Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers

Article *in* Clinical Cancer Research · March 2020

DOI: 10.1158/1078-0432.CCR-19-3249

CITATIONS
133

READS
592

12 authors, including:

Hongyi Zhang
University of Pennsylvania/The Children's Hospital of philadelphia
29 PUBLICATIONS **925** CITATIONS
SEE PROFILE

Longchao Liu
Institute of Microbiology, Chinese Academy of Sciences
26 PUBLICATIONS **1,170** CITATIONS
SEE PROFILE

Sachet A Shukla
Broad Institute of MIT and Harvard
153 PUBLICATIONS **46,136** CITATIONS
SEE PROFILE

Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers

Hongyi Zhang[1], Longchao Liu[2], Jian Zhang[3], Jiahui Chen[1], Jianfeng Ye[1], Sachet Shukla[4], Jian Qiao[2], Xiaowei Zhan[5], Hao Chen[2], Catherine J. Wu[4], Yang-Xin Fu[2], and Bo Li[1,6]

ABSTRACT

Purpose: Cancer antigen–specific T cells are key components in antitumor immune response, yet their identification in the tumor microenvironment remains challenging, as most cancer antigens are unknown. Recent advance in immunology suggests that similar T-cell receptor (TCR) sequences can be clustered to infer shared antigen specificity. This study aims to identify antigen-specific TCRs from the tumor genomics sequencing data.

Experimental Design: We used the TRUST (Tcr Repertoire Utilities for Solid Tissue) algorithm to assemble the TCR hypervariable CDR3 regions from 9,700 bulk tumor RNA-sequencing (RNA-seq) samples, and developed a computational method, iSMART, to group similar TCRs into antigen-specific clusters. Integrative analysis on the TCR clusters with multi-omics datasets was performed to profile cancer-associated T cells and to uncover novel cancer antigens.

Results: Clustered TCRs are associated with signatures of T-cell activation after antigen encounter. We further elucidated the phenotypes of clustered T cells using single-cell RNA-seq data, which revealed a novel subset of tissue-resident memory T-cell population with elevated metabolic status. An exciting application of the TCR clusters is to identify novel cancer antigens, exemplified by our identification of a candidate cancer/testis gene, *HSFX1*, through integrated analysis of HLA alleles and genomics data. The target was further validated using vaccination of humanized HLA-A*02:01 mice and ELISpot assay. Finally, we showed that clustered tumor-infiltrating TCRs can differentiate patients with early-stage cancer from healthy donors, using blood TCR repertoire sequencing data, suggesting potential applications in noninvasive cancer detection.

Conclusions: Our analysis on the antigen-specific TCR clusters provides a unique resource for alternative antigen discovery and biomarker identification for cancer immunotherapies.

Introduction

Antigen-specific tumor-infiltrating T lymphocytes (TIL) play a central role in cancer immunity (1–3), with demonstrated applications in cancer immunotherapies, including checkpoint blockade (4–6) and adoptive cell transfer therapies (7, 8). Therefore, identification of antigen-specific TIL is critical to understanding tumor–immune interactions and designing individualized treatments. However, this task remains challenging despite extensive clinical efforts (9, 10). First, cancer antigens may come from diverse sources, including missense mutations (11, 12), frameshift insertions or deletions (13, 14), tissue-specific gene overexpression (15, 16), and other antigenic processes (17–20), making it difficult to profile all the possible targets. In addition, the antigen-binding CDR3 region on the T-cell receptor (TCR) is extremely diverse (21) and their targets are usually unknown. Thus, limited progress has been made in the analysis of TIL repertoire despite pressing clinical needs. This is because statistical significance is usually difficult to reach in such analyses unless using a large cancer cohort.

Efforts have recently been made to partition the immune repertoire into groups linking to antigen specificity (GLIPH; ref. 22) or evaluate the similarity of CDR3s with known specificity for functional predictions (TCRdist; ref. 23). These studies set the stage for using clustering for detection of antigen-specific TCRs when the knowledge of the antigens is unavailable. GLIPH was applied to the data of infectious disease and reliably defined TCR groups that can contact with antigenic peptides, while TCRdist provided a quantitative measure of TCR similarity. Both computational frameworks have the potential to study the repertoire of cancer-associated TCRs. However, due to the extremely diverse interactions between cancer antigens and tumor-reactive TILs, their specificity in recognizing antigen-specific TCRs need to be further evaluated.

In this work, we systematically compared different clustering methods and introduced a new tool to identify the potentially antigen-specific T cells using a novel CDR3 dataset profiled from over 9,700 tumor RNA-seq samples from the Cancer Genome Atlas (TCGA). Similar studies have been conducted to investigate patterns of shared TCRs in human and mice (24, 25). However, they were not intended to provide in-depth analysis on the TCR clusters or their associated antigens in cancer. Our analysis integrated information of the antigen-specific TIL clusters, cancer genomics data, patient HLA genotypes, single-cell RNA-seq data, and immune repertoire sequencing data (26, 27). This pan-cancer multi-omics analysis led to several interesting findings, which might provide novel targets for late-stage cancer treatment, and suggest alternative avenues for cancer diagnosis. Specifically, we explored the phenotypes of antigen-specific T cells in the tumor microenvironment, and identified a previously unreported

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, Texas. [2]Department of Pathology, UT Southwestern Medical Center, Dallas, Texas. [3]Beijing Institute of Basic Medical Sciences, Beijing, China. [4]Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts. [5]Department of Clinical Science, UT Southwestern Medical Center, Dallas, Texas. [6]Department of Immunology, UT Southwestern Medical Center, Dallas, Texas.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

H. Zhang and L. Liu contributed equally to this article.

Corresponding Authors: Bo Li, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390. Phone: 214-648-1654; Fax: 214-648-4067; E-mail: bo.li@utsouthwestern.edu; and Yang-Xin Fu, yang-xin.fu@utsouthwestern.edu

Clin Cancer Res 2020;26:1359–71

doi: 10.1158/1078-0432.CCR-19-3249

Translational Relevance

This analysis extensively identified cancer-associated T-cell receptors using a novel computational algorithm and large genomics data integration. We made three clinically relevant discoveries: (i) A subset of antigen-specific tumor-infiltrating T cells carry the tissue-resident memory (T_{rm}) phenotype and express cytotoxic molecules. We showed that the presence of these T_{rm} cells can be used as a predictor for patient outcome in multiple cancer types. (ii) Using T-cell receptor (TCR) clusters, we predicted HSP transcription factor X-linked 1 (*HSFX1*) to be a novel cancer-associated antigen with restricted expression in endometrial or colorectal tumors. *HSFX1* can serve as a candidate for therapeutic vaccine development. (iii) Using independent study cohorts, we observed significantly elevated levels of cancer-associated TCRs in the blood of patients with cancer compared with healthy donors. This result might suggest an alternative avenue to develop noninvasive diagnostic approaches using blood TCR repertoire.

tissue-resident memory T-cell (T_{rm}) subpopulation with altered metabolic program. In addition, we identified novel candidate cancer-associated antigens and performed *in vivo* validation. Our findings may expand the current pool of cancer antigens for future therapeutic vaccine development. Finally, we demonstrated that the cancer-associated TCRs derived from the TCGA samples could also be observed in the peripheral blood repertoire of patients with both early- and late-stage cancer of independent cohorts. We then performed a proof-of-principle analysis to distinguish patients with early-stage cancer from healthy individuals using the content of cancer antigen–specific T cells.

Materials and Methods

Data resources information

TCGA level-2 RNA-seq data aligned to hg19 human reference genome by MapSplice (28) were downloaded from GDC legacy archive (https://portal.gdc.cancer.gov/legacy-archive/). Gene expression data (TPM), mutation annotation files, and patient clinical information of the TCGA cohort were downloaded from GDAC broad firehose (https://gdac.broadinstitute.org/). Tumor purity information was downloaded from the Cistrome TIMER website (http://cistrome.org/TIMER/misc/AGPall.zip). TCR repertoire data and patient information for the HCMV cohort, late-stage melanoma, and early breast cancer samples, were downloaded from AdaptiveBiotechnology immunoSeq Analyzer (https://www.adaptivebiotech.com/). Antigen-specific CDR3 sequence information for benchmarking iSMART were downloaded from VDJdb (https://vdjdb.cdr3.net/). GLIPH software package was accessed from GitHub (https://github.com/immunoengineer/gliph). Single-cell gene expression data and matched TCR information were downloaded from GEO database (accession number GSE114724).

Materials and animal model

HSFX1-derived 9-mer peptide was synthesized by GenScript; CpG oligonucleotide ODN 1826 was purchased from InvivoGen with catalog number 1826-1; polyinosinic-polycytidylic acid sodium salt, or poly (I:C), was ordered from Millipore Sigma with catalog number P1530-25MG. Immunocompetent C57BL/6J and transgenic C57BL/6-Mcph1Tg(HLA-A2.1)1Enge/J mice were obtained from Jackson Laboratory (JAX:000664 and JAX:003475). All mice were maintained under specific pathogen–free conditions, and all animal procedures were performed in accordance with the experimental animal guidelines set by the Institutional Animal Care and Use Committee of the UT Southwestern Medical Center (Dallas, TX) under animal protocol number (APN): 2015-101350.

iSMART for pairwise CDR3 alignment and clustering

iSMART takes *M* complete CDR3 sequences as input, where complete CDR3 region is defined as the last cysteine in the variable gene to the first amino acid in the FGXG motif in the joining gene (29). It first orders the CDR3s by length and then performs pairwise comparisons. For CDR3s with different lengths, iSMART allows at most one insertion in the comparison, and imposes a gap penalty (default 6). Alignment scores are calculated on the basis of BLOSUM62 matrix, with individual matched score capped at 4. The third to $(n\text{-}3)^{th}$ positions of the CDR3s are used for scoring, where *n* is the CDR3 amino acid sequence length. Pairwise alignment score is normalized by the length of the longer CDR3 sequence (*n*-4, excluding first and last 2 amino acids). After calculation of the *M*-by-*M* pairwise scoring matrix, a predefined cut-off value (default 3.5) is applied to filter out all the low scoring comparisons. iSMART then performs a depth-first search on the matrix to identify all the connected CDR3 clusters, and output all the CDR3s with empirical cluster IDs. iSMART is written in Python and the source code is publicly available.

Although iSMART is benchmarked to run without variable gene assignment, it supports the input with variable gene information. In this mode, the pairwise alignment on the CDR3 regions is the same except that iSMART uses the 5^{th} to $(n\text{-}3)^{th}$ positions of the sequence for scoring. As the first 4 amino acids of the CDR3s are mainly determined by the variable gene, we made this change to avoid repeated use of variable gene information. In the pairwise sequence comparison step, the CDR1 and CDR2 regions of two TCRs are also used to calculate alignment scores under the same rules. The total score is scaled to 8, where CDR3 and variable gene contribute equally, and a cut-off value (default 7.5) is used to generate the CDR3 clusters. iSMART in variable gene mode was tested using the 15 antigen benchmark dataset, which is described in section below, and reached a higher specificity of 94.3% (100/106 clusters have unique antigen assignment) than without variable gene input.

iSMART, TCRdist, and GLIPH performance evaluation

Both iSMART and GLIPH can predict antigen-specific CDR3 clusters without variable gene information. TCRdist can also be used to perform this task, but requires additional codes to identify closely related TCR groups. We programmed the codes according to the descriptions in the original article and added the input/output modules to allow TCR β chain clustering without variable gene usage. We evaluated the performances of all three methods using TCRs of known antigen specificity in VDJdb (30). We selected 15 9-mer human antigens with balanced number (K) of associated TCRβ CDR3s (100<K<1,000; Supplementary Data S1). CDR3s associated with more than one antigens were excluded, resulting in a total of 2,347 unique sequences. Both iSMART and GLIPH were run on this dataset with default parameters.

The command line for iSMART is python iSMART.py –f human15aa.txt –v, where –v option is applied to disable the use of variable gene. For GLIPH, the command line is./gliph-group-discovery.pl –tcr human15aa.txt.

Interestingly, although iSMART performs time-consuming pairwise sequence alignments, its computational time (1 second) is less

than TCRdist (24 seconds) or GLIPH (approximately 1 hour) on MacBook Pro with 3.1 GHz Intel core i7 and 16 GB DDR3 memory. Therefore, iSMART has the computational efficiency to scale up for larger TCR repertoire datasets.

The clustering of TCRdist relies on the cutoff of pairwise TCR distances. Lower cutoff results in higher specificity, but will reduce sensitivity and cluster size. In this work, we chose the cutoff to be 12 to match the fraction of large clusters (>3 TCRs). At this cutoff, it calls 20.2% large clusters, comparable with iSMART (18.4%) and GLIPH (22.0%). If we reduce the cutoff by 1, this fraction drops down to 7%, which is not useful for downstream analysis.

As each CDR3 is uniquely linked to one antigen in the benchmark dataset, we defined cluster purity (p) as the number of the most abundant antigen divided by the number of CDR3s in a cluster. We use the percent of completely pure ($P = 1$) clusters as a measure for specificity. To make visualization of the clustering specificity, we computed the cross-antigen classification errors as follows: the 15-by-15 cross-antigen matrix (M) is initialized by 0, and for each cluster, let A denote the set of antigens associated with the CDR3s in this cluster, we add 1 to all the entries in M[A, A]. Therefore, if A contains only one antigen, the diagonal values for M will increase by 1. Otherwise the off-diagonal values will increase by 1, which are considered as classification errors. We looped through all the clusters and used the final output to plot the heatmaps in Supplementary Figure S2.

Noncancerous public TCR identification

A critical preprocessing procedure in our analysis is to exclude noncancerous public TCRs to reduce false positives in our downstream analysis. We used a cohort of noncancer individuals with TCR repertoire sequencing data available (31). There are two batches of this cohort, with the first batch containing 666 human cytomegalovirus (HCMV) infected ($n = 289$) or normal individuals. The HCMV-infected individuals can be used as control samples for our purposes. The second batch contains 120 individuals. We will use the first batch to remove public TCRs and the second for downstream analysis, to avoid systematic bias. It is known that cancer-specific T cells are also present in healthy individuals in the form of low abundant naïve T cells (32). Therefore, to prevent false removal of bona fide cancer-specific CDR3s, we restricted our analysis within the top 10,000 most abundant sequences, sufficient to cover all the clones with $\geq$5 copies that are expected to be effector T cells. We combined all the sequences as normal CDR3s to be removed in the TCGA data before iSMART clustering. The resulting dataset as well as samples used in this analysis are available as Supplementary Data S2.

We removed public sequences from the 170,516 complete CDR3s and obtained 82,427 nonpublic sequences for downstream analysis. As the TCR repertoire data in the public domain are mainly β chain sequences, currently we do not have enough data to eliminate public α-chain CDR3s from the analysis. We will rely on future efforts to sequence more TCR α-chain repertoire samples to define public α-chain CDR3 sequences.

Analysis of single-cell sequencing data

Postprocessed gene expression data in sparse matrix format (mtx) and TCR hypervariable CDR3 sequences with matched cell barcodes were downloaded directly from the GEO database. In total, there are 5 samples from 3 patients, BC09, BC10, and BC11. BC10 has the largest overlap with TCGA-derived CDR3 clusters. For BC10, we selected 1,103 genes with SD $\geq$1 and performed t-Distributed Stochastic Neighbor Embedding (tSNE) analysis on the 4,926 cells using these genes for dimension reduction. This filter is purely for visualization purposes. Two-dimensional scatter plot using tSNE values were generated to visualize the distributions of genes of interest. On the basis of the locally enriched pattern of 18 clustered cells, we defined a subgroup of 44 cells. For each of the 1,103 genes, we performed Wilcoxon rank sum test between this group and other cells and used Benjamini–Hochberg method to evaluate FDR. These results, including the cell barcodes for the selected group, are available in Supplementary Data S1. *ZNF683* expression levels in the TCGA samples were split into two groups by the median level. Survival analysis for *ZNF683* was performed using Cox proportional hazard model on the binary variable corrected for patient age.

We performed cell trajectory analysis for selected clonotypes in the breast cancer samples. For sample BC10, we selected 418 cells with CDR3 sequences found in the $CD8^+$ subgroup identified in the tSNE plot, and used R package monocle (33) to perform cell ordering by pseudotime. As the direction of pseudotime is arbitrary, we used representative biomarkers for T-cell activation to determine the beginning of the trajectory, and identified the T_{pre} population. The T_{rm} clusters were then selected at the end of the trajectory. Spearman correlation between each gene expression level and pseudotime was calculated, and we selected important biomarkers for cell identity (*IL7R, TCF7, CCR7*), cytotoxicity (*GZMB, PRF1, IFNG*), exhaustion (*PD-1, LAG3, TIM-3*), resident memory signature (*SELL, KLRG1, CD103*), and metabolic enzymes. For BC11, we first merged the two biological replicates into one dataset and selected 31 cells with *IL7R*$\leq$1, *TCF7*$\leq$1, *GZMB*$\geq$5, *ZNF683*$\geq$5, and CD103$\geq$10 as tissue-resident T cells and used all the 728 cells sharing the same CDR3s with these 31 cells to perform the pseudotime trajectory analysis. These cells in total come from 11 clonotypes, but for the individual clonotype evolution analysis, we removed two clonotypes with $n = 1$ and showed the remaining 9 in Supplementary Fig. S9C. We did not identify enough cells using the same selection criteria for T_{rm} cells for sample BC09.

Gene expression analysis

We performed a genome-wide correlative analysis to identify genes associated with counts (K) of clustered CDR3s in each individual. TCGA samples with both gene expression and genomic estimated tumor purity information were further selected, and 10 cancer types with more than 100 samples were kept. For each cancer, we calculated partial Spearman correlation between K and the expression level for each gene. Tumor purity is corrected in this analysis as it is expected to impact gene expression profiles (34) and is correlated to T-cell infiltration. False discovery rate is estimated using Benjamini–Hochberg procedure by pooling all the P values. After FDR calculation, we selected genes with correlation ≤ -0.1 and FDR ≤ 0.05. We further removed genes significant in only one cancer type. This step resulted in 414 genes, as shown in the Heatmap in **Fig. 2**. The representative clusters were manually selected to include cancer–gene blocks with significantly negative associations between K and gene expression levels across multiple cancer types, and was intended for visualization purposes only. Unbiased GO term enrichment analysis was performed for all 414 genes using GSEA.

We also performed differential gene expression analysis to identify novel cancer-associated antigens (**Fig. 4**). 120 clusters with CDR3 length 20$\geq$L$\geq$13 and with $\geq$10 sequences were selected. For each cluster, we performed one-tailed Wilcoxon rank sum test for each gene between clustered and nonclustered individuals from all cancers,

pooled all the *P* values and estimated FDR using Benjamini–Hochberg correction. This step selected 3,524 significant results (FDR<0.05 and fold change ≥10), including 1,409 unique genes spanning 115 clusters. Fold change was calculated for each cluster, as the median expression value of the samples in the CDR3 cluster divided by that of those not in the cluster. If the denominator is zero, we used an arbitrarily small number 10^{-13}. Of all the protein-coding genes, *HSFX1* has the top significant value, and is associated with clusters 1,724 and 1,767. We performed a second differential gene expression analysis to visualize the top highly expressed genes, by combining samples in the two clusters.

Vaccination of naïve and transgenic mice and ELISpot assay

C57BL/6J and C57BL/6J-HLA-A2.1Tg mice were purchased from the Jackson Laboratory. All mice were maintained under specific pathogen-free conditions at UT Southwestern Medical Center (Dallas, TX). Ten micrograms of VMF or VRF peptide was mixed with 50 μg ODN1826 and 100 μg poly (I:C) in 100 μL PBS and then subcutaneously injected to the mouse on day 0 and day 14. Single-cell suspensions were prepared on day 18 post first vaccination. Splenocytes were seeded at 4×10^5 per well and stimulated with either 10 μg peptide or PMA + Ionomycin for 36 hours. ELISpot assay was performed using an IFNγ ELISpot assay kit (BD Biosciences) according to the manufacturer's instructions. Spots were enumerated by ImmunoSpot Analyzer (CTL).

HLA allele–binding prediction

All the HLA allele–binding predictions in this work were performed using either NetMHC or NetMHCpan online server. We implemented NetMHCpan for less common HLA alleles not covered in NetMHC. For missense mutations, the input peptide is a 17-mer peptide with mutated amino acid in the middle. For frameshift mutations, we included 8-mer before and all the amino acid sequence after the mutation locus. For cancer-associated antigens, we downloaded the complete protein sequence from Uniprot (www.uniprot.org), and input the fasta file to NetMHC/NetMHCpan server. Biding of the control peptide VRF for *in vivo* validation to HLA-A*02:01 was predicted using NetMHC server. Default rank cutoffs were applied to define weak (≤2) or strong binders (≤0.5). We exhaustively predicted the binding affinities of 415 9-mer peptides produced by the *HSFX1* protein. The HLA alleles ($n = 32$) were selected to match the 9 individuals with *HSFX1* mRNA expression. In total, we identified 6 epitopes (FQRDSPHLL, SAPPATPVM, AAVPGPAAL, YVPGSPTQM, NSYGPVVAL, and VMFPHLPAL) binding to all the individuals. However, only VMF binds to A*02:01 with high affinity. On the basis of our search on the Jackson laboratory catalog, the available transgenic humanized mouse models can only test the binding for 3 MHC-I alleles: A*02:01, A*11:01, and B*27:05. Among them, A*02:01 is the only allele carried by a subset of the 9 individuals. Therefore, we used VMF for downstream experimental validation for the immunogenicity of the *HSFX1* protein. Same analysis was performed for *TSSK2* protein sequence. We predicted the binding affinity of 349 9-mer peptides to 22 alleles from the 5 patients in cluster 189. We identified two peptides (SAYSERLKF and GRIYIIMEL) as strong binders to all 5 individuals.

HXFX1 IHC staining of endometrial tumor and healthy tissue samples

Formalin-fixed paraffin-embedded (FFPE) tissue slides of endometrial serous carcinoma samples and healthy tissue array were obtained from UTSW Department of Pathology under Institutional review board protocol STU 072018-066. *HSFX1* IHC antibody was purchased from LiveSpan BioSciences Inc with catalog number: LS-C165049. Four-micron–thick paraffin sections of formalin-fixed samples were cut, deparaffinized in xylene, and rehydrated through graded concentrations of ethanol. Heat-induced antigen retrieval was performed by boiling in sodium citrate buffer (10 mmol/L sodium citrate, pH 6.0) for 10 minutes at high power and then 15 minutes at low power in the microwave oven rated 200–2,200 W. Endogenous peroxidase activity was blocked by incubating with Dual Endogenous Enzyme Block (Dako, K4065) for 30 minutes at room temperature, followed by the incubation overnight at 4°C with the primary antibody to *HSFX1* 1:100. The subsequent staining was developed using the Dako EnVision + Dual Link System-HRP (DAB+) kits (Dako, K4065). Briefly, the sections were incubated with Labeled Polymer-HRP Rabbit/Mouse for 30 minutes and staining was achieved by adding 100 μL of DAB+ Chromogen diluted 1:50 in substrate buffer for 5 to 10 minutes. Nuclei were counterstained with hematoxylin. All IHC staining results were verified by pathologist (H. Chen)at Department of Pathology, USTW (Dallas TX).

Prediction of cancer disease status

In this analysis, we compared 3 TCR repertoire datasets from different studies, including pre/post anti-CTLA4 treatment late-stage melanoma (melanoma), early breast cancer (breast cancer), and HCMV cohort (HCMV) as normal control. To avoid systematic bias after public TCR removal, we randomly sampled 50 individuals from the second batch ($n = 120$) of the HCMV cohort in this analysis. Direct comparison between different study cohorts will be biased toward sequencing depth. Therefore, we conducted a down-sampling procedure to ensure the comparability. The targeted capture protocol applied for TCR repertoire sequencing allowed one read to completely cover the whole CDR3 region, and read count is used to estimate clonal abundance. We first calculate the size for each TCR-seq library (N), which is the summation of the read counts (m) for all the CDR3s. A combined vector of CDR3s with length N was made, with each CDR3 sequence i repeated by m_i times, where m_i is the read count for CDR3 sequence i. For the melanoma cohort, we used all the cancer samples ($n = 21$ for either pre- or posttreatment), and randomly sampled 100 individuals with replacement as normal control. For each sample, we downsampled the library to K = 100,000 reads, each read being a CDR3 amino acid sequence. The read count (m') for each unique CDR3 was then calculated. For each sample, CDR3s with identical sequence to one of the cancer-associated CDR3s were selected, and the summation of m' for all the selected CDR3s was used as predictor (CDR3 abundance) for cancer status. For breast cancer cohort, same strategy was applied, except that we used K = 60,000 for peripheral blood mononuclear cells (PBMC) and 20,000 for TIL samples.

Statistical analysis

Statistical analyses were performed using R statistical programming language (35). Survival analysis was implemented using Cox proportional hazard model in R package *survival*. All survival analyses performed in this work were restricted to one cancer type. ROCs and AUC calculations were performed with R package *AUC*. tSNE plots for single-cell analysis were generated using *Rtsne* (36). Single-cell pseudotime trajectory analysis was performed using *cellrangerRkit* and *monocle*. Two-way ANOVA test for comparing different treatment groups of vaccinated mice was performed using commercial software GraphPad Prism.

Results

Detection of antigen-specific CDR3 clusters with iSMART

We have previously described the TRUST algorithm (37) for sensitive detection of TCR hypervariable CDR3 sequences using bulk tissue RNA-seq data. In this work, we applied a later version of TRUST (38) with improved sensitivity to 9,709 TCGA tumor RNA-seq samples and assembled 1.5 million CDR3 sequences (**Fig. 1**). Of these, 170,000 were complete productive CDR3s, following the IMGT nomenclature (29). A sizeable fraction of the human T-cell repertoire is public, derived from biased V(D)J recombination (39), and are present in both healthy and diseased individuals. To exclude the irrelevant public TCRs that are prevalent in healthy individuals (24), we compared the TCGA TIL CDR3s with a large cohort of TCR repertoire data from noncancer individuals (31) (Materials and Methods). CDR3s observed in these samples with high abundances were excluded, leaving 82,000 nonpublic sequences.

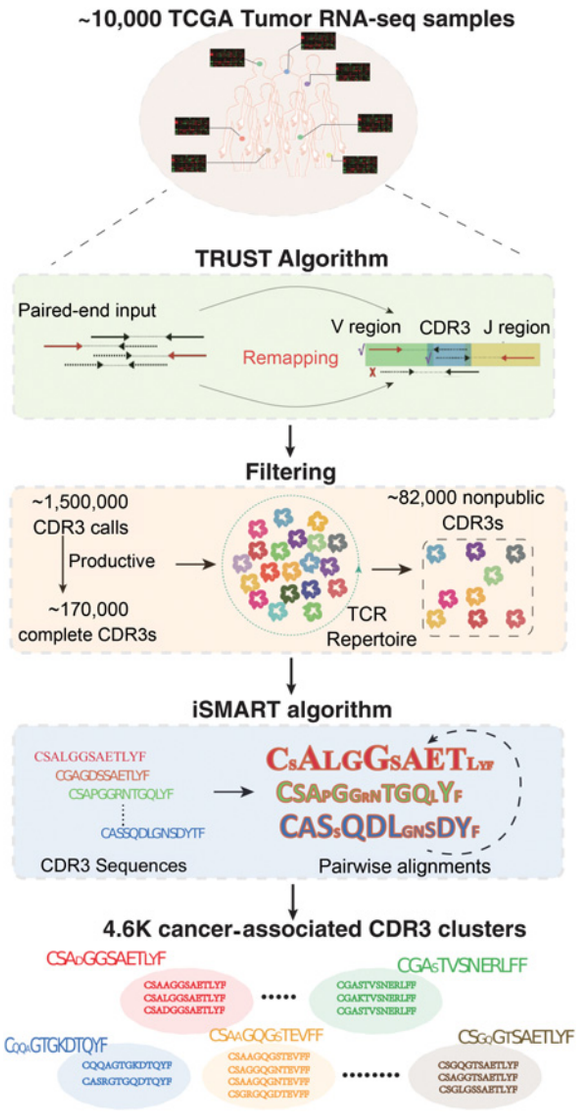

Figure 1.

Methodology summary and performance evaluation for iSMART. Flowchart illustrating the analytic procedures carried out in this work to generate CDR3 clusters using TCR receptor sequences extracted from the tumor RNA-seq data.

Identification of antigen-specific CDR3 groups from TCR repertoire data is highly desirable, yet challenging due to the high diversity of CDR3 regions (21) and promiscuous binding between T-cell receptors and antigenic peptides (40–42). Previous approaches (22, 23) demonstrated that CDR3s grouped into motif-sharing clusters are expected to recognize the same antigens, and reported two methods, GLIPH and TCRdist, to perform antigen-specific clustering of the β chain CDR3 regions. In this work, we developed a fast clustering method, immunoSimilarity Measurement by Aligning Receptors of T cells, or iSMART (Materials and Methods). TCRs specific to 15 known epitopes validated in previous experiments were used as benchmark data. Each of the three methods was applied to this dataset to compare the performances. Implementation of iSMART and GLIPH used their default parameters, where a distance cutoff of 12 was applied for TCRdist to ensure equally high fraction of large TCR clusters (with >3 sequences; Supplementary Fig. S1A).

Specificity was measured by cluster purity, the percentage of the most abundant antigen assignment in the cluster. Perfect purity (= 1) is reached only when all the TCRs in a cluster are specific to the same antigen. We observed that iSMART clustering generally achieves the highest purity across different sizes, yet with slightly lower cluster counts (Supplementary Fig. S1B). In this study, we are interested in large clusters with >3 TCRs, as they yield higher statistical power to detect shared antigens compared with smaller ones. Of all three methods, the mean purity of large clusters is the highest for iSMART (Supplementary Fig. S1C). We also investigated if the higher specificity of iSMART is antigen-dependent. Antigen-specificity of the clustered TCRs was visualized by heatmaps (Supplementary Fig. S2A). Off-diagonal signal indicates TCRs specific to different antigens were grouped together. We noted that iSMART has the lowest amount of nonspecific assignment. This observation was further quantified by cluster purities, where iSMART clustering has the highest purity for 14 of the 15 antigens (Supplementary Fig. S2B).

From these results, we concluded that iSMART is a more specific clustering method, and applied it to group the 82,000 nonpublic TIL CDR3 sequences from the TCGA data. We detected a total of 4,657 clusters (**Fig. 1**). As most clusters contain more than one individual, we also used the term "CDR3 cluster" to denote the subset of patients carrying the CDR3s in a given cluster. A total of 18,113 CDR3 sequences were grouped into these clusters, and were expected to be enriched for cancer-associated TCRs. To note, each CDR3 is unique to individual patient, but may be presented in multiple patients in the dataset. These sequences will be clustered as well.

Features of CDR3 clusters and association with tumor gene expression profiles

The number of sequences in the clusters spans two orders of magnitude (Supplementary Fig. S3A), and for each sample, the number of clustered CDR3s (***K***) also spans two orders of magnitude (Supplementary Fig. S3B). Higher value of ***K*** is expected to be associated with higher abundance of antigen-specific T cells in the tumor microenvironment; accordingly, lower K might be related to higher level of immunosuppression. For each gene, we calculated the partial Spearman correlation between ***K*** and its expression levels (Supplementary Data S1), controlled for tumor purity, which is expected to influence both values (ref. 43; Materials and Methods). Among the genes with top positive correlations are putative T-cell activation markers, including *TBX21 (T-bet), ICOS, TIGIT*, and granzymes (Supplementary Fig. S4). Gene ontology enrichment (44)

analysis suggested that the top 500 genes are strongly enriched for immune cell activation and immune responses (Supplementary Table S1). Interestingly, on the top of the list, there is a pair of lysophosphatidylserine receptors, *GPR174* and *P2RY10*, which have been identified as suppressors for regulatory T-cell function (ref. 45; Supplementary Fig. S4). These results strongly suggest that CDR3s clustered by iSMART are enriched for activated T cells in the tumor microenvironment.

We next investigated genes negatively correlated with $\boldsymbol{K}$ as potential regulators for T-cell inactivation and exclusion (**Fig. 2**). The 414 genes with correlation < -0.1 and FDR < 0.05 were significantly enriched in mitochondrial-related biological processes (GO term enrichment analysis, FDR $= 1.4 \times 10^{-36}$). We observed four interesting clusters. Cluster (i) contains a putative oncogene *MAPK3* (46), inhibition of which has been linked to enhanced antitumor immune response (47). This cluster also harbors a key glycolysis enzyme, *ALDOA*, which has recently been shown to impair T-cell infiltration and cytotoxicity (48). Cluster (ii) contains two oncogenes, *RHOD* and *PKP3*, the former recently being implicated in immune suppression (49). We also identified a number of other metabolic enzymes, including protein metabolic enzyme *POMGNT1*, cytochrome c enzymes *COX6A1* and *UQCRQ*, lipid metabolic enzymes *DGAT1* and *FAAH*, etc, supporting the recently elucidated immunosuppressive role of cancer metabolism pathways (50).

Identification of T_{rm} subpopulations with distinct metabolic status

To further elucidate the phenotypes of the T-cell clonotypes with clustered CDR3s, we analyzed a recently generated single-cell RNA-seq (scRNA-seq) data with matched TCR information (51). Using the TCGA-derived CDR3s as clonotype markers, we identified a number of clustered T-cell clones in the three breast tumor scRNA-seq samples. We first studied sample BC10, which has the largest amount ($n = 55$) of cells carrying clustered CDR3s. The selected 55 cells were visualized on the background of all 4,926 cells using tSNE (52) plots, and observed a local clustering of 18 events in a restricted region (**Fig. 3A**). All 18 events share the same β chain CDR3 sequence, and we delineated the region containing these cells as a separate $CD8^+$ subgroup ($n = 44$). Differential gene expression analysis on the cells in this group against all the others (Supplementary Data S1) revealed upregulated genes both involved in T-cell cytotoxicity (*GZMB, PRF1, IFNG*) and exhaustion (*PDCD1, LAG3*; Supplementary Fig. S5A). Interestingly, the top targets showed high consistency with a recently reported T_{rm} signature (53), including upregulation of *CD103* (*ITGAE*), *TIGIT*, and *GZMB* and downregulation of *SELL (CD62L), KLF2*, and *KLRG1*. This group also expresses a number of other previously reported T_{rm} markers (54) (Supplementary Fig. S6), including transcription factor *ZNF683*, or *HOBIT* (homolog of *Blimp-1* in T cells), a key regulator for T_{rm} differentiation (55). We observed significant association of

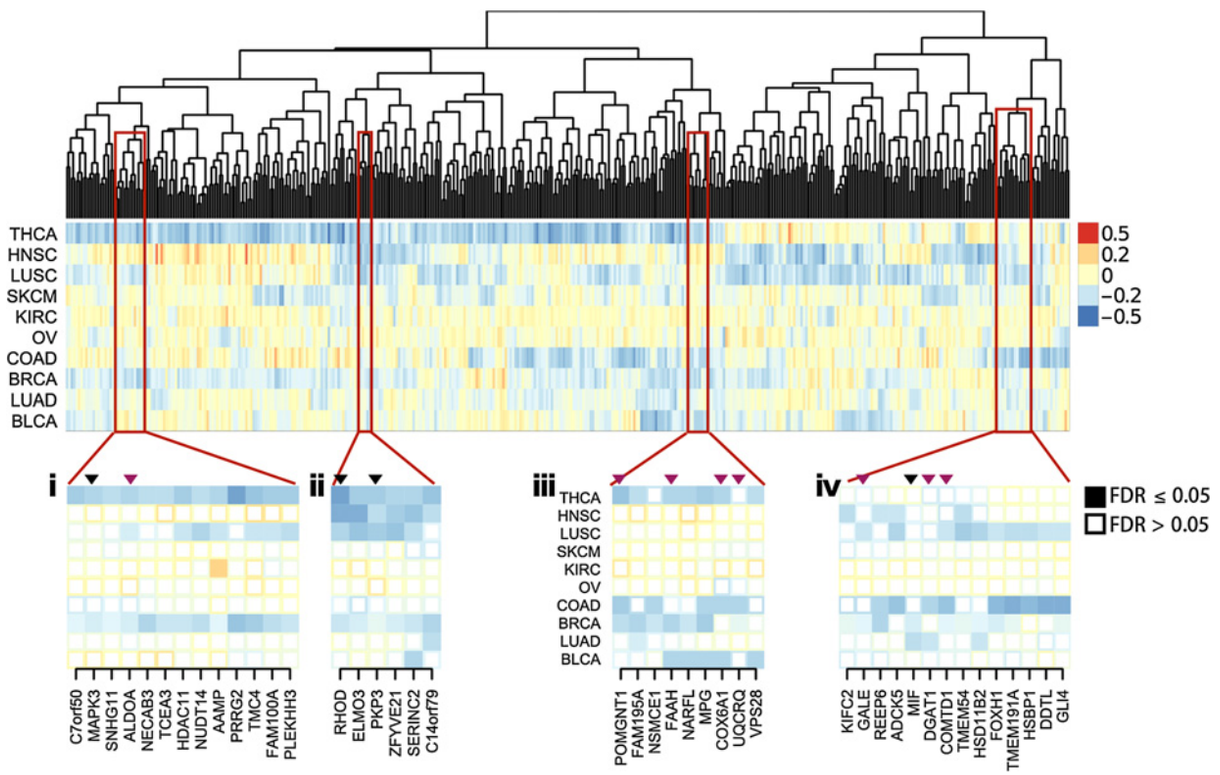

Figure 2.

Potential negative regulators for T-cell activation in the tumor microenvironment. Genes with Spearman correlation $\rho \leq -0.1$ and FDR ≤ 0.05 were selected for visualization in the heatmap. Hierarchical clustering on ρ was performed to order the genes into similar groups across different cancer types. Four representative clusters with putative oncogenes (labeled by black arrows) or recently identified metabolic enzymes (red arrows) were displayed as smaller heatmaps in the bottom panels. Statistical significance was evaluated using partial Spearman correlation test correcting for tumor purity, and FDR was performed using Benjamini–Hochberg procedure.

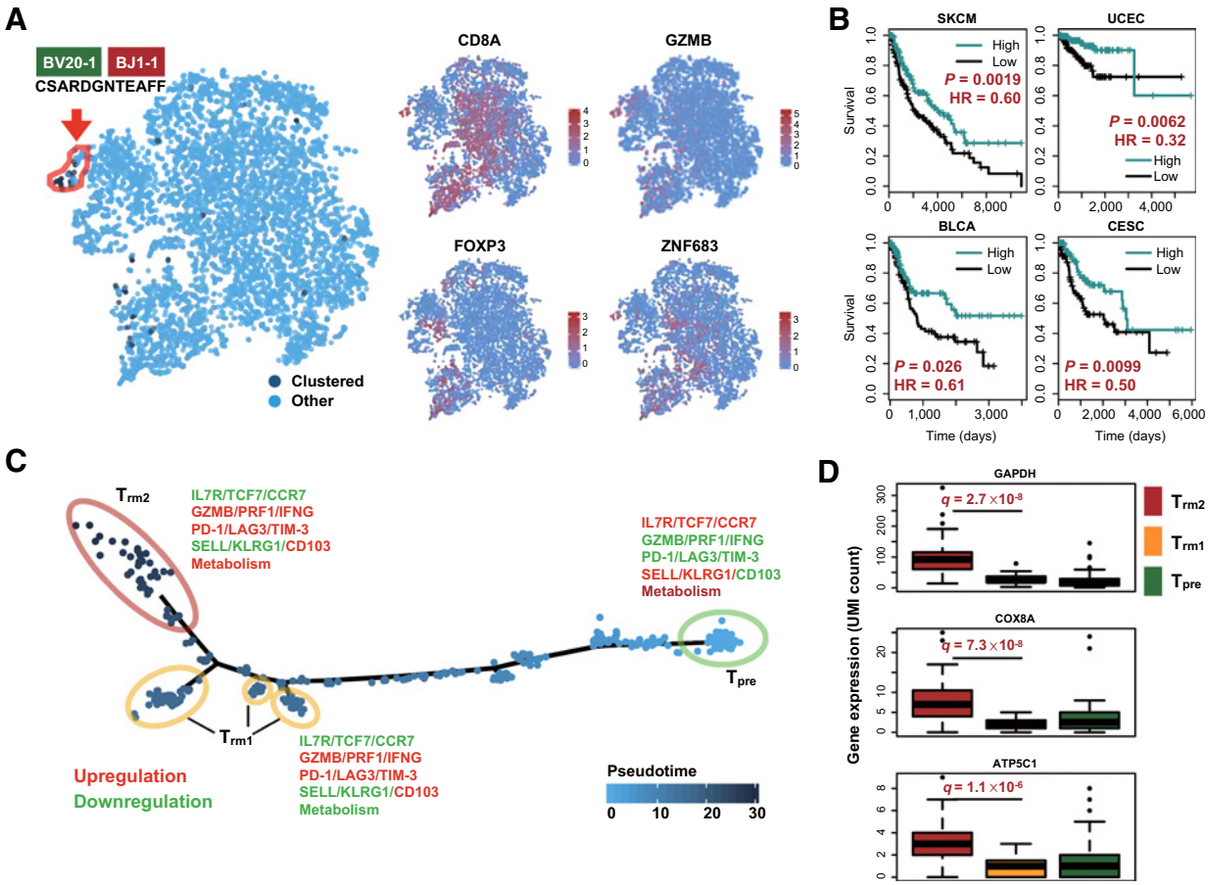

Figure 3.

iSMART-clustered clonotypes showing tissue-resident memory phenotype. **A,** tSNE plots showing the distributions for clustered clonotypes in the TIL population (left), and the expression levels of selected putative markers for cell identity (*CD8A/FOXP3/ZNF683*) or function (*GZMB*; right four panels). All selected markers passed FDR = 0.05. Color legends for gene expression are in log scale. **B,** Kaplan–Meier curves for four TCGA cancers showing the survival benefit for *ZNF683* high expression. For each cancer, median value was applied to define high or low groups. Statistical significance and HRs were evaluated using Cox proportional hazards model. **C,** Pseudotime trajectory plot illustrating the inferred evolutionary path. Cell clusters located on the beginning or end of the trajectory were manually selected. Representative markers significantly correlated (Spearman correlation test, FDR < 0.05) with pseudotime inference were labeled for each cluster, with red for negative (high in T_{pre}) and green for positive (high in T_{rm}) correlations. **D,** Boxplots showing the distributions for selective metabolic enzymes in the three cell clusters shown in **C**. Statistical significance for differential gene expression between T_{rm1} and T_{rm2} was evaluated using Wilcoxon rank sum test, with FDR corrected by Benjamini–Hochberg method.

ZNF683 expression with better outcomes in multiple cancer types (**Fig. 3B**), supporting the antitumor role for T_{rm} cells.

T cells undergo profound differentiations in the tumor microenvironment, and it is unclear which evolutionary path T cells have taken to become resident memory cells. The 44 cells in the subgroup come from 20 productive clonotypes, which in total contain 418 cells. We performed single-cell trajectory analysis (33) to infer the progression of these TILs (**Fig. 3C** and Materials and Methods). The pseudotime trajectory starts from a group of cells (T_{pre}) expressing high levels of *IL7R, SELL*, and *KLRG1*, with low expression of effector molecules (*GZMB, PRF1, IFNG*) and exhaustion markers (*PDCD1, LAG3, TIM-3*). These markers agree with the signatures of effector T cells primed by antigens and still maintained mobility, and will later differentiate into memory cells (56). We designated them as precursor cells, as effector memory T cells will lose mobility markers (*KLRG1* and *SELL*; ref. 53) after tissue homing. Two clusters were observed at the end of the trajectory, both carrying the resident memory markers, and we named them T_{rm1} and T_{rm2}. Notably, the T_{rm1} cluster largely overlaps with the previously identified $CD8^+$ subgroup. Differential expression analysis revealed that compared with T_{rm1}, the newly identified T_{rm2} population significantly upregulates metabolic enzymes, including *GAPDH, COX8A, ATP5C1*, etc (**Fig. 3D**; Supplementary Fig. S5B). The high expression of metabolic related genes in the T_{rm2} population is not a consequence of dying cells (Supplementary Fig. S7). Pseudotime trajectories for individual clonotypes revealed that the differentiation of T cells into resident memory status might be receptor dependent (Supplementary Fig. S8). Specifically, we observed different distributions across the three populations for different T cell clones: (i) mainly in T_{pre} and T_{rm1}; (ii) mainly in T_{rm1} and T_{rm2}. Few clones were distributed only in T_{pre} and T_{rm2}, without the presence of T_{rm1}. This observation suggested that infiltrating T cells may undergo sequential evolution (T_{pre} to T_{rm1} then T_{rm1} to T_{rm2}) to reach the terminal status.

We analyzed other scRNA-seq samples to see whether this observation is reproducible, and indeed, a strikingly similar pseudotime trajectory for resident memory T cells was observed in sample BC11 (Supplementary Fig. S9A). Representative markers observed in sample

BC10 also showed significant differences across the three cell groups with consistent trends. Higher expression of metabolic genes was also observed in the T_{rm2} group (Supplementary Fig. S9B). In addition, the corresponding clonotypes also displayed two modes of distributions (Supplementary Fig. S9C), consistent with our findings for BC10. Differentially expressed genes between two resident memory T-cell groups are available in Supplementary Data S1. These results indicate that the T_{rm} cells further divide into two populations distinguished by low or high metabolic statuses, and the differentiation from their precursors into these populations is dependent on the TCRs.

Identification of novel cancer-associated antigen candidates

Most current studies focus on searching for tumor antigens from mutated genes with matched HLA alleles combining the elution of peptides from the MHC molecules (57). However, malignant cells may overexpress a number of genes that are usually silenced in most normal tissues, resulting in novel antigenic targets for cancer treatment. This is exemplified by the clinical use of cancer/testis antigens, that have restrictive expression in the male germ cells (15, 16). We performed a genome-wide differential gene expression analysis on each of the 120 qualifying CDR3 clusters, and identified a total of 1,409 significant (FDR<0.05) genes from 115 clusters (Materials and Methods; Supplementary Data S1). In this analysis, the gene expression levels of individuals within a given cluster were compared with those of other patients. Of these, two clusters (1724 and 1767) showed an interesting enrichment in colon and endometrial cancers, with distinct CDR3 conservation patterns (**Fig. 4A**). We performed differential expression analysis on the combined samples from the two clusters, and identified Heat Shock Transcription Factor X-linked 1 (*HSFX1*) as the top hit (**Fig. 4B**). This gene has extremely low expression (median TPM$\leq$0.02) in all the tissue types covered in the GTEx data (58), while expressed (TPM$\geq$1) in 13% colorectal and 73% endometrial cancers, with higher expression (TPM$\geq$2) in 27% patients (Supplementary Fig. S10). There is over 100-fold change in the expression levels between some tumor samples and the normal tissues. It is also a favorable predictor of survival for endometrial cancer (**Fig. 4C**). Therefore, we hypothesized that the tissue-specific overexpression for *HSFX1* may be a trigger for antitumor immune response.

Of the 17 colon or endometrial cancer samples from cluster 1724 and 1767, 9 express *HSFX1* (Supplementary Fig. S11) and have solved HLA genotypes (59). Computational prediction for HLA allele binding suggests that *HSFX1* protein generates six 9-mer peptides with high affinity to all 9 individuals. Among them, VMFPHLPAL (VMF) is the only one binding to HLA*02:01, which is selected for experimental validation (Supplementary Table S2 and S3 and Materials and Methods). To test whether VMF can activate T cells *in vivo*, we synthesized the 9-mer antigen peptide and injected it into HLA-A*02:01 humanized transgenic mice (Materials and Methods). We used peptide

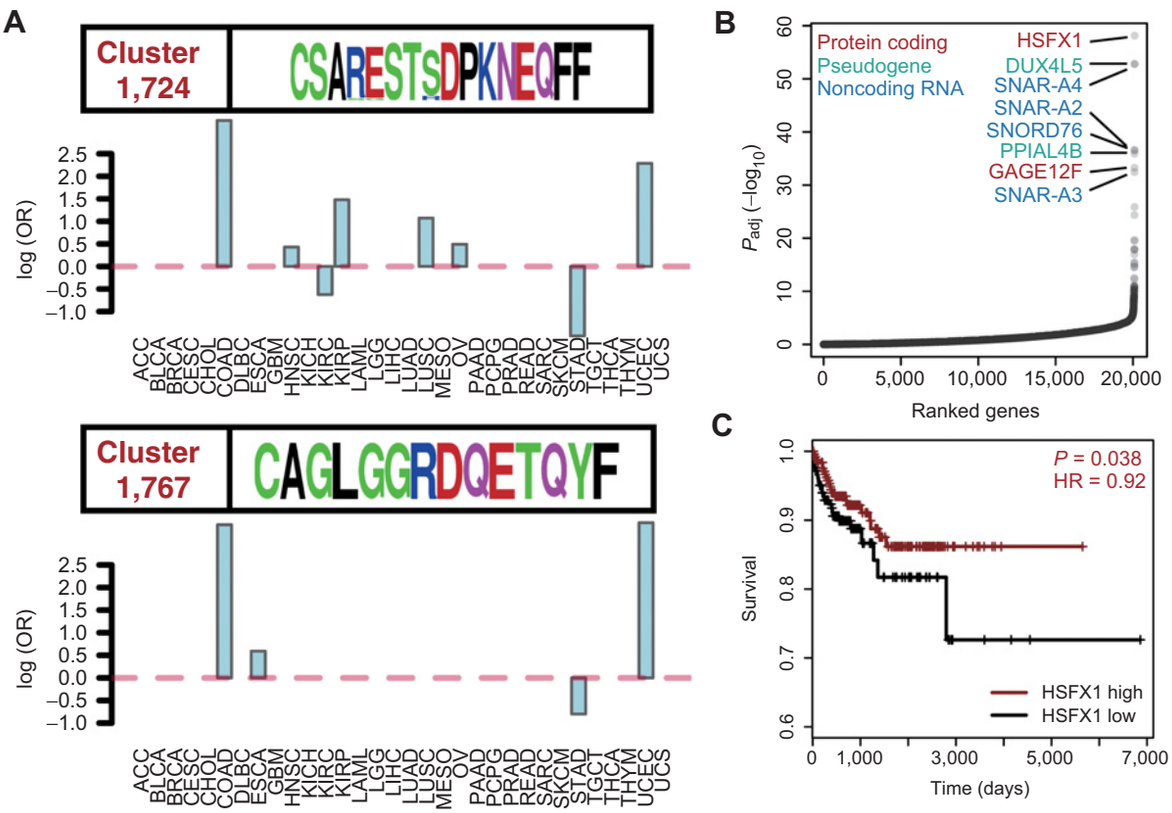

Figure 4.

Identification of *HSFX1* as a candidate cancer-associated antigen. **A,** Selective enrichments in colon and endometrial cancers of samples in CDR3 clusters 1724 and 1767. CDR3 amino acid conservation patterns were displayed in the top panel for each barplot. **B,** Genes ranked by *P* values from differential gene expression analysis, with top hits labeled in colored texts. *HSFX1* has the most significant *P* value among all the genes. Statistical significance was evaluated using Wilcoxon rank sum test with FDR correction. **C,** Kaplan–Meier survival curves for patients with endometrial cancer with or without *HSFX1* expression, separated by median expression value. Statistical significance and HRs for *HSFX1* levels were estimated using Cox proportional hazards model on binary input of *HSFX1* groups, corrected for patient age.

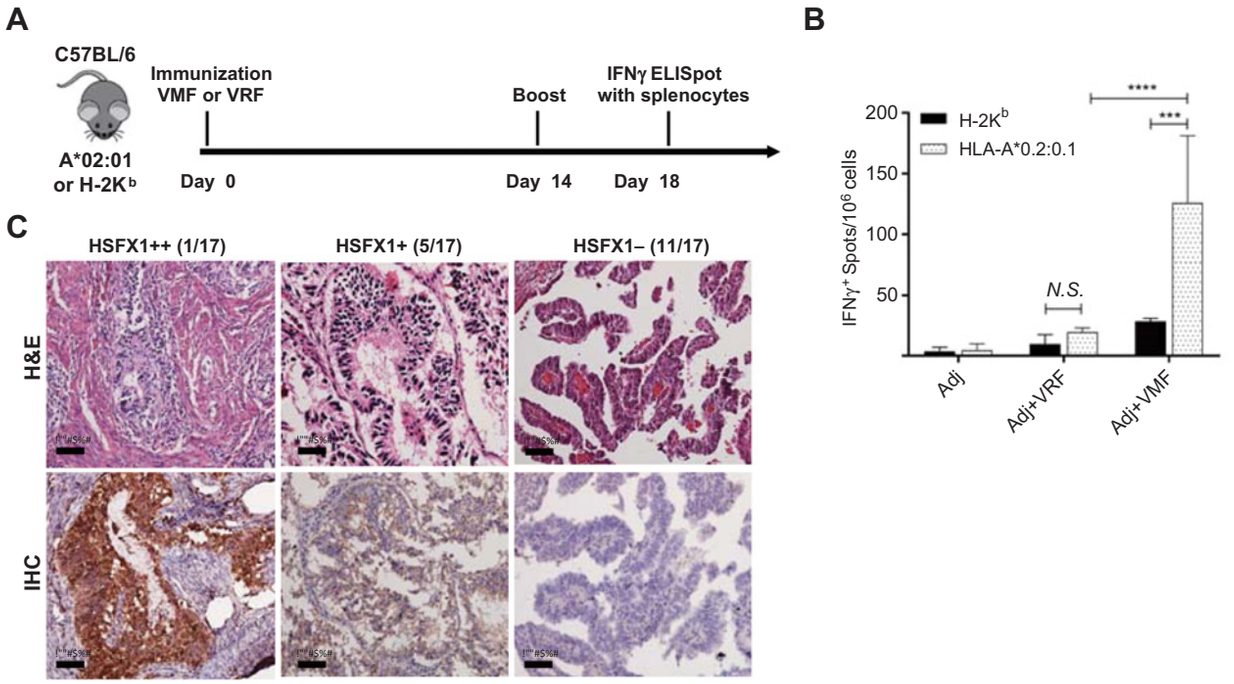

Figure 5.

Immunogenicity of the 9-mer peptide derived from HSFX1 protein in HLA-A*02:01 transgenic mice. **A,** HLA-A*02:01 transgenic mice (female, $n = 4$) were subcutaneously immunized with 10 μg peptide mixed with 100 μg poly(I:C) and 50 μg CpG1826. Fourteen days after vaccination, mice were boosted with the same vaccine. Four days later, splenocytes were isolated for IFNγ ELISpot assay. Significant difference of antigen-specific T-cell response from the control peptide was observed (**B**). Data are expressed as the means ± SD, and representative results from two independent experiments are shown. Statistical analysis was performed by two-way ANOVA. ***, $P < 0.001$; ****, $P < 0.0001$. Adj is short for adjuvant injected during vaccination. VMF, antigen peptide VMFPHLPAL; VRF, control peptide V**R**FPHLPAL. **C,** Hematoxylin and eosin (H&E, top) staining of cytoplasm and nucleus, and IHC (bottom) staining of HSFX1 protein expression in selected endometrial serous carcinoma samples. High-grade ESC pathology was confirmed with H&E slides. Title for each panel describes the strength of the signal (++, +, or −) and the number of corresponding samples out of the 17 total.

VRFPHLPAL, which has one amino acid difference, as control, because it is predicted not to bind HLA-A*02:01. After 18 days, splenocytes of the vaccinated mice were collected to perform an IFNγ ELISpot assay for antigen-specific T-cell responses (**Fig. 5A**). Compared with the control peptide (VRF), we observed significantly higher IFNγ response in the transgenic mice, but not in identically primed immunocompetent mice with H-2K^b genotype (**Fig. 5B**; Supplementary Fig. S12).

This result indicates that VMF is a bona fide binder to human HLA-A*02:01 allele, and can be recognized by mouse TCRs to elicit an *in vivo* immune response. As human TCR repertoire is significantly larger than the mouse repertoire (32), it is likely that the epitope can be engaged by human T cells as well. Notably, VMF may not be the only peptide produced by *HSFX1* that can elicit an immune response. Our experiment proves that at least one epitope from *HSFX1* can be immunogenic. Nonetheless, *in vivo* immunogenicity of *HSFX1* as a self-antigen in humans requires evasion from central tolerance (60), given tissue-restricted gene overexpression. Such genes have been used in cancer vaccine therapies, including *MAGEA3, NY-ESO-1, MLANA*, etc. Our analysis demonstrated the cancer-specific overexpression of *HSFX1* at mRNA level, but it remains unclear whether protein expression is also restricted to malignant tissues. To confirm, we collected 17 endometrial serous carcinoma (ESC) samples and a panel of healthy tissues from different organs as control (Materials and Methods). IHC staining of *HSFX1* was performed on all samples. Six of 17 ESC tumors showed positive IHC staining, with 1 strong positive sample (**Fig. 5C**). In contrast, no healthy tissue except placenta, which is an immune privileged (61) organ, expressed *HSFX1* proteins (**Fig. 5D**). In particular, *HSFX1* staining for benign proliferative endometrium and myometrium are negative, suggesting that *HSFX1* is a specific marker for the malignant cells. These data indicated that *HSFX1* protein has restricted expression in a subset of ESCs, supporting our hypothesis that *HSFX1* is a novel cancer-associated antigen.

In addition to *HSFX1*, we also identified a putative cancer/testis antigen, *TSSK2*, with expression restricted to esophageal and stomach tissues (Supplementary Fig. S13). *TSSK2* also generates two peptides binding to HLA alleles from all 5 patients with expressing *TSSK2* from cluster 189 (Supplementary Tables S4 and S5 and Materials and Methods). These results suggest that genes with tumor-specific overexpression might produce cancer-associated antigens and elicit T-cell responses. Our analysis revealed a number of such unmutated genes as promising targets for cancer vaccine development.

Existence of cancer-associated CDR3s in the blood repertoire of patients with cancer

In the above analysis, we observed multiple potential tumor antigens showing significant associations to the iSMART identified CDR3 clusters, suggesting that the clustered CDR3s are enriched for cancer-associated T cells. We therefore investigated whether it is feasible to

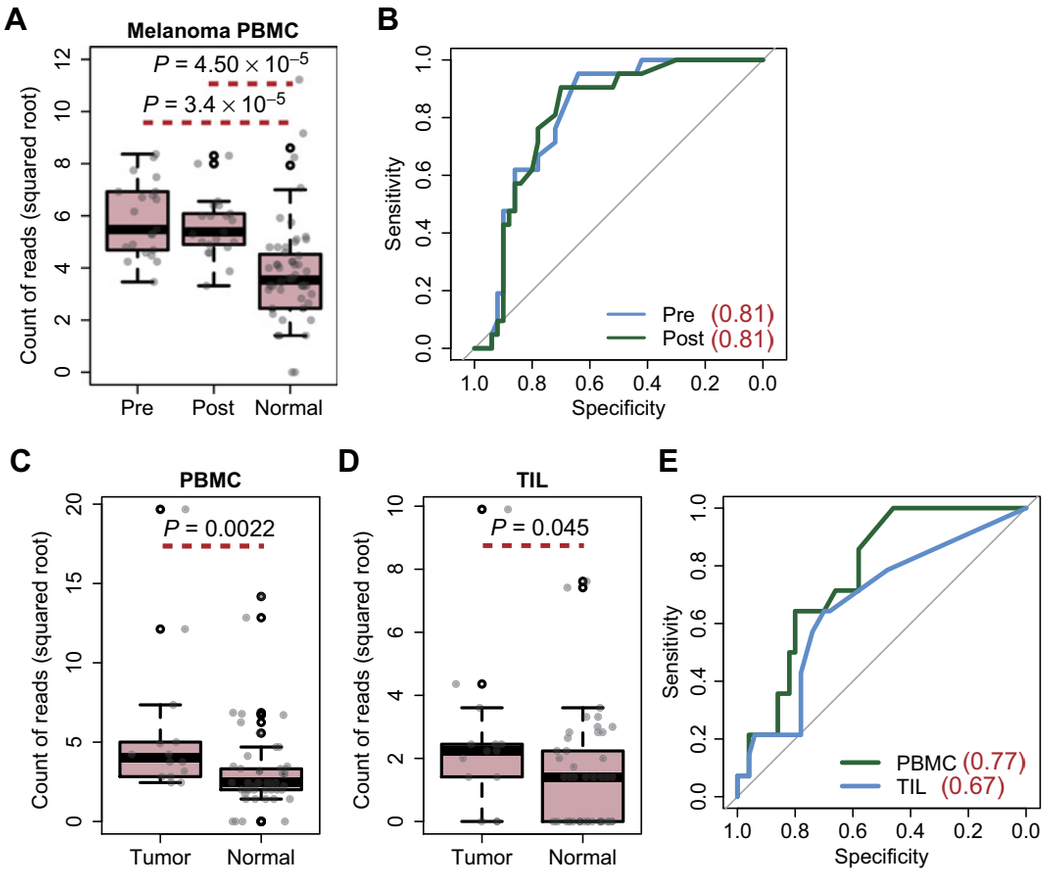

Figure 6.

Prediction of late- and early-stage cancers using cancer-associated CDR3s. **A,** Boxplot showing the distributions of the read counts for cancer-associated CDR3s for pre– or post–anti-CTLA4 treatment late-stage melanoma and normal control samples. TCR repertoire data from all the samples were derived from PBMCs. **B,** ROC curves for using CDR3 read count as a predictor for late-stage melanoma ($n = 21$). Numbers in the figure legend are area under the curve (AUC) values. **C** and **D,** Cancer-associated CDR3 read count distributions for early-stage breast cancers compared with normal samples, with cancer samples being PBMCs (**C**) or TILd (**D**). **E,** ROC curves for using the abundance of cancer-associated CDR3s PBMC or TIL samples ($n = 16$) as predictors for early breast cancer onset. AUC values are shown in the legend. Statistical significance was evaluated using Wilcoxon rank sum test between labeled groups.

detect these CDR3s in the TCR repertoire sequencing data profiled from the PBMCs of the patients with cancer. In this analysis, none of the CDR3s in the TCR sequencing samples were used to generate the TCGA TCR clusters, the latter serving as the reference cancer-associated TCRs. We used all the TCRs in the clusters as reference sequences and searched for their abundance in the additional TCR-seq data. Neither age ($P = 0.96$, Spearman correlation test) nor gender ($P = 0.25$, two-sided Wilcoxon rank sum test) was associated with CDR3 abundance in the healthy donor cohort. We then studied a cohort of 21 patients with late-stage melanoma before and after anti-CTLA4 treatment (27). When compared with the healthy donors, we identified significantly higher abundance of cancer-associated CDR3s in the patients' blood samples (**Fig. 6A**; Materials and Methods). Using cancer-associated CDR3 counts as a disease predictor, pre- and post- PBMC samples reached similar AUC of 0.81 (**Fig. 6B**).

We next evaluated the performance of the above approach on the challenging yet more useful task of predicting early cancer status via PBMC repertoire. We applied the same method to study early-stage breast cancer samples with both PBMC and TIL repertoires sequenced (26). Indeed, both repertoires showed significantly higher levels of cancer-associated CDR3s than healthy donors (**Fig. 6C** and **D**), indicating that the abundance of cancer-associated CDR3s is able to distinguish healthy individuals from patients with early-stage cancer as well. Using iSMART-clustered CDR3 counts as a predictor, we observed an AUC of 0.77 for PBMC samples (**Fig. 6E**). At threshold of 13, it reached 64% sensitivity and 80% specificity. On the basis of a previous study on cancer and inflammation (62), we speculated that during early cancer development, the immune system is able to recognize and respond to a few shared cancer associated antigens (such as *HSFX1*), and produce a significant amount of effector T cells in the circulation, allowing for cancer detection from the peripheral T-cell repertoire.

Discussion

Despite extensive efforts and critical clinical applications, antigen-specific TILs remain largely uncharacterized, mainly because it is experimentally challenging to identify the immunogenic cancer antigens and to profile the tumor-reactive T cells. In this work, we extracted CDR3s from the tumor RNA-seq data, and identified a large

number of CDR3 clusters with high sequence similarity. Because of the excessive diversity of the TCR repertoire, the probability that different individuals independently produce near-identical non-public TCRs is extremely low, suggesting that shared antigen-specificity is the main cause for the generation of these CDR3 clusters. Previous studies have also shown that TCRs sharing motifs on the CDR3 region may recognize the same antigen (22, 23). Therefore, we used iSMART-identified CDR3 clusters (Supplementary Data S1) as surrogates for TCR antigen specificity, and comprehensively analyzed the tumor-specific TILs using a large human cancer cohort.

We leveraged the iSMART-clustered clonotypes to perform an in-depth analysis of a tumor scRNA-seq dataset with solved TCR sequences, and observed an interesting group of $CD8^+$ T cells. The marker set for this group is highly consistent with a recent study on T_{rm} (53), suggesting reproducible identification of T_{rm} in triple-negative breast tumor microenvironment. Using CDR3 as clonotype markers, we further identified two subpopulations of T_{rm} with distinct metabolic states, and observed divergent evolutionary paths to these states among different TIL clonotypes. Our results suggest that after initial homing to the target tissue, T_{rm} may switch to a high metabolic status. This result is potentially linked to the immunosuppressive roles for metabolic enzymes in the malignant cells, which they use to compete for resources required for T-cell survival and cytotoxic functions.

It has been shown from protein structure studies that one antigenic peptide may bind to dissimilar CDR3 sequences with different docking strategies (63, 64), suggesting that individuals responding to the same antigen may carry divergent TCR sequences. Indeed, we observed two distinct CDR3 sequences from clusters 1724 and 1767, which were both predicted to recognize the same antigen derived from cancer-associated antigen *HSFX1*. We performed *in vivo* experiments using transgenic humanized mice to show that a 9-mer peptide derived from a predicted antigen *HSFX1* is able to bind HLA-A*02:01, and induce reliable T-cell responses. These results, combined with the observation that *HSFX1* has restricted expression in selected cancers, and its positive clinical relevance, strongly indicated that it might escape central tolerance in humans and become an immunogenic cancer-associated antigen.

A fraction of the CDR3 clusters remain unassociated with any potential targets, likely due to the unexplored categories of cancer-associated antigens. In our gene expression analysis, we observed significant associations of some clusters with noncoding RNAs (Supplementary Data S1), such as lncRNA, pseudogenes, and small nucleolar RNAs (snoRNA). Ribosome profiling data suggests that many noncoding RNAs are actually translated (65), which may serve as valid cancer antigens when overexpressed in the tumor tissues. snoRNAs participate in many biological processes, including RNA splicing. Thus, their abnormal expression in selected cancer types may produce new antigenic targets from alternative splicing. Post-translational modification (PTM) may also generate foreign peptide products that are subject to immunosurveillance (17). However, due to the insufficiency of related data, it is currently challenging to study the antigenic potentials of these mechanisms in cancer immunity.

Our study has several limitations. First, the use of tumor RNA-seq data to profile the infiltrating TCR repertoire has limited statistical power to call TCR clonotypes, resulting in smaller and lower number of clusters. With deep TCR-seq profiling of TCGA samples, more cancer-associated TCRs and antigens would be identified. Second, to benchmark iSMART and other methods, we applied a dataset of epitopes mainly from infectious diseases. Ideally, TCRs specific to cancer antigens will be optimal to test the specificity of the clustering methods. Unfortunately, due to limited known cancer-associated antigens, the related TCRs are insufficient to evaluate the methods. Third, in this study, we used TCR clusters as surrogates for shared antigen-specificity. Because of the complexity of the adaptive immune system, it is unclear whether tumor is the main contributor of shared antigens. This may explain the large fraction of clusters without any association in our analysis. Fourth, using bulk tissue RNA-seq data to call TCR is not possible to pair the TCR αβ chains. With one chain, it is not feasible to directly validate the CDR3s predicted to recognize certain antigens through *in vitro* synthetic TCRs. The *in vivo* experiment for the VMF peptide demonstrated the TCR recognition potential of the epitope, but is not conclusive evidence for its immunogenicity in humans. Immunospot assays using fresh white blood cells from qualifying cancer patients will be needed to provide definitive validations of the predicted antigens. Finally, the analysis of cancer status using blood TCR repertoire is preliminary, with prediction performance insufficient for practical clinical applications. A more sophisticated machine learning algorithm on a larger cancer-associated TCR set as training data will be needed to further improve the prediction accuracy. Should this method be developed, additional clinical samples from patients with early-stage cancer and healthy donors will be needed to test its sensitivity and specificity, and evaluate its potential utilities in cancer screens.

In summary, we provided a comprehensive analysis to characterize cancer antigens and tumor-reactive T cells. The tool and datasets from this study can be applied to the rapidly generated tumor single-cell sequencing and RNA-seq data to expand the current repertoire of cancer-associated TCRs. Therefore, we anticipate broad utilities of our work for future studies to identify more antigens and biomarkers for cancer immunotherapies.

Disclosure of Potential Conflicts of Interest

C.J. Wu is an employee/paid consultant for Neon Therapeutics. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: B. Li, L. Liu, Y.-X. Fu
Development of methodology: B. Li, L. Liu, J. Zhang
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): B. Li, L. Liu, J. Chen, J. Ye, H. Chen, C.J. Wu
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): B. Li, H. Zhang, L. Liu, J. Chen, S. Shukla, J. Qiao, X. Zhan, H. Chen
Writing, review, and/or revision of the manuscript: B. Li, H. Zhang, L. Liu, J. Chen, C.J. Wu
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): B. Li, J. Zhang
Study supervision: B. Li

Acknowledgments

This work is supported by CPRIT grant RR170079 (to B. Li), Circle of Friends Cancer Center Grant 2018 (to B. Li), and CPRIT grants RR150072 (to Y.-X. Fu). The authors thank Dr. James Brugarolas for helpful discussions during manuscript preparation. The authors acknowledge the TCGA research network for providing publicly available cancer genomics data that enabled this analysis.

Received October 4, 2019; revised November 26, 2019; accepted December 5, 2019; published first December 12, 2019.

References

1. Ahmadzadeh M, Johnson LA, Heemskerk B, Wunderlich JR, Dudley ME, White DE, et al. Tumor antigen-specific CD8 T cells infiltrating the tumor express high levels of PD-1 and are functionally impaired. Blood 2009;114:1537–44.
2. Coulie PG, Van den Eynde BJ, van der Bruggen P, Boon T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. Nat Rev Cancer 2014;14:135–46.
3. Gooden MJ, de Bock GH, Leffers N, Daemen T, Nijman HW. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. Br J Cancer 2011;105:93–103.
4. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature 2014;515:577–81.
5. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer 2012;12:252–64.
6. Tang H, Wang Y, Chlewicki LK, Zhang Y, Guo J, Liang W, et al. Facilitating T cell infiltration in tumor microenvironment overcomes resistance to PD-L1 blockade. Cancer Cell 2016;30:500.
7. Kalos M, June CH. Adoptive T cell transfer for cancer immunotherapy in the era of synthetic biology. Immunity 2013;39:49–60.
8. Rosenberg SA, Restifo NP, Yang JC, Morgan RA, Dudley ME. Adoptive cell transfer: a clinical path to effective cancer immunotherapy. Nat Rev Cancer 2008; 8:299–308.
9. Stronen E, Toebes M, Kelderman S, van Buuren MM, Yang W, van Rooij N, et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. Science 2016;352:1337–41.
10. Tran E, Turcotte S, Gros A, Robbins PF, Lu YC, Dudley ME, et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science 2014;344:641–5.
11. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446: 153–8.
12. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010;463:191–6.
13. Saeterdal I, Bjorheim J, Lislerud K, Gjertsen MK, Bukholm IK, Olsen OC, et al. Frameshift-mutation-derived peptides as tumor-specific antigens in inherited and spontaneous colorectal cancer. Proc Natl Acad Sci U S A 2001;98:13255–60.
14. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol 2017;18:1009–21.
15. Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev 2002;188: 22–32.
16. Simpson AJ, Caballero OL, Jungbluth A, Chen YT, Old LJ. Cancer/testis antigens, gametogenesis and cancer. Nat Rev Cancer 2005;5:615–25.
17. Doyle HA, Mamula MJ. Post-translational protein modifications in antigen recognition and autoimmunity. Trends Immunol 2001;22:443–9.
18. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med 2015;7:45.
19. Sahu A, Singhal U, Chinnaiyan AM. Long noncoding RNAs in cancer: from function to translation. Trends Cancer 2015;1:93–109.
20. Cascio S, Zhang L, Finn OJ. MUC1 protein expression in tumor cells regulates transcription of proinflammatory cytokines by forming a complex with nuclear factor-kappaB p65 and binding to cytokine promoters: importance of extracellular domain. J Biol Chem 2011;286:42248–56.
21. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. Science 1999;286:958–61.
22. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. Nature 2017;547:94–8.
23. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 2017;547:89–93.
24. DeWitt WS 3rd, Smith A, Schoch G, Hansen JA, Matsen FAt, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. eLife 2018;7:e38358.
25. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I, et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. eLife 2017;6:e22057.
26. Beausang JF, Wheeler AJ, Chan NH, Hanft VR, Dirbas FM, Jeffrey SS, et al. T cell receptor sequencing of early-stage breast cancer tumors identifies altered clonal structure of the T cell repertoire. Proc Natl Acad Sci U S A 2017;114:E10409–E17.
27. Robert L, Tsoi J, Wang X, Emerson R, Homet B, Chodon T, et al. CTLA4 blockade broadens the peripheral T-cell receptor repertoire. Clin Cancer Res 2014;20:2424–32.
28. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 2010;38:e178.
29. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic Acids Res 2015;43:D413–22.
30. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res 2018;46:D419–D27.
31. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet 2017;49:659–65.
32. Jenkins MK, Moon JJ. The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude. J Immunol 2012;188: 4135–40.
33. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–6.
34. Li B, Senbabaoglu Y, Peng W, Yang ML, Xu J, Li JZ. Genomic estimates of aneuploid content in glioblastoma multiforme and improved classification. Clin Cancer Res 2012;18:5595–605.
35. R Core Team. R: a language and environment for statistical computing; 2015. Vienna, Austria: R Core Team.
36. Krijthe JH. Rtsne: T-distributed stochastic neighbor embedding using a barnes-hut implementation. Available from: https://rdrr.io/cran/Rtsne.
37. Wallin JJ, Bendell JC, Funke R, Sznol M, Korski K, Jones S, et al. Atezolizumab in combination with bevacizumab enhances antigen-specific T-cell migration in metastatic renal cell carcinoma. Nat Commun 2016;7:12624.
38. Li B, Li T, Wang B, Dou R, Zhang J, Liu JS, et al. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data. Nat Genet 2017; 49:482–3.
39. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, 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 2014;24:1603–12.
40. Adams JJ, Narayanan S, Liu B, Birnbaum ME, Kruse AC, Bowerman NA, et al. T cell receptor signaling is limited by docking geometry to peptide-major histocompatibility complex. Immunity 2011;35:681–93.
41. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-MHC specificity of T cell recognition. Cell 2014;157: 1073–87.
42. Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, et al. Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. Cell 2018;172:549–63.
43. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol 2016; 17:174.
44. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 2005;102:15545–50.
45. Barnes MJ, Li CM, Xu Y, An J, Huang Y, Cyster JG. The lysophosphatidylserine receptor GPR174 constrains regulatory T cell development and function. J Exp Med 2015;212:1011–20.
46. Burotto M, Chiou VL, Lee JM, Kohn EC. The MAPK pathway across different malignancies: a new perspective. Cancer 2014;120:3446–56.
47. Ebert PJR, Cheung J, Yang Y, McNamara E, Hong R, Moskalenko M, et al. MAP kinase inhibition promotes T cell and anti-tumor activity in combination with PD-L1 checkpoint blockade. Immunity 2016;44:609–21.
48. Cascone T, McKenzie JA, Mbofung RM, Punt S, Wang Z, Xu C, et al. Increased tumor glycolysis characterizes immune resistance to adoptive T cell therapy. Cell Metab 2018;27:977–87.
49. Chaker M, Minden A, Chen S, Weiss RH, Chini EN, Mahipal A, et al. Rho GTPase effectors and NAD metabolism in cancer immune suppression. Expert Opin Ther Targets 2018;22:9–17.

50. Renner K, Singer K, Koehl GE, Geissler EK, Peter K, Siska PJ, et al. Metabolic hallmarks of tumor and immune cells in the tumor microenvironment. Front Immunol 2017;8:248.
51. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 2018;174:1293–308.
52. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res 2008;9:2579–605.
53. Savas P, Virassamy B, Ye C, Salim A, Mintoff CP, Caramia F, et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat Med 2018;24:986–93.
54. Kumar BV, Ma W, Miron M, Granot T, Guyer RS, Carpenter DJ, et al. Human tissue-resident memory T cells are defined by core transcriptional and functional signatures in lymphoid and mucosal sites. Cell Rep 2017;20:2921–34.
55. Mackay LK, Minnich M, Kragten NA, Liao Y, Nota B, Seillet C, et al. Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. Science 2016;352:459–63.
56. Herndler-Brandstetter D, Ishigame H, Shinnakasu R, Plajer V, Stecher C, Zhao J, et al. KLRG1(+) Effector CD8(+) T cells lose KLRG1, differentiate into all memory T cell lineages, and convey enhanced protective immunity. Immunity 2018;48:716–29.
57. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature 2014;515:572–6.
58. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobanking 2015;13:311–9.
59. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 2015;33:1152–8.
60. Hogquist KA, Baldwin TA, Jameson SC. Central tolerance: learning self-control in the thymus. Nat Rev Immunol 2005;5:772–82.
61. Kanellopoulos-Langevin C, Caucheteux SM, Verbeke P, Ojcius DM. Tolerance of the fetus by the maternal immune system: role of inflammatory mediators at the feto-maternal interface. Reprod Biol Endocrinol 2003;1:121.
62. Iheagwara UK, Beatty PL, Van PT, Ross TM, Minden JS, Finn OJ. Influenza virus infection elicits protective antibodies and T cells specific for host cell antigens also expressed as tumor-associated antigens: a new view of cancer immuno-surveillance. Cancer Immunol Res 2014;2:263–73.
63. Valkenburg SA, Josephs TM, Clemens EB, Grant EJ, Nguyen TH, Wang GC, et al. Molecular basis for universal HLA-A*0201-restricted CD8+ T-cell immunity against influenza viruses. PNAS 2016;113:4440–5.
64. Yang X, Chen G, Weng NP, Mariuzza RA. Structural basis for clonal diversity of the human T-cell response to a dominant influenza virus epitope. J Biol Chem 2017;292:18618–27.
65. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife 2015;4:e08890.