T cell receptor convergence is an indicator of antigen-specific T cell response in cancer immunotherapies

Ming Yao Pan and Bo Li

Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center,
Dallas, TX 75390, USA

Correspondence: Bo.Li@UTSouthwestern.edu

Abstract:

T cells are potent at eliminating pathogens and playing a crucial role in the adaptive immune response. T cell receptor (TCR) convergence describes T cells that share identical TCRs with the same amino acid sequences but have different DNA sequences due to codon degeneracy. We conducted a systematic investigation of TCR convergence using single-cell immune profiling and bulk TCRβ-sequence (TCR-seq) data, and uncovered a strong link between antigen-specificity and convergence. This association was stronger than T cell expansion, a putative indicator of antigen-specific T cells. By using flow sorted tetramer^{+} single T cell data, we discovered that convergent T cells were enriched for a neoantigen-specific CD8^{+} effector phenotype in the tumor microenvironment. Moreover, TCR convergence demonstrated better prediction accuracy for immunotherapy response than the existing TCR repertoire indexes. In conclusion, convergent T cells are likely to be antigen-specific and might be a novel prognostic biomarker for anti-cancer immunotherapy.

Keywords: T cell receptor; antigen-specificity; TCR convergence; immunotherapy; TCR sequencing

Introduction:

T lymphocytes, or T cells, are one of the most important components of the adaptive immune system. T cell receptors (TCRs) are protein complexes found on the surface of T cells that can specifically recognize antigens(*1, 2*). Through the combinational somatic rearrangement of multiple variable (V), diversity (D) (only for β chains), joining (J) and constant (C) gene segments, theoretically, the diversity of unique TCR α and β chains pairs can reach to $2*10^{19}$(*3*). Having such a wide variety of TCRs allows the recognition of numerous endogenous and exogenous antigens. However, due to codon degeneracy, which means multiple codons can encode the same amino acid(*4*), different V(D)J rearrangements can end up encoding the same TCR proteins and this phenomenon is called TCR convergence(*5*). TCR convergence can be observed universally in almost every individual, but very few studies have been conducted to explain its biological significance.

While little is known about the functional impact of TCR convergence, codon degeneracy has been widely studied(*6, 7*). The fact that the number of three-nucleotides codons (64) exceeds the number of encoding amino acids (20) provides the basis of codon degeneracy. The distribution of degeneracy among the 20 encoding amino acids is uneven. Arginine, leucine, and serine are among the amino acids with six corresponding codons, while others have one to four codons(*8*), and the origin of codon degeneracy remains the subject of debate. Some believe that codon degeneracy is the result of co-evolution in which codon assignment occurred by organisms inheriting parts of the codon set from precursor amino acids(*9*). Some have tried to explain it through stereochemical interaction, where stereochemical specificity enables codons to selectively bind to assigned amino acids(*10*). There are also other theories pertaining to this topic(*11-13*). Apart from its elusive origins, the biological significance of codon degeneracy also varies in different scenarios. In the context of T cell immunity, the T cell repertoire is not only results from somatic recombination but also is shaped by multiple selective pressures(*14*). Therefore, it is possible that TCR convergence is more than the mere consequence of codon degeneracy. It may provide unique insights into the antigen-driven TCR selection process.

T cells can be activated by their cognate antigens through the binding of the TCR to the peptide/MHC complex(*15*). Hence, antigen-specific T cells are crucial parts of the T cell 'army' and play dominant roles in eliminating pathogens or tumor cells. There is an increasing number of immunotherapies based on antigen-specific T cells, such as immune checkpoint blockade or neoantigen vaccine therapies(*16, 17*), which further emphasizes the necessity to study T cell antigen-specificity. However, the enormous diversity of epitopes targeted by T cells and the highly polymorphic nature of MHC make it extremely challenging to identify the specificity of given T cells(*18*). Several studies have attempted to predict antigen-specific T cell responses by utilizing TCR characteristics such as diversity, clonality, or evenness scores. While there were cases where these signals were associated with the patient's survival(*19, 20*), in most cases these signals did not correlate with survival(*21-24*). This could be attributed to the weak correlation between these parameters and the antigen selection process. Here, we speculate that TCR convergence is a better indicator of antigen-driven selection. This is because compared with a non-convergent TCR, a convergent TCR consumes more rearrangement resources, and the principle of parsimony(*25*) implies a greater impact on the adaptive immune response for these convergent TCRs.

Over the last few years, high-throughput bulk TCRβ-sequencing (bulk TCR-seq) and single-cell immune profiling have become powerful tools for unraveling a variety of biological problems(*26, 27*). Previously generated datasets are excellent resources for re-analysis. Utilizing three single-cell immune profiling and ten bulk TCR-seq datasets, we conducted a comprehensive analysis of TCR convergence. We began by investigating the basic features of convergent TCRs, including their degeneracies, publicity, and amino acid usage. Next, we explored the correlation between convergent and antigen-specific TCRs from different perspectives. Significant overlaps were found between convergent T cells and antigen-specific T cells across independent datasets. We proceeded to examine the phenotypes of convergent T cells in both mouse and human samples. As a result, convergent T cells exhibited gene signatures associated with cytotoxicity, memory, and exhaustion, related to antigen-specific T cells. Finally, we found that TCR convergence predicts improved clinical outcomes for cancer patients who received

immune checkpoint blockade therapies. Our work presents a new angle to study T cell repertoire and deepens our understanding of antigen-specific T cell selection.

Materials and Methods:

Datasets preparation

All the TCR-seq or single cell RNA-seq (scRNA-seq) datasets used were accessed from public resources. This study analyzed three single-cell immune profiling datasets and ten bulk TCRβ sequencing datasets, which were subjected to different analyses based on the features of the datasets. All the TCR-seq samples were downloaded from the immuneACCESS online database (https://clients.adaptivebiotech.com/immuneaccess). The detailed information for all datasets is described in supplementary table 1.

Definition of TCR convergence

T cells with identical CDR3 amino acid sequences and variable genes, but different CDR3 nucleotide reads were defined as convergent T cells. Single cell immune profiling data analysis included pairing the α-β chain CDR3 sequences and variable genes of each T cell to form a complete TCR sequence. As for the bulk TCR-seq data, the β chain CDR3 sequences and their variable genes were used to represent their TCR sequences.

The definition of public TCRs

To define the public TCRs for the COVID-19 cohort and cohort representing different cancer types, we first constructed the datasets for each group. 50 samples were randomly picked from the Nolan COVID-19 cohort(*28*). The kidney cancer cohort (n=19) is a combination of data from our internal database and samples from the Chow cohort(*29*). Ovarian cancer data was also sourced from our in-house data (n=45). The immune sequencing data for lymphoma (n=53), lung cancer (n=50), and melanoma (n=29) came from the Cader cohort(*30*), Reuben cohort(*31*), and Riaz cohort(*32*), respectively. Each of these cohorts was mixed with the 666 samples from the Emerson cohort(*33*) to form new cohorts with sample sizes of 716, 685, 711, 719, 716, and 695. The TCR sequences shared by at least 5% of different individuals within a cohort were defined as the public TCRs of that cohort.

Definition of TCR clonality and diversity

TCR clonality was defined as 1-Pielou's evenness and was calculated as:

$$TCR\ clonality = 1 + \frac{\sum_i^n p_i log_2\ (p_i)}{log_2(n)},$$

where p_i is the proportional of that unique TCR clonotype and n is the total number of TCR clonotypes within a given sample. TCR diversity was calculated as n/N, where n is the total number of TCR clonotypes and N is the total reproductive sequence reads.

Analysis of scRNA-seq data

The scRNA-seq data were analyzed primarily using R/4.1.1 and Seurat 4.0.6. Data were processed with the conventional scRNA-seq pipeline before the analysis of TCR convergence. In brief, all data first had to pass quality control, where cells with a high ratio of mitochondria and ribosomes were excluded since they were likely to be dying cells. Cells with too many or too few RNA reads were also excluded from studies because they might be doublets (a single barcode identifies more than one cell) or background noise. Data scaling and normalization were then performed. We regressed out the cell-cycle associated genes in our scaled data. Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) algorithm were applied to reduce the dimensions and visualize the data. Based on the markers of the cells within a cluster, cells were grouped into an appropriate number of clusters and assigned T cell types. Combining the scTCR-seq data with the scRNA-seq data allowed us to study the distribution of convergent T cells and their phenotypes.

Statistical analysis

The two-tailed exact binomial tests were performed with the built-in function of R/4.1.1 to estimate the statistical differences in amino acid composition between convergent TCRs and non-convergent TCRs. Fisher's exact tests

were employed to analyze the association between convergent TCRs and antigen-specific TCRs, as well as the relationship between clonally expanded TCRs and antigen-specific TCRs. Using ggstatsplot v.0.8.0, the Fisher's exact test results of the scTCR-seq data were visualized. Ggsignif v.0.6.3 was used to calculate the statistical difference between different groups in all boxplots using Welch's t-test. All survival analyses were conducted using the R package survival v.3.2.13. Survival curves were derived using the Kaplan-Meier method. Cox multivariate regressions were performed to assess the association between the patients' outcomes and TCR convergence, clonality, diversity, and different treatments.

Results:

TCR convergence is different from publicity

We defined the convergent TCRs as above described (Figure 1A) and analyzed their basic characteristics, including the degeneracies of the TCRs, sequence lengths, variable gene usage and amino acid components using a bulk TCR dataset containing 666 blood samples from healthy donors(*34*). Before analysis, quality control was performed to exclude CDR3 reads that did not meet the following two standards: 1) between 8 and 23 amino acids in length; 2) starting with cysteine and ending with phenylalanine. We randomly selected 50 samples, with a total of 9,567,751 that passing quality control (QC) unique TCR used in the following analysis.

We defined the degeneracy of a TCR protein as the number of distinct clonotypes (defined by DNA sequences) encoding that TCR within one sample. In other words, a degeneracy equal to 1 means non-convergent TCRs, while greater than 1 indicates convergent TCRs. As expected, convergent TCRs only comprised a small proportion (5.40%) of the total T cell population (Figure 1B). As TCR degeneracy increased from two to four, the number of detected TCRs decreased approximately tenfold with each unit increase in degeneracy. Most convergent TCRs had a degeneracy of two (90.65%) or three (7.16%), while larger degeneracies were rare.

We studied the length distribution of the convergent TCRs, which peaked at 11 amino acids (AA) (Figure S1A). While TCRs with 14 to 15 AA made up the greatest percentage of all the unique TCRs (Figure S1B), the proportion of convergent TCRs continually decreased after reaching 11 AA or longer. This result suggested that shorter TCR sequences appeared to be more conducive to TCR convergence. In terms of the correlation between TCR degeneracy and CDR3 lengths, CDR3s of 9 AA long displayed the greatest degeneracy on average (~2.5) (Figure S1C). From 11 amino acids onward, the longer the sequences, the smaller the degeneracy. After reaching 17 amino acids, the degeneracy remained constant. This is an unexpected observation, as the probability of the occurrence of codon degeneracy should increase with longer sequence lengths. We further analyzed the variable gene usage of convergent TCRs and observed that the usage of every TCRβ chain variable gene (TCRBV) superfamily was relatively even (Figure S1D). TCR degeneracy was not significantly impacted by variable gene usage (Figure S1E) either.

We then calculated the average percentage of a given amino acid in TCRs with degeneracy from one to four. The first and last three AA in the CDR3 sequences were not included since they are determined by the types of V or J genes. The result showed that, generally, convergent TCR favors polar neutral amino acids and acidic amino acids, particularly tyrosine (Y), serine (S), and glutamic acid (E), with alkaline amino acids and nonpolar hydrophobic amino acid less favorable (Figure 1C). Arginine (R), leucine (L), and serine (S), each of which has 6 corresponding codons, have the largest codon degeneracy. However, leucine and arginine did not show the same level of relevance to TCR convergence as serine. While the ratio of leucine slightly increased in convergent TCR, the percentage of arginine even decreased as the TCR degeneracy increased. Tyrosine and glutamic acid, on the other hand, have only two codons each but were found in a higher percentage of convergent TCRs than non-convergent TCRs. The possibility of each above-mentioned amino acid (L, Y, S, E, R) in each degree of TCR degeneracy has been validated by two-sided binomial exact tests with p-values less than 2.2×10^{-16}. These results indicate that the level of codon degeneracy is not a determinant of TCR degeneracy. Instead, the physiochemical properties of a given amino acid might have a greater impact on TCR convergence. Therefore, even though TCR convergence is the result of codon degeneracy, it exhibits an independent distribution and may carry a biological significance that is distinct from codon degeneracy.

Public TCRs are generated from VDJ recombination biases and are shared across different individuals, which might target common antigens, such as viral epitopes(*33*). In contrast, TCRs targeting cancer neoantigens are mostly 'private' that are unique to each individual(*35*). As TCR convergence and publicity are conceptually similar, we next investigated the differences between the two by comparing the fractions of public TCRs within the convergent TCRs from patients with COVID-19 or different cancers (Figure 1D). As expected, we observed a high overlap (46%) between the two in COVID-19 patients. Whereas for the cancer patients, the overlap between TCR convergence and publicity decreases for cancer types with higher mutation burden, which was estimated from

previous studies(*36*). The fraction of public TCRs within the convergence sequences dropped from 34% for the low mutation burden kidney cancer to 22% for the highly mutated melanoma samples. As the public TCRs in these cohorts are consistently defined by the Emerson 2017 cohort(*33*), this difference is unlikely caused by unknown batch effects. Together, our results indicated that TCR convergence and publicity represent two different biological processes that diverge in cancer patients by tumor mutation load.

Convergent TCRs are more likely to be antigen-specific

To explore the relationship between convergent T cells and antigen-specific T cells, we began by examining datasets with known TCR antigen specificity. To be precise, since all the T cells are specific to some antigen(s) during positive thymic selection, the term "antigen-specific" means T cells with an ongoing or memory antigen-specific immune response in this context. In 2020, 10x Genomics detected the antigen specificity of T cells using highly multiplexed peptide-MHC (pMHC) multimers(*37*) from PBMC samples of four healthy individuals. We picked out all convergent TCRs within each donor and examined their overlaps with antigen-specific multimer$^+$ TCRs. Interestingly, while convergent TCRs only consisted of a small proportion of all unique clones, they were dominantly enriched for antigen-specific TCRs (Figure 2A). In these four donors, only 12.91%, 33.32%, 71.65%, and 6.24% of the unique TCRs were multimer positive, respectively. However, 88.00%, 88.28%, 93.10%, and 22.22% of convergent TCRs were antigen-specific, suggesting that convergent TCRs were much more likely to be antigen-specific than their non-convergent counterpart. We next estimated the statistical significance of this observation for each donor using Fisher's exact test(*38*) (Figure 2B). The odds ratio for donor 1 reached 50.16, indicating that convergent TCRs are much more likely to be antigen-specific. Similar conclusions also held for donors 2 and 3, while the p-value (0.105) of donor 4 was less significant. Considering the small percentage of antigen-specific T cells detected in donor 4, it is possible that most antigens specific to donor 4's T cells did not fall within the spectrum of tested antigens in this experiment, which caused the lower overlap. Overall, the findings from this dataset supported the hypothesis that convergent T cells are more likely to be antigen-specific than other T cells.

We further tested this hypothesis using high-throughput bulk TCRβ-seq data. TCR-seq data were downloaded from ImmunoAccess, containing TCRβ sequences from over 1,400 subjects that had been exposed to or infected with the SARS-CoV-2 virus, along with over 160,000 SARS-CoV-2-specific TCRs(*39*). Similarly, we calculated the odds ratio of convergent TCRs and SARS-CoV-2-specific TCRs using Fisher's exact test for each sample. In comparison, we also investigated the overlap between clonally expanded T cells and antigen-specific T cells in these datasets, as T cell clonal expansion is a commonly used criterion for antigen-specificity in recent clinical studies(*24, 40*). In each sample, TCR sequences with the highest read numbers were selected as expanded TCRs, and their number was restricted to the same number as convergent TCRs. Overall, the odds ratio of convergent TCR was almost unanimously higher than that of the clonally expanded T cells (Figure 2C). The average odds ratio of TCR convergence was 8.94, significantly higher than TCR clonal expansion, which was only 1.74 (Figure 2D). These results confirmed that converged TCRs are more likely to be antigen-experienced, and indicated that TCR convergence is a better indicator of antigen-specificity than clonal expansion.

Convergent T cells exhibit a CD8$^+$ cytotoxic gene signature

In the next step, we studied the gene expression signatures of the convergent T cells using single cell RNA sequencing (scRNA-seq) data in the tumor samples to test the hypothesis that convergence is associated with an antigen-experienced effector phenotype. In the tumor microenvironment, antigen-specific T cells typically express cytotoxicity, memory, and/or exhaustion gene signatures(*41, 42*). We, therefore, checked the phenotypes of convergent T cells using recent single cell RNA-seq data generated from cancer studies.

The first dataset we investigated contained tetramer-labeled neoantigen-specific T cells in the MC38 tumor mouse model treated with neoantigen vaccine and anti-PD-L1(*26*). The CD4$^+$ and CD8$^+$ T cells were analyzed separately using the same cell cluster annotations as previously described(*26*). We then defined and analyzed the

convergent TCRs using the scTCR-seq data of the same T cells. Only 1.86% (n=92) of the 4,945 passed-QC $CD4^+$ cells had convergent TCRs (Figure S2A), which were not enriched in any cluster. In contrast, 9,503 $CD8^+$ T cells passed the quality control and 15.48% (n=1,471) of them were convergent T cells (Figure 3A). Following this result, we noticed two interesting clusters: CD8-05 with significant enrichment of convergent $CD8^+$ T cells and CD8-04 with none of the T cells having a convergent TCR. To further investigate these two groups, we re-clustered the $CD8^+$ cells with a higher resolution (0.3) and divided the T cells into 5 new groups (Figure 3B). 87.5% (419 out of 479) of T cells from group 4 were convergent T cells, whereas none of the group 5 (n=201) were convergent. Group 4 T cells expressed both effector and exhaustion markers, such as Tox(*43, 44*), Tigit(*45*), Eomes(*46*), and inhibitory receptors like Il10ra(*47*). On the other hand, group 5 mainly expressed naïve T cell gene signatures(*47*), such as Sell, Tcf7, and Ccr7 (Figure 3C), and down-regulated genes associated with effector function, like Gzmb, Lgals1(*48*), and T cell activation or exhaustion(*49*), like Pdcd1 and Lag3 (Figure 3C). Finally, by comparing the TCRs of group 4 T cells and tetramer sorted T cells specific to tumor neoantigen *Adpgk*, we observed a significant enrichment of convergent TCRs for neoantigen-specificity (OR=4.41, $p<0.001$) (Figure S2B). Based on these results, we concluded that convergent T cells are enriched for a neoantigen-activated effector phenotype in the tumor microenvironment.

To verify this observation in humans, we analyzed the second dataset of pan-cancer scRNA-seq data of nine cancer types(*50*). There were samples collected from tumor tissues, normal adjacent tissues, and peripheral blood from human patients. Similar to the mouse data, convergent T cells were rare in $CD4^+$ T cells (n=74,554), accounting for 0.16% of the $CD4^+$ population (Figure S2C), whereas 10.15% of the $CD8^+$ T cells (n=69,618) were convergent (Figure 3D). 14 $CD4^+$ clusters and 15 $CD8^+$ clusters were defined using the Seurat R package, and the cluster annotations were assigned by the gene markers of each cluster shown in Supplementary Table 2. Similar to the mouse data, convergent T cells were enriched in clusters CD8-01-XCL1 and CD8-05-FGFBP2, yet remained at low levels in the naïve clusters, CD8-09-IL7R and CD8-13-CCR7 (1.27% and 0.00%) (Table 1). CD8-01-XCL1 cluster expressed gene signatures of activated T cells, including Il7r(*51*), ZFP36L2(*52*), XCL1(*53*), and CXCR6(*54*) (Figure 3E), as well as proliferation and mobility markers LDLRAD4(*55*) and CAPG(*56*). CD8-05-FGFBP2 cluster exhibited a $CD8^+$ effector phenotype, upregulating genes involved in cytotoxicity and anti-tumor activity, FGFBP2, GNLY, FCGR3A(*27*), GZMH, PRF1, TGFBR3, and GZMB (Figure 3E). In conclusion, the phenotypes of convergent T cells in human samples resembled those discovered in the mouse model. Convergent T cells were consistently enriched in the activated $CD8^+$ T cells clusters with the gene signatures of cytotoxicity, memory, and exhaustion, and demonstrated the phenotypes of antigen-specific T cells. To validate the conclusion that TCR convergence was more common in $CD8^+$ T cells than in $CD4^+$ T cells, we used another bulk TCRβ data collected from patients with Hodgkin lymphoma(*30*). As a result, $CD8^+$ T cells consistently exhibited a higher level of TCR convergence throughout different collection time points (Figure S3A and S3B).

TCR convergence is associated with the clinical outcome of ICB treatment

Immune checkpoint blockade (ICB) has seen great success in treating late-stage cancer patients(*57, 58*). While ICB treatment is effective for some patients, a significant proportion of patients remain unresponsive, and the reason for this is not completely understood. The discovery of a biomarker that can assist in the prognosis of immunotherapy remains one of the top clinical priorities. As antigen-specific T cells play a crucial role in defending tumor cells, we next investigated whether TCR convergence is predictive of the immunotherapy outcomes. We used bulk TCRβ-seq data generated by Snyder A, et al(*59*) which included blood samples from 29 urothelial cancer patients collected on the first day of the anti-PD1 treatment. We observed a significant association between TCR convergence level and both the overall survival (OS) and progression-free survival (PFS) (Figures 4A-4B). In the low TCR convergence group, approximately half of the patients deceased within 110 days, whereas over half of the patients in the high TCR convergence group remained alive after 500 days (Figure 4A).

To validate the result, we examined another independent melanoma data set containing 30 samples collected from melanoma patients treated with sequential ICB reagents(*60*). Patients were randomly divided into two arms

with the flipped ordering of anti-PD1 or anti-CTLA4 treatments(*60*). TCR convergence significantly predicted patient outcome ($p=0.00012$) (Figure 4C), with higher TCR convergence associated with longer survival. Responders of the ICB treatment have significantly higher levels of TCR convergence (Figure 4D). To determine whether this link is confounded by other signals, we also included three additional variables that might influence the clinical outcomes in the multivariate Cox regression model, including different sequential treatments, TCR clonality, and TCR diversity. TCR convergence remained significant while adjusted for the additional variables ($p<0.001$) (Figure 4E). Together, our results indicated that TCR convergence is an independent prognostic predictor for patients receiving ICB treatments.

Discussion

Functional redundancy in crucial molecular pathways, such as cell cycle, metabolism, apoptosis, etc., is one of the key mechanisms that keeps biological systems robust and stable against genetic variations and environmental changes(*61, 62*). Redundancy may occur at both molecular and cellular levels, the latter being particularly exploited by the immune system(*63*). In this study, we investigated the convergence of TCRs, which could be a potentially unreported mechanism of redundancy in the adaptive immunity. The selection of multiple clones of T cells with identical TCRs during antigen encounters might be a fail-safe mechanism to ensure the expansion of the antigen-specific T cells. This speculation is in line with our observations in this work: convergent TCRs are enriched for antigen-experienced receptors, and these T cells exhibited a neoantigen-specific, cytotoxic $CD8^+$ phenotype in the tumor microenvironment.

There are conceptual overlaps between TCR publicity and convergence, as both describe the redundancy of TCRs. However, our data suggested that these two processes could be fundamentally different. Currently, the functional consequence of V(D)J recombination bias remains unclear, but it has been speculated that this conserved process is evolved to eliminate common foreign pathogens, such as viruses and bacteria, that are frequently encountered in the environment(*64, 65*). This process is genetically encoded and arises intrinsically. In contrast, TCR convergence might reflect the magnitude of clonal selection, which is highly contingent on the antigen landscape. Our observation that convergent and public TCRs diverge in cancers with high neoantigen load (Figure 1C) strongly supports this speculation.

This indication of antigen-driven selection makes TCR convergence an attractive biomarker for disease diagnosis and/or prognosis. Many previous clinical studies have attempted to use the TCR repertoire to monitor the response of immunotherapies(*66-69*). TCR diversity indexes, such as Shannon's entropy, richness, clonality, etc. are routinely used in these studies, yet a strong association with the patient outcome is rarely observed. This is potentially because these indexes only reflect the overall dynamics of the repertoire, which cannot specifically capture the antigen-specific response. As a new summary statistic of the immune repertoire, TCR convergence hinges on the (neo)antigen-specific T cell responses in late-stage tumors, and achieved better prediction accuracy compared to the diversity indexes. Awaiting further clinical validation, we anticipate that TCR convergence will be a powerful new biomarker to predict ICB therapy responses.

There are also limitations of this study. First, as TCR convergence is identified by small DNA changes, it relies heavily on sequencing accuracy. Improper handling of sequencing errors may result in the overestimation of TCR convergence. Second, since convergent T cells constitute only a small proportion of the total population of T cells, the study of TCR convergence requires a large number of sequenced T cells from each sample. Therefore, an accurate and deep sequencing approach is required for the study of TCR convergence. Third, although the phenotypic signatures of the convergent T cells were confirmed using both mouse and human scRNA-seq/scTCR-seq data, additional neoantigen-specific single T cell datasets would be necessary to consolidate our conclusions. Finally, our observation that TCR convergence is a favored prognostic predictor for ICB therapy was based on only three cohorts, due to the lack of qualified datasets. Future clinical studies with TCR repertoire profiles will be required to confirm TCR convergence as a new biomarker.

Antigen-specific T cells are central to a T cell immune response and determine the efficacy of immunotherapy, but their identification is challenging. However, the detection of convergent T cells is an easier task that only requires the profiling of the immune repertoire. The insights regarding TCR convergence gained from this work might provide an alternative angle to study the dynamics of antigen-specific T cells and a better understanding of the TCR repertoire.

Acknowledgements
This work is supported by NCI 1R01CA245318 (B.L.) and 1R01CA258524 (B.L.).

Author contributions
Conceptualization, B.L.; Methodology, B.L. and M.Y.P.; Investigation, M.Y.P.; Writing - Original Draft, M.Y.P. and B.L.; Writing-Review & Editing, B.L. and M.Y.P.; Funding Acquisition, B.L.; Supervision, B.L..

Data and code availability
The Python code related to TCR convergence calculation are available at: https://github.com/Mia-yao/TCR-convergence/tree/main. Associated datasets are uploaded to Zenodo, with DOI: 10.5281/zenodo.6603757.

Competing interests
No competing interests.

Figure Legends:

Figure 1. TCR convergence is different from publicity

A. Cartoon illustration of TCR convergence. T cells with different DNA sequences encode the same TCR due to codon degeneracy.

B. log_{10}-transformed counts of TCRs with different degrees of degeneracies. The X-axis indicates the number of distinct DNA sequences corresponding to the same TCR, or, the "degree of degeneracy". "D1" refers to the sequences without any degeneracy.

C. The proportion of twenty amino acids in TCRs with degeneracies from one to four. Each of the 20 natural amino acids is represented on the x-axis by a one-letter abbreviation.

D. Tumor mutation burden (TMB) and the proportion of public TCRs in convergent TCRs among different cancer types. The upper panel shows the TMB of kidney, ovarian, lymphoma, lung, and melanoma cancer types. Lower panel shows the ratio of public TCRs among convergent TCRs in the respective cancer types.

Figure 2. Convergent TCRs are more likely to be antigen-specific

A. The overlaps between all unique TCRs, viral antigen-specific TCRs, and convergent TCRs of four donors respectively.

B. The results of Fisher's exact tests regarding the association of convergent TCRs and viral antigen-specific TCRs of four donors.

C. Log_{10}-scaled odds ratio of TCR convergence and expansion with respect to antigen-specificity of each sample. The samples were sorted in order of TCR convergence odds ratio from small to large. The horizontal stripes on the left represent the density of samples, colored by the TCR signal.

D. The difference of log_{10}-transformed odds ratio between TCR convergence and TCR expansion. The p-value was calculated by the two-sided Welch's t-test.

Figure 3. Convergent T cells exhibit a $CD8^+$ cytotoxic gene signature

A. t-distributed stochastic neighbor embedding (t-SNE) plots showing cell type annotations and convergent T cell distribution of $CD8^+$ T cells collected from mouse tumor tissues. Each dot represents a T cell. The area within the red dashed line indicates enrichment of convergent T cells and cells within the blue dashed line were naive T cells.

B. t-SNE displaying the re-cluster result of $CD8^+$ T cells. The red dashed line and blue dashed line highlighted the same groups of cells as in Fig 3A.

C. Volcano plots showing the differentially expressed genes of cluster 4 and cluster 5.

D. t-SNE plots showing the cell clusters and convergent T cell distribution of $CD8^+$ T cells collected from human tumors and normal adjacent tissues. An area within the red dashed line was enriched with convergent T cells and the T cells within the blue dashed line were naive T cells.

E. The gene markers of cluster 01 and cluster 05 of human pan-cancer data. (Only the up-regulated genes of each cluster are shown)

Figure 4. TCR convergence is associated with the clinical outcome of ICB treatment

A. Kaplan-Meier overall survival curves of urothelial cancer patients (n=29) with different levels of convergent T cells. The convergence levels were calculated by the number of convergent T cells divided by the number of all the detected T cells. High, above median; low, below the median. For TCR convergence, all the survival analyses in this analysis followed this definition.

B. Progression-free survival curve of urothelial cancer patients (n=29) with different levels of convergent T cells. The progression of the tumor was determined by RECIST 1.1 test.

C. Overall survival curves of melanoma patients (n=30) with different levels of convergent T cells.

D. The differences in the TCR convergence levels among patients with the distinct best overall response towards anti-PD1 immunotherapy. The number of patients with each BORs are as follows: CR (n=1), PR (n=9), PD (n=18), SD (n=2).

E. Cox model multivariate analysis for melanomas cohort with covariate adjustments.

Supplementary Figure 1. The length and variable gene usage of convergent TCRs

A. The proportion of convergent TCRs among TCRs of a certain length. The x-axis represents the amino acid length of TCR CDR3. Each dot indicates one sample.

B. The length distribution of all unique TCR CDR3 reads.

C. The average degeneracy of convergent TCRs with different CDR3 lengths.

D. The proportion of convergent TCRs among TCRs with a certain variable gene. The x-axis refers to the TCRβ chain variable gene (TCRBV) superfamily.

E. The average degeneracy of convergent TCRs with different variable gene usages.

Supplementary Figure 2. TCR convergence in $CD4^+$ T cells

A. Cell type annotations and convergent T cell distribution of $CD4^+$ T cells collected from mouse tumor tissues.

B. The results of Fisher's exact tests regarding the association of $tetramer^+$ T cells with T cells from cluster 4 or cluster 5.

C. Cell clusters and convergent T cell distribution of $CD4^+$ T cells collected from human tumors and normal adjacent tissues.

Supplementary Figure 3. The convergent T cell levels in $CD4^+$ and $CD8^+$ T cells

A. The different levels of convergent T cells in $CD4^+$ and $CD8^+$ T cells. The samples (n=17) were collected at the baseline. Each dot represents one sample.

B. The different levels of convergent T cells in $CD4^+$ and $CD8^+$ T cells. The samples (n=16) were collected on day 1 of anti-PD1 treatment cycle 4. Each dot represents one sample.

References

1. P. Marrack, J. Kappler, The T-Cell Receptor. *Science* **238**, 1073-1079 (1987).
2. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395-402 (1988).
3. J. A. Pai, A. T. Satpathy, High-throughput and single-cell T cell receptor sequencing technologies. *Nat Methods* **18**, 881-892 (2021).
4. L. E. Trainor, G. W. Rowe, V. L. Szabo, A tetrahedral representation of poly-codon sequences and a possible origin of codon degeneracy. *J Theor Biol* **108**, 459-468 (1984).
5. T. J. Looney *et al.*, TCR Convergence in Individuals Treated With Immune Checkpoint Inhibition for Cancer. *Front Immunol* **10**, 2985 (2019).
6. D. L. Gonzalez, S. Giannerini, R. Rosa, On the origin of degeneracy in the genetic code. *Interface Focus* **9**, 20190038 (2019).
7. D. A. McClellan, The Codon-Degeneracy Model of Molecular Evolution. *Journal of Molecular Evolution* **50**, 131-140 (2000).
8. F. H. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, General nature of the genetic code for proteins. *Nature* **192**, 1227-1232 (1961).
9. J. T.-F. Wong, A Co-Evolution Theory of the Genetic Code. *Proceedings of the National Academy of Sciences* **72**, 1909-1912 (1975).
10. S. D. Copley, E. Smith, H. J. Morowitz, A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proceedings of the National Academy of Sciences* **102**, 4442-4447 (2005).
11. E. V. Koonin, A. S. Novozhilov, Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99-111 (2009).
12. F. H. Crick, The origin of the genetic code. *J Mol Biol* **38**, 367-379 (1968).
13. M. Di Giulio, The origin of the tRNA molecule: implications for the origin of protein synthesis. *J Theor Biol* **226**, 89-93 (2004).
14. S. C. Jameson, M. J. Bevan, T-cell selection. *Current Opinion in Immunology* **10**, 214-219 (1998).
15. J. E. Smith-Garvin, G. A. Koretzky, M. S. Jordan, T Cell Activation. *Annual Review of Immunology* **27**, 591-619 (2009).
16. M. A. Postow, M. K. Callahan, J. D. Wolchok, Immune Checkpoint Blockade in Cancer Therapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **33**, 1974-1982 (2015).
17. M. Peng *et al.*, Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular Cancer* **18**, 128 (2019).
18. E. W. Newell, M. M. Davis, Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. *Nat Biotechnol* **32**, 149-157 (2014).
19. N. Riaz *et al.*, Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949.e916 (2017).
20. P. C. Tumeh *et al.*, PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568-571 (2014).
21. D. B. Johnson *et al.*, Targeted Next Generation Sequencing Identifies Markers of Response to PD-1 Blockade. *Cancer Immunology Research* **4**, 959-967 (2016).
22. L. Robert *et al.*, CTLA4 Blockade Broadens the Peripheral T-Cell Receptor Repertoire. *Clinical Cancer Research* **20**, 2424-2432 (2014).
23. R. N. Amaria *et al.*, Neoadjuvant immune checkpoint blockade in high-risk resectable melanoma. *Nature Medicine* **24**, 1649-1654 (2018).
24. J. Kidman *et al.*, Characteristics of TCR Repertoire Associated With Successful Immune Checkpoint Therapy Responses. *Frontiers in Immunology* **11**, (2020).
25. C. B. Stewart, The powers and pitfalls of parsimony. *Nature* **361**, 603-607 (1993).

26. L. Liu *et al.*, Concurrent delivery of immune checkpoint blockade modulates T cell dynamics to enhance neoantigen vaccine-generated antitumor immunity. *Nature Cancer*, (2022).
27. C. Zheng *et al.*, Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169**, 1342-1356 e1316 (2017).
28. S. Nolan *et al.*, A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq*, rs.3.rs-51964 (2020).
29. J. Chow *et al.*, Radiation induces dynamic changes to the T cell repertoire in renal cell carcinoma patients. *Proc Natl Acad Sci U S A* **117**, 23721-23729 (2020).
30. F. Z. Cader *et al.*, A peripheral immune signature of responsiveness to PD-1 blockade in patients with classical Hodgkin lymphoma. *Nat Med* **26**, 1468-1479 (2020).
31. A. Reuben *et al.*, Comprehensive T cell repertoire characterization of non-small cell lung cancer. *Nature Communications* **11**, 603 (2020).
32. N. Riaz *et al.*, Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949 e916 (2017).
33. R. O. Emerson *et al.*, Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics* **49**, 659-665 (2017).
34. R. O. Emerson *et al.*, Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* **49**, 659-+ (2017).
35. A. Madi *et al.*, T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* **24**, 1603-1612 (2014).
36. A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, H. P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756 (2018).
37. x. Genomics, A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype.
38. L. Blevins, C. J. McDonald, Fisher's Exact Test: an easy-to-use statistical test for comparing outcomes. *MD Comput* **2**, 15-19, 68 (1985).
39. T. M. Snyder *et al.*, Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. *medRxiv*, (2020).
40. S. K. Subudhi *et al.*, Clonal expansion of CD8 T cells in the systemic circulation precedes development of ipilimumab-induced toxicities. *Proceedings of the National Academy of Sciences* **113**, 11919-11924 (2016).
41. J. X. Caushi *et al.*, Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* **596**, 126-132 (2021).
42. G. Oliveira *et al.*, Phenotype, specificity and avidity of antitumour CD8(+) T cells in melanoma. *Nature* **596**, 119-125 (2021).
43. Y. Bordon, TOX for tired T cells. *Nature Reviews Immunology* **19**, 476-476 (2019).
44. T. Sekine *et al.*, TOX is expressed by exhausted and polyfunctional human effector memory CD8⁺ T cells. *Science Immunology* **5**, eaba7918 (2020).
45. D. Ostroumov *et al.*, Transcriptome Profiling Identifies TIGIT as a Marker of T-Cell Exhaustion in Liver Cancer. *Hepatology* **73**, 1399-1418 (2021).
46. J. Li, Y. He, J. Hao, L. Ni, C. Dong, High Levels of Eomes Promote Exhaustion of Anti-tumor CD8+ T Cells. *Frontiers in Immunology* **9**, (2018).
47. F. A. Al-Abbasi *et al.*, Computational Protein Phenotype Characterization of IL10RA Mutations Causative to Early Onset Inflammatory Bowel Disease (IBD). *Frontiers in Genetics* **9**, (2018).
48. Y. Li *et al.*, Unraveling LGALS1 as a Potential Immune Checkpoint and a Predictor of the Response to Anti-PD1 Therapy in Clear Cell Renal Carcinoma. *Pathology & Oncology Research* **26**, 1451-1458 (2020).
49. R. Saleh *et al.*, Expression of immune checkpoints and T cell exhaustion markers in early and advanced

stages of colorectal cancer. *Cancer Immunology, Immunotherapy* **69**, 1989-1999 (2020).
50. L. Zheng *et al.*, Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, abe6474 (2021).
51. N. Seddiki *et al.*, Expression of interleukin (IL)-2 and IL-7 receptors discriminates between human regulatory and activated T cells. *Journal of Experimental Medicine* **203**, 1693-1700 (2006).
52. G. Petkau *et al.*, The timing of differentiation and potency of CD8 effector function is set by RNA binding proteins. *bioRxiv*, 2021.2006.2003.446738 (2021).
53. D. Ordway *et al.*, XCL1 (lymphotactin) chemokine produced by activated CD8 T cells during the chronic stage of infection with Mycobacterium tuberculosis negatively affects production of IFN-γ by CD4 T cells and participates in granuloma stability. *Journal of Leukocyte Biology* **82**, 1221-1229 (2007).
54. B. Wang *et al.*, CXCR6 is required for antitumor efficacy of intratumoral CD8⁺ T cell. *Journal for ImmunoTherapy of Cancer* **9**, e003100 (2021).
55. Z. Liu *et al.*, Low density lipoprotein receptor class A domain containing 4 (LDLRAD4) promotes tumorigenesis of hepatic cancer cells. *Exp Cell Res* **360**, 189-198 (2017).
56. J. Wei, L. Feng, L. Wu, Integrated analysis identified CAPG as a prognosis factor correlated with immune infiltrates in lower-grade glioma. *Clin Transl Med* **10**, e51-e51 (2020).
57. S. Bagchi, R. Yuan, E. G. Engleman, Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. *Annu Rev Pathol* **16**, 223-249 (2021).
58. R. W. Jenkins, D. A. Barbie, K. T. Flaherty, Mechanisms of resistance to immune checkpoint inhibitors. *British Journal of Cancer* **118**, 9-16 (2018).
59. A. Snyder *et al.*, Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. *PLoS Med* **14**, e1002309 (2017).
60. E. Yusko *et al.*, Association of Tumor Microenvironment T-cell Repertoire and Mutational Load with Clinical Outcome after Sequential Checkpoint Blockade in Melanoma. *Cancer Immunol Res* **7**, 458-465 (2019).
61. G. S. Duncan *et al.*, Genetic Evidence for Functional Redundancy of Platelet/Endothelial Cell Adhesion Molecule-1 (PECAM-1): CD31-Deficient Mice Reveal PECAM-1-Dependent and PECAM-1-Independent Functions. *The Journal of Immunology* **162**, 3022-3030 (1999).
62. P. Chambon, The retinoid signaling pathway: molecular and genetic analyses. *Seminars in Cell Biology* **5**, 115-125 (1994).
63. K. E. Dombrowski, S. E. Wright, Construction of a multiple mucin tandem repeat with a mutation in the tumor-specific epitope by a solid-phase gene assembly protocol. *Nucleic Acids Res* **20**, 6743-6744 (1992).
64. W. Huisman *et al.*, Public T-Cell Receptors (TCRs) Revisited by Analysis of the Magnitude of Identical and Highly-Similar TCRs in Virus-Specific T-Cell Repertoires of Healthy Individuals. *Frontiers in Immunology* **13**, (2022).
65. H. Li, C. Ye, G. Ji, J. Han, Determinants of public T cell responses. *Cell Res* **22**, 33-42 (2012).
66. D. B. Page *et al.*, Deep Sequencing of T-cell Receptor DNA as a Biomarker of Clonally Expanded TILs in Breast Cancer after Immunotherapy. *Cancer Immunology Research* **4**, 835-844 (2016).
67. M. A. Postow *et al.*, Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. *Journal for ImmunoTherapy of Cancer* **3**, 23 (2015).
68. A. C. Hopkins *et al.*, T cell receptor repertoire features associated with survival in immunotherapy-treated pancreatic ductal adenocarcinoma. *JCI Insight* **3**, e122092 (2018).
69. W. Roh *et al.*, Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci Transl Med* **9**, (2017).

Figure 1

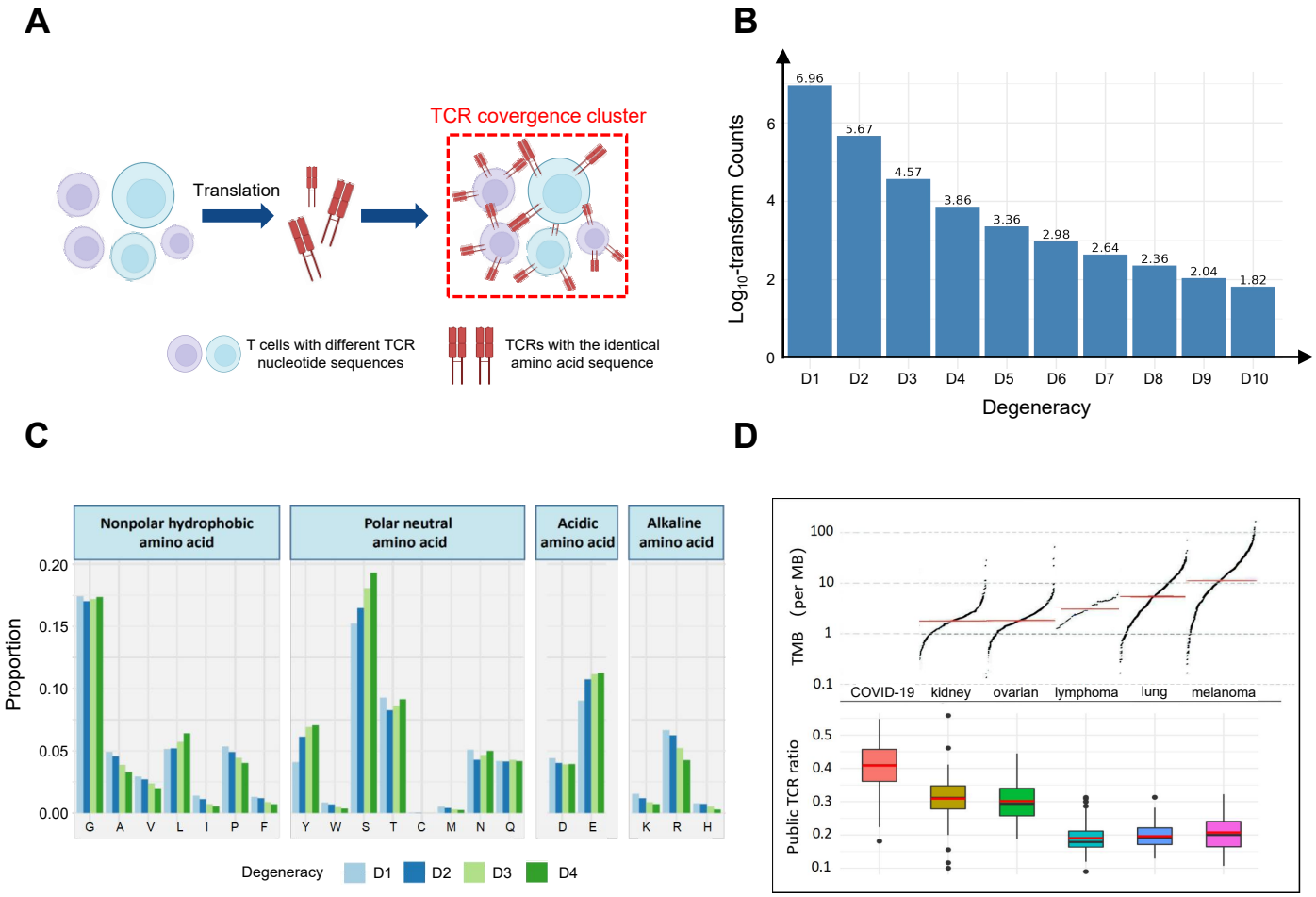

A
Translation
TCR covergence cluster
T cells with different TCR nucleotide sequences
TCRs with the identical amino acid sequence
B
Log10-transform Counts
6.96
5.67
4.57
3.86
3.36
2.98
2.64
2.36
2.04
1.82
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
Degeneracy
C
Nonpolar hydrophobic amino acid
Polar neutral amino acid
Acidic amino acid
Alkaline amino acid
Proportion
G A V L I P F
Y W S T C M N Q
D E
K R H
Degeneracy D1 D2 D3 D4
D
TMB (per MB)
COVID-19 kidney ovarian lymphoma lung melanoma
Public TCR ratio

Figure 2

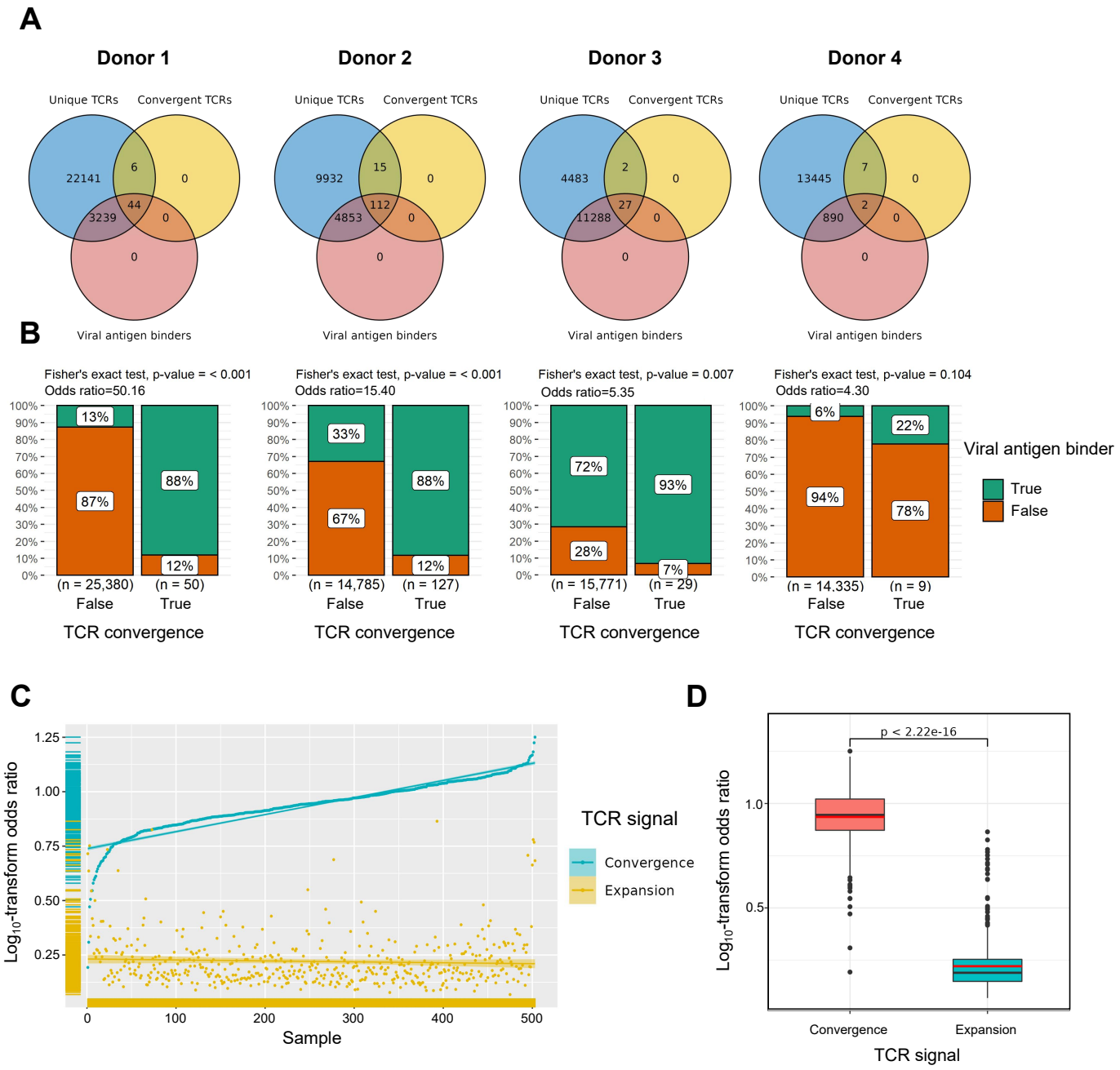

A
Donor 1
Donor 2
Donor 3
Donor 4
Unique TCRs
Convergent TCRs
Viral antigen binders
22141
6
0
3239
44
0
0
9932
15
0
4853
112
0
0
4483
2
0
11288
27
0
0
13445
7
0
890
2
0
0
B
Fisher's exact test, p-value = < 0.001
Odds ratio=50.16
Fisher's exact test, p-value = < 0.001
Odds ratio=15.40
Fisher's exact test, p-value = 0.007
Odds ratio=5.35
Fisher's exact test, p-value = 0.104
Odds ratio=4.30
13%
87%
88%
12%
33%
67%
88%
12%
72%
28%
93%
7%
6%
94%
22%
78%
(n = 25,380)
(n = 50)
(n = 14,785)
(n = 127)
(n = 15,771)
(n = 29)
(n = 14,335)
(n = 9)
False
True
TCR convergence
Viral antigen binder
True
False
C
Log10-transform odds ratio
Sample
TCR signal
Convergence
Expansion
D
p < 2.22e-16
Log10-transform odds ratio
Convergence
Expansion
TCR signal

Figure 3

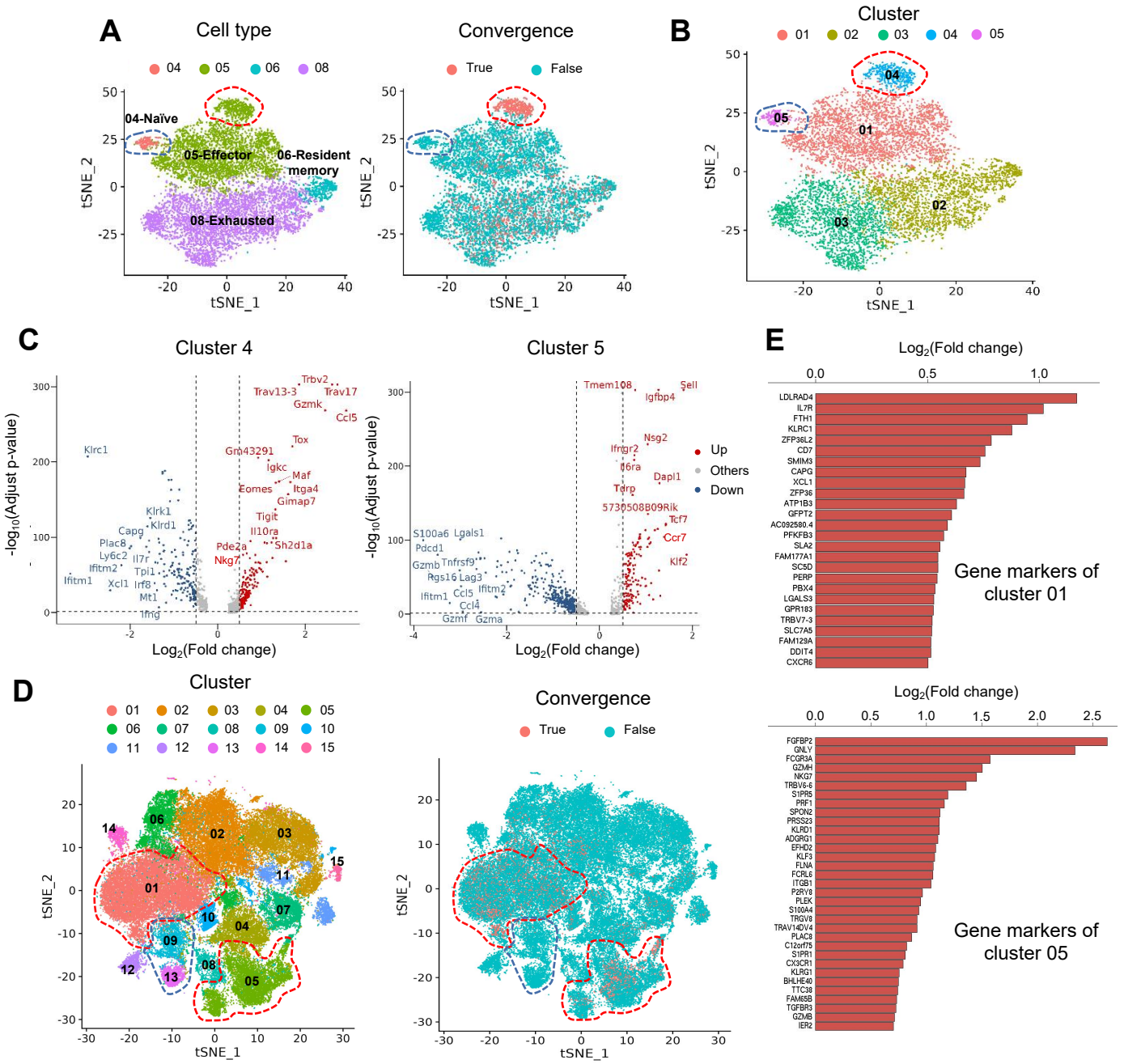
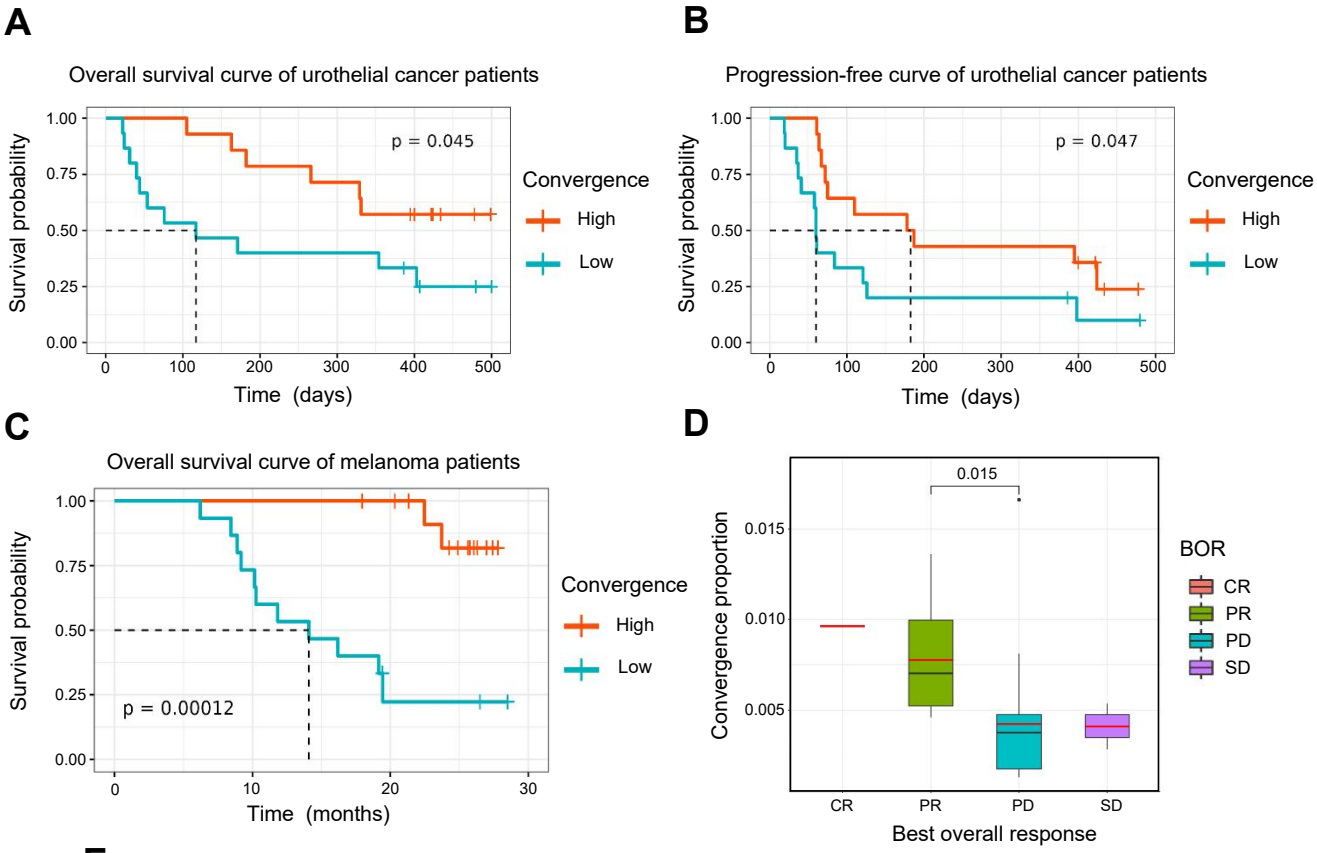

A
Cell type
04
05
06
08
04-Naïve
05-Effector
06-Resident memory
08-Exhausted
tSNE_1
tSNE_2
Convergence
True
False
B
Cluster
01
02
03
04
05
C
Cluster 4
Cluster 5
-log10(Adjust p-value)
Log2(Fold change)
Up
Others
Down
D
Cluster
Convergence
E
Log2(Fold change)
Gene markers of cluster 01
Gene markers of cluster 05

Figure 4



E

Variable		Hazard ratio	p
Convergence	High (N=15)	reference	
	Low (N=15)	140.700 (7.6490 - 2588.12)	*** <0.001
Treatment	Ipilimumab (N=17)	reference	
	Nivolumab (N=13)	1.207 (0.3492 - 4.17)	0.766
Clonality	High (N=15)	reference	
	Low (N=15)	0.059 (0.0046 - 0.76)	0.03 *
Diversity	High (N=15)	reference	
	Low (N=15)	1.054 (0.3122 - 3.56)	0.933

0.01 0.1 1 10 100 1000

Table 1 The Fisher's exact test result of $CD8^+$ clusters

clusters	normal adjacent tissues			tumor tissues		
	convergence ratio	odds ratio	p-value	convergence ratio	odds ratio	p-value
CD8-01-XCL1	22.95%	5.07	<2.2e-16	15.03%	5.66	<2.2e-16
CD8-02-GZMK	2.59%	0.24	1	3.48%	0.85	0.9949
CD8-03-TIGIT	3.69%	0.37	1	2.12%	0.49	1
CD8-04-ZNF683	7.43%	0.79	0.9984	3.72%	1.62	<2.2e-16
CD8-05-FGFBP2	32.44%	6.19	<2.2e-16	26.69%	12.25	<2.2e-16
CD8-06-IFNG	5.80%	0.59	1	8.71%	2.34	9.725E-15
CD8-07-IHSPA6	22.75%	3.93	<2.2e-16	4.99%	1.27	0.004411
CD8-08-PRF1	4.84%	0.5	1	3.11%	0.77	0.9404
CD8-09-IL7R	1.02%	0.1	1	1.44%	0.35	1
CD8-10-IFI6	11.11%	1.23	0.2523	5.54%	1.42	0.0003609
CD8-11-STMN1	6.07%	0.64	0.9603	3.93%	0.98	0.5834
CD8-12-KLRB1	0.48%	0.03	1	0.13%	0.03	1
CD8-13-CCR7	0.00%	0	1	0.00%	0	1
CD8-14-CXCR6	0.45%	0.01	1	5.00%	1.26	0.4325
CD8-15-CTLA4	0.00%	0	1	1.68%	0.41	0.2