Direct Reconstruction of Gene Regulatory Networks underlying Cellular state Transitions without Pseudo-time Inference

Ruosi Wan[a,1], Yuhao Zhang[b,1], Yongli Peng[a,1], Feng Tian[b], Ge Gao[b], Fuchou Tang[b], Xiaoliang S. Xie[b,d,2], Jinzhu Jia[c,2], and Hao Ge[a,b,2]

[a]Beijing International Center for Mathematical Research, Peking University, Beijing, 100871, China
[b]Biomedical Pioneering Innovation Center, Peking University, Beijing, 100871, China
[c]School of Public Health and Center for Statistical Science, Peking University, Beijing, 100871, China
[d]Beijing Advanced Innovation Center for Genomics, Peking University, Beijing, 100871, China

May 13, 2021

[1]R.W., Y.Z. and Y.P. contributed equally to this work.
[2]To whom correspondence should be addressed. E-mail: haoge@pku.edu.cn(H.G.) or jzjia@math.pku.edu.cn(J.Z.J.) or sunneyxie@pku.edu.cn(X.S.X.)

Abstract

Nowadays the advanced technology for single-cell transcriptional profiling enables people to routinely generate thousands of single-cell expression data, in which data from different cell states or time points are derived from different samples. Without transferring such time-stamped cross-sectional data into pseudo-time series, we propose COSLIR (COvariance restricted Sparse LInear Regression) for directly reconstructing the gene regulatory networks (GRN) that drives the cell-state transition. The differential gene expression between adjacent cell states is modeled as a linear combination of gene expressions in the previous cell state, and the GRN is reconstructed through solving an optimization problem only based on the first and second moments of the sample distributions. We apply the bootstrap strategy as well as the clip threshold method to increase the precision and stability of the estimation. Simulations indicate the perfect accuracy of COSLIR in the oracle case as well as its good performance and stability in the sample case. We apply COSLIR separately to two cell lineages in a published single-cell qPCR dataset during mouse early embryo development. Nearly half of the inferred gene-gene interactions have already been experimentally reported and some of them were even discovered during the past decade after the dataset was published, indicating the power of COSLIR. Furthermore, COSLIR is also evaluated on several single-cell RNA-seq datasets, and the performance is comparable with other methods relying on the pseudo-time reconstruction.

Single-cell omics measurements is one major breakthrough in experimental bio-technologies during the past decade, which has generated a massive amount of data and provided lots of important insights into biological systems and complex diseases [27, 32, 49, 50, 56, 39, 29, 54]. However, single-cell omic experiments sacrifice the cell in each assay, and thus single cells measured at different time points or stages of development have to come from different batches of cells, which is independent of each other. So until now it can only be able to produce time-stamped cross-sectional (TSCS) data rather than longitudinal time series data containing the real-time information [25], as shown in Fig. 1A.

Although reconstructing temporal information from single-cell transcriptomic measurements has become an emerging field [44], the methods for gene regulatory network (GRNs) inference based on the construction of pseudo-

time series has recently been shown to be sensitive to the accuracy of pseudo-time series construction, making them less stable [40]. Many other approaches without pseudo-time construction have also been proposed to infer GRNs from TSCS single-cell expression data [30, 33, 1, 10, 28, 26, 36]. However, they are mostly only confined within one single cell stage or time point, the network inferred from which is the mechanism for sustaining the current cell stage rather than the mechanism responsible for driving the cell-state transition along cell lineage.

Moreover, the potential upstream regulatory genes picked up through differential expressed gene (DEG) analysis is typically a lot, and the GRN responsible for maintaining a single cell state should also be dense due to biological complexity. However, the GRN driving cell-state transition is believed to be sparse [18]. Once we can identify the much fewer potential upstream regulatory genes and possible regulatory relationships among genes, further biological functional analysis is easy to be performed.

Hence in this paper, we directly model the GRN that is directly responsible for the temporal evolution of gene expression, and propose a novel optimization method called Covariance Restricted Sparse Linear Regression (COSLIR), which only requires the first and second moments of the measured samples. The output of COSLIR is a directed GRN with signs and weights on each gene-gene interaction. We use the alternative direction method of multipliers algorithm(ADMM, [6]) to solve this optimization problem, and apply the bootstrapping([37]) and clip thresholding techniques for selecting significant gene-gene interactions to improve the precision and stability of the estimator. Published single-cell qPCR and RNA-seq gene-expression datasets during mouse and human early embryo development are used to evaluate COSLIR. The performance of COSLIR is comparable to previous methods using pseudo-time construction, but with fewer assumptions and requirements.

Our approach: Covariance restricted sparse linear regression (COSLIR)

Model

Let X_t and X_{t+1} be two p-dimensional vectors representing the expression values of p genes of a same single cell at stages t and $t+1$ respectively. In

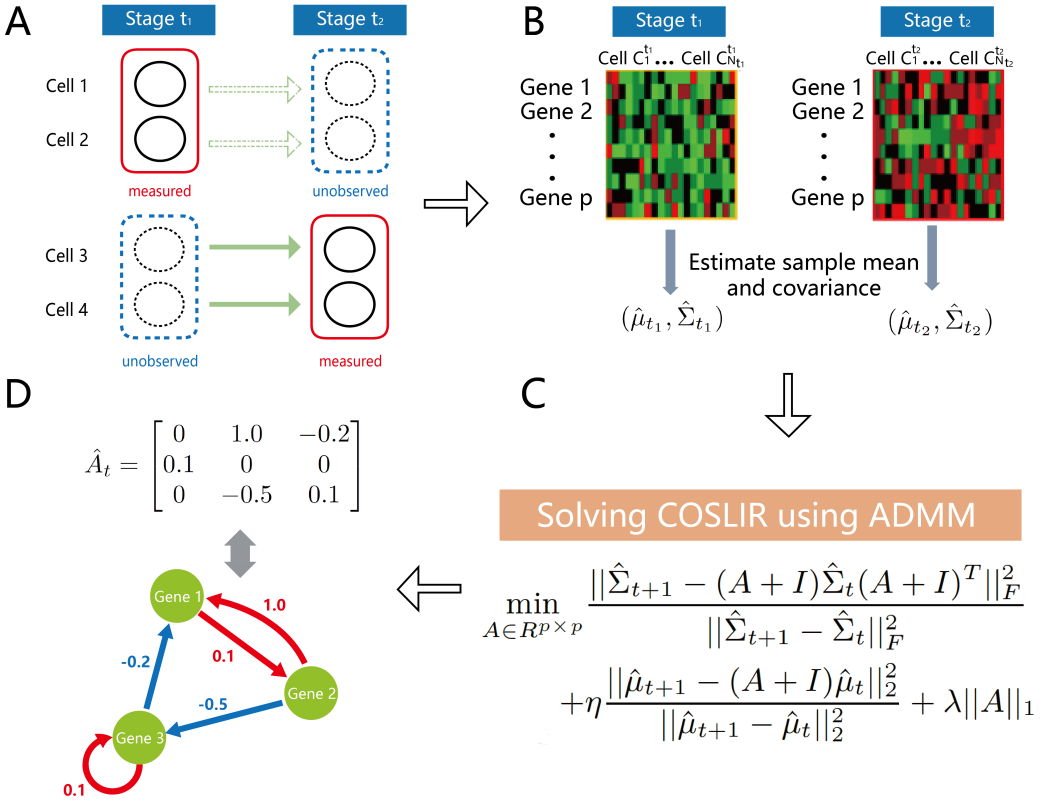

Figure 1: Overview of the COSLIR method for GRN reconstruction using time-stamped cross-sectional single-cell expression data. (A) The illustration of time-stamped cross-sectional data generated in single-cell experiments. Only the data in solid red rectangles has been measured. (B) Estimation of sample mean and co-variance matrix. (C) Solving the optimization problem in COSLIR through ADMM. (D) Inferred gene regulatory network among genes represented by the estimator $\hat{A}$.

order to model the dynamic evolution from X_t to X_{t+1}, we introduce a p by p matrix A_t, the element $(A_t)_{ij}$ at the i-th row and j-th column of which represents the regulatory strength from gene j to gene i.

For the purpose of simplicity, we propose the following linear regression model in which the dependent variable is the difference between X_{t+1} and X_t, and the explanatory variable is X_t:

$$X_{t+1} - X_t = A_t X_t + l_t + \epsilon_t, \tag{1}$$

where the noise term ϵ_t is a p-dimensional random vector independent of X_t, and l_t is certain possible external perturbation.

However, as we demonstrated in introduction, once we measure X_t or X_{t+1}, the other one is missing. There is no correspondence between the cross-sectional single cell data collected at stages t and $t+1$. Even if the sample size tends to infinity, what we can finally obtained is only the exact distributions of X_t and X_{t+1}, still not sufficient to uniquely determine A_t. To overcome such difficulties, we further assume A_t is sparse, and l_t as well as ϵ_t is rather small. It's a commonly used assumption in the field of statistical learning [53, 9, 16], which is consistent with discoveries in single-cell biology [18]. Then we could obtain the sparse matrix A_t by solving the following non-convex optimization problem (Fig. 1C, Supplementary Information)

$$\begin{aligned}\min_{A\in R^{p\times p}} & \frac{||\hat{\Sigma}_{t+1}-(A+I)\hat{\Sigma}_t(A+I)^T||_F^2}{||\hat{\Sigma}_{t+1}-\hat{\Sigma}_t||_F^2} \\ & +\eta\frac{||\hat{\mu}_{t+1}-(A+I)\hat{\mu}_t||_2^2}{||\hat{\mu}_{t+1}-\hat{\mu}_t||_2^2}+\lambda||A||_1,\end{aligned} \tag{2}$$

where $\hat{\Sigma}_\cdot$ and $\hat{\mu}_\cdot$ are the estimators of co-variance and mean of $X_\cdot$ (Fig. 1B). $||\cdot||_F$ denotes the Frobenius norm, and $||\cdot||_1$ denotes the sum of the absolute value of all the elements in a matrix. There are two tuning parameters η and λ. The first two terms in (2) control the quantitative relations between the mean and co-variance of X_t and X_{t+1} according to (1), while λ controls the sparsity of A_t. We used an efficient numerical algorithm, Alternating Direction Method of Multipliers (ADMM [6]), to solve (2). After solving the optimization problem, the non-zero elements in A_t constitute the regulatory relations among genes (Fig. 1D). See Materials and Methods for more details.

It is worth mentioning that COSLIR should be applied to each pair of adjacent cell stages or time points, not requiring the GRN to be invariant with time t.

Estimation of mean and co-variance

In oracle cases, i.e. if we know the true mean $\mu_\cdot$ and true co-variance $\Sigma_\cdot$ of $X_\cdot$, the estimator of A_t could be obtained by directly solving (2). However, the true mean and co-variance is usually unknown, thus we need to estimate

the mean and co-variance of X_t in advance (Fig. 1B). More accurate estimators of mean and co-variance certainly will lead to a better estimator of the interacting matrix A_t. The naive sample mean and co-variance estimators are recommended when the sample size is large enough. But when sample size is small compared with the number of components, sample mean and co-variance may contain large variance and result in quite inaccurate estimator of A_t, thus other high dimensional techniques [4, 7, 31, 15] should be applied to obtain more accurate estimations of mean and co-variance. Also for single-cell RNA-seq data, one may first apply the imputation techniques to address the technical noise such as dropout or batch effect [52], before estimating the mean and covariance.

Bootstrapping

With only sample data in hand, we have to apply the estimated co-variance and mean of X_t and X_{t+1} in (2), which probably will lead to the inaccurate estimation of A_t. Thus we apply the non-parametric bootstrapping technique [14] to increase the robustness of estimator as well as the precision, which is more interested to experimentalists [40]. We repeat the following procedure multiple times and take an ensemble of all the estimators for A_t we got together: first perform randomly sampling with replacements from the two collections of samples at different stages or time points, and constitute two new collections of observations; then apply COSLIR to the new collections of the samples to obtain a new estimator of A_t. Finally only keep those non-zero elements whose confidence (repetition ratio) is above certain threshold, into the final estimator of the interacting matrix A_t.

Simulation Study

We first evaluate the performance of COSLIR in a simulation study, in the oracle case where the true mean and co-variance matrix are known as well as the sample case with only random samples in hand.

In oracle cases we have evaluated the criteria we proposed for model selection (Materials and Methods), i.e. determining the two tuning parameters λ and η in (2). The results are quite robust with respect to the value of η, and our criteria can usually help to find the optimal or sub-optimal value of λ (see Supplementary Information).

In oracle cases, the recovery of the estimator obtained through COSLIR is almost exact (Fig.2A), even when the dimension of the data is up to 500 and the number of gene-gene interactions in A_t that should be inferred is as high as 2.5×10^5. Not only the precision (the fraction of inferred interactions that are correct) and recall (the fraction of correct interactions that are inferred) are nearly 100%, the exact values of the estimator are nearly identical to the correct ones that we generated (See Supplementary Information). This gives us the confidence to apply this method to the simulated sample data, in which all of Σ_t, Σ_{t+1}, μ_t and μ_{t+1} should be estimated from data separately at first.

Fig. 2B-D shows how the performance of COSLIR varies as a function of the number of genes, the number of cells and the threshold of confidence in the sample cases. Even when the number of genes are high, high degree of precision can still be achieved, as long as the threshold of confidence from bootstrapping is set to be high (Fig. 2B, 2D). As a price, the recall will decrease much as compared to the low gene-number case, which is already lower than the oracle case (Fig. 2B). However, noticing the number of gene-gene interactions in the matrix A_t that need to be inferred is very high(square of dimension), the number of successfully recovered interactions among genes is not small at all. The time to run COSLIR once is less than 1 minutes for dimension 100 and about 40 minutes for dimension 500 in a typical personal computer. Bootstrapping procedure typically needs to run about 50 times. It should be much faster to run COSLIR parallel on computer clusters.

Sample size is another important factor to determine the performance of COSLIR. When the sample size increases, so do the precision and recall (Fig. 2C), approaching the performance of the oracle case. There is also a trade off between precision and recall when tuning the threshold of confidence in the bootstrapping procedure (Fig. 2D). It can be determined case by case in real applications, but the rule of thumb here is to choose a high threshold of confidence if one cares more about precision such as in most studies of experimental sciences. Furthermore, the more sparse the true matrix A_t is, the better the performance is and the less samples one needs (Supplementary Information).

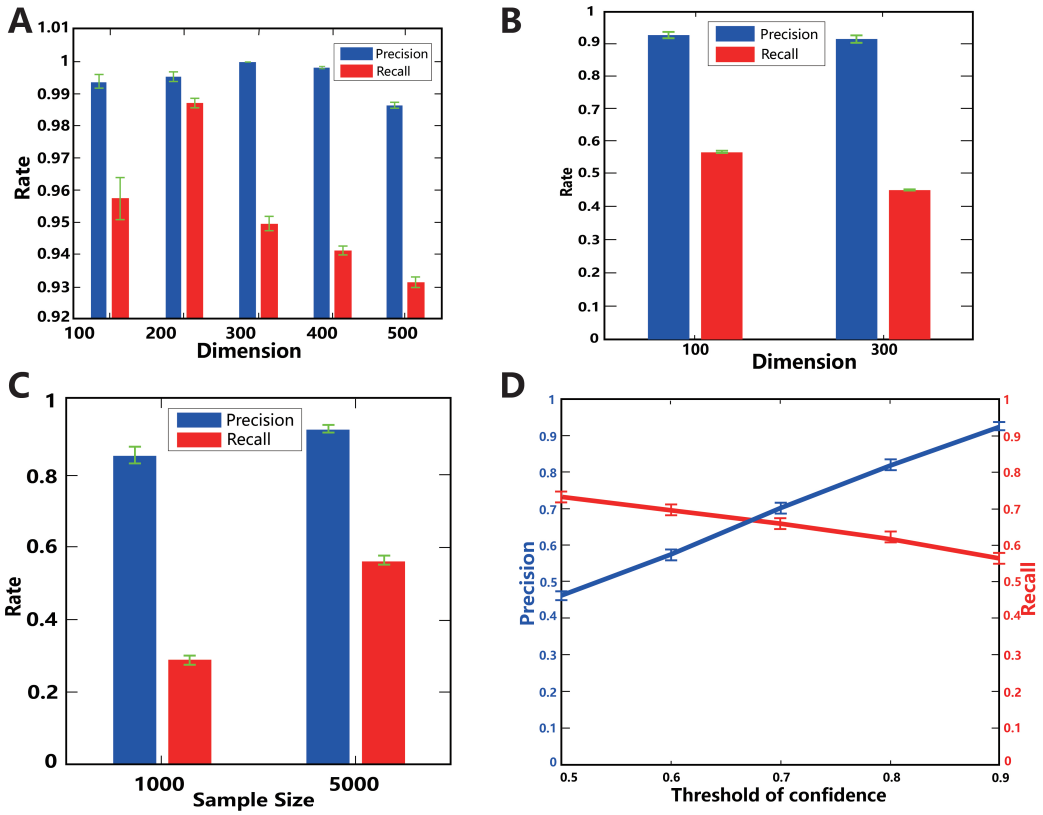

Figure 2: Performance of COSLIR validated through a simulation study, both in oracle cases where the true mean and co-variance matrix are known (A), and in sample cases where only random samples are available (B-D). Detailed simulation settings is in Materials and Methods as well as Supplementary Information. The precision and recall vary with the number of genes(A,B), the number of cells (C) and the threshold of confidence(D). The number of cells are 5000 in (B) and (D). The number of genes is 100 in (C) and (D). The threshold of confidence is 0.9 in (B) and (C). Clip threshold is always 0.01.

Results on experimental TSCS Single-Cell datasets during Early Embryo Development

Uncovering the gene expression patterns in early embryos is crucial for understanding cellular developmental processes [21, 45, 38, 11, 48]. However, due

to the limitation of real-time experimental measurements, people are still not clear about the detailed regulatory mechanism driving this very beginning process of life. We are going to use COSLIR to infer the gene regulatory networks responsible for the control of the cell fate decisions.

Single-cell qPCR dataset

We first analyzed a set of published single-cell gene expression data during mouse embryo development obtained using the qPCR technique[21].

In [21], 442 single cells were selectively collected from early mouse embryonic development, containing 7 developmental stages: Zygote, 2-cell stage, 4-cell stage, 8-cell stage, 16-cell stage, Morula stage and Blastocyst stage. Blastocyst embryos are found to be made up of 3 different types of cells, i.e. trophectoderm(TE), primitive endoderm(PE), epiblast(EPI) which have diverse gene markers and expression patterns[43].

We apply the nonlinear dimension reduction method ISOmap [51] to the blastocyst data starting from the 8 cell stage (Fig. 3B), instead of the linear dimension reduction method PCA used in [21]. It indicates that there are two cell-fate decisions in this data set: (1) the setting apart of inner and outer cells at the 16-cell stage and consequently forming ICM and TE, and (2) the subsequent formation of primitive endoderm and epiblast from ICM.

Since during the first cell-fate decision, maternal gene degradation still dominates the variation of gene expression, which violate the requirement of COSLIR method, here we are only able to analyze the second cell-fate decision applying COSLIR.

Fig. 3C-D gives an overview of the two inferred GRNs using Cytoscape [47]. We have inferred a certain amount of directed gene regulatory relationships using COSLIR, i.e. the positive or negative elements in those matrices A_t's, indicating the activated or inhibited regulatory relations between genes. The inferred upstream regulatory genes and main regulatory relationships are consistent very well with the existing knowledge about the second cell-fate decision during early embryo development [17, 35, 20, 45]. $Sall4$, $Sox2$, $Pou5f1$, $Gata6$, $Tcfap2c$ and $Pdgfra$ serve as the main upstream regulators of the inferred GRN driving the ICM cells towards the EPI fate, the main task of which is to activate the EPI markers and inhibit the PE markers (Fig. 3C). Similarly, $Sall4$, $Sox2$, $Pou5f1$, $Gata4$, $Tcfap2c$ and $Nanog$ serve as the main upstream regulators of the inferred network driving the ICM cells towards the PE fate, the main task of which is to inhibit the

EPI markers and activate the PE markers. All of these inferred upstream regulatory genes as well as their targeted markers are well known for their functions during early embryo development [55, 21].

Furthermore, through searching the literature and databases(ChIP-atlas(ESC), BioGRID and TRRUST), we found out that nearly half of the inferred regulatory relationships have already been experimentally reported, some of which were uncovered many years after the used data set from [21] was published (Figs. S5-S6 in the Supplementary). Many of them have even been validated by more than one database.

The inferred regulatory relationships that have not been included in the three databases may still be correct predictions. For instance, the regulatory relation from $Gata3$ to $Gata6$ has just recently been reported in [24], which is not contained in the three databases but predicted by COSLIR.

Notice that we only used the data sets of ICM and EPI or ICM and PE to infer the gene regulatory networks separately. For example, even though the fold change of $Gata6$ from the ICM stage to the PE stage is only 1.03, it is known to be an important marker gene of the PE stage compared with its expression in the EPI stage. Now it is correctly inferred by COSLIR, even without referring to the gene expression data in the EPI stage, indicating the power of COSLIR.

Single-cell RNA-seq datasets

We compare COSLIR with three existing regression-based GRN reconstruction algorithms, SCODE, SINCERITIES, SINGE [41]. Two experimental scRNA-seq datasets are analyzed, one in human cells (hESC, [12]) and one in mouse cells(mESC, [23]). These datasets contain multiple time points. Hence we reconstruct the GRNs underlying each pair of adjacent time points. Functional interaction networks (STRING) and cell-type-specific ChIP-Seq data [12, 23] are used as the ground-truth networks for evaluation.

We utilized a slightly modified early precision rate (EPR) from BEELINE [41] to assess these algorithms. A weighted network is generated by each algorithm. EPR is just the fraction of true positives in the most significant k edges. We chose these k's in Fig. 4 when the averaged EPR as a function of k obtained from the GRN generated by COSLIR has attained a stable value (See Supplementary Information). In Fig. 4, the differential EPRs among algorithms with respect to COSLIR across different datasets are shown.

Among these totally 18 times of comparisons, COSLIR is much better

than all the other algorithms in 7 times, and ranks the second in 8 times. Therefore we can come to the conclusion that COSLIR is at least comparable with these existing methods, without the step of pseudo-time reconstruction.

Discussion

In this era, the intersection between machine learning and network biology is offering invaluable opportunities and challenges, including gene regulatory network inference [8]. In this paper, we proposed an optimization model COSLIR for the inference of gene regulatory network, with only the time-stamped cross-sectional single-cell expression data in hand. Our approach minimizes the requirements for the inputs, i.e. using just the estimated mean and co-variances of the samples, even without the construction of pesudotime trajectories, but outputs directed GRNs with weights and sign. Our model does not assume any specific category of sample distributions, broadly enhancing its applicability. In the simulation study, we showed that COSLIR is able to exactly recover the true network in the oracle cases, while its performance is still quite good in the sample cases, especially the precision with the help of bootstrapping. Moreover, in the real-data analysis, COSLIR is applied to the single-cell qPCR and RNA-seq datasets. COSLIR is able to recovers crucial upstream regulatory genes as well as gene-gene interactions during early mouse and human embryonic development. It implies that the GRN information has long been hidden in the TSCS data itself, even in the absence of real time-series data.

In real applications, several other issues will influence the performance of COSLIR. For example, if the data is multi-scaled across the genes, certain normalization is necessary before applying COSLIR. We recommend using correlation matrix instead of co-variance matrix, though it may cause some bias. Gene selection is also a problem. Similar to many other GRN reconstruction methods, the computation time of COSLIR would be significant if the number of genes exceeds 1000 [40]. Typical strategy for selecting genes is choosing the highly varying ones together with transcriptional factors. Better and automatic strategy is still in demand.

Another important issue is whether the data strictly follows the linear model (1). Although linear regression model is the most commonly used model in GRN reconstruction using single-cell expression data [19, 34, 13, 46, 2], the information on linearity is actually missing in the time-stamped cross-

sectional data. However, in statistical learning, we always chose the linear model except there is strong evidence that it is highly nonlinear, because empirically, as long as the data does not highly deviate from the linearity assumption, the performance of statistical inference by the linear model is still pretty good [22]. Also linear model is much easier to be interpreted. People believe that the simpler is often better.

The reconstruction method of GRN can be further combined with existing prior knowledge such as additional databases or supervision [1] to increase the accuracy. And very recently, it was shown that RNA velocity analysis might restore certain amount of temporal information [42]. Hence combining COSLIR with existing database knowledge and bioinformatic methods will be our future research directions.

Materials and Methods

Derivation of the Model

Denote the mean and co-variance of the p-dimensional random vector X_t as μ_t and Σ_t. Note that X_t doesn't have to be drawn from normal distribution. ϵ_t is a noise term with mean l_t and co-variance D_t which is small compared with Σ_t, then the following equations could be derived from (1):

$$\begin{aligned} \Sigma_{t+1} &= (A_t + I)\Sigma_t(A_t + I)^T + D_t; \\ \mu_{t+1} &= (A_t + I)\mu_t + l_t, \end{aligned} \tag{3}$$

in which I is the identity matrix.

It is under-determined as an equation of A_t, i.e. there are infinite number of solutions of A_t, even if $\Sigma_\cdot, \mu_\cdot, l_\cdot, D_\cdot$ have all been known. Hence additional assumption should be added to reduce the number of variables. One popular approach is to assume the sparsity of parameters [53, 9]. Hence here we assume A_t is sparse, i.e. only a few entries of A_t are non-zero, then we propose an estimator for A_t following the idea of compressed sensing with nonlinear observations [5]

$$\begin{aligned} & \min_{A \in R^{p \times p}} ||A||_1 \\ s.t. \quad & \Sigma_{t+1} - (A + I)\Sigma_t(A + I)^T = D_t \\ & \mu_{t+1} - (A + I)\mu_t = l_t. \end{aligned} \tag{4}$$

In practice, the exact co-variance $\Sigma_\cdot$ and mean $\mu_\cdot$ can be replaced by their estimators $\hat{\Sigma}_\cdot$ and $\hat{\mu}_\cdot$ from samples. Also the noise mean l_t and co-variance D_t are usually unknown in real data analysis, thus applying the penalty method as well as the idea of least square estimation, the optimization problem (4) could be turned into the unconstrained minimization problem (2). (2) is a non-convex approximation problem, hence we used the ADMM method to solve it (See Supporting Information), which may have many local optimum. We found out that setting zero matrix as the start point in the ADMM algorithm can always obtain a reasonable solution.

Co-variance matrix estimation

We first obtain the naive co-variance matrix estimator $\hat{\Sigma}$, and then corrected it by

$$\tilde{\Sigma} = (1-\alpha)\hat{\Sigma} + \alpha I, \tag{5}$$

where I is an identity matrix, $\alpha = 0.01$, to ensure positive definiteness.

Model selection and evaluation

Notice there are two hyper-parameters needed to be set in COSLIR: λ to controlling the sparsity of A_t and η controlling the noise term in (3). Our model is an unsupervised-learning model due to the lack of simultaneous measured X_t and X_{t+1}, hence cross-validation does not work. To overcome such a problem, empirically, we used three indexes of criterion to help select the model:

$$\begin{aligned} e_\Sigma(\hat{A}) &= ||\hat{\Sigma}_{t+1} - (\hat{A}_t + I)\hat{\Sigma}_t(\hat{A}+I)^T||_F / ||\hat{\Sigma}_{t+1} - \hat{\Sigma}_t||_F^2 \\ e_\mu(\hat{A}) &= ||\hat{\mu}_{t+1} - (\hat{A}_t + I)\hat{\mu}_t||_2 / ||\hat{\mu}_{t+1} - \hat{\mu}_t||_2^2 \\ s_0(\hat{A}) &= \frac{\#\{\text{non-zero elements of } \hat{A}\}}{p^2}. \end{aligned} \tag{6}$$

e_Σ, e_μ measure the small noise terms in (3), while $s_0(\hat{A})$ measures the sparsity of $\hat{A}_t$. We chose the model with all three indexes small, namely, choosing the sparsest matrix A_t with e_Σ, e_μ below a reasonable level. Our simulation study illustrates the effectiveness of this criterion we proposed(See Supplementary Information for more details). Also to make things easier, during bootstrapping, we first determine the hyper-parameters λ and η using the

whole sample, and then just use these determined values for the subsequent analysis.

In real applications, figuring out the positions of nonzero elements and their sign(positive or negative) in A_t is enough, and we call such elements as *properly recovered*. Therefore we used the commonly used criterion **precision** and **recall** in machine learning to evaluate the performance in the simulation study, i.e.

$$precision = \frac{\#\{\text{properly recovered nonzero elements}\}}{\#\{\text{nonzero elements in } \hat{A}_t\}}; \tag{7}$$

$$recall = \frac{\#\{\text{properly recovered nonzero elements}\}}{\#\{\text{nonzero elements in true } A_t\}}. \tag{8}$$

Clip thresholding

In practice, we always need a clip thresholding procedure on the output of the numerical algorithm, to eliminate those non-zero entries with very small absolute values below certain threshold. In the oracle case, it is quite simple, since there is always a huge gap among the values of non-zero elements. However, in the sample case, it is not that clear which value of clip threshold we should chose. We suggest to determine the clip threshold based on the trade off among the three indexes of criterion in (6), i.e let $\hat{A}_t(\varepsilon)$ be the adjusted estimator after taking ε as the clip threshold, then we determine the clip threshold ε by choosing the sparsest $\hat{A}_t(\varepsilon)$ with $e_\Sigma(\hat{A}_t(\varepsilon))+\eta e_\mu(\hat{A}_t(\varepsilon))$ below certain reasonable level. Similar to the hyper-parameters λ and η, the clip threshold is also determined in advance before bootstrapping using the whole sample.

Confidence of bootstrapping

After clip thresholding, for each matrix element in A_t, if the ratio of non-zero estimators with the same sign(confidence) during the bootstrapping procedure is greater than a given threshold, then we regard the estimated sign of this element as statistically significant, and the mean value of these estimators generated through bootstrapping makes the final estimator.

Simulation settings in Fig. 2

In the oracle case, the co-variance matrix Σ_t is generated by the formula $\Sigma_t = P\Lambda P^T$ in which $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix with $\Lambda_{ii} = \exp\left(\frac{i}{p}\right), i = 1, 2, ...p$ and $P = I + R$ where I is the identity matrix and the elements of R are all sampled from independent standard normal distributions. The true interacting matrix $A_t \in \mathbb{R}^{p\times p}$ is a sparse matrix with only 10% elements are non-zero and randomly generated from $\mathcal{N}(0, 1)$ independently. The elements of the $\mu_t \in \mathbb{R}^p$ are all generated from $\mathcal{N}(0, 100)$ independently. The mean of each element in the noise term $\epsilon_t \in \mathbb{R}^p$ is set to be 0.1 and the co-variance matrix D_t of ϵ_t is set to be a diagonal matrix with $(D_t)_{ii} = 0.01$ for each i. Then μ_{t+1} and Σ_{t+1} are calculated through Eq. 3. Given the mean and covariance matrix, the sample data is generated using normal distribution, with only one of X_t and X_{t+1} available.

Numerical experiments have been repeated for 100 times in the oracle case and 50 times in the sample case. The bootstrapping procedure is repeated for 50 times in each sample experiment. The determined values of η, λ and clip threshold are summarized in Supplementary Information.

Implementation in real-data analysis

The qPCR dataset from [21] contains mRNA expression levels of 48 genes (including 27 transcriptional factors, 19 known marker genes and 2 housekeeping genes for normalization) in over 500 individual cells, during the first two cell fate decisions of the early mouse embryo. We only use the data of the 46 non-housekeeping genes to do the analysis. We re-scaled the raw data by log transformation:

$$\tilde{x} = \log(x + 1). \tag{9}$$

The bootstrapping procedure is repeated for 500 times.

The genes in the RNA-seq datasets are selected using the strategy in BEELINE [3]. The mESC datasets using STRING as the ground-truth network contains 646 genes, with 90 (0h), 68 (12h), 90 (24h), 82 (48h) and 91 (72h) cells during 5 time points; the mESC datasets using cell-type-specific ChIP-Seq data as the ground-truth network contains 970 genes with the same number of cells during the 5 time points. The hESC datasets using STRING contains 517 genes with 92 (0h), 102 (12h), 66 (24h), 172 (36h), 138 (72h) and 188 (96h) cells during 6 time points; the hESC datasets using cell-type-

specific ChIP-Seq contains 814 genes with the same number of cells during the 6 time points. Note the data from [3] has already been normalized so we just use it after selecting genes. The bootstrapping procedure in COSLIR is repeated for 50 times.

Re-scaled values for selecting most influential regulatory interactions

In single-cell expression data analysis, people may be more interested in those interactions contributing more to the changes in the mean expression, i.e. DEGs. Therefore, we define a rescaled value

$$(\tilde{A})_{ij} = \frac{(|\hat{A}|)_{ij}|\hat{\mu}_1|_j}{|(\hat{\mu}_1)_i - (\hat{\mu}_2)_i|}, i, j = 1, 2, ...p,$$

for selecting the most influential regulatory gene-gene interactions. We only perform this technique on the final estimator obtained after bootstrapping.

Software and data

Code, simulation data, and single cell gene expression data are available at https://github.com/Ge-lab-pku/COSLIR.

Author Contribution

H.G. designed research; X.S.X., J.J. and H.G. supervised research; R.W., Y.Z., Y.P., J.J. and H.G. performed research. R.W., Y.Z. and Y.P. analyzed the single-cell data with the help of F.Tian, G.G. and F.Tang; W.R., Y.Z., Y.P., X.S.X., J.J. and H.G. wrote the paper.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank Yunuo Mao, Xiaojie Qiu, Yu Xue, Zaiwen Wen, Wei Lin and Ruibin Xi for helpful discussions.

References

[1] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

[2] P.C. Aubin-Frankowski and J.P. Vert. Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference. *Preprint at https://doi.org/10.1101/464479*, 2020+.

[3] Mikhail Belkin and Partha Niyogi. Using manifold stucture for partially labeled classification. In *Advances in neural information processing systems*, pages 929–936, 2002.

[4] Peter J Bickel, Elizaveta Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[5] T. Blumensath. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 59:3466–3474, 2013.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[7] Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[8] Diogo M. Camacho, Katherine M. Collins, Rani K. Powers, James C. Costello, and James J. Collins. Next-generation machine learning for biological networks. *Cell*, 173:1581–1592, 2018.

[9] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[10] T.E. Chan, M.P.H. Stumpf, and A.C. Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems*, 5:251C267, 2017.

[11] C. Chazaud and Y. Yamanaka. Lineage specification in the mouse preimplantation embryo. *Development*, 143:1063–1074, 2016.

[12] Li Fang Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17(1):173, 2016.

[13] A. Deshpande, L.F. Chu, R. Stewart, and A. Gitter. Network inference with granger causality ensembles on single-cell transcriptomic data. *Preprint at https://doi.org/10.1101/534834*, 2020+.

[14] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[15] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[17] T. Frum and A. Ralston. Cell signaling and transcription factors regulating cell fate during formation of the mouse blastocyst. *Trends in Genetics*, 31:402–410, 2015.

[18] L.A. Furchtgott, Melton S., V. Menon, and S. Ramanathan. Discovering sparse transcription factor codes for cell states and state transitions during development. *Elife*, 6:e20488, 2017.

[19] N.P. Gao, S.M. Minhaz Ud-Dean, O. Gandrillon, and R. Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34:258C266, 2018.

[20] M. Goolam, A. Scialdone, S.J.L. Graham, I.C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J.C. Marioni, and M. Zernicka-Goetz. Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*, 165:61–74, 2016.

[21] G. Guo, M. Huss, G.Q. Tong, C. Wang, L.L. Sun, N.D. Clarke, and P. Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, 18(4):675–685, 2010.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[23] T. Hayashi, H. Ozaki, Y. Sasagawa, M. Umeda, H. Danno, and I. Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature Communications*, 9:90, 2018.

[24] P. Home, R.P. Kumar, A. Ganguly, B. Saha, J. Milano-Foster, B. Bhattacharya, S. Ray, S. Gunewardena, A. Paul, S.A. Camper, P.E. Fields, and S. Paul. Genetic redundancy of gata factors in the extraembryonic trophoblast lineage ensures the progression of preimplantation and postimplantation mammalian development. *Development*, 144:876–888, 2017.

[25] W. Huang, X.Y. Cao, F. H. Biase, P.F. Yu, and S. Zhong. Time-variant clustering model for understanding cell fate decisions. *Proc. Nat. Acad. Sci.*, 111:E4797–E4806, 2014.

[26] V.A. Huynh-Thu, A. Irrthum, L. Wehenke, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *Plos One*, 5:e12776, 2010.

[27] J.P. Junker and A. van Oudenaarden. Every cell is special: Genome-wide studies add a new dimension to single-cell biology. *Cell*, 157:8–11, 2014.

[28] S. Kim. ppcor: An r package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22:665C674, 2015.

[29] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[30] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[31] Han Liu, Lie Wang, and Tuo Zhao. Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, 23(2):439–459, 2014.

[32] W. Lu and F.C. Tang. Single-cell sequencing in stem cell biology. *Genome Biology*, 17:71, 2016.

[33] D. Marbach, J.C. Costello, R. Kffner, R.J. Vega, N.M. andPrill, D.M. Camacho, K.R. Allison, The DREAM5 Consortium, M. Kellis, J.J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796C804, 2012.

[34] H. Matsumoto, H. Kiryu, C. Furusawa, M.S.H. Ko, S.B.H. Ko, N. Gouda, T. Hayashi, and I. Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33:2314C2321, 2017.

[35] S. Menchero, T. Rayon, M.J. Andreu, and M. Manzanares. Signaling pathways in mammalian preimplantation development: Linking cellular phenotypes to lineage decisions. *Dev. Dyn.*, 246:245–261, 2017.

[36] T. Moerman, S.A. Santos, C.B. Gonzalez-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35:2159C2161, 2019.

[37] Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 94-95. Sage, 1993.

[38] Y. Nakai-Futatsugi and H. Niwa. Epiblast and primitive endoderm differentiation: Fragile specification ensures stable commitment. *Stem Cell*, 16:346–347, 2015.

[39] Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35, 2018.

[40] A. Pratapa, A.P. Jalihal, J.N. Law, A. Bharadwaj, and T.M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17:147C154, 2020.

[41] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(5):147–154, 2020.

[42] X.J. Qiu, A. Rahimzamani, L. Wang, Q. Mao, T. Durham, J.L. McFaline-Figueroa, L. Saunders, C. Trapnell, and S. Kannan. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Systems*, page In press, 2020.

[43] J. Rossant and P.P. Tam. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development*, 136:701–713, 2009.

[44] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 39:547C554, 2019.

[45] N. Saiz and B. Plusa. Early cell fate decisions in the mouse embryo. *Reproduction*, 145:R65–80, 2013.

[46] M. Sanchez-Castillo1, D. Blanco1, I.M. Tienda-Luna, M. C. Carrion, and Y. Huang. A bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*, 34:964–970, 2018.

[47] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–504, 2003.

[48] B. Sozen, A. Can, and N. Demir. Cell fate regulation during preimplantation development: A view of adhesion-linked molecular interactions. *Developmental Biology*, 395:73–83, 2014.

[49] Oliver Stegle, Sarah Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews in Genetics*, 16:133–145, 2015.

[50] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865, 2017.

[51] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[52] L. Tian, X. Dong, S. Freytag, K. L Cao, S. Su, A. JalalAbadi, D. Amann-Zalcenstein, T.S. Weber, A. Seidi, J.S. Jabbari, S.H. Naik, and M.E. Ritchie. Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16:479C487, 2020.

[53] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[54] Betsabeh Khoramian Tusi, Samuel L Wolock, Caleb Weinreb, Yung Hwang, Daniel Hidalgo, Rapolas Zilionis, Ari Waisman, Jun R Huh, Allon M Klein, and Merav Socolovsky. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, 555(7694):54–60, 2018.

[55] M. Zernicka-Goetz, S.A. Morris, and A.W. Bruce. Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. *Nat Rev Genetics*, 10:467–477, 2009.

[56] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

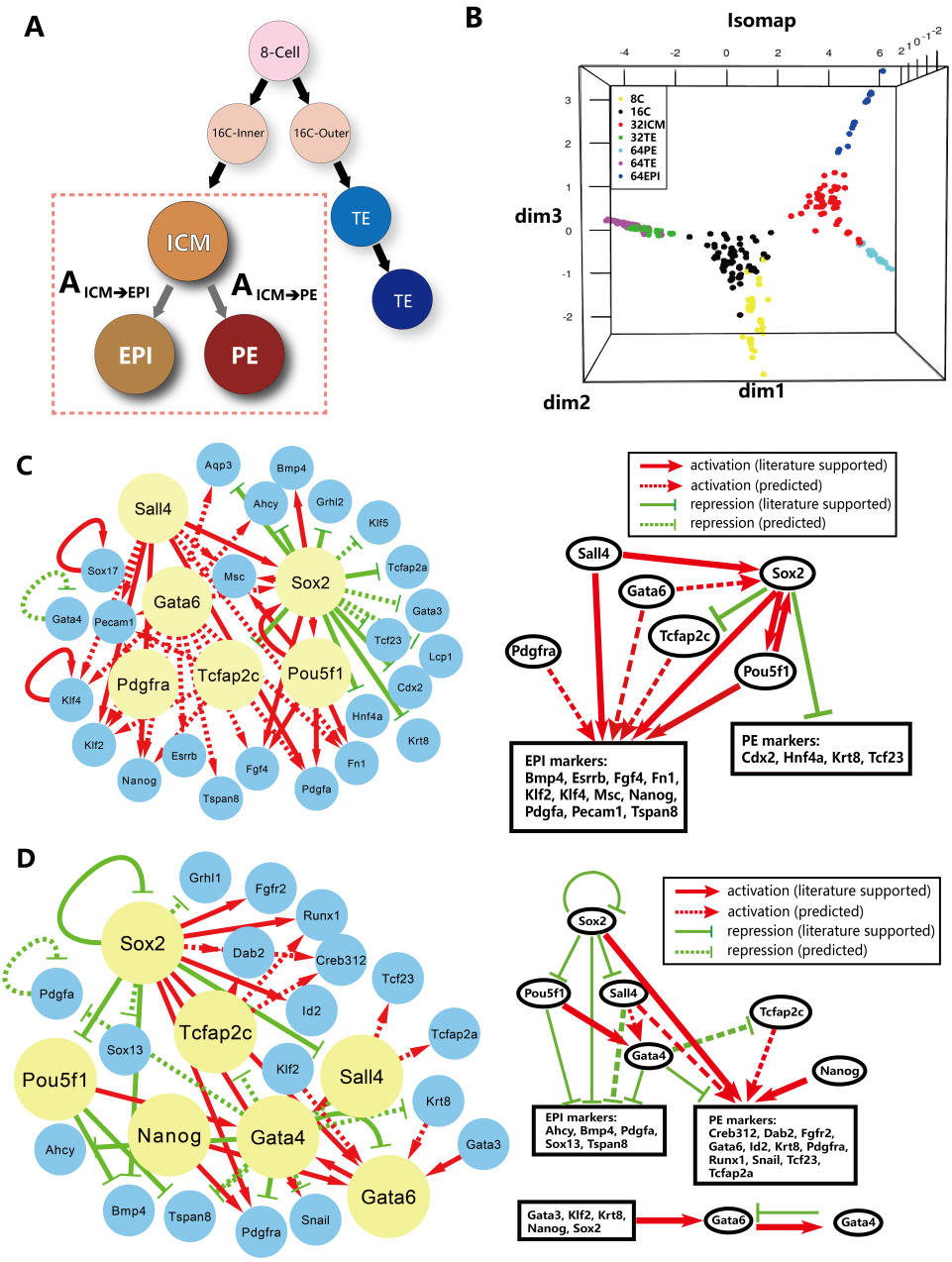

Figure 3: Overview of the real single-cell RT-PCR expression data analysis applying COSLIR. (A) The developmental lineage tree for mouse early embryo development. (B) Data illustration using Isomap. (C) Inferred GRNs driving the ICM cells towards the fate of EPI, with its sketch map. (D) Inferred GRNs driving the ICM cells towards the fate of PE, with its sketch map. Here clip threshold is 0.01, the threshold of confidence is 0.75, and the threshold of rescaled values we chosen is 0.1.

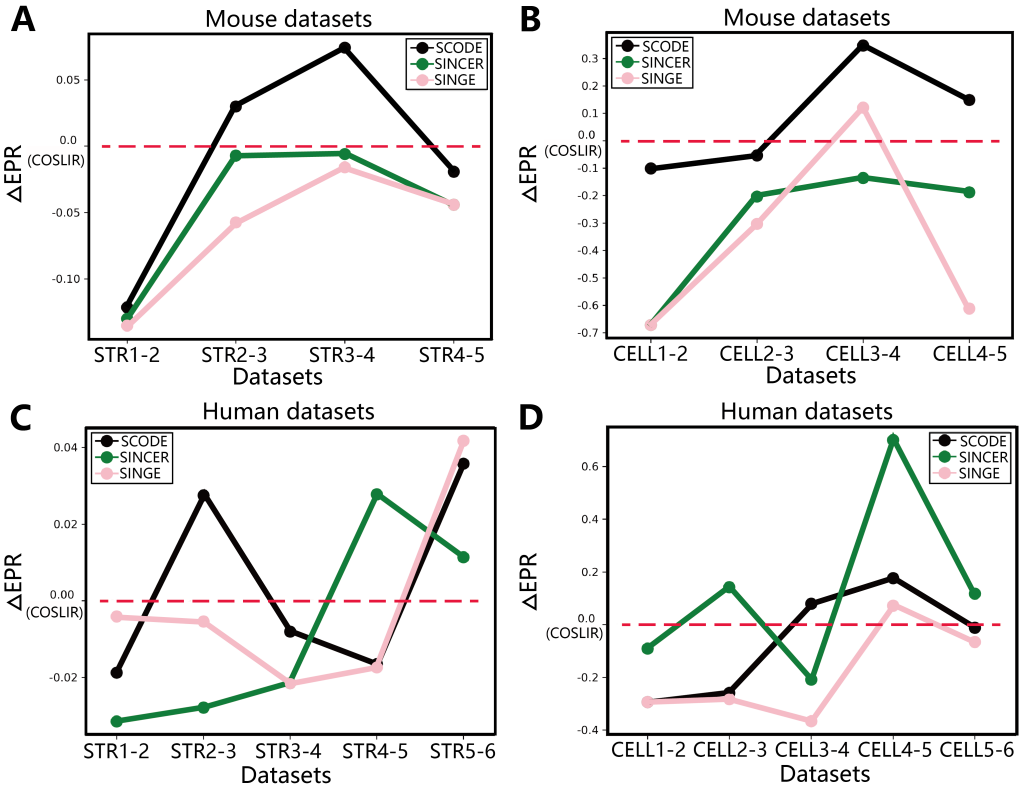

Figure 4: Early precision rate (EPR) across different algorithms in mESC (A, B) and hESC (C, D) datasets, using STRING (A, C) and cell-type-specfic ChIP-Seq (B, D) as the ground-truth. mESC has 5 time points and hESC has 6 time points. We here plot the differential EPR of existing algorithms with respect to COSLIR.