nature portfolio

Corresponding author(s): Soumya Raychaudhuri and Kazuyoshi Ishigaki

Last updated by author(s): Dec 22, 2021

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 for data collection.
Data analysis	All data were analyzed with open-source software available online and detailed in Methods: cellranger (version 5.0.1, GRCh38-3.0.0), R (version >= 3.6.1). R packages: stats, lme4, broom, broom.mixed, glmnet, CCA, uwot, car, VIF, boot, dplyr, DescTools, infotheo, singlecellmethods, harmony, symphony, Seurat, class, ggseqlogo, ggplot2, ggrastr, ggpubr, ggrepel, pheatmap, RColorBrewer, pals, patchwork, corrplot2, officer, rvg.

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 analyzed in this study were previously deposited in the following locations:

immuneACCESS

DOI: https://doi.org/10.21417/B73S3K

DOI: https://doi.org/10.21417/B7C88S

DOI: https://doi.org/10.21417/AMT2019EJI

DOI: https://doi.org/10.21417/CS2020CR

DOI: https://doi.org/10.21417/B7001Z

Gene Expression Omnibus (GEO)

GSE158769

GSE123813

GSE114724

Github

URL: https://github.com/aleksobrad/humanized-mouse-data

Zenodo

DOI: https://doi.org/10.5281/zenodo.3711134

ArrayExpress

E-MTAB-8581

10X Genomics

URL: https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz

McPAS-TCR

URL: http://friedmanlab.weizmann.ac.il/McPAS-TCR

VDJdb

URL: https://vdjdb.cdr3.net

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 allow for ample statistical power, we conducted our main analyses with bulk sequencing data. Aggregating several publicly available datasets of TCR sequences from human peripheral blood flow-sorted into Treg and Tconv populations amounted to more than 2.4E7 TCR observations, affording >0.99 power to detect a mean difference as small as 0.005 standard deviations.
Data exclusions	For bulk sequencing data, we considered each unique TCR sequence within each donor to be a single observation and ignored duplicates. To define a tractable number of TCR positions for study, we excluded CDR3b sequences shorter than 12 amino acids or longer than 17 amino acids. For training our Treg vs Tconv statistical model, we included only data from donors in which both T cell phenotypes of interest were

	collected. To focus on Treg-Tconv distinctions, we excluded TCR sequences that were observed as both Treg and Tconv within the same donor.
Replication	To replicate our findings, we fit mixed-effects logistic regression models for the V-, J-, and middle regions of the TCR with the same covariates in an independent cohort of bulk TCR sequencing from flow-sorted Tregs and Tconvs. We further replicated our findings by applying the TiRP scoring system to two more independent cohorts of bulk TCR sequencing as well as three independent cohorts of scRNA sequencing. The findings replicated in six out of six cohorts examined.
Randomization	We considered each TCR sequence to be randomly assigned by the process of V(D)J recombination in the thymus. To control for possible confounds in this pseudo-randomization, we model donor and batch ID as random effects and a donor-individualized thymic selection rate parameter (Supplementary Note) as a fixed effect.
Blinding	Data from each individual were analyzed in the same manner. For our outcome of interest (Treg vs Tconv), the level of observation is the T cell, for which blinding is not meaningful.

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
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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