Coupled Co-clustering-based Unsupervised Transfer Learning for the Integrative Analysis of Single-Cell Genomic Data

Pengcheng Zeng[1] and Zhixiang Lin *[1]

[1]Department of Statistics, The Chinese University of Hong Kong, HK

March 29, 2020

Abstract

Unsupervised methods, such as clustering methods, are essential to the analysis of single-cell genomic data. Most current clustering methods are designed for one data type only, such as scRNA-seq, scATAC-seq or sc-methylation data alone, and a few are developed for the integrative analysis of multiple data types. Integrative analysis of multimodal single-cell genomic data sets leverages the power in multiple data sets and can deepen the biological insight. We propose a coupled co-clustering-based unsupervised transfer learning algorithm (*couple*CoC) for the integrative analysis of multimodal single-cell data. Our proposed *couple*CoC builds upon the information theoretic co-clustering framework. We applied *couple*CoC for the integrative analysis of scATAC-seq and scRNA-seq data, sc-methylation and scRNA-seq data, and scRNA-seq data from mouse and human. We demonstrate that *couple*CoC improves the overall clustering performance and matches the cell subpopulations across multimodal single-cell genomic data sets. The software and data sets are available at https://github.com/cuhklinlab/coupleCoC

Keywords: Clustering, Unsupervised transfer learning, Information-theoretic co-clustering, Single-cell genomic

*Corresponding author: zhixianglin@sta.cuhk.edu.hk

1 Introduction

Single-cell sequencing technologies have been developed to profile genomic features for many cells in parallel. Methods have been developed that can profile different types of genomic features, including single-cell RNA sequencing (scRNA-seq) that profiles transcription, single-cell ATAC sequencing (scATAC-seq) that identifies accessible chromatin regions, single-cell methylation assays that profile methylated regions and other methods. The available data sets are growing in the area of single-cell genomics (Rotem et al., 2015; Cusanovich et al., 2015; Rozenblatt-Rosen et al., 2017). Compared to bulk genomic data, single-cell genomic data is characterized by high technical variation and high noise level, due to the minimal amount of genomic materials isolated from individual cells (Kharchenko et al., 2014; Lun et al., 2016; Vallejos et al., 2017; Zhu et al., 2018; Hicks et al., 2018). These experimental factors make it challenging to analyze single-cell genomic data, and they can impact the result and interpretation of unsupervised learning methods, including dimension reduction and clustering (Jaitin et al., 2014; Usoskin et al., 2015; Lafon and Lee, 2006; Vandermaaten, 2008).

Most of the clustering methods developed for single-cell genomic data are focused on one data type. Here we first review the methods for clustering scRNA-seq, scATAC-seq or methylation data alone. DIMM-SC (Sun et al., 2017) builds upon a Dirichlet mixture model and aims to cluster droplet-based single cell transcriptomic data. RaceID (Grün et al., 2016) uses an iterative k-means clustering algorithm based on a similarity matrix of Pearson's correlation coefficients. SIMLR (Wang et al., 2017) implements a kernel-based similarity learning algorithm, where RBF kernel was used with Euclidean distance. SC3 (Kiselev et al., 2017) and SAFE-clustering (Yang et al., 2018) belong to the class of ensemble clustering algorithms. The former combines the clustering outcomes of several other methods and the latter embeds several other methods and utilizes hypergraph-based partitioning algorithms. SOUP (Zhu et al., 2019) first identifies the set of pure cells by exploiting the block structures in cell-cell similarity matrix, uses them to build the membership matrix, and then estimates the soft memberships for the other cells. To cluster scATAC-seq data, *chrom*VAR (Schep et al., 2017) evaluates groups of peaks that share the same motifs or functional annotations together. *sc*ABC (Zamanighomi et al., 2018) weights cells by sequencing depth and applies weighted K-medoid clustering to reduce the impact of noisy cells with low sequencing depth. *cis*Topic (Gonzalez-Blas et al., 2019), applies latent Dirichlet allocation model on scATAC-seq data to identify *cis*-regulatory topics and simultaneously clusters cells and accessible regions based on the cell-topic and region-topic distributions. Xiong et al. (2019) proposed the SCALE method which combines a deep generative framework and a probabilistic Gaussian mixture model to learn latent features for scATAC-seq data. Clustering methods have also been proposed for methylation data (Shen et al., 2007; Siegmund et al., 2004). Siegmund et al. (2004) argued that model-based clustering techniques are often superior to nonparametric

approaches. Houseman et al. (2008) proposed a model-based clustering method for methylation data produced by Illumina GoldenGate arrays. Their method incorporates a beta mixture model (Ji et al., 2005). A weighted model-based clustering called LmiWCluster for Illumina BeadArray was introduced in Kuan et al. (2010). It weights each observation according to the detection P-values systematically and avoids discarding subsets of the data. Kapourani and Sanguinetti (2019) proposed Melissa, a Bayesian hierarchical method to cluster cells based on local methylation patterns, and Melissa was shown to discover patterns of epigenetic variability between cells.

High technical variation and high noise level in single-cell genomic data sets motivate the development of integrative analysis methods, which combines multiple data sets. To identify sub-populations of cells that are present across multiple data sets, Butler et al. (2018) introduced an analytical strategy for integrating scRNA-seq data sets based on common sources of variation, enabling the identification of shared populations across data sets and downstream comparative analysis. In order to compare data sets across experiments, Kiselev et al. (2018) presented a method called SCMAP for projecting cells from an scRNA-seq data set onto cell types or individual cells from other experiments. Also, scVDMC (Zhang et al., 2018), a method of variance-driven multi-task clustering of single-cell RNA-seq data, was proposed to cluster single cells in multiple scRNA-seq experiments of similar cell types and markers but varying expression patterns, by utilizing multiple single-cell populations from biological replicates or different samples. Duren et al. (2018) developed *couple*NMF algorithm to address the increasingly common situation where two or more types of sc-genomics experiments are performed on different sub-samples from the same cell population. *couple*NMF is based on extensions of non-negative matrix factorization. The connection between chromatin accessibility and gene expression builds upon prediction models trained from bulk data with diverse cell types. Stuart et al. (2019) developed a strategy to "anchor" diverse data sets together with the capability of integrating single-cell measurements not only across scRNA-seq technologies, but also across different modalities. Their work presented a strategy for the assembly of harmonized references and transfer of information across data sets. Lin et al. (2019) proposed *sc*ACE, a model-based approach to jointly cluster single-cell chromatin accessibility and single-cell gene expression data. The model does not rely on training data to connect the two data types and allows for statistical inference of the cluster assignment. A more comprehensive discussion on integration of single-cell data across samples, experiments, and types of measurement is presented in David et al. (2020).

In this paper, we try to answer the following questions: can we cluster one type of single-cell genomic data (scATAC-seq data or single-cell methylation data) by incorporating the knowledge from another type (scRNA-seq data)? And can we cluster single-cell genomic data from one species (mouse scRNA-seq data) by learning patterns from another species (human scRNA-seq data)? These questions belong to the field of unsupervised transfer learning. Unsupervised transfer learning aims to improve the learning of one data set, called the target data, using the knowledge

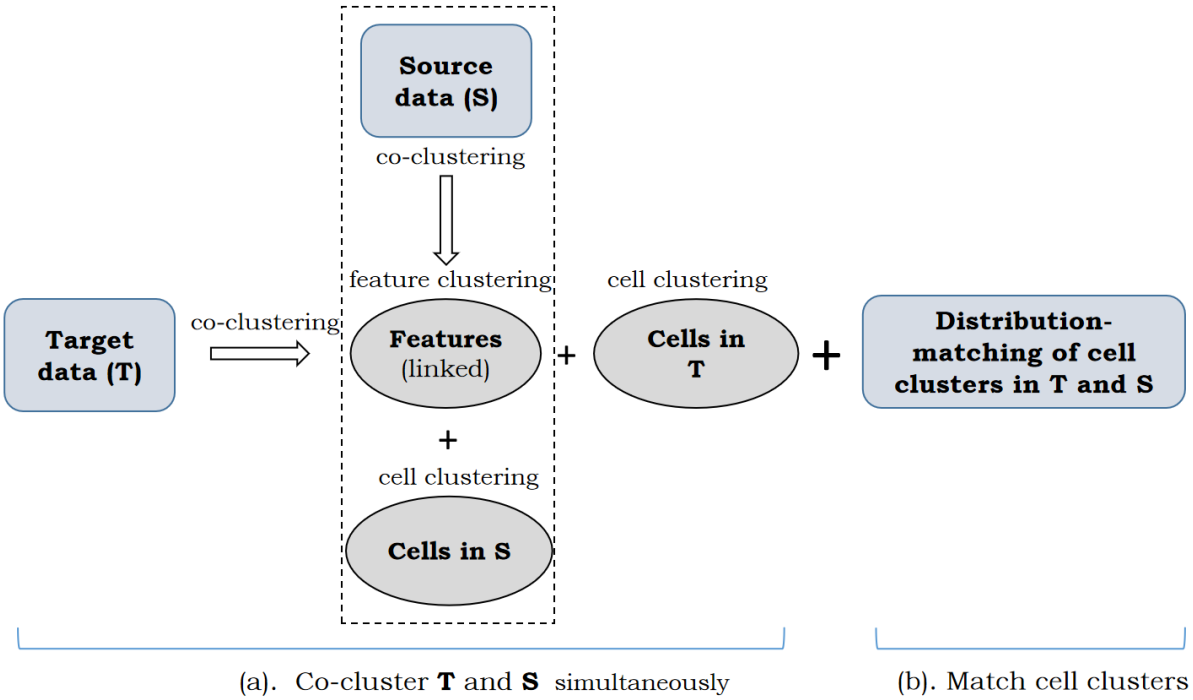


Figure 1: The schematic of the coupled co-clustering-based (*couple*CoC) unsupervised transfer learning model

learnt in another data set, called the source data(Pan and Yang, 2009). Wang et al. (2008) proposed a transferred discriminative analysis (TDA) algorithm to solve the transfer dimensionality reduction problem. As a clustering version of the self-taught learning (Raina et al., 2007), Dai et al. (2008) proposed a self-taught learning algorithm (STC) to learn a common feature space across data sets, which improves clustering for the target data. Details on transfer learning are provided in the survey by Pan and Yang (2009).

Here we propose a novel coupled co-clustering-based (*couple*CoC) unsupervised transfer learning model for the integrative analysis of single-cell genomic data sets, as shown in Fig.1. Our goal is to use one data set, the source data, to improve clustering of another data set, the target data. We assume that features in the two data sets are linked: scRNA-seq and scATAC-seq data are connected by gene expression and promoter accessibility; scRNA-seq and single-cell methylation data are connected by gene expression and gene body hypomethylation; human and mouse scRNA-seq data are connected by homologs. We conduct co-clustering for the source data S and the target data T simultaneously (Fig.1(a)), where the linked features enable the transfer of knowledge between the source data and the target data. We also introduce a distribution-matching term of cell clusters that provides another level of knowledge sharing, and encourages the obtained clusters to be similar in source data and target data (Fig.1(b)). Last, the degree of knowledge transfer is learnt adaptively from the data.

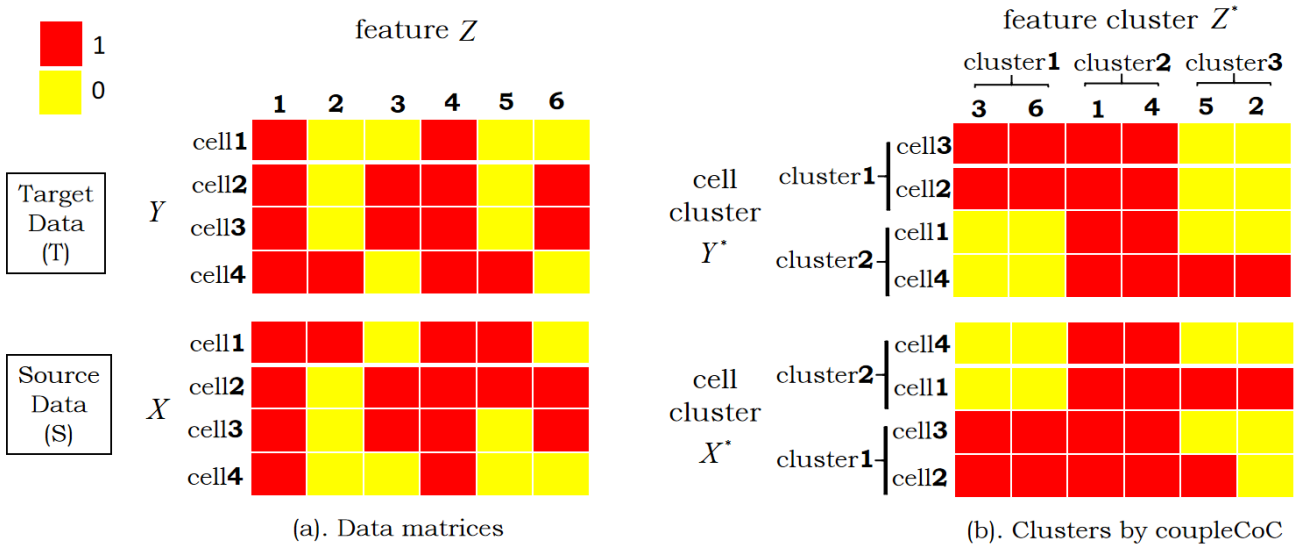


Figure 2: Toy example of our method (coupleCoC) for the analysis of single-cell data. (a). Red color means active, and yellow color means inactive. (b). Data after clustering: cells and features are reordered such that similar cells and similar features are grouped together. Note that the clusters of cells are matched in the target data and source data by our method.

The rest of the paper is organized as the following. We present the method *couple*CoC in Section 2. Simulation studies are presented in Section 3. Three real data examples are presented in Section 4. The conclusion is presented in Section 5.

2 The methodology

2.1 Problem Formulation

We binarize the data matrices of scRNA-seq, scATAC-seq and single-cell methylation (details presented in Section 2.3). Depending on the implementation and the strength of the signal in separating the cells, we choose one data set as the source data and the other data set as the target data. More details on choosing the source data and target data are presented later in the real data session. Our goal is to use source data to improve the clustering of target data. We denote the target data as a matrix $\mathrm{T} = (\mathrm{T}_{yz})_{n_\mathrm{T} \times q}$, where $\mathrm{T}_{yz} = 1$ if the z-th feature is active in the y-th cell, and $\mathrm{T}_{yz} = 0$ otherwise. In scRNA-Seq data, an active feature means the gene is expressed; in scATAC-Seq data, an active feature means the promoter is accessible as accessible promoter is associated with active gene expression; in single-cell methylation data, an active feature means the gene body is hypomethylated as hypomethylated gene body is associated with active gene expression, Similarly, we denote the source data as a binary matrix $\mathrm{S} = (\mathrm{S}_{xz})_{n_\mathrm{S} \times q}$, and the active

feature is defined the same as that in target data. Here we assume that the features in target data and source data are linked. For simplicity, we use a toy example to illustrate the coupled data set (Fig.2(a)). We regard two different 4×6 matrices as the target data (T) and the source data (S), respectively.

Let Y be a discrete random variable, which takes value from the set of cell indexes $\{1, \ldots, n_{\mathrm{T}}\}$ in target data T. Let X be a discrete random variable, which takes value from the set of cell indexes $\{1, \ldots, n_{\mathrm{S}}\}$ in source data S. Further, let Z be a discrete random variable, which takes value from the set of feature indexes $\{1, \ldots, q\}$. Take the target data with 4 cells and 6 features as an example (see Fig.2(a)): $Y = y, y = 1, \ldots, 4$, means the selected cell is the y-th cell among 4 cells, and $Z = z, z = 1, \ldots, 6$ means the selected feature is the z-th feature among 6 features.

Let $p_{\mathrm{T}}(Y, Z)$ be the joint probability distribution for Y and Z, which can be represented by a $n_{\mathrm{T}} \times q$ matrix. $p_{\mathrm{T}}(Y = y, Z = z)$ is the probability of the z-th feature being active in the y-th cell, and it is estimated from the observed target data T:

$$p_{\mathrm{T}}(Y = y, Z = z) = \frac{\mathrm{T}_{yz}}{\sum_{y=1}^{n_{\mathrm{T}}} \sum_{z=1}^{q} \mathrm{T}_{yz}}, \tag{1}$$

where $y \in \{1, \ldots, n_{\mathrm{T}}\}, z \in \{1, \ldots, q\}$. The marginal probability distributions are then expressed as $p_{\mathrm{T}}(Y = y) = \frac{\sum_{z=1}^{q} \mathrm{T}_{yz}}{\sum_{y=1}^{n_{\mathrm{T}}} \sum_{z=1}^{q} \mathrm{T}_{yz}}$, and $p_{\mathrm{T}}(Z = z) = \frac{\sum_{y=1}^{n_{\mathrm{T}}} \mathrm{T}_{yz}}{\sum_{y=1}^{n_{\mathrm{T}}} \sum_{z=1}^{q} \mathrm{T}_{yz}}$. Take the target data in Fig.2(a) as an example. We have

$$p_{\mathrm{T}}(Y, Z) = \frac{1}{14} \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}, \tag{2}$$

$p_{\mathrm{T}}(Y) = \frac{1}{14}[2, 4, 4, 4]'$ and $p_{\mathrm{T}}(Z) = \frac{1}{14}[4, 1, 2, 4, 1, 2]'$ ("A'" means transpose of A).

For the source data, $p_{\mathrm{S}}(X, Z)$ is the joint probability distribution for X and Z; $p_{\mathrm{S}}(X)$ and $p_{\mathrm{S}}(Z)$ are the marginal probabilities. The distributions $p_{\mathrm{S}}(X, Z)$, $p_{\mathrm{S}}(X)$ and $p_{\mathrm{S}}(Z)$ are calculated similarly to that in the target data.

We aim at grouping similar cells into clusters and at grouping similar features into clusters as well. Suppose we want to cluster the cells into N clusters and the features into K clusters. Let the random variables Y^* and X^* take values from the set of cell cluster indexes $\{1, \ldots, N\}$. We use $C_Y(\cdot)$ to represent the clustering function for target data and $C_Y(y) = i$ $(i = 1, \ldots, N)$ indicates that cell y belongs to cluster i. For source data, the clustering function $C_X(\cdot)$ is defined similarly as that for the target data. Let the random variable Z^* take values from the set of feature cluster indexes $\{1, \ldots, K\}$. We use $C_Z(\cdot)$ to represent the clustering function of the features and $C_Z(z) = j$ $(j = 1, \ldots, K)$ indicates that feature z belongs to cluster j. The tuples $(C_Y(y), C_Z(z))$ and $(C_X(x), C_Z(z))$ are referred as co-clustering (Dhillon et al., 2003).

Let $p_{\mathrm{T}}(Y^*, Z^*)$ be the joint probability distribution for Y^* and Z^*, which can be represented as $N \times K$ matrix. Based on the Equation (1), this distribution can be expressed as

$$p_{\mathrm{T}}(Y^* = i, Z^* = j) = \sum_{y \in \{C_Y(y)=i\}} \sum_{z \in \{C_Z(z)=j\}} p_{\mathrm{T}}(Y = y, Z = z), \tag{3}$$

Taking the clusters of target data in Fig.2(b) as an example, we have

$$p_{\mathrm{T}}(Y^*, Z^*) = \frac{1}{14} \begin{bmatrix} 4 & 4 & 0 \\ 0 & 4 & 2 \end{bmatrix}, \tag{4}$$

and the marginal probability distributions $p_{\mathrm{T}}(Y^*) = \frac{1}{14}[8, 6]'$ and $p_{\mathrm{T}}(Z^*) = \frac{1}{14}[4, 8, 2]'$. For the source data, $p_{\mathrm{S}}(X^*, Z^*)$, $p_{\mathrm{S}}(X^*)$ and $p_{\mathrm{S}}(Z^*)$ are defined and calculated analogously to those for the target data.

The goal of coupled co-clustering-based unsupervised transfer learning in this work is to find the optimal cell clustering function C_Y by utilizing the information of both the cell clustering function C_X and the feature clustering function C_Z.

2.2 The CoupleCoC Algorithm

According to the information theoretic co-clustering framework (Dhillon et al., 2003), the objective function of co-clustering between instances and features is defined as minimizing the loss in mutual information before and after co-clustering. For target data T and source data S, their objective functions of co-clustering can be expressed as

$$L_{\mathrm{T}} = M(Y, Z) - M(Y^*, Z^*),$$

and

$$L_{\mathrm{S}} = M(X, Z) - M(X^*, Z^*)$$

respectively, where $M(\cdot, \cdot)$ denotes mutual information between two random variables: $M(A, B) = \sum_a \sum_b p(a, b) \log \frac{p(a,b)}{p(a)p(b)}$. Practically, we only need to consider the elements satisfying $p(a, b) \neq 0$.

We propose the following objective function for coupled co-clustering of the target data T and source data S:

$$L = L_{\mathrm{T}} + \lambda_1 L_{\mathrm{S}} + \lambda_2 D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*) || p_{\mathrm{S}}(X^*)), \tag{5}$$

where $D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*) || p_{\mathrm{S}}(X^*))$ is the Jensen-Shannon Divergence, defined as

$$D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*) || p_{\mathrm{S}}(X^*)) = \frac{1}{2} D_{\mathrm{KL}}(p_{\mathrm{T}}(Y^*) || p(O)) + \frac{1}{2} D_{\mathrm{KL}}(p_{\mathrm{S}}(X^*) || p(O)),$$

where $p(O) = \frac{1}{2}(p_{\mathrm{T}}(Y^*) + p_{\mathrm{S}}(X^*))$, and $D_{\mathrm{KL}}(\cdot || \cdot)$ denotes the Kullback-Leibler divergence (KL divergence) between two probability distributions (Cover and Thomas, 1991), where $D_{\mathrm{KL}}(g || h) =$

$\sum_x g(x)\log\frac{g(x)}{h(x)}$. The term L_{T} is specific to the target data T and the term L_{S} is specific to the source data S. The two terms L_{T} and L_{S} share the same feature clusters C_Z, and Z^* can be viewed as a bridge to transfer knowledge between T and S. λ_1 is a tuning parameter that controls the contribution of the source data S to the target data T. Another component in Equation (5) that bridges T and S is the distribution-matching term $D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))$. The term $D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))$ encourages Y^* and X^*, the clusters in target and source data, to be similar. On top of improving the clustering accuracy, this term can offer better matching of the clusters in target and source data. Without the term $D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))$, the objective function $L = L_{\mathrm{T}} + \lambda_1 L_{\mathrm{S}}$ is invariant to permutations of the cluster labels and this will lead to mismatch of the clusters in target and source data. λ_2 is a tuning parameter on how well the clusters are matched in target and source data. We call this model *couple*CoC, which is the abbreviation of coupled co-clustering.

We need to solve the optimization problem:

$$\tilde{C}_Y = \underset{C_Y}{\operatorname{argmin}}\, L.$$

This optimization problem is non-convex and challenging to optimize. We can rewrite the terms L_{T} and L_{S} in the objective function L into the forms of KL divergence and the reformulated objective function is easier to optimize. To be more specific, for the target data T,

$$L_{\mathrm{T}} = D_{\mathrm{KL}}(p_{\mathrm{T}}(Y,Z)||p^*_{\mathrm{T}}(Y,Z)), \tag{6}$$

where $p^*_{\mathrm{T}}(Y,Z)$ is defined as

$$\begin{aligned} p^*_{\mathrm{T}}(Y=y,Z=z) =& p_{\mathrm{T}}(Y^*=C_Y(y),Z^*=C_Z(z)) \\ &\times \frac{p_{\mathrm{T}}(Y=y)}{p_{\mathrm{T}}(Y^*=C_Y(y))}\frac{p_{\mathrm{T}}(Z=z)}{p_{\mathrm{T}}(Z^*=C_Z(z))}, \end{aligned} \tag{7}$$

The probabilities $p_{\mathrm{T}}(Y^*=C_Y(y),Z^*=C_Z(z))$, $p_{\mathrm{T}}(Y=y)$, $p_{\mathrm{T}}(Y^*=C_Y(y))$, $p_{\mathrm{T}}(Z=z)$, $p_{\mathrm{T}}(Z^*=C_Z(z))$ are estimated as in the previous sections. Further, we have

$$\begin{aligned} D_{\mathrm{KL}}(p_{\mathrm{T}}(Y,Z)||p^*_{\mathrm{T}}(Y,Z)) &= \sum_{i=1}^{N}\sum_{y\in\{y:C_Y(y)=i\}} p_{\mathrm{T}}(Y=y)D_{\mathrm{KL}}(p_{\mathrm{T}}(Z|Y=y)||p^*_{\mathrm{T}}(Z|Y^*=i,Y=y)) \\ &= \sum_{j=1}^{K}\sum_{z\in\{z:C_Z(z)=j\}} p_{\mathrm{T}}(Z=z)D_{\mathrm{KL}}(p_{\mathrm{T}}(Y|Z=z)||p^*_{\mathrm{T}}(Y|Z^*=j,Z=z)), \end{aligned} \tag{8}$$

where $p^*_{\mathrm{T}}(Z=z|Y^*=i,Y=y) \triangleq \frac{p^*_{\mathrm{T}}(Y=y,Z=z)}{p_{\mathrm{T}}(Y=y)}$, $z=1,\ldots,q$, for any $y\in\{y:C_Y(y)=i\}$, and $p^*_{\mathrm{T}}(Y=y|Z^*=j,Z=z) \triangleq \frac{p^*_{\mathrm{T}}(Y=y,Z=z)}{p_{\mathrm{T}}(Z=z)}$, $y=1,\ldots,n_{\mathrm{T}}$, for any $z\in\{z:C_Z(z)=j\}$. The relations in Equations (6) and (8) have been proven by Dhillon et al. (2003) and Dai et al. (2008). For

the source data S, we can rewrite L_{S} similarly as rewriting L_{T} in Equations (6), (7) and (8). Therefore, the objective function in Equation (5) can be rewritten as:

$$\begin{aligned} L =& D_{\mathrm{KL}}(p_{\mathrm{T}}(Y,Z)||p^*_{\mathrm{T}}(Y,Z)) + \lambda_1 D_{\mathrm{KL}}(p_{\mathrm{S}}(X,Z)||p^*_{\mathrm{S}}(X,Z)) \\ &+ \lambda_2 D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*)), \end{aligned} \tag{9}$$

To minimize L, we can iteratively update clustering functions C_Y, C_X and C_Z as follows:

- Given C_X and C_Z, update C_Y. Minimizing L is equivalent to minimizing

$$\sum_{i=1}^{N} \sum_{y \in \{y: C_Y(y)=i\}} p_{\mathrm{T}}(Y=y) Q_{\lambda_2}(Y^*=i, Y=y|C_X, C_Z),$$

  where

$$\begin{aligned} & Q_{\lambda_2}(Y^*=i, Y=y|C_X, C_Z) \\ \triangleq & D_{\mathrm{KL}}(p_{\mathrm{T}}(Z|Y=y)||p^*_{\mathrm{T}}(Z|Y^*=i)) + \frac{\lambda_2 D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))}{n_{\mathrm{T}} N p_{\mathrm{T}}(Y=y)}. \end{aligned}$$

  We iteratively update the cluster assignment $C_Y(y)$ for each cell y ($y = 1, \ldots, n_{\mathrm{T}}$) in the target data, fixing the cluster assignment for the other cells:

$$C_Y(y) = \underset{i \in \{1,\ldots,N\}}{\operatorname{argmin}} Q_{\lambda_2}(Y^*=i, Y=y|C_X, C_Z). \tag{10}$$

- Given C_Y and C_Z, update C_X. Minimizing L is equivalent to minimizing

$$\sum_{j=1}^{N} \sum_{x \in \{x: C_X(x)=j\}} p_{\mathrm{S}}(X=x) Q_{\lambda_2}(X^*=j, X=x|C_Y, C_Z),$$

  where

$$\begin{aligned} & Q_{\lambda_2}(X^*=j, X=x|C_Y, C_Z) \\ \triangleq & D_{\mathrm{KL}}(p_{\mathrm{S}}(Z|X=x)||p^*_{\mathrm{S}}(Z|X^*=j)) + \frac{\lambda_2 D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))}{n_{\mathrm{S}} N p_{\mathrm{S}}(X=x)}. \end{aligned}$$

  We iteratively update the cluster assignment $C_X(x)$ for each cell x ($x = 1, \ldots, n_{\mathrm{S}}$) in the source data, fixing the cluster assignment for the other cells:

$$C_X(x) = \underset{j \in \{1,\ldots,N\}}{\operatorname{argmin}} Q_{\lambda_2}(X^*=j, X=x|C_Y, C_Z). \tag{11}$$

- Given C_Y and C_X, update C_Z. Minimizing L is equivalent to minimizing

$$\sum_{s=1}^{K} \sum_{z \in \{z: Z^*(z)=s\}} R_{\lambda_1}(Z^*=s, Z=z|C_Y, C_X),$$

where

$$
\begin{aligned}
&R_{\lambda_1}(Z^* = s, Z = z | C_Y, C_X) \\
\triangleq &\lambda_1 p_{\mathrm{T}}(Z = z) D_{\mathrm{KL}}(p_{\mathrm{T}}(Y|Z = z)||p^*_{\mathrm{T}}(Y|Z^* = s)) \\
&+ p_{\mathrm{S}}(Z = z) D_{\mathrm{KL}}(p_{\mathrm{S}}(X|Z = z)||p^*_{\mathrm{S}}(X|Z^* = s)).
\end{aligned}
$$

We iteratively update the cluster assignment $C_Z(z)$ for each feature z ($z = 1, \ldots, q$), fixing the cluster assignment for the other features:

$$
C_Z(z) = \underset{s \in \{1,\ldots,K\}}{\operatorname{argmin}} R_{\lambda_1}(Z^* = s, Z = z | C_Y, C_X). \tag{12}
$$

The procedure of optimization for the global objective function (5) are summarized as follows:

(i) Initialize. First, calculate $p_{\mathrm{T}}(Y, Z)$ and $p_{\mathrm{S}}(X, Z)$ using the target data T and source data S. Second, initialize the clustering functions $C_Y^{[0]}, C_X^{[0]}$ and $C_Z^{[0]}$. Third, initialize $p_{\mathrm{T}}^{*[0]}(Y, Z)$ and $p_{\mathrm{S}}^{*[0]}(X, Z)$. We set the number of iterations as I.

(ii) Update.

  (a) Fix $p_{\mathrm{T}}^{*[t-1]}(Y, Z)$ and $p_{\mathrm{S}}^{*[t-1]}(X, Z)$ and sequentially update $C_Y^{[t]}$, $C_X^{[t]}$ and $C_Z^{[t]}$ based on the Equations (10), (11) and (12).

  (b) Fix $C_Y^{[t]}$, $C_X^{[t]}$ and $C_Z^{[t]}$ and update $p_{\mathrm{T}}^{*[t]}(Y, Z)$ and $p_{\mathrm{S}}^{*[t]}(X, Z)$.

  (c) Iterate these updates until $t = I$.

(iii) Output $C_Y^{[I]}$.

The objective function (Equation(9)) is non-increasing in the updates of C_Y, C_X and C_Z, and the algorithm will converge to a local minimum. Finding the global optimal solution is NP-hard. Additionally, the *couple*CoC algorithm converges in a finite number of iterations due to the finite search space. In practice, this algorithm works well in both simulations and real data analysis.

2.3 Parameter Tuning and Variable Selection

We assume that N, the number of clusters for the cells, is given. Our algorithm *couple*CoC has three tuning parameters:λ_1, λ_2 and K, the number of clusters for the features. We use grid search to choose the combination of parameters, which minimizes the objective function (Equation (9)). In practice, we choose the search domains $\lambda_1 \in [0, 5]$, $\lambda_2 \in [0, 5]$ and $K \in (0, 10)$. Grid search performs well in both simulated data and real data.

We implement variable selection before performing clustering for single-cell genomic data sets. This preprocessing step speeds up computation and reduces the noise level. For the data sets

in Sections 4.1 and 4.2, we applied clustering to the individual data sets (Zamanighomi et al., 2018) and then selected cluster-specific features (Love et al., 2014; Zamanighomi et al., 2018). We selected 500 cluster-specific features in each of scRNA-seq and scATAC-seq data in 4.1, and 500 in each of scRNA-seq and methylation data in Section 4.2. For the data set in Section 4.3, we selected the 500 most variable homologs from each of the mouse and human scRNA-seq data. After variable selection, we binarize the data matrices by setting the non-zero entries to 1.

3 Simulation Studies

3.1 Simulation Setting

We follow the simulation setup given in Lin et al. (2019). We set the scRNA-seq data as the target data T: $\mathrm{T}_{yz} = 1$ if gene z is expressed in cell y, and $\mathrm{T}_{yz} = 0$ otherwise. We set the scATAC-seq data as the source data S: $\mathrm{S}_{xz} = 1$ if the promoter region for feature z is accessible in cell x, and $\mathrm{S}_{xz} = 0$ otherwise. The steps of generating T and S are given in A. We set the number of cells in both target data T and source data S as $n_{\mathrm{T}} = n_{\mathrm{S}} = 100$, and set the number of features as $q = 100$. We varied the differential degree (w) across clusters and the standard deviation (σ) of the generative distribution. Larger w leads to better separation of the cell clusters, and larger σ leads to higher noise level. We consider three different simulation settings, varying the parameters w and σ.

3.2 Simulation Results

Table 1 presents the simulation results. We use four criteria to quantify the clustering results: purity, Rand index (RI), adjusted Rand index (ARI) and normalized mutual information (NMI). The tuning parameters are set as $\lambda_1 = \lambda_2 = 1.5, K = 3$ by grid search. The trends for different clustering criteria are similar over the three settings.

Compared with the first setting ($w = 0.8, \sigma = 0.6$), the second setting has a larger variance ($\sigma = 1.2$) for the target data, and the third setting has lower differential ability across the cell clusters ($w = 0.7$). Setting 2 and 3 have lower overall clustering performance compared with that in setting 1. As expected, our model (*couple*CoC) achieves better clustering results, compared with k-means, spectral clustering and hierarchial clustering. This is likely because of the fact that our model transfers the knowledge from the source data, which helps to improve the clustering performance on the target data; while the other three methods do not utilize the information from the source data.

Table 1: The simulation results of clustering scRNA-seq data for 50 independent runs are summarized.

Settings	Methods	Purity	RI	ARI	NMI
	*couple*CoC	0.891	0.806	0.612	0.531
Setting 1	k-means	0.811	0.706	0.411	0.352
$w = 0.8, \sigma = 0.6$	spectral clustering	0.548	0.501	0.001	0.037
	hierarchical clustering	0.549	0.501	0.001	0.035
	*couple*CoC	0.838	0.729	0.459	0.381
Setting 2	k-means	0.701	0.594	0.188	0.162
$w = 0.8, \sigma = 1.2$	spectral clustering	0.555	0.502	-0.000	0.036
	hierarchical clustering	0.557	0.503	0.002	0.040
	*couple*CoC	0.693	0.585	0.169	0.137
Setting 3	k-means	0.620	0.533	0.067	0.061
$w = 0.7, \sigma = 0.6$	spectral clustering	0.548	0.501	0.000	0.036
	hierarchical clustering	0.550	0.501	0.001	0.037

4 Application to Real Data

4.1 Example 1: Human scRNA-seq and scATAC-seq Data

In this section, we evaluate our *couple*CoC method by clustering human scRNA-seq and scATAC-seq data. 223 K562 and 91 HL60 scATAC-seq cells are obtained from Buenrostro et al. (2015), and 42 K562 and 54 HL60 deeply sequenced scRNA-seq cells are obtained from Pollen et al. (2014). We compare our proposed *couple*CoC with another unsupervised transfer learning method STC (Dai et al., 2008), and other commonly used clustering methods proposed for single-cell genomic data, including SC3 (Kiselev et al., 2017), and SIMLR (Wang et al., 2017). The true cell labels are used as a benchmark for evaluating the performance of the clustering methods. The tuning parameters in *couple*CoC are set as $\lambda_1 = 1.5, \lambda_2 = 2, K = 3$ by grid search.

Example 1 in Table 2 shows that all four methods perfectly separated the cell types in the source data (scRNA-Seq data). SIMLR do not cluster the target data (scATAC-seq data) well, compared with *couple*CoC and STC. This may be due to the fact that *couple*CoC and STC utilize information from scRNA-seq data to improve the clustering of scATAC-seq data. SIMLR and SC3 are designed for clustering one data type, thus they are not able to match cell types across the source data and target data. STC has good clustering performance, but the cell types are mismatched across the data types, likely because STC does not consider the distribution-matching

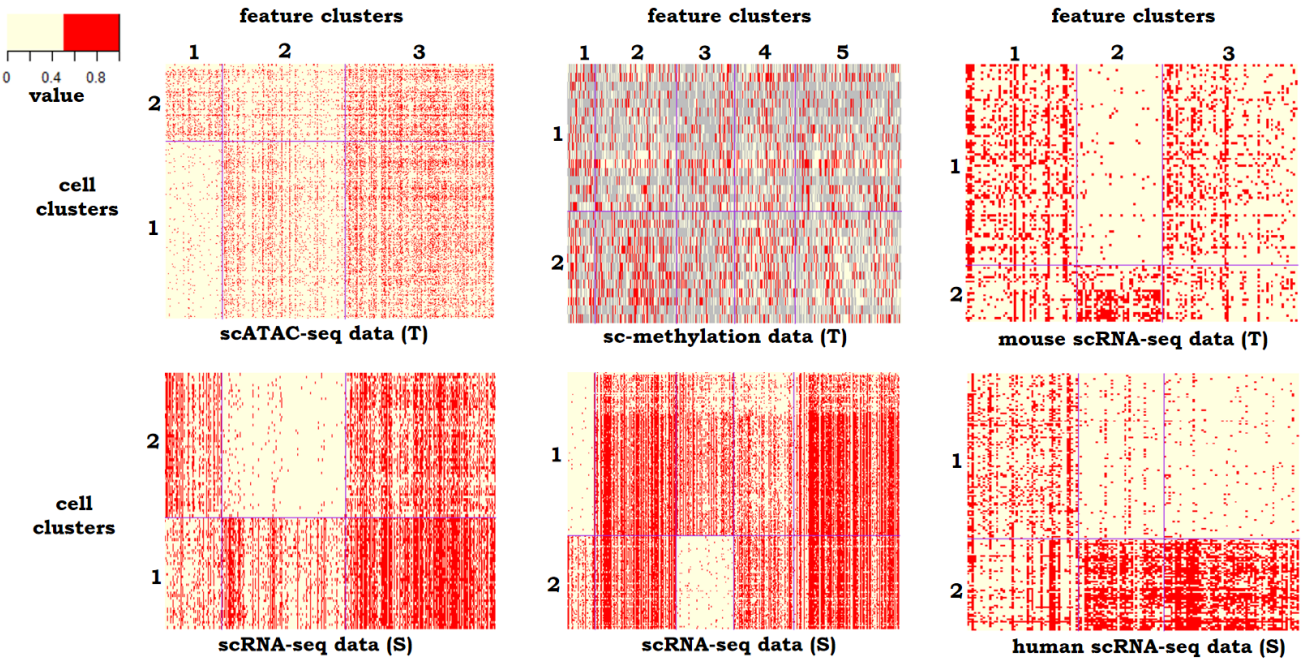
term $D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))$. Our proposed method, *couple*CoC, has good clustering performance and the cell types are correctly matched. Visualization of the co-clustering results by *couple*CoC is presented in Fig. 3(a), where we grouped and segmented cell clusters and feature clusters by blue lines. Our proposed *couple*CoC clearly clusters similar cells and features.

Table 2: Clustering tables for *couple*CoC, STC, SC3 and SIMLR on both target data T and source data S for three real data examples. Example 1: K562 and HL60 scATAC-seq and scRNA-Seq data. Example 2: GM12878 and K562 sc-methylation and scRNA-Seq data. Example 3: Clara cells and Ependymal cells, mouse and human scRNA-seq data.

		*couple*CoC		STC		SC3		SIMLR	
Example 1		K562	HL60	K562	HL60	K562	HL60	K562	HL60
scATAC-seq data	cluster1	6	217	212	11	3	220	209	14
$(\mathrm{T}, n_{\mathrm{T}} = 314, q = 500)$	cluster2	89	2	4	87	85	6	83	8
scRNA-seq data	cluster1	0	42	0	42	42	0	41	1
$(\mathrm{S}, n_{\mathrm{S}} = 96, q = 500)$	cluster2	54	0	54	0	0	54	0	54
		*couple*CoC		STC		SC3		SIMLR	
Example 2		GM12878	K562	GM12878	K562	GM12878	K562	GM12878	K562
sc-methylation data	cluster1	16	3	11	8	0	19	3	16
$(\mathrm{T}, n_{\mathrm{T}} = 30, q = 500)$	cluster2	1	10	5	6	11	0	9	2
scRNA-seq data	cluster1	128	0	0	128	128	0	110	18
$(\mathrm{S}, n_{\mathrm{S}} = 201, q = 500)$	cluster2	0	73	73	0	0	73	66	7
		*couple*CoC		STC		SC3		SIMLR	
Example 3		Ep	Cl	Ep	Cl	Ep	Cl	Ep	Cl
Mouse scRNA-seq data	cluster1	7	92	22	77	9	90	74	25
$(\mathrm{T}, n_{\mathrm{T}} = 113, q = 500)$	cluster2	14	0	14	0	0	14	12	2
Human scRNA-seq data	cluster1	3	110	112	1	112	1	71	42
$(\mathrm{S}, n_{\mathrm{S}} = 171, q = 500)$	cluster2	58	0	0	58	3	55	2	56

4.2 Example 2: Human scRNA-seq and sc-methylation Data

Next, we evaluate our method *couple*CoC by jointly clustering human scRNA-seq and single-cell methylation data. For the scRNA-seq data, we obtained 128 GM12878 cells and 73 K562 cells under the accession number GSE81861 (Li et al., 2017). For the single-cell methylation data, we obtained 19 GM12878 cells and 11 K562 cells under the accession number GSE83882 (Pott, 2017). We impute each missing value in the sc-methylation data matrix with the same Bernoulli distribution with outcome: mean of non-missing values (with probability p) or 0 (with probability = 1 - p), where p is estimated by the frequency of non-zeros in the sc-methylation data matrix. Because methylated gene body implicates repressed gene expression, we swapped values 0 and 1 for the sc-methylation data before we implemented *couple*CoC and STC. The true cell labels are used as a benchmark for evaluating the performance of the clustering methods. The tuning



(a). Example 1 (b). Example 2 (c). Example 3

Figure 3: Heatmaps of clustering results on both target T and source data S (all of data matrices are binarized) for three real data examples. From the color key, the input value 0 and 1 are corresponding to the color orange and red, respectively. The color grey in the heatmap of sc-methylation data corresponds to the missing values. The cell clusters and feature clusters are segmented by blue lines. (a) K562 and HL60 scATAC-seq and scRNA-Seq data. (b) GM12878 and K562 sc-methylation and scRNA-Seq data. (c) Clara cells and Ependymal cells, mouse and human scRNA-seq data.

parameters in the *couple*CoC are set as $\lambda_1 = 1, \lambda_2 = 1.2, K = 5$ by grid search.

Results for example 2 is presented in Table 2, where we compared the performance of coupleCoC, STC, SC3 and SIMLR. First, all four methods performed quite well on scRNA-seq data except for SIMLR. Second, for the sc-methylation data, SC3 performs the best, but it cannot match the cell types across the two data types. Third, because of the distribution-matching term $D_{\mathrm{JS}}(p_{\mathrm{T}}(Y^*)||p_{\mathrm{S}}(X^*))$, our proposed method *couple*CoC has better performance than STC: cells in sc-methylation data are better clustered and the cell types are correctly matched in sc-methylation and scRNA-seq data. Fig. 3(b) is the corresponding heatmap after clustering by *couple*CoC. For the scRNA-seq data, our proposed *couple*CoC clearly clusters similar cells and features. However, no clear pattern is observed in sc-methylation data. This is likely because sc-methylation data is very noisy and more than half of the entries are missing in sc-methylation data.

4.3 Example 3: Human and Mouse scRNA-seq Data

The last coupled real data set includes human and mouse scRNA-seq data, and the data sets are downloaded from *panglaodb.se* (Fran et al., 2019). We obtained 99 Clara cells and 14 Ependymal cells under the accession number SRS4237518 for the mouse data, and we obtained 113 Clara cells and 58 Ependymal cells under the accession number SRS4660846 for the human data. We use the cell-type annotation (Angelidis et al., 2019) as a benchmark for evaluating the performance of the clustering methods. The tuning parameters in the *couple*CoC are set as $\lambda_1 = 3, \lambda_2 = 0.2, K = 3$ by grid search.

The results for example 3 are presented in Table 2. SIMLR does not perform well on both the target data (mouse scRNA-seq data) and the source data (human scRNA-seq data). SC3 method performs well on source data but does not perform well on target data. Both STC and *couple*CoC have good performance on the source data, while STC does not work well on the target data. Moreover, the cell types are correctly matched in *couple*CoC. Fig. 3(c) shows the heatmap after clustering by *couple*CoC. Our proposed *couple*CoC clearly clusters similar cells and features.

5 Conclusion

Unsupervised methods including dimension reduction and clustering are essential to the analysis of single-cell genomic data as the cell types need to be inferred from the data. In addition, the matching of cell types in multimodal data cannot be disregarded, because it can provide rich functional information. In this work, we have developed an unsupervised transfer learning approach for the integrative analysis of multimodal single-cell genomic data, and we have demonstrated its performance on simulated data and real data analysis. Through the linked features across multimodal single-cell data, our proposed method can effectively transfer knowledge across multiple data sets. The degree of knowledge transfer learning is learnt adaptively from the data sets. Our proposed method is applicable in other areas including text mining.

Funding

This work has been supported by the Chinese University of Hong Kong direct grant 2018/2019 No. 4053360, the Chinese University of Hong Kong startup grant No. 4930181, and Hong Kong Research Grant Council Grant ECS No. CUHK 24301419.

References

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., and Mayr, C. H. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun*, 10(963).

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, (523):486–490.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satijia, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*, 36(5):411–420.

Cover, T. M. and Thomas, J. A. (1991). Elements of information theory. *Wiley-Interscience*.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348:910–914.

Dai, W. Y., Yang, Q., Xue, G. R., and Yu, Y. (2008). Self-taught clustering. *Proceedings of the 25th international Conference on Machine Learning*.

David, L., Johannes, K., Ewa, S., and the others (2020). Eleven grand challenges in single-cell data science. *Genome Biol*, 21(31).

Dhillon, I. S., Mallela, S., and Modha, D. S. (2003). Information-theoretic co-clustering. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98.

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A., Chang, H., Wang, Y., and Wong, W. H. (2018). Integrative analysis of single cell genomics data by coupled non-negative matrix factorizations. *Proc. Natl. Acad. Sci.*, (115):7723–7728.

Fran, O., Gan, G. M., and Johan, L. M. B. (2019). Panglaodb:a web serer for exploration of mouse and human single-cell rna sequencing data. *Database*.

Gonzalez-Blas, C. B. et al. (2019). cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nat. Methods*, 16:397–400.

Grün, D., Muraro, M. J., Boisset, S. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., Vandenborn, M., Vanes, J., et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19:266–277.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578.

Houseman, E. A., Christensen, B. C., Yeh, R. F., et al. (2008). Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 9(365).

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343:776–779.

Ji, Y., Wu, C., Liu, P., Wang, J., and Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21:2118–2122.

Kapourani, C. A. and Sanguinetti, G. (2019). Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biol*, 20(69).

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11:740–742.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., et al. (2017). Sc3: Consensus clustering of single-cell rna-seq data. *Nat. Methods*, 14(483).

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell rna-seq data across data sets. *Nat Methods*, 15:359–362.

Kuan, P. F., Wang, S., Zhou, X., and Chu, H. (2010). A statistical framework for lllumina dna methylation arrays. *Bioinformatics*, 26(22):2849–2855.

Lafon, S. and Lee, A. B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1393–1403.

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*, 49(5):708–718.

Lin, Z. X., Zamanighomi, M., Daley, T., Ma, S., and Wong, W. H. (2019). Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. *Stat. Sci.*

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(550).

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biol*, 15(75).

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning.

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., and Leyrat, A. A. (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol*, (32):1053–1058.

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *elife*, 6.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. (2007). Self-taught learning: transfer learning from unlabeled data. *Proceedings of the Twenty-fourth International Conference on Machine Learning*, pages 759–766.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol*, 33:1165–1172.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: From vision to reality. *Nat. News*, 550(451).

Schep, N. A., Wu, B., Buenrostro, J. D., and Greenleaf, W. J. (2017). chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, 14:975–978.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., et al. (2007). Genome-wide profiling of dna methylation reveals a class of normally methylated cpg island promoeters. *PLOS Genet*, 3.

Siegmund, K. D., Laird, P. W., and Laird-Offringa, I. A. (2004). A comparison of cluster analysis methods using dna methylation data. *Bioinformatics*, 20:1896–1904.

Stuart, T., Butler, A., Hoffman, P., and the others (2019). Comprehensive integration of single-cell data. *Cell*, (177):1888–1902.

Sun, Z., Wang, T., Deng, K., Wang, X. F., Lafyatis, R., Ding, Y., Hu, M., and Chen, W. (2017). Dimm-sc: A dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, (34):139–146.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P. V., Linnarsson, S., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nat. Neurosci.*, 18:145–153.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell rna sequencing data: challenges and opportunities. *Nat. Methods*, 14:565–571.

Vandermaaten, L. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9:2579–2605.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nat. Methods*, 14:414–416.

Wang, Z., Song, Y., and Zhang, C. (2008). Transferred dimensionality reduction. *Machine Learning and Knowledge Discovery in Databases, Eurpean Conference, ECML/PKDD.*

Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., and Zhang, Q. C. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nat. Commun*, 10(4576).

Yang, Y., Huh, R., Culpepper, H. W., Lin, Y., Love, M. I., and Li, Y. (2018). Safe-clustering: Single-cell aggregated(from ensemble)clustering for single-cell rna-seq data. *Bioinformatics.*

Zamanighomi, M., Lin, Z., Daley, T., Chen, X., Duren, Z., Schep, A., Greenleaf, W. J., and Wong, W. H. (2018). Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun*, 9(2410).

Zhang, H., Lee, C. A. A., Li, Z., and the others (2018). A multitask clustering approach for single-cell rna-seq analysis in recessive dystrophic epidermolysis bullosa. *PLoS Comput Biol*, 14(4).

Zhu, L., Lei, J., Klei, L., Devlin, B., and Roeder, K. (2019). Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. USA*, 116:466–471.

Zhu, L., Lei, J., and Roeder, K. A. (2018). A unified statistical framework for single cell and bulk rna sequencing data. *Ann. Appl. Stat.*, 12(1):609–632.

References

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., and Mayr, C. H. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun*, 10(963).

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, (523):486–490.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satijia, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*, 36(5):411–420.

Cover, T. M. and Thomas, J. A. (1991). Elements of information theory. *Wiley-Interscience*.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348:910–914.

Dai, W. Y., Yang, Q., Xue, G. R., and Yu, Y. (2008). Self-taught clustering. *Proceedings of the 25th international Conference on Machine Learning*.

David, L., Johannes, K., Ewa, S., and the others (2020). Eleven grand challenges in single-cell data science. *Genome Biol*, 21(31).

Dhillon, I. S., Mallela, S., and Modha, D. S. (2003). Information-theoretic co-clustering. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98.

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A., Chang, H., Wang, Y., and Wong, W. H. (2018). Integrative analysis of single cell genomics data by coupled non-negative matrix factorizations. *Proc. Natl. Acad. Sci.*, (115):7723–7728.

Fran, O., Gan, G. M., and Johan, L. M. B. (2019). Panglaodb:a web serer for exploration of mouse and human single-cell rna sequencing data. *Database*.

Gonzalez-Blas, C. B. et al. (2019). cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nat. Methods*, 16:397–400.

Grün, D., Muraro, M. J., Boisset, S. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., Vandenborn, M., Vanes, J., et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19:266–277.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578.

Houseman, E. A., Christensen, B. C., Yeh, R. F., et al. (2008). Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 9(365).

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343:776–779.

Ji, Y., Wu, C., Liu, P., Wang, J., and Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21:2118–2122.

Kapourani, C. A. and Sanguinetti, G. (2019). Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biol*, 20(69).

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11:740–742.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., et al. (2017). Sc3: Consensus clustering of single-cell rna-seq data. *Nat. Methods*, 14(483).

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell rna-seq data across data sets. *Nat Methods*, 15:359–362.

Kuan, P. F., Wang, S., Zhou, X., and Chu, H. (2010). A statistical framework for lllumina dna methylation arrays. *Bioinformatics*, 26(22):2849–2855.

Lafon, S. and Lee, A. B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1393–1403.

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*, 49(5):708–718.

Lin, Z. X., Zamanighomi, M., Daley, T., Ma, S., and Wong, W. H. (2019). Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. *Stat. Sci.*

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(550).

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biol*, 15(75).

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning.

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., and Leyrat, A. A. (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol*, (32):1053–1058.

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *elife*, 6.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. (2007). Self-taught learning: transfer learning from unlabeled data. *Proceedings of the Twenty-fourth International Conference on Machine Learning*, pages 759–766.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol*, 33:1165–1172.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: From vision to reality. *Nat. News*, 550(451).

Schep, N. A., Wu, B., Buenrostro, J. D., and Greenleaf, W. J. (2017). chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, 14:975–978.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., et al. (2007). Genome-wide profiling of dna methylation reveals a class of normally methylated cpg island promoeters. *PLOS Genet*, 3.

Siegmund, K. D., Laird, P. W., and Laird-Offringa, I. A. (2004). A comparison of cluster analysis methods using dna methylation data. *Bioinformatics*, 20:1896–1904.

Stuart, T., Butler, A., Hoffman, P., and the others (2019). Comprehensive integration of single-cell data. *Cell*, (177):1888–1902.

Sun, Z., Wang, T., Deng, K., Wang, X. F., Lafyatis, R., Ding, Y., Hu, M., and Chen, W. (2017). Dimm-sc: A dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, (34):139–146.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P. V., Linnarsson, S., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nat. Neurosci.*, 18:145–153.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell rna sequencing data: challenges and opportunities. *Nat. Methods*, 14:565–571.

Vandermaaten, L. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9:2579–2605.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nat. Methods*, 14:414–416.

Wang, Z., Song, Y., and Zhang, C. (2008). Transferred dimensionality reduction. *Machine Learning and Knowledge Discovery in Databases, Eurpean Conference, ECML/PKDD.*

Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., and Zhang, Q. C. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nat. Commun*, 10(4576).

Yang, Y., Huh, R., Culpepper, H. W., Lin, Y., Love, M. I., and Li, Y. (2018). Safe-clustering: Single-cell aggregated(from ensemble)clustering for single-cell rna-seq data. *Bioinformatics.*

Zamanighomi, M., Lin, Z., Daley, T., Chen, X., Duren, Z., Schep, A., Greenleaf, W. J., and Wong, W. H. (2018). Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun*, 9(2410).

Zhang, H., Lee, C. A. A., Li, Z., and the others (2018). A multitask clustering approach for single-cell rna-seq analysis in recessive dystrophic epidermolysis bullosa. *PLoS Comput Biol*, 14(4).

Zhu, L., Lei, J., Klei, L., Devlin, B., and Roeder, K. (2019). Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. USA*, 116:466–471.

Zhu, L., Lei, J., and Roeder, K. A. (2018). A unified statistical framework for single cell and bulk rna sequencing data. *Ann. Appl. Stat.*, 12(1):609–632.

Appendices

A Steps of data generation in simulation study

Similar to the simulation scheme designed by (Lin et al., 2019), we generate data in simulation study as follows:

1. Generate $\mathbf{w}^{acc}$ and $\mathbf{w}^{exp}$.

$$w_{rj}^{acc} = \begin{cases} w, & r = 1, j = 1, \ldots, k(1-w)/2; \\ & r = 2, j = k(1-w)/2 + 1, \ldots, k(1-w) \\ 1-w, & r = 2, j = 1, \ldots, k(1-w)/2; \\ & r = 1, j = k(1-w)/2 + 1, \ldots, k(1-w). \end{cases}$$

   $w_{1j}^{acc} = w_{2j}^{acc} \sim Beta(0.5, 2), j = k(1-w)+1, \ldots, k$. $w_{cj}^{exp} \sim Beta(w_{cj}^{acc}, 10), j = 1, \ldots, k(1-w)$;$w_{1j}^{exp} = w_{2j}^{exp} \sim Beta(w_{1j}^{acc}, 10), j = k(1-w)+1, \ldots, k$.

2. Generate z^{acc} and z^{exp}. The cluster labels are generated with equal probability 0.5.

3. Generate u^{acc} and $\tilde{u}^{acc}$. $\tilde{u}_{ij}^{acc} \sim Bernoulli(0.5)$ if $u_{ij}^{acc} = 1$, where $u_{ij}^{acc} \sim Bernoulli(w_{cj}^{acc})$ if $z_{ic}^{acc} = 1, i = 1, \ldots, m$;$\tilde{u}_{ij}^{acc} = 0$ otherwise.

4. Generate u^{exp} and $\tilde{v}^{exp}$. $\tilde{v}_{lj}^{exp} \sim Bernoulli(0.8)$ if $u_{lj}^{exp} = 1$, where $u_{lj}^{exp} \sim Bernoulli(w_{lj}^{exp})$ if $z_{lc}^{exp} = 1$;$\tilde{v}_{lj}^{exp} \sim Bernoulli(0.1)$ otherwise; $l = 1, \ldots, n$.

5. Generate C and G. $C_{ij} \sim N(0, 0.6^2)$ if $\tilde{u}_{ij}^{acc} = 0$ and $C_{ij} \sim N(2, 0.6^2)$ if $\tilde{u}_{ij}^{acc} = 1$; $G_{lj} \sim N(0, \sigma^2)$ if $\tilde{v}_{lj}^{exp} = 0$ and $G_{lj} \sim N(2, \sigma^2)$ if $\tilde{v}_{lj}^{exp} = 1$.

6. Generate T and S. $\mathrm{T}_{lj} = 1$ if $\mathrm{G}_{lj} > 0$ and $\mathrm{T}_{lj} = 0$ otherwise; $\mathrm{S}_{ij} = 1$ if $C_{ij} > 0$ and $\mathrm{S}_{ij} = 0$ otherwise.

Note that the notations $\mathbf{w}^{acc}$, $\mathbf{w}^{exp}$, z^{acc}, z^{exp}, u^{acc}, $\tilde{u}^{acc}$, u^{exp}, $\tilde{v}^{exp}$, C, G shown above have their special meanings in the model-based method by (Lin et al., 2019). Because they are beyond the topic of this work, we ignore the related details. In our simulation, the difference on the steps of data generation from their simulation scheme is the addition of Step 6. The target data T and the source data S are then obtained on the basis of chromatin accessibility data C and gene expression level data G afterwards.