nature portfolio

Corresponding author(s): Hongyi Xin and Kong Chen

Last updated by author(s): Nov 2, 2022

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 commercial, open source or custom software or code was used for data collection.
Data analysis	Scripts for data analysis are available at https://github. com/QiuyuLian/TOMAS. Cell Ranger (v 2.2.0) was used for alignment and UMI counting of single-cell library. STAR (v 2.7.8a) and featureCounts (v 2.0.0) were used for alignment and feature counting of bulk RNA-seq data, respectively. Analysis was performed using Python (v 3.8.5), R (v 4.1.0) and packages Scanpy (v 1.7.1), Seurat (v 4.0.1), TOMAS (v 1.0.0), pyDIMM (v 0.0.2), GMM-Demux (v 0.2.1.3) and DoubletCollection (v 1.1.0).

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

Raw and pre-processed scRNA-seq and bulk-RNA-seq data will be publicly released to GEO upon manuscript acceptance. Temporary access for raw and pre-processed fastq data (during the reviewing phase) is granted through: https://tinyurl.com/2hu8d465. CellRanger processed gene-UMI counts data and the raw bulk-RNA-seq fastqdata for the in-house naive-and-activated CD4+ T cell dataset is deposited on Mendeley at https://data.mendeley.com/datasets/4whftgzxjj/draft?a=89c1109d-e81c-4f4b-8fe0-102563e710cd. Public CITE-seq human PBMC dataset [35] can be found at the GEO (GSE152981).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No human research participants were involved in this study. The in-house bulk and single-cell RNA sequencing data are from mouse models.
Population characteristics	n/a. See above.
Recruitment	n/a. See above.
Ethics oversight	n/a. See above.

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	To generate enough heterotypic doublets for total mRNA ratio inference, we collect around 24k naive T cells and 24k activated T cells for cell-hashing and scRNA-seq.
Data exclusions	We totally excluded 1,359 droplets, including poor-quality droplets detected using scanpy with default settings and triplets detected with GMM-Demux, from subsequent analysis, as described in Methods.
Replication	Two biologically independent replicates were performed independently and measured with scRNA-seq and bulk RNA-seq respectively. We observed high concordance among the two biological replicates in terms of differential expressed genes.
Randomization	For evaluating TOMAS's accuracy in total mRNA ratio inference, we randomly sampled datasets with different setting of ground truth of total mRNA ratios, as described in Figure 5, Figure S3 and Methods. We also randomly sampled various numbers of singlets and heterotypic doublets from real datasets to test TOMAS's sensitivity to cell-type size imbalance, as described in Figure S6. No other randomization strategies were applied.
Blinding	No blinding was performed as this was 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

Methods

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

Antibodies

Antibodies used

In vitro stimulus: Ultra-LEAF™ Purified anti-mouse CD28 Antibody 102116 biolegned, Ultra-LEAF™ Purified anti-mouse CD3 Antibody 100238.
FACS: Anti-mouse CD3 (BioLegend, Cat# 100216), anti-mouse CD4 (BioLegend, Cat# 100531), anti-mouse CD69 (BioLegend, Cat# 104505), anti-mouse CD62L (BioLegend, Cat# 161203), anti-mouse CD44 (BioLegend, Cat# 103011), anti-mouse CD27 (BioLegend, Cat# 124207).
Single-cell sequencing: six anti-mouse Cell Hashing antibodies (BioLegend, Cat# 155801, 155803, 155805, 155807, 155809, 155811).

Validation

All antibodies applied in this study are the widely-used clones in this field. We selected these antibodies which have been validated by the manufacturer and all information is available at the manufacturer website.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	C57BL/6J female mice were obtained from the Jackson Laboratory.
Wild animals	The study did not involve wild animals.
Reporting on sex	This study focuses on controlling total mRNA differences between cell types, which is a fundamental problem in comparative transcriptomic measurements. Total mRNA differences are widely observed in diverse tissues, organs and species, independent of sex. We provide a computational method harnessing heterotypic doublets as internal controls of total mRNA difference. Here, we use mouse CD4 T cells to showcase the importance of considering the total mRNA differences in scRNA-seq analysis. Sex is not relevant to this study.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	All mouse experiments were approved by the University of Pittsburgh Institutional Animal Care and Use Committee. See Methods, section 'Naive and Activated T Cell Preparation'.

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

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Total murine CD4 T cells were isolated from mouse spleen using magnetic beads. Cells were either left unstimulated or activated using plated-bounded anti-CD3/anti-CD28 overnight then harvested and stained with cell surface markers for T cell activation. Based on FACS analysis (BD FACSDiva), we sorted naïve CD4 T cells (CD4+CD69-CD62L+CD44lo) and activated CD4T cells (CD4+CD69+CD62L-CD44hi).
Instrument	Cell populations were sorted with ARIA-III instrument (BD Biosciences, San Jose, CA)
Software	Analysis was performed by using BD FACSDIVA software (BD Bioscience) and FlowJo software (FlowJo, LLC)

Cell population abundance

The abundance of the naive and activated CD4 T cells were calculated as the frequency of CD3+CD4+CD69-CD62L+CD44lo and CD3+CD4+CD69+CD62L-CD44hi cells, respectively, as described in Figure Sx.

Gating strategy

FSC/SSC gate -> gating on CD4+ T cells -> CD69/CD62L gate -> gating on CD69-CD62L+ naive cells and CD69+CD62L- activated cells,
CD69-CD62L+ naive cells -> CD44 gate -> gating on CD44 low cells to obtain pure CD4+ naive T cells,
CD69+CD62L- activated cells -> CD44 gate -> gating on CD44 high cells to obtain pure CD4+ activated T cells.
A visual representation is provided in Supplementary Figure S10.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.