Supplemental Information

SERGIO: a single-cell expression simulator guided by gene regulatory networks

Payam Dibaeinia[1], Saurabh Sinha[1,2,3,4*]

[1] Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA
[2] Carl R. Woese Institute of Genomic Biology, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA
[3] Cancer Center at Illinois, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA
[4] Lead Contact
[*] Correspondence: sinhas@illinois.edu

Supplemental Figures

Supplemental Figure S1

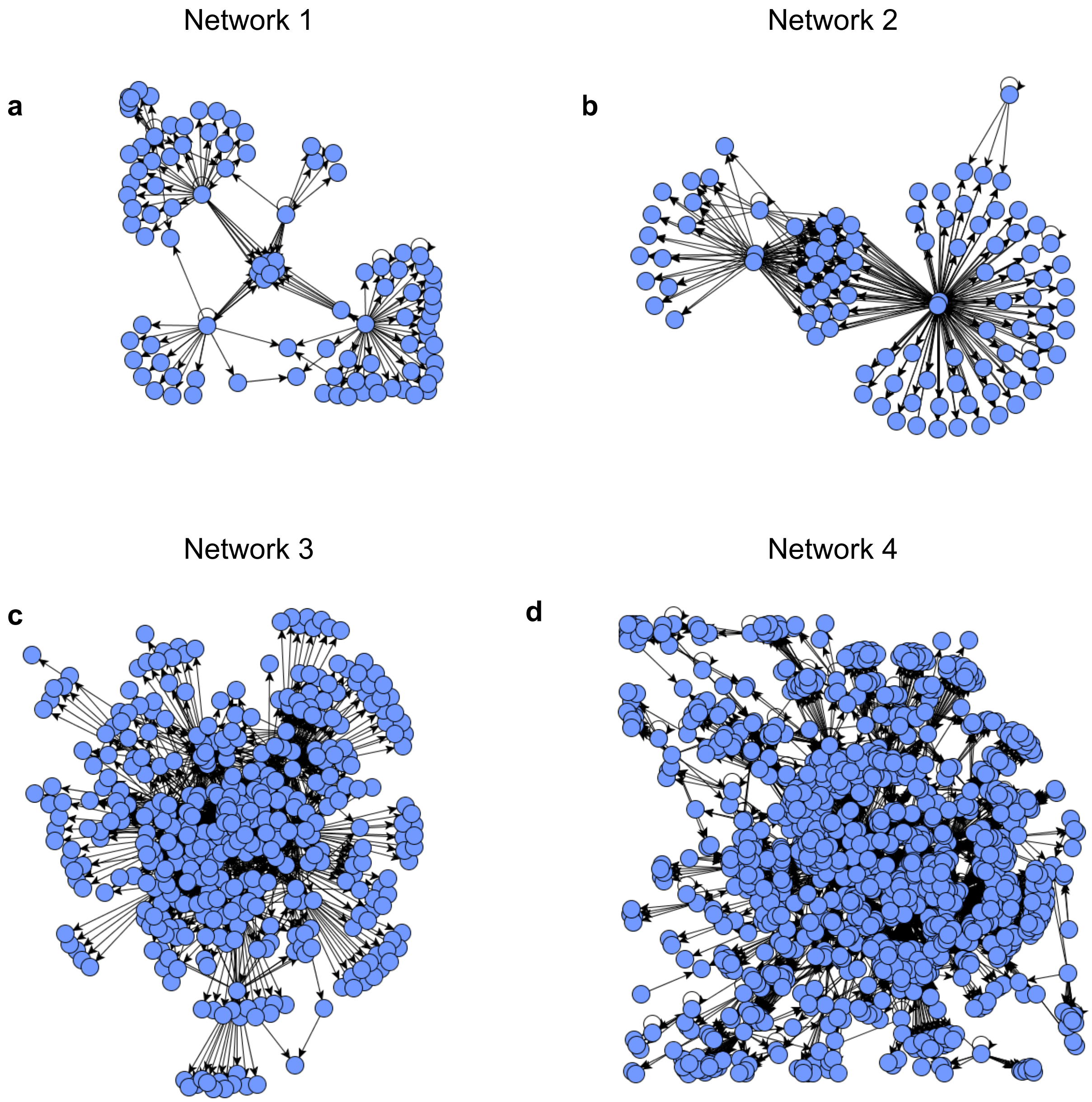

Figure S1: *Related to Figure 2.* The structure of four gene regulatory networks used in this study titled by their network ID. These figures were generated using GNW package (Schaffter et al., 2011). Note that all the auto-regulatory edges as well as cycles were removed prior to feeding networks to Sergio although they are present in this figure. **(a)** Shows network 1, sampled from E.coli, containing 100 genes and 137 regulatory edges. **(b)** Shows network ID 2, sampled from E.coli, containing 100 genes and 258 regulatory

edges. **(c)** Shows network ID 3, sampled from S. cerevisiae, containing 400 genes and 1155 regulatory edges. **(d)** Shows network ID 4, sampled from E. coli, containing 1200 genes and 2713 regulatory edges.

Supplemental Figure S2

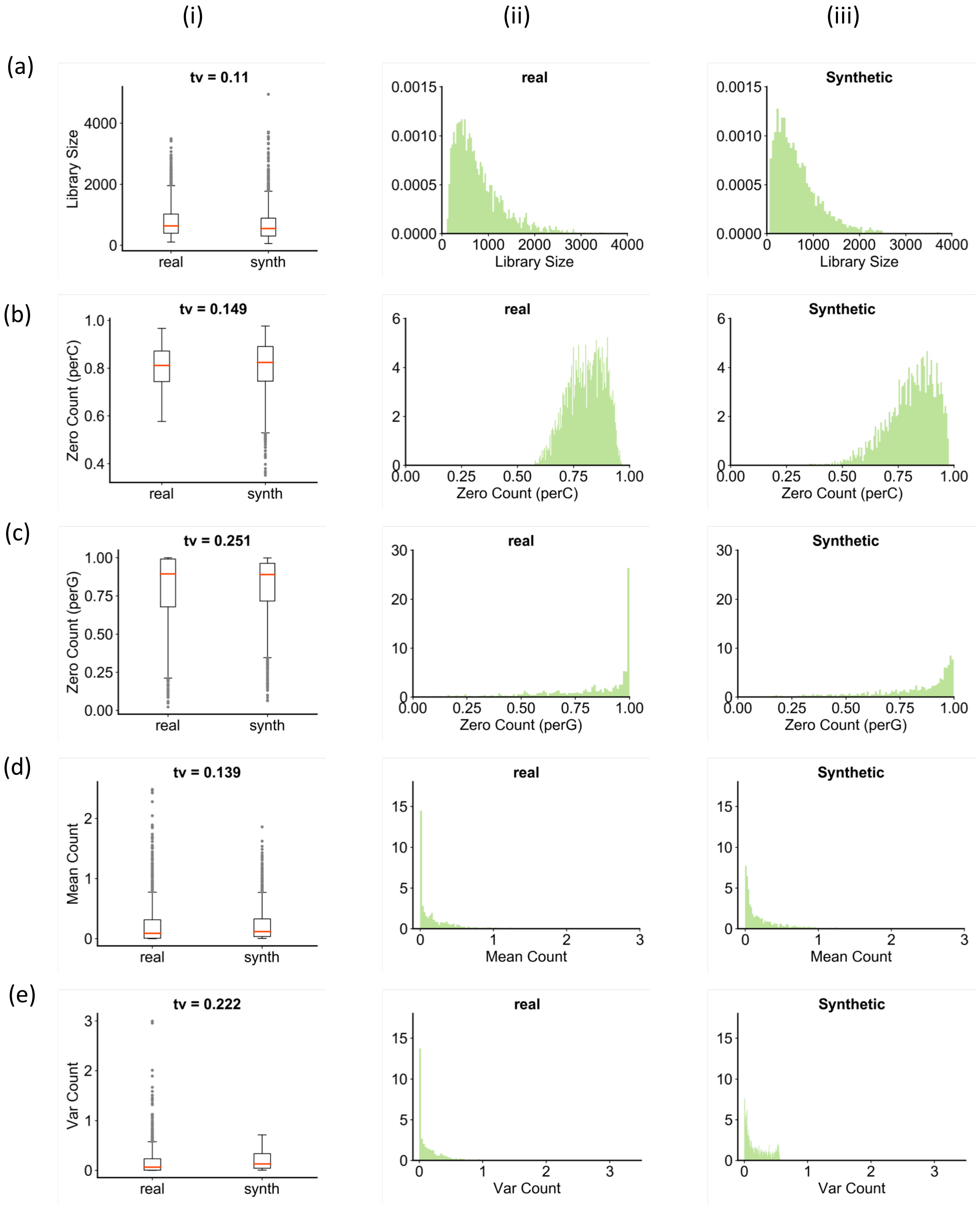

Figure S2: *Related to Figure 2.* To interpret the total variation values and assess the quality of match between the real and synthetic data, for each statistic we looked at a pair of simulated replicate of DS3 and a real sample which their total variation is close to the median of the total variations of the corresponding statistic. This figure gives a qualitative understanding of the total variation score, which is a number between 0 and 1 reflecting how well two distributions match. Each row represents one of the quantities studied in Figure 2, and shows the distribution of that quantity in one of the simulated replicates and one of the real data sets; the two data sets selected for display here have a total variation ("tv") that is typical for that quantity. **(i)** This column compares the distribution of synthetic against the real data as a box plot. **(ii)** This column shows an alternative visualization of the distribution of the quantity of interest in real data. **(iii)** This column shows an alternative visualization of the distribution of quantity of interest in the synthetic data. The quantities examined include **(a)** library sizes **(b)** zero counts per cell **(c)** zero counts per gene **(d)** mean mRNA counts and **(e)** variance of mRNA counts.

Supplemental Figure S3

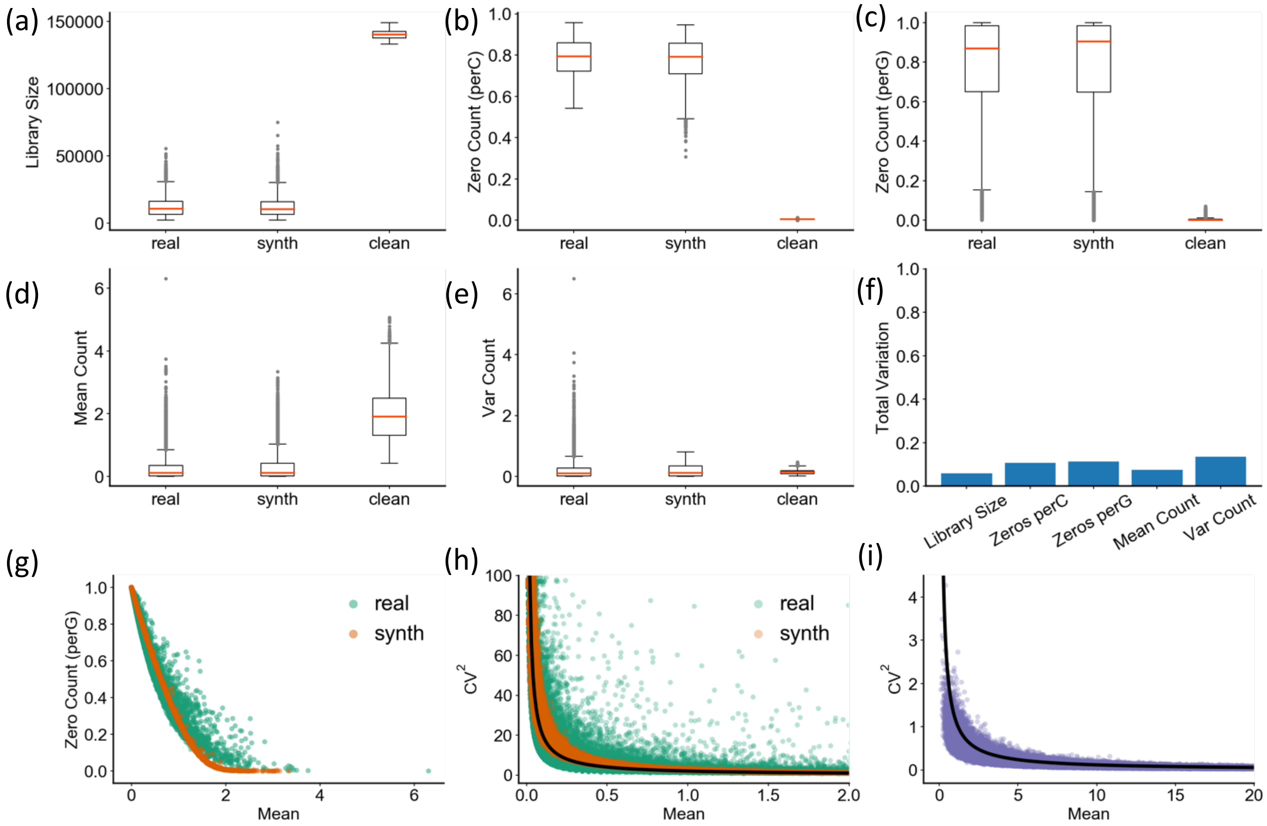

Figure S3: *Related to Results.* A curated gene regulatory network for mouse was obtained from RegNetwork database (Liu et al., 2015). After preprocessing and excluding the genes that are not present in the mouse brain scRNA-seq data (Zeisel et al., 2015) we obtained a gene regulatory network (GRN) containing 15272 genes and 76483 gene-gene interactions. Due to the absence of prior knowledge about the regulatory role of the majority of these interactions, for each interaction we randomly assigned either an activation (probability of 75%) or a repression role (probability of 25%). The interaction strengths were uniformly sampled from the range 1 to 5 (similar to DS1-15) and master regulators' production rates were sampled using the same settings used for DS2-8 to represent nine established cell types. We simulated this GRN using SERGIO to obtain one synthetic expression data containing 15272 genes and 3600 single-cells. Subsequently, we added technical noise by comparing this data against the mouse brain scRNA-seq data set (Zeisel et al., 2015) which contains the expression of the same 15272 genes (genes that are not present in the RegNetwork's GRN were excluded) in 3005 single-cells. The quantities examined for adding technical noise include **(a)** library sizes **(b)** zero counts per cell **(c)** zero counts per gene **(d)** mean mRNA counts and **(e)** variance of mRNA counts. **(f)** Total variation between the real and simulated data after adding technical noise (synth) are small (<0.2) and are as good as total variations for the same statistics in DS1-8. **(g)** The inverse relation between genes' mean expression and zero counts (per gene) present in the real data was reproduced after adding technical noise.

(h) Inverse relation between squared coefficient of variation and mean expression of genes over all single-cells is matched between real and simulated data after adding technical noise. The black line shows an arbitrary function of form $y \sim 1/x$ which matches with the observed behavior in both real and synthetic data. As is evident from this plot, highly variable genes in the real data with mean expression > 0.15 were not captured in the simulated data. Addition tuning of parameters of SERGIO might help tune the variance of genes and improve the quality of match between the real and synthetic data. **(i)** The inverse relation of form $y \sim 1/x$ is not a result of technical noise and is also observed in clean simulated data.

Supplemental Figure S4

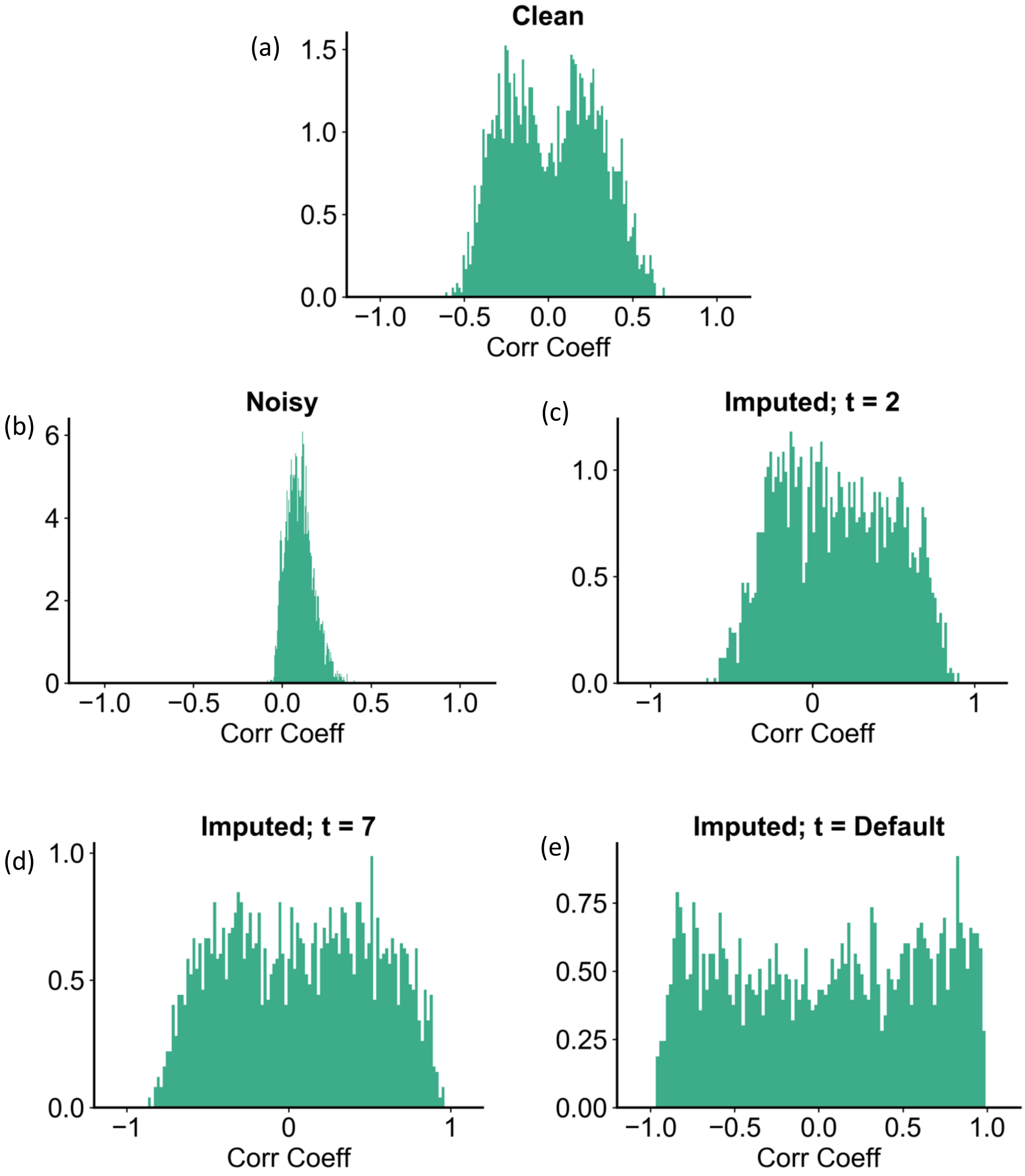

Figure S4: *Related to Figure 4.* Shows the distribution of correlation coefficients between all pairs of interacting genes (regulator – target pairs present in the "ground truth" GRN that was used for simulations) in clean and noisy data of one simulated replicate of DS3, as well as in data imputed by MAGIC (van Dijk et al., 2018). **(a)** Represents the distribution of TF-gene expression correlation coefficients in the clean simulated data. **(b)** Represents correlation coefficients in the noisy data. After adding technical noise, the co-expression signal in the data (panel a) is severely distorted. **(c)** Distribution of correlation coefficients

in the data underlying panel b, after imputation with MAGIC (van Dijk et al., 2018) using parameter setting t = 2. Even upon setting *t* to such a small value, several spurious co-expression signals (right tail of distribution as compared to panel a) emerged in the data, compared to the ground truth shown in panel a. **(d)** Distribution of correlation coefficients after imputation with MAGIC using t = 7. This introduces even more false co-expression signals compared to panel c. **(e)** MAGIC imputed data with default *t* setting. We observe almost a uniform distribution over the whole range of correlation coefficients, showing a large number of false positives of co-expressed TF-gene pairs.

Supplemental Figure S5

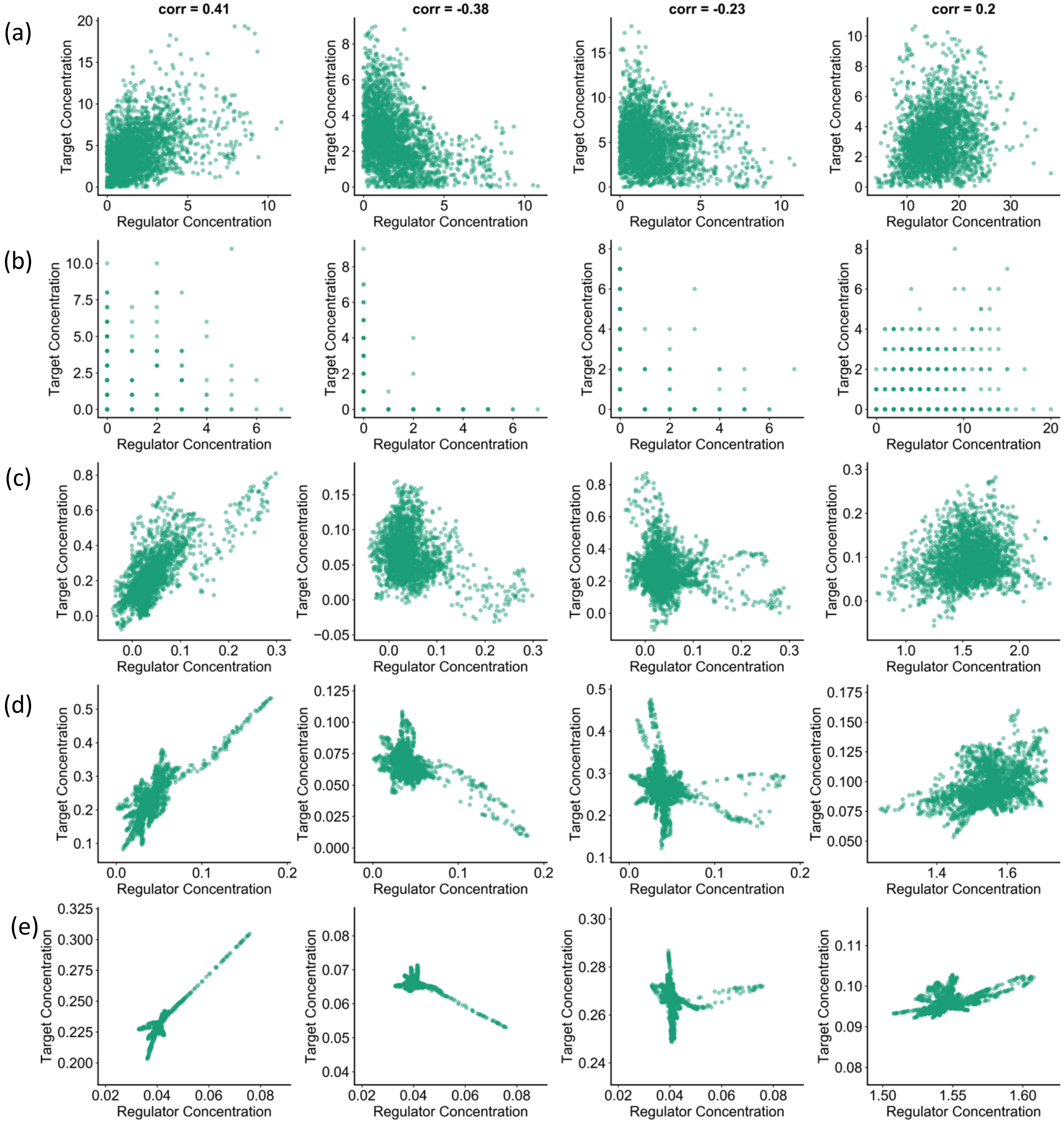

Figure S5: *Related to Figure 4.* Correlation structures in clean and noisy simulated data sets as well as imputed versions of the latter. Columns correspond to four arbitrarily selected regulatory interactions (TF-gene pairs) in DS3 (network 4). **(a)** Clean simulated data. Each panel shows the expressions of the chosen regulator and target pair, in single cells, and the Pearson correlation coefficient between these two observables is noted in

caption at the top. **(b)** TF and target gene expression values for the same TF-gene pairs as in (a), after technical noise has been added. The simulated UMI counts are shown. **(c)** TF and target gene expression values for the same TF-gene pairs as in (b), after imputed using MAGIC with *t* = *2*. Note that level of co-expression appears greater than that in clean data ("ground truth"). **(d-e)** Same as (c), but with MAGIC run using *t* = *7* and *t* = default respectively.

Supplemental Figure S6

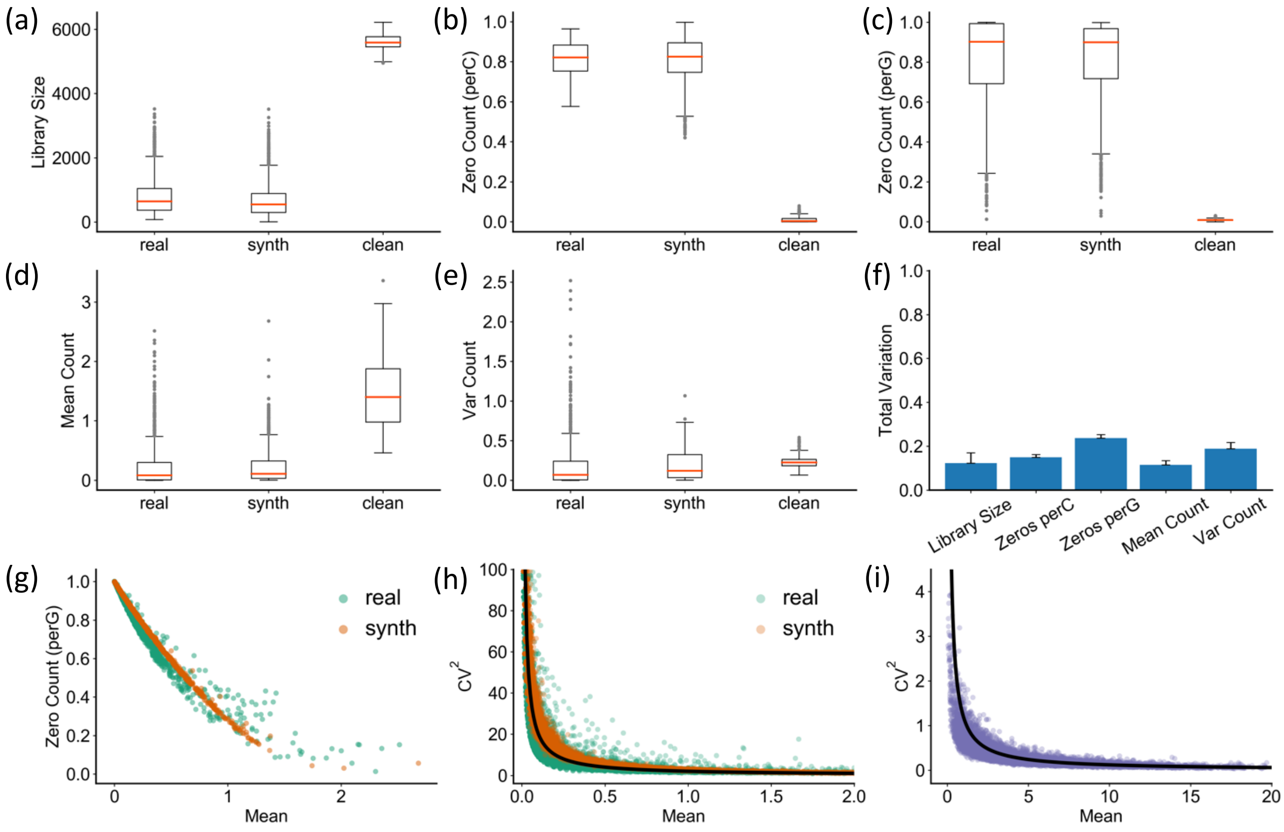

Figure S6: *Related to Results and STAR Methods.* We used the GRN containing 1200 genes (the same network as that used for DS3-8) to simulate data sets with 15 replicates using the mode of SERGIO that includes activator-activator cooperative regulation. Subsequently, we added technical noise by comparing this data set against 50 samples obtained from the mouse brain scRNA-seq data set (Zeisel et al., 2015) (the same samples as those used for adding noise to DS3). The quantities examined for adding technical noise include **(a)** library sizes, **(b)** zero counts per cell, **(c)** zero counts per gene, **(d)** mean mRNA counts, and **(e)** variance of mRNA counts. **(f)** Total variation between each sample and simulated replicate after adding technical noise. **(g)** The inverse relation between genes' mean expression and zero counts (per gene) present in the real data was reproduced after adding technical noise. **(h)** Inverse relation between squared coefficient of variation and mean expression of genes over all single-cells is matched between real and simulated data after adding technical noise. The black line shows an arbitrary function of form $y \sim 1/x$ which completely matches with the observed behavior in both real and synthetic data. **(i)** The inverse relation of form $y \sim 1/x$ is not a result of technical noise and is also observed in clean simulated data.

Supplemental Figure S7

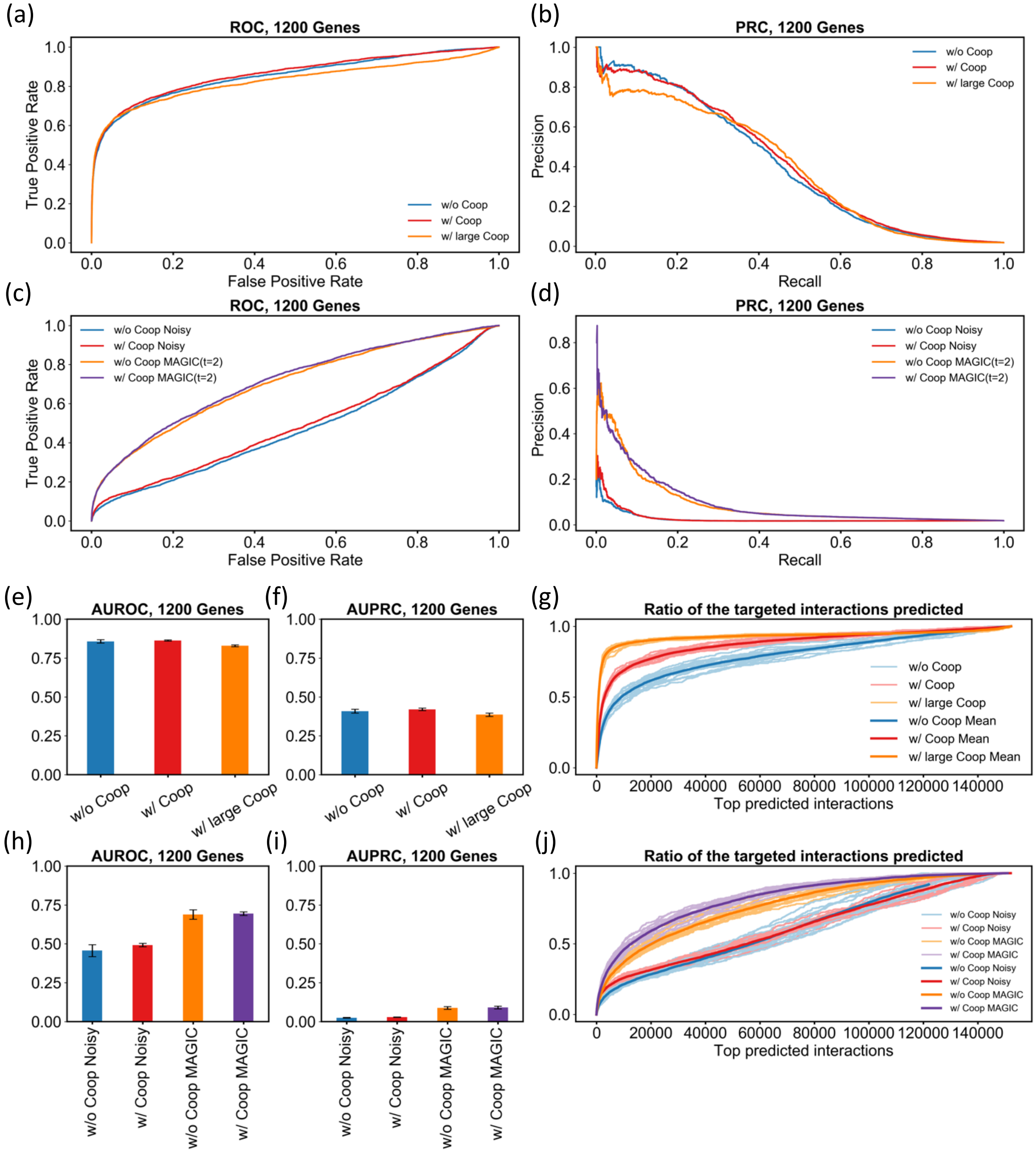

Figure S7: *Related to Results and STAR Methods.* Comparing the performance of GENIE3 on three clean simulated data sets, namely a data set without cooperative regulation (DS3; called "w/o Coop" here), a data set with moderate cooperative regulation ("w/ Coop"), and a data set with large cooperative regulation ("w/ large Coop") dominating non-cooperative effects. All three data sets were simulated with the same underlying GRN (Network ID 4) and in 15 replicates. **(a)** ROC and **(b)** PRC of GRN prediction by GENIE3

(on one simulated replicate from each data set) shows that inclusion of cooperative regulation does not impact GRN inference. We next added technical noise to the data set with moderate cooperativity (w/ Coop), in a way that matches the noise in a mouse brain scRNA-seq data set (Zeisel et al., 2015) (Figure S6). GENIE3 was applied on this noisy data set before and after imputation by MAGIC (t=2). **(c)** ROC and **(d)** PRC of GRN predicition by GENIE3 (on one simulated replicate from each data set) confirms that without imputation, GRN inference from noisy data is not impacted by inclusion of cooperative regulation (compare “w/o Coop noisy” to “w/ Coop noisy”). Moreover, similar to our observations on data without cooperative regulation, using MAGIC (t=2) to impute noisy data increases signal for GRN inference by GENIE3, even when the data was generated by a GRN with cooperative regulation (compare “w/ Coop noisy” to “w/ Coop MAGIC(t=2)”)). Mean AUROC **(e)** and AUPRC **(f)** over 15 replicates of each of the three clean simulated data sets and mean AUROC **(h)** and AUPRC **(i)** over 15 replicates of each of the four noisy simulated data sets (two before imputation and two after imputation by MAGIC) show the same trend discussed in (a-d). We next evaluated the enrichment of regulator-target interactions that are affected by cooperativity (e.g., interactions B-A and C-A are said to be affected by cooperativity if B and C cooperatively regulate A) among the top-k predictions of GENIE3 obtained from data sets simulated with cooperative regulation. Also, to assess the impact of cooperativity on this enrichment, we collected the interactions affected by cooperativity (in cooperativity simulations) and evaluated their enrichments among GENIE3 predictions obtained from simulated data in the absence of cooperativity. Note that this is feasible because we used the same GRN topology in the two modes of simulation. **(g)** Shown is the fraction of such interactions recovered in the top-k predictions (x axis) of GENIE3 applied to clean simulated data sets. Although inclusion of cooperative regulation does not impact GRN inference from simulated data (e.g., Figure S7 a-b), interactions that were affected by cooperativity are more enriched among the top GENIE3 predictions as compared to the same interactions in the absence of cooperativity. **(j)** Shows the enrichment of interactions affected by cooperativity among the top-k predictions of GENIE3 applied to noisy simulated data sets before and after imputation by MAGIC. Imputation by MAGIC increases the enrichment of such interactions (purple versus red curve).

Supplemental Figure S8

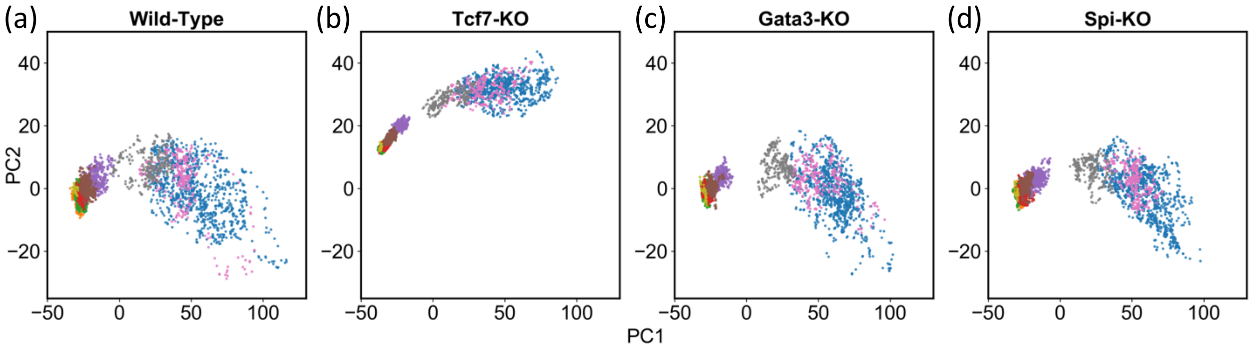

Figure S8: *Related to Figure 5.* PC representation of wild-type (WT) and knockout (KO) simulated trajectories using GRN obtained by GENIE3 **(a)** Two-dimensional PC representation of WT trajectory (identical to Figure 5b, right). **(b)** Projection of Tcf-KO simulated trajectory on the PC space of WT trajectory. The average Euclidian distance between cluster centers of Tcf7-KO and WT trajectories in 10 dimensional PC space is 8.2. **(c)** Projection of Gata3-KO simulated trajectory on the PC space of WT trajectory. The average Euclidian distance between cluster centers of Gata3-KO and WT trajectories in 10 dimensional PC space is 1.0. **(d)** Projection of Spi-KO simulated trajectory on the PC space of WT trajectory. The average Euclidian distance between cluster centers of Spi-KO and WT trajectories in 10 dimensional PC space is 1.2.

Supplemental Figure S9

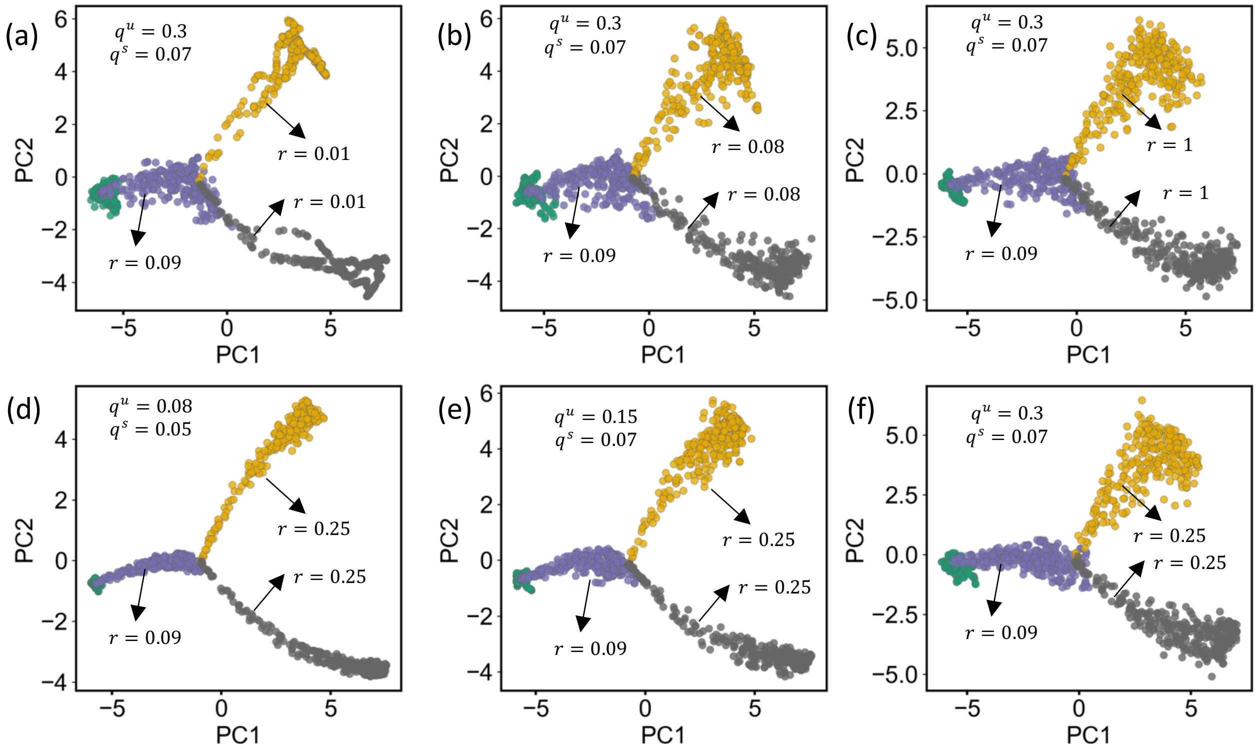

Figure S9: *Related to Figure 6.* User can control the thickness of differentiation path and the dispersion of cells around the trajectory. The user defined migration rate r controls the number of paths that are simulated between two cell types (each edge in the provided differentiation graph). For a given number of cells per cell type ($nCells$) and migration rate r, a total number of $r \times nCells$ paths is simulated between the two cell types. Finally, single-cells are randomly sampled from the aggregation of all simulated paths. **(a,b,c)** For fixed unspliced and spliced noise amplitudes q^u and q^s respectively, increasing the migration rate r increases the thickness of the simulated differentiation path as single-cells are sampled from a bigger pool of cells in between the two origin and end cell types. **(d,e,f)** For fixed migration rates r, increasing the spliced and unspliced noise amplitudes increases the dispersion of single cells because the higher stochastic noise increases the variance among single-cells.

Supplemental Figure S10

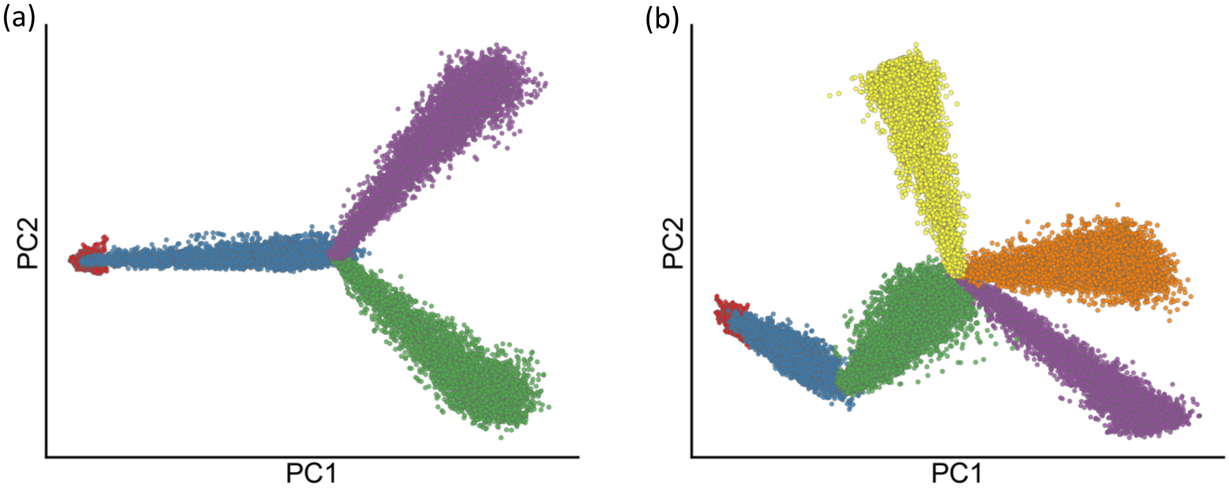

Figure S10: *Related to Figure 6.* **(a)** PCA representation of single cells in the clean simulated version of DS13. This data set contains 24000 cells in total. For simulating DS13 we used the same GRN, parameter settings, and differentiation graph as we used for DS10. **(b)** PCA representation of single cells in the clean simulated version of DS14. This data set contains 36000 cells in total. For simulating DS14 we used the same GRN, parameter settings, and differentiation graph as we used for DS11.

Supplemental Figure S11

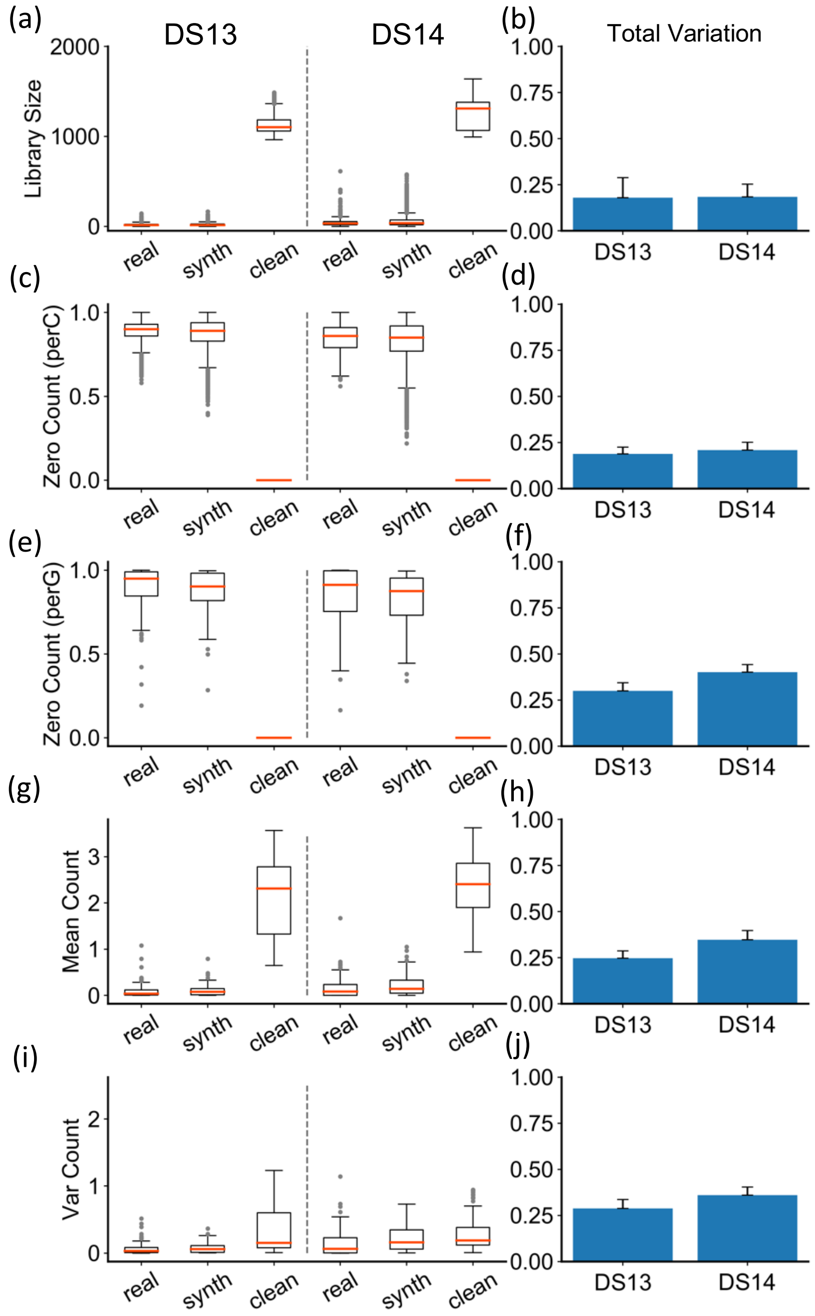

Figure S11: *Related to Figure 6.* Comparisons between differentiation data sets generated by SERGIO and real scRNA-seq data sets. We show the distributions of per-cell quantities in (a,c), and per-gene quantities in (e, g, i), for DS13 and DS14 separated by dashed lines. These comparisons are shown between one sample from the real data set ("real"), the clean simulated data ("clean"), and its technical noise-added version ("synth"). The real data used for DS13 is a published 10X genomics single-cell data of

dentate gyrus of mouse hippocampus (Hochgerner et al., 2018), and for DS14 we used a single-cell RNA-seq data set from the mouse cerebral cortex (Zeisel et al., 2015). More comprehensive comparisons – between the noisy simulated data and every real sample – are shown in panels to the right: the total variation (see METHODS) is calculated to compare the real and synthetic distributions and the average total variation across all comparisons is shown in panels (b, d, f, h, j). **(a,b)** Distributions and total variation of library sizes. **(c,d)** Distributions and total variation of zero counts per cell (normalized by number of genes). **(e,f)** Distributions and total variations of zero counts per gene (normalized by total number of cells). **(g,h)** Distributions and total variations of genes' mean expression. **(i,j)** Distributions and total variations of genes' expression variances.

Supplemental Figure S12

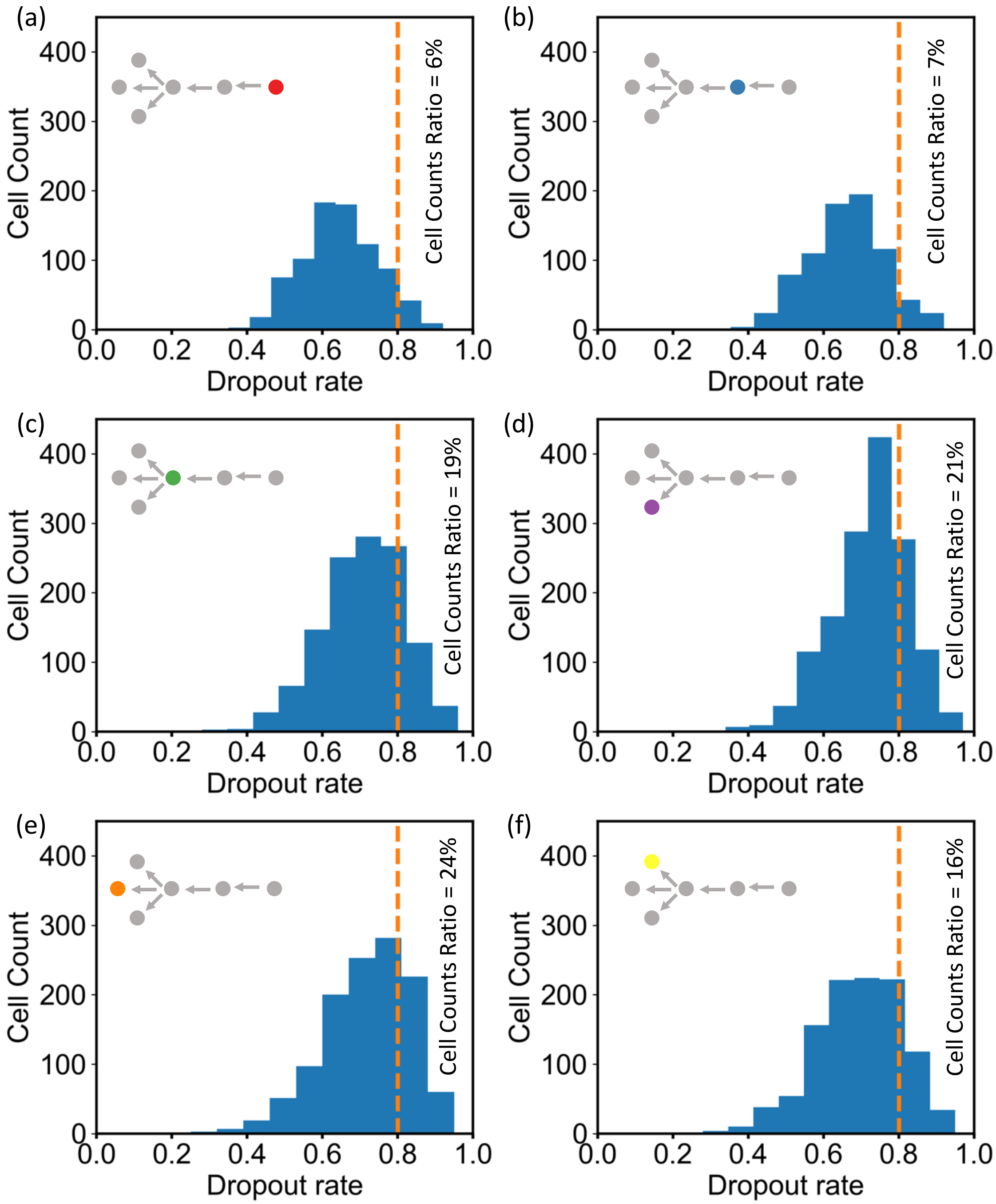

Figure S12: *Related to Figure 7.* Distributions of dropout rates in single cells belonging to **(a)** red, **(b)** blue, **(c)** green, **(d)** purple, **(e)** orange, and **(f)** yellow, cell types. For each cell type, the ratio of cells which have 80% or more dropout rate is denoted. The orange cell type suffers the most from dropout and its distribution shows the most skewness toward large dropout rates as compared to other cell types. This is consistent with the poor correlation observed for this cell type, between the inferred pseudotime from the noisy and clean expression matrices (Figure 7g).

Supplemental Tables

Supplemental Table S1: *Related to STAR Methods.* Technical noise parameters used in this study

DS-ID	Outlier Genes			Library Size		Dropouts		Low Quality threshold *
	π^O	μ^O	σ^O	μ^L	σ^L	k	q	τ
1	0.01	0.8	1	4.8	0.3	20	82	5
2	0.01	0.8	1	6	0.4	12	80	5
3	0.01	0.8	1	7	0.4	8	80	5
4	0.01	3	1	6	0.3	8	74	5
5	0.01	3	1	6	0.4	8	82	5
6	0.01	5	1	4.5	0.7	8	45	5
7	0.01	3	1	4.4	0.8	8	85	5
8	0.01	4.5	1	10.8	0.55	2	92	2500
13	0.01	0.8	1	3.6	0.4	8	70	5
14	0.01	0.8	1	5	0.4	4	80	5

(*) Cells with a total count $< \tau$ were considered as low quality cells and were removed from both real samples and synthetic replicates.

Supplemental Table S2: *Related to STAR Methods.* Parameter settings used for running Singe (Deshpande et al., 2019)

Parameter	Value
λ	0, 0.1, 0.01
(dT,num_lags)	(3,5), (5,9), (9,5), (5,15), (15,5)
kernel_width	0.5, 1, 2, 4
prob_zero_removal	0
prob_remove_samples	0.2
num_replicates	10

Supplemental Table S3: *Related to STAR Methods.* Low and high expression ranges from which the master regulators' production rates were sampled

DS-ID	Low Expression Range	High Expression Range
DS1	[0.2 0.5]	[0.7 1]
DS2-8	[0 2]	[2 4]
DS9-15	[0 1]	[3 4]

Supplemental Table S4: *Related to STAR Methods.* A comparison between running times of SERGIO and BoolODE

Experiment	Network ID	Number Genes	Number Cell Type/Simulation Time	SERGIO time (s)	BoolODE time (s)
1	4	1200	9	4607	8179
2	4	1200	1	733	1058
3	3	400	1	132	1145
4	2	100	1	28	75