nature genetics

Landscape of tumor-infiltrating T cell repertoire of human cancers

Bo Li[1,2,11], Taiwen Li[1,3,11], Jean-Christophe Pignon[4], Binbin Wang[5], Jinzeng Wang[5], Sachet A Shukla[6], Ruoxu Dou[7], Qianming Chen[3], F Stephen Hodi[8], Toni K Choueiri[9], Catherine Wu[6], Nir Hacohen[10], Sabina Signoretti[4], Jun S Liu[2] & X Shirley Liu[1,2]

We developed a computational method to infer the complementarity-determining region 3 (CDR3) sequences of tumor-infiltrating T cells in 9,142 RNA-seq samples across 29 cancer types. We identified over 600,000 CDR3 sequences, including 15% that were full length. CDR3 sequence length distribution and amino acid conservation, as well as variable gene usage, for infiltrating T cells in many tumors, except in brain and kidney cancers, resembled those for peripheral blood cells from healthy donors. We observed a strong association between T cell diversity and tumor mutation load, and we predicted SPAG5 and TSSK6 as putative immunogenic cancer/testis antigens in multiple cancers. Finally, we identified three potential immunogenic somatic mutations on the basis of their co-occurrence with CDR3 sequences. One of them, a *PRAMEF4* mutation encoding p.Phe300Val, was predicted to result in peptide binding strongly to both MHC class I and class II molecules, with matched HLA types in its carriers. Our analyses have the potential to simultaneously identify immunogenic neoantigens and tumor-reactive T cell clonotypes.

The T cell receptor (TCR) consists of a heterodimer of two chains (αβ or γδ), both of which are products of V(D)J recombination[1]. This somatic rearrangement only occurs in the T cell genome and produces an extremely diverse repertoire of TCRs. The most variable region in the TCR is CDR3, which has a critical role in antigen recognition[2].

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Statistics, Harvard University, Boston, Massachusetts, USA. [3]State Key Laboratory of Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, China. [4]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA. [5]School of Life Science and Technology, Tongji University, China, Shanghai, China. [6]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [7]Department of Colorectal Surgery, Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. [8]Center for ImmunoOncology, Harvard Medical School, Boston, Massachusetts, USA. [9]Kidney Cancer Center, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [10]Center for Cancer Immunotherapy, Massachusetts General Hospital, Boston, Massachusetts, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to J.S.L. (jliu@stat.harvard.edu) or X.S.L. (xsliu@jimmy.harvard.edu).

Received 31 December 2015; accepted 4 May 2016; published online 30 May 2016; doi:10.1038/ng.3581

The lower limit for the number of distinct TCRs in the peripheral blood of a healthy individual is around 1.1 million[3], and the theoretical diversity for αβ T cells, the most abundant T cell type in humans, includes up to 1×10^{16} TCRs[4]. This large repertoire of T cells with structurally divergent TCRs is required to recognize cells expressing foreign or mutated proteins, including neoantigens from cancer cells. Therefore, characterizing the repertoire of tumor-infiltrating T cells can help identify tumor-reactive T cell clones and facilitate the clinical practice of cancer immunotherapies.

The current commonly used strategy for characterizing CDR3 sequences is TCR profiling, which amplifies cDNA or genomic DNA from the TCR β-chain CDR3 (β-CDR3) locus using predesigned PCR primers, followed by deep sequencing. Recent developments of cancer immunotherapies[5–7] have seen TCR sequencing applied to monitor differences in the T cell repertoire before and after therapy in humans or animal models[8–11]. Although these studies found exciting mechanisms of tumor immunity and the pharmacology of drugs leading to checkpoint blockade, they were limited by small sample size and thus had low power to detect important features shared among individuals. Efforts have been made to design methods to study the repertoires of T and B cells in non-solid or solid tumors using unselected RNA-seq data[12,13], which could potentially scale up to large cohorts. However, these studies adopted computational methods not specifically designed for unselected RNA-seq data[14–16], resulting in poor detection of CDR3 sequences and limited power in downstream characterization of the tumor-infiltrating T cell repertoires of the cohorts.

In this study, we developed a new computational method for *de novo* assembly of sequences from CDR3 regions using paired-end RNA-seq data and applied it to 9,142 samples from The Cancer Genome Atlas (TCGA). In comparison to a previous RNA-seq-based analysis[13], we assembled an order of magnitude more distinct CDR3 sequences, which gave us enough power to perform deeper analyses on the TCR repertoire of the tumor microenvironment. We observed interesting interactions between tumors and the host immune system and identify potential therapeutic targets that might be useful for multiple immunotherapies.

RESULTS

De novo assembly of CDR3 sequences and method validation

We developed a method to assemble *de novo* the CDR3 sequences generated from TCR locus transcripts in paired-end RNA-seq data

(Online Methods and **Supplementary Fig. 1**). In brief, this method first maps reads to the human genome and searches for read pairs with one mate properly mapped to a TCR gene and the other mate unmappable to the genome, potentially owing to V(D)J recombination. It then initiates pairwise comparisons between the unmapped reads and constructs a read-overlap matrix, represented by an undirected graph, with each node corresponding to a read and edges indicating partial sequence overlap between two connected reads. This graph is further divided into disjoint cliques to represent potentially different CDR3 sequences. Finally, the method assembles all the reads in each clique to obtain contigs of DNA sequence and annotates them with information such as amino acid sequence and associated variable (V) and joining (J) genes. Contigs not annotated as CDR3 regions were discarded to reduce false positive calls (Online Methods). Counts of the reads and contigs kept at each step of the method for one example are summarized in **Supplementary Figure 2**.

To validate the above approach, we first selected three kidney renal clear cell carcinoma (KIRC) samples from TCGA with available RNA-seq data, extracted genomic DNA from the corresponding formalin-fixed and paraffin-embedded (FFPE) tumors and sent the DNA for TCRβ sequencing (immunoSEQ). Although the number of CDR3 sequences assembled from RNA-seq data was much smaller than that from immunoSEQ analysis, over 60% of the CDR3 sequences identified from RNA-seq data were also observed among the immunoSEQ CDR3 sequences, which partially validates the accuracy of our method. It is worth noting that, because of DNA fragmentation in FFPE samples, only a subpopulation (~25–50%) of the infiltrating T cells can be recovered; thus, it is not surprising that a fraction of our assemblies from RNA-seq data were not present in the immunoSEQ results. Also, as expected, the CDR3 assemblies from RNA-seq data were enriched for those from abundant T cell clones, with over 50% of the most abundant clones (in the 99.9% quantile) recovered (Online Methods and **Supplementary Fig. 3**).

Because immunoSEQ analysis using FFPE samples cannot retrieve the complete repertoire of CDR3 sequences for infiltrating T cells, we conducted *in silico* simulations to systematically evaluate the performance of our method. To this end, we generated pseudo RNA-seq data for tumor samples by *in silico* mixing of TCR transcript reads from a deeply sequenced immunoSEQ sample[3] and RNA-seq reads from a TCR-negative cancer cell line (K562) (Online Methods and **Supplementary Fig. 4**). Comparison of the results from our method with those from this procedure demonstrated that our method achieves high accuracy in CDR3 sequence assembly over a large range of T cell infiltration levels (**Supplementary Fig. 5a**). The results also confirmed the above observation that the assembled CDR3 sequences were enriched for sequences from T cell clonotypes present at high frequencies (**Supplementary Fig. 5b**,**c**).

In both the immunoSEQ validation and *in silico* mixing simulation, our method assembled a small fraction (0.5–5%) of the total CDR3 repertoire (**Supplementary Figs. 3** and **5**). This is because the actual coverage of the TCR region in an RNA-seq sample is estimated to be as low as 0.04 (Online Methods). We introduced additional simulations to investigate how the CDR3 assembly rate changes with sequencing depth (Online Methods and **Supplementary Fig. 4**). Surprisingly, we found that, at coverage of 1 (library size = 5 billion reads), our method achieved 33% recall of the simulated 'true' CDR3 transcripts with high precision (97.2%), whereas for a competing method (iSSAKE[16]) the recall was only 0.7% with precision of 7.1% (**Supplementary Fig. 6**). The above results indicate that our method is a highly sensitive and accurate CDR3 assembler for tumor RNA-seq data. It outperforms a competing method by at least an order of magnitude, making it statistically powerful in analyzing the immune repertoires of large-scale RNA-seq sample cohorts.

Distribution of TCR gene usage and T cell type abundance

We applied the CDR3 assembly method to study 9,142 samples across 29 cancer types from TCGA; the resulting CDR3 sequences are available in the **Supplementary Data**. We first used mapped reads to estimate the usage of different TCRα variable (TRAV) and TCRβ variable (TRBV) genes across all tumor samples. The three most abundant transcripts for each class corresponded to 30, 13-1 and 12-2 for TRAV genes and 20-1, 5-1 and 6-5 for TRBV genes (**Fig. 1a**,**b**). Although there are few studies on TRAV gene usage to validate our estimates, our observations for TRBV genes are consistent with previous reports[17,18] examining peripheral blood from healthy donors[16,17]. This result suggests that TRBV gene usage in tumor-infiltrating T cells does not deviate substantially from that in peripheral blood. For most cancer types, the distribution of TRAV or TRBV gene usage among the tumor samples was similar, except in brain and kidney cancers (**Fig. 1c**,**d**). Brain cancer samples displayed very different patterns of both TRAV and TRBV gene usage, whereas kidney cancer samples were only different in TRAV gene usage, as compared to the majority of TCGA tumors (**Fig. 1a**,**b** and **Supplementary Fig. 7**). These differences in gene usage might be due to potentially different immune regulation in brain cancer and expression of endogenous retrovirus in kidney cancer[19].

We next investigated the CDR3 assemblies. Here CDR3 regions are defined as encompassing all amino acids between the last cysteine encoded by the V gene and the phenylalanine in the FGXG motif encoded by the J gene, as was previously described[17]. In total, we identified 683,418 CDR3 sequences, including 650,496 from αβ T cells and 32,922 from γδ T cells. Of all the CDR3 assemblies, the vast majority (95.8%) had read counts of less than 10, with a median of 1.7 (**Supplementary Fig. 8**). Of these assemblies, 77,060 β-CDR3 assemblies and 1,060 TCR δ-chain CDR3 (δ-CDR3) assemblies were complete sequences harboring the conserved N-terminal four amino acids and C-terminal phenylalanine. On the basis of sequence count, γδ T cells accounted for ~4.8% of the total T cell population, consistent with previous observations[20]. However, the γδ T cell fraction varied among cancer types (**Fig. 1e**), potentially owing to different neoantigens presented on different tumor cells.

Features of β– and δ-chain CDR3 sequences

The TCR β and δ chains have undergone V(D)J recombination and are responsible for most antigen recognition. β-CDR3 sequences had lengths ranging from 6 to 31 amino acids, with a median of 14 amino acids (**Fig. 2a**). The sequence pattern[21] for the most frequent 14-amino-acid β-CDR3 sequences (**Fig. 2b**) was very similar to that from the peripheral blood of healthy donors determined by TCR sequencing[17], with minor differences in the first 4 residues. These differences are potentially due to reduced TRBV20-1 abundance in our data, an observation consistent with previous studies[18,22].

δ-CDR3 sequences also had lengths ranging from 6 to 31 amino acids, although they had a longer median of 17 amino acids (**Fig. 2c**). Besides having a larger median length, δ-CDR3 sequences also had greater variation in length (s.d. = 4.6) than β-CDR3 sequences (s.d. = 1.9). The more constrained distribution of β-CDR3 lengths potentially reflects the functional requirement for the TCR β chain to contact the peptide major histocompatibility complex (pMHC), which is not a requirement for the δ chain in γδ T cells. These observations agree with previous reports[20,23], supporting the validity of our CDR3 calls. In addition, the β-CDR3 and δ-CDR3 sequences

showed no strong over-representation of specific amino acids except for the small prevalence of glycine residues in the middle of the sequence logo (**Fig. 2d**).

Identification of public and private β-CDR3 sequences

Despite the extreme amino acid sequence diversity, 4,252 of the complete β-CDR3 sequences appeared in more than one tumor (**Fig. 3a**),

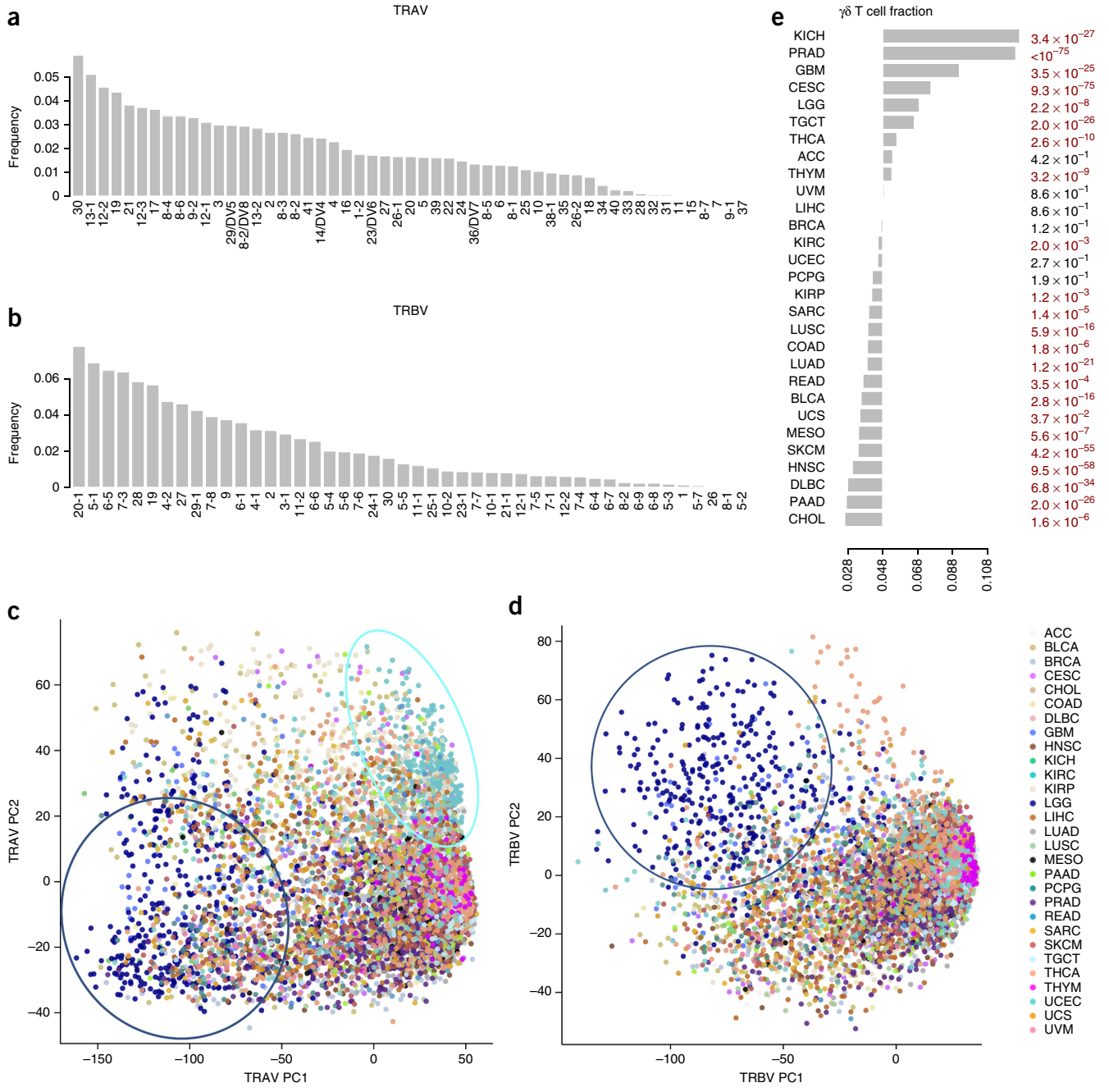


Figure 1 Distribution of αβ T cell variable gene usage and γδ T cell abundance across multiple cancer types. (**a**,**b**) Frequencies of TRAV (**a**) and TRBV (**b**) gene usage shown in order of decreasing frequency. IMGT functional genes were selected for the display. (**c**,**d**) Principal-component analysis (PCA) of TRAV (**c**) and TRBV (**d**) gene usage across different cancer types. For TRAV genes, PC1 was driven by the difference between brain cancer (LGG; dark blue) and other tumors, whereas PC2 was driven by the difference between kidney cancer (KIRC; cyan) and other cancers. (**e**) γδ T cell fraction (*x* axis) in multiple cancer types shown in order of decreasing fraction. The mean γδ T cell fraction across all samples was 4.8%. For each cancer, we used the binomial test with an expected probability of 0.048 to calculate the statistical significance of deviation from the mean. We applied Benjamini–Hochberg adjusted *P* values to determine the FDR. The numbers on the right margin of the plot are *q* values; red color indicates significance at $q < 0.05$. ACC, adrenocortical carcinoma; BLCA, bladder carcinoma; BRCA, breast carcinoma; CESC, cervical squamous cell carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, diffuse large B cell lymphoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, lower-grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO. mesothelioma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; TGCT, testicular germ cell tumor; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carsinosarcoma; UVM, uveal melanoma.

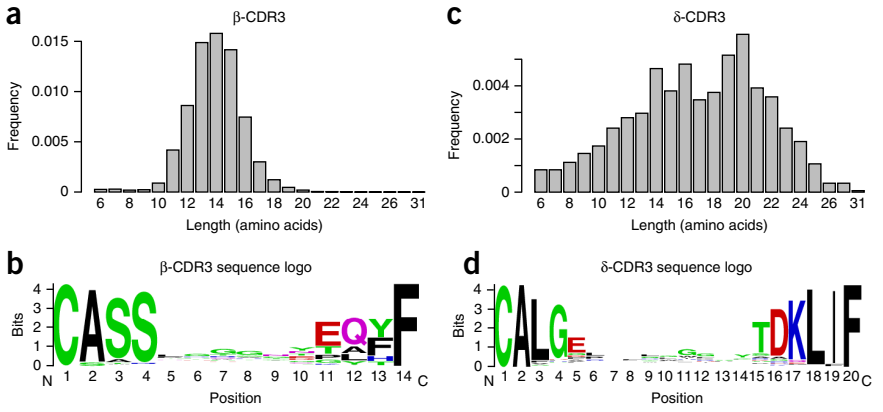


Figure 2 Length and amino acid conservation of β- and δ-chain CDR3 sequences in tumor-infiltrating T cells. (**a**–**d**) The length distribution for complete CDR3 calls is estimated using a histogram for the β chain (**a**) and δ chain (**c**). We selected 14-amino-acid β-CDR3 (**b**) and 20-amino-acid δ-CDR3 (**d**) sequences for WebLogo analysis. The *y* axis in the sequence logo plots shows the conservation score. For a given position, the height of a letter reflects the relative frequency of that amino acid.

and the number of individuals sharing these sequences could be as high as 65 (**Fig. 3b**). A large fraction of the shared, or public, CDR3 sequences are potentially products of convergent recombination in the thymus[24]. We compared our β-CDR3 calls to the peripheral blood repertoire of healthy individuals as determined by deep TCR sequencing[3], and 2,059 of the shared β-CDR3 sequences in the TCGA cohort were also present in this data set (**Fig. 3a**). This result suggests that a large fraction of the shared sequences might not be related to tumor antigens but might be derived from public T cells with a potential role in responses to common antigens, such as those present in persistent viral infections[24]. Interestingly, 10,249 β-CDR3 sequences private to individual TCGA tumors also overlapped with the peripheral blood CDR3 repertoire. These sequences are likely also shared but were missed in other TCGA tumor(s) because of the limited power for identifying T cell clonotypes of low abundance from RNA-seq data. Therefore, we merged these 10,249 sequences with the previous 4,252 sequences to generate the final set of public β-CDR3 sequences. It is worth noting that our final set of private sequences in TCGA data may still contain a substantial number of truly public β-CDR3 sequences.

A previous study reported that the β-CDR3 sequences of private T cells are significantly longer than those of public T cells in peripheral blood[3], and we observed the same for tumor-infiltrating T cells (**Fig. 3c**). As β-CDR3 sequences contained a highly conserved four-amino-acid sequence at their N terminus and phenylalanine at their C terminus (**Fig. 2b**), we defined the 'CDR3 motif' as the amino acid sequence in between these conserved regions of the complete β-CDR3 sequence. Interestingly, the middle three amino acids of private β-CDR3 motifs contained a significantly greater fraction of hydrophobic residues than those of public β-CDR3 motifs (**Fig. 3d**). A recent study reported that hydrophobicity is a hallmark of immunogenic neoepitopes[25], and, hence, our results suggest that private β-CDR3 sequences might have greater potential for tumor antigen recognition.

Association of T cell diversity with neoantigen load

The clonotype diversity of the T cell repertoire is an important property of the immune system and is closely related to the capacity of T cells to recognize antigens. As each T cell clone possesses a unique TCR, CDR3 sequences are often used as proxies to represent clonotype diversity. In our data, the number of unique CDR3 calls in each tumor was linearly correlated with the total number of TCR reads (**Fig. 4a**), an expected observation because tumors with higher levels of T cell infiltrates have more TCR reads, resulting in the assembly of more CDR3 sequences. We therefore used the number of unique CDR3 calls in each sample normalized by the total read count in the TCR region, which we call clonotypes per thousand (kilo) reads (CPK), as a measure of clonotype diversity (Online Methods). In kidney, lung and pancreatic cancers, female patients had significantly higher CPK values than male patients ($P = 0.0015$, 0.0075 and 0.0016, respectively, Wilcoxon rank test), consistent with the long-standing knowledge of

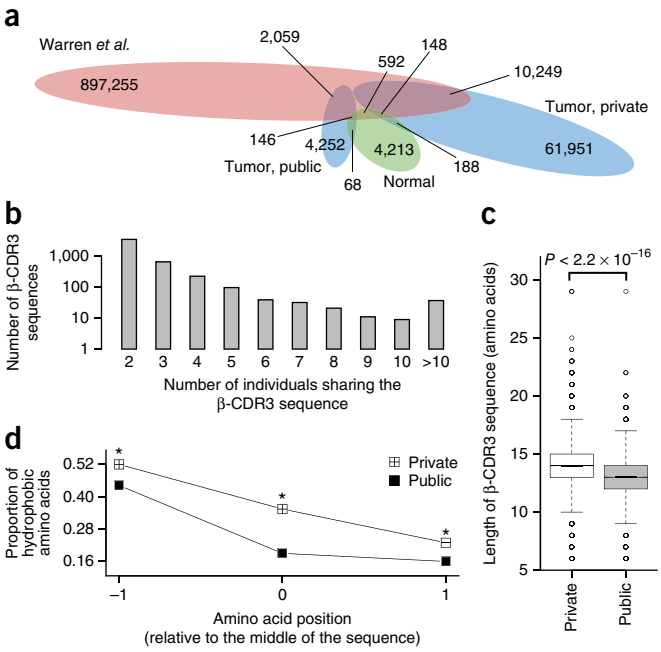


Figure 3 Public and private β-CDR3 amino acid sequences have different lengths and proportions of hydrophobic residues. (**a**) Sharing of β-CDR3 sequences among TCGA tumor, TCGA normal sample and peripheral blood[3] repertoires displayed in a Venn diagram. The numbers inside the ellipses are the overall calls for each category. The numbers outside the ellipses that are connected to colored regions are the counts of calls overlapping between categories. (**b**) Distribution of the frequencies of public β-CDR3 sequences. Sequence sharing was determined using only TCGA data, not including the 10,249 sequences shared with the peripheral blood repertoire. (**c**) Comparison of the lengths of private and public β-CDR3 sequences. The *P* value was calculated using the Wilcoxon test. Each box includes data between the 25th and 75th percentiles, with the horizontal line representing the median; the upper whisker is the third quartile + 1.5 times the IQR and the lower whisker is the first quartile − 1.5 times the IQR, where IQR is interquartile range. There were 51,583 and 14,443 sequences in the private and public groups, respectively. (**d**) Hydrophobicity analysis of the middle three amino acids in private and public β-CDR3 sequences. For each position, a binomial test was applied to estimate the significance of the difference in the fraction of hydrophobic amino acids between the groups, using the fraction in the public group as the expected probability. The difference in hydrophobicity for all three positions was significant at FDR = 0.05 (indicated by an asterisk).

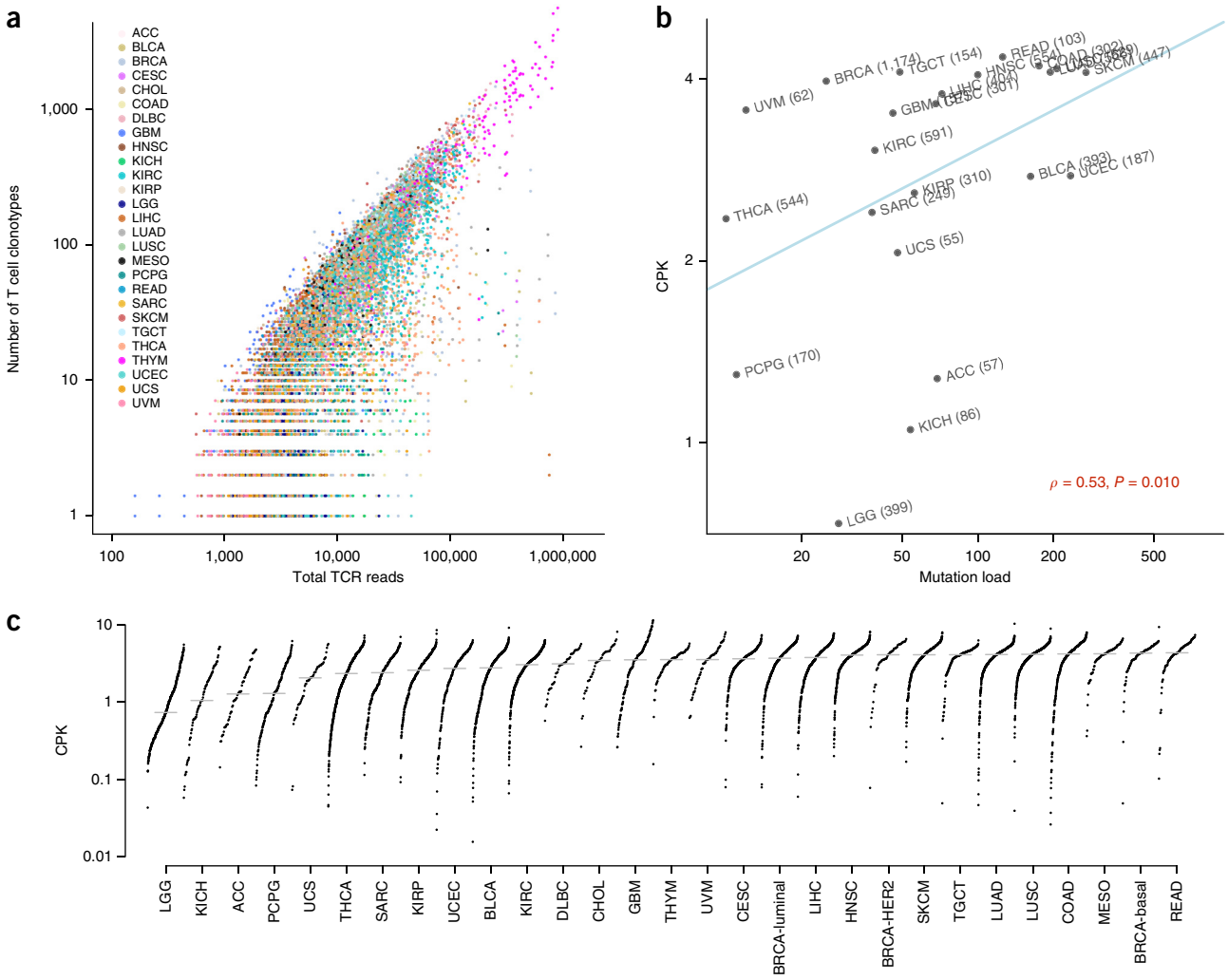


Figure 4 The diversity of T cell clonotypes positively associates with cancer somatic mutation load. (**a**) Scatterplot of the number of CDR3 calls in each sample plotted against the total number of reads extracted from the three TCR regions. Prostate and pancreatic cancers were excluded because of high expression of non-TCR genes from these regions (Online Methods). (**b**) CPK is positively correlated with tumor mutation load. Median CPK values are plotted against median somatic mutation load for each cancer type in the scatterplot. Cancers with <50 samples were excluded. Significance was estimated using Spearman's correlation test. (**c**) Distribution of CPK values across all cancer types. PAM50 subtypes of breast cancer[39] are displayed to show between-tumor heterogeneity in this disease. Horizontal bars represent the median CPK for each cancer type.

elevated immune responses in females[26] (**Supplementary Table 1**). In addition, the expression level of granzyme A, an indicator of immune-mediated cytolysis[18,25], was also positively correlated with CPK, even after correcting for tumor purity[19,27] (**Supplementary Fig. 9**). These observations support the validity of using CPK as a measure of immune responses.

We next calculated CPK for each tumor sample and observed a strong positive association between the CPK value and the load of non-synonymous somatic muta tions (**Fig. 4b** and **Supplementary Fig. 10**). When ranking the cancer types by median CPK, we found that breast cancer showed surprisingly high levels of between-tumor heterogeneity, where the CPK of the basal breast cancer subtype was 1.2-fold that of the luminal subtype. Besides basal breast cancer, testicular cancer (TGCT) also had unusually high CPK, which might be related to the high level of alternative splicing during spermatogenesis[28]. The remaining cancers with the highest T cell clonotype diversity included colorectal cancers, non-small-cell lung carcinomas, mesothelioma and melanoma (**Fig. 4c**). These cancers are known to be associated with external stimuli, such as microbiota, smoking, and carcinogen and UV exposure, respectively. There are at least two possible explanations for our observations in these cancer types: (i) tumors with a higher mutation load present more neoantigens to the immune system, which recruit antigen-specific infiltrating T cells, and (ii) external stimuli such as UV light or carcinogens directly interact with the immune system to increase the diversity of the T cell repertoire. If the second explanation is valid, we expect a higher fraction of public β-CDR3 sequences in these cancer types, due to the presence of public T cells in response to common stimuli. Among the above cancers, this was true only in melanoma (**Supplementary Fig. 11**), suggesting that the diversity of infiltrating T cells in most cancers might be regulated through tumor-specific somatic mutations.

Cancer/testis (CT) antigens are derived from a family of genes whose expression is normally restricted to germ cells but that can also be expressed in tumors as a result of epigenetic instability. CT antigens are not subject to thymus tolerance like antigens derived from genes expressed in other tissues and can be recognized as foreign antigens by the immune system. Efforts have been made to explore the possibility of using CT antigens as cancer vaccine targets[29,30], and clinical trials

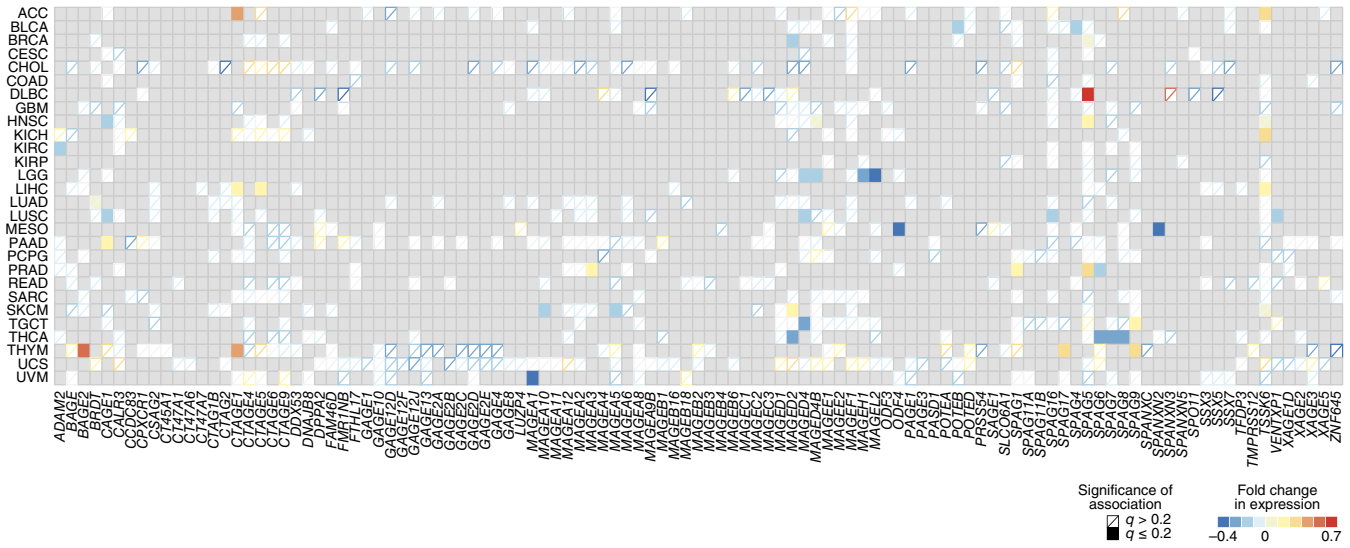


Figure 5 Association of T cell diversity with expression of cancer/testis antigens identifies SPAG5 and TSSK6 as vaccine targets. The association between CPK and CT antigen expression was evaluated using partial Spearman correlation corrected for tumor purity. Solid cells highlight genes with an association between CPK and differential expression in tumors significant at FDR ≤ 0.2. Gray cells correspond to genes whose expression was not altered in tumor cells relative to normal samples, as suggested by correlative analysis with tumor purity.

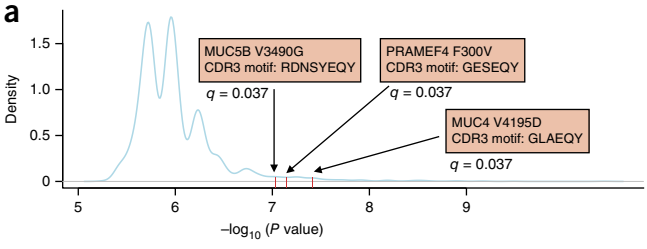
have been conducted using a number of CT antigens[31]. We examined whether the expression of CT antigens is associated with infiltrating T cell diversity. Of the 109 known CT antigens[19,32], SPAG5 and TSSK6 had expression levels that positively correlated with CPK in multiple cancers (**Fig. 5**). We further analyzed all nine-amino-acid peptides in the SPAG5 and TSSK6 protein sequences for binding affinity to MHC class I molecules using NetMHC4.0 (ref. 33) and identified 25 peptides for SPAG5 and 7 peptides for TSSK6 with strong (rank <0.5%) binding to common human leukocyte antigen (HLA) alleles (**Supplementary Fig. 12**). Together, this evidence supports SPAG5 and TSSK6 as potential vaccine targets in multiple cancer types.

Joint prediction of neoantigens and tumor-reactive T cells

According to the mechanism of immunoediting[34], tumors with the same immunogenic somatic mutation might harbor tumor-reactive T cells with similar antigen-recognizing TCR domains. To explore the recurrent patterns in the CDR3 sequences, we studied the CDR3 motifs as defined above. From the complete β-CDR3 sequences, we identified a total of 64,824 unique motifs. For each motif, we searched all 683,418 CDR3 calls for sequences containing the motif and documented the corresponding individuals. Extremely short motifs are not informative, as their recurrence can occur by chance. Additionally, highly recurrent motifs may come from public T cells, whose evaluation was not the intent of this analysis. Therefore, we focused on β-CDR3 motifs with recurrence in 5 to 20 tumors and more than five amino acids in length, constituting 5,347 high-quality motifs. Next, we obtained the somatic mutation profiles from exome sequencing data for each tumor with these motifs. We filtered out 5′UTR, 3′UTR, nonsense and read-through mutations because they do not result in altered peptide sequences or the generation of potential neoantigens in tumors. The remaining 2,353 nonsynonymous mutations occurring in more than three tumors were kept for downstream analyses.

We then examined the co-occurrence of β-CDR3 motifs and nonsynonymous mutations in patients with cancer, using Fisher's exact test to estimate statistical significance. We applied a heuristic method to find the most promising β-CDR3 motif–mutation pairs and used permutation tests to correct for the false discovery (FDR) rate (Online Methods). On the basis of over 1.5 million null permutation tests and a stringent selection criterion, we identified the top three pairs that were statistically significant (FDR = 0.05), involving mutations in *MUC4*, *PRAMEF4* and *MUC5B* (**Fig. 6a**). Examining the nine-amino-acid peptides affected by these mutations for binding to MHC class I molecules with NetMHC4.0 (ref. 35), we found that all three mutations corresponded to at least one predicted binding peptide (**Supplementary Table 2**). We compared the mutated peptides with their wild-type counterparts and excluded mutated peptides with lower binding affinity than the corresponding wild-type peptide. The remaining peptides were all derived from the Phe300Val mutant of PRAMEF4 (**Fig. 6b**). Interestingly, Phe300Val PRAMEF4 was also

a



b

Gene	Peptide	MHC I allele	Rank	Predicted binding
PRAMEF4	CLKTSLK**V**L	B*08:01	0.4%	Strong
PRAMEF4	KTSLK**V**LTI	B*58:01	0.4%	Strong
PRAMEF4	K**V**LTITNCV	A*02:01	1%	Weak
PRAMEF4	K**V**LTITNCV	A*30:01	0.7%	Weak

Figure 6 Nonsynonymous mutations co-occur with CDR3 motifs. (**a**) Three pairs of nonsynonymous substitutions and CDR3 motifs co-occurred more often than expected by random chance, with statistical significance (FDR ≤ 0.05) based on permutation tests (Online Methods). (**b**) Predictions of MHC I binding affinity were made for all possible nine-amino-acid peptides derived from the three nonsynonymous changes in **a** using NetMHC4.0; mutated peptides binding more strongly than the corresponding wild-type peptides are shown. Mutated amino acids are shown in bold.

predicted by NetMHC-II2.2 (refs. 33,36) to produce peptides with high-affinity binding to MHC class II molecules (**Supplementary Fig. 13**). *PRAMEF4* mutation encoding p.Phe300Val co-occurred with the β-CDR3 motif GESEQY in three patients, TCGA-4K-AA1I, TCGA-2G-AAGE and TCGA-2G-AAKO, all of whom had testicular germ cell tumors. We did not find this β-CDR3 motif in two other tumors carrying a mutation encoding p.Phe300Val; one possible reason for this is that the β-CDR3 motif was present in these tumors but was not identified from the RNA-seq data. *PRAMEF4* is expressed in cholangiocarcinoma and liver, ovarian, endometrial and testicular cancers, but its expression is almost absent in other cancers and normal tissues (**Supplementary Fig. 14**). The three individuals with a *PRAMEF4* mutation encoding p.Phe300Val all had *PRAMEF4* expression in the tumor, and mutant PRAMEF4 peptides are therefore produced in these tumors. We next annotated the HLA types for these three individuals using POLYSOLVER[27] (**Supplementary Table 3**). One had an HLA-A*30:01 allele, the exact allele predicted to bind K**V**LTITNCV peptide. Two of the individuals, TCGA-2G-AAGE and TCGA-2G-AAKO, also had an HLA-B*08:01 allele, which is predicted to bind CLKTSLK**V**L peptide. Therefore, all three patients harbored at least one HLA allele binding the mutant peptides. These results support the notion that the *PRAMEF4* mutation encoding p.Phe300Val is a potential immunogenic mutation in testicular cancer. In addition, our analysis suggests that, if the p.Phe300Val substitution is truly immunogenic, the corresponding tumor-reactive T cell clonotypes are likely to carry the GESEQY motif in their CDR3 sequences.

DISCUSSION

Understanding the crosstalk between cancer antigens and host adaptive immunity is critical to finding therapeutic targets and developing effective immunotherapies. Improved characterization of the tumor-infiltrating T cell repertoire is highly desirable yet has been limited to small-scale samples owing to technical and cost barriers. In this study, we developed a computational method to extract the TCR sequence from unselected tumor RNA-seq data and applied it to over 9,000 TCGA samples across 29 cancer types. To our best knowledge, this work is among the first to analyze the infiltrating T cell repertoire in a large cancer cohort. In comparison to a similar work relying on reads covering the complete CDR3 region[13], our method assembled an order of magnitude more CDR3 sequences, leading to stronger associations and improved statistical power.

Our observations on TRAV and TRBV gene usage, CDR3 sequence length and amino acid conservation, γδ T cell fraction, and features of public and private T cells were similar to those previously reported for the peripheral blood repertoire[3,17]. These results suggest that the population of infiltrating T cells maintains a large fraction of public clonotypes, which are also present in the peripheral repertoires of healthy donors. In comparison to CDR3 regions from private T cells, the CDR3 regions of public clonotypes were shorter and less likely to bind neoepitopes, according to the hydrophobicity analyses[25]. Future efforts are needed to elucidate the potential functional impact of public T cells in the tumor microenvironment.

According to our results, the presence of cancer antigens, including ones derived from somatic mutations and CT genes, might increase the diversity of the infiltrating T cell repertoire. Specifically, we identified SPAG5 and TSSK6 as candidate cancer vaccine targets on the basis of their association with CPK, a metric for T cell clonotype diversity. It must be emphasized that our CDR3 calls only represent prevalent clonotypes in infiltrating T cells because of limited detection power from RNA-seq data. Therefore, CPK is a simplified measure of diversity for abundant T cell clonotypes, which is potentially the reason that CPK was not associated with patient survival (**Supplementary Table 1**).

Our analysis identified three instances of strong co-occurrence of recurrent tumor mutations and CDR3 sequence motifs and provided HLA typing evidence supporting the idea that at least one of the mutations, a *PRAMEF4* mutation encoding p.Phe300Val, is a putative immunogenic mutation. Our analysis also identified the corresponding antigen-recognizing CDR3 motifs. Unfortunately, we could not conduct TCR sequencing on the remaining mutation-positive but CDR3-motif-negative samples or other experimental validation at this point to further substantiate the association. Our computational approach is potentially useful for the development of cancer vaccine and adoptive T cell[5] as well as chimeric antigen receptor T cell therapies[37,38]. One possible reason that we were not able to find additional significant pairs is our limited CDR3 detection power. Another possible reason is that TCRs with different CDR3 sequences might be able to bind the same neoepitope, and the same somatic mutation may be recognized by multiple CDR3 motifs. Therefore, future efforts might be made to group different CDR3 sequences with similar biochemical properties and match the groups to somatic mutations to identify additional immunogenic somatic mutations.

In this study, we demonstrated the feasibility of using unselected RNA-seq data to characterize the tumor-infiltrating T cell repertoire. Although the scale and power of our analysis were sometimes still limited by low coverage and insufficient sample size, we were able to observe interesting associations between the T cell repertoire and tumor clinical and molecular features in the TCGA cohort. With the rapid decrease in sequencing cost and increase in tumor profiling efforts, we anticipate further analyses of tumor–immune interactions on more high-quality RNA-seq data to yield better biological insights in the near future.

URLs. Cancer Genomics Hub, https://cghub.ucsc.edu/; TCGA data portal, https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp; GDAC Firehose, https://gdac.broadinstitute.org/; simNGS, http://www.ebi.ac.uk/goldman-srv/simNGS/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. ImmunoSEQ data generated in this study are accessible through Adaptive Biotechnologies at https://adaptivebiotech.com/pub/Liu-2016-NatGenetics.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank G. Freeman for helpful discussion during manuscript preparation. We also acknowledge the following funding sources for supporting our work: NCI grant 1U01 CA180980, National Natural Science Foundation of China grant 31329003 and a Chinese Scholarship Council Fellowship. This work was supported in part by NIH/NCI DF/HCC Kidney Cancer SPORE P50 CA101942 to S.S. and T.K.C.

AUTHOR CONTRIBUTIONS

B.L. conceived this project, developed the CDR3 calling method, processed the data sets and performed statistical analysis. T.L. performed statistical analysis, generated a subset of the figures and helped write the manuscript. B.W., J.W. and R.D. helped with analysis of CDR3 sequences. S.A.S. performed analyses using POLYSOLVER. Q.C. helped analyze the data. J.-C.P., S.S. and T.K.C. conducted experimental validation. F.S.H., C.W. and N.H. conceived some of the analyses and contributed to the manuscript. X.S.L. and J.S.L. supervised the whole study and wrote the manuscript with B.L.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Alt, F.W. *et al.* VDJ recombination. *Immunol. Today* **13**, 306–314 (1992).
2. Davis, M.M. & Bjorkman, P.J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
3. Warren, R.L. *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* **21**, 790–797 (2011).
4. Robins, H.S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).
5. Rosenberg, S.A., Restifo, N.P., Yang, J.C., Morgan, R.A. & Dudley, M.E. Adoptive cell transfer: a clinical path to effective cancer immunotherapy. *Nat. Rev. Cancer* **8**, 299–308 (2008).
6. Sharma, P., Wagner, K., Wolchok, J.D. & Allison, J.P. Novel cancer immunotherapy agents with survival benefit: recent successes and next steps. *Nat. Rev. Cancer* **11**, 805–812 (2011).
7. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
8. Savage, P.A. *et al.* Recognition of a ubiquitous self antigen by prostate cancer-infiltrating $CD8^+$ T lymphocytes. *Science* **319**, 215–220 (2008).
9. Obenaus, M. *et al.* Identification of human T-cell receptors with optimal affinity to cancer antigens using antigen-negative humanized mice. *Nat. Biotechnol.* **33**, 402–407 (2015).
10. Tumeh, P.C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
11. Twyman-Saint Victor, C. *et al.* Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer. *Nature* **520**, 373–377 (2015).
12. Blachly, J.S. *et al.* Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA* **112**, 4322–4327 (2015).
13. Brown, S.D., Raeburn, L.A. & Holt, R.A. Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Med.* **7**, 125 (2015).
14. Bolotin, D.A. *et al.* MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* **10**, 813–814 (2013).
15. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
16. Warren, R.L., Nelson, B.H. & Holt, R.A. Profiling model T-cell metagenomes with short reads. *Bioinformatics* **25**, 458–464 (2009).
17. Freeman, J.D., Warren, R.L., Webb, J.R., Nelson, B.H. & Holt, R.A. Profiling the T-cell receptor β-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).
18. van Heijst, J.W. *et al.* Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. *Nat. Med.* **19**, 372–377 (2013).
19. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
20. Chien, Y.H. & Hampl, J. Antigen-recognition properties of murine γδ T cells. *Springer Semin. Immunopathol.* **22**, 239–250 (2000).
21. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
22. Dean, J. *et al.* Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome Med.* **7**, 123 (2015).
23. Rock, E.P., Sibbald, P.R., Davis, M.M. & Chien, Y.H. CDR3 length in antigen-specific immune receptors. *J. Exp. Med.* **179**, 323–328 (1994).
24. Venturi, V., Price, D.A., Douek, D.C. & Davenport, M.P. The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238 (2008).
25. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic $CD8^+$ T cell epitopes. *Proc. Natl. Acad. Sci. USA* **112**, E1754–E1762 (2015).
26. Schuurs, A.H. & Verheul, H.A. Effects of gender and sex steroids on the immune response. *J. Steroid Biochem.* **35**, 157–172 (1990).
27. Shukla, S.A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
28. He, C. *et al.* Genome-wide detection of testis- and testicular cancer–specific alternative splicing. *Carcinogenesis* **28**, 2484–2490 (2007).
29. Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T. & Old, L.J. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* **5**, 615–625 (2005).
30. Caballero, O.L. & Chen, Y.T. Cancer/testis (CT) antigens: potential targets for immunotherapy. *Cancer Sci.* **100**, 2014–2021 (2009).
31. Drake, C.G., Lipson, E.J. & Brahmer, J.R. Breathing new life into immunotherapy: review of melanoma, lung and kidney cancer. *Nat. Rev. Clin. Oncol.* **11**, 24–37 (2014).
32. Siliņa, K. *et al.* Sperm-associated antigens as targets for cancer immunotherapy: expression pattern and humoral immune response in cancer patients. *J. Immunother.* **34**, 28–44 (2011).
33. Andreatta, M., Schafer-Nielsen, C., Lund, O., Buus, S. & Nielsen, M. NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PLoS One* **6**, e26781 (2011).
34. Dunn, G.P., Bruce, A.T., Ikeda, H., Old, L.J. & Schreiber, R.D. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).
35. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).
36. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238 (2007).
37. Grupp, S.A. *et al.* Chimeric antigen receptor–modified T cells for acute lymphoid leukemia. *N. Engl. J. Med.* **368**, 1509–1518 (2013).
38. Porter, D.L., Levine, B.L., Kalos, M., Bagg, A. & June, C.H. Chimeric antigen receptor–modified T cells in chronic lymphoid leukemia. *N. Engl. J. Med.* **365**, 725–733 (2011).
39. Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

ONLINE METHODS

Data preparation and preprocessing. We downloaded the RNA-seq data of 9,142 TCGA samples from the Cancer Genomics Hub. Level 2 Affymetrix SNP 6.0 array data for TCGA tumors were downloaded from the TCGA data portal. Other information, including somatic mutation profiles, gene expression data and clinical annotations, were downloaded from GDAC Firehose. The purity of all TCGA tumors was inferred using the R package CHAT[40].

The RNA-seq data were in BAM format, which had been aligned to human reference genome hg19 using MapSplice[41]. For each file, we first extracted all reads mapping to the three major TCR regions: TCRα (chr. 14: 22,090,057–23,021,075), TCRβ (chr. 7: 141,998,851–142,510,972) and TCRγ (chr. 7: 38,279,625–38,407,656). The TCRδ gene region (chr. 14: 22,891,537–22,935,569) is embedded in the TCRα region, and therefore the reads for this region will be obtained along with those for TCRα. All coordinates correspond to the hg19 genome build. Counts of mapped reads in each variable gene region for the TCRα and TCRβ chains were used to estimate the usage of different genes and PCA. Of all the mapped reads extracted in the above step, a subset had unmapped mates, which were potentially generated from the CDR3 regions and could not be aligned to the reference genome. We then screened all the reads in the BAM file and searched for the mate for each such read until all mapped reads in the TCR regions were paired. The unmapped reads found in this step were used for CDR3 *de novo* assembly.

CDR3 *de novo* assembly and annotation workflow. *Pairwise read comparison.* For all unmapped reads we obtained from the previous step, we first allocated them to each gene according to the locations of their mapped mates. For unmapped reads assigned to the same gene, including variable (V), joining (J) and constant (C) genes, we exhaustively compared the sequences of each pair of reads, searching for sequence overlap larger than ten bases. We used two-bit coding for the nucleotide sequences to accelerate computation. We allowed one mismatch in the overlapped sequence, to tolerate sequencing errors. To note, the maximum probability of two random sequences sharing at least ten bases with a maximum of one mismatch is $(1 + 3 \times 10)/4^{10}$, or $\sim 3 \times 10^{-5}$, implying that a comparison of 1,000 reads may produce $1{,}000 \times 999/2 \times (3 \times 10^{-5}) = 15$ false positive assemblies. To reduce random sequence sharing and improve computational speed, when there were more than 1,000 unmapped reads in a gene region, we linearly split the reads into K smaller sets ($K = (N/200)$, where N is the number of unmapped reads) and performed pairwise comparison within each read set. This procedure prevents an extreme amount of reads generating numerous false positive assemblies, with the drawback that potential connections between read sets might be missed. As described below, a secondary assembly between contigs was introduced to rescue connections missing from the primary approach. The output of this step is read-sharing matrices individually documented for each gene.

Clique identification and contig assembly. Each matrix defined above can be represented by an undirected graph. Each node in the graph represents an unmapped read, and each edge connecting two nodes represents read sequence overlap. We traversed the graph to find all disjoint cliques. Within each clique, we then ordered the reads by the direction of their overlap. For example, if the 3′ end of read X overlapped with the 5′ end of read Y, X was placed in the 5′ end of Y and vice versa. Reverse-complement reads were converted before this analysis. After ordering, reads were assembled into one contig. It is possible for one read to overlap two reads in the same direction without these two reads overlapping. This situation suggests that there are two possible CDR3 sequences sharing the same upstream variable gene. In this scenario, we separately assemble each CDR3 sequence.

Merging contigs. In the above step, we assembled contigs using unmapped reads assigned to individual genes. This assignment is based on the mapping locations of the pair mates of the unmapped reads. Because a complete TCR transcript contains a variable, a joining and a constant gene, it is possible that a contig assigned to a variable gene comes from the same CDR3 sequence as a contig assigned to a joining or constant gene. Therefore, it is necessary to merge contigs from upstream genes (V) with those from downstream genes (J or C), because contigs of different genes may be produced by the same transcript. Such a scenario requires partial overlap between an upstream contig and a downstream contig. We next compared each pair of upstream–downstream contigs, and if the two overlapped by at least ten bases, with one mismatch tolerated, we merged the two contigs into one if these criteria were met.

Annotation of CDR3 sequences. To annotate assembled contigs, we downloaded the amino acid sequences of all the V and J genes from the international immunogenetics information system (IMGT)[42] in fasta format. We first translated the DNA contigs assembled using the above procedures into six reading frames (three for each strand). Sequences containing a stop codon were removed. For each remaining reading frame, we matched the consensus motif and its variations of the V gene (CAXS, CASX, CSVE, CAVX, CAGX, CAMX, CALX, CAFX, CAYX, CAEX) and J gene (FGXG, LGGG, VGPG). To minimize false positive calls, we only kept contigs containing the motifs of either the V or J gene.

For each contig kept, we extracted the residual sequence before the V gene motif or after the J gene motif (or both, if both were matched). If the residual sequence was longer than four amino acids, we further aligned it to the IMGT reference sequences and report the alignment with the highest score (S). Finally, we measured the quality of the assembly by $S + 4I_V + 4I_J$, where I_V was 1 if the motif for the V gene was matched and 0 otherwise and I_J was 1 if the motif for the J gene was matched and 0 otherwise. Intuitively, this quality measure is the mapping score including the conserved V or J gene sequence motifs. Our final report is in fasta format of contig DNA sequence, containing the V gene, J gene and assembly score as well as the amino acid sequence of the CDR3 region in the information line. Source code is available upon request.

Genomic DNA extraction and TCRβ repertoire sequencing. Three cases of clear cell renal cell carcinoma (TCGA-CZ-5463-01A, TCGA-CZ-5985-01A and TCGA-CZ-4862-01A) from Brigham and Women's Hospital previously analyzed in the KIRC TCGA cohort were selected for DNA extraction. Institutional review board approval was obtained before analysis (protocols DFCI 01-130 and DFCI 07-336). FFPE tissue sections stained with hematoxylin and eosin were reviewed by an expert genitourinary pathologist (S.S.), and for each case a tissue block with histopathological features comparable to the block analyzed by TCGA was chosen. FFPE tissue sections (8 μm) were prepared, and tumor areas of at least 6 cm^2 were dissected for each case, using a surgical blade. DNA extraction was performed using the QIAamp DNA FFPE Tissue kit (Qiagen) according to the manufacturer's guidelines. TCRβ repertoire sequencing of the genomic DNA from the three TCGA tumors was performed using immunoSEQ at survey level. The raw TCR reads were preprocessed with the immunoSEQ Analyzer.

Validation based on *in silico* simulations. To thoroughly evaluate the performance of our CDR3 assembly method, we introduced two *in silico* simulation approaches (**Supplementary Fig. 4**). In the first simulation, we generated pseudo tumor samples with various levels of T cell infiltration using an *in silico* mixing approach. Specifically, we downloaded the raw reads (paired end, 150 bp) from a peripheral blood sample profiled by deep TCRβ sequencing[3] (TCR-seq). In total, there were approximately 70,000 clonotypes as estimated by MiTCR[14]. Meanwhile, we downloaded raw reads (paired end, 75 bp) from an ENCODE cell line, K562 (ENCFF002DKI and ENCFF002DKF), derived from leukemia cells that do not express TCR transcript. Lack of TCR expression in the K562 cell line was further confirmed by zero CDR3 calls when we applied our method to this data set. We sampled a subset of the reads from each data set, truncated the reads to 50 bp and mixed the two sources into one fastq file. In total, we sampled 130 million reads from the RNA-seq data. For the TCR-seq data, we sampled approximately 8,000, 40,000, 120,000 and 250,000 reads to simulate T cell infiltration levels of 2%, 10%, 30% and 60%, respectively. To be clear, an infiltration level of X% does not imply that X% of the complete T cell repertoire is included in the mixed data set but indicates that the amount of TCR reads reaches the equivalent of X% of T cells in the whole tumor tissue. We then applied our method to the pseudo tumor samples and compared our CDR3 assemblies with the arbitrarily designated true TCR transcripts. The assembly rates, false positive call rates and clonal frequencies of the CDR3 sequences called using our method were thoroughly evaluated using the true TCR sets, and the results are discussed in the main text and **Supplementary Figure 5**.

For both real biological samples and the pseudo tumor samples generated using the *in silico* mixing approach, the actual coverage for the TCR transcripts was as low as 0.04. To explore the relationship between our CDR3 assembly rate and sequence depth across a wider spectrum, we introduced a second simulation approach as described below. We simulated 100 TCRβ transcripts using IMGT reference sequences through random V(D)J recombination. The full-length transcripts were used as templates to generate Illumina paired-end short reads of 50 bp by simNGS. This tool first produces a simulated library containing sequence fragments, from which it samples short reads using a noise model to generate sequencing errors. Our method was then applied to the resulting data sets with different coverage settings (1, 2 and 3). Here coverage of 1 was defined as each transcript being covered on average once. At each coverage setting, we repeated the simulation 100 times to estimate the uncertainty of the true and false positive rates. The output CDR3 calls were then compared to the simulated set of true transcripts to evaluate the performance of our method (main text and **Supplementary Fig. 6**).

Comparison to a competing method. iSSAKE[16] is a previously developed method that was applied to analyze TCR sequences through *de novo* assembly of short reads. We compared its performance with our method using the first simulation approach. At low coverage, our method retrieved at least an order of magnitude more CDR3 sequences than iSSAKE (**Supplementary Fig. 6**). This difference potentially occurs because the latter approach applied a *k*-mer assembly approach and limited its search depth to a given threshold, preventing the algorithm from comprehensively traversing all possible combinations.

Clonotypes per kilobase of reads calculation. According to a previous analysis of the TCR repertoire in peripheral blood using deep TCR sequencing[17], the number of distinct TCRβ sequences identified increases with read counts, and the number of TCRβ sequences plateaued at 2×10^8 reads for one blood draw. At low read coverage, the number of distinct CDR3 sequences increases approximately linearly with read count. Comparing to our observation in **Figure 4a**, we concluded that the read counts we obtained from RNA-seq data were far from saturation and stayed in the linear phase. This result suggests that the ratio of distinct CDR3 sequence calls over the number of reads is independent of sequence library size, as long as the read count does not reach the saturation level, which is typically larger than 100 million reads. The maximum read count extracted from RNA-seq data in our study was around 1 million reads, orders of magnitude smaller than this threshold. Therefore, in our analysis, it was not necessary to control for library size in the CPK analysis.

The total read count in each tumor consisted of reads from all three TCR regions (α, β and γ). In most cancer types, TCR transcripts were the only product from these regions. However, there were two exceptions. In pancreatic cancer, *PRSS2* (encoding protease serine 2), located in the TCRβ region, was highly expressed, as was *TARP* (encoding TCRγ alternative reading frame protein), located in the TCRγ region, in prostate cancer. Thus far, our CDR3 assembly was not affected by the expression of non-TCR genes, but the median levels of CPK were affected. Therefore, we excluded pancreatic and prostate cancers from our analysis in **Figure 4b**,**c**. Analyses of correlation between CPK and other factors (such as gene expression in **Fig. 5** and **Supplementary Fig. 9**) were not affected.

Estimation of reads/cell ratio. As mentioned in the *in silico* simulation analyses, we applied low coverage of 1, 2 and 3 for our simulated data sets for method validation and demonstrated that our method performs essentially better than a competing approach. To justify this low coverage for real RNA-seq samples, we will discuss how an experimental factor, the reads/cell ratio, is related to coverage and T cell infiltration levels. It is estimated that TCR transcripts account for 5×10^{-4} of the total transcripts in a T cell[13]. The median T cell infiltration in TCGA samples was 2%. For an RNA-seq library with 200 million reads (the median for TCGA data), the expected number of reads contributed by the TCR regions is 200 million × 0.02 × (5×10^{-4}) = 2,000. A tumor with a volume of 1 cm^3 typically contains approximately 100 million cells[43], and with 2% infiltration the expected number of T cells is 2 million. The estimated reads/cell ratio for complete TCR transcripts was 2,000/2 million = 0.001. In each cell, the TCR transcripts are identical. Given 2 million cells, the median number of cells per T cell clone is 40, as estimated using data from a previous study[44]. Therefore, the median coverage per transcript in a tumor of 1 cm^3 in size with 2% T cell infiltration and sequenced with 200 million reads is 40 × 0.001 = 0.04.

From the above discussion, we can see that the reads/cell ratio is directly related to library size and is linearly related to coverage. In our simulations, we have thoroughly explored the true and false positive rates of CDR3 assembly under different levels of coverage. Therefore, estimation of the reads/cell ratio is helpful in interpreting the results. Specifically, in our first simulation, with infiltration of 60%, at a library size of 100 million reads, the reads/cell ratio is only 0.0005 and median coverage is 0.02, which explains why our assembly rate (4%) is much lower than that of the scenario with coverage = 1 (33%).

Permutation analysis of CDR3 motifs and recurrent somatic mutations. We used the middle part of the complete β-CDR3 sequences as CDR3 motifs and searched for recurrent motifs in all CDR3 assemblies among TCGA tumors. We further selected the motifs that recurred between 5 and 20 times. The upper limit was chosen to exclude public T cell motifs, and the lower limit was chosen to reduce the chance of selecting random motif overlaps. We applied these filters to keep motifs with potentially high statistical power. This analysis resulted in an $S \times M$ motif sharing matrix, where S is the number of individuals (S = 8,984) and M is the number of distinct motifs that passed filtering (M = 5,347). Meanwhile, we selected 2,353 recurrent nonsynonymous somatic mutations, involving 7,449 tumors. Both matrices were binary coded, taking a value of 1 when the individual carried the motif or mutation and a value of 0 otherwise. The two data sets had 5,689 individuals in common. Our goal was to find mutation–motif pairs that co-occurred more often than expected from random chance. Theoretically, we needed to explore every pair, estimate the statistical significance of independence and correct the FDR. However, this would have required us to perform over 12 million tests, which would be computationally expensive. Therefore, we took an alternative approach as described below.

We first calculated the number of co-occurrences for each mutation–motif pair, which was computationally efficient in the R programming language. We selected the top 1,822 pairs with more than two co-occurrences and tested significance using Fisher's exact test for each pair. To estimate whether the observed significance could be achieved by random arrangements, we performed a rigorous permutation analysis. We randomly permutated the individual IDs of the motif sharing matrix, computed the number of co-occurrences for each pair using the permutated matrix, selected the top pairs with more than two co-occurrences and used Fisher's exact test to obtain P values. We repeated this analysis 1,000 times (resulting in over 1.5 million tests) and estimated the null distribution by pooling all the top K significant P values. We compared this null distribution of $K \times 1{,}000$ ordered statistics (top K null P values) with our true signals and estimated the corrected significance level using the fraction of null P values smaller than the given true P values. The advantage of this approach is that it automatically corrects for multiple tests because it compares the distribution of P values under an alternative hypothesis with that generated under the null hypothesis. Therefore, instead of requiring millions of comparisons, our final analysis only involved K tests. We found that when $K \leq 3$, all K P values were significant, and we chose K = 3 in our final report.

Other statistical analysis. To study the distribution of TRAV and TRBV gene usage, we first pooled all the reads mapped to each gene region across individuals, counted the reads and normalized the read counts by gene length. The proportion for each gene was then calculated as the ratio between the normalized count for this gene and the summation of all normalized counts. To investigate the patterns of TRAV and TRBV gene usage in different cancers, we calculated the fraction of each AV or BV gene by the ratio of normalized counts to total counts within each individual. This analysis resulted in a gene × sample matrix. We applied rank transformation to the fraction of each gene within a sample to remove batch effects (in the case of ties, we used the minimum rank). We then performed PCA on ranks to evaluate between-tumor heterogeneity (**Fig. 1c**,**d**). Benjamini–Hochberg correction was applied to all the FDR analyses. Other analysis in this work, including Wilcoxon tests, Fisher's exact tests, binomial tests, survival analysis, linear regression and correlation tests, was applied using the R programming language[45].

doi:10.1038/NG.3581

40. Li, B. & Li, J.Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.* **15**, 473 (2014).
41. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
42. Lefranc, M.P. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb. Protoc.* **2011**, 595–603 (2011).
43. Del Monte, U. Does the cell number 10(9) still really fit one gram of tumor tissue? *Cell Cycle* **8**, 505–506 (2009).
44. Emerson, R.O. *et al.* High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. *J. Pathol.* **231**, 433–440 (2013).
45. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).