Supplement to netNMF-sc: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis

Contents

S1 netNMF-sc with Euclidean distance

We also formulated netNMF-sc with a Euclidean distance cost function. This cost function is equivalent to maximizing the Gaussian likelihood the data $\mathbf{X}$ given its factors $\mathbf{W}$ and $\mathbf{H}$ [Févotte and Cemgil, 2009]. Graph-regularized NMF [Cai et al., 2008] is the following:

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} ||\mathbf{X}-\mathbf{WH}||_F^2 + \lambda Tr(\mathbf{W}^T\mathbf{LW}), \tag{1}$$

where λ is a positive real constant, $\mathbf{L}$ is the Laplacian matrix of the gene-gene interaction network, and $Tr(\cdot)$ indicates the trace of the matrix. We allow for zero inflation using a binary matrix $\mathbf{M}$ that masks zero entries in $\mathbf{X}$, such that a non-zero entry in a_{ij} in $\mathbf{WH}$ is not penalized when the corresponding entry x_{ij} of $\mathbf{X}$ is equal to 0. $\mathbf{M}$ has the same dimensions as $\mathbf{X}$ with entries

$$m_{ij} = \begin{cases} 1 & \text{if } x_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Incorporating the mask, the final formulation of netNMF-sc with a Euclidean distance cost function is

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} ||\mathbf{M}\circ\mathbf{X}-\mathbf{M}\circ\mathbf{WH}||_F^2 + \lambda Tr(\mathbf{W}^T\mathbf{LW}), \tag{3}$$

where $\circ$ indicates element-wise multiplication (or Schur product of matrices).

To meet the Gaussian assumptions of this model, we set $\mathbf{X}$ to be the log-transform of the transcript counts with a pseudocount of 1, as in many scRNA-seq models which assume an underlying Gaussian distribution [Prabhakaran et al., 2016, Li and Li, 2018]. The zero entry mask is not implemented in many commonly used NMF methods [Pedregosa et al., 2011, MATLAB, 2018], but has a profound effect on improving clustering performance and imputation accuracy at high dropout rates (Fig S3(a-b))

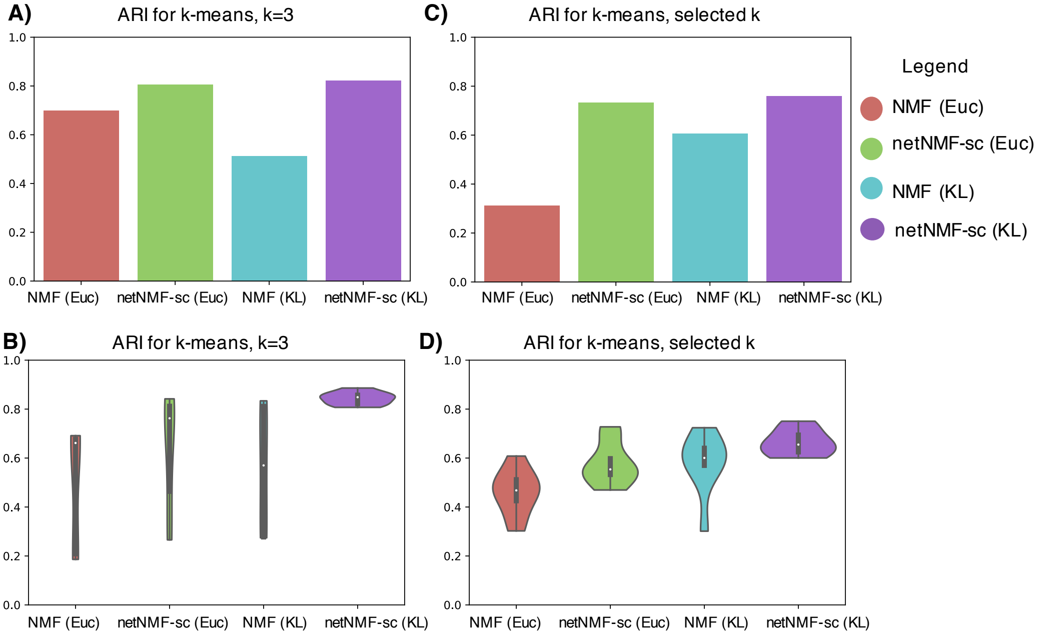


Figure S1: A) Clustering performance of NMF and netNMF-sc on scRNA-seq of 182 cells from Buettner et al. [2015] with Euclidean (Euc) and KL divergence cost functions, and k-means clustering with $k = 3$. The factor matrices $\mathbf{W}$ and $\mathbf{H}$ are randomly initialized by sampling i.i.d from the standard normal distribution, taking the absolute value of each entry to ensure non-negativity. The result that minimizes the netNMF-sc objective value across 10 random initializations is displayed. B) Variance in clustering performance across 10 initializations of NMF or netNMF-sc. C) Clustering performance of NMF and netNMF-sc with Euclidean and KL divergence distance functions clustered with k-means. For each initialization, the k which produces the highest silhouette score within the range $2 \leq k \leq 20$ is selected. D) Variance in clustering performance across 10 initializations of NMF or netNMF-sc with k selected using silhouette score.

S2 Parameter selection via holdout validation

We use the following holdout validation procedure to select the number of latent dimensions d and the regularization parameter λ.

1. Select 20% of the entries of $\mathbf{X}$ to be held-out at random. Let $\mathcal{V}$ denote the indices of these data in $\mathbf{X}$.

2. Run netNMF-sc for a range of latent dimensions d with $\lambda = 0$, masking out held-out entries using the matrix $\mathbf{M}$

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{(\mathbf{M} \circ \mathbf{WH}|_{ij})} - x_{ij} + (\mathbf{M} \circ \mathbf{WH}|_{ij})|_{ij} \right) + \lambda Tr(\mathbf{W}^T \mathbf{LW}), \tag{4}$$

   where $\mathbf{M}$ contains zeros for $m_{ij} \in \mathcal{V}$ and ones for $m_{ij} \notin \mathcal{V}$. We hold out random entries rather than rows or columns to prevent overfitting as proposed by Owen et al. [2009].

3. Calculate root mean squared error (RMSE) $= \sqrt{\frac{\sum_{(i,j)\in\mathcal{V}} (\mathbf{WH}_{ij} - \mathbf{X}_{ij})^2}{|\mathcal{V}|}}$ between the held-out data from $\mathbf{X}$ and the reconstructed data $\mathbf{WH}$, where $|\mathcal{V}|$ denotes the number of held-out entries

4. Select the value of d which results in the lowest RMSE

We perform the analogous procedure to select the regularization parameter λ using the value of d selected in the previous step.

S3 Library size normalization

For a transcript count matrix $\mathbf{X}$, the library size l_j of each cell j is the sum of all transcript counts across every gene,

$$l_j = \sum_{i \in n} x_{ij}.$$

To normalize $\mathbf{X}$, we divide each entry x_{ij} in a cell's expression profile by the cell's library size and then multiply x_{ij} by the median library size q across all cells,

$$\bar{x}_{ij} = q \frac{x_{ij}}{l_j},$$

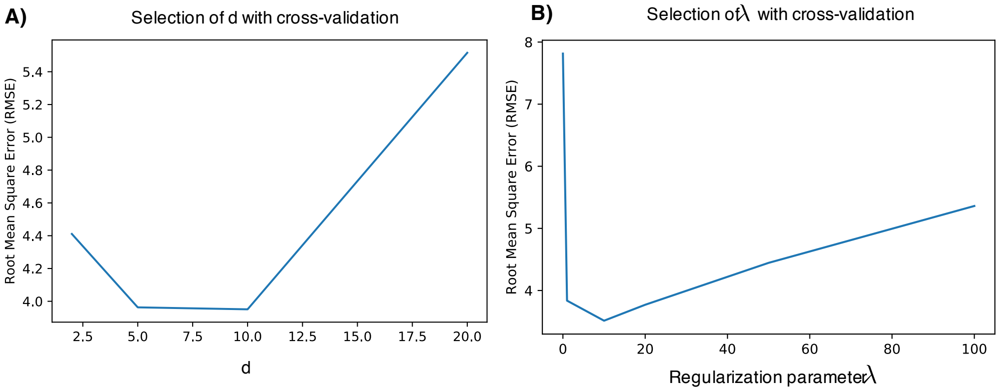


Figure S2: A) RMSE between held-out entries of $\mathbf{X}$ and corresponding imputed entries of $\mathbf{WH}$ on simulated data. Here $d = 10$ has the lowest root mean squared error. B) RMSE between held-out entries of $\mathbf{X}$ and corresponding imputed entries of $\mathbf{WH}$ with $d = 10$. Here $\lambda = 10$ has the lowest RMSE.

where $\bar{x}_{ij}$ is an entry in the normalized transcript count matrix $\bar{\mathbf{X}}$.

S4 Clustering low-dimensional cell matrices

To compare the results of the dimensionality reduction and imputation methods PCA, scNBMF, NMF, netNMF-sc, MAGIC, and scImpute, we cluster cells by running k-means on the output from each method. For dimensionality reduction methods (scNBMF, NMF, netNMF-sc) we cluster by running k-means on the low-dimensional cell matrix, $\mathbf{H}$, where the number of dimensions d is selected using holdout validation (Section S2). For PCA we cluster by running k-means on the top principal components which explain 90% of the variance in the data. For imputation methods (MAGIC and scImpute) we run PCA on the imputed matrices to reduce the dimensionality of the data and cluster by running k-means on the top principal components which explain 90% of the variance in the data. For each method, k-means is run with 100 random initializations and the clusters corresponding to the optimal objective value are reported.

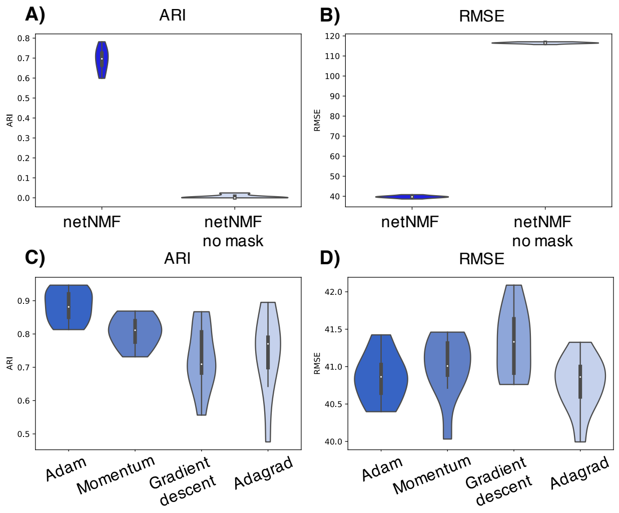


Figure S3: (A-B) Adjusted rand index (ARI) and Root mean square error (RMSE) of netNMF-sc with Euclidean distance on simulated data with and without masking of zero entries. (C-D) Clustering performance (ARI) and imputation error (RMSE) of netNMF-sc with Euclidean distance using different optimizers (Adam, Momentum, Gradient descent, and Adagrad).

S5 Data simulation

We use a real gene-gene co-expression network obtained from COEXPEDIA [Yang et al., 2016] and randomly select 5000 genes to be retained using the *random.sample* command. To define differentially expressed genes, for each of the k clusters, we randomly sample 5 genes and their neighbors to be differentially expressed. If this results in more than 10% of genes being differentially expressed in each cluster, we downsample, at random, these selected genes such that at most 10% of the genes in each cluster are differentially expressed. Each differentially expressed gene is scaled by a *differential expression factor* as described by Splatter [Zappia et al., 2017], however we ensure that if a gene is overexpressed in a cluster (differential expression factor > 1), then its selected neighbors are also overexpressed. The same is true for for underexpressed genes (differential expression factor < 1). Dropout of transcripts is performed following either the double exponential or the multinomial dropout model.

Table S1: Methods for analyzing scRNA-seq data

Method	Imputation	Dimensionality reduction	Clustering	GRN inference
BISCUIT [Azizi et al., 2017]	✓		✓	
MAGIC [Van Dijk et al., 2018]	✓			
scImpute [Li and Li, 2018]	✓			
drImpute [Gong et al., 2018]	✓			
SAVER [Huang et al., 2018]	✓			
CIDR [Lin et al., 2017]		✓	✓	
SC3 [Kiselev et al., 2017]			✓	
Seurat [Butler et al., 2018]			✓	
BackSPIN [Zeisel et al., 2015]			✓	
PhenoGraph [Levine et al., 2015]			✓	
ZIFA [Pierson and Yau, 2015]		✓		
ZINB-WaVE [Risso et al., 2018]	✓	✓		
SIMLR [Wang et al., 2017]		✓		
pCMF [Durif et al., 2018]		✓		
scNBMF [Sun et al., 2019]		✓		
netNMF-sc	✓	✓		
SCODE [Matsumoto et al., 2017]				✓
Sinova [Li et al., 2016]				✓
SINCERA [Guo et al., 2015]			✓	✓
SCENIC [Aibar et al., 2017]			✓	✓

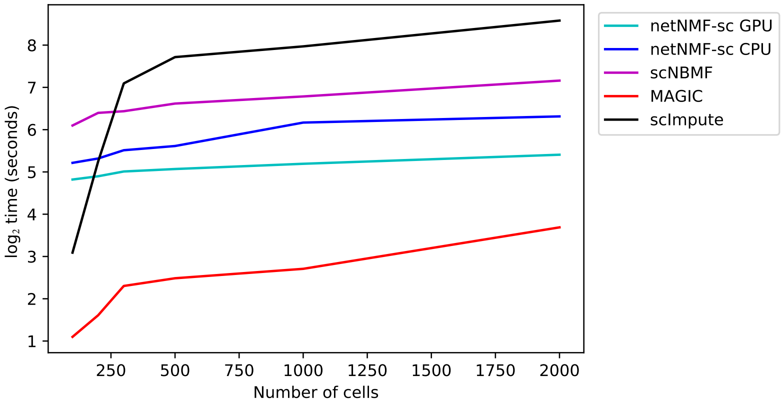


Figure S4: Runtime ($\log^2$)of imputation methods as a function of the number of cells (with 5000 genes).

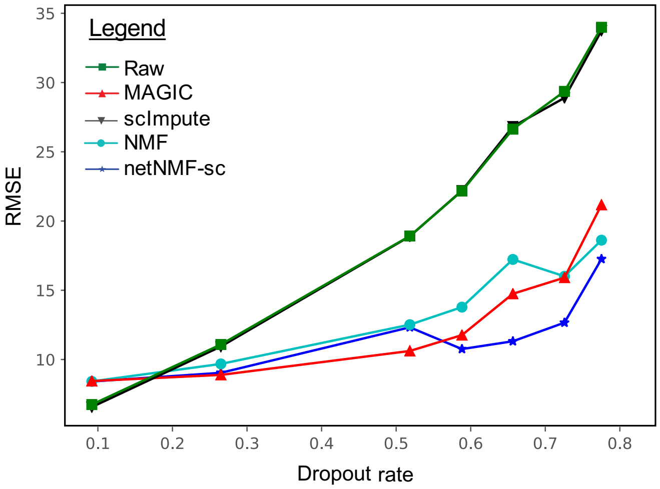


Figure S5: Root mean square error (RMSE) on simulated data using the *multinomial dropout model*.

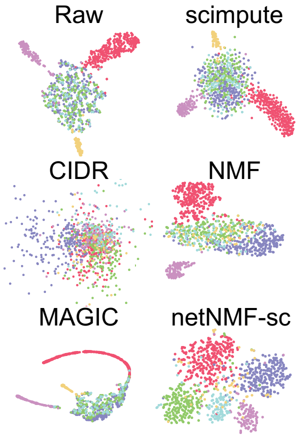


Figure S6: t-SNE projections of imputed simulated data with 5 simulated cell clusters.

S6 Clustering on cell cycle data

To quantify the effect our choice of network has on the performance of netNMF-sc, we ran netNMF-sc with two different external networks as well as a network containing randomized edges. The first network is the previously described network obtained from the ESCAPE database [Xu et al., 2014]. The second network is a generic gene-gene co-expression network which is the result of combining expression data from $2,486$ mouse microarray experiments [Yang et al., 2016]. Next, we constructed a k-nearest neighbors network, constructed by representing the 10 nearest neighbors of each gene in the input data matrix as edges with weight 1 in the network. Finally, we constructed a randomized network that maintains the same node degree as the ESCAPE network by performing the double_edge_swap procedure from the python library networkx.

We found that all networks besides the random network significantly improved clustering results compared to NMF (Fig S9A-B), with the mESC-specific network obtained from the ESCAPE database performing the best.

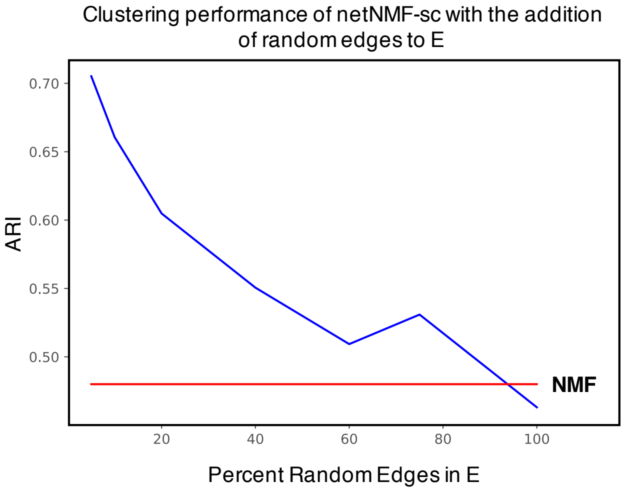


Figure S7: Clustering performance of netNMF-sc run on simulated data with 5000 genes, 1000 cells, and 6 clusters. Dropout was simulated using the multinomial dropout model with a dropout rate of 0.7. The x-axis measures the number of random edges added to the original graph $G = (V, E)$, where the number of random edges is $x|E|$. The red line shows the performance of NMF on the same data.

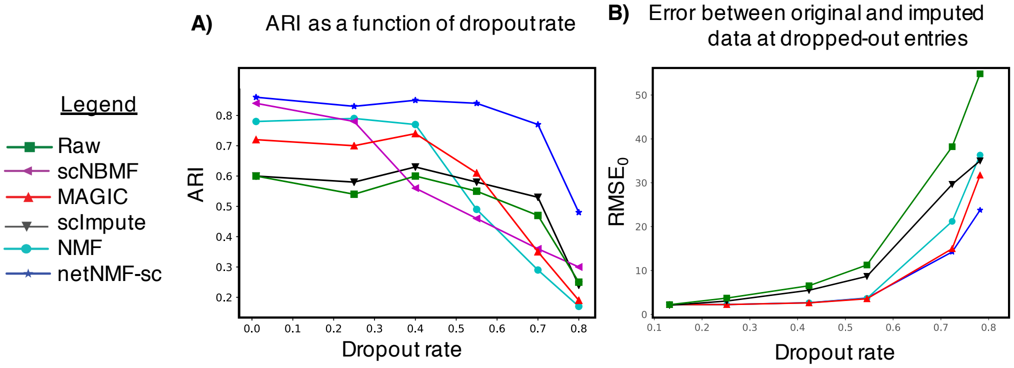


Figure S8: Comparison of netNMF-sc and other methods on clustering and imputation for a simulated scRNA-seq dataset containing 1000 cells and 5000 genes, with dropout simulated using a double exponential model. (A) Clustering results for several scRNA-seq methods on simulated data with different dropout rates. (B) Imputation results with different dropout rates.

S7 Recovering marker genes and gene-gene correlations from EMT data

Using a set of 16 canonical EMT marker genes (3 genes overexpressed in epithelial cells and 13 genes overexpressed in mesenchymal cells) [Gibbons and Creighton, 2018], we defined the set of all 120 gene pairs as our gold standard. We note that this set includes several gene pairs not investigated in the MAGIC paper [van Dijk et al., 2017]. To validate our approach, we looked for positive correlations between pairs of mesenchymal or epithelial genes and negative correlations between pairs containing one epithelial and one mesenchymal gene.

We clustered cells by *CDH1* and *VIM* expression, two canonical marker genes for epithelial (*CDH1*) and mesenchymal cells (*VIM*), respectively. We labeled the 200 cells with the highest *CDH1* expression epithelial and the 200 cells with the highest *VIM* expression mesenchymal. We compared the ranked list of differentially expressed genes from data imputed by netNMF-sc to the ranked lists of differentially expressed genes from the raw data and data imputed NMF, MAGIC, scImpute. We found that the EMT marker genes ranked very highly in netNMF-sc results ($p \leq 1.4 \times 10^{-5}$, Wilcoxon

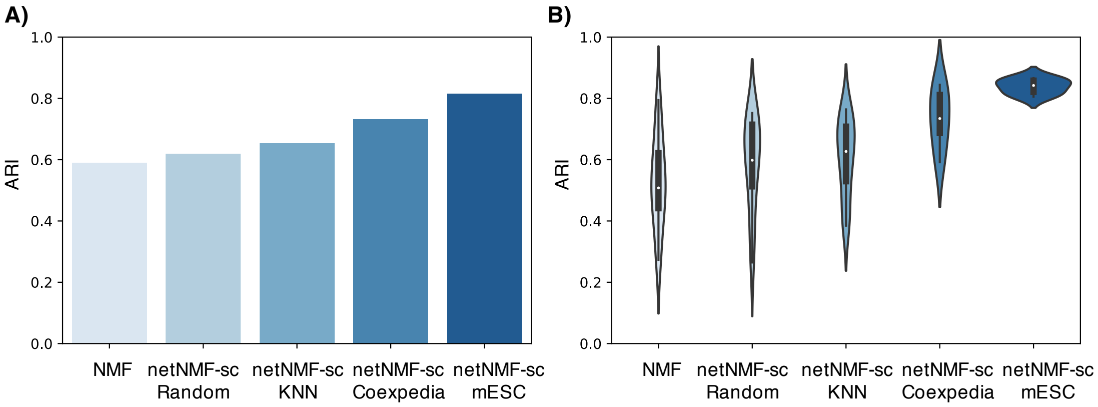


Figure S9: (A) Clustering results for cell cycle data from [Buettner et al., 2015]. The result that minimizes the netNMF-sc objective value across 10 random initializations is displayed. NMF is compared with netNMF-sc run with different networks used as input. COEXPEDIA is a generic gene-gene co-expression network, ESCAPE is a gene-gene co-expression network specific to mESCs, and KNN is a k-nearest neighbors network constructed from the 10 nearest neighbors of each gene in the input data matrix. Random is a random network constructed to have the same number of edges and degree as the ESCAPE network. (B) Variance in clustering performance across 10 random initializations.

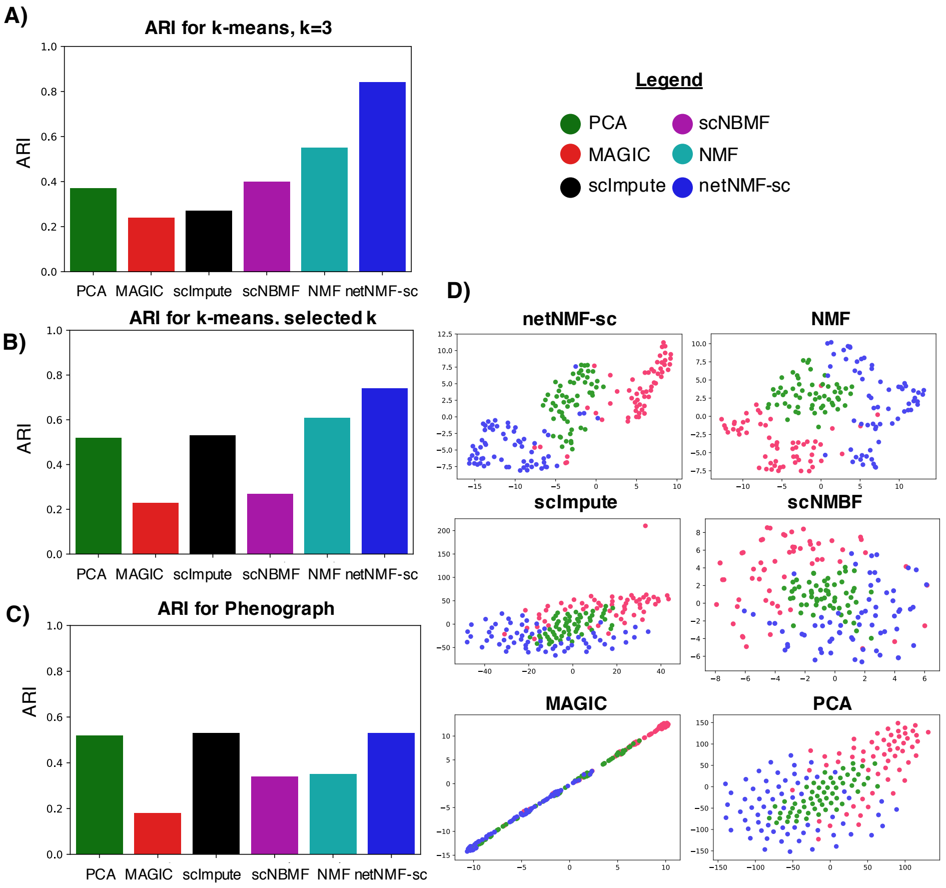


Figure S10: Clustering results on the mouse embryonic stem cell (mESC) dataset from Buettner et al. [2015], which has 3 clusters of cell determined by flow-sorting according to 3 cell cycle stages. (A) k-means clustering results for $k = 3$. (B) k-means clustering results for the value k that produced the highest silhouette score in the range $2 \leq k \leq 20$ for each method. (C) Phenograph clustering results. (D) t-SNE projections of k-means clustering results for $k = 3$.

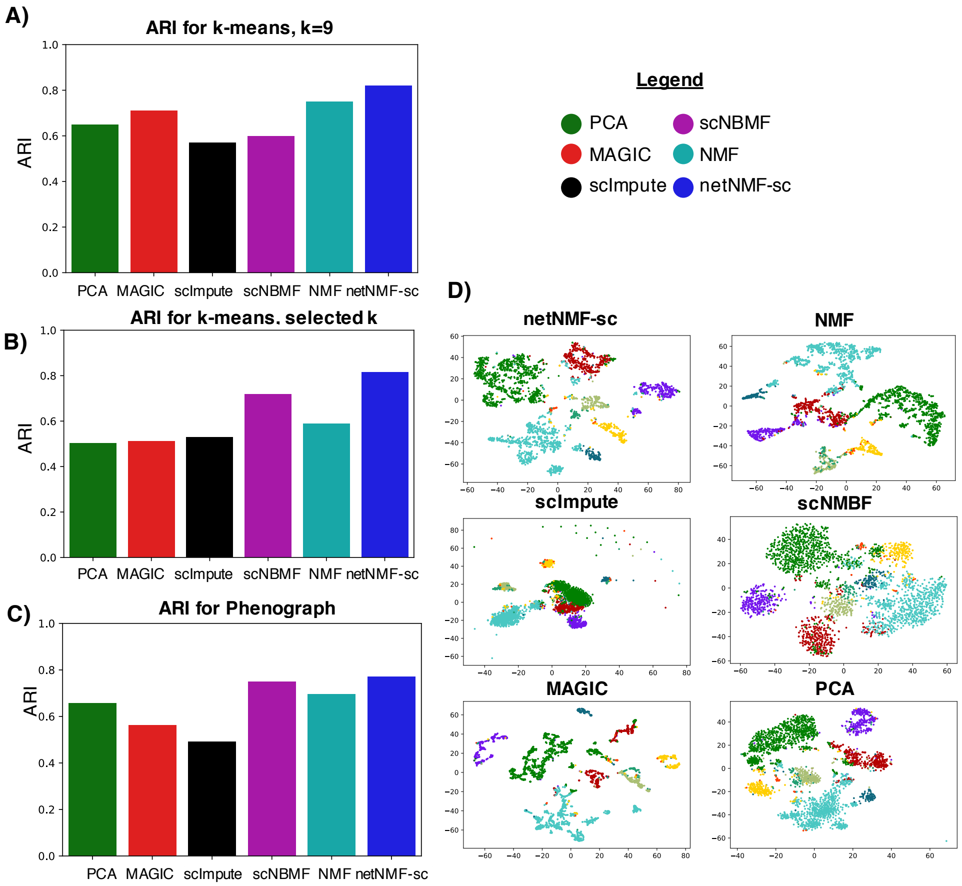


Figure S11: Clustering results on brain cell dataset from Zeisel et al. [2015] who identified 9 cell types. (A) k-means clustering results for $k = 9$. (B) k-means clustering results for the value k that produced the highest silhouette score in the range $2 \leq k \leq 20$ for each method. (C) Phenograph clustering results. (D) t-SNE projections of k-means clustering results for $k = 9$.

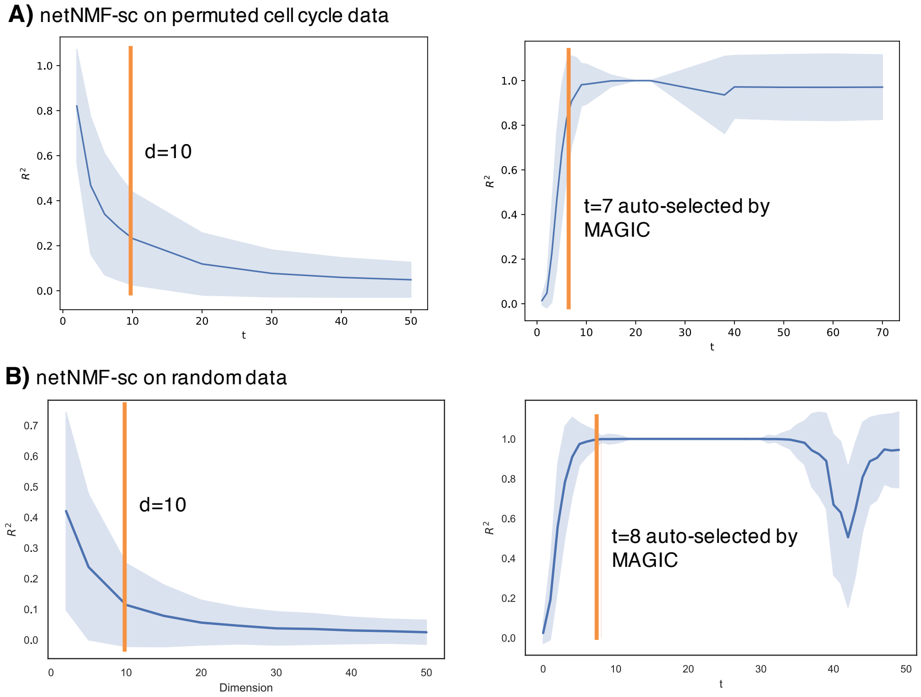
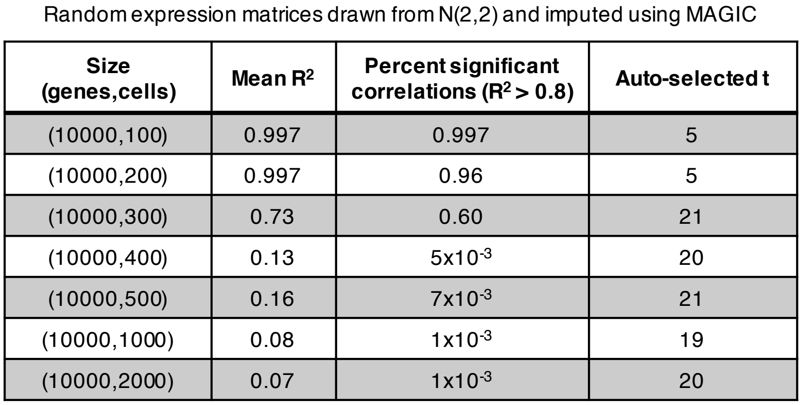


Figure S12: (A) (left)Average R^2 correlation over gene pairs on permuted cell cycle data as a function of the number d of dimensions in the matrix factorization from netNMF-sc. (right) netNMF-sc run on random data drawn from $N(2, 2)$. (B) (left) Average R^2 correlation over gene pairs on permuted cell cycle data as a function of the diffusion operator, t, used by MAGIC (light blue indicates standard deviation). $t = 5$ is auto-selected by MAGIC according to the Procrustes disparity of the diffused data. (right) MAGIC run on random data drawn from $N(2, 2)$.



Random expression matrices drawn from N(2,2) and imputed using MAGIC

Size (genes,cells)	Mean R^2	Percent significant correlations ($R^2 > 0.8$)	Auto-selected t
(10000,100)	0.997	0.997	5
(10000,200)	0.997	0.96	5
(10000,300)	0.73	0.60	21
(10000,400)	0.13	$5\text{x}10^{-3}$	20
(10000,500)	0.16	$7\text{x}10^{-3}$	21
(10000,1000)	0.08	$1\text{x}10^{-3}$	19
(10000,2000)	0.07	$1\text{x}10^{-3}$	20

Figure S13: Gene-gene correlations introduced by MAGIC on expression matrices simulated from a $N(2,2)$ distribution.

rank sum), a significant improvement compared to their ranking in the raw data ($p \leq 3.1 \times 10^{-3}$, Wilcoxon rank sum) (Fig S14(a)). In contrast, the next best performing method MAGIC had a smaller improvement in the ranking of EMT marker genes compared to the raw data ($p \leq 1.1 \times 10^{-4}$, Wilcoxon rank sum).

We observed that in data imputed by MAGIC, the E marker gene *TJP1* had higher average expression in M cells than E cells ($p = 1.5 \times 10^{-33}$) (Fig S14(b)). This resulted in *TJP1* being negatively correlated ($R = -0.57, p = 3.4 \times 10^{-50}$) with another E marker gene, *CDH1* in the MAGIC imputed data; in contrast, these E marker genes showed positive correlation ($R = 0.66, p = 6.4 \times 10^{-78}$) in the netNMF-sc imputed data, correlation that was not apparent in the raw data (Fig S14(c)). We also investigated whether netNMF-sc could recover gene-gene correlations between EMT marker genes in E and M cells. We expect that pairs of E or M genes would exhibit positive correlation, while pairs containing one E and one M gene would exhibit negative correlations. In data imputed by netNMF-sc, 12% of the EMT gene pairs were significantly correlated ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$), with all gene pairs correlated in the expected orientation (Fig S15). In data imputed by MAGIC, 23% of EMT gene pairs were significantly correlated, but 5% were correlated in the *opposite* direction than expected (Fig S14(d)).

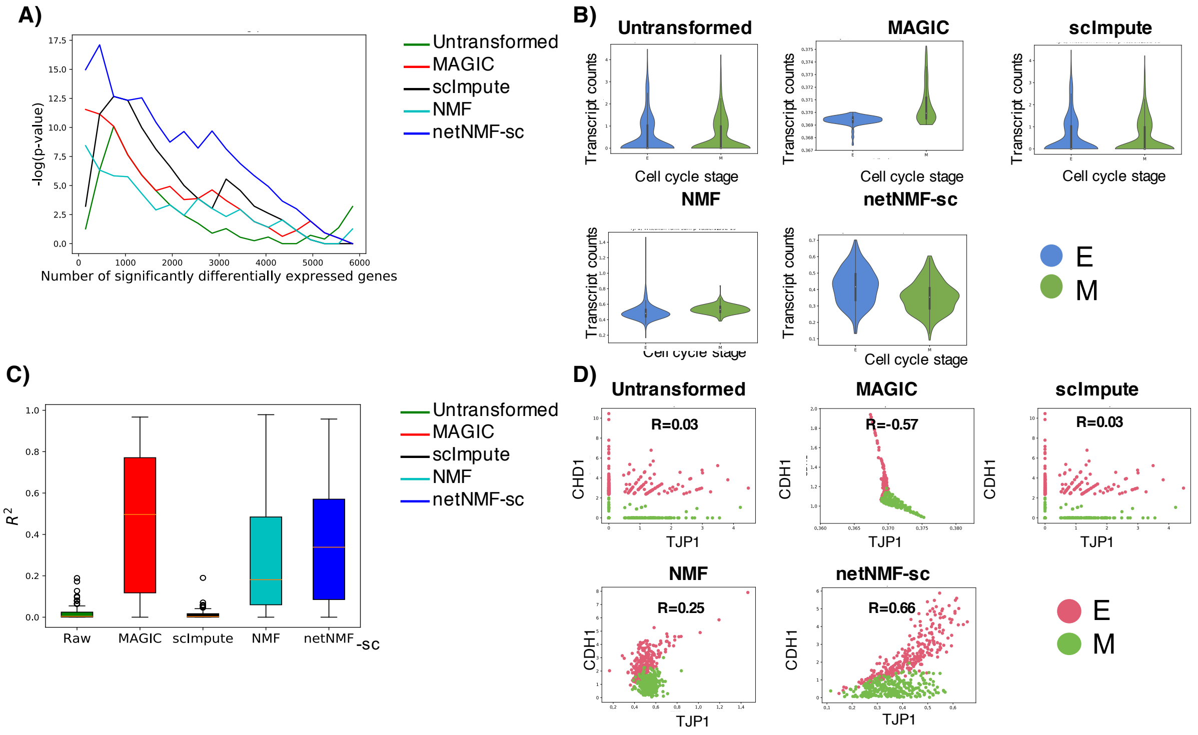


Figure S14: Comparison of gene-gene correlations and differential gene expression in raw data from [Van Dijk et al., 2018] and data imputed using netNMF-sc , NMF, scImpute, and MAGIC. (A) Overlap between differentially expressed genes and EMT marker genes (log *p*-values from Fisher's exact test). (B) Expression of the E marker gene *TJP1* in cells labeled as E (blue) and cells labeled as M (green) in data imputed by each method. In netNMF-sc inputed data, *TJP1* is overexpressed in E cells compared to M cells ($p = 1.4 \times 10^{-12}$), as expected. In contrast, in data imputed by MAGIC, *TJP1* is *underexpressed* in E cells compared to M cells ($p = 1.5 \times 10^{-33}$), and shows no significant difference in expression in raw and scImpute data. (C) Correlation between pairs of periodic genes in cell cycle data. (D) Scatter plot of two E phase genes: *CDH1* and *TJP1*. The genes are positively correlated in data imputed by netNMF-sc ($p = 6.3 \times 10^{-78}$) but negatively correlated in data imputed by MAGIC ($p = 3.4 \times 10^{-50}$).

Method	Gene pairs with significant ($R^2 \geq .8$) correlation	Periodic gene pairs with significant ($R^2 \geq .8$) correlation in correct/incorrect orientation
Raw	6e-8	0.00 / 0.00
MAGIC	0.05	0.18 / 0.05
scImpute	6e-8	0.00 / 0.00
NMF	0.02	0.06 / 0.00
netNMF-sc	6e-3	0.12 / 0.00

Figure S15: Fraction of all gene pairs and EMT gene pairs (defined by Gibbons and Creighton [2018]) with significant correlations ($R^2 \geq 0.8, p \leq 2.2 \times 10^{-16}$) in the EMT dataset. *Correct* orientation means that a pair of E-E or M-M genes have positive correlation while E-M genes have negative correlation.

References

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083, 2017.

Elham Azizi, Sandhya Prabhakaran, Ambrose Carr, and Dana Pe'er. Bayesian inference for single-cell clustering and imputing. *Genomics and Computational Biology*, 3(1):e46–e46, 2017.

Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 63–72. IEEE, 2008.

Ghislain Durif, Laurent Modolo, Jeff E Mold, Sophie Lambert-Lacroix, and Franck Picard. Probabilistic count matrix factorization for single cell expression data analysis. In *RECOMB*, pages 254–255. Springer, 2018.

Cédric Févotte and A Taylan Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. In *2009 17th European Signal Processing Conference*, pages 1913–1917. IEEE, 2009.

Don L Gibbons and Chad J Creighton. Pan-cancer survey of epithelial–mesenchymal transition markers across the cancer genome atlas. *Developmental Dynamics*, 247(3):555–564, 2018.

Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J. Garry. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC Bioinformatics*, 19(1):220, Jun 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2226-y. URL `https://doi.org/10.1186/s12859-018-2226-y`.

Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575, 2015.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, page 1, 2018.

Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483, 2017.

Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

Junxiang Li, Haofei Luo, Rui Wang, Jidong Lang, Siyu Zhu, Zhenming Zhang, Jianhuo Fang, Keke Qu, Yuting Lin, Haizhou Long, et al. Systematic reconstruction of molecular cascades regulating gp development using single-cell rna-seq. *Cell reports*, 15(7):1467–1480, 2016.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9 (1):997, 2018.

Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):59, 2017.

MATLAB. *version 1.8.0 (R2018b)*. The MathWorks Inc., Natick, Massachusetts, 2018.

Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

Art B Owen, Patrick O Perry, et al. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The annals of applied statistics*, 3(2): 564–594, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1): 241, 2015.

Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Peer. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079, 2016.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):284, 2018.

Shiquan Sun, Yabo Chen, Yang Liu, and Xuequn Shang. A fast and efficient count-based matrix factorization method for detecting cell types from single-cell rnaseq data. *BMC systems biology*, 13(2):28, 2019.

David van Dijk, Juozas Nainys, Roshan Sharma, Pooja Kathail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *BioRxiv*, page 111591, 2017.

David Van Dijk, Roshan Sharma, Juoas Nainys, Kristina Yim, Pooja Kathail, Ambrose Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. 2018.

Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods*, 14(4):414, 2017.

Huilei Xu, Yen-Sin Ang, Ana Sevilla, Ihor R Lemischka, and Avi Ma'ayan. Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS computational biology*, 10(8):e1003777, 2014.

Sunmo Yang, Chan Yeong Kim, Sohyun Hwang, Eiru Kim, Hyojin Kim, Hongseok Shim, and Insuk Lee. Coexpedia: exploring biomedical hypotheses via co-expressions associated with medical subject headings (mesh). *Nucleic acids research*, 45(D1):D389–D396, 2016.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.