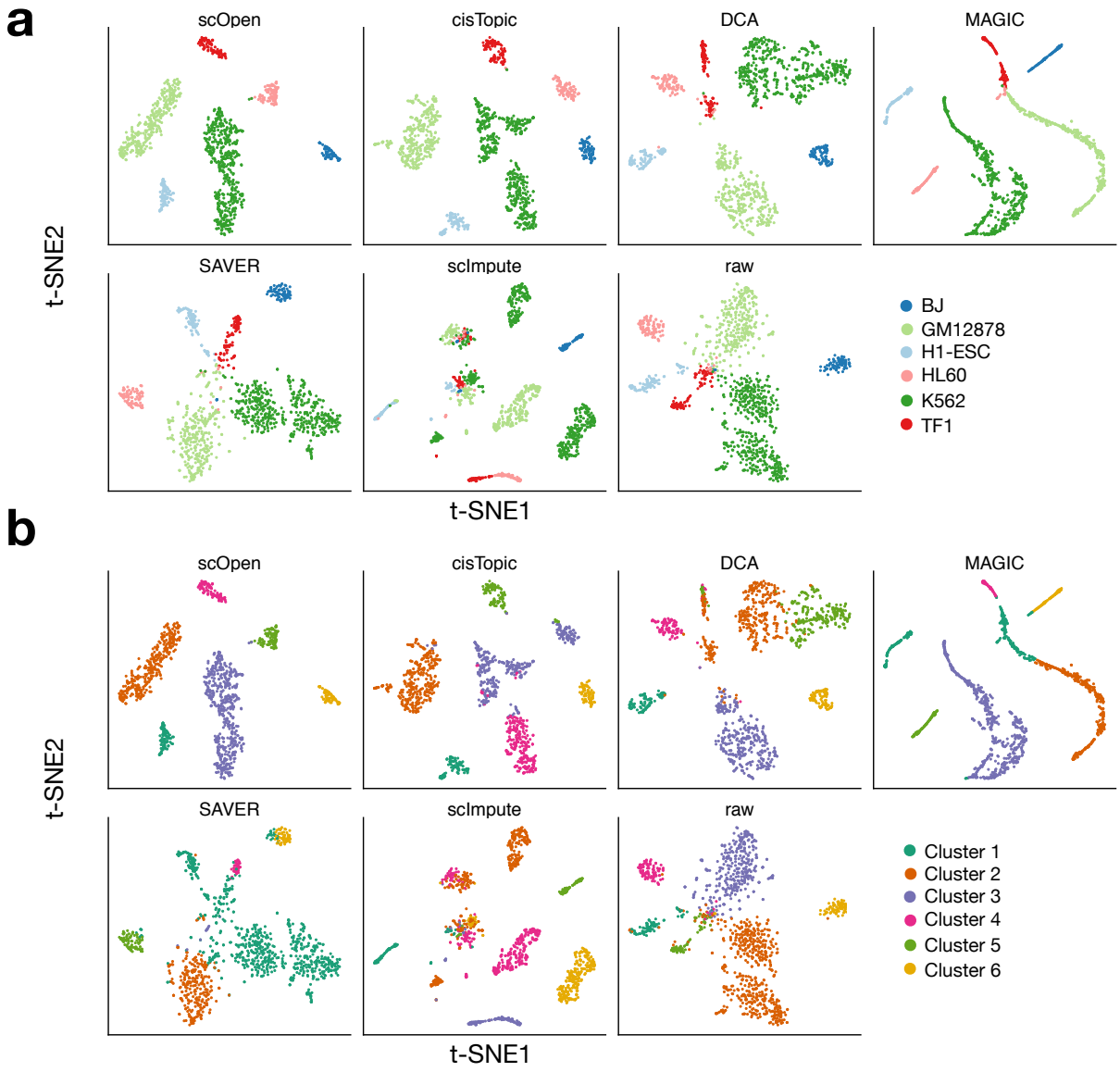
scOpen: chromatin-accessibility estimation of single-cell ATAC data

Zhijian Li[1,+], Christoph Kuppe[2,+], Mingbo Cheng[1], Sylvia Menzel[2], Martin Zenke[3,4], Rafael Kramann[2,*], and Ivan G. Costa[1,*]

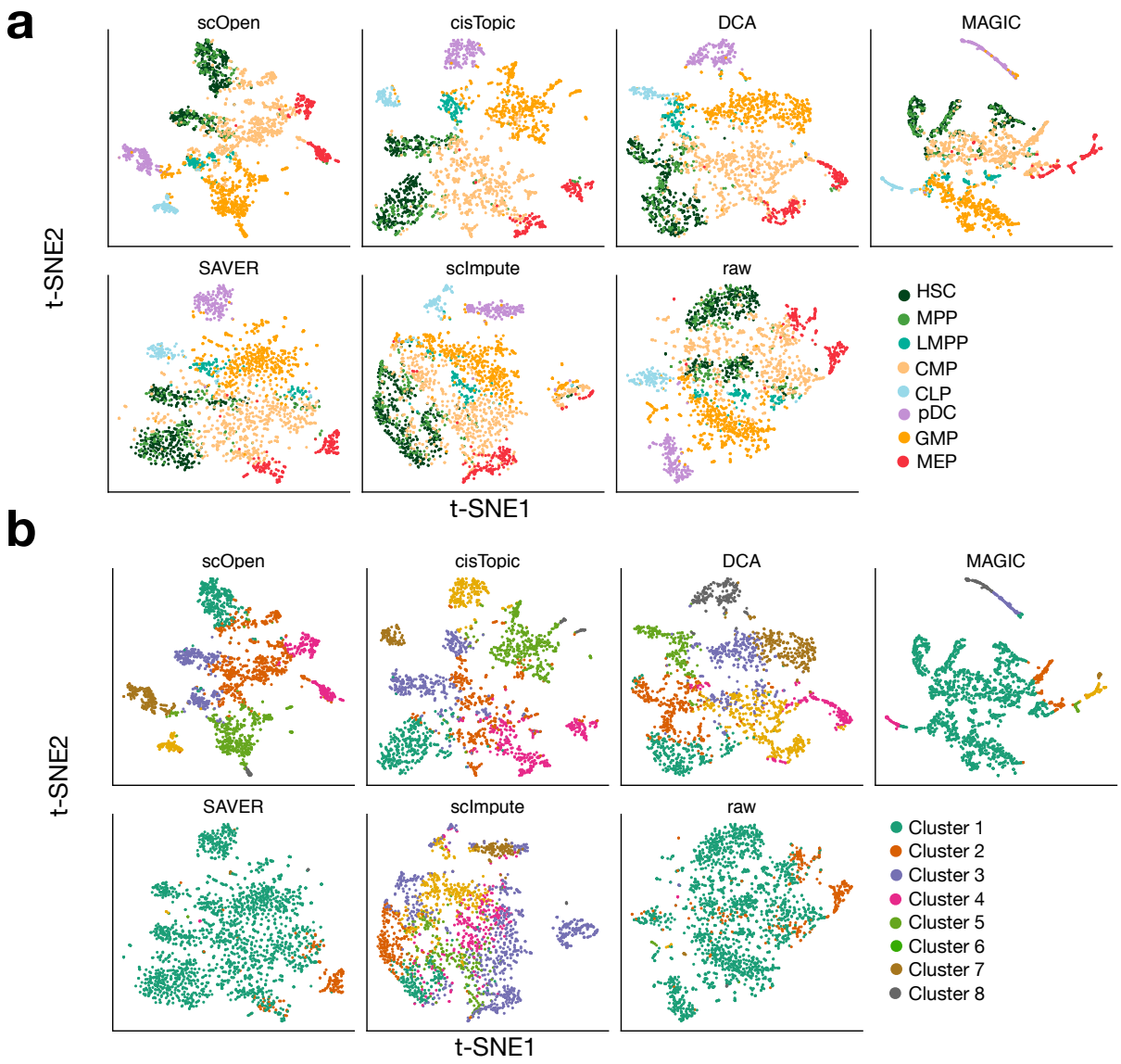
[1]Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, 52074 Aachen, Germany
[2]Division of Nephrology and Clinical Immunology, RWTH Aachen University, 52074 Aachen, Germany
[3]Department of Cell Biology, Institute of Biomedical Engineering, RWTH Aachen University Medical School, Aachen, 52074, Germany
[4]Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Aachen, Germany
[*]corresponding authors: rkramann@ukaachen.de, ivan.costa@rwth-aachen.de
[+]these authors contributed equally to this work

December 5, 2019

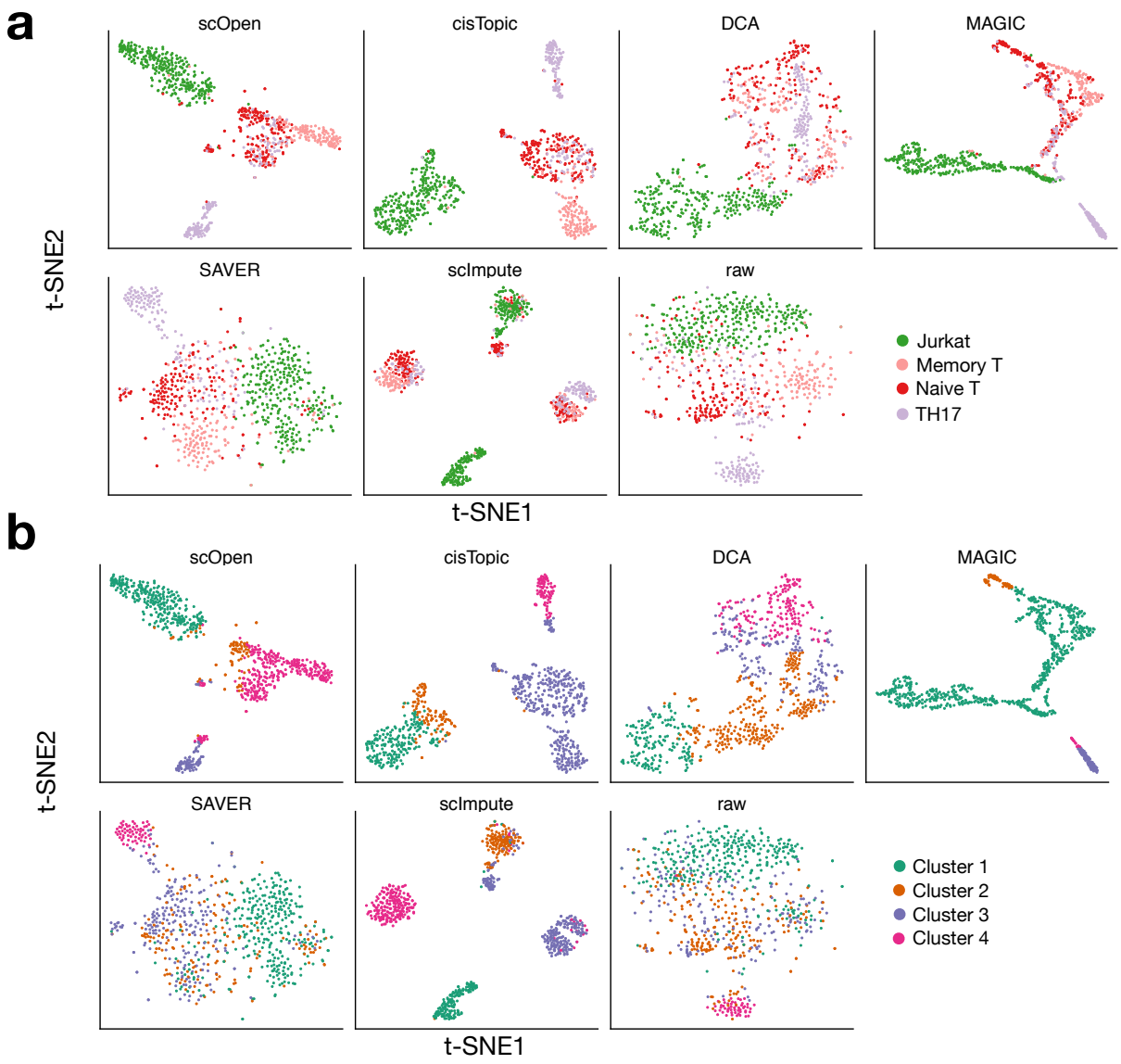
Supplementary Figures



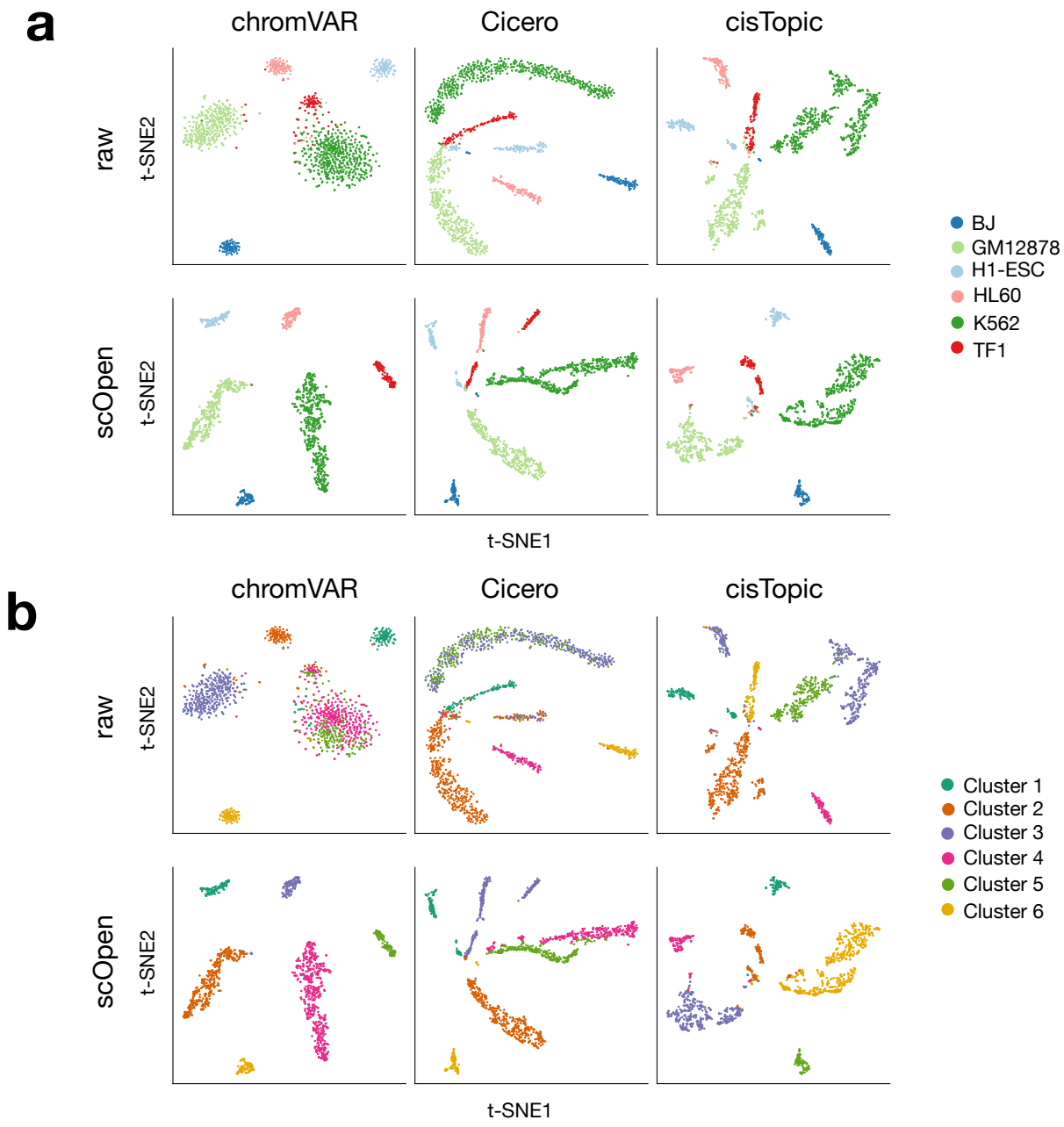
Supplementary Fig. 1. Imputation methods on cell lines dataset. t-SNE visualisation of scOpen, cisTopic, DCA, MAGIC, SAVER, scImpute and the raw data for cell lines with true class labels (**a**) and k-medoids clustering results (**b**).



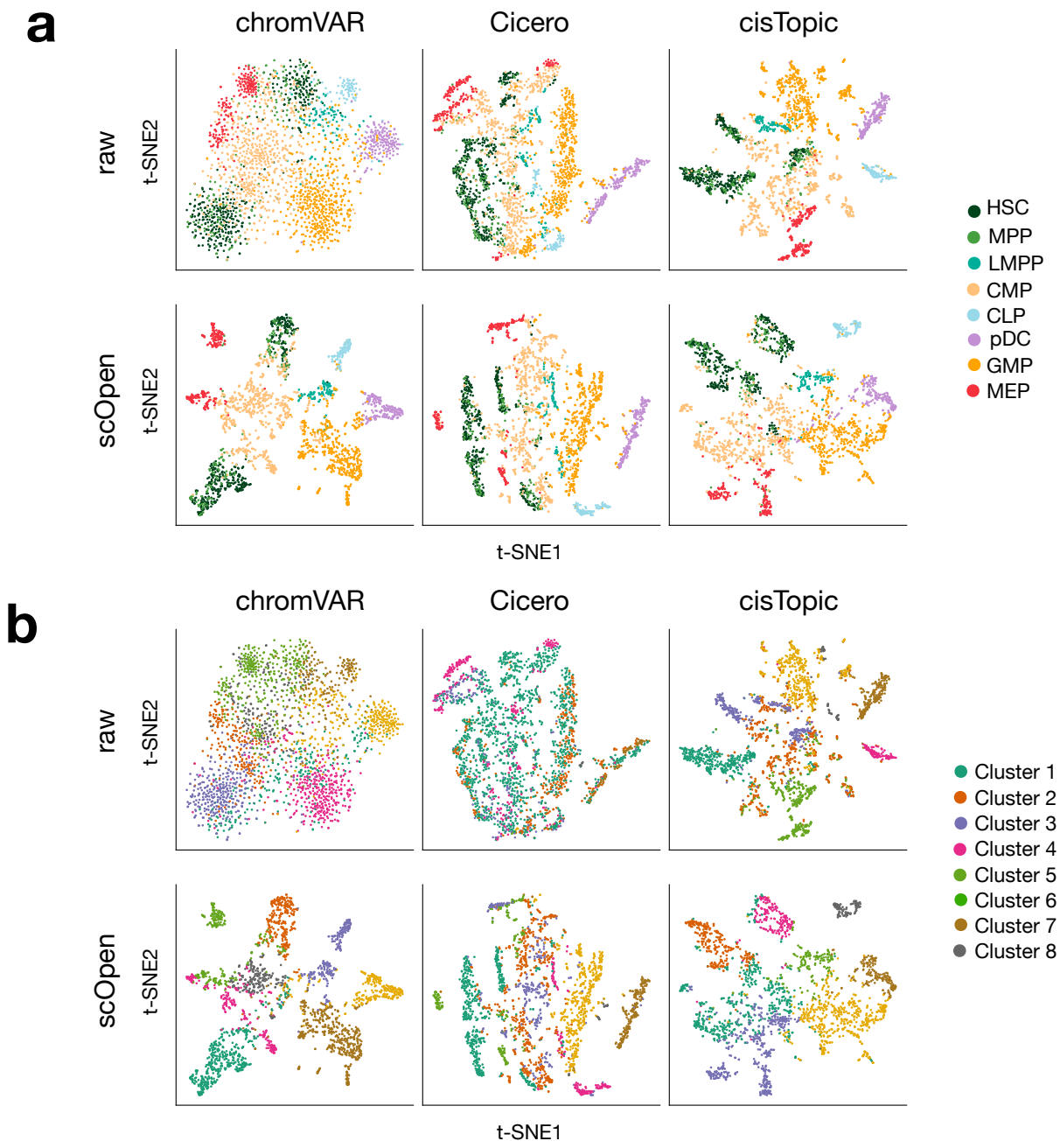
Supplementary Fig. 2. Imputation methods on hematopoiesis dataset. t-SNE visualisation of scOpen, cisTopic, DCA, MAGIC, SAVER, scImpute and the raw data for hematopoiesis with true class labels (**a**) and k-medoids clustering results (**b**).



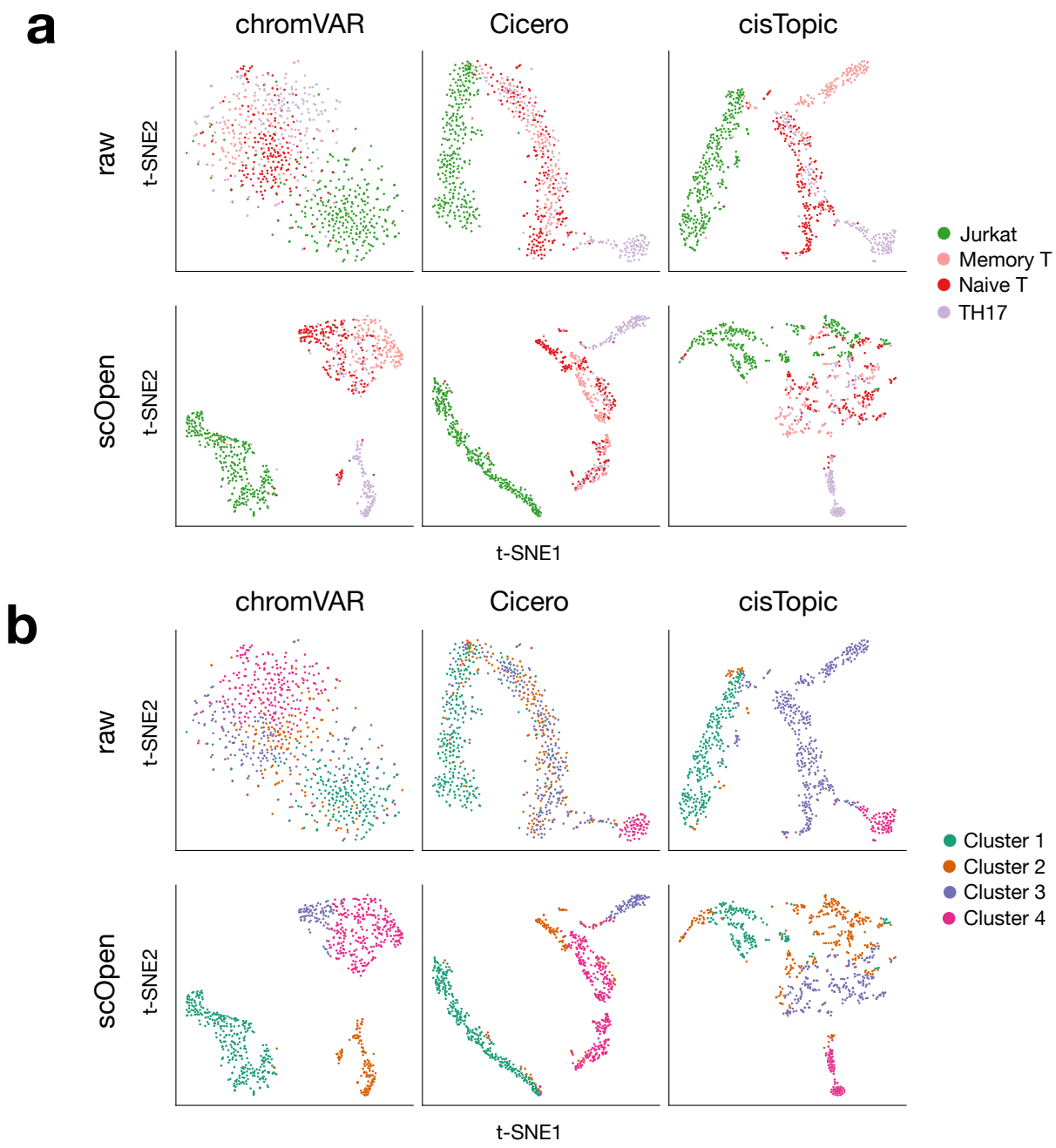
Supplementary Fig. 3. Imputation methods on T cells dataset. t-SNE visualisation of scOpen, cisTopic, DCA, MAGIC, SAVER, scImpute and the raw data for T cells with true class labels (**a**) and k-medoids clustering results (**b**).



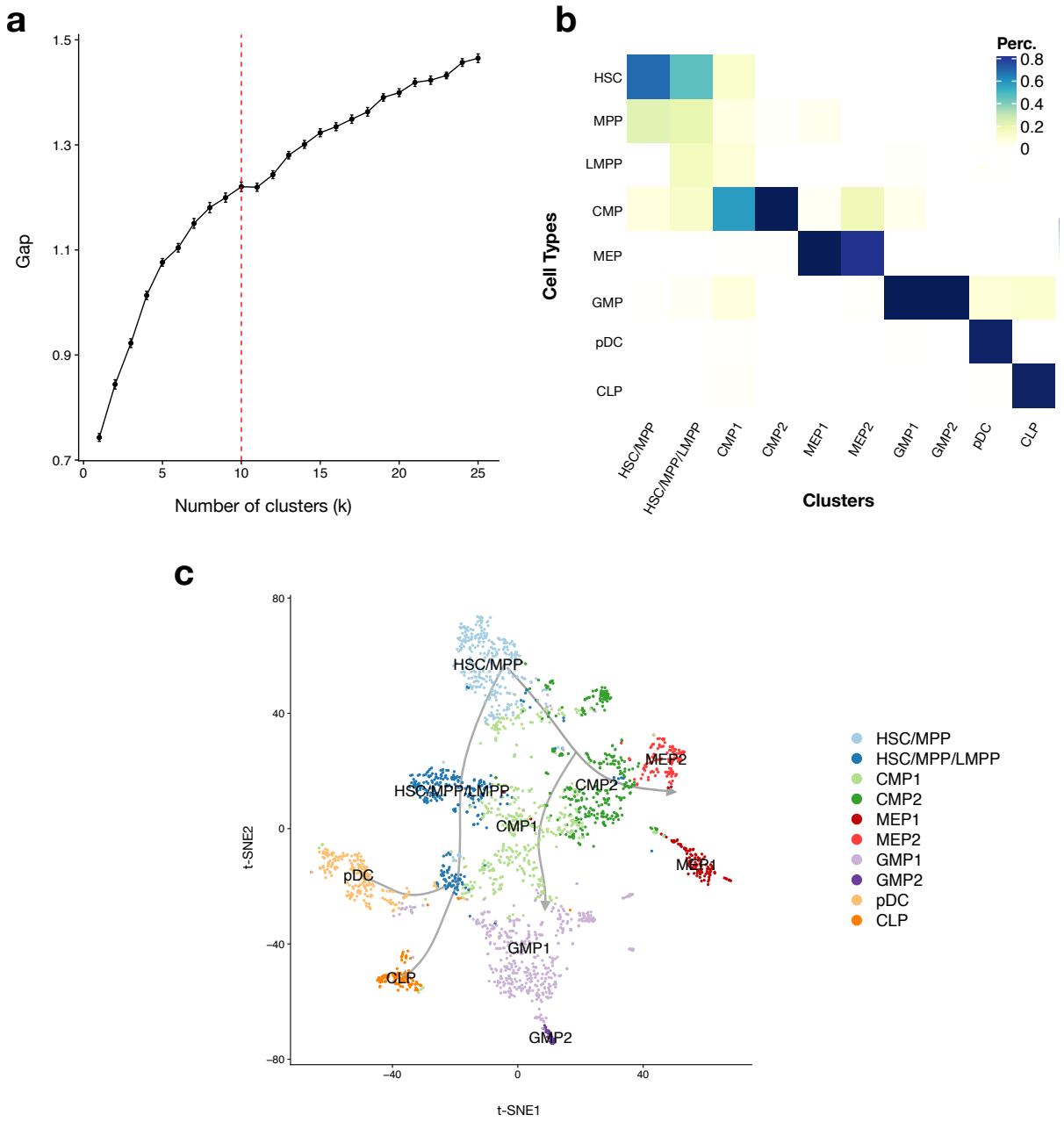
Supplementary Fig. 4. scATAC-seq methods on cell lines dataset. t-SNE visualisation of chromVAR, Cicero and cisTopic using the raw (top) and scOpen (bottom) estimated matrices for cell lines with true class labels (**a**) and clustering results (**b**).



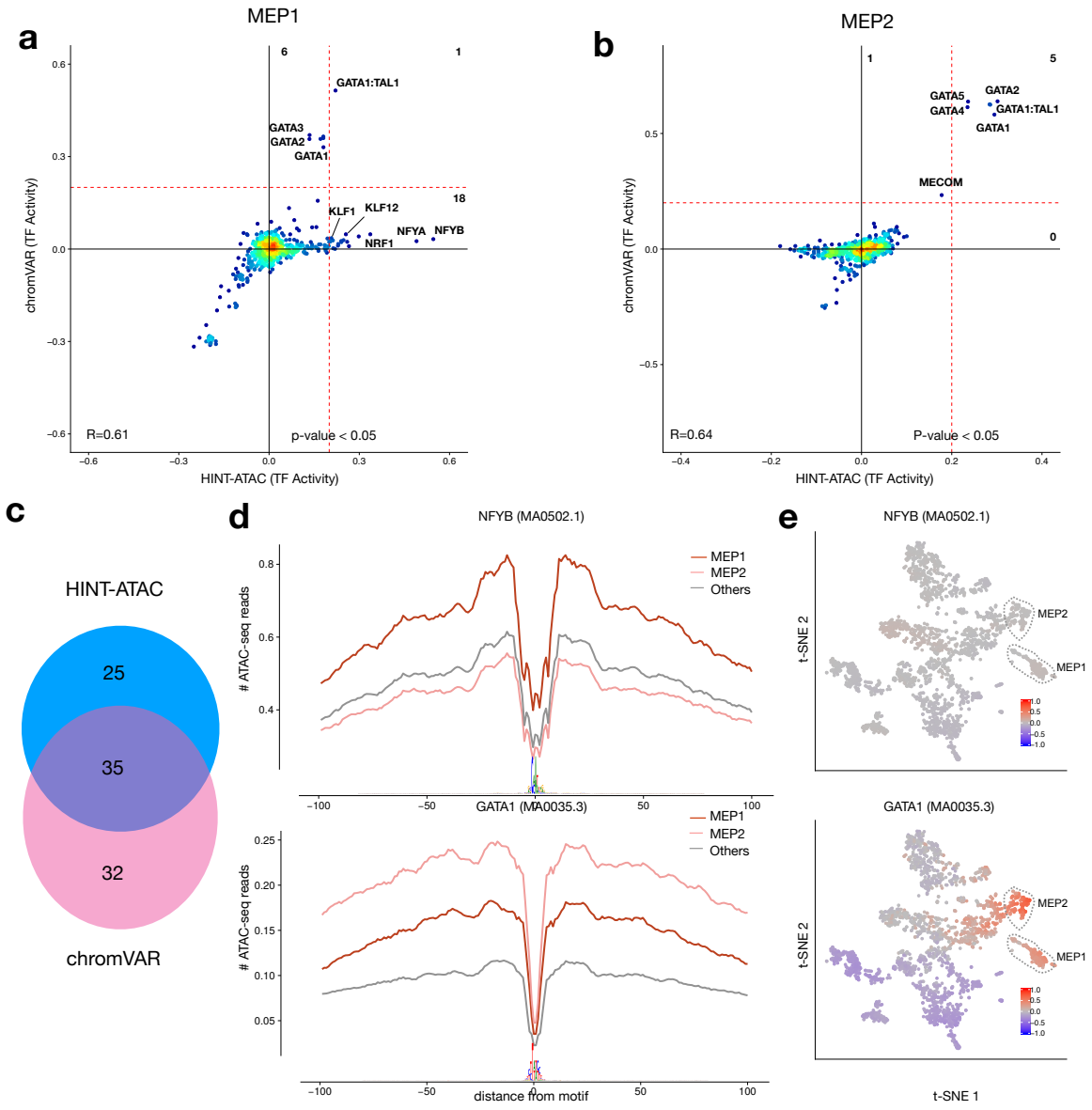
Supplementary Fig. 5. scATAC-seq methods on hematopoiesis dataset. t-SNE visualisation of chromVAR, Cicero and cisTopic using the raw (top) and scOpen (bottom) estimated matrices for hematopoiesis with true class labels (**a**) and clustering results (**b**).



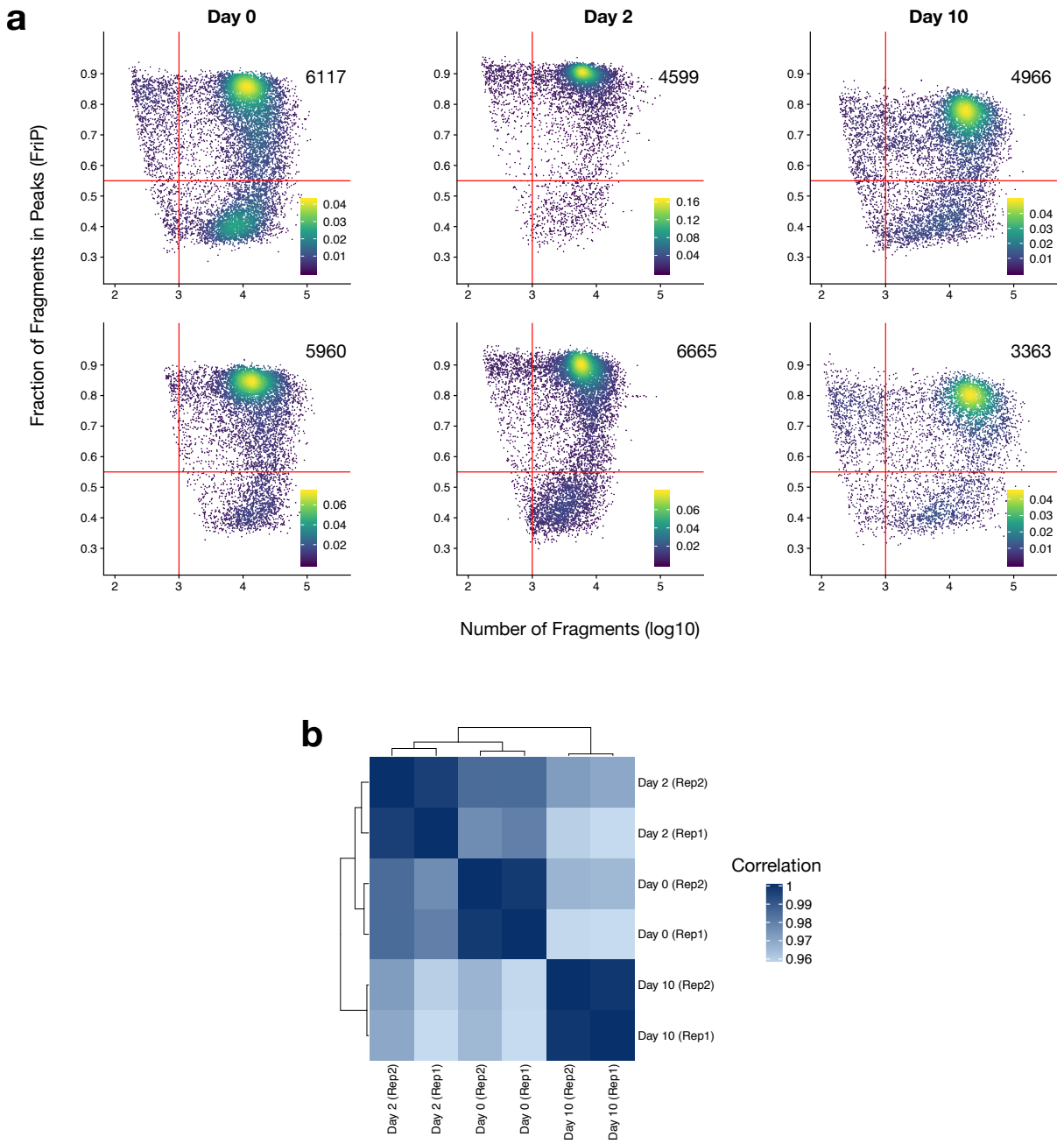
Supplementary Fig. 6. scATAC-seq methods on T cells dataset. t-SNE visualisation of chromVAR, Cicero and cisTopic using the raw (top) and scOpen (bottom) estimated matrices for T cells with true class labels (**a**) and clustering results (**b**).



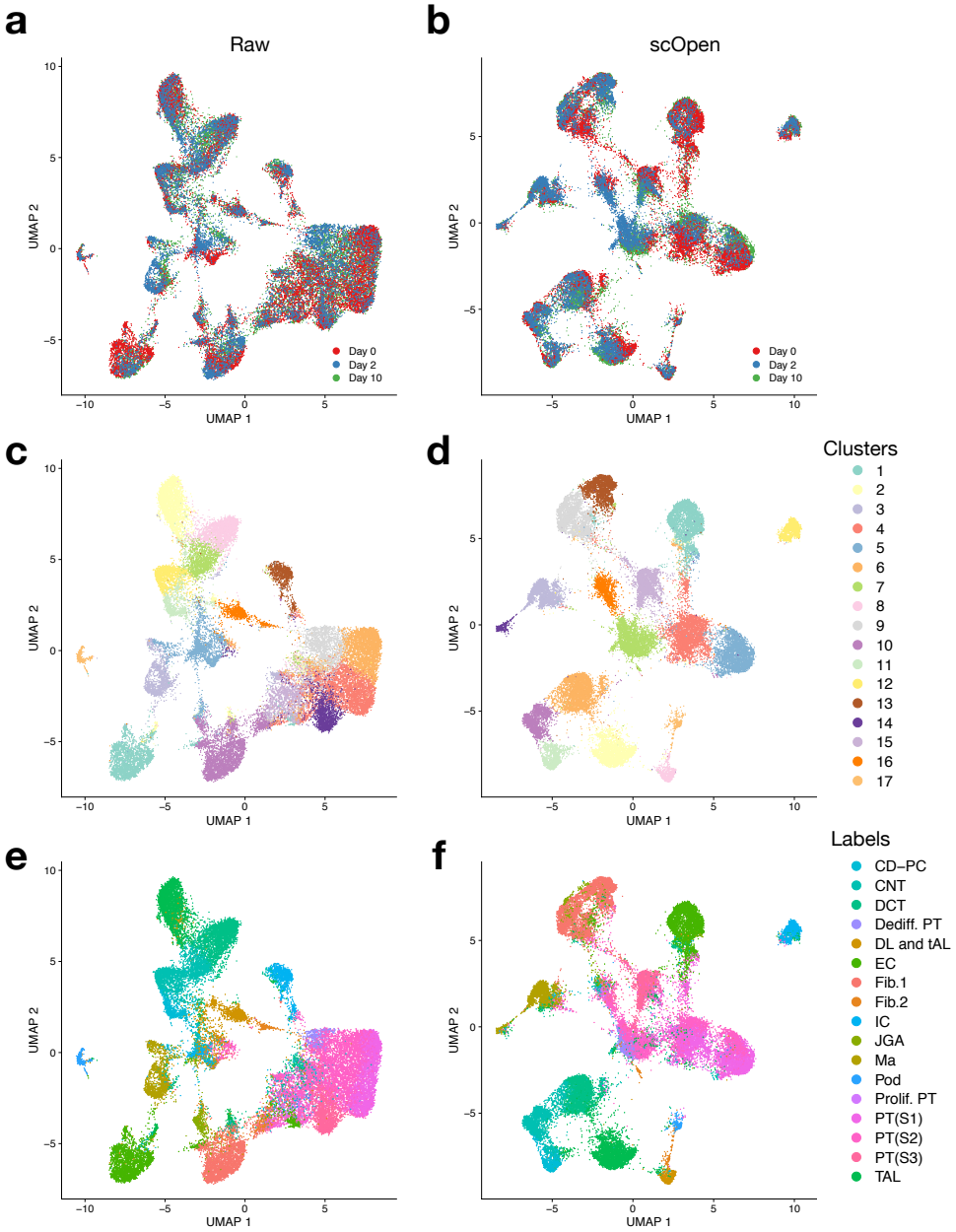
Supplementary Fig. 7. Unsupervised clustering of hematopoiesis dataset. **a,** Line plot showing the gap statistic results of different number of clusters (k). Error bars represent the standard deviation (SD) for B = 10 simulated reference data sets. Following the strategy from [2], we select the smallest k such that $k > k+1-\mathrm{SD}(\mathrm{k}+1)$, which is $k = 10$. **b,** Confusion matrix with percentage of cells regarding true labels (y-axis) for each of the 10 detected clusters (x-axis). Cluster names are based on most frequent cell types in a cluster. **c,** t-SNE projection of hematopoiesis dataset with cluster labels. Most of the detected sub-clusters, i.e. GMP1 and GMP2 or CMP1 and CMP2, are localised within differentiation paths (grey lines), which supports that they consist of cells at distinct differentiation stages as previously shown in [1]. A clear exception is sub-groups of MEP1 and MEP2 cells, which appears to form two previously uncharacterised differentiation branches.



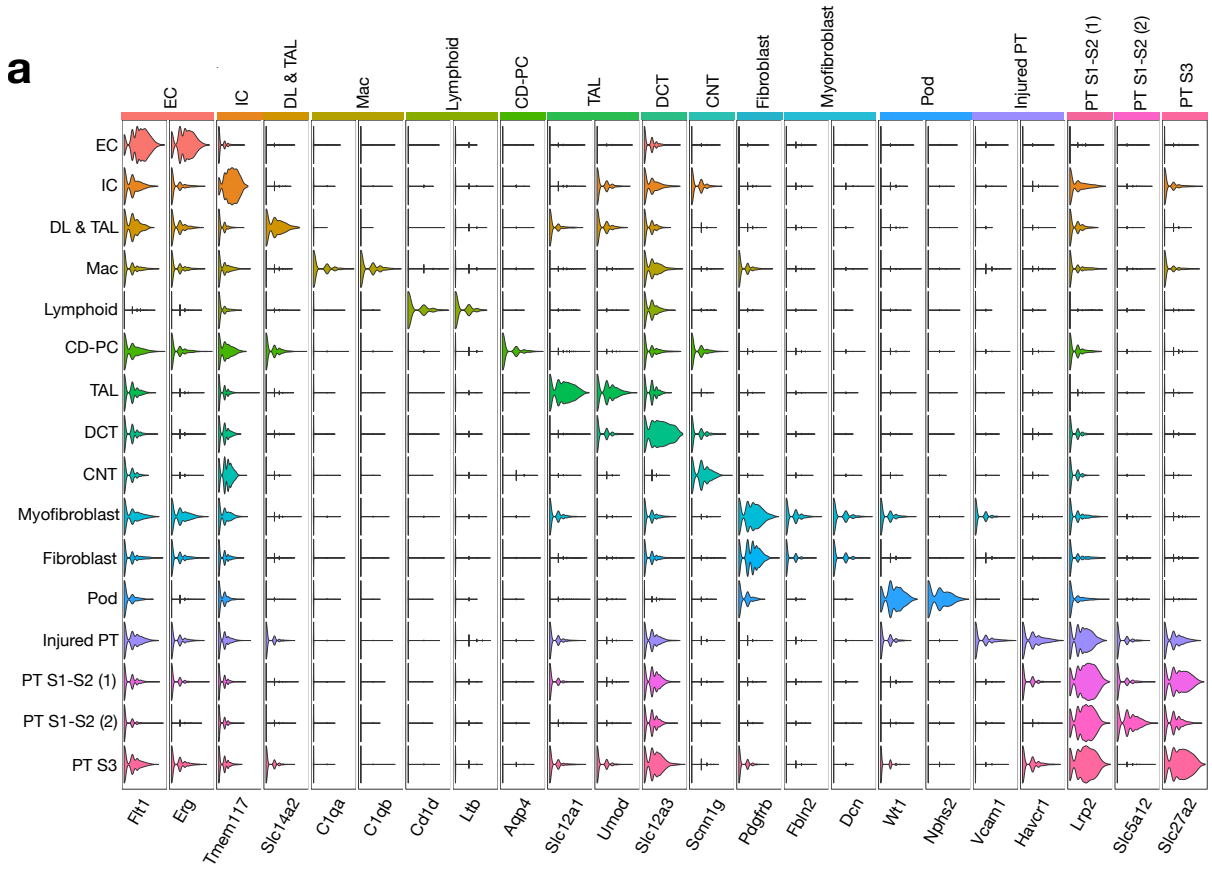
Supplementary Fig. 8. Analysis of transcription factors describing MEP1 and MEP2 clusters. a-b, Scatter plots indicating the transcription factor activity scores from chromVAR (y-axis) and HINT-ATAC (x-axis) for MEP1 and MEP2 cells. High positive values indicate factors with specific activity in these clusters. We focus here on factors with differential activity > 0.2 (p-value < 0.05) regarding differential activity in HINT-ATAC. Both HINT-ATAC and chromVAR indicate GATA motifs as relevant in both clusters. Interestingly, only HINT-ATAC detects particular TFs with MEP1 specific activity including KLF1 and NFY family factors. **c,** Number of TFs with differential activity > 0.2 from HINT-ATAC and chromVAR in all 10 hematopoietic clusters. **d,** Line plots showing average ATAC-seq read profiles around footprints supported by motifs with high activity scores in MEP1 (NFYB) and MEP2 (GATA1) cells. **e,** Cell specific chromVAR scores for NFYB and GATA1 motifs.



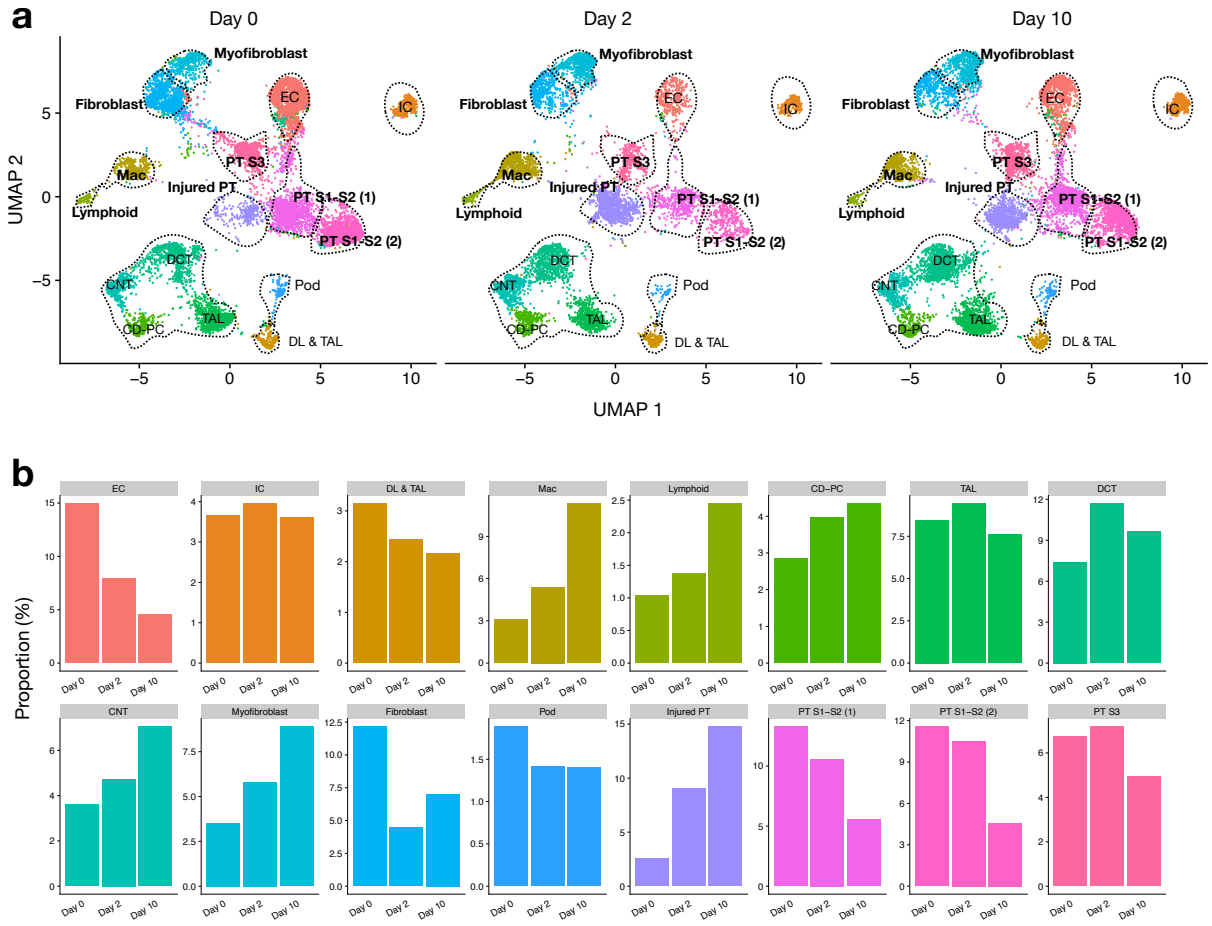
Supplementary Fig. 9. Cell filtering of UUO scATAC-seq data. Scatter plots showing number of unique fragments (reads) (x-axis) versus fraction of fragments (reads) in peaks (FRIP) (y-axis) from the aggregate data of day 0, 2 and 10 for replicate 1 (top) and 2 (bottom). Each dot represents a cell identified by cellranger and red lines represent cut-off used for final cell detection. Number on right-upper corner represents the number of cells that pass filtering. **b,** Heatmap with the correlation between sequencing libraries. Replicates are highly correlated to each other.



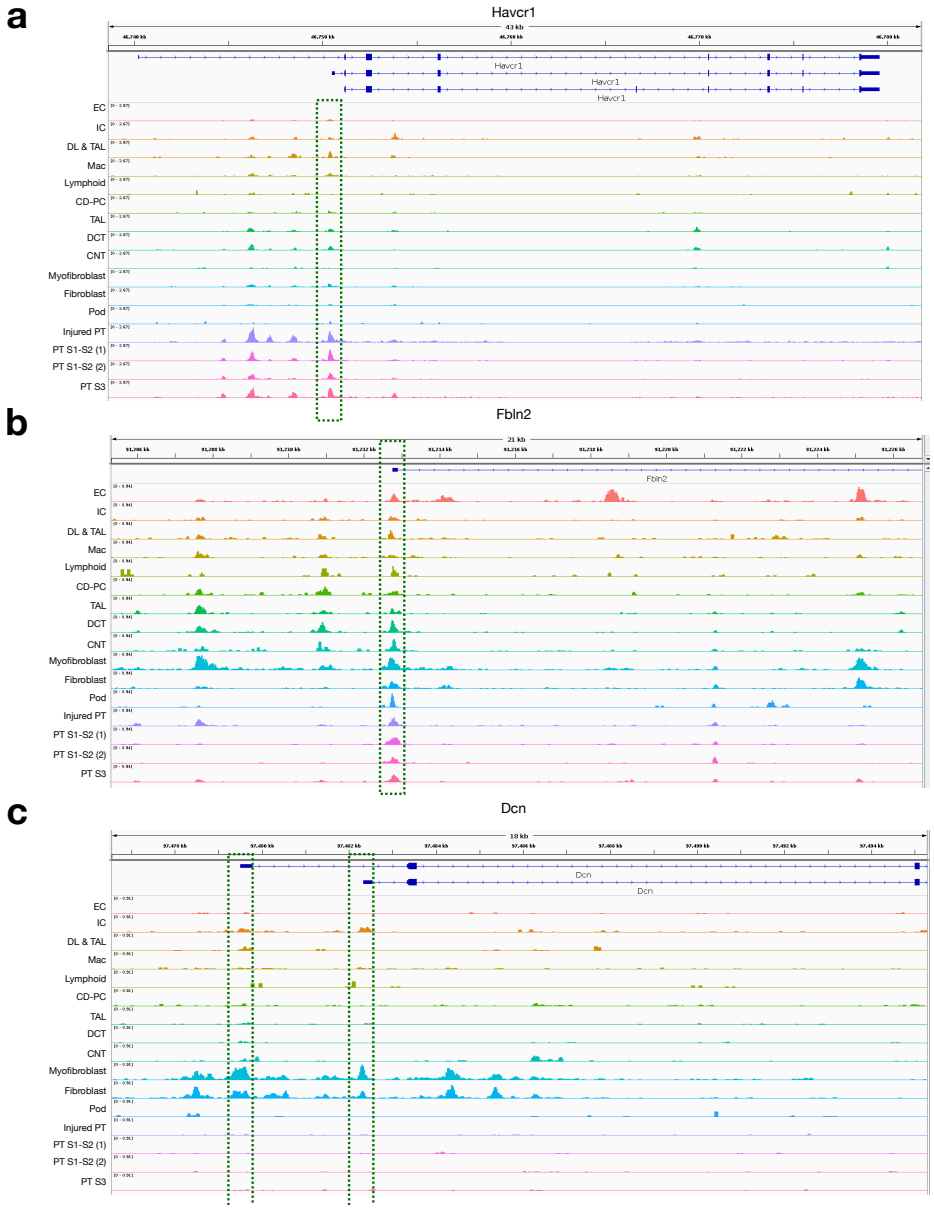
Supplementary Fig. 10. UUO scATAC-seq data integration. a-b, UMAP visualisation of the integrated matrix using either raw matrix or scOpen estimated matrix as input. **c-d,** Clustering results of the integrate data. **e-f,** Predicted cell types using label transfer approach from Seurat.



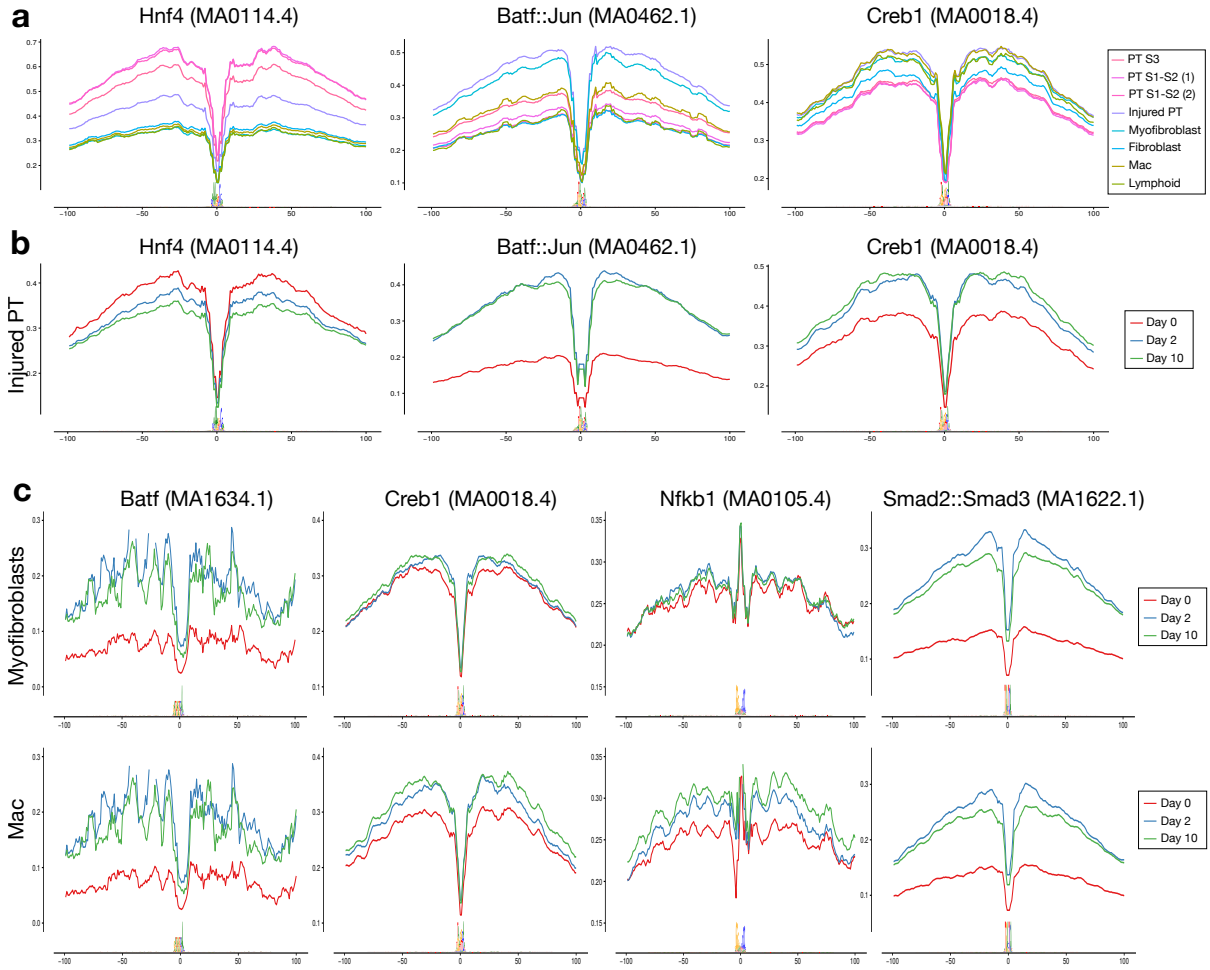
Supplementary Fig. 11. UUO cluster-specific marker genes. a, Violin plot showing cluster-specific (y-axis) gene accessibility score associated to known marker genes (x-axis).



Supplementary Fig. 12. UUO per day. a, Scatter plots showing condition-specific UMAP visualization of UUO scATAC-seq data for each day, associated to **Fig. 3b**. **b,** Bar plots showing proportion of each cell type across different days, associated to **Fig. 3c**.



Supplementary Fig. 13. Chromatin accessibility around marker genes. IGV screenshot showing increased accessibility around the injury marker Havcr1 (**a**), which is expanded in injured PTs in contrast to other PT clusters; and myofibroblasts markers Fbln2 (**b**) and Dcn (**c**), which have higher accessibility in myofibroblast cluster in comparison to fibroblasts.



Supplementary Fig. 14. ATAC-seq profiles around footprints. a, Transcription factor footprints for Hnf4, Batf::Jun and Creb1 factors for selected clusters. **b,** Transcription factor footprints for Hnf4, Batf::Jun and Creb1 factors for injured PT cells in Day 0, Day 2 and Day 10. **c,** Transcription factor footprints for Batf, Creb1, Nfkb1 and Smad2::Smad3 factors for Myofibroblast cells and macrophages in Day 0, Day 2 and Day 10.

Supplementary Tab. 1. Statistics of data sets used in this study. The number of detected cells, number of regions (peaks), fraction of non-zero entries, average number of reads per cell, fraction of reads in peaks (FRIP) and total number of valid reads are shown below. For comparison purposes, we also included similar statistics for a scRNA-seq data corresponding to the Hematopoiesis scATAC-seq data. We observed a higher sparsity in scATAC-seq (0.039 vs 0.119) despite the fact scATAC-seq data has 7 times more reads per cell than scRNA-seq.

Dataset	Type	Cells	Regions/Genes	Non-zeros	Reads per cell	FRIP	Reads
Cell lines	scATAC-seq	1,224	125,647	0.036	41,467,80	0.248	50,756,587
T cells	scATAC-seq	765	49,344	0,033	14,963,39	0.418	11,446,993
Hematopoiesis	scATAC-seq	2,210	109,418	0.039	34.656,15	0.272	76,590,091
Hematopoiesis	scRNA-seq	14,432	12,558	0,119	5.209,45	NA	75,182,840
UUO	scATAC-seq	31,670	252,146	0.032	14,752.04	0.789	467,197,173

References

[1] J. D. Buenrostro, M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang, and W. J. Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548, 2018.

[2] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.