Supplemental Material for

Dissecting and improving gene regulatory network inference using single-cell transcriptome data

Lingfeng Xue[1], Yan Wu[1,2], Yihan Lin[1,2,3,]*

Table of Contents

Supplemental Notes 1

Supplemental Note S1. On concerns regarding the biological relevance of our model simulations and parameter choice 2

Supplemental Note S2. On the comparison between AEP and AUPR metrics........................ 4

Supplemental Note S3. Analysis using transcription factor induction datasets 4

Supplemental Figures 6

Supplemental Figure S1. Using model simulations to compare pre-mRNA and mRNA dynamics.. 6

Supplemental Figure S2. Schematics illustrating the pipeline of simulation using the dyngen package.. 7

Supplemental Figure S3. Network backbones and types of dynamics in the simulation 8

Supplemental Figure S4. Factor-dependency analysis in simulated datasets........................ 9

Supplemental Figure S5. Network motif analysis for the simulated datasets of the four different backbones.. 10

Supplemental Figure S6. Additional comparisons between pre-mRNA-based and mRNA-based methods using experimental single-cell datasets........................ 11

Supplemental Figure S7. Extended characterizations for factor-dependency analysis of GRNs inferred from experimental datasets 12

Supplemental Tables 13

Supplemental Table S1. Detailed information on datasets used in the study 13

Supplemental Table S2. Analysis of top five hub TFs from six different datasets................ 14

Supplemental Note S1. On concerns regarding the biological relevance of our model simulations and parameter choice.

We used simulations of kinetic models to compare the two methods in Fig. 2, i.e., pre-mRNA-based and mRNA-based. Using model simulation, we compared them under a range of parameter sets. Concerns may arise as the simulations of the model may not accurately capture the behaviors of real networks, making it challenging to validate the enhanced performance of our proposed pre-mRNA-based method compared to the mRNA-based method. We would like to address such concerns with the following clarifications.

We would like to clarify that the kinetic rates used in our study were not arbitrary. These rates were based on experimental measurements from a previous study (Rabani et al. 2014), which allowed us to select default kinetic parameters according to the median values measured in that study. We further collected experimentally measured RNA half-lives and splicing rates from several literature studies (Table SN1 and Table SN2) to justify that the ranges of parameters used in Fig. 2 are biologically reasonable. Importantly, from these experimentally determined parameters, it is apparent that mRNA half-life is typically on the order of hours and the pre-mRNA splicing time is typically on the order of minutes, and such ranges of parameters were covered in our analysis in Fig. 2. Although we cannot provide a gene-specific estimation of parameters, we believe that it is appropriate and acceptable practice to computationally enumerate possible combinations of parameters (as in Fig. 2) in order to provide a quantitative understanding of the model behavior. More generally, in our simulations, we intentionally varied these parameters to explore the trends and principles underlying the determination of GRN inference accuracy. By doing so, we aimed to provide useful rules and principles that indicate when the inference would be accurate or inaccurate. It is important to note that these simulations are not intended to model the networks in real experimental datasets but rather serve as a means for understanding general rules and principles.

Cell type	Median mRNA half-life	Reference DOI
HeLa TO cells	3.4 h	10.1101/gr.130559.111
Human K562 cells	50 min	10.1126/science.aad9841
Mouse ESC	3.9 h	10.1038/nmeth.4435

Human K562 cells	8.5 h	10.1021/jacs.8b08554
Mouse ESC	7.1 h	10.1093/dnares/dsn030
Human B cell	315 min	10.1093/nar/gkp542
Mouse fibroblasts	274 min	10.1093/nar/gkp542
Mouse DC	86.1 min	10.1016/j.cell.2014.11.015

Table SN1. Experimentally measured mRNA half-lives from 8 separate literatures.

Cell type	Typical splicing time	Reference DOI
Drosophila	2 min	10.7554/eLife.32537
Human neuroblastoma cell	5-10 min	10.1038/nsmb.1666
Human U2OS cell	0.4-7 min	10.1083/jcb.201009012
Human HEK293	2.5-3 min	10.1016/j.celrep.2013.08.013
Mouse DC	14 min	10.1016/j.cell.2014.11.015

Table SN2. Experimentally measured pre-mRNA splicing time from 5 separate literatures.

We acknowledged the concern that kinetic rates are highly dependent on gene and cell-type or tissue context. Our intention in varying the parameters was to explore a wide range of possible kinetic scenarios and not to restrict ourselves to a specific biological context. This approach allowed us to gain insights into the general principles governing GRN inference accuracy, which can be applicable across different biological settings. It should also be noted that for a range of parameter combinations, we showed that pre-mRNA is not necessarily better than mRNA in terms of capturing upstream regulatory activity dynamics, thus illustrating potential biological settings where pre-mRNA would perform worse than mRNA for GRN inference.

We next explained why we cannot simply use rates inferred from scVelo in the analysis of our model. We have carefully examined the validity of the parameters inferred from a typical scRNA-seq dataset using the scVelo tool, such as the splicing rate parameter and the mRNA degradation rate parameter, and found that the software cannot provide an accurate estimation of such rate parameters. More specifically, as illustrated in the tutorial from the scVelo website (https://scvelo.readthedocs.io/en/stable/DynamicalModeling/), the inferred pre-mRNA splicing rates are on the same order of magnitude as the mRNA degradation rates. Notably, such a result is in sharp contrast with the experimentally determined results (Table SN1 and Table SN2), as the measured splicing rate is much faster than the mRNA degradation rate.

Supplemental Note S2. On the comparison between AEP and AUPR metrics

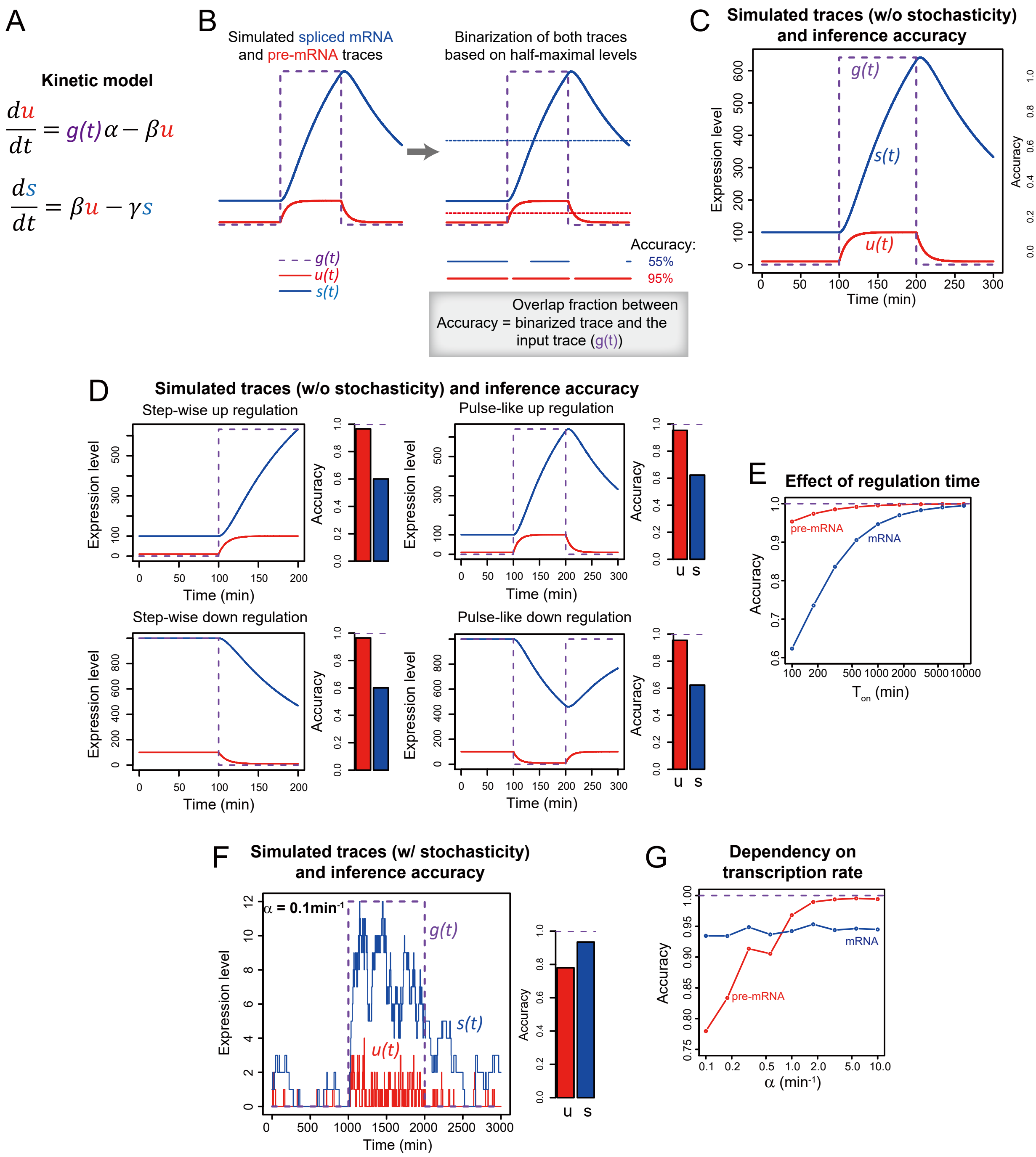
We have analyzed the performance of our inferred networks using both AEP (Average Early Precision) and AUPR as two separate metrics. Although we used AUPR to evaluate networks inferred from experimental single-cell datasets, we believed that AEP is a more appropriate and informative metric, which focuses on the most confident links in the inferred network. More specifically, we observed that AUPR is generally less sensitive than AEP, particularly when evaluating the results using the Motif database as the ground truth (Fig. 4E-F). For a few datasets evaluated using AUPR, the mRNA-based method performed better than the pre-mRNA-based method, and it was unclear what factors might have contributed to this observation, which necessitates further investigations. Nevertheless, we believe that including AUPR as an additional performance metric can provide valuable insights into the global performance of our inferred networks.

Supplemental Note S3. Analysis using transcription factor induction datasets.

In order to further demonstrate the enhanced performance of the pre-mRNA-based method over the mRNA-based method, we chose to compare the inferred GRNs from datasets where exogenously induced TFs were used to program cell fates. The rationale is that we should be able to recover the TFs being induced in the inferred network without needing to rely on GRN databases as the ground truth. More specifically, we utilized scRNA-seq data generated by Hersbach et al., 2022, particularly the Ascl1-Hnf1a-Myod1-Oct4_induction_24h_biol_rep1 dataset (GSM6504514) and the Ascl1-Hnf1a-Myod1-Oct4_induction_48h_biol_rep1 dataset (GSM6504515). In these experiments, the authors induced cell fate transition in mouse embryonic fibroblast cells using several TFs.

We first compared the inferred GRNs of the first dataset, i.e., at 24 h post-induction using the pre-mRNA-based or the mRNA-based method (Supplemental Fig. S6C). For the GRN inferred using the pre-mRNA-based method, two of the largest hub TFs (i.e., Ascl1 and Myod1) belonged to the TFs being exogenously expressed, which are as expected. For the GRN inferred using the mRNA-based method, one of the hub nodes was the ribosomal protein Rps4x, which was comparable in size to the TF being induced (i.e., Ascl1). Thus, it appears that the pre-mRNA-based method better captured the immediate regulatory effects caused by the exogenous TFs at the 24 h time point. We next compared the inferred GRNs at 48 h post-induction (Supplemental Fig. S6D). The top hub nodes in the GRN inferred using the pre-mRNA-based method became ribosomal proteins while the top hub nodes for the mRNA-based method were still the TFs being induced (i.e., Myod1 and Ascl1).

Based on the results from the two post-induction time points, we reasoned that the pre-mRNA-based method captured the transient up-regulation of the target genes of Ascl1 and Myod1, which led to much more enhanced pre-mRNA levels of target genes at 24 h compared to at 48 h. In contrast, because mRNA levels of target genes were used for inference for the mRNA-based method, relatively persistent activities of Ascl1 and Myod1 were inferred, instead of transient bursts of activities. Because it is known that a step increase in TF expression typically leads to an adaptive response in downstream gene expression (i.e., a transient pulse of target activation), we reasoned that the pre-mRNA-based method appeared to capture such an adaptive response more accurately compared to the mRNA-based method.



Supplemental Figure S1. Using model simulations to compare pre-mRNA and mRNA dynamics.

(A) Kinetic model used for simulating transcriptional regulation. In the first equation, unspliced mRNA level (pre-mRNA level) is increased by transcription at the rate of g(t)α and is then reduced by splicing at the rate of βu. In the second equation, spliced mRNA level is increased by splicing at the rate of βu and is reduced by degradation at the rate of γs.

(B) Schematics illustrating the calculation of inference accuracy. Simulated pre-mRNA (red) or mRNA (blue) trace was used to calculate inference accuracy, defined by the overlap between the binarized expression trace and the regulatory activity trace (i.e., *g*(t)).

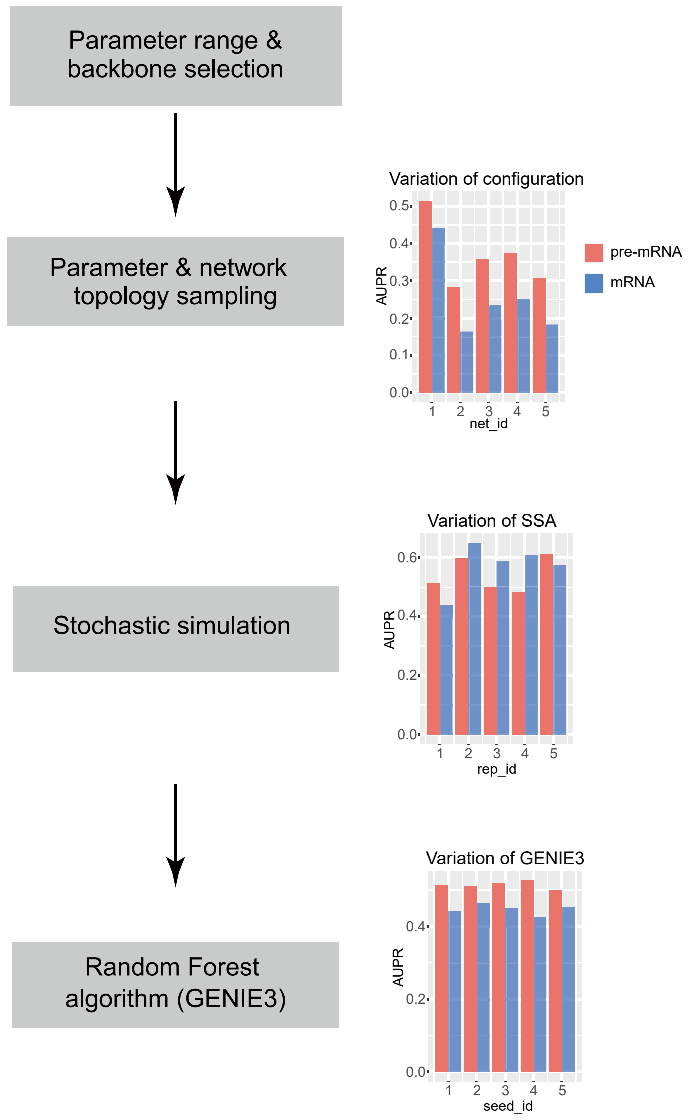
(C) Simulated pre-mRNA and mRNA traces for a gene with high expression level and fast regulation (i.e., short T_{on} as in **Fig. 2A**). The bar graph shows the calculated inference accuracies of the two traces.

(D) Simulated pre-mRNA and mRNA traces for a gene undergoing four different types of transcriptional regulation and the corresponding inference accuracies.

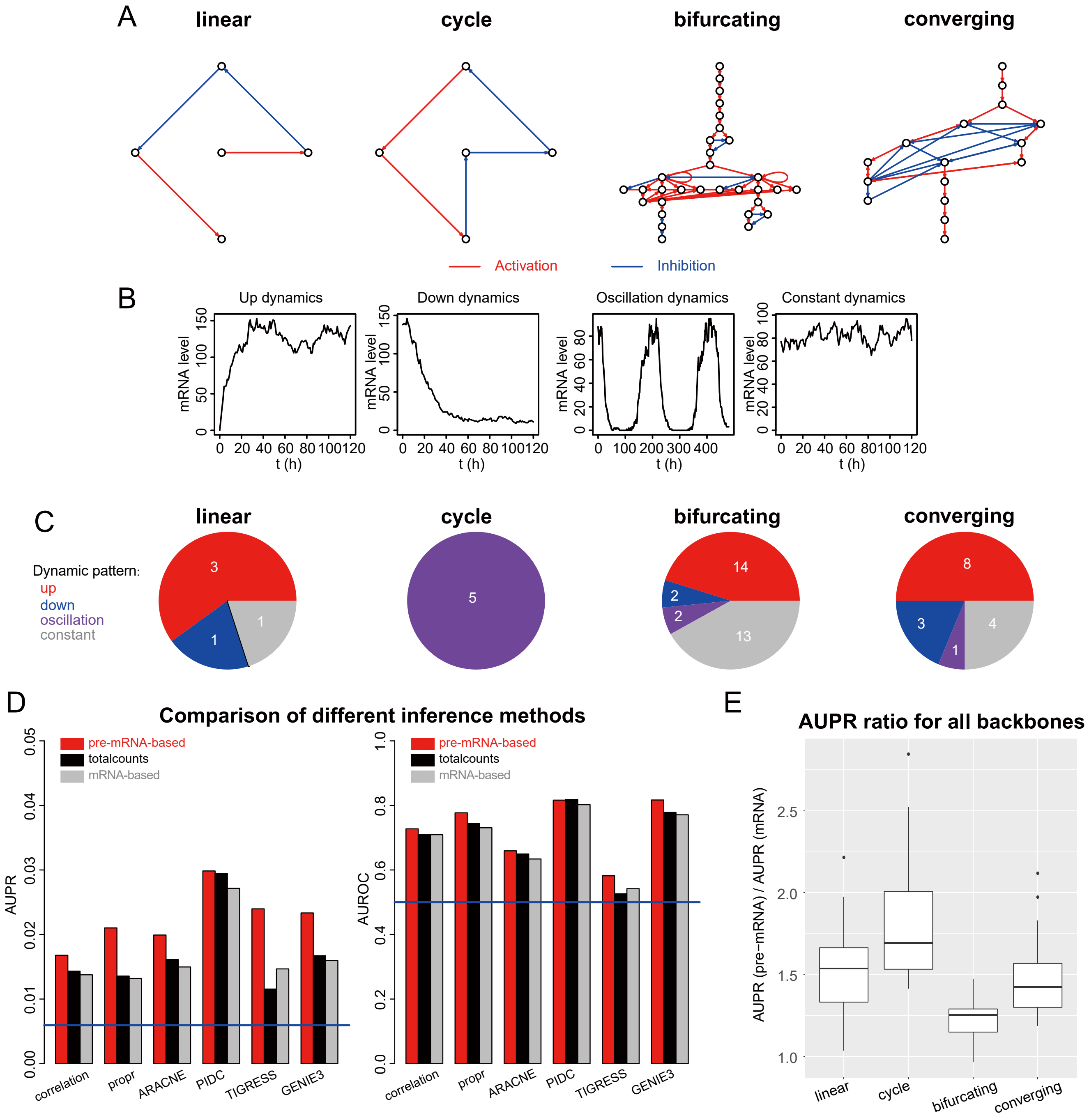
(E) The effect of regulation time (T_{on}) on the accuracies of pre-mRNA-based and mRNA-based methods.

(F) Stochastically simulated pre-mRNA and mRNA traces for a gene with low expression level and slow regulation (i.e., long T_{on}). Note that with this parameter set, the inference accuracy of pre-mRNA is worse than mRNA.

(G) The effect of transcription rate on the accuracies of pre-mRNA-based and mRNA-based methods in the presence of stochasticity.



Supplemental Figure S2. Schematics illustrating the pipeline of simulation using the dyngen package. Simulation started with the choice of network backbones and kinetic parameters, and the dyngen package was used for performing stochastic dynamic simulations. The output trajectories were converted into count matrices (see **Methods**), which were then used for network inference using GENIE3. Panels on the right contain example outputs showing how AUPR can be different across simulation replications.



Supplemental Figure S3. Network backbones and types of dynamics in the simulation.

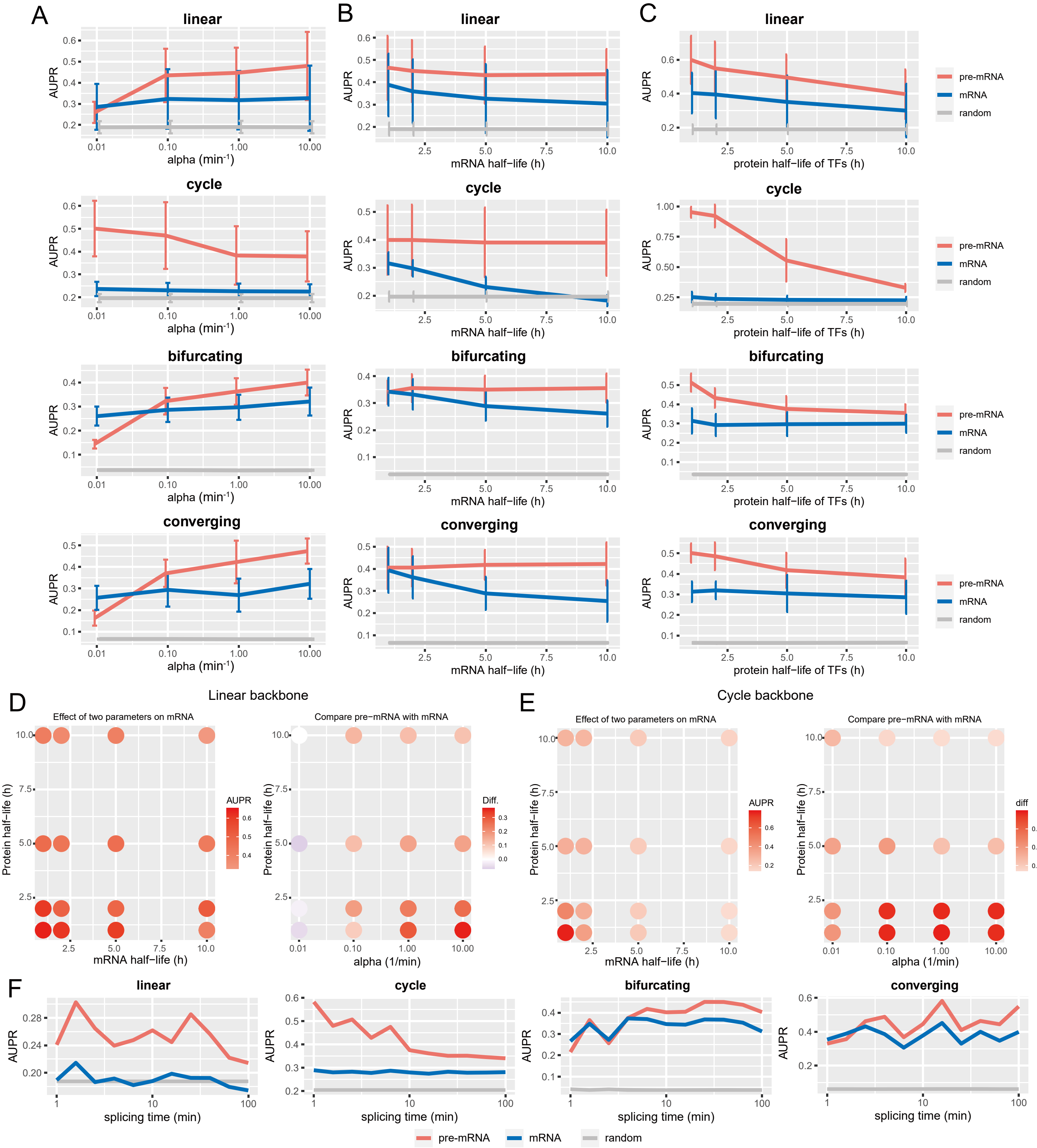
(A) Four types of network backbones used in simulations. Example networks for each type are shown. Red edges represent activation, while blue edges represent inhibition.

(B) Examples of four typical types of dynamics observed in simulations.

(C) Pie chart illustrating the numbers of TFs exhibiting different types of activity dynamics in each backbone. Note that the classification was based on dynamics shown in **B**.

(D) Comparative analysis of GRN inference algorithm performance on a synthetic dataset (with bifurcating backbone, 200 genes, 750 cells) generated by dyngen, demonstrating the impact of using different input matrices, including pre-mRNA (target)/mRNA (regulator), total counts (mRNA plus pre-mRNA) for both target and regulator, and mRNA for both target and regulator. Six algorithms were tested: (1) correlation, (2) propr, (3) ARACNE, (4) PIDC, (5) TIGRESS, and (6) GENIE3.

(E) Boxplots showing the comparison between pre-mRNA-based method and mRNA-based method in four backbones. The ratio of AUPR between pre-mRNA and mRNA was calculated for each backbone (related to **Fig. 3B**).



Supplemental Figure S4. Factor-dependency analysis in simulated datasets.

(A) The effect of transcription rate on network inference accuracy in four backbones. Four different transcription rates were used for separate simulations. Error bars indicate S.D., n = 10.

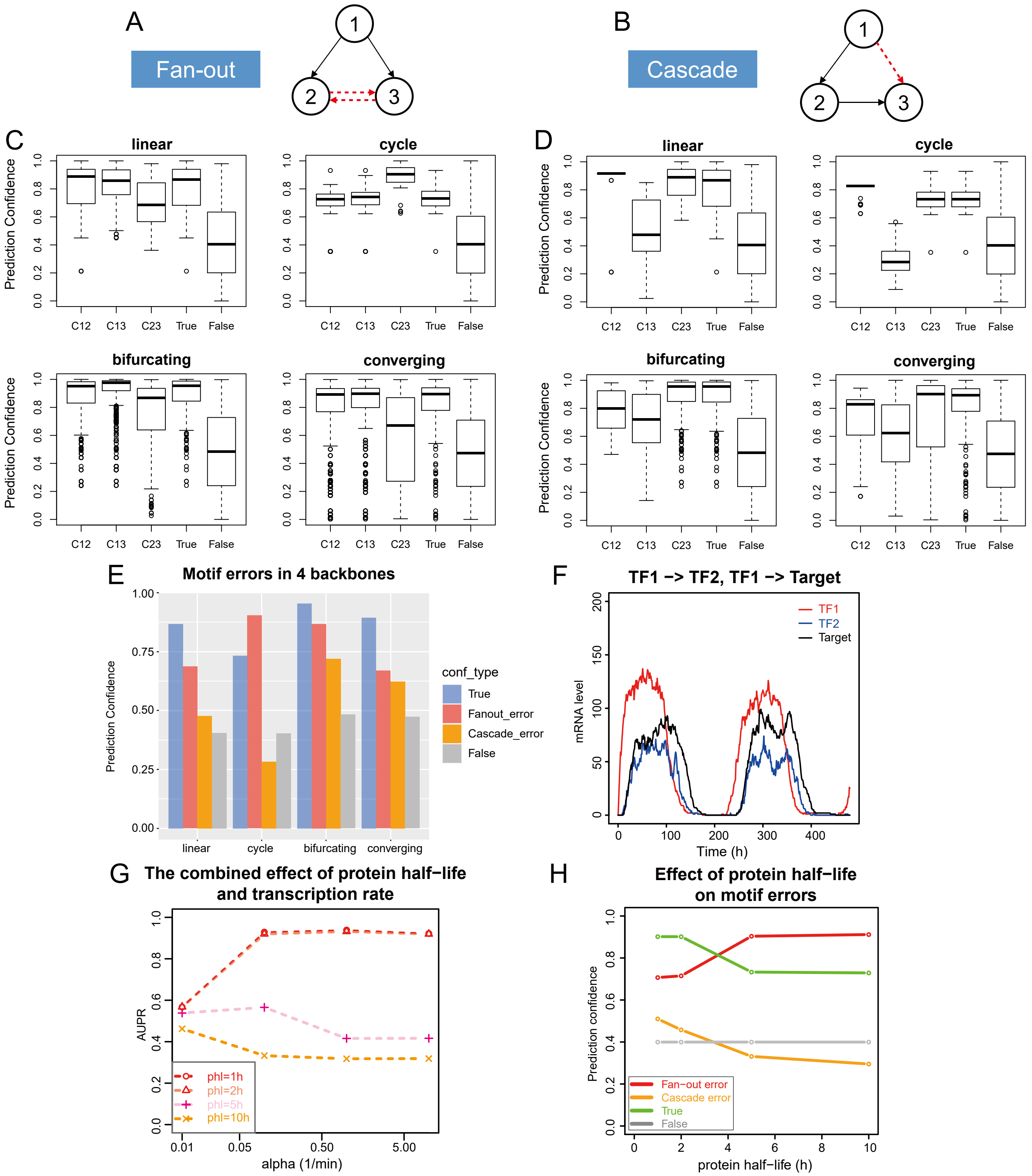
(B) The effect of mRNA half-life on network inference accuracy in four network backbones. Four different mRNA half-lives were used for separate simulations. Error bars indicate S.D., n = 10.

(C) The effect of protein half-life on network inference accuracy in four network backbones. Four different protein half-lives were used for separate simulations. Error bars indicate S.D., n = 10.

(D) The combined effects of two parameters on mRNA-based method (left) and on the relative performance between the two methods (right) for the linear backbone. For the right panel, difference in AUPR was calculated by AUPR (pre-mRNA) minus AUPR (mRNA).

(E) Analogous as (**D**) for the cycle backbone.

(F) The effect of splicing time on network inference accuracy in four backbones. Splicing time was varied from 1 to 100 min. Note that the default splicing time in our simulations is 10 min.



Supplemental Figure S5. Network motif analysis for the simulated datasets of the four different backbones.

(A) Schematic illustrating fan-out error in network inference. When gene 1 regulates both genes 2 and 3, it often occurs that the inference algorithm erroneously considered gene 3 being regulated by gene 2 (and/or vice versa).

(B) Schematic illustrating cascade error in network inference. When gene 1 regulates gene 2, and gene 2 regulates gene 3, it often occurs that the inference algorithm erroneously considered gene 1 regulating gene 3.

(C) Boxplots showing fan-out errors in four backbones. Prediction confidence was estimated for each type of links (1->2, 1->3 and 2<->3), and was compared with the background levels (for all links in the synthetic network, i.e., TRUE, and for links that are not in the synthetic network, i.e., FALSE).

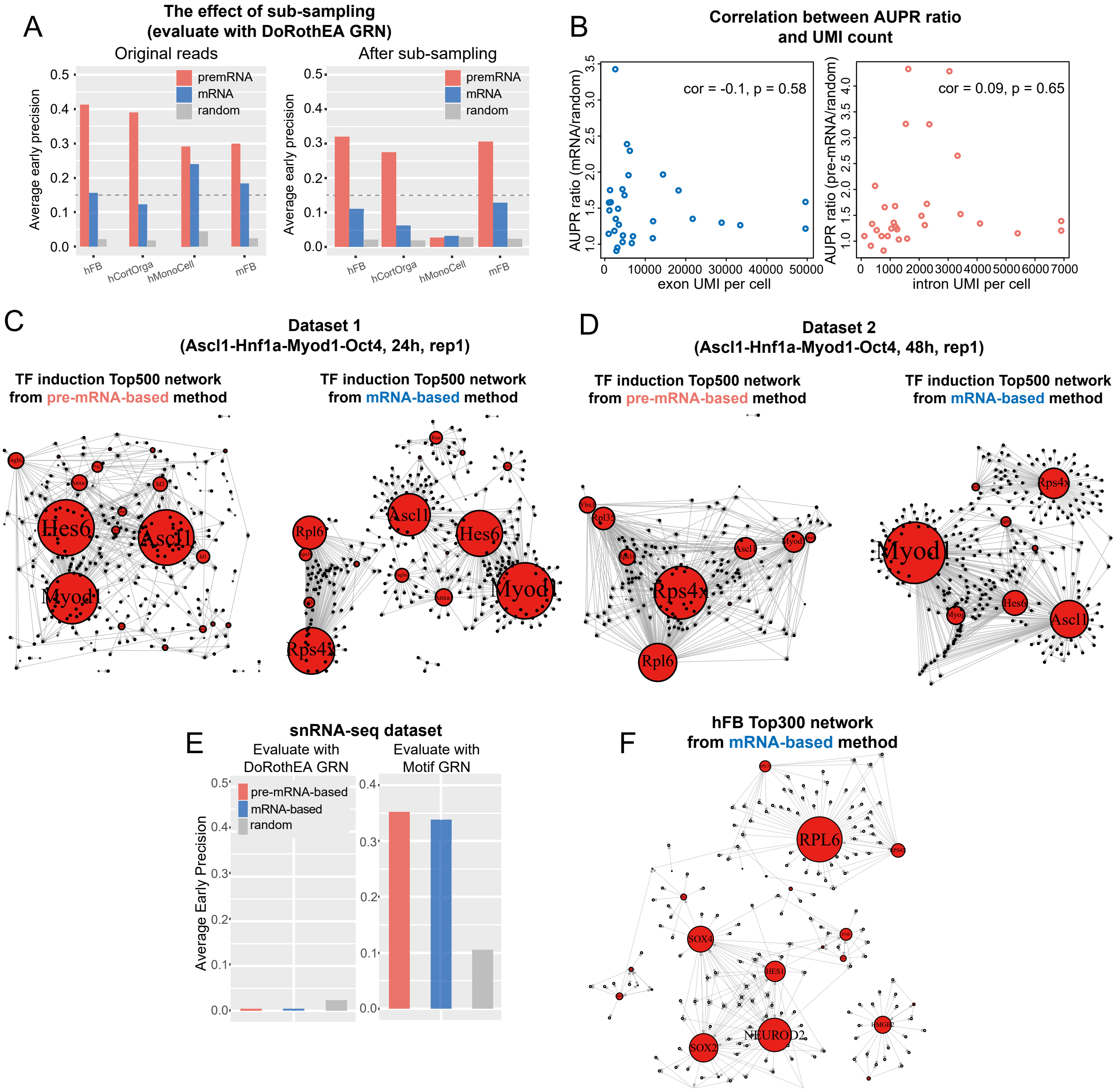
(D) Analogous to (**C**) for cascade errors in the four backbones.

(E) Summary of fan-out errors and cascade errors in the four backbones.

(F) Representative simulated traces showing fan-out error in the cycle backbone.

(G) The combined effects of protein half-life and transcription rate on network inference for the cycle backbone.

(H) The effect of protein half-life on motif errors for the cycle backbone.



Supplemental Figure S6. Additional comparisons between pre-mRNA-based and mRNA-based methods using experimental single-cell datasets.

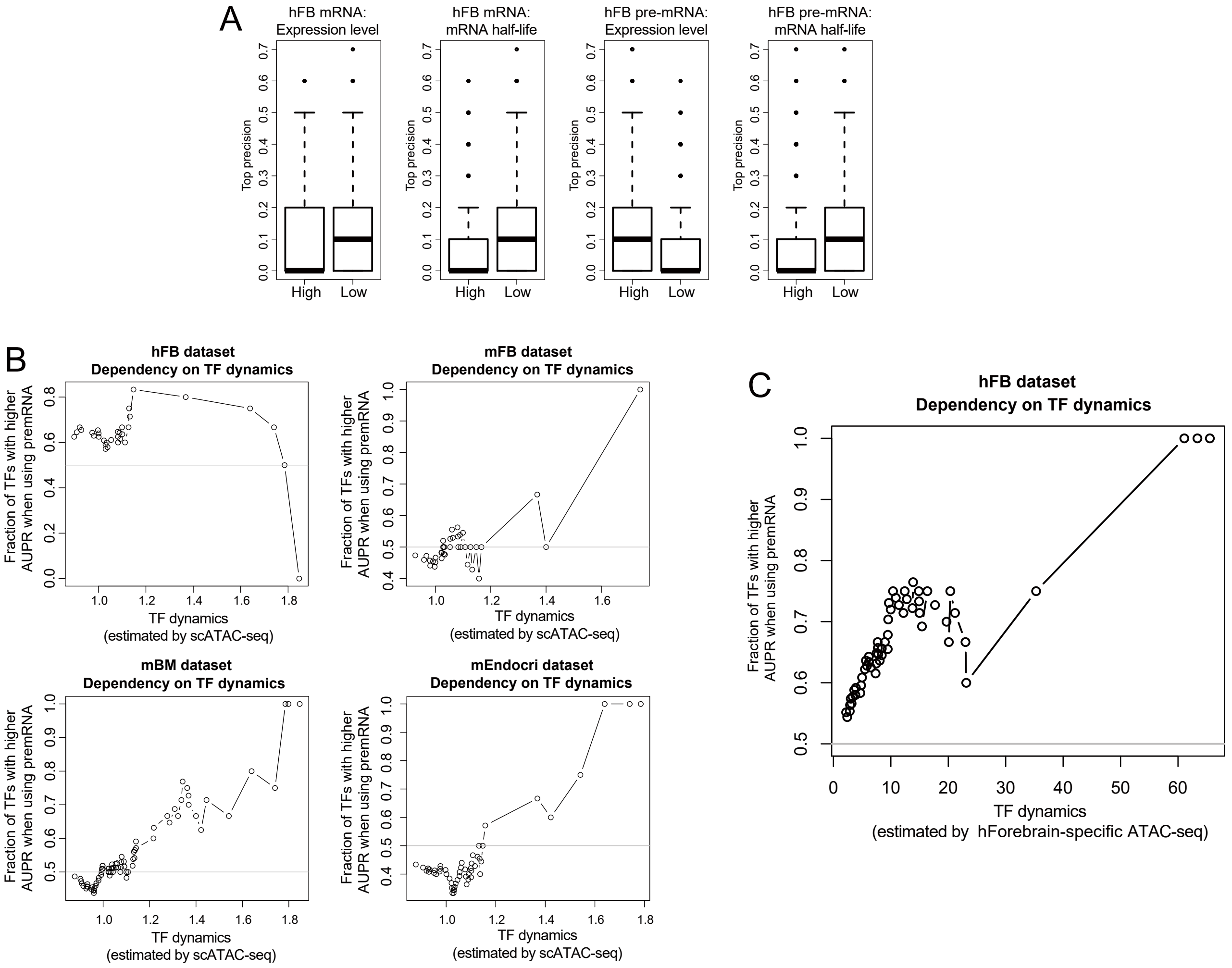
(A) The effect of subsampling on inference accuracy for four different datasets. Reads from each dataset were subsampled to ~30% of the original read numbers and both pre-mRNA-based and mRNA-based methods were implemented for GRN inference. Average early precisions were shown for GRNs inferred from both methods before (left) and after (right) subsampling. Data on the left was from **Fig. 4A**.

(B) AUPR ratio for mRNA-based or pre-mRNA-based method versus UMI counts per cell (exon or intron) for individual datasets. AUPR ratio was calculated by diving the AUPR of the network inferred from either method by that of the network from a random predictor Inferred GRN was evaluated with the DoRothEA database.

(C-D) Inferred GRNs from single mouse embryonic fibroblast cells induced by reprograming transcription factors at 24h **(C)** or 48 h **(D)** post-induction using either the pre-mRNA-based method or the mRNA-based method. Raw sequencing data were downloaded from GSM6504514 **(C)** and GSM6504515 **(D)**. In these networks, the edges represent the inferred transcriptional regulation from one TF (transcription factor) to one target gene. The size of the node represents the number of inferred target genes for the TF. 500 interactions (edges) of the highest confidence were shown (i.e., Top500 network).

(E) Inference accuracy using a snRNA-seq dataset. A snRNA-seq dataset for mouse skeletal myofibers (SRX7939765) was used. Cell number was 8064 , UMI count per cell: intron 3622, exon 1247. AEP evaluated using either DoRothEA database or Motif GRN.

(F) Inferred GRN for the human forebrain dataset using mRNA-based method. Top300 network was shown. See also **Fig. 5B**.



Supplemental Figure S7. Extended characterizations for factor-dependency analysis of GRNs inferred from experimental datasets.

(A) The effect of the expression level or the mRNA half-life on inference accuracy using the mRNA-based or the pre-mRNA-based method. For each panel, target genes were first divided into two groups according to the expression level (of pre-mRNA or mRNA) or the mRNA half-life, and the top-10-precision of each target (i.e., mean inference precision of the top 10 inferred TFs of each target) within each bin was calculated and shown in the boxplots. p-values from Wilcoxon tests were plotted in **Fig. 5B**.

(B) The dependency of the network inference accuracy on the TF dynamics for individual datasets. The dynamics of TFs were approximated using cell-to-cell variabilities of TF activities measured by public single-cell ATAC-seq data. The mean variability from scATAC-seq data of multiple cell lines was used (**Methods**). TFs from each data-set were sorted according to TF dynamics, and the fraction of TFs more accurately inferred by pre-mRNA-based method than mRNA-based method was calculated for TFs above the indicated value of TF dynamics on x-axis (i.e., we focus on TFs with high dynamics).

(C) Analogous to **(B)** except that the TF dynamics were from ATAC-seq data measured in human forebrain.

Supplemental Table S1. Detailed information on datasets used in the study.

Name in the study	Species	Cell type	SRA accession	GEO accession	url	Data source
hFB	human	human week 10 fetal forebrain dataset	SRR6470906,SRR6470907			Velocyto
hESC_1	human	hESC	SRR6328624			SRR
hESC_2	human	hESC	SRR9117953			SRR
PBMC	human	PBMC			https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k	10x Genomics
A549	human	A549	SRR8144883, SRR8144884, SRR8144885, SRR8144886, SRR8144887			SRR
hBM	human	Bone marrow	SRR6192408			SRR
hEK	human	Embryonic kidney cortex kidney 1	SRR6921770			SRR
hLungPro	human	Lung progenitors	SRR6042036			SRR
hPancIslet	human	Pancreatic islets	SRR7142646			SRR
hPancPro	human	Pancreatic progenitor cells	SRR7905628			SRR
hBCell	human	B cell	SRR7340856			SRR
hCortOrga	human	Cortical organoids	SRR6996081			SRR
hLiverPro	human	Hepatocyte-derived liver progenitor-like cells	SRR7721684,SRR7721685,SRR7721686,SRR7721687			SRR
hLiverHomo	human	Patient 1 Total Liver Homogenate	SRR7276474			SRR
hNKCell	human	CD56Neg NK cells	SRR7293994			SRR
hCD4TCell	human	Precursors of human CD4+ cytotoxic T lymphocytes	SRR6260183,SRR6260184			SRR
hMonoCell	human	Peripheral blood mononuclear cell	SRR6260181,SRR6260182			SRR
hPlacenta	human	Placenta	SRR7895963			SRR
hSpleen	human	spleen	SRR8073185			SRR
hTCell	human	T cells	SRR7797510			SRR
mEndocri	mouse	endocrinogenesis_day15		GSE132188		scVelo
mEpi	mouse	intestinal epithelium		GSE92332		Velocyto
mBM	mouse	Bone Marrow		GSE109989		Velocyto
mDenGy	mouse	hippocampal dentate gyrus neurogenesis		GSE95753		scVelo
mFB	mouse	Forebrain	SRR11966461			SRR
mESC_1	mouse	mESC 2I	SRR12318312			SRR
mESC_2	mouse	mESC Serum	SRR12318318			SRR
mESC_3	mouse	mESC 2I	SRR11394540			SRR
mE3.5	mouse	E3.5	SRR82562xx			SRR
mEK	mouse	E15.5 embryonic kidney	SRR7689139			SRR

Supplemental Table S2. Analysis of top five hub TFs from six different datasets.

Network	Method	TF1	TF2	TF3	TF4	TF5	Note
hFB	mRNA-based	RPL6	NEUROD2	SOX2	SOX4	HES1	In the mRNA-based network, the top 1 hub TF is RPL6, a component of the large ribosomal subunit, which appears unrelated to neuronal development. In contrast, the pre-mRNA-based network identifies MEF2C as a top 5 hub TF, which is crucial for normal neuronal development.
	pre-mRNA-based	NEUROD2	SOX2	HES1	SOX4	MEF2C	
mBM	mRNA-based	Ltf	Anxa1	Mxd1	Rps4x	Hmgb2	The mRNA-based network identifies Rps4x and Hmgb2 as hub TFs. Rps4x is a ribosomal protein and unlikely to be a TF, while Hmgb2 acts as a cytoplasmic promiscuous immunogenic DNA/RNA sensor. The pre-mRNA-based network identifies Cebpb and Pou2af1 as hub TFs. Cebpb is an essential transcription factor regulating the expression of genes involved in immune and inflammatory responses, and Pou2af1 regulates transcription in various tissues, including activating immunoglobulin gene expression.
	pre-mRNA-based	Ltf	Cebpb	Anxa1	Mxd1	Pou2af1	
mEpi	mRNA-based	Tff3	Rps4x	Ckmt1	Gm2000	Rps10	The mRNA-based network identifies Rps4x and Rps10 as hub TFs, both of which are ribosomal proteins. The pre-mRNA-based network identifies Creb3l4 and Hmgn3 as hub TFs. Creb3l4 is a transcriptional activator potentially involved in the unfolded protein response, while Hmgn3 binds to nucleosomes, regulating chromatin structure and associated processes such as transcription, DNA replication, and DNA repair.
	pre-mRNA-based	Ckmt1	Tff3	Gm2000	Creb3l4	Hmgn3	
mDenGy	mRNA-based	Ybx1	Gm10269	Sox4	Zbtb20	Nfib	The mRNA-based network identifies Nfib as a hub TF, a transcriptional activator of GFAP essential for proper brain development. The pre-mRNA-based network identifies Rps4x as a hub TF, a ribosomal protein seemingly unrelated to neuronal development.
	pre-mRNA-based	Ybx1	Gm10269	Sox4	Zbtb20	Rps4x	
mEndocri	mRNA-based	Rps4x	Neurog3	Sox4	Hspa5	Gadd45a	The hub TFs inferred by both intron and mRNA-based methods are the same.
	pre-mRNA-based	Neurog3	Sox4	Hspa5	Rps4x	Gadd45a	
mND	mRNA-based	Hmgb2	Rps4x	Kif22	Ybx1	Sox11	The mRNA-based network identifies Rps4x, Kif22, and Ybx1 as hub TFs. Rps4x is a ribosomal protein, Kif22 is a kinesin family member involved in spindle formation and chromosome movements

	pre-mRNA-based	Sox11	Hmgb2	Celf4	Tbr1	Neurod2	during mitosis and meiosis, and Ybx1 is a DNA- and RNA-binding protein involved in various processes. The pre-mRNA-based network identifies Celf4, Tbr1, and Neurod2 as hub TFs. Celf4 mediates exon inclusion/exclusion in pre-mRNA that are subjected to tissue-specific and developmentally regulated alternative splicing, Tbr1 is a transcriptional repressor involved in multiple aspects of cortical development, and Neurod2 is a transcriptional regulator implicated in neuronal determination.