nature portfolio

Corresponding author(s): Linghua Wang

Last updated by author(s): Apr 12, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	A detailed description of softwares and code as well as parameter settings used for bioinformatics data analysis are provided in the Methods. All other statistical analyses were performed using statistical software R v3.6.0 and v4.0.3, which are also described in the Methods. The R script TCellMap is available at GitHub (https://github.com/Coolgenome/TCM). An open-source implementation of the TESLA algorithm in Python can be downloaded from https://github.com/jianhuupenn/TESLA. The custom script used to overlay the spatial locations of the hypoxia signal and TSTR cells on the same histology image is available at GitHub (https://github.com/Coolgenome/TCM/blob/main/res_largerT.py#L230). In addition, we have built a user-friendly and interactive on-line data portal, Single Cell Research Portal (SCRP, https://singlecell.mdanderson.org/), for visualizing scRNA-seq data. All scRNA-seq data used to build T-cell reference maps in this study can be visualized and queried via SCRP at https://singlecell.mdanderson.org/TCM/.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data availability: A detailed description of data availability including data sources and accession numbers of the scRNA-seq datasets included in the original data collection was provided in Supplementary Tables 1 and 2. In this study, we utilized 10 newly generated scRNA-seq datasets (labeled as "in-house" Supplementary Tables 1, column I). Specifically, the AML dataset can be downloaded from the European Genome-Phenome Archive (EGA) database under the accession number EGAD00001007672. The lung cancer (LC_1) dataset can be downloaded from EGA under the accession number EGAS00001005021. The lymphoma dataset (LN_2) can be downloaded from EGA under the accession number EGAS00001006052. The scRNA-seq data generated on PBMC samples from healthy donors (PBMC_3) can be downloaded from EGA under the accession number EGAD00001006994. The GBM datasets can be downloaded from the Gene Expression Omnibus (GEO) database under the accession number GSE222522. The breast cancer (BRCA_2) dataset, the lung cancer (LC_5) dataset, the ovarian cancer (OV) dataset, and the STAD dataset can be downloaded from GEO under the accession number GSE222859. The scRNA-seq data generated on reactive lymph nodes from healthy donors (LN_1) can be downloaded from GEO under the accession number GSE203610. For the six scRNA-seq datasets generated from patients who received ICB therapy, their data accession numbers, references, and detailed clinical information are provided in Supplementary Table 15. For the four scRNA-seq datasets used as a demonstration of TCellMap, their data accession numbers and references are provided in Supplementary Table 16. The CosMx SMI datasets generated on NSCLC and HCC samples can be downloaded from https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/. For Visium spatial transcriptomics datasets used in this study, the expression count matrices for the BRCA study can be downloaded from https://github.com/almaan/her2st. The melanoma Visium data can be downloaded from https://www.spatialresearch.org/resources-published-datasets/doi-10-1158-0008-5472-can-18-0747/. The CSCC Visium data can be obtained from GEO under the accession number GSE144240. The ccRCC Visium data can be obtained from GEO under the accession number GSE175540. The LUAD Visium data can be obtained from EGA under the accession numbers EGAS00001005021. Further information and requests should be directed to and will be fulfilled by the corresponding author Dr. Wang (LWang22@mdanderson.org). All requests for data and materials will be promptly reviewed by The University of Texas MD Anderson Cancer Center to verify if the request is subject to any intellectual property or confidentiality obligations. Any data and materials that can be shared will be released via a Material Transfer Agreement.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex was not considered in the study design/not sufficient statistical power to perform sex-specific analyses
Reporting on race, ethnicity, or other socially relevant groupings	No socially relevant categorization variables or terms used.
Population characteristics	See above
Recruitment	No patient recruitment involved, as this not a prospective study. This is not a clinical trial.
Ethics oversight	All experiments were compliant with the review board of the University of Texas MD Anderson Cancer Center (MDACC), and the studies were conducted in accordance with the Declaration of Helsinki. For the LUAD (LC_1) study, all samples were obtained under the waiver of consent from banked or residual tissues approved by MDACC internal review board (IRB) protocols (PA14-0077 and LAB90-020). For the rest of the cohorts, written informed consent was provided by all patients. Tumor specimens were collected with informed consent in accordance with the MDACC IRB-approved protocols (LN_1, LN_2, and BRCA_2: PA19-0420; GBM: 2012-0441; AML: PA12-0305; HNSC_2: 2019-1059, LAB02-039, and PA18-0782; LUAD LC_5: PA14-0276; OV: 2017-0264). For the STAD dataset, the study was approved by the Ethics Committee of Zhejiang Cancer Hospital (# IRB-2020-109) and all patients provided written informed consent to participate.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We obtained scRNA-seq data on T cells in 486 samples from 324 cancer patients and healthy donors. Ten out of 27 scRNA-seq datasets were generated internally. Following a rigorous quality control process, a total of 308,048 high-quality transcriptomes were retained for

	subsequent analyses. In Figs. 2i and 3d, downsampling analysis was performed. For CDS T cells, the down sampled cell number is 11,592 cells for each tissue group. For CD4 T cells, the down sampled cell number is 10,703 cells for each tissue group. For TCGA cohorts, 11,051 tumours with bulk RNA-seq data available were included. Among the 1,008 patients from the CPIIOOO+ cohorts, 562 patients with genomic data, expression data, and clinical response data available were included. For the CosMx SMI dataset, 8 tissue sections from 5 non-small-cell lung cancer (NSCLC) samples were included. To assess the clinical significance of identified T cell subsets in the context of ICB therapy, scRNA-seq data generated on a total of 247 samples from 133 patients across 6 cohorts and 4 cancer types in neoadjuvant and adjuvant settings were included. In addition, as a demonstration of TcellMap, 4 additional scRNA-seq datasets (as listed in the Supplementary Table 16) were included.
Data exclusions	For scRNA-seq datasets, the merged scRNA-seq data matrix was subjected to a multi-step filtering process to remove low-quality cells, likely cell debris and doublets. Cells with low complexity libraries (in which detected transcripts are aligned to less than 200 genes) were filtered out and excluded from subsequent analyses. Likely dying or apoptotic cells where >15% of transcripts derived from the mitochrondria were also excluded. In addition, cells with high-complexity libraries in which detected transcripts are aligned to> 6,500 genes were removed. The resulting matrix was further filtered to clean additional possible doublets. More details on doublet identification and removal are provided in Methods. Samples that had< 200 T cells were excluded from sample-level analysis. For TCGA cohorts and the CPIIOOO+ Cohorts, only patients with bulk expression data available were included. For T cell deconvolution analyses, samples with low abundance of T cells (the bottom 25% of the ranked data, Supplementary Table 12) were excluded. For correlation analysis in the CPIIOOO+ cohorts, 446 patient without expression or clinical response data were excluded. For group-level correlation analysis in six scRNA-seq cohorts received ICB therapy, T cell subsets with <100 total cells were excluded(< 30 tumor- or viral-specific CDS T cells for the scRNA-seq dataset from Caushi et al) were excluded when calculating Ro/e values to quantify their tissue prevalence between groups. In addition, samples with response information not available or not evaluated as described in the Supplementary Table 15 were also excluded. No sex- or gender-based analyses have been performed in this study due to incomplete information collected from public studies.
Replication	Three consecutive sections were included in the lung cancer CosMx dataset. In addition, the BRCA and CSCC spatial transcriptomics datasets included consecutive sections. The detailed information on replicates is provided in Supplementary Table 14. We observed very consistent results across these replicates as described in Extended Data Figures 8, 9, and Supplementary Figure 12.
Randomization	Randomization is not relevant to this study.
Blinding	Blinding is not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging