bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Highly parallelized scRNA-seq pipelines are now becoming the standard. In many current and proposed studies, thousands to millions of cells are sequenced, with each cell receiving low coverage. At low coverages of 500-1000 unique molecular identifiers (UMI) per cell, precise distinctions between cell states are blurred and genes with low expression cannot be accurately quantified. To address this challenge, methods have been developed to de-noise and impute scRNA-seq data [4–7]. These methods, however, may not perform well when sequencing is done at extremely low depth, or when applied to cell types that are rare. Notably, existing denoising techniques act solely upon the data from a given study and ignore existing datasets in public domain, which may contain similar cell types.

In light of the Human Cell Atlas initiative [8], the scientific community will soon have detailed atlases for each anatomic organ in the human body; for the laboratory mouse, such an atlas (*Tabula Muris*) was recently unveiled [9]. Accumulation of publicly available scRNA-seq data presents an opportunity to leverage existing data in the denoising of a new scRNA-seq data set. Yet, it is unclear how much information can be borrowed across datasets which might be generated using different platforms, wherein samples are processed differently or at different coverages. Moreover, such transfer learning must guarantee that the denoising process will not introduce bias or force the new data to lose its distinctive features and conform to the patterns in existing data.

Here we describe a denoising framework, called Single-cell Analysis via Expression Recovery harnessing eXternal data (SAVER-X). It uses the deep autoencoder, a neural network that achieves noise reduction by means of an information bottleneck [10]. Consider a target dataset to be denoised. The autoencoder can be trained on this data starting either from random initialization of the weights, as in other denoising tools like DCA [7], or from weights obtained by training on existing public data sets (pre-training data; Figure 1b) with related cell types. The latter – initialization by pre-trained weights followed by refinement on the test data – transfers information from public data to a user's current dataset.

A pivotal challenge in transfer learning lies in finding the balance between transferring extensively and/or scarcely from the pre-training data. While the former would result in the user data losing its

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

own distinctive features, the latter risks producing inappreciable improvements. SAVER-X adaptively achieves an appropriate amount of information transfer by refining and updating weights to fit the test data (Figure 1a, item 2A). This drives the model away from the pre-training data. Subsequently, cross-validation is used to identify genes that are poorly fit by the autoencoder, and the autoencoder output for these genes is replaced by their mean expression values (Figure 1a, item 2B). Finally, for each gene in every cell, SAVER-X computes a weighted average of the fitted value and the observed normalized count [4] (Figure 1a, item 2C). This weighted average is the posterior mean of the gene's expression in the given cell based on a Bayesian hierarchical model, assuming the Poisson-alpha technical noise model [11]. SAVER-X outputs the mean denoised values, which can then feed into downstream analyses.

Many core cell types and essential pathways are shared between human and mouse, and, importantly, experiments can be performed more readily on mice than on human subjects [8,12,13]. Thus, effective mouse to human transfer learning invites new ways to use mouse as a model organism. To enable cross-species data sharing, the autoencoder in SAVER-X consists of three sub-networks with human-specific, mouse-specific and human-mouse shared nodes (Figure 1c). The human-mouse shared network receives human-mouse homologous genes as input. To adjust for the differences between data generated using non-UMI- and UMI-based technologies, an indicator node at the input layer feeds into each sub-network.

SAVER-X is publicly available at http://singlecell.wharton.upenn.edu/saver-x/, where users can choose from models pre-trained on 31 mouse tissues and human immune cells. Models jointly pretrained on cells from both species are also available for brain and pancreatic tissues. For additional details on SAVER-X architecture and estimation, see Online Methods.

Cells that constitute the immune system are implicated in virtually every disease. While understanding the features of infiltrating immune cells in an inflamed tissue is of critical importance, their representation is often small in scRNA-seq studies in the absence of flow sorting. Thus, denoising the observed values without relying on external data becomes especially challenging [14,15]. We demonstrate that SAVER-X can perform transfer learning for immune cells between healthy and

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

disease conditions. By pre-training SAVER-X on scRNA-seq data from the Human Cell Atlas (HCA) project [8] (500,000 immunocytes from umbilical cord blood and bone marrow) and 10X Genomics website [16] (200,000 peripheral blood Mmononuclear cells), we were able to meaningfully improve the data quality in other scRNA-seq studies that profiled immune cells.

First, we evaluated SAVER-X with and without pre-training against existing denoising methods on a set of purified cells from 9 immune cell types [16]. We created a "test" dataset by randomly selecting 100 cells for each cell type (Online Methods). Among this set of 900 immune cells, with an average UMI count of roughly 1200 per cell, neither is it easy to visually distinguish NK cells from T-cells, nor can T-cell subtypes be separated (Figure 2a). SAVER-X imputation of this dataset, without using any existing data for pretraining, enhances the separation of NK cells from T-cells. Although the visualization of intra T-cell subtype heterogeneity also improves, the subtypes remain difficult to identify. The impact of transfer learning becomes apparent when we denoise the test data using SAVER-X pre-trained on the HCA data (Figure 2a). The Adjusted Rand Index (ARI) improves significantly as $CD8^+$ T-cells clearly separate from $CD4^+$ T-cells, and naïve $CD4^+$ T-cells become distinguishable from other subtypes. This observation suggests that significant information about cell type-specific transcriptional signatures can be transferred between datasets even when the cell types belong to different tissues and are prepared in different laboratories. Furthermore, SAVER-X pre-trained on both HCA and 10X data led to a distinct separation between $CD4^+$ memory T-cells and regulatory T-cells (T_{regs}). The 10X dataset obtained from peripheral blood mononuclear cells (PBMCs) contains ~120,000 T-cells, and a SAVER-X model trained specifically on these cells further improves the separation of T-cells subtypes (Fig. 2a). Reliable detection of T-cell subtypes is crucial to the characterization of a tissue's immune environment. For instance, naïve $CD4^+$ T-cells help maintain immune competence throughout life [17], and yet the mechanisms underlying their establishment and maturation remain elusive. SAVER-X allows us to confidently identify this sub-population and study its homeostasis, which is ultimately critical for clinical applications in both vaccination and immune reconstitution.

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The potential of transfer learning in biology hinges on its ability to adapt to diverse and practical settings. Thus, we explored if SAVER-X can effectively learn from healthy HCA cells in the denoising of immune cells sequenced from primary breast carcinoma samples from eight treatment-naïve patients [18]. SAVER-X, pre-trained on publicly available immune cell datasets to denoise the tumor tissue-resident immune cells, not only allowed us to better characterize immune cell types, but also clarified the expression patterns of marker genes (Figure 2b, Figure S3) in these patients. This improved reconstruction of the tumor immune microenvironment typifies the potential gains achievable by transfer learning from accumulating public data.

We further assessed the utility of SAVER-X in scenarios where either the number of cells sequenced could be small (less than 100), or the sequencing depth might be too low (60 UMIs per cell; Table S1). Currently, cells with such low coverage are typically discarded. We show that SAVER-X not only salvages such data, but also extracts useful information about gene-gene relationships (Figure S2). We benchmarked SAVER-X against other scRNA-seq denoising methods that do not employ transfer learning, *viz.,* DCA [7], scImpute [5] and MAGIC [6], and found that SAVER-X significantly outperforms existing methods (Figure 2c, Figure S1) in most scenarios. As expected, for separating major cell types, the benefits of transfer learning diminish when a large number of cells are sequenced at a sufficiently high depth. As a denoising method, SAVER-X enables improved gene-level analysis; for instance, as gene expression becomes less sparse, selection of important regulatory and marker genes becomes more reliable (Figure 2d).

Having demonstrated that SAVER-X effectively transfers information across labs and from healthy to disease settings, we next examined the feasibility of transfer learning across species. Mouse models have helped scientists understand the basis of several human disorders, and although transcriptomic patterns in mouse might not always provide a direct route to the cognate human condition, similarities and disparities of genetic programs, once understood, are likely to provide a deeper understanding of the fundamental architecture underlying cellular development and physiology. In this regard, the ability to harness mouse data in the denoising of human data represents a new mode of cross-species learning. We examined scRNA-seq data from cells in the

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

developing ventral midbrain of both human and mouse, and found that, indeed, SAVER-X pre-trained on mouse scRNA-seq data enhances the quality of the human data (Figure 3).

First, we reduced the high coverage human ventral midbrain scRNA-seq data by sampling only 10% of the reads [13], to a median per cell coverage of 452 UMIs. To compare the gains achievable by intra- and inter-species transfer learning, we split the human cells randomly into two groups, down-sampled one group and used the other group as the pre-training data (Figure 3a). SAVER-X pre-trained on the matched mouse brain cells led to a distinct improvement in cell type identification for human compared with the un-pretrained model, affirming the potential of transfer learning across species (Figure 3b). We found that a model jointly pre-trained on both human and mouse data further augments the human scRNA-seq data quality compared with pre-training on the human cells alone. Remarkably, pre-training SAVER-X on cells from regions other than the ventral mid-brain using the *Tabula Muris* [9] also improved the ARI (Figure 3b). We then pre-trained SAVER-X on three human non-UMI datasets [19–21], and found that the model jointly pre-trained using both the non-UMI human cells and mouse cells outperforms training on either species alone (Figure S4a). These observations suggest that SAVER-X prevents negative transfer of information between species by harnessing the heterogeneity among public datasets. Data heterogeneity forces SAVER-X to learn robust low-dimensional representation of information, which likely contains the true biological signals that are shared across studies.

To further demonstrate that SAVER-X does not unnaturally bias data denoising, we examined whether a model pre-trained on mouse data affects human-specific patterns. We denoised human scRNA-seq data using the matched mouse data, and then compared the log fold-change of the genes differentially expressed between human and mouse for each cell type before and after denoising. We found that the fold changes are indeed preserved, suggesting that SAVER-X introduces negligible bias (Figure 3c). On the other hand, simply relying on an autoencoder, without gene filtering or Bayesian shrinkage, reduces the fold change between human and mouse for some genes in some cell types (Figure S4b). This highlights the importance of balancing the autoencoder predictions against the observed data to prevent bias.

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Taken together, our results demonstrate that the transfer learning framework employed by SAVER-X can leverage existing scRNA-seq datasets to improve the quality of new scRNA-seq data across UMI-based sequencing platforms, species, organs and cell types. At its core, SAVER-X trains a deep neural network on scRNA-seq data across a range of study designs and applies this model to new data to strengthen shared biological patterns. This general framework for inferring "true" relationships from raw and error-prone experimental data will be broadly applicable in other high-throughput settings. Through applications in immunology and developmental neuroscience, we show that SAVER-X can improve cell type classification and gene expression characterization in both healthy and disease settings. With increasing accumulation of publicly available data, SAVER-X will increase in generalization accuracy and in tissue- and cell-type specificity. A technology like SAVER-X changes the approach to scRNA-seq data analysis from a process of study-specific quality control and statistical modeling to an automated process of cross-study data integration and information sharing.

Methods

Collection of public datasets

The Human Cell Atlas (HCA) dataset was downloaded from the HCA data portal (https://preview.data.humancellatlas.org/) and the PBMC data was downloaded from the 10X website (https://support.10xgenomics.com/single-cell-gene-expression/datasets, Table S2). The purified data for each immune cell type was also downloaded from the 10X website [16]. The breast cancer data [18] was downloaded from GEO (GSE114725). The developing midbrain data [13] was downloaded from GEO (GSE76381). For other mouse developing brain datasets, we include cells from neonatal and fetal brain tissues in the *Tabula Muris* [9] data (GSE108097). For the other non-UMI human developing brain datasets, we include three: GSE75140 [20] , GSE104276 [21] and SRP041736 [19]. No filtering is done on the original data and all genes and cells provided in the original datasets were used by SAVER-X.

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A complete list of the pre-training datasets used for pre-training the models on the SAVER-X website is provided in Table S2.

Details of SAVER-X

SAVER-X uses a Bayesian hierarchical model to combine evidence from the raw read counts of a new data set with predictions made by an autoencoder. The autoencoder can be trained exclusively on the new data set, or first pre-trained on existing data sets and then on the new data set. The autoencoder used by SAVER-X has three subnetworks, as shown in Figure 1a, with one subnetwork taking human genes as input, one subnetwork taking mouse genes as input, and one subnetwork taking shared human-mouse homologous genes as input. 21183 and 21122 genes are used for human and mouse (Supplementary Data, Supplementary Note), respectively, as input and output nodes of the autoencoder. By current annotations using the `getLDS()` function in the `bioMaRt` R package, 15494 genes have homologs shared between the two species (Supplementary Data, Supplementary Note). For each sub-network, the number of nodes in the encoding and decoding layers are, successively, 128, 64, 32, 64, and 128. If only human data are available, only the human and shared sub-network weights are updated. Similarly, if only mouse data are available, only the mouse and shared sub-network weights are updated.

For UMI datasets, let the raw UMI count for each cell c and gene g be x_{cg}, then the input expression levels are normalized by library size, re-scaled and log-transformed using formula: $\tilde{x}_{cg} = \log\left(x_{cg}/l_c \times 10000 + 1\right)$ where $l_c = \sum_g x_{cg}$ is the library size of cell c. For non-UMI datasets, TPM for each cell c and gene g are denoted as x_{cg}, and then the x_{cg} are transformed using the same formula as that for UMI. If a gene is missing in the dataset, the input is set to 0 while the corresponding output node is not accounted for in the loss function. Specifically, let the output value for gene g be defined as $\hat{x}_{cg}$, which we refer to as the autoencoder prediction. Conditional on the prediction $\hat{x}_{cg}$, the observed UMI count is assumed to follow a Negative Binomial distribution. Thus, for each cell c, the loss function for UMI-based counts is defined as the sum of log likelihoods:

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

On the other hand, TPM data is assumed to approximately follow a zero inflated Negative Binomial distribution (although TPM is not integer-valued, the likelihood function can still be computed) and the loss is defined as:

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

where $\mathrm{NB}(x; \mu, \theta)$ and $\mathrm{ZINB}(x; \mu, \theta, \pi)$ are the density of Negative Binomial and zero-inflated Negative Binomial distributions (see Supplementary Note). A separate gene-specific dispersion parameter θ_g^U and θ_g^{NU} is dedicated for UMI and non-UMI input, respectively. For non-UMI data, the gene- and cell-specific zero inflation parameter is defined as

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

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

Although SAVER-X accepts pre-training data both with and without UMI, **the target data must have UMI**. When SAVER-X is applied to the denoising of a UMI-based target data matrix, the following steps are applied (Figure 1a): (1) The autoencoder is fit on the target data, optionally starting with a user-selected pre-trained model. (2) Cross-validation is applied to filter out genes that cannot be predicted well by the autoencoder. Specifically, the target data is randomly split into held-in and held-out cell sets, the autoencoder is trained on the held-in set and then used to make predictions on the held-out set. For a specific gene g, let the normalized predictions using the held-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 normalized counts be μ_g. Then a gene is unpredictable if the Poisson deviance of the predictions and original UMI counts of the held-out samples is larger than that of the held-in sample mean and held-out original UMI counts, equivalently:

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

where $\varepsilon = 10^{-10}$ to avoid taking the log of zeros. After unpredictable genes are identified, the autoencoder is trained again on all the cells, but the predicted values of the unpredictable genes or genes that are not present in the nodes are replaced with the sample mean of library size normalized UMI counts. (3) After we obtain these predicted values, we apply empirical Bayes shrinkage, following the model used by SAVER [4]. In SAVER, let λ_{cg} be the true relative expression level of the gene that we want to recover, then we assume

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

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 autoencoder prediction. The final denoised expression level of a gene in each cell is a weighted average of the autoencoder predicted value and its observed UMI count:

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

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

Data denoising using other bench-marking methods

MAGIC [6] was performed using the R version 1.3.0 on the square root transformed mean library-size normalized expression. scImpute [5] version 0.0.9 was performed on the unnormalized expression values setting Kcluster = 9. DCA [7] version 0.2.2 was performed on the unnormalized expression values and the library-size normalized expression output was used for downstream analysis.

Generating down-sampled datasets

For an observed UMI count data matrix, we down-sample the reads to obtain a data set of the same gene and cell numbers but with lower quality. For cell c and gene g, the down-sampled value y_{cg} is generated by independently drawing from a Poisson distribution with $y_{cg} \sim \text{Poisson}(\tau_c x_{cg})$ where τ_c

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

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

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

t-SNE visualization and cell clustering

We used Seurat version 2.0 to perform cell clustering and t-SNE visualization according to the workflow detailed at (https://satijalab.org/seurat/pbmc3k_tutorial.html). For all analyses, we set the number of principal components to 15. For cell clustering using Seurat, resolution is set to be 1.6, 1.2, 0.8 and 0.8 for each of the four experiments (90 cells, 900 cells, 9000 cells and 9000 cells with down-sampled reads) of the PBMC data and kept the same on all the methods compared. The resolution is set to 1 for the cell clustering of the midbrain data [13]. The adjusted Rand Index (ARI) is computed using R package `mclust`.

Differential expression analysis

Differentially expressed genes between human and mouse for each cell type of the developing midbrain are obtained also using Seurat 2.0, where the Wilcoxon rank sum test is used. P-value adjustment is performed using Bonferroni correction based on the total number of genes in the dataset. A gene is selected as differentially expressed if its adjusted p-value is ≤ 0.05 and the absolute log fold change is ≥ 0.25.

Competing Interests

The authors declare no competing interests

References

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/457879; this version posted October 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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).

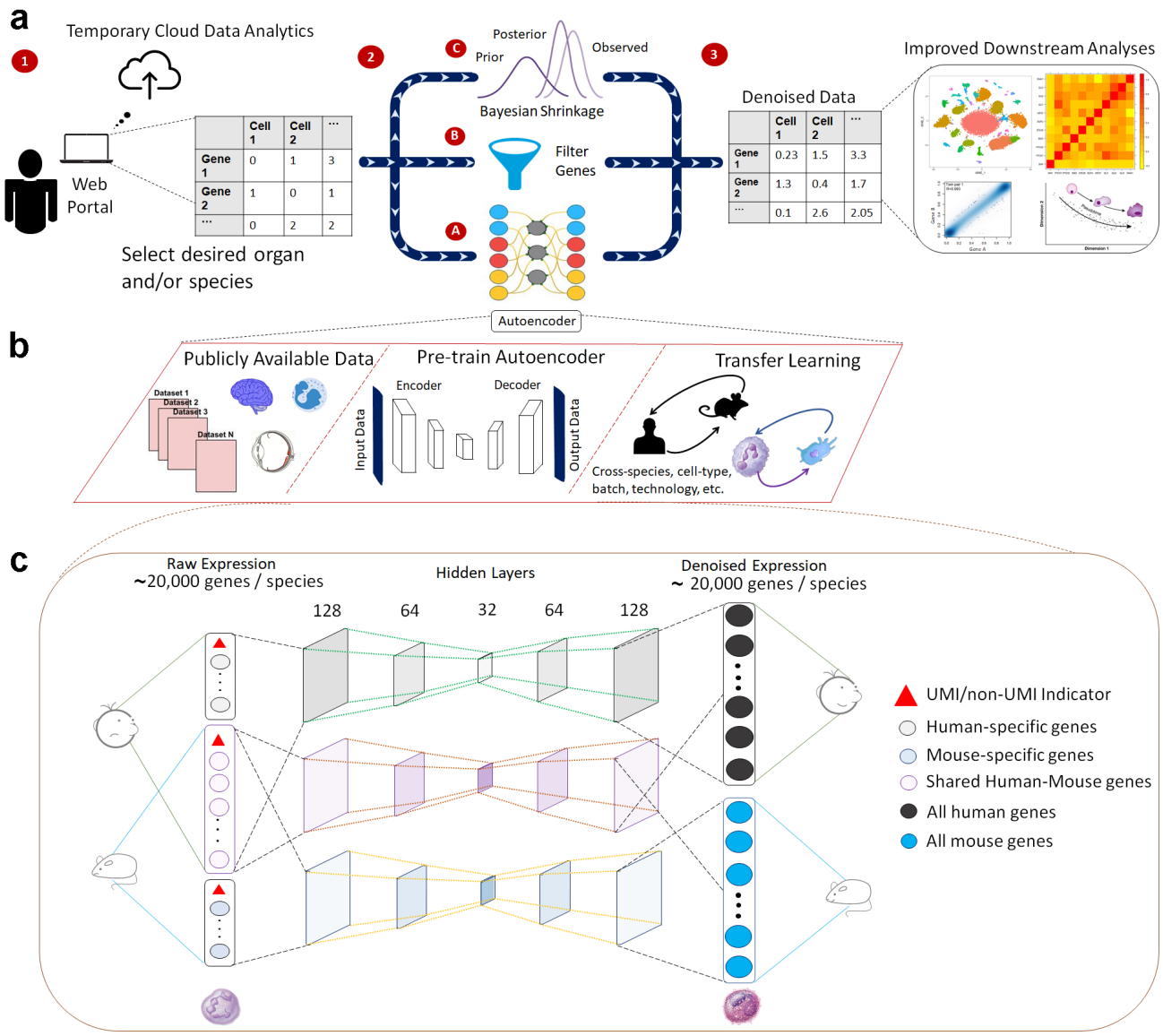

Figure 1: Outline of the SAVER-X transfer learning framework.
a) Online learning: User uploads data (UMI count matrix) to the SAVER-X web portal (1) and chooses a desired pre-trained model (2A). The model is trained with the user data. Filtering of unpredictive genes (2B) and empirical Bayes shrinkage (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.

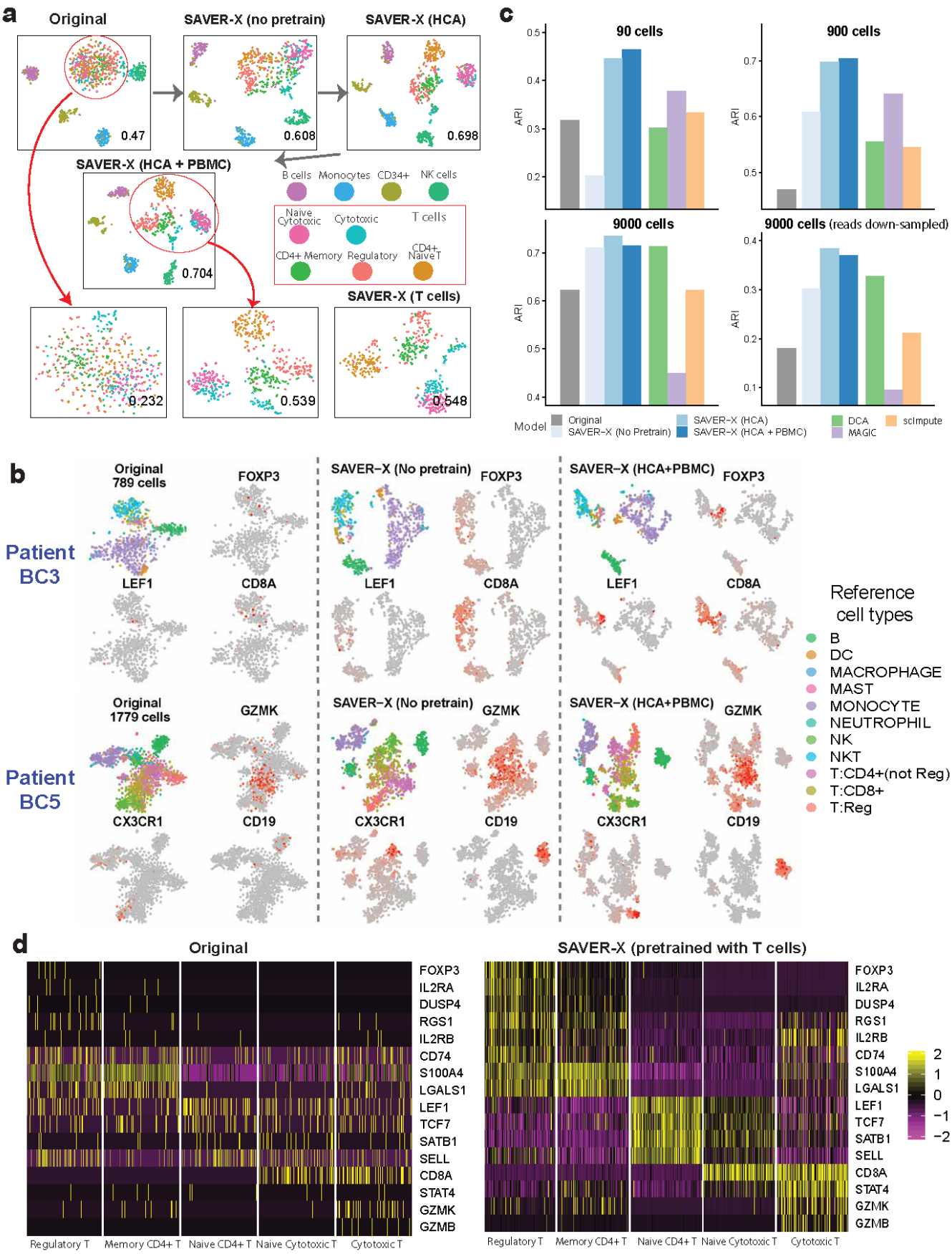

Figure 2: SAVER-X de-noising of immune cell subsets

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.

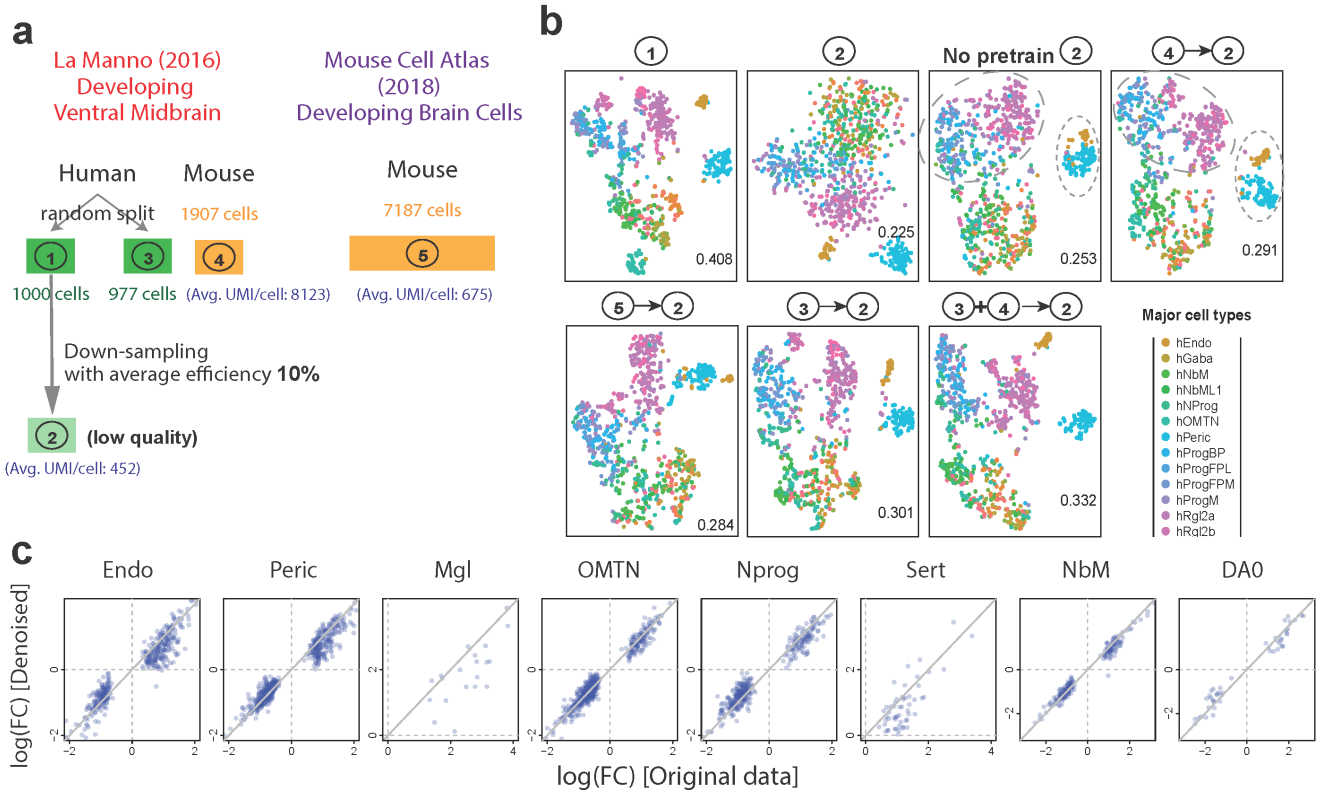

Figure 3: Mouse to human transfer learning improves denoising of single cell expression within the developing ventral midbrain.
a) Summary of datasets and the split-sample experiment design. **b**) t-SNE plots for the original data, the down-sampled data, and the SAVER-X denoised data are shown. No-pretrain 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.