts with any residues of pMHCs. P value is calculated by one-way Wilcoxon Signed Rank Test. (c) Same analysis done as in (a) and (b) except for using alanine scan. For boxplots in (a)-(c), box boundaries represent interquartile ranges, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median. (d) Predicted rank changes of amino acid residues in the CDR3 of one example TCR-pMHC structure (PDB id:5hhm). The top panel shows the results for 0-setting and the bottom panel shows the results for alanine scan. (e) 3D structure of 5hhm. Blue: CDR3 of TCRβ chain; yellow: TCRα chain; tints: other regions of the TCRβ chain; magenta: antigen; green: HLA.



Antigen Type	Immunogenic Proportion
Neoantigen	$\frac{12}{157+12}$ =0.071
Self-antigen	$\frac{2}{34+2}$ =0.056
HERV-E	$\frac{4}{25+4}$ =0.138

Fig. 5.

Characterizing the TCR-pMHC interactions in human tumors. (a) The number of immunogenic and non-immunogenic antigens of different classes for one example ccRCC patient (percentile rank cutoff=1%). The lower table shows the immunogenic percentage calculation process for this patient, which is applied to every patient in Fig. 4b. (b) The average percentage of immunogenic neoantigens, self-antigens (excluding HERV-E), and HERV-E peptides in each patient cohort. A series of binding cutoffs on the predicted pairing strength is applied. And with each cutoff, the immunogenic percentage is calculated for each patient and averaged within each cohort. (c) TCR clonal fractions of binding and non-binding TCRs identified in one example patient. “Binding” refers to the predicted binding of TCRs to any of the neoantigens, self-antigens, or HERV-Es, with the binding rank cutoff being 1%. The box boundaries represent interquartile ranges, and the line in the middle of the box represents the median. (d) The ratio of the number of patients with binding T cells having a higher average clonal fraction over the number of patients with non-binding T cells having a larger average clonal fraction. This ratio is calculated with a series of binding rank cutoffs. The dashed horizontal line indicates the ratio of 1.



Variable		Hazard ratio	p
NIES	High	Reference	
	Low	2.04 (1.37, 3.05)	<0.001
Stage	I	Reference	
	II	2.46 (1.60, 3.77)	<0.001
	III	1.98 (1.18, 3.31)	0.010
	IV	2.41 (1.05, 5.52)	0.038
Gender	Female	Reference	
	Male	1.08 (0.74, 1.57)	0.703
Age	<70	Reference	
	>=70	1.39 (0.96, 2.02)	0.080
TCR Diversity	High	Reference	
	Low	1.18 (0.80, 1.74)	0.404
Type	LUAD	Reference	
	LUSC	1.26 (0.81, 1.96)	0.316
	SKCM	0.43 (0.24, 0.76)	0.004

Cohort	Binding Rank Cutoff	Binding Rank P Value	P Value (Neoantigen Load)	P Value (T Cell Infiltration)	P Value (TCR Diversity)
LUAD	0.01%	0.017	0.963	0.55	0.350
	0.05%	0.002			
	0.2%	0.006			
LUSC	0.01%	0.006	0.421	0.201	0.215
	0.05%	0.100			
	0.2%	0.028			
SKCM	0.01%	0.025	0.212	0.223	0.709
	0.05%	0.005			
	0.2%	0.0004			
RCC	0.01%	0.332	0.289	0.260	0.196
	0.05%	0.339			
	0.2%	0.038			

Fig. 6.

Efficiencies of TCR-neoantigen interactions impact tumor progression. (a-e) Kaplan-Meier estimator was used to visualize patient overall survival for each cohort. P values for log-rank tests are shown for testing the separation of the survival curves of high NIES and low NIES patients within the high T cell infiltration subsets. Patients were split on the median of T cell infiltration and median of NIES. (a) LUAD (b) LUSC, (c) SKCM, (d) RCC, and (e) combined cohort of LUAD, LUSC, and SKCM. There are 427, 389, 401, and 366 patients in LUAD, LUSC, SKCM, and RCC cohort respectively. (f) Multivariate analysis for the cohort in (e) with adjustment of several important covariates. The results shown in (a-f) use the cutoff of 1%. (g) The prognosis power of NIES calculated with TCRs assigned to neoantigens with a series of cutoffs on predicted binding ranks. The same analyses for neoantigen loads, T cell infiltrations and TCR diversity were also carried out as the control.