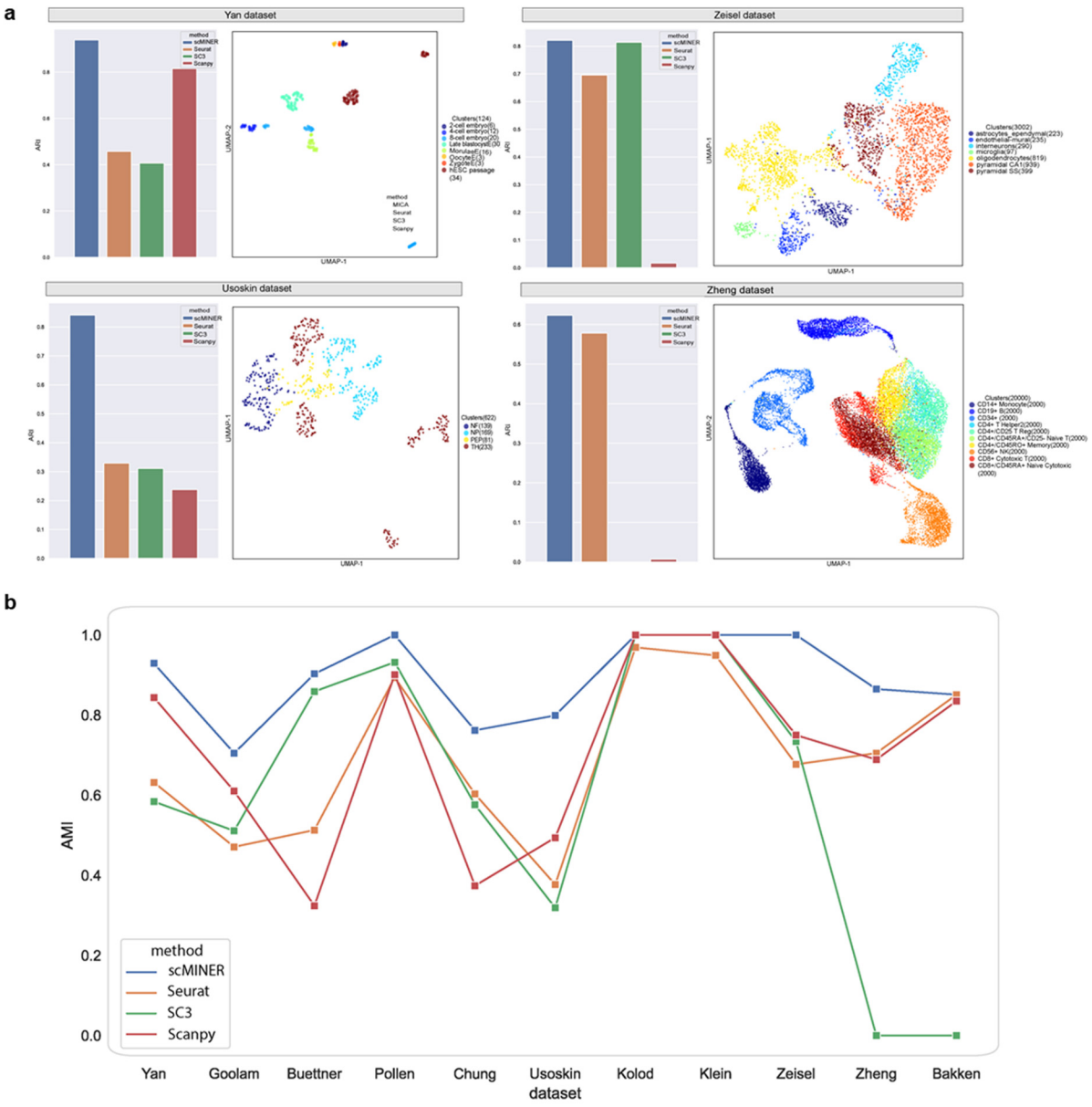
Supplementary Information

scMINER: a mutual information-based framework for identifying hidden drivers from single-cell omics data

Supplementary Figures 1-9.

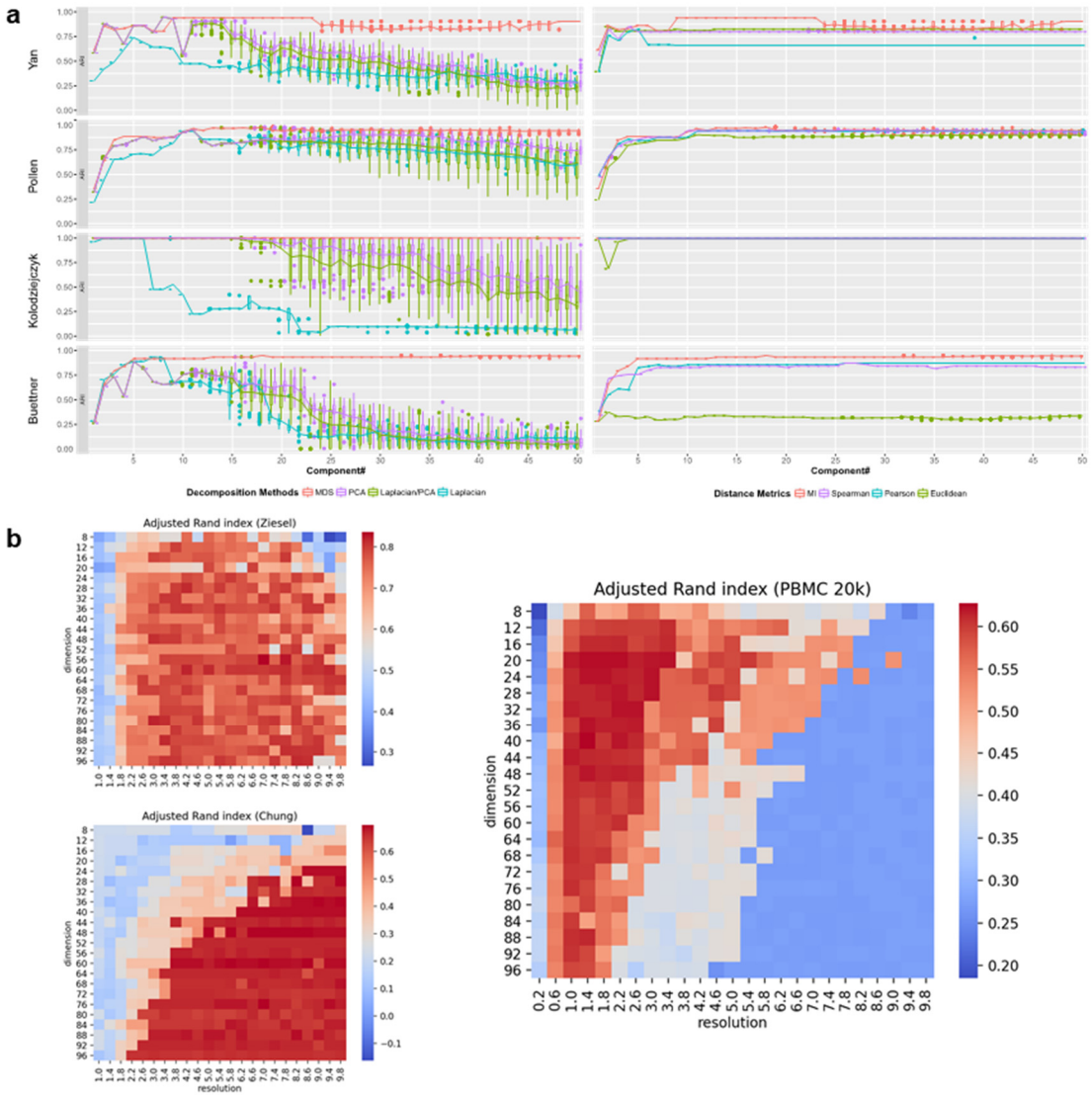
Supplementary Tables 1-2.

Supplementary Note.



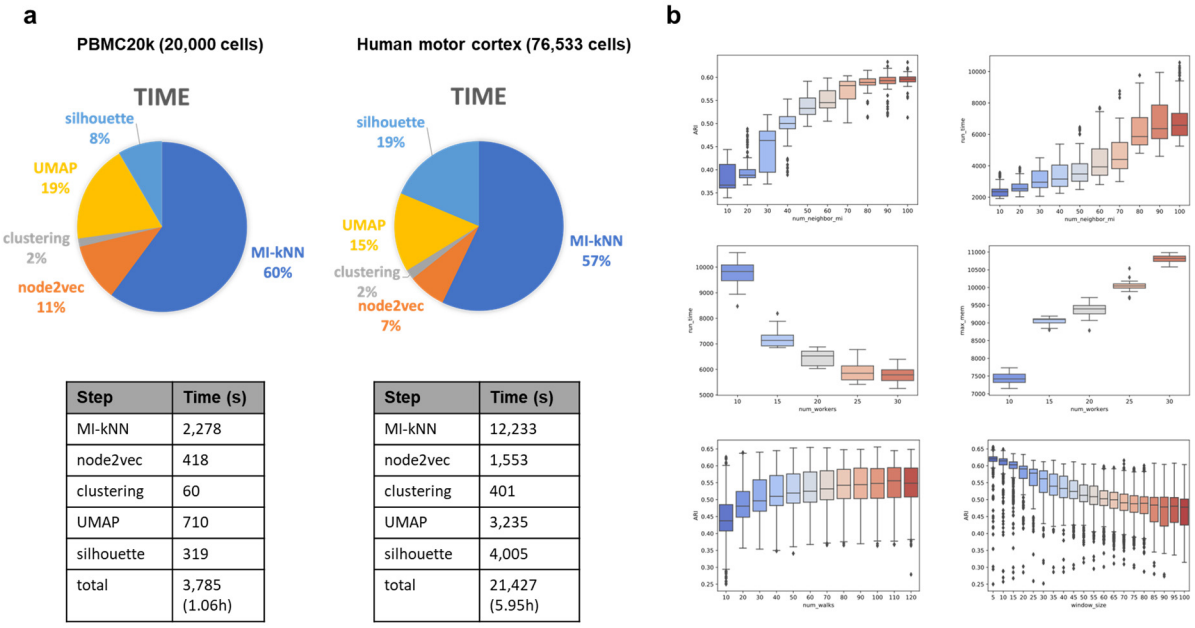
Supplementary Figure 1. scMINER clustering performance evaluation using AMI and true label projection on four datasets.

a, ARI bar plots and UMAP plots of scMINER clustering results annotated using true labels on Yan, Zeisel, Usoskon, and Zheng datasets. **b,** Clustering performance of scMINER, Seurat, SC3 and Scanpy measured by adjusted mutual information (AMI).



Supplementary Figure 2. Effect of distance metrics and parameters on the clustering performance.

a, Clustering performance comparison using four distance metrics (left) and four dimension reduction methods (right) on Yan, Pollen, Kolodziejczyk, and Buettner datasets. **b,** Clustering performance in term of ARI with respect to dimension and resolution parameters.

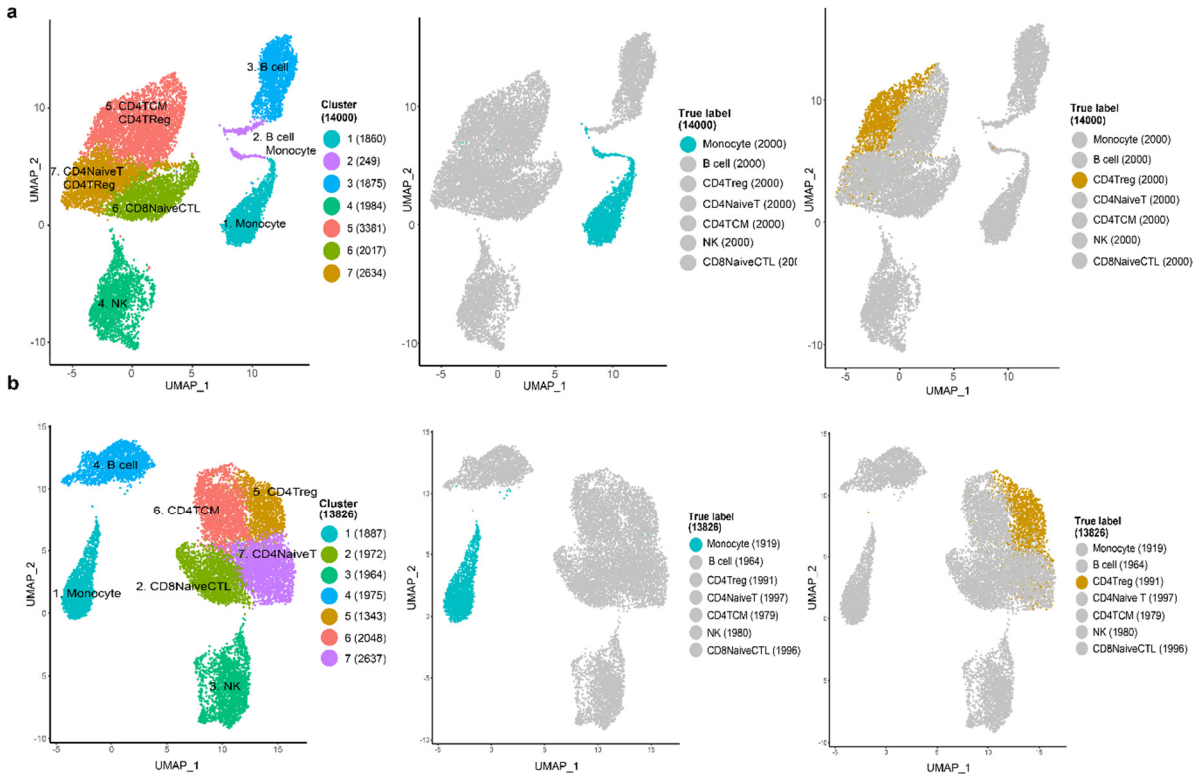


Step	Time (s)
MI-kNN	2,278
node2vec	418
clustering	60
UMAP	710
silhouette	319
total	3,785 (1.06h)

Step	Time (s)
MI-kNN	12,233
node2vec	1,553
clustering	401
UMAP	3,235
silhouette	4,005
total	21,427 (5.95h)

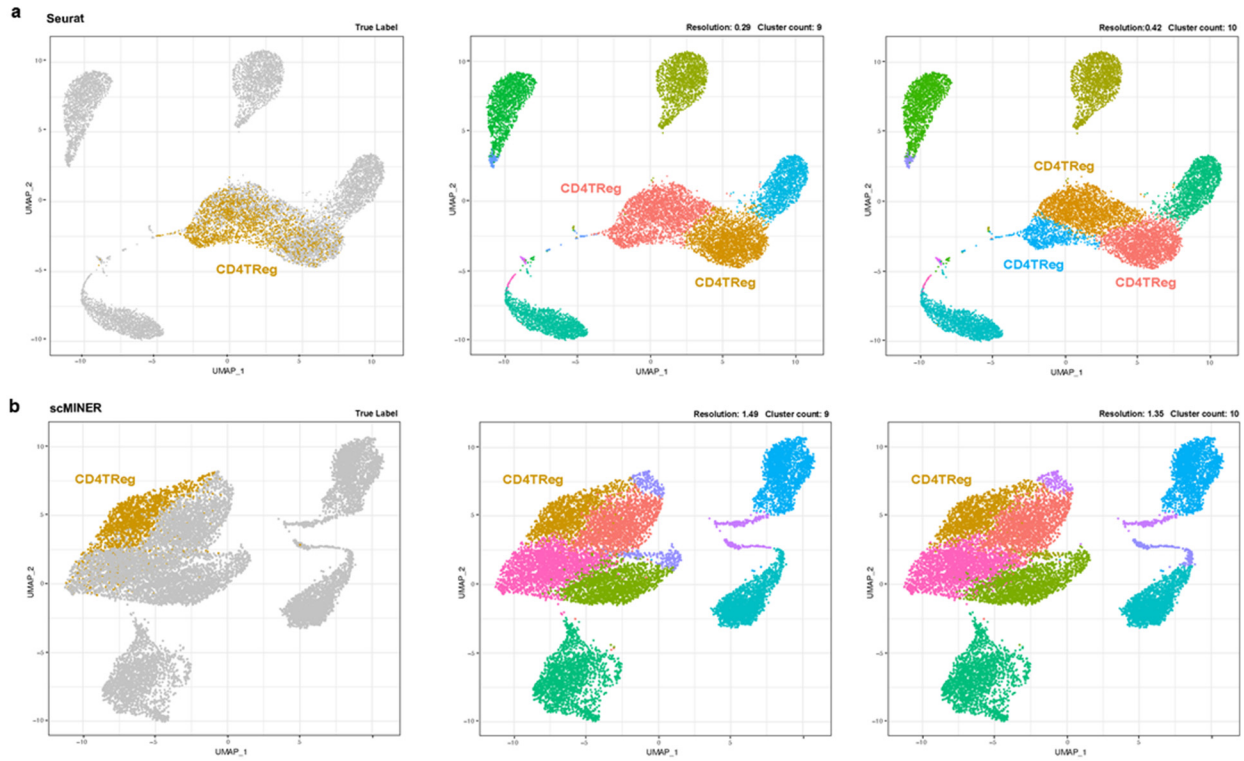
Supplementary Figure 3. MICA computing resource usage analysis for PBMC (Zheng) and Human Motor Cortex (Bakken) datasets.

a, Run time for each step of MICA for PBMC20k and human motor cortex datasets using 25 cores. **b,** ARI, run time and memory consumption for PBMC with respect to some important parameters, e.g., number of workers, number of neighbors in building MI-kNN, and node2vec window size, etc.

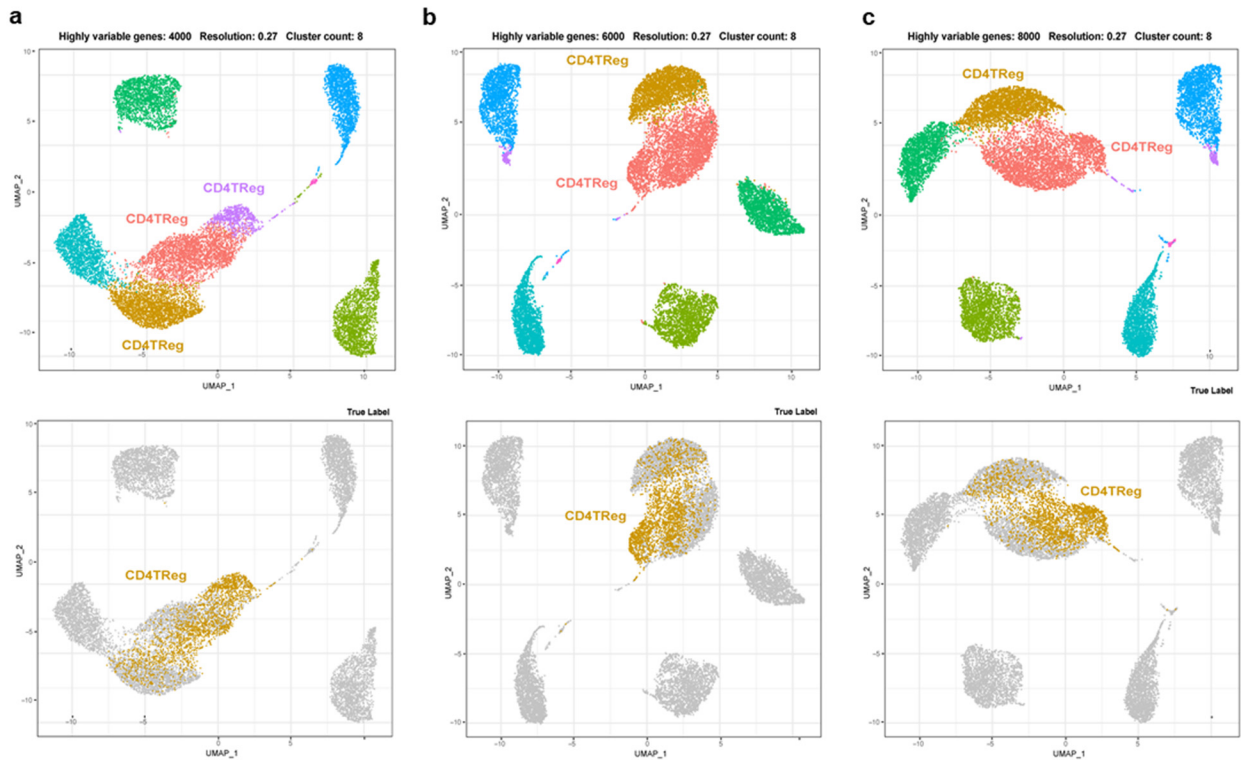


Supplementary Figure 4. Effect of CP10K and CPM normalization on the clustering result of Zheng dataset.

a, UMAP plots of all 7 clusters using count per 10K (CP10K) for normalization. **b,** UMAP plots of all 7 clusters using count per million (CPM) for normalization.

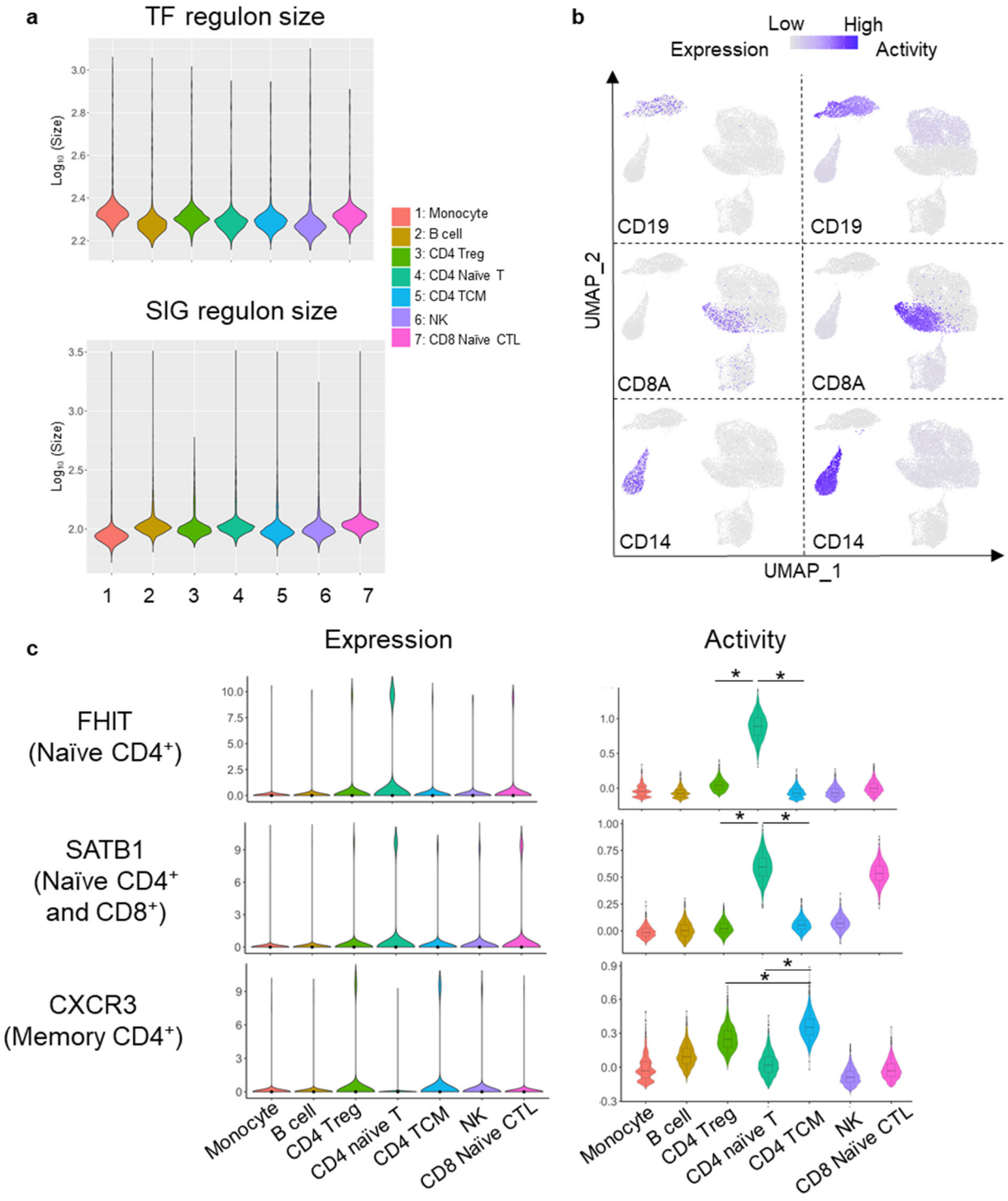


Supplementary Figure 5. Comparison of scMINER and Seurat CD4Treg cell distribution on UMAPs with respect to the changing of clustering resolution.
a, CD4Treg cell distribution on Seurat clusters with respect to the increasing number of resolution and cluster count. **b,** CD4Treg cell distribution on scMINER clusters with respect to the increasing number of resolution and cluster count.



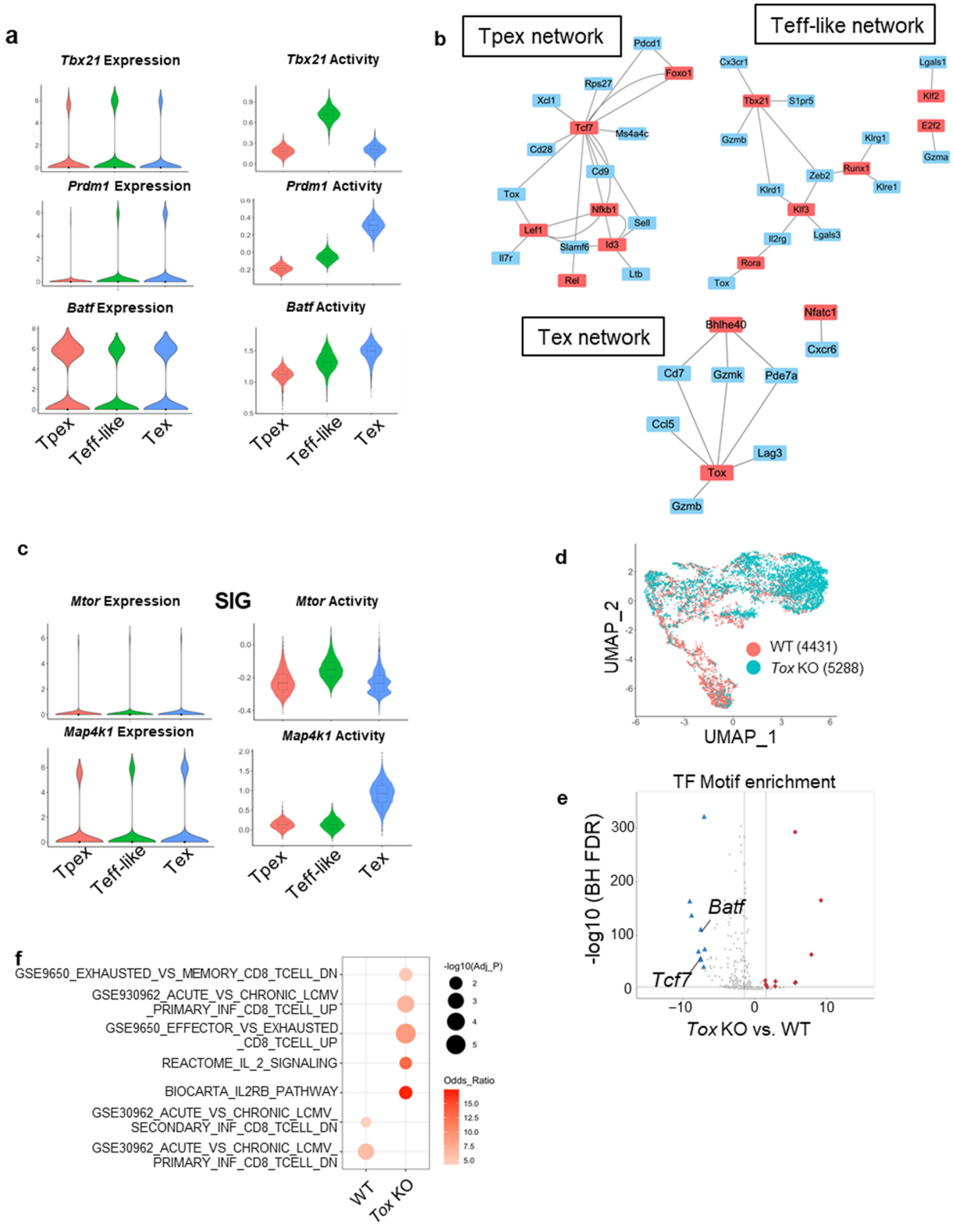
Supplementary Figure 6. CD4Treg cell distribution on Seurat clusters with respect to the changing of the number of highly variable genes.

a-c, CD4Treg cells are distributed in three Seurat clusters with 4,000 (a), 6,000 (b) and 8,000 (c) highly variable genes.



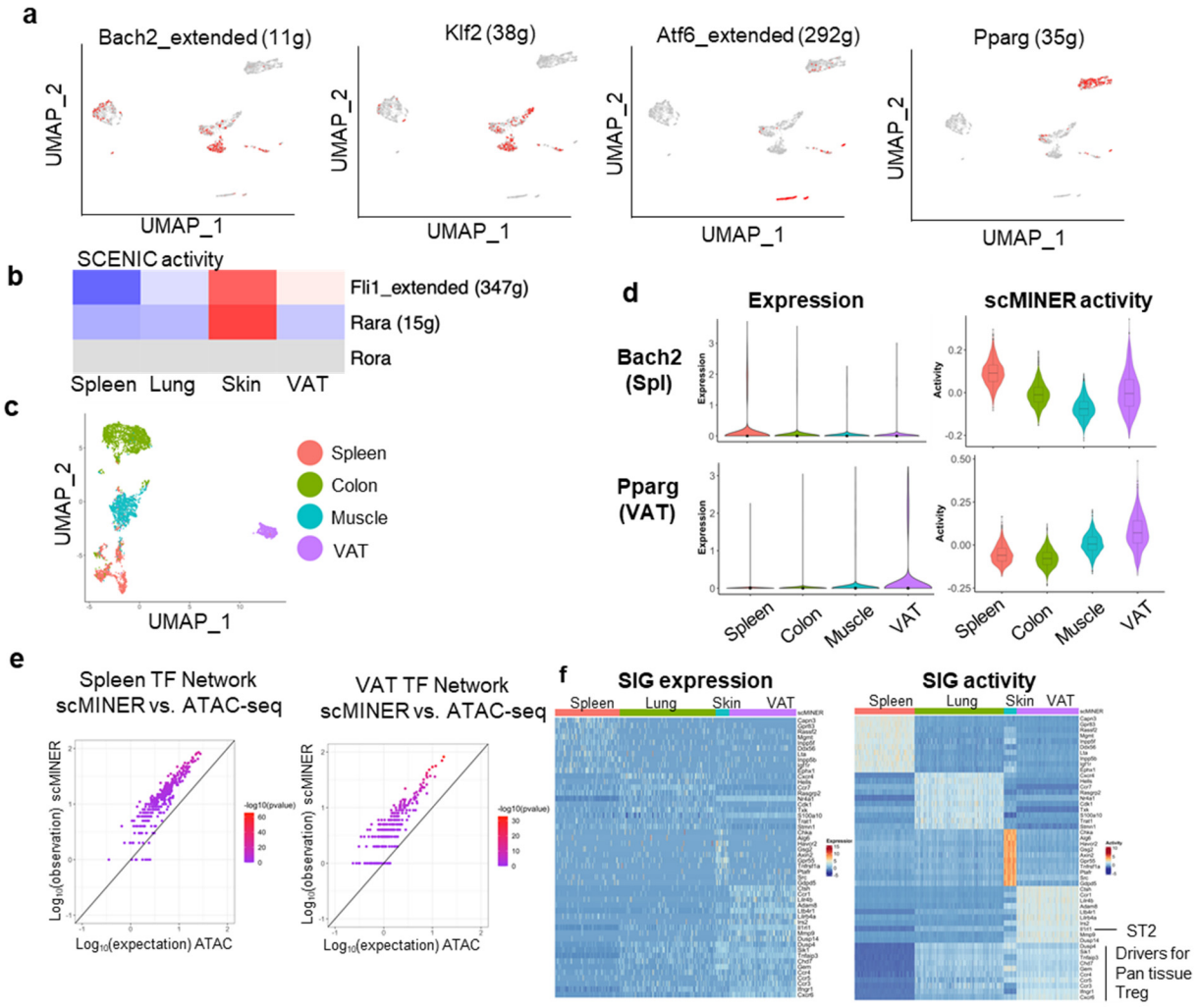
Supplementary Figure 7. scMINER measures the activity of cell type-specific markers in PBMC.

a, Mean TF and SIG regulon sizes in 7 sorted cell populations from PBMC scRNA-seq data. **b,** Expression and activity of CD19, CD8A and CD14 on UMAP using PBMC scRNA-seq data. **c,** Violin plot visualization of FHIT, SATB1 and CXCR3 expression and scMINER activity in 7 sorted cell populations from PBMC scRNA-seq data. *, $P < 2e\text{-}16$.



Supplementary Figure 8. scMINER reveals drivers in wild-type and gene-perturbed CD8$^+$ T cells during chronic infection.

a, Violin plot visualization of Tbx21, Blimp1 and Batf expression and activity in 3 subsets of $CD8^+$ T cells. **b,** GRNs for Tpex cells, Teff-like cells and Tex cells. Key TFs shown in Fig. 5b for each $CD8^+$ T cell subset are highlighted in red. **c,** Violin plot visualization of *Mtor* and *Map4k1* expression and activity in 3 subsets of $CD8^+$ T cells. **d,** UMAP visualization of wild type and *Tox* deficient $CD8^+$ T cells in chronic infection (GSE119940). The numbers in the bracket indicates the cell numbers of each genotype. **e,** TF motif enrichment analysis for *Tox* deficient vs. wild-type $CD8^+$ T cells using an ATAC-seq dataset (GSE132986). BH FDR, the Benjamini-Hochberg false discovery rate. **f,** Functional pathway enrichment of a union of top 50 TFs and top 200 SIGs predicted by scMINER for wild type and *Tox* deficient $CD8^+$ T cells.



Supplementary Figure 9. scMINER showed reproducibility in unravelling drivers in tissue specific Treg cells from different datasets.

a, UMAP visualization of SCENIC binary activity of Bach2, Klf2, Atf6 and Pparg. **b,** Heatmap of average SCENIC activity of FLI1, RARA and RORA in Treg cells from each tissue. Grey indicates that the TF activity could not be predicated by SCENINC. **c,** MICA MDS clustering of mouse Foxp3$^+$ regulatory CD4$^+$ T cells (GSE109742) isolated from spleen, colon, muscle and visceral adipose tissue (VAT). **d,** Violin plot visualization of *Bach2* and *Pparg* expression and scMINER activity in spleen, colon, muscle and VAT Treg cells from GSE109742. **e,** Similarity of TF regulon in spleen and VAT Treg cells (GSE109742) generated by SJARACNe and footprint genes detected by ATAC-seq data (GSE112731) in corresponding tissues. Expected number of genes in intersection of ATAC-seq footprints as reference (log10 scale, x axis) with regard to hypergeometric distribution vs. observed intersection (log10 scale, y axis). For all genes, the observed intersection is significantly higher than expectation (black line). The color of the dots represents the -log$_{10}$ (P-value) according to Fisher's exact test. **f,** Heatmap visualization of SIG expression in each cell clustered by mouse Foxp3$^+$ regulatory CD4$^+$ T cells isolated from

spleen, lung, skin and VAT. Drivers for Pan tissue Treg, drivers that have higher activity in Treg cells from the lung, skin and VAT than from spleen.

Supplementary Table 1. Summary of 11 single-cell datasets used for the evaluation of clustering methods.

Dataset	Protocol	Size	Class	Taxonomy	Tissue	Accession ID
Yan (2013)[1]	Tang	124	8	Human	Embryonic stem	GSE36552
Goolam (2016)[2]	Smart-Seq2	124	5	Mouse	Development	E-MTAB-3321
Buettner	C1	182	3	Mouse	Embryonic stem	E-MTAB-2805
Pollen (2014)[4]	SMARTer	301	11	Human	Cerebral cortex	SRP041736
Chung (2017)[5]	SMARTer	515	5	Human	Breast cancer	GSE75688
Usoskin (2015)[6]	STRT-seq	622	4	Mouse	Sensory neurons	GSE59739
Kolod (2015)[7]	SMARTer	704	3	Mouse	Embryonic stem	E-MTAB-2600
Klein (2015)[8]	inDrop	2,717	4	Mouse	Embryonic Stem	GSE65525
Zeisel (2015)[9]	STRT-seq	3,005	7	Mouse	Cortex, hippocampus	GSE60361
Zheng (2017)[10]	10x Genomics	20,000	10	Human	Sorted peripheral	SRP073767
Bakken (2020)[11]	10x Genomics	76,533	20	Human	Motor cortex	Azimuth

Supplementary Table 2. Summary of scRNA-seq and ATAC-seq datasets used for scMINER applications.

Accession ID	Data type	Cell types	Protocol
GSE122712	scRNA-seq	$CD8^+$ T cells from chronic infection[12]	10x Genomics
GSE130879	scRNA-seq	Tissue (spleen, lung, skin, and VAT) Treg cells[13]	10x Genomics
GSE130879	scRNA-seq	Tissue Treg precursors[13]	10x Genomics
GSE109742	scRNA-seq	Tissue (spleen, colon, muscle, and VAT) Treg cells[14]	InDrop
GSE119940	scRNA-seq	$CD8^+$ T cells from WT and Tox KO in chronic infection (day 7)[15]	10x Genomics
GSE123236	ATAC-seq	Tpex and Tex in LCMV infection[12]	Bulk
GSE132986	ATAC-seq	WT and Tox KO $CD8^+$ T cells[16]	Bulk
GSE112731	ATAC-seq	Tissue (spleen and VAT) Treg cells[14]	Bulk
GSE156112	scATAC-seq	Tissue (spleen, lung, skin, and VAT) Treg cells[17]	10x Genomics

Supplementary Note: Comprehensive scMINER documentation and tutorial with examples is publicly accessible via https://jyyulab.github.io/scMINER.

References

1. Yan, L. et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology* **20**, 1131-1139 (2013).
2. Goolam, M. et al. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165**, 61-74 (2016).
3. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155-160 (2015).
4. Pollen, A.A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology* **32**, 1053-1058 (2014).
5. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications* **8**, 15081 (2017).
6. Usoskin, D. et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience* **18**, 145-153 (2015).
7. Kolodziejczyk, Aleksandra A. et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell stem cell* **17**, 471-485 (2015).
8. Klein, Allon M. et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187-1201 (2015).
9. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).
10. Zheng, G.X.Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
11. Bakken, T.E. et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119 (2021).
12. Miller, B.C. et al. Subsets of exhausted $CD8^+$ T cells differentially mediate tumor control and respond to checkpoint blockade. *Nat Immunol* **20**, 326-336 (2019).
13. Delacher, M. et al. Precursors for Nonlymphoid-Tissue Treg Cells Reside in Secondary Lymphoid Organs and Are Programmed by the Transcription Factor BATF. *Immunity* **52**, 295-312 e211 (2020).
14. DiSpirito, J.R. et al. Molecular diversification of regulatory T cells in nonlymphoid tissues. *Sci Immunol* **3** (2018).
15. Yao, C. et al. Single-cell RNA-seq reveals TOX as a key regulator of $CD8^+$ T cell persistence in chronic infection. *Nat Immunol* **20**, 890-901 (2019).
16. Khan, O. et al. TOX transcriptionally and epigenetically programs $CD8^+$ T cell exhaustion. *Nature* **571**, 211-218 (2019).
17. Delacher, M. et al. Single-cell chromatin accessibility landscape identifies tissue repair program in human regulatory T cells. *Immunity* **54**, 702-720 e717 (2021).