dyngen: a multi-modal simulator for spearheading new single-cell omics analyses

Robrecht Cannoodt* Wouter Saelens* Louise Deconinck Yvan Saeys

01 June 2020

Abstract

Purpose: A recent breakthrough in single-cell omics is the ability to perform multi-modal measurements at single-cell level. However, in many cases insufficient metadata is available to accurately benchmark new types of algorithms. Existing generators of scRNA-seq are valuable in benchmarking novel computational tools more rigorously, but adding additional modalities or experimental conditions usually requires significant methodological alterations.
Results: We introduce dyngen, a simulator of synthetic single cells by simulating gene regulation, splicing and translation at a single molecule level. From a simulation many layers of information can be extracted, including molecule abundance, progression along a dynamic process, activation strength of individual regulatory interactions, and a ground-truth RNA velocity. We demonstrate dyngen by showcasing its use on three novel computational approaches where no benchmarking of tools have been described for up to date: RNA velocity, cell-specific gene regulatory network inference and trajectory alignment methods. In itself, each of these three application cases constitutes a novel result that has not yet been described in the literature so far, and which we believe will spearhead research on development of novel computational tools for single-cell data.
Conclusion: Ultimately, dyngen allows anticipating technological developments in single-cell multi-omics. It is thus possible to design and evaluate the performance and robustness of new types of computational analyses before experimental data becomes available. In addition, dyngen could also be used to compare which experimental protocol is the most cost-effective in producing qualitative and robust results in downstream analysis.

Introduction

Continuous technological advancements to single-cells omics are having profound effects on how researchers can validate biological hypotheses. Early experimental technologies typically only allowed profiling a single modality (e.g. DNA sequence, RNA or protein expression). However, recent developments permit profiling multiple modalities simultaneously, and every modality added allows for new types of analyses that can be performed.

This confronts method developers with several challenges. The majority of the >250 peer-reviewed computational tools for analysing single-cell omics data were published without a quantitative assessment of the accuracy of the tool. This is partially due to low availability of suitable benchmarking datasets; even if there are sufficient suitable input datasets available, these are often not accompanied by the necessary metadata to serve as ground-truth for a benchmark.

Here, synthetic data plays a crucial role in asserting minimum performance requirements for novel tools in anticipation of adequate real data. Generators of scRNA-seq data (e.g. splatter [1], powsimR [2], PROSSTT [3] and SymSim [4]) have already been used extensively to explore the strengths and weaknesses of computational tools, both by method developers [5, 6, 7, 8] and independent benchmarkers [9, 10, 11]. However, a limitation of existing scRNA-seq profiles generators is that they would require significant methodological alterations to add additional modalities or experimental conditions.

An ideal experiment would be able to observe all aspects of a cell, including a full history of its molecular states, spatial positions and environmental interactions [12]. While this falls outside the reach of current experimental technologies, generating synthetic data in anticipation of new experimental technologies would allow developing the next wave of computational tools.

We developed a multi-modality simulator of single cells called dyngen (Figure 1). dyngen uses an optimised version of Gillespie's stochastic simulation algorithm [13] to simulate gene regulation, splicing, and translation at a single-molecule level. Its methodology allows tracking of many layers of information throughout the simulation, including the abundance of any molecule in the cell, the progression of the cell along a dynamic process, and the activation strength of individual regulatory interactions. dyngen can simulate a large variety of dynamic processes (e.g. cyclic, branching, disconnected) as well as a broad range of experimental conditions (e.g. batch effects and time-series, perturbation and knockdown experiments). The fine-grained controls over simulation parameters allow dyngen to be applicable to a broad range of use-cases. We demonstrate this by performing the first quantitative evaluations of three types of novel computational approaches: RNA velocity, cell-specific network inference, and trajectory alignment methods.

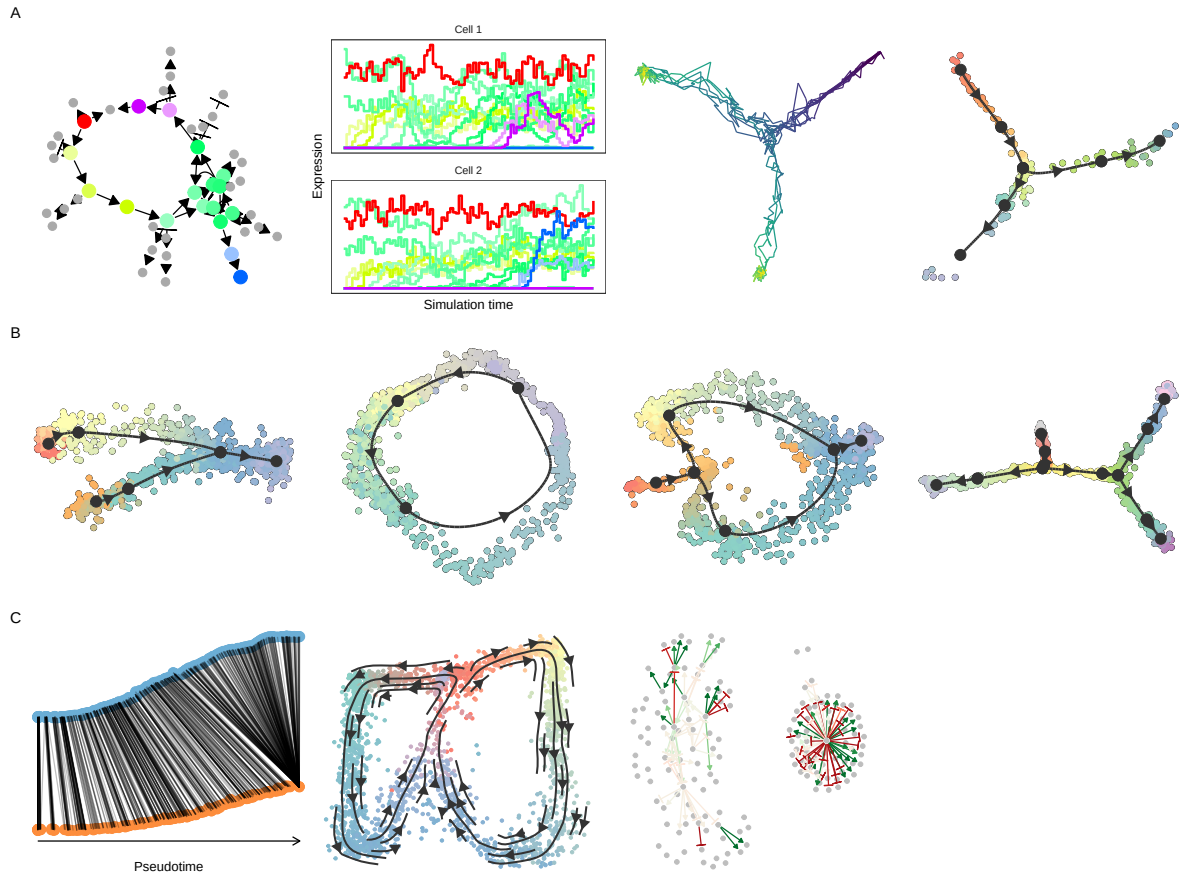

Figure 1: **Showcase of dyngen functionality.** **A:** The typical process of generating a bifurcating dataset with dyngen. Step 1: generating a gene regulatory network. Step 2: simulating changes in expression over time. Step 3: combining multiple simulations and mapping to the ground-truth trajectory. Step 4: sampling cells and molecules. **B:** dyngen includes different pre-defined trajectories, including converging, cyclic, bifurcating converging, and a binary tree. **C:** From the different types of ground-truth information, dyngen allows – amongst others – suppleming quantitative benchmarks of trajectory alignment, RNA velocity, and cellwise network inference methods.

Results

A cell consists of a set of molecules, the abundance of which is affected by a set of reactions: transcription, splicing, translation, and degradation (Figure 2A). A gene regulatory network (GRN) defines the reactions that are allowed to occur (Figure 2B), which is constructed in such a way that cells slowly develop over time (Figure 2C,D). With every time step dt in the simulation, the probability of a reaction occurring is computed (not shown). From the probabilities are sampled which reactions occur during this time step dt (Figure 2E).

dyngen returns many modalities throughout the whole simulation: molecular abundance, number of reaction firings, reaction likelihoods, and regulation activations (Figure 2C–F). These modalities can serve both as input data and ground truth for benchmarking many types of computational approaches. For example, a network inference method could use mRNA abundance and regulation activities as inputs and its output could be benchmarked against the gold standard GRN.

Depending on how the GRN is designed, different cellular developmental processes can be simulated. dyngen includes generators of GRNs which result in many different developmental topologies (Figure 3), including branching, converging, cyclic and even disconnected. Custom-defined GRNs offer more fine-grained control over the simulation.

Together, these qualities allow it to be applicable in benchmarking a broad range of use-cases. In practice, dyngen has already successfully been used to evaluate trajectory inference [10], trajectory-based differential expression [14], and network inference [15] methods. To demonstrate this point even further, we apply dyngen on several promising novel computational approaches for which quantitative assessment of the performance was until now lacking.

RNA velocity

In eukaryotes, a gene is first transcribed to a pre-mRNA and subsequently spliced into mature mRNA. Because reads coming from both unspliced and spliced transcripts are observed in expression data, the relative ratio between the two can tell us something about which genes are increasing, decreasing or remaining the same [16, 17]. To determine this, some parameters have to be estimated to determine which fraction of unspliced and spliced mRNAs correspond to an increase or decrease. The estimation of these parameters makes some assumptions and can be handled in different ways in the two main algorithms that are now available for RNA velocity estimation: velocyto [17] and scvelo [18]. It can be difficult to obtain ground truth data to benchmark these algorithms, given that it would require continuous data of transcriptional dynamics in individual cells. On the other hand, the ground truth velocity is rapidly extracted from the dyngen model, by looking at whether each transcript is currently increasing or decreasing in expression.

We tested scvelo and velocyto on 8 datasets containing linear, bifurcating, disconnected and cyclic trajectories, and varied the main parameter settings in which they estimate the velocity. We found that the original velocyto implementation, which assumes that the velocity remains constant in some cells, performed the best across all datasets. The dynamical estimation of velocyto, as implemented in scvelo, performed the worst of all parameter settings. This was mainly due to scvelo overestimating the dynamics of a gene, especially towards upregulation, while velocyto correctly estimated not only when a gene changes, but also when it remained in a steady state.

Cell-specific network inference

Cell-specific network inference (CSNI) methods[1] predict not only which transcription factors regulate which target genes (Figure 5A, top left), but also how active each interaction is in every case (Figure 5A).

While a few pioneering CSNI approaches have already been developed [19, 20, 21], a quantitative assessment of the performance is until now lacking. This is not surprising, as neither real nor in silico

[1] Different terms are commonly used when dealing with data from a particular source. For example, single-cell NI when applied to single-cell transcriptomics data; sample-specific NI when applied to bulk transcriptomics; patient-derived NI when applied to bulk profiles of patients. A more generalised variant of CSNI is casewise NI, which does not specify the type of data which is being analysed.

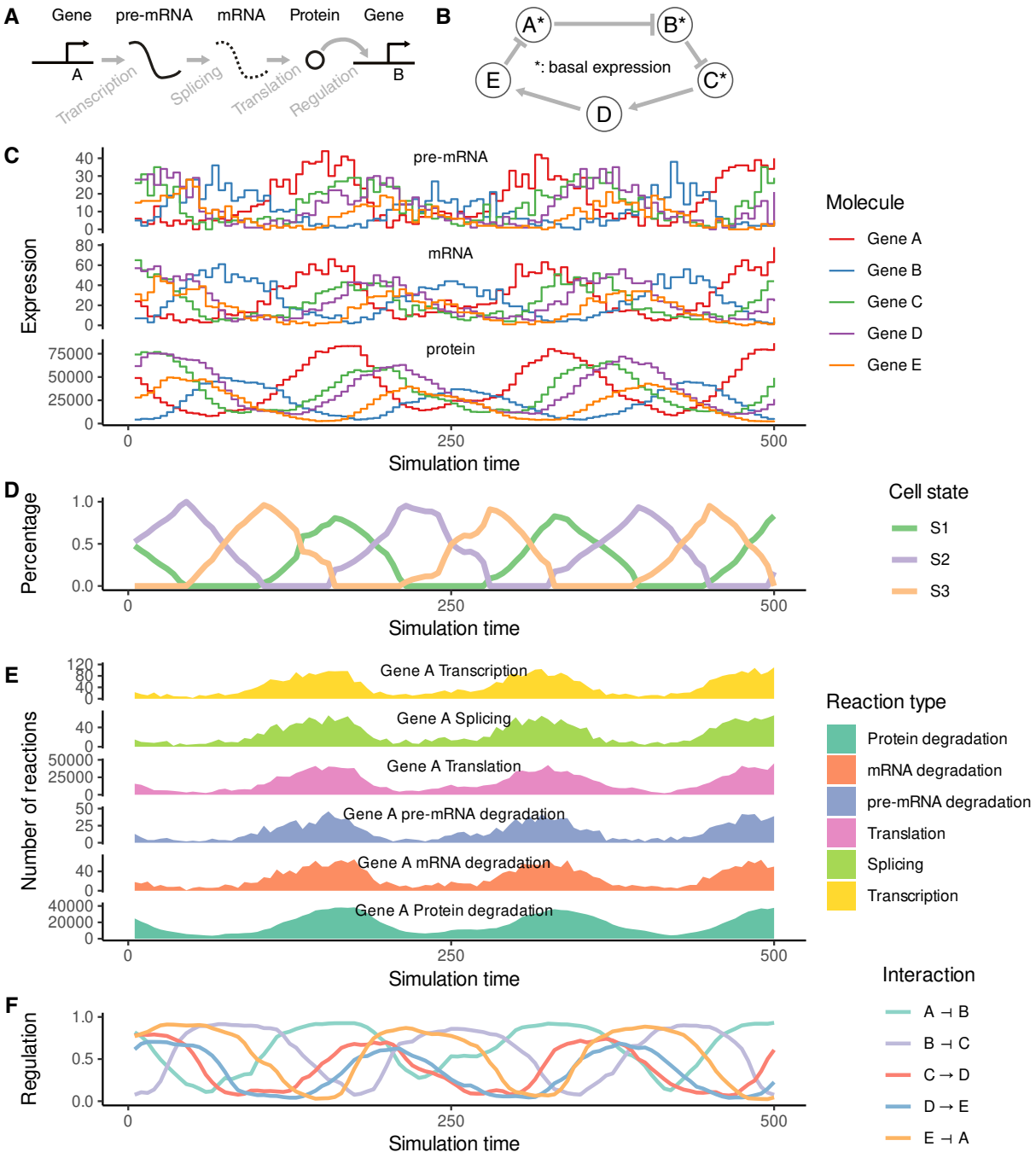

Figure 2: **dyngen models reactions at a single-molecule level and keeps track of multiple levels of information throughout a simulation. A:** Changes in abundance levels are driven strictly by gene regulatory reactions. **B:** The input GRN is defined such that it models a dynamic process of interest. **C:** The reactions define how abundance levels of molecules change at any particular time point. **D:** Firing many reactions can significantly alter the cellular state over time. **E:** dyngen keeps track of the reactions that were fired during small intervals of time. **F:** Similarly, dyngen can also keep track of the regulatory activity of every interaction.

datasets of cell-specific or even cell-type-specific interactions exist that are large enough so that it can be used as a ground-truth for evaluating CSNI methods. Extracting the ground-truth dynamic network in dyngen is straightforward though, given that we can calculate how target gene expression would change without the regulator being present.

We used this ground-truth to compare the performance of three CSNI methods (Figure 5B). We calculated the AUROC and AUPR score – which are common metrics for NI benchmarking – for each cell

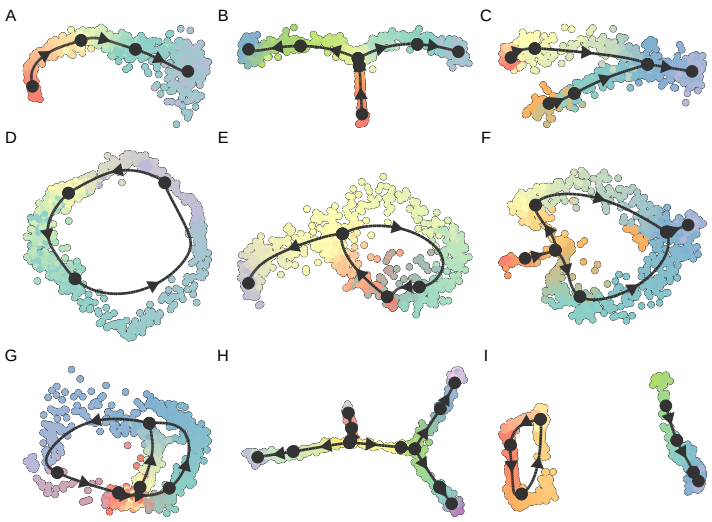

Figure 3: **Multiple executions of dyngen with different predefined backbones.** From each simulation of about 200 genes, 1000 cells were sampled. **A:** Linear. **B:** Bifurcating. **C:** Converging. **D:** Cyclic. **E:** Bifurcating loop. **F:** Bifurcating converging. **G:** Bifurcating cycle. **H:** Consecutive branching. **I:** Disconnected.

individually. Computing the mean AUROC and AUPR per dataset showed that pySCENIC significantly outperforms LIONESS + Pearson, which in turn outperforms SSN*.

Trajectory alignment

Trajectory alignment allows studying the differences between the same trajectories from different samples, as shown in Figure 6 A. For example, the cell developmental process of a patient could be compared to that of a healthy control to detect the transcriptomic differences of a particular lineage.

Dynamic Time Warping (DTW) [22] is most commonly used to align linear trajectories. DTW is a technique originating in the field of speech recognition and aligns temporal sequences by creating a warping path between the sequences that indicate which sequence must be dilated or contracted to best match the other one. DTW can also be used in combination with smoothed pseudo-cells [23].

Trajectory alignment, using DTW, has been used to compare gene expression kinetics resulting from different biological processes [24], to compare human, chimpanzee and macaque neuronal development [25], to find differences in gene regulation in the presence of certain growth factors [26], and to compare human and mouse embryogenesis [23].

The pseudotime values produced by dyngen are comparable across experiments, which allows us to evaluate the accuracy of an alignment technique. We use DTW, but process the trajectories in 3 different ways: we either used the original cells, we used pseudocells with a smoothed gene expression, or we used every tenth cell in the trajectory. An example of the different results obtained is shown in 6 B, C and D. We can see that the smoothed way of processing recovers the signal in the data best, it finds a diagonal warping path, which indicates that the two trajectories develop in a similar fashion. We test this on 10 pairs of two datasets, where an increasing amount of noise was added to the count matrix. As shown in Figure 6 E, the smoothed pseudocells perform best.

Discussion

As is, dyngen's single-cell simulations can be used to evaluate common single-cell omics computational methods such as clustering, batch correction, trajectory inference, and network inference. However, the combined effect of these advantages results in a framework that is flexible enough to adapt to a broad range of applications. This may include methods that integrate clustering, network

inference, and trajectory inference. In this respect, dyngen may promote the development of new tools in the single-cell field similarly as other simulators have done in the past [27, 28].

dyngen ultimately allows anticipating technological developments in single-cell multi-omics. In this way, it is possible to design and evaluate the performance and robustness of new types of computational analyses before experimental data becomes available. In addition, it could also be used to compare which experimental protocol is the most cost-effective in producing qualitative and robust results in downstream analysis.

Currently, dyngen focuses on simulating cells as standalone entities that are well mixed. Splitting up the simulation space into separate subvolumes could pave the way to better study key cellular processes such as cell division, intercellular communication, and migration [29].

Methods

The workflow to generate *in silico* single-cell data consists of six main steps (Figure 7).

Defining the module network

One of the main processes involved in cellular dynamic processes is gene regulation, where regulatory cascades and feedback loops lead to progressive changes in expression and decision making. The exact way a cell chooses a certain path during its differentiation is still an active research field, although certain models have already emerged and been tested *in vivo*. One driver of bifurcation is mutual antagonism, where two genes strongly repress each other [30, 31], forcing one of the two to become inactive [32]. Such mutual antagonism can be modelled and simulated [33, 34]. Although the two-gene model is simple and elegant, the reality is frequently more complex, with multiple genes (grouped into modules) repressing each other [35].

To start a dyngen simulation, the user needs to define a module network. The module network describes how sets of genes regulate each other and is what mainly determines which dynamic processes occur within the simulated cells.

A module network consists of modules connected together by regulatory interactions, which can be either up- or down-regulating. A module may have basal expression, which means genes in this module will be transcribed without the presence of transcription factor molecules. A module marked as "active during the burn phase" means that this module will be allowed to generate expression of its genes during an initial warm-up phase (See section). At the end of the dyngen process, cells will not be sampled from the burn phase simulations. Interactions between modules have a strength (which is a positive integer) and an effect (+1 for upregulating, -1 for downregulating).

Several examples of module networks are given (Figure 8). A simple chain of modules (where one module upregulates the next) results in a *linear* process. By having the last module repress the first module, the process becomes *cyclic*. Two modules repressing each other is the basis of a *bifurcating* process, though several chains of modules have to be attached in order to achieve progression before and after the bifurcation process. Finally, a *converging* process has a bifurcation occurring during the burn phase, after which any differences in module regulation is removed.

Note that these examples represent the bare minimum in terms of the number of modules used. Using longer chains of modules is typically desired. In addition, the fate decisions made in this example of a bifurcation is reversible, meaning cells can be reprogrammed to go down a different differentiation path. If this effect is undesirable, more safeguards need to be put in place to prevent reprogramming from occurring.

Generating the gene regulatory network

The GRN is generated based on the given module network in four main steps (Figure 9).

Step 1, sampling the transcription factors (TF). The TFs are the main drivers of the molecular changes in the simulation. The user provides a backbone and the number of TFs to generate. Each TF is assigned to a module such that each module has at least x parameters (default $x = 1$). A TF inherits the 'burn' and 'basal expression' from the module it belongs to.

Step 2, generating the TF interactions. Let each TF be regulated according to the interactions in the backbone. These interactions inherit the effect, strength, and independence parameters from the interactions in the backbone. A TF can only be regulated by other TFs or itself.

Step 3, sampling the target subnetwork. A user-defined number of target genes are added to the GRN. Target genes are regulated by a TF or another target gene, but are always downstream of at least one TF. To sample the interactions between target genes, one of the many FANTOM5 [36] GRNs is sampled. The currently existing TFs are mapped to regulators in the FANTOM5 GRN. The targets are drawn from the FANTOM5 GRN, weighted by their page rank value. For each target, at most x regulators are sampled from the induced FANTOM5 GRN (default $x = 5$). The interactions connecting a target gene and its regulators are added to the GRN.

Step 4, sampling the housekeeping subnetwork. Housekeeping genes are completely separate from any TFs or target genes. A user-defined set of housekeeping genes are also sampled from the FANTOM5 GRN. The interactions of the FANTOM5 GRN are first subsampled such that the maximum in-degree of each gene is x (default $x = 5$). A random gene is sampled and a breadth-first-search is performed to sample the desired number of housekeeping genes.

Convert gene regulatory network to a set of reactions

Simulating a cell's GRN makes use of a stochastic framework which tracks the abundance levels of molecules over time in a discrete quantity. For every gene G, the abundance levels of three molecules are tracked, namely of corresponding pre-mRNAs, mature mRNAs and proteins, which are represented by the terms x_G, y_G and z_G respectively. The GRN defines how a reaction affects the abundance levels of molecules and how likely it will occur. Gibson and Bruck [37] provide a good introduction to modelling gene regulation with stochastic frameworks, on which many of the concepts below are based.

For every gene in the GRN a set of reactions are defined, namely transcription, splicing, translation, and degradation. Each reaction consists of a propensity function – a formula $f(.)$ to calculate the probability $f(.) \times \mathrm{d}t$ of it occurring during a time interval $\mathrm{d}t$ – and the effect – how it will affect the current state if triggered.

The effects of each reaction mimic the respective biological processes (Table 1, middle). Transcription of gene G results in the creation of a single pre-mRNA molecule x_G. Splicing turns one pre-mRNA x_G into a mature mRNA x_G. Translation uses a mature mRNA y_G to produce a protein z_G. Pre-mRNA, mRNA and protein degradation results in the removal of a x_G, y_G, and z_G molecule, respectively.

The propensity of all reactions except transcription are all linear functions (Table 1, right) of the abundance level of some molecule multiplied by a per-gene constant (Table 2). The propensity of transcription of a gene G depends on the abundance levels of its TFs. The per-gene and per-interaction constants are based on the median reported production-rates and half-lives of molecules measured of 5000 mammalian genes [38], except that the transcription rate has been amplified by a factor of 10.

Table 1: **Reactions affecting the abundance levels of pre-mRNA x_G, mature mRNA y_G and proteins z_G of gene G.** Define the set of regulators of G as R_G, the set of upregulating regulators of G as R_G^+, and the set of downregulating regulators of G as R_G^-. Parameters used in the propensity formulae are defined in Table 2.

Reaction	Effect	Propensity
Transcription	$\emptyset \rightarrow \mathrm{x}_G$	$\mathrm{xpr}_G \times \frac{\mathrm{bas}_G - \mathrm{ind}_G^{\lvert \mathrm{R}_G^+ \rvert} + \prod_{H \in \mathrm{R}_G^+} (\mathrm{ind}_G + \mathrm{bind}_{G,H})}{\prod_{H \in \mathrm{R}_G} (1 + \mathrm{bind}_{G,H})}$
Splicing	$\mathrm{x}_G \rightarrow \mathrm{y}_G$	$\mathrm{ysr}_G \times \mathrm{x}_G$
Translation	$\mathrm{y}_G \rightarrow \mathrm{y}_G + \mathrm{z}_G$	$\mathrm{zpr}_G \times \mathrm{y}_G$
Pre-mRNA degradation	$\mathrm{x}_G \rightarrow \emptyset$	$\mathrm{ydr}_G \times \mathrm{x}_G$
Mature mRNA degradation	$\mathrm{y}_G \rightarrow \emptyset$	$\mathrm{ydr}_G \times \mathrm{y}_G$
Protein degradation	$\mathrm{z}_G \rightarrow \emptyset$	$\mathrm{zdr}_G \times \mathrm{z}_G$

Table 2: **Default parameters defined for the calculation of reaction propensity functions.**

Parameter	Symbol	Definition
Transcription rate	xpr_G	$\in U(10, 20)$
Splicing rate	ysr_G	$= \ln(2) \,/\, 2$
Translation rate	zpr_G	$\in U(100, 150)$
(Pre-)mRNA half-life	yhl_G	$\in U(2.5, 5)$
Protein half-life	zhl_G	$\in U(5, 10)$
Interaction strength	$\mathsf{str}_{G,H}$	$\in 10^{U(0,2)}$ *
Hill coefficient	$\mathsf{hill}_{G,H}$	$\in U(0.5, 2)$ *
Independence factor	ind_G	$\in U(0, 1)$ *
(Pre-)mRNA degradation rate	ydr_G	$= \ln(2) \,/\, \mathsf{yhl}_G$
Protein degradation rate	zdr_G	$= \ln(2) \,/\, \mathsf{zhl}_G$
Dissociation constant	dis_H	$= 0.5 \times \frac{\mathsf{xpr}_H \times \mathsf{ysr}_H \times \mathsf{zpr}_H}{(\mathsf{ydr}_H + \mathsf{ysr}_H) \times \mathsf{ydr}_H \times \mathsf{zdr}_H}$
Binding strength	$\mathsf{bind}_{G,H}$	$= \mathsf{str}_{G,H} \times (\mathsf{z}_H \,/\, \mathsf{dis}_H)^{\mathsf{hill}_{G,H}}$
Basal expression	bas_G	$= \begin{cases} 1 & \text{if } \mathsf{R}_G^+ = \emptyset \\ 0.0001 & \text{if } \mathsf{R}_G^- = \emptyset \text{ and } \mathsf{R}_G^+ \neq \emptyset \text{ *} \\ 0.5 & \text{otherwise} \end{cases}$

*: unless G is a TF, then the value is determined by the backbone.

The propensity of the transcription of a gene G is inspired by thermodynamic models of gene regulation [39], in which the promoter of G can be bound or unbound by a set of N transcription factors H_i. Let $f(\mathsf{z}_1, \mathsf{z}_2, \ldots, \mathsf{z}_N)$ denote the propensity function of G, in function of the abundance levels of the transcription factors. The following subsections explain and define the propensity function when $N = 1$, $N = 2$, and finally for an arbitrary N.

Propensity of transcription when $N = 1$

In the simplest case when $N = 1$, the promoter can be in one of two states. In state S_0, the promoter is not bound by any transcription factors, and in state S_1 the promoter is bound by H_1. Each state S_j is linked with a relative activation α_j, a number between 0 and 1 representing the activity of the promoter at this particular state. The propensity function is thus equal to the expected value of the activity of the promoter multiplied by the pre-mRNA production rate of G.

$$f(y_1, y_2, \ldots, y_N) = \mathsf{xpr} \cdot \sum_{j=0}^{2^N-1} \alpha_j \cdot P(S_j) \tag{1}$$

(2)

For $N = 1$, $P(S_1)$ is equal to the Hill equation, where k_i represents the concentration of H_i at half-occupation and n_i represents the Hill coefficient. Typically, n_i is between [1,10]

$$P(S_1) = \frac{y_1^{n_1}}{k_1^{n_1} + y_1^{n_1}} \tag{3}$$

$$= \frac{(y_1/k_1)^{n_1}}{1 + (y_1/k_1)^{n_1}} \tag{4}$$

The Hill equation can be simplified by letting $\nu_i = \left(\frac{y_i}{k_i}\right)^{n_i}$.

$$P(S_1) = \frac{\nu_1}{1 + \nu_1} \tag{5}$$

Since $P(S_0) = 1 - P(S_1)$, the activation function is formulated and simplified as follows.

$$f(y_1) = \text{xpr} \cdot (\alpha_0 \cdot P(S_0) + \alpha_1 \cdot P(S_1)) \tag{6}$$
$$= \text{xpr} \cdot \left(\alpha_0 \cdot \frac{1}{1+\nu_1} + \alpha_1 \cdot \frac{\nu_1}{1+\nu_1} \right) \tag{7}$$
$$= \text{xpr} \cdot \frac{\alpha_0 + \alpha_1 \cdot \nu_1}{1+\nu_1} \tag{8}$$
$$\tag{9}$$

Propensity of transcription when $N = 2$

When $N = 2$, there are four states S_j. The relative activations α_j can be defined such that H_1 and H_2 are independent (additive) or synergistic (multiplicative). In order to define the propensity of transcription $f(.)$, the Hill equation $P(S_j)$ is extended for two transcription factors.

Let w_j be the numerator of $P(S_j)$, defined as the product of all transcription factors bound in that state:

$$w_0 = 1 \tag{10}$$
$$w_1 = \nu_1 \tag{11}$$
$$w_2 = \nu_2 \tag{12}$$
$$w_3 = \nu_1 \cdot \nu_2 \tag{13}$$

The denominator of $P(S_j)$ is then equal to the sum of all w_j. The probability of state S_j is thus defined as:

$$P(S_j) = \frac{w_j}{\sum_{j=0}^{j<2^N} w_j} \tag{14}$$
$$= \frac{w_j}{1 + \nu_1 + \nu_2 + \nu_1 \cdot \nu_2} \tag{15}$$
$$= \frac{w_j}{\prod_{i=1}^{i \leq N} (\nu_i + 1)} \tag{16}$$

Substituting $P(S_j)$ and w_j into $f(.)$ results in the following equation:

$$f(y_1, y_2) = \text{xpr} \cdot \sum_{j=0}^{2^N - 1} \alpha_j \cdot P(S_j) \tag{17}$$
$$= \text{xpr} \cdot \frac{\sum_{j=0}^{2^N-1} \alpha_j \cdot w_j}{\prod_{i=1}^{i \leq N} (\nu_i + 1)} \tag{18}$$
$$= \text{xpr} \cdot \frac{\alpha_0 + \alpha_1 \cdot \nu_1 + \alpha_2 \cdot \nu_2 + \alpha_3 \cdot \nu_1 \cdot \nu_2}{(\nu_1 + 1) \cdot (\nu_2 + 1)} \tag{19}$$
$$\tag{20}$$

Propensity of transcription for an arbitrary N

For an arbitrary N, there are 2^N states S_j. The relative activations α_j can be defined such that H_1 and H_2 are independent (additive) or synergistic (multiplicative). In order to define the propensity of transcription $f(.)$, the Hill equation $P(S_j)$ is extended for N transcription factors.

Let w_j be the numerator of $P(S_j)$, defined as the product of all transcription factors bound in that state:

$$w_j = \prod_{i=1}^{i \leq N} (j \bmod i) = 1 \; ? \; \nu_i : 1 \tag{21}$$

The denominator of $P(S_j)$ is then equal to the sum of all w_j. The probability of state S_j is thus defined as:

$$P(S_j) = \frac{w_j}{\sum_{j=0}^{j<2^N} w_j} \tag{22}$$

$$= \frac{w_j}{\prod_{i=1}^{i \leq N} (\nu_i + 1)} \tag{23}$$

Substituting $P(S_j)$ into $f(.)$ yields:

$$f(y_1, y_2, \dots, y_N) = \text{xpr} \cdot \sum_{j=0}^{2^N - 1} \alpha_j \cdot P(S_j) \tag{24}$$

$$= \text{xpr} \cdot \frac{\sum_{j=0}^{2^N - 1} \alpha_j \cdot w_j}{\prod_{i=1}^{i \leq N} (\nu_i + 1)} \tag{25}$$

Propensity of transcription for a large N

For large values of N, computing $f(.)$ is practically infeasible as it requires performing 2^N summations. In order to greatly simplify $f(.)$, α_j could be defined as 0 when one of the regulators inhibits transcription and 1 otherwise.

$$\alpha_j = \begin{cases} 0 & \text{if } \exists i : j \bmod i = 1 \text{ and } H_i \text{ represses } G \\ 1 & \text{otherwise} \end{cases} \tag{26}$$

Substituting equation 26 into equation 25 and defining $R = \{1, 2, \dots, N\}$ and $R^+ = \{i | H_i \text{ activates } G\}$ yields the simplified propensity function:

$$f(y_1, y_2, \dots, y_N) = \text{xpr} \cdot \frac{\prod_{i \in R^+} (\nu_i + 1)}{\prod_{i \in R} (\nu_i + 1)} \tag{27}$$

Independence, synergism and basal expression

The definition of α_j as in equation 26 presents two main limitations. Firstly, since $\alpha_0 = 1$, it is impossible to tweak the propensity of transcription when no transcription factors are bound. Secondly, it is not possible to tweak the independence and synergism of multiple regulators.

Let ba $\in [0, 1]$ denote the basal expression strength G (i.e. how much will G be expressed when no transcription factors are bound), and sy $\in [0, 1]$ denote the synergism of regulators H_i of G, the transcription propensity becomes:

$$f(y_1, y_2, ..., y_N) = \text{xpr} \cdot \frac{\text{ba} - \text{sy}^{|R^+|} + \prod_{i \in R^+}(\nu_i + \text{sy})}{\prod_{i \in R}(\nu_i + 1)} \tag{28}$$

Simulate single cells

dyngen uses Gillespie's stochastic simulation algorithm (SSA) [13] to simulate dynamic processes. An SSA simulation is an iterative process where at each iteration one reaction is triggered.

Each reaction consists of its propensity – a formula to calculate the probability of the reaction occurring during an infinitesimal time interval – and the effect – how it will affect the current state if triggered. Each time a reaction is triggered, the simulation time is incremented by $\tau = \frac{1}{\sum_j prop_j} \ln\left(\frac{1}{r}\right)$, with $r \in U(0, 1)$ and $prop_j$ the propensity value of the jth reaction for the current state of the simulation.

GillespieSSA2 is an optimised library for performing SSA simulations. The propensity functions are compiled to C++ and SSA approximations can be used which allow triggering many reactions simultaneously at each iteration. The framework also allows storing the abundance levels of molecules only after a specific interval has passed since the previous census. By setting the census interval to 0, the whole simulation's trajectory is retained but many of these time points will contain very similar information. In addition to the abundance levels, also the propensity values and the number of firings of each of the reactions at each of the time steps can be retained, as well as specific sub-calculations of the propensity values, such as the regulator activity level $reg_{G,H}$.

Simulate experiment

From the SSA simulation we obtain the abundance levels of all the molecules at every state. We need to replicate technical effects introduced by experimental protocols in order to obtain data that is similar to real data. For this, the cells are sampled from the simulations and molecules are sampled for each of the cells. Real datasets are used in order to achieve similar data characteristics.

Sample cells

In this step, N cells are sampled the simulations. Two approaches are implemented: sampling from an unsynchronised population of single cells (snapshot) or sampling at multiple time points in a synchronised population (time series).

Snapshot The backbone consists of several states linked together by transition edges with length L_i, to which the different states in the different simulations have been mapped (Figure 10A). From each transition, $N_i = N/\frac{L_i}{\sum L_i}$ cells are sampled uniformly, rounded such that $\sum N_i = N$.

Time series Assuming that the final time of the simulations is T, the interval $[0, T]$ is divided into k equal intervals of width w separated by $k-1$ gaps of width g. $N_i = N/k$ cells are sampled uniformly from each interval (Figure 10B), rounded such that $\sum N_i = N$. By default, $k = 8$ and $g = 0.75$. For usual dyngen simulations, $10 \leq T \leq 20$. For larger values of T, k and g should be increased accordingly.

Sample molecules

Molecules are sampled from the simulation to replicate how molecules are experimentally sampled. A real dataset is downloaded from a repository of single-cell RNA-seq datasets [40]. For each *in silico* cell i, draw its library size ls_i from the distribution of transcript counts per cell in the real dataset. The capture rate cr_j of each *in silico* molecule type j is drawn from $N(1, 0.05)$. Finally, for each cell

i, draw ls_i molecules from the multinomial distribution with probabilities $cr_j \times ab_{i,j}$ with $ab_{i,j}$ the molecule abundance level of molecule j in cell i.

Determining the ground-truth trajectory

To construct the ground-truth trajectory, the user needs to provide the ground-truth state network alongside the initial module network (Figure~11). Each edge in the state network specifies which modules are allowed to change in expression in transitioning from one state to another. For each edge, a simulation is run using the end state of an upstream branch as the initial expression vector, and only allowing the modules as predefined by the attribute to change.

As an example, consider the cyclic trajectory shown in Figure~11. State S0 begins with an expression vector of all zero values. To simulate the transition from S0 to S1, regulation of the genes in modules A, B and C are turned on. After a predefined period of time, the end state of this transition is considered the expression vector of state S1. To simulate the transition from S1 to S2, regulation of the genes in modules D and E are turned on, while the regulation of genes in module C is turned off. During this simulation, the expression of genes in modules A, B, D, and E is thus allowed to change. The end state of the simulation is considered the expression vector of state S2.

For each of the branches in the state network, an expression matrix and the corresponding progression time along that branch are retained. To map a simulated cell to the ground-truth, the correlation between its expression values and the expression matrix of the ground-truth trajectory is calculated, and the cell is mapped to the position in the ground-truth trajectory that has the highest correlation.

Determining the cell-specific ground-truth regulatory network

Calculating the regulatory effect of a regulator R on a target T (Figure 7F) requires determining the contribution of R in the propensity function of the transcription of T (section) with respect to other regulators. This information is useful, amongst others, for benchmarking cell-specific network inference methods.

The regulatory effect of R on T at a particular state S is defined as the change in the propensity of transcription when R is set to zero, scaled by the inverse of the pre-mRNA production rate of T. More formally:

$$\text{regeffect}_G = \frac{\text{proptrans}_G(S) - \text{proptrans}_G(S[\text{z}_T \leftarrow 0])}{\text{xpr}_G}$$

Determining the regulatory effect for all interactions and cells in the dataset yields the complete cell-specific ground-truth GRN (Figure 12). The regulatory effect lie between $[-1, 1]$, where -1 represents complete inhibition of T by R, 1 represents maximal activation of T by R, and 0 represents inactivity of the regulatory interaction between R and T.

Comparison of cell-specific network inference methods

14 datasets were generated using the 14 different predefined backbones. For every cell in the dataset, the transcriptomics profile and the corresponding cell-specific ground-truth regulatory network was determined (Section).

Several cell-specific NI methods were considered for comparison: SCENIC [19], LIONESS [41, 20], and SSN [21].

LIONESS [20] runs a NI method multiple times to construct cell-specific GRNs. LIONESS first infers a GRN with all of the samples. A second GRN is inferred with all samples except one particular profile. The cell-specific GRN for that particular profile is defined as the difference between the two GRN matrices. This process is repeated for all profiles, resulting in a cell-specific GRN. By default, LIONESS uses PANDA [42] to infer GRNs, but since dyngen does not produce motif data and motif data is required by PANDA, PANDA is inapplicable in this context. Instead, we used the lionessR [43] implementation of LIONESS, which uses by default the Pearson correlation as a NI method. We marked results from this implementation as "LIONESS + Pearson".

SSN [21] follows, in essence, the exact same methodology as LIONESS except that it specifically only uses the Pearson correlation. It is worth noting that the LIONESS preprint was released before the publication of SSN. Since no implementation was provided by the authors, we implemented SSN in R using basic R and tidyverse functions [44] and marked results from this implementation as "SSN*".

SCENIC [19] is a pipeline that consists of four main steps. Step 1: classical network inference is performed with arboreto, which is similar to GENIE3 [45]. Step 2: select the top 10 regulators per target. Interactions are grouped together in 'modules'; each module contains one regulator and all of its targets. Step 3: filter the modules using motif analysis. Step 4: for each cell, determine an activity score of each module using AUCell. As a post-processing of this output, all modules and the corresponding activity scores are combined back into a cell-specific GRN consisting of (cell, regulator, target, score) pairs. For this analysis, the Python implementation of SCENIC was used, namely pySCENIC. Since dyngen does not generate motif data, step 3 in this analysis is skipped.

The AUROC and AUPR metrics are common metrics for evaluating a predicted GRN with a ground-truth GRN. To compare a predicted cell-specific GRN with the ground-truth cell-specific GRN, the top 10'000 interactions per cell were retained. For each cell-specific network, the AUROC and AUPR were calculated.

Comparison of RNA velocity methods

For each of the 14 different predefined backbones, nine datasets were generated with three difficulty settings and three different seeds. The different difficulty settings were obtained by multiplying the transcription rate by a factor of 25 (easy), 5 (medium) and 1 (hard). After manual quality control, the easy and medium datasets for backbones "bifurcating_converging", "bifurcating_loop", "converging" and "disconnected" were removed, resulting in a final collection of 102 datasets.

We extracted a ground truth RNA velocity by subtracting for each mRNA molecule the propensity of its production by the propensity of its degradation. If the expression of an mRNA will increase in the future, this value is positive, while it is negative if it is going to decrease. For each gene, we compared the ground truth velocity with the observed velocity by calculating the Spearman rank correlation.

We compared two RNA velocity methods. The velocyto method [17], as implemented in the velocyto.py package, in which we varied the "assumption" parameter between "constant_unspliced" and "constant_velocity". The scvelo method [18], as implemented in the python scvelo package scvelo.de, in which we varied the "mode" parameter between "deterministic", "stochastic", "dynamical", "dynamical_residuals". For both methods, we used the same normalized data as provided by dyngen, with no extra cell or feature filtering, but otherwise matched the parameters to their respective tutorial vignettes as well as possible.

To visualize the velocity on an embedding, we used the "velocity_embedding" function, implemented in the scvelo python package.

Comparison of trajectory alignment with added noise

We generated 10 base datasets that contain a linear trajectory using dyngen. For each dataset, we performed the cell generation and experiment generation twice. This results in 20 datasets, with 10 pairs of similar trajectories. For each of these pairs of datasets, we generated 10 progressively noisier pairs, in which we added noise to the complete count matrix. Let q be the 75th quantile of the non-zero values in the count matrix. The noise added to the count matrix is calculated as follows, with i being the noise parameter.

$$countmatrix + \frac{1}{\sqrt{2\pi}} qie^{\frac{x^2}{2(qi)^2}} \quad \forall i \in \{0.1, 0.2, ...1.0\} \tag{29}$$

We aligned each of these 150 pairs of trajectories (the first trajectory in the pair with the second trajectory in the pair) in 3 different ways. Either we used the original count matrix, we used 1/10th of the cells (according to the ordering of the trajectory, so the 1st, 10th, 20th, ... cell of the trajectory) or we used the smoothed pseudocells. The pseudocells are generated at regular points across the pseudotime of the trajectory. The gene expression of each pseudocell is inferred using a Gaussian kernel with a window size of 0.05.

Each of the cells or pseudocells received a pseudotime from dyngen. This time can be compared: cells from the first trajectory in the pair should be aligned closely to cells from the second trajectory with similar pseudotimes. We used the sum of differences between the pseudotimes of the aligned cells to determine the distance between the two trajectories, which should ideally be 0. In order to compare the alignments with 1000 cells per trajectory and the ones with 100 cells per trajectory, we divide the distance of those alignments by 10.

Availability

dyngen is available as an R package on GitHub github.com/dynverse/dyngen. The dyngen codebase is still under development and is thus likely subject to change, including the adding or removing of functionality, renaming functions or parameters, and changing default parameter values. The analyses performed in this manuscript are available on GitHub github.com/dynverse/dyngen_manuscript.

Author contributions

- W.S. and R.C. designed the study.
- R.C., W.S., and L.D. performed the experiments and analysed the data.
- R.C. and W.S. implemented the dyngen software package.
- R.C. and W.S. wrote the original manuscript.
- L.D. wrote the section on trajectory alignment.
- R.C., W.S., L.D., and Y.S. reviewed and edited the manuscript.
- Y.S. supervised the project.

References

[1] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Splatter: Simulation of Single-Cell RNA Sequencing Data". In: *Genome Biology* 18 (Sept. 2017), p. 174. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1305-0.

[2] Beate Vieth et al. "powsimR: power analysis for bulk and single cell RNA-seq experiments". In: *Bioinformatics* 33.21 (Nov. 1, 2017), pp. 3486–3488. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx435. URL: https://academic.oup.com/bioinformatics/article/33/21/3486/3952669 (visited on 08/28/2019).

[3] Nikolaos Papadopoulos, Rodrigo Gonzalo Parra, and Johannes Soeding. "PROSSTT: Probabilistic Simulation of Single-Cell RNA-Seq Data for Complex Differentiation Processes". In: *bioRxiv* (Jan. 2018), p. 256941. DOI: 10.1101/256941.

[4] Xiuwei Zhang, Chenling Xu, and Nir Yosef. "Simulating multiple faceted variability in single cell RNA sequencing". In: *Nature Communications* 10.1 (June 13, 2019), pp. 1–16. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10500-w. URL: https://www.nature.com/articles/s41467-019-10500-w (visited on 09/09/2019).

[5] Kelly Street et al. "Slingshot: Cell Lineage and Pseudotime Inference for Single-Cell Transcriptomics". In: *BMC Genomics* 19.1 (June 2018), p. 477. ISSN: 1471-2164. DOI: 10.1186/s12864-018-4772-0.

[6] R Gonzalo Parra et al. "Reconstructing Complex Lineage Trees from scRNA-Seq Data Using MERLoT". In: *bioRxiv* (Feb. 2018), p. 261768. DOI: 10.1101/261768.

[7] Edroaldo Lummertz da Rocha et al. "Reconstruction of complex single-cell trajectories using CellRouter". In: *Nature Communications* 9.1 (Mar. 1, 2018), p. 892. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03214-y. URL: https://doi.org/10.1038/s41467-018-03214-y.

[8] Yingxin Lin et al. "scClassify: hierarchical classification of cells". In: *bioRxiv* (Jan. 1, 2019), p. 776948. DOI: 10.1101/776948. URL: http://biorxiv.org/content/early/2019/09/26/776948.abstract.

[9] Angelo Duò, Mark D. Robinson, and Charlotte Soneson. "A systematic performance evaluation of clustering methods for single-cell RNA-seq data". In: *F1000Research* 7 (2018), p. 1141. ISSN: 2046-1402. DOI: 10.12688/f1000research.15666.2.

[10] Wouter Saelens et al. "A comparison of single-cell trajectory inference methods". In: *Nature Biotechnology* 37 (May 2019). ISSN: 15461696. DOI: 10.1038/s41587-019-0071-9. URL: http://dx.doi.org/10.1038/s41587-019-0071-9.

[11] Charlotte Soneson and Mark D. Robinson. "Bias, robustness and scalability in single-cell differential expression analysis". In: *Nature Methods* 15.4 (2018), pp. 255–261. ISSN: 1548-7105. DOI: 10.1038/nmeth.4612.

[12] Tim Stuart and Rahul Satija. "Integrative single-cell analysis". In: *Nature Reviews Genetics* 20.5 (May 2019), pp. 257–272. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0093-7. URL: https://www.nature.com/articles/s41576-019-0093-7 (visited on 12/09/2019).

[13] Daniel T. Gillespie. "Exact stochastic simulation of coupled chemical reactions". In: *The Journal of Physical Chemistry* 81.25 (Dec. 1, 1977), pp. 2340–2361. ISSN: 0022-3654. DOI: 10.1021/j100540a008. URL: https://doi.org/10.1021/j100540a008 (visited on 08/31/2019).

[14] Koen Van den Berge et al. "Trajectory-based differential expression analysis for single-cell sequencing data". In: *bioRxiv* (Jan. 1, 2019), p. 623397. DOI: 10.1101/623397. URL: http://biorxiv.org/content/early/2019/05/02/623397.abstract.

[15] Aditya Pratapa et al. "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data". In: *bioRxiv* (June 4, 2019), p. 642926. DOI: 10.1101/642926. URL: https://www.biorxiv.org/content/10.1101/642926v3 (visited on 08/24/2019).

[16] Amit Zeisel et al. "Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli". In: *Molecular Systems Biology* 7.1 (Jan. 1, 2011), p. 529. ISSN: 1744-4292. DOI: 10.1038/msb.2011.62. URL: https://www.embopress.org/doi/full/10.1038/msb.2011.62 (visited on 12/05/2019).

[17] Gioele La Manno et al. "RNA Velocity of Single Cells". In: *Nature* 560.7719 (Aug. 2018), pp. 494–498. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0414-6.

[18] Volker Bergen et al. "Generalizing RNA velocity to transient cell states through dynamical modeling". In: *bioRxiv* (Oct. 29, 2019), p. 820936. DOI: 10.1101/820936. URL: https://www.biorxiv.org/content/10.1101/820936v1 (visited on 12/09/2019).

[19] Sara Aibar et al. "SCENIC: single-cell regulatory network inference and clustering". In: *Nature Methods* (Oct. 2017). ISSN: 1548-7091. DOI: 10.1038/nmeth.4463. URL: http://www.nature.com/doifinder/10.1038/nmeth.4463.

[20] Marieke Lydia Kuijjer et al. "Estimating Sample-Specific Regulatory Networks". In: *iScience* 14 (Mar. 28, 2019), pp. 226–240. ISSN: 2589-0042. DOI: 10.1016/j.isci.2019.03.021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6463816/ (visited on 10/14/2019).

[21] Xiaoping Liu et al. "Personalized characterization of diseases using sample-specific networks". In: *Nucleic Acids Research* 44.22 (2016), e164–e164. ISSN: 0305-1048. DOI: 10.1093/nar/gkw772.

[22] Toni Giorgino. "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package". In: *Journal of Statistical Software* 7 (Sept. 2009). DOI: 10.18637/jss.v031.i07.

[23] Ayelet Alpert et al. "Alignment of single-cell trajectories to compare cellular expression dynamics". In: *Nature Methods* 15.4 (Apr. 2018), pp. 267–270. ISSN: 1548-7105. DOI: 10.1038/nmeth.4628. URL: https://www.nature.com/articles/nmeth.4628 (visited on 12/09/2019).

[24] Davide Cacchiarelli et al. "Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Myogenic Reprogramming Outcome". In: *Cell Systems* 7.3 (Sept. 26, 2018), 258–268.e3. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.07.006. URL: https://www.cell.com/cell-systems/abstract/S2405-4712(18)30314-4 (visited on 12/09/2019).

[25] Sabina Kanton et al. "Organoid single-cell genomic atlas uncovers human-specific features of brain development". In: *Nature* 574.7778 (Oct. 2019), pp. 418–422. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1654-9. URL: https://www.nature.com/articles/s41586-019-1654-9 (visited on 12/09/2019).

[26] José L. McFaline-Figueroa et al. "A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition". In: *Nature Genetics* 51.9 (Sept. 2019), pp. 1389–1398. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0489-5. URL: https://www.nature.com/articles/s41588-019-0489-5 (visited on 12/09/2019).

[27] Thomas Schaffter, Daniel Marbach, and Dario Floreano. "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods." In: *Bioinformatics* 27.16 (Aug. 2011), pp. 2263–2270. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr373. URL: http://www.ncbi.nlm.nih.gov/pubmed/21697125.

[28] Adam D. Ewing et al. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection". In: *Nature Methods* 12.7 (July 2015), pp. 623–630. ISSN: 1548-7105. DOI: 10.1038/nmeth.3407. URL: https://www.nature.com/articles/nmeth.3407 (visited on 08/28/2019).

[29] Stephen Smith and Ramon Grima. "Spatial Stochastic Intracellular Kinetics: A Review of Modelling Approaches". In: *Bulletin of Mathematical Biology* 81.8 (Aug. 1, 2019), pp. 2960–3009. ISSN: 1522-9602. DOI: 10.1007/s11538-018-0443-1. URL: https://doi.org/10.1007/s11538-018-0443-1.

[30] N. Rekhtman et al. "Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells". In: *Genes & Development* 13.11 (June 1, 1999), pp. 1398–1411. ISSN: 0890-9369. DOI: 10.1101/gad.13.11.1398.

[31] Heping Xu et al. "Regulation of bifurcating {B} cell trajectories by mutual antagonism between transcription factors {IRF4} and {IRF8}". In: *Nat. Immunol.* 16.12 (Dec. 2015), pp. 1274–1281.

[32] Thomas Graf and Tariq Enver. "Forcing Cells to Change Lineages". In: *Nature* 462.7273 (Dec. 2009), p. 587. ISSN: 1476-4687. DOI: 10.1038/nature08533.

[33] Jin Wang et al. "Quantifying the {{Waddington}} Landscape and Biological Paths for Development and Differentiation". In: *Proceedings of the National Academy of Sciences* 108.20 (May 2011), pp. 8257–8262. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1017017108.

[34] James E Ferrell. "Bistability, Bifurcations, and Waddington's Epigenetic Landscape". In: *Current Biology* 22.11 (June 2012), R458–R466. ISSN: 0960-9822. DOI: 10.1016/j.cub.2012.03.045.

[35] Nir Yosef et al. "Dynamic regulatory network controlling {TH17} cell differentiation". In: *Nature* 496.7446 (2013), pp. 461–468.

[36] Marina Lizio et al. "Gateways to the FANTOM5 promoter level mammalian expression atlas". In: *Genome Biology* 16.1 (Jan. 5, 2015), p. 22. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0560-6. URL: https://doi.org/10.1186/s13059-014-0560-6.

[37] Michael A. Gibson and Jehoshua Bruck. "A probabilistic model of a prokaryotic gene and its regulation". In: *Computational Methods in Molecular Biology: From Genotype to Phenotype, MIT press, Cambridge* (2000).

[38] Björn Schwanhäusser et al. "Global quantification of mammalian gene expression control". In: *Nature* 473.7347 (May 1, 2011), pp. 337–342. ISSN: 1476-4687. DOI: 10.1038/nature10098. URL: https://doi.org/10.1038/nature10098.

[39] Maria J. Schilstra and Chrystopher L. Nehaniv. "Bio-Logic: Gene Expression and the Laws of Combinatorial Logic". In: *Artificial Life* 14.1 (Jan. 1, 2008), pp. 121–133. ISSN: 1064-5462. DOI: 10.1162/artl.2008.14.1.121. URL: https://doi.org/10.1162/artl.2008.14.1.121 (visited on 12/11/2019).

[40] Robrecht Cannoodt et al. "Single-cell -omics datasets containing a trajectory". In: *Zenodo* (Oct. 2018). DOI: 10.5281/zenodo.1211532. URL: https://doi.org/10.5281/zenodo.1211532.

[41] Marieke Lydia Kuijjer et al. "Estimating sample-specific regulatory networks". In: (2015), pp. 1–19. URL: http://arxiv.org/abs/1505.06440.

[42] Kimberly Glass et al. "Passing Messages between Biological Networks to Refine Predicted Interactions". In: *PLOS ONE* 8.5 (May 31, 2013), e64832. DOI: 10.1371/journal.pone.0064832. URL: https://doi.org/10.1371/journal.pone.0064832.

[43] Marieke L. Kuijjer et al. "lionessR: single sample network inference in R". In: *BMC Cancer* 19.1 (Oct. 25, 2019), p. 1003. ISSN: 1471-2407. DOI: 10.1186/s12885-019-6235-7. URL: https://doi.org/10.1186/s12885-019-6235-7 (visited on 12/05/2019).

[44] Hadley Wickham et al. "Welcome to the Tidyverse". In: (Nov. 21, 2019). DOI: 10.21105/joss.01686. URL: https://joss.theoj.org (visited on 11/26/2019).

[45] Vân Anh Huynh-Thu et al. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods". In: *PLoS ONE* 5.9 (Jan. 2010), e12776. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0012776. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2946910&tool=pmcentrez&rendertype=abstract.

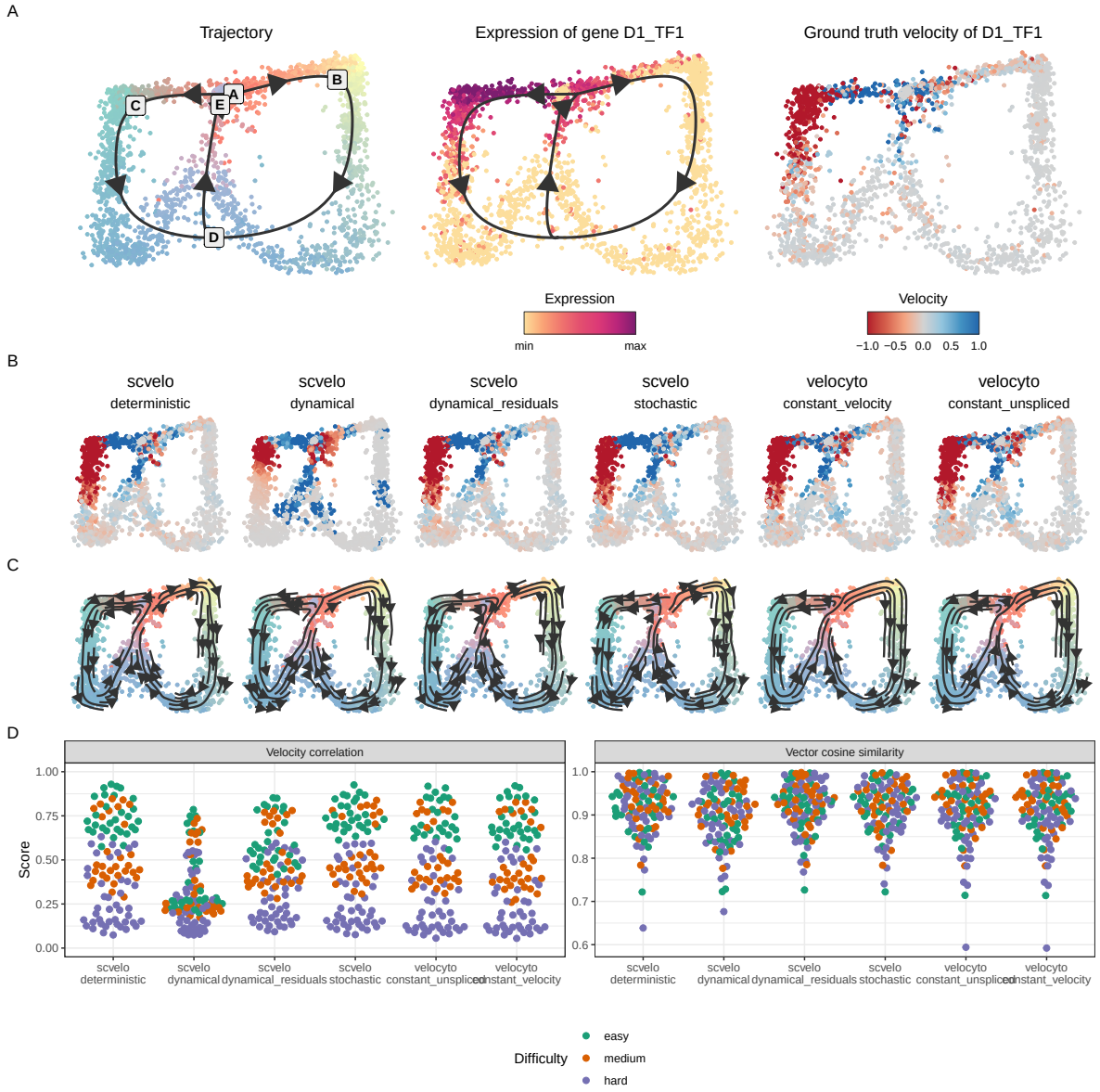

Figure 4: **dyngen allows benchmarking of RNA velocity methods. A:** An example bifurcating cycle dataset, with as illustration the expression and ground truth velocity of a gene D1_TF1 that goes up and down in one branch of the trajectory. **B:** The RNA velocity estimates of gene D1_TF1 by the different methods. **C:** The velocity stream plots produced from the predictions of each method, as generated by scvelo. **D:** The predictions scored by two different metrics, the velocity correlation and the vector cosine similarity. The velocity correlation is the correlation between the ground-truth velocity (A, right) and the predicted velocity (B). The vector cosine similarity is the cosine similarity between the direction of segments of the ground-truth trajectory (A, left) and the RNA velocity values calculated at those points (C).

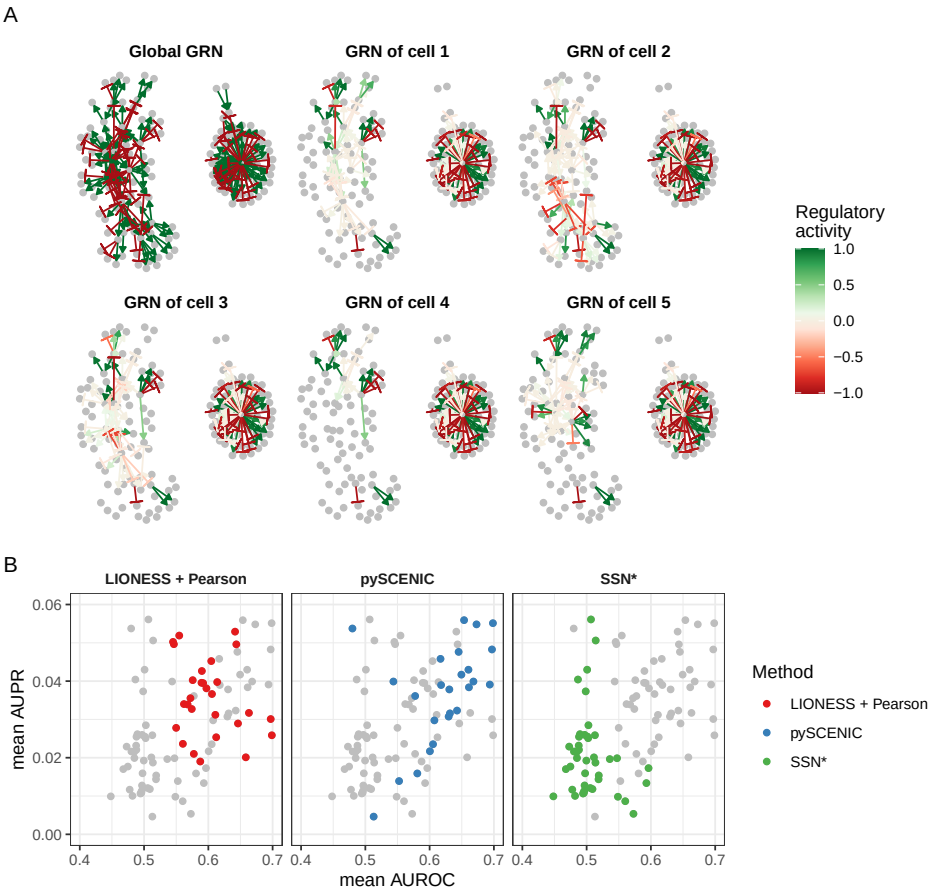

Figure 5: **dyngen allows benchmarking Cell-specific Network Inference (CSNI) methods. A:** A cell is simulated using the global gene regulatory network (GRN, top left). However, at any particular state in the simulation, only a fraction of the gene regulatory interactions are active. **B:** CSNI methods were executed to predict the regulatory interactions that are active in each cell specifically. Using the ground-truth cell-specific GRN, the performance of each method was quantified on 14 dyngen datasets.

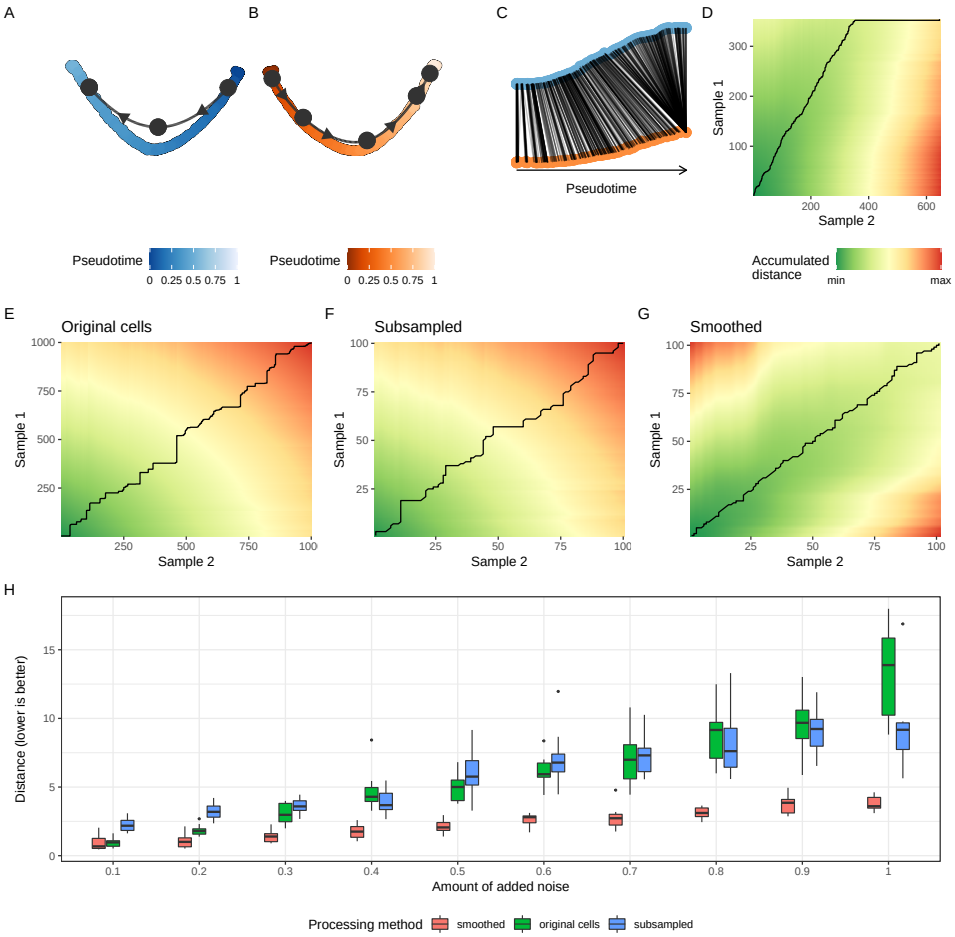

Figure 6: **dyngen allows benchmarking of trajectory alignment methods. A**: Example of a trajectory alignment process. The pseudotimes of individual inferred trajectories can be aligned to indicate differences between the trajectories. **B, C, D**: Shows the accumulated distance matrices obtained after using DTW on two trajectories where noise (noise level of 0.4) was added to the count matrix. In B the complete count matrices were used to perform the alignment. In C, each 10th cell was used to perform the alignment. In D, smoothed pseudocells were used. **E:** Shows the influence of added noise to the different processing methods. We can see that smoothing performs best in noisy circumstances.

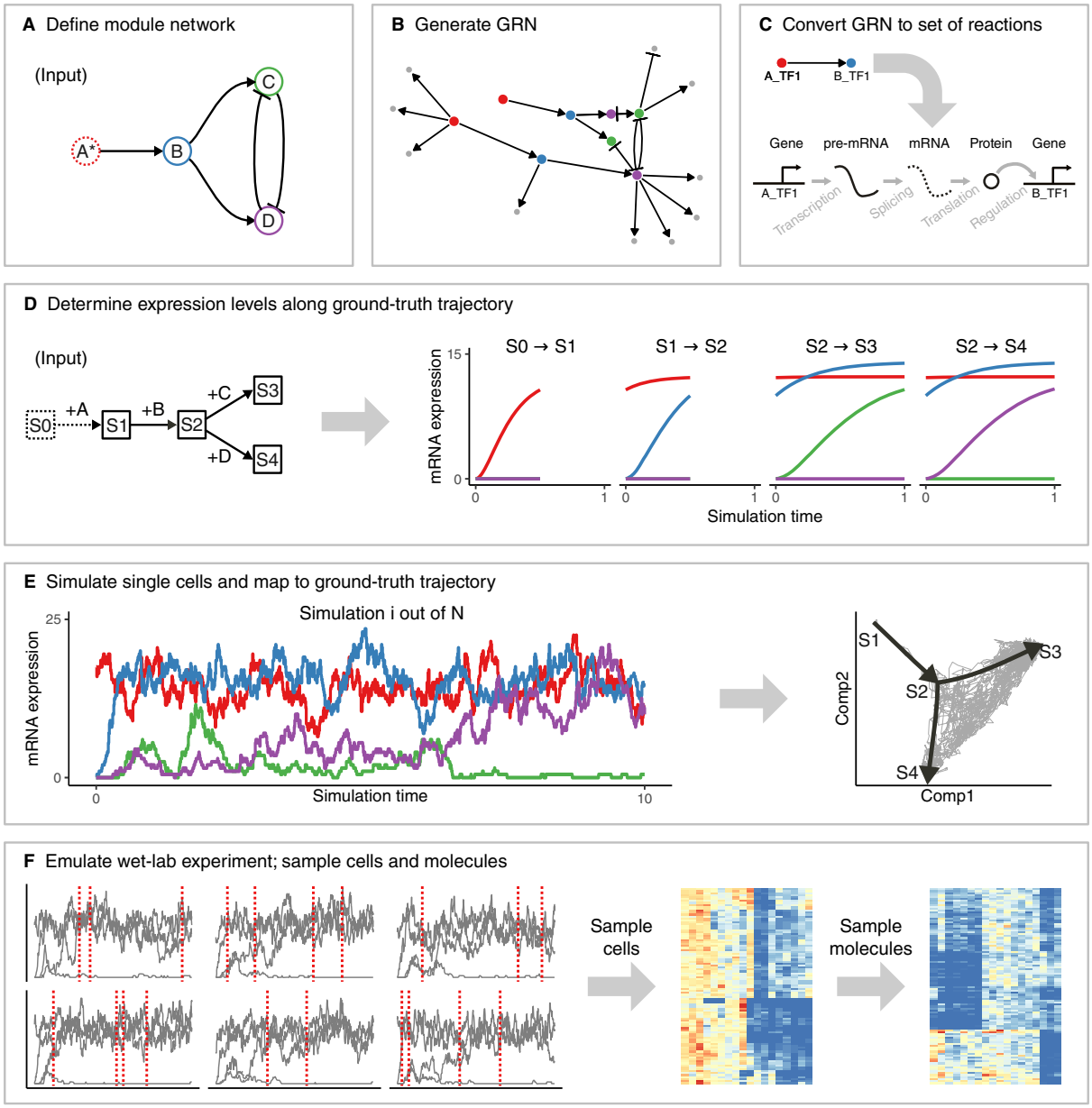

Figure 7: **The workflow of dyngen consists of six main steps. A:** The user needs to specify the desired module network or use a predefined module network. The module network is what determines the dynamic behaviour of simulated cells. **B:** The number of desired transcription factors (which drive the desired dynamic process) are amongst the given modules and adds regulatory interactions according to the module network. Additional target genes (which do not influence the dynamic process) are added by sampling interactions from GRN interaction databases. **C:** Each gene regulatory interaction in the GRN is converted to a set of biochemical reactions. **D:** Along with the module network, the user also needs to specify the backbone structure of expected cell states. The average expression of each edge in the backbone is simulated by activating a restricted set of genes for each edge. **E:** Multiple Gillespie SSA simulations are run using the reactions defined in step C. The counts of each of the molecules at each time step are extracted. Each time step is mapped to a point in the backbone. **F:** The molecule levels of multiple simulations are shown over time (left). From each simulation, multiple cells are sampled (from left to middle). Technical noise from profiling is simulated by sampling molecules from the set of molecules inside each cell (from middle to right).

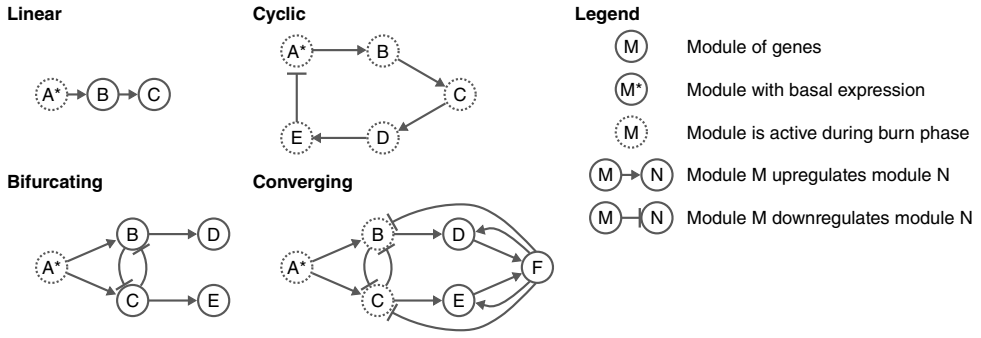

Figure 8: **The module network determines the type of dynamic process which simulated cells will undergo.** A module network describes the regulatory interactions between sets of transcription factors which drive the desired dynamic process.

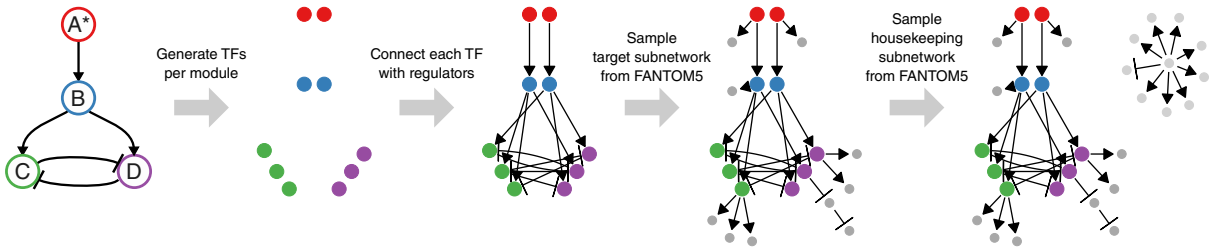

Figure 9: **Generating the feature network from a backbone consists of four main steps.**

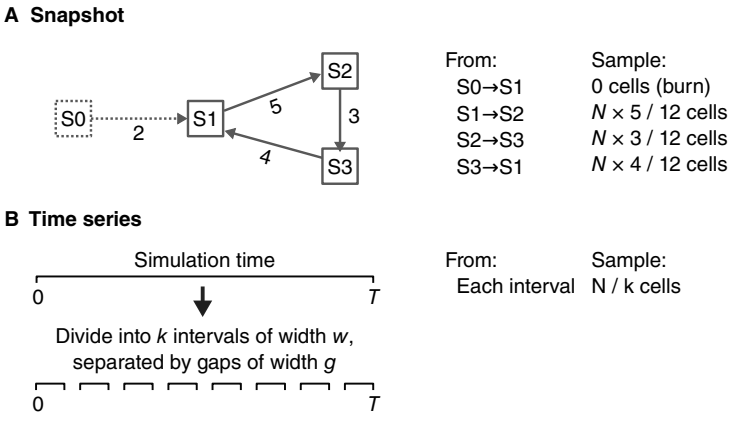

Figure 10: **Two approaches can be used to sample cells from simulations: snapshot and time-series.**

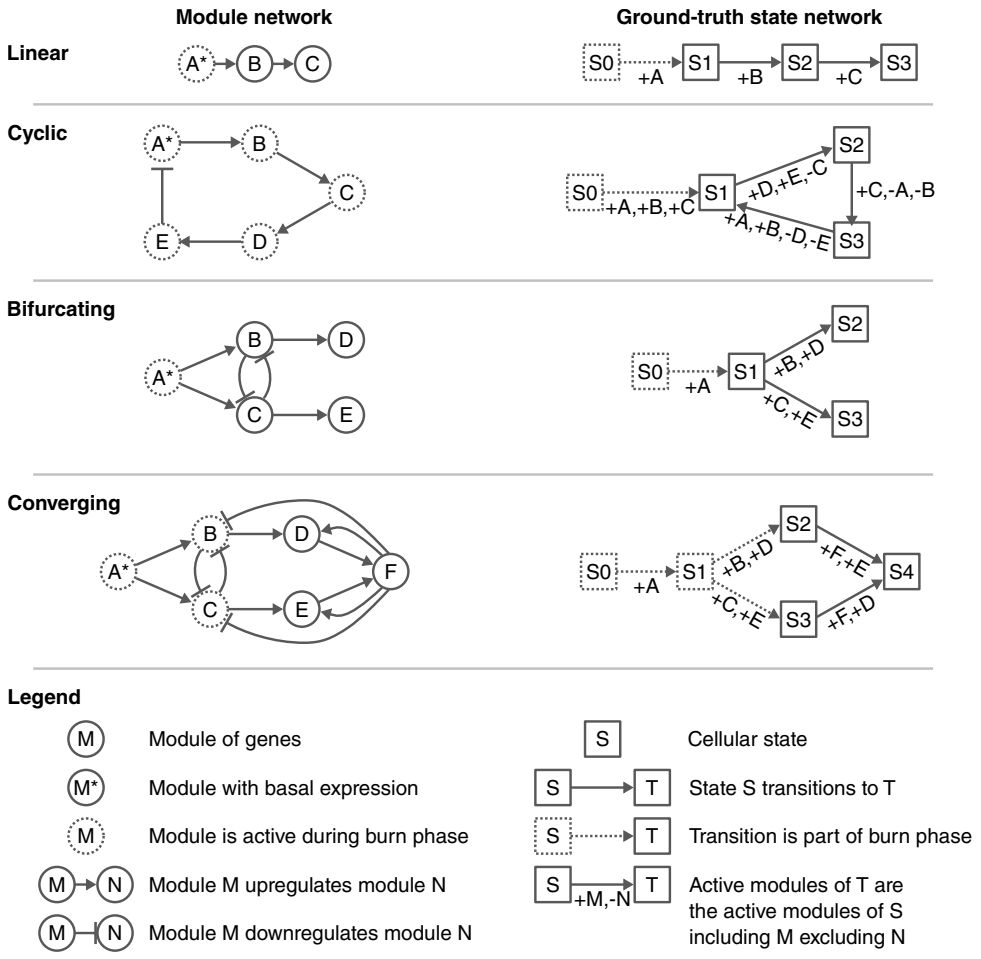

Figure 11: Examples of the ground-truth state networks which need to be provided alongside the module network.

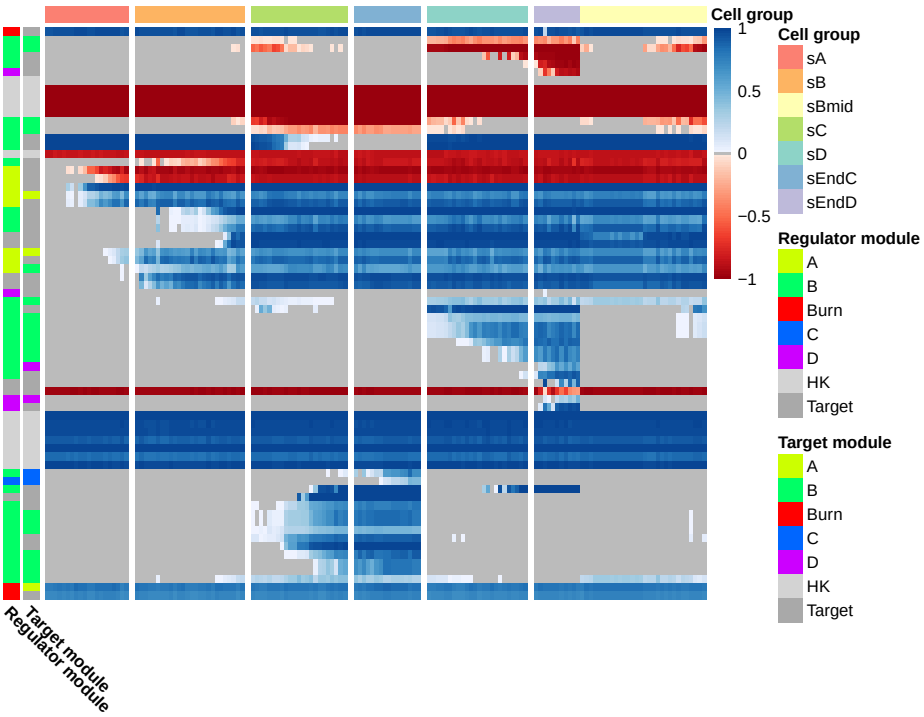

Figure 12: **The cell-specific regulatory effects of all interactions, computed on cells part of a bifurcation trajectory.** Negative values correspond to inhibitory interactions, positive values to activating interactions, and zero values correspond to inactive interactions.