Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery

Jingshu Wang[1], Divyansh Agarwal[2], Mo Huang[1], Gang Hu[3], Zilu Zhou[2], Vincent Conley[4], Hugh MacMullan[4], Nancy R. Zhang[1*]

1) Department of Statistics, University of Pennsylvania, Philadelphia, PA
2) Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA
3) School of Mathematical Sciences, Nankai University, Tianjin, China
4) High Performance Computing and Research IT, The Wharton School, Philadelphia, PA

* Correspondence:
Nancy R. Zhang
nzh@wharton.upenn.edu
(215) 898-8007
Department of Statistics
The Wharton School
University of Pennsylvania

Although single-cell RNA sequencing (scRNA-seq) technologies have shed light on the role of cellular diversity in human pathophysiology[1–3], the resulting data remains noisy and sparse, making reliable quantification of gene expression challenging. Here, we show that a deep autoencoder coupled to a Bayesian model remarkably improves UMI-based scRNA-seq data quality by transfer learning across datasets. This new technology, SAVER-X, outperforms existing state-of-the-art tools. The deep learning model in SAVER-X extracts transferable gene expression features across data from different labs, generated by varying technologies, and obtained from divergent species. Through this framework, we explore the limits of transfer learning in a diverse testbed and demonstrate that future human sequencing projects will unequivocally benefit from the accumulation of publicly available data. We further show, through examples in immunology and neurodevelopment, that SAVER-X can harness existing public data to enhance downstream analysis of new data, such as those collected in clinical settings.

1   Highly parallelized scRNA-seq pipelines are now becoming the standard. In many current and

2   proposed studies, thousands to millions of cells are sequenced, with each cell receiving low

3   coverage. At low coverages of 500-1000 unique molecular identifiers (UMI) per cell, precise

4   distinctions between cell states are blurred and genes with low expression cannot be accurately

5   quantified. To address this challenge, methods have been developed to de-noise and impute

6   scRNA-seq data [4–7]. These methods, however, may not perform well when sequencing is done at

7   extremely low depth, or when applied to cell types that are rare. Notably, existing denoising

8   techniques act solely upon the data from a given study and ignore existing datasets in public

9   domain, which may contain similar cell types.

10   In light of the Human Cell Atlas initiative [8], the scientific community will soon have detailed atlases

11   for each anatomic organ in the human body; for the laboratory mouse, such an atlas (*Tabula Muris*)

12   was recently unveiled [9]. Accumulation of publicly available scRNA-seq data presents an opportunity

13   to leverage existing data in the denoising of a new scRNA-seq data set. Yet, it is unclear how much

14   information can be borrowed across datasets which might be generated using different platforms,

15   wherein samples are processed differently or at different coverages. Moreover, such transfer

16   learning must guarantee that the denoising process will not introduce bias or force the new data to

17   lose its distinctive features and conform to the patterns in existing data.

18   Here we describe a denoising framework, called Single-cell Analysis via Expression Recovery

19   harnessing eXternal data (SAVER-X). It uses the deep autoencoder, a neural network that achieves

20   noise reduction by means of an information bottleneck [10]. Consider a target dataset to be denoised.

21   The autoencoder can be trained on this data starting either from random initialization of the weights,

22   as in other denoising tools like DCA [7], or from weights obtained by training on existing public data

23   sets (pre-training data; Figure 1b) with related cell types. The latter – initialization by pre-trained

24   weights followed by refinement on the test data – transfers information from public data to a user's

25   current dataset.

26   A pivotal challenge in transfer learning lies in finding the balance between transferring extensively

27   and/or scarcely from the pre-training data. While the former would result in the user data losing its

2

1   own distinctive features, the latter risks producing inappreciable improvements. SAVER-X

2   adaptively achieves an appropriate amount of information transfer by refining and updating weights

3   to fit the test data (Figure 1a, item 2A). This drives the model away from the pre-training data.

4   Subsequently, cross-validation is used to identify genes that are poorly fit by the autoencoder, and

5   the autoencoder output for these genes is replaced by their mean expression values (Figure 1a,

6   item 2B). Finally, for each gene in every cell, SAVER-X computes a weighted average of the fitted

7   value and the observed normalized count [4] (Figure 1a, item 2C). This weighted average is the

8   posterior mean of the gene's expression in the given cell based on a Bayesian hierarchical model,

9   assuming the Poisson-alpha technical noise model [11]. SAVER-X outputs the mean denoised

10 values, which can then feed into downstream analyses.

11 Many core cell types and essential pathways are shared between human and mouse, and,

12 importantly, experiments can be performed more readily on mice than on human subjects [8,12,13].

13 Thus, effective mouse to human transfer learning invites new ways to use mouse as a model

14 organism. To enable cross-species data sharing, the autoencoder in SAVER-X consists of three

15 sub-networks with human-specific, mouse-specific and human-mouse shared nodes (Figure 1c).

16 The human-mouse shared network receives human-mouse homologous genes as input. To adjust

17 for the differences between data generated using non-UMI- and UMI-based technologies, an

18 indicator node at the input layer feeds into each sub-network.

19 SAVER-X is publicly available at http://singlecell.wharton.upenn.edu/saver-x/, where users can

20 choose from models pre-trained on 31 mouse tissues and human immune cells. Models jointly

21 pretrained on cells from both species are also available for brain and pancreatic tissues. For

22 additional details on SAVER-X architecture and estimation, see Online Methods.

23 Cells that constitute the immune system are implicated in virtually every disease. While

24 understanding the features of infiltrating immune cells in an inflamed tissue is of critical importance,

25 their representation is often small in scRNA-seq studies in the absence of flow sorting. Thus,

26 denoising the observed values without relying on external data becomes especially challenging [14,15].

27 We demonstrate that SAVER-X can perform transfer learning for immune cells between healthy and

1   disease conditions. By pre-training SAVER-X on scRNA-seq data from the Human Cell Atlas (HCA)

2   project [8] (500,000 immunocytes from umbilical cord blood and bone marrow) and 10X Genomics

3   website [16] (200,000 peripheral blood Mmononuclear cells), we were able to meaningfully improve

4   the data quality in other scRNA-seq studies that profiled immune cells.

5   First, we evaluated SAVER-X with and without pre-training against existing denoising methods on a

6   set of purified cells from 9 immune cell types [16]. We created a "test" dataset by randomly selecting

7   100 cells for each cell type (Online Methods). Among this set of 900 immune cells, with an average

8   UMI count of roughly 1200 per cell, neither is it easy to visually distinguish NK cells from T-cells, nor

9   can T-cell subtypes be separated (Figure 2a). SAVER-X imputation of this dataset, without using

10   any existing data for pretraining, enhances the separation of NK cells from T-cells. Although the

11   visualization of intra T-cell subtype heterogeneity also improves, the subtypes remain difficult to

12   identify. The impact of transfer learning becomes apparent when we denoise the test data using

13   SAVER-X pre-trained on the HCA data (Figure 2a). The Adjusted Rand Index (ARI) improves

14   significantly as CD8$^+$ T-cells clearly separate from CD4$^+$ T-cells, and naïve CD4$^+$ T-cells become

15   distinguishable from other subtypes. This observation suggests that significant information about

16   cell type-specific transcriptional signatures can be transferred between datasets even when the cell

17   types belong to different tissues and are prepared in different laboratories. Furthermore, SAVER-X

18   pre-trained on both HCA and 10X data led to a distinct separation between CD4$^+$ memory T-cells

19   and regulatory T-cells (T$_{regs}$). The 10X dataset obtained from peripheral blood mononuclear cells

20   (PBMCs) contains ~120,000 T-cells, and a SAVER-X model trained specifically on these cells

21   further improves the separation of T-cells subtypes (Fig. 2a). Reliable detection of T-cell subtypes is

22   crucial to the characterization of a tissue's immune environment. For instance, naïve CD4$^+$ T-cells

23   help maintain immune competence throughout life [17], and yet the mechanisms underlying their

24   establishment and maturation remain elusive. SAVER-X allows us to confidently identify this sub-

25   population and study its homeostasis, which is ultimately critical for clinical applications in both

26   vaccination and immune reconstitution.

1   The potential of transfer learning in biology hinges on its ability to adapt to diverse and practical

2   settings. Thus, we explored if SAVER-X can effectively learn from healthy HCA cells in the

3   denoising of immune cells sequenced from primary breast carcinoma samples from eight treatment-

4   naïve patients [18]. SAVER-X, pre-trained on publicly available immune cell datasets to denoise the

5   tumor tissue-resident immune cells, not only allowed us to better characterize immune cell types,

6   but also clarified the expression patterns of marker genes (Figure 2b, Figure S3) in these patients.

7   This improved reconstruction of the tumor immune microenvironment typifies the potential gains

8   achievable by transfer learning from accumulating public data.

9   We further assessed the utility of SAVER-X in scenarios where either the number of cells

10   sequenced could be small (less than 100), or the sequencing depth might be too low (60 UMIs per

11   cell; Table S1). Currently, cells with such low coverage are typically discarded. We show that

12   SAVER-X not only salvages such data, but also extracts useful information about gene-gene

13   relationships (Figure S2). We benchmarked SAVER-X against other scRNA-seq denoising methods

14   that do not employ transfer learning, *viz.,* DCA [7], scImpute [5] and MAGIC [6], and found that SAVER-X

15   significantly outperforms existing methods (Figure 2c, Figure S1) in most scenarios. As expected,

16   for separating major cell types, the benefits of transfer learning diminish when a large number of

17   cells are sequenced at a sufficiently high depth. As a denoising method, SAVER-X enables

18   improved gene-level analysis; for instance, as gene expression becomes less sparse, selection of

19   important regulatory and marker genes becomes more reliable (Figure 2d).

20   Having demonstrated that SAVER-X effectively transfers information across labs and from healthy

21   to disease settings, we next examined the feasibility of transfer learning across species. Mouse

22   models have helped scientists understand the basis of several human disorders, and although

23   transcriptomic patterns in mouse might not always provide a direct route to the cognate human

24   condition, similarities and disparities of genetic programs, once understood, are likely to provide a

25   deeper understanding of the fundamental architecture underlying cellular development and

26   physiology. In this regard, the ability to harness mouse data in the denoising of human data

27   represents a new mode of cross-species learning. We examined scRNA-seq data from cells in the

1   developing ventral midbrain of both human and mouse, and found that, indeed, SAVER-X pre-

2   trained on mouse scRNA-seq data enhances the quality of the human data (Figure 3).

3   First, we reduced the high coverage human ventral midbrain scRNA-seq data by sampling only 10%

4   of the reads [13], to a median per cell coverage of 452 UMIs. To compare the gains achievable by

5   intra- and inter-species transfer learning, we split the human cells randomly into two groups, down-

6   sampled one group and used the other group as the pre-training data (Figure 3a). SAVER-X pre-

7   trained on the matched mouse brain cells led to a distinct improvement in cell type identification for

8   human compared with the un-pretrained model, affirming the potential of transfer learning across

9   species (Figure 3b). We found that a model jointly pre-trained on both human and mouse data

10   further augments the human scRNA-seq data quality compared with pre-training on the human cells

11   alone. Remarkably, pre-training SAVER-X on cells from regions other than the ventral mid-brain

12   using the *Tabula Muris* [9] also improved the ARI (Figure 3b). We then pre-trained SAVER-X on three

13   human non-UMI datasets [19–21], and found that the model jointly pre-trained using both the non-UMI

14   human cells and mouse cells outperforms training on either species alone (Figure S4a). These

15   observations suggest that SAVER-X prevents negative transfer of information between species by

16   harnessing the heterogeneity among public datasets. Data heterogeneity forces SAVER-X to learn

17   robust low-dimensional representation of information, which likely contains the true biological

18   signals that are shared across studies.

19   To further demonstrate that SAVER-X does not unnaturally bias data denoising, we examined

20   whether a model pre-trained on mouse data affects human-specific patterns. We denoised human

21   scRNA-seq data using the matched mouse data, and then compared the log fold-change of the

22   genes differentially expressed between human and mouse for each cell type before and after

23   denoising. We found that the fold changes are indeed preserved, suggesting that SAVER-X

24   introduces negligible bias (Figure 3c). On the other hand, simply relying on an autoencoder, without

25   gene filtering or Bayesian shrinkage, reduces the fold change between human and mouse for some

26   genes in some cell types (Figure S4b). This highlights the importance of balancing the autoencoder

27   predictions against the observed data to prevent bias.

1   Taken together, our results demonstrate that the transfer learning framework employed by SAVER-

2   X can leverage existing scRNA-seq datasets to improve the quality of new scRNA-seq data across

3   UMI-based sequencing platforms, species, organs and cell types. At its core, SAVER-X trains a

4   deep neural network on scRNA-seq data across a range of study designs and applies this model to

5   new data to strengthen shared biological patterns. This general framework for inferring "true"

6   relationships from raw and error-prone experimental data will be broadly applicable in other high-

7   throughput settings. Through applications in immunology and developmental neuroscience, we

8   show that SAVER-X can improve cell type classification and gene expression characterization in

9   both healthy and disease settings. With increasing accumulation of publicly available data, SAVER-

10 X will increase in generalization accuracy and in tissue- and cell-type specificity. A technology like

11 SAVER-X changes the approach to scRNA-seq data analysis from a process of study-specific

12 quality control and statistical modeling to an automated process of cross-study data integration and

13 information sharing.

14

Methods

Collection of public datasets

17 The Human Cell Atlas (HCA) dataset was downloaded from the HCA data portal

18 (https://preview.data.humancellatlas.org/) and the PBMC data was downloaded from the 10X

19 website (https://support.10xgenomics.com/single-cell-gene-expression/datasets, Table S2). The

20 purified data for each immune cell type was also downloaded from the 10X website [16]. The breast

21 cancer data [18] was downloaded from GEO (GSE114725). The developing midbrain data [13] was

22 downloaded from GEO (GSE76381). For other mouse developing brain datasets, we include cells

23 from neonatal and fetal brain tissues in the *Tabula Muris* [9] data (GSE108097). For the other non-

24 UMI human developing brain datasets, we include three: GSE75140 [20], GSE104276 [21] and

25 SRP041736 [19]. No filtering is done on the original data and all genes and cells provided in the

26 original datasets were used by SAVER-X.

1   A complete list of the pre-training datasets used for pre-training the models on the SAVER-X

2   website is provided in Table S2.

Details of SAVER-X

4   SAVER-X uses a Bayesian hierarchical model to combine evidence from the raw read counts of a

5   new data set with predictions made by an autoencoder. The autoencoder can be trained exclusively

6   on the new data set, or first pre-trained on existing data sets and then on the new data set. The

7   autoencoder used by SAVER-X has three subnetworks, as shown in Figure 1a, with one

8   subnetwork taking human genes as input, one subnetwork taking mouse genes as input, and one

9   subnetwork taking shared human-mouse homologous genes as input. 21183 and 21122 genes are

10   used for human and mouse (Supplementary Data, Supplementary Note), respectively, as input and

11   output nodes of the autoencoder. By current annotations using the getLDS() function in the bioMaRt R

12   package, 15494 genes have homologs shared between the two species (Supplementary Data,

13   Supplementary Note). For each sub-network, the number of nodes in the encoding and decoding

14   layers are, successively, 128, 64, 32, 64, and 128. If only human data are available, only the human

15   and shared sub-network weights are updated. Similarly, if only mouse data are available, only the

16   mouse and shared sub-network weights are updated.

17   For UMI datasets, let the raw UMI count for each cell c and gene g be x_{cg}, then the input expression

18   levels are normalized by library size, re-scaled and log-transformed using formula: $\tilde{x}_{cg} =$

19   $\log(x_{cg}/l_c \times 10000 + 1)$ where $l_c = \sum_g x_{cg}$ is the library size of cell c. For non-UMI datasets, TPM

20   for each cell c and gene g are denoted as x_{cg}, and then the x_{cg} are transformed using the same

21   formula as that for UMI. If a gene is missing in the dataset, the input is set to 0 while the

22   corresponding output node is not accounted for in the loss function. Specifically, let the output value

23   for gene g be defined as $\hat{x}_{cg}$, which we refer to as the autoencoder prediction. Conditional on the

24   prediction $\hat{x}_{cg}$, the observed UMI count is assumed to follow a Negative Binomial distribution. Thus,

25   for each cell c, the loss function for UMI-based counts is defined as the sum of log likelihoods:

8

1
$$L(x_c, \hat{x}_c) = \sum_{\text{node } g \text{ presents in the dataset}} \log[\text{NB}(x_{cg}; l_c\hat{x}_{cg}, \theta_g^U)]$$

2 On the other hand, TPM data is assumed to approximately follow a zero inflated Negative Binomial

3 distribution (although TPM is not integer-valued, the likelihood function can still be computed) and

4 the loss is defined as:

5
$$L(x_c, \hat{x}_c) = \sum_{\text{node } g \text{ presents in the dataset}} \log[\text{ZINB}(x_{cg}; \hat{x}_{cg}, \theta_g^{NU}, \pi_{cg})]$$

6 where $\text{NB}(x; \mu, \theta)$ and $\text{ZINB}(x; \mu, \theta, \pi)$ are the density of Negative Binomial and zero-inflated

7 Negative Binomial distributions (see Supplementary Note). A separate gene-specific dispersion

8 parameter θ_g^U and θ_g^{NU} is dedicated for UMI and non-UMI input, respectively. For non-UMI data, the

9 gene- and cell-specific zero inflation parameter is defined as

10
$$\frac{\pi_{cg}}{1 - \pi_{cg}} = w_g \log \hat{x}_{cg} + b_g$$

11 Our implementation of the autoencoder builds on top of the source code of DCA [7], using its library

12 functions.

13 Although SAVER-X accepts pre-training data both with and without UMI, **the target data must**

14 **have UMI**. When SAVER-X is applied to the denoising of a UMI-based target data matrix, the

15 following steps are applied (Figure 1a): (1) The autoencoder is fit on the target data, optionally

16 starting with a user-selected pre-trained model. (2) Cross-validation is applied to filter out genes

17 that cannot be predicted well by the autoencoder. Specifically, the target data is randomly split into

18 held-in and held-out cell sets, the autoencoder is trained on the held-in set and then used to make

19 predictions on the held-out set. For a specific gene g, let the normalized predictions using the held-

20 in set trained model on a held-out cell c is $\hat{x}_{cg}$ and let the held-in sample mean for the library-size

21 normalized counts be μ_g. Then a gene is unpredictable if the Poisson deviance of the predictions

22 and original UMI counts of the held-out samples is larger than that of the held-in sample mean and

23 held-out original UMI counts, equivalently:

1. $$-\sum_{c \text{ is a held-out cell}} \left[x_{cg} \log(l_c \hat{x}_{cg} + \varepsilon) - l_c \hat{x}_{cg}\right] > -\sum_{c \text{ is a held-out cell}} \left[x_{cg} \log(l_c \mu_g + \varepsilon) - l_c \mu_g\right]$$

2. where $\varepsilon = 10^{-10}$ to avoid taking the log of zeros. After unpredictable genes are identified, the

3. autoencoder is trained again on all the cells, but the predicted values of the unpredictable genes or

4. genes that are not present in the nodes are replaced with the sample mean of library size

5. normalized UMI counts. (3) After we obtain these predicted values, we apply empirical Bayes

6. shrinkage, following the model used by SAVER [4]. In SAVER, let λ_{cg} be the true relative expression

7. level of the gene that we want to recover, then we assume

8. $$x_{cg} \sim \text{Poisson}(l_c \lambda_{cg}), \qquad \lambda_{cg} \sim \text{Gamma}(\alpha_{cg}, \beta_{cg})$$

9. where β_{cg} is the rate parameter, $\alpha_{cg} = \hat{x}_{cg} \beta_{cg}$ is the shape parameter and $\hat{x}_{cg}$ is the filtered

10. autoencoder prediction. The final denoised expression level of a gene in each cell is a weighted

11. average of the autoencoder predicted value and its observed UMI count:

12. $$\hat{\lambda}_{cg} = \frac{l_c}{l_c + \hat{\beta}_{cg}} \frac{x_{cg}}{l_c} + \frac{\hat{\beta}_{cg}}{l_c + \hat{\beta}_{cg}} \hat{x}_{cg}$$

13. Where $\hat{\beta}_{cg}$ is obtained by maximizing the likelihood in SAVER.

14. **Data denoising using other bench-marking methods**

15. MAGIC [6] was performed using the R version 1.3.0 on the square root transformed mean library-size

16. normalized expression. scImpute [5] version 0.0.9 was performed on the unnormalized expression

17. values setting Kcluster = 9. DCA [7] version 0.2.2 was performed on the unnormalized expression

18. values and the library-size normalized expression output was used for downstream analysis.

19. **Generating down-sampled datasets**

20. For an observed UMI count data matrix, we down-sample the reads to obtain a data set of the same

21. gene and cell numbers but with lower quality. For cell c and gene g, the down-sampled value y_{cg} is

22. generated by independently drawing from a Poisson distribution with $y_{cg} \sim \text{Poisson}(\tau_c x_{cg})$ where τ_c

1   is a cell-specific efficiency loss. To mimic variation in efficiency across cells, we sampled τ_c as

2   follows:

3   1. 10% efficiency: $\tau_c \sim \text{Gamma}(10, 100)$, used on the mouse midbrain data [13]

4   2. 5% efficiency: $\tau_c \sim \text{Gamma}(5, 100)$, used on the 10X PBMC data [16]

5   **t-SNE visualization and cell clustering**

6   We used Seurat version 2.0 to perform cell clustering and t-SNE visualization according to the

7   workflow detailed at (https://satijalab.org/seurat/pbmc3k_tutorial.html). For all analyses, we set the

8   number of principal components to 15. For cell clustering using Seurat, resolution is set to be 1.6,

9   1.2, 0.8 and 0.8 for each of the four experiments (90 cells, 900 cells, 9000 cells and 9000 cells with

10   down-sampled reads) of the PBMC data and kept the same on all the methods compared. The

11   resolution is set to 1 for the cell clustering of the midbrain data [13]. The adjusted Rand Index (ARI) is

12   computed using R package mclust.

13   **Differential expression analysis**

14   Differentially expressed genes between human and mouse for each cell type of the developing

15   midbrain are obtained also using Seurat 2.0, where the Wilcoxon rank sum test is used. P-value

16   adjustment is performed using Bonferroni correction based on the total number of genes in the

17   dataset. A gene is selected as differentially expressed if its adjusted p-value is ≤ 0.05 and the

18   absolute log fold change is ≥ 0.25.

19

20   **Competing Interests**

21   The authors declare no competing interests

22

23   **References**

1.   1.   Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heteogeneity in primary glioblastoma. *Science (80-.).* **344,** 1396–1402 (2014).

2.   2.   Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24,** 593–607 (2016).

3.   3.   Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24,** 978–985 (2018).

4.   4.   Huang, M. *et al.* SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15,** 539–542 (2018).

5.   5.   Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9,** 1–9 (2018).

6.   6.   van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174,** 716–729.e27 (2018).

7.   7.   Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. DCA: Single cell RNA-seq denoising using a deep count autoencoder. *bioRxiv* 300681 (2018). doi:10.1101/300681

8.   8.   Regev, A. *et al.* The human cell atlas. *Elife* **6,** 1–30 (2017).

9.   9.   Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172,** 1091–1097.e17 (2018).

10.   10.   Hinton, G. E. . & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science (80-.).* **313,** 504–507 (2006).

11.   11.   Wang, J. *et al.* Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci.* **115,** E6437–E6446 (2018).

12.   12.   Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3,** 346–360 (2016).

13.   13.   La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and

12

Stem Cells. *Cell* **167,** 566–580.e19 (2016).

14. Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. Single-Cell Genomics: Approaches and Utility in Immunology. *Trends Immunol.* **38,** 140–149 (2017).

15. Nguyen, A., Khoo, W. H., Moran, I., Croucher, P. I. & Phan, T. G. Single cell RNA sequencing of rare immune cell populations. *Front. Immunol.* **9,** (2018).

16. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** 1–12 (2017).

17. Silva, S. L. & Sousa, A. E. Establishment and Maintenance of the Human Naïve CD4+ T-Cell Compartment. *Front. Pediatr.* **4,** 1–10 (2016).

18. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174,** 1293–1308.e36 (2018).

19. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–1058 (2014).

20. Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci.* **112,** 201520760 (2015).

21. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555,** 524–528 (2018).

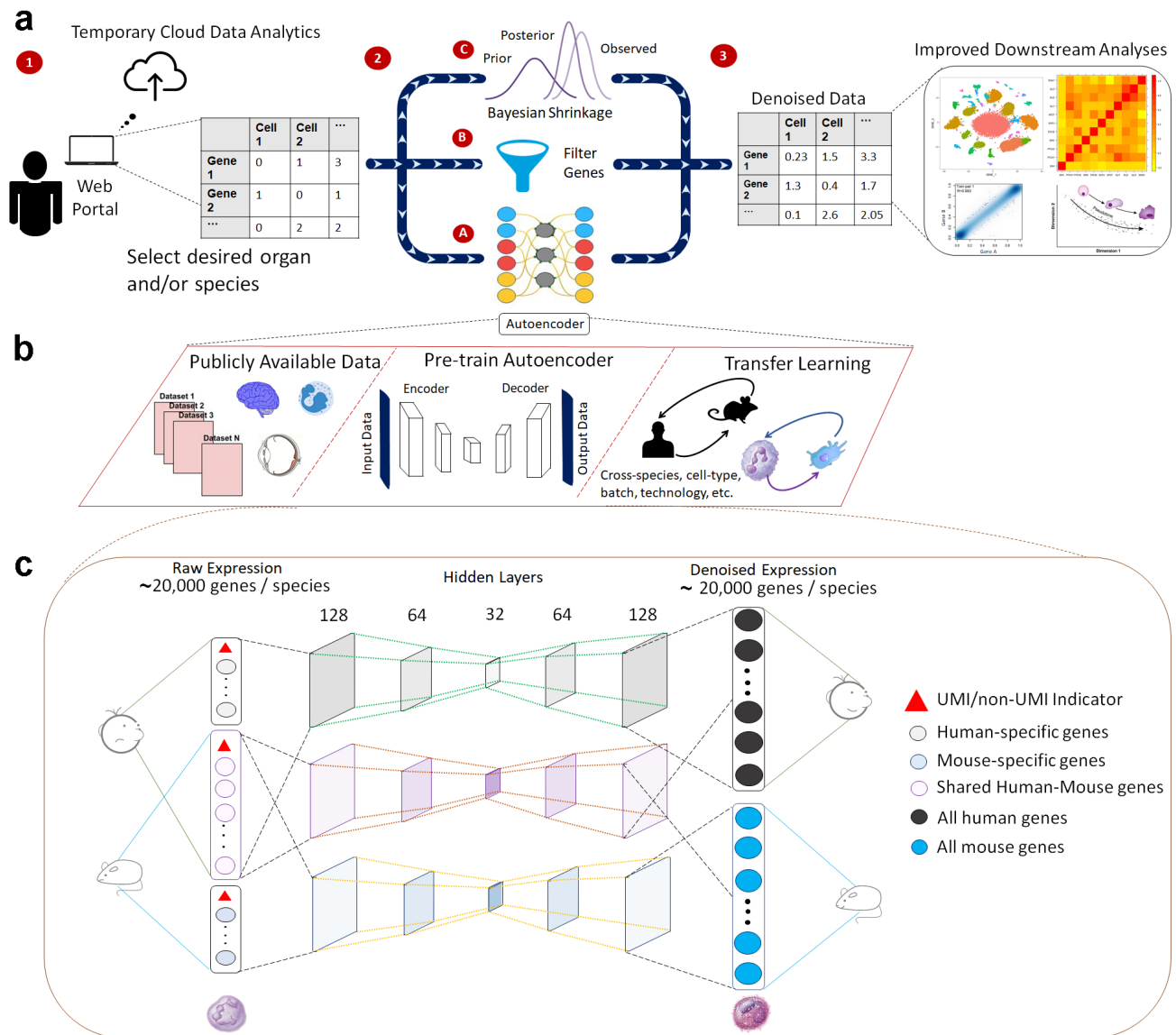**Figure 1: Outline of the SAVER-X transfer learning framework**.
a) Online learning: User uploads data (UMI count matrix) to the SAVER-X web portal (1) and chooses a desired pre-trained model (2A). The model is trained with the user data. Filtering of unpredictive genes (2B) and empirical Bayes shrinkage (2C) are then performed to prevent overfitting. User finally receives a denoised data matrix of the same size as the input data matrix (3), which can be used for downstream analysis. **b**) Offline autoencoder pre-training: for each species and each organ/cell-type, public datasets are collected and combined to train an autoencoder to generate pre-trained weights. c) Architecture of the autoencoder: The autoencoder allows cells from both human and mouse, both with and without UMI, to be used for pre-training. However, For each species, there are about 20,000 input nodes accepting raw gene expression values; approximately two-thirds of the input nodes are shared between species for genes with homologs.
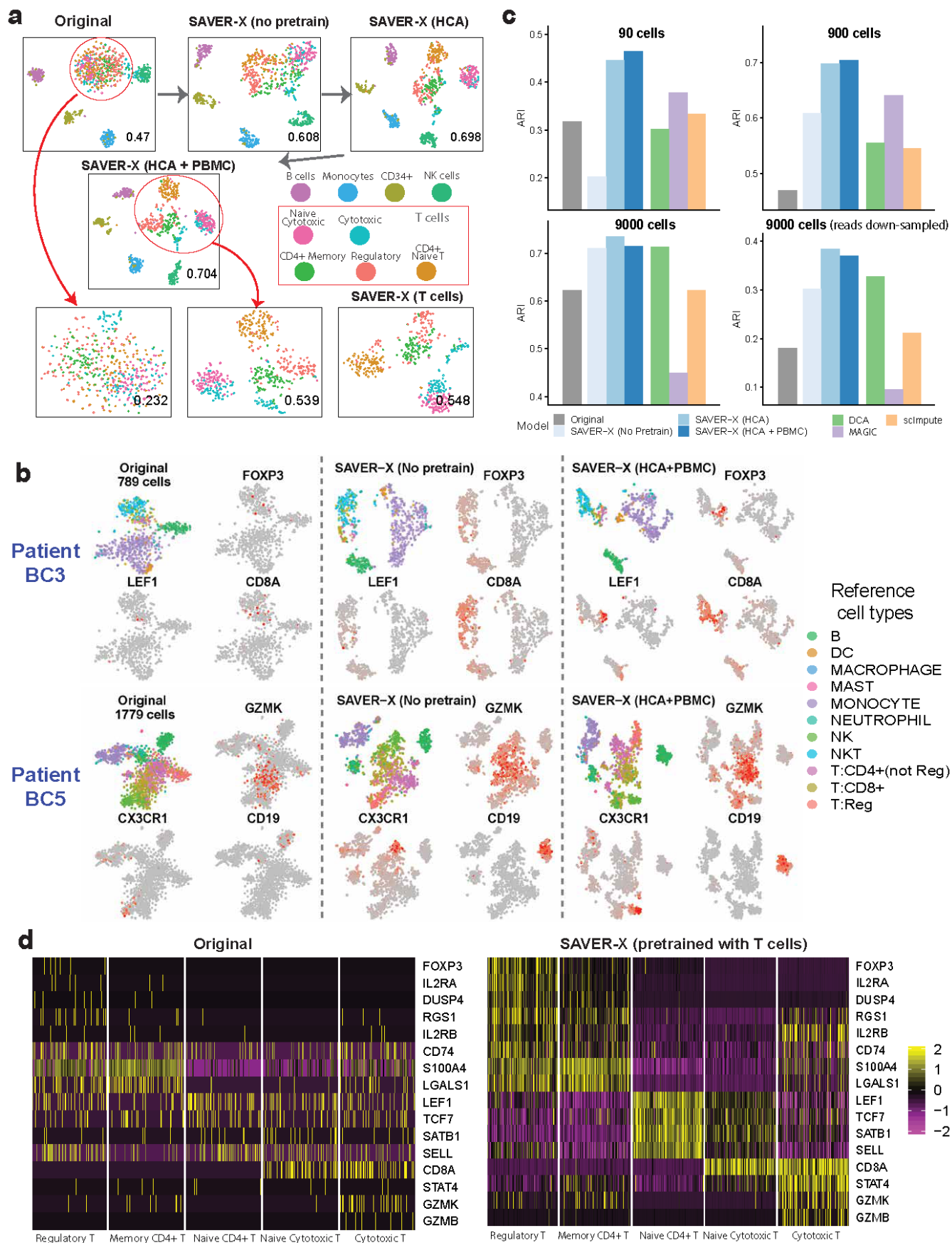
Figure 2: SAVER-X de-noising of immune cell subsets

a) t-SNE plots of 900 immune cells, colored by known cell-type labels, from the original data, and the denoised data using de-noising models with and without transfer learning are shown. The number at the right-bottom corner of each plot is the adjusted rand index (ARI). **b**) Infiltrating immune cells in resected breast carcinoma from two breast cancer patients from Azizi et al. The three panels show visualizations using original data, denoised values by SAVER-X without pre-training, and denoised values by SAVER-X pre-trained on HCA and 10X PBMC data. The t-SNE plots show separation between cell types (cell labels are obtained from the original paper). Feature plots show the expression of known marker genes, and a darker red color represents a relatively higher expression level. **c**) The performance of SAVER-X is benchmarked against existing imputation methods using clustering ARI. Each experiment varies the number of cells and the sequencing coverage. **d**) Relative expression of known marker genes in 5000 PBMC T cells (from the 9000 PBMC cells experiment) before and after denoising using SAVER-X.
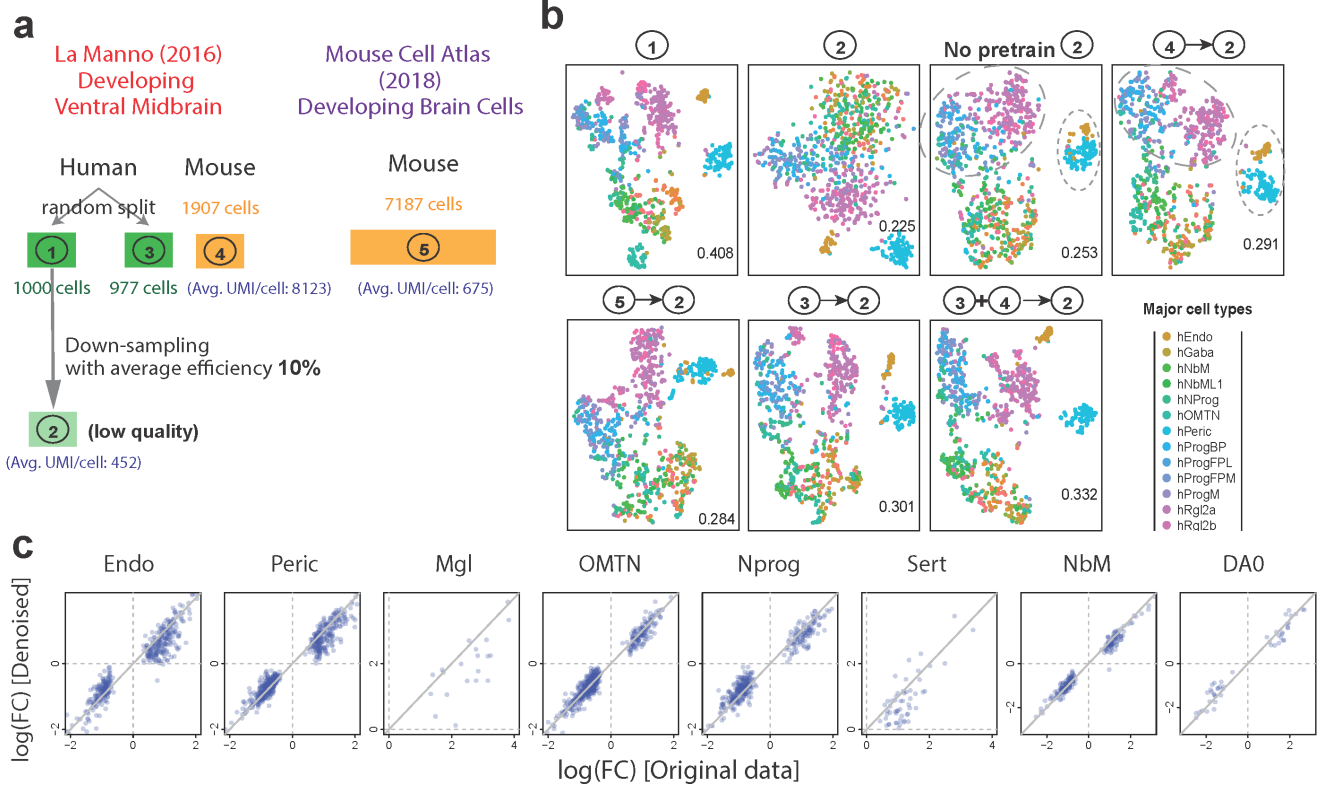
Figure 3: Mouse to human transfer learning improves denoising of single cell expression within the developing ventral midbrain.
a) Summary of datasets and the split-sample experiment design. **b**) t-SNE plots for the original data, the down-sampled data, and the SAVER-X denoised data are shown. No-pretrain denotes that the model uses randomly-initialized weights. Other scenarios where weights are initialized on mouse ventral midbrain cells (4->2), developing brain cells from mouse cell atlas (5->2), human ventral midbrain cells (3->2), and human and brain ventral midbrain cells (3+4->2) are also shown. Reference cell labels are obtained from the original paper. The ARI displayed at the bottom-right corner of each plot is computed against the original labels. **c**) Log fold-change between human and mouse data in 8 major cell types. The X-axis shows the log fold-change computed using the original human data, and the Y-axis denotes the log fold-change computed using the denoised and down-sampled human data, wherein the denoising is done by SAVER-X pretrained with mouse ventral midbrain cells (4->2 model in b). Each dot corresponds to a differentially expressed gene between human and mouse in that tissue.