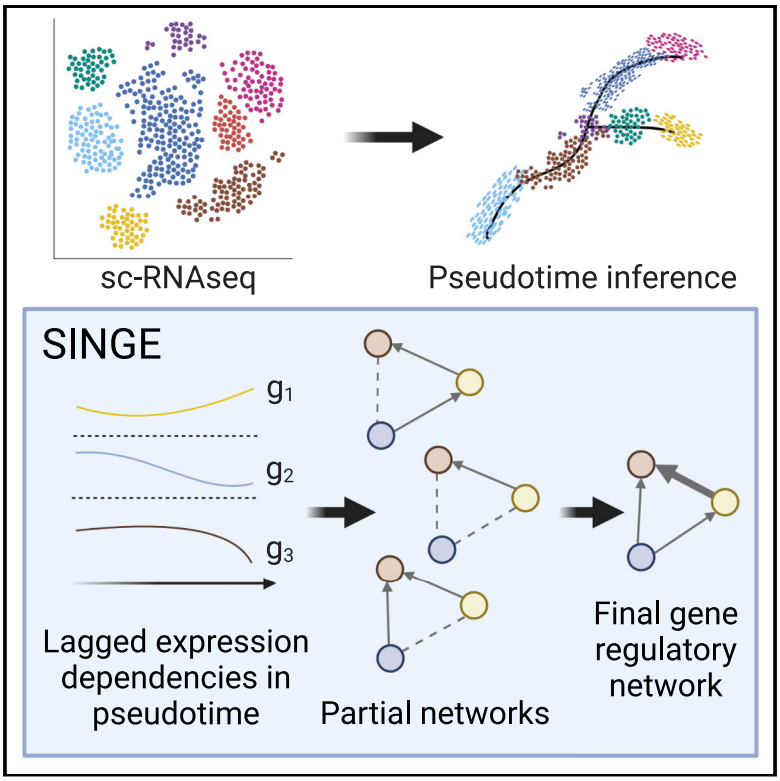
Resource

Cell Reports

Network inference with Granger causality ensembles on single-cell transcriptomics

Graphical abstract



Authors

Atul Deshpande, Li-Fang Chu, Ron Stewart, Anthony Gitter

Correspondence

gitter@biostat.wisc.edu

In brief

Deshpande et al. present SINGE, an algorithm to infer gene regulatory networks from ordered single-cell gene expression data. SINGE uses kernel-based regression to smooth noisy, ordered single-cell data and ensembling to prioritize reliable regulatory relationships.

Highlights

- Pseudotime estimates order cells in a dynamic process using single-cell gene expression
- SINGE infers gene regulatory networks from gene expression trends over pseudotime
- SINGE's ensembling considers many smoothed versions of irregular pseudotemporal data
- Uninformative pseudotime values can be detrimental to network reconstruction

Deshpande et al., 2022, Cell Reports *38*, 110333
February 8, 2022
https://doi.org/10.1016/j.celrep.2022.110333

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Network inference with Granger causality ensembles on single-cell transcriptomics

Atul Deshpande,[1,2,4] Li-Fang Chu,[2,5] Ron Stewart,[2] and Anthony Gitter[2,3,6,*]
[1]Department of Electrical and Computer Engineering, University of Wisconsin – Madison, Madison, WI 53706, USA
[2]Morgridge Institute for Research, Madison, WI 53715, USA
[3]Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison, Madison, WI 53792, USA
[4]Present address: Department of Oncology, Johns Hopkins University, Baltimore, MD 21205, USA
[5]Present address: Department of Comparative Biology and Experimental Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada
[6]Lead contact
*Correspondence: gitter@biostat.wisc.edu
https://doi.org/10.1016/j.celrep.2022.110333

SUMMARY

Cellular gene expression changes throughout a dynamic biological process, such as differentiation. Pseudotimes estimate cells' progress along a dynamic process based on their individual gene expression states. Ordering the expression data by pseudotime provides information about the underlying regulator-gene interactions. Because the pseudotime distribution is not uniform, many standard mathematical methods are inapplicable for analyzing the ordered gene expression states. Here we present single-cell inference of networks using Granger ensembles (SINGE), an algorithm for gene regulatory network inference from ordered single-cell gene expression data. SINGE uses kernel-based Granger causality regression to smooth irregular pseudotimes and missing expression values. It aggregates predictions from an ensemble of regression analyses to compile a ranked list of candidate interactions between transcriptional regulators and target genes. In two mouse embryonic stem cell differentiation datasets, SINGE outperforms other contemporary algorithms. However, a more detailed examination reveals caveats about poor performance for individual regulators and uninformative pseudotimes.

INTRODUCTION

Identifying the underlying gene regulatory networks (GRNs) that dictate cell fate decisions is important for understanding biological systems. Although RNA sequencing (RNA-seq) experiments on populations of cells 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 (scRNA-seq), have enabled observation of the gene expression states of individual cells (Tanay and Regev, 2017; Trapnell, 2015; Bacher and Kendziorski, 2016). Although 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 (Fiers et al., 2018; Blencowe et al., 2019; Nguyen et al., 2021), taking advantage of the large sample sizes. In the strictest sense, GRNs only include regulation of genes by transcription factors (TFs). However, we use the term GRN to mean a network of directed causal relationships between any regulators (not necessarily TFs) and their target genes. GRN inference requires identifying relationships between transcriptional regulators and their target genes or gene modules (De Smet and Marchal, 2010; Marbach et al., 2012a; Chasman et al., 2016). 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 (Chan et al., 2017), linear regression (Intosalmi et al., 2018), or other approaches. Methods like GENIE3 (Huynh-Thu et al., 2010), which was originally designed to infer GRNs from bulk transcriptomics data using tree-based ensembles, can be easily adapted for single-cell datasets. When single-cell expression data are collected at multiple time points, they provide more information that can be used for GRN inference. GRN reconstruction methods originally designed for bulk time-series transcriptomics data (Bar-Joseph et al., 2012) can be repurposed to analyze time-stamped single-cell data. For example, Jump3 (Huynh-Thu and Sanguinetti, 2015), a hybrid machine learning and model-based approach, has been adapted in this manner (Matsumoto et al., 2017). Time-stamped single-cell data also enable analyzing the evolution of gene expression distributions over time (Papili Gao et al., 2017), 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 a greater understanding of the causal regulatory relationships between genes. The dozens of algorithms for ordering cells and assigning pseudotimes (Gitter, 2018), also referred to as trajectory inference, can be distinguished by their use of prior knowledge, treatment of pseudotime uncertainty, and the supported trajectory types (Saelens et al., 2019). Pseudotime algorithms can target cyclic (Leng et al., 2015; Liu et al., 2017), linear (Bendall et al., 2014; Shin et al., 2015), bifurcating (Setty et al., 2016), multifurcating (Matsumoto and Kiryu, 2016), or tree-structured (Qiu et al., 2017; Zhang et al., 2018) trajectories. With most of these methods, a numeric pseudotime that represents the cell's progress along the trajectory is assigned to each cell.

Similar to time series data, 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 (Papili Gao et al., 2017), originally designed to infer GRNs using Granger causality-inspired ridge regression on time-stamped expression data, also admits pseudotime-labeled cells. SCODE (Matsumoto et al., 2017), GRISLI (Aubin-Frankowski and Vert, 2020), and Ocone et al. (2015) infer GRNs by modeling the cell dynamics as ordinary differential equations with pseudotime as the temporal reference. Other strategies involve Gaussian process regression for smoothing pseudotemporal data (Wei et al., 2017), time-lagged correlation (Specht and Li, 2016), variational Bayesian inference on a first-order autoregressive moving average model (Sanchez-Castillo et al., 2017), modified restricted directed information (Qiu et al., 2020), unsupervised classification using Gaussian mixture models (Tsakanikas et al., 2018), empirical Bayes-based thresholding (Chan et al., 2018), modeling information propagation through genes as a cascade (Bonnaffoux et al., 2019), and transfer entropy (Kim et al., 2021). 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 (Matsumoto and Kiryu, 2016) or Gaussian mixtures with continuous parameters (Cordero and Stuart, 2017). 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 (SINGE) algorithm, an ensemble-based GRN reconstruction technique that uses modified Granger causality on single-cell data annotated with pseudotimes. Granger causality (Granger, 1969, 1980) is a powerful approach for detecting specific types of causal relationships in long time series data. It has been used with bulk times series gene expression data (Fujita et al., 2010; Mukhopadhyay and Chatterjee, 2006; Shojaie and Michailidis, 2010; Finkle et al., 2018; Heerah et al., 2021; Lu et al., 2021), but these time series are typically short because of experimental limitations, making it more difficult to detect reliable gene-gene dependencies. The longer (pseudo)time series obtained from 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. SINGE 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 SINGE to reconstruct GRNs of two mouse embryonic stem cell (ESC) differentiation processes characterized with single-cell RNA-seq. SINGE compares favorably with existing GRN inference methods when evaluated using chromatin immunoprecipitation sequencing (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.

RESULTS

SINGE and Granger causality overview

SINGE 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 far that cell has progressed through a dynamic biological process such as differentiation. For each target gene, SINGE assesses which past expression values are most predictive of its expression; that is, the candidate regulators of each gene. The lagged dependencies are detected using a specialized form of Granger causality that is framed as a regularized regression problem. The past expression values are determined using the pseudotimes.

The Granger causality (Granger, 1969, 1980) test at SINGE'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 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 (Qiu et al., 2020). SINGE instead uses an alternative solution proposed by Bahadori and Liu, the generalized lasso Granger (GLG) test (Bahadori and Liu, 2012). GLG modifies the Lasso Granger test (Arnold et al., 2007) to support irregular time series. Within SINGE, 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.

SINGE depends on hyperparameters that control the kernel smoothing, sparsity, and which window of previous expression

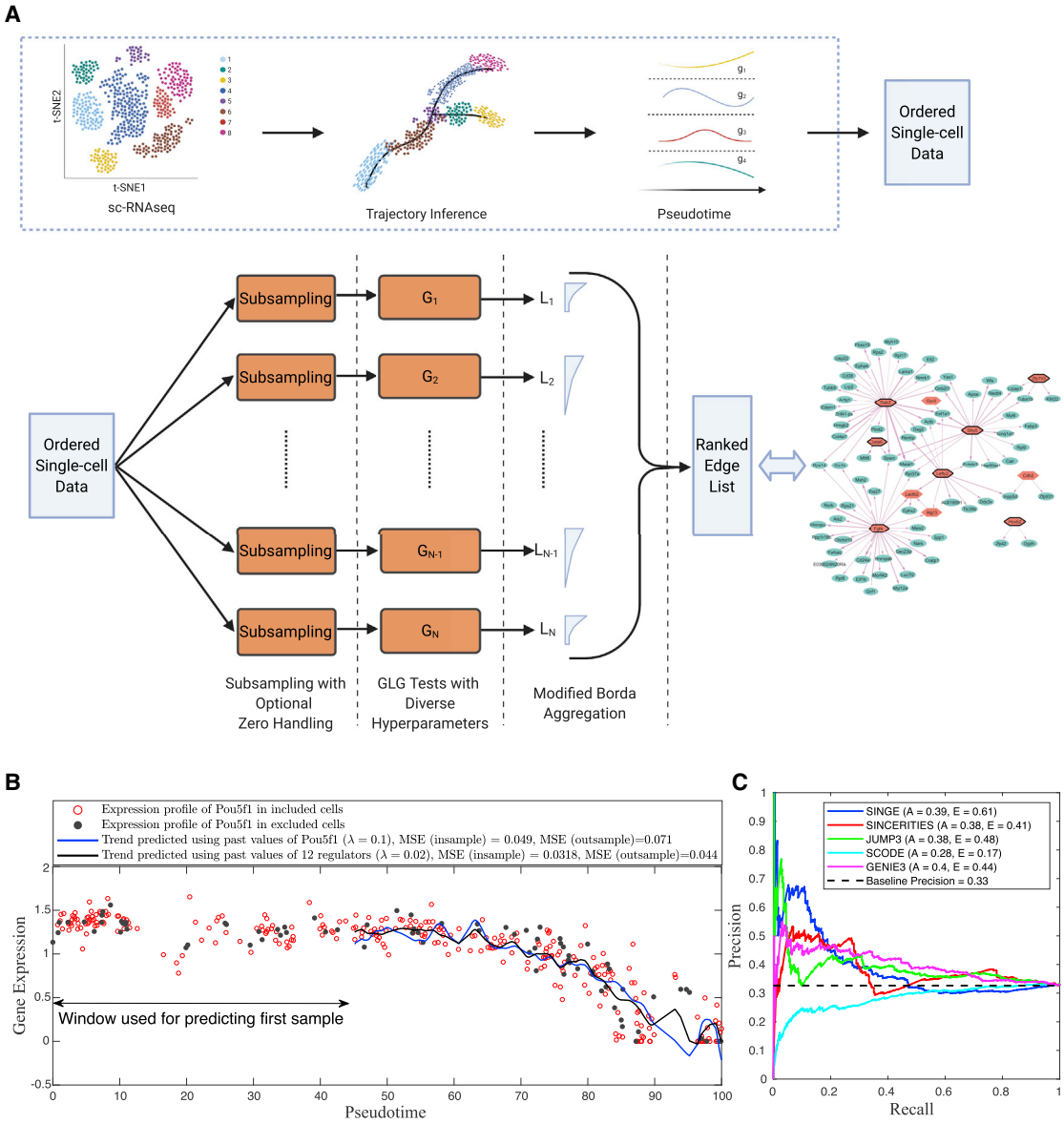


Figure 1. SINGE overview

(A) SINGE takes pseudotemporally ordered single-cell gene expression data as input and predicts a ranked list of regulator-target gene interactions. SINGE conducts multiple GLG tests with different hyperparameter combinations to control network sparsity, kernel smoothing, pseudotemporal resolution, and history for the GLG tests. Partial networks from individual GLG tests are aggregated to obtain an ensemble GRN prediction.

(B) Using past information from more GLG-identified regulators improves the predicted expression trend of Pou5f1 along pseudotime. The GLG model trains on insample cells and is evaluated with outsample cells.

(C) Precision-recall evaluation of SINGE and four other methods on ESC-to-endoderm differentiation data shows SINGE tied with Jump3 and SINCERITIES as having best average precision. SINGE has better average early precision than all other methods. Baseline precision is the expected precision from a random ordering of all regulator-target interactions. A, average precision; E, average early precision ($\leq$ 0.1 recall).

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 SINGE network is obtained from an ensemble of all of the individual predicted networks using different hyperparameters and cell subsamples (Figure 1).

Network inference case studies

Mouse embryonic stem cell-to-endoderm differentiation

Our first application tracks the differentiation of mouse ESCs to primitive endoderm cells over 72 h (Hayashi et al., 2018; STAR Methods). We use this small dataset with 100 TFs and 356 cells to optimize and tune aspects of the SINGE algorithm, such as the

modified Borda aggregation (STAR Methods). We assess how well it recovers known regulator-gene interactions that are relevant in mouse ESC differentiation from the embryonic stem cell atlas from pluripotency evidence (ESCAPE) database (Xu et al., 2013). The ESCAPE gold standard is incomplete because of a lack of experimental data for many of the relevant TFs (STAR Methods). Therefore, the gold standard only contains an 11 × 99 subset of the 100 × 99 regulator-gene interactions that SINGE scores. SINGE does not score self-edges.

The SINGE regulatory network is a ranked list of scored regulator-gene interactions (Deshpande and Gitter, 2021). SINGE ranks Foxd3, Gli2, and Nanog as the three most influential regulators in the 100-gene subnetwork. To illustrate a GLG-inferred regulatory edge, we consider Pou5f1 as an example target gene (Figure 1). We divide the cells into disjoint insample and outsample groups. We predict Pou5f1 expression using a GLG model trained on the insample cells and evaluate it on the outsample cells. As the regularization strength λ is reduced, the number of selected regulators increases, and the insample mean-squared error (MSE) and outsample MSE of Pou5f1 expression decrease (Figure 1). Setting $\lambda = 0.1$ identifies only one regulator, Pou5f1, whereas $\lambda = 0.02$ selects 12 regulators, including Pou5f1. The trend in MSE is consistent for additional values of the regularization strength, with $\lambda = 0.05$ producing an insample MSE of 0.045 and outsample MSE of 0.053 and $\lambda = 0.01$ generating an insample MSE of 0.027 and outsample MSE of 0.035.

To assess whether SINGE can match or exceed the state-of-the-art performance after dataset-specific tuning, we compare its predicted GRN with four existing network inference methods: SINCERITIES (Papili Gao et al., 2017), which uses ridge regression motivated by Granger causality; SCODE (Matsumoto et al., 2017), which is based on ordinary differential equations; Jump3 (Huynh-Thu and Sanguinetti, 2015), based on decision trees on temporal transcriptomics data; and its predecessor, GENIE3 (Huynh-Thu et al., 2010), the best-performing method in the dialogue on reverse engineering assessment and methods (DREAM) 4 *in silico* multifactorial challenge (DREAM Challenges, 2009), which does not use temporal information. We emphasize that this particular evaluation is not indicative of which method would perform best on new data because of SINGE's tuning. Nevertheless, SINGE is effectively tied for the highest average precision with Jump3 and SINCERITIES and has much higher average early precision than all other methods (Figure 1C). Average early precision emphasizes the most confident, top-ranked interactions (STAR Methods). Even though SCODE has been evaluated previously using these gene expression data (Matsumoto et al., 2017), it performs worse than random when assessed using the condition-specific ESCAPE gold standard.

Mouse retinoic acid-driven differentiation

We further test SINGE on a second dataset that tracks retinoic acid-driven differentiation from mouse ESCs to extraembryonic endoderm and neuroectoderm cells over 96 h (Semrau et al., 2017). SINGE is not tuned for this dataset or the subsequent applications. It uses the same algorithm and hyperparameters from the ESC-to-endoderm differentiation analysis. We infer a trajectory for the differentiation process using Monocle 2 (Qiu et al., 2017) and select 1,886 cells from cell states 1 and 2 (Figure S1A; STAR Methods). 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. SINGE returns a ranked list of all 626 × 625 possible regulatory relationships, excluding self-edges (Deshpande and Gitter, 2021).

SINGE identifies key regulators reflecting the differentiation trajectory required for mouse ESCs exiting the pluripotent state, transitioning through the epiblast, where lineage segregations take place (Semrau et al., 2017; Table 1; Figure 2). We use g:Profiler (Reimand et al., 2016) to identify Gene Ontology (GO) biological process terms that are significantly enriched among the ranked SINGE regulators (Deshpande and Gitter, 2021). 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 (Table 1).

There are two ways to explore the SINGE predictions in greater detail: the top regulators ranked by SINGE influence (Table 1), which aggregates influence over all target genes, and the top-ranked edges (Figures 2 and S1B). Table 1 shows the top 20 regulators. Ten of the top predicted regulators are associated with regulation of gene expression (GO:0010468), as are other regulators with high SINGE influence that are beyond the top 20 (Deshpande and Gitter, 2021). 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 (Kuusisto et al., 2017) text mining reveals significant associations between the top 20 regulators and terms related to this developmental process: "ESCs," "neural development," and "endoderm development" (Deshpande and Gitter, 2021).

Figure 2 shows high-confidence regulator-gene edges from the SINGE network, directed from the regulators (hexagons) to the target genes (ellipses). This representative subnetwork comprises 11 unique regulators and 76 unique targets. All 11 regulators are also found among the top 20 regulators by SINGE influence (Table 1), of which seven regulators are known to be associated with regulation of gene expression. Dab2 and Fgf4 are the most influential regulators overall (Table 1) and hub regulators among the high-confidence edges (Figure 2). Fgf4 governs exit from the pluripotent state. *Fgf4*-null mouse ESCs resist neural and mesodermal lineage induction (Kunath et al., 2007). 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) cause implantation lethality and lack of primitive endoderm development (Yamanaka et al., 2010). Moreover, Fgf4 also governs neural induction in ESC differentiation at a later stage of development (Krawchuk et al., 2013). Interactions targeting Rn45s are

Table 1. GO biological process terms, loss-of-function phenotypes, and KinderMiner associations related to the top 20 SINGE regulators

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 (Morris et al., 2002)	ESC, EndoDev
2	Fgf4	✓		✓		EL (Feldman et al., 1995)	ESC, EndoDev, NeurDev
3	Sfrp5	✓		✓	✓	normal (Leaf et al., 2006)	EndoDev
4	Lefty2	✓				EL (Meno et al., 1999)	ESC, EndoDev
5	Zfp703	✓		✓	✓	N/A	
6	Hoxb2	✓				NL (Barrow and Capecchi, 1996)	ESC, NeurDev
7	Gata6	✓		✓		EL (Morrisey et al., 1998)	ESC, EndoDev
8	Cdh2		✓		✓	EL (Radice et al., 1997)	ESC, NeurDev
9	Alg13					EL (Skarnes et al., 2011)	
10	Mdm4	✓				EL (Parant et al., 2001)	
11	Gpx3					others (Olson et al., 2010)	
12	Igf2	✓				others (DeChiara et al., 1990)	ESC, EndoDev, NeurDev
13	Ccnd2					S (Sicinski et al., 1996)	ESC
14	Wdr1		✓			EL (Xiao et al., 2017)	
15	Ilk	✓	✓		✓	EL (Sakai et al., 2003)	ESC
16	Flrt3		✓			EL (Egea et al., 2008)	EndoDev
17	Lactb2					others (Sollars et al., 2002)	
18	Wls				✓	EL (Carpenter et al., 2010)	
19	Fzd3		✓			NL (Wang et al., 2002)	ESC, NeurDev
20	Crabp1					normal (Gorry et al., 1994)	ESC, NeurDev

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; ESC, embryonic stem cell; NeurDev, neural development; EndoDev, endoderm development.

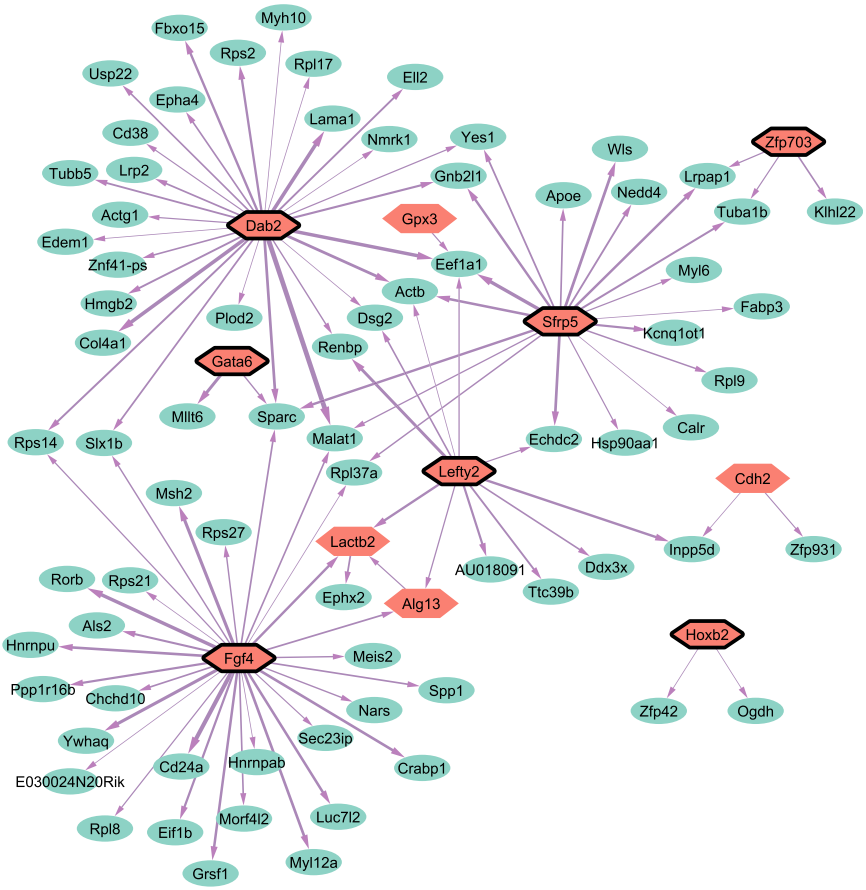


Figure 2. The network obtained from the top 100 edges ranked according to SINGE scores after removing edges involving Rn45s

These top SINGE predictions contain 11 unique regulators (hexagonal nodes; the seven with solid boundaries correspond to known regulators of gene expression listed in Table 1 and 76 unique targets, including Alg13 and Lactb2, which act as regulators and targets. The higher-ranked edges are represented by thicker arrows. See also Figures S1 and S2.

frequently highly ranked by SINGE, but we exclude them from the network in Figure 2 to emphasize more biologically relevant predictions. These Rn45s edges are likely false positive predictions. Rn45s is 45S pre-ribosomal RNA, and its expression levels and variance are much higher than any of the other 625 genes in this dataset. Figure S1B shows the top 100 edges from SINGE, including those targeting Rn45s.

The predicted GRN in Figure 2 also provides hypotheses for future experimental tests. For example, Meis1 and Meis2 are homeobox proteins that directly regulate Pax6 expression during eye development (Zhang et al., 2002). SINGE predicts that Fgf4 regulates Meis2. Thus, Fgf4 could potentially act upstream of Meis1 and Meis2 to regulate Pax6 expression, contributing to neuroectoderm differentiation (Pankratz et al., 2007). Other key primitive endoderm regulators are also highlighted in SINGE predictions, such as Gata6, a TF necessary and sufficient for primitive endoderm lineage differentiation and establishment of extraembryonic endoderm cell lines (Shimosato et al., 2007). Dab2, Sfrp5, and Lefty2 are all expressed in the primitive endodermal lineages, including visceral endoderm and extraembryonic endoderm cell lines (Stavridis et al., 2007; Finley et al., 2003; Takaoka et al., 2017; Cai et al., 2008).

Inspecting the regulator and target expression trends can build confidence in the predicted interactions. For example, in the predictions Dab2 → Yes1 and Fgf4 → Meis2 (Figure 2), we observe that the target gene expression is negatively correlated with the past values of the regulator's expression (Figure S2), which may indicate that these predictions are worth further investigation. In other cases, like Dab2 → Rn45s (Figure S1B), there is no obvious relationship between the regulator and target expression (Figure S2C). As noted above, this is likely a false positive prediction because of Rn45s's outlier expression levels and role as a pre-ribosomal gene.

Many expected GO terms and regulators are represented in Table 1 and Figure 2. However, classic neuroectoderm regulators like Sox1, Nes, and Pax6 (Semrau et al., 2017) are missing because they are excluded from the limited shortlist of genes in the SINGE input. We only run SINGE on the top 626 significantly differentially expressed genes along the differentiation trajectory detected by Monocle 2.

Retinoic acid ESCAPE evaluation

The retinoic acid study can be used to benchmark the relative performance of SINGE and other network inference methods because none of the methods, including SINGE, were optimized or tuned based on the ESCAPE evaluation results. Figure 3A shows the precision-recall performance of SINGE, SINCERITIES, Jump3, SCODE, and GENIE3 when ranking edges in the 626-gene network. Because of Jump3's runtime, we run it on a reduced dataset (STAR Methods), which may affect 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 × 625 possible edges. SINGE, SINCERITIES, and Jump3 have the best average early precision, but SINGE stands out from the other methods in its average precision (Figure 3A). SINCERITIES prioritizes ESCAPE gold standard interactions well at the top of its ranked list, but the performance degrades quickly. Jump3 is effectively tied with SINGE for average early precision but has a near-random precision for recall of greater than 0.2. GENIE3 and SCODE are worse than random. The performance depends on the type of regulator-gene interaction in the ESCAPE database. SINGE can recover loss-of-function (LoF) or gain-of-function (GoF) relationships but struggles to identify

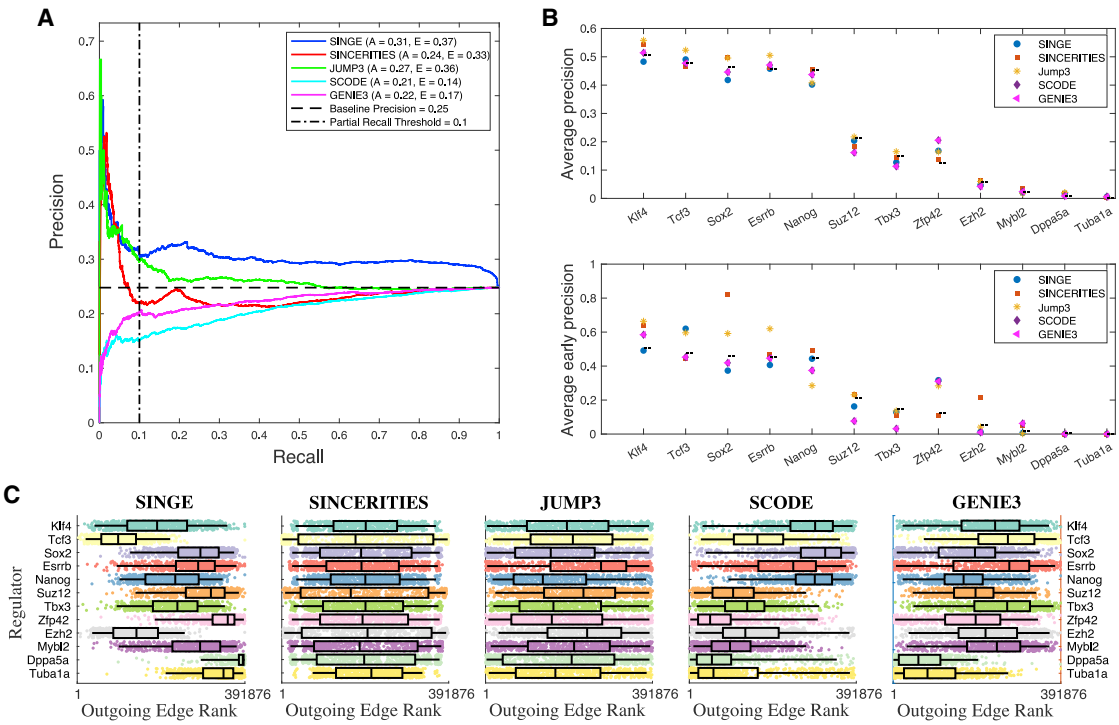


Figure 3. Evaluation of the predicted GRNs for the retinoic acid dataset

(A) Precision-recall performance of SINGE, SINCERITIES, Jump3 (which uses a reduced dataset), SCODE, and GENIE3 when predicting a 626-gene retinoic acid regulatory network. SINGE, SINCERITIES, and Jump3 have the best average early precision, but SINGE has better average precision.

(B) Average precision and average early precision evaluated for individual regulators in the ESCAPE database. The dashed line indicates the expected performance of a random ranking, given by the ratio (total number of true outgoing edges)/(total number of genes − 1). Regulator-specific performance of all five methods is near or below random for most regulators.

(C) The ranking of regulator-specific interactions has a strong effect on the overall precision-recall curve. 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 converse is also true.

See also Figures S3 and S5.

ChIP-based protein-DNA binding interactions (Figure S3). In contrast, Jump3 and GENIE3 recover ChIP-based protein-DNA binding interactions quite well but struggle to identify LoF/GoF relationships.

Visualizing the expression trends over pseudotime can show the types of errors SINGE makes with respect to the ESCAPE gold standard. For example, the interaction Esrrb → Actb was detected with ChIP but is not part of ESCAPE's LoF/GoF dataset. There is no apparent lag between the expression trends of the regulator and target (Figure S2D). This edge was ranked highly by SCODE but not by SINGE, which searches for lagged expression dependencies by design.

A regulator-specific evaluation partially explains the overall precision-recall performance of the GRN methods and demonstrates that it can be somewhat misleading. Figure 3B shows the average precision and average early precision with respect to each regulator in the ESCAPE database. These metrics are obtained from the regulator-specific precision-recall curves in a manner similar to the average precision and average early precision obtained in Figure 3A. The regulator-specific average precision of all five methods is at or below random for most regulators, with a few exceptions.

Because some regulators are more prevalent in the ESCAPE gold standard than others, the overall precision-recall curve is influenced by the 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 is a proxy for the regulator's influence on the evaluation, and generate boxplots of the regulator-specific edge ranks in the GRNs (Figure 3C). SINGE 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. SCODE and GENIE3 rank edges from the regulators with fewer outgoing edges higher than those with many target genes, contributing to their poor overall performance.

These regulator-specific results provide insights into Figure 3A. SINGE's relatively high average early 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, such as Sox2 and Esrrb. From the perspective of individual regulators, all five methods have near-random performance for most regulators (Figure 3B), but this does not necessarily translate into near-random precision-recall performance for the entire GRN (Figure 3A). This is because some of the methods rank certain regulator-specific interactions above others (Figure 3C), either to their benefit (SINGE) or to their detriment (SCODE and GENIE3).

Mouse bone marrow mesenchyme-to-erythrocyte differentiation

As an additional SINGE case study, we choose an scRNA-seq dataset from the Mouse Cell Atlas, profiling the heterogeneity of adult mouse bone marrow (Han et al., 2018). This dataset helps assess SINGE's scalability to more genes (3,025 genes) and cells (3,105 cells). We also generate the pseudotimes (Figure S4) from an alternative trajectory inference method, Embeddr (Campbell et al., 2015). We hypothesize that the bone marrow scRNA-seq data should shed light on regulators of hematopoiesis or its associated diseases. Indeed, SINGE identifies relevant regulators in this context. For example, Asxl2 and Rtel1 are among the top 20 regulators (Deshpande and Gitter, 2021). Asxl2 knockout mice display a phenotype that skews the differentiation potential of hematopoietic stem cells and causes myeloid-lineage cancer (Li et al., 2016). In humans, Asxl2 mutation is known to be associated with acute myeloid leukemia (Micol et al., 2017). Rtel1 is a DNA helicase important for protecting telomeres. Its mutations are associated with a number of clinical phenotypes related to bone marrow failure (Marsh et al., 2018; Balakumaran et al., 2015). Although other top regulators do not have direct experimental data to show their role in hematopoiesis, the SINGE GRN provides a base for future investigations.

Dyngen simulation evaluation

We use dyngen (Cannoodt et al., 2021) to evaluate GRN inference on simulated single-cell datasets. We simulate single-cell gene expression data from a biological process with a linear trajectory. This simulated dataset has 1,000 cells generated from a regulatory network with 140 genes: 25 TFs, 15 housekeeping genes, and 100 target genes. We use SINGE and the four other GRN algorithms to infer networks from this dataset. We first evaluate the precision-recall performance of each inferred network using the known direct regulatory interactions (Deshpande and Gitter, 2021). Most methods, including SINGE (average precision, 0.0082; average early precision, 0.0048), perform as poorly as the random baseline precision of 0.0084, with Jump3 performing best (average precision, 0.015; average early precision, 0.032). If we include indirect regulator-gene interactions in the gold standard, the precision-recall performance of SINGE improves, surpassing the random baseline and other GRN methods, which remain near or worse than random. Thus, in the dyngen simulation, SINGE performs poorly at distinguishing direct interactions from indirect interactions. This is in part because the simulated gene expression trends of a regulator's direct and indirect targets can be quite similar, as exemplified by the cascade from B1_TF1 to Target1 to Target48 (Deshpande and Gitter, 2021).

Analyzing features of the SINGE workflow

We use the retinoic acid dataset to analyze various SINGE features. This is the largest real dataset from our case studies that has a gold standard available.

Effects of subsampling and zero handling

SINGE's ensembling can 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 a probability of removal of 0.2. The default SINGE setting uses 10 subsampled replicates per hyperparameter combination. The average precision and average early precision have only modest improvements when using more than five replicates (Figure 4A). With less than five replicates, the runtime improves considerably, but the average early precision decreases. Reducing the number of replicates from the default 10 to five maintains similar average early precision and still reduces the runtime.

The support for irregular time series also allows us to remove zero-valued data points that may correspond to technical dropouts. As a proof of concept for zero handling, SINGE uses a dropout probability hyperparameter, *prob-zero-removal*, for all genes. For each GLG instance, we remove zero-valued expression samples (and their corresponding timestamps) from each gene's expression series with a user-specified constant probability for each zero value. If the true dropout probabilities for each gene were estimated instead, then this strategy could be modified accordingly.

Figure 4B shows SINGE's precision-recall summaries as the value of *prob-zero-removal* increases. As more zeros are dropped from the dataset, the average precision and especially the average early precision are only marginally affected. Zero dropping could potentially be used to effectively reduce the size of the regression problem for large but extremely sparse datasets without negatively affecting GRN inference. However, filtering too many zeros in such an arbitrary manner could remove genuine zero expression values along with the dropouts. We currently recommend using SINGE without dropping zeros unless it is required to speed up analysis of large datasets.

Benefits of ensembling

The optimal GLG parameters that best identify relationships between 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 set optimal hyperparameters for the GLG test SINGE uses. SINGE attempts to overcome this with an ensemble of reasonable hyperparameters, aggregating the results to obtain the final SINGE score of each GRN edge (STAR Methods). Figure S5A compares the performance of individual GLG hyperparameter combinations with the complete ensembled SINGE GRN for the retinoic acid dataset. Although the ensembled SINGE network does not have 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. Inspecting the performance of the GRNs for individual hyperparameters

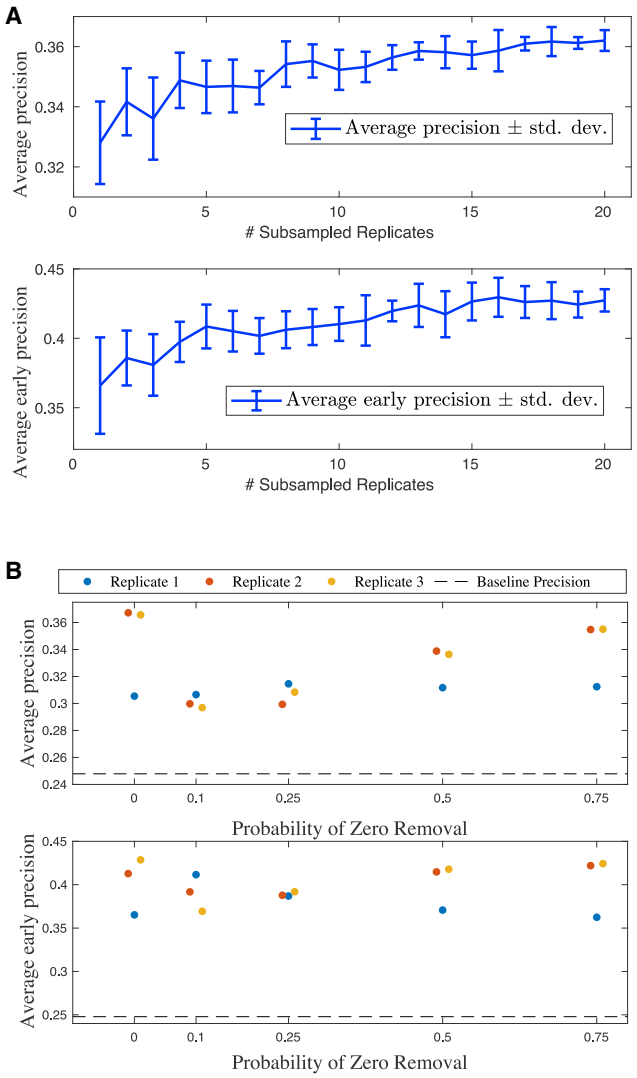


Figure 4. Effect of subsampling and zero removal on GRN reconstruction

(A) Effect of the number of subsampled replicates for each hyperparameter combination in the SINGE ensemble on the retinoic acid dataset.

(B) SINGE precision-recall performance for multiple values of *prob-zero-removal*, the probability of removing a zero value. Zero removal has a limited effect on SINGE's precision-recall performance on the retinoic acid dataset. On this dataset, SINGE's runtime can be improved by dropping large proportions of zero-valued samples with a negligible effect on performance.

(Figures S5B–S5F) shows that the sparsity hyperparameter λ has the strongest effect.

Assessing whether pseudotimes improve GRN reconstruction

We assess how the estimated pseudotime values affect the three methods designed to reconstruct GRNs from pseudotemporal single-cell gene expression: SINGE, SINCERITIES, and SCODE. We exclude GENIE3 and Jump3 because the former does not use any cell ordering information, and the latter uses only the cell ordering but not pseudotime values. For this assessment, we create variants of the ESC-to-endoderm differentiation and retinoic acid datasets as follows:

- *Pseudotime*: the default mode using ordered cells with pseudotimes assigned by an algorithm like Monocle.
- *Order Only*: obtained from the *Pseudotime* dataset by removing the assigned pseudotime values but maintaining the cell order. The cells are assumed to be regularly spaced along the trajectory.
- *Rand. Order*: 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 from the original dataset.

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

For variants of the ESC-to-endoderm differentiation dataset, only SINGE's performance decreases substantially for the *Rand. Order* dataset, as expected (Figures 5A and 5B). 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 1C), so we cannot draw strong conclusions from its performance trend across the dataset variants.

SINGE still outperforms SINCERITIES and SCODE in all variants of the retinoic acid dataset, but the performance trend does not follow the expected pattern. SINGE shows higher performance on the *Order Only* dataset, in which the pseudotime values are removed. Its performance for the *Rand. Order* variants is worse than for the *Order Only* dataset but comparable with the *Pseudotime* dataset. SINCERITIES has similar performance on the *Pseudotime* and *Order Only* datasets, with a slight improvement for *Order Only*. SCODE again performs poorly in all cases. Because the performance improves for SINGE and, to a lesser extent, SINCERITIES with *Order Only*, the Monocle 2 pseudotimes may not provide much information for GRN inference. Regulator-specific analysis using the *Order Only* dataset (Figure S6B) shows that the regulator-specific average precision and average early precision metrics of SINGE and SINCERITIES improve compared with the *Pseudotime* dataset (Figure 3B). Like Jump3, which does not use the pseudotime values, these two methods now have substantially better-than-random average early precision for several regulators. Similarly, the SINGE and SCODE average rankings of the outgoing interactions from the regulators using the *Order Only* dataset better match the regulators' prevalence in the ESCAPE database (Figure S6C). Regulators with more interactions in ESCAPE tend to have higher rankings in these predicted GRNs. These two phenomena combine to improve SINGE's overall precision-recall curve (Figure S6A).

We further investigate the relationship between the quality of the SINGE-inferred GRN and the trajectory inference method used to generate pseudotimes. We infer another trajectory

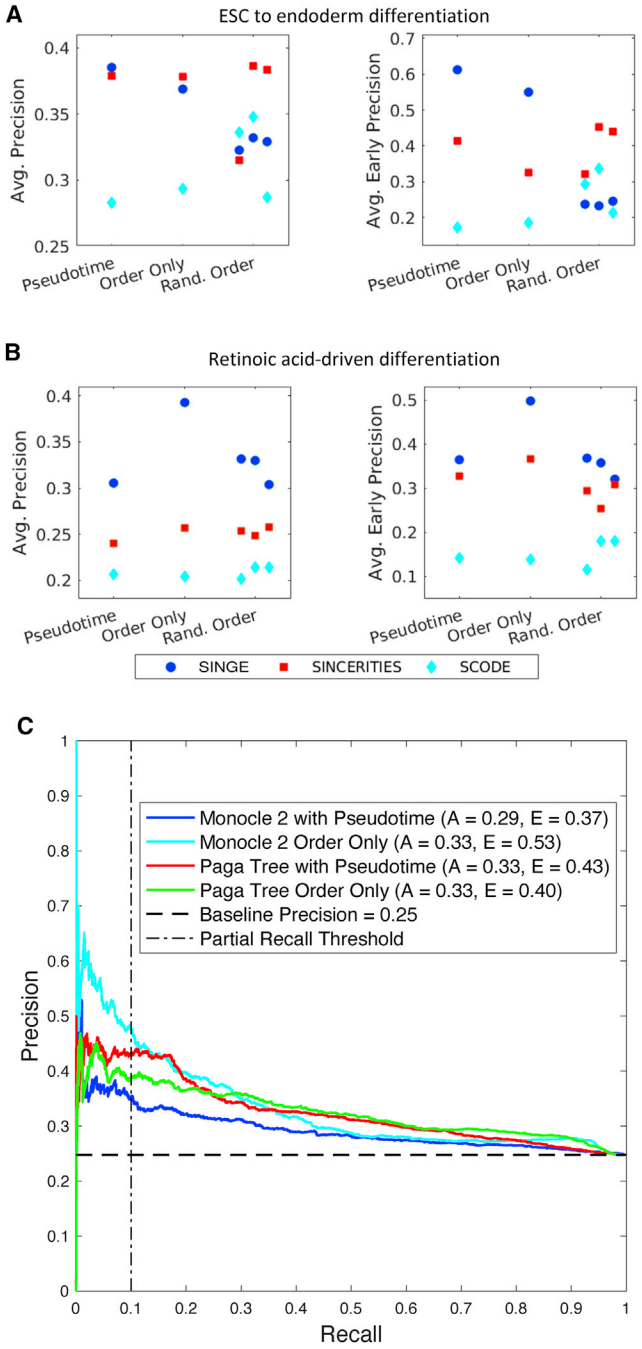


Figure 5. Effect of cell ordering, pseudotime, and trajectory inference method on network inference

(A) Performance of SINGE, SINCERITIES, and SCODE on variants of the ESC-to-endoderm differentiation dataset. Average precision metrics using Monocle pseudotimes are comparable with those using *Order Only*.

(B) Performance of the three methods on variants of the retinoic acid dataset. SINGE's average precision metrics improve when using the *Order Only* dataset instead of Monocle 2 pseudotime values.

(C) Precision-recall comparison of SINGE (version 0.3.0) using *Order Only* and *Pseudotime* datasets from Monocle 2 and PAGA Tree.

See also Figures S6 and S7.

from the retinoic acid dataset using PAGA Tree (Wolf et al., 2019; STAR Methods). We limit our study to the longest branch of the trajectory, which has 2,631 cells, of which 737 cells are common with the Monocle 2 branch (Figure S7A). Thus, the cell populations in the two analyses have some overlap and some unique cells.

The quality of the inferred GRN depends on the type of pseudotimes. Figure 5C evaluates SINGE with *Pseudotime* and *Order Only* versions of the Monocle 2 and PAGA Tree datasets. Importantly, the networks inferred with the original *Pseudotime* are not always better than the *Order Only* version. The average early precision of the SINGE GRN obtained with Monocle 2 *Order Only* is substantially better than all others. The benefits of the Monocle 2 *Order Only* dataset persist across different versions of the SINGE software (Figure S7B). For PAGA Tree, the performance differences between *Pseudotime* and *Order Only* are negligible.

Computational runtime

We designed SINGE to take advantage of high-throughput computing resources, such as the OSG (Pordes et al., 2007). We compare the GRN methods' runtimes on the retinoic acid dataset. SCODE and SINCERITIES require the least computational resources and can be run on a single workstation with an Intel i5-4590 processor and 8 gigabytes of random access memory. On this workstation, SCODE with 100 repetitions requires approximately 6 h to complete. SINCERITIES takes approximately 111 h.

In contrast, SINGE and Jump3 require more extensive computing resources. In the case of Jump3, inferring the GRN from 626 candidate regulators to one target gene takes between 11 min and 74 h to run, with an average runtime of 21.7 h. This is repeated for each target gene. In a typical application, SINGE uses 100 different hyperparameter settings on 10 subsampled expression datasets. Running SINGE for five λ hyperparameter values on one subsampled replicate takes 6.1 h (Table 2). The entire SINGE workflow for all hyperparameters and replicates requires 1,219.4 h. However, Jump3 and SINGE are highly parallelizable. We deployed them on our local high-throughput computing cluster using HTCondor (Erickson et al., 2018), which connects to the OSG (Pordes et al., 2007). In this high-throughput setting, we can run the entire SINGE algorithm in 36 h and the Jump3 algorithm in 72 h.

SINGE can also be configured to run on a single workstation with appropriate changes to the hyperparameters. For example, a complete SINGE run on the bone marrow dataset takes 19 h using a dedicated 16-core server with Intel Xeon Silver 4110 processors. For this run, we limit the number of regulator genes to only 149 TFs from AnimalTFDB 3.0 (Hu et al., 2018) and reduce the subsampled replicates per hyperparameter combination from 10 to two to reduce the runtime.

DISCUSSION

SINGE is a GRN reconstruction algorithm that adapts Granger Causality to detect dependencies in single-cell gene expression data annotated with pseudotimes. Although it was designed for single-cell data, the kernel-based smoothing could also be valuable for bulk time series gene expression data when the time

Table 2. Computational runtime of SINGE for datasets of various sizes

Dataset	Replicates	Genes	Cells	SINGE version	Average runtime for all λ values (h)	Parallel runtime (h)	Serial runtime (h)
Retinoic acid with Monocle 2	10	626	1,886	0.1.0	19.1	–	3,813.0
				0.3.0	6.1		1,219.4
Retinoic acid with PAGA Tree	10	626	2,540	0.3.0	7.8	22.4	1,566.3
dyngen dataset	10	140	1,000	0.3.0	0.1	0.4	16.8
Larger dyngen dataset	10	140	20,000	0.3.0	9.5	32.1	19,038.0
Bone marrow dataset (all genes as regulators)	10	3,025	3,105	0.3.0	23.8	62.7	4,759.2
Bone marrow dataset fast mode (149 TFs as regulators)	2	3,025	3,105	0.5.0	5.2	9.1	207.8

Average runtime for all λ values is calculated as the average compute time taken to run the GLG test for all λ values. SINGE version 0.1.0 runs independent GLG tests for each λ, but starting with version 0.3.0, it runs GLG for all λ values in one batch. Parallel runtime is calculated using the longest individual runtime for a GLG test. This is the approximate SINGE runtime in the hypothetical scenario where all GLG tests are run in parallel simultaneously. Serial runtime is calculated as the aggregate compute time of all GLG tests. This is the approximate SINGE runtime in the hypothetical scenario where it is run serially on a single machine corresponding to the average node in the high-throughput computing pool. SINGE versions 0.3.0 and 0.5.0 have performance improvements not present in version 0.1.0.

points are irregularly spaced or individual expression samples are noisy. SINGE can prioritize regulators for future DNA binding or functional studies. For example, many of the top-ranked SINGE regulators in the retinoic acid study (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 case study, in which none of the GRN methods' settings were tuned to optimize performance on this dataset, SINGE has better precision-recall performance than four existing methods. However, we caution that single metrics like average precision can be misleading. Closer inspection reveals that SINGE's better-than-random precision-recall performance in Figure 3A is driven by its ability to identify important regulators and assign them a higher rank (Figure 3C). In contrast, Jump3's better-than-random precision-recall performance in Figure 3A is driven by its better-than-random performance for many individual regulators (Figure 3B). Because the precision-recall curve for the entire GRN can mask near-random performance for many individual regulators, we recommend regulator-specific visualizations (Figures 3B and 3C) to provide more context.

We designed SINGE 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 SINGE 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 (Marbach et al., 2012a). Our use of modified Borda aggregation for ensembling emphasizes the top-ranked, most confident predictions.

Benchmarking and evaluation

Inferring GRNs from single-cell gene expression data remains a difficult task. Evaluations of network inference algorithms on simulated (Chen and Mar, 2018) and real (Stone et al., 2021) single-cell datasets reported that predictions were generally only slightly better than random edge ordering. New single-cell gene expression simulators designed specifically for simulating GRNs (Pratapa et al., 2020; Dibaeinia and Sinha, 2020) can help inform which GRN inference methods are best for different types of biological trajectories. Both of these studies evaluated SINGE performance with their GRN simulators but with different ensembling strategies. BEELINE (Pratapa et al., 2020) optimized SINGE's hyperparameters and constructed much smaller ensembles than we do in this study. SERGIO's ensembling (Dibaeinia and Sinha, 2020) was much closer to ours except for differences in treatment of the λ hyperparameter. In the SERGIO evaluation, SINGE was the best GRN method when inferring networks from simulated datasets with added technical noise.

An important aspect when evaluating network inference on experimental data is the relevance of the gold standard. BEELINE showed that the choice of a cell-type-specific versus non-specific gold standard has a substantial effect on GRN inference performance metrics (Pratapa et al., 2020). The gold standard in the SCODE evaluation (Matsumoto et al., 2017) 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 (2018) benchmarking of stem cell datasets, the gold standard consisted of all interactions from the STRING database (Szklarczyk et al., 2014). These included interaction types that are not directly informative about transcriptional regulation and were not limited to the specific cell types of interest. In contrast, our evaluation considers only interactions from mouse ESCs.

Future work and extensions

SINGE can currently model all biological processes with acyclic trajectories. However, SINGE provides just one regulatory network for the entire trajectory. In the future, it could identify branch-specific networks and interactions for different parts of the trajectory. An alternative approach would be to adapt SINGE

to treat each branch as a task in a multi-task GRN inference problem (Castro et al., 2019). In addition, the kernel could be modified so that certain pseudotime intervals can be considered to be more informative; for example, the interval around a major bifurcation point in the trajectory.

Because each trajectory inference method orders the cells based on different assumptions and algorithms, no two methods will produce the same cell ordering and pseudotimes. Some trajectories are better suited for a particular biological process, but we cannot objectively verify the correctness of the trajectory and cell ordering. Closer integration with the dynverse framework could help in two ways. First, GRN reconstruction accuracy could complement other benchmarking metrics for pseudotimes (Saelens et al., 2019). We could integrate dynverse's pseudotime benchmarking (Saelens et al., 2019) with GRN benchmarking frameworks (Pratapa et al., 2020) to systematically evaluate combinations of trajectory inference and GRN methods on different datasets. This would also enable empirically assessing the types of GRN motifs that cannot be unambiguously recovered from single-cell expression data (Weinreb et al., 2018). In addition, we could expand the SINGE ensemble to run GLG on pseudotimes from multiple trajectory inference methods in dynverse. This could potentially mitigate the effect of any single misspecified pseudotime estimation method on network inference.

Other elements of the GLG regression framework can be adapted as well. These include placing monotonicity constraints on the temporal lag coefficients (Nguyen and Braun, 2018), adapting the kernel so that more recent samples are assigned higher weights, or implementing kernel-based generalizations of group lasso (Yuan and Lin, 2006; Lozano et al., 2009) to regularize all coefficients from individual regulators as a group instead of independent variables. The kernel-based approach at the core of SINGE provides great flexibility to adapt it to emphasize different aspects of dynamic biological processes.

Limitations of the study

Granger causality has a precise statistical meaning built on the idea that causes must come before their effects (Granger, 1980). It is useful in practice but does not necessarily satisfy more general definitions of causality (Granger, 1980). The main assumption when using GLG, that expression dynamics are obtained from linear and stationary vector autoregressive (VAR) models, is too simplistic for modeling complex biological processes. Violating the assumptions of linearity and stationarity can have a significant effect on the performance of individual GLG tests. Furthermore, Granger causality tests result in false positives in scenarios with hidden variables (Bahadori and Liu, 2013). These discrepancies between theory and practice are commonly accepted in biological applications of Granger causality (Valdés-Sosa et al., 2005; Shojaie and Michailidis, 2010). However, our ESCAPE and dyngen evaluations suggest that SINGE may detect more indirect regulatory relationships than direct TF binding. SINGE's precision-recall performance improves relative to the baseline precision when adding indirect gene interactions to the dyngen gold standard.

Some of the Granger causality-related drawbacks could potentially be addressed by integrating SINGE with complementary data types. The relationship between TF concentration and transcriptional activity represents only one type of transcriptional dynamics, neglecting epigenomic modifications, TF post-translational modifications, TF localization, and transcriptional co-factors (Swift and Coruzzi, 2017). 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 (Siahpirani and Roy, 2016; Greenfield et al., 2013; reviewed in Chasman et al., 2016). Other single-cell GRN inference algorithms have incorporated priors (Gibbs et al., 2021). SOMatic (Jansen et al., 2019) and Symphony (Burdziak et al., 2019) use single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq), and scdiff (Ding et al., 2018) integrates TF-gene interactions. To model prior information in SINGE, we could assign different penalty factors λ_j for the jth regulator of target gene i based on the prior probability of the edge p_{ij}. An alternative would be to use SINGE output in conjunction with the supplementary sources of information and aggregate all information after the fact (Ciofani et al., 2012; Marbach et al., 2012b). However, the current version of SINGE intentionally uses only gene expression data because integrative approaches can benefit from understanding the best ways to infer GRNs from expression data alone. This also makes SINGE widely applicable under conditions and in species where suitable priors are not available.

Another assumption of SINGE, SINCERITIES, and SCODE is that the pseudotime values are biologically meaningful. Assigning uninformative pseudotime values to ordered cells can be detrimental to network inference performance (Figure 5B). Qiu et al. (2020) proposed that RNA velocity (La Manno et al., 2018) may help overcome some of the limitations faced when using pseudotime for GRN reconstruction.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - ESC to endoderm differentiation dataset
  - Retinoic acid dataset
  - Mouse bone marrow mesenchyme to erythrocyte differentiation dataset
  - Dyngen synthetic dataset
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Generalized Lasso Granger test
  - Single-cell inference of networks using Granger ensembles
  - Case study hyperparameters
  - Existing GRN methods

- Evaluation
- ESCAPE database
- Average early precision
- KinderMiner and Gene Ontology enrichment

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.110333.

ACKNOWLEDGMENTS

We thank Christina Kendziorski for feedback on the SINGE algorithm, Alireza Fotuhi Siahpirani for assistance with the ESCAPE database, Rafael Feliciano for discussion of the GRN evaluation, Scott Swanson for testing the SINGE software, Robrecht Cannoodt and Wouter Saelens for help with the dyngen simulation, Eric Bolden for support compiling SINGE for macOS, and members of the Gitter, Stewart, and Kendziorski research groups for their helpful comments. The graphical abstract and Figure 1A were created with BioRender.com. This research was supported by NSF CAREER award DBI 1553206; NIH grants UH3TR000506, U01HL099773, and P50DE026787; the Morgridge Institute for Research; the UW-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation; a grant from Marv Conney; and the compute resources and assistance of the UW-Madison Center for High Throughput Computing and the OSG, which is supported by NSF award OAC 2030508.

AUTHOR CONTRIBUTIONS

Conceptualization, A.D. and A.G.; data curation, A.D.; formal analysis, A.D., L.-F.C., and R.S.; investigation, A.D., L.-F.C., and R.S.; methodology, A.D. and A.G.; software, A.D. and A.G.; validation, A.D.; visualization, A.D.; funding acquisition, A.G.; supervision, A.G.; writing – original draft, all authors; writing – review & editing, all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 5, 2019
Revised: February 19, 2021
Accepted: January 12, 2022
Published: February 8, 2022

REFERENCES

Ahsen, M.E., Vogel, R.M., and Stolovitzky, G.A. (2019). Unsupervised evaluation and weighted aggregation of ranked classification predictions. J. Mach. Learn. Res. *20*, 1–40.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086. https://doi.org/10.1038/nmeth.4463.

Andrews, T., and Hemberg, M. (2018). False signals induced by single-cell imputation [version 1; referees: 4 approved with reservations. F1000Res. *7*. https://doi.org/10.12688/f1000research.16613.1.

Arnold, A., Liu, Y., and Abe, N. (2007). Temporal causal modeling with graphical Granger methods. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (ACM), pp. 66–75.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

Aubin-Frankowski, P.-C., and Vert, J.-P. (2020). Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. Bioinformatics *36*, 4774–4780. https://doi.org/10.1093/bioinformatics/btaa576.

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. *17*, 63.

Bahadori, M.T., and Liu, Y. (2012). Granger causality analysis in irregular time series. In Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 660–671. https://doi.org/10.1137/1.9781611972825.57.

Bahadori, M.T., and Liu, Y. (2013). An examination of practical Granger causality inference. In Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 467–475. https://doi.org/10.1137/1.9781611972832.52.

Balakumaran, A., Mishra, P.J., Pawelczyk, E., Yoshizawa, S., Sworder, B.J., Cherman, N., Kuznetsov, S.A., Bianco, P., Giri, N., Savage, S.A., et al. (2015). Bone marrow skeletal stem/progenitor cell defects in dyskeratosis congenita and telomere biology disorders. Blood *125*, 793–802. https://doi.org/10.1182/blood-2014-06-566810.

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. Nat. Rev. Genet. *13*, 552.

Barrow, J.R., and Capecchi, M.R. (1996). Targeted disruption of the Hoxb-2 locus in mice interferes with expression of Hoxb-1 and Hoxb-4. Development *122*, 3817–3828.

Bendall, S.C., Davis, K.L., Amir, E.-a. D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Peer, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell *157*, 714–725.

Blencowe, M., Arneson, D., Ding, J., Chen, Y.-W., Saleem, Z., and Yang, X. (2019). Network modeling of single-cell omics data: challenges, opportunities, and progresses. Emerg. Top. Life Sci. *3*, 379–398. https://doi.org/10.1042/ETLS20180176.

Bonnaffoux, A., Herbach, U., Richard, A., Guillemin, A., Gonin-Giraud, S., Gros, P.-A., and Gandrillon, O. (2019). WASABI: a dynamic iterative framework for gene regulatory network inference. BMC Bioinf. *20*, 220. https://doi.org/10.1186/s12859-019-2798-1.

Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140. https://doi.org/10.1007/BF00058655.

Bult, C.J., Blake, J.A., Smith, C.L., Kadin, J.A., and Richardson, J.E.; Mouse Genome Database Group (2019). Mouse genome database (MGD) 2019. Nucleic Acids Res. *47*, D801–D806. https://doi.org/10.1093/nar/gky1056.

Burdziak, C., Azizi, E., Prabhakaran, S., and Peer, D. (2019). A nonparametric multi-view model for estimating cell type-specific gene regulatory networks. https://arxiv.org/abs/1902.08138.

Cai, K.Q., Capo-Chichi, C.D., Rula, M.E., Yang, D.-H., and Xu, X.-X. (2008). Dynamic GATA6 expression in primitive endoderm formation and maturation in early mouse embryogenesis. Dev. Dyn. *237*, 2820–2829.

Campbell, K., Ponting, C.P., and Webber, C. (2015). Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell rna-seq profiles. bioRxiv. https://doi.org/10.1101/027219.

Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K.D., and Saeys, Y. (2016). Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. bioRxiv. https://doi.org/10.1101/079509.

Cannoodt, R., Saelens, W., Deconinck, L., and Saeys, Y. (2021). Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nat. Commun. *12*. https://doi.org/10.1038/s41467-021-24152-2.

Carpenter, A.C., Rao, S., Wells, J.M., Campbell, K., and Lang, R.A. (2010). Generation of mice with a conditional null allele for Wntless. Genesis *48*, 554–558. https://doi.org/10.1002/dvg.20651.

DREAM Challenges (2009). DREAM4 in Silico Network Challenge. https://dreamchallenges.org/dream-4-in-silico-network-challenge/.

Castro, D.M., de Veaux, N.R., Miraldi, E.R., and Bonneau, R. (2019). Multi-study inference of regulatory networks for more accurate models of gene regulation. PLoS Comput. Biol. *15*, e1006591. https://doi.org/10.1371/journal.pcbi.1006591.

Chan, T.E., Pallaseni, A., Babtie, A.C., McEwen, K., and Stumpf, M.P. (2018). Empirical Bayes meets information theoretical network reconstruction from single cell data. bioRxiv. https://doi.org/10.1101/264853.

Chan, T.E., Stumpf, M.P., and Babtie, A.C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. *5*, 251–267. https://doi.org/10.1016/j.cels.2017.08.014.

Chasman, D., Siahpirani, A.F., and Roy, S. (2016). Network-based approaches for analysis of complex biological systems. Curr. Opin. Biotechnol. *39*, 157–166. https://doi.org/10.1016/j.copbio.2016.04.007.

Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinf. *19*, 232. https://doi.org/10.1186/s12859-018-2217-z.

Ciofani, M., Madar, A., Galan, C., Sellars, M., Mace, K., Pauli, F., Agarwal, A., Huang, W., Parkurst, C.N., Muratet, M., et al. (2012). A validated regulatory network for Th17 cell specification. Cell *151*, 289–303. https://doi.org/10.1016/j.cell.2012.09.016.

Cordero, P., and Stuart, J.M. (2017). Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories. In Pacific Symposium on Biocomputing 2017. World Scientific, pp. 576–587. https://doi.org/10.1142/9789813207813_0053.

DeChiara, T.M., Efstratiadis, A., and Robertsen, E.J. (1990). A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. Nature *345*, 78–80. https://doi.org/10.1038/345078a0.

Deshpande, A., and Gitter, A. (2021). SINGE Supplemental Information. https://doi.org/10.5281/zenodo.3627325. https://github.com/gitter-lab/SINGE-supplemental.

Dibaeinia, P., and Sinha, S. (2020). SERGIO: a single-cell expression simulator guided by gene regulatory network. Cell Syst. *11*, 252–271.e11. https://doi.org/10.1016/j.cels.2020.08.003.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

Ding, J., Aronow, B.J., Kaminski, N., Kitzmiller, J., Whitsett, J.A., and Bar-Joseph, Z. (2018). Reconstructing differentiation networks and their regulation from time series single-cell expression data. Genome Res. *28*, 383–395. https://doi.org/10.1101/gr.225979.117.

Egea, J., Erlacher, C., Montanez, E., Burtscher, I., Yamagishi, S., Hess, M., Hampel, F., Sanchez, R., Rodriguez-Manzaneque, M.T., Bösl, M.R., et al. (2008). Genetic ablation of FLRT3 reveals a novel morphogenetic function for the anterior visceral endoderm in suppressing mesoderm differentiation. Genes Dev. *22*, 3349–3362. https://doi.org/10.1101/gad.486708.

Erickson, R.A., Fienen, M.N., McCalla, S.G., Weiser, E.L., Bower, M.L., Knudson, J.M., and Thain, G. (2018). Wrangling distributed computing for high-throughput environmental science: an introduction to HTCondor. PLoS Comput. Biol. *14*, e1006468. https://doi.org/10.1371/journal.pcbi.1006468.

van Erp, M., and Schomaker, L. (2000). 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. (International Unipen Foundation), pp. 443–452.

Feldman, B., Poueymirou, W., Papaioannou, V.E., DeChiara, T.M., and Goldfarb, M. (1995). Requirement of FGF-4 for postimplantation mouse development. Science *267*, 246–249. https://doi.org/10.1126/science.7809630.

Fiers, M.W., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. Brief. Funct. Genomics *18*, 1925–1938. https://doi.org/10.1093/bfgp/elx046.

Finkle, J.D., Wu, J.J., and Bagheri, N. (2018). Windowed Granger causal inference strategy improves discovery of gene regulatory networks. Proc. Nat. Acad. Sci. U S A *115*, 2252–2257. https://doi.org/10.1073/pnas.1710936115.

Finley, K.R., Tennessen, J., and Shawlot, W. (2003). The mouse secreted frizzled-related protein 5 gene is expressed in the anterior visceral endoderm and foregut endoderm during early post-implantation development. Gene Expr. Patterns *3*, 681–684. https://doi.org/10.1016/S1567-133X(03)00091-7.

Fraenkel, J., and Grofman, B. (2014). The Borda count and its real-world alternatives: comparing scoring rules in Nauru and Slovenia. Aust. J. Polit. Sci. *49*, 186–205. https://doi.org/10.1080/10361146.2014.900530.

Fujita, A., Severino, P., Sato, J.R., and Miyano, S. (2010). Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models. In Brazilian Symposium on Bioinformatics (Springer), pp. 13–24.

Gibbs, C.S., Jackson, C.A., Saldi, G.-A., Tjärnberg, A., Shah, A., Watters, A., Veaux, N.D., Tchourine, K., Yi, R., Hamamsy, T., et al. (2021). High performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. bioRxiv. https://doi.org/10.1101/2021.05.03.442499.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. *81*, 2340–2361. https://doi.org/10.1021/j100540a008.

Gitter, A. (2018). Single-cell RNA-Seq Pseudotime Estimation Algorithms. https://doi.org/10.5281/zenodo.1297422. https://github.com/agitter/single-cell-pseudotime.

Gitter, A., Siegfried, Z., Klutstein, M., Fornes, O., Oliva, B., Simon, I., and Bar-Joseph, Z. (2009). Backup in gene regulatory networks explains differences between binding and knockout results. Mol. Syst. Biol. *5*. https://doi.org/10.1038/msb.2009.33.

Gorry, P., Lufkin, T., Dierich, A., Rochette-Egly, C., Décimo, D., Dollé, P., Mark, M., Durand, B., and Chambon, P. (1994). The cellular retinoic acid binding protein I is dispensable. Proc. Nat. Acad. Sci. U S A *91*, 9032–9036. https://doi.org/10.1073/PNAS.91.19.9032.

Granger, C.W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica J. Econ. Soc. *37*, 424–438.

Granger, C.W.J. (1980). Testing for causality: a personal viewpoint. J. Econ. Dynam. Control *2*, 329–352. https://doi.org/10.1016/0165-1889(80)90069-X.

Greenfield, A., Hafemeister, C., and Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics *29*, 1060–1067. https://doi.org/10.1093/bioinformatics/btt099.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by microwell-seq. Cell *172*, 1091–1107.e17. https://doi.org/10.1016/j.cell.2018.02.001.

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). TIGRESS: trustful inference of gene regulation using stability selection. BMC Syst. Biol. *6*, 145. https://doi.org/10.1186/1752-0509-6-145.

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. (2018). Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat. Commun. *9*, 619. https://doi.org/10.1038/s41467-018-02866-0.

Heerah, S., Molinari, R., Guerrier, S., and Marshall-Colon, A. (2021). Granger-causal testing for irregularly sampled time series with application to nitrogen signalling in Arabidopsis. Bioinformatics *37*, 2450–2460. https://doi.org/10.1093/bioinformatics/btab126.

Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., and Guo, A.-Y. (2018). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. *47*, D33–D38. https://doi.org/10.1093/nar/gky822.

Huynh-Thu, V.A., and Sanguinetti, G. (2015). Combining tree-based and dynamical systems for the inference of gene regulatory networks. Bioinformatics *31*, 1614–1622. https://doi.org/10.1093/bioinformatics/btu863.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS ONE *5*, e12776. https://doi.org/10.1371/journal.pone.0012776.

Intosalmi, J., Mannerstrom, H., Hiltunen, S., and Lahdesmaki, H. (2018). SCHiRM: single cell hierarchical regression model to detect dependencies in read count data. bioRxiv. https://doi.org/10.1101/335695.

Jansen, C., Ramirez, R.N., El-Ali, N.C., Gomez-Cabrero, D., Tegner, J., Merkenschlager, M., Conesa, A., and Mortazavi, A. (2019). Building gene

regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. PLoS Comput. Biol. *15*, e1006555. https://doi.org/10.1371/journal.pcbi.1006555.

Kim, J., Jakobsen, S., Natarajan, K.N., and Won, K.-J. (2021). TENET: gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data. Nucleic Acids Res. *49*, e1. https://doi.org/10.1093/nar/gkaa1014.

Krawchuk, D., Honma-Yamanaka, N., Anani, S., and Yamanaka, Y. (2013). FGF4 is a limiting factor controlling the proportions of primitive endoderm and epiblast in the ICM of the mouse blastocyst. Dev. Biol. *384*, 65–71. https://doi.org/10.1016/j.ydbio.2013.09.023.

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development *134*, 2895–2902. https://doi.org/10.1242/dev.02880.

Kuusisto, F., Steill, J., Kuang, Z., Thomson, J., Page, D., and Stewart, R. (2017). A simple text mining approach for ranking pairwise associations in biomedical applications. In AMIA Joint Summits on Translational Science Proceedings 2017, pp. 166–174.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity in single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.

Leaf, I., Tennessen, J., Mukhopadhyay, M., Westphal, H., and Shawlot, W. (2006). Sfrp5 is not essential for axis formation in the mouse. Genesis *44*, 573–578. https://doi.org/10.1002/dvg.20248.

Leng, N., Chu, L.-F., Barry, C., Li, Y., Choi, J., Li, X., Jiang, P., Stewart, R.M., Thomson, J.A., and Kendziorski, C. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nat. Methods *12*, 947–950. https://doi.org/10.1038/nmeth.3549.

Li, J., He, F., Zhang, P., Chen, S., Shi, H., Sun, Y., Ying, G., Yang, H., Nimer, S.D., Wang, Q.-F., et al. (2016). ASXL2 is required for normal hematopoiesis and loss of asxl2 leads to myeloid malignancies in mice. Blood *128*, 1509. https://doi.org/10.1182/blood.V128.22.1509.1509.

Linderman, G.C., Zhao, J., and Kluger, Y. (2018). Zero-preserving imputation of scRNA-seq data using low-rank approximation. bioRxiv. https://doi.org/10.1101/397588.

Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O.M., Zhang, M.Q., Jiang, R., and Chen, T. (2017). Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nat. Commun. *8*, 22. https://doi.org/10.1038/s41467-017-00039-z.

Lozano, A.C., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery. Bioinformatics *25*, i110–i118. https://doi.org/10.1093/bioinformatics/btp199.

Lu, J., Dumitrascu, B., McDowell, I.C., Jo, B., Barrera, A., Hong, L.K., Leichter, S.M., Reddy, T.E., and Engelhardt, B.E. (2021). Causal network inference from gene transcriptional time-series response to glucocorticoids. PLoS Comput. Biol. *17*, e1008223. https://doi.org/10.1371/journal.pcbi.1008223.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Aderhold, A., Bonneau, R., Chen, Y., et al. (2012a). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804. https://doi.org/10.1038/nmeth.2016.

Marbach, D., Roy, S., Ay, F., Meyer, P.E., Candeias, R., Kahveci, T., Bristow, C.A., and Kellis, M. (2012b). Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. Genome Res. *22*, 1334–1349. https://doi.org/10.1101/gr.127191.111.

Marsh, J.C.W., Gutierrez-Rodrigues, F., Cooper, J., Jiang, J., Gandhi, S., Kajigaya, S., Feng, X., Ibanez, M.d.P.F., Donaires, F.S., Lopes da Silva, J.P., et al. (2018). Heterozygous RTEL1 variants in bone marrow failure and myeloid neoplasms. Blood Adv. *2*, 36–48. https://doi.org/10.1182/bloodadvances.2017008110.

Matsumoto, H., and Kiryu, H. (2016). SCOUP: a probabilistic model based on the Ornstein–Uhlenbeck process to analyze single-cell expression data during differentiation. BMC Bioinf. *17*, 232. https://doi.org/10.1186/s12859-016-1109-3.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S., Ko, S.B., Gouda, N., Hayashi, T., and Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. Bioinformatics *33*, 2314–2321. https://doi.org/10.1093/bioinformatics/btx194.

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. J. Roy. Stat. Soc. B Stat. Methodol. *72*, 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Meno, C., Gritsman, K., Ohishi, S., Ohfuji, Y., Heckscher, E., Mochida, K., Shimono, A., Kondoh, H., Talbot, W.S., Robertson, E.J., et al. (1999). Mouse lefty2 and zebrafish antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. Mol. Cell *4*, 287–298. https://doi.org/10.1016/S1097-2765(00)80331-7.

Micol, J.-B., Pastore, A., Inoue, D., Duployez, N., Kim, E., Lee, S.C.-W., Durham, B.H., Chung, Y.R., Cho, H., Zhang, X.J., et al. (2017). ASXL2 is essential for haematopoiesis and acts as a haploinsufficient tumour suppressor in leukemia. Nat. Commun. *8*, 15429. https://doi.org/10.1038/ncomms15429.

Morris, S.M., Tallquist, M.D., Rock, C.O., and Cooper, J.A. (2002). Dual roles for the Dab2 adaptor protein in embryonic development and kidney transport. EMBO J. *21*, 1555–1564. https://doi.org/10.1093/emboj/21.7.1555.

Morrisey, E.E., Tang, Z., Sigrist, K., Lu, M.M., Jiang, F., Ip, H.S., and Parmacek, M.S. (1998). GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. Genes Dev. *12*, 3579–3590. https://doi.org/10.1101/gad.12.22.3579.

Mukhopadhyay, N.D., and Chatterjee, S. (2006). Causality and pathway search in microarray time series experiment. Bioinformatics *23*, 442–449. https://doi.org/10.1093/bioinformatics/btl598.

Nguyen, P., and Braun, R. (2018). Time-lagged ordered lasso for network inference. BMC Bioinf. *19*, 545. https://doi.org/10.1186/s12859-018-2558-7.

Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Brief. Bioinf. *22*, bbaa190. https://doi.org/10.1093/bib/bbaa190.

Ocone, A., Haghverdi, L., Mueller, N.S., and Theis, F.J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics *31*, i89–i96. https://doi.org/10.1093/bioinformatics/btv257.

Olson, G.E., Whitin, J.C., Hill, K.E., Winfrey, V.P., Motley, A.K., Austin, L.M., Deal, J., Cohen, H.J., and Burk, R.F. (2010). Extracellular glutathione peroxidase (Gpx3) binds specifically to basement membranes of mouse renal cortex tubule cells. Am. J. Physiol. Ren. Physiol. *298*, F1244–F1253. https://doi.org/10.1152/ajprenal.00662.2009.

Pankratz, M.T., Li, X.-J., LaVaute, T.M., Lyons, E.A., Chen, X., and Zhang, S.-C. (2007). Directed neural differentiation of human embryonic stem cells via an obligated primitive anterior stage. Stem Cell. *25*, 1511–1520. https://doi.org/10.1634/stemcells.2006-0707.

Papili Gao, N., Ud-Dean, S.M., Gandrillon, O., and Gunawan, R. (2017). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics *34*, 258–266. https://doi.org/10.1093/bioinformatics/btx575.

Parant, J., Chavez-Reyes, A., Little, N.A., Yan, W., Reinke, V., Jochemsen, A.G., and Lozano, G. (2001). Rescue of embryonic lethality in Mdm4-null mice by loss of Trp53 suggests a nonoverlapping pathway with MDM2 to regulate p53. Nat. Genet. *29*, 92–95. https://doi.org/10.1038/ng714.

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., Avery, P., Blackburn, K., Wenaus, T., Würthwein, F., et al. (2007). The open science grid. J. Phys. *78*, 012057. https://doi.org/10.1088/1742-6596/78/1/012057.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-

cell transcriptomic data. Nat. Methods *17*, 147–154. https://doi.org/10.1038/s41592-019-0690-6.

Qian, J., Hastie, T., Friedman, J., Tibshirani, R., and Simon, N. (2013). GLMNET for MATLAB.http://www.stanford.edu/~hastie/glmnet_matlab/.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979. https://doi.org/10.1038/nmeth.4402.

Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2020). Inferring causal gene regulatory networks from coupled single-cell expression dynamics using Scribe. Cell Syst. *10*, 265–274.e11. https://doi.org/10.1016/j.cels.2020.02.003.

Radice, G.L., Rayburn, H., Matsunami, H., Knudsen, K.A., Takeichi, M., and Hynes, R.O. (1997). Developmental defects in mouse embryos lacking N-Cadherin. Dev. Biol. *181*, 64–78. https://doi.org/10.1006/DBIO.1996.8443.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. *44*, W83–W89. https://doi.org/10.1093/nar/gkw199.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.

Sakai, T., Li, S., Docheva, D., Grashoff, C., Sakai, K., Kostka, G., Braun, A., Pfeifer, A., Yurchenco, P.D., and Fässler, R. (2003). Integrin-linked kinase (ILK) is required for polarizing the epiblast, cell adhesion, and controlling actin accumulation. Genes Dev. *17*, 926–940. https://doi.org/10.1101/gad.255603.

Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I.M., Carrion, M., and Huang, Y. (2017). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. Bioinformatics *34*, 964–970. https://doi.org/10.1093/bioinformatics/btx605.

Schrynemackers, M., Kueffner, R., and Geurts, P. (2013). On protocols and measures for the validation of supervised methods for the inference of biological networks. Front. Genet. *4*, 262. https://doi.org/10.3389/fgene.2013.00262.

Semrau, S., Goldmann, J.E., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., and Van Oudenaarden, A. (2017). Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. Nat. Commun. *8*, 1096. https://doi.org/10.1038/s41467-017-01076-4.

Setty, M., Tadmor, M.D., Reich-Zeliger, S., Angel, O., Salame, T.M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat. Biotechnol. *34*, 637–645. https://doi.org/10.1038/nbt.3569.

Shimosato, D., Shiki, M., and Niwa, H. (2007). Extra-embryonic endoderm cells derived from ES cells induced by GATA factors acquire the character of XEN cells. BMC Dev. Biol. *7*, 80. https://doi.org/10.1186/1471-213x-7-80.

Shin, J., Berg, D., Zhu, Y., Shin, J., Song, J., Bonaguidi, M., Enikolopov, G., Nauen, D., Christian, K., Ming, G.-l., and Song, H. (2015). Single-cell RNA-Seq with Waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell *17*, 360–372. https://doi.org/10.1016/J.STEM.2015.07.013.

Shojaie, A., and Michailidis, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics *26*, i517–i523. https://doi.org/10.1093/bioinformatics/btq377.

Siahpirani, A.F., and Roy, S. (2016). A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Res. *45*, e21. https://doi.org/10.1093/nar/gkw1160.

Sicinski, P., Donaher, J.L., Geng, Y., Parker, S.B., Gardner, H., Park, M.Y., Robker, R.L., Richards, J.S., McGinnis, L.K., Biggers, J.D., et al. (1996). Cyclin D2 is an FSH-responsive gene involved in gonadal cell proliferation and oncogenesis. Nature *384*, 470–474. https://doi.org/10.1038/384470a0.

Skarnes, W.C., Rosen, B., West, A.P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A.O., Thomas, M., Harrow, J., Cox, T., et al. (2011). A conditional knockout resource for the genome-wide study of mouse gene function. Nature *474*, 337–342. https://doi.org/10.1038/nature10163.

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. Nat. Rev. Microbiol. *8*, 717. https://doi.org/10.1038/nrmicro2419.

Sollars, V.E., McEntee, B.J., Engiles, J.B., Rothstein, J.L., and Buchberg, A.M. (2002). A novel transgenic line of mice exhibiting autosomal recessive male-specific lethality and non-alcoholic fatty liver disease. Hum. Mol. Genet. *11*, 2777–2786. https://doi.org/10.1093/hmg/11.22.2777.

Specht, A.T., and Li, J. (2016). Leap: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics *33*, 764–766. https://doi.org/10.1093/bioinformatics/btw729.

Stavridis, M.P., Lunn, J.S., Collins, B.J., and Storey, K.G. (2007). A discrete period of FGF-induced Erk1/2 signalling is required for vertebrate neural specification. Development *134*, 2889–2894. https://doi.org/10.1242/dev.02858.

Stone, M., Li, J., McCalla, S.G., Siahpirani, A.F., Periyasamy, V., Shin, J., and Roy, S. (2021). Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. bioRxiv. https://doi.org/10.1101/2021.06.01.446671.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. *19*, 477. https://doi.org/10.1186/s12864-018-4772-0.

Swift, J., and Coruzzi, G.M. (2017). A matter of time — how transient transcription factor interactions create dynamic gene regulatory networks. Biochim. Biophys. Acta *1860*, 75–83. https://doi.org/10.1016/j.bbagrm.2016.08.007.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452. https://doi.org/10.1093/nar/gku1003.

Takaoka, K., Nishimura, H., and Hamada, H. (2017). Both nodal signalling and stochasticity select for prospective distal visceral endoderm in mouse embryos. Nat. Commun. *8*, 1492. https://doi.org/10.1038/s41467-017-01625-x.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331. https://doi.org/10.1038/nature21350.

Thattai, M., and Oudenaarden, A.v. (2001). Intrinsic noise in gene regulatory networks. Proc. Nat. Acad. Sci. U S A *98*, 8614–8619. https://doi.org/10.1073/pnas.151588598.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498. https://doi.org/10.1101/gr.190595.115.

Tsakanikas, P., Manatakis, D.V., and Manolakos, E.S. (2018). Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions. bioRxiv. https://doi.org/10.1101/264671.

Valdés-Sosa, P.A., Sánchez-Bornot, J.M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. Phil. Trans. Roy. Soc. Lond. B *360*, 969–981. https://doi.org/10.1098/rstb.2005.1654.

Wang, Y., Thekdi, N., Smallwood, P.M., Macke, J.P., and Nathans, J. (2002). Frizzled-3 is required for the development of major fiber tracts in the rostral CNS. J. Neurosci. *22*, 8563–8573. https://doi.org/10.1523/JNEUROSCI.22-19-08563.2002.

Wei, J., Hu, X., Zou, X., and Tian, T. (2017). Reverse-engineering of gene networks for regulating early blood development from single-cell measurements. BMC Med. Genom. *10*, 72. https://doi.org/10.1186/s12920-017-0312-z.

Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc. Nat. Acad. Sci. U S A *115*, E2467–E2476. https://doi.org/10.1073/pnas.1714723115.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction

reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59. https://doi.org/10.1186/s13059-019-1663-x.

Xiao, Y., Ma, H., Wan, P., Qin, D., Wang, X., Zhang, X., Xiang, Y., Liu, W., Chen, J., Yi, Z., and Li, L. (2017). Trp-Asp (WD) repeat domain 1 is essential for mouse peri-implantation development and regulates Cofilin phosphorylation. J. Biol. Chem. *292*, 1438–1448. https://doi.org/10.1074/jbc.m116.759886.

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., and Ma'ayan, A. (2013). ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database *2013*, bat045. https://doi.org/10.1093/database/bat045.

Yamanaka, Y., Lanner, F., and Rossant, J. (2010). FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. Development *137*, 715–724. https://doi.org/10.1242/dev.043471.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B Stat. Methodol. *68*, 49–67.

Zhang, L., and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. IEEE ACM Trans. Comput. Biol. Bioinf. *17*, 376–389. https://doi.org/10.1109/TCBB.2018.2848633.

Zhang, X., Friedman, A., Heaney, S., Purcell, P., and Maas, R.L. (2002). Meis homeoproteins directly regulate Pax6 during vertebrate lens morphogenesis. Genes Dev. *16*, 2097–2107. https://doi.org/10.1101/gad.1007602.

Zhang, J., Zhou, T., and Nie, Q. (2018). Topographer reveals dynamic mechanisms of cell fate decisions from single-cell transcriptomic data. bioRxiv. https://doi.org/10.1101/251207.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RamDA-seq mouse ESC to endoderm differentiation data	Hayashi et al., 2018; https://doi.org/10.1038/s41467-018-02866-0	GEO: GSE98664
Processed version of the mouse ESC to endoderm differentiation data	Matsumoto et al., 2017; https://doi.org/10.1093/bioinformatics/btx194	https://github.com/hmatsu1226/SCODE/tree/master/data
ESCAPE database	Xu et al., 2013; https://doi.org/10.1093/database/bat045	http://www.maayanlab.net/ESCAPE/
SCRB-seq retinoic acid-driven differentiation data	Semrau et al., 2017; https://doi.org/10.1038/s41467-017-01076-4	GEO: GSE79578
Mouse bone marrow data from Mouse Cell Atlas	Han et al., 2018; https://doi.org/10.1016/j.cell.2018.02.001	GEO: GSE108097 https://figshare.com/s/865e694ad06d5857db4b
Software and algorithms		
SCODE@28acad6	Matsumoto et al., 2017; https://doi.org/10.1093/bioinformatics/btx194	https://github.com/hmatsu1226/SCODE
SINCERITIES v1.0	Papili Gao et al., 2017; https://doi.org/10.1093/bioinformatics/btx575	http://www.cabsel.ethz.ch/tools/sincerities.html
Jump3@03a7e86	Huynh-Thu and Sanguinetti, 2015; https://doi.org/10.1093/bioinformatics/btu863	https://github.com/vahuynh/Jump3
GENIE3 v1.6.0	Huynh-Thu et al., 2010; https://doi.org/10.1371/journal.pone.0012776	https://doi.org/10.18129/B9.bioc.GENIE3
KinderMiner v1.5.4	Kuusisto et al. 2017; www.ncbi.nlm.nih.gov/pmc/articles/PMC5543342/	https://www.kinderminer.org
SINGE (multiple versions)	This paper.	https://github.com/gitter-lab/SINGE https://doi.org/10.5281/zenodo.2549817
Supplemental scripts, analyses, and files (v3.0)	This paper.	https://github.com/gitter-lab/SINGE-supplemental https://doi.org/10.5281/zenodo.3627325
dyno v0.1.1; dynmethods v1.0.5	Saelens et al., 2019; https://doi.org/10.1038/s41587-019-0071-9	https://github.com/dynverse/dyno https://github.com/dynverse/dynmethods
g:GOSt/g:Profiler vr1760_e93_eg40	Reimand et al., 2016; https://doi.org/10.1093/nar/gkw199	https://biit.cs.ut.ee/gprofiler/gost
Monocle 2.12.0	Qiu et al., 2017; http://doi.org/10.1038/nmeth.4402	https://github.com/cole-trapnell-lab/monocle-release
PAGA Tree (scanpy v1.4.3)	Wolf et al., 2019; https://doi.org/10.1186/s13059-019-1663-x	https://github.com/theislab/scanpy
embeddr v0.99.0	Campbell et al., 2015; https://doi.org/10.1101/027219	https://github.com/kieranrcampbell/embeddr
dyngen@73192c8	Cannoodt et al., 2021; https://doi.org/10.1038/s41467-021-24152-2	https://github.com/dynverse/dyngen

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Anthony Gitter (email: gitter@biostat.wisc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- A MATLAB implementation of SINGE is available at https://github.com/gitter-lab/SINGE under the MIT license and archived on Zenodo (https://doi.org/10.5281/zenodo.2549817). The GitHub repository contains the datasets and default hyperparameter settings used in the manuscript, scripts to generate custom hyperparameter files, and compiled code so that SINGE can be run without a MATLAB license. We also include a Docker image (https://hub.docker.com/r/agitter/singe). Supplemental scripts, analyses, and files are available at https://github.com/gitter-lab/SINGE-supplemental and archived on Zenodo (https://doi.org/10.5281/zenodo.3627325) (Deshpande and Gitter, 2021).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

ESC to endoderm differentiation dataset

The first dataset is from Hayashi et al. Hayashi et al. (2018), who collected single-cell RNA-seq data from 456 cells at five time points over a 72 h duration in which primitive endoderm cells were differentiated from mouse embryonic stem cells. Matsumoto et al. Matsumoto et al. (2017) used Monocle to order these cells along the differentiating process, assigning a pseudotime to each cell. We use their Monocle results in our analyses. The expression dataset is limited to 100 TFs exhibiting the highest variance in expression and 356 cells.

Retinoic acid dataset

The second dataset was obtained from Semrau et al. Semrau et al. (2017), where SCRB-seq data was collected at nine times during a 96 h period from mouse embryonic stem cells differentiating into neuroectoderm and extraembryonic endoderm-like cells. We order the cells using Monocle 2 (Qiu et al., 2017), 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. Matsumoto et al. (2017) applied the original Monocle to the first dataset and we retain their pseudotimes, we prefer Monocle 2, the most recent version available at the time of the analysis, for this case study. After ordering, we limit the scope of the analysis to the 1,886 cells along the longest trajectory of the differentiation process (Figure S1A) exhibiting non-trivial expression levels. Once Monocle 2 orders the cells along a pseudotemporal reference, it can find genes that change in expression as cells progress along pseudotime. We retain the top 626 differentially expressed genes ($q<10^{-5}$) along the pseudotime ranked by Monocle 2 for testing the GRN algorithms.

To obtain another version of the retinoic acid dataset, we use dynverse to select and run an appropriate alternative trajectory inference method. Based on the expected branching topology of the trajectory, we use the graphical user interface dynguidelines to narrow our search to the top four recommended methods: Slingshot (Street et al., 2018), SCORPIUS (Cannoodt et al., 2016), PAGA, and PAGA Tree (Wolf et al., 2019). Running trajectory inference on the entire dataset does not yield any meaningful branching trajectories for any these methods. When we instead limit the genes to the same 626 differentially expressed genes from our Monocle 2 analysis, we obtain a reasonable branching trajectory with PAGA Tree (Figure S7A). We choose the longest branch (backbone) of this inferred trajectory as an alternate input for SINGE network inference.

Mouse bone marrow mesenchyme to erythrocyte differentiation dataset

Starting with the dataset available from the Mouse Cell Atlas (Han et al., 2018),we use the dynverse utility to select an appropriate trajectory inference method. Based on the expected topology of the trajectory, we use the graphical user interface dynguidelines to prioritize four appropriate trajectory inference methods: Slingshot (Street et al., 2018), Embeddr (Campbell et al., 2015), SCORPIUS (Cannoodt et al., 2016), and PAGA (Wolf et al., 2019). Of these, the Embeddr inferred trajectory (Figure S4) was most consistent with the reference trajectory provided in the dynverse database. The ordered single-cell dataset has 3,025 genes and 3,105 cells in a linear trajectory. We normalize the count-based expression data using a log(x+1) transformation (Luecken and Theis, 2019). The default recommendation for using SINGE with count-based data would be to use the hyperparameter '--family poisson.' However, we discovered that the glmnet package for MATLAB suffers a high rate of memory segmentation violations when invoked for larger datasets with the Poisson distribution. Log-transforming the count-based data and using the hyperparameter '--family gaussian' mitigates this issue. Finally, because the dataset is large but sparse, we use *prob-zero-removal* = 0.75, which makes the regression problem smaller by dropping many zero-valued samples and speeds up SINGE.

Dyngen synthetic dataset

We generate two simulated datasets using the dyngen package (Cannoodt et al., 2021). Both datasets come from the same simulated GRN that has 140 genes — 25 TFs, 15 housekeeping genes, and 100 target genes — and 164 edges, most of them originating

from TFs. The first dataset contains 1,000 cells, and the second contains 20,000 cells. The script used for generating the datasets, the datasets in SINGE input format, and the gold standard GRN are available in our supplemental repository (Deshpande and Gitter, 2021). The 20,000 cell version was used solely to evaluate the computational runtime for larger datasets.

QUANTIFICATION AND STATISTICAL ANALYSIS

SINGE infers the GRN that underlies a biological process by aggregating ranked edge lists obtained from an ensemble of GLG tests conducted on ordered single-cell transcriptomic data (Figure 1A). The GLG test is a kernel-based generalization (Bahadori and Liu, 2012) of the Lasso Granger Causality test to facilitate the analysis of causal relationships between irregular time series obtained from a linear stationary VAR model. We first describe the GLG test and then the complete SINGE algorithm.

Generalized Lasso Granger test

The GLG test is used to discover temporal causal networks from irregularly-spaced time series data based on concepts of Granger Causality. In GRN inference, 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 VAR process such that

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

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 $\varepsilon_i(t)$ is measurement error, represented by independently distributed Gaussian random variables. More generally speaking, the unknown $P \times P \times L$ matrix $\mathbf{a}$ comprising 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 (Arnold et al., 2007) for an individual regularly-spaced target series is characterized by the optimization problem

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

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 the 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 GLG test proposed by Bahadori and Liu (Bahadori and Liu, 2012) 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

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

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

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

N_x and N_y can differ, which SINGE exploits for its dropout handling and subsampling.

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 of length N_i is sampled at irregularly-spaced timestamps t_i such that $t_i(n+1) \geq t_i(n)$ for $n = 1, 2, \ldots, N_i - 1$. 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 time series $\mathbf{a}'_{i,j}(t)$ with respect to a given timestamp t. This is accomplished by assigning a new timestamp $t_a(l) = t - l\Delta(t)$ to each lagged coefficient $a_{i,j}(l)$, where Δt represents the time-lag between successive lagged coefficients in $a_{i,j}$. Thus, this time series can be viewed as the sequence of (lagged time, coefficient value) pairs.

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

Note that for $t_1 \neq t_2$, the two time series $\mathbf{a}'_{i,j}(t_1)$ and $\mathbf{a}'_{i,j}(t_2)$ would have the same coefficient values $a_{i,j}(l)$ but different lagged timestamps. For example, if we select hyperparameter values $\Delta(t) = 5$, and $L = 3$, then, for $t_1 = 50$ and $t_2 = 75$, we have

$$\mathbf{a}'_{i,j}(50) = \{(35, a_{i,j}(3)), (40, a_{i,j}(2)), (45, a_{i,j}(1))\}$$
$$\mathbf{a}'_{i,j}(75) = \{(60, a_{i,j}(3)), (65, a_{i,j}(2)), (70, a_{i,j}(1))\}$$

respectively.

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

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

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 $\left|\sum_l a_{i,j}(l)\right|$, 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 (Qian et al., 2013), it supports count-based expression data (e.g. from unique molecular identifiers) by assuming a Poisson distribution for the expression levels.

SINGE's Granger Causality formulation models a strict delay between the regulator and target gene expression in pseudotime. The time-lagged expression relationships between regulators and target genes are motivated by simulations of transcription kinetics that naturally induce such lags (Gillespie, 1977; Thattai and Oudenaarden, 2001). Therefore, SINGE is not intended to identify any 'instantaneous' regulatory relationships.

Single-cell inference of networks using Granger ensembles

In this section, we describe how the SINGE algorithm, which has the GLG test at its core, infers GRNs from single-cell expression data. The SINGE algorithm takes ordered single-cell RNA-seq 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 count, with an optional subsampling stage increasing the effective ensemble size.

SINGE input

The input to SINGE is ordered single-cell gene expression data, with a pseudotime assigned to each cell that represents its position along the biological process. 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. 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.

If the single-cell dataset is not already ordered, any cell ordering method that assigns continuous pseudotimes can be used to annotate the cells before running SINGE. For the retinoic acid dataset, we apply Monocle 2 (Qiu et al., 2017), which uses reverse graph embedding to identify branching processes. For other trajectory inference methods, we use the dynverse package (Saelens et al., 2019), which provides a streamlined approach to benchmark and use trajectory inference methods and can prioritize algorithms based on trajectory type, dataset size, and other criteria.

By default, SINGE assumes that all genes are potential regulators. It can optionally limit the possible regulators using regulator indices regix$\subseteq\{1,2,\ldots,P\}$ provided by the user. SINGE will infer a GRN with potential regulator genes limited to only those corresponding to the *regix* indices. Using regulator indices can substantially improve SINGE's runtime.

By default, SINGE also assumes that the ordered single-cell data correspond to a linear biological process. However, if the inferred trajectory has a branching topology, this information can be passed to SINGE by creating an optional binary matrix called *branches* with N_{cells} rows and $N_{branches}$ columns. If *branches*[*i*,*b*] = 1, it indicates that the *i*-th cell is a member of the *b*-th branch of the trajectory. This allows SINGE to handle any type of acyclic trajectory.

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 (Linderman et al., 2018). There have been efforts to overcome this problem by imputing the missing values (van Dijk et al., 2018; Linderman et al., 2018). However, inappropriate imputation can negatively impact differential expression testing (Andrews and Hemberg, 2018) and can have a positive, neutral, or negative effect on Monocle's pseudotimes depending on the choice of algorithm (Zhang and Zhang, 2018).

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, SINGE 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*.

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 the 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 can be 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.

Subsampling stage

SINGE includes an optional stage that increases the effective ensemble size by subsampling versions of the original single-cell data. The subsampling can make the inferred GRN more robust to outliers in the gene expression data. Specifically, for each hyperparameter combination, we generate $N_{subsample}$ (default 10) data replicates. Because GLG can handle irregular time series, we have the option to use two different strategies for subsampling. The simplest strategy would be to randomly remove a small subset of cells from the dataset. This ensures that all genes' pseudotime series have the same pseudo-timestamps. However, removing entire cells could ignore important cells in rare transient states.

We instead use an alternate strategy that randomly removes samples independently from each gene's pseudotime series. Using this strategy, the probability of removing an entire pseudo-timestamp (cell) is greatly diminished. However, no two genes have the same series of pseudo-timestamps, each has a unique irregular time series with high probability. In our experiments, we independently remove samples for each gene using a probability of sample removal of 0.2, the SINGE default. The SINGE subsampling is similar to bagging (Breiman, 1996) except that the sampling is without replacement and it uses a different aggregation approach.

GLG runs and modified Borda aggregation

After enumerating all hyperparameter combinations and subsampled replicates, SINGE runs GLG on each subsampled replicate using the different hyperparameter combinations. At the end of each GLG test, we obtain an adjacency matrix **A** using

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

where **a** is the $P\times P\times L$ coefficient matrix output from the GLG test. The matrix **A** represents one candidate GRN, with the magnitude of each element representing the edge weight assigned to the corresponding regulator-gene interaction. These edge weights are used to rank the possible regulator-gene interactions. A rank is only assigned to those interactions that correspond to a nonzero element of **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 (van Erp and Schomaker, 2000). The Borda count aggregates ranked lists by defining a scoring rule that assigns weights to the items in each ranked list and summing the weights to obtain a final consensus ranking. The goal is to favor items, in our case regulator-gene interactions, that are consistently ranked high over those that are ranked high only

occasionally or not at all. The traditional Borda count scoring rule assigns a weight of N for the first ranked item in a list, $N-1$ to the second, and so on (van Erp and Schomaker, 2000). Alternative scoring rules, such as the Dowdall rule (Fraenkel and Grofman, 2014), assign weights that decay more quickly, placing more relative importance on the top-ranked items.

We use a scoring rule that assigns weights of $1/i^2$ for the i-th ranked interaction within each individual ranked list from a single GLG test. The weight is zero for an unranked regulator-gene interaction. This scoring rule was selected based on empirical tests with the ESC to endoderm differentiation dataset. The final SINGE score of each interaction is obtained by summing the weights assigned to that interaction across all ranked lists. This score is subsequently used for the final GRN edge ranking. We also obtain the top regulators of the biological process by summing the SINGE scores of all outgoing edges for each regulator and sorting the regulators in order of decreasing magnitude. In the case of branching processes, SINGE's default behavior is to perform the modified Borda aggregation on all branches together to obtain one output for the overall branching process. Alternatively, a user can obtain branch-specific SINGE outputs by storing the individual GLG test results in separate branch-specific directory and performing the modified Borda aggregation on each set of results separately.

Similar ensembling and aggregation strategies are widely used in GRN inference in order to improve the robustness of the predicted networks, reduce sensitivity to noise, and avoid false positives. SINGE's modified Borda count aggregation is one specific strategy among many related ideas. It emphasizes the interaction ranking instead of the magnitude of the $\mathbf{A}_{ij}$ coefficients, which are difficult to compare directly when combining results from GLG runs that use different degrees of regularization λ. SINGE's aggregation is closely related to the stability selection (Meinshausen and Bühlmann, 2010) in TIGRESS (Haury et al., 2012) except SINGE aggregates predictions over many hyperparameter combinations and its randomization comes from the randomly removed observations during the subsampling stage instead of randomly rescaling TF expression. Unlike other unsupervised aggregation approaches (Ahsen et al., 2019), SINGE's modified Borda counts do not assume that the ranked interaction lists are conditionally independent. Furthermore, SINGE's aggregation does not require generating a null distribution of $\mathbf{A}_{ij}$ coefficients from permuted data (Lu et al., 2021), which is computationally more expensive but has the benefit of providing interaction false discovery rates.

Versions

We used SINGE version 0.1.0 for nearly all analyses, except Figures 5C and S7B and Table 2, which also include results from newer versions of SINGE as indicated. In addition, the mouse bone marrow and dyngen analyses were performed using only version 0.3.0, which included code optimizations for improving stability and compute time for large datasets. SINGE version 0.5.0 reported in Table 2 was pre-release commit 5630ed3. See https://github.com/gitter-lab/SINGE for the latest release notes and usage recommendations.

Case study hyperparameters

The hyperparameter values used to generate the GLG ensembles for all case studies are tabulated here. Only specific pairs of Δt and L are considered instead of all possible combinations. 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 dataset. For most of the main case studies, we use the default mode with *prob-zero-removal* = 0, only changing it when we analyze its effect on GRN performance (Figure 4B) or to reduce runtime on the mouse bone marrow dataset. The total number of GLG tests, accounting for hyperparameter diversity and subsampling, is

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

SINGE hyperparameter combinations

Hyperparameter(s)	Property	Values
λ	Sparsity	$0, 0.01, 0.02, 0.05, 0.1$
$(\Delta t, L)$	(Time resolution, num lags)	$(3,5); (5,9); (9,5); (5,15); (15,5)$
σ	Kernel width	$0.5, 1, 2, 4$

Existing GRN methods

In addition to SINGE, we use GENIE3, SCODE, SINCERITIES, and Jump3 to infer GRNs. The GENIE3 algorithm (Huynh-Thu et al., 2010), which was originally designed for bulk transcriptomics, can also be applied to single-cell data (Aibar et al., 2017) and acts as a reference method that does not use pseudotemporal ordering information. We used the default settings for GENIE3 and SINCERITIES. We used the same SCODE settings as in Matsumoto et al. (Matsumoto et al., 2017) 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 dataset to account for the much larger network of 626 genes. Jump3 only uses cell ordering information, not pseudotimes. We used *noiseVar.obsnoise* = 0.1, but all other Jump3 settings were the defaults. Because Jump3 did not terminate in

a reasonable amount of time on the full retinoic acid 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, 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 dataset.

Evaluation

To evaluate the GRNs from the ESC to endoderm differentiation and retinoic acid datasets, we use the ESCAPE database (Xu et al., 2013) as a gold standard, namely the cataloged ChIP-chip, ChIP-seq, loss-of-function (LoF), and gain-of-function (GoF) experiments. Each GRN 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 average early precision (**E**) for comparison. Precision-recall is preferable to the receiver operating characteristic for evaluating biological network inference due to the sparsity of the gold standard (Schrynemackers et al., 2013).

We evaluate the GRNs from the simulated dyngen dataset using the dyngen model as a gold standard. This simulated GRN only contains direct edges from regulators to targets. We also evaluate the predicted networks with respect to extended gold standards that include indirect interactions, which considers transitive interactions. That is, if interactions $A \rightarrow B$ and $B \rightarrow C$ exist, whether direct or indirect, then $A \rightarrow C$ also exists in the modified gold standard. Starting with the original direct dyngen GRN with 164 interactions, we iteratively apply this transitive property to obtain gold standards with different levels of indirect gene interactions. These indirect networks contain 352 edges after the first iteration, 744 edges after the second iteration, and 921 edges after the third iteration. Further iterations do not expand the gold standard GRN.

When assessing the MSE for Pou5f1 expression predictions in the ESC to endoderm differentiation dataset, all MSE values are the mean of 10 GLG runs with independent insample-outsample splits.

ESCAPE database

The ESCAPE database (Xu et al., 2013) is a repository cataloging numerous experiments conducted on human and mouse ESCs. We use the ChIP-chip, ChIP-seq, LoF, and GoF experiments as a gold standard to evaluate the inferred GRNs. ChIP-based and perturbation-based data are known to have low overlap (Gitter et al., 2009). Despite ESCAPE being one of the most comprehensive repositories of such experimental results, it does not have reference data for all predicted regulators. 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 interactions in the ChIP-chip/ChIP-seq and LoF/GoF databases related to the genes from the single-cell data being analyzed. Interactions not documented in ESCAPE 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 not 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 1,000 target genes in the ChIP-chip/ChIP-seq data and 500 target genes in the LoF/GoF data.

Average early precision

Because a majority of SINGE's hyperparameter sets predict a sparse regulatory network, it is better suited to ranking the top GRN interactions instead of ranking all of them. Average precision, which summarizes the entire precision-recall curve, 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. For consistency across the evaluations, we generally assess whether there is a 10% improvement in average precision or average early precision when comparing two predicted GRNs. If the performance of one GRN is not at least 10% better than the other, we consider them to be approximately similar.

KinderMiner and Gene Ontology enrichment

We performed KinderMiner (v1.5.4) (Kuusisto et al., 2017) analysis on the SINGE 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 26,877,474 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. We provide the raw KinderMiner results obtained using the search setting 'anySpeciesSEP' in a supplementary repository (Deshpande and Gitter, 2021). 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 SINGE using the g:GOSt tool in g:Profiler (Reimand et al., 2016) version r1760_e93_eg40. We consider only Gene Ontology (Ashburner et al., 2000) biological process terms and specify 'mus musculus' as the organism. The candidate regulator list from SINGE 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. We provide the complete output of the g:GOSt test in a supplementary repository (Deshpande and Gitter, 2021). 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 Database's Mammalian Phenotype Ontology Annotations (Bult et al., 2019).