

---

# CORRELATION IMPUTATION IN SINGLE CELL RNA-SEQ USING AUXILIARY INFORMATION AND ENSEMBLE LEARNING

---

A PREPRINT

**Luqin Gan**  
Rice University  
luqin.gan@rice.edu

**Giuseppe Vinci**  
University of Notre Dame  
gvinci@nd.edu

**Genevera I. Allen**  
Rice University  
gallen@rice.edu

September 3, 2020

## ABSTRACT

Single cell RNA sequencing is a powerful technique that measures the gene expression of individual cells in a high throughput fashion. However, due to sequencing inefficiency, the data is unreliable due to dropout events, or technical artifacts where genes erroneously appear to have zero expression. Many data imputation methods have been proposed to alleviate this issue. Yet, effective imputation can be difficult and biased because the data is sparse and high-dimensional, resulting in major distortions in downstream analyses. In this paper, we propose a completely novel approach that imputes the gene-by-gene correlations rather than the data itself. We call this method SCENA: Single cell RNA-seq Correlation completion by ENsemble learning and Auxiliary information. The SCENA gene-by-gene correlation matrix estimate is obtained by model stacking of multiple imputed correlation matrices based on known auxiliary information about gene connections. In an extensive simulation study based on real scRNA-seq data, we demonstrate that SCENA not only accurately imputes gene correlations but also outperforms existing imputation approaches in downstream analyses such as dimension reduction, cell clustering, graphical model estimation.

**Keywords** Single Cell RNA-seq, Imputation, Correlation Completion, Ensemble Learning, Auxiliary Information, Clustering, Dimension Reduction, Graphical modeling

## 1 Introduction

In genomics, researchers are interested in discovering the relationships between genes, monitoring changes of gene expression, and understanding the influence of genes on the organism. Bulk RNA sequencing (bulk RNA-seq) is a sequencing technology that lets us analyze gene expression from samples that contain a large number of cells by revealing the presence and quantity of RNA. With bulk RNA-seq data, significant results on gene-to-gene connection and gene-to-disease relationship can be obtained by machine learning methods, including dimension reduction, clustering models, and graphical models.

However, bulk RNA-seq only measures average gene expression levels across all cells in the sample, and it cannot detect gene expression differences across different types of cells. Single cell RNA sequencing (scRNA-seq) solves this problem by measuring the gene expression of individual cells, allowing us to clarify the critical difference among cells from the same organism. This genomic technology has helped discovering rare cells in different tissues by gene expression patterns and therefore is an important and powerful tool for transcriptome analysis [19].

Yet, data quality of scRNAseq is poorer than that of the bulk RNAseq, especially because of the presence of *dropouts*, technical artifacts where genes erroneously appear to have zero expression due to sequencing inefficiency. The loss of information is significant in scRNA-seq data, and can lead to major problems in downstream analyses.

Numerous *imputation* methods have been developed to fill in the dropout values in the scRNA-seq data. The SAVER model [14] predicts gene expressions under the assumption that the measured gene expressions follow Poisson-Gamma

distributions, where the latent Gamma random variables are the true gene expressions. Based on the similarity among cells' gene expressions, both drImpute [9] and PRIME [16] impute the dropouts of a cell by using the gene expressions of the cells belonging to the same cluster. The scRMD methodology [1] infers the gene expressions of cells by robust matrix decomposition, where the dropouts are encoded in a sparse matrix, and the matrix of true gene expressions is low rank. Other approaches include [30] and [37].

Yet, effective imputation of the missing values in scRNA-seq data can be difficult and biased because the data is sparse and high-dimensional – the number of genes is typically over 20,000 and the number of cells is usually only a few hundreds. In fact, researchers are interested in gene-to-gene connections and interactions, clustering of cells or principal component analysis of the scRNA-seq, and all these analyses require a well estimated correlation matrix of the gene expressions. Unfortunately, the presence of dropouts generates several challenges. For instance, if all zeros are assumed to be true values, the sample correlation matrix is corrupted. On the other hand, assuming that all zeros are dropouts, i.e. missing values, the Pearson correlation of two genes may be computed empirically only if there are enough pairwise-complete observations, otherwise it is infeasible. That is, the presence of dropouts can cause missingness in the sample covariance matrix. For this last problem there are plenty of covariance matrix completion methodologies that may be used [26, 21, 10, 17], but these, just like the data imputation methods [14, 9, 16, 1], perform ideally under different assumptions about the data.

However, in this challenging situation *auxiliary information* could be very helpful. Indeed, improvements in estimation performance due to the use of auxiliary information have been observed in various contexts [5, 11, 25, 8, 22, 31, 27, 23]. For instance, Hecker et al. [11] discuss improvements in network inference allowed by the incorporation of genome sequence and protein-DNA interaction data. Moreover, Lin et al. [25] study age-related macular degeneration by incorporating prior knowledge from previous linkage and association studies. Furthermore, Gao et al. [8] and Li et al. [22] use information about Gene Ontology annotation to improve network estimation. Finally, Novianti et al. [27] use gene pathway databases and genomic annotations to improve prediction accuracy, and Liang et al. [23] use auxiliary information about gene length and test statistics from microarray studies for the analysis of differential expression of genes.

In this paper we propose a novel approach, **SCENA** (*Single cell RNA-seq Correlation completion by ENsemble learning and Auxiliary information*), which estimates the gene-by-gene correlation matrix by incorporating auxiliary information about the underlying biological connections and other data sources together with the dropout-corrupted scRNA-seq data of interest. The auxiliary information we use includes the gene pathway database from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [18], gene interaction networks from the Biological General Repository for Interaction Datasets (BioGRID) [28], protein-protein interaction networks from STRING [29], bulk cell RNA-seq data of 39 different tissues from the Encyclopedia of DNA Elements (ENCODE) [3], and other scRNA-seq data collected from cells of the same type of organism as the scRNA-seq dataset of interest. To implement SCENA, we first convert all these auxiliary data sources into a collection of correlation matrices, in addition to the correlation matrices recovered from scRNA-seq data via various imputation approaches and matrix completion strategies. Then, we ensemble all obtained correlation matrices into a final gene-by-gene correlation matrix estimate by *model stacking*.

We show that SCENA outperforms other existing methods in terms of correlation matrix completion, dimension reduction, clustering, and graphical modeling with an extensive simulation study (Section 3). Finally, we apply the methods to the analysis of a massive scRNA-seq data set of embryonic stem cells [2] (Section 4).

## 2 Method: SCENA

Let  $X_s$  be an  $N \times M$  scRNA-seq data matrix of gene expressions of  $N$  genes measured over  $M$  cells. Because of technology limitations of the sequencing process, the data matrix  $X_s$  typically contains numerous false zeros called “dropouts”, where the transcript was not detected at the sequencing process. Thus,  $X_s$  must be seen as a corrupted version of some underlying data matrix  $X$  that is free from dropouts. To obtain estimates of the gene-by-gene correlations, in traditional approaches the data matrix  $X_s$  is typically first subject to some *imputation* process  $\mathcal{I}$ , which identifies the dropouts and predicts their values. The resulting imputed data matrix  $\mathcal{I}(X_s)$  is then used to compute correlation estimates. Alternatively to this approach, we may complete or repair directly the correlation estimates obtained from  $X_s$  by standard *matrix completion* approaches, or approximate them by using various sources of *auxiliary information*.

Thus, several possible useful estimates of the single-cell gene-by-gene correlations are available, and a combination of them may let us obtain an ultimate reliable estimate of the gene-by-gene correlation matrix. Our proposed approach, **SCENA**, builds upon this strikingly simple but powerful idea of optimally combining multiple correlation matrix estimates. SCENA estimates the gene expression correlation matrix of scRNA-seq data by combining multiple genetic correlation matrices derived from various sources of information. In Section 2.1 we describe the derivation of several

correlation matrix estimates, and in Section 2.2 we combine them via model stacking. In the rest of the paper, gene expressions are transformed according to  $x \mapsto \log_2(1 + x)$  before computing correlations.

## 2.1 Single correlation matrix estimates

SCENA combines the following four groups of single correlation matrix estimates.

1. **BLIND CORRELATION ESTIMATE (“ALL ZEROS ARE TRUE”)** This is the sample correlation matrix  $\hat{\Sigma}_s$  of the scRNA-seq data matrix  $X_s$  assuming all zeros are real zeros.

2. **IMPUTATION (“SOME ZEROS ARE DROPOUTS AND WE TRY TO CORRECT THEM”).** The imputation methods [14, 9, 16, 1] let us impute the scRNA-seq data, and thereby obtain correlation matrices.

3. **CORRELATION MATRICES BASED ON AUXILIARY DATA.** Auxiliary is any kind of data that is beyond the scRNA-seq data of interest. We consider the following three kinds of auxiliary data, which can be used to compute correlation matrices.

- (a) *Bulk RNA-seq data.* Given a matrix of bulk RNA-seq data, we calculate its sample correlation matrix. In this paper we use auxiliary bulk RNA-seq data of [3].
- (b) *Other scRNA-seq data.* It is possible that another scRNA-seq data set  $X_s^*$  presents less dropouts for some of the genes in the main data matrix  $X_s$ , so the lost information of those genes might be found in  $X_s^*$ . The sample correlation matrix of such additional data matrix  $X_s^*$  is computed. In this paper we use auxiliary scRNA-seq data from [35, 20].
- (c) *Biological networks.* The KEGG pathway database [18] provides information about  $n = 6860$  genes and  $c = 239$  gene pathways which can be summarized in an  $n \times c$  matrix  $K = [K_{ij}]$  where  $K_{ij} = 1$  if gene  $i$  is in pathway  $j$ ,  $K_{ij} = 0$  otherwise. We compute the  $n \times n$  sample correlation matrix of  $K$ . From the BioGRID network [28] we extract an adjacency matrix  $A \in \{0, 1\}^{N \times N}$  of gene connections, and obtain the correlation matrix  $\text{diag}(L^{-1})^{-\frac{1}{2}} L^{-1} \text{diag}(L^{-1})^{-\frac{1}{2}}$ , where  $L$  is the Laplacian matrix  $L = D - A$ , and  $D$  is a diagonal matrix with  $D_{ii} = \sum_{j=1}^N A_{ji}$ , i.e. the degree of gene  $i$ . Finally, in the STRING network [29], we construct a correlation matrix by treating gene-by-gene combined connection scores as correlations.

4. **SKEPTICAL CORRELATION ESTIMATES (“ALL ZEROS ARE DROPOUTS”)** Assuming all zeros are dropouts, we obtain the matrix  $X_s^{\text{NA}}$  which corresponds to  $X_s$  with all zeros replaced by missing values NAs. From this matrix, it is possible to compute the *pairwise complete* sample correlation matrix  $\hat{\Sigma}_O$ , which contains missing entries for all those gene pairs with no jointly nonzero measured scRNA-seq expressions. We obtain completed versions of  $\hat{\Sigma}_O$  as follows.

- (a) *Matrix completion.* We use [26] to produce a complete correlation matrix.
- (b) *Convex combinations with matrices in (3).* Given a correlation matrix derived from auxiliary data, say  $\hat{\Sigma}_{aux}$ , we can obtain a completed version of the pairwise complete scRNA-seq sample correlation matrix  $\hat{\Sigma}_O$  as  $\hat{\Sigma} = \alpha \odot \hat{\Sigma}_O^* + (1 - \alpha) \odot \hat{\Sigma}_{aux}$ , where  $\hat{\Sigma}_O^*$  is a version of  $\hat{\Sigma}_O$  with all NA’s replaced by zeros, and  $\alpha = [\alpha_{ij}]$  is a  $N \times N$  weight matrix. We consider two types of weights:

(i) *Simple replacement*

$$\alpha_{ij} = \begin{cases} 1, & \text{if } \hat{\Sigma}_{O,ij} \text{ is not NA} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

(ii) *Signal-to-noise ratio*

$$\alpha_{ij} = \frac{1}{M} \sum_{k=1}^M I(X_{s,ik} \neq 0) I(X_{s,jk} \neq 0), \quad (2)$$

which is the proportion of cells where genes  $i$  and  $j$  have jointly nonzero read counts.

## 2.2 Model stacking

Let  $\hat{\Sigma}_1, \dots, \hat{\Sigma}_p$  be the single correlation matrix estimates derived in Section 2.1. We aim to obtain a final correlation matrix estimate  $\tilde{\Sigma}$  by model stacking in the form

$$\tilde{\Sigma} = F^{-1} \left( \sum_{q=1}^p \beta_q F(\hat{\Sigma}_q) \right), \quad (3)$$

where  $F : \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^{N \times N}$  is an invertible mapping, and  $\beta_1, \dots, \beta_q \in \mathbb{R}$ . The simplest choice of  $F$  is the identity mapping  $F(A) = A, \forall A \in \mathbb{R}^{N \times N}$ , which however does not guarantee  $\tilde{\Sigma}$  to be a positive semi-definite correlation matrix, even if all  $\hat{\Sigma}_1, \dots, \hat{\Sigma}_p$  are positive semi-definite correlation matrices, unless we impose appropriate constraints on  $\beta_1, \dots, \beta_q$ . For instance, a sufficient condition is  $\beta_q > 0, \forall q$ , with  $\sum_q \beta_q = 1$ , which specifies a convex linear combination. Another possible mapping is  $F(A) = A^{-1}$ , which requires  $\hat{\Sigma}_p \succ 0, \forall p$ . In any case, if  $\tilde{\Sigma}$  is not a positive semi-definite correlation matrix, we replace it with the nearest correlation matrix [13] as per  $\tilde{\Sigma} := \arg \min_{\Psi \in C} \|\tilde{\Sigma} - \Psi\|_F$ , where  $C$  is the set of positive semi-definite correlation matrices.

There are many possible ways to specify Equation (3). We consider the following ones.

### 2.2.1 Simple average.

The simple average is obtained by setting  $F =$  identity mapping and  $\beta_q = \frac{1}{p}$ , for all  $q$ , yielding the convex linear combination

$$\tilde{\Sigma} = \frac{1}{p} \sum_{q=1}^p \hat{\Sigma}_q \quad (4)$$

Since the weights are prespecified, this approach requires no additional tuning or validation steps. Also, if  $\hat{\Sigma}_1, \dots, \hat{\Sigma}_p$  are all positive semi-definite correlation matrices, so is  $\tilde{\Sigma}$ . We will denote this solution by **SCENA<sub>average</sub>**.

### 2.2.2 Regression.

We assume a linear relationship between the true underlying correlation and the single correlation estimates,

$$f(\Sigma_{ij}) = \sum_{q=1}^p \beta_q f(\hat{\Sigma}_{q,ij}) + \epsilon_{ij}, \quad (5)$$

where  $f : (-1, 1) \rightarrow \mathbb{R}$  is an invertible function, e.g. the Fisher transformation  $f(x) = \frac{1}{2} \log((1+x)/(1-x))$ , and  $\epsilon_{ij}$  is an error component. The vector of coefficients  $\beta = (\beta_1, \dots, \beta_q)^T$  is then estimated by solving the penalized optimization problem

$$\hat{\beta} = \arg \min_{\beta} \sum_{i < j} \left( f(\Sigma_{ij}) - \sum_{q=1}^p \beta_q f(\hat{\Sigma}_{q,ij}) \right)^2 + \lambda \mathcal{P}(\beta) \quad (6)$$

where  $\mathcal{P}$  is a penalty and  $\lambda \geq 0$  is selected via cross-validation. Cross-validation lets us reduce the risk of overfitting, and is implemented by creating multiple held-out data subsets that are iteratively removed from training and used instead to validate prediction accuracy. Setting  $\mathcal{P}(\beta) = \sum_{q=1}^p \beta_q^2$  produces the *ridge estimator*, and the resulting final correlation matrix  $\tilde{\Sigma}$  which will be denoted by **SCENA<sub>ridge</sub>**.

Of course, we do not know  $\Sigma_{ij}$  in Equation (5), but we can identify a small subset of genes and cells which we may assume to contain very few dropouts and could give us reliable estimates of  $\Sigma_{ij}$ . Thus, to fit the regression model in Equation (5), we first extract a reference data matrix  $Y'$  from the scRNA-seq data matrix  $X_s$  (Algorithm A), then compute the sample correlation matrix  $\hat{\Sigma}_{Y'}$  of  $Y'$ , and finally extract the off-diagonal entries which will be used as the response vector of the regression. Then, we obtain multiple perturbed versions of  $Y'$  by creating artificial dropouts (Algorithm A). The off-diagonal entries of the matrices  $\hat{\Sigma}_q$ , for  $q = 1 \dots p$ , based on the perturbed data, are used as predictors. Algorithm A summarizes the full procedure.

	<b>chu</b>	<b>chu_time</b>	<b>darmanis</b>
<b>Tissue</b>	embryonic stem cells	definitive endoderm cell	brain cell
<b># cell types</b>	7	6	5
<b># cells</b>	1,018	758	366
<b># genes</b>	21,413	18,294	17,738
<b>% zeros</b>	47.43%	51.15%	80.06%
<b># reference cells</b>	951	689	332
<b># reference genes</b>	2,522	2,160	2,306
<b>citation</b>	[2]	[2]	[4]
<b>GEO accession code</b>	GSE75748	GSE75748	GSE67835

Table 1: Human scRNA-seq data sets used in simulations.

### 3 Simulations

In this section we present an extensive simulation study showing that SCENA is superior to other methods in terms of correlation matrix completion, dimensionality reduction, clustering, and conditional dependence graphical modeling. In Section 3.1 we describe how we generate realistic artificial scRNA-seq data based on real data sets. Specifically, given a real scRNA-seq data set, we first extract a *reference data set*, a subset of data where all zeros can be safely assumed to be true values and not dropouts. Then, we generate *downsampled data* by creating dropouts in the reference scRNA-seq data according to the Poisson-Gamma scheme in Algorithm A. Finally, in Section 3.2 we assess the performance of SCENA and other existing methods at recovering the correlation structure of the reference data based on the corrupted downsampled data and other available auxiliary data.

#### 3.1 Generating scRNA-seq data

##### 3.1.1 Original data sets

We use three human scRNA-seq data sets in this simulation study:

1. **chu**: human embryonic stem cells [2].
2. **chu\_time**: human definitive endoderm cells (time-series sequencing) [2].
3. **darmanis**: human brain cells [4].

The number of cells and number of cell types are reported in Table 1 (genes with zero expression in all cells are removed).

##### 3.1.2 Reference data sets

For each of the three data sets, we first match the genes with those available in the auxiliary data (Section 2.1), and then apply Algorithm A to perform quality control by filtering out low quality genes and cells, and finally extract reference data. All values in the reference data are treated as true gene expressions, i.e. the reference data is free from dropouts. The dimensions of the three resulting reference data sets are reported in Table 1.

##### 3.1.3 Downsampled data

For each of the three reference data sets, we apply Algorithm A to generate downsampled versions of the reference data. We set  $s = 10$ , and  $r = 3000, 1000, 1000$  for **chu**, **chu\_time** and **darmanis** data, respectively, to ensure the expected percentage of zeros in the downsampled data to be similar to the percentage of zeros in the original scRNA-seq data (Table 1).

#### 3.2 Models comparison

In this section we assess the performance of SCENA and other existing methods at recovering the correlation structure of the reference data based on the corrupted downsampled data and other available auxiliary data. We show that SCENA<sub>average</sub> and SCENA<sub>ridge</sub> (Section 2.2) outperform SAVER, drImpute, scRMD, and PRIME in terms of correlation matrix completion, dimension reduction, clustering, and graphical modeling. The results shown are averaged across multiple downsampled data sets.

	MSE			CMD		
	chu	chu_time	darmanis	chu	chu_time	darmanis
Downsample	<b>0.01008</b>	0.00796	0.00804	0.21217	0.15855	0.22897
SAVER	0.01949	0.01850	0.01476	0.44604	0.40991	0.52235
drImpute	0.01181	0.00796	0.01371	0.23595	0.16050	0.33769
scRMD	0.01125	0.00868	0.00879	0.23598	0.17621	0.26170
PRIME	0.03707	0.03508	0.02661	0.47846	0.42137	0.33218
SCENA_average	0.01285	0.01119	0.01313	0.27283	0.21596	0.42946
SCENA_ridge	0.01212	<b>0.00708</b>	<b>0.00513</b>	<b>0.20497</b>	<b>0.12942</b>	<b>0.12474</b>

Table 2: Correlation completion accuracy. MSE and Correlation Matrix Distance (CMD) between the reference correlation and the estimated correlation derived from various methods. SCENA<sub>ridge</sub> has the lowest MSE and CMD among imputation methods in chu\_time and darmanis data, and is lower than the baseline (“Downsample”), which is the sample correlation matrix of the downsampled data. All other imputation approaches perform worse than the baseline.

### 3.2.1 Correlation completion

We measure the similarity between the correlation matrix estimators considered and the reference correlation matrix  $\hat{\Sigma}_{\text{ref}}$  in terms of *mean squared error* (MSE) and average *correlation matrix distance* (CMD) [12]. Table 2 shows that SCENA outperforms all other data imputation methods in terms of both MSE and CMD. The baseline is set to be the MSE and CMD between  $\hat{\Sigma}_{\text{ref}}$  and the sample correlation of the downsampled data (blind estimate  $\hat{\Sigma}_s$ , Section 2.1), treating all 0s as true gene expressions. SCENA<sub>ridge</sub> has the lowest MSE and CMD among all methods as well as the baseline in chu\_time and darmanis data. SCENA<sub>average</sub> is also better in correlation completion than SAVER and PRIME.

### 3.2.2 Dimension reduction

For each correlation matrix estimate from inputted data and SCENA, we compute the matrix of eigenvectors  $V$ , and obtain the principal component scores  $U = Z^T V$ , where  $Z$  is a standardized version of the log transformed scRNA-seq data  $X_s$ . In Figure 1, we compare the scatterplots of the top two PC scores of the cells against each other. The plots are colored by cell type labels, and PCs are derived from sample correlations of reference data, SAVER imputed data, and correlation estimations of SCENA<sub>ridge</sub> and SCENA<sub>average</sub>. Both SCENA<sub>ridge</sub> and SCENA<sub>average</sub> recover the reference data structure better than data imputation, and yield scatterplots with a clear separation among different types of cells indicated by the cell type labels.

### 3.2.3 Clustering

We perform hierarchical cells clustering (Ward’s minimum variance method with Manhattan distance; [7]) based on the standardized principal components of the downsample scRNA-seq data obtained from the different approaches considered (quantities  $U$  computed in Section 3.2.2). For each method, we use the top PCs with proportion of variance explained within the range (90%,99%), and set the number of clusters equal to the number of true cell labels in the scRNA-seq data. To assess clustering performance, we measure the similarity between cluster assignments and true cell labels by calculating the *adjusted rand index* (ARI). This metric takes values in the interval  $[0, 1]$ , with large values indicating stronger similarity. In Figure 2 we can see that SCENA<sub>average</sub> yields the best clustering performance over all other methods in all three datasets. Interestingly, in the chu data, SCENA<sub>average</sub> has even better performance than the clustering obtained from reference data, in accordance with the fact that SCENA exploits auxiliary information besides the scRNA-seq data.

### 3.2.4 Conditional dependence graphs

Conditional dependence graph estimation in the case where several pairs of variables are never observed jointly is a major statistical problem that has gained strong interest recently; a thorough theoretical investigation of the so called *graph quilting problem* can be found in [32]. Such problem is strictly related to ours, where an extremely large number of gene pairs have no reliable empirical correlation estimates. Here, we investigate the graph recovery performance based on the various correlation matrix estimates via simulations. Specifically, we plug the correlation matrix estimates into the graphical lasso [36] to obtain gene-by-gene conditional dependence graphs via sparse precision matrix estimation. For simplicity, we compute graphs about only the top 50 most variable genes among cell types, identified by applying ANOVA to gene expression of reference data adjusted by cells’ library sizes. To evaluate the

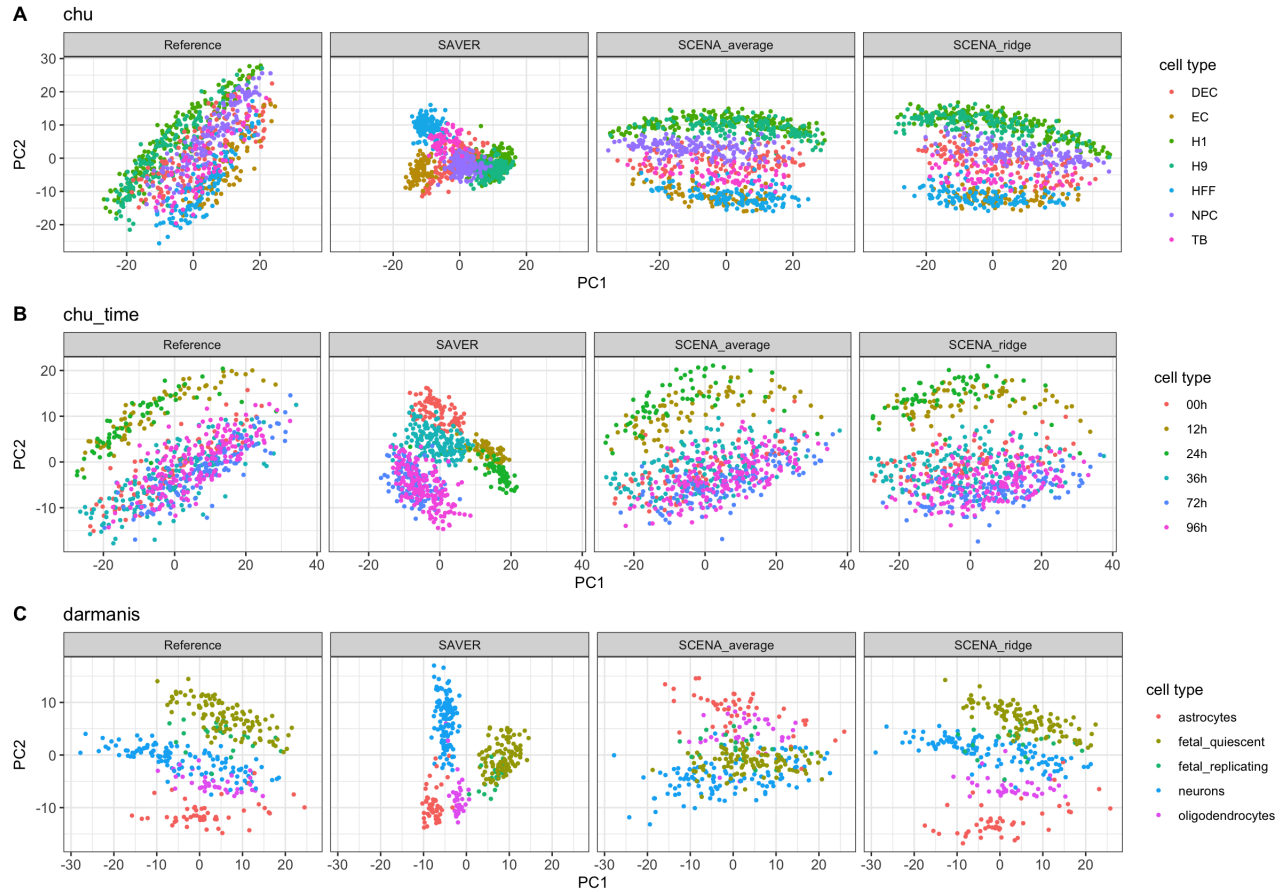


Figure 1: Dimension reduction accuracy. Scatterplots of the top two PC scores of the cells colored by cell type. Both  $SCENA_{ridge}$  and  $SCENA_{average}$  appear to recover the reference data structure better than SAVER, yielding scatterplots with a clear separation among different types of cells.

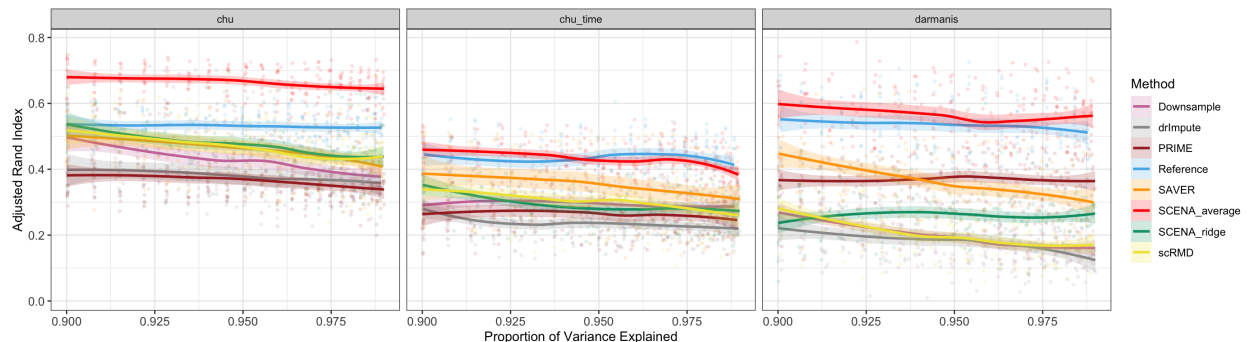


Figure 2: Clustering performance. Adjusted rand index (higher is better) of cell type grouping via hierarchical clustering after dimension reduction via PCA explaining various proportions of variance.  $SCENA_{average}$  yields the best clustering performance over all other methods in all data sets, and even better than the clustering obtained from the reference data in the chu data.

graph recovery performance of a method, we compute the F1-score with respect to the graph estimated from reference data. In Figure 3A we plot F1-score versus number of graph edges for all methods and data sets.  $SCENA_{ridge}$  is superior in recovering the reference graph than other methods in chu and darmanis data, and it produces a similarly high F1-score in chu\_time data as SAVER method. For illustration, in Figure 3B we also display conditional dependence graphs relative to the chu data with 50 edges.

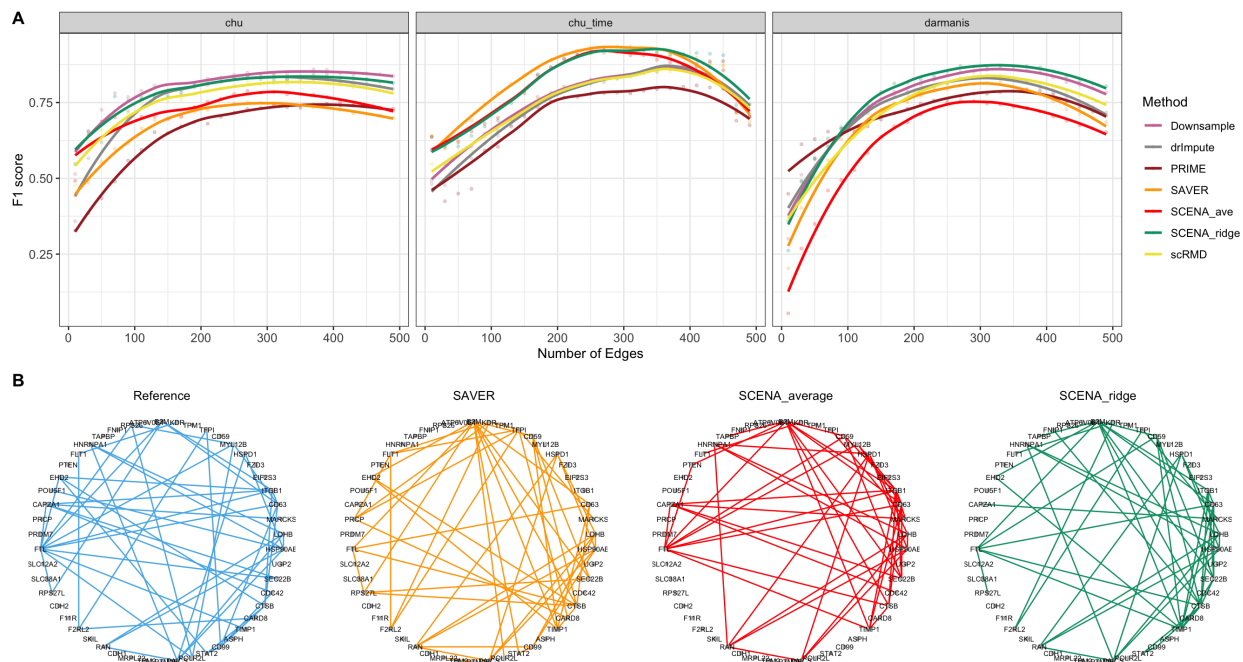


Figure 3: Genetic graph recovery. A: F1 score (higher is better) quantifying the performance of methods at recovering the reference conditional dependence graphs of 50 most variable genes for various numbers of edges.  $SCENA_{ridge}$  exhibits strong performance for all data sets, while other methods' performance dramatically changes across different data sets. B: Conditional dependence graphs of chu data, setting the number of edges to 50.

## 4 Application to stem cell data

We now apply the methods to the analysis of the chu data set (Table 1) containing the gene expression of 6,038 genes (largest genes set that matched available auxiliary information) measured in 1,018 human embryonic stem cells. In Figure 4A we plot the first two principal components based on  $SCENA_{average}$  and  $SCENA_{ridge}$ , while in Figure 4B we compare the cell clustering performance of SCENA with other methods in terms of ARI. The hierarchical clustering based on  $SCENA_{average}$  performs the best at recovering true cell type labels, in accordance with simulation results (Section 3.2.3). Finally, in Figure 4C we display the conditional dependence graph (graphical lasso; [36]) of the 30 most variable genes among cell types (ANOVA criterion as in Section 3.2.4) based on  $SCENA_{ridge}$ , with number of edges 163 selected via Extended Bayesian Information Criterion (EBIC, [6]). The protein coding gene *DNMT3B* is the hub node with largest number of connections (20 edges). This result is reasonable because *DNMT3B* is a catalytically active DNA methyltransferase [24], and is specifically expressed in totipotent embryonic stem cells [34]. Moreover, *DNMT3B* is one of the pluripotency markers with high level of expression in the cell type H1, as demonstrated by [2]. Besides, genes *IFI16* [15] and *HAND1* [33] are marker genes in cell type EC and cell type TB, respectively, and correspondingly have relatively large numbers of connections.

## 5 Discussion

We have proposed and studied SCENA, a novel methodology for gene-by-gene correlation matrix estimation from dropout-corrupted single cell RNA-seq data. SCENA builds upon the strikingly simple but powerful idea of optimally combining multiple gene-by-gene correlation matrices derived from various sources of information, besides the scRNA-seq data of interest. This combination is implemented efficiently via model stacking techniques.

We have demonstrated that SCENA can provide superior estimation performance compared to traditional data imputation methods. In our analyses,  $SCENA_{ridge}$  remarkably recovered the information underlying the corrupted scRNA-seq data in terms of correlation completion, dimension reduction, and graphical modeling, while the hierarchical clustering based on  $SCENA_{average}$  yielded cell groupings that best reflected true cell type heterogeneity in terms of adjusted rand index. Indeed, although both variants combine the same single correlation matrices via model stacking, the



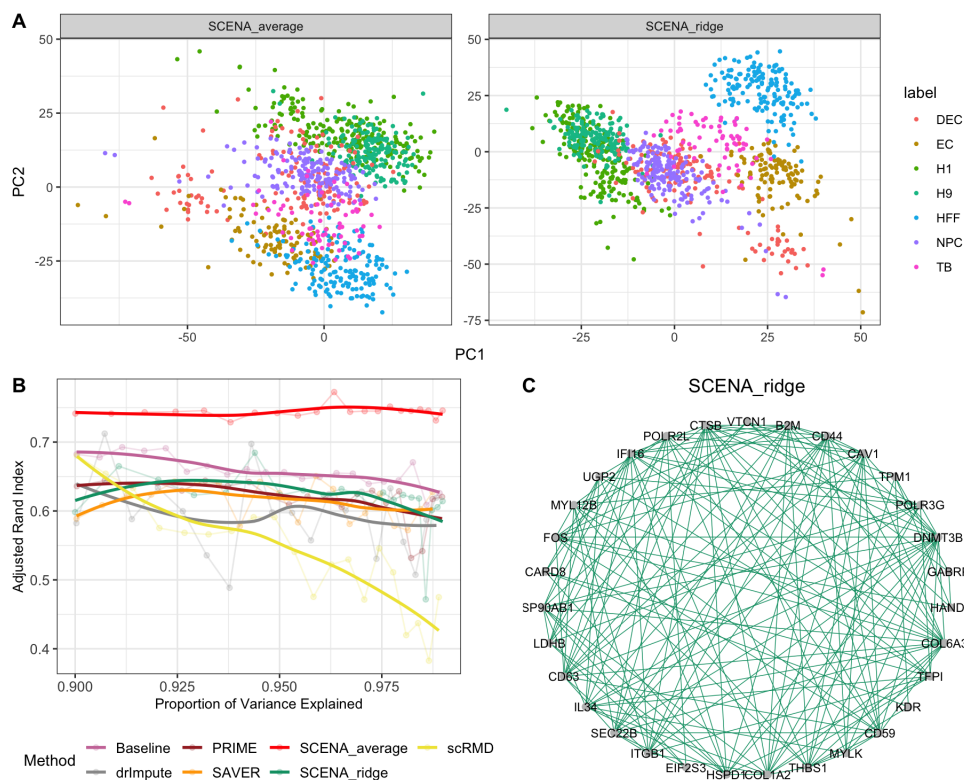


Figure 4: Real data application results for the chu data. A. Dimension reduction: scatterplots of the top two PC scores of the cells colored by cell type. B. Clustering: adjusted rand index of cell type grouping via hierarchical clustering after dimension reduction using PCA explaining various proportions of variance. SCENA<sub>average</sub> yields the best clustering performance over all other methods. C: Conditional dependence graph (graphical lasso; 163 edges selected via EBIC) based on SCENA<sub>ridge</sub> correlation estimate. Gene *DNMT3B* is the hub node with the largest number of connections (20 edges). This result is supported by the scientific literature as *DNMT3B* is one of the pluripotency markers with high level of expression in the cell type H1 [2]. Also, genes *IFI16* and *HAND1* are marker genes in cell type EC and cell type TB, respectively, and correspondingly have relatively large numbers of connections.

weighting coefficients of SCENA<sub>ridge</sub> are calibrated for the optimal recovery of the true correlation structure of the corrupted gene expression data, while SCENA<sub>average</sub> simply assigns uniform weights, presumably upweighting auxiliary biological network structures that are more informative about cell characteristics. SCENA<sub>average</sub> is computationally cheaper than SCENA<sub>ridge</sub>, because the estimation of the weighting coefficients of SCENA<sub>ridge</sub> involves multiple additional imputation and optimization steps that are computationally expensive. For instance, in the application presented in Section 4, the model stacking step for SCENA<sub>ridge</sub> took about 40 minutes, while only about 30 seconds for SCENA<sub>average</sub>, on a laptop with 16GB of RAM (2133 MHz) and dual-core processor (3.1 GHz). Given all these considerations, we recommend to use SCENA<sub>average</sub> for the analysis of massive scRNA-seq data sets.

While we have demonstrated our approach using specific auxiliary sources, SCENA is general and conducive to many different types of correlation imputation approaches and additional sources of auxiliary information on genetic interactions. Additionally, our approach can be further optimized using different machine learning approaches to model stacking and ensemble learning. Overall, we expect SCENA to become an important instrument for downstream analyses of massive scRNA-seq data that powerfully incorporates known auxiliary information on genetic interactions.

L.G., G.V., and G.A. are supported by NIH 1R01GM140468, NSF DMS-1554821 and NSF NeuroNex-1707400. G.V. is additionally supported by a Rice Academy Postdoctoral Fellowship and the Dan L. Duncan Foundation.

## References

- [1] C. Chen, C. Wu, L. Wu, Y. Wang, M. Deng, and R. Xi. scrmid: Imputation for single cell rna-seq data via robust matrix decomposition. *bioRxiv*, page 459404, 2018.

- [2] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.
- [3] E. P. Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [4] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. H. Gephart, B. A. Barres, and S. R. Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.
- [5] E. Ferkingstad, A. Frigessi, H. Rue, G. Thorleifsson, A. Kong, et al. Unsupervised empirical bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 2(2):714–735, 2008.
- [6] R. Foygel and M. Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612, 2010.
- [7] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
- [8] S. Gao and X. Wang. Quantitative utilization of prior biological knowledge in the bayesian network modeling of gene expression data. *BMC bioinformatics*, 12(1):359, 2011.
- [9] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, 19(1):220, 2018.
- [10] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.
- [11] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96(1):86–103, 2009.
- [12] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek. Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels. In *2005 IEEE 61st Vehicular Technology Conference*, volume 1, pages 136–140. IEEE, 2005.
- [13] N. J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.
- [14] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539, 2018.
- [15] T. P. Hurst, A. Aswad, T. Karamitros, A. Katzourakis, A. L. Smith, and G. Magiorkinis. Interferon-inducible protein 16 (ifi16) has a broad-spectrum binding ability against ssdna targets: an evolutionary hypothesis for antiretroviral checkpoint. *Frontiers in microbiology*, 10:1426, 2019.
- [16] H. Jeong and Z. Liu. Prime: a probabilistic imputation method to reduce dropout effects in single cell rna sequencing. *bioRxiv*, 2020.
- [17] J. Josse, S. Sardy, and S. Wager. denoiser: A package for low rank matrix estimation. *arXiv preprint arXiv:1602.01206*, 2016.
- [18] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462, 2016.
- [19] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.
- [20] B. B. Lake, S. Chen, B. C. Sos, J. Fan, G. E. Kaeser, Y. C. Yung, T. E. Duong, D. Gao, J. Chun, P. V. Kharchenko, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology*, 36(1):70–80, 2018.
- [21] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.
- [22] Y. Li and S. A. Jackson. Gene network reconstruction by integration of prior biological knowledge. *G3: Genes, Genomes, Genetics*, 5(6):1075–1079, 2015.
- [23] K. Liang et al. Empirical bayes analysis of rna sequencing experiments with auxiliary information. *The Annals of Applied Statistics*, 13(4):2452–2482, 2019.
- [24] J. Liao, R. Karnik, H. Gu, M. J. Ziller, K. Clement, A. M. Tsankov, V. Akopian, C. A. Gifford, J. Donaghey, C. Galonska, et al. Targeted disruption of dnmt1, dnmt3a and dnmt3b in human embryonic stem cells. *Nature genetics*, 47(5):469, 2015.

- [25] W.-Y. Lin and W.-C. Lee. Incorporating prior knowledge to facilitate discoveries in a genome-wide association study on age-related macular degeneration. *BMC research notes*, 3(1):26, 2010.
- [26] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.
- [27] P. W. Novianti, B. C. Snoek, S. M. Wilting, and M. A. Van De Wiel. Better diagnostic signatures from rnaseq data through use of auxiliary co-data. *Bioinformatics*, 33(10):1572–1574, 2017.
- [28] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, et al. The biogrid interaction database: 2011 update. *Nucleic acids research*, 39(suppl\_1):D698–D704, 2010.
- [29] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.
- [30] S. Tracy, G.-C. Yuan, and R. Dries. Rescue: imputing dropout events in single-cell rna-sequencing data. *BMC bioinformatics*, 20(1):388, 2019.
- [31] M. A. Van De Wiel, T. G. Lien, W. Verlaat, W. N. van Wieringen, and S. M. Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, 35(3):368–381, 2016.
- [32] G. Vinci, G. Dasarathy, and G. I. Allen. Graph quilting: graphical model selection from partially observed covariances. *arXiv preprint arXiv:1912.05573*, 2019.
- [33] V. Wagh, A. Pomorski, K. J. Wilschut, S. Piombo, and H. S. Bernstein. Microrna-363 negatively regulates the left ventricular determining transcription factor hand1 in human embryonic stem cell-derived cardiomyocytes. *Stem cell research & therapy*, 5(3):75, 2014.
- [34] D. Watanabe, I. Suetake, T. Tada, and S. Tajima. Stage-and cell-specific expression of dnmt3a and dnmt3b during embryogenesis. *Mechanisms of development*, 118(1-2):187–190, 2002.
- [35] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131, 2013.
- [36] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.
- [37] L. Zhu, J. Lei, B. Devlin, and K. Roeder. A unified statistical framework for single cell and bulk rna sequencing data. *The annals of applied statistics*, 12(1):609, 2018.

## A Algorithms

[Reference data selection]

INPUT:  $N \times M$  data matrix  $X$ ; parameter vector  $a$ .

1. Filter out cells with library size greater than  $a_1$ -th percentile.
2. Remove genes with mean expression less than  $a_2$ -th percentile.
3. Remove genes with less than  $a_3$ -th percentile non-zero cells.
4. Keep cells with library size greater than the  $a_4$ -th percentile.
5. Keep genes with non-zero proportion greater than  $a_5$ -th percentile.

OUTPUT:  $N' \times M'$  reference data matrix  $Y$ . We use default values  $a_1 = 95$ ,  $a_2 = 25$ ,  $a_3 = 15$ ,  $a_4 = 5$ ,  $a_5 = 50$ .

[Poisson-Gamma downsampling]

INPUT:  $N \times M$  data matrix  $X$ ; parameters  $s, r > 0$ .

1. Draw  $Z_1, \dots, Z_M \stackrel{\text{i.i.d.}}{\sim} \Gamma(s, r)$ .
2. Draw  $\tilde{X}_{ij} \sim \text{Poisson}(X_{ij}Z_j)$ , for  $i = 1, \dots, N$  and  $j = 1, \dots, M$ .

OUTPUT:  $N \times M$  downsampled data matrix  $\tilde{X}$ .

[Stacking regression validation]

INPUT:  $N \times M$  scRNA-seq data matrix  $X$ ; reference parameter vector  $a$ ; downsampling parameters  $s, r > 0$ ; collection of  $N \times N$  single correlation matrices (Section 2.1); number of downsampling repeats  $B$ ; transform function  $f$ .

1. Obtain reference  $N' \times M'$  data matrix  $X'$  via Algorithm A with parameter  $a$ .
2. Construct response vector  $y \in \mathbb{R}^{\frac{N'(N'-1)}{2}}$  by extracting off-diagonal entries from  $\hat{\Sigma}_{X'} = \text{cor}(X')$ .
3. For  $b = 1, \dots, B$ :
  - (a) Generate downsampled  $N' \times M'$  data matrix  $\tilde{X}_b$  from  $X$  via Algorithm A with parameters  $s, r$ .
  - (b) Obtain all  $N' \times N'$  single correlation estimates based on  $\tilde{X}_b$  and auxiliary correlation matrices  $\hat{\Sigma}_1^{(b)}, \dots, \hat{\Sigma}_p^{(b)}$ .
  - (c) Construct predictors matrix  $W^{(b)} \in \mathbb{R}^{\frac{N'(N'-1)}{2} \times p}$  by extracting off-diagonal entries from each  $\hat{\Sigma}_1^{(b)}, \dots, \hat{\Sigma}_p^{(b)}$ .
4. Compute  $\hat{\beta}$  by regressing  $f(\mathbf{y})$  on  $f(\mathbf{W})$  via Equation (6), where  $\mathbf{y} = (y^T, y^T, \dots, y^T)^T$  and  $\mathbf{W} = (W^{(1)T}, \dots, W^{(B)T})^T$ .

OUTPUT:  $N \times M$  correlation matrix  $\tilde{\Sigma}$  via Equation (3) with vector of coefficients  $\hat{\beta}$ .