國立臺灣大學電機資訊學院
生物資訊學國際研究生博士學位學程
博士論文
Taiwan International Graduate Program on Bioinformatics
College of Electrical Engineering and Computer Science
National Taiwan University
Doctoral Dissertation

利用未剪切轉錄建構情境依賴的基因調控網路可解釋調控動態及細胞軌跡

Context-Dependent Gene Regulatory Network Explains Regulation Dynamics and Cell Trajectories Using Unspliced Transcripts

杜岳華
Yueh-Hua Tu

指導教授：黃宣誠 博士、阮雪芬 博士
Advisor: Hsuan-Cheng Huang, Ph.D., Hsueh-Fen Juan, Ph.D.

中華民國 111 年 12 月
December, 2022

doi:10.6342/NTU202210156

Abstract

Gene regulatory networks govern the complex gene expression programs in various biological phenomena, including cell development, cell fate decision, and oncogenesis. Single-cell techniques provide higher resolution in gene expression than traditional bulk RNA sequencing, but also incur more noise and sparser expression measurements, making it challenging to infer gene regulatory networks from such profiles. Inference of a complete gene regulatory network across different cell types is also difficult. Here, we propose to address the problem by constructing context-dependent gene regulatory networks (CDGRN) from single-cell RNA sequencing data. A gene regulatory network is decomposed into subgraphs that correspond to distinct transcriptomic contexts. Each subgraph is composed of the consensus active regulation pairs of transcription factors and their target genes shared by a group of cells. The activities of each regulation pair in different cell groups are inferred by a Gaussian mixture model using both the spliced and unspliced transcript expression levels. We find that the union of gene regulation pairs in all contexts provides sufficient information for the reconstruction of differentiation trajectories. CDGRN allows establishing the connection between gene regulation at the molecular level and cell differentiation at the macroscopic level. Functions specific to the cell cycle, cell differentiation, or tissue-specific functions are enriched throughout the developmental progression in each context. Surprisingly, we observe that the network entropy of CDGRN decreases with differentiation progression, implying directionality in differentiation. In conclusion, we leverage the advantage of single-cell RNA sequencing and establish a connection between molecular regulation and differentiation trajectory. Context-dependent network entropy may indicate the maturity of cells in certain contexts. The CDGRN model is available at `https://github.com/yuehhua/CDGRNs.jl`.

Keywords: *Gene Regulatory Networks*, *Unspliced RNA*, *Single-cell RNA Sequencing Data Analysis*, *Gaussian Mixture Model*, *Cell Trajectory*

doi:10.6342/NTU202210156

摘要

在多樣的生物現象中，基因調控網路掌控複雜的基因表現，包含細胞發育、決策細胞命運，以及癌化。單細胞定序技術，比起以往大批RNA定序，提供基因表現較高的解析度，但是同時測量到更多的雜訊，以及更稀疏的表現量，這讓基因調控網路的推論更加有挑戰性。跨不同細胞型態要推論完整的基因調控網路也是相當困難。這邊我們提出情境依賴基因調控網路（CDGRN），它可以從單細胞RNA定序資料來解決這個問題。基因調控網路可以被拆解成子圖，它對應到不同的轉錄情境。每個子圖是由共同活躍的調控配對組成，其中包含由一群細胞共享的轉錄因子，以及他們的目標基因。在不同細胞群體，每個調控配對的活性是由高斯混合模型推得，當中使用了剪切及未剪切轉錄的表現量。我們發現在所有情境下基因表現的聯集提供了足夠的資訊以建構細胞分化軌跡。CDGRN建立了分子層級基因調控與巨觀層級細胞分化之間的連結。在整個發育過程的各個情境中，細胞週期、細胞分化，或是組織特有功能有過度表現這些功能。更令人驚訝的是，我們發現CDGRN的網路亂度會隨著分化過程下降，這暗示了分化的方向。總結而言，我們利用了單細胞RNA定序技術的優勢，並建立了分子調控與分化軌跡之間的連結。情境依賴的網路亂度或許暗示了在特定情境下的細胞成熟度。CDGRN模型被釋出在https://github.com/yuehhua/CDGRNs.jl。

關鍵字: 基因調控網路、未剪切轉錄、單細胞轉錄定序資料分析、高斯混合模型、細胞軌跡。

doi:10.6342/NTU202210156

Contents

doi:10.6342/NTU202210156

List of Tables

doi:10.6342/NTU202210156

List of Figures

doi:10.6342/NTU202210156

Chapter 1

Introduction

Gene regulation plays a central role in cellular biology, governing complex gene expression and cellular functions. RNA sequencing techniques have been developed for measuring gene activity, and single-cell sequencing techniques are extending measurement resolution towards the single-cell level. A large number of applications [1, 2, 3, 4] for single-cell RNA sequencing (scRNA-seq) data analysis have been published. Integration of multi-omics single-cell data [5] can be achieved through multimodal integration. However, while gene regulation can be easily inferred from bulk RNA-seq data, this is more difficult using scRNA-seq data. In contrast, trajectory inference on cell differentiation progression can be made from scRNA-seq data but not from bulk RNA-seq data.

Trajectory inference (TI) analysis and pseudo-temporal ordering are frequent targets for single-cell techniques. The approach provides a macroscopic point of view of cell fate decision and developmental processes. Multiple algorithms have been proposed to address the problem of inferring developmental trajectories, including Monocle 3 [6], Palantir [7], Slingshot [8], STREAM [9], and PAGA [10]. TI tries to identify developmental trajectories from transcriptional states, but the developmental direction in the transcriptional landscape can usually not be derived in this manner [11]. RNA velocity models [12] have been proposed to give an indication of

doi:10.6342/NTU202210156

developmental direction. However, anomalous gene regulation patterns, including multiple kinetics and transcriptional boosting, can make this a difficult undertaking [13], and the inferred developmental direction is sometimes inconsistent with biological sense. Disturbances from a mixture of distinct gene regulation patterns are the main obstacle in inferring reliable trajectories and developmental directions from machine-learning models, leading to a loss of connection with underlying gene regulation.

The gene regulatory network (GRN) is situated at the microscopic level of cell differentiation and constitutes the underlying driving force of the system. GRNs have been modelled by various approaches, including differential equations (SCOUP [14], SCODE [15], GRISLI [16]), machine learning tree-based model (GENIE3 [17], GRNBoost2 [18], SCENIC [19, 20]), deep learning (SIGNET [21], BiRGRN [22]), information measures (PIDC [23]), causation (Scribe [24], SINGE [25]), and statistics (PPCOR [26], GRNVBEM [27], LEAP [28]). Traditionally, the differential equations approach has been applied for reconstruction of GRNs in terms of dynamical systems theory. Alternatively, random forest (GENIE3 [17]) and gradient boost tree (GRNBoost2 [18]) are proposed to infer GRN from scRNA-seq data and they enjoy the several advantages, including adapting to directed, nonlinear relationship, high accuracy, not requiring time labels, and allowing feedback loops. Many approaches, such as SCODE, GRISLI, BIRGRN, GRNVBEM, etc., require time labels or pseudotime for inferring GRNs from scRNA-seq data. Despite time information is provided for GRN inference, it is reported that algorithms not requiring time labels, such as GENIE3 and GRNBoost2, pose higher accuracy on predicting regulation relationships [29]. Hence, modelling realistic GRNs not requiring time information from high-dimensional data remains an open issue [30], and owing to the inherent noise and sparseness of the data, it is still challenging to reconstruct a full GRN from scRNA-seq data, especially when multiple cell types are involved.

TI algorithms are regarded as a separate avenue of investigation and have seen

doi:10.6342/NTU202210156

independent development. Currently, TI algorithms generally provide less connection between developmental trajectories and gene regulation. TI and GRN reconstruction algorithms have tended to be conducted independently, followed by later compilation and interpretation. This approach is unsystematic and underlines the crucial need for an integrated explanation of the connection between developmental trajectories and GRNs, which would be required for a consistent macroscopic and microscopic interpretation based on the same model and dataset.

We here propose a context-dependent gene regulatory network (CDGRN) to simultaneously identify GRNs and visualize developmental trajectories in certain contexts. It allows integrated explanation of molecular mechanisms and corresponding developmental trajectories. To address the issue of mixed regulation patterns, the idea of decomposing mixture patterns is applied to identify components of gene regulation patterns for each regulation pair. Since the identified regulation patterns exhibit cellular behaviors and dynamics in certain transcriptional contexts, patterns can be used to identify classes of contexts and assign cells to these. GRNs can then be inferred from cells with homogeneous transcriptional profiles for certain transcriptional contexts, and the developmental trajectories can be visualized from the same set of profiles.

doi:10.6342/NTU202210156

Chapter 2

Materials and Methods

2.1 Preprocessing datasets

Pancreatic, dentate gyrus, mouse gastrulation, and human bone marrow datasets were imported from scVelo package [31]. Data were preprocessed following the default scVelo pipeline. For each gene, both spliced and unspliced count matrices were filtered by a minimum count of 20. Spliced and unspliced count matrices were normalized for each cell by total counts over all genes. The 5000 most highly variable genes were preserved, and matrices were log-transformed by $\log(1 + x)$. Principal component analysis (PCA) was applied to reduce the dimensions to 30 principal components (PC). Neighbor graphs were established with 30 nearest neighbors by Euclidean distance in PCA space, and used to compute first-/second-order moments for each cell over its nearest neighbors. First-order moments of spliced and unspliced matrices were used in the downstream modeling.

2.2 RNA velocity and latent time inference

We followed the default scVelo pipeline for generalized RNA velocity inference. The RNA velocity and velocity graph were computed first. A dynamical model was fitted by calling `scv.tl.recover_dynamics` to derive the latent time, which was

doi:10.6342/NTU202210156

then fetched for each dataset. Genes with an RNA velocity model likelihood higher than 0.1 were selected.

2.3 Selection of regulation pairs

Transcription factor (TF)–target gene pairs were compiled from a transcription factor binding site (TFBS) list downloaded from the FANTOM5 data portal: `https://fantom.gsc.riken.jp/5/datafiles/phase1.3/`. Genes were mapped to corresponding HGNC id's, and those that were successfully mapped were retained. Regulations between TFs and their target genes were modeled with GMM models; details are described in Section 2.4. After regulation pairs were selected, they were mapped to the CHEA database [32] and regulation pairs present in the database were retained. The selected regulation pairs were then used in downstream CDGRN modeling and analysis.

2.4 Context-dependent gene regulatory network

The process for establishing a context-dependent gene regulatory network can be divided into three stages. First, a single regulatory pattern should be identified from a mixture of regulation patterns. This requires identifying contextual regulation patterns from mixed regulation patterns for the whole dataset. Second, transcriptional contexts are identified from the profile of contextual regulation patterns for each cell. Third and finally, gene regulatory networks are inferred for each context and developmental trajectories are visualized. In the first stage, a GMM model is used to model the mixture of regulation patterns, and a single component can be extracted as the contextual regulation pattern for each pair of TF and its target gene. Any regulation relationship can be described by the expression of TF x_i and its specific target gene y_i for each cell i. Assuming that there are K distinct kinds of components involved in a regulation relationship for a certain TF and target gene

doi:10.6342/NTU202210156

pair, then, for each component, the distribution is determined by their mean vector μ_k and covariance matrix Σ_k for the specific k-th component:

$$P(x_i, y_i \mid \mu_k, \Sigma_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i, y_i \mid \mu_k, \Sigma_k) \tag{2.1}$$

The spliced mRNA expression for the TF gene and the unspliced mRNA expression for the target gene are used to train the GMM model. A GMM is trained across all observations and GMM clusters are identified as context-dependent motifs for each regulation pair. Clusters can be identified by calculating a hard clustering from the posterior probability:

$$z_i^* = \arg\max_k \frac{P(z_i = k \mid \theta) P(x_i, y_i \mid z_i = k, \theta)}{\sum_{k'=1}^{K} P(z_i = k' \mid \theta) P(x_i, y_i \mid z_i = k', \theta)} \tag{2.2}$$

where θ is the set of μ_k and Σ_k for all $k \in [1, K]$. The hyperparameter K denotes number of components for GMM and it corresponds to number of regulatory patterns in a regulatory pair. Empirically, we observed that number of potential components in a regulatory pair often falls below 5. Therefore, it is selected from model selection ranging from 1 to 5. The GMM model with the lowest Akaike information criterion (AIC) score is selected. The AIC score is calculated as

$$aic = 2\omega - 2 \ln \mathcal{L}^* \tag{2.3}$$

where ω denotes the number of parameters estimated from the GMM model and $\mathcal{L}^*$ is the model's maximum likelihood value. If the best GMM model contains only a single component ($k = 1$), then the corresponding TF and target gene pair are considered unregulated. Selected TFs and their target gene sets, as well as the corresponding contexts, are then used in downstream modeling.

In the second stage, to identify contexts, the profile of contextual regulation patterns for each cell is collected for the whole dataset. The profile can be expressed as an observation-motif matrix. An observation-motif matrix is filled with GMM

doi:10.6342/NTU202210156

clusters for each regulation pair in columns and observations in rows, then used as a feature matrix. Contexts are identified as cell clusters by using hierarchical clustering (based on cluster dissimilarity) with Ward linkage over all context-dependent motifs to calculate the Hamming distance. Context-dependent motifs can be regard as distinct entities, with the distance between observations equal to the Hamming distance. The variance of cluster dissimilarity is considered and Ward linkage is used to minimize increase in total within-cluster variance after merging two clusters. Ward linkage uses the objective function of minimizing the sum of square errors to optimize clustering. The initial cluster distances are defined as:

$$d_{ij} = d(\{X_i\}, \{X_j\}) = ||X_i - X_j||^2 \tag{2.4}$$

where the cluster dissimilarity d_{ij} between clusters i and j is defined as the distance between two singleton clusters $\{X_i\}$ and $\{X_j\}$. Cluster dissimilarity $d_{(ij)k}$ can then be calculated after merging clusters. For distinct clusters C_i, C_j, and C_k with sizes n_i, n_j, and n_k, respectively:

$$\begin{aligned} d_{(ij)k} &= d(C_i \cup C_j, C_k) \\ &= \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j). \end{aligned} \tag{2.5}$$

After clustering, contexts can be extracted by dividing the dendrogram into distinct clusters C_i at defined distances. Observations are dissected into several contexts.

In the final stage, CDGRNs can be inferred by using a multiple regression model for each context of each regulatory pair. In each context, a gene expression profile with selected spliced and unspliced mRNA levels against corresponding observations in the context is used to train the model. For each regulatory pair, a multiple regression model for a target gene and its upstream TFs is trained on the corresponding gene expression profile for a given context. Thus, a set of multiple regression models forms a context-dependent gene regulatory network for that context. The regula-

doi:10.6342/NTU202210156

tion relationship can be determined by the correlation between TFs and their target genes in the context.

2.5 Data visualization for trajectory

After removing uncorrelated regulatory pairs, TFs and their target gene expression profiles from spliced and unspliced transcripts are merged into a feature matrix. The feature matrix is then reduced to the top five dimensions by PCA, and trajectories from selected dimensions are plotted in 2D or 3D space.

2.6 Network visualization

Visualization of regulatory networks is done using Cytoscape v3.9.1. The network is visualized by coloring edges by correlation sign and sizing edges by correlation strength. The correlation strength and sign for selected TF–target gene pairs in each context are written to CSV files.

2.7 Statistical methods

To compare contextual regulation patterns to global regulation patterns, the absolute value of correlation for each TF–target gene pair in a given context was computed. To resolve the difference in sample size between contextual regulation patterns to global regulation patterns, the dataset was randomly sub-sampled at the sample size of the contextual regulation pattern. A two-sample Wilcoxon rank-sum test and a Kolmogorov–Smirnov test were applied to sub-sampled data using HypothesisTests.jl.

doi:10.6342/NTU202210156

2.8 Functional enrichment analysis

ConsensusPathDB [33] (`http://cpdb.molgen.mpg.de/`) was used for the functional enrichment analysis. To investigate the biological processes activated in different CDGRNs, we excluded low correlation (< 0.3) TF–target gene regulation pairs. Gene sets were compiled from each CDGRN and uploaded to the website to query all significant gene ontology (GO) terms from levels 3 to 5.

doi:10.6342/NTU202210156

Chapter 3

Results

3.1 Unspliced mRNA reveals regulatory patterns in TF-target gene pairs

Unspliced mRNAs can be derived by calling from single-cell RNA sequencing (scRNA-seq) data [12]. As a central concept of molecular biology, mRNA is transcribed and spliced by the spliceosome in eukaryotic cells. Mature mRNAs then undergo translation, and regulation is applied to TFs binding to the promoter region of a target gene. For this reason, the spliced mRNA level of a TF gene should be in a regulatory relationship with the unspliced mRNA level from its target genes. To demonstrate the spliced and unspliced mRNA levels reveal such a regulatory relationship and form a regulation pair, we investigated this relationship, which at least should then constitute a stronger correlation than the relationships of spliced mRNA levels to target genes. To this end, regulation gene pairs were selected from a ChIP-X experiments CHEA database [32]; a pancreatic dataset from embryonic mice including cell fate commitment to four kinds of pancreas islet cells was used. We calculated gene regulatory connections between TFs and their target genes using unspliced mRNA levels and compared these to the same metric using spliced mRNA levels (Fig. 3.1). Surprisingly, we found that, unlike the case for spliced mRNA levels and

doi:10.6342/NTU202210156

all genes, gene regulations were stronger and there were more robust correlations between spliced mRNA levels for TF genes and unspliced mRNA levels for their target genes. We further found that the dataset contained different cell types at unbalanced proportions. As distinct cell types show different transcriptional behaviors, the dataset mixed multiple regulation patterns from distinct cell types for each pair of gene regulations. To address the issue of multiple regulations, components were decomposed from specific gene regulation pairs by a Gaussian mixture model (GMM), with each component representing a specific regulation pattern in a given context. This allowed the definition of a transcriptional context for each regulation dynamic.

3.2 Context-dependent gene regulatory network

We propose a context-dependent gene regulatory network (Fig. 3.2a) that decomposes cell transcriptional states at the molecular level for different contexts. Cells carry out their diverse functions, or stay in phase in the cell cycle, because they remain in distinct contexts. Gene regulations govern complex cellular functions, and regulations change if the context changes. This context could be distinct cell types, cell transition dynamics, or even cell transcriptional states. It is usually determined by upstream gene regulation of TFs and their target genes. We therefore modeled the gene regulation for distinct patterns and constructed contexts based on the combination of distinct regulation patterns (Fig. 3.2b).

To infer regulation effects more directly, target gene expressions could be measured from unspliced transcript levels. A GRN is inferred from TF and their target genes using spliced and unspliced transcript levels (Fig. 3.3). Regulation between TF and target genes is then inferred from correlation. However, gene regulations are extracted from datasets containing mixture pattern made up of many different cell types, which impedes reliable inferral of regulations from scRNA-seq data. A GMM

doi:10.6342/NTU202210156

is thus used to identify and isolate distinct components for each TF–target gene pair, which represent single regulation tendencies. Context can then be identified from the combination of regulations across TFs and their target genes; cells sharing a similar combination of regulations should be in similar distinct contexts, which in turn are identified by cluster analysis. Cells in the same cluster are considered to be in the same context, and are used to infer CDGRNs for that context.

We constructed CDGRNs (Fig. 3.4) for four independent real datasets. First, a pancreatic dataset describing embryonic mouse pancreas cell fate commitment to alpha, beta, delta, and epsilon cell lineages was used. A total of 3,696 cells with 27,998 expressed genes were fetched from scVelo and preprocessed. The 5,000 most highly variable genes were selected, for which 11,610 TF–target gene pairs were identified by GMM model. Of these models, 8,734 corresponded to TF–target gene pairs showing two or more components in their regulation patterns. The remaining single-component pairs were discarded. To further remove spurious regulation pairs, the selected TF–target gene pairs were mapped to the CHEA database [32], which collects experimentally curated transcriptional factor binding site profiles and their targets, and contains 199 TFs and 21,585 target genes (198 TFs included), thus forming 386,776 pairs. Mapping yielded a match for 830 TF–target gene pairs consisting of six TF genes and 609 target genes, corresponding to a total of 2,270 components.

Second, a dentate gyrus neurogenesis dataset was fetched from scVelo and analyzed. It contained 2,930 cells with 13,913 expressed genes. A total of 2,688 TF–target gene pairs were identified by GMM modeling of the 5,000 most highly variable genes, yielding 1,063 models corresponding to multiple-component TF–target gene pairs. After mapping to the CHEA database, 371 TF–target gene pairs consisting of four TF genes and 259 target genes were retained.

Third, a dataset describing mouse gastrulation to erythroid lineages was analyzed, which contained 9,815 cells and 53,801 expressed genes. A total of 1,208

doi:10.6342/NTU202210156

TF–target gene pairs were identified, which included 785 multiple-component TF–target gene pairs. After mapping to the CHEA database, 268 TF–target gene pairs consisting of four TF genes and 141 target genes were retained.

Fourth, a human bone marrow dataset describing the haematopoiesis process in bone marrow and consisting of 5,780 cells with 14,319 expressed genes was used. A total of 8,727 TF–target gene pairs were identified, which included 7,643 multiple-component TF–target gene pairs. After mapping to the CHEA database, 893 TF–target gene pairs consisting of eight TF genes and 461 target genes were retained.

3.3 Extracting contextual regulation pattern as a single component from global mixture regulations

To investigate a single component of a regulation pattern in a given context, cells in that context are extracted from the complete dataset (Fig. 3.5a). Extracted cells share the same single regulation component, which corresponds to a component in the respective GMM model (Figures 3.5de, 3.6). The single contextual regulation pattern represents a shared dynamic of gene regulations, e.g., the estimated relationship between TF gene expression and target gene expression. To verify that the contextual regulation pattern provides a simple and more robust descriptor of regulation than global regulation, we tested correlation strengths for cells in a given context against all cells across all regulation pairs. To this end, an equal number of global regulation patterns was matched to TF–target gene pairs and correlation strength was calculated. We found that in the pancreatic dataset, contextual regulation patterns yielded higher correlation strength than global regulation patterns (Fig. 3.5b, p value $< 10^{-32}$; Wilcoxon rank-sum test), and that the empirical cumu-

doi:10.6342/NTU202210156

lative distribution functions of correlation differed significantly between these levels (Fig. 3.5c, p value $< 10^{-7}$; Kolmogorov–Smirnov test). This suggests that dissecting global mixture regulations into several components is a suitable approach to create a simple and robust basis for analysis.

3.4 Explaining differentiation trajectory from regulatory pairs

TI algorithms are typically developed independently to GRN algorithms, and the connection between macroscopic or cellular phenomena and microscopic or molecular mechanisms remains unclear in most analyses. To explain the connection between differentiation progression and gene regulation, we used gene expression profiles of previous selected TF–target gene pairs derived from spliced and unspliced mRNA levels to visualize differentiation trajectories (Fig. 3.3). We found that these trajectories are suitable for determining cell differentiation progression and describe useful cell types well in eigenspaces.

In the pancreatic dataset, ductal cells (Fig. 3.7a) exhibited DNA replication and mitosis in the five highest-ranked enriched Gene Ontology (GO) terms (Table 3.1). *Ngn3*-low EP cells derived from the trunk domain [34] progressed towards pre-endocrine cells, showing chromosome condensation in context 4 (Fig. 3.7b) and gland morphogenesis and development in context 5 (Fig. 3.7b, Tables 3.1, 3.2). Cells committed to terminal alpha, beta, epsilon, and delta cells and progressed to endocrine system development in context 1 (Fig. 3.7b, Tables 3.1, 3.2).

In the dentate gyrus dataset, a neurogenesis trajectory was revealed (Fig. 3.8a) from nIPC and neuroblasts to granule (mature) cells. Initial nIPC and neuroblasts corresponding to context 3 (Fig. 3.8b) changed to partial neuroblasts at the turning corner. Granule maturation can be observed in context 1 (Fig. 3.8b). Immature cells were aligned along the trajectory and mature cells terminated at the end of context

doi:10.6342/NTU202210156

1 (Fig. 3.8c). The five highest-ranked enriched GO terms in contexts 1 and 3 only partially consisted of terms related to nervous system development (Table 3.3).

The mouse gastrulation dataset demonstrated that blood progenitors differentiate into erythroid cells. The trajectory indicates that differentiation progressed significantly from blood progenitors to erythroid cells (Fig. 3.9a). The five highest-ranked enriched GO terms for blood progenitors 1 and 2 in context 1 (Fig. 3.9c) contained terms like blood vessel morphogenesis and development, cardiovascular system development, and circulatory system development (Tables 3.5, 3.6). For erythroid 1, these terms indicate that myeloid leukocyte activation and differentiation occurred in context 2. Myeloid leukocytes may undergo further cell migration. Myeloid cell differentiation remains active until the erythroid 2 and 3 stages (context 3) (Fig. 3.9c). In context 3, the regulation for systematic anatomical structure morphogenesis and fine-grained cellular component organization takes place. This dataset demonstrates that cells in different contexts shift progressively from coarse, early-stage to detailed, late-stage cellular functions.

The differentiation landscape of human hematopoiesis in bone marrow showed a progression from human stem cells to erythrocytes, dendritic cells, monocytes, and megakaryocytes (Fig. 3.9b). In the monocyte lineage, cells originating from stem cells HSC_1 and HSC_2 (context 2 and partially context 7) were enriched in the regulation of hemopoiesis and hematopoietic or lymphoid organ development (Table 3.7). In contrast to context 2, cells in context 7 were further enriched in leukocyte differentiation (Fig. 3.9d). Cells in context 6 covered most precursors, and Mono_1 monocytes were active in the regulation of immune system processes. Meanwhile, Mono_2 monocytes in context 3 showed enrichment unrelated to monocytes or the immune system. Data from the Reactome database indicates that contexts 2, 6, and 7 all were involved in the regulation of granulopoiesis (Table 3.8).

Differentiation progression also aligned well with latent time, which was inferred from generalized RNA velocity using a dynamical model. Trajectories from the

doi:10.6342/NTU202210156

pancreatic dataset were biologically consistent with latent times (Fig. 3.7c). Note that, while in gastrulation to erythroids, latent time was reversed in the model (Fig. 3.10); this was still consistent with inverse trajectories.

3.5 Revealing regulation network dynamics by progression of contexts

A set of contextual regulation patterns describes cellular behavior at the molecular regulation level. Contexts describe cellular regulation states and can be identified by clustering cells against contextual regulation patterns. Cells in similar contexts tend to have similar regulatory networks. To investigate the underlying regulatory network in a given context, the regulatory network is inferred from the relevant cells. Contexts are inferred by applying hierarchical clustering against the contextual regulation patterns, and CDGRNs are in turn inferred by calculating the correlation between spliced mRNA levels for TF and unspliced mRNA levels for target genes in each context (Fig. 3.11). Each context then corresponds to its underlying CDGRN. Regulation pairs with high correlation strength (e.g., > 0.3) are then selected from each CDGRN, and genes involved in these pairs are collected as a gene set for enrichment. The dynamics of the underlying gene regulatory network can be explained by rewiring gene regulations from one CDGRN to another.

In the pancreatic dataset, we inferred five contexts. Ductal cells and a very small portion of Ngn3 with low EP underwent DNA replication in context 3 (Fig. 3.11a). *POLA1*, *CCNE2*, and *CDT1* genes, which are polymerases and key factors involved in DNA synthesis, were positively regulated by the *E2F1* gene. Another portion of ductal and low-EP Ngn3 cells played roles in spindle localization and microtubule organization in the M phase, in which *NUSAP1* genes are involved and are regulated by *PAX6*, *PDX1*, and *E2F1* genes (Fig. 3.11b). In early- to middle-stage high-EP Ngn3 cells, complicated regulation processes occurred simultaneously in context 4, includ-

doi:10.6342/NTU202210156

ing cell cycle regulation and regulation of hormone levels and metabolic processes (Fig. 3.11c). Peptide hormone processing, transport, and regulation were enriched in context 4, including *PCSK1*, *HADH*, *CPE*, *NR3C1*, *PDX1*, and *SNAP25* genes. Surprisingly, late-stage high-EP Ngn3 cells switched their behaviors in context 5 (Fig. 3.11d). In addition to the behaviors observed in context 4, gland morphogenesis, cell proliferation, and multicellular organ development processes were enriched in context 5, involving other, more complicated groups of genes. Finally, cell development went through pre-endocrine stages and terminated in four types of islet cells in context 1 (Fig. 3.11e). Unexpectedly, these cell types shared similar CDGRNs for context 1, and the remaining regulations were relatively simple. *PDX1*, *NR3C1*, and *PAX6* genes were involved in gland development and islet cell functions, including regulation of hormone levels and responses to nutrient and fatty acid levels. The developed CDGRNs allowed explanation of cellular behaviors in each context in terms of gene regulations and functional enrichment analysis, and provided insights into sub-population behaviors within a given cell type.

3.6 Shrinkage of regulation network size shrinks during cell differentiation process

We also discovered that the size of CDGRNs decreased as cell differentiation progressed. Network entropy as a measure of regulation network complexity declined gradually in parallel with cell maturation (Fig. 3.12). This may indicate that the activity of a regulation network simplifies during maturation. More detailed context dissections showed that this decline fluctuated to some degree. During phases of rising network entropy, cells progressed from one stable cell type to another, and entropy declined again when the next stable type was reached. In other words, network entropy indicated not only network complexity but also the stability of transcriptional states. Evaluated over a longer time frame, the network entropy of

doi:10.6342/NTU202210156

a CDGRN may be an indicator of cell maturity, and differences in network entropy may imply varying differentiation directions.

doi:10.6342/NTU202210156

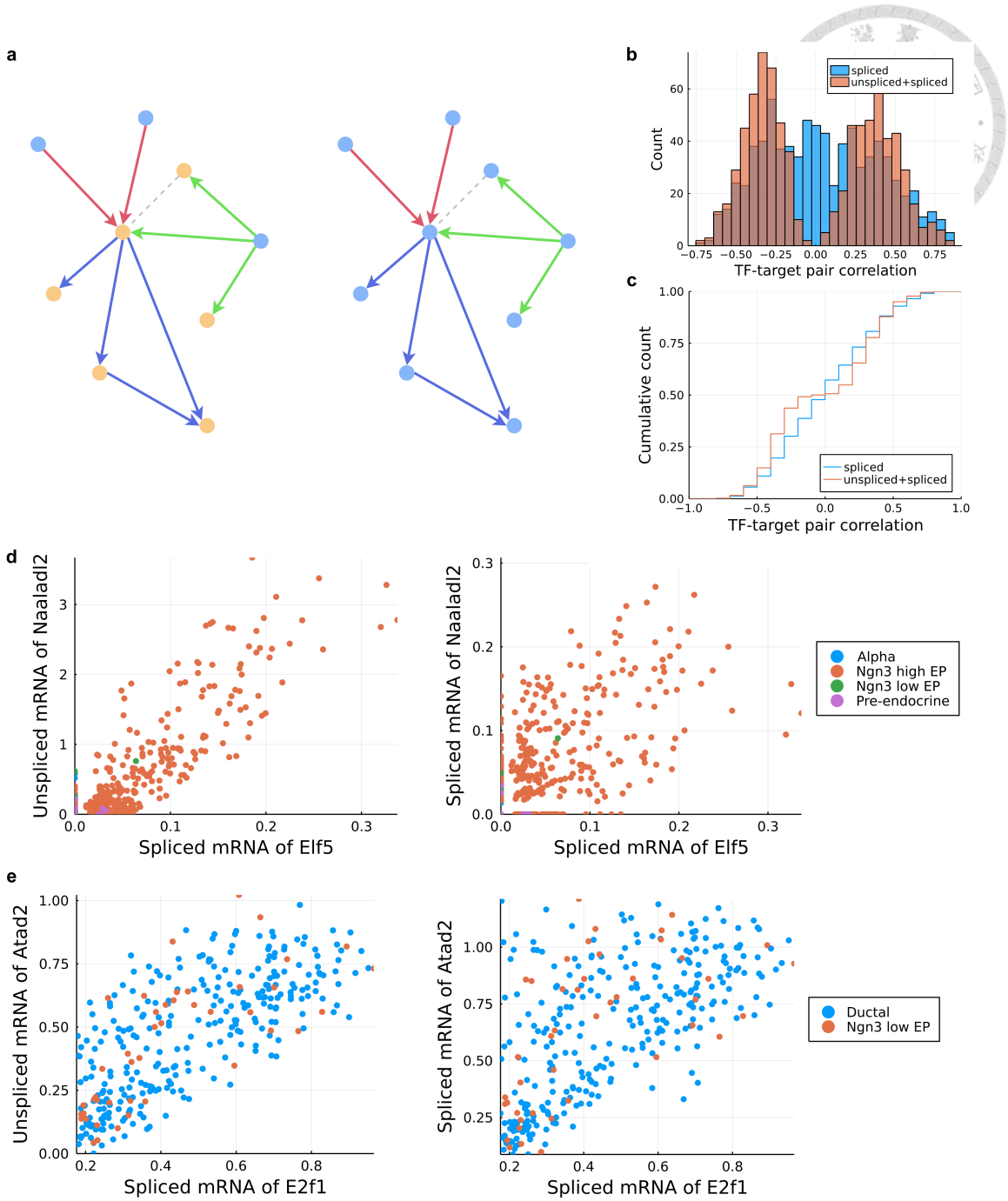

Figure 3.1: Comparison of regulatory network inference from unspliced and spliced mRNA levels. (**a**) Gene regulation between TFs and their target genes. Blue points represent spliced mRNA, yellow points represent unspliced mRNA. Two scenarios are compared. (**b**) Histogram of correlations from different mRNA levels in the pancreatic dataset. (**c**) Empirical cumulative distribution function of correlations from different mRNA levels in the pancreatic dataset. (**d**) Gene regulation pattern between *Naaladl2* and *Elf5* from unspliced (left) and spliced (right) mRNA levels in a given context. (**e**) Gene regulation pattern between *Atad2* and *E2f1* from unspliced (left) and spliced (right) mRNA levels in a given context.

doi:10.6342/NTU202210156

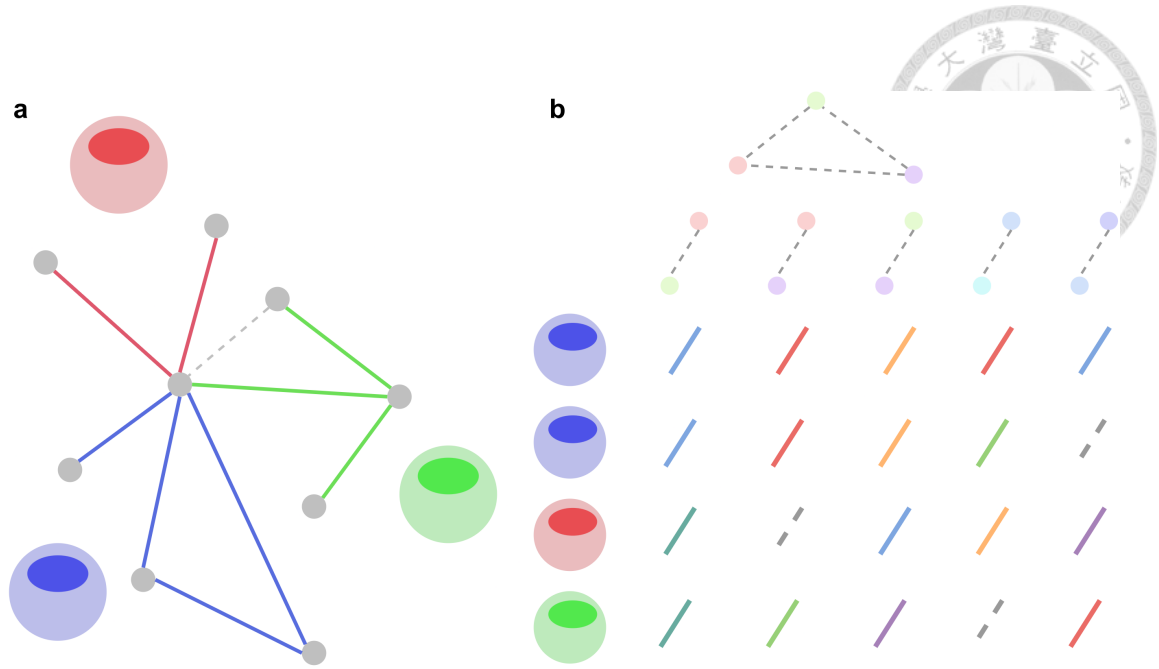

Figure 3.2: Context-dependent gene regulatory networks. (**a**) A CDGRN is derived by decomposing a GRN into several sub-networks for distinct contexts. (**b**) Regulation patterns are used to cluster cellular contexts.

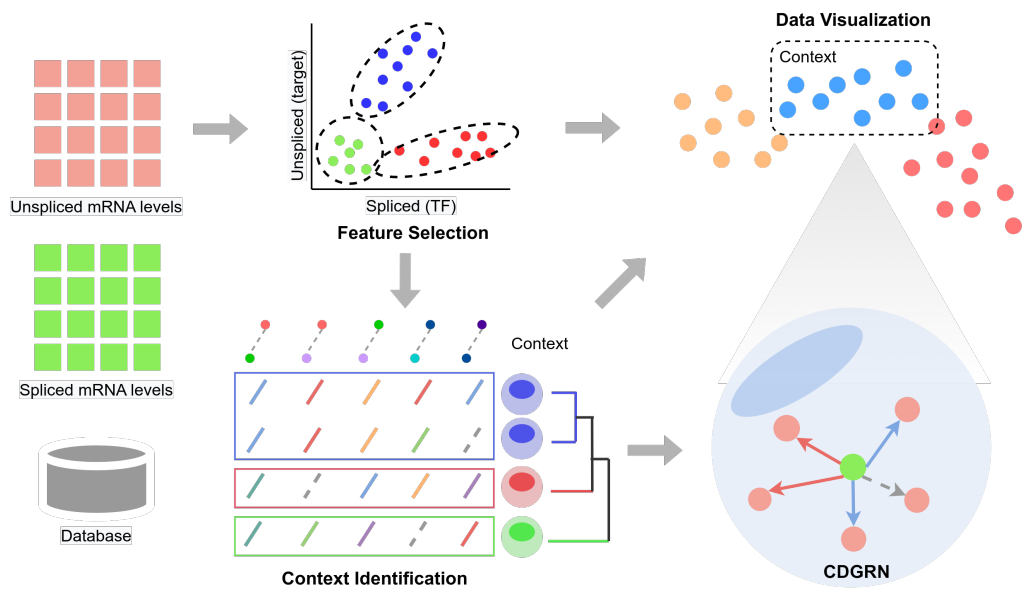

Figure 3.3: An overview of the CDGRN framework. Input of (un)spliced transcripts are used for GMM feature selection for significant regulatory patterns. Contexts are then identified from regulatory pattern profiles by clustering. Each GRN can be inferred for each context and developmental trajectory can also be inferred from selected gene expression profiles.

doi:10.6342/NTU202210156

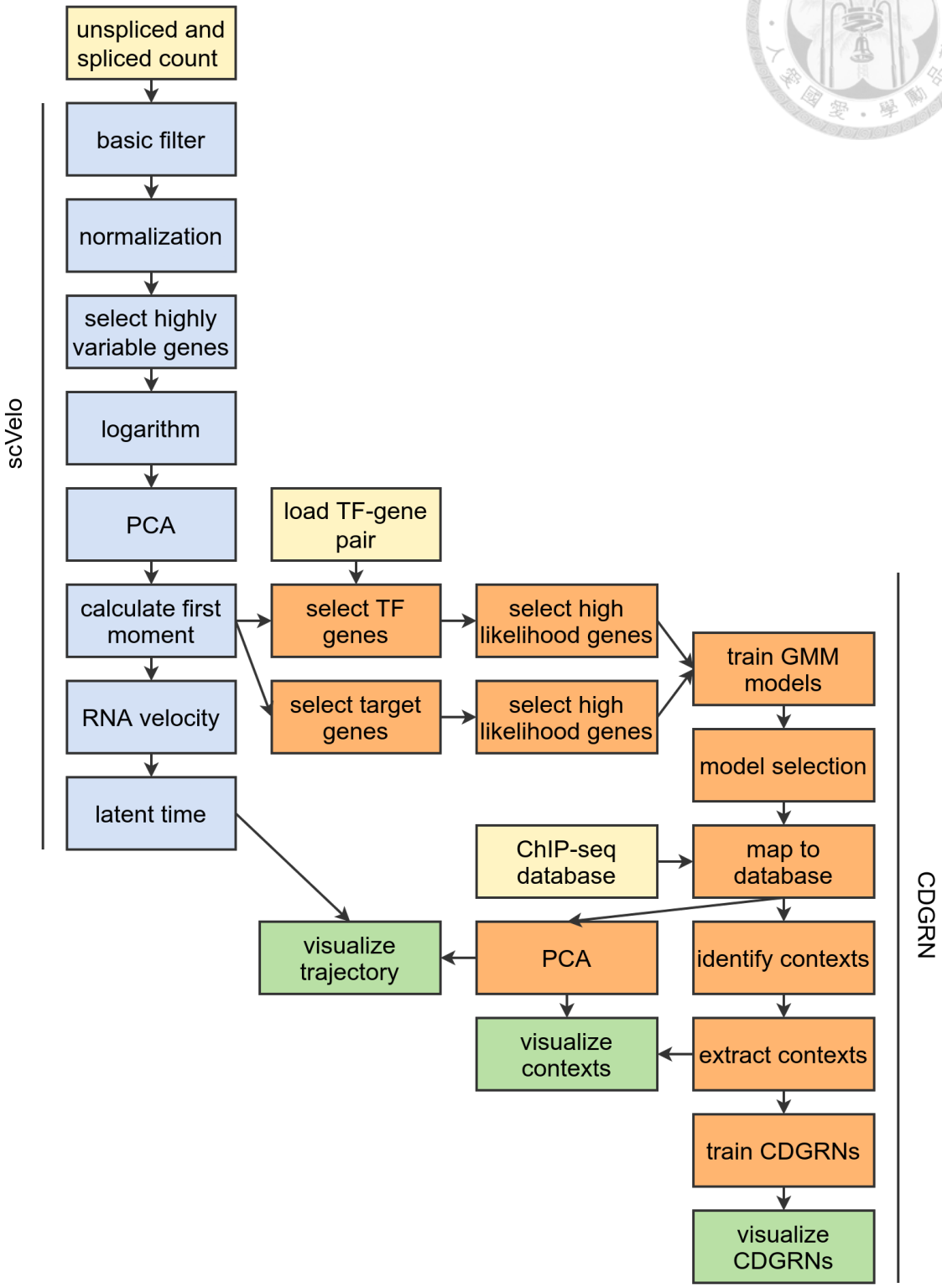

Figure 3.4: The detailed workflow of CDGRN.

doi:10.6342/NTU202210156

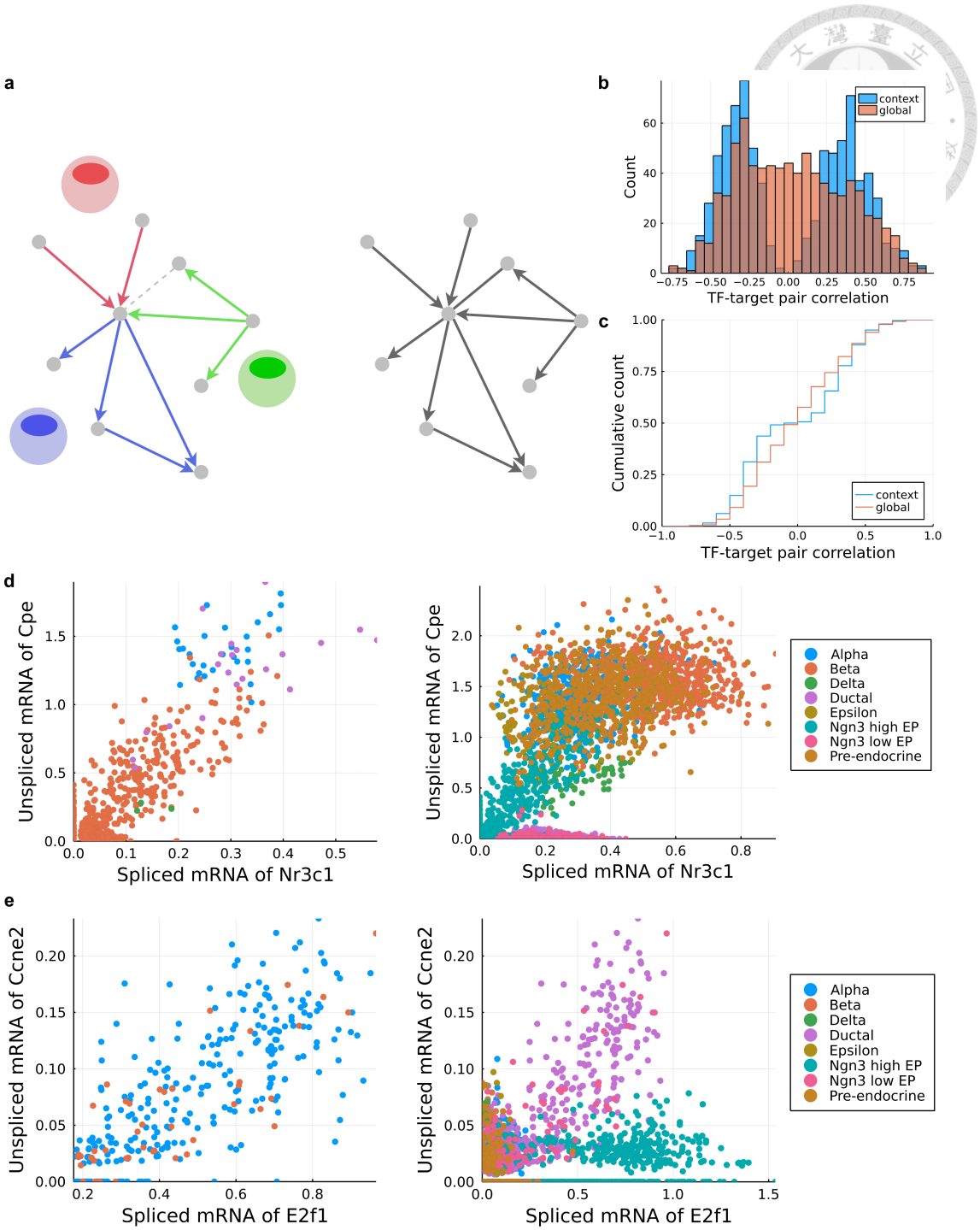

Figure 3.5: Comparison of regulatory network inference from distinct and global contexts. (**a**) Comparison of cellular contexts with global context. (**b**) Histogram of correlations between distinct and global contexts in the pancreatic dataset. (**c**) Empirical cumulative distribution function of correlations between distinct and global contexts in the pancreatic dataset. (**d**) Gene regulation patterns between *Nr3c1* and *Cpe* in distinct (left) and global (right) contexts. (**e**) Gene regulation patterns between *E2f1* and *Ccne2* in distinct (left) and global (right) contexts.

doi:10.6342/NTU202210156

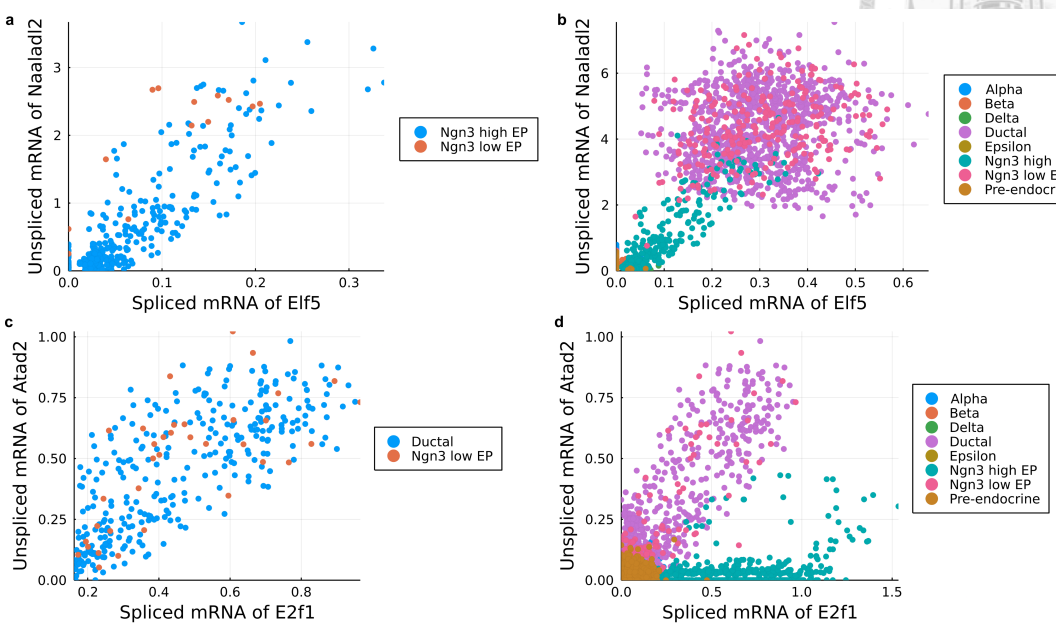

Figure 3.6: Cases of gene regulations for regulatory network inference for specific and global context. The gene regulation pattern between *Elf5* and *Naaladl2* in **a**, specific and **b**, global context. The gene regulation pattern between *Atad2* and *E2f1* in **c**, specific and **d**, global context.

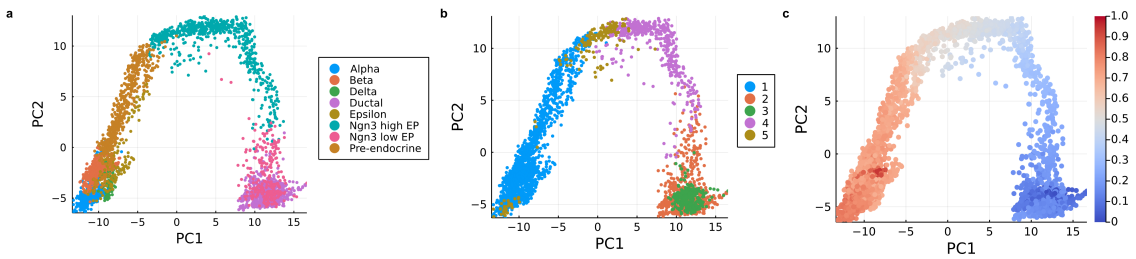

Figure 3.7: Inference and visualization of landscape for the pancreatic dataset. (**a**) Developmental trajectory visualized after GMM feature selection for the pancreatic dataset. (**b**) Distinct contexts are identified, revealing regulation dynamics in the developmental trajectory. (**c**) The developmental trajectory aligns well with latent time inferred from the generalized RNA velocity model.

doi:10.6342/NTU202210156

Table 3.1: Ten highest-ranked enriched GO terms for pancreatic islet cell development.

Context	**Level 4 biological process**	**q-value**
1	gland development	0.001092
	endocrine system development	0.001092
	gland morphogenesis	0.001092
	cellular response to oxygen-containing compound	0.001092
	cellular response to oxygen levels	0.001092
	regulation of cell proliferation	0.002197
	regulation of cell death	0.002197
	cellular response to nutrient levels	0.002952
	hexose metabolic process	0.002952
	response to fatty acid	0.002952
2	establishment of spindle localization	0.000811
	regulation of DNA binding	0.003040
	microtubule cytoskeleton organization involved in mitosis	0.003040
	cellular response to organic substance	0.053478
	brain development	0.058915
3	DNA metabolic process	8.92×10^{-6}
	macromolecule biosynthetic process	8.92×10^{-6}
	cellular macromolecule biosynthetic process	0.000012
	nuclear DNA replication	0.000123
	nucleic acid metabolic process	0.000123
	mitotic DNA replication	0.000123
	chromosome organization	0.000149
	heterocycle biosynthetic process	0.000696
	aromatic compound biosynthetic process	0.000696
	negative regulation of cellular process	0.001006
4	positive regulation of metabolic process	0.001059
	positive regulation of cellular process	0.001738
	mitotic chromosome condensation	0.002228
	hormone transport	0.007280
	regulation of cellular metabolic process	0.009952
	meiotic chromosome condensation	0.009952
	mitotic sister chromatid segregation	0.011022
	regulation of nitrogen compound metabolic process	0.011622
	macromolecule biosynthetic process	0.013475
	hormone secretion	0.013642
5	gland morphogenesis	0.000012
	gland development	0.000023
	nervous system development	0.000023
	regulation of cell proliferation	0.000024
	cell projection morphogenesis	0.000027
	neuron development	0.000037
	plasma membrane bounded cell projection organization	0.000037
	cell part morphogenesis	0.000037
	neurogenesis	0.000037
	axon guidance	0.000081

doi:10.6342/NTU202210156

Table 3.2: Ten highest-ranked enriched pathway terms for pancreatic islet cell development.

Context	**Pathway terms**	**Source**	**q-value**
1	Reelin signalling pathway	KEGG	0.000290
	SUMOylation of intracellular receptors	Reactome	0.005005
	Maturity onset diabetes of the young	KEGG	0.005005
	Nuclear Receptor transcription pathway	Reactome	0.012816
	Chemical carcinogenesis - receptor activation	KEGG	0.018803
2	(none)		
3	DNA Replication	Reactome	7.32×10^{-10}
	Synthesis of DNA	Reactome	9.19×10^{-9}
	Mitotic G1 phase and G1/S transition	Reactome	2.57×10^{-8}
	Activation of the pre-replicative complex	Reactome	2.94×10^{-8}
	S Phase	Reactome	3.67×10^{-8}
	G1/S Transition	Reactome	3.67×10^{-8}
	DNA replication - Mus musculus	KEGG	3.67×10^{-8}
	DNA Replication Pre-Initiation	Reactome	3.84×10^{-8}
	Cell Cycle, Mitotic	Reactome	6.03×10^{-6}
	Lagging Strand Synthesis	Reactome	6.03×10^{-6}
4	Cell Cycle, Mitotic	Reactome	0.000303
	Cell Cycle	Reactome	0.000698
	M Phase	Reactome	0.003657
	Thyroid hormone signaling pathway	KEGG	0.004907
	Insulin secretion	KEGG	0.013079
	Mitotic Prometaphase	Reactome	0.022410
	Carbohydrate digestion and absorption - Mus musculus	KEGG	0.022490
	Growth hormone synthesis, secretion and action - Mus musculus	KEGG	0.023886
	Mitotic Prophase	Reactome	0.023886
	GnRH secretion - Mus musculus	KEGG	0.034641
5	Prostate cancer	KEGG	0.006888
	DNA Replication	Reactome	0.059728
	Cocaine addiction	KEGG	0.059728
	Insulin secretion	KEGG	0.059728
	SUMOylation of intracellular receptors	Reactome	0.059728
	Small cell lung cancer - Mus musculus	KEGG	0.059728
	Maturity onset diabetes of the young - Mus musculus	KEGG	0.062375
	Activation of the pre-replicative complex	Reactome	0.078607
	Amphetamine addiction - Mus musculus	KEGG	0.078607
	DNA replication - Mus musculus	KEGG	0.078607

doi:10.6342/NTU202210156

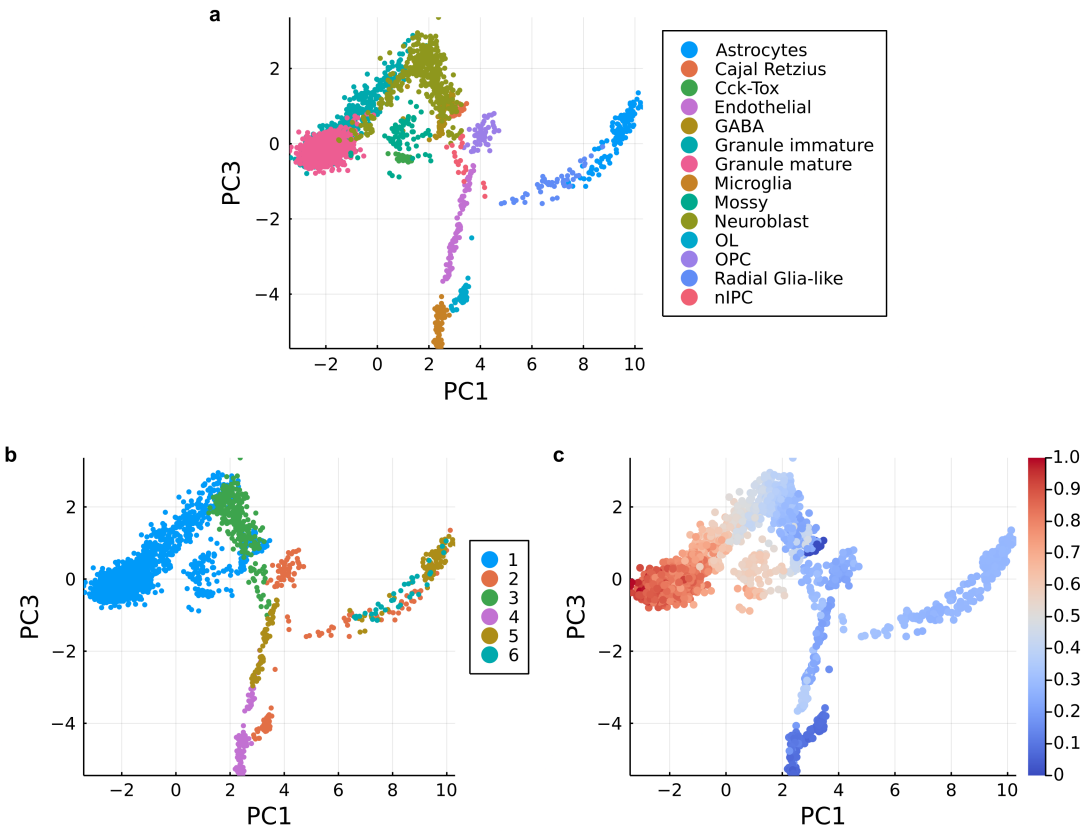

Figure 3.8: Inference and visualization of CDGRNs for dentate gyrus dataset. **a**, Developmental trajectory visualized from CDGRN for dentate gyrus dataset. **b**, Distinct contexts are clustered and reveals regulation dynamics in developmental trajectory. **c**, Developmental trajectory aligns well with latent time inferred from generalized RNA velocity model.

doi:10.6342/NTU202210156

Table 3.3: Ten highest-ranked enriched GO terms for dentate gyrus dataset.

Context	**Level 4 biological process**	**q-value**
1	sensory organ morphogenesis	0.002534
	cell morphogenesis	0.002534
	nervous system development	0.002534
	actin filament organization	0.002534
	muscle organ development	0.002534
	neuron development	0.002534
	muscle tissue development	0.002534
	skeletal muscle cell differentiation	0.002534
	cell migration	0.002701
	neuron differentiation	0.002701
3	regulation of neuronal synaptic plasticity	0.000767
	nervous system development	0.000767
	neurogenesis	0.000767
	regulation of vesicle-mediated transport	0.001770
	regulation of cell proliferation	0.001851
	positive regulation of cellular process	0.001851
	regulation of multicellular organismal development	0.002432
	cell projection morphogenesis	0.002458
	positive regulation of developmental process	0.002458
	cell part morphogenesis	0.002725
5	B cell lineage commitment	0.006214
	glial cell migration	0.008199
	cognition	0.008199
	cell migration	0.009270
	positive regulation of multicellular organismal process	0.012891
	regulation of transmembrane transporter activity	0.016714
	trans-synaptic signaling	0.018080
	positive regulation of cellular process	0.020871
	regulation of trans-synaptic signaling	0.020871
	response to light stimulus	0.021355

doi:10.6342/NTU202210156

Table 3.4: Ten highest-ranked enriched pathway terms for dentate gyrus dataset.

Context	**Pathway terms**	**Source**	**q-value**
1	AGE-RAGE signaling pathway in diabetic complications	KEGG	0.001212
	Parathyroid hormone synthesis, secretion and action	KEGG	0.001212
3	Glioma - Mus musculus	KEGG	0.018286
	ErbB signaling pathway - Mus musculus	KEGG	0.018286
	GnRH signaling pathway - Mus musculus	KEGG	0.018286
	AGE-RAGE signaling pathway in diabetic complications - Mus musculus	KEGG	0.018286
	Cholinergic synapse - Mus musculus	KEGG	0.018286
	Trafficking of AMPA receptors	Reactome	0.018286
	Glutamate binding, activation of AMPA receptors and synaptic plasticity	Reactome	0.018286
	HIF-1 signaling pathway - Mus musculus	KEGG	0.018286
	Neurotrophin signaling pathway - Mus musculus	KEGG	0.018286
	Regulation of TP53 Activity through Acetylation	Reactome	0.018286
5	Post-translational protein phosphorylation	Reactome	0.036114
	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	Reactome	0.036114
	Hedgehog signaling pathway - Mus musculus	KEGG	0.060524
	Glycerolipid metabolism - Mus musculus	KEGG	0.060524
	Focal adhesion - Mus musculus	KEGG	0.060524
	p53 signaling pathway - Mus musculus	KEGG	0.060524

doi:10.6342/NTU202210156

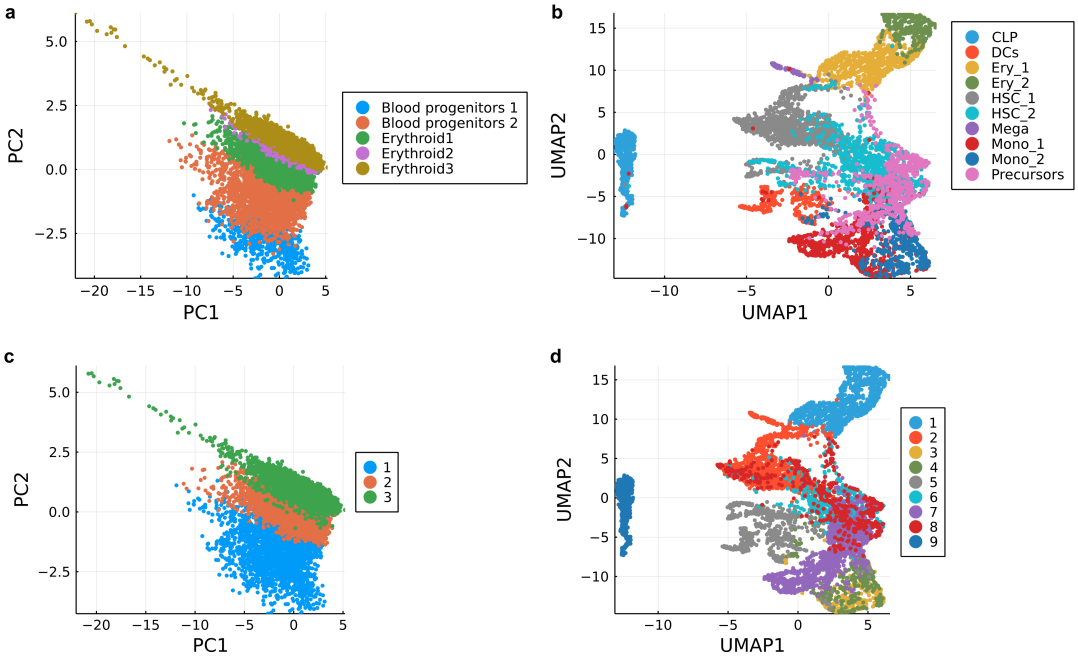

Figure 3.9: Inference and visualization of landscapes for mouse gastrulation to erythroid lineage and human bone marrow datasets. Developmental trajectory visualized after GMM feature selection for (**a**) mouse gastrulation to erythroid lineage and (**b**) human bone marrow. Distinct contexts reveal regulation dynamics for (**c**) mouse gastrulation to erythroid lineage, and (**d**) human bone marrow.

doi:10.6342/NTU202210156

Table 3.5: Ten highest-ranked enriched GO terms for mouse gastrulation to erythroid lineage.

Context	**Level 4 biological process**	**q-value**
1	blood vessel morphogenesis	1.10×10^{-6}
	vasculature development	4.23×10^{-6}
	cardiovascular system development	4.23×10^{-6}
	circulatory system development	0.000008
	cell migration	0.000018
	hematopoietic or lymphoid organ development	0.000033
	enzyme linked receptor protein signaling pathway	0.000081
	response to laminar fluid shear stress	0.000094
	small GTPase mediated signal transduction	0.000156
	myeloid cell differentiation	0.000225
2	cytoskeleton organization	0.014426
	myeloid leukocyte activation	0.015613
	cell migration	0.045814
	myeloid cell differentiation	0.045814
	regulation of cell motility	0.045814
	hematopoietic or lymphoid organ development	0.045814
	positive regulation of cellular process	0.045814
	regulation of cellular component movement	0.045814
	hematopoietic progenitor cell differentiation	0.045814
	regulation of cellular component organization	0.045814
3	regulation of anatomical structure morphogenesis	0.107952
	regulation of cellular component organization	0.107952
	myeloid cell differentiation	0.107952
	regulation of protein complex assembly	0.107952

doi:10.6342/NTU202210156

Table 3.6: Ten highest-ranked enriched pathway terms for mouse gastrulation to erythroid lineage.

Context	**Pathway terms**	**Source**	**q-value**
1	Reelin signalling pathway	Reactome	0.000037
	Platelet activation, signaling and aggregation	Reactome	0.000045
	GPVI-mediated activation cascade	Reactome	0.000122
	PECAM1 interactions	Reactome	0.000198
	Transcriptional misregulation in cancer	KEGG	0.001402
	Signal Transduction	Reactome	0.001785
	DAP12 signaling	Reactome	0.001812
	Hemostasis	Reactome	0.002553
	Signaling by VEGF	Reactome	0.002576
	Interleukin-3, Interleukin-5 and GM-CSF signaling	Reactome	0.002576
2	Transcriptional misregulation in cancer	KEGG	0.004060
	Acute myeloid leukemia	KEGG	0.004801
	Chronic myeloid leukemia	KEGG	0.004801
	Pathways in cancer	KEGG	0.014636
	Axon guidance	Reactome	0.021230
	Nervous system development	Reactome	0.021230
3	Transcriptional misregulation in cancer	KEGG	0.001959

doi:10.6342/NTU202210156

Table 3.7: Ten highest-ranked enriched GO terms for human bone marrow.

Context	**Level 4 biological process**	**q-value**
2	regulation of cell-cell adhesion	0.000062
	regulation of hemopoiesis	0.001356
	leukocyte differentiation	0.001356
	hematopoietic or lymphoid organ development	0.001356
	regulation of cell activation	0.002040
	neurogenesis	0.002040
	leukocyte cell-cell adhesion	0.002147
	lymphocyte differentiation	0.002147
	positive regulation of multicellular organismal process	0.002147
	regulation of cell differentiation	0.002949
7	hematopoietic or lymphoid organ development	0.013494
	leukocyte differentiation	0.027460
	bone cell development	0.034438
	nucleobase metabolic process	0.034438
	cellular response to xenobiotic stimulus	0.034438
	myeloid cell differentiation	0.034438
	bone development	0.034438
	regulation of multicellular organismal development	0.035925
	lipopolysaccharide-mediated signaling pathway	0.035925
	regulation of cell proliferation	0.035925
6	hematopoietic or lymphoid organ development	0.004979
	leukocyte differentiation	0.004979
	negative regulation of erythrocyte differentiation	0.009821
	negative regulation of immune system process	0.009821
	T cell activation	0.009821
	regulation of hemopoiesis	0.009821
	lymphocyte differentiation	0.015608
	defense response to protozoan	0.022174
	glomerulus vasculature development	0.022174
	myeloid cell differentiation	0.022174
3	cellular response to oxygen-containing compound	0.012110
	response to muscle stretch	0.012110
	cellular response to organonitrogen compound	0.012110
	cellular response to drug	0.012110
	response to decreased oxygen levels	0.012110
	cellular response to nitrogen compound	0.012110
	positive regulation of metabolic process	0.012110
	pigment cell differentiation	0.012116
	cellular response to xenobiotic stimulus	0.015304
	response to peptide hormone	0.015304

doi:10.6342/NTU202210156

Table 3.8: Ten highest-ranked enriched pathway terms for human bone marrow.

Context	**Pathway terms**	**Source**	**q-value**
2	Transcriptional regulation of granulopoiesis	Reactome	0.000043
	Signaling by EGFR	Reactome	0.004850
	Regulation of lipolysis in adipocytes	KEGG	0.005304
	Developmental Biology	Reactome	0.007104
	RUNX1 regulates transcription of genes involved in differentiation of HSCs	Reactome	0.008266
	Platelet activation - Homo sapiens	KEGG	0.012526
	Hepatitis C - Homo sapiens	KEGG	0.015672
	Axon guidance - Homo sapiens	KEGG	0.017749
7	Transcriptional regulation of granulopoiesis	Reactome	0.000174
	Plasma lipoprotein clearance	Reactome	0.005398
	Cholesterol metabolism	KEGG	0.009530
	Plasma lipoprotein assembly, remodeling, and clearance	Reactome	0.011499
	RUNX1 regulates transcription of genes involved in differentiation of HSCs	Reactome	0.014227
	Cell junction organization	Reactome	0.014227
	Developmental Biology	Reactome	0.021473
6	Transcriptional regulation of granulopoiesis	Reactome	0.001445
	Rap1 signalling	Reactome	0.007776
	NGF-stimulated transcription	Reactome	0.024503
	Signal Transduction	Reactome	0.039711
	Nuclear Events (kinase and transcription factor activation)	Reactome	0.039711
	Rap1 signaling pathway - Homo sapiens	KEGG	0.039757
	Cell surface interactions at the vascular wall	Reactome	0.039757
3	Signal Transduction	Reactome	0.001591
	Intracellular signaling by second messengers	Reactome	0.002737
	Parathyroid hormone synthesis, secretion and action - Homo sapiens	KEGG	0.002737
	Transcriptional regulation of granulopoiesis	Reactome	0.003185
	Integrin signaling	Reactome	0.003185
	Apelin signaling pathway - Homo sapiens	KEGG	0.003185
	SUMOylation of intracellular receptors	Reactome	0.003185
	Hemostasis	Reactome	0.003926
	Platelet Aggregation (Plug Formation)	Reactome	0.004316
	Nuclear Receptor transcription pathway	Reactome	0.006615

doi:10.6342/NTU202210156

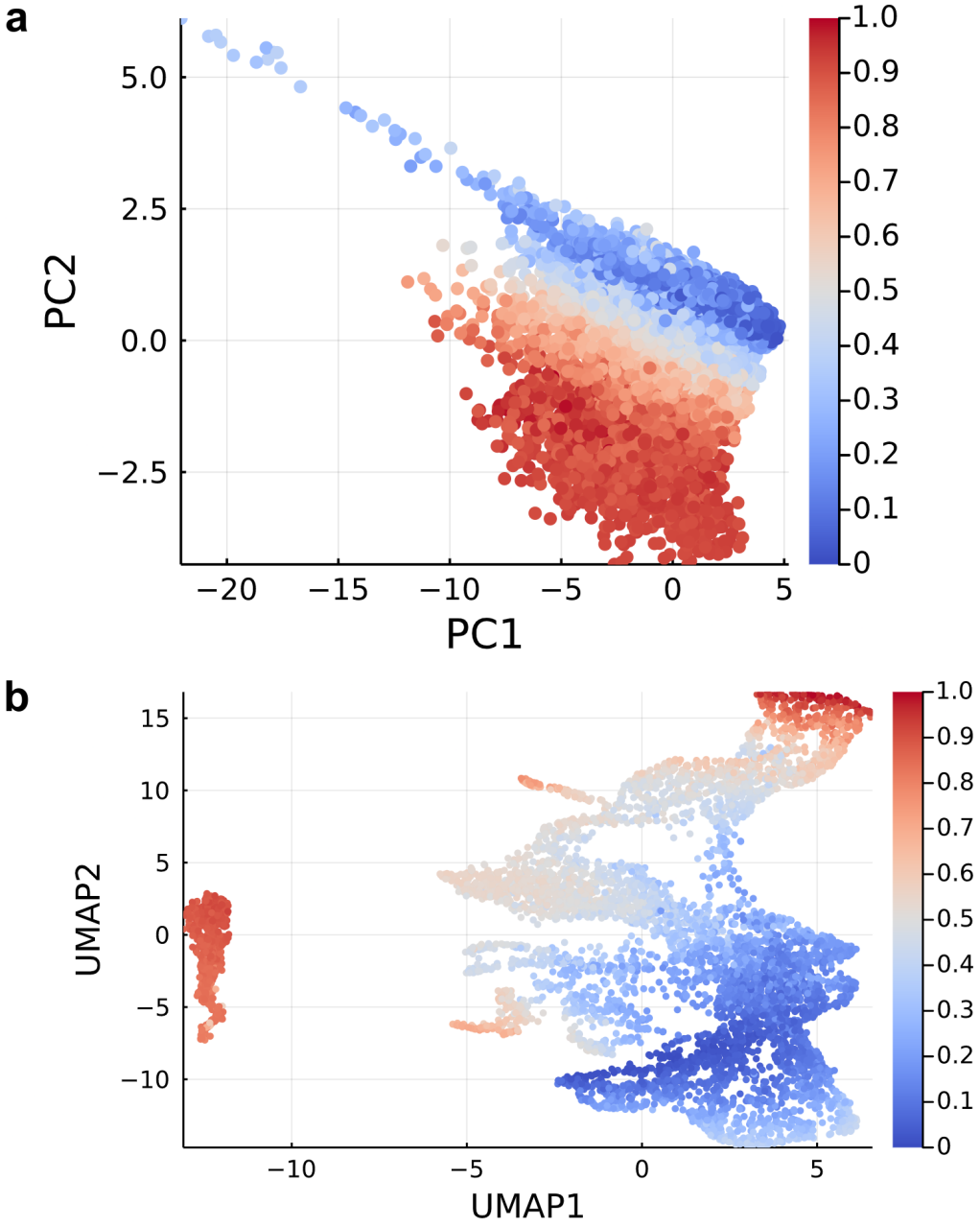

Figure 3.10: Inversed latent time inferred from generalized RNA velocity model for **a**, mouse gastrulation to erythroid lineage and **b**, human bone marrow.

doi:10.6342/NTU202210156

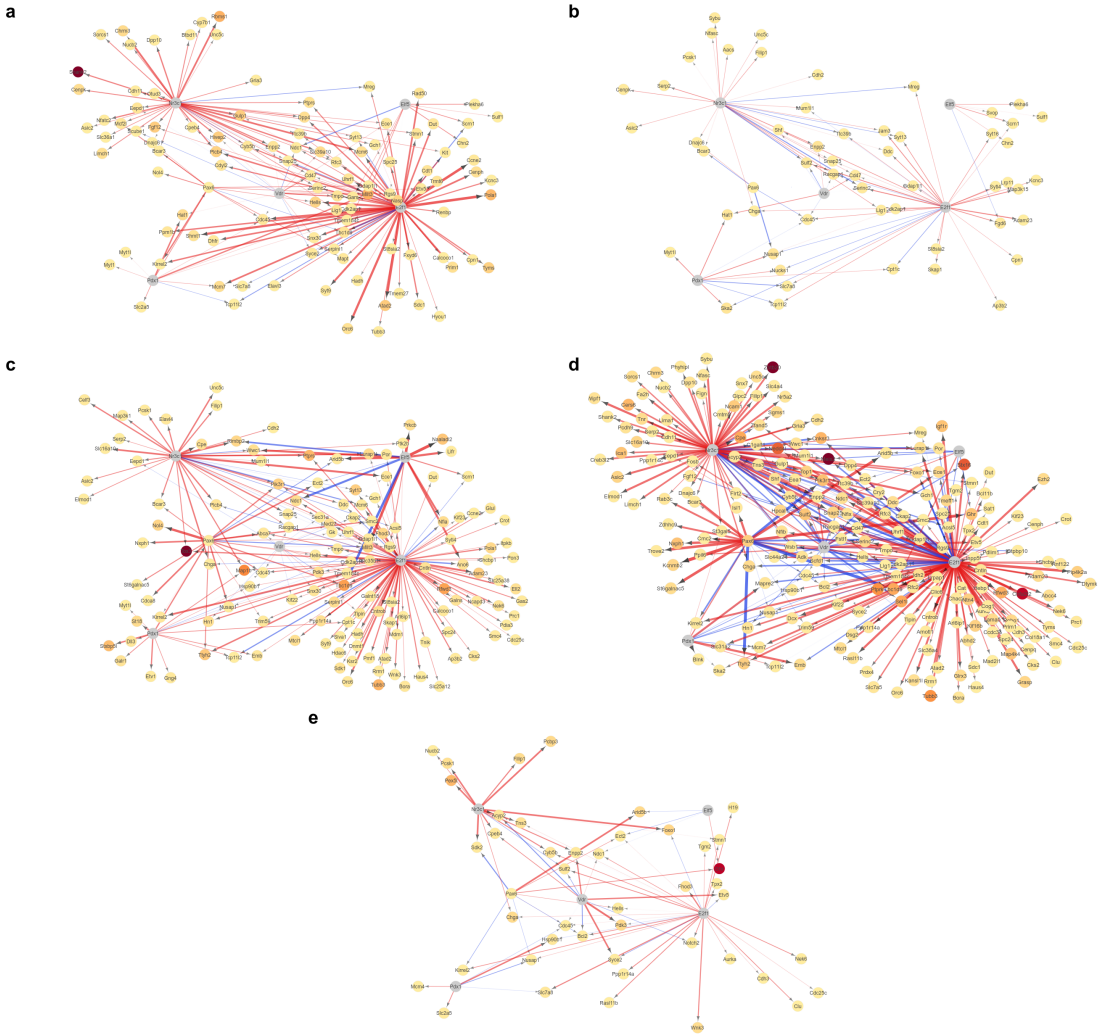

Figure 3.11: The visualization of CDGRN for (**a**) context 3, (**b**) context 2, (**c**) context 4, (**d**) context 5, and (**e**) context 1 in the pancreatic dataset. Each node represents a gene with its expression level in color from yellow (low) to dark red (high). Regulations are shown as directed edges with their colors in red (positive correlation) and blue (negative correlation). Directed edges with greater line with pose higher (absolute) correlations.

doi:10.6342/NTU202210156

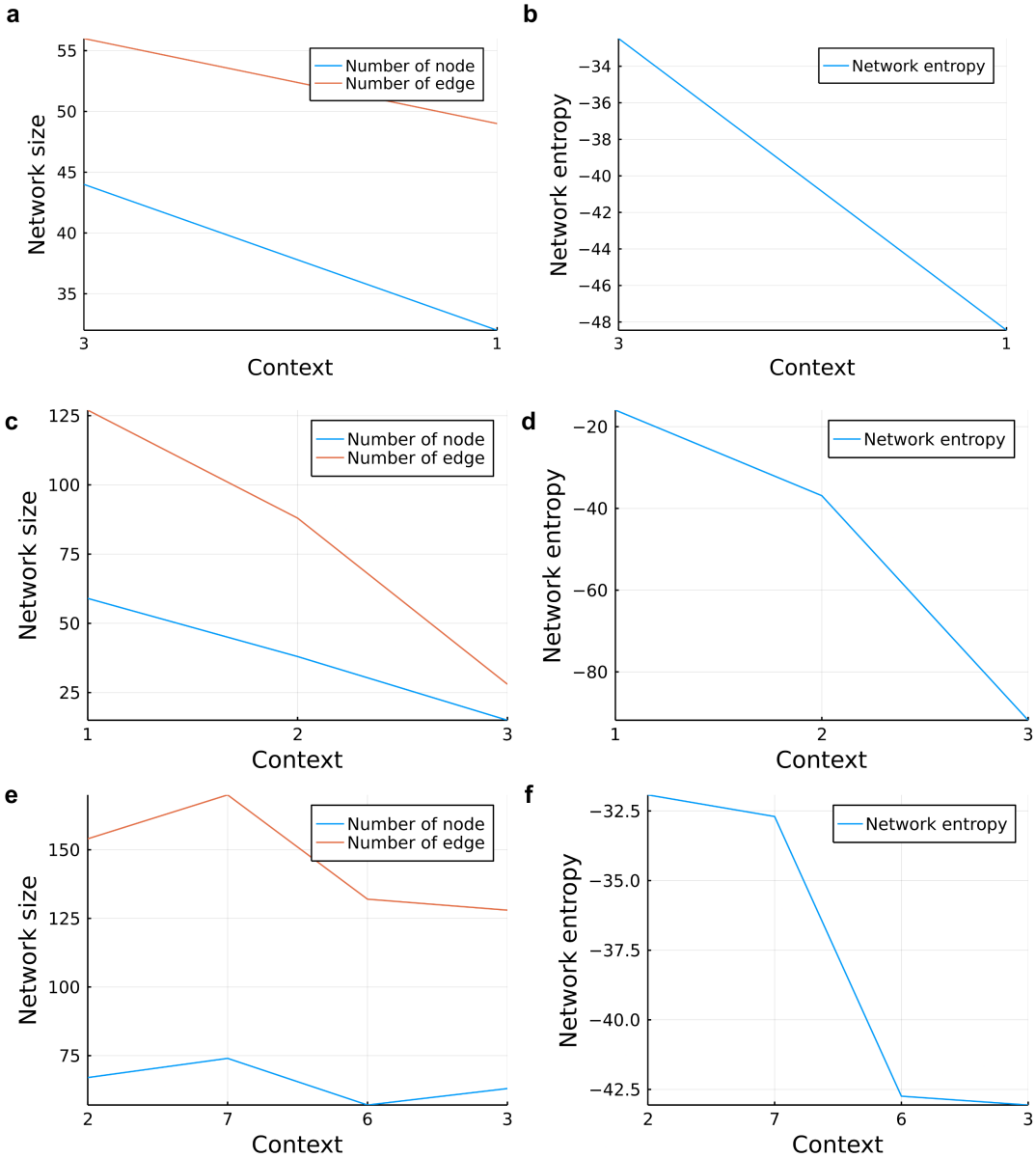

Figure 3.12: Network statistics for CDGRNs in each dataset. (**a**) Numbers of nodes and edges and (**b**) CDGRN network entropy for the dentate gyrus neurogenesis dataset. (**c**) Numbers of nodes and edges and (**d**) CDGRN network entropy for the mouse gastrulation to erythroid lineage dataset. (**e**) Numbers of nodes and edges and (**f**) CDGRN network entropy for the human bone marrow.

doi:10.6342/NTU202210156

Chapter 4

Discussion

We have investigated the GRN construction issue for single-cell sequencing data. A prespective of mixed regulatory patterns is revealed and can be decomposed by GMM into components. Mixed regulatory patterns represent an extent of nonlinearity in regulation dynamics, which can be decomposed into several linear patterns, across all cell types. Machine learning approaches often formulate regulation relationship prediction into a regression problem. Tree-based models like GENIE3 [17], GRNBoost2 [18], SCENIC [19, 20] also decomposed nonlinear features into piecewise linear patterns. While tree-based models leverage the power of approximation to nonlinearity for prediction, CDGRN dissects whole dataset into contexts based on these components. This shows CDGRN have ability to resolve a degree of nonlinearity for GRN construction problem.

Theoretically, some properties of CDGRN can be carried out. The use of Gaussian mixture model in CDGRN provides property of approximation to arbitrary distributions in general [35]. Arbitrary mixture patterns can be decomposed into several components in terms of linear patterns (lines) or cluster patterns (spots) from GMM. Contexts can be identified by clustering cells against regulatory pattern profiles. This provides the ability of CDGRN for capturing any kinds of regulatory patterns or even mixed regulatory patterns. This reasonably generalizes CDGRN

doi:10.6342/NTU202210156

to any dataset for gene regulatory network inference. Additionally, CDGRN also shows its robustness statistically. Suppose the dataset is random sampled from a certain population. In CDGRN, the population of any regulatory pair is modeled by GMM. Thus, the estimation of GMM would be the same from their population for the regulatory pair and the inferred CDGRN will be the same. The robustness of CDGRN is ensured statistically.

We also observed that CDGRNs may enable the determination of master regulators while not having sufficient evidence. The *Pdx1* gene has been reported as a unique master regulator in embryonic development and pancreatic cancer [36, 37]. The *Pax6* gene acts as a developmental regulator for maintenance of islet cell function and beta cell identity [38, 39, 40]. Given the possible TF–target gene pairs, a CDGRN extracts regulation relationships from single-cell transcriptome data. Since target genes are usually regulated by higher level TFs, hub TFs regulate more target genes, and in this context may then be candidates for master regulators.

It is essential to validate inferred results from CDGRN through biological experiments. There are some thoughts enable validating the results from CDGRN. A RNA-seq or in vivo fluorescent protein biological experiment can be made to measure and get a time-course cell differentiation data which provides insight into cell transition between states. Over the duration of transition between cell states, this measures the expression of genes or proteins across different CDGRNs. Thus, changes of gene regulations can be validated while cells change their contexts. Another more detailed experiment can be designed using a reporter system to validate gene regulation of interest in a more sophisticated setup.

GMM is employed to identify and extract regulation relationships, and Hamming distances are then computed to represent the distance among cells for following hierarchical clustering. Hamming distance regards regulation relationships as distinct classes, which neglects fine structure in-between the spectrum of distinct classes. A continuous clustering methods like fuzzy c-means or other distance methods can

doi:10.6342/NTU202210156

be considered to improve estimating distances. Therefore, it could provide more fine-grained contexts identification from cells.

While the proposed model establishes connections between macroscopic developmental trajectories and microscopic gene regulations, some further developments are likely desirable. For example, the currently employed method for identification of contexts is hierarchical clustering, which provides a simple method to identify transcriptional contexts from regulatory patterns. However, this approach only considers different regulatory patterns as distinct regulatory dynamics, and nuances like positive/negative regulations or regulation strength are not taken into account. The development of a more easily interpretable and meaningful method to identity transcriptional contexts may lead the way for describing discrete cellular contexts as continuous contexts.

CDGRNs can serve as a general approach for analyzing not only developmental trajectories but also cell clusters. While we only treat the former case in this study, there are no conceptual limitations for the latter, e.g., in the analysis of data from the PBMC dataset.

As noted, network complexity in the form of entropy in a CDGRN could work as an indicator for cell maturity; however, this remains a imprecise metric. It may be worthwhile to determine a more robust descriptor of network complexity that allows more reliable predictions of cell maturity and thus developmental directions.

The proposed model is subject to some limitations. Because the TF–target gene lists are fetched from a ChIP-seq experimental database, the use of a pure TFBS for modeling is limited. This issue could be addressed by pooling several ChIP-seq databases and thus enlarging the available data space; however, coverage of TF–target gene pairs would remain problematic. To resolve the issue thoroughly, an approach for modeling from a pure TFBS would be needed.

CDGRNs infer regulations by calculating correlations for gene expression. However, correlation does not equal causation, and spurious regulation relationships

doi:10.6342/NTU202210156

may be represented in a CDGRN. To reduce such cases, partial correlation networks could be used. Furthermore, the use of experimental databases rather than pure TFBS information may lower this risk.

In future research, the integration of scRNA-seq and scATAC-seq data is likely to become important. Taking into account chromatin openness in gene regulations may avoid a large proportion of falsely estimated positive regulations. From the perspective of epigenetics, the memory effect of chromatin openness explains how gene regulation differs from case to case. Individual or environmental factors may interact in their contributions at each level from the epigenome to gene regulations. Epigenetic information enables construction of a Waddington epigenetic landscape [41], which acts as a theoretical model for understanding how cell fates are determined and combines several advantages in one model. These include describing and explaining developmental trajectories, providing an explanation of the underlying gene regulation for macroscopic phenomena, evaluating the direction of cell differentiation, and predicting cell types. Some of these goals are achieved by the use of CDGRNs, making them possible building blocks for modeling epigenetic landscapes. Once cell maturity can be predicted correctly, modeling of the Waddington epigenetic landscape will become possible.

doi:10.6342/NTU202210156

Chapter 5

Conclusions

We propose a model intended to allow simultaneous inference of a cell population's gene regulatory network in a given context and the identification of the different contexts within the population. The model provides solid evidence for the interpretation of biological phenomena. We applied this model on four real datasets and show that the revealed trajectory is consistent with current biological knowledge. CDGRN explains gene regulation coupled with functional enrichment analysis in each context. Contexts dissect developmental trajectory into disjoint parts, and we found that subpopulation behaviors could be differ from other cells within the same cell types. We further show that the network entropy of CDGRN indicates cell maturity along the developmental trajectory.

doi:10.6342/NTU202210156

Bibliography

[1] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).

[2] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015). Seurat v1.

[3] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018). Seurat v2.

[4] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019). Seurat v3.

[5] Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021). Seurat v4.

[6] Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

[7] Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with palantir. *Nature Biotechnology* **37**, 451–460 (2019).

[8] Street, K. *et al.* Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19** (2018).

doi:10.6342/NTU202210156

[9] Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. *Nature Communications* **10** (2019).

[10] Wolf, F. A. *et al.* Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* **20**, 1–9 (2019).

[11] Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37**, 547–554 (2019).

[12] Manno, G. L. *et al.* Rna velocity of single cells. *Nature* **560**, 494–498 (2018).

[13] Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. Rna velocity-current challenges and future perspectives. *Molecular Systems Biology* **17** (2021).

[14] Matsumoto, H. & Kiryu, H. Scoup: Probabilistic model based on the ornstein-uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics* **17** (2016).

[15] Matsumoto, H. *et al.* Scode: An efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).

[16] Frankowski, P. C. A. & Vert, J. P. Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference. *Bioinformatics* **36**, 4774–4780 (2020).

[17] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5** (2010).

[18] Moerman, T. *et al.* Grnboost2 and arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).

doi:10.6342/NTU202210156

[19] Aibar, S. *et al.* Scenic: Single-cell regulatory network inference and clustering. *Nature Methods* **14**, 1083–1086 (2017).

[20] de Sande, B. V. *et al.* A scalable scenic workflow for single-cell gene regulatory network analysis. *Nature Protocols* **15**, 2247–2276 (2020).

[21] Luo, Q., Yu, Y. & Lan, X. Signet: single-cell rna-seq-based gene regulatory network prediction using multiple-layer perceptron bagging. *Briefings in Bioinformatics* **23** (2022).

[22] Gan, Y., Hu, X., Zou, G., Yan, C. & Xu, G. Inferring gene regulatory networks from single-cell transcriptomic data using bidirectional rnn. *Frontiers in Oncology* **12** (2022).

[23] Chan, T. E., Stumpf, M. P. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems* **5**, 251–267.e3 (2017).

[24] Qiu, X. *et al.* Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Systems* **10**, 265–274.e11 (2020).

[25] Deshpande, A., Chu, L. F., Stewart, R. & Gitter, A. Network inference with granger causality ensembles on single-cell transcriptomics. *Cell Reports* **38** (2022).

[26] Kim, S. ppcor: An r package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods* **22**, 665–674 (2015).

[27] Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C. & Huang, Y. A bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* **34**, 964–970 (2018).

doi:10.6342/NTU202210156

[28] Specht, A. T. & Li, J. Leap: Constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics* **33**, 764–766 (2017).

[29] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods* **17**, 147–154 (2020).

[30] Teschendorff, A. E. & Feinberg, A. P. Statistical mechanics meets single-cell biology. *Nature Reviews Genetics* **22**, 459–476 (2021).

[31] Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology* **38**, 1408–1414 (2020).

[32] Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).

[33] Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with consensuspathdb. *Nature Protocols* **11**, 1889–1907 (2016).

[34] Bastidas-Ponce, A. *et al.* Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development (Cambridge)* **146** (2019).

[35] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016). Http://www.deeplearningbook.org.

[36] Zhu, Y., Liu, Q., Zhou, Z. & Ikeda, Y. Pdx1, neurogenin-3, and mafa: Critical transcription regulators for beta cell development and regeneration. *Stem Cell Research and Therapy* **8** (2017).

doi:10.6342/NTU202210156

[37] Vinogradova, T. V. & Sverdlov, E. D. Pdx1: A unique pancreatic master regulator constantly changes its functions during embryonic development and progression of pancreatic cancer. *Biochemistry (Moscow)* **82**, 887–893 (2017).

[38] Hart, A. W., Mella, S., Mendrychowski, J., van Heyningen, V. & Kleinjan, D. A. The developmental regulator pax6 is essential for maintenance of islet cell function in the adult mouse pancreas. *PLoS ONE* **8** (2013).

[39] Swisa, A. *et al.* Pax6 maintains β cell identity by repressing genes of alternative islet cell types. *Journal of Clinical Investigation* **127**, 230–243 (2017).

[40] Gosmain, Y. *et al.* Pax6 is crucial for β-cell function, insulin biosynthesis, and glucose-induced insulin secretion. *Molecular Endocrinology* **26**, 696–709 (2012).

[41] Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: A landscape takes shape. *Cell* **128**, 635–638 (2007).

doi:10.6342/NTU202210156