ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

A Computational Tool to Predict T Cell Tumor Reactivity from Single-cell Transcriptomics Data

AUTOR/A: Changsheng Li	**NIA: 105532**
GRAU: Bioinformàtica	
CURS ACADÈMIC: 3	
DATA: 20 de juny, 2022	
TUTOR/S: Santiago J. Carmona; Massimo Andreatta	

Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

A Computational Tool to Predict T Cell Tumor Reactivity from Single-cell Transcriptomics Data

Changsheng Li

Scientific director: Santiago J. Carmona; co-director: Massimo Andreatta

Department of Oncology, UNIL CHUV and Ludwig Institute for Cancer Research Lausanne, University of Lausanne, 1011 Lausanne, Switzerland

Agora Cancer Research Center, 1011 Lausanne, Switzerland

Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract

Motivation:

$CD8^+$ T cells are one of the most important players of anti-tumor immune responses, and the main target of immunotherapies for cancer. T cell receptors (TCRs) expressed by T cells can recognize tumor-specific antigens, leading to the destruction of tumor cells. There is a great interest in the identification of tumor-reactive TCRs, as they present great therapeutic opportunities. In recent years, single-cell sequencing technologies have enabled exploration of T cells responses at high resolution. Here we want to provide a computational tool that can predict tumor-reactive TCRs based on single-cell transcriptome and T cell receptor sequencing data.

Results:

Here we developed "TCRanker", a computational tool that takes as input single-cell RNA-sequencing (scRNA-seq) data paired with single-cell TCR-sequencing (scTCR-seq) data, and ranks TCRs in a tumor sample according to their likelihood to recognize tumor antigens. To rank TCRs, TCRanker evaluates T cell clonal expansion and T cell transcriptional features associated with tumor reactivity in each T cell clonotype. We compared TCRanker with other tools, and evaluated its predictive performance using publicly available data, demonstrating the utility of our approach.

TCRanker is available on GitHub at carmonalab/TCRanker. A demonstration is also publicly available at carmonalab/TCRanker.demo.

Supplementary information:

Supplementary material 1: Script of FUN test and ProjecTILs validation.

Supplementary material 2: Script of TCRanker demonstration with B16 data set[1].

Supplementary material 3: Script of comparison of TCRs in tumor and peripheral blood[2].

1 Introduction

T cell plays a crucial role in anti-tumor immunity. Among them, the cluster of differentiation 8 ($CD8^+$) T cell, with its cytotoxicity, takes core action to kill cancer cells. They can recognize the unique tumor antigens presented by cancer cells after mutations because of T cell receptors (TCRs). As the structure on the surface of T cells that recognizes and binds antigens, TCR has two chains (alpha and beta) that contain highly variable segments. The enormous possibility of variation happens with gene rearrangement when T cells mature in the thymus and would be kept afterward. Once encountered

matching antigen, the T cell would start to proliferate and expand to a clone. T cells in a clone lineage have the same TCR clonotype. Thus, we can classify $CD8^+$ T cells by TCRs.

There is a great interest in the identification of tumor-reactive TCRs, as they present great therapeutic opportunities. The genetic engineering of TCR is also a mature field for modifying T cells (including T stem cells). Currently, various therapies with different strategies apply to T cells to treat cancer[3]. But the TCR selection itself remains a complicated and expensive problem.

While no repertoire of tumor-reactive TCR across tumors and individuals exists due to immune heterogeneity, the tumor-infiltrating lymphocytes (TIL) could provide a personalized pre-selection. Although tumors are thought to be enriched in tumor-reactive T cells (compared e.g., with peripheral blood), not all TILs are tumor-reactive. Some have estimated tumor-reactive TCRs among TILs are around 10%[4]. In this work, we aim at providing a tool to identify them, or at least "concentrate" them to a significant level.

T-cell exhaustion is one potential indicator of tumor reactivity, a state in which T-cells dysfunction occurs hierarchically after chronic infections, including cancer[5,6]. The higher the antigen exposure level is, the higher the T cells' exhaustion level would be[7,8]. Although it would also relate to the tumor developmental stage, we could validate the relativity between highly exhausted $CD8^+$ T cell clonal populations and tumor reactivity in vitro later. But to do so, we need to screen them first.

The Next-generation sequencing technologies (NGS) have enabled RNA sequence analysis at single-cell resolution (scRNA-seq)[9]. The ability also comes with single-cell T-cell receptor sequencing (scTCR-seq)[10]. They could provide us with the data and, most importantly, the ability to apply efficient bioinformatics methods to traditionally complicated problems like TCR selection.

2 Objectives

We want to create a tool to predict tumor-reactive TCRs from single-cell transcriptomics data of tumor biopsies. The algorithm will quantitatively profile the expansion, proliferation, and exhaustion level of $CD8^+$ TILs clonotypes, to provide a criterion to predict tumor reactivity.

3 Methods

3.1 TCRanker

To standardize and facilitate the TCR clonal state profiling workflow, we automated the process by an R package[11]. It is named "TCRanker".

3.2 Supported methods

Currently, we want TCRanker to support methods for both Seurat[12] and SingleCellExperiment[13] (SCE) objects as input.

3.3 Cell filtering

Generally, when sequencing the sample from tumor tissue, $CD8^+$ T cells and other immune cells, stromal cells, and cancer cells, would be sequenced indiscriminately. However here we are only interested in $CD8^+$ T cells for its cytotoxicity. Thus, a cell-type determination should be settled (optionally) before further processing to purify the clonotypes analyzed according to demands. Here scGate[14], an immune cell subtype assignment tool, is implemented to accomplish the cell type purification by the scRNA-seq.

When we set to filter $CD8^+$ T cells, scGate would tag those assigned to be the kind as "pure" and others as "impure".

We would expect the same clonotype coming from the same cell type. But some cells with the same clonotype might be tagged differently due to being on the fringe state or just noises. Only clonotypes with more than 50% pure cells would be considered valid and pass the filter in such cases. As for the cells in the valid clonotype but marked as "impure",

they, however, could be opted in but would be filtered out by default.

scGate supports gating not only $CD8^+$ T cells but also other immune cell types[14]. It gives true potential to generalize our package to other T cells even B cells.

3.4 Information gathering

- Expression matrix

  scRNA-seq expression data

- Clonotypes

  scTCR-seq clonotype data.

  In general practice, it would be presented in the format of sequences of joint cdr3 segments of alpha and beta chains (delta and gamma chains are also applicable).

- Groups

  In the case of a query containing several samples or individuals, a grouping function should also be included. When applied, TCR clones should be categorized by both clonotype and group.

3.5 Gene signatures

We want to embed default gene signatures of two transcriptomics characteristics: exhaustion and proliferation. The gene signatures are from Andreatta et al., 2021[15].

Note that regarding the gene signatures of exhaustion level, CD8 protein-coding genes should be taken out (Cd8a and Cd8b1 genes in mouse signature; CD8A and CD8B genes in human signature). They were included in the original study because these gene signatures were aimed simply at $CD8^+$ T cells. In our pipeline, the cell type would be purified in the cell filtering step, before utilizing the gene signature, so it is no longer vital here. Plus, to generalize our tool to more T cell sub-types, our default gene signature shouldn't be $CD8^+$ T cell-specific.

Furthermore, we wanted the gene signatures to be customizable to profile more cell states.

3.6 Individual scores

Single-cell sequencing technology drives us down to the resolution of individual cells. With UCell[16], we can make continuous measurements of cells from scRNA-seq data. Given a gene signature, it would return a score between 0 and 1, measuring the transcriptomics level of a particular cell in a specific state. Thus, we could quantify the transcriptome state of all cells at our disposal for further analysis of any state we are interested in if we could provide a proper gene signature.

3.7 Clonal aggregation

Identified by their unique cdr3 segments, $CD8^+$ T cells could be classified by their TCR clonotypes.

It remains unclear what the optimal distribution model of T cell exhaustion levels distribution would be. Therefore, the problem of how to aggregate the individual scores of all cells in a clonotype must be taken care of. It's natural to assume that the scores follow a normal distribution, and taking the average score is the best option. But without further study, we tried multiple functions.

In our case, considering that a higher exhaustion level potentially indicates higher tumor responsibility, it might suit our interests more to lean towards the higher end of scores distribution when determining a representative clonal score. We tested 5 different functions:

- max
- mean
- median
- ninetyPT: the ninety percentiles of scores distribution.
- top: the mean of top 10 percent scores, at least 5 cells.

We have an existing dataset from Xiong et al. (Cancer Immunology Research, 2019)[17] And we have five clonotypes validated in vitro. Clonotype 1 is reactive, while clonotypes 2 to 5 are non-reactive.

We will test all five functions and determine the best default one by the results. The workflow is available at supplementary material 1.

3.8 Minimal clonal size

In many scTCR sequencing results, we would usually have a limited number of cells in each clonotype. We assume that the T cells with relevant clonotype should have expanded in tumor tissue to be potentially tumor-reactive. So, when the sample was taken from tumor biopsies, we should have a minimal demand of cell numbers in a clonotype, i.e., clonal size, to dispose of irrelevant clonotypes from the ranking.

3.9 Cross validation

To assess the accuracy of exhaustion measurement of TCRanker, we used ProjecTILs[15] to make cross-validation. It is a tool that projects new data onto a reference cell atlas, allowing automated cell state annotation. We also used the object after the cell filter to project, as those not-filtered were not considered when ranking the TCRs. We projected the object before and after the filter, and cells with the top 3 clonotypes in Xiong's dataset[17] to ensure that TCRanker functions appropriately.

3.10 Generalization

We want to extend our tool's functionality to analyze more immune cell sub-types and more transcriptomics states. The general direction is to make TCRanker as customizable as possible.

3.11 Case studies

To illustrate the functionality and the potential utilization of TCRanker, we performed two case studies.

Firstly, as a basic illustration of TCRanker prediction. we analyzed paired scRNA and scTCR data from wildtype and PMEL mice that were both engraft with B16 melanoma (data from Carmona et al., 2020)[1]. The accession number is GSE116390[1].

The PMEL group was used as control of tumor reactivity since we knew these $CD8^+$ T cells are reactive. We used the exhaustion level of the clonotype that had the highest frequency in PMEL mice as a positive control. We predicted that TCRs from wildtype mice with higher exhaustion scores than the positive control would have a significant possibility to be reactive. The workflow is available at supplementary material 2.

Furthermore, we analyzed paired scRNA and scTCR samples from mice MC38 tumor tissue and peripheral blood (data from Pauken et al., 2021)[2]. The data contains "tumor-matching" $CD8^+$ T cells with the same TCR in both tumor and blood. The accession number is GSE158520[2]. We took two integrated data sets and run TCRanker separately over blood and tumor samples. The TCR information is stored in "TCR" and grouped in "Sample". When running the blood sample, the minimal clonal size was reduced to 1 (no demand regarding clonal size) to maximize the allowance of having enough matching TCRs in the result, as we don't expect having $CD8^+$ T cells significantly expanded in peripheral blood without other infections.

With the ranking results of both blood and tumor samples, we could find matching TCRs by the intercept of two clonotype lists. We compared the exhaustion level between matching TCRs in blood and tumor, and we expect to see a clear progressive difference between the exhaustion level. A paired one-tailed t-test accompanies the comparison.

We are more interested in comparison with TCRs in blood. We want to know that is there any evidence of the development of those $CD8^+$ T cells with tumor-matching TCRs after being activated by dendritic cells in secondary lymphoid organs and before tumor infiltration solely among the $CD8^+$ T cells in the blood. An independent two-tailed t-test accompanies the comparison. The workflow is available at supplementary material 3.

4 Results

4.1 TCRanker

We developed TCRanker, an R package that takes scRNA-seq data and paired scTCR-seq data as input, filters and keeps only $CD8^+$ T cells (by default), classifies them by their TCR clonotypes and rank the clones by their exhaustion levels. The tool is well customizable. The basic workflow is shown in fig. 1. The source code of TCRanker is available at carmonalab/TCRanker. There is also a repository of demonstrations available at carmonalab/TCRanker.demo.

You can find explanations about parameters and output format in detail in the Appendix section.

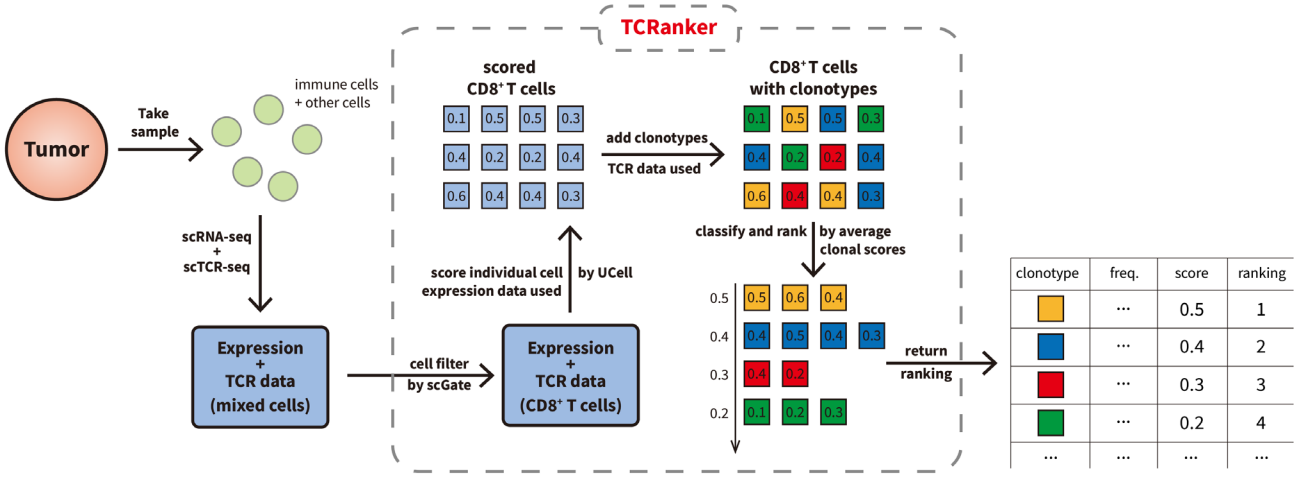

Fig. 1 | TCRanker workflow

4.2 ProjecTILs validation

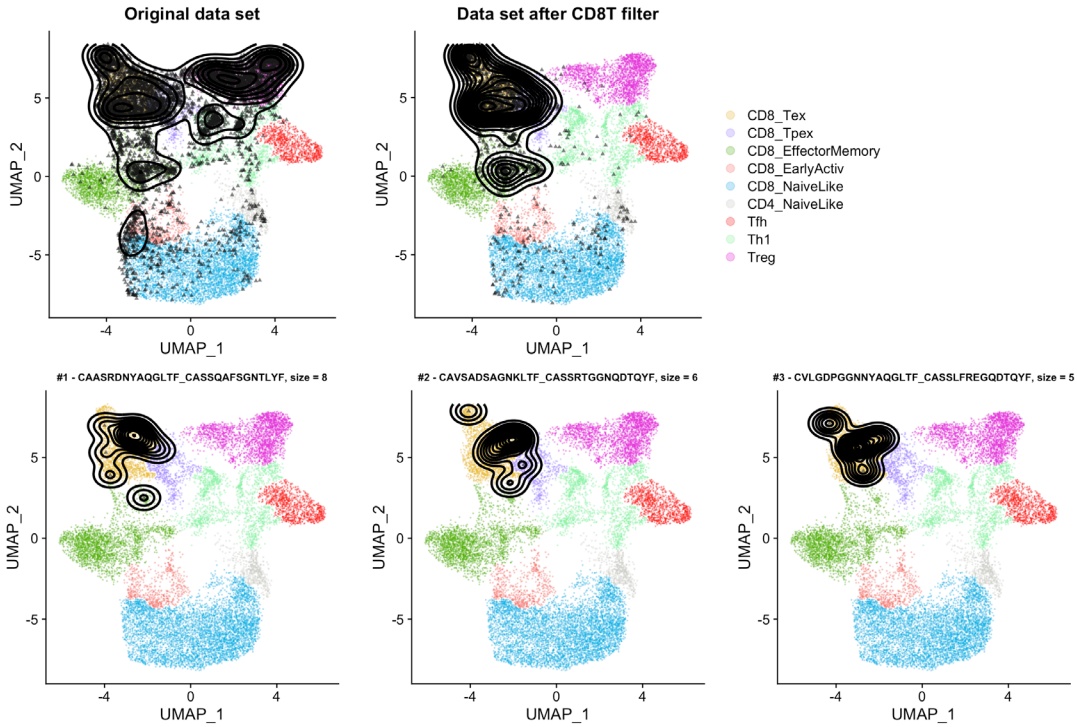

Fig. 2.1-2.2 | projection of the object before and after cell filter

Fig. 2.3-2.5 | cells projection with top 3 clonotypes

In fig. 2.1 and 2.2, we projected the data set onto the atlas before and after the cell filter (see supplementary material 1). The clusters of regulatory T (Treg) cells and T helper 1 (Th1) cells were removed as expected, and the clusters of $CD8^+$ T cells were correctly retained.

In fig. 2.3-2.5, cells with the top 3 ranked clonotypes were projected separately. They all fell into the area of exhausted $CD8^+$ T (CD8_Tex) cells. Thus, the correctness of the basic functionality of TCRanker was validated.

4.3 Case study 1: predicting $CD8^+$ TIL B16 tumor reactivity

After running TCRanker over B16 samples from both wildtype and PMEL mice, we have the clonal scores and ranking, and we wanted to predict the tumor reactivity in wildtype mice. We see in fig. 3.1 the clear maximal expansion of one clonotype in PMEL that has exhaustion score of 0.3444799. The number was then used as a positive control in fig. 3.2. By then, we predicted the 3 TCR clonotypes above the control should be tumor reactive.

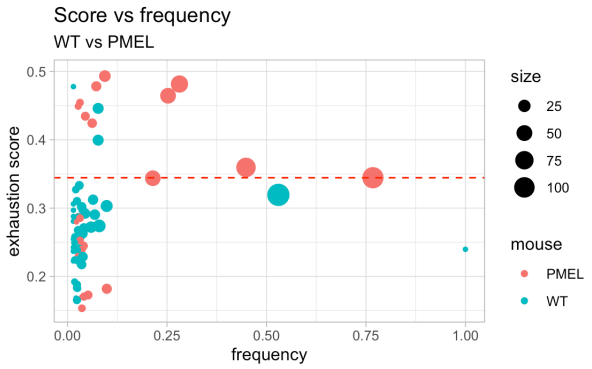

Fig. 3.1 | positive control determination

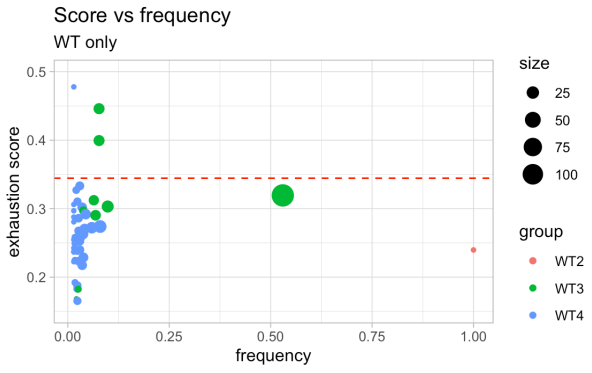

Fig. 3.2 | tumor-reactivity of TCRs in WT

4.4 Case study 2: comparing mouse tumor vs blood $CD8^+$ TIL clonotypes

From the tumor sample paired with peripheral blood sample, we obtained 69 tumor-matching (presented both in blood and tumor) and 8795 unmatching TCRs.

4.4.1 Between blood and tumor samples

In general, $CD8^+$ T cells showed higher exhaustion levels in tumors than in blood (fig. 4.1). Note that here the pseudo correlation between exhaustion score and frequency of tumor TCRs is not representative as it shows only the matching TCRs.

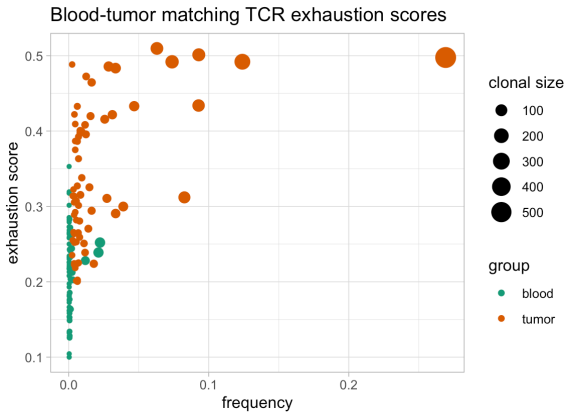

Fig. 4.1 | exhaustion scores of tumor-matching TCRs

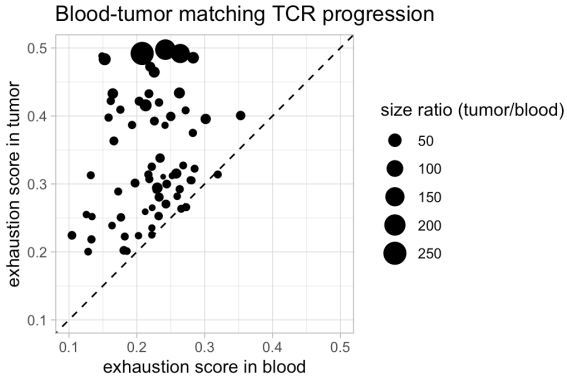

Fig. 4.2 | exhaustion score from blood to tumor

When pairing the same TCRs from two samples, almost all TCRs had developed more exhaustion in the tumors comparing with the same clonotype in the blood. And they are also well expanded (fig. 4.2).

The paired one-tailed t-test justified the exhaustion progress from blood to tumors with mean difference is 0.1207918, $P < 2.2 \times 10^{-16}$.

4.4.2 Within Blood samples

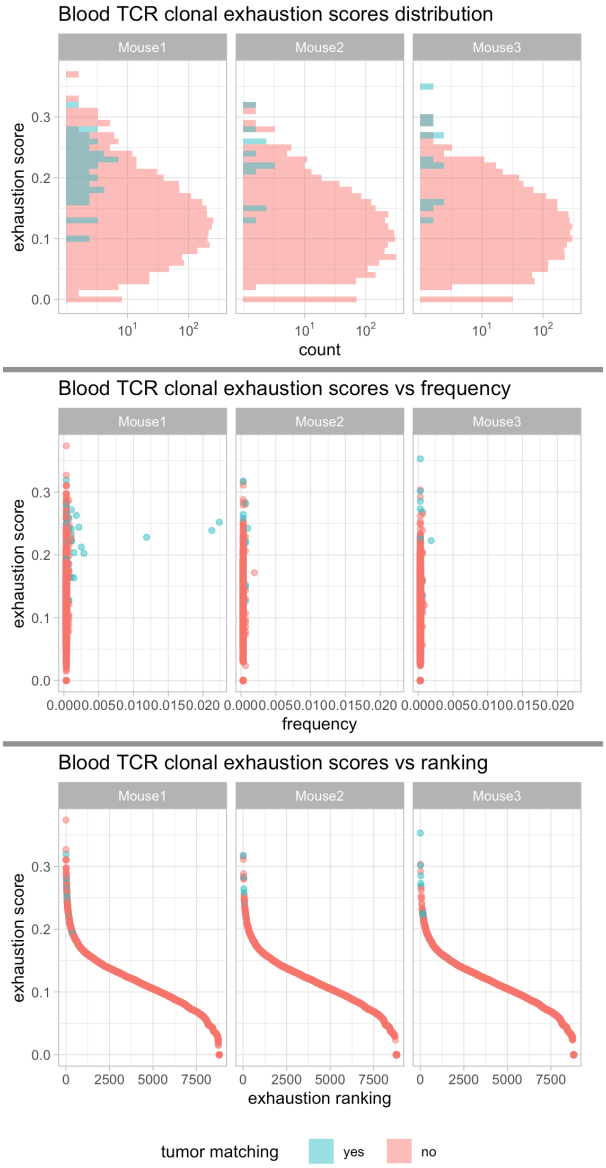

Fig. 5.1 | Blood TCR clonal exhaustion scores distribution

Fig. 5.2 | Blood TCR clonal exhaustion scores vs frequency

Fig. 5.3 | Blood TCR clonal exhaustion scores vs ranking

When comparing tumor-matching TCRs with unmatching ones (fig. 5), although unmatching TCRs spread far away to overlap the exhaustion range of matching ones in fig. 5.1, the two peaks were clearly staggered. It was proven by the independent one-tailed t-test: both coming from the blood sample, tumor-matching TCRs were found to be more exhausted than unmatching TCRs, mean of matching TCRs is 0.2179006 while mean of unmatching TCRs is 0.1128823, $P < 2.2 \times 10^{-16}$.

5 Discussion

To validate the efficacy of TCRanker, we will need data sets to be accompanied by in vitro experiment of tumor reactivity. Also new data will allow us to learn additional signatures of tumor-reactivity, complementary to that of T cell exhaustion. But currently, there are a minimal amount of data that satisfies our demand.

To tackle the data shortage, tumor reactivity lab validation could be conducted. When sufficient data is available, TCRanker can be configured with the optimal default parameters, and the viability of prediction would be confirmed.

In this work we have focused on CD8 TILs, since most tumor antigens are expressed in MHC class I. However, we envision extending our work to predict MHC class II-restricted tumor-specific TCRs, expressed in CD4+ T cells.

Since UCell could score cells against heterogeneity[16], we could expect clonal exhaustion scores derived from UCell should also be robust to heterogeneity. Thus, there might be an arbitrary numeric threshold of exhaustion scores (possibly around 0.5 given our current tests) that we could use as a positive control when there is no other control or comparing groups available.

6 Conclusions

To enable automated prediction of tumor-reactive T cells from single-cell transcriptomics data of tumors, we developed TCRanker. This computational tool that calculates relevant transcriptional features of $CD8^+$ T cell clonotypes from scRNA-seq and scTCR-seq data, including exhaustion and proliferation, and prioritizes TCRs according to these features. TCRanker is publicly available as an R package. We envision TCRanker will contribute towards faster identification of tumor-reactive TCRs, facilitating the development of personalized cancer immunotherapies.

7 Appendix

7.1 Algorithm Parameters

7.1.1 Query (indispensable)

It is an S4 object representing single-cell expression data for R. Currently, TCRanker supports methods for both Seurat and SingleCellExperiment (SCE) object as input.

An error would be raised with any other object format to inform users to use only supported formats.

Ideally, other metadata like clonotype and group should be within the query. But separated metadata is also allowed—details in later illustration.

7.1.2 TCR (indispensable)

It is the metadata of clonotypes. If it is within the query as a column, we would need only the column name here.

In rare cases, even though not recommended, users can also use a simple TCR vector separated from an S4 object. But the vector of clonotypes must be either the same length as the cell number or be named with cell barcodes for correspondence. Otherwise, such practice is not allowed.

7.1.3 signature

It is the gene signature that represents a transcriptomics characteristic or cell state. By default, we have the signatures of exhaustion and proliferation level of $CD8^+$ T cells.

Custom gene signatures are also allowed to extend the functionality and analyze more wanted cell states. Custom signatures should be wrapped in a list with entries appropriately named to avoid ambiguity in the output.

Having only one signature, a simple vector would also be acceptable, and the state would be called “user” in the output.

7.1.4 assay

It is the name of the scRNA-seq expression data assay. By default, TCRanker would seek "counts" in SCE and "RNA" in the Seurat object.

7.1.5 group

It is the extra group information to be aggregated and included in the output, and the same clonotype in different groups would be aggregated separately. It is a column name of metadata in the query object by default.

Even though optional, it's still recommended to be included when analyzing.

A vector format is also allowed, but the same requisites as parameter TCR in vector format apply here.

7.1.6 exhaustion

It is a logical parameter to include exhaustion scores and ranking in the output or not. TRUE by default.

7.1.7 proliferation

It is a logical parameter to include proliferation scores and ranking in the output or not. TRUE by default.

It is mandatory to have at least one set of gene signature. So, if you exclude both exhaustion and proliferation analysis, provide a custom gene signature instead.

7.1.8 species

It is a character parameter and will be auto-detected by the gene names if omitted.

Currently, TCRanker only supports mouse and human default cell filters and gene signatures. You may change it when using only custom signatures and the cell filter disabled.

7.1.9 FUN

It is the function used to aggregate scores of the same clonotype. It could be mean, median, or customized functions that take a numeric vector or list as input and return a single numeric result. By default, mean will be implemented.

The mean produced the optimal ranking among all the functions tested (see supplementary material 1), as shown in the fig. 6. It is still negotiable which the optimal function would be since we haven't had more in vitro results. So, with the mean being the default function, the choice of parameter FUN is still open to users.

Fig. 6 shows us another result: despite the function used, clonotype 1, the validated reactive one, would consistently be ranked among the top ones, at least the highest among these 5. It gave out some evidence of the viability of exhaustion level being an indicator of tumor reactivity.

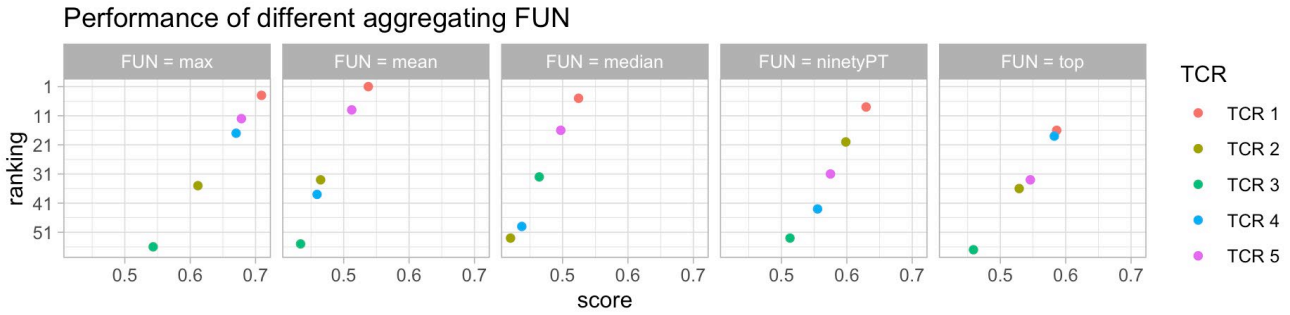

Fig. 6 | performance of different aggregation functions

7.1.10 minClonSize

It is a numeric parameter. It is the threshold of the minimal clonal size that would be considered as meaningfully expanded, 5 by default.

7.1.11 filterCell

Cells without clonotype information would be filtered first for efficiency.

It is the name of the sub-cell type to filter using scGate. The default cell type is "CD8T", and the filter depends on scGate.

If you have a data set with a pure cell type or don't want to filter the data set, set it to NULL to disable the filter.

7.1.12 strictFilter

It is a logical parameter. Concerning cell filter, if there are any impure cells (not wanted cell type) in pure clonotype (more than 50% of cells in this clonotype are the wanted cell type.), they would be removed by default.

The parameter is only valid when filterCell is on.

7.1.13 keepObject

It is a logical parameter. If you want to inspect the SCE or Seurat object after cell filtering, set it to TRUE. FALSE by default.

7.2 Output

7.2.1 Default output

The output data frame would have the following structure in table 1.

Columns	Description
clonotype	Clonotypes, the form depends on the input
group	Group info., optional, depends on whether in the input
size	Clonal size, the count of cells, also group-related (if provided)
freq	Clonal frequency, also group-related (if provided)
score 1	Clonal state score by the first gene signature
ranking 1	Clonal state ranking by the first gene signature
score 2	Clonal state score by the second gene signature
ranking 2	Clonal state ranking by the second gene signature
…	More scores and rankings… (if more custom gene signatures provided)

Table 1 | columns in the output data frame

The existence of column "group" depends on whether it was provided in the input and will affect the number of clonotypes obtained in two ways: 1. The same clonotype in different groups would be ranked separately and thus might increase the number of entries; 2. The clonotype sub-divided by the group would potentially fall beneath the minimal clonal size required and thus might decrease the number of entries.

7.2.2 Output with keepObject option on

In this case, TCRanker would return the auxiliary SCE/Seurat object after the cell filtering step and the ranking data frame in a list. The ranking data frame would be named "TCRanking", and the object would be named "filteredQuery" in the list.

8 Bibliography

1. Carmona, S. J., Siddiqui, I., Bilous, M., Held, W. & Gfeller, D. Deciphering the transcriptomic landscape of tumor-infiltrating CD8 lymphocytes in B16 melanoma tumors with single-cell RNA-Seq. *OncoImmunology* **9**, (2020).
2. Pauken, K. E. *et al.* Single-cell analyses identify circulating anti-tumor CD8 T cells and markers for their enrichment. *Journal of Experimental Medicine* **218**, (2021).
3. Guedan, S., Ruella, M. & June, C. H. Emerging Cellular Therapies for Cancer. *Annual Review of Immunology* **37**, 145–171 (2019).
4. Scheper, W. *et al.* Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. *Nature Medicine* **25**, 89–94 (2019).
5. Blank, C. U. *et al.* Defining 'T cell exhaustion.' *Nature Reviews Immunology* **19**, 665–674 (2019).
6. Oliveira, G. *et al.* Phenotype, specificity and avidity of antitumour CD8+ T cells in melanoma. *Nature* **596**, 119–125 (2021).
7. Day, C. L. *et al.* Proliferative Capacity of Epitope-Specific CD8 T-Cell Responses Is Inversely Related to Viral Load in Chronic Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology* **81**, 434–438 (2007).
8. Wherry, E. J. T cell exhaustion. *Nature Immunology* **12**, 492–499 (2011).
9. Eberwine, J., Sul, J. Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nature Methods* vol. 11 25–27 (2014).
10. de Simone, M., Rossetti, G. & Pagani, M. Single cell T cell receptor sequencing: Techniques and future challenges. *Frontiers in Immunology* vol. 9 (2018).
11. R Core Team. R: A Language and Environment for Statistical Computing. (2022).
12. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
13. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nature Methods* **17**, 137–145 (2020).
14. Andreatta, M., Berenstein, A. J. & Carmona, S. J. Gene expression scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets. doi:10.5281/zenodo.6202614.
15. Andreatta, M. *et al.* Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nature Communications* **12**, (2021).
16. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Computational and Structural Biotechnology Journal* **19**, 3796–3798 (2021).
17. Xiong, H. *et al.* Coexpression of inhibitory receptors enriches for activated and functional CD8+ T cells in murine syngeneic tumor models. *Cancer Immunology Research* **7**, 963–976 (2019).