Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Hot Topics in Computational Biology
Tsinghua University

OXFORD

Gene expression

SCNet: Unsupervised Multi-Channel Genome Network Inference from Single-Cell RNA Sequences

Shiyin Wang [1,*]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, China

* To whom correspondence should be addressed.

Abstract

Motivation: The rapid advancement of single-cell technologies has shed new light on the complex mechanisms of cellular heterogeneity, as the previous protein networks were ambiguously constructed. Taking advantage of the single-cell expression information can contribute to building reliable next-generation signaling protein networks and conducting causality inference.
Results: We proposed an unsupervised causality inference algorithm to extract explainable genome networks from single-cell RNA sequencing results.
Availability: The implementation of algorithm SCNet, including source codes and tutorial, is accessible at https://github.com/shiyinw/singlecell_casuality with the final version.
Contact: wangshiy16@mails.tsinghua.edu.cn
Supplementary information: Supplementary data have been uploaded along with the final version to the Github repo.

1 Introduction

Understanding heterogeneous diseases like cancer on the molecular level are crucial to quantifying the relevant genes and applying precise therapies accordingly.

With the rapid development of computing and biology, advances in sequencing technology have enabled us to profile the RNA sequence in the single-cell level when analyzing genomics, transcriptomics, proteomics, and metabolomics. The rapid advancement of single-cell technologies has shed new light on the complex mechanisms of cellular heterogeneity. Individual single-cell RNA sequencing (scRNA-seq) experiments have already been performed to discover new cell states and reconstruct cellular differentiation trajectories.

The adequacy of experimental data provides opportunities to extract knowledge for the functional relations of genomic entities. The existing protein interaction networks have been widely used in both industry and academia. However, most of the relations are defined in an ambiguous way. Text mining plays a vital role in defining the weights of edges for the sufficiency of text resources. Consequently, we propose employing **Single-Cell Network (SCNet)** to utilize single-cell expression profiles for inferring the topology of the protein interaction networks.

Numerous methods have been developed to automatically generate causal directed acyclic graphs (DAGs) upon samples. In reference to those existing techniques, we model single-cell sequences as the samples of the set of transcript genes. We define a probabilistic network model to quantify the likelihood of single-cell sequences on networks. Based on this model, we can then regard it as an optimization problem and solve it accordingly.

We conduct several experiments to test the functionality of SCNet. The result of SCNet is dependent on the provided single-cell sequences. SCNet statistically captures the temporary protein association based on the single-cell expression data. We observe the variation of networks on distinct cell types and find that the dynamic transition of networks aligns with previous experimental traits.

2 Methods

2.1 Probabilistic network model

We model the experimental data with a hierarchical Bayesian model. For each pair of variables among genes, the relationship of their expression profiles can show their relationship. We have analyzed the distribution of gene profiles in Figure1. From the data's perspective, we apply the log scale to model the gene expression data in Equation 1. Because of the existence of 0 expressed genes, we consider $log(X+1)$ in this equation.

$$\begin{aligned} log(X_i+1) &= k_{ij} log(X_j+1) + b \\ k_{ij} &\sim N(\theta_{ij}, \sigma_{ij}^2) \end{aligned} \tag{1}$$

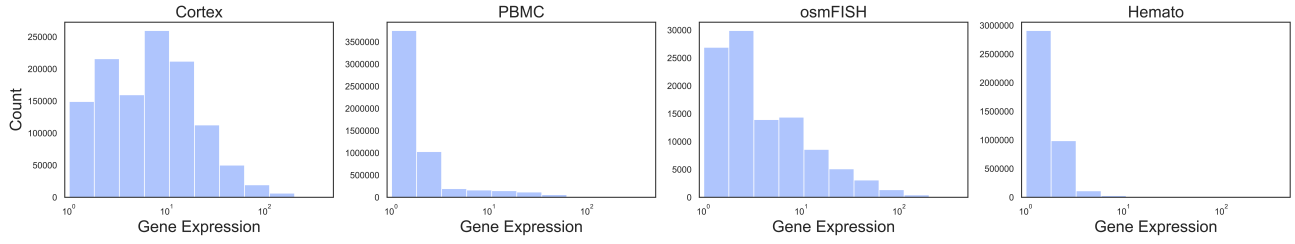

Figure 1. The count plots for the datasets we used in this project. The x-axis is on the log scale. As we can see, we should use a log scale to consider the gene expression level.

Equation 1 models the pairwise joint distribution of gene expression records. We interpret k as the relation between X and Y, and it corresponds to the edge weight in the network. Therefore, we can get a likelihood representation $Pr(experi|network)$ given the network and gene expression records by calculating the likelihood of k.

$$Pr(net|experi) \propto Pr(experi|net)Pr(net) \tag{2}$$

In Equation 3, we assume the prior distribution is Poisson depending on the number of edges in the network. The reason why choose Poisson is because of its nice property of derivative, as shown in Figure 2.

$$Pr(net) \sim Possion(\lambda) \tag{3}$$

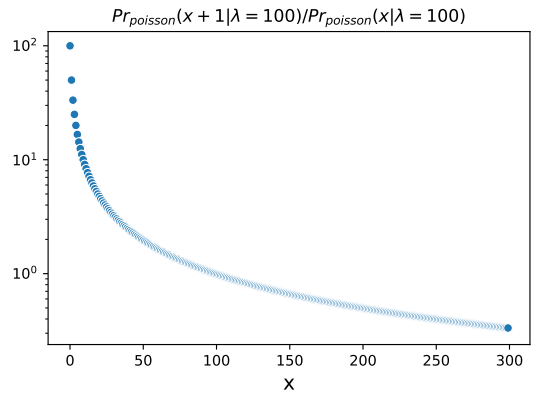

Figure 2. This fraction changes rapidly when x is far away from λ, while not significant when x is close to λ. This property can make the number of edges converges fast to the λ. Therefore, by adjusting parameter λ, we can guide the sparsity of the network.

Thus, we can use Bayesian Rule Equation 5 to get the posterior distribution of the network. And we call the optimized network as **Single-Cell Network (SCNet)**. This method can automatically capture the latent relations beneath single-cell gene expression profiles. Moreover, the introduction of prior distribution can flexibly penalize the connection rate of the network.

2.2 Relaxation on the variance estimation

If we want to strictly calculate the probability $Pr(experi|net)$, we should estimate both θ and σ^2 from the network weights. It is easy to regard the weights as θ, but it is ambiguous to get the variance σ^2. Alternatively, we assume $\sigma^2_{experi} \approx \sigma^2_{net}$ and acquire the estimation of variance from the gene expression records.

Lemma 1 (Jackknife Estimate for Variance). *Given n observations $k_0, k_1, \ldots, k_{n-1}$, we want to get the estimate of variance v. We can compute the sample median for each subsample omitting the i-th observation $\hat{\theta_i} = \frac{1}{n-1}(\sum_{j=0}^{n-1} k_j - k_i)$. Then the Jackknife estimate for variance can be acquired by the following function:*

$$v_{jack} = \frac{n-1}{n}\sum_{i=0}^{n-1}(\hat{\theta} - \hat{\theta_i})^2 \tag{4}$$

2.3 Optimization with MCMC

We use Markov Chain Monte Carlo (MCMC) algorithm to optimize parameters. The transition probability is define as Equation **??**. Under the condition that $V(f_1|f_2)$, we can simplify the formula. Finally, we can apply the Equation 5.

$$\begin{aligned} a(_1 \to _2) &= max\{1, \frac{Pr_{net|experi}(net_2|experi)V(net_1|net_2)}{Pr_{net|experi}(net_1|experi)V(net_2|net_1)}\} \\ &= max\{1, \frac{Pr_{net|experi}(net_2|experi)}{Pr_{net|experi}(net_1|experi)}\} \\ &= max\{1, \frac{Pr_{experi|net}(experi|net_2)Pr_{net}(net_2)}{Pr_{experi|net}(experi|net_1)Pr_{net}(net_1)}\} \end{aligned} \tag{5}$$

Algorithm 1 Metropolis-Hastings

Compute $log(X+1)$ from gene expression records
Compute weight means $\theta_{ij} = Pearson(log(X_i+1), log(X_j+1))$
Compute weight variances v_{jack} by Equation 4
Initialize network $g^{(0)}$ connections and weights
for t from 1 to N **do**
  Randomly choose among actions $\{delete\_an\_edge, add\_an\_edge, change\_a\_weight\}$ to get a new network g'
  Calculate $a = max\{1, \frac{Pr_{y|g}(y|g')Pr_g(g')}{Pr_{y|g}(y|g^{(t-1)})Pr_g(g^{(t-1)})}\}$
  if $random() <= a$ **then**
    Accept this new state $g^{(n)} = g'$ with probability a
  else
    $g^{(n)} = g^{(n-1)}$
  end if
end for

2.4 Datasets and processing

We use multiple annotated datasets in this work. To avoid nasty analysis and amplification, analyze the sequence in the level of genes. We apply the preprocessed data from [Lopez *et al.*, 2018] and used the Python packet scVI [Lopez *et al.*, 2018, Xu *et al.*, 2019, Gayoso *et al.*, 2019, Clivio *et al.*,

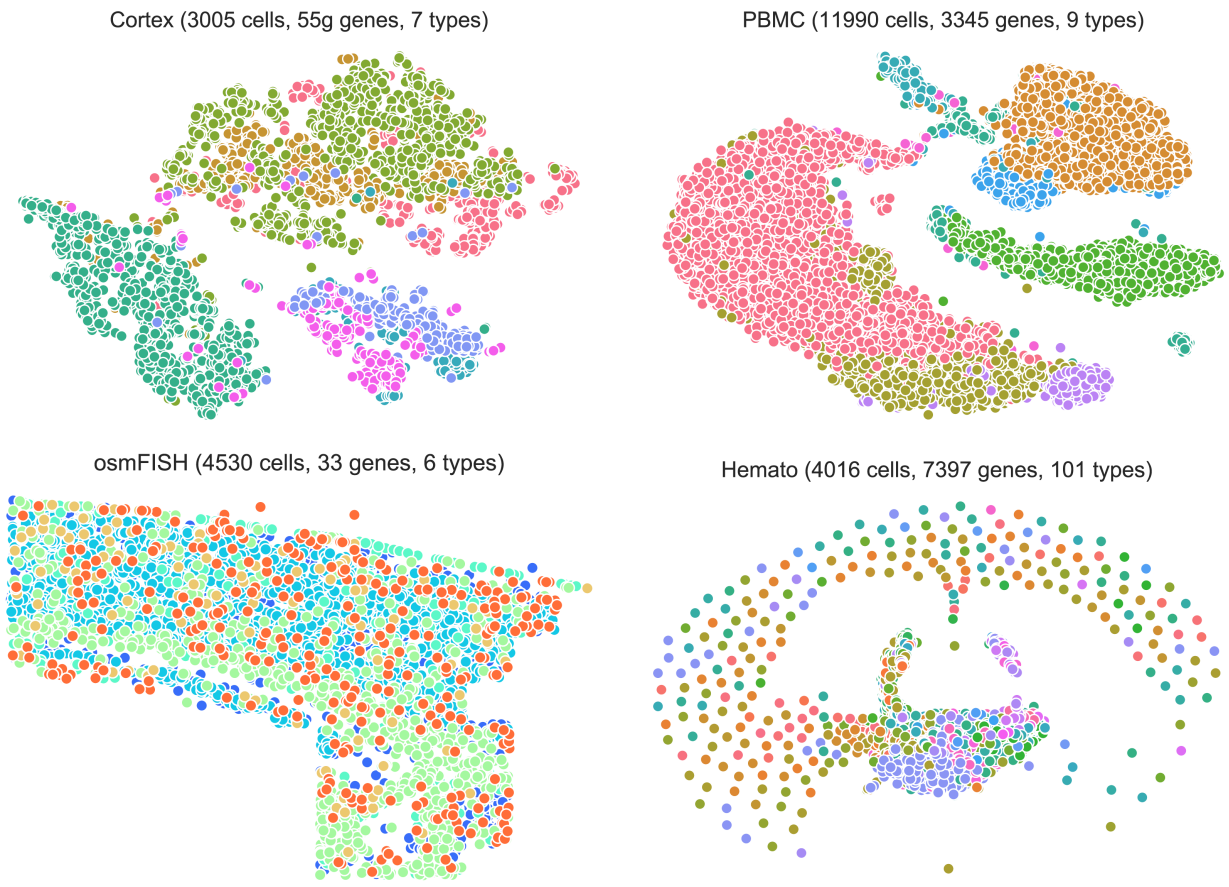

Figure 3. We use the above four datasets in this project. Except for smFISH which has provided 2-dim positions, all the other datasets are positioned by tSNE results.

2019, Boyeau *et al.*, 2019] to process the data. We visualize the datasets in Figure 3. The osmFISH dataset provides the values of x and y coordinates for each sample. But the other three datasets don't have positions, so we applied tSNE to map the points into 2-dim for visualization.

CORTEX The Mouse Cortex Cells dataset from [Zeisel *et al.*, 2015] contains 3,005 mouse cortex cells and gold-standard labels for seven distinct cell types. Each cell type corresponds to a cluster to recover. The researchers in [Lopez *et al.*, 2018] retained the top 558 genes ordered by variance.

Hemato The Hemato datasetTusi *et al.* [2018] captures continuous gene expression variations. It includes 4,016 cells of 101 gene types from two batches that were profiled using in-drop. This data provides a snapshot of hematopoietic progenitor cells differentiating into various lineages.

PBMC Peripheral blood mononuclear cells (PBMCs) dataset from 10x Genomics[x10, 2017] consists of 11990 cells on 3346 genes. The work Lopez *et al.* [2018] extracted 12,039 cells with 10,310 sampled genes and filtered genes that could not match with the bulk data.

osmFISH The research project Codeluppi *et al.* [2018] developed a cyclic smFISH protocol, thermed ouroboros smFISH, to quantify the expression of a large number of genes in tissue sections, with the goal to build a cell type map of the mouse somatosensory cortex. We applied the supporting dataset consists of 4530 cells and 33 genes.

3 Results

3.1 Convergence Analysis

The algorithm converges well, as shown in Figure 4. Because the network can have multiple variations, it is acceptable that the equilibrium does not stay on a certain network skeleton. Instead, it changes among several network skeletons.

Under appropriate λ Poisson prior, the acceptance rate after achieving equilibrium becomes smaller. As discussed in the Figure 2, the λ represents the expected number of edges in the network. When the networks are encouraged to have more edges, this method will introduce more noise to the network, and thus changes rapidly under equilibrium.

3.2 Compare with Gold Standard Provided by Human Experts

We compare the inferred networks of osmFISH dataset with the existing database STRING to acquire supporting evidence for our method. Figure 5 is the result provided by the STRING database. Our network has a similar topology with this result. There are ten correctly inferred edges in our SCNet. There are 55 edges in the SCNet and 39 edges in the string network. The true positive is 10. Many difference occurs in the text mining channel. We interpret that our SCNet method should be equipped with expert knowledge, and it also has the potential to capture more hidden information.

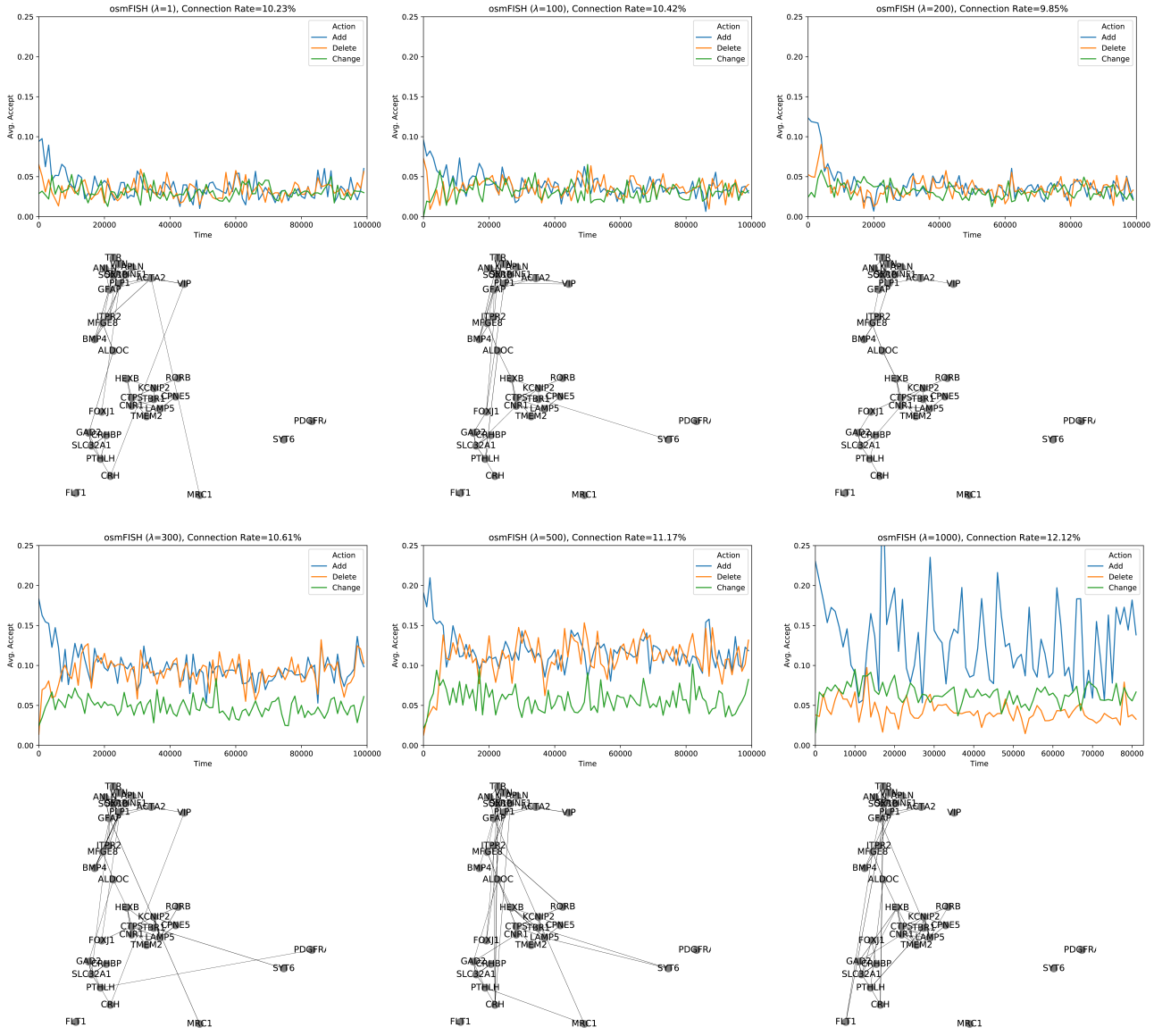

Figure 4. The average acceptance rate of every 1000 consecutive time steps. "Accept" is recorded as 1 and "Reject" is recorded as 0.

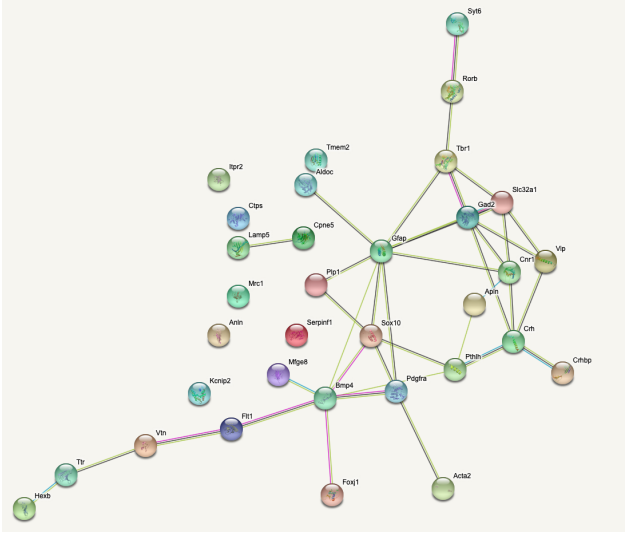

Figure 5. The search result for the 33 genes shown in the osmFISH dataset.

4 Discussion

In this paper, we proposed an unsupervised causality inference algorithm to extract explainable genome networks from single-cell RNA sequencing results.

We first statistically build a probabilistic model theoretically bridge the experimental records and network skeleton. To simplify the optimization process, we apply relaxation for the variance estimation. We also discuss the distribution choice of Poisson in our model. We finally apply the MCMC estimation algorithm to calculate the optimum networks.

Then we apply our method to several datasets, especially on osmFISH, which has considerably fewer gene sets. We compare the result of SCNet with STRING database and discuss the performance. Though the performance is not persuasive enough to predict the connections, it provides insights on which edge may have errors.

There are several future directions following our project:

- Apply the network to see gene types variations.
- Deploy methods to correct existing network databases automatically.

- Theoretical analysis to compare with other methods that derive networks from the covariance matrix directly.

Acknowledgements

This is the final course project of "Hot Topics in Computational Biology" supervised by Professor Jianyang Zeng at Tsinghua University in the 2019 autumn semester.

References

(2017). 10x genomics. https://support.10xgenomics.com/single-cell-gene-expression/datasets.

Boyeau, P., Lopez, R., Regier, J., Gayoso, A., Jordan, M. I., and Yosef, N. (2019). Deep generative models for detecting differential expression in single cells. *bioRxiv*, page 794289.

Clivio, O., Lopez, R., Regier, J., Gayoso, A., Jordan, M. I., and Yosef, N. (2019). Detecting zero-inflated genes in single-cell transcriptomics data. *bioRxiv*, page 794875.

Codeluppi, S., Borm, L. E., Zeisel, A., La Manno, G., van Lunteren, J. A., Svensson, C. I., and Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osmfish. *Nature methods*, **15**(11), 932.

Gayoso, A., Lopez, R., Steier, Z., Regier, J., Streets, A., and Yosef, N. (2019). A joint model of rna expression and surface protein abundance in single cells. *bioRxiv*, page 791947.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, **37**(6), 685.

Jiang, H., Sohn, L. L., Huang, H., and Chen, L. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics*, **34**(21), 3684–3694.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, **15**(12), 1053.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, **15**(6).

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S., Ko, S. B., Gouda, N., Hayashi, T., and Nikaido, I. (2017a). Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, **33**(15), 2314–2321.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., Hayashi, T., and Nikaido, I. (2017b). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, **33**(15), 2314–2321.

Neu, K. E., Tang, Q., Wilson, P. C., and Khan, A. A. (2017). Single-cell genomics: approaches and utility in immunology. *Trends in immunology*, **38**(2), 140–149.

Tusi, B. K., Wolock, S. L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J. R., Klein, A. M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, **555**(7694), 54.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. (2019). Harmonization and annotation of single-cell transcriptomics data with deep generative models.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., *et al.* (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, **347**(6226), 1138–1142.

Zhang, W., Li, W., Zhang, J., and Wang, N. (2019). Data integration of hybrid microarray and single cell expression data to enhance gene network inference. *Current Bioinformatics*, **14**(3), 255–268.