Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data

Atul Deshpande[1,2], Li-Fang Chu[2], Ron Stewart[2], and Anthony Gitter[2,3]

[1] *Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA*

[2] *Morgridge Institute for Research, Madison, WI 53715, USA*

[3] *Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA.*

Abstract

Advances in single-cell transcriptomics enable measuring the gene expression of individual cells, allowing cells to be ordered by their state in a dynamic biological process. Many algorithms assign 'pseudotimes' to each cell, representing the progress along the biological process. Ordering the expression data according to such pseudotimes can be valuable for understanding the underlying regulator-gene interactions in a biological process, such as differentiation. However, the distribution of cells sampled along a transitional process, and hence that of the pseudotimes assigned to them, is not uniform. This prevents using many standard mathematical methods for analyzing the ordered gene expression states. We present Single-Cell Inference of Networks using Granger Ensembles (SCINGE), an algorithm for gene regulatory network inference from single-cell gene expression data. Given ordered single-cell data, SCINGE uses kernel-based Granger Causality regression, which smooths the irregular pseudotimes and missing expression values. It then aggregates the predictions from an ensemble of regression analyses with a modified Borda method to compile a ranked list of candidate interactions between transcriptional regulators and their target genes. In two mouse embryonic stem cell differentiation case studies, SCINGE outperforms other contemporary algorithms for gene network reconstruction. However, a more detailed examination reveals caveats about transcriptional network reconstruction with single-cell RNA-seq data. Network inference methods, including SCINGE, may have near random performance for predicting the targets of many individual regulators even if the aggregate performance is good. In addition, in some cases including cells' pseudotime values can hurt the performance of network reconstruction methods. A MATLAB implementation of SCINGE is available at `https://github.com/gitter-lab/SCINGE`.

1. Introduction

Identifying the underlying gene regulatory networks (GRNs) that dictate cell-fate decisions is important for understanding biological systems. Although RNA-seq experiments on populations of cells undergoing a process of interest have been used to study cellular decision making, averaging transcriptional information from a heterogeneous population of cells can obscure biological signals. Advances in single-cell transcriptomics, such as single-cell RNA-seq, have enabled observing the gene expression states of individual cells [1–3]. While these solve the averaging problem faced by bulk transcriptomics, they are beset with new technical challenges, including measurement dropouts and a lower signal-to-noise ratio. Despite the technical problems, snapshots of the gene expression states of individual cells provide larger sample sizes and a finer understanding of the gene expression and regulatory dynamics during a biological process.

Many algorithms use single-cell RNA-seq data to infer GRNs [4], taking advantage of the large sample sizes. GRN inference requires identifying relationships between transcriptional regulators and their target genes or gene modules [5–7]. One strategy is to search gene expression datasets for dependencies among mRNA expression levels, making the simplifying assumption that a regulator's mRNA level approximates its regulatory activity. Single-cell datasets offer more data from which to learn these gene-gene relationships using multivariate information theory [8], linear regression [9], or other approaches. When single-cell expression data are collected at multiple times points, it provides more information that can be used for GRN inference. GRN reconstruction methods originally designed for bulk time-series transcriptomic data [10] can be repurposed to analyze time-stamped single-cell data. For example, Jump3 [11], a hybrid machine learning and model-based approach, has been adapted in this manner [12]. Time-stamped single-cell data also enables analyzing the evolution of gene expression distributions over time [13], which is not possible with bulk time series data or single-cell data collected at one time point.

When single-cell RNA-seq samples are not collected at multiple time points, computationally ordering cells along a biological process based on their expression states can approximate each cell's position along the process. These inferred times, called 'pseudotimes', can potentially lead to greater understanding of the causal regulatory relationships between genes. The dozens of algorithms for ordering cells and assigning pseudotimes [14], also referred to as trajectory inference, can be distinguished by their use of prior knowledge, treatment of pseudotime uncertainty, and the supported trajectory types [15]. Pseudotime algorithms can target trajectory types such as cyclic [16, 17], linear [18, 19], bifurcating [20], multifurcating [21], or tree-structured [22, 23]. In most of these methods, a pseudotime is assigned to each cell, which represents the cell's progress along the trajectory.

Similar to time series data, the pseudotemporal ordering provides an understanding of the gene expression trends along the biological process, which can support more accurate GRN reconstruction. Strategies for GRN inference with pseudotemporal data are related to those for time-stamped data with additional specializations to account for the technical differences. For example, SINCERITIES [24], originally designed to infer GRNs using ridge regression on time-stamped expression data, also admits pseudotime-labelled cells. SCODE [12], GRISLI [25], and Ocone et al. [26] infer GRNs by modelling the cell dynamics as ordinary differential equations with pseudotime as the temporal reference. Other strategies involve Gaussian processes regression for smoothing pseudotemporal data [27], time-lagged correlation [28], variational Bayesian inference on a first-order autoregressive moving average model [29], modified Restricted Directed Information [30], unsupervised classification using Gaussian Mixture Models [31], empirical Bayes-based thresholding [32], and modeling information propagation through genes as a cascade [33]. These strategies require estimating the cell trajectories before GRN inference. An alternative approach is to perform joint trajectory and co-expression network inference, for example, using Ornstein-Uhlenbeck models [21] or Gaussian mixtures with continuous parameters [34]. Despite these algorithmic advances, in case studies on real data the GRN reconstruction performance has often been disappointing and sometimes not substantially better than random networks.

In this study, we adapt Granger Causality for pseudotemporally-ordered single-cell expression data to assess whether this causal framework can overcome the difficulties faced by prior pseudotime-based GRN inference methods. We introduce our Single-Cell Inference of Networks using Granger Ensembles (SCINGE) algorithm, an ensemble-based GRN reconstruction technique that uses modified Granger Causality on single-cell data annotated with pseudotimes. Granger Causality [35] is a powerful approach for detecting causal relationships in long time series data. It has been used with bulk times series gene expression data [36–39], but these time series are typically short due to experimental limitations, making it more difficult to detect reliable gene-gene dependencies. The longer (pseudo) time series obtained from ordered single-cell datasets make them appealing for Granger Causality-based GRN reconstruction. However, single-cell challenges such as dropouts and irregular sampling along the biological trajectory counteract the benefits of the longer pseudotime series. SCINGE addresses these concerns by using a kernel-based Granger Causality method that smooths the expression data and ensembling to improve GRN prediction robustness.

We apply SCINGE to reconstruct GRNs of two mouse embryonic stem cell differentiation processes characterized with single-cell RNA-seq. SCINGE compares favorably with existing GRN inference methods designed for temporal or pseudotemporal gene expression data when evaluated using ChIP-seq, ChIP-chip, loss-of-function, and gain-of-function data. However, our evaluation reveals important caveats about GRN evaluation and the value of pseudotime for GRN inference that are broadly applicable for pseudotime-based GRN reconstruction.

2. Results

2.1. SCINGE and Granger Causality Overview

SCINGE takes ordered single-cell gene expression data as input and provides a ranked list of regulator-gene relationships as its primary output. It requires the single-cell dataset to be annotated with pseudotimes. This assigns a numeric pseudotime to each cell in the dataset that represents how much that cell has progressed through a dynamic biological process such as differentiation. For each target gene, SCINGE assesses which past expression values are most predictive of its expression, that is, the candidate regulators of each gene. This is achieved using a specialized form of Granger Causality, which is framed as a regularized regression problem. The past expression values are determined using the pseudotimes.

The Granger Causality [35] test at SCINGE's core is a hypothesis test to ascertain predictive causality between a 'source' and 'target' time series. A series x is said to Granger-cause y if past values of x contain information that helps predict future values of y. The primary complication of applying Granger Causality to single-cell expression data with inferred pseudotimes is that the distribution of cells along the trajectory, and the pseudotimes assigned to them, is not uniform. Standard Granger Causality is not an effective analytical tool with irregularly-spaced pseudotimes [30]. One potential workaround is to resample the irregularly-spaced pseudotime series to obtain a regular time series. However, resampling introduces interpolation errors in the form of a low-pass filtering, which could be detrimental to analysis of highly non-linear biological processes. SCINGE instead uses an alternative solution proposed by Bahadori and Liu, the Generalized Lasso Granger (GLG) test [40]. GLG modifies the Lasso Granger test [41] to support irregular time series. Within SCINGE, GLG uses a kernel function to smooth the past expression values of candidate regulators, mitigating the irregularly-spaced pseudotimes and zero values that are prevalent in single-cell expression data.

SCINGE depends on hyperparameters that control the kernel smoothing, sparsity, and which window of previous expression is considered. We do not search for a single optimal set of hyperparameters but rather consider many regulator-gene predictions obtained under different hyperparameters. In addition, we subsample the expression data many times to further improve robustness. The final SCINGE network is obtained from an ensemble of all of the individual predicted networks using different hyperparameters and cell subsamples (Figure 1).

2.2. ESC to Endoderm Differentiation

Our first case study tracks the differentiation of mouse embryonic stem cells (ESC) to primitive endoderm cells over 72 hours [42]. Matsumoto et al. [12] previously pre-processed this dataset to benchmark their SCODE GRN algorithm. We reuse their processed version of the data, which included expression data for only 100 transcription factors (TFs) and assigned pseudotimes to the 356 single-cell RNA-seq measurements using Monocle. We use this ESC to endoderm differentiation dataset to design, optimize, and tune the SCINGE algorithm, assessing how well it recovers known regulator-gene interactions that are relevant in mouse embryonic stem cell differentiation from the ESCAPE database [43]. The ESCAPE gold standard is incomplete due to lack of experimental data for many of the relevant TFs (see Section 3.4.2). Therefore, the gold standard only contains an 11×100 sub-set of the 100×99 regulator-gene interactions that SCINGE scores. SCINGE does not score self-edges.

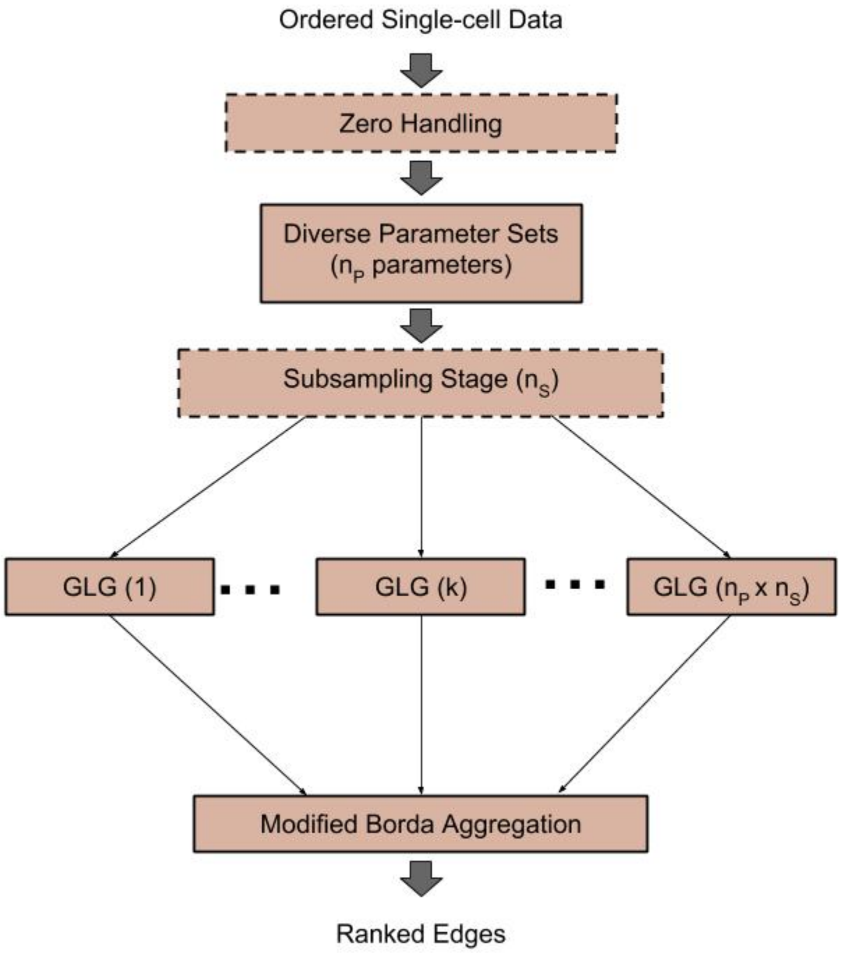

Figure 1: SCINGE takes single-cell gene expression data that has been annotated with pseudotimes as input and predicts a ranked list of regulator-target gene edges. Dashed boxes in the pipeline are optional steps. Zero handling removes a portion of the 0 values for each gene so that they do not have too strong an influence on the smoothed expression values. Many combinations of hyperparameters that control sparsity, the expression smoothing, and the pseudotemporal history are considered. Subsampling repeatedly removes a fraction of the cells at random. For each hyperparameter combination and subsampled dataset, GLG regression predicts the regulators of each target gene. These results are aggregated into a final ensemble network prediction with a modified version of Borda aggregation.

We obtain the SCINGE-inferred regulatory network as a ranked list of predicted regulator-gene interactions (Supplementary File 1). SCINGE ranks Foxd3, Gli2, and Nanog as the three most influential regulators in the 100-gene subnetwork. To illustrate the notion of a GLG-inferred regulatory edge, we consider Pou5f1 as an example target gene. Figure 2 shows that using additional past information from GLG-identified regulators improves the predicted expression trend of Pou5f1 along pseudotime. Indeed, as more genes are added in decreasing order of the 'edge weight,' the predicted expression trends becomes more accurate.

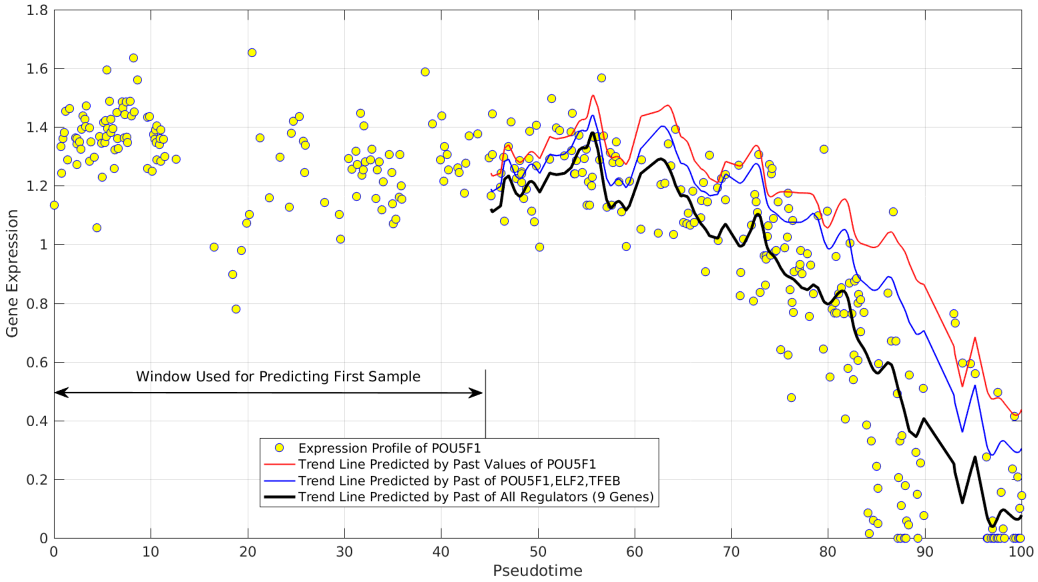

Figure 2: Generalized Lasso Granger example. The Pou5f1 expression profile is predicted more accurately using all regulator genes detected by the GLG test compared to its own history or its history and the top two regulators.

To assess whether SCINGE can match or exceed the state-of-the-art performance after dataset-specific tuning, we compare its predicted GRN with three existing network inference methods — SINCERITIES [24], which uses ridge regression; SCODE [12], which is based on ordinary differential equations; and Jump3 [11], which is based on decision trees. We emphasize this particular evaluation is not indicative of which method would perform best on new data because of SCINGE's tuning. Nevertheless, SCINGE performs much better than the other methods with respect to the average precision (**A**) and average early precision (**E**), which both summarize a precision-recall curve (Figure 3). Average early precision emphasizes the most-confident, top-ranked interactions. Even though SCODE was previously evaluated using this gene expression data [12], it performs worse than random when assessed using the condition-specific ESCAPE gold standard (see Section 4.2).

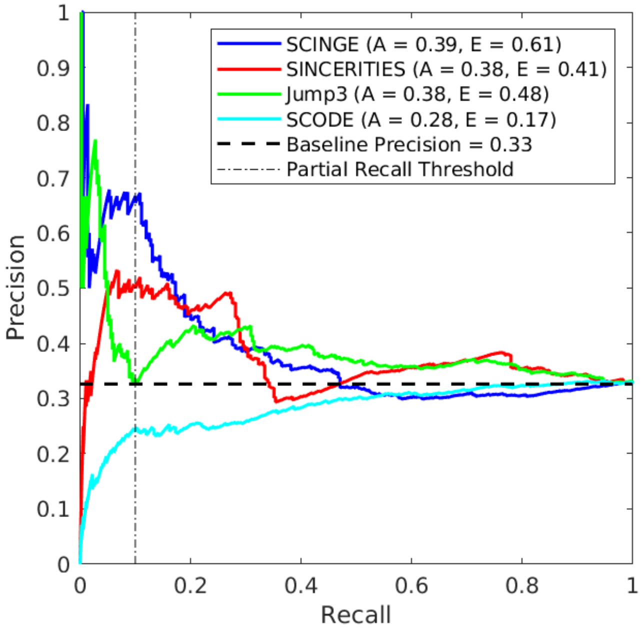

Figure 3: Precision-recall performance of SCINGE compared to SINCERITIES, Jump3, and SCODE when predicting the GRNs involved in primitive endoderm differentiation. The Baseline Prediction is the expected precision obtained by randomly ordering all regulator-gene interactions. Key: **A** - Average Precision, **E** - Average Early Precision ($\leq$ 0.1 recall).

2.3. Retinoic Acid-driven Differentiation

We further test SCINGE on a second dataset that tracks retinoic acid-driven differentiation from mouse embryonic stem cells to extraembryonic endoderm and neuroectoderm cells over 96 hours [44]. We infer a trajectory for the biological process using Monocle 2 and select 1886 cells from cell states 1 and 2 (Figure S1 and Section 3.3.2). Monocle 2 also identifies 626 genes whose expression changes substantially as a function of pseudotime, which we use for GRN reconstruction. These genes are not filtered to include only TFs or other known expression regulators. SCINGE returns a ranked list of all 626 × 625 possible regulatory relationships, excluding self-edges.

SCINGE identifies key regulators reflecting the differentiation trajectory required for mouse embryonic stem cells exiting the pluripotent state, transitioning through the epiblast where lineage segregations take place [44] (Table 1 and Figure 4). We use g:Profiler [45] to identify Gene Ontology (GO) biological process terms that are significantly enriched among the ranked SCINGE regulators (Supplemental File 3). This searches for GO terms that are enriched at the top of the ranked list, assessing all possible rank thresholds. The g:Profiler analysis identifies relevant significantly enriched biological processes in the sorted regulator list including cellular response to growth factor stimulus (GO:0071363), cell morphogenesis involved in differentiation (GO:0000904), neuron differentiation (GO:0030182), and additional terms depicted in Table 1.

There are two ways to explore the SCINGE predictions in greater detail: the top regulators ranked by SCINGE influence (Table 1), which aggregates influence over all target genes, and the top-ranked edges (Figure 4). Table 1 shows the top 20 regulators ranked by SCINGE influence. Ten of the top predicted regulators are associated with regulation of gene expression (GO:0010468), as are other regulators with high SCINGE influence that are beyond the top 20 (Supplemental File 3). The top 20 regulators also include essential genes that cause embryonic lethality in mouse embryos harboring homozygous null alleles. Others show phenotypes ranging from postnatal lethality to growth retardation (Table 1). Three of the predicted regulators (Alg13, Gpx3, and Lactb2) are known for their roles in metabolic processes but are not known to participate in regulation of early embryonic lineage specification. In addition, KinderMiner [64] text mining reveals significant associations between the top 20 regulators and terms related to this developmental process: ‘embryonic stem cells,’ ‘neural development,’ and ‘endoderm development’.

Figure 4 illustrates the most-confident 100 regulator-gene edges from the SCINGE network, directed from the regulators (hexagons) to the target genes (ellipses). This representative subnetwork comprises 18 unique regulators and 65 unique targets. Fourteen of these regulators are also found among the top 20 regulators by SCINGE influence (Table 1), including all 10 known to be associated with regulation of gene expression. The other four regulators participate in one or more high-confidence edges but do not have high ag-

Rank	Gene name	Regulation of gene expression	Neurogenesis	Regulation of cellular response to growth factor stimulus	Regulation of canonical Wnt signaling pathway	Loss-of-function phenotypes	KinderMiner associations
1	Dab2	✓		✓	✓	EL [46]	ESC EndoDev
2	Fgf4	✓		✓		EL [47]	ESC EndoDev NeurDev
3	Sfrp5	✓		✓	✓	Normal [48]	EndoDev
4	Lefty2	✓				EL [49]	ESC EndoDev
5	Zfp703	✓		✓	✓	N/A	
6	Hoxb2	✓				NL [50]	ESC NeurDev
7	Gata6	✓		✓		EL [51]	ESC EndoDev
8	Cdh2		✓		✓	EL [52]	ESC NeurDev
9	Alg13					EL [53]	
10	Mdm4	✓				EL [54]	
11	Gpx3					Others [55]	
12	Igf2	✓				Others [56]	ESC EndoDev NeurDev
13	Ccnd2					S [57]	ESC
14	Wdr1		✓			EL [58]	
15	Ilk	✓	✓		✓	EL [59]	ESC
16	Flrt3		✓			EL [60]	EndoDev
17	Lactb2					N/A	
18	Wls				✓	EL [61]	
19	Fzd3		✓			NL [62]	ESC NeurDev
20	Crabp1					Normal [63]	ESC NeurDev

Table 1: GO biological process terms, loss-of-function phenotypes, and KinderMiner associations related to the top 20 SCINGE regulators. **Phenotype key** — EL: Embryonic lethality, NL: Neonatal lethality, S: Sterile, Normal: Homozygous mutant mice are phenotypically normal and fertile, Others: Homozygous mutant mice display other physiological phenotypes, N/A: No knockout mice reported in peer-reviewed studies. **KinderMiner key** — ESC: Embryonic Stem Cells, NeurDev: Neural development, EndoDev: Endoderm development

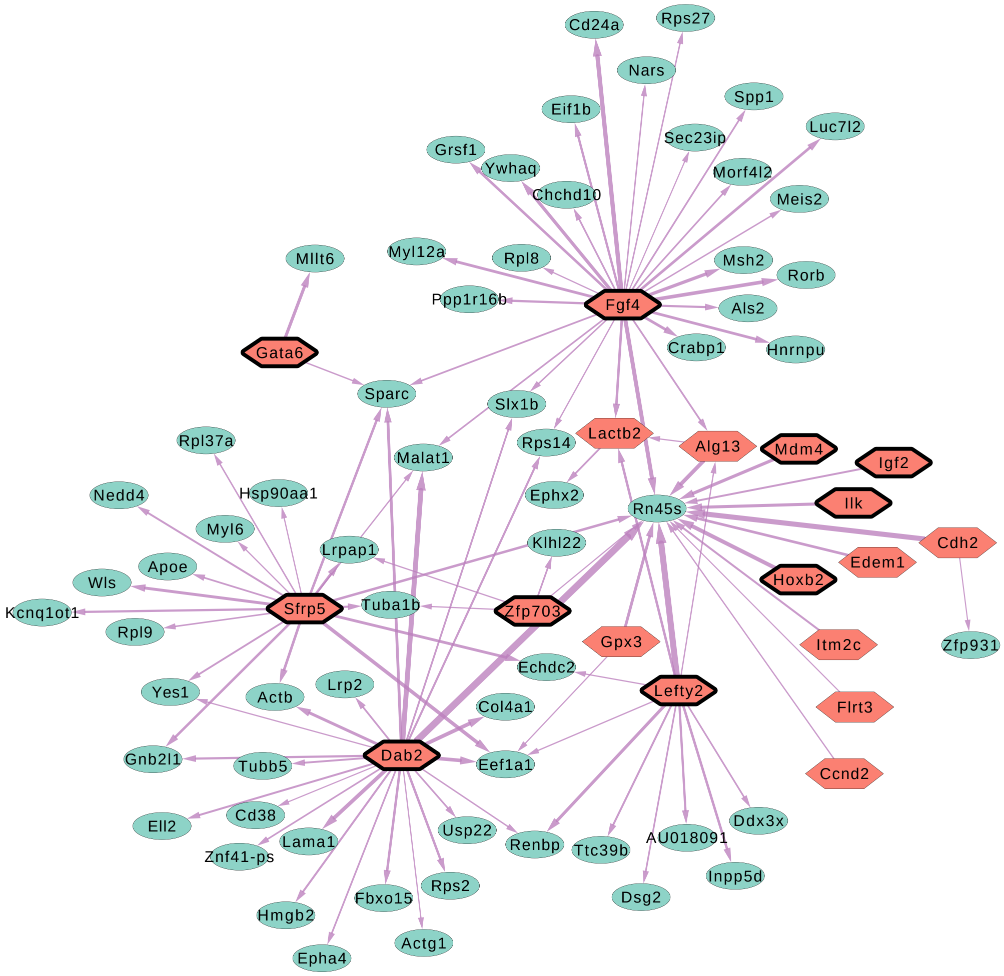

Figure 4: The network obtained from the top 100 edges ranked according to SCINGE scores shows 18 unique regulators (hexagonal nodes, the ten with solid boundaries corresponding to known regulators of gene expression listed in Table 1) and 65 unique targets (elliptical nodes). The higher ranked edges are represented by thicker arrows.

gregate influence. Dab2 and Fgf4 are the most influential regulators overall (Table 1) and hub regulators among the top 100 edges (Figure 4). Rn45s is a frequently-regulated target gene. Fgf4 governs the exit from the pluripotent state. *Fgf4-null* mouse embryonic stem cells resist neural and mesodermal lineage induction [65]. Indeed, the Fgf/Map kinase signaling pathway plays multiple roles during mouse blastocyst development, and mutations of the signaling components (e.g., Fgf4, Fgfr2, and Grb2) all cause implantation lethality and lack of primitive endoderm development [66]. Moreover, Fgf4 also governs neural induction in embryonic stem cell differentiation at a later stage of development [67].

The predicted GRN in Figure 4 also provides hypotheses for future experimental tests. For example, Meis1 and Meis2 are homeobox proteins that directly regulate Pax6 expression during eye development [68]. SCINGE predicts that Fgf4 regulates Meis2. Thus, Fgf4 could potentially act upstream of Meis1 and Meis2 to regulate Pax6 expression, contributing to neuroectoderm differentiation [69]. Other key primitive endoderm regulators are also highlighted in SCINGE predictions such as Gata6, a transcription factor necessary and sufficient for primitive endoderm lineage differentiation and establishment of extraembryonic endoderm cell lines [70]. Dab2, Sfrp5, Lefty2, and Igf2 are all expressed in the primitive endodermal lineages, including visceral endoderm and extraembryonic endoderm cell lines [71–75].

Many expected GO terms and regulators are represented in Table 1 and Figure 4. However, classic neuroectoderm regulators like Sox1, Nes, and Pax6 [44] are missing because they are excluded from the limited shortlist of genes in the SCINGE input. We only run SCINGE on the top 626 significantly differentially expressed genes along the differentiation trajectory detected by Monocle 2.

2.4. Retinoic Acid-driven Differentiation ESCAPE Evaluation

The retinoic acid-driven differentiation study can be used to benchmark the relative performance of SCINGE with respect to the other network inference methods because none of the methods, including SCINGE were optimized or tuned based on the ESCAPE evaluation results. Figure 5 shows the precision-recall performance of SCINGE compared with SINCERITIES, Jump3, and SCODE when ranking edges in the 626-gene network. Due to Jump3's runtime, we run it on a reduced dataset (Section 3.4.1), which may impact its performance. As with the ESC to endoderm differentiation dataset, the ESCAPE database had only partial information (12 regulators), thus limiting the gold standard to a submatrix of 12 × 626 possible edges. SCINGE is the best method overall in terms of average precision and average early precision (Figure 5). Jump3 is effectively tied with SCINGE for average early precision but has near-random precision for recall > 0.2. SINCERITIES prioritizes ESCAPE gold standard interactions well at the top of its ranked list, but the performance degrades quickly. SCODE is worse than random. The performance depends on the type of regulator-gene interaction in the ESCAPE database. SCINGE can recover loss-of-function or gain-of-function (*lof/gof*) relationships but struggles to identify ChIP-based protein-DNA binding interactions (Figure S2).

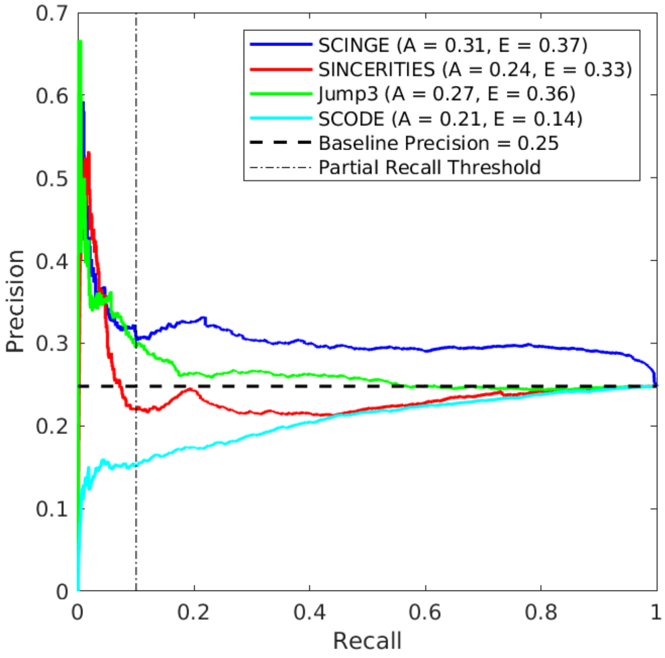

Figure 5: Precision-recall performance of SCINGE, SINCERITIES, Jump3 (which uses a reduced data set), and SCODE when predicting a 626-gene retinoic acid-driven differentiation regulatory network [44]. Key: **A** - Average Precision, **E** - Average Early Precision (≤ 0.1 recall).

Visualizing the expression trends over pseudotime can illustrate the types of errors SCINGE makes with respect to the ESCAPE gold standard. For example, the interaction Esrrb→Actb was detected with ChIP but is not part of the ESCAPE's *lof/gof* dataset. There is no apparent lag between the expression trends of the regulator and target (Figure S3). This edge was ranked highly by SCODE but not by SCINGE, which searches for lagged expression dependencies by design.

A regulator-specific evaluation partially explains the overall precision-recall performance of the four GRN methods and demonstrates that it can be somewhat misleading. Figure 6 shows the average precision and average early precision for all four methods with respect to each regulator in the ESCAPE database. The regulator-specific average precision of all four methods is at or below random for most regulators, with a few exceptions.

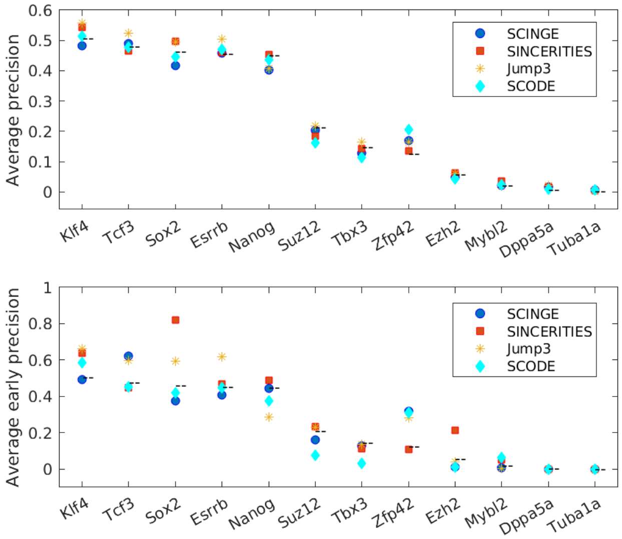

Figure 6: Average precision and average early precision evaluated for individual regulators in the ESCAPE database. The dashed line (−−) indicates random performance.

Because some regulators are more prevalent in the ESCAPE gold standard than others, the overall precision-recall curve is influenced by regulator-specific precision and the relative ordering of the regulators in the ranked edge list. We can sort these 12 regulators in decreasing order by their number of outgoing edges in the ESCAPE gold standard, which informs the regulator's influence on the evaluation, and generate boxplots of the regulator-specific edge ranks in the GRNs (Figure 7). SCINGE ranks outgoing edges from ESCAPE's most prevalent regulators (Klf4 and especially Tcf3) higher on average than the regulators with fewer target genes (Dppa5a and Tuba1a). The distributions of rankings from SINCERITIES and Jump3 are widely dispersed for each regulator. Meanwhile, SCODE ranks edges from the regulators with few outgoing edges higher than those with many target genes, contributing to its poor overall performance.

These regulator-specific results provide insights into Figure 5. SCINGE's relatively high early average precision is influenced by how it ranks regulators in accordance with their prevalence in the ESCAPE database. On the other hand, Jump3 ranks all regulators uniformly but has better than random average precision on multiple individual regulators. Unlike the other three GRN methods, Jump3 does not use the pseudotime values, which may boost its regulator-specific performance (Section 2.7).

2.5. Benefits of Ensembling

The optimal GLG parameters that best identify causal relationships between two genes can vary from gene to gene and for different biological processes. In the absence of prior information about the regulatory network, it is difficult to predict optimal hyperparameters for the GLG test. Furthermore, it is also plausible that different transcriptional regulators have different kinetics and consequently different optimal hyperparameters.

SCINGE attempts to overcome this with an ensemble of hyperparameters, aggregating the results to obtain the final SCINGE score of each GRN edge. Figures S4–S7 compare the performance of individual GLG hyperparameters to the complete ensembled SCINGE GRN for both datasets. Although SCINGE does not result in the best average precision or average early precision, it performs better than the majority of the individual hyperparameters. Ensembling reduces the risk of choosing a single set of hyperparameters that would perform poorly for a particular dataset.

2.6. Effects of Subsampling and Zero Handling

SCINGE's ensembling can also improve performance by supporting subsampling and zero handling. Because the core GLG test is compatible with irregular time series, we can create randomly subsampled time series from each gene's expression data to generate multiple instances of the original dataset. In these experiments, subsampled replicates are created by removing individual expression data samples with probability of removal 0.2. Both types of precision-recall summaries, average precision and average early precision, tend to increase as more subsampled replicates are added to the ensemble (Figure 8).

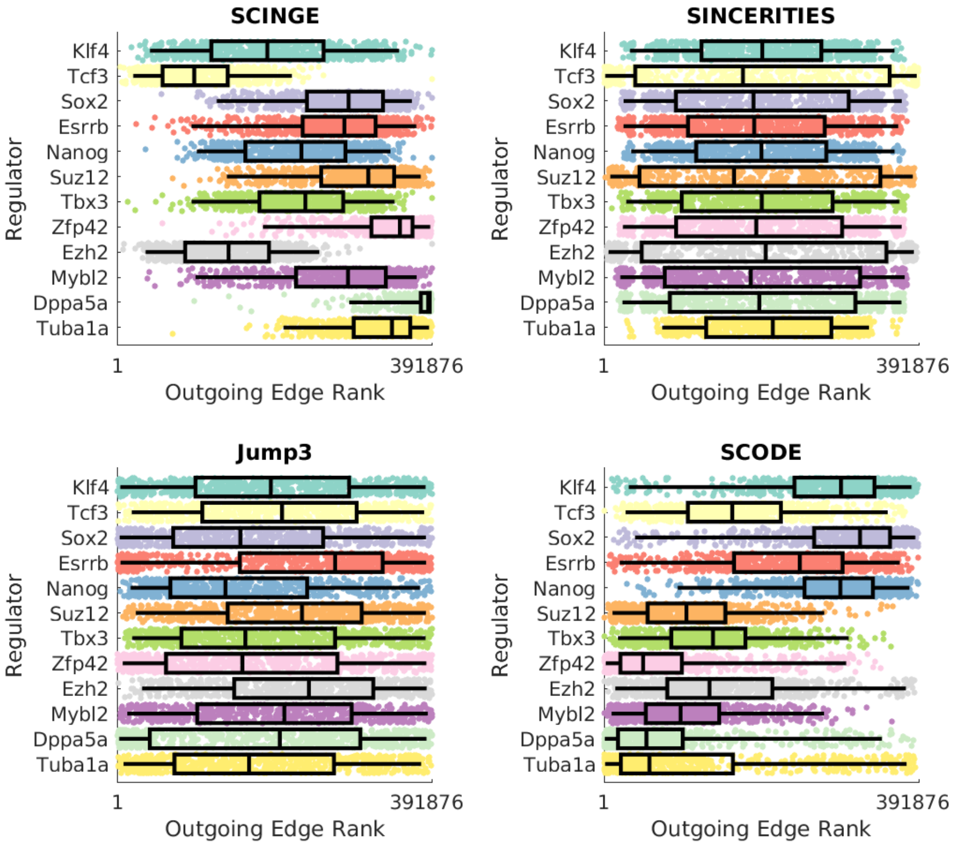

Figure 7: The ranking of regulator-specific interactions has a strong effect on the overall precision-recall curve (Figure 5). The boxplots show the outgoing edge ranks for each regulator in each predicted GRN, in decreasing order of regulator prevalence in the ESCAPE database. Ranking regulator-gene interactions involving the predominant ESCAPE regulators (e.g., Klf4) above those involving the less frequent ESCAPE regulators (e.g., Tuba1a) improves the precision-recall performance, and the inverse is also true.

The support for irregular time series also allows us to remove zero-valued data points corresponding to technical dropouts. The true dropout probability is gene dependent and can be estimated by methods like SCONE [76]. As a proof of concept of SCINGE's support for zero handling, we incorporate a simpler strategy that uses a constant dropout probability hyperparameter *prob_zero_removal* for all genes. For each GLG instance, we remove zero-valued expression samples (and their corresponding timestamp) from each

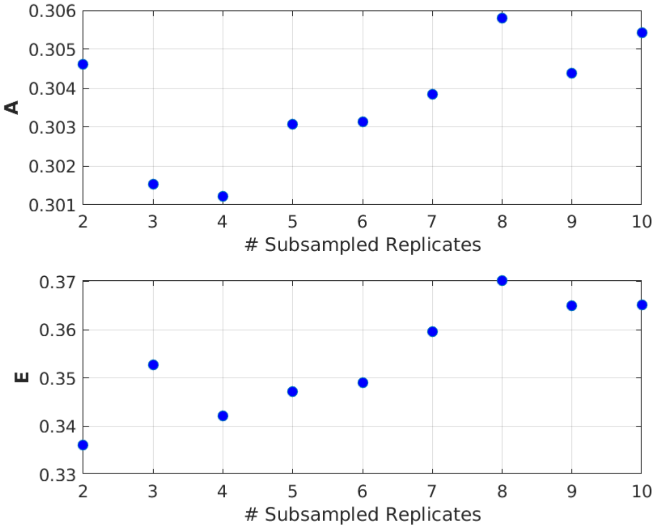

Figure 8: Effect of subsampling on SCINGE performance on the retinoic acid-driven differentiation dataset. The probability of removing each expression sample when creating a subsampled dataset is 0.2. Key: **A** - Average Precision, **E** - Average Early Precision (≤ 0.1 recall).

gene's expression series with a *prob_zero_removal* probability of removal for each zero.

Figure 9 shows the precision-recall performance of SCINGE as the value of *prob_zero_removal* increases. A moderate approach to zero handling increases the average precision and average early precision, but as it becomes more aggressive, performance degrades. Filtering too many zeros may remove genuine zero expression values along with the dropouts. We currently recommend using SCINGE with a moderate constant dropout probability and will explore directly supporting gene-dependent dropout (Section 4.3).

2.7. Assessing whether Pseudotimes and Cell Ordering Improve GRN Reconstruction

We assess the impact of using assigned cell order and pseudotime values on the performance of the three methods designed to reconstruct GRNs from pseudotemporal single-cell gene expression — SCINGE, SINCERITIES and SCODE. We exclude Jump3 because it uses only the cell ordering and does not use pseudotime values. For this assessment, we create variants of both

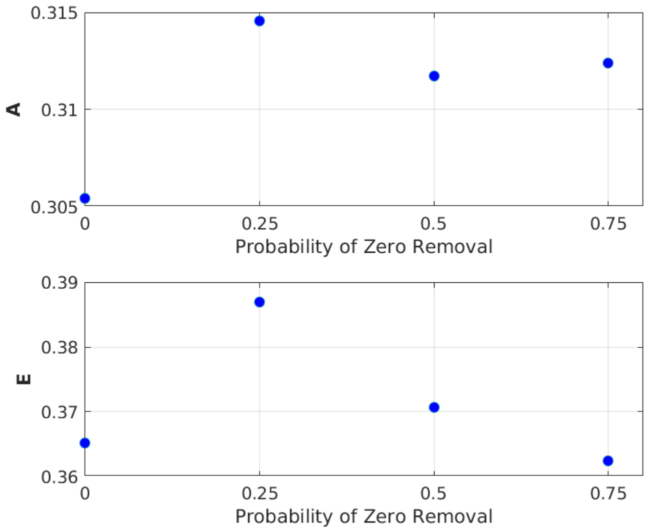

Figure 9: Effect of zero removal on SCINGE performance for multiple values of $prob\_zero\_removal$, the probability of removing a zero value. SCINGE ensembles the results from 10 zero-filtered replicates on the retinoic acid-driven differentiation dataset. Key: **A** - Average Precision, **E** - Average Early Precision ($\leq$ 0.1 recall).

the ESC to endoderm differentiation and retinoic acid-driven differentiation datasets as described below:

- *Pseudotime:* The default mode using ordered cells with Monocle or Monocle 2 assigned pseudotimes.
- *Order Only:* Obtained from the *Pseudotime* variant by removing the assigned pseudotime values but maintaining the cell order. The cells are assumed to be regularly-spaced along the trajectory.
- *Rand. Order (3):* Three replicates obtained from random permutation of the regularly-spaced cells from the *Order Only* variant. The randomized data have neither pseudotime annotations nor ordering information.

If estimated pseudotimes contribute high-quality information for GRN reconstruction, the three methods should have highest performance on the *Pseudotime* dataset, with less accurate predictions from the *Order Only* and *Rand. Order* datasets.

Figure 10 shows the average precision and early average precision of SCINGE, SINCERITIES, and SCODE when run on the three variants of each dataset above. For variants of the ESC to endoderm differentiation dataset, only SCINGE's performance decreases substantially for the *Rand. Order* datasets as expected. Its performance on the *Order Only* dataset is only slightly worse than the original *Pseudotime* dataset. SINCERITIES is less consistent on the *Rand. Order* datasets, with some randomized cell orders providing better GRNs than the real *Order Only* or *Pseudotime* datasets. SCODE performs poorly even on the original *Pseudotime* dataset (Figure 3) so we cannot draw strong conclusions from its performance trend across the dataset variants.

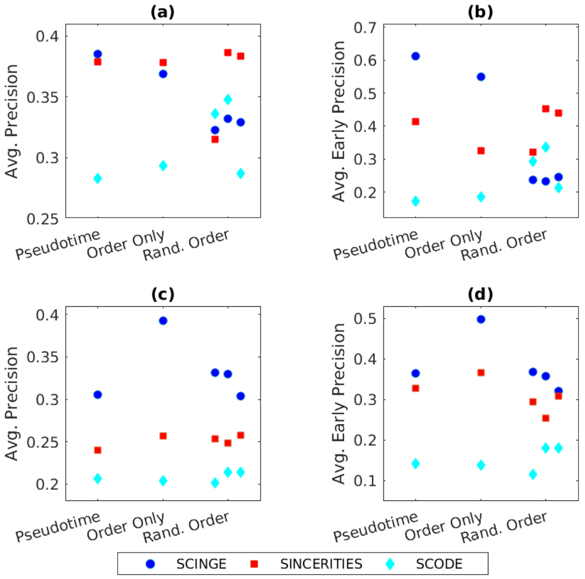

Figure 10: Effects of pseudotimes and cell ordering on the performance of SCINGE, SINCERITIES, and SCODE. (a) and (b) show the performance of the three methods when analyzing variants of the ESC to endoderm differentiation dataset. (c) and (d) show the performance of the three methods on variants of the retinoic acid-driven differentiation dataset.

On the other hand, for variants of the retinoic acid-driven differentiation dataset SCINGE still outperforms SINCERITIES and SCODE in all cases, but the performance trend does not follow the expected pattern. Both SCINGE and SINCERITIES show higher performance on the *Order Only* dataset in which the pseudotime values are removed. The performance for the *Rand. Order* variants is worse than the *Order Only* dataset but comparable to the *Pseudotime* dataset. SCODE again performs poorly in all cases. Because the performance improves for both SCINGE and SINCERITIES when only the cell ordering is used, one possible explanation is that the assigned pseudotime values themselves are low fidelity, counteracting any potential benefits from the additional information. Indeed, further analysis of the regulator-specific performance of these three methods using the *Order Only* dataset in Figure S8 shows that the regulator-specific average precision and average early precision metrics of SCINGE and SINCERITIES improve compared to Figure 6. Like Jump3, these two methods now have substantially better than random early average precision for several regulators. Similarly, Figure S9 shows that the SCINGE and SCODE average rankings of the outgoing interactions from the regulators using the *Order Only* dataset are more commensurate with their ESCAPE prevalence. These two phenomena combine to improve SCINGE's overall precision-recall curve with respect to its *Pseudotime* dataset performance and those of other GRN methods for either form of the dataset (Figure S10).

2.8. Computational Runtime

We designed SCINGE to take advantage of the high-throughput computing resources that are readily available to computational researchers, such as the free Open Science Grid [77]. We compare the computational runtime of this strategy to the other three GRN inference methods on the retinoic acid-driven differentiation dataset. SCODE and SINCERITIES require the least computational resources. It was possible to run them on a single workstation with a 64-bit Intel i5-4590 CPU and 8 GB RAM. Specifically, on this workstation, the SCODE algorithm with 100 repetitions requires approximately 6 hours to complete, whereas the SINCERITIES algorithm takes approximately 111 hours.

In contrast, both SCINGE and Jump3 require more varied and extensive computing resources. In the case of Jump3, inferring the GRN from 626 regulators to one target gene takes between 11 minutes to 74 hours to run, with an average runtime of 21.7 hours. This is repeated for each target gene. Within SCINGE, obtaining the ranked edge list of the entire 626×626 subnetwork for a single GLG test on one subsampled replicate takes approximately 3.8 hours. In a typical application, SCINGE uses 100 different hyperparameter settings on 10 subsampled expression datasets, which translates to approximately 3813 hours of computation. However, both Jump3 and SCINGE are highly parallelizable. We deployed them on our local high-throughput computing cluster using HTCondor [78], which connects to the Open Science Grid [77]. In this high-throughput setting we can run the entire SCINGE algorithm in 36 hours and the Jump3 algorithm in 72 hours.

3. Methods

As illustrated in Figure 1, SCINGE infers underlying gene regulatory networks of a biological process by aggregating ranked edge lists obtained from an ensemble of Generalized Lasso Granger tests conducted on ordered single-cell transcriptomic data. The GLG test is a kernel-based generalization [40] of the Lasso Granger Causality test to facilitate the analysis of causal relationships between irregular time series obtained from a linear stationary vector autoregressive (VAR) model. We first describe the GLG test and then the structure of the SCINGE algorithm.

3.1. Generalized Lasso Granger Test

The Generalized Lasso Granger test is used to discover temporal causal networks from irregularly-spaced time series data based on concepts of Granger Causality. In the GRN inference domain, the time series correspond to temporal gene expression measurements.

Assume P regularly-spaced time series $x_1, x_2, \ldots, x_P$ are obtained at timestamps $\{t\} \doteq 1, 2, \ldots, T$. These time series are assumed to be governed by a linear and stationary vector autoregressive process such that

$$x_i(t) = \sum_{j=0}^{P} \sum_{l=1}^{L} \mathbf{a}_{i,j}(l) x_j(t-l) + \varepsilon_i(t), \tag{1}$$

for $i = 1, 2, \ldots, P$, where $\mathbf{a}_{i,j}(l)$ corresponds to the l-th lagged coefficient from source time series x_j to target time series x_i and $\epsilon_i(t)$ is measurement error, represented by independently distributed Gaussian random variables. More generally speaking, the unknown $P \times P \times L$ matrix $\mathbf{a}$ comprising of L lagged coefficient matrices $\mathbf{a}(1), \mathbf{a}(2), \ldots, \mathbf{a}(L)$ represents the evolutionary mechanism of $x_1, x_2, \ldots, x_P$.

The Lasso Granger Causality test [41] for an individual regularly-spaced target series x_i is characterized by the optimization problem

$$\min_{\{\mathbf{a}_i\}} \sum_{t=L+1}^{T} \left| x_i(t) - \sum_{j=1}^{P} \sum_{l=1}^{L} \mathbf{a}_{i,j}(l) \cdot x_j(t-l) \right|^2 + \lambda \sum_{j=1}^{P} ||\mathbf{a}_{i,j}||_1, \qquad (2)$$

and provides a sparse estimation of the $P \times L$ coefficient matrix $\mathbf{a}_i$ representing the VAR process that relates each source series $x_{j \neq i}$ to the target series x_i. Specifically, if elements of j-th column $\mathbf{a}_{i,j}$ of the matrix are statistically significant, then we claim that x_j Granger-causes x_i (represented by $j \rightarrow i$). The Lagrange multiplier λ dictates the sparsity of the learned matrix $\mathbf{a}_i$.

The Generalized Lasso Granger (GLG) test proposed by Bahadori and Liu [40] is a kernel-based modification of Equation 2 to facilitate the analysis of irregular time series. Irregular means that the time between consecutive time points can vary. Given two timestamps t_1 and t_2, Bahadori and Liu define a Gaussian kernel function, instrumental to the generalization process, as

$$w(t_1, t_2) = exp\left(\frac{-(t_1 - t_2)^2}{\sigma^2}\right),$$

where σ represents the effective kernel width. Based on this kernel function, the operator $\odot$ defined below generalizes the inner product for two 'irregular' time series — x, sampled at times $t_x(1), t_x(2), \ldots, t_x(N_x)$, and y, sampled at times $t_y(1), t_y(2), \ldots, t_y(N_y)$ — as

$$x(t_x) \odot y(t_y) \doteq \sum_{n=1}^{N_x} \frac{\sum_{m=1}^{N_y} x(n) y(m) w(t_x(n), t_y(m))}{\sum_{m=1}^{N_y} w(t_x(n), t_y(m))}.$$

N_x and N_y can differ, which SCINGE exploits for its dropout handling and subsampling (Section 3.2).

We now have P irregular time series $x_1, x_2, \ldots, x_P$ obtained from a linear and stationary VAR process. Each series x_i of length N_i is sampled at irregularly-spaced timestamps t_i such that $t_i(n+1) \geq t_i(n)$ for $n = 1, 2, \ldots, N_i$. As with the Lasso Granger Causality test, the objective of the GLG test is to obtain the sparse coefficient matrix $\mathbf{a}_i$, which represents the underlying VAR model for the target series x_i. To overcome the irregularity of the time series, we follow Bahadori and Liu by visualizing each vector $\mathbf{a}_{i,j}$ of the coefficient matrix as a quasi-time series $\mathbf{a}'_{i,j}(t)$ with respect to timestamp t as

$$\mathbf{a}'_{i,j}(t) = \{(t_a(l), a_{i,j}(l)) | l = 1, 2, \ldots, L, t_a(l) = t - l\Delta t\},$$

where Δt represents the time resolution of the quasi-time series $\mathbf{a}'_{i,j}(t)$. Note that for $t_1 \neq t_2$, $\mathbf{a}'_{i,j}(t_1)$ and $\mathbf{a}'_{i,j}(t_2)$ would have the same observation variables $a_{i,j}(l)$ but different timestamps.

Next, for a given timestamp $t_i(n)$ corresponding to a sample in x_i, we generalize the inner product in Equation 2 by using

$$\sum_{l=1}^{L} \mathbf{a}'_{i,j}(l) \odot x_j(t - l)$$

defined on $\mathbf{a}'_{i,j}(t_i(n))$ and $x_j(t_j)$ using their respective timestamps to calculate the kernel weights. Substituting this generalized inner product in Equation 2, we obtain the optimization problem for GLG, given by

$$\min_{\{\mathbf{a}_i\}} \sum_{t_i(n) \geq L\Delta t} \left| x_i(t_i(n)) - \sum_{j=1}^{P} \mathbf{a}'_{i,j}(t_i(n)) \odot x_j(t_j) \right|^2 + \sum_{j=1}^{P} \lambda_j ||\mathbf{a}_{i,j}||_1. \quad (3)$$

The first term represents the mean-squared error between the sample values $x_i(t_i(n))$ of the i-th series at each timestamp $t_i(n) \geq L\Delta t$ and its corresponding prediction from the generalized inner product $\mathbf{a}'_{i,j}(t_i(n)) \odot x_j(t_j)$, which uses the kernel defined above to 'smooth over' the mismatched irregular timestamps. The second term is a sparsity constraint on the coefficient matrix $\mathbf{a}_i$. The minimizer $\mathbf{a}_i$ of the objective function in Equation 3 provides the coefficient matrix that represents the VAR model of the target series x_i from all available source time series x_j, with λ determining the sparsity of the coefficient matrix. If the time series represent irregularly-spaced gene expression data, $\mathbf{a}_i$ can be interpreted as an estimate of the regulatory effect of other genes on the i-th gene. The presence of edges in the regulatory network for the i-th gene is indicated by significant non-zero values in the matrix $\mathbf{a}_i$. The 'edge weight' of $j \rightarrow i$ can be quantified by $||\mathbf{a}_{i,j}||_2$, $||\mathbf{a}_{i,j}||_\infty$, or $|\sum_l a_{i,j}(l)|$, the latter aiming to capture the net impact of gene j on gene i. In the default GLG setup, the individual weights in the l_1-constraint of the above equation are assigned the same value, with $\lambda_j = \lambda$. However, because we are not interested in the auto-regulation of x_i, we remove the sparsity constraint on the autoregressive edge ($\lambda_i = 0$) in order to reduce the number of false positives in the cross-regulatory relationships, where sparsity is typically enforced with a positive $\lambda_{j \neq i} = \lambda$.

The optimization problem in Equation 3 can be solved P separate times to infer the regulators of all P genes in the network. The GLG-identified regulators are obtained as the smallest group of genes whose past expression values are most predictive of gene i's time series expression values. Because the core algorithm of the GLG test is implemented using the *glmnet* package [79], it supports count-based expression data (e.g. from unique molecular identifiers) by assuming a Poisson distribution for the expression levels.

3.2. Single-Cell Inference of Networks using Granger Ensembles

In this section, we describe how the SCINGE algorithm, which has the GLG test at its core, infers gene regulatory networks from single-cell expression data. The SCINGE algorithm takes ordered single-cell data as input, with an optional zero-handling pre-processing step to mitigate the effect of dropouts. The data are analyzed using multiple GLG instances with different hyperparameters, each inferring possibly differently ranked regulator-gene interactions. These ranked inferences are aggregated using a modified Borda method, with an optional subsampling stage increasing the effective ensemble size for the aggregation step.

3.2.1. Ordered Single-Cell Data

The input to SCINGE is ordered single-cell data, with a pseudotime assigned to each cell that represents its position along the biological process. If the single-cell dataset is not already ordered, any cell-ordering method that assigns continuous pseudotimes (Section 1) can be used to order the data before providing it as input to SCINGE. We apply Monocle 2 [22], which uses reverse graph embedding to identify branching processes.

Given ordered single-cell data, the pseudotimes are first normalized to a scale of 0–100. Thus, the first cell represents 0% progress, and the last cell represents 100% progress through the biological process represented by the single-cell data. The distribution of cells' pseudotimes is not uniform. As a result, each gene's expression data is an irregularly-spaced time series in the pseudotemporal reference. We represent each gene's expression trend along the pseudotemporal reference as an augmented series with both the pseudotimes and the gene expression values. That is, for the i-th gene, we create the series (t_i, x_i), where x_i is the time-series representing the gene's expression and t_i represents the pseudotime of the corresponding cell. Currently, SCINGE only handles trajectories without major branches. See Section 4.3 for strategies to analyze branching processes.

3.2.2. Zero (Dropout) Handling

One of the most prominent technical artifacts in single-cell RNA-seq is dropout. This is manifested as a large number of zero readings due to inefficiencies in mRNA capture in the measurement process. Dropout causes the measured expression data to contain a higher number of zeros than the true biological zeros [80]. There have been efforts to overcome this problem by imputing the missing values [80, 81]. However, inappropriate imputation can negatively impact differential expression testing [82] and can have a positive, neutral, or negative effect on Monocle's pseudotimes depending on the choice of algorithm [83]. If we remove the zero-valued measurements altogether from the dataset, GLG effectively imputes the missing values without an external imputation algorithm by virtue of its kernel-based approach for analyzing irregular time series. Thus, depending on the severity of the dropout, SCINGE contains an optional step of removing some of the zeros and the corresponding pseudotime values. This can be achieved through an additional hyperparameter $prob\_zero\_removal$. For each gene, each zero-valued sample and its corresponding pseudotime are removed with probability $prob\_zero\_removal$.

3.2.3. Hyperparameter Diversity

The primary hyperparameters in the GLG tests include the sparsity constraint λ, the time resolution Δt between the elements of the vector $\mathbf{a}_{i,j}$, the length L of the vector $\mathbf{a}_{i,j}$ (which determines the extent of the lagged time series for the GLG analysis), and kernel width σ. The zero-handling stage introduces another optional hyperparameter $prob\_zero\_removal$.

If the process being studied is a stationary process containing simplistic regulatory networks, the above hyperparameters could potentially be tuned to optimize cross-validation performance. However, transcriptional regulation is non-linear and non-stationary in nature. A single GLG test, however optimal its settings, can produce false positives due to the assumption of linear and stationary causal relationships. In addition, there may not be a single set of hyperparameters that are optimal for all regulatory interactions. To overcome this, we analyze the data using multiple GLG tests with diverse hyperparameters and aggregate the rankings obtained from the individual GLG tests. Our assumption is that the top-ranked regulatory edges that consistently appear for many hyperparameter combinations are enriched for true positive interactions.

3.2.4. Subsampling Stage

This optional stage increases the effective ensemble size in SCINGE by obtaining subsampled versions of the original single-cell data. Specifically, for each hyperparameter combination above, we generate $N_{subsample}$ (default 10) data replicates by arbitrarily removing pseudotime-gene expression pairs (t_i, x_i) with probability of removal 0.2. Because each gene series is independently subsampled, we obtain uniquely irregular time series for each gene with a high probability. This also means that for any given cell, the probability of all genes' expression values from that cell being disregarded is extremely low. The SCINGE subsampling is similar to bagging [84] except that the sampling is without replacement and it uses a different aggregation approach.

3.2.5. GLG Runs and Modified Borda Aggregation

After enumerating all hyperparameter combinations and subsampled replicates, SCINGE runs GLG on each subsampled replicate using the different hyperparameter combinations. At the end of each GLG test, we obtain an adjacency matrix $\mathbf{A}$ using

$$\mathbf{A}_{ij} = \left| \sum_{l} \mathbf{a}_{i,j}(l) \right|,$$

where $\mathbf{a}$ is the $P \times P \times L$ coefficient matrix output from the GLG test. The matrix $\mathbf{A}$ represents one inference of the GRN, with the magnitude of each element representing the edge weight assigned to the corresponding regulator-gene interaction. These edge weights are used in forming a ranked list of the regulator-gene interactions. The ranking is assigned to only those interactions that correspond to a nonzero element of $\mathbf{A}$.

Once the rankings from the GLG tests on all hyperparameter combinations and subsampled replicates are obtained, we aggregate them using a modification of the Borda count aggregation method [85], which favors those edges that are consistently ranked high by multiple GLG tests over those that are ranked high only occasionally or not at all. The aggregation process involves assigning weights of $1/n^2$ for the n-th ranked interaction within each individual ranked list, with a weight of zero for an unranked regulator-gene interaction. The final SCINGE score of each interaction is obtained by summing the weights assigned to that interaction across all individual ranked lists. This score is subsequently used for the final GRN edge ranking. After the final ranking of regulator-gene interactions, we can also obtain the 'top

N regulators' of the biological process by summing the SCINGE scores of all outgoing edges for each regulator and sorting the regulators in order of decreasing magnitude.

3.2.6. Case Study Hyperparameters

Table 2 lists the hyperparameter values used to generate the GLG ensembles for both case studies. The subsampling stage creates 10 replicates for each hyperparameter setting by removing samples from individual gene expression values with probability of sample removal 0.2. Thus, not only is each time series irregular, but it has partially different time references compared to the other time series in the data set. For the main case studies, we use the default mode with $prob\_zero\_removal = 0$. All figures in Section 2 use this default setting, with the exception of Figure 9. The total number of GLG tests, accounting for hyperparameter diversity and subsampling, is

$$N = N_{\lambda}(5) \times N_{(\Delta t, L)}(5) \times N_{\sigma}(4) \times N_{subsample}(10) = 1000.$$

The subsampling approach and the consensus-rewarding nature of the modified Borda aggregation stage reduces the need to optimize the hyperparameter combinations for each dataset. The outputs from GLG tests on each subsampled replicate will have stronger consensus for a meaningful hyperparameter for the dataset.

Table 2: Hyperparameter combinations considered. λ is the sparsity parameter, Δt determines the time resolution, L represents the number of time lags under consideration, and σ represents the kernel width used for GLG. Only specific pairs of Δt and L are considered instead of all possible combinations.

Hyperparameter(s)	Values
λ	$0, 0.01, 0.02, 0.05, 0.1$
$(\Delta t, L)$	$(3, 5); (5, 9); (9, 5); (5, 15); (15, 5)$
σ	$0.5, 1, 2, 4$

3.3. Datasets

In this subsection, we describe the two single-cell datasets with which we evaluate the performance of SCINGE and compare it to existing GRN methods.

3.3.1. ESC to Endoderm Differentiation

The first dataset is obtained from Hayashi et al. [42], where single-cell RNA-seq data was collected from 456 cells at five time points over a 72 hour duration in which primitive endoderm cells were differentiated from mouse embryonic stem cells. Matsumoto et al. [12] used Monocle [22] to order these cells along the differentiating process, assigning a pseudotemporal reference to each cell in the process. We used their Monocle results in our analyses. The expression dataset is limited to 100 transcription factors exhibiting the highest variance in expression value and 356 cells.

3.3.2. Retinoic Acid-driven Differentiation

The second dataset was obtained from Semrau et al. [44], where SCRB-seq data was obtained at nine collection times during 96 hours from mouse embryonic stem cells differentiating into neuroectoderm and extraembryonic endoderm-like cells. We order the cells using Monocle 2 [22], with the ordering genes chosen by Monocle 2 in an unsupervised manner by identifying genes that are differentially expressed in response to the introduction of the growth medium. Although Matsumoto et al. [12] applied the original Monocle to the first dataset and we retain their pseudotimes, we prefer Monocle 2 for this case study, the most recent version available at the time of the analysis. Post ordering, we limit the scope of the analysis to the 1886 cells along the longest trajectory of the differentiation process (Figure S1) exhibiting non-trivial expression levels. Once Monocle 2 orders the cells along a pseudotemporal reference, it allows the analysis of cells to find genes that change in expression as cells progress along pseudotime. We shortlist the top 626 differentially expressed genes ($q < 10^{-5}$) along the pseudotime ranked by Monocle 2 for testing the GRN algorithms.

3.4. Evaluation

To test the regulatory network inferred by SCINGE and the other GRN methods, we use information from the ESCAPE database [43] as a gold standard, namely the cataloged ChIP-chip, ChIP-seq, loss-of-function (*lof*) and gain-of-function (*gof*) experiments. Each method ranks the possible edges in the network in order of confidence. We plot the respective precision-recall curves and compute the average precision (**A**) and early average precision (**E**) for comparison.

3.4.1. Existing GRN Methods

In addition to SCINGE, we use SCODE, SINCERITIES, and Jump3 to infer networks using the same data and evaluate their performance using ESCAPE. The SINCERITIES toolbox dated 16 December 2016 was obtained from `http://www.cabsel.ethz.ch/tools/sincerities.html` and the default settings were used for both the ESC to endoderm differentiation and retinoic acid-driven differentiation datasets. We downloaded SCODE from the GitHub repository `https://github.com/hmatsu1226/SCODE` (git commit 28acad67893c0fba7eeee670c339809d45ae6377) and used the same settings as in Matsumoto et al. [12] for the ESC to endoderm differentiation dataset with $D = 4$ degrees of freedom in the expression dynamics. We used $D = 20$ for the retinoic acid-driven differentiation dataset to account for the much larger network of 626 genes. An equivalent version of the Jump3 code we used can be obtained from `https://github.com/vahuynh/Jump3` (git commit 03a7e86d82f2383c56fd11c658dfce574fbf1a1a). In contrast to the other methods, Jump3 uses only ordering information. We used $noiseVar.obsnoise = 0.1$, but all other settings were the defaults. Because Jump3 did not terminate in a reasonable amount of time on the full retinoic acid-driven differentiation dataset, we reduced the dataset by arbitrarily dropping cells with probability 0.5. Despite this reduction in the data size, the Jump3 algorithm did not converge for two target genes, namely Tdh and Vdac1. As a result, we rank the corresponding edges at the bottom of the ranked list, which could affect the quality of the Jump3 results for the retinoic acid-driven differentiation dataset.

3.4.2. ESCAPE Database

The ESCAPE database [43] is a repository of data from numerous experiments conducted on human and mouse embryonic stem cell lines. Of particular interest to us are the gene interactions obtained from ChIP-chip/ChIP-seq experiments and loss-of-function/gain-of-function (*lof/gof*) experiments, which we use as a gold standard to evaluate the inferred GRNs. Despite ESCAPE being one of the most comprehensive repositories of such experimental results, it may not have reference data for all regulators under consideration. Therefore, we evaluate the inferred networks using the sub-matrix for which the gold standard is available.

To generate the gold standard, we combine all gene interactions in the ChIP-chip/ChIP-seq and *lof/gof* databases related to the genes from the single-cell data being analyzed. Gene interactions not documented in the

ESCAPE databases are assumed to not exist. However, this approach can lead to a high number of false zeros in the gold standard if a particular regulator was studied genome-wide. For example, whereas ESCAPE documents thousands of ChIP-chip/ChIP-seq interactions for most TFs, two of the TFs report less than 200 interactions. To avoid false zeros in the gold standard, we generate our gold standard using only regulators with at least 1000 gene interactions in the ChIP-chip/ChIP-seq database and 500 gene interactions in the *lof/gof* database.

3.4.3. Average Early Precision

Because a majority of SCINGE's hyperparameter sets predict a sparse regulatory network, it is better suited to rank the top gene interactions instead of ranking all of them. Average precision may not be the ideal performance metric for evaluating such methods. In addition, the top-ranked regulator-gene interactions are the most relevant for prioritizing experimental studies. Therefore, we also consider the average early precision, which evaluates the inferred network by calculating the average precision up to a partial recall threshold. We use a partial recall threshold of 0.1. That is, average early precision evaluates the ranking performance of GRN inference methods up to the point where they identify 10% of known gene interactions according to the gold standard.

3.4.4. KinderMiner and Gene Ontology Enrichment

We performed KinderMiner (v1.5.4) [64] analysis on the SCINGE top 20 regulators to search for known associations of these genes with the three keyphrases 'embryonic stem cells,' 'neural development,' and 'endoderm development' in a local collection of 26877474 PubMed abstracts downloaded from NCBI in December 2018. We report the statistically significant associations ($p < 10^{-4}$) in Table 1 using the labels 'ESC,' 'NeurDev,' and 'EndoDev,' respectively. The significance threshold corresponds to a family-wise error rate of $FWER < 6 \times 10^{-3}$, accounting for a family size of 60 gene-keyphrase pairs. In Supplemental File 4, we provide the raw KinderMiner results obtained using the search setting **anySpeciesSEP**. This corresponds to a species agnostic search in which words of keyphrase can be anywhere in the PubMed abstract.

We also performed functional profiling of the ordered 626-gene list from SCINGE using the g:GOSt tool in g:Profiler [45] version r1760_e93_eg40. We consider only Gene Ontology [86] biological process terms and specify 'mus musculus' as the organism. The candidate regulator list from SCINGE is ordered, so we use the 'ordered query' option, which allows g:Profiler to perform incremental enrichment analysis over the gene list. The significance threshold used was Fisher's one-tailed test, the default test for g:GOSt, with multiple testing correction using the default g:SCS method. Supplemental File 3 provides the complete output of the g:GOSt test. The significance test considers the entire ranked regulator list, but we highlight only the top 20 regulators in Table 1. In addition, we derived the loss-of-function phenotypes in Table 1 from the Mouse Genome Databases Mammalian Phenotype Ontology Annotations [87].

3.4.5. SCINGE Software Availability

A MATLAB implementation of SCINGE is available at `https://github.com/gitter-lab/SCINGE` under the MIT license and archived on Zenodo (`https://doi.org/10.5281/zenodo.2549817`). We used SCINGE version 0.1.0 for these analyses.

4. Discussion

SCINGE is a GRN reconstruction algorithm that adapts the Granger Causality test to detect dependencies in temporal data for single-cell gene expression dataset. It has the potential to prioritize regulators for future DNA-binding or functional studies. For example, in the retinoic acid-driven differentiation study, many of the top-ranked SCINGE regulators (Table 1) are enriched for relevant differentiation process and regulatory annotations but have not yet been characterized in the ESCAPE database.

When assessed in the retinoic acid-driven differentiation case study, in which none of the GRN methods' settings were tuned to optimize performance on this dataset, SCINGE has better precision-recall performance than three existing methods. However, we caution that single metrics like average precision can be misleading. Closer inspection reveals SCINGE's better performance is in part because it successfully prioritizes the regulators that are more dominant in the ESCAPE database. Because the precision-recall curve can mask near-random performance for many individual regulators, we recommend regulator-specific visualizations (Figures 6 and 7) to provide more context.

We designed SCINGE for a high-throughput computing environment, ensembling many GLG tests under different hyperparameters and using data subsampling to improve robustness and performance. This approach makes SCINGE more resilient to dropout in the single-cell gene expression data and less sensitive to the hyperparameter ranges tested. Ensembling strategies have proven effective in a variety of GRN inference settings, such as DREAM challenges [6]. Our use of modified Borda aggregation for ensembling emphasizes the top-ranked, most-confident predictions. Borda aggregation is also capable of ensembling the related networks we obtain from subsampling the same dataset. Unlike other unsupervised aggregation approaches [88], it does not assume they are conditionally independent.

4.1. Caveats and Limitations

The main assumption we make by using GLG is that the expression data are obtained from a linear and stationary VAR model. However, complex biological systems have dynamic, non-linear gene interactions and are expected to generate non-stationary expression trends. Violating the assumptions of linearity and stationarity can have a significant impact on the performance of individual GLG tests. Furthermore, Granger Causality tests result in false positives in scenarios with hidden variables [89]. However, these discrepancies between theory and practice are commonly accepted in biological applications of Granger Causality [38, 90]. In addition, SCINGE's Borda aggregation helps to push the most robust edges in the network to the top of the final ranked list of edges.

Some of the Granger Causality-related drawbacks potentially could be addressed by integrating SCINGE with complementary data types. GRN inference can be more accurate when using ChIP-chip, ChIP-seq, protein-protein interactions, regulator *lof/gof* experiments, or DNA binding motifs as prior knowledge on the network structure [91, 92] (reviewed in Chasman et al. [7]). Priors for single-cell GRN inference have been incorporated by scdiff [93], which uses TF-gene interactions, and SOMatic [94], which uses single-cell ATAC-seq. To model prior information in SCINGE, we could assign different penalty factors λ_j for the j-th regulator of target gene i based on the prior probability of the edge p_{ij}. An alternative would be to use SCINGE output in conjunction with the supplementary sources of information and aggregate all the information after-the-fact [95, 96]. In this version of SCINGE, we intentionally model only gene expression data. This makes SCINGE widely applicable in conditions and species where suitable priors are not available.

Another assumption of SCINGE, SINCERITIES, and SCODE is that the pseudotime values assigned to individual cells have a high fidelity. Our results show that in the ESC to endoderm differentiation dataset, incorporating pseudotime values improves the precision-recall performance for all three methods. However, in the retinoic acid-driven differentiation dataset, these methods perform as well or better in the absence of the pseudotime values (Figure 10). For this dataset, assigning uninformative pseudotime values to ordered cells is more detrimental to the network inference performance than simply using the order without pseudotimes (Figures S8–S10). The relatively poor quality of the pseudotimes, in the context of GRN reconstruction, could be attributed to the type of single-cell data in the retinoic acid-driven differentiation dataset or aspects of this biological process. We propose that pseudotimes' impact on GRN accuracy could be used to evaluate pseudotime inference algorithms, complementing other benchmarking metrics for pseudotimes [15]. Integrating GRN methods like SCINGE into the dynverse [15] benchmarking framework would enable us to systematically evaluate which types of pseudotimes best support network inference and empirically assess the types of GRN motifs that cannot be unambiguously recovered from single-cell expression data [97]. Qiu et al. [30] proposed that RNA velocity [98] may help overcome limitations of pseudotime for GRN reconstruction.

4.2. Benchmarking and Evaluation

Inferring gene regulatory networks only from single-cell gene expression data is a difficult task. An evaluation of network inference algorithms on simulated single-cell datasets reported that their performances were only slightly better than a random edge ordering [99]. Network inference accuracy on experimental datasets cannot be calculated perfectly because there is no comprehensive gold standard. However, it may be even worse than simulated performance due to the additional biological and technical noise and confounding encountered in real expression data.

An important aspect when evaluating network inference on experimental data is the relevance of the gold standard. In the SCODE evaluation [12], the gold standard was TF binding interactions estimated from DNaseI footprints and sequence motifs. However, it was merged across all human and mouse cell types instead of only those relevant to the mouse ESC to endoderm differentiation process. In the Chen and Mar benchmarking of stem cell datasets [99], the gold standard consisted of all interactions from the STRING database [100]. These included interaction types that are not directly informative about transcriptional regulation and were not limited to the specific cell types of interest. For our evaluation, we limit the gold standard to data from mouse embryonic stem cells obtained from ChIP-chip/ChIP-seq and *lof/gof* studies cataloged by ESCAPE [43], which are more relevant to the biological processes we study and more directly indicative of transcriptional relationships.

There remain open questions regarding appropriate evaluation methodologies. For example, we combine ChIP-chip/ChIP-seq and *lof/gof* information, but the precision-recall performance of the four GRN methods is quite different when examining ChIP-chip/ChIP-seq or *lof/gof* data alone (Figure S2). These two types of data are known to have low overlap [101], and our evaluation suggests the SCINGE's search for lagged gene expression dependencies may detect more indirect regulatory relationships than direct TF binding.

Although the GRN methods we evaluate have better than random average precision when assessing the entire network, they are only marginally better than random when ranking outgoing edges from individual regulators. For the regulator-specific early average precision, each GRN method is better than random for only some regulators. Precision-recall is preferable to the receiver operating characteristic for evaluating biological network inference due to the sparsity of the gold standard [102], but the average precision for the entire network may overestimate the utility of GRN inference methods for studying individual regulators.

4.3. Future Work and Extensions

The current version of SCINGE is limited to biological processes that have a single path in the trajectory, without any major branches. One way to infer the GRN from a branching process is to select the cells from each branched path and apply SCINGE to these datasets independently. A better approach would be to adapt SCINGE to treat each branch as a task in a multi-task GRN inference problem [103]. In addition, the kernel could be modified so that certain pseudotime intervals can be considered more informative, for example, the interval around a major bifurcation point.

SCINGE accommodates a common *prob_zero_removal* for all genes, but the algorithm can easily be modified to incorporate gene-specific zero removal probabilities. Future work may involve a more sophisticated zero-handling approach, which would remove only those zeros that are inconsistent with other non-zero measurements from similar cells. Methods like SCONE [76] can provide additional information for removing zeros more selectively than the current approach.

Other elements of the GLG regression framework can be adapted as well. Nguyen and Braun [104] place a monotonicity constraint on the coefficients $\mathbf{a}_{i,j}$ such that the more recent coefficients have higher magnitude than the more distant ones. Similarly, we could adapt the kernel to give higher weight to more recent samples in the pseudotime than more distant ones. Another possible direction involves exploration of the kernel-based generalizations to the Group Lasso [105, 106]. This would enable SCINGE to regularize all coefficients from a regulator as a group instead of treating different lagged coefficients as separate variables. In general, the kernel-based approach at the core of SCINGE provides great flexibility to adapt our GRN reconstruction algorithm to emphasize different aspects of dynamic biological processes.

Acknowledgements

We thank Christina Kendziorski for feedback on the SCINGE algorithm, Alireza Fotuhi Siahpirani for assistance with the ESCAPE database, Rafael Feliciano for discussion of the GRN evaluation, Scott Swanson for testing the SCINGE software, and members of the Gitter, Stewart, and Kendziorski research groups for their helpful comments.

Funding

This research was supported by NSF CAREER award DBI 1553206, the Wisconsin Head and Neck SPORE grant NIH P50 DE026787, NIH grants UH3 TR000506 and U01 HL099773, the Morgridge Institute for Research, a grant from Marv Conney to Ron Stewart, and the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation and is an active member of the Open Science Grid. The Open Science Grid is supported by NSF award 1148698 and the U.S. Department of Energy's Office of Science.

Supplementary Files

- Supplementary Information containing Figures S1 to S10
- Supplementary File 1. Predicted GRNs from SCINGE, SINCERITIES, Jump3, and SCODE on the ESC to endoderm differentiation dataset
- Supplementary File 2. Predicted GRNs from SCINGE, SINCERITIES, Jump3, and SCODE on the retinoic acid-driven differentiation dataset
- Supplementary File 3. List of genes ordered according to SCINGE influence for the retinoic acid-driven differentiation dataset with corresponding g:Profiler enrichment results
- Supplementary File 4. KinderMinder associations for the top 20 regulators according to SCINGE influence for the retinoic acid-driven differentiation dataset

References

[1] A. Tanay and A. Regev, "Scaling single-cell genomics from phenomenology to mechanism," *Nature*, vol. 541, no. 7637, p. 331, 2017.

[2] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome Research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[3] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell RNA-sequencing experiments," *Genome Biology*, vol. 17, no. 1, p. 63, 2016.

[4] M. W. Fiers, L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak, and S. Aerts, "Mapping gene regulatory networks from single-cell omics data," *Briefings in Functional Genomics*, 2018.

[5] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, no. 10, p. 717, 2010.

[6] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, A. Aderhold, R. Bonneau, Y. Chen, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, pp. 796–804, Aug 2012.

[7] D. Chasman, A. F. Siahpirani, and S. Roy, "Network-based approaches for analysis of complex biological systems," *Current Opinion in Biotechnology*, 2016.

[8] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene regulatory network inference from single-cell data using multivariate information measures," *Cell Systems*, vol. 5, no. 3, pp. 251–267, 2017.

[9] J. Intosalmi, H. Mannerstrom, S. Hiltunen, and H. Lahdesmaki, "SCHiRM: Single cell hierarchical regression model to detect dependencies in read count data," *bioRxiv*, 2018.

[10] Z. Bar-Joseph, A. Gitter, and I. Simon, "Studying and modelling dynamic biological processes using time-series gene expression data," *Nature Reviews Genetics*, vol. 13, no. 8, p. 552, 2012.

[11] V. A. Huynh-Thu and G. Sanguinetti, "Combining tree-based and dynamical systems for the inference of gene regulatory networks," *Bioinformatics*, vol. 31, no. 10, pp. 1614–1622, 2015.

[12] H. Matsumoto, H. Kiryu, C. Furusawa, M. S. Ko, S. B. Ko, N. Gouda, T. Hayashi, and I. Nikaido, "SCODE: An efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation," *Bioinformatics*, p. btx194, 2017.

[13] N. P. Gao, S. M. Ud-Dean, and R. Gunawan, "Gene regulatory network inference using time-stamped cross-sectional single cell expression data," *IFAC-PapersOnLine*, vol. 49, no. 26, pp. 147–152, 2016.

[14] A. Gitter, "Single-cell RNA-seq pseudotime estimation algorithms," *doi:10.5281/zenodo.1297422*, Jun 2018.

[15] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods: towards more accurate and robust tools," *bioRxiv*, 2018.

[16] N. Leng, L.-F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski, "Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments," *Nature Methods*, vol. 12, no. 10, pp. 947–950, 2015.

[17] Z. Liu, H. Lou, K. Xie, H. Wang, N. Chen, O. M. Aparicio, M. Q. Zhang, R. Jiang, and T. Chen, "Reconstructing cell cycle pseudo time-series via single-cell transcriptome data," *Nature Communications*, vol. 8, no. 1, p. 22, 2017.

[18] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Peer, "Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[19] J. Shin, D. Berg, Y. Zhu, J. Shin, J. Song, M. Bonaguidi, G. Enikolopov, D. Nauen, K. Christian, G.-l. Ming, and H. Song, "Single-cell RNA-Seq with Waterfall reveals molecular cascades underlying adult neurogenesis," *Cell Stem Cell*, vol. 17, pp. 360–372, Sep 2015.

[20] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.

[21] H. Matsumoto and H. Kiryu, "SCOUP: a probabilistic model based on the Ornstein–Uhlenbeck process to analyze single-cell expression data during differentiation," *BMC Bioinformatics*, vol. 17, no. 1, p. 232, 2016.

[22] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories," *Nature Methods*, vol. 14, no. 10, p. 979, 2017.

[23] J. Zhang, T. Zhou, and Q. Nie, "Topographer reveals dynamic mechanisms of cell fate decisions from single-cell transcriptomic data," *bioRxiv*, 2018.

[24] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan, "Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles," *Bioinformatics*, vol. 34, no. 2, pp. 258–266, 2017.

[25] P.-C. Aubin-Frankowski and J.-P. Vert, "Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference," *bioRxiv*, 2018.

[26] A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis, "Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data," *Bioinformatics*, vol. 31, no. 12, pp. i89–i96, 2015.

[27] J. Wei, X. Hu, X. Zou, and T. Tian, "Reverse-engineering of gene networks for regulating early blood development from single-cell measurements," *BMC Medical Genomics*, vol. 10, no. 5, p. 72, 2017.

[28] A. T. Specht and J. Li, "Leap: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering," *Bioinformatics*, vol. 33, no. 5, pp. 764–766, 2016.

[29] M. Sanchez-Castillo, D. Blanco, I. M. Tienda-Luna, M. Carrion, and Y. Huang, "A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data," *Bioinformatics*, vol. 34, no. 6, pp. 964–970, 2017.

[30] X. Qiu, A. Rahimzamani, L. Wang, Q. Mao, T. Durham, J. L. McFaline-Figueroa, L. Saunders, C. Trapnell, and S. Kannan, "Towards inferring causal gene regulatory networks from single cell expression measurements," *bioRxiv*, 2018.

[31] P. Tsakanikas, D. V. Manatakis, and E. S. Manolakos, "Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions," *bioRxiv*, 2018.

[32] T. E. Chan, A. Pallaseni, A. C. Babtie, K. McEwen, and M. P. Stumpf, "Empirical Bayes meets information theoretical network reconstruction from single cell data," *bioRxiv*, 2018.

[33] A. Bonnaffoux, U. Herbach, A. Richard, A. Guillemin, S. Giraud, P.-A. Gros, and O. Gandrillon, "Wasabi: a dynamic iterative framework for gene regulatory network inference," *bioRxiv*, 2018.

[34] P. Cordero and J. M. Stuart, "Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories," in *Pacific Symposium on Biocomputing 2017*, pp. 576–587, World Scientific, 2017.

[35] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[36] A. Fujita, P. Severino, J. R. Sato, and S. Miyano, "Granger causality in systems biology: Modeling gene networks in time series microarray data using vector autoregressive models," in *Brazilian Symposium on Bioinformatics*, pp. 13–24, Springer, 2010.

[37] N. D. Mukhopadhyay and S. Chatterjee, "Causality and pathway search in microarray time series experiment," *Bioinformatics*, vol. 23, no. 4, pp. 442–449, 2006.

[38] A. Shojaie and G. Michailidis, "Discovering graphical granger causality using the truncating lasso penalty," *Bioinformatics*, vol. 26, no. 18, pp. i517–i523, 2010.

[39] J. D. Finkle, J. J. Wu, and N. Bagheri, "Windowed Granger causal inference strategy improves discovery of gene regulatory networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 9, pp. 2252–2257, 2018.

[40] M. T. Bahadori and Y. Liu, "Granger causality analysis in irregular time series," in *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 660–671, 2012.

[41] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical Granger methods," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 66–75, ACM, 2007.

[42] 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*, vol. 9, no. 1, p. 619, 2018.

[43] H. Xu, C. Baroukh, R. Dannenfelser, E. Y. Chen, C. M. Tan, Y. Kou, Y. E. Kim, I. R. Lemischka, and A. Ma'ayan, "Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells," *Database*, vol. 2013, 2013.

[44] S. Semrau, J. E. Goldmann, M. Soumillon, T. S. Mikkelsen, R. Jaenisch, and A. Van Oudenaarden, "Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells," *Nature Communications*, vol. 8, no. 1, p. 1096, 2017.

[45] J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, and J. Vilo, "g:Profiler-a web server for functional interpretation of gene lists (2016 update)," *Nucleic Acids Research*, vol. 44, no. W1, pp. W83–W89, 2016.

[46] S. M. Morris, M. D. Tallquist, C. O. Rock, and J. A. Cooper, "Dual roles for the Dab2 adaptor protein in embryonic development and kidney transport," *The EMBO Journal*, vol. 21, pp. 1555–1564, Apr 2002.

[47] B. Feldman, W. Poueymirou, V. E. Papaioannou, T. M. DeChiara, and M. Goldfarb, "Requirement of FGF-4 for postimplantation mouse development," *Science*, vol. 267, no. 5195, pp. 246–249, 1995.

[48] I. Leaf, J. Tennessen, M. Mukhopadhyay, H. Westphal, and W. Shawlot, "Sfrp5 is not essential for axis formation in the mouse," *Genesis*, vol. 44, pp. 573–578, Dec 2006.

[49] C. Meno, K. Gritsman, S. Ohishi, Y. Ohfuji, E. Heckscher, K. Mochida, A. Shimono, H. Kondoh, W. S. Talbot, E. J. Robertson, A. F. Schier, and H. Hamada, "Mouse Lefty2 and Zebrafish Antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation," *Molecular Cell*, vol. 4, pp. 287–298, Sep 1999.

[50] J. R. Barrow and M. R. Capecchi, "Targeted disruption of the Hoxb-2 locus in mice interferes with expression of Hoxb-1 and Hoxb-4," *Development*, vol. 122, no. 12, pp. 3817–3828, 1996.

[51] E. E. Morrisey, Z. Tang, K. Sigrist, M. M. Lu, F. Jiang, H. S. Ip, and M. S. Parmacek, "GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo," *Genes & Development*, vol. 12, pp. 3579–3590, Nov 1998.

[52] G. L. Radice, H. Rayburn, H. Matsunami, K. A. Knudsen, M. Takeichi, and R. O. Hynes, "Developmental defects in mouse embryos lacking N-Cadherin," *Developmental Biology*, vol. 181, pp. 64–78, Jan 1997.

[53] W. C. Skarnes, B. Rosen, A. P. West, M. Koutsourakis, W. Bushell, V. Iyer, A. O. Mujica, M. Thomas, J. Harrow, T. Cox, D. Jackson, J. Severin, P. Biggs, J. Fu, M. Nefedov, P. J. de Jong, A. F. Stewart, and A. Bradley, "A conditional knockout resource for the genome-wide study of mouse gene function," *Nature*, vol. 474, pp. 337–342, Jun 2011.

[54] J. Parant, A. Chavez-Reyes, N. A. Little, W. Yan, V. Reinke, A. G. Jochemsen, and G. Lozano, "Rescue of embryonic lethality in Mdm4-null mice by loss of Trp53 suggests a nonoverlapping pathway with MDM2 to regulate p53," *Nature Genetics*, vol. 29, pp. 92–95, Sep 2001.

[55] G. E. Olson, J. C. Whitin, K. E. Hill, V. P. Winfrey, A. K. Motley, L. M. Austin, J. Deal, H. J. Cohen, and R. F. Burk, "Extracellular glutathione peroxidase (Gpx3) binds specifically to basement membranes of mouse renal cortex tubule cells," *American Journal of Physiology-Renal Physiology*, vol. 298, pp. F1244–F1253, May 2010.

[56] T. M. DeChiara, A. Efstratiadis, and E. J. Robertsen, "A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting," *Nature*, vol. 345, pp. 78–80, May 1990.

[57] P. Sicinski, J. L. Donaher, Y. Geng, S. B. Parker, H. Gardner, M. Y. Park, R. L. Robker, J. S. Richards, L. K. McGinnis, J. D. Biggers, J. J. Eppig, R. T. Bronson, S. J. Elledge, and R. A. Weinberg, "Cyclin D2 is an FSH-responsive gene involved in gonadal cell proliferation and oncogenesis," *Nature*, vol. 384, pp. 470–474, Dec 1996.

[58] Y. Xiao, H. Ma, P. Wan, D. Qin, X. Wang, X. Zhang, Y. Xiang, W. Liu, J. Chen, Z. Yi, and L. Li, "Trp-Asp (WD) repeat domain 1 is essential for mouse peri-implantation development and regulates Cofilin phosphorylation," *The Journal of Biological Chemistry*, vol. 292, pp. 1438–1448, Jan 2017.

[59] T. Sakai, S. Li, D. Docheva, C. Grashoff, K. Sakai, G. Kostka, A. Braun, A. Pfeifer, P. D. Yurchenco, and R. Fässler, "Integrin-linked kinase (ILK) is required for polarizing the epiblast, cell adhesion, and controlling actin accumulation," *Genes & Development*, vol. 17, pp. 926–40, Apr 2003.

[60] J. Egea, C. Erlacher, E. Montanez, I. Burtscher, S. Yamagishi, M. Hess, F. Hampel, R. Sanchez, M. T. Rodriguez-Manzaneque, M. R. Bösl, R. Fässler, H. Lickert, and R. Klein, "Genetic ablation of FLRT3 reveals a novel morphogenetic function for the anterior visceral endoderm in suppressing mesoderm differentiation," *Genes & Development*, vol. 22, pp. 3349–62, Dec 2008.

[61] A. C. Carpenter, S. Rao, J. M. Wells, K. Campbell, and R. A. Lang, "Generation of mice with a conditional null allele for Wntless," *Genesis*, vol. 48, pp. 554–558, Aug 2010.

[62] Y. Wang, N. Thekdi, P. M. Smallwood, J. P. Macke, and J. Nathans, "Frizzled-3 is required for the development of major fiber tracts in the rostral CNS," *Journal of Neuroscience*, vol. 22, pp. 8563–73, Oct 2002.

[63] P. Gorry, T. Lufkin, A. Dierich, C. Rochette-Egly, D. Décimo, P. Dollé, M. Mark, B. Durand, and P. Chambon, "The cellular retinoic acid binding protein I is dispensable," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, pp. 9032–6, Sep 1994.

[64] F. Kuusisto, J. Steill, Z. Kuang, J. Thomson, D. Page, and R. Stewart, "A simple text mining approach for ranking pairwise associations in biomedical applications," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2017, pp. 166–174, 2017.

[65] T. Kunath, M. K. Saba-El-Leil, M. Almousailleakh, J. Wray, S. Meloche, and A. Smith, "FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment," *Development*, vol. 134, no. 16, pp. 2895–2902, 2007.

[66] Y. Yamanaka, F. Lanner, and J. Rossant, "FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst," *Development*, vol. 137, no. 5, pp. 715–724, 2010.

[67] D. Krawchuk, N. Honma-Yamanaka, S. Anani, and Y. Yamanaka, "FGF4 is a limiting factor controlling the proportions of primitive endoderm and epiblast in the ICM of the mouse blastocyst," *Developmental Biology*, vol. 384, no. 1, pp. 65–71, 2013.

[68] X. Zhang, A. Friedman, S. Heaney, P. Purcell, and R. L. Maas, "Meis homeoproteins directly regulate Pax6 during vertebrate lens morphogenesis," *Genes & Development*, vol. 16, no. 16, pp. 2097–2107, 2002.

[69] M. T. Pankratz, X.-J. Li, T. M. LaVaute, E. A. Lyons, X. Chen, and S.-C. Zhang, "Directed neural differentiation of human embryonic stem cells via an obligated primitive anterior stage," *Stem Cells*, vol. 25, no. 6, pp. 1511–1520, 2007.

[70] D. Shimosato, M. Shiki, and H. Niwa, "Extra-embryonic endoderm cells derived from ES cells induced by GATA factors acquire the character of XEN cells," *BMC Developmental Biology*, vol. 7, no. 1, p. 80, 2007.

[71] M. P. Stavridis, J. S. Lunn, B. J. Collins, and K. G. Storey, "A discrete period of FGF-induced Erk1/2 signalling is required for vertebrate neural specification," *Development*, vol. 134, no. 16, pp. 2889–2894, 2007.

[72] K. R. Finley, J. Tennessen, and W. Shawlot, "The mouse secreted frizzled-related protein 5 gene is expressed in the anterior visceral endoderm and foregut endoderm during early post-implantation development," *Gene Expression Patterns*, vol. 3, no. 5, pp. 681–684, 2003.

[73] K. Takaoka, H. Nishimura, and H. Hamada, "Both nodal signalling and stochasticity select for prospective distal visceral endoderm in mouse embryos," *Nature Communications*, vol. 8, no. 1, p. 1492, 2017.

[74] K. Q. Cai, C. D. Capo-Chichi, M. E. Rula, D.-H. Yang, and X.-X. Xu, "Dynamic GATA6 expression in primitive endoderm formation and maturation in early mouse embryogenesis," *Developmental Dynamics*, vol. 237, no. 10, pp. 2820–2829, 2008.

[75] J. Artus, P. Douvaras, A. Piliszek, J. Isern, M. H. Baron, and A.-K. Hadjantonakis, "BMP4 signaling directs primitive endoderm-derived XEN cells to an extraembryonic visceral endoderm identity," *Developmental Biology*, vol. 361, no. 2, pp. 245–262, 2012.

[76] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, "Normalization of RNA-seq data using factor analysis of control genes or samples," *Nature Biotechnology*, vol. 32, no. 9, pp. 896–902, 2014.

[77] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, and R. Quick, "The open science grid," *Journal of Physics: Conference Series*, vol. 78, p. 012057, Jul 2007.

[78] R. A. Erickson, M. N. Fienen, S. G. McCalla, E. L. Weiser, M. L. Bower, J. M. Knudson, and G. Thain, "Wrangling distributed computing for high-throughput environmental science: An introduction to HTCondor," *PLoS Computational Biology*, vol. 14, no. 10, p. e1006468, 2018.

[79] J. Qian, T. Hastie, J. Friedman, R. Tibshirani, and N. Simon, "GLMNET for MATLAB." `http://www.stanford.edu/~hastie/glmnet_matlab/`, 2013.

[80] G. C. Linderman, J. Zhao, and Y. Kluger, "Zero-preserving imputation of scRNA-seq data using low-rank approximation," *bioRxiv*, 2018.

[81] D. van Dijk, J. Nainys, R. Sharma, P. Kathail, A. J. Carr, K. R. Moon, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er, "MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data," *BioRxiv*, 2017.

[82] T. Andrews and M. Hemberg, "False signals induced by single-cell imputation [version 1; referees: 4 approved with reservations]," *F1000Research*, vol. 7, no. 1740, 2018.

[83] L. Zhang and S. Zhang, "Comparison of computational methods for imputing single-cell RNA-sequencing data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2018.

[84] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, Aug 1996.

[85] M. van Erp and L. Schomaker, "Variants of the Borda count method for combining ranked classifier hypotheses," in *Proceedings 7th International Workshop on Frontiers in Handwriting Recognition (7th IWFHR)* (L. Schomaker and L. Vuurpijl, eds.), pp. 443–452, International Unipen Foundation, 2000.

[86] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, pp. 25–29, May 2000.

[87] C. J. Bult, J. A. Blake, C. L. Smith, J. A. Kadin, J. E. Richardson, and the Mouse Genome Database Group, "Mouse Genome Database (MGD) 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D801–D806, 2019.

[88] M. E. Ahsen, R. Vogel, and G. Stolovitzky, "Unsupervised evaluation and weighted aggregation of ranked predictions," *arXiv*, Feb 2018.

[89] M. T. Bahadori and Y. Liu, "An examination of practical Granger causality inference," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 467–475, 2013.

[90] P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 360, no. 1457, pp. 969–981, 2005.

[91] A. F. Siahpirani and S. Roy, "A prior-based integrative framework for functional transcriptional regulatory network inference," *Nucleic Acids Research*, vol. 45, no. 4, pp. e21–e21, 2016.

[92] A. Greenfield, C. Hafemeister, and R. Bonneau, "Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks," *Bioinformatics*, vol. 29, no. 8, pp. 1060–1067, 2013.

[93] J. Ding, B. J. Aronow, N. Kaminski, J. Kitzmiller, J. A. Whitsett, and Z. Bar-Joseph, "Reconstructing differentiation networks and their regulation from time series single-cell expression data," *Genome Research*, 2018.

[94] C. Jansen, R. Ramirez, N. El-Ali, D. Gomez-Cabrero, J. Tegner, M. Merkenschlager, A. Conesa, and A. Mortazavi, "Building gene regulatory networks from single-cell ATAC-seq and RNA-seq using linked self-organizing maps," *bioRxiv*, 2018.

[95] M. Ciofani, A. Madar, C. Galan, M. Sellars, K. Mace, F. Pauli, A. Agarwal, W. Huang, C. N. Parkurst, M. Muratet, K. M. Newberry, S. Meadows, A. Greenfield, Y. Yang, P. Jain, F. K. Kirigin, C. Birchmeier, E. F. Wagner, K. M. Murphy, R. M. Myers, R. Bonneau, and D. R. Littman, "A validated regulatory network for Th17 cell specification," *Cell*, vol. 151, no. 2, pp. 289–303, 2012.

[96] D. Marbach, S. Roy, F. Ay, P. E. Meyer, R. Candeias, T. Kahveci, C. A. Bristow, and M. Kellis, "Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks," *Genome research*, vol. 22, pp. 1334–49, Jul 2012.

[97] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein, "Fundamental limits on dynamic inference from single-cell snapshots," *Proceedings of the National Academy of Sciences*, vol. 115, no. 10, pp. E2467–E2476, 2018.

[98] A. Gioele, L. Manno, R. Soldatov, H. Hochgerner, A. Zeisel, Z. Liu, D. V. Bruggen, J. Guo, E. Sundström, G. Castelo-branco, P. Cramer, I. Adameyko, and S. Linnarsson, "RNA velocity in single cells," *Nature*, vol. 560, no. 7719, pp. 494–8, 2018.

[99] S. Chen and J. C. Mar, "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data," *BMC Bioinformatics*, vol. 19, no. 1, p. 232, 2018.

[100] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, Oct 2014.

[101] A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, I. Simon, and Z. Bar-Joseph, "Backup in gene regulatory networks explains differences between binding and knockout results," *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[102] M. Schrynemackers, R. Kueffner, and P. Geurts, “On protocols and measures for the validation of supervised methods for the inference of biological networks,” *Frontiers in Genetics*, vol. 4, p. 262, 2013.

[103] D. M. Castro, N. de Veaux, E. R. Miraldi, and R. Bonneau, “Multi-study inference of regulatory networks for more accurate models of gene regulation,” *bioRxiv*, 2019.

[104] P. Nguyen and R. Braun, “Time-lagged Ordered Lasso for network inference,” *BMC Bioinformatics*, vol. 19, p. 545, Dec 2018.

[105] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[106] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, “Grouped graphical Granger modeling for gene expression regulatory networks discovery,” *Bioinformatics*, vol. 25, no. 12, pp. i110–i118, 2009.

Supplementary Information

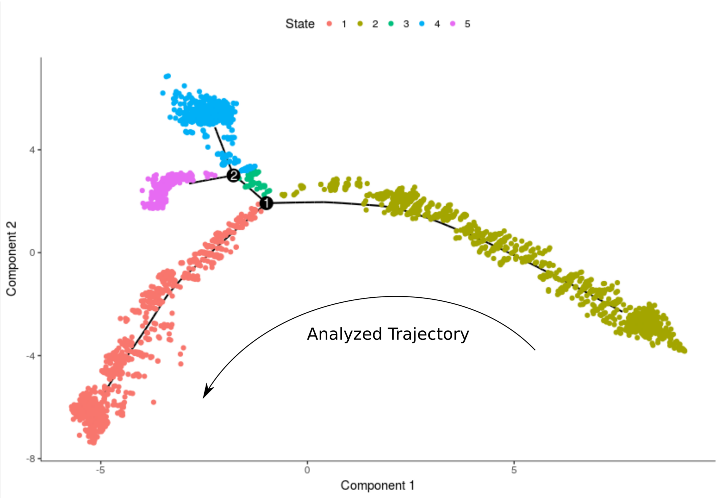

Figure S1: Monocle 2 trajectory of the retinoic acid-driven differentiation process. The trajectory constituting states 2 (early part comprising mostly data collected at 0h) and 1 (later part including cells collected at 96h) was analyzed using SCINGE and other network inference methods.

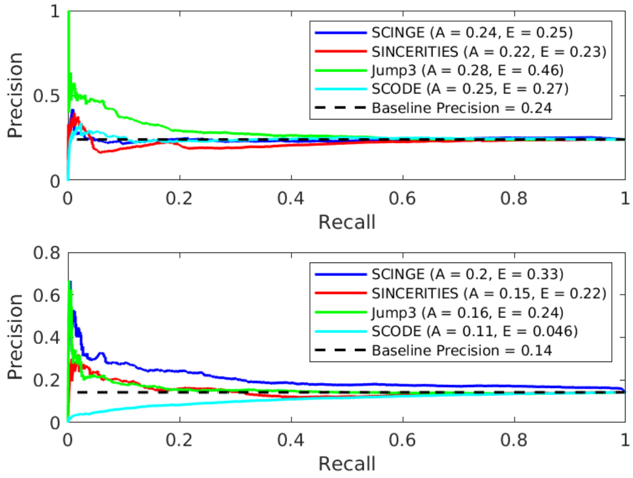

Figure S2: Precision-recall performance of network inference methods on the retinoic acid-driven differentiation dataset for ESCAPE gold standard interactions from (a) ChIP-chip/ChIP-seq and (b) *lof/gof* studies.

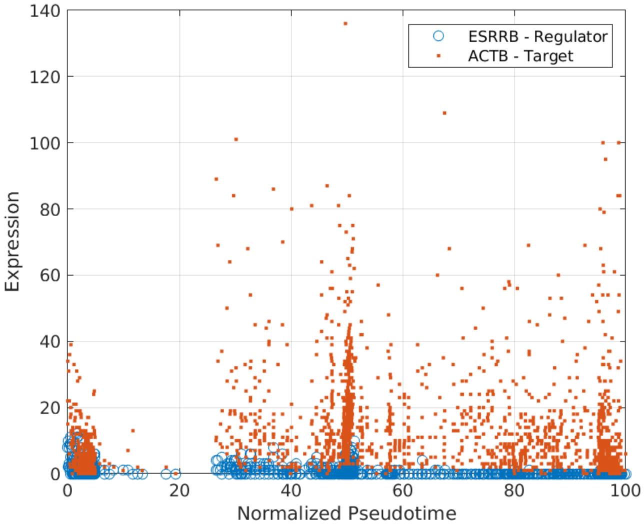

Figure S3: Expression trends of Esrrb and Actb show no apparent lag between regulator (Esrrb) and target expression (Actb). The interaction Esrrb→Actb is ranked highly by SCODE but not by SCINGE.

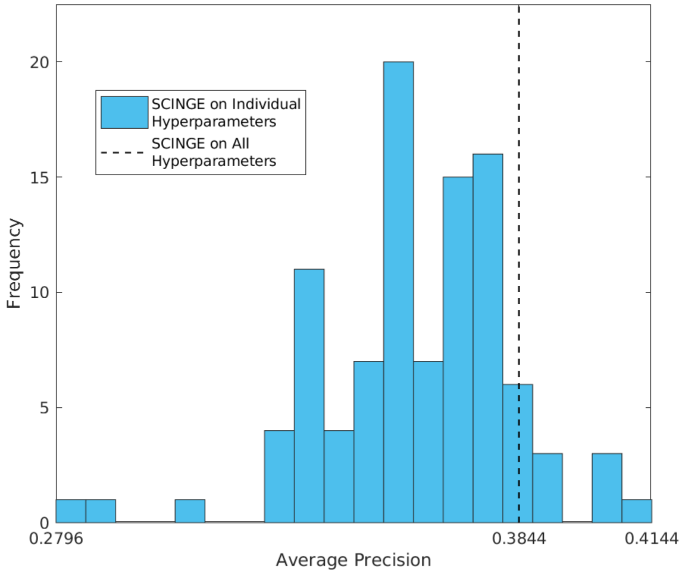

Figure S4: Histogram of average precision for individual hyperparameters for the ESC to endoderm differentiation dataset.

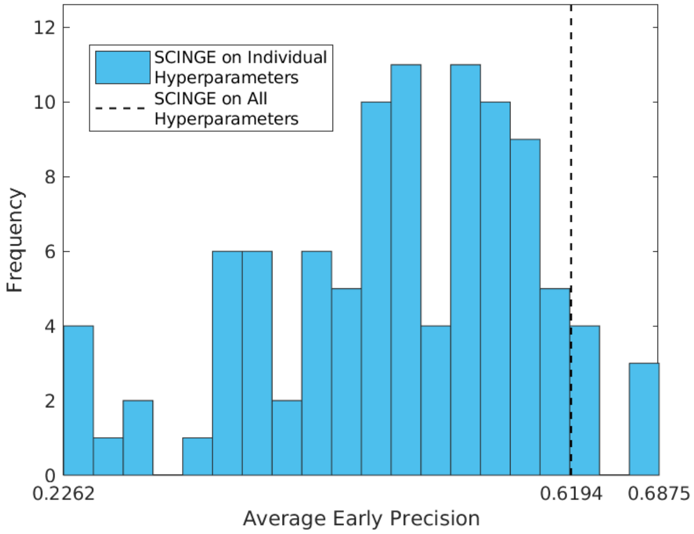

Figure S5: Histogram of average early precision for individual hyperparameters for the ESC to endoderm differentiation dataset.

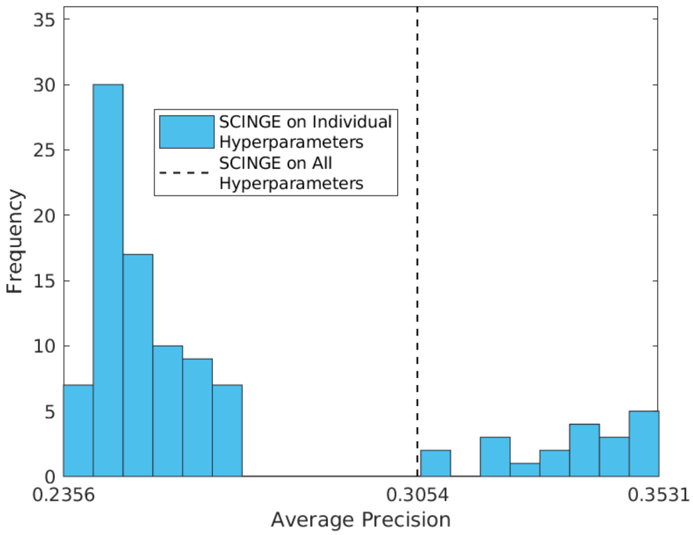

Figure S6: Histogram of average precision for individual hyperparameters for the retinoic acid-driven differentiation dataset.

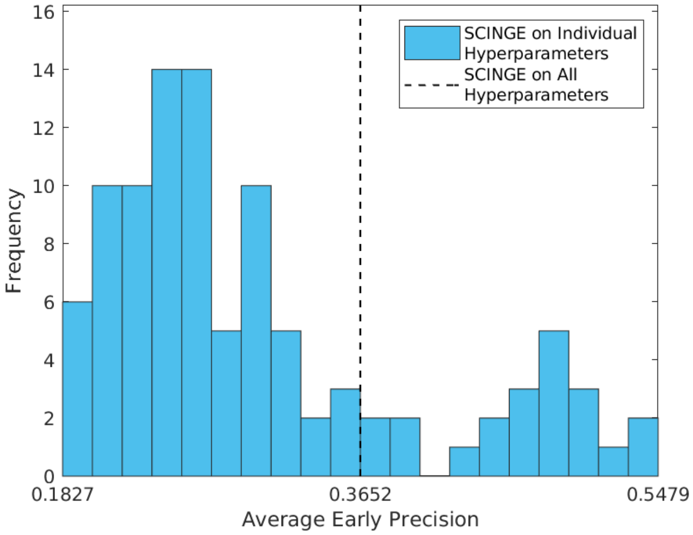

Figure S7: Histogram of average early precision for individual hyperparameters for the retinoic acid-driven differentiation dataset.

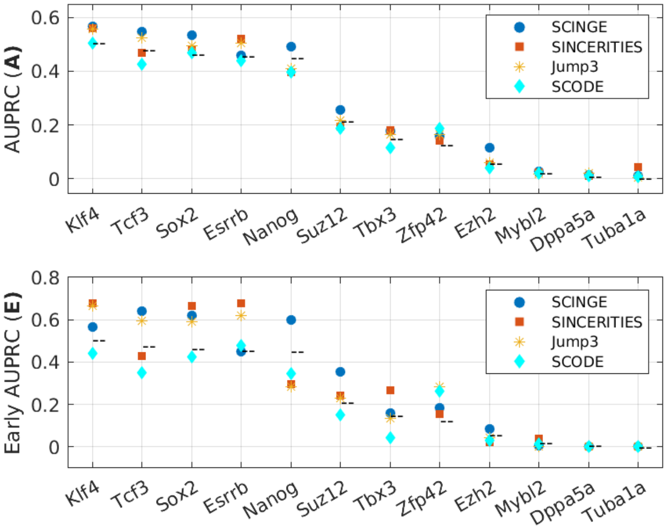

Figure S8: Average precision and average early precision evaluated for individual regulators for rankings obtained using the *Order Only* dataset. The dashed line (−−) indicates random performance. This shows that the performance of SCINGE and SINCERITIES in Figure 6 is hampered due to the use of unreliable pseudotimes. The Jump3 results are the same as in Figure 6 because it does not use pseudotime values.

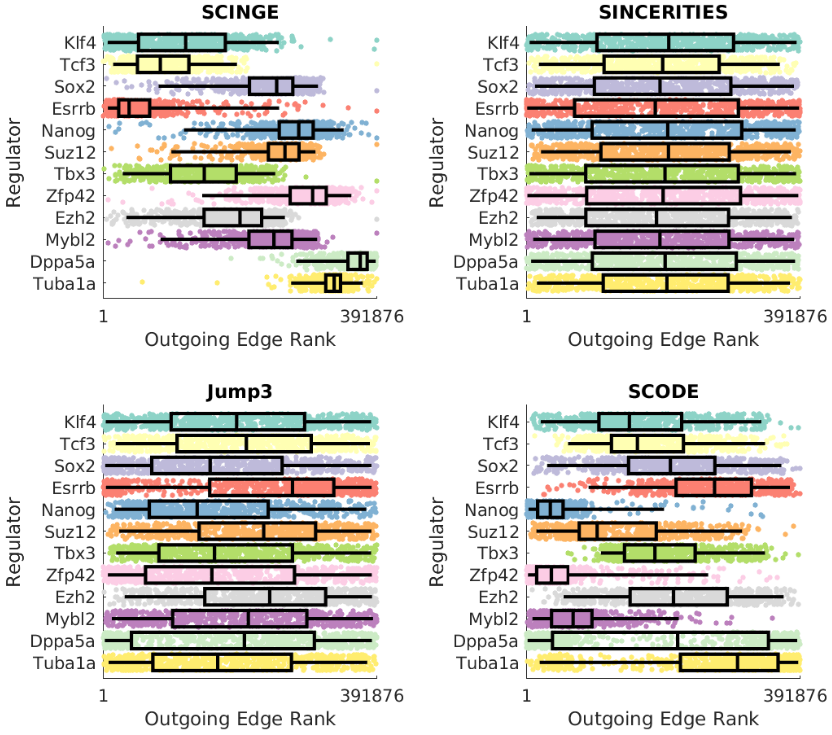

Figure S9: Boxplots of outgoing edge ranks for each regulator in each predicted GRN obtained using the *Order Only* dataset. The Jump3 results are the same as in Figure 7 because it does not use pseudotime values.

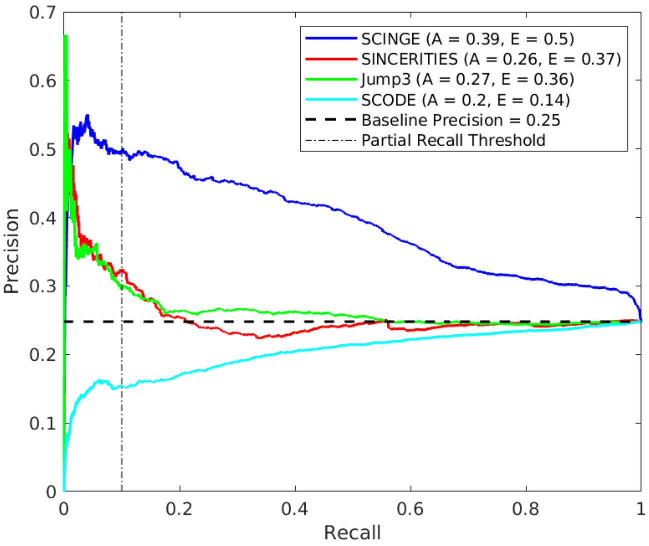

Figure S10: Precision-recall performance comparison of the four methods when using *Order Only* dataset. Key: **A** - Average Precision, **E** - Average Early Precision (≤ 0.1 recall). The Jump3 results are the same as in Figure 5 because it does not use pseudotime values.