nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-00979-3



# MARS: discovering novel cell types across heterogeneous single-cell experiments

Maria Brbić[1], Marinka Zitnik[2], Sheng Wang[3], Angela O. Pisco[4], Russ B. Altman[3,4], Spyros Darmanis[4] and Jure Leskovec[1,4] ✉

**Although tremendous effort has been put into cell-type annotation, identification of previously uncharacterized cell types in heterogeneous single-cell RNA-seq data remains a challenge. Here we present MARS, a meta-learning approach for identifying and annotating known as well as new cell types. MARS overcomes the heterogeneity of cell types by transferring latent cell representations across multiple datasets. MARS uses deep learning to learn a cell embedding function as well as a set of landmarks in the cell embedding space. The method has a unique ability to discover cell types that have never been seen before and annotate experiments that are as yet unannotated. We apply MARS to a large mouse cell atlas and show its ability to accurately identify cell types, even when it has never seen them before. Further, MARS automatically generates interpretable names for new cell types by probabilistically defining a cell type in the embedding space.**

High-throughput single-cell transcriptional profiling has enabled remarkable progress in our understanding of cellular mechanisms of disease and development[1–4]. Cell atlas datasets, including the Mouse Cell Atlas[5,6] and Human Cell Atlas[7], systematically measure the transcriptome of individual cells in multiple sites in the organism and at several time points during growth and development. These datasets have contributed to the discovery of new cell types and cell transcriptional states[8–11]. However, to assist with the identification of new cell types, there is currently a big gap as this requires techniques that (1) harmonize heterogeneous and time-varying datasets, (2) learn dataset-invariant cell representations and (3) use the learned representations to decide whether groups of measured cells represent previously uncharacterized cell types and cell states. Such techniques would have the power to reveal new cell types, enable investigation of biology that underlies those cell types and their cellular activity, and would thus form a crucial tool in an expanding single-cell computational toolbox.

Existing single-cell tools train deep neural network models to learn how to embed cells into a vector space. The structure of the space is optimized during model training to reflect geometry of the training dataset[12–17]. After the method learns cell embeddings, it clusters them to find groups of cells with similar gene expression programs. Finally, the method then annotates/assigns each group to a cell type for which enough annotated cells already exist in the training dataset[18,19]. However, current methods are unable to annotate cells that are not characterized in existing datasets or have not been measured before. Also these methods cannot classify cells into new cell types that do not exist in the training data. While recent semisupervised and supervised methods[20–23] have made initial steps toward empowering single-cell analyses by reusing previously annotated datasets, these methods require that all cell types have many annotated examples in the training data. As a result, current methods are unable to identify new/unseen cell types.

Here we introduce MARS, an approach for annotating known/seen as well as new/unseen cell types in heterogeneous and time-varying single-cell datasets. MARS uses meta-learning, a paradigm in machine learning that focuses on efficient use of limited annotations[24–27]. In particular, MARS first constructs a meta-dataset by integrating (1) any number of single-cell experiments in which cells are annotated (that is, labeled) by a cell type, and (2) an unannotated experiment, which does not necessarily share any cell types with the labeled data. Using the meta-dataset, MARS jointly learns a set of cell-type landmarks and an embedding function that projects cells into a shared embedding space, such that cells are close to their cell-type landmarks. The embedding space, learned by a deep neural network, identifies gene expression programs and leverages commonalities between experiments in the meta-dataset. This gives MARS a unique ability to generalize to unannotated experiments and identify cell types that were never seen during training. We apply MARS to Tabula Muris[6] and Tabula Muris Senis[28] cell atlases. We find that MARS successfully transfers knowledge between diverse tissues and aligns the same cell types, even when they originate from different tissues. Further, we find that MARS learns meaningful cell-type-specific signatures of aging in a mouse. Our results show that MARS considerably outperforms current techniques for cell-type classification. MARS is able to accurately identify cell types it has never seen during training and can probabilistically recommend interpretable names for them.

## Results

**Meta-learning in MARS.** MARS takes as input single-cell gene expression profiles from heterogeneous or time-varying experiments, such as different tissues or stages of development. MARS creates a meta-dataset that consists of (1) experiments in which cells are annotated according to their cell types, and (2) a completely unannotated experiment in which cell types are unknown. The unannotated experiment can originate from different source and does not need to share any cell types with the annotated experiments. The goal then is to annotate cells in the unannotated experiment, such as never-before-seen tissue or stage of development. This is a new setup not considered by previous single-cell methods.

**Overview of MARS.** Given a meta-dataset as input, MARS learns a set of cell-type landmarks and a nonlinear embedding function.

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Informatics, Harvard University, Boston, MA, USA. [3]Department of Bioengineering, Stanford University, Stanford, CA, USA. [4]Chan Zuckerberg Biohub, San Francisco, CA, USA. ✉e-mail: jure@cs.stanford.ed

The embedding function projects a high-dimensional expression profile of each cell to a low-dimensional vector (that is, cell embedding), which directly captures the cell-type identity (Fig. 1a). Cell-type landmarks are defined as cell-type representatives and are learned for both annotated and unannotated experiments. The embedding function is a deep neural network that maps cells to the embedding space. The embedding space is defined, such that cells embed close to their cell-type landmarks. The embedding function is shared between all experiments in the meta-dataset, which gives MARS the ability to generalize to an unannotated experiment and to capture the similarity of cell types across annotated and unannotated experiments.

Mathematically, MARS uses regularization in the form of pretraining the neural network with a deep autoencoder that minimizes a data reconstruction error (Methods). The pretraining step serves as a prior for the parameter space, which is useful for generalization to an unannotated dataset. Using the pretrained network as initialization, MARS then learns mapping of all cells to the shared embedding space such that similar cells are close to each other, while dissimilar cells are far way. Equipped with the concept of cell-type landmarks, we design an objective function that aims to learn a representation in which cells group close to their corresponding landmarks (Methods). The objective function consists of three parts (Fig. 1b): (1) in the annotated experiments, the distance between cell embeddings and ground-truth cell-type landmark is minimized; (2) in the unannotated experiment, the distance between cell embeddings and the nearest cell-type landmark is minimized and (3) distance between cell-type landmarks within each experiment is maximized. The rationale is to encourage cells from the same cell type to have similar representations, while representations of cells from different cell types are far apart. MARS does not impose any constraint on the radius of a discovered cell type, so cell types can form clusters that reflect their transcriptional similarity to other cell types.

**MARS identifies cell-type-specific signatures of aging.** We assess MARS's ability to infer cell-type trajectories on the Tabula Muris Senis dataset[28], covering the life span of a mouse. In particular, we analyze whether the same cell types from different time points are embedded close together (that is, aligned) in the embedding space. We use the brown adipose tissue (BAT) data from 3-, 18- and 24-month-old mice as annotated experiments. We regard BAT data from each time point as a separate experiment; therefore, MARS assigns different landmarks to the same cell types across time points. We then evaluate MARS on a different tissue by using BAT from three time points as three annotated experiments. We find that natural killer (NK) cells change their position at every time point (Fig. 1c), indicating the MARS detects the existence of transcriptional changes. On the contrary, in the joint low-dimensional embedding, inferred using principal component analysis with the same number of components as the dimensionality of MARS, NK cells are joined with T cells and aligned across different time points (Extended Data Fig. 1). To confirm that the motion of NK cells as detected by MARS is meaningful, we further analyze the variability in gene expression of differentially expressed genes across three time points. Populations of NK cells indeed show higher variability than other cell types with a coefficient of determination ($R^2$) of 0.80 between 3- and 18-month-old mice, and 0.58 between 18- and 24-month-old mice (Fig. 1d). In contrast, the median of $R^2$ of other cell types is 0.93 (Q1–Q3, 0.89–0.95) and 0.89 (Q1–Q3, 0.84–0.89), respectively. Furthermore, populations of NK cells share 6% of differentially expressed genes across three time points compared to the average of 26.8% shared genes on other cell types in BAT, confirming that the representation learned by MARS captures transcriptional changes in aging NK cells. Moreover, this finding has been well-characterized experimentally[29–31], suggesting that cellular functions of NK cells are impaired in aging mice and can lower the resistance to cancer and pathogenic microorganisms.

**MARS outperforms other methods by a large margin.** To demonstrate the performance of MARS on a cell-type annotation task, we use the manually curated Tabula Muris dataset[6]. We consider each tissue as a separate experiment (Methods and Supplementary Note 1). We leave one tissue out as unannotated and use all others as annotated experiments. We then test the performance on the unannotated held-out tissue experiment. Note that often the unannotated held-out tissue shares no cell types with the annotated tissues, which means that MARS has to be able to identify entirely new cell types it has never seen during training.

We compare MARS to four methods that can also apply to this task: deep generative model ScVi[12], kernel-learning approach SIMLR[32] and two community detection approaches, Leiden[33] and Louvain[34], which are used in two popular single-cell analysis tools, Scanpy[35] and Seurat[36] (Supplementary Note 2). MARS achieves a 45% gain in adjusted Rand index (ARI) score over the second-best performing SIMLR (Fig. 2a). When measuring performance using various other classification or clustering metrics, MARS retains substantially better performance than all other methods. In particular, MARS achieves 20, 27, 30, 21 and 21% improvement over the second-best baseline in terms of adjusted mutual information, accuracy, macro-F1 score, macro-precision and macro-recall, respectively (Extended Data Fig. 2a–e and Supplementary Note 3). To directly measure the effect of our objective function that jointly learns landmarks and cell embedding function across independent experiments, we compare MARS to the *K*-means clustering applied in the autoencoder's latent space the end of the MARS pretraining. MARS achieves 20–52% relative improvement in performance across all evaluation metrics, clearly demonstrating the advantage of our meta-learning setting (Extended Data Fig. 2f). Of note, MARS uses the same set of parameters across all tissues and shows high robustness to the choice of the neural network architecture. In particular, MARS's average performance across tissues is not affected when the embedding dimension changes (Extended Data Fig. 2g).

We further compare cell-type-level F1 score between MARS and the second-best performing SIMLR on never-before-seen cell types (Supplementary Note 3). MARS outperforms SIMLR by a large margin and performs exceptionally well on cell types with very few annotated cells (Fig. 2b), and cell types with very few differentially expressed genes (Extended Data Fig. 3a). Across all previously unseen cell types, MARS achieves a 14% median improvement in F1 score over SIMLR. A similar trend is observable when considering all cell types (Extended Data Fig. 3b,c). When comparing performance on individual tissues, MARS performs better than SIMLR on 20 out of 21 tissues and achieves 34% higher area under the curve than SIMLR, and 44% compared to ScVi (Extended Data Fig. 4). For instance, for heart tissue that contains seven out of 11 never-before-seen cell types, MARS improves SIMLR's ARI score by 25.8%.

Additionally, we assess MARS performance on three benchmark datasets: (1) two CellBench datasets[37,38] consisting of lung cancer cells sequenced with different sequencing protocols (10X and CEL-Seq2); (2) three Allen brain datasets[38,39] consisting of different species (mouse and human), as well as single-cell RNA-seq and single-nucleus RNA-seq datasets and (3) two clustering benchmark datasets consisting of diverse human cell types[40] and mouse pluripotent cells[41]. Within each benchmark dataset, we regard each dataset as a separate experiment and train MARS in a leave-one-experiment-out manner. MARS substantially outperforms other baselines and effectively transfers information across sequencing technologies and species, even when experiments consist of a small number of annotated cells (Extended Data Fig. 5 and Supplementary Note 4).

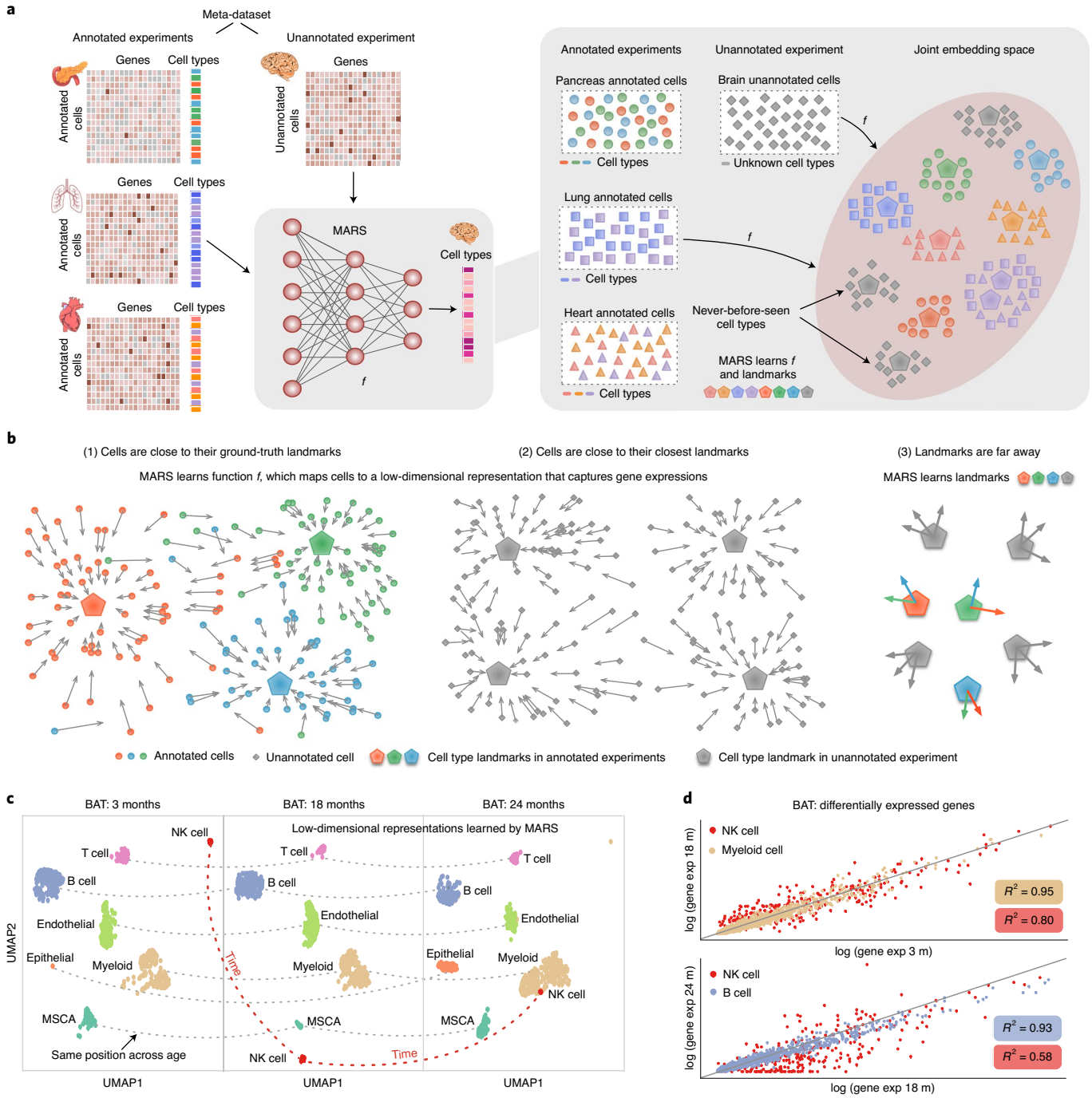



**Fig. 1 | MARS is a meta-learning approach for discovery of new cell types across heterogeneous single-cell experiments. a**, Illustration of the MARS method. Given a set of heterogeneous annotated experiments (for example, pancreas, lung, heart tissues), MARS aims to annotate a new, completely unannotated experiment (for example, brain tissue), even if it does not have any cell type in common with annotated experiments. Using deep neural networks, MARS projects all cells in the meta-dataset (annotated and unannotated) to the shared embedding space and learns nonlinear embedding function *f* such that cells from the same cell types are embedded close to each other, while cells from different cell types are embedded far away. **b**, MARS relies on the notion of a cell-type landmarks. Objective function of MARS simultaneously optimizes three parts: (1) within annotated experiment, distance to the ground-truth landmark is minimized; (2) within unannotated experiment, distance to the closest landmark is minimized and (3) within each experiment, distance between landmarks is maximized. Cell-type landmarks and experiment-invariant cell representations are learned jointly and in an end-to-end fashion. **c**, MARS reconstructs a trajectory of BAT cell types during the life span of a mouse. All BAT cell types except NK cells retain the same position across three different time points. **d**, Comparison of gene expressions of differentially expressed genes in BAT across different time points. Top plot shows average gene expression of differentially expressed genes of 3- and 18-month-old mice for NK cells and myeloid cells. Bottom plot shows average gene expression of 18- and 24-month-old mice for NK cells and B cells. Average is calculated over $n = (17, 27, 4, 168, 201, 211, 120)$ for NK cells in 3-, 18- and 24-month-old mice, myeloid cells in 3- and 18-month-old mice and B cells in 18- and 24-month-old mice, respectively.

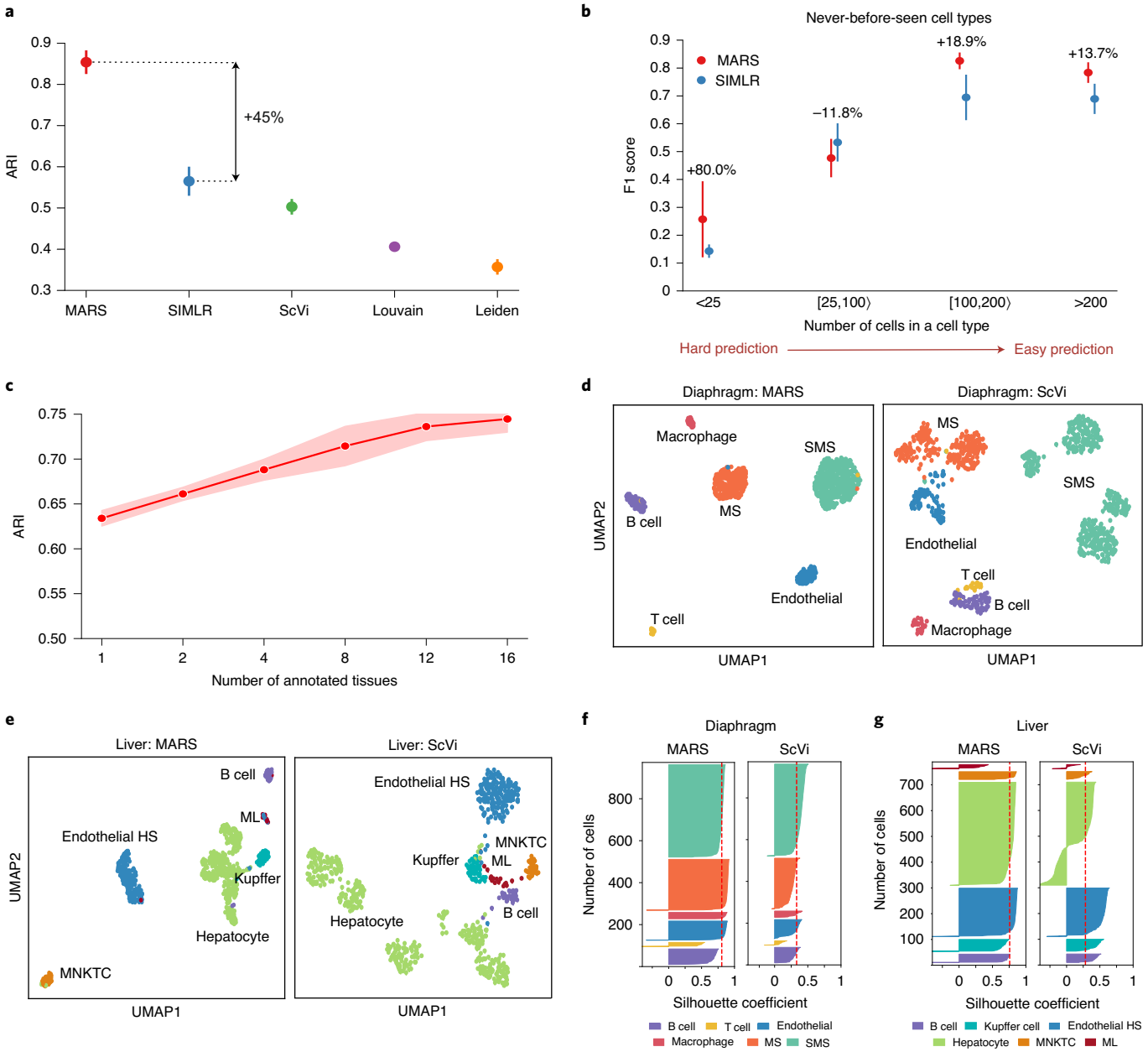



**Fig. 2 | MARS achieves positive learning transfer and accurately annotates cells. a**, Median performance of MARS and four baseline methods evaluated using ARI score across 21 different tissues (Methods). Higher value indicates better performance, where 1 is perfect performance and 0 indicates random clustering. Error bars are standard errors estimated as a standard deviation of the mean by bootstrapping cells within tissue with $n=20$ iterations. MARS is trained in leave-one-tissue-out manner, and the held-out tissue was completely unannotated (Methods). **b**, Cell-type-level comparison of MARS's F1 score with the SIMLR on cell types that have never been seen in the annotated experiments. Standard errors are estimated as a standard deviation of the mean by bootstrapping cells within each tissue with $n=20$ iterations. Cell types are grouped based on the number of cells in the Tabula Muris annotations, where cell types with fewer number of cells are harder to recognize as a separate cluster. **c**, Effect of the number of annotated tissues in the meta-dataset on MARS's performance. Performance is measured as average ARI. Error bands are standard deviation across 20 runs of the method. Annotated tissues are selected according to their similarity to an unannotated tissue. **d**,**e**, UMAP visualizations of deep variational autoencoder ScVi's and MARS's embeddings for diaphragm tissue (**d**) and liver tissue (**e**). SMS stands for skeletal muscle cell, MS for mesenchymal stem, HS for hepatic sinusoid, ML for myeloid leukocyte and MNKTC for mature NK T cell. Color indicates Tabula Muris cell-type annotations. Only cell types with more than five annotated cells are shown. **f**,**g**, Quality of the neural embeddings of MARS and ScVi measured as silhouette coefficient on diaphragm tissue (**f**) and liver tissue (**g**).

**MARS achieves positive knowledge transfer across tissues.** We show that MARS achieves better performance as the number of annotated experiments increases. Specifically, we start with the meta-dataset consisting of only one annotated experiment, and then gradually add more annotated experiments in the meta-dataset based on their similarity to the unannotated experiment (Methods). We find that MARS performs considerably better on large meta-datasets (Fig. 2c). In particular, when using heart and mesenteric fat as the unannotated experiments, MARS improves by 64.1 and 34.5%, respectively, between using one and all tissues (Extended Data Fig. 6). Although subcutaneous fat, mesenteric fat, heart and BAT do not have any cell types in common with large intestine

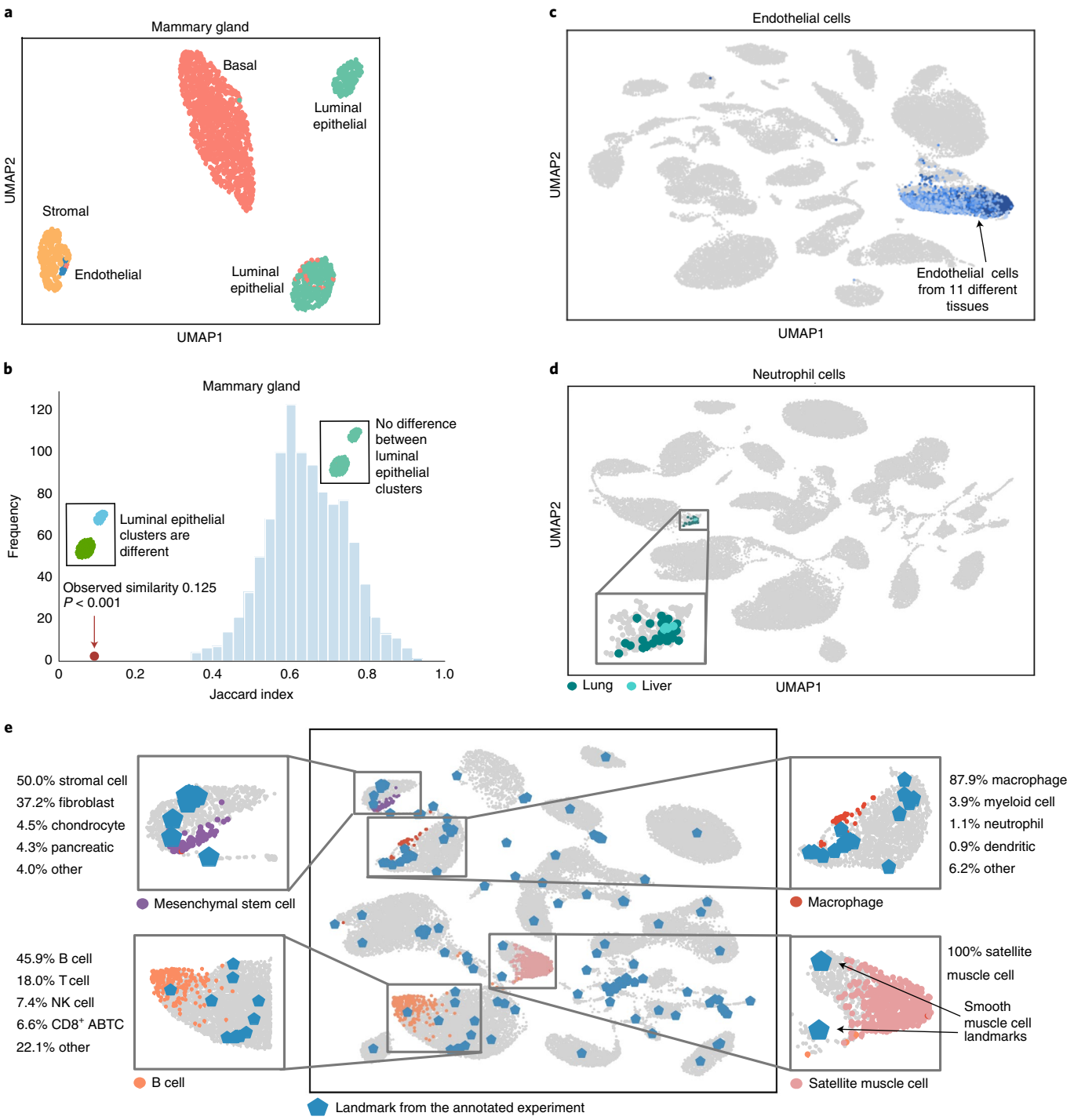



**Fig. 3 | MARS accurately identifies cell types, even when tissues have no cell types in common, and automatically generates interpretable names for new cell types. a**, UMAP visualization of MARS's embedding of mammary gland tissue cells. Color indicates Tabula Muris cell-type annotations. **b**, Results of a permutation test under the null hypothesis that there is no difference between luminal epithelial cells. We define the test statistic to be a Jaccard similarity of enriched GO terms of differentially expressed genes between two groups (Methods). The observed value is the similarity between two clusters of luminal epithelial cells found by MARS, while distribution is obtained by randomly permuting luminal epithelial cells into two groups with $n = 1{,}000$ iterations. The observed difference between two clusters found by MARS is significant with $P < 10^{-3}$. **c**,**d**, UMAP visualizations of MARS joint embedding space of all tissues for endothelial cells (**c**) with thymus tissue as unannotated tissue, and a small cluster of neutrophil cells (**d**) with lung as unannotated tissue. **e**, Overview of the MARS cell-type naming approach. For an unannotated cell type that we want to name, MARS determines distances to all landmarks from the annotated experiments and for each of them outputs probability that discovered cell type should receive the same name (Methods). In the example, limb muscle is used as an unannotated tissue. PDC stands for plasmacytoid dendritic cell and CD8$^+$ ABTC for CD8-positive alpha–beta T cell.

tissue, including them into meta-dataset improves performance by 20.6% when predicting cell types of large intestine. This analysis demonstrates that MARS effectively reuses annotated experiments, even when they differ in their gene expression profiles from the unannotated experiment. Our results suggest that more annotated experiments yield higher-quality cell embeddings.

**MARS discovers new cell types and subtypes.** We visualize representations of cells learned by MARS in the two-dimensional uniform manifold approximation and projection (UMAP)[42] space. MARS learns to embed similar cells close to each other, while dissimilar cells are embedded far, agreeing well with the Tabula Muris annotations. In contrast, in the ScVi embedding space, different cell types are often mixed without a clear decision boundary between cell types (Fig. 2d,e). To quantitatively evaluate the quality of the neural embeddings, we use the silhouette coefficient, which compares inter- and intracluster distance of data points, indicating how well is a data point matched to its own cluster with −1 as the lowest and 1 as the highest score. In both tissues, MARS achieves a silhouette coefficient of 0.8, whereas ScVi achieves a score of 0.3 (Fig. 2f,g). Additionally, we compare latent space at the end of the MARS pretraining step and the final MARS model. While some cell types form clusters after a pretraining step, most cell types can only be separated with the final MARS model (Extended Data Fig. 7).

We further observe that MARS discovers new cell subtypes. In particular, we analyze mammary gland tissue for which the cell types discovered by MARS differ from the Tabula Muris annotations. MARS separates cells annotated as luminal epithelial cells by Tabula Muris into two different clusters (Fig. 3a). To check whether luminal epithelial cells in two clusters detected by MARS are indeed different, we run a permutation test, comparing Jaccard similarity of Gene Ontology (GO)[43] enriched terms of differentially expressed genes in the sampling distribution to Jaccard similarity of clusters detected by MARS (Methods). Results confirm that luminal epithelial cells in clusters detected by MARS differ significantly ($P < 10^{-3}$; Fig. 3b), indicating that MARS discovers subtypes of luminal epithelial cell. We also compare discovered subtypes to free annotations that provide additional cell-type resolution for mammary gland tissue. We find that MARS annotations entirely agree with the free annotations, and a discovered subtype represents luminal progenitor cells (Extended Data Fig 8a,b). Using these free annotations, we additionally evaluated whether MARS can separate cell subtypes of basal cell of the epidermis and dendritic cells and obtained perfect performance (Extended Data Fig. 8c,d).

**MARS correctly aligns and annotates cell types across tissues.** MARS uses a meta-dataset to learn embedding space, which effectively generalizes to never-before-seen experiments. Next, we examine whether the same cell types across tissues in the annotated and unannotated experiments are embedded close to each other. We first investigate endothelial cells, which appear in 11 tissues. We use thymus tissue as an unannotated experiment and 21 other tissues as annotated experiments. According to the tissue-level performance, we select thymus as the most challenging tissue with endothelial cells (Extended Data Fig. 4b). We find that endothelial cells are exceptionally well aligned across diverse tissues, even in the unannotated thymus tissue (Fig. 3c). We observe near-perfect alignment for other cell types that appear across many tissues, such as B cells (Extended Data Fig. 9). We further evaluate a small neutrophil cell type that appears in only lung and liver tissues by using the lung as an unannotated experiment. Remarkably, neutrophils from unannotated lung tissue align well to only four liver neutrophil cells (Fig. 3d). Finally, we note that MARS does not explicitly correct for batch effects, but it is complementary to integrative approaches for batch correction, including refs. [17,22,23,44]. MARS can be applied to batch-corrected datasets returned by these methods.

**MARS can name new cell types.** Last, we demonstrate the ability of MARS to assign interpretable names to discovered groups of cells. MARS relies on the cell-type landmarks in the annotated experiments to probabilistically define cell type based on its region in the low-dimensional embedding space. Probabilities are assigned to landmarks in proportion to their probability density under a Gaussian distribution centered at a target unannotated cell type (Methods). To demonstrate our approach, we analyze whether cell types with more than ten cells from the limb muscle tissue are correctly assigned. Indeed, MARS accurately identifies satellite muscle cells and endothelial cells with 100% probability, macrophages with over 87% probability and B cells with more than 45% probability (Fig. 3e). At first glance, it may look like MARS misclassifies mesenchymal stem cells (MSCs) by assigning them to stromal cells with high confidence; however, MSCs are adherent stromal cells[45]. Furthermore, with a 37.2% of probability, MSCs are assigned to the fibroblast cell type, which is indistinguishable from MSCs using morphology and cell-surface markers[45,46]. Hence, distances in MARS's embedding space can also be used to infer the similarity between cell types. Even if datasets are not corrected for batch effects, MARS can be used to discover new cell types; however, our post hoc naming approach relies on the distances across experiments. Therefore, if batch effects across experiments are present, datasets need to be corrected first with existing approaches for batch correction[17,22,23,44] for our naming approach to return meaningful results.

## Discussion

MARS has a unique ability to transfer knowledge of cell embeddings across heterogeneous experiments that possibly do not have any cell types in common. In doing so, MARS introduces a practical setting for the analysis of single-cell data, in which the experiment of interest can be completely new and unannotated, thereby requiring generalization to never-before-seen cell types.

MARS addresses this challenge by learning cell-type-specific landmarks and a nonlinear embedding function that maps all cells in a joint low-dimensional embedding space shared by annotated and unannotated experiments. Using the learned landmarks to identify new cell types, MARS provides a framework for annotation of discovered cell types by probabilistically assigning cell types to the neighborhood of the annotated landmarks. As a result, MARS can considerably alleviate the post hoc manual analyses of cell types. However, post hoc annotation relies on distances and MARS does not perform batch correction. Therefore, for annotation to be effective, the datasets need to be batch-corrected beforehand.

MARS allows for knowledge transfer across tissues, time-varying experiments, species and sequencing protocols. Our approach has important implications for other types of knowledge transfer, including the transfer of cell representations across different omics measurements and transfer of cell states across related diseases.

Finally, MARS is complementary to tools for correcting batch effects and data integrative studies, including Scanorama[44], Harmony[17] and Seurat V3 (ref. [22]). Results returned by these tools can be directly used as input to MARS. As new comprehensive atlas datasets are generated in line with Human Cell Atlas[7] efforts, we envision that MARS will become a useful tool to help in unraveling an unknown cellular diversity of healthy and diseased tissues.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-00979-3.

Received: 21 February 2020; Accepted: 15 September 2020;
Published online: 19 October 2020

## References

1. Park, J. et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018).

2. McKenna, A. & Gagnon, J. A. Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730 (2019).
3. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotech.* **33**, 495–502 (2015).
4. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
5. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).
6. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: the Tabula Muris consortium. *Nature* **562**, 367–372 (2018).
7. Regev, A. et al. Science forum: the Human Cell Atlas. *eLife* **6**, e27041 (2017).
8. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
9. Suo, S. et al. Revealing the critical regulators of cell identity in the mouse cell atlas. *Cell Rep.* **25**, 1436–1445 (2018).
10. Aevermann, B. D. et al. Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Hum. Mol. Genet.* **27**, R40–R47 (2018).
11. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed infibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).
12. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
13. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).
14. Wang, J. et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16**, 875–878 (2019).
15. Amodio, M. et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).
16. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
17. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
18. Tyssowski, K. M. et al. Different neuronal activity patterns induce different gene expression programs. *Neuron* **98**, 530–546 (2018).
19. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
20. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
21. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Preprint at *bioRxiv* https://doi.org/10.1101/532895 (2020).
22. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
23. Wang, T. et al. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol.* **20**, 1–15 (2019).
24. Schmidhuber, J. *Evolutionary Principles in Self-referential Learning*. Diploma thesis, Technische Univ. München (1987).
25. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. Meta-learning with memory-augmented neural networks. In *Proc. International Conference on Machine Learning* **33** (eds Balcan, M. F. et al.), 1842–1850 (PMLR, 2016).
26. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. *Proc Adv. Neural Inform. Proc. Syst.* **31** (eds Guyon, I. et al.), 4077–4087 (Curran Associates, 2017).
27. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. International Conference on Machine Learning* **34** (eds Precup, D. et al.) 1126–1135 (PMLR, 2017).
28. The Tabula Muris Consortium. A single cell transcriptomic atlas characterizes aging tissues in the mouse. *Nature* **583**, 590–595 (2020).
29. Albright, J. W. & Albright, J. F. Age-associated impairment of murine natural killer activity. *Proc. Natl Acad. Sci. USA* **80**, 6371–6375 (1983).
30. Nogusa, S., Ritz, B. W., Kassim, S. H., Jennings, S. R. & Gardner, E. M. Characterization of age-related changes in natural killer cells during primary influenza infection in mice. *Mech. Ageing Dev.* **129**, 223–230 (2008).
31. Nair, S., Fang, M. & Sigal, L. J. The natural killer cell dysfunction of aged mice is due to the bone marrow stroma and is not restored by IL-15/IL-15Rα treatment. *Aging Cell* **14**, 180–190 (2015).
32. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-Seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
33. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
34. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
35. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
36. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnology* **36**, 411–420 (2018).
37. Tian, L. et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16**, 479–487 (2019).
38. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
39. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
40. Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotech.* **32**, 1053–1058 (2014).
41. Kolodziejczyk, A. A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
42. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
43. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
44. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotech.* **37**, 685–691 (2019).
45. Haniffa, M. A., Collin, M. P., Buckley, C. D. & Dazzi, F. Mesenchymal stem cells: the fibroblasts new clothes? *Haematologica* **94**, 258–263 (2009).
46. Hematti, P. Mesenchymal stromal cells and fibroblasts: a case of mistaken identity? *Cytotherapy* **14**, 516–521 (2012).

## Methods

**Dataset preprocessing.** We downloaded raw read Tabula Muris[6] and Tabula Muris Senis[28] datasets with cell-type annotations (see Data availability). We filtered low-quality cells with fewer than 5,000 reads and 500 genes, as well as genes expressed in less than five cells. We used Scanpy[35] to normalize each cell to 10,000 read counts, and then log transformed the data. Finally, we scaled the dataset to unit variance and zero mean, and we truncated values with maximum value set to ten. The normalization and scaling steps remove experiment-specific differences and enable alignment based on the relative gene expression values. We find that jointly scaling data is an important preprocessing step. After preprocessing, the number of retained genes was 22,903. The number of annotated cells was 105,960 in Tabula Muris Senis and 44,516 in Tabula Muris. The number of cells per dataset ranged from 906 to 13,417 cells in Tabula Muris Senis, and 366 to 5,067 cells in Tabula Muris. To demonstrate the ability of MARS to detect aging signatures, we used Tabula Muris Senis dataset. For all other analyses, we used Tabula Muris dataset with reannotations from Tabula Muris Senis. Additional details are provided in Supplementary Note 1.

**Overview of MARS.** The key idea in the MARS model is that representation that encourages clustering of cells in one experiment also helps in learning to separate cells in a distinct experiment. We aim to accomplish the goal of learning experiment-invariant representations by transferring knowledge of the right distance metric from previously annotated experiments to a new, completely unannotated experiment. We refer to the set of all experiments (annotated and unannotated) over which MARS learns as a meta-dataset; that is, a dataset for learning to learn representation that can easily adapt to new tasks. To achieve transferable features, MARS learns shared representation across all experiments in the meta-dataset. Specifically, given gene expression profiles and cell-type annotations in the annotated experiments, and gene expression profiles of an unannotated target experiment, MARS learns the nonlinear mapping function $f_\theta$ that maps cells from all experiments into a joint embedding space such that cells are grouped according to their cell types. The function $f$ is parameterized by learnable feature mapping parameters $\theta$ of a deep neural network. MARS consists of two stages: (1) pretraining on an unannotated target experiment with deep autoencoder, and (2) learning cell-type landmarks and nonlinear cell embedding with deep neural network. MARS optimizes cell-type landmarks and parameters $\theta$ in an end-to-end manner.

1. Pretraining. We first pretrain MARS with an autoencoder. An autoencoder network takes as input normalized gene expression profiles of unannotated experiment $X^u \in \mathbb{R}^{N\times G}$, where $N$ denotes number of cells and $G$ denotes number of genes. Input is mapped to a lower-dimensional dense representation vector (that is, encoding). The decoder part maps encoding vector to the reconstruction of the input $\hat{X}^u$. Autoencoder is trained to minimize reconstruction loss $\mathcal{L}(X^u, \hat{X}^u)$, given as the mean squared error between $X^u$ and $\hat{X}^u$. After pretraining, we remove the decoder part and use learned weights to initialize neural network.
2. Initialization of cell-type landmarks. To initialize cell-type landmarks, we first map all cells into a lower-dimensional representation vector learned by autoencoder. Then, for each experiment in the meta-dataset we separately run $K$-means clustering in the embedding space. We use ten random initializations and take the best one in terms of the sum of squared distances of cells to their closest cluster landmark.
3. Loss function. Let $\mathcal{D}_{\text{meta}} = \{\mathcal{D}^{(i)}\}_{i=1}^{M} \cup \mathcal{U}$ be a set of $(M+1)$ distinct experiments to which we refer to as a meta-dataset. We assume that each experiment $\mathcal{D}^{(i)}$ consists of a matrix of normalized gene expression profiles $X^{(i)} = \{\mathbf{x}_j^{(i)} \in \mathbb{R}^G\}_{j=1}^{N_i}$, and a vector of cell-type annotations $\mathbf{y}^{(i)} = \{y_j \in \{1, \dots, K_i\}\}_{j=1}^{N_i}$, where $G$ denotes number of genes, $N_i$ number of cells and $K_i$ number of cell types in the experiment $\mathcal{D}^{(i)}$. Furthermore, let $\mathcal{U}$ consists of a matrix of gene expression profiles $X^{(u)} = \{\mathbf{x}_j \in \mathbb{R}^G\}_{j=1}^{N}$, with unknown cell annotations, where $N$ denotes number of cells in $\mathcal{U}$. Given a meta-dataset $\mathcal{D}_{\text{meta}}$, MARS learns a set of cell-type landmarks in the annotated experiments $\left\{\left\{\mathbf{p}_k^{(i)} \in \mathbb{R}^Z\right\}_{k=1}^{K_i}\right\}_{i=1}^{M}$, a set of cell-type landmarks in the unannotated experiment $\{\mathbf{p}_k \in \mathbb{R}^Z\}_{k=1}^{K}$ and a nonlinear mapping function $f_\theta : \mathbb{R}^G \to \mathbb{R}^Z$, where $K$ denotes number of cell types in the unannotated experiment, $Z$ is dimension of the embedding space and $\theta$ are learnable parameters. In MARS, we seek to find a joint embedding space such that within each experiment cells group around a single-cell-type landmark and landmarks are far away. Therefore, the mapping function $f_\theta$ is shared between all experiments in the meta-dataset and maps all cells into the joint embedding space.
   In the annotated meta-dataset, cell-type annotations are known and MARS encourages cells to be close to their ground-truth cell-type landmarks. For each annotated experiment $\mathcal{D}^{(i)} \in \mathcal{D}_{\text{meta}}$, MARS incorporates the following part in the objective function:

$$\mathcal{L}_i = \frac{1}{N_i}\sum_{k=1}^{K_i}\sum_{j=1}^{N_i} \mathbb{I}_{\{y_j^{(i)}=k\}}\, d\left(f_\theta\left(\mathbf{x}_j^{(i)}\right), \mathbf{p}_k^{(i)}\right) - \frac{\lambda}{K_i(K_i-1)}\sum_{k_1=1}^{K_i}\sum_{k_2=1}^{K_i} d\left(\mathbf{p}_{k_1}^{(i)}, \mathbf{p}_{k_2}^{(i)}\right), \quad (1)$$

where $\lambda$ is a regularization constant, $\mathbb{I}$ denotes the indicator function and $d$ is a distance function. We use squared Euclidean distance as a distance function, but others can be easily incorporated. Of note, all distances are calculated in the low-dimensional embedding space. The first part of the equation measures intracluster distance between cells and ground-truth landmarks, whereas the second part measures intercluster distance between all pairs of landmarks. Intracluster distance is minimized to achieve compact representations within a cluster, whereas intercluster distance is maximized to push representations of distinct landmarks far away from each other.
   Next, we include in the objective function term that encourages clustering structure of the unannotated experiment $\mathcal{U}$. With the same intuition as above, we again measure intra- and intercluster distance. However, in this case cell-type assignments are unknown, so MARS minimizes the distance to the closest cell-type landmark in the unannotated experiment. Formally, for $\mathcal{U} \in \mathcal{D}_{\text{meta}}$, MARS extends the objective function with the following term:

$$\mathcal{L}_u = \frac{1}{N}\sum_{j=1}^{N} \min_{k=1,\dots,K} d\left(f_\theta(\mathbf{x}_j), \mathbf{p}_k\right) - \frac{\lambda}{K(K-1)}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} d\left(\mathbf{p}_{k_1}, \mathbf{p}_{k_2}\right). \quad (2)$$

   The final objective function optimizes for the annotated and unannotated experiments jointly:

$$\mathcal{L}_{\text{MARS}} = \min_{\theta, \{\mathbf{p}_k^{(i)}\}_{i,k}, \{\mathbf{p}_k\}_k} \sum_{i=1}^{M} \mathcal{L}_i + \tau\mathcal{L}_u. \quad (3)$$

   The objective function balances between intracluster minimization and intercluster maximization. Both parts are optimized within each experiment, allowing clusters across experiments to align with each other. Cluster landmarks and representation parameters $\theta$ learned by deep neural network are optimized simultaneously. In each iteration, we first optimize for landmarks while fixing the parameters $\theta$. Then, we optimize for $\theta$ while fixing the landmarks. In the annotated experiments, landmarks are obtained as a closed-form solution of equation (1). In the unannotated experiment, we update landmarks with the Adam optimizer.
4. Inference. Embeddings of cells in the meta-dataset are obtained by the representation learned in the last layer of the neural network. At the inference time, we annotate cells from the unannotated experiment. In particular, MARS embeds cells from the unannotated experiment into the learned shared embedding space and assigns them to the cluster of the closest cell-type landmark from the unannotated dataset.
5. Cell-type naming. MARS probabilistically assigns interpretable names to discovered clusters by relying on the annotated cell-type landmarks in the meta-dataset. Probabilities are estimated for every cell type seen in the annotated experiments in proportion to their probability density under a Gaussian distribution centered at the mean of a discovered cluster. Then, annotations are assigned to the discovered cluster based on the annotations of the most similar annotated landmarks. Formally, given cell-type landmarks $\left\{\left\{\mathbf{p}_k^{(i)} \in \mathbb{R}^Z\right\}_{k=1}^{K_i}\right\}_{i=1}^{M}$ in the annotated experiments, conditional probability that $j$th cluster in the unannotated experiment adds $k$th landmark from the annotated experiments in the set of the most similar landmarks is calculated as follows:

$$p_{k|j} = \frac{\exp\left(-||\mathbf{p}_k - \boldsymbol{\mu}_j||^2/2\sigma_j^2\right)}{\sum_{i=1}^{M}\sum_{k'}^{K_i}\exp\left(-||\mathbf{p}_{k'}^{(i)} - \boldsymbol{\mu}_j||^2/2\sigma_j^2\right)},$$

   where $\boldsymbol{\mu}_j$ is the mean of cell embedding vectors assigned to target cluster $j$ and $\sigma_j$ is estimated based on the standard deviation of pairwise Euclidean distances of cells assigned to cluster $j$. Empirically, we observe that embedding data points beforehand in the low-dimensional space with UMAP improves the results. We used ten UMAP components.

**Architecture and hyperparameters.** The neural network used in MARS consists of two fully connected layers. We used 1,000 neurons in the first layer, and 100 neurons in the second layer of the neural network. On the Tabula Muris data, the input is given by gene expression profiles of 22,903 genes. During pretraining, we used a mirror-image of this neural network as a decoder. During meta-learning stage, we removed decoder part and optimized the parameters with the loss introduced in MARS. Best parameters were found in a small grid search according to the best mean performance across all tissues. We used Adam optimizer with learning rate 0.001 for pretraining and fine tuning. Activities of the neurons were normalized using layer normalization that estimates the normalization statistics over all hidden units in the same layer. The ELU function, defined as $\text{ELU}(x) = \max(0, x) + \min(0, \alpha(\exp(x) - 1))$, was used as a nonlinear activation with $\alpha$ set to 1. We pretrained the network for 25 epochs, and fine-tuned for 30 epochs. Regularizers $\lambda$ and $\tau$ in the MARS's objective function were set to 0.2 and 1, respectively. We assessed the robustness of MARS to the selection of architecture by varying embedding dimension across a range of possible values, while keeping

all other parameters fixed (Extended Data Fig. 2g), as well as robustness to the regularizers $\lambda$ and $\tau$ (Extended Data Fig. 10a,b). Additionally, we evaluated performance when training for more or fewer epochs (Extended Data Fig. 10c).

**Number of clusters.** MARS requires the number of cell types for the unannotated dataset to be predefined as a parameter. By varying the number of cell types, MARS can be used for a multi-resolution exploration and more fine-grained annotation of the cell types. Empirically, we find that if the number of clusters is set to a slightly too high value, MARS does not use unneeded landmarks. In particular, in our experiments we find that if during optimization none of the cells chooses some landmark as its cell-type representative, then the initial number of clusters can be reduced.

**Pretraining step.** Pretraining is a required step of the MARS model, and it gives a substantial boost in the performance compared to starting from the random weights. MARS can be pretrained on only an unannotated experiment, or jointly on annotated and unannotated experiments. Empirically, we find that on the Tabula Muris dataset the performance is not boosted by adding annotated experiments during pretraining, while the pretraining time increases. Including an unannotated experiment during pretraining is crucial to initialize the model toward configurations of the parameter space that are useful for learning a good representation of the unannotated experiment.

**Performance evaluation.** We evaluated MARS performance in leave-one-tissue-out manner. We used all except one tissue as the set of annotated experiments, and held-out tissue as an unannotated experiment. We evaluated performance by comparing cell-type assignments of the unannotated experiment to the ground-truth clusters. To evaluate how the number of annotated experiments in the meta-dataset affects performance, we used as annotated experiments $n$ most similar tissues to unannotated tissue, while varying $n$ from 1 to 16. Similarity between tissues was computed as the Euclidean distance of their mean gene expression profiles. More details on evaluation are provided in Supplementary Note 3.

**Choice of baselines.** MARS is designed as an inherently unsupervised technique. Opposed to the existing supervised and semisupervised methods[20–23] that transfer annotations across experiments, MARS uses annotated experiments solely to learn a good embedding space. Therefore, annotated and unannotated experiments do not need to have any cell type in common. Even if a same cell type appears in the annotated and unannotated experiments, MARS will assign a new landmark to that cell type in the unannotated experiment. Our naming approach is the only part that transfers annotations across experiments and requires that the experiments are batch-corrected. Therefore, it is designed as a post hoc step that users can decide whether to use. Tasks that can be uniquely solved by MARS can be compared only to existing clustering methods; however, clustering methods cannot transfer information across datasets.

**Visualization.** We visualized cell embeddings using UMAP[42]. Cell neighborhood graph was calculated with the number of neighbors set to 30. For the visualization of the alignment and our naming approach (Fig. 3c–e), we calculated neighborhood graph and performed UMAP on MARS's cell embeddings across all tissues.

**Differential gene expression.** We performed differential gene expression analysis using Scanpy package[35]. We used a $t$-test as the statistical test, and Benjamini–Hochberg method for the adjustment of $P$ values. For Figs. 1d and 3b we consider all genes with a Benjamini–Hochberg false-discovery rate adjusted $P$ value $<0.1$ as differentially expressed (two-tailed $t$-test). Maximum number of genes was set to 100, which is the default value in Scanpy.

**Permutation test and functional enrichment analyses.** To check whether two clusters of luminal epithelial cells in Fig. 3a are significantly different, we performed permutation test. We chose Jaccard similarity of enriched GO[43] terms between differentially expressed genes of two samples as the test statistic. To calculate differential gene expression, the reference set of cells consisted of all cells that are not annotated as luminal epithelial cells (stromal, basal and endothelial cells). The observed value of the test statistic was Jaccard similarity of enriched GO terms between differentially expressed genes of two clusters of luminal epithelial cells detected by MARS. Sampling distribution of the test statistic was estimated by randomly permuting luminal epithelial into two groups and calculating Jaccard similarity between the groups. GO-enriched terms were calculated using GOATOOLS package[47]. GO terms were propagated to parent terms before functional enrichment tests were calculated.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The Tabula Muris Senis dataset is publicly available at https://figshare.com/projects/Tabula_Muris_Senis/64982. The Tabula Muris dataset is publicly available at https://doi.org/10.6084/m9.figshare.5829687.v8. We retrieved data from the website on 2 November 2019. We made Tabula Muris and Tabula Muris Senis datasets in h5ad format available at https://snap.stanford.edu/mars/data/tms-facs-mars.tar.gz. The Pollen dataset[40] is available in the NCBI Sequence Read Archive under accession number SRP041736. Kolodziejczyk[41] sequencing data are available in the ArrayExpress database under accession number E-MTAB-2600. CellBench[37] and Allen Brain datasets[39] are downloaded from https://doi.org/10.5281/zenodo.3357167. Originally, three brain datasets—Allen Mouse Brain (AMB), VISp and ALM—were from the Allen Institute Brain Atlas (http://celltypes.brain-map.org/rnaseq) and are available under accession number GSE115746. The CellBench 10X dataset is available under accession number GSM3618014, while CellBench CEL-Seq2dataset is from three datasets (GSM3618022, GSM3618023, GSM3618024). The project website with links to data and code can be accessed at http://snap.stanford.edu/mars/.

## Code availability

MARS is written in Python using the PyTorch library. The source code is available on Github at https://github.com/snap-stanford/mars.

## References

47. Klopfenstein, D. et al. GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).

## Acknowledgements

We gratefully acknowledge the support of DARPA under nos. FA865018C7880 (ASED) and N660011924033 (MCS); ARO under nos. W911NF-16-1-0342 (MURI) and W911NF-16-1-0171 (DURIP); the National Science Foundation (NSF) under nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions) and IIS-2030477 (RAPID); the Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, Boeing, Chase, Docomo, Hitachi, JD.com, NVIDIA, Dell. J.L. is a Chan Zuckerberg Biohub investigator. M.Z. is supported, in part, by NSF grant nos. IIS-2030459 and IIS-2033384, and by the Harvard Data Science Initiative. A.O.P is supported by CZ Biohub. R.B.A. is supported by CZ Biohub and grant no. NIH GM102365.

## Author contributions

M.B., M.Z. and J.L. conceived the study, designed and performed research, contributed new analytical tools, analyzed data and wrote the manuscript. M.B. also developed the software, performed experiments and developed the metrics. S.W. discussed the results and contributed to the writing. A.O.P. helped procure and interpret the datasets. S.D. and R.B.A. supervised research and contributed to the writing. J.L. supervised the research and the entire project.

## Competing interests

The authors declare no competing interests.

## Additional information

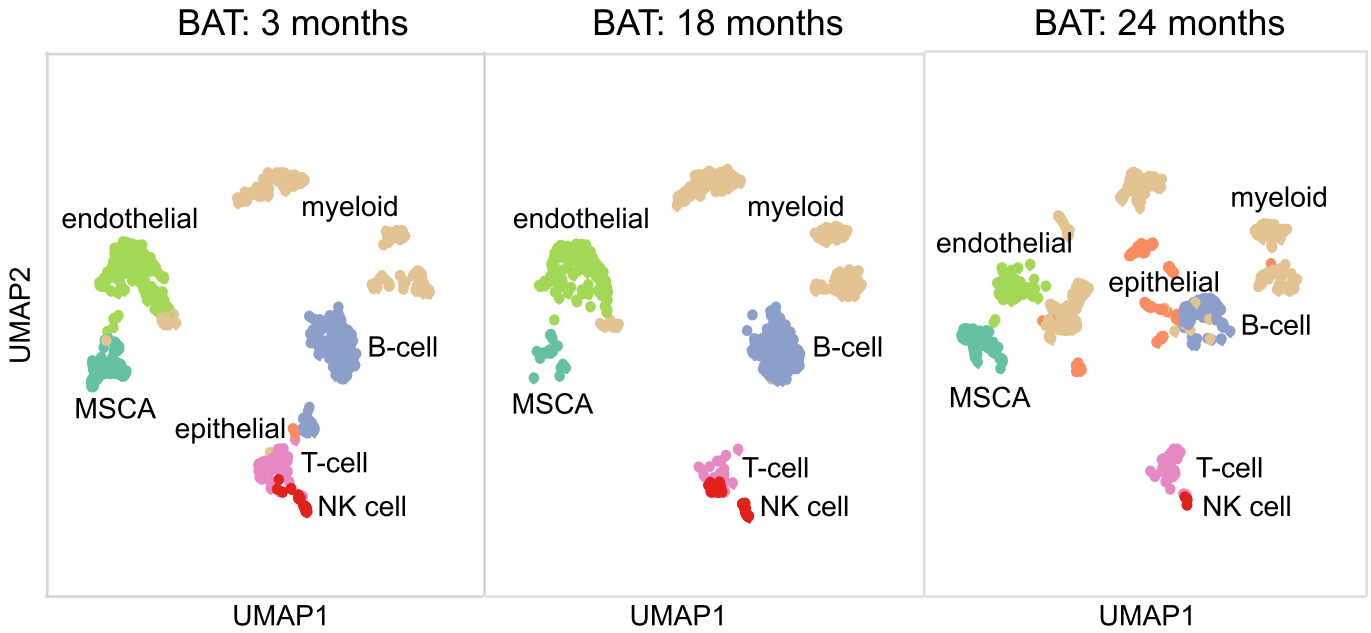
**Extended data** is available for this paper at https://doi.org/10.1038/s41592-020-00979-3.

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41592-020-00979-3.

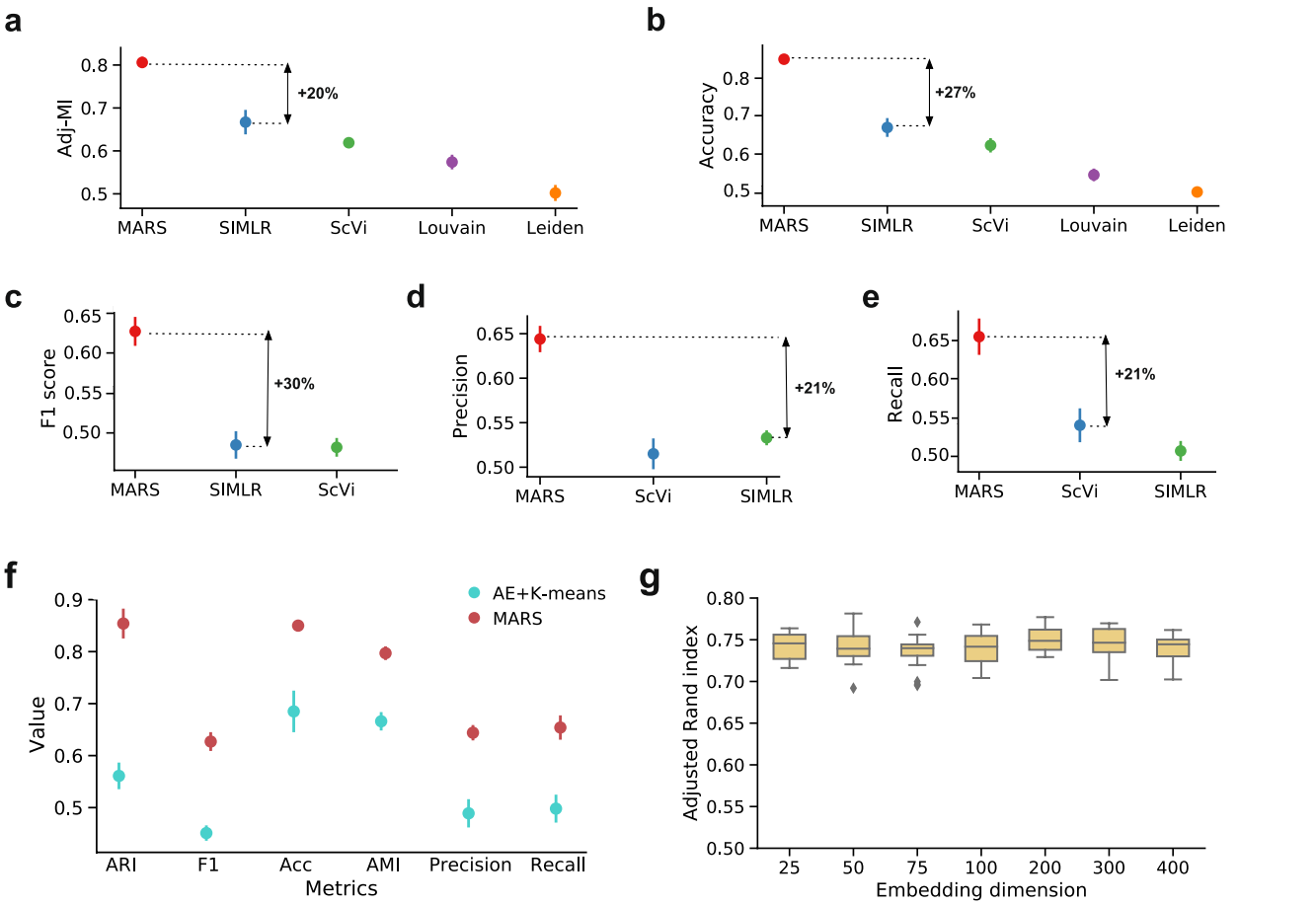
**Correspondence and requests for materials** should be addressed to J.L.

**Peer review information** Lin Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

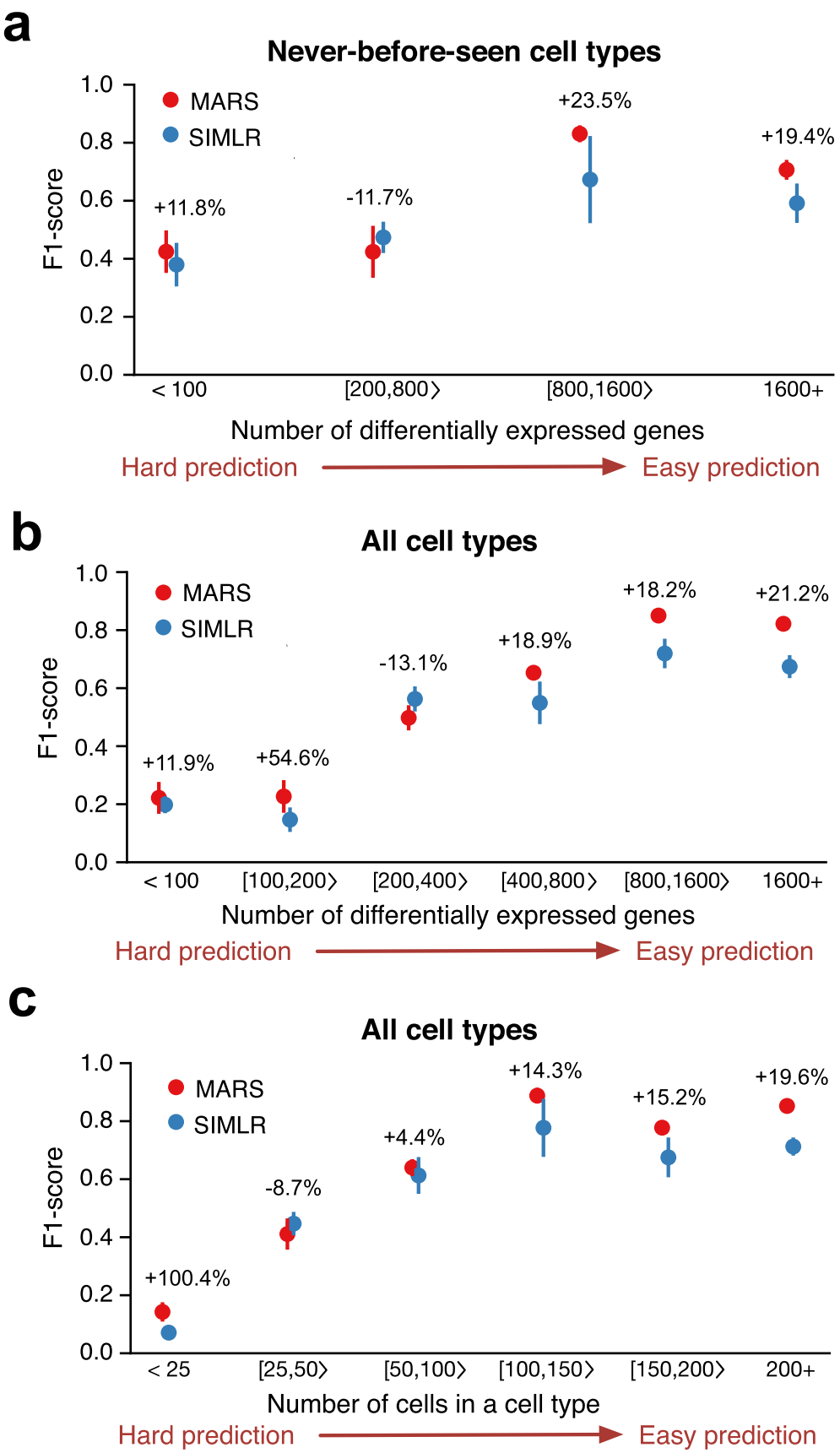
**Reprints and permissions information** is available at www.nature.com/reprints.




**Extended Data Fig. 1 | Brown adipose tissue embedding using PCA.** Joint low-dimensional embedding of brown adipose tissue (BAT) cell types during the life span of a mouse obtained using the PCA. We performed PCA using 100 components, corresponding to the dimensionality of low-dimensional MARS's embeddings. Opposed to the MARS embedding space, NK cells do not change their position across different time points and are joined with the T-cells.

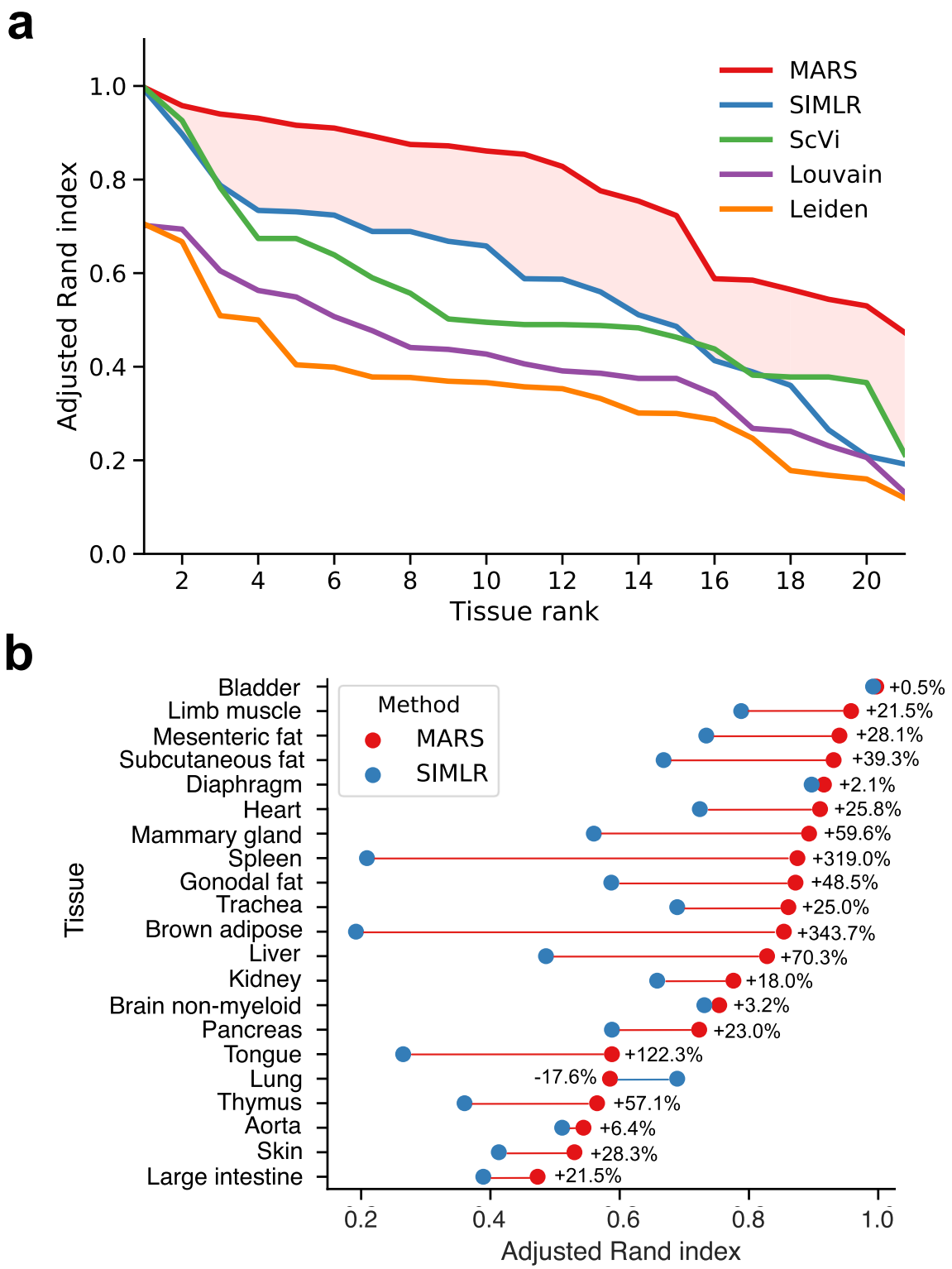



**Extended Data Fig. 2 | MARS outperforms other baselines and it is robust to embedding dimension.** Median performance of MARS and baseline methods evaluated using (**a**) adjusted mutual information (**b**) accuracy, (**c**) macro-F1-score, (**d**) macro-precision, and (**e**) macro-recall. For Leiden[33] and Louvain[34] we report adjusted mutual information and accuracy (Supplementary Note 3). Median is calculated across 21 different tissues. Error bars are standard errors estimated as a standard deviation of the mean by bootstrapping cells within tissue with n = 20 iterations. **f**, Median performance of MARS and K-means clustering applied in the latent space of the autoencoder at the end of the MARS pretraining. ARI stands for adjusted Rand index, F1 for macro-F1 score, and AMI for adjusted mutual information. Median is calculated across 21 different tissues. Error bars are standard errors estimated as a standard deviation of the mean by bootstrapping cells within tissue with n = 20 iterations. **g**, Performance of MARS when varying number of neurons in the last layer of the neural network which corresponds to the dimension of learned low-dimensional cell representation. Distribution is estimated with n = 20 runs of the method with different initial random seeds.

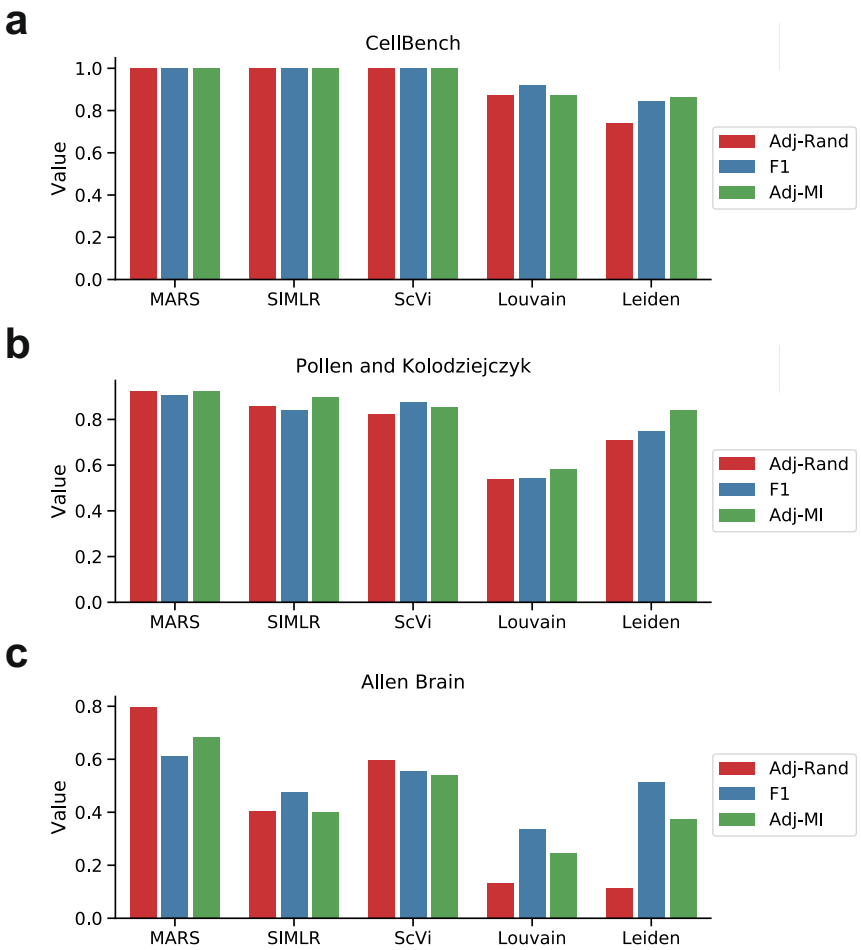



**Extended Data Fig. 3 | See next page for caption.**

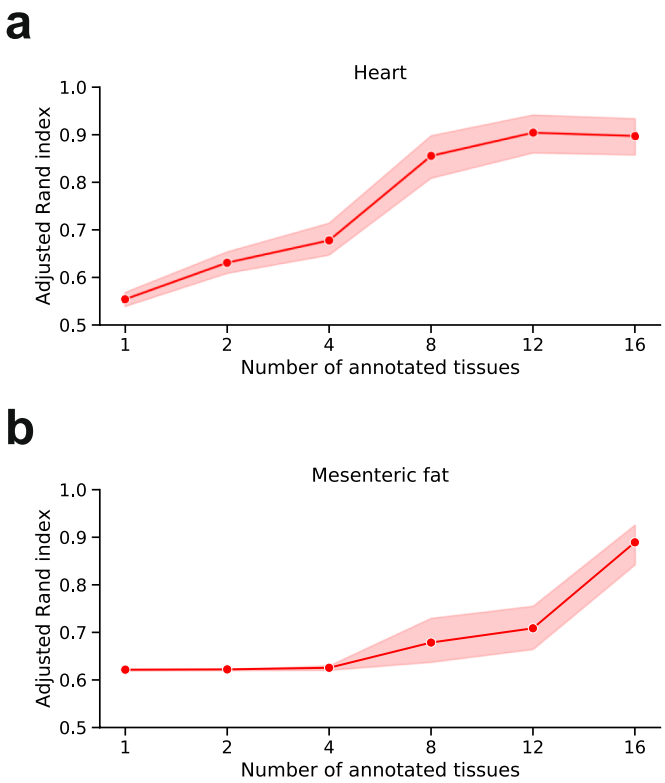
**Extended Data Fig. 3 | Cell-type level performance.** Cell-type level comparison of MARS's F1-score with the SIMLR[32] on (**a**) cell types that appear in only one tissue grouped by the number of differentially expressed genes in a cell type, (**b**) all cell types grouped by the number of differentially expressed genes in a cell type, and (**c**) all cell types grouped by the number of cells in a cell type. Standard errors are estimated as a standard deviation of the mean by bootstrapping cells within each tissue with n = 20 iterations. Number of differentially expressed genes is calculated using the Tabula Muris annotations by taking all genes with Benjamini-Hochberg FDR adjusted p-value<0.01 (two-tailed t-test).



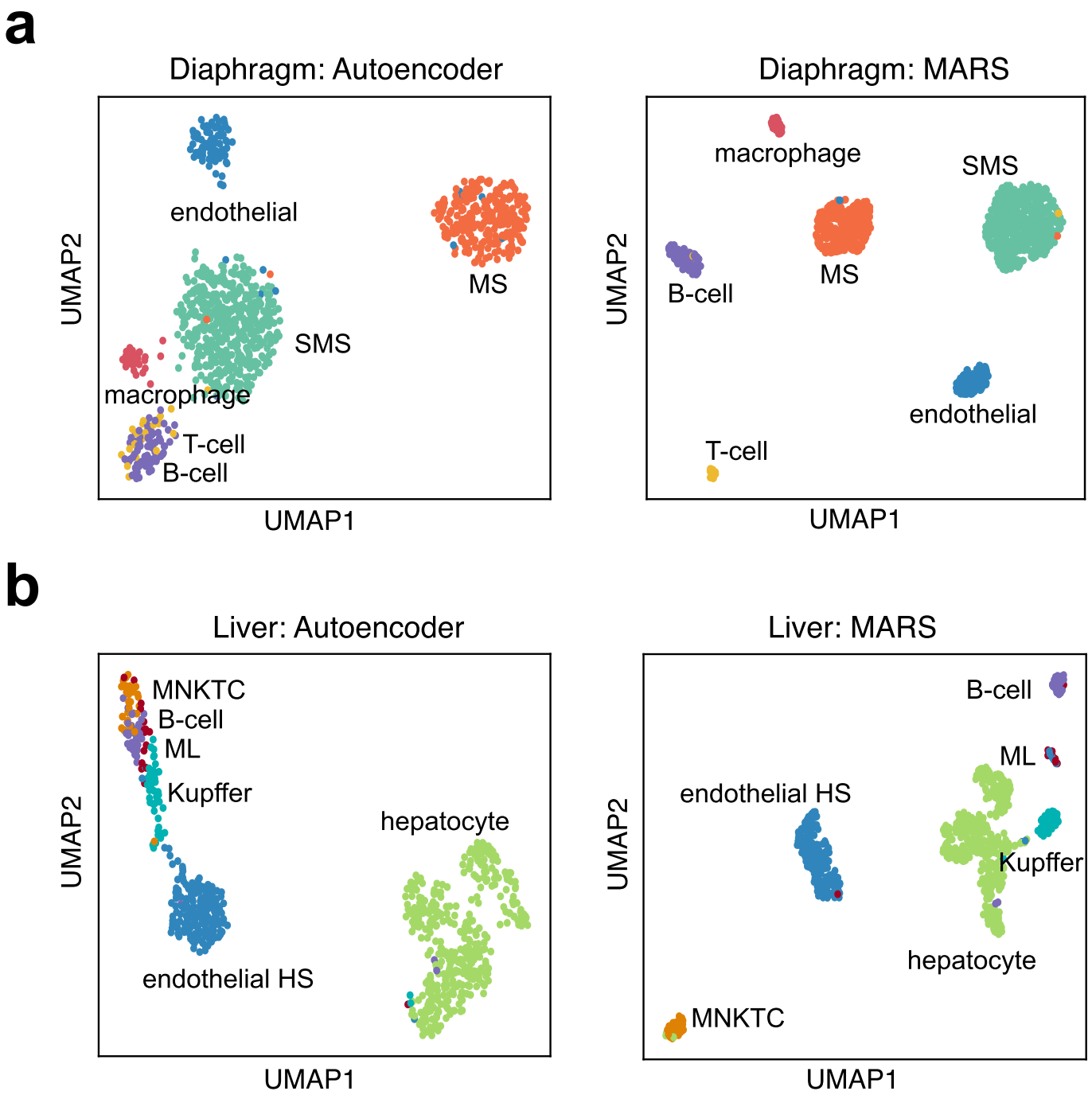

**Extended Data Fig. 4 | Tissue-level performance.** Comparison of the MARS's performance on individual tissues with the baseline methods. Performance is measured as adjusted Rand index score. **a**, Across all tissues, MARS achieves 34.3% higher area under the curve compared to the SIMLR, and 44.3% higher compared to the ScVi baseline. For each method, tissues are ranked in the decreasing order of the achieved score. **b**, Comparison with the second best performing method SIMLR[32] on individual tissues. MARS significantly outperforms SIMLR ($p = 1e-4$; two-tailed Wilcoxon signed-rank test). Tissues are ranked according to the MARS's ARI score. Performance is measured in a single run for both methods.



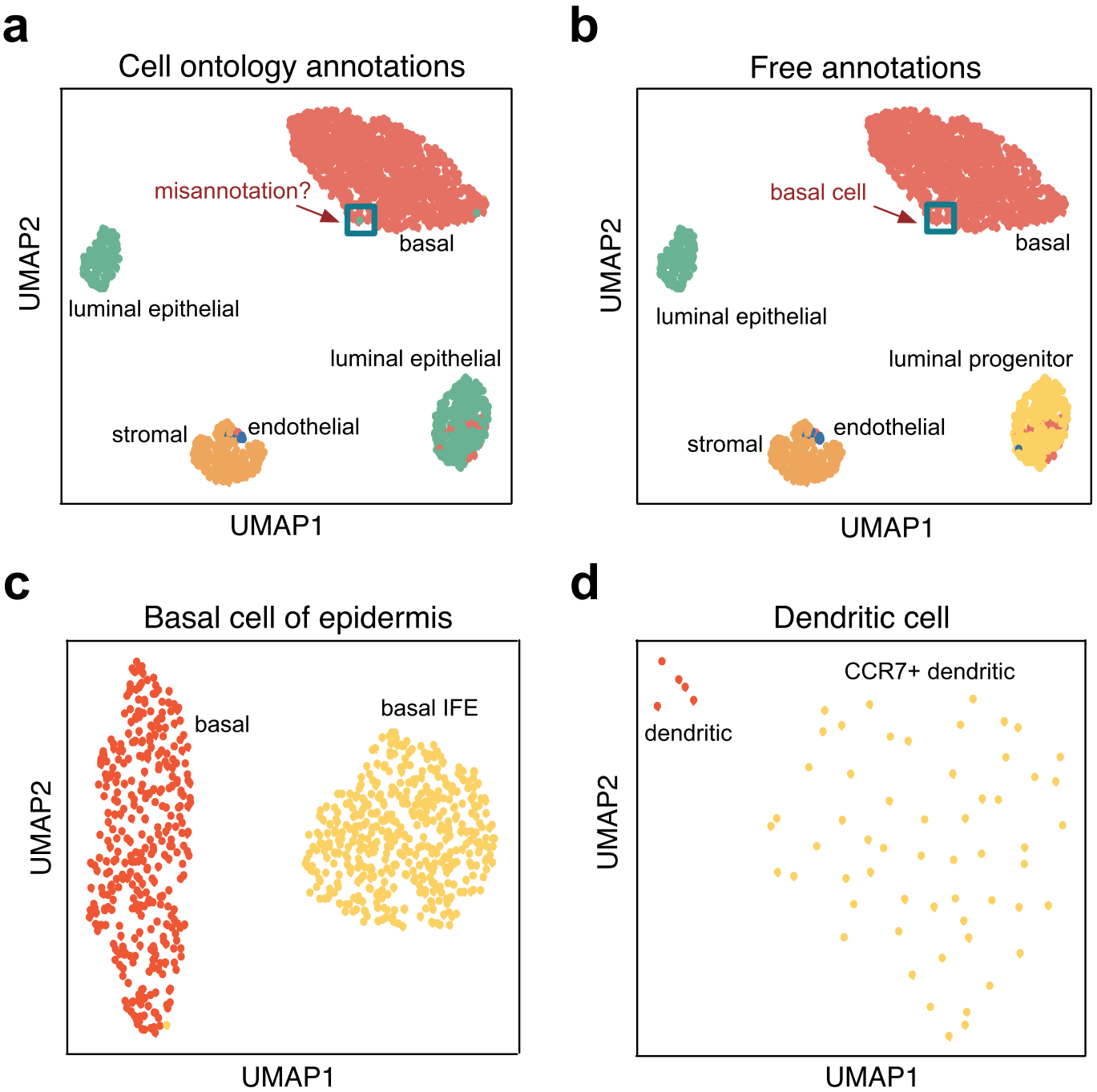

**Extended Data Fig. 5 | Performance on other datasets.** Mean performance of MARS and four baseline methods evaluated using adjusted Rand index (Adj-Rand), F1-score (F1) and adjusted mutual information (Adj-MI) on (**a**) two CellBench datasets[37,38] (**b**) Pollen[40] and Kolodziejczyk clustering benchmark datasets[41], and (**c**) three Allen Brain datasets[38,39]. For all metrics, higher value indicates better performance. MARS is trained in leave-one-dataset-out manner, and the held out dataset was completely unannotated.



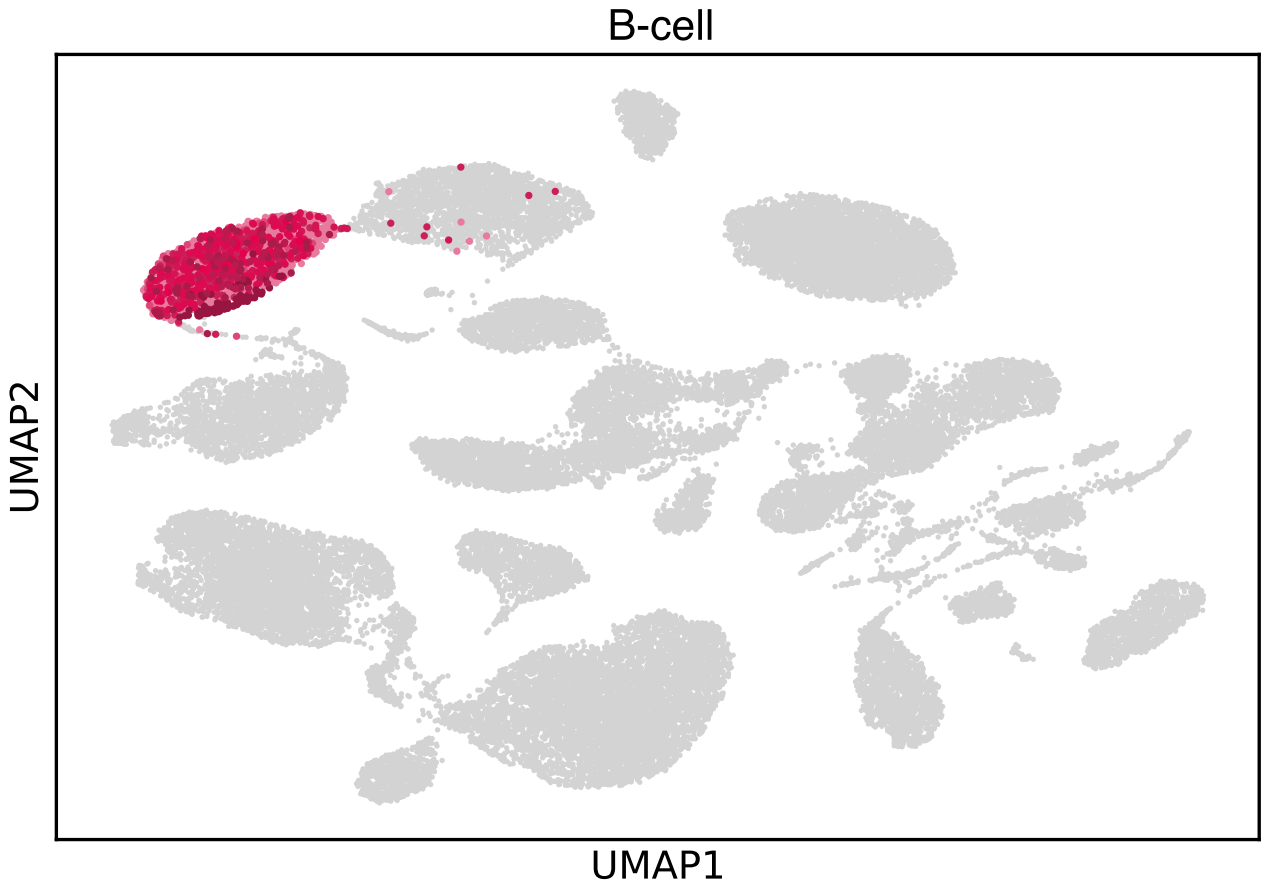

**Extended Data Fig. 6 | Positive knowledge transfer on heart and mesenteric fat tissues.** Effect of the number of annotated tissues in the meta-dataset on MARS's performance when using (**a**) heart tissue as unannotated experiment, and (**b**) mesenteric fat as unannotated experiment. Performance is measured as average adjusted Rand index across 20 runs of the method. Error bands are confidence intervals (95%) determined across 20 runs.



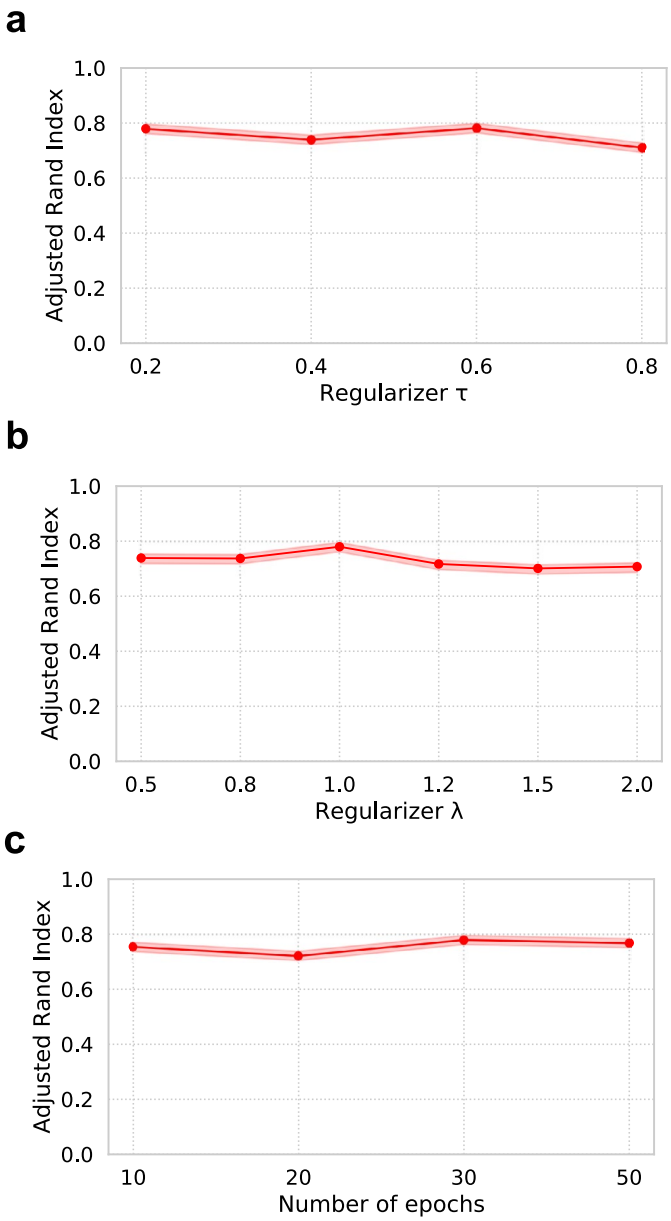

**Extended Data Fig. 7 | Embeddings after pretraining step.** UMAP visualizations of embeddings obtained with the MARS autoencoder (left), and the final MARS model (right) on (**a**) diaphragm tissue, and (**b**) liver tissue. Color indicates Tabula Muris cell type annotations. Autoencoder embeddings are obtained at the end of the MARS pretraining. Network parameters of the encoder and cluster centers from the K-means clustering are used to initialize MARS network and landmarks, respectively. MARS then learns new embeddings and new landmarks. SMS stands for skeletal muscle cell, MS for mesenchymal stem cell, HS for hepatic sinusoid, and MNKTC for mature NK T-cell.




**Extended Data Fig. 8 | MARS discovers cell subtypes. a**,**b**,UMAP visualization of MARS's embedding of mammary gland tissue cells. Colors indicate (**a**) Tabula Muris cell type annotations according to Cell Ontology class, and (**b**) free annotations in Tabula Muris that provide additional cell type resolution. Separation of cells labeled as luminal epithelial cells into two different clusters agrees perfectly with the free annotations and separate cluster found by MARS is labeled as luminal progenitor cells. MARS also correctly assigns one basal cell misannotated as luminal epithelial cells by Cell Ontology class annotations. **c**,**d**, UMAP visualization of MARS's embedding of subtypes of (**c**) basal cells of epidermis, and (**d**) dendritic cells. Colors indicate free annotations in Tabula Muris. We use all tissues as annotated experiments except the ones in which basal cells of epidermis or dendritic cells appear, and test the MARS ability to separate subtypes of these cell types. Clusters discovered by MARS perfectly agree with the free annotations.




**Extended Data Fig. 9 | Alignment of B cells.** Using MARS, B-cells in Tabula Muris data are extremely well aligned across 11 different tissues, including brown adipose tissue, diaphragm, gonodal fat, heart, kidney, limb muscle, lung, liver, mesenteric fat, subcutaneous fat, and spleen. Limb muscle is used as an unannotated tissue.




**Extended Data Fig. 10 | Robustness to hyperparameters.** MARS's performance when varying (**a**) regularizer λ, (**b**) regularizer τ, and (**c**) number of epochs. Performance is measured as average adjusted Rand index score. Average is calculated over all tissues by including each tissue as an unannotated dataset and using other tissues as annotated experiments. Error bars are standard errors estimated as a standard deviation of the mean by bootstrapping cells within tissue with n = 20 iterations. For each value, we train MARS with all other parameters fixed.

nature research

https://doi.org/10.1038/s41592-020-00979-3

## Supplementary information

# MARS: discovering novel cell types across heterogeneous single-cell experiments

In the format provided by the authors and unedited

Supplementary materials for

# MARS: Discovering Novel Cell Types across Heterogeneous Single-cell Experiments

Maria Brbić[1], Marinka Zitnik[2], Sheng Wang[3], Angela O. Pisco[4], Russ B. Altman[3,4], Spyros Darmanis[4], Jure Leskovec[1,4,*]

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[2]Department of Biomedical Informatics, Harvard University, Boston, MA 02115, USA
[3]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA
[4]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
[*]Corresponding author. Email: jure@cs.stanford.edu

This PDF file includes:

## Supplementary Note 1 Datasets and preprocessing

We downloaded Tabula Muris Senis [1] datasets with annotations from https://figshare.com/projects/Tabula_Muris_Senis/64982. Raw data for the Tabula Muris dataset is obtained from https://doi.org/10.6084/m9.figshare.5829687.v8. Due to the updated Tabula Muris annotations on the Tabula Muris Senis dataset, we used annotations for both datasets from Tabula Muris Senis. In particular, we used field *cell_ontology_class_reannotated* as cell type labels and selected only 3 months old data from Tabula Muris Senis to obtain Tabula Muris data. For cross-age transfer analysis, we used Tabula Muris Senis dataset for 3 months, 18 months and 24 months old mouse. For cross-tissue transfer analysis, we used Tabula Muris dataset. We observed that all methods are incapable to distinguish cell types in brain myeloid tissue that consists of microglial and macrophage cell types. The difficulty is biologically explainable by many shared molecular markers between these cell types, among which TMEM119 is the only known stable marker highly expressed by microglial cells but not expressed by macrophages [2]. Furthermore, microglial cells cover 99% of cells in brain myeloid of Tabula Muris dataset [3], making it hard to detect small macrophage cluster. For that reason, we did not include brain myeloid tissue in the analysis. At the time of writing the paper, marrow tissue annotations were not validated by expert so we did not perform any experiments on the marrow tissue. Pretraining of the MARS was performed on the unannotated tissue of the Tabula Muris Senis dataset.

## Supplementary Note 2 Baseline methods

We compared MARS to four unsupervised methods used for clustering single-cell data: Louvain [4], Leiden [5], SIMLR [6] and ScVi [7].

For Louvain and Leiden, we first performed PCA and retained $43$ principal components[3]. We computed neighborhood graph with number of neighbors set to $30$. We used Scanpy's implementations of Louvain and Leiden with the resolution parameter set to $1.0$.

For SIMLR and ScVi, we used implementations provided by the authors. For SIMLR we first performed PCA and retained $500$ principal components. Number of neighbors for constructing cell-cell similarity graph was set to $30$. For ScVi we first pretrained network with variational autoencoder for $150$ epochs. We tried two pretraining strategies: (i) pretraining on the same data as MARS (only unannotated tissues from Tabula Muris Senis), and (ii) pretraining on the Tabula Muris data from all tissues. The latter achieved better performance, so we used that setting in all our analysis and comparison with the ScVi method. After pretraining, we used all parameters recommended by the authors. Specifically, we trained network for $200$ epochs with learning rate $0.001$ and Adam optimizer. Neural network consisted of two layers with widths $128$ and $32$. To obtain clustering assignments, we applied K-means clustering on the learned cell embeddings. Since K-means depends on initialization, we reported mean score across $20$ runs.

## Supplementary Note 3 Evaluation

We evaluated MARS and other baselines on the Tabula Muris dataset [3] using six different metrics: adjusted Rand index, adjusted mutual information, accuracy, macro-F1-score, macro-precision and macro-recall. MARS is solving unsupervised/clustering task on the unannotated dataset and each cell type in the unannotated dataset (previously seen or unseen) gets a new landmark. Therefore, each discovered cell type corresponds to a new cluster and we evaluated MARS as a clustering method. For adjusted Rand index and adjusted mutual information, we compared clusters obtained using MARS and other baselines to ground truth cell type annotations where each cell type corresponds to a different cluster. For accuracy, F1-score, precision and recall we first solve optimal assignment problem using Hungarian algorithm [8]. Once estimated clusters are assigned to the ground-truth cell type annotations, we calculated accuracy, macro-F1-score, macro-precision and macro-recall.

For Louvain and Leiden we reported only adjusted Rand index, adjusted mutual information and accuracy. The reason is that these methods often lead to overclustering and additionally assigned clusters are not matched to ground truth annotations during assignment with the Hungarian algorithm. Clustering metrics such as adjusted Rand index and adjusted mutual information do not suffer from this drawback, and clearly demonstrate that all other baselines significantly outperform Louvain and Leiden.

When evaluating performance on never-before-seen cell types, we selected 63 cell types that appear in only one tissue, and conservatively filtered all cell types with 'endothelial','basal','smooth muscle','epithelial','B cell','T cell','fibroblast','mesenchymal','macrophage' in their name, except 'kidney collecting duct epithelial cell' and 'kidney loop of Henle ascending limb epithelial cell'. For instance, we filtered 'fibroblast of lung','lung macrophage','regulatory T cell','respiratory basal cell','pancreatic B cell','smooth muscle cell of trachea','epithelial cell of thymus'.

## Supplementary Note 4 Benchmark datasets

We tested MARS on three benchmark datasets: (i) two CellBench datasets [9], (ii) three Allen brain datasets [10], and (iii) two clustering benchmark datasets consisting of diverse human cell types (Pollen [11]) and mouse pluripotent cells (Kolodziejczyk [12]).

Two CellBench datasets consist of five sorted lung cancer cell lines sequenced with 10X and CEL-Seq2 protocols. Three Allen brain datasets VISp, ALM, and MTG, consist of mouse and human species, as well as single-cell RNA-seq and single-nucleus RNA-seq datasets. We use coarse-grained cell type annotations (excitatory, inhibitory and non-neuronal). Pollen dataset consists of 11 diverse human cell types: skin cells, pluripotent stem cells, blood cells, and neural cells. Kolodziejczyk dataset consists of three cell types of mouse pluripotent cells.

Within each benchmark dataset, we joined individual datasets by taking common subset of genes. For CellBench dataset we obtained 4,373 cells and 10, 217 genes, for Pollen and Kolodziejczyk we obtained 953 cells and 8,138 genes, whereas for Allen Brain datasets were already joined [13], resulting in 34,735 cells and 16,024 genes. We normalized each cell to 10,000 read counts, and then jointly scaled the datasets to unit variance and zero mean, truncating values with maximum value set to 10. We used Scanpy [14] for preprocessing. In all three benchmarks, we considered each individual dataset as a separate experiment and trained MARS in leave-one-experiment-out manner. First two benchmarks have exactly the same set of cell types across experiments, while the third benchmark requires generalization to novel cell types.

CellBench and Allen Brain datasets [13] were downloaded from https://doi.org/10.5281/zenodo.3357167). Originally, three brain datasets, Allen Mouse Brain (AMB), VISp, ALM (GSE115746), are from the Allen Institute Brain Atlas http://celltypes.brain-map.org/rnaseq. The CellBench 10X dataset is from (GSM3618014), and the CellBench CEL-Seq2 dataset is from 3 datasets (GSM3618022, GSM3618023, GSM3618024) and joined into one dataset [13]. We downloaded Pollen and Kolodziejczyk datasets from https://github.com/BatzoglouLabSU/SIMLR.

## Supplementary references


1. Almanzar, N. *et al.* A single cell transcriptomic atlas characterizes aging tissues in the mouse. *Nature* **583**, 590–595 (2020).

2. Bennett, M. L. *et al.* New tools for studying microglia in the mouse and human cns. *Proceedings of the National Academy of Sciences* **113**, E1738–E1746 (2016).

3. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature* **562**, 367 (2018).

4. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

5. Traag, V. A., Waltman, L. & van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9** (2019).

6. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods* **14**, 414 (2017).

7. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053 (2018).

8. Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly* **2**, 83–97 (1955).

9. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods* **16**, 479–487 (2019).

10. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

11. 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 (2014).

12. Kolodziejczyk, A. A. *et al.* Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).

13. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology* **20**, 194 (2019).

14. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).