MODELING TUMOR NEOANTIGENS FOR PREDICTING PATIENTS' CLINICAL OUTCOMES

APPROVED BY SUPERVISORY COMMITTEE

Tao Wang, Ph.D. (Mentor)
Guanghua Xiao Ph.D. (Mentor)
Yujin Hoshida Ph.D.
Todd Aguilera M.D. Ph.D.
Chul Ahn Ph.D.

DEDICATION

I would like to thank my mentors Dr. Tao Wang and Dr. Guanghua Xiao. They provided unwavering support and direction on every research project through my PhD study at UT Southwestern Medical Center. I would also like to thank my thesis committee members, Dr. Yujin Hoshida, Dr. Chul Ahn, Dr. Todd Aguilera, who provided me with insightful suggestions and guidance on my research work. Furthermore, I am grateful to have the opportunities to collaborate with scientists on campus and out of campus on exciting research projects. Finally, I am thankful for my parents, my husband, and my friends who are supportive and encouraging to me during my PhD study.

MODELING TUMOR NEOANTIGENS FOR PREDICTING PATIENTS' CLINICAL OUTCOMES

by

TIANSHI LU

DISSERTATION / THESIS

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

December, 2021

MODELING TUMOR NEOANTIGENS FOR PREDICTING PATIENTS' CLINICAL OUTCOMES

Tianshi Lu, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2021

Supervising Professor: Tao Wang, Ph.D. & Guanghua Xiao, Ph.D.

Abstract

Tumor neoantigens are critical targets of the host antitumor immune response and their presence play an important role in affecting tumor progressions and immunotherapy treatment response. Neoantigens showed a lot of potential of being applied to clinical treatment. However, systematic study of neoantigens' impact on tumors and patients is still challenging due to the huge diversity of neoantigens, heterogeneity within tumors, and the model to study the pairing

between neoantigen-MHC and T cells to identify the neoantigens that truly elicit T cell response. To study the impact of neoantigen-T cell interaction on tumorigenesis, I developed a Bayesian hierarchical model to infer the history of neoantigen-cytotoxic T cell interactions in tumors.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

First or co-first author publications

1. Zhu M*, **Lu T*,** Jia Y*, Luo X*, Gopal P, Li L, Odewole M, Renteria V, Singal AG, Jang Y, Ge K, Wang SC, Sorouri M, Parekh JR, MacConmara MP, Yopp AC, Wang T , Zhu H . Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. ***Cell***. 2019; PubMed [journal] PMID: 30955891 **(Co-first author)**

2. **Lu T**, Wang S, Xu L, Zhou Q, Singla N, Gao J, Manna S, Pop L, Xie Z, Chen M , Luke J, Brugarolas J, Hannan R, and Wang T. Tumor Neoantigenicity Assessment with CSiN Score Incorporates Clonality and Immunogenicity to Predict Immunotherapy Outcomes. 2020. ***Science Immunology.*** Sci. Immunol. 5, eaaz3199. PMID: 32086382 **(First author)**

3. **Lu T**, Park S, Zhu J, Wang Y, Zhan X, Wang X, Wang L, Zhu H, and Wang T. Overcoming Expressional Drop-outs in Lineage Reconstruction from Single-cell RNA Sequencing Data. ***Cell Reports***. 2021. Jan 5;34(1):108589. PMID: 33406427. **(First author)**

4. **Lu, T**., Zhang, Z., Zhu, J. et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. Nat Mach Intell 3, 864–875 (2021). https://doi.org/10.1038/s42256-021-00383-2 **(First author)**

Co-author publications

5. Lin Y, Zhang S, Zhu,M, **Lu,T**, Chen K, In,Z, Wang S, Xiao G, Luo,D, Jia Y, Li L, MacConmara,M, Hoshida Yu, Singal A, Yopp A, Wang T, Zhu H. Mice With Increased Numbers of Polyploid Hepatocytes Maintain Regenerative Capacity but Develop FeIr Tumors Following Chronic Liver Injury. 2020. ***Gastroenterology*** 158:1698-1712 e14

6. Park S, Wang X, Lim, J Xiao G, **Lu T**, and Wang T. Bayesian Multiple Instance Regression Model for Modeling Immunogenic Neoantigens. 2020. ***Statistical Methods in Medical Research.*** 13;962280220914321.

7. Hsieh D, Hsieh A, Samstein R, **Lu T**, Beg M, Gerber D, Wang T, Morris L, Zhu H. DNA repair gene mutations as predictors of immune checkpoint inhibitor response beyond tumor mutation burden. 2020. ***Cell Reports Medicine*** 1:100034

8. Chung A, Mettle M, Ganguly D, **Lu T**, Wang T, Brekken R, Hsiehchen D, and Zhu H. Immune Checkpoint Inhibition is Safe and Effective for Liver Cancer Prevention in a Mouse Model of Hepatocellular Carcinoma. 2020. ***Cancer Prevention Research***: canprev res.0200.2020.

9. Lu C, Guan J, Lu S, Jin Q, Rousseau B, **Lu T**, Stephens D, Zhang H, Zhu J, Yang M, Ren Z, Liang Y, Liu Z, Han C, Liu L, Cao X, Zhang A, Qiao J, Batten K, Chen M, Castrillon DH, Wang T, Li B, Diaz LA Jr, Li GM, Fu YX. DNA Sensing in Mismatch Repair-Deficient Tumor Cells Is Essential for Anti-tumor Immunity. ***Cancer Cell***. 2021 Jan 11;39(1):96-108.e6. doi: 10.1016/j.ccell.2020.11.006. Epub 2020 Dec 17. PMID: 33338425.

10. Zhu M, Li L, **Lu T**, Yoo H, Zhu J, Gopal P, Wang SC, Porembka MR, Rich NE, Kagan S, Odewole M, Renteria V, Waljee AK, Wang T, Singal AG, Yopp AC, Zhu H. Uncovering Biological Factors That Regulate Hepatocellular Carcinoma Growth Using Patient-Derived Xenograft Assays. ***Hepatology***. 2020 Sep;72(3):1085- 1101. doi: 10.1002/hep.31096. Epub 2020 Sep 2. PMID: 31899548; PMCID: PMC7332388.

LIST OF FIGURES

LIST OF TABLES

LIST OF DEFINITIONS

DNA – Deoxyribonucleic acid

RNA – Ribonucleic acid

MHC – Major Histocompatibility Class

TCR – T cell receptor

NGS – Next generation sequencing

TMB – Tumor mutation burden

TNB – Tumor neoantigen burden

FACS – Fluorescence-activated Cell Sorting

IEDB – The Immune Epitope Database and Analysis Resource

PD-1 – Programmed cell death protein 1

PD-L1– Programmed death-ligand 1

CTLA4 – A protein receptor that functions as an immune check point and downregulates immune responses

CSiN – The model named Caauchy-Schwarz index of Neoantigens

ccRCC- Clear cell renal cell carcinoma

NSCLC – Non-small cell lung cancer

NDB – Non-durable clinical benefit

DCB – Durable clinical benefit

TCGA – The cancer genome atlas

LSTM – Long short-term memory

RELU – Rectified linear unit

CHAPTER ONE
INTRODUCTION

1.1 The history of neoantigen identification and understanding

Neoantigens were firstly discovered to simulate T cells in mouse tumor models by cDNA library screening (De Plaen et al. 1988; Jiang et al. 2019). They found that even one different amino acid between normal and tumor peptide could result in positive T cell responses. Researchers overexpress neoantigens and MHC in the antigen presenting cells and co-culture with T cells to identify the antigens which positively interact with T cells. But this method is time-consuming and costly. After the development of next generation sequencing (NGS), researchers can identify antigens from tumor cells by comparing the differences between tumor and normal cells sequenced by NGS (Richters et al. 2019, Lancaster et al. 2020). The identification of neoantigens from tumor cells by NGS makes it necessary to develop computational pipelines. Neoantigens generated by somatic mutations were identified in various human tumors including melanoma and renal cell carcinoma (Jiang et al. 2019). Researchers found that the infusion of neoantigen-responsive T cells in the tumor could result in regression of tumors (Ali et al 2019), which indicates the importance of neoantigens in the tumor-immune system interactions.

1.2 The interaction between neoantigen and T cells

Neoantigens are the small peptides that are transcribed and translated from somatic mutations in tumor cells. They are presented by Major Histocompatibility Class (MHC) on the surface of tumor cells and serve as the targets of T cells by interacting with T cells through T cell receptors (TCRs). The binding of T cells with neoantigens initiates tumor cytotoxic effects. Immunotherapies are applied to cancer patients and showed promising ways to cure cancer. But only a small proportion of patients showed clinical benefit after being treated by immunotherapy.

People found that the CD8+ T cell level is positively correlated with patients' responses to immunotherapy. CD8+ T cells iterate with tumor cells through the recognition of neoantigens and neoantigens are one of the biggest differences between cancer cells and normal cells. Immunotherapies have highlighted the role of neoantigens in checkpoint inhibitor-induced immune response. Thus, the interaction between neoantigen and T cell is central to understanding the neoantigens' impact on tumorigenesis, prognosis, and treatment response. Unfortunately, researchers in this field are far from clearly understanding the difference between responders and nonresponders and neoantigens' impact on clinical outcomes.

In this field, one major impediment of current research is the way of correlating neoantigens with immunotherapy treatment response by only considering neoantigen/mutation load.

In my thesis work, I studied neoantigens and the interaction between neoantigen and T cells in the following aspects. An assessment method for evaluating the neoantigen repertoires for tumors was developed in order to better understand the relationship between neoantigens and patient's clinical outcomes and predict the patients' clinical responses. I developed a deep learning model for predicting the binding between TCR and peptide-MHC pairing for identifying the neoantigens which can truly elicit T cell responses. I also constructed a model to study the interaction between neoantigens and T cells and their impact on the evolution of cancerous masses.

1.3 Existing analysis methods studying neoantigens

Bioinformatics tools and large-scale screenings were developed in order to better understand the relationship between neoantigens and T cells. The quick development of antigen research also revealed challenges and problems in the field. In this section, I will give an overview of existing bioinformatics tools and large-scale screening methods for studying neoantigen-T cell interactions.

1.3.1 Tumor mutation load/tumor neoantigen load

After neoantigens were discovered, researchers tried to link neoantigens with patients' clinical phenotypes. The earliest and most popular way of correlating neoantigens with phenotypes is tumor mutation burden/tumor neoantigen burden, which is the count of mutations or the total number of neoantigens in each patient. But the correlation between mutation burden/neoantigen load and immunotherapy response is significant in some cohorts but not in other cohorts. The problem of this method is that it simplistically treats all mutations and neoantigens equally. This overly simplistic approach fails to take full advantage of the wealth of information contained in the entire repertoire of neoantigens. Neoantigens are associated with mutations that can be truncal or subclonal. Some neoantigens are more immunogenic than others. These details are not fully captured by the basic neoantigen/mutation load approach but could be critical for understanding the responsiveness of patients with cancer to immunotherapy treatment.

1.3.2 Neoantigen Fitness Model

A more sophisticated neoantigen-based predictive metric is the neoantigen fitness model (Łuksza et al. 2017) developed on the basis of evolutionary modeling of patient neoantigen profiles. This model only considered the neoantigen class I major histocompatibility complex (MHC) binding affinity but did not consider the neoantigen class II MHC binding affinity. It only retained the top neoantigen resulting from missense mutations with the highest binding affinity within each tumor clone. The neoantigens generated by other types of mutations, such as insertion, deletion is not considered. This metric demonstrated excellent predictive power for survival of patients after immunotherapy treatment in a few cohorts; however, its predictive values and prognostic values have not been widely evaluated.

1.3.3 netTCR

netTCR is a model based on convolutional networks to predict epitopes recognized by T cells. This model was trained by data from VDJdb and IEDB. However, netTCR only accommodates for the HLA-A:0201 allele, epitopes shorter than 10 amino acids, and CDR3s shorter than 10 amino acids. This method has not been comprehensively validated in independent testing datasets.

1.3.4 TCRGP

TCRGP (Jokinen et al. 2021) is a Gaussian process (GP) classification model with the substitution method of modified BLOSUM62 for sequence representation. They used the data from VDJdb for training. This model is limited to 22 known antigens including, BMLF, M1, pp65, ATDALMTGY, CINGVCWTV, etc.

1.3.5 TCRex

TCRex (Gielis et al. 2019) is a predictive model for antigen-TCR binding based on a random forest model. This model could only be applied to a handful of antigens, including 7 CMV viral antigens, 1 DENV1 viral antigen, 1 DENV2 viral antigen, 1 DENV3/4 viral antigen, 6 EBV viral antigens, 5 HCV viral antigens, 17 HIV viral antigens, 1 HSV2 viral antigen, 3 Influenza viral antigens, 5 cancer antigens. This model will not be applicable for novel neoantigens.

1.3.6 IEDB

The Immune Epitope Database (Fleri et al. 2017) and Analysis Resource is a comprehensive database with 26000 available human TCR sequences and antigen sequences with experiment information from 18000 references. This database also collected around 30 pairs of interactive TCRs and antigens with their protein structures.

1.3.7 VDJdb

VDJdb (Bagaev et al. 2020) is a database with TCR sequences and antigen sequences with known specificities. The database collected 61049 pairs of TCRs and antigens from 155

published studies. It also takes all the available experiment information into account to assign a confidence score to highlight reliable antigen-T cell pairs in the database. The data is used by most of the bioinformatics tools studying antigens and T cells.

1.3.8 T-Scan

T-Scan (Kula et al. 2019) is a cell-based high-throughput method for studying the targets for cytotoxic T cells. They expressed a library of antigens that are processed and presented endogenously on MHC molecules in the target cells. These cells are co-cultured with T cells from patients or donors.

These target cells, which can be recognized by T cells, can be differentiated by a GzB reporter for granzyme B activity through fluorescence-activated cell sorting (FACS). They used PCR and NGS to sequence and identify the antigens which have positive interactions with T cells. This large-scale screening method is labor-intensive and costly due to the restricted number of T cell receptors per experiment. But it provided a lot of positively interactive T cell-antigen data for the community to study.

CHAPTER TWO

Tumor Neoantigenicity Assessment with CSiN score

Background and rationale

Immunotherapies shed light on the treatment of cancer in modern era clinical sciences. However, most immunotherapies are only beneficial for a small proportion of patients. For example, only 10% to 50% response rate showed for melanoma patients and non-small lung cancer patients treated by anti-PD-1 and anti-PD-L1. 30% response rate was shown for renal cell carcinoma

patients treated by anti-PD1. This field is in need of understanding how to distinguish responders from non-responders and the mechanism behind it. All forms of immunotherapy, such as checkpoint inhibitors and neoantigen vaccines, seek to activate the host immune system to attack the tumor cells. These forms of immunotherapy have different modes of actions, but most are intended to mobilize the cytotoxicity of T cells in cancer patients. Neoantigens are one of the biggest differences between tumor cells and normal cells and the primary targets of T cells. So, the profiles of tumor neoantigens in each patient play a crucial role in determining the responsiveness to immunotherapy treatment.

Most existing method for correlating neoantigens with immunotherapy treatment response is by using total neoantigen/mutation load (namely, the total number of neoantigens or mutations). The neoantigen load and mutation load showed a good correlation with patients' phenotypes in some cohorts but not others. For example, Van Allen. et al found that higher mutation load and neoantigen was correlated with better immunotherapy (anti-CTLA 4 treatment) responses. But Matsushita et al did not find a good correlation between the neoantigen load or mutation load and patients' overall survival rate. The reason that the model could not be successfully validated across multiple cohorts with various tumor types is that it failed to take advantage of the rich information of the entire repertoire. The wealth of information contained in the neoantigen repertoire includes the concentration of neoantigens on truncal/subclonal mutations, the immunogenicity of neoantigens, etc. These details are not Ill captured by the neoantigen load/mutation load.

The only other study that defined a more sophisticated neoantigen-based assessment metric is the neoantigen fitness model based on evolutionary modeling of patient neoantigen profiles. This work only considered the neoantigen class I major histocompatibility complex (MHC) binding

affinity and only retained the top neoantigens resulting from missense mutation with the highest binding affinity within each tumor clone. The neoantigens generated from stoploss, indel mutations, which are found to be more immunogenic neoantigens, are not considered in this work. This method showed good predictive poIr for patients' survival rate after immunotherapy treatment in three cohorts; hoIver, the predictive values and prognostic values have not been widely evaluated.

In order to take advantage of the rich information contained in the neoantigen repertoire and quantitatively characterize the neoantigen profiles in patients, I developed an assessment tool called Cauchy-Schwarz index of Neoantigens (CSiN). The model considered the number of neoantigens, the distribution of neoantigens on truncal or subclonal mutations and the affinity betIen MHC and neoantigen.

Methods

The definition of CSiN

The CSiN score considers the pairing between the repertoire of neoantigens and the tumor mutations to which they belong. I characterized this property by averaging the product of the VAFs of somatic mutations and the number of neoantigens generated by each mutation, normalized by the average VAF and average mutation-specific neoantigen load in each patient. The product of average VAF and average mutation-specific neoantigen load forms the backbone of the CSiN score. The name CSiN was selected because of the pairing of tumor mutations and neoantigens, and its effect on the overall score bear analogy to the Cauchy-Schwarz inequality, which describes the upper bound of the product sum of two vectors of real numbers and the condition for the equality to be achieved.

The fundamental building block of CSiN is $\sum_{i=1..n} \frac{Vaf_i}{Vaf} \times \frac{load_i}{load}$. The variance allele frequency (VAF) is the number of variant reads divided by the total number of reads covering each variant position. The load is the number of neoantigens associated with each mutation. *n* is the total number of missenses, indels, and stop-loss somatic mutations in a tumor sample. Vaf describes the average VAF of all the somatic mutations (to control for tumor purity) and load is the average per-mutation neoantigen load across all somatic mutations (so CSiN is orthogonal to neoantigen load). It is common to see different tumor biopsies have different levels of non-tumor cell contents (immune and stromal cells), and the tumor mutations' VAFs will be influenced by this confounding factor. The procedure of division by Vaf helps to normalize this effect. According to the Cauchy-Schwarz inequality, when the mutations with higher VAFs are also the mutations that generate more neoantigens (our hypothesized favorable distribution), the product value will be larger (higher CSiN score). Therefore, a higher CSiN conforms to a favorable neoantigen clonal structure.

Because the neoantigens vary in quality, and to give more weight to better neoantigens, the value is calculated by the average of the products calculated with different cutoffs on quality of neoantigens, with better neoantigens convolved in more rounds of calculations.

$$CSiN = \frac{\sum_{c=\{c_0,c_1,\ldots c_k\}} \log\left(\frac{\sum_{i=1..n}^{q(i)>c} \frac{Vaf_i}{Vaf_c} \times \frac{load_i}{load_c}}{\sum_{i=1..n} I(q(i)>c)}\right)}{k}$$

In this study, I used the percentile rank variable generated by the IEDB MHC binding affinity prediction software as the quality metric, q(i). This variable measures the binding strength between neoantigens and the MHC molecules, and a smaller percentile rank delineates a greater

affinity. The average VAF and neoantigens load are calculated with their according cutoff value, c, and I used k cutoff values of 0.375, 0.5, 0.625 0.75, 1.25, 1.75, and 2. The upper bound of the cutoff values is 2%, which is the most Ill-established cutoff for an epitope to be considered as an HLA binder, according to netMHCpan. I evaluate to 1 if the statement is true, 0 otherwise. Accordingly, the definition, the definition of the average VA and neoantigen loads area revise as:

$$\overline{Vaf_c} = \frac{\sum_{\substack{i=1..n \\ q(i)>c}} Vaf_i}{\sum_{i=1..n} I(q(i)>c)} \text{ and } \overline{load_c} = \frac{\sum_{\substack{i=1..n \\ q(i)>c}} load_i}{\sum_{i=1..n} I(q(i)>c)}$$

$$CSiN = \frac{\sum_{c=\{c_0,c_1,\ldots c_k\}} \log\left(\frac{\sum_{\substack{i=1..n \\ rank(-Vaf_i)\le M \\ q(i)>c}} \frac{Vaf_i}{\overline{Vaf_c}} \times \frac{load_i}{\overline{load_c}}}{\sum_{\substack{i=1..n \\ rank(-Vaf_i)\le M}} I(q(i)>c)}\right)}{k}$$

To accommodate the patient samples with an extremely large number of mutations, an adjustment is made where the calculation only considers the top M mutations with the largest VAFs when there are more than M mutations (M=500 in this study).

$$CSiN = \frac{a}{k} \times \sum_{c=\{c_0,c_1,\ldots c_k\}} \log\left(\frac{\sum_{\substack{i=1..n \\ rank(-Vaf_i)\le M \\ q(i)>c}} \frac{Vaf_i}{\overline{Vaf_c}} \times \frac{load_i}{\overline{load_c}}}{\sum_{\substack{i=1..n \\ rank(-Vaf_i)\le M}} I(q(i)>c)}\right)$$

The CSiN score defined above is a random variable centered approximately at zero. The final reported CSiN score is multiplied by a fixed constant, a (a=10), to increase the dynamic range for better visualization.

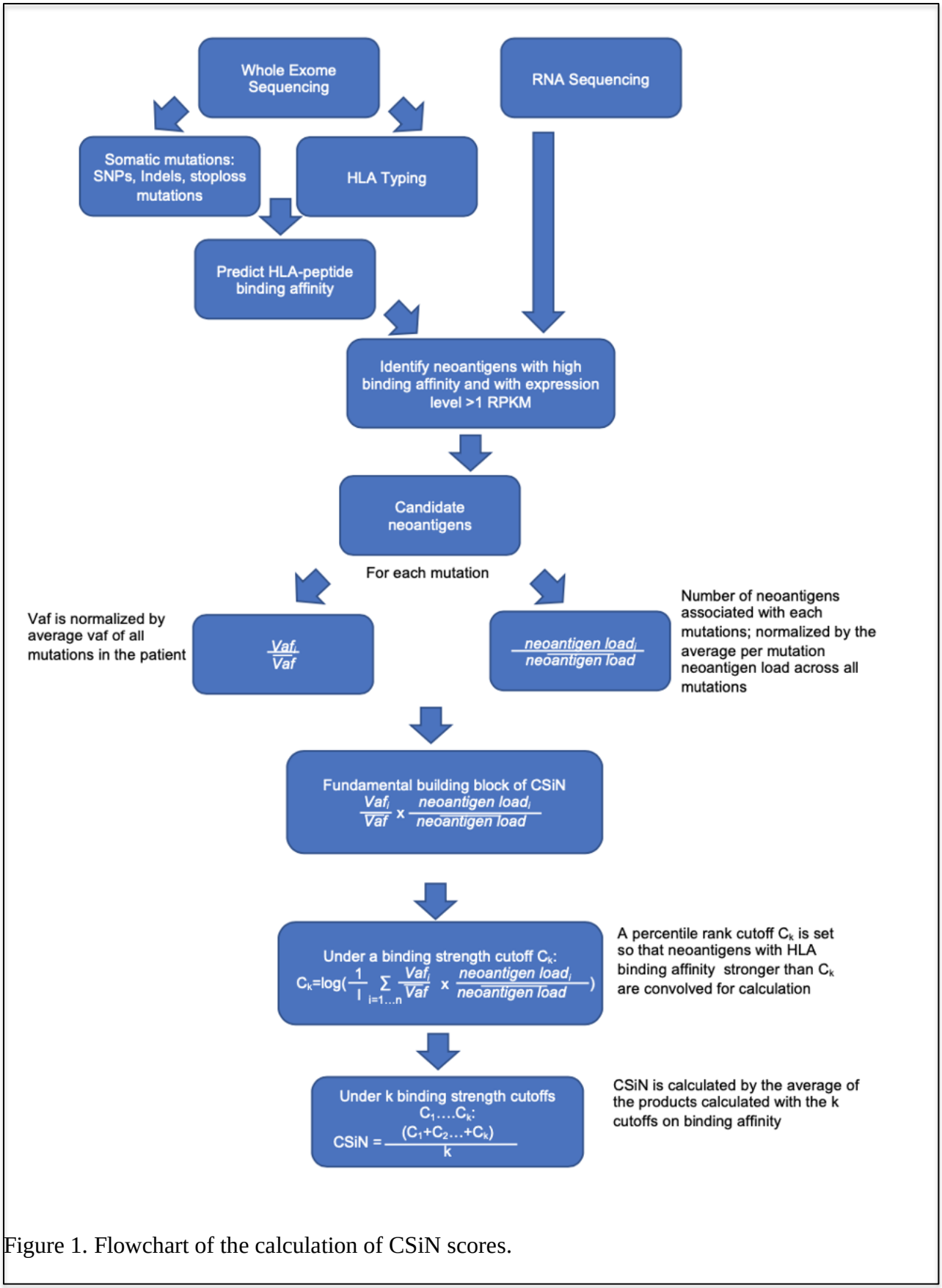


Figure 1. Flowchart of the calculation of CSiN scores.

Result

Better response to checkpoint inhibitors in immunogenic cancers is associated with higher CSiN scores

I gathered nine cohorts of cancer patients treated by immunotherapy (Table. 1) to investigate the implications of CSiN for checkpoint inhibitor treatment response. One cohort of melanoma patients on anti--CTLA-4 therapy from Van Allen *et al. (Van Allen et al. 2015)* was analyzed. The whole exome sequencing data and RNA sequencing data were acquired and processed to get mutations, neoantigens, and HLA types. I observed patients with better responses were more likely to have high CSiN than patients with worse responses (Figure 2A, P=0.009, chi-squared test). Another cohort of melanoma patients (Snyder cohort (Snyder et al. 2014)) treated by anti-CTLA-4 therapy were analyzed. Patients who received a durable clinical benefit had higher CSiN scores than patients with no durable benefit (Figure 2B, P=0.033). A third cohort of melanoma patients (Riaz cohort (Riaz et al. 2017)) treated by anti--PD-1. The association between CSiN score and patients' responses to treatment was significantly positive (Figure 2C, P=0.037). One more cohort (Hugo cohort (Hugo et al. 2016)) of melanoma patients showed the trend of patients with better responses associated with high CSiN scores (Figure 2D, P=0.043). Other than melanoma patients, I also acquired clear cell RCC (ccRCC) patients treated by anti--PD-1/anti--PD-L1 from (Miao et al. 2018). The same significantly positive association of higher CSiN scores with better response was observed (Figure 2E, P=0.036). I analyzed metastatic ccRCC patients treated with atezolizumab, an anti--PD-L1 agent (IMmotion150 cohort) (McDermott et al. 2018). A significant association of higher CSiN scores with better treatment responses for T effector--high patients treated with atezolizumab (Figure 2F, P=0.028). In contrast, this trend was not observed in patients treated by sunitinib (P=0.890). NSCLC patients

(the Hellmann cohort) treated with PD-1 and CTLA-4 inhibitors were available from (Hellmann et al. 2018). The analysis showed that patients with durable clinical benefit had higher CSiN scores than patients with no durable benefit (Figure 2G, P=0.007), whereas this association is insignificant for patients with low PD-L1 expression. Another NSCLC cohort (the Acquired cohort) from (Anagnostou et al. 2017) and (Gettinger et al. 2017). Patients with sustained response are more likely to have higher CSiN scores than patients with short-term progression (Figure 2H, P=0.015). Last, another cohort of NSCLC patients on anti--PD-1 therapies from (Rizvi et al. 2015) had higher CSiN scores than patients with NDB responses (Figure 2I, P=0.058).

Cohort ID	Disease type	Immunotherapy treatment	Raw data	Total # patients
Hugo	Melanoma	Anti-PD1	GSE78220	26
Riaz	Melanoma	Anti-PD1	SRP095809 and SRP094781	65
Snyder	Melanoma	Anti-CTLA4		61
VanAllen	Melanoma	Anti-CTLA4	phs000452.v2.p1	37
Miao	ccRCC	Anti-PD1/anti-PDL1	phs001493.v1.p1	33
IMmotion150	ccRCC	anti-PDL1	EGAS00001002928	149
	Non-Small Cell Lung Cancer	Anti-PD1/anti-PDL1/anti-CTLA4	phs001464.v1.p1	11
Acquired	Lung adenocarcinoma	Anti-PD1/anti-CTLA4		3
Hellmann	Non-Small Cell Lung Cancer	Anti-PD-1 plus anti-CTLA-4		74
Rizvi	Lung adenocarcinoma	Anti-PD1		26

Table 1. The information of patients treated by immune checkpoint inhibitors

To compare the performance of CSiN score with other widely used metrics for neoantigenicity, I also examined the predictive power of neoantigen load and the neoantigen fitness model scores to split the cohorts. I used a bootstrap analysis to evaluate the statistical significance of the improvement of CSiN compared with the other two approaches, which is an accepted methodology for model comparisons (Sieberts et al. 2016) (Costello et al. 2014). The CSiN significantly outperformed neoantigen load in seven of the nine cohorts evaluated, and neoantigen fitness in seven of the nine cohorts (Figure 2J). Overall, the results show that CSiN is capable of predicting clinical response to checkpoint inhibitors in immunogenic cancers and demonstrated a significant improvement over other existing predictive tools.

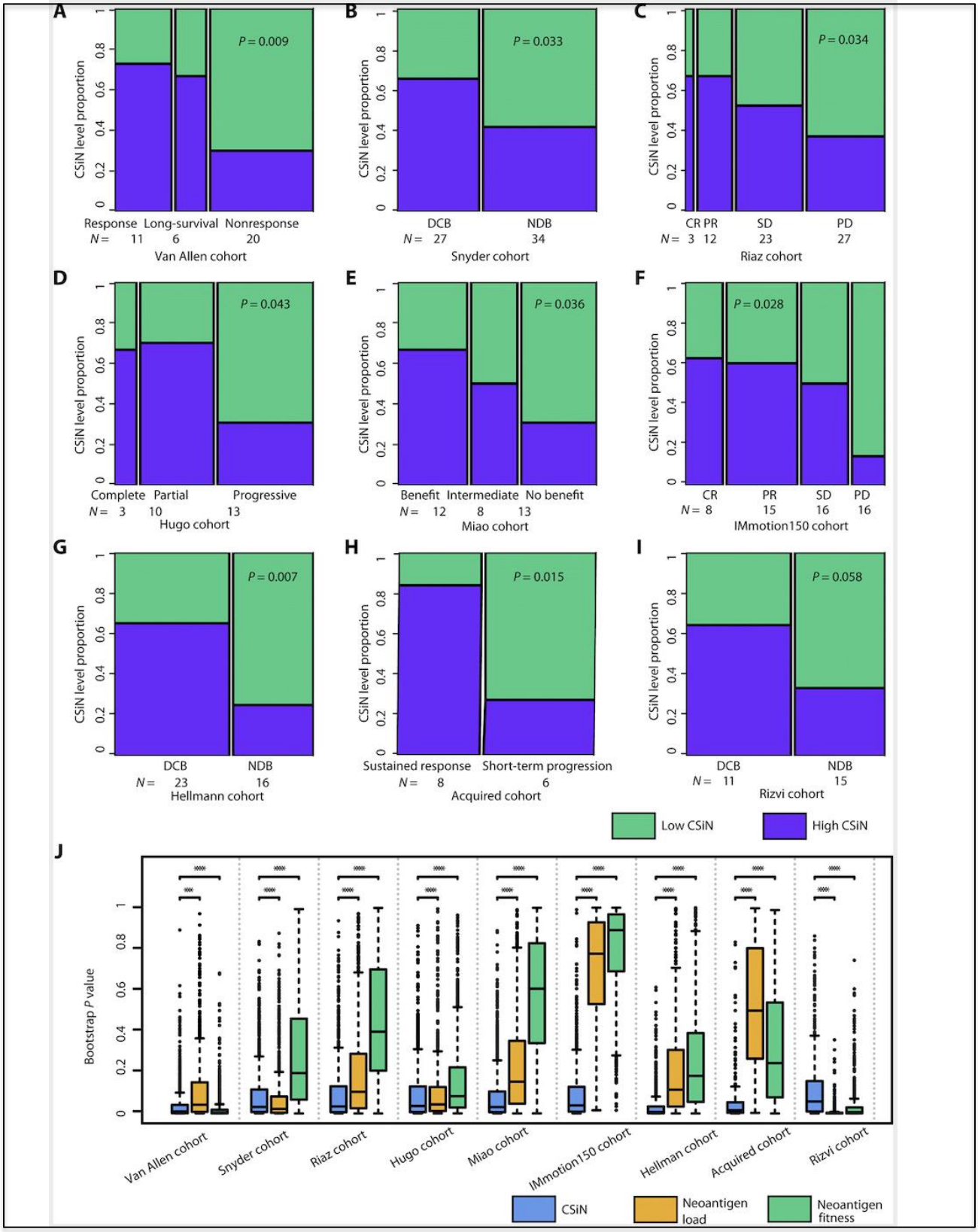

A
P = 0.009
CSiN level proportion
Response Long-survival Nonresponse
N = 11 6 20
Van Allen cohort
B
P = 0.033
DCB NDB
N = 27 34
Snyder cohort
C
P = 0.034
CR PR SD PD
N = 3 12 23 27
Riaz cohort
D
P = 0.043
Complete Partial Progressive
N = 3 10 13
Hugo cohort
E
P = 0.036
Benefit Intermediate No benefit
N = 12 8 13
Miao cohort
F
P = 0.028
CR PR SD PD
N = 8 15 16 16
IMmotion150 cohort
G
P = 0.007
DCB NDB
N = 23 16
Hellmann cohort
H
P = 0.015
Sustained response Short-term progression
N = 8 6
Acquired cohort
I
P = 0.058
DCB NDB
N = 11 15
Rizvi cohort
Low CSiN
High CSiN
J
Bootstrap P value
Van Allen cohort
Snyder cohort
Riaz cohort
Hugo cohort
Miao cohort
IMmotion150 cohort
Hellman cohort
Acquired cohort
Rizvi cohort
CSiN
Neoantigen load
Neoantigen fitness

Figure 2 Association of CSiN score with checkpoint inhibitor treatment response.

(A) The Van Allen cohort. Eleven patients with clinical benefit (response group), 6 patients with long-term survival (long-survival) group, and 20 patients with minimal or NDB (nonresponse) group. (B) The Snyder cohort. Thirty-seven patients with DCB, and 34 patients with NDB. (C) The Riaz cohort. Three patients with complete response (CR), 12 patients with partial response (PR), 23 patients with stable disease (SD), and 27 patients with progressive disease (PD). (D) The Hugo cohort. Three patients with complete response, 10 patients with partial response, and 13 patients with progressive disease. (E) The Miao cohort. TIlve patients with clinical benefit, 8 patients with intermediate benefit, and 13 without clinical benefit. (F) The IMmotion150 cohort. There are 8 patients with CR, 15 patients with PR, 16 patients with SD, and 16 patients with PD. These patients are treated with atezolizumab and have high Teff signature expression. (G) The Hellmann cohort. There are 23 PD-L1+ (IHC $\geq$ 3) patients with DCB, and 16 PD-L1+ patients with NDB. (H) The Acquired cohort. There are 8 patients with sustained response (progression < 12 month) and 6 patients with short-term progression (progression > 12 month). (I) The Rizvi cohort. Eleven patients with DCB and 15 patients with NDB. Biopsy and genomics data are obtained close to the time of progression for all patients, whereas baseline biopsies are lacking for many patients. For (A) to (I), I tested the association of the dichotomized CSiN scores with the ordered response categories using an ordinal $\chi 2$ test. (J) Boxplots of bootstrap P values evaluating the robustness of the predictive performance of CSiN, neoantigen load, and the neoantigen fitness score, with each P value generated from a bootstrap resample of each cohort. Two-sided Wilcoxon signed-rank test was used to compare the bootstrap P values. ***P = 0.0001 to 0.001 and ****P < 0.0001.

Higher CSiN score predicts more favorable prognosis in immunogenic cancers

To understand the implications of neoantigen heterogeneity for patients' long-term survival, I also investigated the association between CSiN and prognosis in the immunogenic tumor types including, RCC, LUAD, LUSC, and SKCM (Table 2). I focused on the patients with high levels of T cell infiltration, in which the neoantigen-T cell axis may have a more active impact on patients' phenotypes. The T cell infiltration was profiled by empirically defined tumor microenvironment gene expression signatures (Wang et al. 2018). In these patients, I observed that higher CSiN scores had a significantly positive association with better survival in patients with high T cell infiltration level for RCC (Figure. 3A, P=0.01), LUAD (Figure. 3B, P=0.036), LUSC (Figure. 3C, P=0.024), SKCM (Figure. 3A, P=0.038). I extracted and combined the higher CSiN scores and had a significantly better overall prognosis (Figure. 3E, P=3.8 x 10-5). To further exclude the effect of clinical confounders, I performed multivariate survival analysis adjusted by disease type, stage, gender, and age in this combined cohort. The significant association between survival and CSiN was retained (Figure. 3F, P<0.001).

Cohort ID	Disease type	Immunotherapy treatment	Raw data	Total # patients
RCC	Renal Cell Carcinoma	Not applicable	EGAS00001000509, TCGA, UTSW KCP	366
LUAD	Lung adenocarcinoma	Not applicable	TCGA	427
LUSC	Lung squamous cell carcinoma	Not applicable	TCGA	389
SKCM	Melanoma	Not applicable	TCGA	401

Table 2 The information of patient cohorts from The cancer genome atlas (TCGA)

In contrast, the same analysis for the neoantigen load and the neoantigen fitness models yielded insignificant associations. I also used bootstrap analysis to evaluate the statistical significance of this comparison. In Fig. 3G, the CSiN significantly outperformed both methods in all four

cohorts evaluated. Overall, the results suggest that the clonal distribution of neoantigens could be more prognostically important.

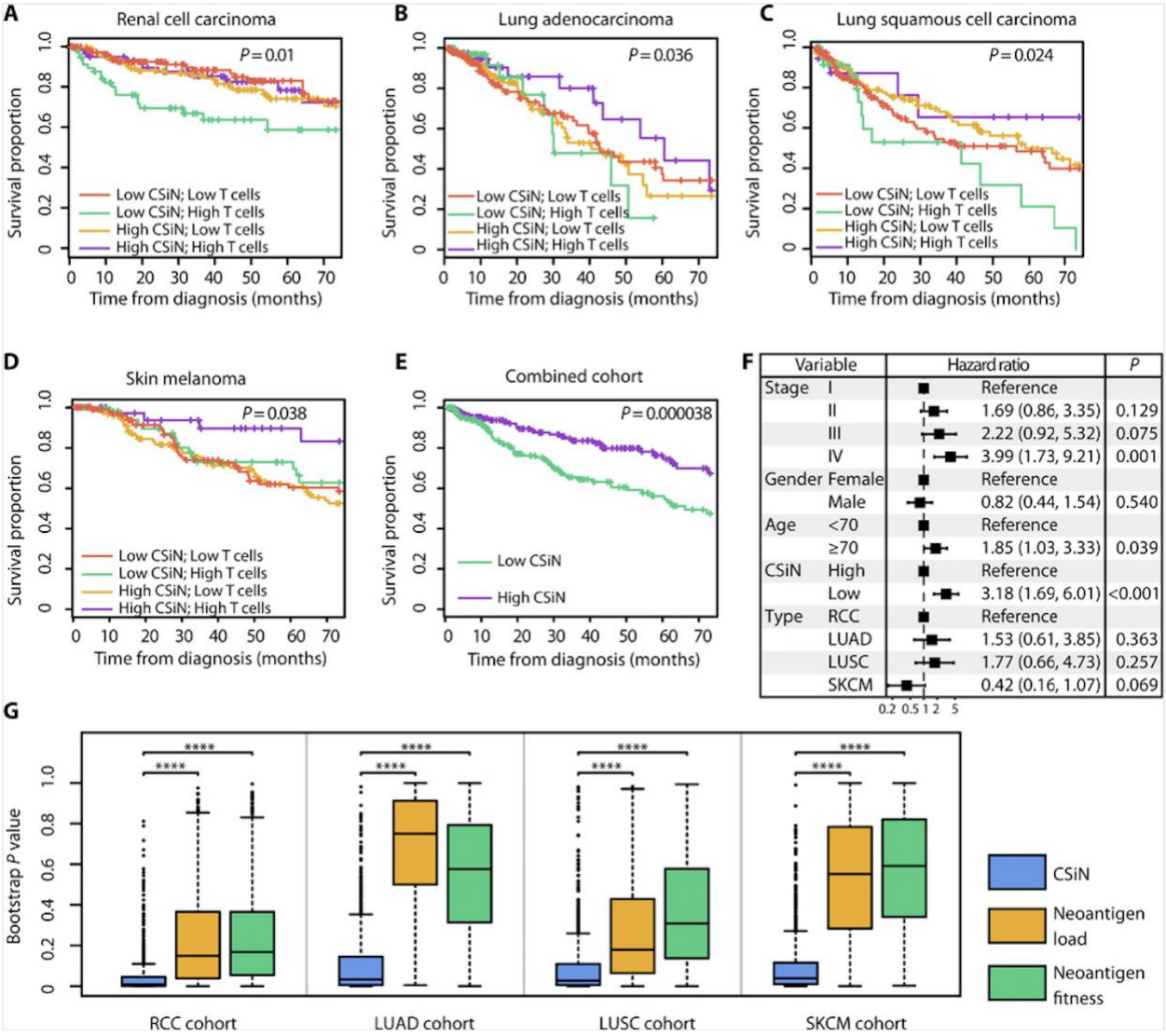


Fig. 3 Association of CSiN score with overall survival of patients.

(A to E) Kaplan-Meier estimator was used to visualize patient overall survival. P values for log-rank tests are shown. (A) The RCC cohort. (B) The LUAD cohort. (C) The LUSC cohort. (D) The SKCM cohort. (E) The patients identified as having “High T cells” are extracted from each cohort, combined, and tested together. The high and low CSiN score designations follow those in

(A to D). (F) Forest plot for the coefficients of the multivariate Cox proportional hazards analysis of the combined cohort in (D). Disease type, pathological stage, gender, age, and the binarized CSiN are included as covariates. The dashed line shows the no effect point. Confidence intervals (95%) Ire shown as bars. (G) Boxplots of bootstrap P values evaluating the robustness of the prognostic performance of CSiN, neoantigen load, and the neoantigen fitness score, with each P value generated from a bootstrap resample of each cohort. Two-sided Wilcoxon signed-rank test was used to compare the bootstrap P values. ****P < 0.0001.

Discussion

The major biological insight from this study is that the neoantigen clonal structures in each tumor specimen and the immunogenicity of the neoantigens are predictive of response to checkpoint inhibitors and prognosis. The comprehensive analyses show that the CSiN score, which describes the properties of the neoantigen profile quantitatively, has substantially better predictive and prognostic performance than other neoantigen-based biomarkers in most of the evaluated cohorts. The implementations of the CSiN, neoantigen load, the neoantigen fitness indices have considered both MHC class I and class II neoantigens also neoantigens generated from insertions/deletions and stop-loss mutations. This is different from the original publication of the neoantigen fitness model (Balachandran et al. 2017; Łuksza et al. 2017) that only considered 9-mer class I neoantigens generated from missense mutations. Inclusion of all these potential sources of neoantigens is important for a complete characterization of neoantigen profiles in each patient. In alignment with the findings in this study, McGranahan *et al. (McGranahan et al. 2016)* made a qualitative observation that CTLA-4--resistant tumors could be enriched for subclonal mutations, which may enhance total neoantigen burden but not elicit an effective antitumor response due to the subclonal nature of these neoantigens. Miao *et al.* (Miao

et al. 2018) also made a similar observation. This study is distinguished from these earlier reports in that I provide a robust quantitative measurement that was subjected to systematic evaluations, and I also evaluated prognosis in addition to treatment response. Overall, CSiN could serve as a valuable predictive tool for medical oncologists treating patients with checkpoint blockade and has addressed some of the limitations of prior neoantigen-based predictive biomarkers.

One limitation of this study is that neoantigens used in this study are predicted from genomics data for correlation with patient phenotypes. Despite the efforts to validate the neoantigen predictions, it is likely that there are still false-positive and false-negative predicted neoantigens that convoluted the analysis. In future studies, incorporating the genomics-based approach with other methods, such as mass spectrometry, may improve the sensitivity and specificity of neoantigen detection and thus further enhance the predictive power of CSiN.

Overall, CSiN offers a new tool to monitor the neoantigen profiles, where different tumor clones could have different growth advantages subject to the pressure of T cell cytotoxicity determined by each clone's neoantigen composition. This work offers a rigorous methodology of predicting response to immunotherapy and prognosis from routing patient samples and could be useful for personalizing medicine in immunotherapy.

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

Background and rationale

Another one of the most fundamental and unsolved questions regarding neoantigens and antigen biology is general is the lack of understanding of why not all neoantigens elicit T cell responses (immunogenic), notwithstanding that they are expressed and presented on the cell surface. In neoantigen vaccine trials, the reported immunogenicity rate ranges from 16%-66% (Linette & Carreno 2017). In adoptive cell transfer experiments by Verdegaal et al (Verdegaal et al. 2016), T cells from two patients with hundreds of somatic mutations only exhibited immunogenicity toward a few predicted neoantigens. Despite the differences in experimental methods and biological systems, all these observations point to an urgent need for discerning truly immunogenic neoantigens.

Even less is known about the TCR binding specificity to immunogenic neoantigens presented by MHC molecules (pMHCs). Linking 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 (Altman et al. 2011), TetTCR-seq (Zhang et al. 2018) and T-scan (Kula et al. 2019), have been developed to detect pairing of TCRs and pMHCs. However, these methods are time-consuming, technically challenging, and costly. Furthermore, each technique has some caveats. Ito el al 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.

Method

Conceptually, I employed a staged approach of dividing the goal of learning the TCR-binding specificity of neoantigens (pMHCs) into three steps, to lower the difficulty level of the prediction task. First, I trained a numeric embedding of pMHCs (class I only) using Long short-term memory (LSTM) network so the protein sequences of neoantigens and MHCs could be represented numerically. Second, I 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. The advantage of using the embeddings of TCR CDR3βs and MHC peptides as the model input instead of gene names is that a new gene's name (e.g. a new MHC allele) unknown to the model during the training phase cannot be handled in testing/prediction, while embedding their protein sequences allows their testing/prediction. At the final stage, I 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. I employed fine-tuning to finalize the prediction model for the pairing between TCRs and pMHCs.

Embedding TCR CDR3β sequences

The method for encoding TCR CDR3β sequences can be found in Zhang *et al. (Zhang et al. 2021)*.

Embedding pMHCs

The embedding of pMHCs mostly follows the netMHCpan algorithm (Figure 4a). The netMHCpan algorithm uses a pseudo sequence method to encode the MHC proteins (Nielsen et al. 2007). 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 are also encoded by the BLOSUM50 matrix as in netMHCpan. I constructed a deep learning model with the HLA pseudo sequence and the antigen sequence as the input. I 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 the implementation from the original netMHCpan model is that, instead of simple feed-forward neural networks, I 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. I found this change to seem to have increased the speed of reaching model convergence. 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. The same data used for training netMHCpan Ire used to re-train the model, which consists 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, which is comparable with the Pearson Correlation of 0.76 from the original netMHCpan publication (Nielsen & Andreatta 2016). After training is completed, I extracted the immediate 60-dimensional fully connected layer before the single-neuron output layer (again a short numeric vector), as the embedding of pMHCs.

I have also tried to use a feed-forward neural network for encoding pMHCs but the performance I achieved (Pearson Correlation=0.72) is worse than the LSTM network that I finally adopted (Pearson Correlation=0.781, Figure 4b). This is likely because LSTM is quite powerful in digesting sequential data, due to its inherent design and capability to handle recurrent structures. In fact, LSTM networks have been widely used for encoding protein sequences and predicting protein functions (Gao et al. 2020; Guo et al. 2019; Liu & Gong 2019).

Learning TCR binding specificity of pMHCs

Finally, I leveraged the trained numeric vector encodings of TCRs and pMHCs for learning the pairing between them. I 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. I 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 (Figure 4c). I 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.

Based on this integrated model, I 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. I collected a total of 32,607 pairs of binding TCR-pMHCs from a series of peer-reviewed publications (Bagaev et al. 2020; Chen et al. 2017; Glanville et al. 2017; Huth et al. 2019; Joglekar et al. 2019; Tickotsky et al. 2017; Unable to find information for 7266116; Zhang et al. 2018) (N=13,388), and four Chromium Single Cell Immune Profiling

Solution datasets (N=19,219). The details of these data are shown in Sup. File 1. Some databases provided quality metrics, which I used to filter the records to keep only pairs with high confidence. For example, in the VDJdb data, I only included records with vdj.score>0, as is also done in TCRGP (Jokinen et al. 2021). Duplicated records were removed. I created 10 times more negative pairs, by random mismatching TCR and pMHC of these 32,607 pairs. The training was performed for 150 epochs. The loss function of the internal validation decreased smoothly, and the loss function of the independent validation set stumbled but closely followed the decreasing trend, demonstrating a good dynamic of the training of model parameters (Figure 1d). The final model was named as, pMTnet for pMHC-TCR binding prediction network. Following the 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. I use a smaller rank to denote a stronger binding, similar to netMHCpan. Importantly, as I 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.

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 are treated as positive data. And I 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, I developed a novel differential training method to instruct pMTnet to distinguish binding TCRs from non-binding TCRs through comparison. To implement this, I 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 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*), I sampled 10,000 TCR sequences randomly from the databases to form a background distribution, {Tb}. I will calculate the percentile of f(p*,T*) in the whole distribution of {f(p*,Tb)}, where T* is the TCR of interest. The larger this value, the stronger I 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), I also inverted this rank. Therefore, in the final output, a smaller rank between a pMHC and a TCR refers to a stronger binding prediction between them.

Regarding the duplicated-network approach, one could argue that if I simply feed positive and negative pMHC-TCR pairs of the same pMHC together in one mini-batch, the training could also be performed. However, this approach would not work, as it precludes us from explicitly

comparing the positive and negative TCRs of the same pMHC in a well-controlled manner. This is reflected in the loss function above. In this loss function, there is a contrast between the positive output and the negative output (the first part), which is transformed together by a RELU function, and there is also a penalty to regularize the range of both outputs (summation), to limit the outputs to the best dynamic range. The mini-batch approach cannot achieve this effect of delicate control. I have trained the model with this mini-batch approach, and the AUCs of ROC and PR are, 0.604, and 0.397, respectively, on the 619-test cohort.

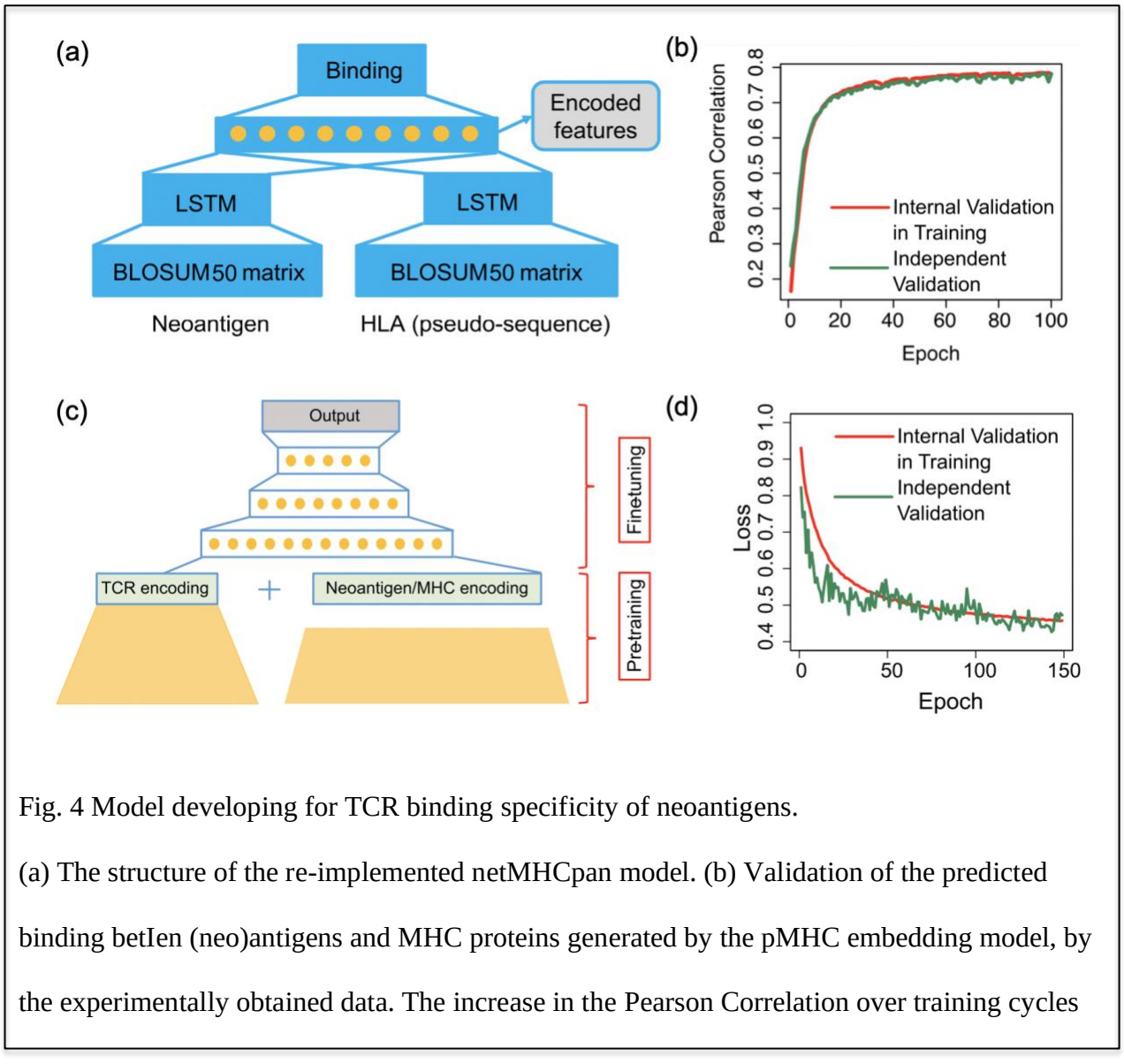


Fig. 4 Model developing for TCR binding specificity of neoantigens.

(a) The structure of the re-implemented netMHCpan model. (b) Validation of the predicted binding betIen (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. (c) Structure of the final pMTnet model. (d) 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.

Results

pMTnet predicts TCR-pMHC pairing in independent experimental data

I performed a series of validation analyses with a large number of known TCR-pMHC binding pairs collected from independent studies. Data was hidden during the training, and pMTnet was no longer modified when validation was performed with this data.

First, I collected 619 experimentally validated TCR-pMHC binding pairs (Attaf et al. 2018; Berger et al. 2011; Borbulevych et al. 2011; Bourcier et al. 2001; Brennan et al. 2007; Burrows et al. 1995; Cole et al. 2014, 2017; Gee et al. 2018; Grant et al. 2016; Klinger et al. 2015; Kløverpris et al. 2015; Lee et al. 2004; Leslie et al. 2006; Lichterfeld et al. 2006; Liu et al. 2013; Motozono et al. 2014; Ogunshola et al. 2018; Ott et al. 2018; Purbhoo et al. 2007; Shimizu et al. 2013; Tran et al. 2015; Unable to find information for 7598048; Valkenburg et al. 2016; Yu et al. 2007). 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. I used two metrics, Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) and Precision-Recall (PR). Values closer to 1 indicate better performance. Strikingly, the AUC of ROC reached 0.827

in this cohort and AUC of PR reached 0.565 (Figure. 5a). To test whether pMTnet truly "learned" the features that determine binding, or is simply "remembering" pairing cases, I looked at the prediction performance for TCRs with different degrees of similarity to the training TCRs (Figure. 5b, left group). For calculating "similarity", I calculated the minimum of each testing TCR's Euclidean Distances to all the training TCRs based on the TCR embeddings. 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, I performed the same analyses, and made similar observations (Figure. 5b, right group).

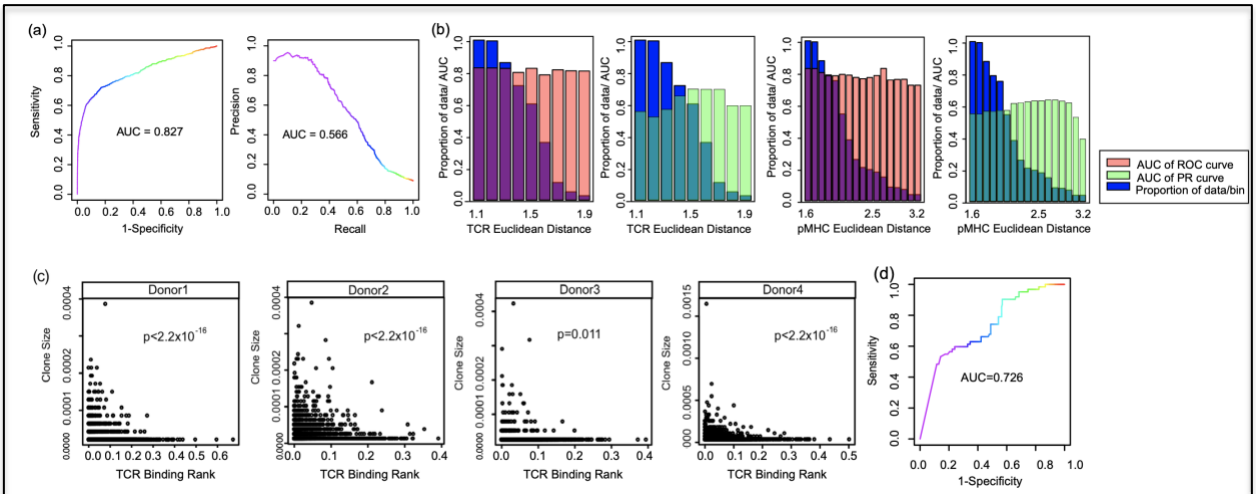


Figure 5. Validation of pMTnet. (a) AUCs of Receiver operating characteristic (ROC) and precision-recall (PR) of the predicted binding ranks (smaller ranks refer to stronger binding) were shown for the 619 experimentally validated TCR-pMHC binding pairs and 10 times more randomly shuffled negative pairs. (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. 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.

I also compared the performance of pMTnet with the other software that can predict TCR/epitope pairing, including netTCR, TCRex(Gielis et al. 2019), and TCRGP (Jokinen et al. 2021; Unable to find information for 6439084). 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.

I also validated pMTnet on additional high quality pairing data from VDJdb and Gee et al (Gee et al. 2018) that are not used during the training, and showed that the AUROC of pMTnet achieved >0.8 on them, and out-performed competing software. Interestingly, in these analyses, I took advantage of the differential quality of the binding pairs to show that the performances of

pMTnet and competing software all increased on validation data of higher quality (determined through objective criterion, such as confidence score curated by VDJdb).

Next, I 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. I 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, I recorded the pMHC with the strongest predicted binding strength, by pMTnet, among all 44 pMHCs. Interestingly, I found the clone sizes and predicted ranks for T cell clonotypes were negatively correlated with statistical significance achieved (Figure. 5c). 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 (Table 3) testing enrichment of the expanded T cell clonotypes with high affinity binding antigens. Conversely, I 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.

	Donor	Odds Ratio	CI(95%)	
	Donor1	48.61	31.70	74.55
	Donor2	6.35	3.32	12.16
	Donor3	57.49	22.73	145.41
Rank Cutoff :0.5%	Donor4	26.49	15.65	44.85
	Donor1	55.70	39.43	78.67
	Donor2	7.71	4.44	13.40
	Donor3	42.93	18.37	100.29
Rank Cutoff :1%	Donor4	24.13	15.09	38.57
	Donor1	67.03	48.65	92.34
	Donor2	9.60	5.69	16.19
	Donor3	39.03	18.26	83.45
Rank Cutoff :2%	Donor4	36.13	23.30	56.03

* The Odds ratio is calculated as:

$$\frac{\#\textit{expanded clones with high affinity binding}}{\#\textit{expanded clones with low affinity binding}} \times \frac{\#\textit{unexpanded clones with low affinity binding}}{\#\textit{unexpanded clones with high affinity binding}}$$

Table 3 The T cell clonotypes that have expanded are strongly biased towards the T cells with predicted high affinity antigens, in the 10X single cell datasets. An odds-ratio and its confidence interval is shown for each of the four donors. The cutoff for defining expanded/unexpanded clones is the top 1% clone size of the TCR clones of each donor, and the cutoff for defining high/low affinity is 0.5%/1%/2%.

I further analyzed whether pMTnet is capable of distinguishing the impact of the fine details of peptide sequences on TCR binding specificity. 186 pMHC-TCR pairs are acquired from Liu et al (Liu et al. 2013), Cole et al (Cole et al. 2014), and Tran et al (Tran et al. 2015). 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 Iaker 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 (Figure. 5d, AUC=0.726).

I further validated pMTnet in prospective experimental data. I 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. I 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). I found that TCRs predicted to have stronger binding (smaller ranks) to any of these peptides exhibited higher clonal proportions than the other TCRs (Figure. 6a), in both the blood (left panel) and in vitro expanded T cells (right panel). I 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. I observed a stronger enrichment in both the blood and expanded T cells, while I performed permutations of the predicted binding ranks and observed much smaller odds ratios (Figure. 6b). Then I treated the expanded T cells with each of the viral peptides and performed scRNA-seq with paired TCR-seq, and I also performed vehicle treatment. I 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. I 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, I observed differentially expressed genes enriched in pathways

essential for T cell proliferation, migration, survival, and cytotoxicity (results for GLCTLVAML shown in Figure. 6c as an example). I 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 (Figure. 6d, clonal size ratio >1).

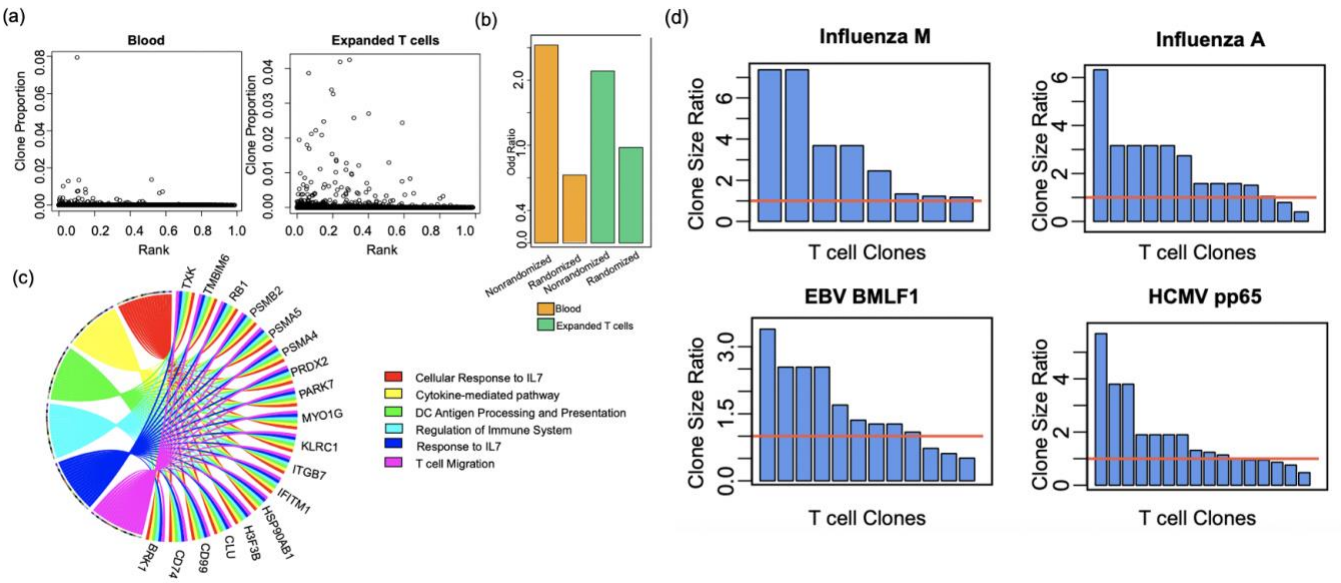


Figure 6 Prospective validation of pMTnet predictions. (a) TCR CDR3s predicted to have smaller binding ranks have higher clonal sizes. Blood cells: left panel and *in vitro* expanded T cells: right panel. X-axis shows the minimum of the binding ranks to any of the four viral pMHCs. Y-axis shows the clonal 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. I extracted the #CDR3s with clonal proportions>0.1% and with predicted rank<2% (H_B); #CDR3s with clonal proportions<0.1% and predicted rank>2% (L_S); #CDR3 with clonal proportions>0.1% and predicted rank>2% (L_B); #CDR3 with clonal proportions<0.1% and predicted rank<2% (H_S). Odds ratios are calculated as (H_B *L_S)/(L_B *H_S). 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.

Structural analyses support predicted TCR-pMHC interactions

I 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, I 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 (Weiss et al. 2000). I 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. I divided each TCR CDR3 into six segments of equal lengths (Fig. 7a), and as expected, residues in the middle segments of CDR3s, which bulge out and are in closer contact with pMHCs, Ire 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, I extracted a total of 13 TCR-pMHC pairs from the IEDB, with 3D crystal structures available in Protein Data Bank (PDB) and whose predicted binding affinity rank was less than 2%. According to the structures, I grouped CDR3 residues by whether or not they formed any direct contact with pMHCs residues within 4Å. I found that the contact residues were more likely to induce larger changes in predicted pMHC binding strength than non-contact residues (Fig. 7b, P value=0.036). I also performed in silico

alanine scanning and found a similar trend (Fig. 7c). 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). HoIver, 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. 7a-c, I 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. The P values for the contact vs. no contact comparisons are relatively large and around the borderline cutoff of 0.05. I believe the small sample size, noises in the structure data and imperfection of the predictions all contributed to the relatively large P values.

In Fig. 7de, I showed an example TCR-pMHC structure with the PDB id of 5hhm, generated by Valkenburg et al. Overall, I found that R98 and S99 had the biggest differences in predicted ranks after the “0-setting” scan (Fig. 7d, upper panel) and alanine scan (Fig. 7d, 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. I observed that S95 formed intra-chain contacts with the small loop formed by Q103 and the side chains of E102 and Y104.

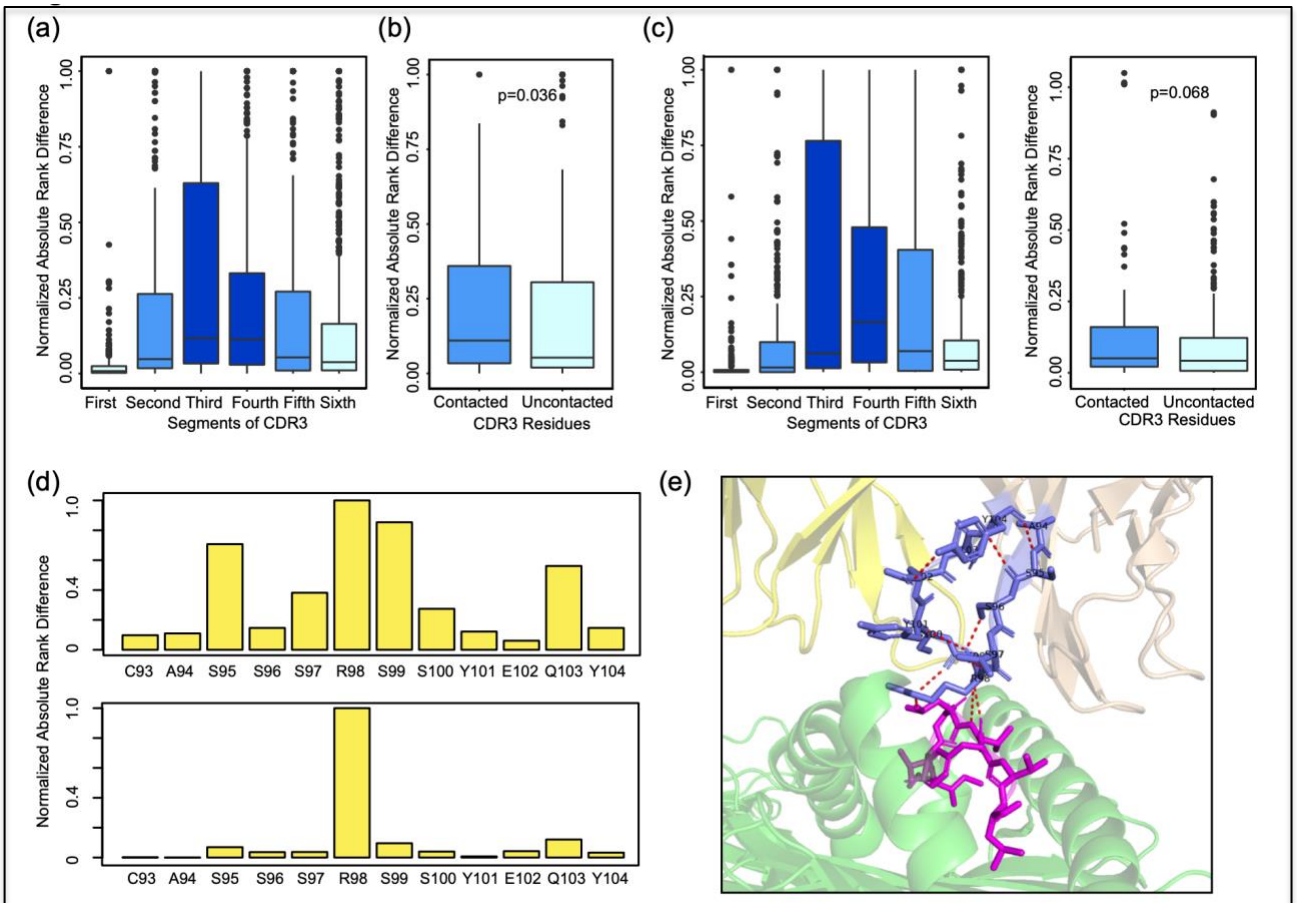


Figure 7 Structural analyses support the predicted TCR-pMHC interactions. (a) Residues in the middle segments of CDR3s are more likely to induce larger changes in predicted binding affinity. I divided each TCR CDR3 into six segments of equal lengths, and plotted the normalized changes in predicted binding 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.

Characterizing the TCR-pMHC interactions in human tumors

To further validate pMTnet and demonstrate the value of pMTnet as a knowledge discovery tool, I 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(Wang et al. 2018). I analyzed the genomics data of The Cancer Genome Atlas (TCGA) and the in-house Renal Cell Carcinoma (RCC) data (Wang et al. 2018). TCGA patients included lung adenocarcinoma patients (LUAD) (Cancer Genome Atlas Research Network 2014), lung squamous cell carcinoma patients (LUSC) (Cancer Genome Atlas Research Network 2012), clear cell renal cell carcinoma patients (KIRC) (Cancer Genome Atlas Research Network 2013) and melanoma patients (SKCM) (Cancer Genome Atlas Network 2015).

I 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 (Lo et al. 2014). 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 (Cherkasova et al. 2016). 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 ansIr this question, I identified candidate neoantigens and self-antigens from the genomic data. For RCCs, I profiled the expression of this very Ill characterized HERV-E found by Cherkasova et al. (Cherkasova et al. 2013) . The TCRs are called by Mixcr from the TCGA fresh frozen RNA-seq data, with an average of 25 unique CDR3 sequences per patient. In each patient sample, I 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 loIr than each one of a series of cutoffs between 0.00% and 2% (otherwise, this TCR will be unassigned). I note that Bolotin et al has confirmed, via TCR-sequencing data, the ability of Mixcr to extract all relatively large TCR clonotypes with consistent clonal frequencies from RNA-seq data (Bolotin et al. 2017). Admittedly, very small TCR clonotypes will be missed by Mixcr, but these should account for a minor proportion of the whole repertoire (Bolotin et al. 2017; da Silva et al. 2017).

For each patient sample, I 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. 8a shows the total and immunogenic antigen numbers for one example RCC patient. Then for all patients of all cancer types, I 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. I observed that neoantigens are generally more immunogenic than self-antigens (higher proportions of neoantigens are predicted to bind TCRs) (Fig. 8b). This is fitting because neoantigens, unlike self-antigens, are mutated peptides that have not been

encountered by T cells during the developmental process. However, I 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 (Cherkasova et al. 2016).

Next, I examined the impact of TCR-pMHC interactions on the clonal expansion of T cells. For each patient, I compared the clonal fractions of TCRs (#specific TCR clonotype/#all TCRs) that Ire 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. 8c), I 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, I 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, I observed that more and more patients demonstrated clonal expansion of their antigen-targeting T cells compared to other T cells (Fig. 8d), with smaller and smaller rank percentile cutoffs (stronger affinity) to define antigen-TCR pairing. Consistent with Fig. 6, this result also shows that more immunogenic tumor antigens induce stronger T cell clonal expansion in human tumors.

Finally, I 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=8.1x10-9).

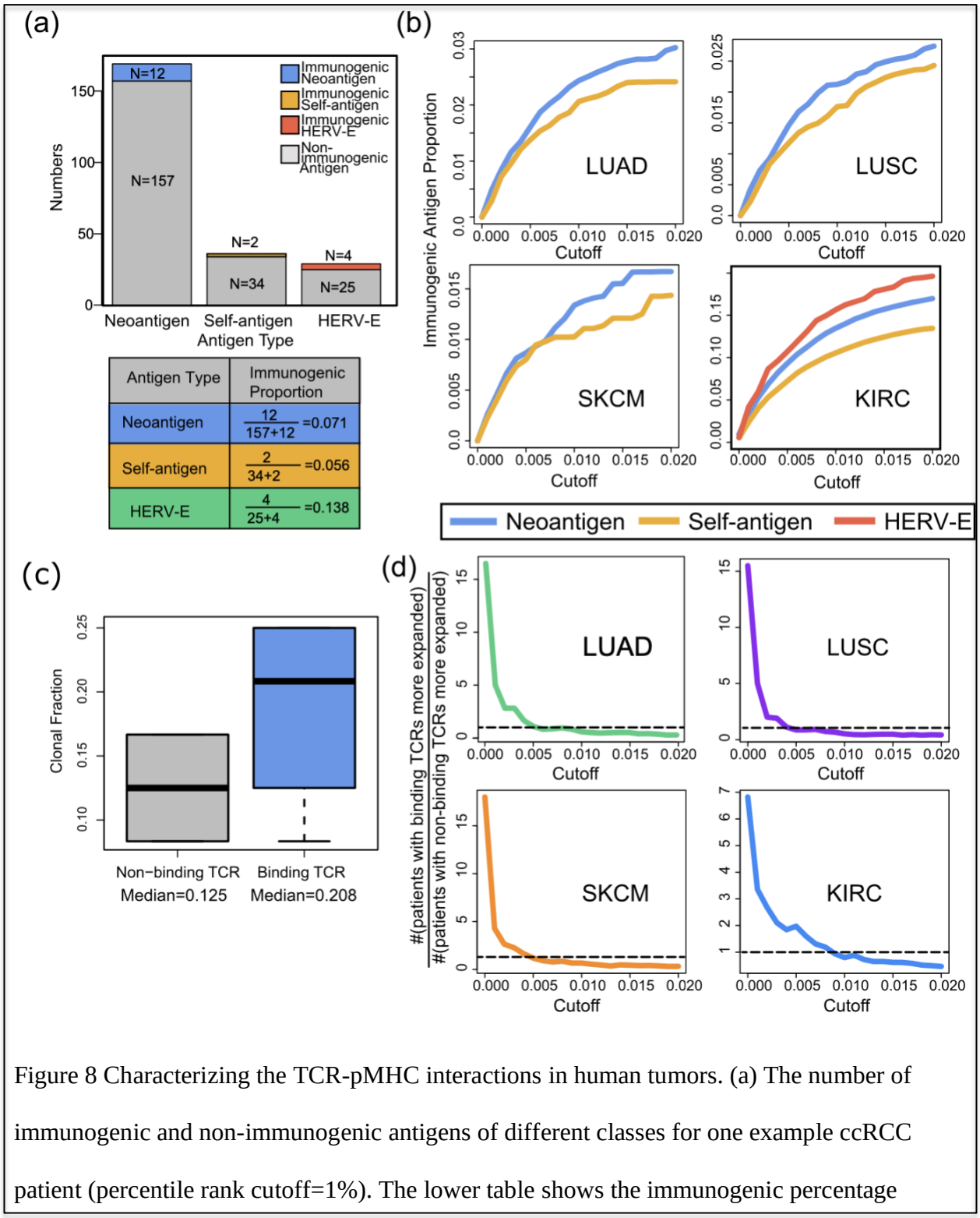


Figure 8 Characterizing the TCR-pMHC interactions in human tumors. (a) The number of immunogenic and non-immunogenic antigens of different classes for one example ccRCC patient (percentile rank cutoff=1%). The lower table shows the immunogenic percentage

calculation process for this patient. (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.

Discussion

This 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. 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.

Furthermore, a suite of genome-wide analyses was now enabled by pMTnet, which has revealed interesting biological discoveries. This 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 (Gee et al. 2018). Surprisingly, three of the four receptors recognized unmutated self-antigens. Consistent with the observations of Gee et al in a limited number of patients, I 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.

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 the training dataset. Admittedly, the training dataset collection has many common epitopes such as those Ill studied ones from CMV. In the future, I 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, I 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 my work. I expect pMTnet to propel tumor immunogenomics research and also to enhance the design and implementation of immunotherapy in the modern era of personalized medicine.

CHAPTER FOUR
Inferring the evolution of neoantigen-T cell interactions in tumors

Background and rationale

Neoantigens are key markers for the recognition of T cells, and the binding of T cells with neoantigens initiates their tumor cytotoxic effects. Unfortunately, researchers within the field are far from clearly understanding neoantigens' impact on tumorigenesis, prognosis, and treatment response. An elucidation of how neoantigens participate in past tumor evolution has been absent but could give us a sneak peek into the behavior of the tumors in the future, particularly their response to immunotherapies.

The survival fitness of cancerous cells diminishes when mutations within tumor DNA arise that give way to neoantigens presented on the surface of these tumor cells. In a tumor microenvironment with actively-infiltrating T cells, these mutations will be recognized and the tumor cells bearing them will be selected against during evolution. With constant external immune selection pressure, the numbers of neoantigens generated by newly occurring tumor somatic mutations are expected to stay constant over the course of tumorigenesis. When anti-tumor T cell immunity is strong, it is anticipated that the mutations that generate more neoantigens will be more strongly selected against. On the contrary, when there is not enough T cell infiltration or there is functional exhaustion/inhibition of the T cells, selection pressure will be substantially lessened for tumor cells with mutations of high neoantigen counts. Thus, ascertaining the dynamics of neoantigen distributions throughout molecular time can reveal the evolutionary history of the anti-tumor immune pressure during tumorigenesis.

To achieve this task, it is critical to time the genetic events that happened to the tumors. Tumors at the time of diagnosis often consist of heterogeneous clones (Deshwar et al. 2015; Miller et al. 2014; Roth et al. 2014), each with a unique set of somatic mutations sharing similar variant cellular prevalence. The tumor clones can be easily detected through the clustering of mutations

via algorithms such as PyClone (Roth et al. 2014), PhyloWGS (Deshwar et al. 2015), and SciClone (Miller et al. 2014). However, detection of the developmental time-ordering of these clones is a much harder problem. Each child clone is grown from a tumor cell within the parent clone, due to the occurrence of a tumor-driving event, along with possible passenger mutations also taking place. One parent clone may yield two or more child clones. Due to the large potential search space, it is difficult to reliably order the clones into a phylogenetic tree of parent-child relationships. Also, the clonal size is an unreliable indicator of the appearance times of the tumor clones, due to sampling bias and the fact that different clones have different proliferating potentials. However, the prevalences of the clones can help distinguish which mutations occurred earlier from those that occurred later.

I employed an innovative approach of treating the intra-clone cellular prevalences of the somatic mutations as a surrogate of a molecular clock within each tumor clone, and developed a Bayesian hierarchical model, named netie, to infer the evolution of neoantigen-$CD8^+$ T cell interactions in tumors by sampling from different clones. Netie is systematically validated by a series of simulation studies and real human tumor data. We utilized netie to evaluate 3,211 tumors of 18 cancer types, and provided the first pan-cancer landscape of the impact of neoantigens on tumors' molecular phenotypes, prognosis, and treatment response to immunotherapies. While most prior studies of neoantigens focus on immunogenic tumors (Lu et al. 2020; Łuksza et al. 2017), such as lung cancer, we also showed an effect of neoantigens on non-immunogenic tumors using netie. Our work achieved an understanding of how neoantigens participate in tumorigenesis and how they impacted the molecular makeup of the tumors, which is neglected by most other works on neoantigens. Translationally, netie revealed a curious synergy between neoantigen distributions and T cell infiltrations for the prediction of patient prognosis and

treatment response to immunotherapies, which advocates for development of future combo biomarkers consisting of these and other potential components.

Method

Each patient is modeled separately. For each patient, there are $S \geq 1$ samples. Let c denote the tumor clones inferred by PyClone for $c = 1, ..., Cs$, for each sample s. Let $a_{s,c}$ denote the anti-tumor immune selection pressure for each clone in each sample. Note that some clones in different samples are essentially the same clone, sharing a very similar set of mutations and similar ranks of variant allele frequencies. Thus, their $a_{s,c}$ values (see below) should be the same. For convenience of notation, we will record them as different clones in different samples. For each clone, we have $k = 1, ..., Ks,c$ mutations. The prevalence of each mutation is $v_{s,c,k}$, which is bounded between 0 and 1. For each mutation, we have the associated number $n_{s,c,k}$ of neoantigens generated from that mutation. In our neoantigen calling pipeline, we only consider mutations with VAF > 0.05, as only these mutations' neoantigens have been called. We only consider clones with at least two mutations and at least one mutation with neoantigen count > 0.

For patients with multiple samples, we define a function, $\varphi(s, c) = 1, 2, ..., \Phi$, to denote whether the different clones of the different samples are actually the same clone. If the $\varphi(s, c)$ returns the same value, they are actually the same clone. Thus, $a_{s,c}$ should be treated as an alias of $a_{\varphi(s,c)}$. If a mutation has variant allele frequency (VAF) > 0.05 in any sample, we will consider it. In some samples, this mutation's VAF may be < 0.05, and thus will not have neoantigens called for that sample. But this mutation's neoantigens should still be considered and its neoantigen count can be obtained from the samples in which this mutation has VAF > 0.05.

Bayesian hierarchical model

Given the observed neoantigen count data and the prevalence of mutations, we aim to infer the anti- tumor selection pressure as,c for each clone and a for the whole tumor. There are two scenarios for modeling: only one unique clone detected across all samples, and more than one clone detected across all samples.

First scenario: only one unique clone detected (degenerate case)

I assume a zero-inflated Poisson distribution to model the number of neoantigens generated for each mutation. That is,

$$\Pr(n_{s,c,k} = y, z_{s,c,k}|\pi, a_{s,c}, b_{s,c}) = \pi \mathrm{I}(z_{s,c,k} = 0)\mathrm{I}(y = 0) + (1 - \pi)\frac{\lambda_{s,c,k}^{y} e^{-\lambda_{s,c,k}}}{y!}\mathrm{I}(z_{s,c,k} = 1)$$

$\lambda_{s,c,k}$ is the expected number of neoantigens of the non-zero inflation part, as a function of the time in the history of the tumor development. The "time in history" is inferred by the prevalence of each mutation, $v_{s,c,k}$, which serves as a molecular clock surrogate. In particular, we assume

$$\lambda_{s,c,k} = \exp(a_{s,c} v_{s,c,k} + b_{s,c})$$

We let $z_{s,c,k}$ donate whether the mutation comes from the first component ($z_{s,c,k}=0$) with probability , or the second component ($z_{s,c,k}=1$). In other words, $\Pr(z_{s,c,k}=0|\pi)=\pi$

Second scenario: more than one unique clone detected

We still assume the same zero-inflated Poisson distribution as in the first scenario. However, different clones should share some similarity in properties. Therefore, we assume a hierarchical structure; for $a_{\varphi(s,c)}$, we have

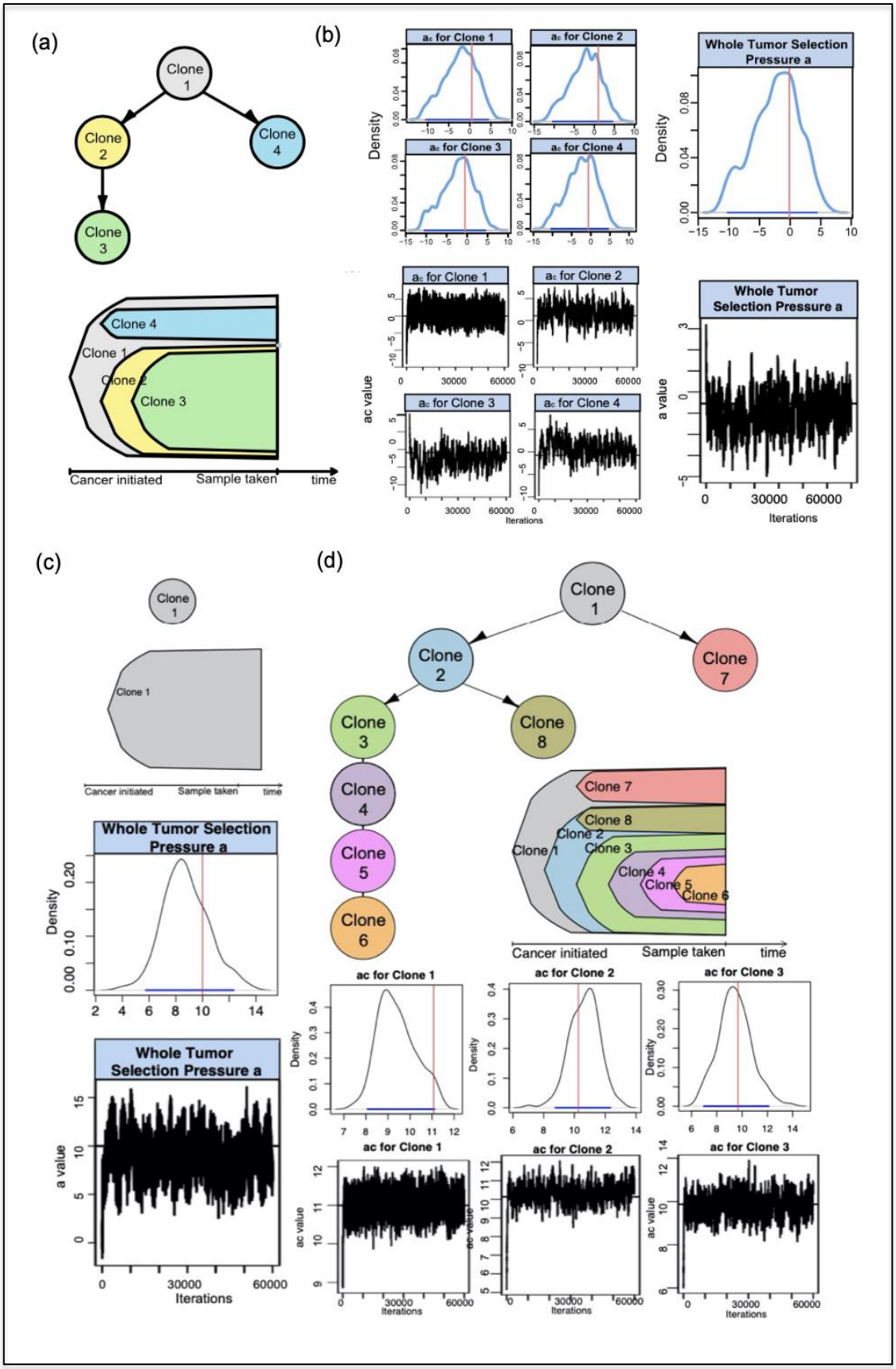
$$a_{\phi(s,c)} = a + \epsilon_{\phi(s,c)}$$

where $\varepsilon_{\varphi(s,c)} \sim N(0, \sigma a2)$ and $\sigma a2$ is a predefined positive number.

Results

Netie is validated by simulation studies

We applied netie to three simulated tumor samples. The first sample had four clones and 100 mutations (Fig. 9a); the second sample had one clone and 100 mutations (Fig. 9c); the third sample had eight clones and 400 mutations (Fig. 9d). Cellular prevalences from the same clone were sampled from normal distributions with the same means and variances. The number of clones and mutations, prevalences, and the clonal structures were simulated to be comparable to those observed in typical real human tumors. The performance of netie was evaluated with respect to the estimation of each variable. For the estimated immune selection pressure “ac” and overall “a”, we compared the posterior estimates with the ground truths of the simulation. The true values were all located within the 95% highest posterior density interval (Fig. 9b), meaning that netie has correctly inferred the trend of variation in immune selection pressure. In Fig. 9b, the traceplot shows the sampled a and ac at each MCMC iteration. The fluctuations of the sampled variables around stable values (minimum upward or downward shift in average) represent a good dynamic of convergence. The potential scale-reducing factors for all the inferred parameters are less than 1.1, which also demonstrates that MCMC converged (Fig. 9e) (Kulmon 2021; Li et al. 2021). All of these indicate dependable performance characteristics of netie.


(a)
Clone 1
Clone 2
Clone 3
Clone 4
Clone 4
Clone 1
Clone 2
Clone 3
Cancer initiated
Sample taken
time
(b)
ac for Clone 1
ac for Clone 2
ac for Clone 3
ac for Clone 4
Density
Whole Tumor Selection Pressure a
ac value
a value
Iterations
(c)
Clone 1
Cancer initiated
Sample taken
time
Whole Tumor Selection Pressure a
Density
Whole Tumor Selection Pressure a
a value
Iterations
(d)
Clone 1
Clone 2
Clone 3
Clone 4
Clone 5
Clone 6
Clone 7
Clone 8
Cancer initiated
Sample taken
time
ac for Clone 1
ac for Clone 2
ac for Clone 3
Density
ac value
Iterations

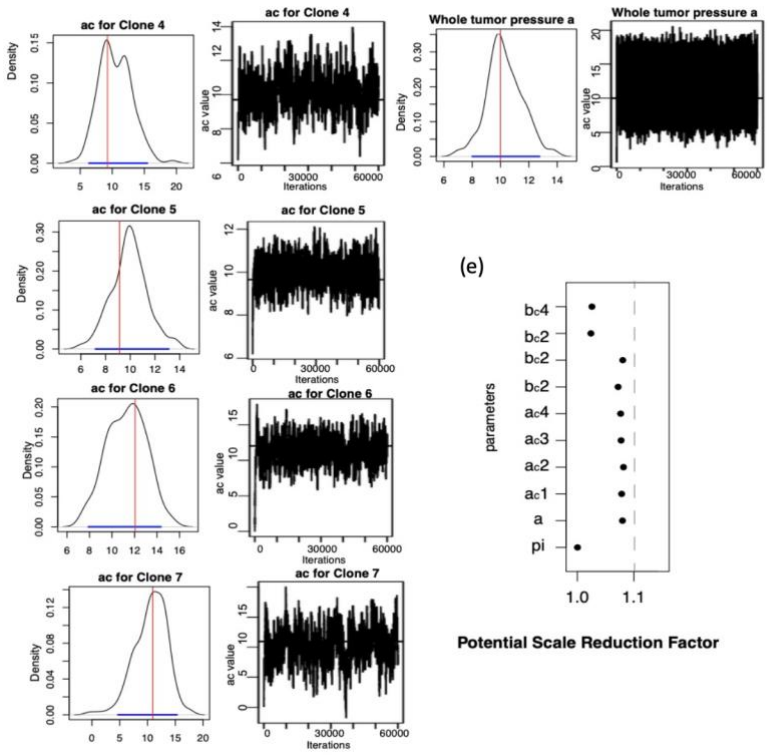


Figure 9. Applying netie on simulation data. (a) The setup of the simulation data, where the assumed clones and their parental relationships were shown. (b) The posterior density curves of the random variables to be estimated, with the 95% highest posterior density intervals presented by blue bars on the x-axes. The vertical red lines are located at the true assumed values. Trace plots showing the convergence of the netie estimates of the random variables around the true values, throughout the MCMC iterations. (c-d) Two more simulation datasets. The same simulation and analysis procedures, as in Figure 9b, were carried out. (e) The potential scale reducing factors (PSRFs) for all the inferred variables of the simulation dataset in a. "*ac*" is the inferred trend of change in anti-tumor selection pressure for each clone. "*bc*" and "pi" are the posterior estimates of the other variables in the Bayesian model.

Patients with increasing immune pressure demonstrate stronger T cell activation

I also validated netie in real data through demonstrating its power in revealing biologically meaningful signals. I applied it to The Cancer Genome Atlas Program (TCGA) and kidney cancer patients from our in-house UTSW Kidney Cancer Program cohort (Wang et al. 2018). We included a total of 17 cancer types in the study. Netie analyses were successfully performed on 2,545 patients' genomics data. Some samples with available genomics data were lost in the analysis pipeline for a number of reasons, such as no somatic mutations nor neoantigens detected, or failure of PyClone's inference to generate clusters in the output. We divided successfully processed patients based on the trends of their tumor immune pressure's variation over time, "a". We first define three groups of patients: patients with high "a" (more than 70% of MCMC iterations have inferred "a">0), patients with low "a" (fewer than 30% of iterations have inferred "a">0), and the other patients in the middle. Fig. 10a shows the proportion of patients in each category, for each tumor type. In every cancer type that was investigated, we observed that there were patients who displayed an increase in anti-neoantigen immune pressure over time, while others exhibited a decrease in this trend, showing that heterogeneous tumor evolutionary processes, as a result of T cell-mediated pressure, exist in all cancer types. In addition, we found that in certain tumor types (Fig. 10a) there were much greater proportions of patients who demonstrated an increased immune selection pressure over time, such as Adrenocortical carcinoma (ACC), Lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), and Uveal melanoma (UVM); meanwhile, other tumor types had many more patients who showed decreasing immune pressure, such as Pancreatic adenocarcinoma (PAAD), and Uterine carcinosarcoma (UCS). This suggests that the nature of the tumor-immune interactions also

varies in a sophisticated manner across different cancer types, which could inform immunotherapy choices for treating these cancers.

It is unclear from present studies as to whether and how neoantigen presentation can alter the molecular phenotypes of the tumors. To answer this question, for each cancer type, we divided patients into a group “INisp”, with increasing immune pressure over time ($a>0$ in more than 50% of iterations), and another group “DEisp”, with decreasing immune pressure over time ($a<0$ in more than 50% of iterations). We compared the expression profiles of INisp patients and DEisp patients, and performed gene ontology (GO) analyses to identify enriched pathways in differentially expressed genes. We identified immune-related pathways in the differentially expressed genes for every tumor type we investigated. Immune-related pathways are defined as GO terms with any keyword related to any type of immune cells, or keyword related to immune/interleukin/cytokine/chemokine/bacteria. The top enriched pathways with the most significant P-values, for Kidney renal clear cell carcinoma (KIRC) and Melanoma (SKCM), are shown as examples in Fig. 10b. In fact, Fig. 10c shows that every tumor type has at least three enriched immune-associated pathways identified. Among the most immunogenic cancer types (Lung squamous cell carcinoma (LUSC), Lung adenocarcinoma (LUAD), Melanoma, and Kidney renal clear cell carcinoma) (Wang et al. 2018), kidney cancer has the highest number of immune-related pathways. Curiously, immune-related pathways are also detected in the differentially expressed genes for other cancer types that are usually considered non-immunogenic, such as Adrenocortical carcinoma (ACC), Bladder urothelial carcinoma (BLCA), and Uveal Melanoma (UVM). This observation suggests that neoantigens broadly impact the tumor evolutionary processes of non-immunogenic cancer types, in addition to immunogenic

cancer types, which the field has been focusing on for neoantigen-related research and applications previously.

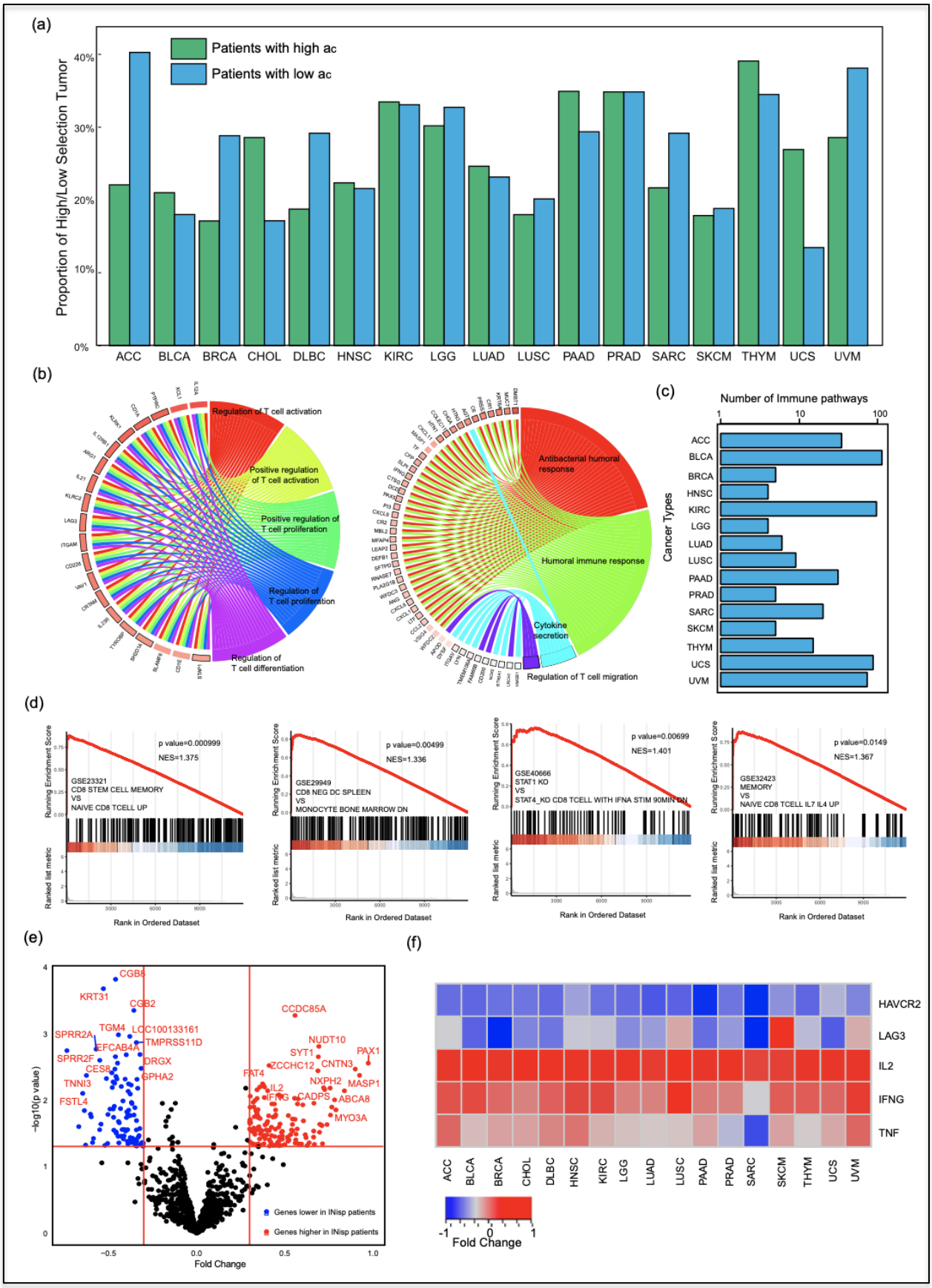

(a)
Patients with high ac
Patients with low ac
Proportion of High/Low Selection Tumor
40%
30%
20%
10%
0%
ACC BLCA BRCA CHOL DLBC HNSC KIRC LGG LUAD LUSC PAAD PRAD SARC SKCM THYM UCS UVM
(b)
Regulation of T cell activation
Positive regulation of T cell activation
Positive regulation of T cell proliferation
Regulation of T cell proliferation
Regulation of T cell differentiation
Antibacterial humoral response
Humoral immune response
Cytokine secretion
Regulation of T cell migration
(c)
Number of Immune pathways
1
10
100
Cancer Types
ACC
BLCA
BRCA
HNSC
KIRC
LGG
LUAD
LUSC
PAAD
PRAD
SARC
SKCM
THYM
UCS
UVM
(d)
p value=0.000999
NES=1.375
GSE23321
CD8 STEM CELL MEMORY
VS
NAIVE CD8 TCELL UP
p value=0.00499
NES=1.336
GSE29949
CD8 NEG DC SPLEEN
VS
MONOCYTE BONE MARROW DN
p value=0.00699
NES=1.401
GSE40666
STAT1 KO
VS
STAT4_KO CD8 TCELL WITH IFNA STIM 90MIN DN
p value=0.0149
NES=1.367
GSE32423
MEMORY
VS
NAIVE CD8 TCELL IL7 IL4 UP
Running Enrichment Score
Ranked list metric
Rank in Ordered Dataset
(e)
−log10(p value)
Fold Change
CGB8
KRT31
CGB2
SPRR2A
TGM4
LOC100133161
TMPRSS11D
EFCAB4A
SPRR2F
DRGX
CES8
GPHA2
TNNI3
FSTL4
CCDC85A
NUDT10
SYT1
PAX1
ZCCHC12
CNTN3
FAT4
NXPH2
MASP1
IL2
IFNG
CADPS
ABCA8
MYO3A
Genes lower in INisp patients
Genes higher in INisp patients
(f)
HAVCR2
LAG3
IL2
IFNG
TNF
ACC BLCA BRCA CHOL DLBC HNSC KIRC LGG LUAD LUSC PAAD PRAD SARC SKCM THYM UCS UVM
-1
0
1
Fold Change

Fig. 10 Immune selection pressure variations correlate with the phenotypes of the tumor and tumor clones. (a) Applying netie on the TCGA plus the in-house KCP data. The percentages of the patients with high "*a*" (*a* >0 in more than 70% MCMC iterations) and low "*a*" (*a*<0 in more than 70% iterations) were shown for each tumor type. The other patients in the middle can be deduced by 1 minus the proportions of patients with high "*a*" and low "*a*", but omitted from plotting to avoid cluttering the figure. (b) Circos plots showing the enriched pathways in the genes that are differentially expressed between INisp and DEisp patients (*a*> or <0 in more than 50% of iterations, unlike the (a) panel). Left: KIRC; right: SKCM. Only the top pathways are shown in each panel for ease of presentation. (c) The number of enriched immune-related pathways found in the genes differentially expressed between INisp and DEisp patients, for each cancer type. (d) The top differentially enriched pathways between INisp and DEisp patients, detected by GSEA. For this analysis, all patients regardless of cancer types were combined. (e) Volcano plot showing the genes that are differentially expressed between INisp and DEisp patients of SARC. A positive value on the X axis means the gene is up-regulated in the INisp patients. (f) A heatmap showing the differential expression of *HAVCR2*, *LAG3*, *IL-2*, *IFNG*, and *TNF*, in all cancer types. Red refers to higher expression in INisp patients, and blue refers to higher expression in DEisp patients.

To confirm our findings above, I also performed Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) between the expression profiles of INisp patients and DEisp patients. We observed that a large number of immune-related pathways, especially T cell related pathways, are differentially enriched between INisp and DEisp patients (the top pathways shown in Fig. 10d).

Then I focused on studying individual genes that were directly related to T cell functions. The volcano plot of differential gene expression of Sarcoma (SARC) is shown in Fig. 10e as an example. I found that IL-2 and IFNG, which are landmark genes up-regulated in activated T cells (Yi et al. 2010), have higher expression levels in sarcoma INisp patients. I also found another gene, TNF, which promotes T cell activation (Unable to find information for 2006230; Yi et al. 2010), is up-regulated in INisp patients in many other cancer types such as Head and Neck squamous cell carcinoma (HNSCC) and Adrenocortical carcinoma (ACC). I systematically demonstrated the differential expression of these genes between INisp and DEisp patients in all the cancer types analyzed (Fig. 10f). Across almost all cancer types, we observed higher expression of IL-2, IFNG, and to a lesser extent, TNF in INisp patients. In Fig. 10f, we also included two genes, LAG-3 and HAVCR2, which are markers of T cell exhaustion (Anderson 2012; Unable to find information for 6662975), and displayed their consistent up-regulation in DEisp patients.

We also examined whether the patient neoantigen repertoires were correlated with tumors' epigenetic profiles. By comparing the methylation levels between INisp patients and DEisp patients, we identified two cancer types, kidney renal clear cell carcinoma and Head and Neck squamous cell carcinoma (HNSCC), with 16 and 2 enriched immune-related pathways in the differentially methylated genes.

Overall, netie built a link between patient neoantigens and tumor molecular profiles, which is capable of yielding novel mechanistic insights into the complicated process of host-tumor interactions.

Immune pressure variations are correlated with the genotypes of tumor clones

Next, I investigated whether the immune pressure variations inferred by netie are correlated with the genotypes of the tumor clones by looking at whether the tumor clones with and without somatic mutations in each gene display any difference in the immune pressure variations a_c. For this analysis, I pooled all tumor clones from all patients and employed a two-sided Wilcoxon test to compare ac. I plotted a histogram of the Wilcoxon test P values of all genes, and interestingly observed an enrichment towards small P values, while the null distribution will be a uniform distribution of P values between 0 and 1. This suggests that there are indeed some genes whose mutations impact the performance of cytotoxic T cells in the tumor microenvironment and affect the immune selection pressure, inferred by netie. In Fig. 11a, I display the top genes with the smallest P values. Interestingly I noticed two genes, SETDB1 and FN1, with prior implications of their involvement in tumor-T cell interactions (Dang et al. 2017; Griffin et al. 2021; Unable to find information for 11382656). In Fig. 11b, we showed that when SETDB1 is mutated, the tumor clones will more likely demonstrate an increase in immune pressure. Fig. 11c shows the same phenotype for FN1.

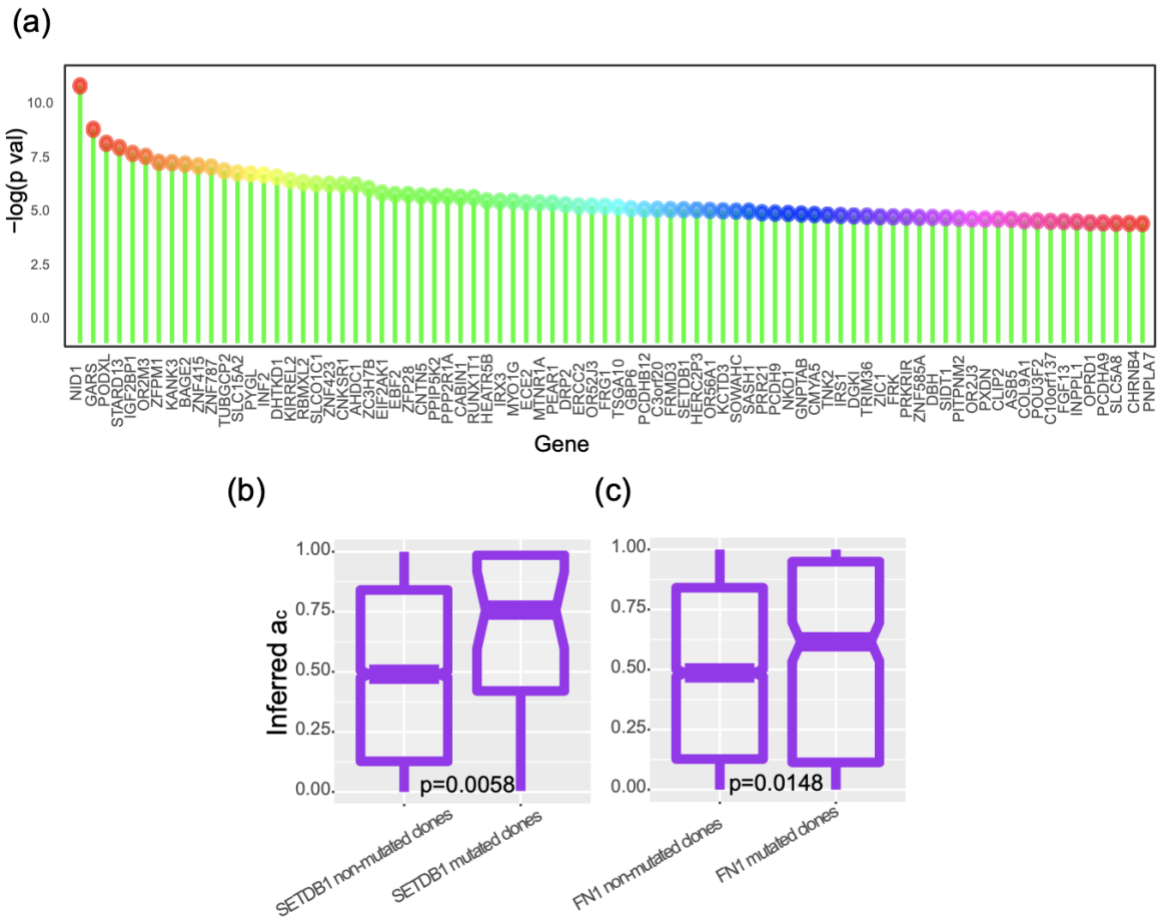


Fig. 11 Immune selection pressure variations correlate with the genotypes of the tumor and tumor clones. (a) The top genes with smallest Wilcoxon test P values comparing the immune pressure variations in the tumor clones with and without mutations in each gene. (b,c) Boxplots of the immune pressure variation (*ac*) in the tumor clones with and without mutations in SETDB1 (b) and FN1 (c).

Multi-region sampling reveals intra-tumor heterogeneity of immune pressure

Netie is also applicable for joint-analysis of multiple samples from the same tumor. In the prior analyses with only one sample per patient, netie infers the clone-specific immune selection pressure and reports an overall tumor-wise average. The availability of multiple samples per

patient and the unique composition of tumor clones in each sample allows for closely examining the differences between different samples and individual clones, providing a more fine-grained insight into the intra-tumor heterogeneity of immune selection pressure. I generated WES and RNA-seq data for four non-small cell lung cancer patients (NSCLC), for each of whom three samples from different regions of the tumors were collected. The phylogenetic tree for each patient was reconstructed by Pyclone (Roth et al. 2014) and Clonevol (Dang et al. 2017). One example showed one patient's phylogenetic tree in Fig. 12a. There were a total of 13 clones found in the three samples of this patient (one common clone, and 12 private clones). Interestingly, netie inferred that the private clones demonstrated an enhancement of immune selection pressure over time, while the sole shared clone demonstrated the opposite trend (Fig. 12b). I also found a stronger decrease in immune pressure for the shared clones of each of the other three patients' multi-region sampled tumors, compared with their private tumor clones (**Fig. 12c**). One possible explanation for this curious observation could be the different levels of immune selection pressure inflicted upon the distinct tumor clones. The tumor clones with stronger decrease of immune responses, due to certain unknown reasons, are more likely to persist and evolve in more regions of the tumor (and thus become the observed "shared" clone).

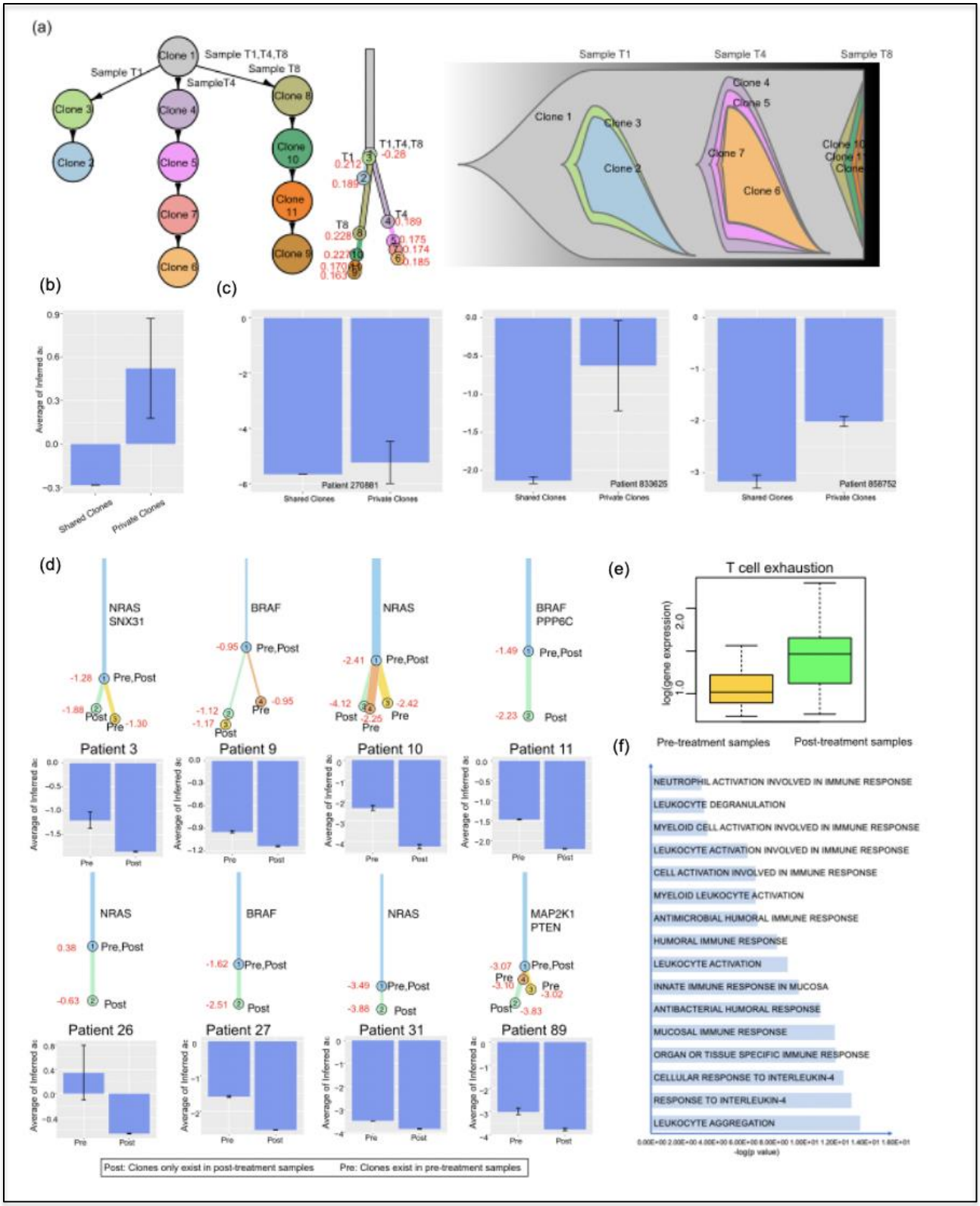

(a)
Clone 1
Sample T1,T4,T8
Sample T1
Sample T8
SampleT4
Clone 3
Clone 4
Clone 8
Clone 2
Clone 5
Clone 10
Clone 7
Clone 11
Clone 6
Clone 9
T1,T4,T8
-0.28
T1
0.212
0.189
T4
0.189
T8
0.228
0.175
0.174
0.185
Sample T1
Sample T4
Sample T8
Clone 1
Clone 3
Clone 2
Clone 4
Clone 5
Clone 7
Clone 6
(b)
Average of Inferred ac
Shared Clones
Private Clones
(c)
Patient 270881
Patient 833625
Patient 858752
Shared Clones
Private Clones
(d)
NRAS
SNX31
-1.28
Pre,Post
-1.88
Post
Pre
-1.30
BRAF
-0.95
Pre,Post
-1.12
-1.17
Post
Pre
-0.95
NRAS
-2.41
Pre,Post
-4.12
Post
-2.25
Pre
-2.42
BRAF
PPP6C
-1.49
Pre,Post
-2.23
Post
Patient 3
Patient 9
Patient 10
Patient 11
NRAS
0.38
Pre,Post
-0.63
Post
BRAF
-1.62
Pre,Post
-2.51
Post
NRAS
-3.49
Pre,Post
-3.88
Post
MAP2K1
PTEN
-3.07
Pre,Post
Pre
-3.10
Pre
-3.02
Post
-3.83
Patient 26
Patient 27
Patient 31
Patient 89
Post: Clones only exist in post-treatment samples
Pre: Clones exist in pre-treatment samples
(e)
T cell exhaustion
log(gene expression)
Pre-treatment samples
Post-treatment samples
(f)
NEUTROPHIL ACTIVATION INVOLVED IN IMMUNE RESPONSE
LEUKOCYTE DEGRANULATION
MYELOID CELL ACTIVATION INVOLVED IN IMMUNE RESPONSE
LEUKOCYTE ACTIVATION INVOLVED IN IMMUNE RESPONSE
CELL ACTIVATION INVOLVED IN IMMUNE RESPONSE
MYELOID LEUKOCYTE ACTIVATION
ANTIMICROBIAL HUMORAL IMMUNE RESPONSE
HUMORAL IMMUNE RESPONSE
LEUKOCYTE ACTIVATION
INNATE IMMUNE RESPONSE IN MUCOSA
ANTIBACTERIAL HUMORAL RESPONSE
MUCOSAL IMMUNE RESPONSE
ORGAN OR TISSUE SPECIFIC IMMUNE RESPONSE
CELLULAR RESPONSE TO INTERLEUKIN-4
RESPONSE TO INTERLEUKIN-4
LEUKOCYTE AGGREGATION
-log(p value)

Fig. 12 Netie is capable of performing multi-sample joint analyses. (a) Netie analysis of the multi-site samples of one MDACC lung cancer patient (Patient ID 886403). The tumor clones were visualized in the phylogenetic tree plot and the fish plot. (b) The immune selection pressure scores of the shared and the private tumor clones of this patient in (a). (c) The immune selection pressure scores of the shared and the private tumor clones of the other MDACC lung cancer patients with multi-region sampling. (d) Netie analysis of the pre-treatment and post-treatment samples from the Riaz cohort. The immune selection pressure scores were also visualized in barplots for comparison between clones that occurred in pre-treatment samples and new clones that occurred only in the post-treatment samples. The P value of paired T test for testing the differences in "*a*" (before *vs.* after treatment) for the 8 patients is 0.015. (e) Boxplots of the expression levels of the T cell exhaustion signature, comparing the pre-treatment and post-treatment samples. (f) GO analysis of the genes differentially expressed between the pre-treatment and post-treatment samples. The lengths of the bars are proportional to the -log(P value) of the GO analysis.

Additionally, I analyzed a cohort of melanoma patients treated with checkpoint inhibitors (Riaz et al (Riaz et al. 2017)). There are a total of eight patients from the Riaz cohort for whom both pre- and post-treatment samples were collected and for whom netie, Pyclone, and Clonevol analyses were all successfully performed. These patients were mostly stable disease and progressive disease patients, without any complete response patients present in this dataset. Netie showed that these patients' tumors demonstrated an overall decreasing trend of immune activity (Fig. 12c, $a_C<0$), which seemed to be consistent with the lack of responsiveness in these patients. Due to the availability of both pre- and post-treatment samples, we were able to distinguish

which tumor clones occurred later on. I compared the evolutionary patterns of the immune selection pressure of the clones that were pre-existing in the pre-treatment samples and those that had newly occurred in the post-treatment samples. Interestingly, I found that the clones that newly arose after immune-checkpoint blockade therapy all had stronger waning of anti-tumor immune activity than clones that already existed in the pre-treatment samples (Fig. 12d, Pval=0.015). This observation could be caused by the exhaustion of T cells after checkpoint blockade. To confirm this, I examined the expression of a T cell exhaustion gene signature (Yi et al. 2010) in these samples. I observed that the T cell exhaustion level was indeed higher in post-treatment samples than in pre-treatment samples, with statistical significance achieved (Fig. 3d, Pval=0.037). I also examined the differential expression of pre-treatment and post-treatment samples in an unbiased manner. In Fig. 12f, we showed that the differentially expressed genes were enriched in pathways essential for immune system activation, leukocyte activation, and leukocyte aggregation. Overall, netie analyses revealed, from the perspective of the evolution of neoantigens, an exhaustion of T cell anti-tumor activity after checkpoint blockade.

Overall, multi-sample genomics data, when viewed through the lens of netie, revealed that the out-growth of particular tumor clones is concomitant with a weakening of T cells' immune surveillance on these clones.

Discussion

This work provided a tool, netie, to infer the footprint left by anti-tumor T cells on the evolution of each subclone of a heterogeneous tumor over the course of tumorigenesis. Netie was systematically validated by simulation data with assumed gold-standard and through application in large scale real human tumor data. This is the first study that has explicitly modeled how

tumor neoantigens and T cells shaped the clonal structures of tumors. Interestingly, this study showed that tumor neoantigens shape the intra-tumor heterogeneity structure of not only the most immunogenic cancer types, but also many non-immunogenic cancer types - hinting to the broad opportunities of neoantigen-based immunotherapies (such as neoantigen vaccines) for these cancer types. With the model netie, we were also the first to characterize the extent to which tumor neoantigens impact the transcriptomic states of the tumors. While previous studies focus on studying the relationship between neoantigen loads and patient clinical phenotype, this analysis provided mechanistic insights into the inter-relationship between neoantigen repertoire and tumor genotypes/phenotypes, and the roles of neoantigens during tumorigenesis and clonal evolution. On the other hand, netie revealed that the past history of tumor-immune interactions can inform the prediction of patients' future prognosis and responsiveness to immunotherapy treatment. As netie is built for the inference of tumor-T cell interactions, it seems fitting and logical to observe the strong synergistic effects between T cell infiltration and the netie INisp/DEisp classifications for predictions of prognosis and treatment response. I envision future work to systematically evaluate the incorporation of netie in biomarkers for patient outcomes in prospective clinical trials.

This work was enabled by the innovative design of leveraging the clusters of mutations (representing tumor clones) detected by software such as PyClone, PhyloWGS and SciClone, and also the cellular prevalences of the mutations in the same tumor clone to serve as a molecular clock for timing the occurrence of each genetic event. While the usage of these tumor clone detection software has been commonplace, very few works have examined each individual clone in a manner near-paralleling the thoroughness conducted during this study. I expect this surrogate molecular clock approach to be generally applicable to other domains of tumor

genomic research, and to provide new discoveries beyond the scope of tumor neoantigens. Based on this core rationale, we developed an advanced and robust Bayesian Hierarchical model to infer the history of neoantigen-T cell interactions during tumorigenesis. The Bayesian framework allows netie to probabilistically consider all tumor clones in a random effect model, which respects the characteristics of each individual tumor clone but also digests the shared information across clones at the same time. Furthermore, the netie model is highly flexible and can handle either single or multiple samples from the same patient.

One limitation of our study is that the trend of the variation in the anti-tumor immune pressure is simply categorized as increasing or decreasing in netie. The interaction between tumors and the immune system is usually complicated and the trend of change in the immune selection pressure may not necessarily be linear throughout the evolution process. We hope future works from the field of *in silico* dissection of intra-tumor heterogeneity will develop more reliable software to time the occurrences of different tumor clones and mutations, which will enable netie to model the relationships more comprehensively between tumors and anti-tumor immunity, and to capture their complicated, nonlinear interactions.

Overall, netie bridged the field of neoantigen research and the field of tumor clonal deconvolution research, revealing an exciting and uncharted territory for future studies.

CHAPTER FIVE
Discussion

In this thesis, I introduced and discussed neoantigens and their application to clinical science. The development of genome sequencing technology and bioinformatics has greatly improved our

capability to study the interaction between tumor and immune system and understand the functions of neoantigens in physiological and tumor development. Then I showed my work on methodologies development. I developed three tools, CSiN, pMTnet, netie to study the function of neoantigens. Then, I validated these tools in the real datasets and made novel biological discoveries using the tools. Finally, I presented the impact of neoantigens on tumor development and patients' clinical outcomes from different aspects by using the tools. Overall, I contributed to the tumor immunity research community from the bioinformatics side by describing the neoantigen repertoires in human tumors, understanding the interaction between neoantigens and T cells, and investigating the impact of neoantigens on tumorigenesis and progression.

The development of the next generation sequencing greatly improves the study of tumor immunology. NGS enabled the comparison between tumor and normal cell genome and identification of mutations and neoantigens with high efficiency. Notwithstanding the effort to predict neoantigens from the whole exome sequencing, it is possible that false-positive and false-negative neoantigens convoluted the neoantigen assessment and analyses. Combining genomics sequencing technologies and transcriptomics sequencing technologies with proteomics-based technologies for identifying neoantigens could possibly increase the accuracy of identification of neoantigens and the accuracy for predicting patients' clinical outcomes by neoantigens. Mass spectrometry, mild acid elution (MAE), and immunoprecipitation (IP) (Kote et al. 2020) provided possible ways to isolate MHC-neoantigens from the tumor cell surface of human tissue. The direct detection of neoantigens by proteomics-based technologies would help decrease the errors brought by genomics and transcriptomics-based methods.

With the maturation of both the sequencing technology and bioinformatics tools, more discoveries and predictions based on neoantigens could be more accurate. For example, the

assessment of neoantigens could be improved by validating predicted neoantigens with mass spectrometry. The prediction for T cell-pMHC could also be improved by more validated training data. Another future direction of neoantigen TCR MHC binding is to incorporate TRA sequencing into the training data and collect more training data from the large-scale screening methods. All these efforts will help us better understand tumor immunology and the clinical treatment to cure cancer, and auto-immune diseases.

Bibliography

Altman JD, Moss PAH, Goulder PJR, Barouch DH, McHeyzer-Williams MG, et al. 2011. Phenotypic analysis of antigen-specific T lymphocytes. Science. 1996. 274: 94-96. *J. Immunol.* 187(1):7–9

Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, et al. 2017. Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. *Cancer Discov.* 7(3):264–76

Anderson AC. 2012. Tim-3, a negative regulator of anti-tumor immunity. *Curr. Opin. Immunol.* 24(2):213–16

Attaf M, Malik A, Severinsen MC, Roider J, Ogongo P, et al. 2018. Major TCR Repertoire Perturbation by Immunodominant HLA-B*44:03-Restricted CMV-Specific T Cells. *Front. Immunol.* 9:2539

Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, et al. 2020. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* 48(D1):D1057–62

Balachandran VP, Łuksza M, Zhao JN, Makarov V, Moral JA, et al. 2017. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature.* 551(7681):512–16

Berger CT, Frahm N, Price DA, Mothe B, Ghebremichael M, et al. 2011. High-functional-avidity cytotoxic T lymphocyte responses to HLA-B-restricted Gag-derived epitopes associated with relative HIV control. *J. Virol.* 85(18):9334–45

Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, et al. 2017. Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* 35(10):908–11

Borbulevych OY, Santhanagopolan SM, Hossain M, Baker BM. 2011. TCRs used in cancer gene

therapy cross-react with MART-1/Melan-A tumor antigens via distinct mechanisms. *J. Immunol.* 187(5):2453–63

Bourcier KD, Lim DG, Ding YH, Smith KJ, Wucherpfennig K, Hafler DA. 2001. Conserved CDR3 regions in T-cell receptor (TCR) CD8(+) T cells that recognize the Tax11-19/HLA-A*0201 complex in a subject infected with human T-cell leukemia virus type 1: relationship of T-cell fine specificity and major histocompatibility complex/peptide/TCR crystal structure. *J. Virol.* 75(20):9836–43

Brennan RM, Miles JJ, Silins SL, Bell MJ, Burrows JM, Burrows SR. 2007. Predictable alphabeta T-cell receptor selection toward an HLA-B*3501-restricted human cytomegalovirus epitope. *J. Virol.* 81(13):7269–73

Burrows SR, Silins SL, Moss DJ, Khanna R, Misko IS, Argaet VP. 1995. T cell receptor repertoire for a viral epitope in humans is diversified by tolerance to a background major histocompatibility complex antigen. *J. Exp. Med.* 182(6):1703–15

Cancer Genome Atlas Network. 2015. Genomic classification of cutaneous melanoma. *Cell*. 161(7):1681–96

Cancer Genome Atlas Research Network. 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 489(7417):519–25

Cancer Genome Atlas Research Network. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*. 499(7456):43–49

Cancer Genome Atlas Research Network. 2014. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 511(7511):543–50

Chen G, Yang X, Ko A, Sun X, Gao M, et al. 2017. Sequence and structural analyses reveal distinct and highly diverse human CD8+ TCR repertoires to immunodominant viral

antigens. *Cell Rep.* 19(3):569–83

Cherkasova E, Scrivani C, Doh S, Weisman Q, Takahashi Y, et al. 2016. Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer. *Cancer Res.* 76(8):2177–85

Cherkasova E, Weisman Q, Childs RW. 2013. Endogenous retroviruses as targets for antitumor immunity in renal cell cancer and other tumors. *Front. Oncol.* 3:243

Cole DK, Fuller A, Dolton G, Zervoudi E, Legut M, et al. 2017. Dual Molecular Mechanisms Govern Escape at Immunodominant HLA A2-Restricted HIV Epitope. *Front. Immunol.* 8:1503

Cole DK, Miles KM, Madura F, Holland CJ, Schauenburg AJA, et al. 2014. T-cell receptor (TCR)-peptide specificity overrides affinity-enhancing TCR-major histocompatibility complex interactions. *J. Biol. Chem.* 289(2):628–38

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, et al. 2014. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 32(12):1202–12

Dang HX, White BS, Foltz SM, Miller CA, Luo J, et al. 2017. ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann. Oncol.* 28(12):3076–82

da Silva VL, Fonseca AF, Fonseca M, da Silva TE, Coelho AC, et al. 2017. Genome-wide identification of cancer/testis genes and their association with prognosis in a pan-cancer analysis. *Oncotarget.* 8(54):92966–77

Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. 2015. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 16:35

De Plaen E, Lurquin C, Van Pel A, Mariamé B, Szikora JP, et al. 1988. Immunogenic (tum-)

variants of mouse tumor P815: cloning of the gene of tum- antigen P91A and identification of the tum- mutation. *Proc Natl Acad Sci USA*. 85(7):2274–78

Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, et al. 2017. The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol.* 8:278

Gao W, Mahajan SP, Sulam J, Gray JJ. 2020. Deep learning in protein structural modeling and design. *Patterns (N Y)*. 1(9):100142

Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, et al. 2018. Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. *Cell*. 172(3):549-563.e16

Gettinger S, Choi J, Hastings K, Truini A, Datar I, et al. 2017. Impaired HLA class I antigen processing and presentation as a mechanism of acquired resistance to immune checkpoint inhibitors in lung cancer. *Cancer Discov.* 7(12):1420–35

Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, et al. 2019. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front. Immunol.* 10:2820

Glanville J, Huang H, Nau A, Hatton O, Wagar LE, et al. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature*. 547(7661):94–98

Grant EJ, Josephs TM, Valkenburg SA, Wooldridge L, Hellard M, et al. 2016. Lack of Heterologous Cross-reactivity toward HLA-A*02:01 Restricted Viral Epitopes Is Underpinned by Distinct αβT Cell Receptor Signatures. *J. Biol. Chem.* 291(47):24335–51

Griffin GK, Wu J, Iracheta-Vellve A, Patti JC, Hsu J, et al. 2021. Epigenetic silencing by SETDB1 suppresses tumour intrinsic immunogenicity. *Nature*. 595(7866):309–14

Guo Y, Li W, Wang B, Liu H, Zhou D. 2019. DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC*

Bioinformatics. 20(1):341

Hellmann MD, Nathanson T, Rizvi H, Creelan BC, Sanchez-Vega F, et al. 2018. Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *Cancer Cell*. 33(5):843-852.e4

Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, et al. 2016. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*. 165(1):35–44

Huth A, Liang X, Krebs S, Blum H, Moosmann A. 2019. Antigen-Specific TCR Signatures of Cytomegalovirus Infection. *J. Immunol.* 202(3):979–90

Jiang T, Shi T, Zhang H, Hu J, Song Y, et al. 2019. Tumor neoantigens: from basic research to clinical applications. *J. Hematol. Oncol.* 12(1):93

Joglekar AV, Leonard MT, Jeppson JD, Swift M, Li G, et al. 2019. T cell antigen discovery via signaling and antigen-presenting bifunctional receptors. *Nat. Methods*. 16(2):191–98

Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. 2021. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput. Biol.* 17(3):e1008814

Klinger M, Pepin F, Wilkins J, Asbury T, Wittkop T, et al. 2015. Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PLoS ONE*. 10(10):e0141561

Kløverpris HN, McGregor R, McLaren JE, Ladell K, Harndahl M, et al. 2015. CD8+ TCR Bias and Immunodominance in HIV-1 Infection. *J. Immunol.* 194(11):5329–45

Kote S, Pirog A, Bedran G, Alfaro J, Dapic I. 2020. Mass Spectrometry-Based Identification of MHC-Associated Peptides. *Cancers (Basel)*. 12(3):

Kula T, Dezfulian MH, Wang CI, Abdelfattah NS, Hartman ZC, et al. 2019. T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. *Cell*. 178(4):1016-1028.e13

Kulmon P. 2021. Reversible jump MCMC for deghosting in MSPSR systems. *Sensors*. 21(14):

Lee JK, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, et al. 2004. T cell cross-reactivity and conformational changes during TCR engagement. *J. Exp. Med.* 200(11):1455–66

Leslie A, Price DA, Mkhize P, Bishop K, Rathod A, et al. 2006. Differential selection pressure exerted on HIV by CTL targeting identical epitopes but restricted by distinct HLA alleles from the same HLA supertype. *J. Immunol.* 177(7):4699–4708

Lichterfeld M, Williams KL, Mui SK, Shah SS, Mothe BR, et al. 2006. T cell receptor cross-recognition of an HIV-1 CD8+ T cell epitope presented by closely related alleles from the HLA-A3 superfamily. *Int. Immunol.* 18(7):1179–88

Linette GP, Carreno BM. 2017. Neoantigen vaccines pass the immunogenicity test. *Trends Mol. Med.* 23(10):869–71

Liu J, Gong X. 2019. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC Bioinformatics*. 20(1):609

Liu YC, Miles JJ, Neller MA, Gostick E, Price DA, et al. 2013. 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(22):15442–54

Li Q, Xu J, Wang J, Jing Y, Wang X. 2021. Uncertainty Quantification Enforced Flash Radiography Reconstruction by Two-Level Efficient MCMC. *IEEE Trans. Image Process.* 30:7184–99

Lo AS-Y, Xu C, Murakami A, Marasco WA. 2014. 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

Lu T, Wang S, Xu L, Zhou Q, Singla N, et al. 2020. Tumor neoantigenicity assessment with CSiN score incorporates clonality and immunogenicity to predict immunotherapy outcomes. *Sci. Immunol.* 5(44):

Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, et al. 2017. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*. 551(7681):517–20

McDermott DF, Huseni MA, Atkins MB, Motzer RJ, Rini BI, et al. 2018. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat. Med.* 24(6):749–57

McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, et al. 2016. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. 351(6280):1463–69

Miao D, Margolis CA, Gao W, Voss MH, Li W, et al. 2018. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science*. 359(6377):801–6

Miller CA, White BS, Dees ND, Griffith M, Welch JS, et al. 2014. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* 10(8):e1003665

Motozono C, Kuse N, Sun X, Rizkallah PJ, Fuller A, et al. 2014. Molecular basis of a dominant T cell response to an HIV reverse transcriptase 8-mer epitope presented by the protective allele HLA-B*51:01. *J. Immunol.* 192(7):3428–34

Nielsen M, Andreatta M. 2016. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* 8(1):33

Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, et al. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE*. 2(8):e796

Ogunshola F, Anmole G, Miller RL, Goering E, Nkosi T, et al. 2018. Dual HLA B*42 and B*81-reactive T cell receptors recognize more diverse HIV-1 Gag escape variants. *Nat. Commun.* 9(1):5023

Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, et al. 2018. Corrigendum: An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*. 555(7696):402

Purbhoo MA, Li Y, Sutton DH, Brewer JE, Gostick E, et al. 2007. The HLA A*0201-restricted hTERT(540-548) peptide is not detected on tumor cells by a CTL clone or a high-affinity T-cell receptor. *Mol. Cancer Ther.* 6(7):2081–91

Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, et al. 2017. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*. 171(4):934-949.e16

Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, et al. 2015. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 348(6230):124–28

Roth A, Khattra J, Yap D, Wan A, Laks E, et al. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*. 11(4):396–98

Shimizu A, Kawana-Tachikawa A, Yamagata A, Han C, Zhu D, et al. 2013. Structure of TCR and antigen complexes at an immunodominant CTL epitope in HIV-1 infection. *Sci. Rep.* 3:3097

Sieberts SK, Zhu F, García-García J, Stahl E, Pratap A, et al. 2016. Erratum: Crowdsourced

assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. *Nat. Commun.* 7:13205

Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, et al. 2014. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 371(23):2189–99

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* 102(43):15545–50

Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. 2017. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*. 33(18):2924–29

Tran E, Ahmadzadeh M, Lu Y-C, Gros A, Turcotte S, et al. 2015. Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science*. 350(6266):1387–90

Valkenburg SA, Josephs TM, Clemens EB, Grant EJ, Nguyen THO, et al. 2016. Molecular basis for universal HLA-A*0201-restricted CD8+ T-cell immunity against influenza viruses. *Proc Natl Acad Sci USA.* 113(16):4440–45

Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, et al. 2015. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. 350(6257):207–11

Verdegaal EME, de Miranda NFCC, Visser M, Harryvan T, van Buuren MM, et al. 2016. Neoantigen landscape dynamics during human melanoma-T cell interactions. *Nature*. 536(7614):91–95

Wang T, Lu R, Kapur P, Jaiswal BS, Hannan R, et al. 2018. An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. *Cancer Discov.* 8(9):1142–55

Weiss GA, Watanabe CK, Zhong A, Goddard A, Sidhu SS. 2000. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc Natl Acad Sci USA.* 97(16):8950–54

Yi JS, Cox MA, Zajac AJ. 2010. T-cell exhaustion: characteristics, causes and conversion. *Immunology*. 129(4):474–81

Yu XG, Lichterfeld M, Chetty S, Williams KL, Mui SK, et al. 2007. Mutually exclusive T-cell receptor induction and differential susceptibility to human immunodeficiency virus type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes. *J. Virol.* 81(4):1619–31

Zhang S-Q, Ma K-Y, Schonnesen AA, Zhang M, He C, et al. 2018. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat. Biotechnol.*

Zhang Z, Xiong D, Wang X, Liu H, Wang T. 2021. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods*. 18(1):92–99