Integration of Single-Cell Multi-Omics for Gene Regulatory Network Inference

Xinlin Hu[1], Yaohua Hu[1], Fanjie Wu[2], Ricky Wai Tak Leung[2], Jing Qin[2,*]

[1]Shenzhen Key Laboratory of Advanced Machine Learning and Applications, College of Mathematics and Statistics, Shenzhen University, Shenzhen, 518060, China.
[2]School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University, Shenzhen, 518107, China.

[*]To whom correspondence should be addressed. Email: qinj29@mail.sysu.edu.cn

Abstract

The advancement of single-cell sequencing technology in recent years has provided an opportunity to reconstruct gene regulatory networks (GRNs) with the data from thousands of single cells in one sample. This uncovers regulatory interactions in cells and speeds up the discoveries of regulatory mechanisms in diseases and biological processes. Therefore, more methods have been proposed to reconstruct GRNs using single-cell sequencing data. In this review, we introduce technologies for sequencing single-cell genome, transcriptome, and epigenome. At the same time, we present an overview of current GRN reconstruction strategies utilizing different single-cell sequencing data. Bioinformatics tools were grouped by their input data and mathematical principles for readers' convenience, and the fundamental mathematics inherent in each group will be discussed. Furthermore, the adaptabilities and limitations of these different methods will also be summarized and compared, with the hope to facilitate researchers recognizing the most suitable tools for them.

Keywords: Single-cell sequencing; Gene regulatory network inference; Single-cell multi-omics integration

Gene regulatory networks (GRNs), which describe the regulatory connections between transcription factors (TFs) and their target genes, help researchers to investigate the gene regulatory circuits and underlying mechanisms in various diseases and biological processes. A simple model of gene transcriptional regulation includes two key events: (1) an active TF binds to a *cis*-regulatory element such as a gene promoter; (2) such binding activates/suppresses the expression of the gene, which leads to the increase/decrease of the gene's RNA level. By integrating high-throughput omics data detecting the above two events in genome-wide scale, various powerful methods have been developed for reconstructing GRNs (Hawe et al., 2019; Karlebach and Shamir, 2008; Marbach et al., 2012; Qin et al., 2016). The recent development of technology makes it possible to sequence the single-cell genome, transcriptome, and epigenome. This provides rich datasets for GRN analyses. However, the inference of GRNs from single-cell sequencing data raises new challenges for method development. One of the main challenges is the underlying phenomenon of missing data. For single-cell transcriptome sequencing, the starting amount of RNAs extracted from single cells are often very low, genes with low or moderate expression are thus being omitted from the followed processing and sequencing steps due to inadequate sensitivities. Moreover, stochastic inherence and cell-to-cell variability of gene expression also result in aggravated noises (Gong et al., 2018; Kharchenko et al., 2014). For single-cell genome or epigenome sequencing, each DNA molecule in a diploid genome has only one or two opportunities to be sequenced. When only thousands of distinct reads can be detected per cell, it is impossible to cover all sites in the genome. Therefore, single-cell genome and epigenome sequencing suffer data omission even worse than that of transcriptome sequencing. Despite the challenges mentioned, dozens of methods have been developed to predict GRNs from single-cell sequencing data (Blencowe et al., 2019; Chen and Mar, 2018; Efremova and Teichmann, 2020; Fiers et al., 2018; Pratapa et al., 2020). However, selecting the proper tool according to one's needs is not an easy task for biological/biomedical researchers, as they are usually not very familiar with the mathematical reasoning behind these tools. Thus, understanding the basic principles of the algorithms implemented in these tools and their adaptabilities facilitates researchers making suitable choices according to their needs. In the following sections, we will be introducing, grouping, and discussing current GRN reconstruction strategies. This would also help tool developers to improve their tools by comparing the advantages and disadvantages of different methods. This review focuses on the representative and popular GRN inference approaches which utilize single-cell sequencing data especially on those with multi-omics data integration that can likely improve their performances (Table 1).

1. Single-cell sequencing for GRN reconstruction

Different from bulk sequencing that averages signals from a bulk of cells, single-cell sequencing isolates single cells from cell populations and labels DNA molecules derived from every single cell with unique barcodes before next-generation sequencing (Curtis et al., 2012). Single-cell RNA sequencing (scRNA-seq), the most popular single-cell sequencing technology, sequences RNA molecules in each cell and quantifies their expression levels. It can capture gene expression stochasticity and dynamics while revealing transcriptome-wide cell-to-cell variability at a high resolution (Pratapa et al., 2020). With thousands of genes in hundreds to thousands of single cells being measured by scRNA-seq, TF-gene interactions could be inferred based on the dependency of their expression. Thus, scRNA-seq data becomes one of the major data sources for GRN construction. Single-cell epigenome sequencing is another way to explore the regulatory relationship between TF and gene. Single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) (Buenrostro et al., 2015) detects the chromatin accessibility in single cells. scATAC-seq allows the identification of DNA regulatory elements within accessible genomic DNA regions in single cells. Similarly, single-cell chromatin immunocleavage sequencing (scChIC-seq) profiles histone modifications such as H3K4me3 in single cells, which detects active DNA regulatory regions during gene regulations, for example, regions associated with transcription activations (Ku et al., 2019). Meanwhile other single-cell sequencing techniques such as single-cell reduced representation bisulfite sequencing (scRRBS) (Guo et al., 2013), single-cell whole-genome bisulfite sequencing (scWGBS) (Farlik et al., 2015), genome-wide CpG island (CGI) methylation sequencing for single cells (scCGI-seq) (Han et al., 2017) and single-cell bisulfite sequencing (scBs-seq) (Clark et al., 2017) were developed for detecting DNA methylation profiles throughout single-cell genomes. With these single-cell epigenome data, GRN could be reconstructed by inferring TFs that bind to the genes with open or active DNA regulatory elements and epigenetic modifications, which indicates potential direct regulations between the TFs and the target genes. In addition, single-cell genome sequencing that detects genomic variations among single cells is a powerful tool to explore genetic heterogeneity and reconstruct cell lineage hierarchies of complex samples, such as tumor tissues. Mutations located at genomic DNA regulatory elements are also an important inducer of disease and affect the underlying gene regulatory network (Melton et al., 2015), thus the information of genomic variations in single cells is also valuable for GRN reconstruction. Another method screening genetic perturbation pool after clustered regularly interspaced short palindromic repeats (CRISPR)- mediated gene inactivation is called Perturb-seq, which is very useful for reverse genetics and thus GRN constructions when combined with scRNA-seq (Dixit et al., 2016). It can also be used to verify inferred GRNs by perturbing selected TFs in the network.

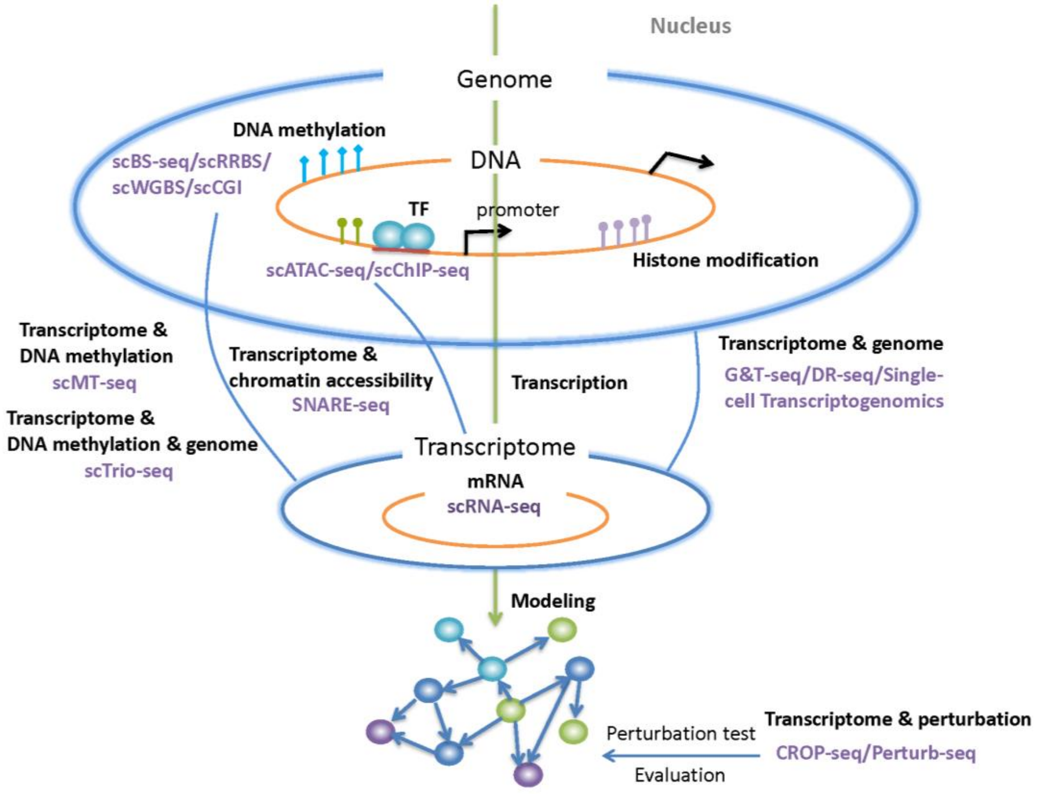

Figure 1 Single-cell sequencing technologies that investigate gene regulatory mechanisms.

Furthermore, there are techniques able to detect more than one type of single-cell omics profiles simultaneously. For example, single-cell genome and transcriptome sequencing (G&T-seq) (Macaulay et al., 2015), gDNA-mRNA sequencing (DR-Seq) (Dey et al.,

2015) and single-cell transcriptogenomics (SCTG) (Li et al., 2015) are techniques examining transcriptome and genome sequences in the same single-cell at the same time. Single-cell DNA methylome and transcriptome sequencing (scMT-seq) (Hu et al., 2016a) and (scM&T-seq) (Angermueller et al., 2016) are able to detect methylome and transcriptome in parallel to explore the cellular connections between epigenetic variation and transcriptional regulation. Single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq) (Chen et al., 2019) draws the combined map of chromatin accessibility and mRNA expression in the same cell. Some technologies can even measure three types of molecules in single cells. For instance, single-cell nucleosome, methylation and transcription sequencing (scNMT-seq) detects chromatin accessibility, DNA methylation and transcriptome profiling in parallel (Clark et al., 2018). Single-cell nucleosome occupancy and methylome sequencing (scNOMe-seq) measures chromatin accessibility and endogenous DNA methylation in single cells (Pott, 2017). Single-cell triple omics sequencing (scTrio-seq) (Hou et al., 2016) combines single-cell genome, methylome and transcriptome. These methods explore how the heterogeneity of genome and epigenome affects transcriptional heterogeneity in the same cells, thus probably enable GRN inference using computational methods originally designed for integrating bulk sequencing of multiple omics (Ritchie et al., 2015).

These single-cell omics and multi-omics technologies give us new opportunities to investigate complex gene regulatory mechanisms in a single-cell resolution (Figure 1). In short, sequencing data of single-cell genome, transcriptome and epigenome provides distinct information for GRN inference. In the following sections, we will discuss several popular strategies and algorithms that incorporate various single-cell sequencing data to construct GRNs (Figure 2).

2. Methods for scRNA-seq data alone

Tools designed for GRN reconstruction from scRNA-seq data alone have been reviewed and evaluated elsewhere (Blencowe et al., 2019; Chen and Mar, 2018; Pratapa et al., 2020). The performance of these tools was compared using simulated and real scRNA-seq data, and results in these studies revealed that there is no one method well accepted to be the best. This may be because different methods are suitable for different types and sources of data. However, in these reviews, the mathematical concepts and basic algorithms implicit in these tools were not discussed in depth. In this section, we introduce four major categories of popular algorithms for inferring GRNs from scRNA-seq data alone: (1) the ordinary differential equation (ODE)-based model, (2) the regression-based model, (3) the correlation/information-based mode and (4) the Boolean network. For each group, the mathematical principle of the algorithm and the representative tools are described to bridge the knowledge gap between method developers and biological/biomedical researchers.

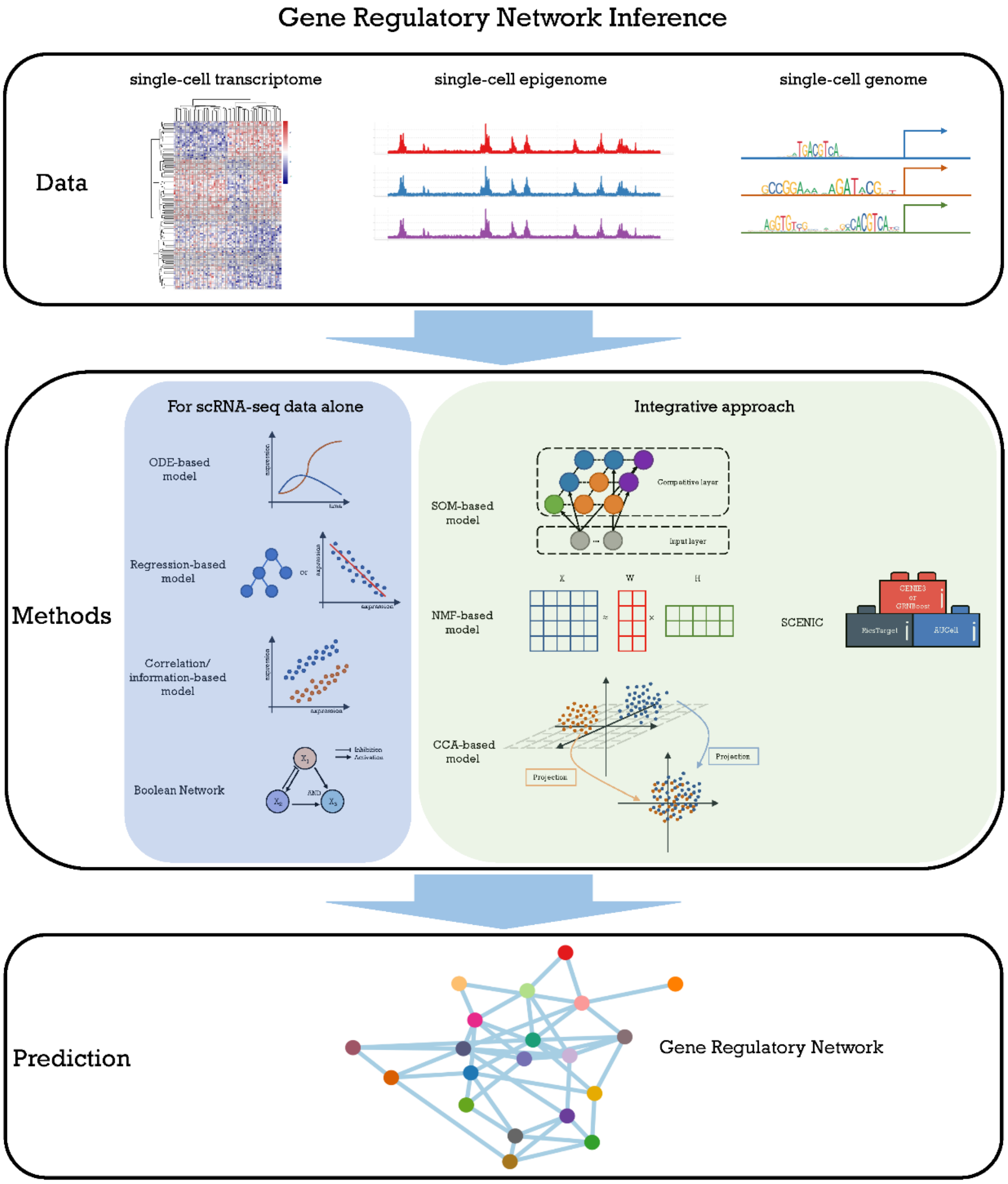

Figure 2 The summary of gene regulatory network inference from single cell sequencing data.

There are two types of scRNA-seq data - with and without temporal information. In a biological process, condition or experiment, cells can be collected from tissues or cell cultures. These cells could be in a process of change or in a steady state. For instance, cells might undergo differentiation, drug treatments, environmental changes, etc., and transit from one condition to another. In these processes, single-cell snapshot data can be obtained by collecting cells at a certain time point. Although each single cell represents a static state at this single time point, cells may have different stochastic behaviors during the same process (Elowitz et al., 2002), some sort of temporal information is still retained in this snapshot of cells. Such temporal information, called pseudo-time, can be inferred by the cell trajectory analysis (Saelens et al., 2019; Tian et al., 2019). Based on scRNA-seq data, cells could be ordered along the trajectory of the cell transition process (Griffiths et al., 2018), which represents the pseudo-time series. When cells are collected from tissues without any treatment or cells under pooled CRISPR screening, these cells are in a relatively static state or in a large number of independent processes. Cell populations in these samples do not show temporal relationships as those mentioned above. Therefore, when choosing the method/tool to reconstruct a GRN, we need to first determine whether there is temporal information in the single-cell sample, as some methods are designed specifically to work with temporal information, and others are more suitable for those without. While, there are also some methods can analyze both types of data.

2.1 ODE-based model

Provided with expression data with temporal information, ODE has been applied to describe expression dynamics and infer GRNs, which is generally formulated as

$$\frac{dy}{dt} = f(x), \tag{1}$$

where x and y represent the expression data of TFs and a target, respectively, and both are time series related to time t. The task is to find the function $f(x)$ and describe the expression change rate of target y, which also depicts how target y is regulated by TFs x.

Assumed that the expression change rate of target y linearly depends on the expression of TFs x, equation (1) is reduced to a simple linear ODE:

$$\frac{dy}{dt} = a_1x_1 + a_2x_2 + \cdots a_nx_n. \tag{2}$$

If parameters $a_1, a_2, \cdots, a_n$ and the initial values of t and y in equation (2) are provided, equation (2) can be solved by integration. However, the parameters are usually unknown in practice. Hence, the major task is to find the parameters $a_1^*, a_2^*, \cdots, a_n^*$ for equation (2) such that the error between estimation $y(a_1^*, a_2^*, \cdots, a_n^*)$ and observation $\hat{y}$ is minimal (Banks and Bihari, 2001). These parameters are also able to imply the regulatory relationships between the target and TF, whose observed expression data are $x_1, x_2, \cdots, x_n$. Several algorithms for solving this problem have been investigated by using least squares (Li et al., 2005; Xue et al., 2010), two-stage methods (Hemker, 1972), and so on (Liang and Wu, 2008; Wu et al., 2019).

2.1.1 SCODE

SCODE is a bioinformatics tool designed for scRNA-seq data by using the linear ODEs with pseudo-time series to describe expression dynamics and infer GRNs (Matsumoto et al., 2017). Two important assumptions are made in the SCODE: (1) all cells are on the same trajectory, that is, all cells are differentiating into the same cell type; and (2) the expression change rate of each TF linearly depends on expression profiles of themselves. Thus, the expression dynamics of TFs can be described for all differentiating cells along the pseudo-time series by using the linear ODEs:

$$\frac{d\mathbf{x}_c}{dt} = A\mathbf{x}_c, \tag{3}$$

where $\mathbf{x}_c := [x_1, x_2, \cdots, x_T]_c^{\mathrm{T}}$ denotes the expression of T TFs in cell c at time t_c, and the square matrix A represents the regulatory network among TFs. More precisely, the ODE (3) for each element x_i in vector $\mathbf{x}_c$ can be reformulated in the form of equation (2):

$$\frac{dx_i}{dt} = A_{i\cdot}\mathbf{x}_c = A_{i1}x_1 + A_{i2}x_2 + \cdots A_{iT}x_T,$$

where dx_i/dt represents the expression change rate of the ith TF. The task of the ODE-based model is to estimate the matrix A such that the expression change rate of the ith TF at the time t_c can be approximately described by all TFs' expression levels.

A major challenge of the ODE-based models is the expensive computational complexity caused by the high dimensionality of samples and genes. To reduce the computational complexity, SCODE alternatively solves an ODE with low-dimensional data by assuming that the high-dimensional data can be linearly expressed in a low-dimensional subspace (Matsumoto et al., 2017). In details, suppose that $\mathbf{x}_c$ can be expressed as a linear regression of a low-dimensional subspace

$$\mathbf{x}_c = W\mathbf{z}_c, \tag{4}$$

where $W \in \mathbb{R}^{T\times D}$ with $D \ll T$, and $\mathbf{z}_c$ obeys an ODE

$$\frac{d\mathbf{z}_c}{dt} = B\mathbf{z}_c. \tag{5}$$

Then the equation (3) is reduced to

$$\frac{d\mathbf{x}_c}{dt} = WBW^{+}\mathbf{x}_c,$$

where W^{+} denotes the pseudo-inverse matrix of W, and thus, A can be generated by

$$A = WBW^{+}.$$

Solving the ODE (5) in a low-dimensional subspace instead of the ODE (3), the SCODE algorithm significantly reduces the computational complexity and consumes much less running time than the traditional ODE (3). Also, this method is capable of dealing with large networks, for instance, a network with 5000 genes (Pratapa et al., 2020). However, the linear relationship in ODE might be too simple to describe the regulatory relationships between TFs. In addition, SCODE cannot directly infer GRN from single-cell expression data without temporal information (Matsumoto et al., 2017). For example, a tissue sample containing various cell types going through different biological processes is not suitable to be analyzed by this method.

2.1.2 GRISLI

GRISLI is another bioinformatics tool for single-cell pseudo-time-series data based on linear ODE (Aubin-Frankowski and Vert, 2018), where the expression dynamics are modeled by ODE (3) as in SCODE. While, different from SCODE, GRISLI designs a fast algorithm via solving a linear regression with a response as $d\mathbf{x}_c/dt$ in ODE (3) instead of integrating the ODE. The inferred GRN is assumed to be sparse, that is, most of elements in matrix A are zero, due to the biological assumption that each gene is regulated by only a few TFs.

Breaking the assumption in SCODE that all cells are in the same trajectory, GRISLI believes that different cells could evolve on different trajectories and focuses on those cells whose trajectories are close to each other. First, the expression change rate, also described as velocity, between cell c and cell e at two close pseudo-time points t_c and t_e is estimated by

$$\hat{\mathbf{v}}_{c,e} = \frac{\mathbf{x}_c - \mathbf{x}_e}{t_c - t_e}.$$

Considering that some data points might live in the past $(t < t_c)$ or the future $(t > t_c)$ of a given data point $(\mathbf{x}_c, t_c)$, the final estimator of velocity $\hat{\mathbf{v}}_c$ of cell c is defined as a weighted average of all velocities between cell c and those cells closed to it, which is written as

$$\hat{\mathbf{v}}_c = \frac{1}{2}\frac{\sum_{e|t_e>t_c} K(\mathbf{x}_e, t_e, \mathbf{x}_c, t_c)\hat{\mathbf{v}}_{c,e}}{\sum_{e|t_e>t_c} K(\mathbf{x}_e, t_e, \mathbf{x}_c, t_c)} + \frac{1}{2}\frac{\sum_{e|t_e<t_c} K(\mathbf{x}_e, t_e, \mathbf{x}_c, t_c)\hat{\mathbf{v}}_{c,e}}{\sum_{e|t_e<t_c} K(\mathbf{x}_e, t_e, \mathbf{x}_c, t_c)},$$

where the spatio-temporal kernel $K(\mathbf{x}_e, t_e, \mathbf{x}_c, t_c)$ measures the significance of a point to the velocity estimation. The velocity matrix $\widehat{\mathbf{V}} := [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_C] \in \mathbb{R}^{G\times C}$ is then estimated with corresponding expression data $\mathbf{X} := [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_C] \in \mathbb{R}^{G\times C}$, where G and C are the numbers of genes and cells, respectively.

The following procedures are repeated to obtain the frequency of nonzero elements in the estimated matrix $\hat{A}$: (1) data $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{V}})$ are generated by randomly subsampling and multiplying each row i of $\mathbf{X}$ by a random number; see section Methods in (Aubin-Frankowski and Vert, 2018) for details; (2) the Lasso regression (Tibshirani, 1996),

$$\min_{A_{j\cdot} \in \mathbb{R}^G} \left\|\widetilde{\mathbf{V}}_{j\cdot} - A_{j\cdot}\widetilde{\mathbf{X}}\right\|^2 + \lambda\left\|A_{j\cdot}\right\|_1,$$

is then solved for each row j to obtain a sparse matrix $\hat{A}$, where $\|\cdot\|$ and $\|\cdot\|_1$ denotes sum of squared values and absolute values, respectively, of all elements in the vector. The penalty parameter λ is set to satisfy the required number of nonzero entries in the row vector of A. After repetition of above procedures, the final GRN can be inferred based on the area score (Haury et al., 2012) or the original stability selection score (Meinshausen and Bühlmann, 2010) calculated from the frequency of occurred regulatory links (nonzero elements in the estimated matrix $\hat{A}$).

As GRISLI describes expression dynamics by linear ODE as SCODE does, the problem is transformed as a sparse regression under the assumption that inferred GRN is sparse. GRISLI is more efficient to estimate the matrix A via solving a convex optimization problem rather than integrating the ODE, and more genes (but less than 1000 genes) can be considered in practice (Pratapa et al., 2020). Moreover, it allows cells to be on different trajectories, which suits for more realistic and general cases. For example, cells may differentiate into two types of cells simultaneously. However, the same as SCODE, GRISLI cannot reconstruct the GRN directly from scRNA-seq data without temporal information.

2.1.3 InferenceSnapshot

InferenceSnapshot is a modular skeleton to extract the temporal information and capture gene expression dynamics directly from scRNA-seq snapshot data (Ocone et al., 2015). By combining the diffusion map algorithm for dimensionality reduction (Coifman et al., 2005) and ad hoc algorithm for clustering, the low-dimensional data can be obtained and separated into several branches with different cellular processes. Pseudo-time series is generated by using the Wanderlust algorithm (Bendall et al., 2014) to order single-cells along discrete paths that represent pseudo-time variables. Two types of ODE-based models are used to describe the interactions between M TFs $x_i(i = 1, \dots, M)$ and target gene y, representing AND and OR logic gates when combining regulatory effects of TFs, which are respectively formulated as

$$\frac{dy}{dt} = \alpha \prod_{m=1}^{M} f_m(x_m(t), \theta_m) - \mu y,$$

$$\frac{dy}{dt} = \alpha \sum_{m=1}^{M} f_m(x_m(t), \theta_m) - \mu y,$$

where α and μ denote the production rate and decay rate of target gene expression, respectively, and

$$f(x(t); \kappa, b) := \begin{cases} \dfrac{x^b}{x^b + \kappa^b}, & \text{if } x \text{ is activating,} \\ \dfrac{\kappa^b}{x^b + \kappa^b}, & \text{if } x \text{ is inhibiting.} \end{cases}$$

Markov chain Monte Carlo based method is used to estimate the parameters in ODE-based models mentioned above. In the model selection process, a coarse GRN is generated by GENIE3 (Vân Anh Huynh-Thu et al., 2010) as prior knowledge, and Bayes' factors are computed to select the ODE model from Bayesian model comparison through thermodynamic integration (Calderhead and Girolami, 2009).

InferenceSnapshot makes it possible to extract pseudo-time series from snapshot data directly and allows the analysis of data with multiple trajectories. Using the nonlinear function and different logic to combine regulatory effects of multiple TFs, InferenceSnapshot can be used to describe more complicated networks and nonlinear expression relationships, but difficult to be scaled up to large networks due to high computational complexity of ODE and Bayesian models (e.g., a network with 18 genes is considered in the original study) (Ocone et al., 2015). Moreover, the final inferred GRN would be limited because of the coarse GRN generated from GENIE3.

2.2 Regression-based model

Different from the ODE that considers expression change rate, the regression-based model is built on the assumption that the expression of a target gene can be predicted by the expression of TFs regulating it. Regression is one of the most common used methods to search for a suitable prediction function f to characterize the underlying networks. For example, if the expression data of gene y can be predicted by the expression data of TFs x, then those TFs jointly regulates gene y. Hence, the regression model is written as

$$y = f(x) + \varepsilon, \tag{6}$$

where ε denotes the noise in data. The function f in the regression model can be either linear or non-linear, depending on the assumption of the structure of the target network.

A significant benefit of the regression model is that it is simple to understand its mathematical principle and convenient to apply to the complicated biological system (Pratapa et al., 2020; Qin et al., 2014). When the type of prediction function f is provided due to the biological process or data observation, the ordinary least squares is a popular method for solving the regression model (6) to estimate the coefficients involved in f, which aims to minimize the sum of squared errors between the true data and the prediction, that is,

$$\min \|f(x) - y\|^2 . \tag{7}$$

The most common form of regression is linear regression and the associated linear least squares method. Furthermore, the structure of the GRNs can be characterized by adding an associated penalty function p in the regression model to improve the accuracy and stability of prediction, that is,

$$\min \|f(x) - y\|^2 + \lambda p(x).$$

For example, ridge regression uses the ℓ_2 penalty (i.e., $p(x) = \|x\|^2$) to measure the magnitude of coefficients (Hoerl and Kennard, 1970); Lasso regression employs the ℓ_1 penalty (i.e., $p(x) = \|x\|_1$) to induce the sparsity of variables (Tibshirani, 1996). Moreover, the low-order penalized Lasso (Qin et al., 2014) and fused Lasso have been used in GRN inference (Omranian et al., 2016).

Another important benefit of the regression model is the exclusive development of optimization algorithms. Several popular and efficient numerical algorithms have been proposed to solve the least squares problem (7) and the ridge regression problem such as gradient descent methods, Newton-type methods and Levenberg-Marquardt method (Bertsekas, 2015; Hu et al., 2016b; Nocedal and Wright, 2006). Many state-of-the-art algorithms have been designed and applied to solve the Lasso-type regression models such as proximal/projected gradient methods, alternative direction method with multipliers, block coordinate descent methods and augmented Lagrange methods (Boyd et al., 2011; Hu et al., 2017; Wright, 2015).

Furthermore, with non-linear functions, other regression-based methods like tree-based method (Vân Anh Huynh-Thu et al., 2010) are also applied to fit expression data.

2.2.1 GENIE3

Gene network inference with ensemble of trees (GENIE3) is a tree-based method to reconstruct GRNs (Vân Anh Huynh-Thu et al., 2010). Although it was originally designed for bulk RNA-seq, it has also been used in scRNA-seq data (Pratapa et al., 2020) because of its good performance in GRN reconstruction from bulk RNA-seq (Marbach et al., 2012). The input expression data is an $N \times G$ matrix, where the expression of G genes are quantified in N experiments (or cells). GENIE3 assumes that the expression of each gene could be described as a function of the expression of some TFs, which means the selected TFs could regulate the target gene. Thus, the inference of GRNs is decomposed into G different regression problem for each target gene.

Denote the expression of gene j and all genes except gene j in the kth experiment (or cell) by $x_{j,k}$ and $\mathbf{x}_{-j,k}$, respectively. The major objective of GENIE3 is to find a suitable function f_j for gene j such that

$$x_{j,k} = f_j(\mathbf{x}_{-j,k}) + \varepsilon_k, \forall k \in [1, N],$$

where ε_k represents a random noise with zero mean. Regression tree (Breiman et al., 1984) is a good candidate to seek such function and identify those TFs that could be used to predict the expression of gene j. Generally, random forests (Breiman, 2001) is able to reduce the variance and improve the performance compared with regression tree (Hastie et al., 2009). In addition, random forests is able to avoid the overfitting phenomenon and requires little tuning parameters. Consequently, random forests is applied in GENIE3 for each gene to identify the TFs used to predict.

In random forests, m variables (e.g., TFs) are randomly selected from G variables as split candidates at each node, and K single regression trees are built by K bootstrapping. Importance measure (IM) is defined to quantify how relevant each TF (input gene) is to the target gene (output gene) and is computed for each single regression tree. The attribute IM is extended by averaging the IMs over K regression trees in random forests; see section Methods in (Vân Anh Huynh-Thu et al., 2010) for details. By ranking G IMs from every single ensembled tree and aggregating them to get global interaction ranking, the final GRN is inferred by setting a threshold to identify the regulatory links.

Benefitting from the fact that few assumptions are required in random forests, GENIE3 owns ability to explain more complex regulatory relationships in GRNs when comparing with linear regression. GENIE3 is a good choice for scRNA-seq data without temporal information, while it might perform worse than other methods if scRNA-seq data contains temporal information. In addition, it may be harder for GENIE3 to infer large networks when it is needed to build $G \times K$ regression trees one by one, while the computational difficulty can be relieved by parallel computation. For example, a large network (e.g., with 5000 genes) could still be inferred in practice (Pratapa et al., 2020).

2.2.2 SINCERITIES

Single-cell regularized inference using timestamped expression profiles (SINCERITIES) applies regularized linear regression and partial correlation analysis to reconstruct GRNs based on temporal changes in the distributions of gene expression (Papili Gao et al., 2018). This method assumes the expression change of a target gene linearly depends on the expression changes of TFs at a time delay.

Such temporal changes in the expression of each gene is measured by the distance of gene expression distributions between two subsequent time points, which is called as the distributional distance (DD). Kolmogorov-Smirnov distance is used to compute the DDs of all genes (Massey Jr, 1951) and $\widehat{DD}_{j,l}$ denotes the normalized DD of gene j at time window l. Based on the assumption mentioned above, SINCERITIES reconstructs GRNs by solving G linear regressions for G genes. More precisely, the linear regression for target gene j at time window $l+1$ is formulated as:

$$\widehat{DD}_{j,l+1} = A_{1,j}\widehat{DD}_{1,l} + A_{2,j}\widehat{DD}_{2,l} + \cdots + A_{G,j}\widehat{DD}_{G,l},$$

where $A_j := [A_{1,j}, A_{2,j}, \cdots, A_{G,j}]^\top$ represents the coefficients in linear regression. Since the number of genes is larger than the number of time windows in general, SINCERITIES applies an ℓ_2 norm penalized linear regression (ridge regression) (Hoerl and Kennard, 1970) to overcome the difficulty of solving the underdetermined equations for target gene j, that is,

$$\min_{A_j} \|Y_j - XA_j\|^2 + \lambda\|A_j\|^2,$$

where

$$Y_j := [\widehat{DD}_{j,2}, \widehat{DD}_{j,3}, \cdots, \widehat{DD}_{j,n-1}]^\top \text{ and } X := \begin{bmatrix} \widehat{DD}_{1,1} & \widehat{DD}_{2,1} & \cdots & \widehat{DD}_{G,1} \\ \widehat{DD}_{1,2} & \widehat{DD}_{2,2} & \cdots & \widehat{DD}_{G,2} \\ \vdots & \vdots & \cdots & \vdots \\ \widehat{DD}_{1,n-2} & \widehat{DD}_{2,n-2} & \cdots & \widehat{DD}_{G,n-2} \end{bmatrix}.$$

After ranking the absolute values of the coefficients of all possible edges, the inferred GRN could be obtained by setting a threshold for the ranked value. The sign of the regulatory edge between each pair of TF and target is determined by the sign of the corresponding partial correlation.

SINCERITIES reconstructs the GRNs with low computational complexity and suits for high-dimensional data (e.g., a network with 5000 genes) (Papili Gao et al., 2018; Pratapa et al., 2020). As the regressions for all genes are independent of each other, the running time could be depleted by employing parallel computation technique. However, temporal information is required in this method, and the relationship between temporal changes in the expression of TFs and target gene may not be linear as assumed to be.

2.3 Correlation/information-based model

The regulatory links in GRNs can also be determined by measuring the relationship between the expression of target genes and TFs. The Pearson's correlation, is the simplest statistic to characterize the association between X and Y:

$$\rho_{X,Y} := \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where μ_X and σ_X denote the mean and variance of variable X, respectively, $\mathrm{cov}(X,Y)$ represents the covariance between X and Y, and $E(\cdot)$ denotes the expectation.

However, the Pearson's correlation is too naive to characterize the complicated regulatory relationship in GRNs. For example, if genes i and j are not connected but both connected to gene k, the correlation between i and j is still possible to be high. Partial correlation (Lawrance, 1976) could be used to avoid the effect of other genes. It can be quickly obtained by computing the correlation between the residuals from two corresponding linear regressions, which means that the linear relationship is assumed.

In information theory, the entropy $H(X)$ is used to measure the uncertainty of random variable X. If the random variable Y is known, one may define another concept called conditional entropy $H(X|Y)$ (Cover and Thomas, 2012). These two basic concepts are defined as

$$H(X) := -\sum_{x \in X} p(x) \log p(x)$$

and

$$H(X|Y) := -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(y)},$$

respectively.

By considering the distributions of genes, mutual information (MI) has the ability to quantify the dependence between two genes based on their distributions. MI for two random variables X and Y is formulated as

$$I(X;Y) := \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) = H(X) - H(X|Y).$$

The second equality shows the relationship between MI and entropy. From the formula mentioned above, MI measures the reduction in uncertainty of a random variable X when the knowledge of variable Y is known. Considering the effect from a third variable Z, conditional MI is used to measure the reduction in the uncertainty of X due to knowledge of Y when Z is given (Cover and Thomas, 2012), which is formulated as

$$I(X;Y|Z) := \sum_{z \in \mathrm{Z}} p(z) \sum_{x \in X} \sum_{y \in Y} p(x,y|z) \log \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right).$$

However, the estimation of MI and conditional MI involves data discretization and estimation of empirical probability distributions (Chan et al., 2017), and thus different choices of discretization method and estimator for MI would affect the performance of MI-based method (de Matos Simoes and Emmert-Streib, 2011; Walters-Williams and Li, 2009; Zhang and Zheng, 2015).

The existence of regulatory links is more reliable when the value of measurements is larger. After computing these measurements mentioned above for all genes, those links with lower values could be removed by choosing a threshold to infer the final GRNs.

2.3.1 LEAP

Lag-based expression association for pseudo-time series (LEAP) is a correlation-based algorithm to infer the GRNs for pseudo-time-series data (Specht and Li, 2017). As LEAP is developed based on the Pearson's correlation, the linear relationship between a pair of genes is always assumed (Lee Rodgers and Nicewander, 1988).

Given expression data $x_{i,t}$ of gene i at time $t \in \{1,2,\cdots,T\}$, the series $\mathbf{X}_{i,l} := \{x_{i,l+1}, x_{i,l+2}, \cdots, x_{i,l+s}\}$ for gene i is extracted by setting windows of size s, where the lag $l \in \{0,1,\cdots,T-s\}$. Instead of Pearson's correlation, LEAP uses maximum absolute correlation (MAC) to measure the regulatory relationship:

$$\rho_{ij}^{*} := \max_{l \in \{0,1,\cdots,T-s\}} |\rho_{ijl}|,$$

where ρ_{ijl} denotes the Pearson's correlation between gene i at lag 0 ($\mathbf{X}_{i,0}$) and gene j at lag l ($\mathbf{X}_{i,l}$). The directional regulatory relationship could be inferred by the value $l^{*} = \arg \max_{l \in \{0,1,\cdots,T-s\}} |\rho_{ijl}|$ and the corresponding MAC value ρ_{ij}^{*}: (1) if $l^{*} \neq 0$, the MAC value $\rho_{ij}^{*} > 0$ and $\rho_{ij}^{*} < 0$ represents that the gene i activates and inhibits gene j, respectively; (2) if $l^{*} = 0$, gene i, and j are both regulated by a third gene. Finally, the statistical significance can be calculated based on the false discovery rate (Benjamini and Hochberg, 1995).

The LEAP provides a strategy to find the regulatory links between genes and define their directional relationship by computed measurements. However, the relationships between all genes are assumed to be linear, where it might not satisfy for most cases. As the temporal information is considered in the method, pseudo-time-series data is required to infer GRNs. In practice, this correlation-based model generally consumes less time because the measurements can be directly computed by the analytical formulas, and it works for a large network. For example, a network with 5000 genes is considered in (Pratapa et al., 2020).

2.3.2 PIDC

Partial information decomposition and context (PIDC) is an information-based algorithm to infer the regulatory relationship between genes (Chan et al., 2017). Partial information decomposition (PID) is used to decompose the multivariate MI, where unique information $\mathrm{Unique}_Z(X;Y)$ is the portion of information provided only by Y (Williams and Beer, 2010). To quantify the information between multiple genes in GRNs, PIDC defines a new measurement called proportional unique contribution (PUC) between genes X and Y, which is the sum of the ratio $\mathrm{Unique}_Z(X;Y)/I(X;Y)$ for all other genes Z in set S. The ratio eliminates the impact from the quantity of MI, and the computation of PUC could be formulated as

$$u_{X,Y} := \sum_{Z \in S_{\setminus\{X,Y\}}} \frac{Unique_Z(X;Y)}{I(X;Y)} + \sum_{Z \in S_{\setminus\{X,Y\}}} \frac{Unique_Z(Y;X)}{I(X;Y)}.$$

A global threshold for PUC scores might bias the result of the inferred GRNs due to the distributions of PUC scores differ between genes (Chan et al., 2017). The confidence of a regulatory link between a pair of genes could be calculated by the empirical probability distribution estimated from PUC scores; see section Results in (Chan et al., 2017) for details.

The PIDC provides an approach to quantify the relationship between a pair of genes considering the effect of other related genes in GRNs. It extracts more information from the expression data. However, the data discretization and MI estimators are required in this method, which might impede the computation of PUC scores. The performance of PIDC might be influenced by the choice of data discretization methods and MI estimators (Chan et al., 2017). Although the method owns high complexity, the problem could be relieved by implementing in Julia programming language to speed up (Bezanson et al., 2017). Moreover, it is capable of dealing with a large network (e.g., with 5000 genes) in practice (Pratapa et al., 2020).

2.3.3 Scribe

Scribe is another information-based toolkit designed for datasets with temporal information to infer causality relationship between genes. It relies on restricted directed information (RDI) (Rahimzamani and Kannan, 2016) to measure the information transmitted from potential regulators to downstream targets. The GRNs can be correctly reconstructed based on the assumption that the underlying processes can be described by a first-order Markov process, which is true in most biological processes (Rahimzamani and Kannan, 2016). To measure information transferred from the regulator X at time $t-d$ to Y at time t with time delay d when the information of Y at time $t-1$ is given, the computation of RDI is formulated in the form of conditional MI:

$$\mathrm{RDI}_d(X \to Y) := I(X_{t-d}; Y_t | Y_{t-1}).$$

Furthermore, conditional RDI (cRDI) is considered to remove the arbitrary effect from other potential regulators Z, and thus the computation of cRDI can be formulated as:

$$\mathrm{RDI}_{d_1}\left(X \to Y | Z_{t-d_2}\right) := I\left(X_{t-d_1}; Y_t | Y_{t-1}, Z_{t-d_2}\right).$$

To correct the sampling bias in computation and improve the performance, uniformized RDI and cRDI scores are computed by replacing the original empirical distribution of the samples with a uniform distribution (Qiu et al., 2020). The final GRN is generated by the RDI-based scores and further refined by context likelihood of relatedness algorithm (Faith et al., 2007) with graph regularization method; see section STAR Methods in (Qiu et al., 2020) for details.

Scribe extracts more intrinsic information from single-cell expression data by considering arbitrary effect delay from regulator X to target Y and the effect from other potential regulators Z. It quantifies the regulatory causality between X and Y based on the time information. Also, Scribe can detect both linear and non-linear causality in GRNs (Qiu et al., 2020). However, as the RDI is one type of conditional MI, the Scribe involves the estimation of RDI-based measurements, which might be time-consuming. In practice, a network with 1000 genes can be reconstructed by this method (Pratapa et al., 2020). In addition, scRNA-seq data with temporal information is required because the time information is needed. As several methods mentioned above, Scribe can analyze pseudo-time series and RNA velocity.

2.4 Boolean network

Unlike the continuous expression values of the nodes in ODE, Boolean network describes the interaction of genes with discrete values for their states along with discrete time points. The nodes and edges of the network represent genes and regulatory relationships between them, respectively. To represent the expression status of genes, the numeric "1" or "0" is used to denote the state of nodes as "on" or "off". In order to characterize the dynamics of the network, Boolean functions with three main operations: AND, OR and NOT are built to update the successive state for each node, where the operators represent the regulatory manners of TFs to their targets. The final successful model can be obtained by verifying the dynamic sequence of system states and comparing with biological evidence. A drawback of Boolean network is that the computation consumes more time when more possible networks are needed to be considered with an increasing number of genes. Thus, the method is limited in a small number of genes in real practice (generally smaller than 100) (Fiers et al., 2018; Liang and Han, 2012). The method would be sensitive to dropouts since the binarization of expression data is required before modeling (Fiers et al., 2018; Wynn et al., 2012). The example showed below simply illustrates the Boolean network for three nodes.

Example 1. Consider the following network with three nodes as X_1, X_2 and X_3:

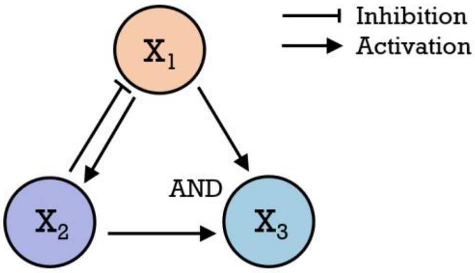

Figure 3 Three nodes network.

The Boolean update functions can be presented as follow:

$$\begin{aligned} X_1(t+1) &= \text{NOT } X_2(t); \\ X_2(t+1) &= X_1(t); \\ X_3(t+1) &= X_1(t) \text{ AND } X_2(t), \end{aligned}$$

where $X_1(t)$ denotes the state of the node X_1 at the time t.

2.4.1 SCNS toolkit

Single cell network synthesis toolkit (SCNS toolkit) is a Boolean network-based toolkit for scRNA-seq data with temporal information to reconstruct and analyze GRNs. The diffusion map method (Coifman et al., 2005) is used to identify the developmental trajectories in gene expression data from different cell stages (Moignard et al., 2015).

The SCNS toolkit firstly discretizes the single-cell gene expression into binary states, where "1" and "0" represent that a gene is expressed or not respectively. According to the Boolean update functions that represent connections of a possible network, the vector bearing "1" or "0" states of all genes at an early time point can transit into the state vector of the next time point. State vectors at two adjacent time points could be connected to form a state transition graph. Boolean functions that fit the state series best are being chosen when the network is being reconstructed; see section Implementation in (Woodhouse et al., 2018) for details.

The SCNS toolkit provides insights into the developmental processes and the interactions between genes in GRNs across time. It considers regulatory logic when reconstructing the GRNs. Yet the method for data discretization in SCNS toolkit might influence the further inference of GRNs. As we mentioned above, the Boolean network-based method can only deal with the small-scale GRNs in real-life computation.

3. Methods for scRNA-seq data with genome

Although scRNA-seq data are widely used for GRN reconstruction, the performance of current tools on this data type is still unsatisfactory (Chen and Mar, 2018; Pratapa et al., 2020). This is because, with similarity to those designed for bulk RNA-seq, these tools are all based on the assumption that the expression relationships between a target gene and its TFs imply transcriptional regulations among them. However, the observed associations between TFs and genes may be due to other biological events or even randomness rather than transcriptional regulations. Given the stochastic variation of gene expression in single cells, the dropouts and technical variations of scRNA-seq data, the signal-to-noise ratio of scRNA-seq is even lower than that of bulk RNA-seq. Besides, based on scRNA-seq data alone, it is also difficult to distinguish between direct and indirect regulations. To overcome these issues and improve the performance of GRN inference, integration of additional data is considered as an improved way. Genome sequences bearing the genomic regulatory codes can be exploited to guide the identification of potential TF binding. A TF binding motif located at the DNA regulatory element of a gene indicates a potential direct regulation between them.

Single-cell regulatory network inference and clustering (SCENIC) is one of such tools

(Aibar et al., 2017). It incorporates the promoter sequences extracted from the reference genome to search direct connection between TFs and their target among the coexpression network modules built by GENIE3 (Vân Anh Huynh-Thu et al., 2010) or GRNBoost (Aibar et al., 2017). By removing the indirect targets lacking enriched motif detected using RcisTarget (Aibar et al., 2017), SCENIC dramatically reduces the false connections in the GRN inferred from scRNA-seq alone (Aibar et al., 2017). It also quantifies the subnetwork activity in each cell by the AUCell algorithm (Aibar et al., 2017), which allows the comparison of the activities of cell-specific networks among different cell types and subpopulations. It enables the combination of coexpression networks with *cis*-regulatory analysis, leading to a better exploration of GRNs and cell states. Thus, the datasets with complex cell states can also achieve good performances. When dealing with very large datasets, GRNBoost, a variant of GENIE3, can advance the efficiency and reduces the time used in GRN reconstructions. The SCENIC provides a strategy to discover interactions between TFs and target genes, yet the inference of the coexpression network might affect the further analysis. The SCENIC might perform better with other methods when it is inferring coexpression networks.

However, when the majority of associated genetic variants locates in regulatory regions of patient genomes in diseases like cancer (Melton et al., 2015), the reference genome is unable to reflect the heterogeneity of regulatory codes in cell populations. Regulatory variants in different cell subpopulations may drive the regulations on diverse patterns of gene expression. Thus, integration of scRNA-seq and single-cell genome sequencing will be a better strategy to understand the heterogeneity of GRNs in a tumor cell population. Although technologies, such as G&T-seq (Macaulay et al., 2015) and DR-seq (Dey et al., 2015), allow parallel sequencing of the genome and transcriptome in the same single cell, the high cost of sequencing covering the whole genomes for thousands of single cells and relatively low resolution of the technique have limited the popularization of this approach. Thus, so far, no bioinformatics tools were especially designed for this analysis. However, it is still worthy to develop such tool especially for cancer research, when targeted genome sequencing may dramatically reduce the sequencing cost by selecting genes and genomic regions of interests (Ng et al., 2009).

4. Methods for scRNA-seq data with single-cell epigenomes

Fortunately, the development of single-cell epigenomic technologies, such as scATAC-seq, allows the identification of DNA regulatory elements in single cells at a reasonable cost. Open chromatin regions detected by scATAC-seq often contain active DNA regulatory elements for TF binding and gene regulations (Buenrostro et al., 2015). Thus, scATAC-seq is able to identify direct regulations in GRNs. The integration of bulk RNA-seq and bulk ATAC-seq (or other epigenomic data) has been proved to improve the accuracy of GRN inference significantly (Ackermann et al., 2016; Qin et al., 2014; Wang et al., 2015). This approach is also applicable to single-cell sequencing data. However, due to cell-type/condition specificity of transcriptome and epigenome profiles, the integration of bulk RNA-seq with bulk ATAC-seq/ChIP-seq usually requires that the two data sets are derived from the same cell type and in the same condition. Although several technologies allow sequencing transcriptome and epigenome simultaneously in the same cell (Angermueller et al., 2016; Chen et al., 2019; Hu et al., 2016a), researchers often conduct scRNA-seq and single-cell epigenome separately, so the major challenge for the integration approach is how to match the cell clusters of the same cell type, condition or cell state for the two sequencing data types respectively. Since scATAC-seq is more commonly used for single-cell epigenome profiling than other techniques like scChIC-seq, three bioinformatics tools have been introduced to combine scRNA-seq and scATAC-seq data for GRN reconstruction. These methods can analyze more than ten thousand genes, and they are applicable to high-dimensional matrices during multi-omics data integration (Table 1).

4.1 SOM

Self-organizing map (SOM), also known as the Kohonen network, is an unsupervised learning method for clustering and visualization (Kohonen, 1982, 1990). The main structure of SOM is separated into two parts: an input layer and a competitive layer (also as output layer). The competitive layer is generally a two-dimensional array of output nodes that are assumed to be a regular hexagonal or rectangular grid.

Denote n nodes in input layer by

$$\mathbf{X} := [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_m] \in \mathbb{R}^{m\times n},$$

where $\mathbf{x}_u \in \mathbb{R}^n$ is the uth input vector (e.g. the uth sample in expression data). Each unit i in competitive layer is connected to input layer by a weight vector

$$\mathbf{w}_i := [w_{i1}, w_{i2}, \cdots, w_{in}]^\top \in \mathbb{R}^n,$$

where w_{ij} denotes the weight for the connection between unit i and node j (e.g., gene j) in input layer. The iterative computation in SOM involves searching a winning unit k in competitive layer based on the minimal Euclidean distance

$$k = \arg\min_i \|\mathbf{w}_i - \mathbf{x}_u\|^2$$

or the maximal inner product

$$k = \arg\max_i \mathbf{w}_i^\top \mathbf{x}_u.$$

Given a random initial weight vector $\mathbf{w}_i(0)$ for each unit i, the weights for the neighborhood of winning unit k are updated by

$$\mathbf{w}_i(l+1) = \mathbf{w}_i(l) + \eta(l) h_{ki}(l)[\mathbf{x}_u - \mathbf{w}_i(l)], \qquad \forall i \in O_k$$

with a learning rate $\eta(l)$, where O_k denotes a set of unit k's neighborhood (based on the structure in the competitive layer), and $h_{ki}(l)$ is the neighborhood function for unit k; see (Kohonen, 1990) for more details.

The SOM has the ability to map data from a high dimension space to a low dimension one. Although the convergence of the algorithm has been proved under some conditions, the SOM might converge until hundreds of thousands of iterations (Bianchi et al., 2007). Thus, SOM is computationally expensive compared with other clustering methods.

4.1.1 LinkedSOMs

Linked self-organizing maps (LinkedSOMs) is a bioinformatics tool developed to infer GRNs by integrating scRNA-seq and scATAC-seq data. The input data for LinkedSOMs are the gene expression data and chromatin data, while the pseudo-time is not required. Two SOMs with the output set of SOM units are available after training the scRNA-seq and scATAC-seq data separately. K-means clustering (Forgy, 1965) is then performed to determine centroids among units, and the cluster of the units, called metaclusters, are built around these centroids based on Akaike information criterion score (Akaike, 1998). To link gene expression and chromatin accessibility, GREAT algorithm (McLean et al., 2010) is implemented to obtain the linked SOM metaclusters (LMs). The underlying GRNs are then inferred after gene ontology analysis and motif analysis on these LMs; see section Methods in (Jansen et al., 2019) for details.

Training two SOMs for scRNA-seq and scATAC-seq datasets makes LinkedSOMs, time-consuming as mentioned above, though it can still analyze large datasets. Even though the original study of LinkedSOMs focuses on integrating scRNA-seq and scATAC-seq data, it is also applicable to multi-omics data analysis incorporating other single-cell sequencing data.

4.2 NMF

Nonnegative matrix factorizations (NMF) aims to decompose a nonnegative matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ into two nonnegative matrices $W \in \mathbb{R}^{n\times r}$ and $H \in \mathbb{R}^{r\times m}$ such that $\mathbf{X} \approx WH$ (Lee and Seung, 1999). The approach to find W and H is by solving the minimization problem

$$\min_{W,H\geq 0} \|\mathbf{X} - WH\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Via the NMF, the matrix $\mathbf{X}$ could be approximately represented as linear combinations of r column vectors in feature matrix W with assignment weight matrix H. The NMF method has been widely applied to GRN inference (Ochs and Fertig, 2012; Wu et al., 2016; Yang and Michailidis, 2016). Many methods are developed to solve the NMF problem, such as simple multiplicative update method (Lee and Seung, 2001) and projected gradient method (Lin, 2007). To the best of our knowledge, the convergence properties of the projected gradient method have been proved, while the convergence properties of simple multiplicative update method are still not clear (Lin, 2007; Takahashi et al., 2018).

4.2.1 Coupled NMF

Coupled nonnegative matrix factorizations (coupled NMF) is an NMF-based approach to reconstruct GRNs via integrative analysis of scRNA-seq and scATAC-seq data. The main assumption in coupled NMF is that the expression of a subset of genes (detected by scRNA-seq) can be linearly predicted from the status of chromatin regions (detected by scATAC-seq).

Coupled NMF aims to cluster the cells in each dataset with information from another one by developing a new optimization problem based on NMF. Denote the scRNA-seq and scATAC-seq data by $\mathbf{X}$ and $\mathbf{O}$, respectively. Borrowing the idea from NMF and introducing the coupling matrix A to connect the clusters W_1 and W_2 of two datasets, the coupled NMF is formulated as

$$\min_{W_1,H_1,W_2,H_2\geq 0} \frac{1}{2}\|\mathbf{O}-W_1H_1\|_F^2+\frac{\delta_1}{2}\|\mathbf{X}-W_2H_2\|_F^2-\delta_2\,\mathrm{tr}\big(W_2^{\mathrm{T}}AW_1\big)+\delta_3(||W_1||_F^2+||W_2||_F^2),$$

where $\delta_k(k=1,2,3)$ are the penalty parameters in this optimization problem. The trace term $\mathrm{tr}\big(W_2^{\mathrm{T}}AW_1\big)$ owns ability to induce the consistency of features W_2 with linear transformed features AW_1. The last term in objective function controls the growth of W_1 and W_2 (Duren et al., 2018). Before solving the coupled NMF mentioned above, the coupling matrix A is firstly obtained by performing the regression model on the paired gene expression and chromatin accessibility data. The coupled NMF is then solved by a modified multiplicative update algorithm (Duren et al., 2018). The method finally generates the cluster-specific expression of genes and accessibilities of regulatory elements, where the cluster-specific expression of genes can be predicted from the cluster-specific accessibilities of regulatory elements by AW_1. After gene ontology analysis and motif analysis on each cluster, in the end the final GRNs can be reconstructed; see section Materials and Methods in (Duren et al., 2018) for details.

Similar to LinkedSOMs discussed above, other single-cell multi-omics data can also be applied in this approach to analyze and infer the GRNs with coupled NMF. Although the numerical behavior of coupled NMF was showed (Duren et al., 2018), the convergence properties have not been established yet.

4.3 CCA

Canonical correlation analysis (CCA) is a method to project two different datasets into a correlated low-dimensional space by maximizing the correlation between two linear combinations of the features in each dataset (Hotelling, 1992). Denote two datasets by $\mathbf{X}$ and $\mathbf{O}$. Introducing the linear combinations as

$$\mathbf{U}:=\mathbf{Xu}\ \text{ and }\ \mathbf{V}:=\mathbf{Ov}$$

with two canonical correlation vectors (CCVs) $\mathbf{u}$ and $\mathbf{v}$, the CCA can be described as pursuing the maximum correlation of linear combinations $\mathbf{U}$ and $\mathbf{V}$:

$$\max_{\mathbf{u},\mathbf{v}} \mathrm{corr}(\mathbf{U},\mathbf{V}).$$

Supposed that the columns of $\mathbf{X}$ and $\mathbf{O}$ have been centered and scaled, the problem can be re-written as

$$\begin{aligned}&\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^{\top}\mathbf{X}^{\top}\mathbf{Ov}\\ &\text{s.t. } \mathbf{u}^{\top}\mathbf{X}^{\top}\mathbf{Xu}\leq 1,\\ &\qquad \mathbf{v}^{\top}\mathbf{O}^{\top}\mathbf{Ov}\leq 1,\end{aligned}$$

The solution ($\mathbf{u}$ and $\mathbf{v}$) of CCA can be obtained by solving a standard eigenvalue problem (Hotelling, 1992; Uurtio et al., 2017). When it comes to high-dimensional application, its performance achieves a better result if it treats the covariance matrix of $\mathbf{X}$ and $\mathbf{O}$ as diagonal matrix (Dudoit et al., 2002; Tibshirani et al., 2003). By replacing the $\mathbf{X}^{\top}\mathbf{X}$ and $\mathbf{O}^{\top}\mathbf{O}$ with the identity matrix, the modified optimization problem called diagonal CCA is reformulated as

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^{\top}\mathbf{X}^{\top}\mathbf{O}\mathbf{v} \\ &\text{s.t. } \|\mathbf{u}\|^2 \leq 1, \\ &\qquad \|\mathbf{v}\|^2 \leq 1, \end{aligned}$$

and it can be solved by penalized matrix decomposition (Witten et al., 2009).

4.3.1 Seurat v3

Seurat v3 is a bioinformatics framework that infer GRNs from scRNA-seq and scATAC-seq data based on CCA. Denote the scRNA-seq and scATAC-seq data by $\mathbf{X}$ and $\mathbf{O}$, respectively. The CCVs $\mathbf{u}$ and $\mathbf{v}$ are generated by performing diagonalized CCA with standard singular value decomposition method, which is followed by ℓ_2-normalization on CCVs to eliminate global differences in scale across datasets. For each cell in one dataset, its K-nearest neighbors (KNNs) in another dataset can be identified in the shared low-dimensional space based on the ℓ_2-normalized CCV. If a pair of cells from each dataset is contained in each other's KNN, the pair of cells is defined as the mutual nearest neighbor (MNN), also called anchor (Haghverdi et al., 2018; Stuart et al., 2019). Then the anchors are scored and filtered to alleviate the effects of any incorrectly identified anchors. After converting scATAC-seq data into a predicted gene expression matrix (Pliner et al., 2018), an integrated expression matrix for scRNA-seq and scATAC-seq is finally computed with the strategy in batch correction (Haghverdi et al., 2018). The GRNs can be inferred with this expression matrix as input via any single-cell GRN inference method; see section Method Details in (Stuart et al., 2019).

The Seurat v3 focuses on the integration of scRNA-seq with different single-cell technologies such as scATAC-seq. It generates an integrated expression matrix in the end, which can be the input in further downstream analysis like GRN inference with any single-cell analytic method. Moreover, the approach in Seurat v3 is extended to assemble multiple datasets, and this would provide a deeper insight into single cells. In addition, based on the principle of CCA and KNN, the Seurat v3 is capable of dealing with high-dimensional datasets.

5. Conclusions

With the development of various single-cell sequencing technologies nowadays, more and more methods for GRNs inference from single-cell sequencing data are proposed (Blencowe et al., 2019; Chen and Mar, 2018; Efremova and Teichmann, 2020; Fiers et al., 2018; Pratapa et al., 2020). Understanding the mathematical background of each method might help researchers use these methods appropriately in different cases. It also benefits the tool developer to design new tools with comprehensive considerations. This review introduces various single-cell sequencing data available for GRN reconstruction. Then mathematical principles and adaptabilities of several popular algorithms that have been applied to scRNA-seq data alone or integrative multiple single-cell data are discussed. For each reprehensive tool, the acceptable data type and underlying assumption are emphasized to point out the specific circumstance where the method could be applied.

As the proverb says, "Essentially, all models are wrong, but some models are useful". Although comparisons on several tools that work on scRNA-seq data have been performed with simulated data and real data in several published reviews (Chen and Mar, 2018; Pratapa et al., 2020), it is still difficult to conclude which method is the best. First, in general, it seems that there is no method that significantly outperforms others in all datasets, especially on real datasets. Second, since GRNs are highly condition-specific and largely unknown, the GRNs inferred by these tools from real scRNA-seq data are hard to be well evaluated. Current comparisons on their performance are usually based on "gold standard" of non-specific networks or very limited known network connections under the benchmarking data. While methods for integrative multiple single-cell data have the same issues. Thus, we only discuss their adaptabilities and limitations based on their basic algorithm here. Further comparison on the accuracy of GRNs that they predict from real data requires more good benchmarking data and corresponding verified gold standard networks, which is not available now.

We also point out that the future direction of method development would be the integration of multiple single-cell sequencing data. Integrations of single-cell multi-omics could reduce the impacts of noise and enhance the performance by cross-validating the regulatory connections in GRNs through multiple datasets. More integrative tools will emerge when more types of single-cell data, such as proteome, metabolome, cell image, et al., become prevalent in the future. They will depict gene regulatory mechanisms underlying disease and biological processes more accurately, and provide a more comprehensive map of GRNs covering multiple biological molecules and regulatory layers. In addition to the integration of multiple data types, combining multiple algorithms and tools has also been shown to improve the accuracy of network inference from bulk-cell data (Marbach et al., 2012). We speculate that the same phenomenon will occur for single-cell data. Thus, new tools considering multiple algorithms may further improve the prediction of GRNs from single-cell sequencing data.

Acknowledgment

This work was supported by the Natural Science Foundation of Guangdong Province of China (2019A1515011917, 2020B1515310008), Project of Educational Commission of Guangdong Province of China (2019KZDZX1007), Natural Science Foundation of Shenzhen (JCYJ20190808173603590, JCYJ20170817100950436, JCYJ20170818091621856) and Interdisciplinary Innovation Team of Shenzhen University. We would like to thank three anonymous reviewers and the editor for their constructive comments.

Table 1 Summary of bioinformatics tools for GRN reconstruction from single-cell sequencing data.

Data	Methods	Name	Reference	Data dimension (cell*gene)	Adaptability
scRNA-seq alone	ODE-based	SCODE	(Matsumoto et al., 2017)	mESC:456*100 Fibroblast:405*100 hESC:758*100	Reduced computational complexity; assume all cells are on the same trajectory; the linear relationship between change rate of target gene and expression of input is assumed; require expression data with temporal information.
		GRISLI	(Aubin-Frankowski & Vert, 2018)	Embryonic:373*40 Hescs:758*49	Consider multiple trajectories; assume that each gene is regulated by only a few TFs; expression change rate of target gene and TFs is assumed to be linearly related; require expression data with temporal information.
		InferenceSnapshot	(Ocone et al., 2015)	HSCs:597*18	Directly extract temporal information from single-cell snapshot data; reconstruct more complicated network; limited ODE-based models are considered; final inferred GRN would be limited because of the coarse GRN generated from GENIE3; limit to small-sized GRNs.
	Regression-based	GENIE3	(Vân Anh Huynh-Thu et al., 2010)	*E. coli*:907*4297	Not require temporal information; explain more complicated underlying GRNs; fast running when using parallel computation.
		SINCERITIES	(Papili Gao et al., 2018)	THP-1:960*45	Low computational complexity; parallel computation is available; the relationship between distributional distance of the regulators and target gene is assumed to be linear; require expression data with temporal information.
	Correlation\ information -based	LEAP	(Specht & Li, 2017)	Dendritic:564*557	Fast and efficient algorithm; identify more interactions; relationships between all genes are assumed to be linear; require expression data with temporal information.
		PIDC	(Chan et al., 2017)	MEP:681*87 Embryonic:3934*20 Hematopoietic:442*46	Not require temporal information; consider more complicated information from data; influenced by the choice of data discretization methods and MI estimators; high computational complexity but could be relieved by Julia.
		Scribe	(Qiu et al., 2020)	*C.* elegans:184442*265	Consider more complicated structure of underlying GRN; assume that underlying processes can be described by a first-order Markov process; high computational complexity; require expression data with temporal information.
	Boolean network	SCNS toolkit	(Moignard et al., 2015)	mESC:3934*42	When applied to different stages of cell population, it can be used to reveal the developmental trajectory of the whole organ from the single-cell level, but the increase of gene number will significantly increase the computation, and it is limited in small-sized GRNs.
scRNA-seq with genome	Motif	SCENIC	(Aibar et al., 2017)	Mouse brain:3005*151 Human neurons:3083*259 Human brain:466*259	GRN can be reconstructed to identify cell states at the same time, which means that it can be applied to data sets with complex cell states.
scRNA-seq with scATAC-seq	SOM	LinkedSOMs	(Jansen et al., 2019)	Mouse pre-B: 128*12380 (scRNA-seq) + 227*25466 (scATAC-seq)	Provides a framework of integration of different type of data; SOM may spend a long time to converge.
	NMF	Coupled NMF	(Duren et al., 2018)	mESCs: 463*21973 (scRNA-seq) + 415*23180 (scATAC-seq)	Provides a framework of integration of different type of data; the expression of a subset of genes is assumed that is can be linearly predicted from the status of chromatin regions; quickly converge but no convergence properties are established.
	CCA	Seurat v3	(Stuart et al., 2019)	Mouse visual cortex: 14249*34617 (scRNA-seq) + 2420*? (scATAC-seq)	Provides a framework of integration of different type of data; the output of method is an integrated expression matrix that could be used in any single-cell GRN inference method.

References

Ackermann, A.M., Wang, Z., Schug, J., Naji, A., and Kaestner, K.H. (2016). Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Molecular metabolism *5*, 233-244.

Aibar, S., González-Blas, C.B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., and van den Oord, J. (2017). SCENIC: single-cell regulatory network inference and clustering. Nature methods *14*, 1083-1086.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike (Springer), pp. 199-213.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S.A., Ponting, C.P., and Voet, T. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature methods *13*, 229-232.

Aubin-Frankowski, P.-C., and Vert, J.-P. (2018). Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. BioRxiv, 464479.

Banks, H.T., and Bihari, K.L. (2001). Modelling and estimating uncertainty in parameter estimation. Inverse Problems *17*, 95.

Bendall, S.C., Davis, K.L., Amir, E.-a.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell *157*, 714-725.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) *57*, 289-300.

Bertsekas, D.P. (2015). Convex optimization algorithms (Athena Scientific Belmont).

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.B. (2017). Julia: A fresh approach to numerical computing. SIAM review *59*, 65-98.

Bianchi, D., Calogero, R., and Tirozzi, B. (2007). Kohonen neural networks and genetic classification. Mathematical and Computer Modelling *45*, 34-60.

Blencowe, M., Arneson, D., Ding, J., Chen, Y.-W., Saleem, Z., and Yang, X. (2019). Network modeling of single-cell omics data: challenges, opportunities, and progresses. Emerging Topics in Life Sciences *3*, 379-398.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning *3*, 1-122.

Breiman, L. (2001). Random forests. Machine learning *45*, 5-32.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). Classification and regression trees (CRC press).

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.

Calderhead, B., and Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics & Data Analysis *53*, 4028-4045.

Chan, T.E., Stumpf, M.P., and Babtie, A.C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell systems *5*, 251-267. e253.

Chen, S., Lake, B.B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature biotechnology *37*, 1452-1457.

Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC bioinformatics *19*, 232.

Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., and Marioni, J.C. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nature communications *9*, 1-9.

Clark, S.J., Smallwood, S.A., Lee, H.J., Krueger, F., Reik, W., and Kelsey, G. (2017). Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). Nature protocols *12*, 534.

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S.W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the national academy of sciences *102*, 7426-7431.

Cover, T.M., and Thomas, J.A. (2012). Elements of information theory (John Wiley & Sons).

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., and Yuan, Y. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352.

de Matos Simoes, R., and Emmert-Streib, F. (2011). Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. PLoS One *6*.

Dey, S.S., Kester, L., Spanjaard, B., Bienko, M., and Van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. Nature biotechnology *33*, 285.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., and Raychowdhury, R. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853-1866. e1817.

Dudoit, S., Fridlyand, J., and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association *97*, 77-87.

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A.T., Chang, H.Y., Wang, Y., and Wong, W.H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proceedings of the National Academy of Sciences *115*, 7723-7728.

Efremova, M., and Teichmann, S. (2020). Computational methods for single-cell omics across modalities. Nature methods *17*, 14.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology *5*.

Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell reports *10*, 1386-1397.

Fiers, M.W., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. Briefings in functional genomics *17*, 246-254.

Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics *21*, 768-769.

Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., and Garry, D.J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. BMC bioinformatics *19*, 220.

Griffiths, J.A., Scialdone, A., and Marioni, J.C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. Molecular systems biology *14*.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome research *23*, 2126-2135.

Haghverdi, L., Lun, A.T., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature biotechnology *36*, 421-427.

Han, L., Wu, H.-J., Zhu, H., Kim, K.-Y., Marjani, S.L., Riester, M., Euskirchen, G., Zi, X., Yang, J., and Han, J. (2017). Bisulfite-independent analysis of CpG island methylation enables genome-scale stratification of single cells. Nucleic acids research *45*, e77-e77.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (Springer Science & Business Media).

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). TIGRESS: trustful inference of gene regulation using stability selection. BMC systems biology *6*, 145.

Hawe, J.S., Theis, F.J., and Heinig, M. (2019). Inferring interaction networks from multi-comics data-a review. Frontiers in genetics *10*, 535.

Hemker, P. (1972). Numerical methods for differential equations in system simulation and in parameter estimation. Analysis and Simulation of biochemical systems *28*, 59-80.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics *12*, 55-67.

Hotelling, H. (1992). Relations between two sets of variates. In Breakthroughs in statistics (Springer), pp. 162-190.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., and Huang,

Y. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell research *26*, 304-319.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., and Fan, G. (2016a). Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome biology *17*, 88.

Hu, Y., Li, C., Meng, K., Qin, J., and Yang, X. (2017). Group sparse optimization via lp, q regularization. The Journal of Machine Learning Research *18*, 960-1011.

Hu, Y., Li, C., and Yang, X. (2016b). On convergence rates of linearized proximal algorithms for convex composite optimization with applications. SIAM Journal on Optimization *26*, 1207-1235.

Jansen, C., Ramirez, R.N., El-Ali, N.C., Gomez-Cabrero, D., Tegner, J., Merkenschlager, M., Conesa, A., and Mortazavi, A. (2019). Building gene regulatory networks from scATAC-seq and scRNA-seq using Linked Self Organizing Maps. PLoS computational biology *15*.

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology *9*, 770-780.

Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nature methods *11*, 740.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological cybernetics *43*, 59-69.

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE *78*, 1464-1480.

Ku, W.L., Nakamura, K., Gao, W., Cui, K., Hu, G., Tang, Q., Ni, B., and Zhao, K. (2019). Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. Nature methods *16*, 323-325.

Lawrance, A. (1976). On conditional and partial correlation. The American Statistician *30*, 146-149.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788-791.

Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. Paper presented at: Advances in neural information processing systems.

Lee Rodgers, J., and Nicewander, W.A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician *42*, 59-66.

Li, W., Calder, R.B., Mar, J.C., and Vijg, J. (2015). Single-cell transcriptogenomics reveals transcriptional exclusion of ENU-mutated alleles. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis *772*, 55-62.

Li, Z., Osborne, M.R., and Prvan, T. (2005). Parameter estimation of ordinary differential equations. IMA Journal of Numerical Analysis *25*, 264-285.

Liang, H., and Wu, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. Journal of the American Statistical Association *103*, 1570-1583.

Liang, J., and Han, J. (2012). Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks. BMC systems biology *6*, 113.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. Neural computation *19*, 2756-2779.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., and Shirley, L.M. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature methods *12*, 519-522.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Aderhold, A., Bonneau, R., and Chen, Y. (2012). Wisdom of crowds for robust gene network inference. Nature methods *9*, 796.

Massey Jr, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association *46*, 68-78.

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

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495.

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *72*, 417-473.

Melton, C., Reuter, J.A., Spacek, D.V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nature genetics *47*, 710.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., and Diamanti, E. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nature biotechnology *33*, 269.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., and Eichler, E.E. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272-276.

Nocedal, J., and Wright, S. (2006). Numerical optimization (Springer Science & Business Media).

Ochs, M.F., and Fertig, E.J. (2012). Matrix factorization for transcriptional regulatory network inference. Paper presented at: 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (IEEE).

Ocone, A., Haghverdi, L., Mueller, N.S., and Theis, F.J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics *31*, i89-i96.

Omranian, N., Eloundou-Mbebi, J.M., Mueller-Roeber, B., and Nikoloski, Z. (2016). Gene regulatory network inference using fused LASSO on multiple data sets. Scientific reports *6*, 20533.

Papili Gao, N., Ud-Dean, S.M., Gandrillon, O., and Gunawan, R. (2018). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics *34*, 258-266.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., and Minkina, A. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Molecular cell *71*, 858-871. e858.

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. Elife *6*, e23203.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods, 1-8.

Qin, J., Hu, Y., Xu, F., Yalamanchili, H.K., and Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. Methods *67*, 294-303.

Qin, J., Yan, B., Hu, Y., Wang, P., and Wang, J. (2016). Applications of integrative OMICs approaches to gene regulation studies. Quantitative Biology *4*, 283-301.

Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2020). Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. Cell Systems.

Rahimzamani, A., and Kannan, S. (2016). Network inference using directed information: The deterministic limit. Paper presented at: 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (IEEE).

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics *16*, 85-97.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nature biotechnology *37*, 547-554.

Specht, A.T., and Li, J. (2017). LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics *33*, 764-766.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888-1902. e1821.

Takahashi, N., Katayama, J., Seki, M., and Takeuchi, J.i. (2018). A unified global convergence analysis of multiplicative update rules for nonnegative matrix factorization. Computational Optimization and Applications *71*, 221-250.

Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T.S., Seidi, A., and Jabbari, J.S. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nature methods *16*, 479-487.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) *58*, 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science, 104-117.

Uurtio, V., Monteiro, J.M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., and Rousu, J. (2017). A tutorial on canonical correlation methods. ACM Computing Surveys (CSUR) *50*, 1-33.

Vân Anh Huynh-Thu, A.I., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PloS one *5*.

Walters-Williams, J., and Li, Y. (2009). Estimation of mutual information: A survey. Paper presented at: International Conference on Rough Sets and Knowledge Technology (Springer).

Wang, P., Qin, J., Qin, Y., Zhu, Y., Wang, L.Y., Li, M.J., Zhang, M.Q., and Wang, J. (2015). ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. Nucleic acids research *43*, W264-W269.

Williams, P.L., and Beer, R.D. (2010). Nonnegative decomposition of multivariate information. arXiv preprint arXiv:10042515.

Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics *10*, 515-534.

Woodhouse, S., Piterman, N., Wintersteiger, C.M., Göttgens, B., and Fisher, J. (2018). SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. BMC systems biology *12*, 59.

Wright, S.J. (2015). Coordinate descent algorithms. Mathematical Programming *151*, 3-34.

Wu, L., Qiu, X., Yuan, Y.-x., and Wu, H. (2019). Parameter estimation and variable selection for big systems of linear ordinary differential equations: A matrix-based approach. Journal of the American Statistical Association *114*, 657-667.

Wu, S., Joseph, A., Hammonds, A.S., Celniker, S.E., Yu, B., and Frise, E. (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proceedings of the National Academy of Sciences *113*, 4290-4295.

Wynn, M.L., Consul, N., Merajver, S.D., and Schnell, S. (2012). Logic-based models in systems biology: a predictive and parameter-free network analysis method. Integrative biology *4*, 1323-1337.

Xue, H., Miao, H., and Wu, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. Annals of statistics *38*, 2351.

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics *32*, 1-8.

Zhang, Z., and Zheng, L. (2015). A mutual information estimator with exponentially decaying bias. Statistical applications in genetics and molecular biology *14*, 243-252.