Novel Computational Methods for Modeling Gene Regulatory Interactions

A thesis submitted by

Benjamin Clauss

in partial fulfillment of the requirements for the degree of

PhD

In

Genetics

Tufts University

Graduate School of Biomedical Sciences

February 2024

Advisor: Mingyang Lu

Abstract

Gaining a mechanistic understanding of cellular state transitions is a fundamental objective of biology. Of interest are the factors that coordinate cell state transitions, the type of regulatory interactions between them, and how both can be impacted to modify the state transition. A powerful tool for modeling gene interactions that underlie cellular state transitions are transcriptional regulatory networks. Some of the key advantages of modeling cellular state transitions with transcriptional networks is their ability to be combined with mathematical modeling to simulate the network, allowing for a dynamical understanding of how the network controls the state transition, as well as the experimental tractability of testing transcriptional regulation. However, accurate networks are difficult to obtain and currently no gold standard networks exist for mammalian biology, and currently there is a need to develop methods to improve network inference and modeling.

The work presented in this thesis consists of novel computational methods for modeling gene regulatory interactions. Chapter 1 consists of an introduction to the work presented herein, Chapter 2 provides details about the methods used for the work contained in the thesis, and in Chapter 3, work is presented that shows how classical systems biology methods can be used to gain information (i.e. which transcriptional regulatory interactions govern specific cellular state distributions or how the behavior of these interactions change in networks of different sizes and types) for genomics approaches. We present the first comprehensive dynamical evaluation of all nonredundant four-node circuits and use this to develop a method that can identify circuits, motifs, and motif

coupling responsible for any dynamical feature that can be described with a score, which is a quantitative measure of how well circuits are capable of achieving a specific dynamical feature. We demonstrate how this method can be used to understand the interactions governing theoretical representations of commonly observed types of gene expression state distributions, different classes of simulated gene expression state distributions, and human glutamatergic neuron differentiation. This approach is then extended to identify the regulatory interactions that underlie fundamental properties of transcriptional networks, such as the ability to generate multiple states and the ability to respond to signals.

In Chapter 4, a method for the inference and dynamical modeling of transcriptional regulatory networks is presented. This method uses both literature and data driven approaches to identify transcription factors that are active, the activity of those transcription factors, and the transcriptional regulatory network governing those important transcription factors, explicitly for the purpose of dynamical modeling. We show how to simulate the resulting network to gain a mechanistic understanding for the regulatory interactions responsible for TGF- ß driven EMT and macrophage activation.

The methods in this thesis help to address key issues in both the fields of Systems Biology and Genomics by showing how aspects of each can be combined to learn important information about gene regulatory interactions. We identify specific regulatory interactions governing multiple different classes of state distributions and human glutamatergic neuron differentiation. We also identify regulatory interactions that allow

for key dynamical aspects that allow regulatory interactions to function and evaluate how the behavior of those interactions change in networks of different sizes and types. Additionally, we identify key TFs and regulatory networks governing both EMT and macrophage activation. The information gained through the combination of these two fields is vital for accurately characterizing the factors involved in cell state transitions, understanding how these factors interact, and predicting the outcome of changes to these interactions.

Dedication

This thesis is dedicated to my friends and family. Without you I would not have been able to make it this far. Thank you for the constant love and support.

Acknowledgements

I would like to acknowledge my PI, Mingyang Lu, for giving the me the chance to join his lab despite never having coded before and being from a completely different field. I would also like to acknowledge my Thesis Advisory Committee, Chris Baker, Amy Yee, and Greg Carter, for supporting me throughout this long journey. Finally, I would like to acknowledge the other members of my lab, Danya Gordin, Ataur Katebi, Cristian Caranica, Dan Ramires, Kaitlyn Ramesh, Lijia Huang, and Yukai You.

Table of Contents

List of Figures

List of Copyright Materials Used

Huang, L., Clauss, B. & Lu, M. What Makes a Functional Gene Regulatory Network? A Circuit Motif Analysis. *J. Phys. Chem. B* **126**, 10374–10383 (2022).

List of Abbreviations

EMT – Epithelial to Mesenchymal transition
TRN – Transcriptional Regulatory Network
TPM – Transcripts per Million
RPKM/FPKM – Reads/Fragments per kilobase per million
TMM – Trimmed mean of M-values
TF – Transcription factor
GBM – Gradient boosting machine
DAG – Directed acyclic graph
ABM – Agent based model
ODE – Ordinary differential equation
mESC – Mouse embryonic stem cells
SCNS – Single cell network synthesis
KNN – K-nearest neighbor
PC – Principle component
DE – Differential expression
GSEA – gene set enrichment analysis
AUC – Area under ROC curve
RACIPE – Random circuit perturbation
scRNA-seq – Single cell RNA-seq
GRN – gene regulatory network
KD – Knockdown
HCA – Hierarchical clustering analysis
MET – Mesenchymal to epithelial transition
hiPSC – Human induced pluripotent stem cell
MI – Mutual information
GMM – Gaussian mixture model

1. Introduction

1.1 Cell states and cell state transitions

Cell states are defined by their cellular ability to achieve a specific function and a transition between cell states necessarily involves a change in the cell function[1,2]. The transitions involve a large number of molecular and phenotypic changes including: activity of important regulatory elements (promoters and enhancers), RNAs that are expressed, modifications of proteins, shape of the cell, levels of cell motility, and reliance on contacts with other cells[2,3]. There are also a number of epigenic changes that occur during cellular state transitions including changes in the accessibility of chromatin at both genes and regulatory regions, DNA methylation, histone modifications, and re-organization of the three-dimensional chromatin landscape[2,3]. These transitions play pivotal roles in normal development and disease progression, with notable examples observed during embryonic development[4], tissue maintenance[5], and cancer[6]. In embryonic development, pluripotent stem cells transition through many different cell states before arriving at terminally differentiated cell types[7]. In homeostasis and repair, different sets of epidermal stem cells are either continuously differentiating to ensure tissue renewal in the absence of injury, or are activated upon injury to provide cells for regeneration and tissue repair[5]. In cancer, a well-studied process, Epithelial to Mesenchymal Transition (EMT), occurs that mimics normal development and wound healing[8]. EMT involves a state transition from an epithelial cell phenotype to mesenchymal phenotype, with mesenchymal states implicated in invasion and

metastasis[6,8]. Cell states were traditionally identified based on observable features, such as morphology and location [1–3]. However, more recently, the classification of cell states has shifted towards molecular characteristics, with gene expression patterns being one of the primary means of characterizing cellular states[1,3].

The initial characterization of cell states through gene expression primarily relied on the expression of a few marker genes that were highly correlated with specific functional states[1,3]. However, high throughput methods have emerged that enable cell type identification based on genome-wide expression patterns to discern distinct cell types. Cell type and cell state are terms that are often used interchangeably, however cell types are generally a more broad term that refers to the biological classification of the type of cell (i.e. neuron, epithelial cell, macrophage, etc.) while cell state refers to the functional status of the cell. Different cell types each occupy different cellular states, but the same cell types can occupy different cell states depending on what it is currently doing. One particularly powerful method that has been developed is single-cell RNA-seq, which profiles the RNA in individual cells, and is great for capturing rare cell states. One downside of single-cell RNA-seq is that genes that are only expressed in a small number of cells, or with a very low overall expression, are often filtered out[9]. This filtering process results in only a small fraction of the genome being quantified. An older method, bulk RNA-seq is still commonly employed to capture genome wide expression patterns. Since bulk RNA-seq represents total transcripts of a cell population, it is not able to provide the same level of information as single-cell RNA-seq when it comes to heterogeneity or the presence of rare cell populations since it relies on population

averaging of transcripts[10]. Despite the differences in techniques, the classification of cell type of both methods are similar, with genome-wide patterns of cells states in question typically being compared to reference populations of known cell types.

Although gene expression capture methods are the predominate way of classifying cellular states, there are some limits to their ability to provide a functional readout of the cell. mRNA is only an intermediate molecule, between DNA and proteins, that often does not have catalytic functions by itself. Despite expression levels being used as a proxy for protein levels, there are several regulatory events that occur after expression and before the protein achieves its function that show mRNA is not the best readout for the function of the cell[11,12]. These regulatory events include: RNA splicing influenced by splicing elements, RNA silencing by miRNA or siRNA, RNA export, mRNA stability in the cytoplasm influenced by elements in the 3' UTR, translational control, post translational modifications required for fully functional proteins (eg. phosphorylation, glycosylation), and folding of the peptide into the correct functional confirmation with the help of chaperones. The fully functional proteins are then potentially subjected to degradation, must be transported to the location where they achieve their function, and finally often have to associated with other factors to achieve their goal. All these steps can be regulated and cause a lack of correlation between RNA and protein levels[13]. However, due to the difficulty of profiling the full protein compliment of the cell, coupled with the abundance of publicly available gene expression data and ease of generating new gene expression data, gene expression is still widely used as a functional readout of the cell.

Classical studies of cell state transitions have led to the idea of discrete cell states with commitment points for cell fates, where a specific point occurs in a transition and the cell becomes irreversibly committed without the ability to transition back[14]. However, more recent studies have supported the idea of continuous cell states, with a distinction between two states arising from statistical cut-offs[15–17].

Cell state transitions are controlled by the coordinated interaction of many intrinsic and/or extrinsic factors[18–20] but it is often unknown exactly how these factors cooperate to achieve cell state transitions. Currently, it is believed that the expression of a few key transcription factors give rise to the maintenance of cellular states and that a change in their expression levels is required to initiate and facilitate cell state transitions[21,22]. The alteration of the core transcription factors leading to cell state transitions is determined by a set of intrinsic regulatory logic between TFs, and this regulatory logic may or may not be influenced by the environment the cell is in; some cells are able to maintain the specific interactions that define their cell type despite environmental conditions, while others will transition to a new cell state in response to different environmental cues[1]. However, identifying the specific factors that are responsible for maintenance of cell identity or driving transitions, as well as the regulatory logic between them, is a difficult problem and is an area of research garnering a great deal of attention. Furthermore, it is not certain if cellular state transitions are controlled by core TF networks or by larger sets of interactions and testing the hypothesis that core TF networks drive cellular state transitions is also a difficult problem. Characterizing the TFs responsible directly responsible for cell state transitions, identifying the regulatory interactions that allow

them to achieve their functions, and predicting the response that direct perturbations have on cellular identity are all key questions in the study of cell state transitions.

1.2 Transcriptional Regulatory Networks

Networks serve as powerful tools for visualizing and modeling the regulatory interactions that govern cell states and their transitions. Transcriptional regulatory networks (TRNs) specifically depict the transcriptional regulatory interactions within a cell. In TRNs, nodes represent transcription factors, while edges symbolize the regulatory relationships between them. These edges can be activating or inhibiting, signifying the nature of the regulatory interaction[23].

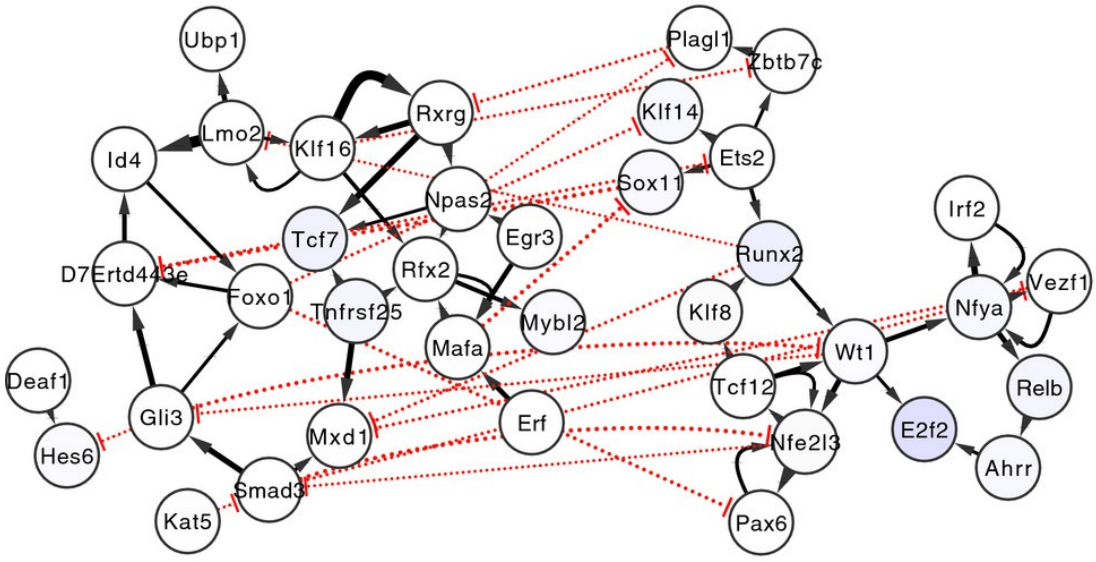

Figure 1.1 Example of a gene regulatory network[24]. Nodes are TFs and edges show activating interactions if arrows and inhibitory interactions if bars.

Various methods exist to infer networks[25], and once inferred, they serve as blueprints of molecular interactions, enabling the derivation of biological hypotheses[26]. For instance,

the identification of a novel transcription factor in the regulation of a specific state transition can be inferred from the network, leading to new insights and hypotheses[23,26]. Additionally, inferred networks can be simulated through mathematical modeling, allowing for the exploration of network perturbations and providing a mechanistic understanding of how changes in the expression of specific genes contribute to the overall functioning of the network[27].

1.3 Top-down and bottom up approaches

When inferring TRNs, there are two general approaches: bottom-up and top-down[28,29]. Bottom-up methods are considered more classical and do not rely on high-throughput sequencing technologies. Instead, they derive their interactions from extensive literature searches. This approach ensures the inclusion of known key regulators in the network, with edges often determined experimentally, thus providing high-confidence and direct interactions. Bottom-up approaches typically employ mathematical modeling to simulate the network, which involves defining equations or rules to represent the regulatory interactions of the network that are then solved to see how the expression of each gene in the system changes over time. A major step in the development of models is the ascertainment of correct parameter values, which are often done with literature searches, optimization algorithms, or empirical measurement from experiments. An important final step in the develop of a biologically relevant model is validation, which requires the comparison of the simulated behavior of the network to known behaviors of the biological system under study, ensuring that the inferred network captures the essential

dynamics of the biological system[28,29]. Mathematical modeling of transcriptional networks can then provide detailed understanding of how the interactions in a network lead to the emergence of distinct cellular states defined by specific mRNA levels behaviors and enables the formulation of new testable hypotheses facilitating a deeper mechanistic understanding of cell state transitions[22,23,26,28].

In contrast, top-down approaches utilize bioinformatic algorithms and whole-genome data, such as transcriptomic data, to directly infer networks. Network topologies are typically constructed by identifying statistical dependencies between genes, such as correlation, mutual information, or regression. The main advantage of top-down approaches lies in their context-specific nature. By leveraging the experimental conditions and biological setting at hand, top-down approaches ensure that the identified interactions directly reflect the given context. Additionally, these approaches have the potential to uncover novel factors and non-canonical interactions involved in the control of cell state transitions[28,29].

Despite the numerous advantages offered by both top-down and bottom-up approaches, several challenges persist in accurately inferring TRNs , and currently, there are no established gold standard TRNs for mammals. In the case of bottom-up approaches, constructing a network based on curated literature information can introduce issues. The literature sources often cover diverse biological contexts and experimental conditions, which can lead to conflicting or incorrect interactions. Additionally, the curation process

itself can be laborious and time-consuming, particularly when dealing with large networks and vast amounts of available information.

On the other hand, top-down approaches encounter a distinct set of challenges. The statistical dependencies used in these approaches do not necessarily represent direct interactions and are susceptible to false positives for interactions between genes[25,30,31]. Moreover, well-known regulators can be overlooked by current top-down methods. Furthermore, these approaches often neglect the evaluation of network dynamics, resulting in network topologies that fit statistical relationships but fail to capture the essential dynamics of the biological system.

Currently there is a critical need to develop new methods for modeling gene regulatory networks/circuits that combine top-down genome-wide approaches with bottom-up systems biology methodologies. By combining these disparate approaches to network modeling, it will be possible to generate context specific networks while also ensuring the networks behave as expected. Only through the combination of the two approaches will it be possible to infer accurate and personalized networks.

1.4 Top-down approaches

1.4.1 RNA-seq processing

mRNA plays a crucial role as an intermediate between DNA and proteins, and is generally used to represent a cell's functional status due to the abundance of publicly

available information as well as the ease in obtaining new information. Therefore, information regarding the abundance of different mRNA transcripts is essential for inferring TRNs. Currently, three types of gene expression information are commonly used for this purpose: microarray, bulk RNA-seq, and single-cell RNA-seq.

Microarray data, although an older technique, consists of an array of probes with known sequence. These probes are complimentary to genes of interest. Experimentally derived transcripts, labeled with a fluorophore, are then hybridized with the probes. The location of each probe is known and the resulting signal intensity is used to measure gene expression levels[32]. However, microarray technology is not frequently used anymore.

Bulk RNA-seq and single-cell RNA-seq are more modern and high-throughput techniques that enable profiling of the entire transcriptome, offering comprehensive information and the potential to identify novel transcripts. Bulk RNA-seq involves averaging gene expression across a population of cells, which may obscure cellular heterogeneity. On the other hand, single-cell RNA-seq captures the expression profile of individual cells, allowing for the identification of distinct cell types within a sample. However, single-cell RNA-seq has limitations such as extensive dropouts (missing data) and heteroskedasticity (uneven variation). Despite these challenges, single-cell RNA-seq is particularly suitable for inferring cell state transitions because it can capture multiple states within the same sample[10].

Typical processing pipelines for both bulk and single-cell RNA-seq data include quality control to filter out low-quality reads, mapping of transcripts to a genome or transcriptome, quantification of mapped reads for each feature, normalization of the resulting counts matrix, and differential gene expression analysis[33].

Normalization is an important intermediary step in RNA-seq processing, and the choice for normalization can have large impacts on downstream analysis results. Normalization is required for accurate comparison of gene expression between samples and is the process of scaling raw counts to account for factors that can such as sequencing depth, gene length, and RNA composition. For Bulk RNA-seq, five methods for normalization are commonly used: 1) Counts Per Million (counts per million) which scales reads based on total number of reads, accounting for sequencing depth. CPM is good for gene count comparisons between replicates but not for within sample comparisons or DE analysis. 2) Transcripts per kilobase million (TPM) scales reads by length of transcript per million reads mapped, accounting for sequencing depth and gene length. TPM is good for comparison of genes within or between samples, but not for DE analysis. 3) Reads/Fragments per kilobase of exon per million reads/fragments mapped (RPKM/FPKM) which scales in a similar way and accounts for the same factors as TPM and is good for comparing genes within a sample but not for between samples or DE analysis. Normalization by RPKM/FPKM results in different numbers of normalized counts for each sample and therefore should not be used for comparison. 4) Median of ratios, developed and used by DeSeq2, which scales counts by the median ratio of counts relative to the geometric mean per gene and accounts for sequencing depth and RNA

composition. Median of ratios is good for comparison of genes between samples and for DE analysis but not for within sample comparisons. 5) trimmed mean of M values (TMM), designed and used by EdgeR, which scales on a weighted trimmed mean of the log expression ratios between samples and accounts for sequencing depth and RNA composition. TMM is good for comparison between samples and for DE analysis but not for comparisons within a sample[34,35].

There are three commonly used methods for the analysis of bulk RNA-seq data, DEseq2[36], edgeR[37], and limma[38]. The input to all three of these methods are a counts matrix and the output is a list of genes that are significantly differentially expressed between samples. Deseq2 first normalizes its data via a scaling factor that is estimated as described above, estimates dispersion for each gene, shrinks the dispersion of genes that are below average or slightly above, fits the counts of genes to a negative binomial model, and identifies statistically significant fold changes via the Wald or maximum likelihood tests[36]. EdgeR normalizes counts using the trimmed mean of M values. Edge R models counts assuming a negative binomial distribution, using an empirical bayes procedure to shrink dispersion, and identifies DE genes an exact test analogous to Fisher's Exact test but for over dispersed data[37]. Normalization for Limma is often performed using EdgeR's TMM, uses empirical bayes statistics to moderate standard errors but does not estimate gene wise dispersion. Limma fits a linear model to the data, without a negative binomial assumption, to identify DE genes using a modified t-statistic[38].

Seurat[39] is a commonly used single-cell RNA-seq analysis toolkit. Seurat takes multiple single-cell RNA-seq datasets as an input, and as a first step, uses canonical correlation analysis to learn a gene correlation structure for each dataset, called a CCA bias vector. The bias vector is then aligned between datasets using non-linear dynamic time warping to create a shared low dimensional space for projection. After the datasets are projected into a shared space, clustering is performed to identify different cell populations which can then be analyzed to identify population differences between conditions, such as cell type proportion shifts and specific transcriptional responses.

1.4.2 Bioinformatic methods for inference of gene regulatory networks

Bioinformatic methods of regulatory network modeling often use gene expression data to construct networks from statistical dependences between genes[40]. These methods take high-throughput gene expression data as an input and output a network as described in section 1.2. Such methods usually stop once the network is identified and do not model dynamics of the network. Multiple evaluations of network inference methods have been performed, including the DREAM4[41] in silico network inference challenge, the DREAM5[42] network inference challenge, and BEELINE[25], among others[30,31,43]. When evaluated, these types of methods often have shortcoming that include failing to infer multiple regulatory inputs of genes and an abundance of false positive edges that represent indirect interactions. Below, some of the top performing methods are described.

GENIE3[44] is a highly effective network inference tool that has demonstrated outstanding performance in the DREAM4 and DREAM5 network inference challenges[43]. It has also proven to be one of the top performers and most stable inference methods in other recent benchmarking studies of network inference tools[25]. The key principle behind GENIE3 is the use of random forest models to identify the most predictive genes for the expression of each target gene in a given gene expression matrix. The models assign weights to transcription factors (TFs), with higher weights indicating stronger TF-target regulatory interactions. By employing random forest models, GENIE3 can capture non-linear interactions from gene expression data. The output of GENIE3 is a table that provides information on genes, their potential regulators, and the corresponding weights.

Random forest models are a form of supervised machine learning that consist of an ensemble of decision trees[45]. Random forest algorithms can be used for both regression and classification problems and work by combining the results of multiple decision trees to increase the overall performance of the model. The random forest algorithm randomly selects observations and features to build its decision trees and then averages the results to make predictions. The main advantages of random forest models are their ability to be used for both regression and classification and the ability to view the relative importance the models assign to the input features. For GENIE3, gene expression serves as the input and each gene's expression profile is treated as a feature. The goal of GENIE3 is to predict the expression of each gene based on the expression of all other genes in the dataset. GENIE3 creates multiple decision trees for each gene using the expression of all other genes. The feature importance calculated by GENIE3 show how much each gene

contributes to the expression of each target gene. Features with higher importance are considered more influential and used to build GRNs.

Another highly performing network inference tool is GRNBoost2[46], which has excelled in the DREAM5[43] challenge and BEELINE[25]. GRNBoost2 builds upon the architecture of GENIE3 but utilizes gradient-boosting machines (GBM) to infer regulatory interactions. It employs shallow regression trees (trees with a depth of 1) that are combined to create a more robust tree. This approach differs from GENIE3, as GENIE3 employs bagging to enhance model performance. Bagging is a machine learning technique that improves model prediction by training multiple instances of a single base learning algorithm using different subset of the training data and then combining their predictions to make a final prediction[47].

SCENIC[48] is a comprehensive workflow that incorporates either GENIE3 and GRNBoost2, along with RcisTarget and AUCell, to infer regulons of TF-target interactions specifically for single-cell data. The SCENIC workflow involves three main steps: 1) construction of a co-expression matrix using either GENIE3 or GRNBoost, with a defined threshold to identify putative regulatory interactions above the threshold. Although co-expression can introduce false positives due to co-regulation and indirect interactions, SCENIC overcomes this limitation by 2) identifying direct interactions through cis-regulatory sequence analysis using RcisTarget. RcisTarget identifies enriched TF-binding motifs and candidate TFs for a given gene list. Modules, identified in step 1, are retained only if they show significant enrichment of the correct upstream regulator,

and edges are removed if the target genes do not contain regulatory motifs for the upstream TF, ensuring the presence of only direct interactions. The retained pruned modules are referred to as regulons. 3) The activity of regulons is then scored using AUCell, which calculates the activity score for the entire regulon, ensuring robustness against dropouts[48].

Despite these methods being top performers in benchmarking studies, further improvements are still needed. Multiple reports have shown that these network inference methods typically have a high level of false positives, low intersections with interactions inferred from ChIP-seq data, and do not generate reproducible results[25,30,49]. One major shortcoming of these methods is that they focus purely on identifying topologies that describe statistical dependencies between genes and do not take into account the dynamical behavior of the networks they generate (i.e. state distribution), leading to networks that are unable to recapitulate the behavior of the biological processes they are trying to model.

1.5 Bottom up methods

1.5.1 Boolean, Bayesian. and ODE network modeling approaches

Various approaches have been employed for modeling gene regulatory networks, including Boolean, Bayesian, and ODE methods, each offering distinct advantages and limitations. Boolean networks are represented as signed and directed graphs and are parameterized with logical functions. In Boolean modeling, gene activation is binary,

either 1 (active) or 0 (inactive). TFs in a Boolean network exert regulatory influence on target genes through logical functions, and the network is simulated using update schemes that reflect the logic of interactions. Simulations can then be analyzed to identify attractor states, which represent different cellular states, transitions between attractor states, and to understand how perturbations to the network can alter the resulting states. The key advantages of Boolean modeling are its requirement for a relatively small number of parameters and the ability to incorporate logic rules[50].

Bayesian modeling consists of two components: 1) a directed acyclic graph (DAG) where nodes represent random variables and directed edges depict stochastic dependencies, and 2) conditional probability distributions assigned to each variable that describe the interactions specified by the edges. The absence of an edge between two nodes implies that they are independent given the values of any intermediate nodes, which is known as the Markov condition. Bayesian models offer the advantage of providing probability distributions as output, which can be more informative than a single value. However, Bayesian modeling has limitations, including the fact that the DAG structure may not reflect biologically relevant TRNs, and it often requires substantial computational resources[51].

Ordinary Differential Equation (ODE) based methods are used to model dynamical systems of genes, often with a different equation for each gene. Each ODE describes the change in the expression of that gene over time and contains numerous parameters that describe different aspects of the system, such as rates for production and degradation,

thresholds for activation or inhibition, and the magnitude of influence a regulator has on its target[52]. Dynamical modeling with ODEs has led to the idea of different types of steady states for biological systems including oscillatory steady states, stable steady states and unstable steady states. Oscillatory steady states in biology are stable states that have reoccurring patterns which do not converge and instead exhibit periodic oscillation around an equilibrium. A common example of an oscillatory steady state is the cell cycle. Stable steady states are states in which slight perturbations cause a return to the systems original state, also known as attractor states. An unstable steady state is a state in which slight perturbations results in the system moving far away from the original state until another steady state is reached. Terminally differentiated cells are an example of stable steady states while bifurcation points that occur during the differentiation process are thought of as unstable steady states[53]. ODEs key benefits are their ability to give detailed and biologically relevant descriptions of a system. The drawback of ODEs however, are that they require many difficult to obtain parameters in order to achieve biological relevancy and that they do not scale well[52,53].

While the approaches described in this section consider the dynamical behavior of networks, they are typically applied to smaller circuits derived from literature interactions. The literature interactions often come from multiple genetic backgrounds and experimental conditions. Because of how the networks they model are constructed, they do not consider context specific interactions and are unable to implicate novel TFs in driving cellular state transitions.

1.5.2 Gene Regulatory Circuit Motifs

Network motifs refer to recurring small circuit topologies observed within larger gene regulatory networks[54]. These motifs have been found to play distinct roles in creating and maintaining circuit states, driving state transitions, and processing signals, as demonstrated through synthetic biology[55], computational systems biology modeling[56,57], and experimental studies[58]. Traditionally, gene circuit motifs were identified by examining the topology of large biological networks, such as those from *E. coli* and *S. cerevisiae*, to detect the presence of smaller circuit motifs[54,59,60]. Motifs are considered significant when they are overrepresented in biological networks compared to randomly generated networks[54].

Recent research has focused on identifying functionally relevant circuit motifs that can produce specific dynamical behaviors[56,57,61,62]. This involves using mathematical modeling to design circuits and then analyzing them for enriched motifs. For example, Ye et al.[57] identified three-node circuits capable of generating stepwise transitions between four states with limited reversibility, shedding light on the regulatory interactions controlling T-lymphocyte development. Schaerli et al.[61] investigated circuits capable of stripe formation and identified incoherent feed-forward loops and a two-node motif involving activation and inhibition as critical motifs for this process.

Through the identification of network motifs we can identify the important types of interactions that define the overall functions of networks. Once circuit motifs are identified, the motifs can then be used to inform the construction of larger networks. Currently, new methods need to be developed for the inference of circuit motifs. Most

circuit motif methods have one of two issues: 1) the method will first identify motifs that are over-represented in networks for a specific biological context and then try to assign a function to the motifs; 2) methods don't robustly explore the dynamical behaviors of the motifs to which they are assigning a function. Instead, methods need to be developed that directly infer motifs capable of producing a specific function and are able to evaluate the dynamic capabilities of the motif under many different conditions.

1.6 Combined top-down and bottom-up methods

Combining top-down and bottom-up approaches in network modeling can provide a more comprehensive understanding of gene regulatory networks, incorporating both known interactions and data-driven inference. Here are a few examples that demonstrate the integration of these methodologies:

In the study by Dunn et al.[63] , the authors aimed to identify the core network controlling pluripotency in mouse embryonic stem cells (mESCs). They started with well-known transcription factors implicated in pluripotency based on literature knowledge and measured gene expression changes in different culture conditions. Using correlation analysis, they inferred potential interactions. To refine their network, they employed a Boolean network modeling approach and constrained the interactions to models that could reproduce known behaviors of the system. The resulting models were then used to predict the response to new experimental perturbations, with iterative refinements to fit the data. In 2019[64], they extended their approach to identify a network controlling both

the maintenance and induction of pluripotency by updating their network with new RNA-seq datasets and following the same framework of defining constraints for observed behaviors and iteratively refining their models based on experimental perturbations. The final model achieved a predictive accuracy of 77.4% for new experimental tests.

Moignard et al.[65] developed a computational method for building a network for hematopoietic development, using a bottom-up approach based on experimental data. They focused on hematopoietic marker genes known from literature and quantified the expression of 33 transcription factors involved in hematopoiesis. Their method, called single-cell network synthesis (SCNS) toolkit, directly derived Boolean rules from the experimental data without prior knowledge of the network structure. SCNS discretized the expression of each transcription factor in single-cell expression profiles and created a transition graph representing developmental expression state changes. Edges were assigned directions based on experimentally observed cell state transitions. SCNS then determined Boolean update functions for each gene consistent with the observed expression changes, resulting in a core network of 20 transcription factors. The validity of the network edges was assessed by comparing them to ChIP-seq data, and the agreement of simulated network perturbations with experimental perturbations was evaluated.

In Sha et al[66]. they infer communication between cells and gene regulatory networks that control EMT. They apply a trajectory analysis and identify multiple intermediate EMT stages with hybrid features. From their analysis they are also able to identify different

types of EMT in response to different triggers, with TGFB1 induced EMT being synchronous and EMT induced by epidermal growth factor and tumor necrosis factor being asynchronous. Furthermore they also implicate the intermediate cell states as playing a dominant role in key EMT signaling pathways, such as TGF-β

These examples demonstrate how combining top-down knowledge and bottom-up data-driven approaches can enhance network modeling by incorporating both known interactions and context-specific inference. By iteratively refining the models and evaluating them against experimental data, these approaches can provide deeper insights into gene regulatory networks and enable the generation of new predictions.

1.7 Our Approach

By combining top-down and bottom-up approaches in multiple novel methods for computational modeling of gene regulatory networks, I have addressed the current issues for TRN inference, which consist of 1) genome-wide approaches failing to ensure their networks can behave as expected and 2) systems biology methods unable to infer context specific interactions. A combined top-down and bottom-up approach allows leveraging of the positives of both while mitigating the negatives. The following work consists of novel methods to infer important network motifs, small circuits, or TRNs directly from experimental data, with topologies that are optimized to ensure dynamical features of the biological setting are able to be recapitulated. These methods will aid in the understanding of cell state transition mechanisms by improving the ways in which

networks are inferred, optimized, and evaluated. The application of our methods will help to ensure inferred networks are both context specific and are able to recapitulate the dynamics of the biological setting.

The work contained herein consists of three finished projects, organized into two chapters, and a discussion of an unfinished project that address other longstanding issues within the field. Chapter 3 discusses novel methods for modeling small circuits to gain important insights for bioinformatically derived TRNs. In this chapter, traditional systems biology approaches are learn important regulatory information for genomics. A novel method to identify topological motifs is presented. This method identifies biologically relevant motifs through a comparison of experimental data to simulated small circuits. This method is then applied to glutamatergic neuron differentiation and identifies important interactions between known transcription factors that control the cell state transition. This method is then extended in another publication to identify motifs and network types that underlie general aspects of biological networks, such as their ability to respond to extrinsic signals and create multiple states. In chapter 4, a novel method for the construction TRNs, based off the inferred activity of well-known TFs is presented. This method is then applied to model both EMT and macrophage activation, with ODE based simulations identifying transcription factors that control the cell-state transitions. In the discussion, a framework utilizing both novel and existing computational methods is presented. This method aims to develop a network inference pipeline for networks capable of producing multiple states, and with its topology optimized to ensure the

dynamical behavior of the inferred network matches the biological setting from which it is derived.

Together, the research presented here shows how traditional systems biology methods can be used to gain information for and be combined with current bioinformatic techniques to develop novel methods for modeling gene regulatory networks. This work moves the field forward by going beyond simply analyzing the topology of small circuits and TRNs to evaluating their dynamics to gain information on gene-gene connections. Methods that ensure inferred interactions and networks can achieve the specific behaviors of the biological system they were inferred for is an essential step forward in network modeling.

2 Materials and Methods

2.1 RACIPE

RACIPE[67] is an ODE based mathematical modeling method to simulate the steady-state gene expression of a gene regulatory circuit. Considering a gene j that is transcriptionally regulated by one or multiple regulators i, we can describe the gene expression levels of gene j by:

$$dx_j/dt = \frac{G_j}{\prod_i {\lambda^+}_{ij}} \prod_i H^S(x_i, x_{ij0}, n_{ij}, \lambda_{ij}) - k_j x_j \quad (1)$$

where G_j is the transcription rate of gene j, x_i or x_j is the current expression level of gene i or j, and k_j is the degradation rate of gene j. H^S is the shifted hill function that describes how regulators impact the expression of their target and is defined as:

$$H^S\left(x_i, x_{ij0}, n_{ij}, \lambda_{ij}\right) = \lambda_{ij} + \left(1 - \lambda_{ij}\right)/\left(1 + \left(\frac{x_i}{x_{ij0}}\right)^{n_{ij}}\right) \quad (2)$$

x_{ij0}, n_{ij} and λ_{ij} are the threshold level, the Hill coefficient of regulation, and the maximum fold change for the regulatory link from i to j. λ_{ij} is denoted as ${\lambda^+}_{ij}$ for an excitatory interaction and takes a value larger than 1. In this case, H^S ranges from (1, ${\lambda^+}_{ij}$). In the case of an inhibitory interaction, λ_{ij} is denoted as ${\lambda^-}_{ij}$ and takes a value smaller than one. In this case, H^S ranges from $({\lambda^-}_{ij}, 1)$.

With these equations RACIPE randomly samples kinetic parameters from uniform distributions, i.e., G_j from (1, 100), k_j from (0.1, 1), n_{ij} (integer) from (1, 6), and ${\lambda^+}_{ij}$ from (1, 100), for each parameter. For an inhibitory interaction, ${\lambda^-}_{ij}$ is sampled from a

uniform distribution of (1,100) and its inverse value is taken. x_{ij0}, is randomly chosen from ($0.02M$, $1.98M$). The half-functional rule allows estimation of the median Hill threshold *M*. Once parameters have been generated, RACIPE simulates the ODEs for the whole network with initial conditions randomly sampled from a logarithmic distribution whose maximum is $\frac{G_j}{k_j}$, and minimum is $\frac{G_j}{k_j}(\frac{\prod_i {\lambda^-}_{ij}}{\prod_i {\lambda^+}_{ij}})$.

The major advantage to ODE based methods, like RACIPE, are that they offer a detailed view of the system they are simulating, outputting continuous dynamics for each node in the network. However, they also require a large number of parameters that are often hard to obtain. RACIPE is able to overcome this general disadvantage in ODE modeling with its consensus approach where many different models are generated for a given topology and the parameters for each model are sampled from biologically relevant ranges. This approach allows the robust exploration of TRN dynamics without requiring many hard to obtain parameters.

2.2 Methods for Modeling Small Circuits for Genomics

2.2.1 Generation of all 4-node circuits

To systematically evaluate the dynamical behavior of circuit motifs, we generated all non-redundant four-node gene circuits according to the following rules. First, we obtained all possible four-node circuits, where any two genes can be connected by either an activating interaction, an inhibiting interaction, or no interaction, and any gene can

have either a self-activating interaction, a self-inhibiting interaction, or no autoregulation. Here, only a maximum of one regulatory interaction was considered from one gene to another; because of the directionality of gene regulation, a maximum of two regulatory interactions is possible between two genes (i.e., from the first one to the second, and vice versa). This first step leads to a total of 43,046,721 circuits. Second, circuits containing floating, signal or target nodes were identified and excluded, as the circuits are equivalent to those with less number of nodes. Here, a floating node is defined as a node with no interaction, neither incoming nor outgoing with another node in the circuit; a signal node is defined as a node with only outgoing interactions with other nodes; a target node is defined as a node with only incoming interactions with other nodes. These definitions hold regardless of the occurrence of autoregulation. This filtering step allows us to perform the analysis on smaller circuits without affecting the outcomes of state distribution. However, it also excludes circuits motifs known to process signaling inputs, such as bi-fan and diamond motifs. Third, for each of the remaining circuits, we constructed an adjacency matrix, with 0 representing no interaction, 1 representing activation, and 2 representing inhibition. We then computed the trace, determinant, and eigenvalues of the adjacency matrix. The purpose to compute eigenvalues of the adjacency matrix is to detect redundant gene circuits due to label swapping. We considered two circuits redundant when these values are identical. We kept one circuit from all redundant circuits, which eventually leads to a total of 60,212 non-redundant four-node gene circuits for further analysis. Nonredundant four-node circuits have been analyzed in previous work[56], however the sign of the interactions (i.e., activation and

inhibition) and autoregulation were not explicitly studied, leading to a much smaller number of circuits.

2.2.2 Triangular and linear scores

We developed a method to characterize the simulated gene expression data of all non-redundant 4-node circuits as either belonging to a triangular or linear distribution of three states. We first performed k-means clustering (k = 3) to the simulated gene expression data, and defined a score for triangular structure, Q_1, as:

$$\mathrm{Q}_1 = \min_{i,j} \frac{D_{ij}}{S_i S_j} \quad (3)$$

where D_{ij} is the Euclidean distance between the centers of clusters i and j, and S_i is the average Euclidean distance of each point in the cluster i to the cluster center. Q_1 takes the minimum of the ratio term in Equation (3) over all three pairs of clusters, *i.e.,* 1-2, 2-3, and 3-1 for i and j. Intuitively, each ratio term measures how separate two clusters are. When the lowest of the three ratios is still high (thus high Q_1), all three clusters should be well separated.

We ranked all non-redundant four-node gene circuits with Q_1, so that we can identify circuits whose gene expression distribution most (or worst) resemble to a triangular structure.

Next, we defined a second score for linear structure, Q_2, as:

$$Q_2 = \min_{i,j,k} \left\| D_{jk}(S_j + S_k) - D_{ij}(S_i + S_j) - D_{ik}(S_i + S_k) \right\| \quad (4)$$

where Q_2 takes the minimum of the new term over any order of the three clusters i, j, and k, *i.e.,* 123, 231, 312 (note that the term in Q_2 is unchanged when swapping the order of j and k). The three clusters were obtained by the above-described k-means clustering. When the three clusters are co-linear, one of these Euclidean distances should be close to zero. The S_i terms are included here to minimize the spread of the clusters. We also ranked all non-redundant four-node gene circuits with Q_2 , so that we can identify circuits gene expression distribution most (or worst) resemble to a linear structure.

2.2.3 Circuit Motif enrichment

After ranking all non-redundant four-node gene circuits with both Q_1 and Q_2, we explored how two-node circuit motifs are enriched in these four-node circuits from top or bottom of either ranking. To do so, we enumerated the occurrence of any two-node circuit motif in all non-redundant four-node circuits. Here, for each circuit, the total number of motifs to count is $C_4^2 = 16$. We defined an enrichment score for each circuit motif as

$$E_a = \log\left(\frac{\sum_l 1 - H^-(Q_a, Q_{a0}, n)}{\sum_l H^-(Q_a, Q_{a0}, n)}\right) (5)$$

where $a = 1\ or\ 2$ for the two scores, $H^-(x, x_0, n) := 1/(1 + (\frac{x}{x_0})^n)$ is the inhibitory Hill function, Q_{a0} is the Hill threshold, selected as the Q_a value of the four-node circuit with the 600th ranking by Q_a. n is the Hill coefficient, selected as 20 to allow a sharp transition

of the factor H^- from 1 to 0 for Q_a near Q_{a0}. H^- is essentially a weighting factor: when n becomes very large, the Equation (5) becomes the log fold change of the occurrence of the circuit motif between the top 600 circuits and the rest of the circuits; a relatively small n, like 20, allows to consider the contributions of circuits with Q_a slightly smaller than Q_{a0}, to avoid the issue of zero counts. The summation from both the numerator and denominator in Equation (5) is over all non-redundant four-node circuits (l). See section "Statistical tests for circuit motif enrichment" for details of the statistical analysis.

A similar approach was applied to identify enriched coupling interactions between two two-node circuit motifs over all non-redundant four-node circuits. The coupled two circuit motifs can be classified as *overlapping*, for those that share same node, and *non-overlapping*, for those that do not share same node.

2.2.4 Grouping Scheme of two-node circuit motifs

We classified all 39 two-node circuit motifs into five groups and investigated how the grouping of the two-node circuit motifs contribute to specific gene expression state distributions (**Figs 3.3 and 3.4)** . The group designations were defined based on the number and sign of interactions between the two nodes. Group 1 (in blue) contains circuits with one activation between genes (motifs 1,2,3,4,5,6,7,8 and, 9). Group 2 (in purple) contains circuits with one inhibition between genes (motifs 10, 11, 12, 13, 14, 15, 16, 17, and 18). Group 3 (in red) contains circuits with mutual activation (motifs 19, 20, 21, 23, 25, and 26). Group 4 (in green) contains circuits with mutual inhibition (motifs 22, 24, 27, 37, 38, and 39). Group 5 (in orange) contains circuits with both activation and

inhibition between genes (motifs 28, 29, 30, 31, 32, 33, 34, 35, and 36). This grouping scheme was annotated on the histograms of single-motif enrichment analysis and the heatmaps for two-motif enrichment analysis.

2.2.5 Ks-distance metric

To quantify differences between RACIPE-simulated gene expression data of two four-node gene circuits (denoted as a and b), we defined a new distance function d_{ab} as:

$$d_{ab} = \sum_{i=1}^{4} D(x_{i,a}, x_{i,b}) + \sum_{i=1}^{3} \sum_{j=i+1}^{4} D(x_{i,a} \odot x_{j,a}, x_{i,b} \odot x_{j,b}) \quad (6)$$

where $x_{i,a}$ is the expression vector of gene i for circuit a, $x_{i,a} \odot x_{j,a}$ is the Hadamard product (element-wise product) between the expression vector of gene i and that of gene j for circuit a, $D(x, y)$ denotes the Kolmogorov-Smirnov statistic[68] between the cumulative distribution of x and y.

Furthermore, as the order of the genes in the circuits are arbitrarily assigned, an additional step was required to map the genes of the two circuits. To do so, we compute all 24 d_{ab} where we used the default gene order for the first circuit and a permutation of gene order for the second. The lowest d_{ab} value was eventually selected as the final distance.

2.2.6 Identifying circuits with similar state distributions

Using the above-defined distance function, we constructed a matrix of pairwise distances for all non-redundant four-node gene circuits. Starting from a circuit a, we can identify other circuits b whose d_{ab} are the among the lowest values – these circuits are supposed to have similar state distributions. See section "Statistical tests for selecting top-ranked circuits associated with a gene expression state distribution" for details of the statistical analysis.

To identify clusters of four-node circuits with similar state distributions, we adopted a subsampling approach to generate an association matrix for all the non-redundant four-node gene circuits. We performed Louvain clustering (cluster_louvain function in the igraph R package [69]) of a randomly-selected subset of 1,000 circuits for 100,000 repeats using d_{ab} as the distance function. For every two-circuit pair, the corresponding element of the association matrix was defined as the ratio of the occurrence of the two circuits appeared in the same subset and the occurrence of the two circuits clustered together. From the association matrix we applied the Louvain clustering method again, from which we identified seven major circuit clusters with more than 500 members. For each major circuit cluster, we defined the most representative circuits as those whose state distributions have d_{ab} of 0.05 or lower to the center circuit (circuit with lowest average KS distance to all other circuits in the community).

2.2.7 Modifying distance metric for single-cell

The scRNA-seq data of human glutamatergic neuron differentiation was processed using the Velocyto pipeline, as described in the original paper [70]. Velocyto is a tool used to study RNA velocity and only the initial RNA-processing steps were used to reproduce the data from the original paper. In brief, genes were initially filtered on the basis of 30 minimum counts and detected in over 20 cells. The top 2000 genes were then selected by non-parametric fit of CV versus mean. Another filtering step was applied to keep cells with more than 25 unspliced counts in at least 20 cells, leading to 1448 genes. Normalization was performed by dividing the counts by the total number of molecules in each cell and then multiplied by the median number of molecules across each cell.

We modified the KS test to allow for comparison of the state distributions with different number of genes. Here, we performed the KS test to compare the first two principal components of the experimental data with each two-gene combination in a four-node circuits. The distance function between a gene circuit and the experimental data was defined as the lowest distance between all the two node combinations of the synthetic circuit and the experimental data. In this way, we compared the experimental data to all the nonredundant four-node circuits, from which we identified the top ranked circuits for motif enrichment analysis. See section "Statistical tests for selecting top-ranked circuits associated with a gene expression state distribution" for details of the statistical analysis.

2.2.8 Statistical tests for motif enrichment

The significance of the enrichment was determined by a permutation test, similar to some previous approaches [71]. A null distribution for each enrichment was created by shuffling

the ranking indices of circuits for each score and applying the enrichment test. This step was repeated 10,000 times and the original enrichment results were then compared to the null distribution to estimate the p-value. Adjusted p-values were then calculated by the BH method [72] for multiple hypothesis testing. The number of hypotheses is the total number of two node motifs 39. No methods were used to determine if the data met assumptions of the statistical approach.

2.2.9 Statistical tests for top-ranked circuit for gene expression state distributions

We estimated a p-value for the fit of the top ranked circuits to a gene expression state distribution (derived from a set of single cell gene expression profiles from either simulated or experimental data) by calculating the z-score (using base R) of the distance of each four-node circuit to the experimental data. Here, we regarded the distribution of the KS distances of all four-node circuits as the null distribution. A p-value ≤ 0.05 is reached when the z-score ≤ -1.64. No methods were used to determine if the data met assumptions of the statistical approach

2.2.10 Scores for Multiplicity and Flexibility

We developed scores to identify circuits that are both able to generate multiple states (multiplicity) and circuits able to respond to signals in their environment (Flexibility). For each gene circuit, we applied RACIPE to generate the steady-state gene expression profiles of 10,000 mathematical models with randomly generated kinetic parameters. As

mentioned above, RACIPE-simulated gene expression profiles from a biological network usually form robust clusters of gene expression patterns. However, the gene expression profiles from random gene networks are usually less structured[73,74]. To reflect the ability of biological GRNs in generating distinct cellular states, we defined a scoring function H, namely multiplicity, by the negative differential entropy[75] of the simulated gene expression distributions of the 10,000 models:

$$H = \langle \log(p_i) \rangle \sim \frac{1}{N} \sum_i \log\left(\frac{k}{NVR(k)^d}\right) \quad (7)$$

where p_i is the local density of model i, $\langle \cdot \rangle$ denotes average, N is the number of models, and the summation is over all simulated models. Here, the local density p_i is computed by the k-nearest neighbors (knn) estimator[76], where $R(k)$ is the Euclidean distance of gene expression profiles between k-th nearest model and the center model, d is the dimension of the gene expression space ($d = 4$ for any four-node gene circuit), and a constant scaling factor $V = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$. The multiplicity defined in Equation (7) can be interpreted by the mean log local density. The higher the overall local densities, the higher the H values. Moreover, in the situation of high local density, more gene expression clusters can be observed. This is consistent with our previous findings that the local densities of the gene expression profiles simulated from a stem cell gene regulatory circuit are overall larger than those from a random gene circuit[74].

We next defined the flexibility of a gene circuit, F, by the extent of changes in the gene expression distributions of 10,000 RACIPE models between the unperturbed and knockdown (KD) conditions. Here, a knockdown indicates the maximum expression of

the knocked down gene has been restricted to 10% in all models. More specifically, the flexibility F is defined as

$$F = \sum_{j=1}^{d} \sum_{l=1}^{d} e_l D(p_{l,0}, p_{l,j}) \qquad (8),$$

where the summations are over all gene nodes j (from 1 to the dimension d) and all principal components (PCs) l (from 1 to d). Here, principal component analysis is performed on the gene expression data of models from the unperturbed condition. e_l is the l-th eigenvalue, which we incorporated here to emphasize the changes along the largest PCs. We quantified the differences in gene expression distributions by D, the KS test[77] of the probability distribution of the data along each PC between the unperturbed condition ($p_{l,0}$ for the l-th PC) and the perturbed condition, in which gene j is knocked down ($p_{l,j}$). Here, we subset 10% models with the lowest production rates of the KD gene j to compute the distribution for the KD condition.

In addition, we also defined another scoring function for the combined multiplicity and flexibility. Since the circuits' multiplicity H and flexibility F have values in different ranges, we chose to rank circuits with a new score G, defined by the product of H and F:

$$\mathrm{G} = \mathrm{HF} \qquad (9).$$

2.2.11 Generating networks of different types and sizes

In order to determine the emergent behavior of motifs when combined in networks of different types and sizes, we programmatically generated three types of large GRNs – random, scale-free, and sequential networks. To generate the random networks, we first

built *skeleton* networks using standard network generation algorithms, and then each node in a skeleton network is replaced with a gene circuit motif of choice. In detail, we used the erdos.renyi.game function and the gnp or gnm method from igraph R package[78] to generate the skeleton networks with gaussian degree distribution and directed edges. For the gnm version of the random networks (denoted as *random ver1*), the total number of edges was set to equal the total number of nodes. Therefore, these GRNs are sparsely connected. For the gmp version of the random networks (denoted as *random ver2*), the probability of an edge occurring between nodes was set to 30%. Therefore, these GRNs are densely connected. From a skeleton network, half of the edges were randomly selected and designated as inhibitory edges, while the other half as excitatory edges. Afterwards, each node in the skeleton network was replaced by a randomly selected two-node circuit motif. Any source/target node from an edge in the skeleton network was replaced with a randomly picked gene from the replaced two-node circuit motif.

To generate the scale-free networks, we followed a very similar procedure. In detail, we used the *sample_pa* function from *igraph* R package to generate the skeleton networks with the power raw degree distribution and directed edges. Afterward, every node in the skeleton networks was replaced with a randomly selected two-node circuit motif.

The sequential networks were constructed by connecting the desired number of two-node circuit motifs one after another. First, an edge (either excitatory or inhibitory) was added to connect a randomly picked gene from the first motif (source) to a randomly picked gene from the second motif (target). Second, another edge (either excitatory or inhibitory)

was added to connect the other gene from the second motif (source) to a randomly picked gene from the third motif (target). We continued the procedure iteratively to connect all motifs.

Altogether, we generated networks of four types (random ver1, random ver2, scale-free, and sequential), five different network sizes (10, 20, 30, 40, and 50 nodes), and with three sets of circuit motifs as the building blocks. The first set of motifs contains the top three enriched motifs by multiplicity; the second set contains the top three enriched motifs by flexibility; and the third set contains all the above six motifs (see Results for details of the motif enrichment analysis). Each kind of networks were generated randomly for ten times; thus, we analyzed on a total of 600 large GRNs ($4 \times 5 \times 3 \times 10$).

2.3 Methods for NetAct

2.3.1 Selecting enriched TFs

For a comparison between two experimental conditions, we obtained a ranked gene list quantified by the absolute value of the test statistics (t statistics in microarray and Wald test statistics in RNA-Seq) from differential expression (DE) analysis[79], followed by gene set enrichment analysis (GSEA)[80] using our optimized TF-target gene set database (details in Chapter 4). Here, for each TF, the corresponding gene set consists of all its target genes. GSEA identifies important TFs whose targets are enriched in DE genes between the two conditions. The significance test is achieved through 10,000 permutations of the gene list names and TFs are kept for further analysis when q value is

below a certain threshold cutoff (0.05 by default). A C++ implementation of this version of GSEA, specifically for gene name permutations, has been provided in NetAct for fast computation. For multiple comparisons, a set of enriched TFs are first identified from each pairwise comparison and then a union of the multiple sets of TFs is considered.

In the database benchmark test, for each database, we computed the sensitivity and specificity values for different q-value cutoffs. The benchmark utilized data sets that contained KDs of individual TFs. Here, for each cutoff value, we defined the sensitivity as the proportion of data sets where the gene sets for the KD TFs were enriched with q-values below the cutoff value. We also defined specificity as the fraction of cases where the gene sets for the other TFs (non-KD TFs in the benchmark) were not enriched with q-values above the cutoff value. We then computed area under the ROC curve (AUC) using the DescTools R package[81].

2.3.2 Inferring activity of transcription factors

TF activity is inferred from the expression of target genes retrieved from the TF-target database. NetAct defines the activity of the selected TFs using two different schemes – one using only the expression of target genes and the other using the expression of both the TF and its target genes. The second scheme is only used for the situation of noisy target gene expression in order to filter out weakly correlated targets. The reason for filtering out lowly correlated targets is to ensure the activity calculation is accurate. For

each TF, the algorithm selects the better scheme according to their performance, as described below.

Without directly using TF expression: For each TF, its downstream targets are first divided into two modules using the Newman's community detection algorithm[82] on the pairwise Spearman correlation matrix of the target genes. Then, within each module some less-correlated genes are filtered out to improve the quality of the inference. Here, the filtering step is achieved as follows: (1) each target gene is assigned a vector of correlations with the other target genes, where the distance between two genes is calculated as the sum of squares of the correlation vectors of two genes. (2) k-mean algorithm (k = 1) is performed within each cluster to determine the center vector. (3) genes are filtered out if the distance between the genes and the center is larger than the average distance.

This step outputs two groups of genes – genes in one group are supposed to be activated by the TF, while genes in the other group are inhibited by the TF. Note, at this stage, the nature of activation/inhibition of the individual group is not yet determined. The activity of the TF is calculated as

$$A(TF) = \frac{\sum_{i=1}^{n} w_i g_i l_i}{\sum_{i=1}^{n} w_i} \quad (10),$$

where g_i is the standardized expression value of a target gene i, w_i is the weighting factor defined as a Hill function:

$$w_i = 1/[1 + (\frac{s_i}{s_0})^n] \quad (11),$$

where s_i is the adjusted p value from DE analysis for gene i, the threshold S_0 is 0.05, and n is set to be 1/5 for best performance. I_i is 1 if the corresponding gene belongs to the first group and -1 if it belongs to the second group. If the calculated TF activity pattern is not consistent with the TF expression trend (evaluated by Spearman correlation), both the sign of the two groups and the sign of the activity are flipped. According to our in-silico benchmark test, we found that majority of the targets in one group are activated by the TF, and majority of those in the other group are inhibited by the TF. For genes in the inhibition group, the higher the TF activity, the more the genes are suppressed. Thus, the formula in Equation (10) captures well the activity of TFs for their effects to both activating and inhibitory targets. We also explored a few other community detection algorithms[80,81,83] and found they produced similar results.

Using TF expression: For each TF, its downstream targets are first divided into two groups according to the sign of the Spearman correlation between the TF expression and the target expression. Similar to the previous scheme, in each group, target genes are filtered out if the correlation value is less than the average correlation of all the targets. The activity of the TF is also calculated using Equation 10.

Sign assignment for DE TF: For any DE TF (*i.e.*, there is significant difference in TF expression across cell type conditions) of interest, NetAct computes the activity values from both the schemes (with or without TF's expression), and selects the better way based on how well the activity values correlate with target expression. To this end, NetAct calculates the absolute value of Spearman correlation between the TF activity and

the expression of each target, and selects the scheme whose activity gives larger average correlations.

Sign assignment for non-DE TF: If the expression patterns of the identified TFs fail to show the significant differences between cell type conditions, a semi-manual method to assign the sign of activity can be adopted. Putative interaction partners between DE and non-DE TFs in the inferred network are identified using the Fisher's Exact Test between TF targets in the NetAct TF-target database. The most significant pairs are then cross referenced with the STRING database to identify instances of PPI. A literature search is then performed to identify the nature of the PPI, and the sign of the non-DE TF is adjusted based on the DE TF and the type of PPI. Note that the last step needs to be done manually for each modeling application.

2.3.3 Network construction

NetAct constructs a TF regulatory network using both the TF-TF regulatory interactions from the TF-target database and the activity values. (1) The network is constructed using mutual information between the activity values of two TFs with the direction of the edge coming from the database. (2) Interactions are filtered out if they cannot be found in the TF-target regulatory database (*i.e.*, D1). (3) The sign of each link is determined by the sign of the Spearman correlation between the activity of two TFs. (4) We keep the interaction between two TFs if their mutual information is higher than a threshold cutoff. With different cutoff values for mutual information, NetAct establishes networks of different sizes. To identify the best network model capturing gene expression profiles, we

apply mathematical modeling to each of the TF networks using RACIPE[67]. RACIPE takes network topology as the input and generates an ensemble of mathematical models with random kinetic parameters. By simulating the network, we expect to obtain multiple clusters of gene expression patterns that are constrained by the complex interactions in the network.

3 Modeling Small Circuits for Genomics

3.1 Introduction

One of the main questions in systems biology is to understand how complex gene regulatory networks perform their functions to control important biological processes, such as cell differentiation and cell division [84,59]. Over the years, researchers have focused on studying gene circuit motifs, defined as reoccurring small circuit topologies within larger biological gene regulatory networks [54]. It has been shown, by approaches in synthetic biology [55], computational systems-biology modeling [56,57], and experimental systems biology [58], that different gene circuit motifs exhibit distinct functions in creating and maintaining circuit states, driving state transitions, and processing signals. For example, an autoregulatory negative feedback loop is known to suppress gene expression noise [85,86]; a two-node toggle switch circuit can generate bistability [87,88]; and an incoherent feed-forward loop can achieve adaptation [89]. Although the dynamical behaviors of individual circuit motifs have been widely studied, it is still challenging to characterize the roles of the circuit motifs when they interact with other motifs, or when they present within a large biological network. Due to the presence of additional gene-gene interactions, circuit motifs may behave differently from the standalone motifs.

Understanding emergent behaviors arising from motif coupling will greatly improve our understanding of functionality of circuit motifs and larger networks in general.

Gene circuit motifs were classically identified by searching the topology of a large biological network, such those from *E. coli* and *S. cerevisiae*, for the presence of smaller circuit motifs [54,59,60]. Motifs are important when they are over-represented in biological networks compared to similarly generated random networks because this indicates that the motifs regulatory interactions don't arise through random chance. This approach usually only considers the frequency of circuit motifs' appearance, but not their functionality, for initial identification. This calls into question whether or not over-representation is important without considering function. To address this issue, recent studies [56,57,61,62] have been focused on identifying functionally relevant circuit motifs capable of producing specific dynamical behaviors using mathematical modeling and then analyzing them for enriched motifs. These types of approaches have been devised and applied to elucidate circuits capable of generating oscillations [90,91,92] and multiple stable steady states [57,93,94]. Ye et al. [57] identified three-node circuits capable of generating stepwise transitions between four states with limited reversibility. Analysis of these circuits allowed them to identify regulatory interactions controlling the development of T-lymphocytes[57]. Schaerli et al. [61] investigated circuits capable of stripe formation, identifying incoherent feed-forward loops and a two-node motif containing activation and inhibition as the critical motifs. However, there are a still a few questions remain to be addressed for a more general applicability. First, mathematical modeling of gene circuits is often performed with a set of fixed kinetic parameters or examined with parameters

sampled from a narrow range, limiting the robustness and accuracy of modeling methods in evaluating circuit behaviors. Second, there lacks a quantitative scoring method allowing the ranking of circuits for *any* desired functionality, or to measure *functional* similarities and differences between two circuits. Third, it is still challenging to evaluate motif coupling, *i.e.*, how one circuit motif interacts with another to produce the desired behaviors. The coupling of circuit motifs has been shown to play important roles in the overall behavior of gene circuits [95,96]. In particular, the role of circuit coupling may depend on the proportion of shared nodes between the two coupled circuit motifs [56]. Another recent study [97] developed a framework to identify over-represented connections of circuit motifs, termed hypermotifs, in existing biological, neuronal, social, linguistic, and electronic networks. To the best of our knowledge, no systematic quantitative analysis is available to statistically evaluate the functionality of circuit coupling.

To overcome these challenges, we devised a computational framework that allows robust discovery of causal gene circuit motifs and patterns of motif coupling by defining a quantitative score to identify circuits capable of achieving specific functions. Circuit functions can be anything related to the circuit dynamics or steady state distributions, e.g., gene expression allowing three state clusters, specific multivariant distribution of gene expression, and gene expression distributions derived from experimental single cell data. We performed the first-ever comprehensive analysis on all non-redundant four-node transcriptional regulatory circuits. Compared to previous studies on three-node circuits [57,61,98], our analysis has the following advantages. First, there are around sixty thousand non-redundant four-node circuits, which is still manageable to perform extensive

computational simulations and is sufficiently large for a robust statistical analysis. Second, analyzing these four-node circuits allows for the evaluation of the roles of individual two-node circuit motifs in larger circuits. Third, analyzing four-node circuits has a major advantage in evaluating the role of the coupling between two two-node circuit motifs. Having four nodes in the larger circuit, we can statistically evaluate whether two two-node motifs are likely to occur in a four-node circuit with or without sharing the same node. The analysis of four node circuits is important from a biological perspective because it allows for the understanding of how multiple gene pairs coordinate to achieve specific biological functions, which is infeasible from the typical analysis of three-node circuits.

To model the dynamical behaviors of all these circuits in a high throughput way, we applied our recently developed method, random circuit perturbation (RACIPE) [99], to simulate an ensemble of ODE models with randomly generated kinetic parameters and analyze the steady-state gene expression distribution from these models. RACIPE has been applied to elucidate the dynamics of synthetic gene circuits [87,95,100], gene networks regulating stem cell differentiation [101], cell cycle [102], B-cell development [103], and EMT [28,104] . These previous studies have shown that, despite having randomly sampled kinetic parameters and initial conditions, steady state solutions of models generally converge to distinct clusters of gene expression patterns representing the functional states of the circuit. Functional states to which most models converge represent state distributions of the circuit and define its overall behavior. Furthermore, these studies also show that topology has an instructive role in defining the state distribution. We have also previously

shown that RACIPE-simulated data resembles single cell gene expression data, yet another advantage for discovering biologically relevant circuits.

In the following, we show how our computational framework can be applied to identify all four-node gene circuits allowing a triangular three-state distribution and a linear three-state distribution. We then show how our enrichment analysis can identify the enriched two-node circuit motifs and the patterns of their coupling. Next, we show how our framework can be applied to identify 1) clusters of circuits with distinct gene expression functions and 2) circuits with similar state distributions to any other starting circuit. Finally, we show how our method can be applied to identify circuits motifs and their coupling responsible for experimentally observed single-cell gene expression state distributions.

We also apply this method to identify motifs for general attributes of functional networks. There are two features of a GRN worth looking into. First, a functional GRN needs to generate rich dynamical behaviors, *e.g.*, multiple steady states (i.e., multistability) and/or oscillatory states. As shown in earlier studies, random GRNs tend to generate less interesting dynamical behaviors than biological networks[73,74]. On the other hand, multistability is often required for a GRN model to capture a variety of cellular states during cell differentiation[74,105–108]. Second, a functional GRN needs to be sufficiently flexible so that the GRN can be controlled by extrinsic cell signaling or an environmental factor[109–112]. It is quite common that the activation of a signaling pathway can drive the transition of cellular states[113].Equally typical are gene knockdown/knockout experiments

designed to understand the functions of genes based on the effects of gene perturbations[114]. Thus, functional GRNs need to be flexible, even in the presence of a certain level of compensation and adaptation due to network redundancy[115,116]. Therefore, it is reasonable to hypothesize that a functional GRN is required to produce rich dynamics and meanwhile be flexible upon perturbations.

Under this conceptual framework, we explore four-node nonredundant gene circuits that are responsible for multiplicity (*i.e.*, being rich in dynamical behavior), flexibility (*i.e.*, being versatile to alter gene expression), or both. From the identified small circuits, we will determine the most reoccurring two-node circuit motifs and the propensity of co-occurrence of two circuit motifs. Furthermore, using the identified circuit motifs, we generated a variety of large GRNs of different types (linear, scale-free, and random) and different sizes, from which we investigated the contributing factors of the multiplicity and flexibility of large GRNs. We hope that the outcomes of these analyses will shed light on the key regulatory interactions that allow GRNs to both generate multiple states and respond to environmental cues.

3.2 A Quantitative Evaluation of Topological Motifs and Their Coupling in Gene Circuit State Distributions

3.2.1 Identifying circuits with specific functions

In this study, we devised a computational framework that enables us to quantitatively evaluate the functionality of transcriptional gene circuit motifs. We used statistical analysis of large ensembles of simulation data to identify circuits best able to perform specific functions, and then analyze those circuits to identify the associated functional units. A schematic overview of the framework is illustrated in **Fig. 3.1**.

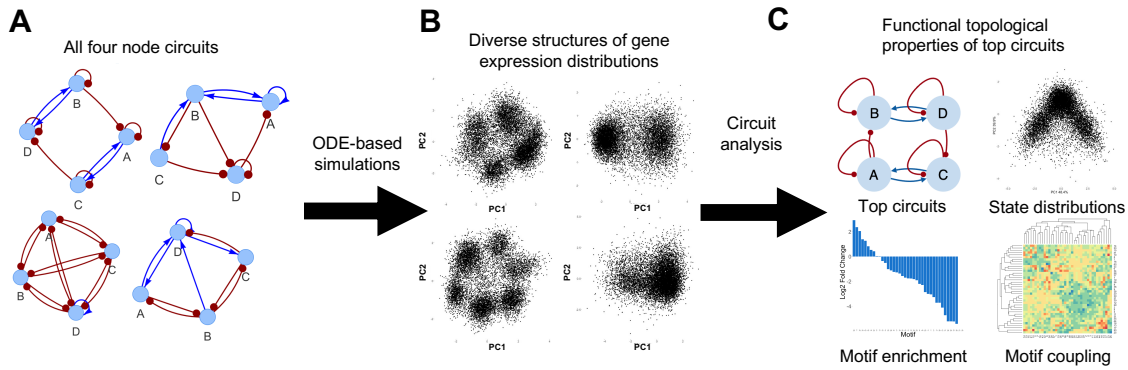

Figure 3.1. A schematic overview of circuit motif analysis. A) All non-redundant four-node gene circuits are first generated.(B) The dynamical behavior of these circuits are then explored using ensemble-based ODE simulations, resulting in diverse structures of state distributions.(C) This rich simulation dataset allows us to (1) identify circuits with a certain structure of state distribution, (2) identify reoccurring two-node circuit motifs and their coupling among these circuits, (3) quantify the similarity of circuit functions from state distributions.

First, we systematically generated all possible four-node gene circuits (**Fig. 3.1A**) (see section 2.2.1). for details of circuit generation) containing regulatory interactions of transcriptional activation/inhibition between nodes and autoregulation of individual nodes. Only circuits containing four functionally connected nodes were considered for analysis, excluding circuits equivalent to three or less nodes. Moreover, for redundant circuits, *i.e.*, circuits with the same topology but switched gene names, only one was included. Second, to each circuit, we applied RACIPE [99] to generate an ensemble of 10,000 ODE models with randomly generated kinetic parameters (see section 2.1 for

details of circuit simulation). Here, for a node that is transcriptionally regulated by multiple nodes, we assume that the effects of the transcriptional regulations from these nodes are independent, resembling AND logic. From the ensemble of mathematical models, we then evaluated the distribution of the steady-state gene expression (**Fig. 3.1B**). Such a state distribution can be interpreted as analogous to single-cell gene expression distributions driven by the specific gene circuit, incorporating the presence of cell-to-cell variability through the sampling of random kinetic parameters [99]. Different circuit topologies can often be associated with a variety of state distributions depending on the range of kinetic parameters explored, highlighting the need to explore a broad parameter space to better characterize the behavior of a circuit. Third, the core of our approach is to perform statistical analysis on the four-node circuits with similar state distributions (**Fig. 3.1C**). The circuit analysis allows the identification of enriched circuit motifs that are functionally associated with state distributions. We also extended the circuit analysis to identify patterns of coupling between two circuit motifs. We mainly focused on circuit motifs of two nodes, but this approach can be readily extended to analyze circuit motifs of other sizes.

3.2.2 Identifying circuits with three-state distributions

To illustrate the application of our circuit motif analysis framework, we evaluated four-node circuit topologies capable of generating a triangular arrangement of three gene expression states, as illustrated in **Fig. 3.2A**.

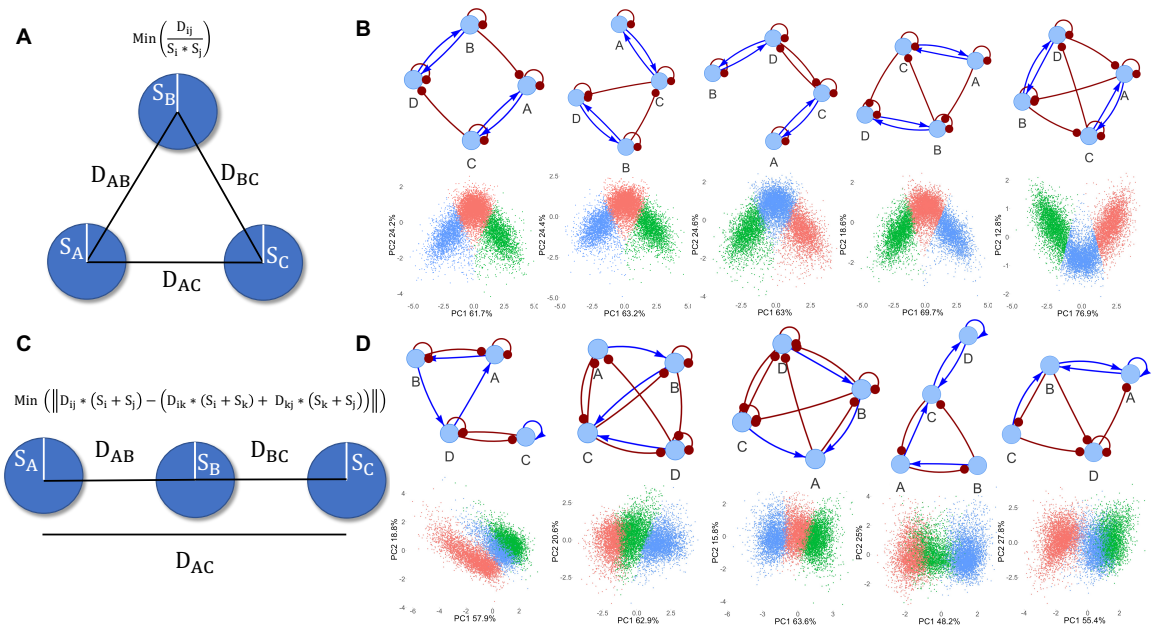

Figure 3.2. Identifying four-node circuits with triangular and linear state distributions (A) Illustration of the score defined to identify circuits with a triangular state distribution.(B) Illustration of the top five circuits with the highest triangularity scores. The plot shows the circuit diagrams (top row) and the scatterplots of the projection of four-dimensional RACIPE simulated gene expression from 10,000 models onto the first two principal components of the same data (bottom row). In the circuit diagrams, the lines and arrows in blue represent activating interactions; the lines and dots in red represent inhibiting interactions. In the scatterplots, red, blue, and green colors show the three clusters of gene expression states identified by k-means clustering (k = 3).(C) Illustration of the score defined to identify circuits with a linear state distribution.(D) Illustration of the top five circuits with the highest linearity scores.

This type of triangular state distributions is frequently observed in biological processes involving distinct cellular state transitions of multiple steps, *e.g.*, multi-lineage differentiation from a progenitor cell type to two distinct differentiated cell types, as is frequently observed in hematopoietic lineages [117,118]. For each four-node circuit, we applied k-means clustering (k = 3) to the RACIPE simulated gene expression profiles of all the non-redundant circuits and calculated the triangularity score Q_1, as defined in Equation (3) in the Method section. Higher Q_1 values indicate state distributions with a greater degree of separation between the three clusters. We then ranked all non-redundant

four-node circuits from high to low Q_1 values, with the top five ranked circuits illustrated in **Fig. 3.2B**. As demonstrated in the PCA projections of the simulated gene expression data of the corresponding circuits (**Fig. 3.2B**, bottom row), these circuits create gene expression state distributions of three states arranged in a triangular shape. Interestingly, the topologies of top ranked circuits are remarkably similar with clear patterns of two-node circuit motifs, such as motif 25 appearing twice in each network without sharing a node and motif 25 and 16 co-occurring while always sharing a node (**Fig 3.3A)** (see below for details).

Next, we explored four-node circuit topologies capable of generating three gene expression states arranged into a linear shape, as illustrated in **Fig. 3.2C**. This type of linear state distributions is frequently observed in the biological processes involving cellular state transitions through an intermediate state, such as transdifferentiation along a singular lineage during Epithelial-mesenchymal transition [119–121]. We performed the same k-means clustering analysis to the RACIPE simulated data, as described above, ranking all non-redundant four-node circuits from low to high Q_2 value, where Q_2 is a linearity score defined in Equation (4) in Chapter 2. The top five ranked circuits of linear state distribution are illustrated in **Fig. 3.2D**. We observed that these circuits can indeed produce a linear distribution of three gene expression clusters. The structures of the circuit topologies are similar among them but distinct from those allowing for a triangular state distribution. We observed repeated motifs containing activating and inhibiting edges between the two nodes, in stark contrast to the motifs observed for the triangular score. Taken together, our results demonstrates that the two scores, Q_1 and Q_2, are effective at detecting circuits capable of producing three states of triangular or linear state

distributions. While we have shown an application of this method to the linear and triangular structures, this analysis can be extended to any scoring function defining a particular state distribution, allowing a similar ranking analysis to identify circuits with novel features.

3.2.3 Novel Enrichment Analysis to identify circuit motifs and their coupling

Next, we evaluated the properties of the circuit topology for those with a triangular state distribution. We first enumerated all possible two-node circuit motifs and identified their occurrence in each four-node circuit. We evaluated the enrichment of each two-node motif in circuits with top triangularity scores (about top 1% circuits, see Chapter 2 for details), as shown in **Fig. 3.3A**.

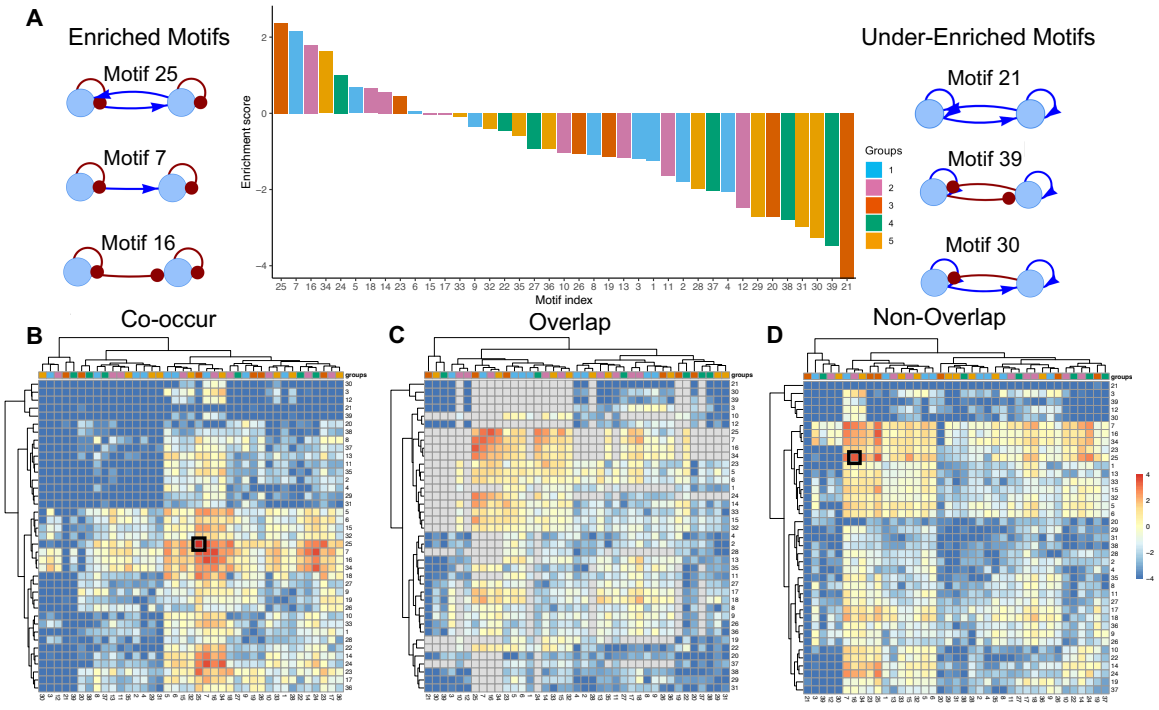

Figure 3.3. Motif enrichment analysis using the triangularity score (A) The enrichment score for all two-node circuit motifs using the triangularity score. All enrichment results are significant (adjusted p value <0.05) except for motifs 6, 15, 17, and 33. Panels.(B–D) show the heatmaps of the enrichment scores for the coupling between two types of two-node circuit motifs. The hierarchical clustering analysis was performed using Euclidean distance and complete linkage method. Interactions between motifs 25 and 16 are highlighted in black. Panel (B) shows the outcomes for all two co-occurring motifs; panel (C) for two co-occurring motifs that share the same node; panel (D) for two co-occurring motifs that do not share the same node. Colors in the column plot and the column annotations of the heatmaps represent different groups of motif types. Groups 1–5 represent two-node motifs with one activation between genes, motifs with one inhibition between genes, motifs with mutual activation, motifs with mutual inhibition, and motifs with both activation and inhibition between genes, respectively.

The topmost enriched circuit motif for the triangular state distribution is a circuit of two genes with both mutual activation and self-inhibition (motif 25). Interestingly, the top three enriched motifs all contain self-inhibition on both genes, suggesting the importance of the inhibitory autoregulation in generating three well separated states. Furthermore, the bottommost enriched circuit motifs are very different from the topmost motifs, in that the motifs are likely to contain activating autoregulation and inhibition between nodes. Interestingly, the most under-enriched motif, motif 21, is similar to motif 25 except that both nodes contain self-activation. Self-inhibition is known to suppress gene expression noise[54], and we observe that removing the negative autoregulation in the top ranked circuits would generate an additional state cluster. This evidence underscores the importance of negative autoregulation in producing the triangular three state distribution.

To understand how circuit motifs cooperate to generate triangular state distributions, we performed a similar enrichment analysis on the co-occurrence of two motifs among the same top ranked circuits. We can visualize the pattens of circuit coupling from the heatmap of enrichment scores for the co-occurrence of two motifs (**Fig. 3.3B**), for co-

occurrence of two motifs with a shared node (motifs with overlapping, **Fig. 3.3C**), and for co-occurrence of two motifs without any overlapping node (**Fig. 3.3D**). Interestingly, the topmost enriched motif coupling patterns are (1) between two of Motif 25 and (2) between Motif 25 and Motif 16. Furthermore, coupling between two of Motif 25 is the highest enriched motif pair for non-overlapping. This is consistent with what we observed in the top four-node circuits with the triangular state distribution – in the top 20 circuits there are 18 cases that contain exactly two motifs of #25 without a shared node. The positive feedback loop in motif 25 is known to generate bistability, and the interactions between two motif 25s in the top ranked circuit are mostly inhibitory interactions allowing the generation of more states. In addition, we observed that motifs 14, 17, 18, and 24 all have relatively higher enrichment for coupling with motifs 7 and 25 (see **Fig. 3.3B**), despite having relatively lower enrichment of as standalone motif (**Fig. 3.3A**). For motif coupling without a shared node (**Fig. 3.3D**), we observed surprisingly high enrichment between motifs 24 and 34, as well as between motifs 23 and 25. The relatively high enrichment of motifs 14, 17, 18, 23, 24, and 34 in the motif coupling indicates emergent behaviors for these motifs that contribute to the triangular state distribution only when coupled with other specific motifs (see **Chapter 2** for more detailed classification of circuit topology that features various configurations of motif coupling). We also analyzed the 600 top-ranking circuits from the triangular score and identified distinct enrichment of certain classes of four-node topologies.

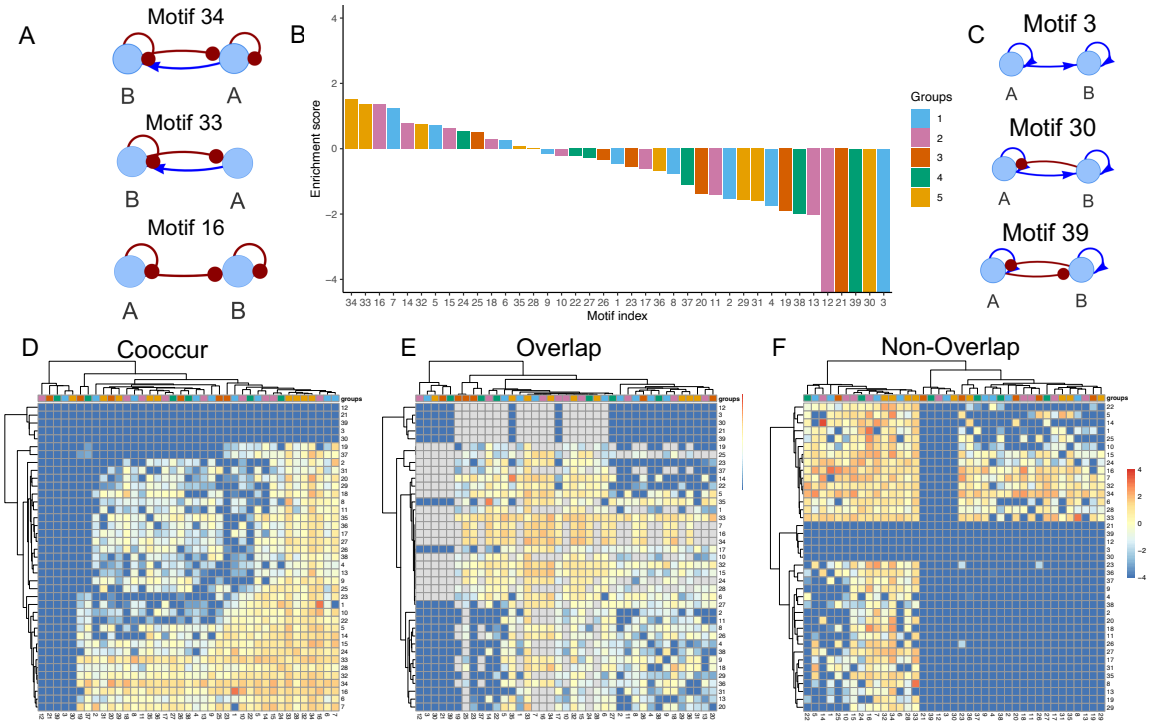

Figure 3.4. Enrichment results for the linear score comparing the top 600 of networks with the bottom 59612. A) Diagrams of the top three enriched two-node circuit motifs **B)** Enrichment scores for all two-node circuit motifs. Over-enriched motifs are present on the left while under-enriched motifs are present on the right. All enrichment was significant (adjusted p-value < 0.05) except for motifs 9, 10, 18, 22, 28, and 35. **C)** Diagrams of the three most under-enriched motifs. Panels **D-E** are heatmaps of the enrichment scores for motif coupling. Panel **D** shows the enrichment for motifs that co-occur in the top 600 networks; panel **E** shows the enrichment for motifs that share a node when co-occurring in the top 600; panel **F** shows the enrichment for motifs that do not share a node when co-occurring in the top 600. Colors annotated in Panels B, D, E and F show the grouping scheme of two-node circuit motifs.

We examined the properties of circuits capable of generating linear state distributions in a similar way to the analysis for the triangular state distribution. Enrichment of motifs in the top 600 of circuits ranked by the linear score were identified, as shown in **Fig. 3.4**. The topmost enriched motif for the linear score, motif 34, is characterized by two nodes with negative autoregulation and one excitatory and one inhibitory edge between nodes. The second top motif, motif 33, is similar to the topmost, however only one node contains negative autoregulation. When compared to the bottom three enriched motifs,

once again we observed a striking difference in the nature of auto-regulation; while the top motifs tend to contain negative autoregulation, the bottom motifs contain positive autoregulation. This may point to a general importance for negative autoregulation in the ability to create three state distributions. The ability of negative autoregulation to decreases gene expression noise[54] may contribute to the ability of the identified motifs to generate separate states. **M**otif 34 by itself is not able to generate multiple states; while from the top five circuits for the linear state distribution (**Fig. 3.2**), motif 34 is coupled with another motif 34 or the other top ranked motifs to generate positive feedback loops, allowing multiple states. We also noted that motif 39, a classic example of the self-activating toggle switch circuit capable of generating tristability [87,88,122,123], is identified as one of the bottom enriched motifs. We believe motif 39 is under enriched because our linear state distributions appear more continuous than those a generated by toggle switch with self-activation alone. These findings demonstrate that our method can detect quantitative differences in state distributions that are qualitatively similar (*i.e.*, continuous three state vs disparate three states) and therefore identify more specifically enriched motifs. The coupling of circuit motifs observed in four-node circuits favoring the linear score was shown in **Fig. 3.4D-F**.

Taken together, our data indicates that we can identify the quantitative contribution of key regulatory interactions, motifs and their coupling, responsible for producing specific structures of gene expression data (**Fig. 3.3**). We show how this can be applied to both linear and triangular arrangements of gene expression states; however, this approach can be expanded to any theoretical state distribution and identify motifs of other sizes.

3.2.4 Biological examples of triangular and linear state distributions

Next, we searched for the occurrence of the top enriched motifs in both the linear and triangular scores in PluriNetWork [124], a manually curated literature-based databases of transcription factor regulations for mouse pluripotency. We identified a total of 57 motifs from both the triangular and linear scores in the pluripotency gene regulatory network – 21 cases of Motif 5, 17 of Motif 6, seven of Motif 14, ten of Motif 15, 1 of Motif 32, and 1 of Motif 25. While our previous analysis suggest that multiples of Motif 25 are required for generating three states, the database may only contain one since not every interaction has been tested. Further experiments may potentially reveal other interactions between TFs in the form of Motif 25. It is also possible that multiple indirect interactions may form a coarse-grained motif 25 therefore allowing for three states. Among these motifs, Sall4 and Oct4 [125] form Motif 25 - a circuit of two genes with both mutual activations and self-inhibitions, as also supported by regulatory interactions from the TRRUST database [126]. Oct4 is a well-known transcription factor that plays critical roles in self-renewal and maintenance of pluripotency[127]. Oct4 has been considered a crucial pluripotency marker for many years with its expression being high in pluripotent stem cells and being down regulated upon differentiation. Furthermore, Oct4 depletion results in the loss of pluripotency in ES cells. Oct4 physically interacts with many proteins to activate or suppress its target genes during differentiation and binding activity of many TFs involved with maintaining the pluripotent state depend on the presence of Oct4[128]. Sall4 is a well-known binding partner with Oct4 and is important in stabilizing ESC self-renewal

[125,129,130]. Sall4 has been proposed to function as a bridge between Oct4 and NuRD complex, one of the major chromatin remodeling complex in cells[129]. The identification of motif 25 as the regulatory interactions between such important TFs in pluripotency supports that motif 25 is important for maintaining a three-state distribution. Furthermore, our findings suggest that the mutual activation between Oct4 and Sall4 plays a major role in maintaining three states in pluripotency. This proposed critical regulatory interaction could be directly tested through experiments that disrupt the transcriptional feedback loop created by the two TFs, potentially with the use of CRISPR/Cas9 tools that disrupt specific TF binding sites.

Tcf7 and Oct4 also form motif 23, a circuit of mutual activation with one node containing positive auto-regulation. Tcf7 is an important transcription factor that alone can restore trilineage differentiation abilities in mouse ESCs lacking all full length TCF/LEFs, demonstrating its importance in generating three states [131]. TCF7 is also known to be a transcriptional regulator downstream of Wnt signaling, and activated Wnt signaling has been shown to be sufficient for self-renewal in both human and mouse ESCs[132]. Once again, the identification of motif 23 between well-known regulators of pluripotency provide further support for its role in driving a three-state distribution. Similar to above, the importance of the regulatory interactions between TCF7 and Oct4 in maintaining the three-state distribution required for pluripotency could be tested by disrupting the transcriptional feedback loop between the two TFs.

Note that the PluriNetWork database may have many missing interactions (*e.g.*, only two genes have negative autoregulation) or annotations (*e.g.*, interactions labeled as unknown), therefore the occurrence of circuit motifs are likely underestimated. It is worth further investigating the roles of these circuit motifs in stem cell differentiation.

3.2.5 Identifying different types of state distributions

In the previous sections, we relied on defining a score to quantify a particular structure of state distribution and then ranking circuits with the score. Here, we aimed to classify gene circuits by structures of state distributions with an unsupervised top-down approach. To achieve this, we defined a distance function to quantify the differences in state distributions between two four-node circuits, and then applied clustering analysis on the resulting distance matrix. The distance function we chose was based on a multivariate Kolmogorov-Smirnov (KS) statistic, which allows the quantification of the differences between the gene expression distributions of two circuits (details in Chapter 2). We then devised a subsampling approach of Louvain clustering (schematic overview in **Fig. 3.5A**, details in Methods) for the distances between all non-redundant four-node circuits, from which we identified 20 circuit clusters with more than 10 members representing distinct classes of state distributions.

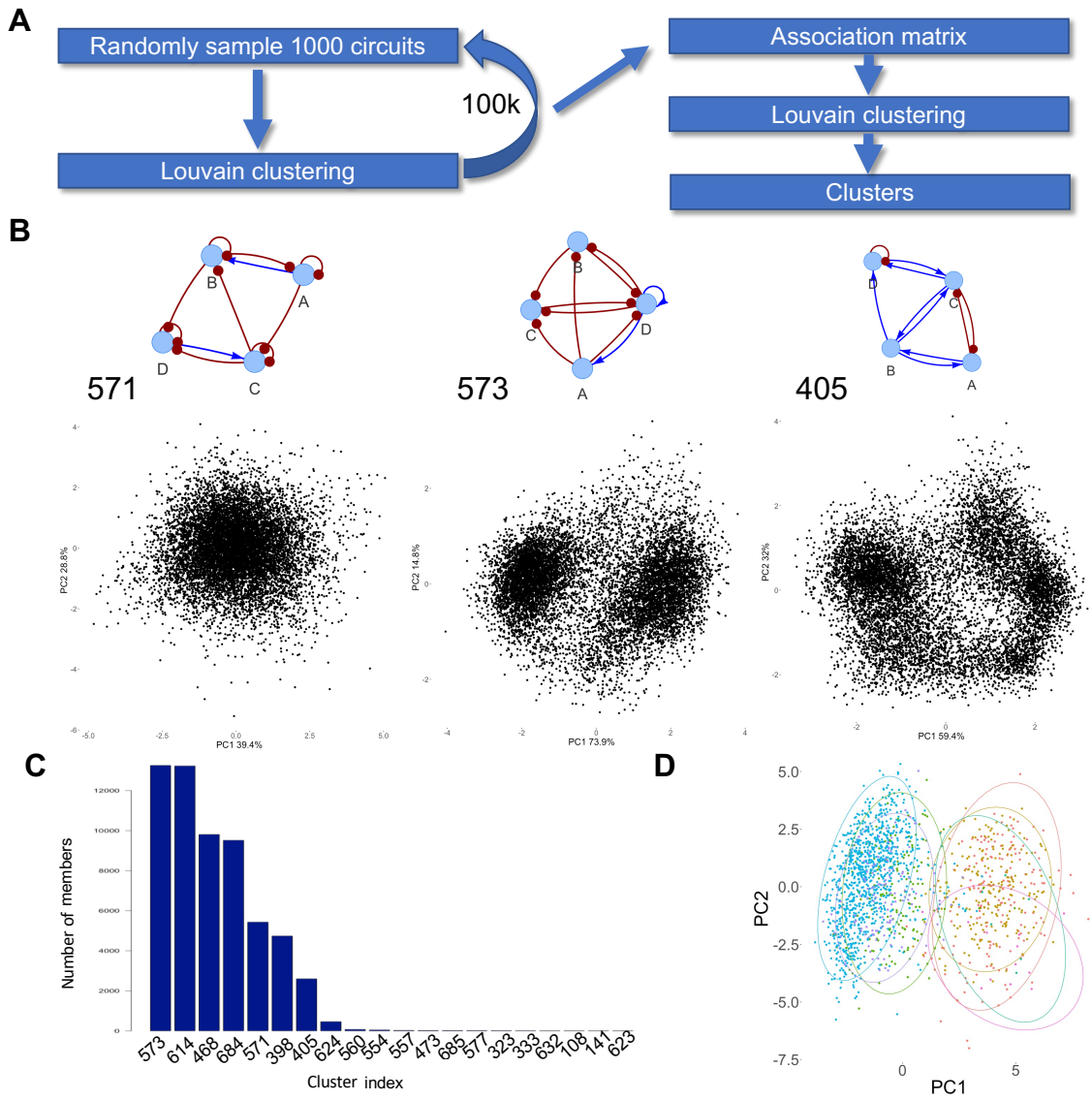

Figure 3.5. Clustering of all non-redundant four-node gene circuits by the similarity of state distributions (A) Flow chart of the clustering analysis. A subset of 1,000 circuits was randomly sampled for Louvain clustering. This step was repeated for 100,000 times to generate sufficient data for constructing an association matrix. The Louvain clustering method was applied again on the association matrix to obtain circuit clusters.(B) The center circuits (circuit diagrams in the top right) for three clusters (leftmost: single state; middle: two states; rightmost: circular state distribution) and the corresponding state distributions from the RACIPE simulations (scatter points in the second row). The numbers at the top left corner are the cluster indices.(C) The histogram of the number of circuits in each community with more than 10 members.(D) Projection of the summary statistics of the most representative circuits of every cluster onto the first two principal components. The clusters are illustrated by the ellipses of different colors.

For example, **Fig. 3.5B** shows representative state distribution from three different clusters – one allowing a single gene expression state (leftmost), one allowing two separated gene expression states (middle), and another allowing a circular state distribution (rightmost). From the histogram of the number of circuits in each cluster (**Fig. 3.5C**), we observe seven major circuit clusters. Interestingly, circuit clusters with more members tend to have simpler state distributions (*i.e.*, distributions with one or two gene expression clusters), while clusters containing more complex structures (*e.g.*, those with six gene expression clusters) often contain less members.

Lastly, we generated an overview of the major circuit clusters using PCA. To do so, we constructed a vector of statistics summarizing the expression of each circuit (details in the Methods section), for the most representative circuits in the largest seven clusters and projected the data to the first two principal components (**Fig. 3.5D**). Different colors and ellipses in the PCA projection illustrate the seven major circuit clusters identified from the Louvain clustering. These circuit clusters form two groups, which are well separated by the first principal axis (PC1). The circuit clusters on the left side of PC1 corresponds to the circuits capable of generating single state distributions, while the circuit clusters on the right side of PC1 corresponds to the circuits capable of generating state distributions with multiple states. Our results from the Louvain clustering seem to provide richer details of circuit behavior than those from the PCA, while the PCA results show the relationship between the identified circuit clusters. Taken together, this top-down approach allows us to identify major classes of circuits associated with distinct state

distributions and identifies multistability as the greatest difference between the four-node circuits.

3.2.6 Identifying multiple circuits with similar state distributions

In the previous section, we had defined a KS statistics-based distance function to quantify the difference between the state distributions from two four-node circuits. This distance function also allows comparison of any two gene expression state distributions, making it possible to identify all non-redundant four-node circuits that have the closest state distributions to any other circuit's state distribution. Two examples are illustrated in **Fig. 3.6**.

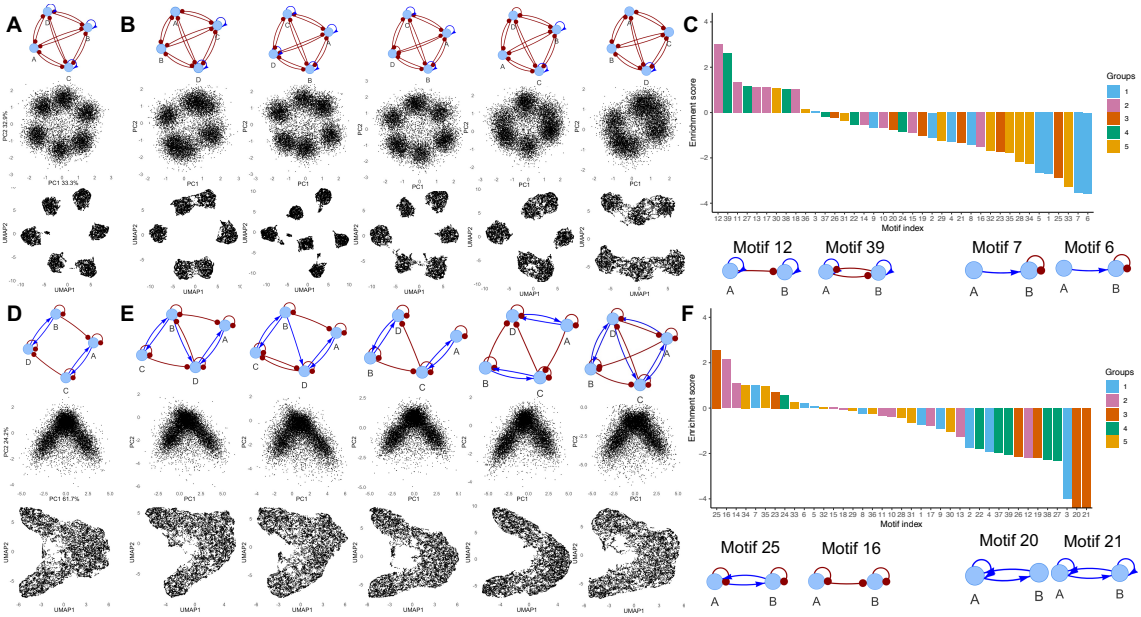

Figure 3.6. Identifying circuits with similar state distributions. A distance function based on Kolmogorov-Smirnov statistic was designed to quantify the similarity of the state distributions of two four-node gene circuits. The distance function allows us to identify other circuits with similar state distributions to a reference circuit. From these

identified circuits, enrichment analysis can be applied to identify reoccurring two-node circuit motifs and their coupling. Two examples are illustrated in the plot. Panels (A) and (D) show two references circuits – (A) for a circuit allowing six states along a circle, and (D) for a circuit allow three states in a triangular shape. Panels (B) and (E) show the five most similar circuits. Each column in panels (A), (B), (D) and (E) shows the circuit diagram (first row), the PCA projection of RACIPE simulations (second row), and the UMAP projection of the same data (third row). Panels (C) and (F) show the enrichment scores of two-node circuit motifs among the top 600 similar circuits (first row) and the circuit diagrams of the most over- (second row, left side) and under- (second row, right side) enriched motifs.. All enrichment results in panel (C) are significant (adjusted pvalue <0.05) except for motif 3, and all enrichment results for panel (F) are significant except for motifs 5, 8, 15, 18, 29, and 32. Colors in the column plots represent different groups of motif types

In the first case, we started with a circuit with a state distribution of a ring of six states (**Fig. 3.6A**). We show the top five circuits with the closest state distributions, based on the described distance function (**Fig. 3.6B**). PCA (2nd row in **Fig. 3.6B**) and UMAP (3rd row in **Fig. 3.6B**) projections show that resulting state distributions from identified circuits indeed contain similar gene expression state distributions.

We note that some of the gene-expression states may overlap in two-dimensional projections (typically with PCA), however, separation of these states usually can be discerned in other dimensions or with the projection of another method, such as UMAP. Strikingly, the identified circuits share very similar topologies – all the top five circuits have mutual inhibiting links between any two nodes and differ only by the autoregulatory links. This is in line with our previous findings that certain circuit topologies are required to generate specific structures of state distributions. Next, we performed enrichment analysis on circuits with lowest distances (top 600 similar circuits, all with p-value ≤ 0.05; details in Methods), from which we identified motifs 39 and 11 to be most enriched in these circuits. Here, motif 39 consists of a toggle switch with self-activation, a motif

well known to generate multiple distinct states. Motif 25, the top enriched motif for the triangularity score, is one of the least enriched motifs in this case. The results indicate that the enrichment analysis allows identification of the circuit motifs responsible for disparate state distributions. Furthermore, we observed an emphasis on positive autoregulation in the top identified motifs, which is a trend that is distinct from what was observed for earlier scores.

As a second example, we started with a circuit with a triangular state distribution of three clusters (**Fig. 3.6 D**). This state distribution was previously described by the triangularity score defined earlier (**Fig. 3.2**). With our current analysis, we successfully identified circuits with the most similar state distributions, without using the triangularity score (**Fig. 3.6E**). In this case, the top identified circuits, despite being very similar to one another, are structurally more different among them than those from the top six-state circuits shown (**Fig. 3.6B**). This could be because the triangular state distributions are more commonly observed state distributions, thus more accessible to circuits of different topologies. The top three enriched motifs, as shown in **Fig. 3.6F**, are the same motifs as identified from our earlier analysis using the triangularity score (**Fig. 3.3**). In particular, the same motif 25 was identified once again as the topmost enriched motif for the triangular state distribution. These outcomes demonstrated the effectiveness of the distance function in identifying circuits and motifs responsible for similar state distributions. Furthermore, we have shown with two orthogonal methods that motif 25 is implicated in generating a triangular distribution of three states.

3.2.7 Identifying circuits, motifs, and coupling of neuron differentiation

We have demonstrated how our approach can identify four-node circuits with similar state distributions to other circuit's state distribution. Now, we extended our analysis to identify four-node circuits with similarities to experimentally observed state distributions from single-cell RNAseq (scRNA-seq) data. This is a conceptually different analyses in that (1) the state distributions from single cell data are commonly derived from many more genes (typically the most variable genes); (2) nodes in the four-node circuits do not necessarily represent individual genes, but a contribution of a group of genes due to the potential modular structure and redundancy observed in large gene networks[133,134]. In other words, the four-node circuits in the current study represent phenomenological models of the data. We considered circuits of four nodes here to take advantage of all simulation data generated in the current study, but this approach can be readily extended to circuits of other sizes. We also revised the KS statistics-based distance functions to enable the circuit and motif analysis for single-cell gene expression data (details in Chapter 2).

Our method was applied to a set of scRNA-seq data from 1,720 cells of human glutamatergic neuron differentiation at week 10 post-conception[70]. **Fig. 3.7A** shows the PCA projection of the expression of 1448 genes to the first two principal components.

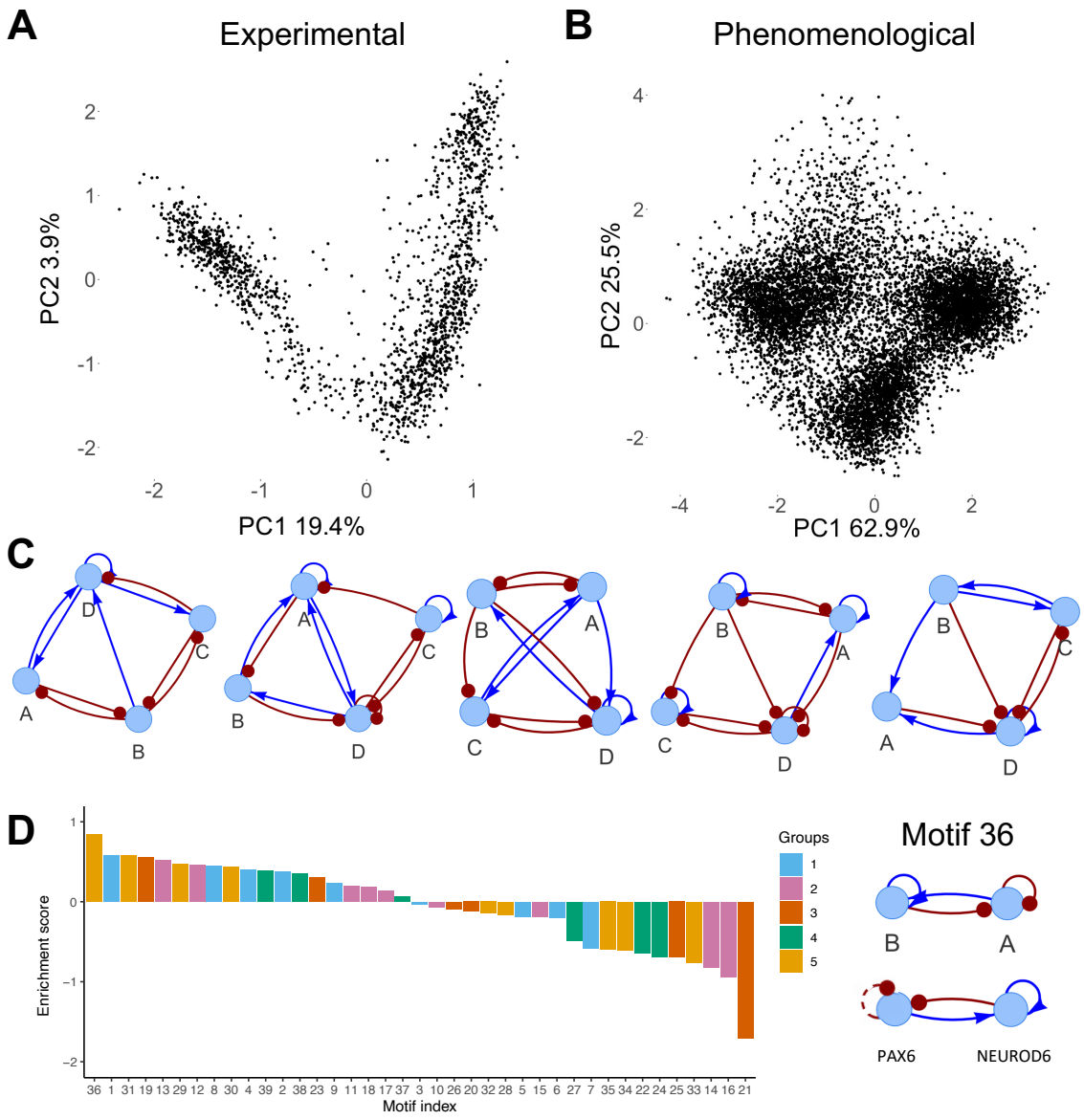

Figure 3.7. Application of the circuit analysis to scRNA-seq data of human glutamatergic neuron differentiation. (A) The projection of the scRNA-seq data to the first two principal components.(B) The projection of the simulated gene expression data of the top ranked four-node circuit to the first two principal components of the simulated data. (C) The diagrams of the top five ranked circuits, with the top ranked circuit on the left and the fifth ranked circuit on the right. (D) The enrichment scores of two-node circuit motifs among the top 218 circuits (left panel), and the diagrams of the most enriched two-node circuit motif and a biological gene circuit (right panel). All enrichment results are significant (adjusted pvalue <0.05) except for motifs 3, 5, 10, 20, 26, and 37. Colors in the column plots represent different groups of motif types.

While a snapshot of gene expression during neuron differentiation, this dataset contains at least three states, consisting of radial glia progenitor cells progressing through intermediate neuroblast stages to differentiated neurons[70]. From the circuit analysis we identified 218 top ranked phenomenological four-node circuits (p-value ≤ 0.05 or z-score ≤ -1.64), the top 5 of whose circuit diagrams are shown in **Fig. 3.7C** and the state distribution of the topmost circuit in **Fig. 3.7B**. The circuit's state distribution resembles that of the single-cell data in that both contain three gene expression clusters with the rightmost two clusters more connected than the leftmost cluster. These clusters potentially correspond to radial glial progenitors, intermediate progenitors, and differentiated neurons. This could be confirmed by comparing the top contributing genes in the PCs that were matched to each node and then identifying which nodes are high in each cluster for the modified KS score (section 2.2.7). If the top contributing genes for the nodes that are high in each cluster match experimental observations, this would indicate that the simulate and experimental distributions are indeed similar. However, the clusters from the circuit simulations appear more spherical while the experimental clusters appear more ellipsoid, presumably because of the wider range of kinetic parameters sampled in RACIPE simulations than those represented for the single cells in the experiment.

From the circuit motif analysis of the top-ranked 218 circuits, we identified the toggle switch (motif 19) among the topmost enriched motifs (**Fig. 3.7D**). Interestingly, although the toggle switch circuit with two-sided self-activations (motif 39) was only moderately enriched, and motif 5 was even under enriched, the coupling of motifs 39 and 5 without a shared node was the most enriched coupling among the top-ranked circuits. The top

identified motif, motif 36, is characterized by a self-inhibiting node A and a self-activating node B, where A activates B, and B inhibits A. Upon analysis of the PCA loadings of the experimental data, we identified *VIM* as the highest negative contributor and *STMN2* and *NEUROD6* as the highest positive contributors to the first principal component. This is consistent with the experimental observation[135] that VIM serves as a marker gene in radial glial population (leftmost cluster), and *NEUROD6* and *STMN2* are important transcription factors in intermediate progenitors and differentiated neuron populations (two clusters from the right side). We also identified *PAX6* as one of the top contributors to the radial progenitor population. *PAX6* is a TF known to play an important role in radial glial cell differentiation that activates neuronal lineages (while repressing others) to ensure correct differentiation to neurons [136]. Furthermore, it has been shown that decreasing Pax6 expression is required to turn off the neural stem-cell self-renewal program [137]. In addition, *NEUROD6* has also been shown to be implicated in sustaining the gene expression program of neurons and for promoting differentiation by triggering cell cycle withdrawal [138,139].

Remarkably, the top identified motif (#36) is consistent with the regulatory interactions responsible for neuronal cell differentiation [135–139]. In this circuit motif (bottom right circuit diagram in **Fig. 3.7D**), one node (network involving *PAX6*) decreases its own expression and activates another node (network involving *NEUROD6),* which activates itself and represses the other node. In glutamatergic neuron differentiation, radial glial cells divide symmetrically or asymmetrically to self-renew or produce neurons, respectively. Pax6 positive cells mark radial glial cells that will go on to differentiate first into intermediate progenitors and then become neurons[140]. The transition to intermediate

progenitor cells is initiated by the expression of pro-neural genes. Pax6 binds to the enhancer region of *Neurog2* and directly induces it expression, which downregulates the expression of Pax6. The activity of Pax6 inhibiting its own expression is reflected by the negative auto regulation observed on node A. The transition to intermediate progenitors is initiated by the elevated expression of Tbr2, which is induced by Pax6. This interaction is reflected in the activating edge from node A to node B (**Fig. 3.7D**). NeuroD6 positive progenitor cells are committed to the glutamatergic neuron fate and NeuroD6 has been shown to be able to independently carry out key steps of neuronal differentiation[138,141]. Expression of NeuroD6 is induced at the start of differentiation and proceeds withdrawal from the mitotic progenitor populations. This is consistent with the cell cycle arresting function of other NeuroD6 family members and is reflected in the inhibitory edge from node B to node A. NeuroD6 has also been shown to sustain the neuronal gene expression program and even promote long term neuron survival upon serum deprivation[142]. NeuroD6 binds to its own promoter and sustains its own expression, reflected in the positive autoregulatory interaction on node B[143]. Taken together, this indicates that the top motif identified as causative for the specific state distribution observed during glutamatergic neuron differentiation agrees with literature evidence. Furthermore, our results suggest that the interaction between the gene expression programs involving NeuroD6 and Pax6 are of paramount importance for maintain the state distribution required for proper differentiation of glutamatergic neurons. We note that the identified circuits/motifs are phenomenological with some of the interactions present in our motif not being direct interactions due to how we modified the KS-score to compare PCs to

each four node circuit. This is significant because each PC represents a combination of genes and not just a single factor, leading to potentially indirect interactions.

Interestingly, despite the general triangular shape of the experimental state distribution, these enriched motifs identified here were not the same as those found in the previous analysis of the triangular state distributions, suggesting that the circuit and motif analysis can recognize subtle aspects of the state distribution, such as the state locality and densities. In summary, we demonstrated the circuit analysis can be applied to experimental scRNA-seq data to identify phenomenological gene circuits capable of recapitulating experimentally observed state distributions.

3.2.8 Discussion for A Quantitative Evaluation of Topological Motifs and Their Coupling in Gene Circuit State Distributions

In this study, we have developed a novel computational framework to identify gene circuits, small circuit motifs, and coupling of motifs responsible for circuit properties by evaluating their gene expression state distributions. This method can be readily generalized to model other dynamical behavior of a circuit, as long as it can be quantified by a scoring function. Our method employs the first comprehensive analysis of all four-node transcriptional regulatory circuits. We have shown how the methodology can be applied to identify circuits allowing triangular or linear state distributions, from which we can further characterize the enriched motifs and motif coupling. We have also defined a KS statistics-based distance function to quantify the differences of the state distributions

between two circuits. Using this distance function, we have identified major classes of circuits with distinct state distributions, circuits with similar state distributions to other circuits, and circuits that recapitulate experimental gene expression distribution from single-cell gene expression data.

Our circuit and motif analysis has the following advantages over existing methods. First, conventional approaches defined motifs as overrepresented small circuit topologies from a large biological network. The function of the identified motifs was then analyzed by mathematical modeling and/or synthetic biology analysis of a standalone circuit motif. While this approach helps to build a fundamental understanding of motifs and their importance, it falls short to discover circuit motifs for a particular function in mind. With our approach we start out by defining a desired circuit property (such as a state distribution) and then identify two-node circuit motifs enriched in all non-redundant four-node circuits with shared features. Other recent studies [57,61] also utilized this motif identification strategy, however the current study provides a more quantitative and generalized methodology. While we demonstrate this with a comprehensive analysis of four-node circuits identifying two-node motifs, the method can be readily adapted to identify larger circuit motifs. Therefore, our approach can alleviate the issues of existing approaches, allowing a more robust evaluation of gene circuits according to their behavior.

Second, our method utilizes RACIPE, an ensemble-based simulation approach, to evaluate circuit behavior. Compared to the earlier methods, RACIPE allows

consideration of variation in kinetic parameters present in different cells. RACIPE-simulated gene expression from an ensemble of random models are usually not randomly scattered in gene expression space but form robust clusters of models. As shown in previous studies, these clusters can usually be associated with biological relevant cellular states [28,101,102,104,144]. This definition of circuit states based on gene expression distribution is more robust compared to the conventional definition based on the steady states of dynamical systems. In this way, our approach ensures a more thorough exploration of circuit behaviors and the associated circuit motifs.

Third, our method can also be applied to infer phenomenological four-node circuit models that capture the gene expression distributions of experimental single cell data. Note that the nodes in the phenomenological models may represent the collective effects of multiple regulators, instead of individual genes[145,146]. We have shown its application to study glutamatergic neuron differentiation through its ability to correctly identify interactions between gene expression programs that are responsible for promoting cellular state transitions from radial glial cells to fully differentiated neurons. The interactions we identified are fully supported by literature and highlight the critical importance of both PAX6 and NeuroD6. From these results we hypothesize that disruption of the specific regulatory interactions we identified between PAX6 and NeuroD6 gene expression programs would lead to a breakdown of neuron differentiation, which could potentially be tested by CRISPR mediated disruption of transcription factor binding sites listed in section 3.2.7. We expect that this approach is invaluable to

elucidate the regulatory mechanism for systems with more complex structures of cellular states.

3.3 What makes a functional gene regulatory network? A circuit motif analysis

Here we apply the method developed in 3.2 to answer new questions: What are the circuits, motifs, and coupling interactions responsible for the ability of networks to have either high multiplicity or flexibility. Multiplicity is the ability for a network to create multiple steady states and is an important quality in networks modeling cellular state transitions, which typically include two or more states. Multiplicity is important in any biological context that includes multiple states with a prime example being development, where progenitor cells often bifurcate into two distinct lineages. Flexibility represents how much the state distribution changes upon perturbation of the network and represents the ability to respond to environmental cues. The ability to respond to environmental cues is important since activation of signaling pathways commonly drive the transition of cellular states.

My contribution to this project consisted of training the lead author on how to use the method, showing her how she could extend it, and developing an algorithm to create different types of networks, of any size, that incorporate any desired motifs.

3.3.1 Identifying circuits with high multiplicity

We applied the multiplicity scoring function to all 60,212 non-redundant four-node gene circuits. As shown in **Fig. 3.8A**, the distribution of multiplicity is a negatively skewed unimodal distribution, with slightly more circuits of high H values.

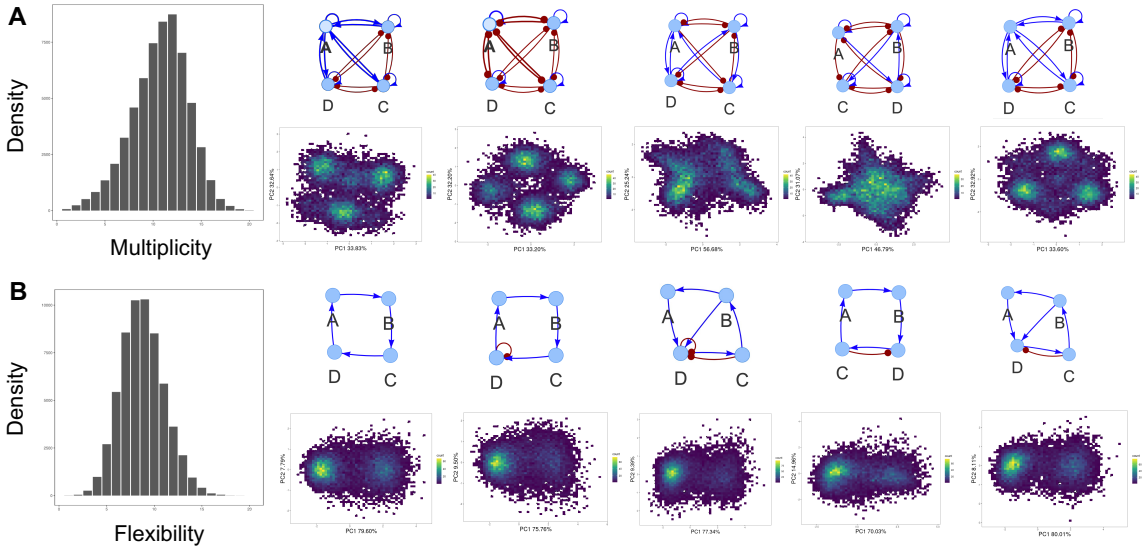

Figure 3.8. Multiplicity and flexibility of gene regulatory circuits. (A) The leftmost plot shows the histogram of multiplicity for all nonredundant four-node circuits. The right panels show, for the top five circuits ranked by multiplicity, the circuit diagrams (top row) and the density maps of RACIPE-simulated gene expression projected onto the first two PCs (bottom row). In the circuit diagrams, the nodes represent genes, labeled as A, B, C, and D. The blue lines and arrows represent excitatory regulations, and the red lines and dots represent inhibitory regulations. Panel B shows the outcomes for the flexibility score.

We found that the scoring function H is indeed effective in capturing circuit multiplicity. We observed that the topmost circuits ranked by H tend to contain regulatory links of mutual activations, mutual inhibitions, and self-activations. The tight regulatory connectivity in those circuits allows them to have a higher number of gene expression clusters, as shown in the density map of the simulated gene expression profiles projected onto the first two principal components (PCs) (**Fig. 3.8A**, right panels).

Next, we applied the circuit motif enrichment analysis by comparing the occurrence of any two-node circuit motifs within the topmost four-node circuits ranked by multiplicity with the occurrence within the rest circuits. As shown in **Fig. 3.9AB**, the top five enriched two-node circuit motifs all contain self-activations, suggesting its dominant role in determining high multiplicity.

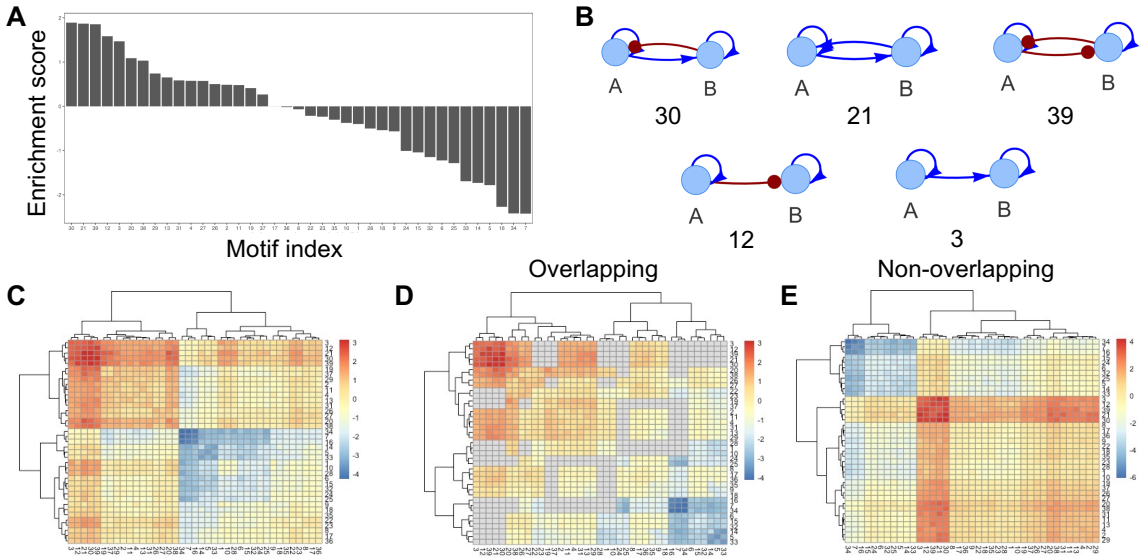

Figure 3.9. Circuit motif enrichment analysis with respect to circuit multiplicity. (A) Enrichment scores for all two-node circuit motifs were computed from all nonredundant four-node gene circuits ranked by multiplicity. The enrichment is significant for most motifs (adjusted *p* values <0.01), except for motifs #17 and 36. (B) Diagrams of the top five enriched circuit motifs. (C–E) Heatmaps of the enrichment scores for the co-occurrence of all pairs of two two-node circuit motifs. Three heatmaps correspond to the overall co-occurrence (C), the co-occurrence of two motifs with a shared node (overlapping, D), and the co-occurrence of two motifs without a shared node (nonoverlapping, E). Hierarchical clustering analysis was applied to each case with the Euclidean distance and complete linkage. There are gray colors in panel D, as some motif combinations do not exist.

The top three motifs (#30, 21, 39), which have similarly high enrichment scores, all contain mutual regulatory links. Contrarily, the bottom three motifs (#7, 34, 16) all contain self-inhibitions. These findings are all consistent with what we observed in the top-ranked four-node circuits in **Fig. 3.8A** and with previous studies showing that self-activation generates multi-stability, and self-inhibition stabilizes gene expression state[147–

[149]. Furthermore, we evaluated the enrichment of the co-occurrence of two circuit motifs within the top-ranked four-node circuits by the multiplicity, as shown in **Fig. 3.9CDE**. We observed again that two motifs with self-activation tend to be more enriched, while two motifs with self-inhibition tend to be less enriched.

3.3.2 Identifying circuits with high flexibility

We applied the flexibility scoring function to rank all 60,212 non-redundant four-node gene circuits. As shown in **Fig. 3.8B**, the distribution of flexibility is close to a symmetric unimodal distribution.

We tested the flexibility score F on a few four-node circuits and found that, for circuits with larger F, gene expression distributions have noticeably larger changes upon gene KD perturbations. We observed that the topmost circuits ranked by F tend to be sparsely connected. Compared to the topmost circuits ranked by H, the topmost circuits by F usually have a mono-directional interaction between two nodes (either the 1st node regulating the 2nd, or the 2nd node regulating the 1st), and there are fewer auto-regulations. Interestingly, the circuits with high flexibility usually have gene expression profiles of two clusters (**Fig. 3.8B**, right panels). This observation can be understood as follows. For circuits allowing only one gene expression cluster, the possible gene expression distribution is limited by the cluster. For circuits with a higher number of gene expression clusters, it is hard to transit through multiple states by perturbation. Thus, circuits with

two gene expression clusters are the most likely to achieve substantial changes in gene expression distributions upon perturbations.

Next, we applied the circuit motif enrichment analysis to circuits ranked by flexibility. As shown in **Fig. 3.10AB**, the first and second most enriched circuit motifs (#1 and 10) all contain a single regulatory link from one gene to the other.

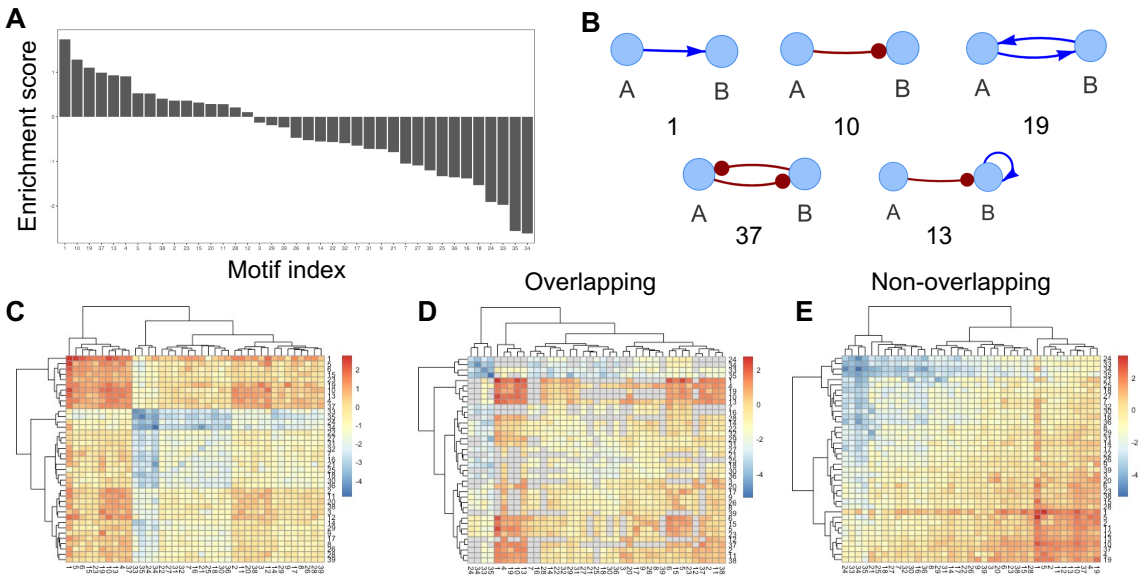

Figure 3.10. Circuit motif enrichment analysis with respect to circuit flexibility. Here, all nonredundant four-node gene circuits are scored and ranked by flexibility. The enrichment of a single motif is significant for all motifs (adjusted *p* values <0.01).

The third and fourth most enriched circuit motifs (#19 and 37) are circuits with mutual activation and mutual inhibition (toggle switch), respectively. We noticed slight differences in the enrichment scores between circuits with excitatory regulations and those with inhibitory regulations, presumably because of sampling deviations and the measurement of flexibility by knockdown perturbations. These most enriched circuit motifs usually do not prefer autoregulation. Interestingly, the toggle-switch-like circuit

motifs (#19 and 37) are frequently observed in the topmost flexible circuits, as they are known to generate bistability[150,151]. These motifs ensure circuits to be sparsely connected and bistable, thus allowing the whole circuit to be also flexible. Furthermore, we evaluated the enrichment of the co-occurrence of two circuit motifs within the top-ranked four-node circuits by the flexibility, as shown in **Fig. 3.10CDE**. Interestingly, we observed frequent co-occurrence of motif #1 with three motifs, #5, #6, and #1 itself. These three circuit motifs all share the same excitatory regulatory link from one gene to the other but differ by just a self-inhibitory link. The co-occurrence of these motifs further demonstrates the sparseness of regulatory interactions as one of the determining factors of flexible circuits.

3.3.4 Identifying circuits with high multiplicity and flexibility

In the previous two sections, we have evaluated the multiplicity and flexibility of all nonredundant four-node gene circuits and applied motif enrichment analysis to identify two-node circuit motifs associated with either high multiplicity or high flexibility. Furthermore, we evaluated the relationship between the multiplicity and flexibility of a circuit. From the density and scatter plot in **Fig. 3.11A**, we observed a weak correlation (0.17275 Pearson correlation coefficient) between these two scores.

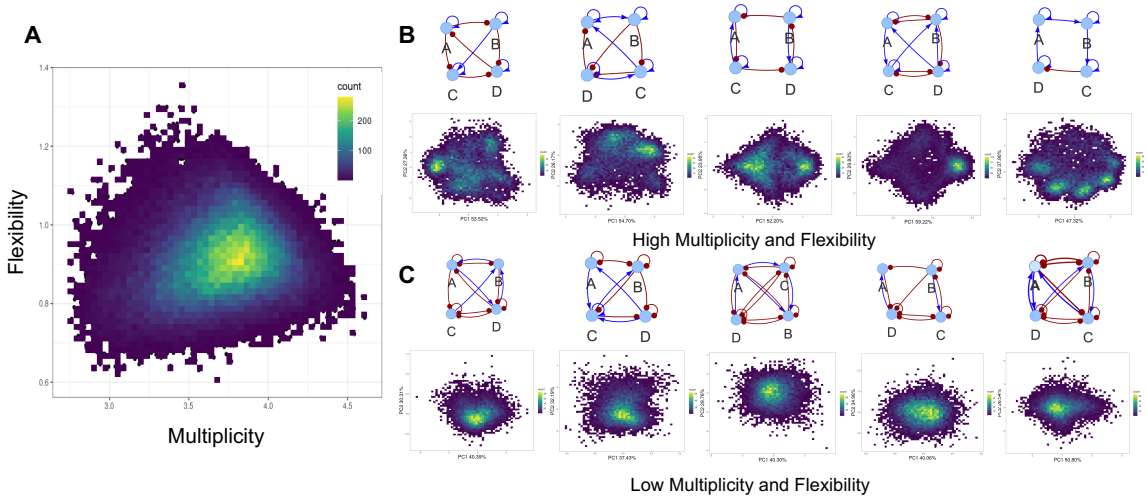

Figure 3.11. Gene regulatory circuits ranked by combined multiplicity and flexibility. (A) The density map of multiplicity (*x*-axis) and flexibility (*y*-axis) for all nonredundant four-node gene circuits. (B) The panels show, for the top five circuits ranked by the product of multiplicity and flexibility, the circuit diagrams (top row) and the density maps of RACIPE-simulated gene expression projected onto the first two PCs (bottom row). Panel C shows the outcomes for the bottom five circuits.

Interestingly, we found circuits rarely have high multiplicity and flexibility simultaneously (only 0.048% of circuits with both H and F higher than 1.5 standard deviations above the mean value). However, much more circuits were found to have high multiplicity and low flexibility (0.144% of circuits with H higher than 1.5 standard deviations above the mean value and F lower than 1.5 standard deviations below the mean value). More circuits were also found to have low multiplicity and high flexibility (0.332% circuits with H lower than 1.5 standard deviations below the mean value and F higher than 1.5 standard deviations above the mean value). It is reasonable that, despite no apparent correlation between multiplicity and flexibility, circuits with the highest multiplicity are less likely to be flexibility, while circuits with the highest flexibility are required to have fewer gene expression clusters, thus being low in multiplicity.

As we discussed earlier, we are interested in functional GRNs with both high multiplicity and flexibility, despite their low occurrence in four-node circuits. We applied the circuit motif analysis with the combined score defined in Equation (3). The top five ranked circuits, as shown in **Fig. 3.11B**, have the following features. First, these circuits all have relatively simpler circuit topologies compared to the topmost circuits ranked by multiplicity. Second, these circuits all contain multiple self-activations, thus generating a high number of gene expression clusters. Third, the gene expression distributions resulting from these circuits seem to be less structured, compared to those from the circuits with the highest multiplicity. All these features contribute to being high in both multiplicity and flexibility. Contrarily, the bottom five ranked circuits, as shown in **Fig. 3.11C**, have more self-inhibitions, allow for gene expression distributions of a single cluster, and have highly connected circuit topologies. These properties are exactly opposite to those of top-ranked circuits, explaining why the bottom-ranked circuits have low multiplicity and flexibility. These observations are also consistent with the outcomes of the circuit motif enrichment analysis, as shown in **Fig. 3.12**.

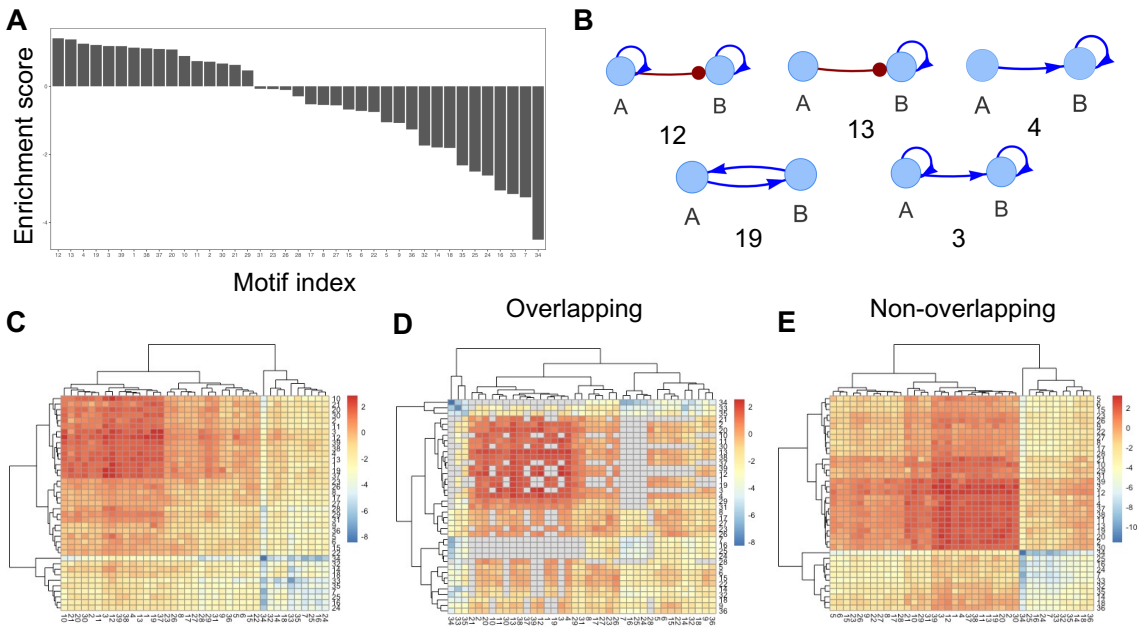

Figure 3.12. Circuit motif enrichment analysis by both multiplicity and flexibility. Here, all nonredundant four-node gene circuits are scored and ranked by the product of multiplicity and flexibility. The enrichment of a single motif is significant for most motifs, except for motif #31 (adjusted *p* values <0.01).

Note that we identified circuit motifs #19 and #39 again as enriched motifs with high multiplicity and flexibility. These toggle-switch-like motifs were observed, presumably because they can generate bistability, thus having more potential to generate more states when coupled with similar motifs, and meanwhile can allow flexible switches among states[39].

3.3.5 Multiplicity and Flexibility in networks of different sizes and types

Lastly, we explored the properties of multiplicity and flexibility in large random GRNs. To generate GRNs of an extended range of multiplicity and flexibility, we selected a list of two-node circuit motifs as the building blocks and synthesized them into large GRNs

of different sizes, with either sequential, scale-free, or random topological structure (see Methods for the detailed implementation). The selected motifs are either (1) the top 3 two-node circuit motifs ranked by multiplicity (i.*e.*, motifs # 30, 21, 39), (2) the top 3 two-node circuit motifs ranked by flexibility (*i.e.*, motifs #1, 10, 19), or (3) the motifs from both (1) and (2). For each motif type, gene regulatory network (GRN) topology type, and GRN size, we randomly generated the topology of ten networks (see the companion GitHub repository[153] for all GRN topologies), followed by RACIPE simulations to generate 10,000 gene expression profiles for each GRN. To calculate the multiplicity and flexibility for a large GRN, we performed principal component analysis on the standardized log transformation gene expression, and applied Equations (7) and (8) (Section 2.2.11) using the data projected onto the first four PCs. Note that in Equation (8), the summation over gene perturbation is still applied to all genes in the GRN. We chose to first project data onto the first four PCs, as the data with reduced dimensions usually capture gene expression states well. Moreover, low dimensional reduction has been widely used in high dimensional gene expression data analysis.

The multiplicity and flexibility of these random GRNs are summarized in **Fig. 3.13**, where we identified the following interesting findings.

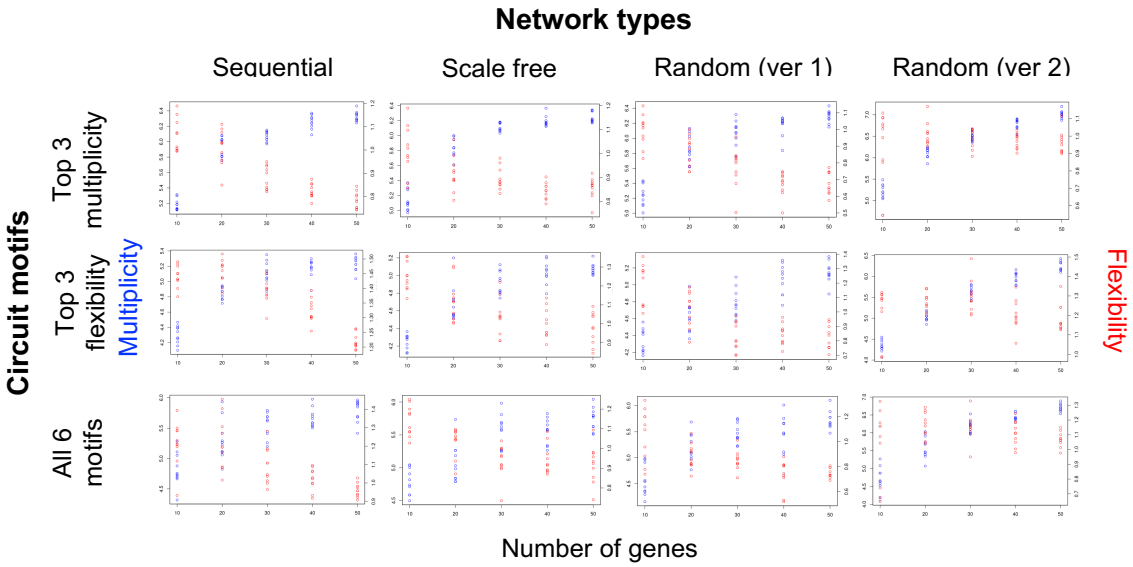

Figure 3.13. Multiplicity and flexibility of large gene regulatory networks. The plots show the multiplicity (blue points) and flexibility (red points) for GRNs of different sizes (number of genes on the *x*-axis) and different network types (each panel). Different columns correspond to sequential networks (1st column), scale-free networks (2nd column), random networks with sparse connectivity (ver 1, 3rd column), and random networks with dense connectivity (ver 2, 4th column). Different rows correspond to networks synthesized with top 3 multiplicity two-node motifs (1st row), top 3 flexibility two-node motifs (2nd row), and all these 6 motifs (3rd row).

First, the overall trends of multiplicity and flexibility for different GRN sizes and types are very similar for different choices of circuit motifs. The multiplicity scores are usually at high levels when using the motifs with the highest multiplicity and at low levels when using the motifs with the highest flexibility. Similarly, the flexibility scores are usually at high levels when using the motifs with the highest flexibility and at low levels when using the motifs with the highest multiplicity. Thus, among GRNs of different sizes, multiplicity and flexibility are anticorrelated. Our finding also suggests that the multiplicity/flexibility properties of a large GRN are largely determined by the properties of the circuit motifs within the GRN.

Second, regardless of the type of GRNs, multiplicity was found to be linearly correlated with the number of genes but saturated for large number of genes (blue points in **Fig. 3.13**). For each category of GRNs (*i.e.*, different sizes and motif types), the variations of multiplicity among ten random networks are mostly small, but slightly larger for the GRNs with mixed motifs. We also computed the multiplicity when the local density was estimated with the gene expression profiles of all dimensions, and, in this case, multiplicity is always linearly correlated with the number network genes. The dependence of multiplicity on GRN sizes can be understood as follows. When the GRNs are very small, the number of distinct states allowed by the GRNs are also limited. When the GRNs become larger, much richer network behaviors can be observed, therefore larger multiplicity. However, when the GRNs get extremely large, although the variations of gene expression still increase (multiplicity for data with full dimensions), the number of distinct gene expression states get saturated (multiplicity for data with reduced dimensions).

Third, flexibility was found to be linearly anti-correlated with the number genes for sequential networks, scale-free networks, and random networks where motifs are sparsely connected with a fixed number of interactions per motif denoted as *random ver1*, see Methods for details) (red points, 1st to 3rd columns in **Fig. 3.13**), despite much larger variations in flexibility among ten networks of the same category. In those situations, we also observed a saturation of flexibility for small GRNs. Because of high variations and saturation of flexibility, we also observed a few small GRNs with low flexibility. Interestingly, for networks where motifs are densely connected with a fixed ratio of

interactions per motif (denoted as *random ver2*), we observed a bell shape of flexibility with respect to the number of genes, *i.e.*, the highest flexibility may occur in GRNs of intermediate sizes.

Taken together, when the network size increases, multiplicity increases while flexibility decreases. Both multiplicity and flexibility tend to be saturated for large and small GRNs, respectively. Based on these findings, we perceive that the GRNs with both high multiplicity and flexibility are likely of intermediate sizes.

3.4 Discussion for What makes a functional gene regulatory network? A circuit motif analysis

In this study, we explored the types of gene circuit motifs that contribute to a functional GRN. We first defined two scoring functions to quantify the multiplicity and flexibility of a gene regulatory circuit based on the circuit's gene expression distribution. We then systematically applied the scores to rank all nonredundant four-node gene circuits. By applying gene circuit motif analysis, we identified reoccurring two-node circuit motifs and the co-occurrence of two motifs that enriched in top-ranked circuits by either multiplicity, flexibility, or a combination of both. Furthermore, using the enriched motifs as the building blocks, we generated many GRNs of different types and sizes and investigated the GRN properties that contribute to high levels of multiplicity and flexibility. We hope this study will improve our understanding of the design of biological GRNs.

The core approach utilized in this study is the circuit motif enrichment analysis introduced earlier in the chapter. We have demonstrated the effectiveness of this approach in identifying not only circuit motifs associated with a particular dynamical behavior but also the coupling of two circuit motifs. Here, we focused on multiplicity, the ability of a GRN in generating a high number of states, and flexibility, the ability of a GRN in altering gene expression upon perturbations. In our view, multiplicity and flexibility are among the most important features of a functional GRN since they are the basic features that a GRN must exhibit to be functional; GRNs that model cellular state transitions must inherently be able to generate multiple states and respond to signals that trigger state transitions as a basic requirement for biological relevance. From the enrichment analysis, circuit motifs with mutual regulations and self-activation tend to have high multiplicity, while circuit motifs with single monodirectional regulation and without autoregulation tend to have high flexibility. Remarkably, two types of circuit motifs allow both high multiplicity and high flexibility—either motifs with sparse connectivity and self-activation or toggle-switch-like motifs.

While it is important to elucidate the types of circuit motifs having high multiplicity and/or flexibility, we also wonder how these circuit motifs contribute to the multiplicity and flexibility of larger GRNs. To address this question, we generated GRNs of different sizes and types using the enriched circuit motifs as the building blocks. From an extensive network analysis, we found that network multiplicity and flexibility indeed are largely impacted by the types of circuit motifs with the GRNs. Overall, GRNs of

intermediate sizes tend to have combined high levels of multiplicity and flexibility. Thus, we hypothesize that a biological GRN, when considered as a functional dynamical system, should be of intermediate sizes. This can be understood by the following: when a GRN is too small, it is not complex enough to robustly generate desired functionality; when a GRN is too large, it could be too rigid to allow sufficient control by external signals or environmental factors[154,155]. Thus, GRNs of intermediate sizes can alleviate the issues of smaller and larger GRNs. In our view, this criterion of network size would be helpful to elucidate the design principle of biological GRNs and improve the effectiveness of GRN inference.

4 NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity

4.1 Introduction

One of the major goals of systems biology is to infer and model complex GRNs which underpin the biological processes of human disease[156–161]. Particularly important are those gene networks that control decisions regarding cellular state transitions (*e.g.*, replicative to quiescent[162–164], EMT[165], pluripotent to differentiated[166,167]), given the central importance of such regulatory processes to both healthy development as well as diseases such as cancer.

In this study, we introduce a computational platform, named NetAct, for inferring a core GRN of key TFs using both transcriptomics data and a literature-based TF-target database. Integrating both resources allows us to take full advantage of the existing knowledgebase of transcriptional regulation. NetAct adopts the combined top-down bioinformatics and bottom-up systems biology approaches, designed specifically to address the following two major issues.

First, many network inference methods rely on correlations of gene expression data, yet the actual transcriptional activities of many master regulators may not be reflected in their gene expression. Instead, the activity may be better associated with either their protein level, the level of a certain posttranslational modification, localization, or their

DNA binding affinity. As a result, the master regulators with weak correlations between the expression level and the transcriptional activity will likely be discarded in the network. Some algorithms have been developed to infer the activities of regulators from transcriptomics data, such as VIPER[168], NCA[169], AUCELL[48]. However, most of these algorithms 1) are not designed for gene network modeling, or 2) still rely on co-expression of a TF and its targeted genes, or 3) do not take advantage of known regulatory interactions from the literature, hindering their applicability as automated algorithms for generic use in systems biology.

Second, conventional mathematical modeling approaches have been applied over the years to simulate the dynamics of a GRN, yet they are not particularly effective in analyzing core GRNs. A popular method models the gene expression dynamics of a system using the chemical rate equations that govern the associated gene regulatory processes. However, it is difficult to directly measure most of the kinetic parameters of a GRN. Although some parameter values can be learned from published results, many others are often based on educated guesses which significantly limits the predictive power of mathematical modeling. Moreover, a core GRN is not an isolated system. Thus, an ideal modeling paradigm should also consider other genes that interact with the core network. To address this infamous parameter issue, we have developed the modeling algorithm RACIPE[67,170,171] in previous work that analyzes a large ensemble of mathematical models with random kinetic parameters. RACIPE has been applied to model the dynamical behavior of gene regulatory networks of different biological

processes, such as epithelial-mesenchymal transition[67,172], cell cycle[171], stem cell differentiation[173].

The new NetAct platform addresses the above-mentioned issues by (1) inferring the activities of TFs for individual samples using the gene expression levels of their targeted genes, (2) identifying the regulatory interactions between two TFs based on their activities rather than their expressions, (3) and subsequently simulating the constructed core GRN with RACIPE to validate and evaluate the gene expression dynamics of the core GRN. In this paper, we describe in detail the NetAct platform, extensive benchmark tests for TF-target databases, TF activity inference, and network construction, and two examples of applications to model GRNs with time series gene expression data.

4.2 Results

We developed a computational systems-biology platform, named NetAct, to construct TF-based GRNs using TF activity. The method uniquely integrates both generic TF-target relationships from literature-based databases and context-specific gene expression data. NetAct also integrates the previously developed mathematical modeling algorithm RACIPE to evaluate whether a constructed network functions properly as a dynamical system. It evaluates the roles of every gene in the network by in-silico perturbation analysis. NetAct has three major steps: (1) identifying the core TFs using GSEA[174] with an optimized TF-target gene set database (**Fig. 15A**). Here, the custom gene set consists of a transcription factor and its targets. GSEA then indicates which TFs are active by

identifying which TF targets are enriched from the DE analysis of gene expression data. A q-value cutoff is used to identify significantly enriched TFs. (2) Inferring TF activity (**Fig. 4.1B**); (3) constructing a core TF network (**Fig. 4.1C**). Then, the network is validated and analyzed by simulating its dynamics using mathematical modeling by RACIPE.

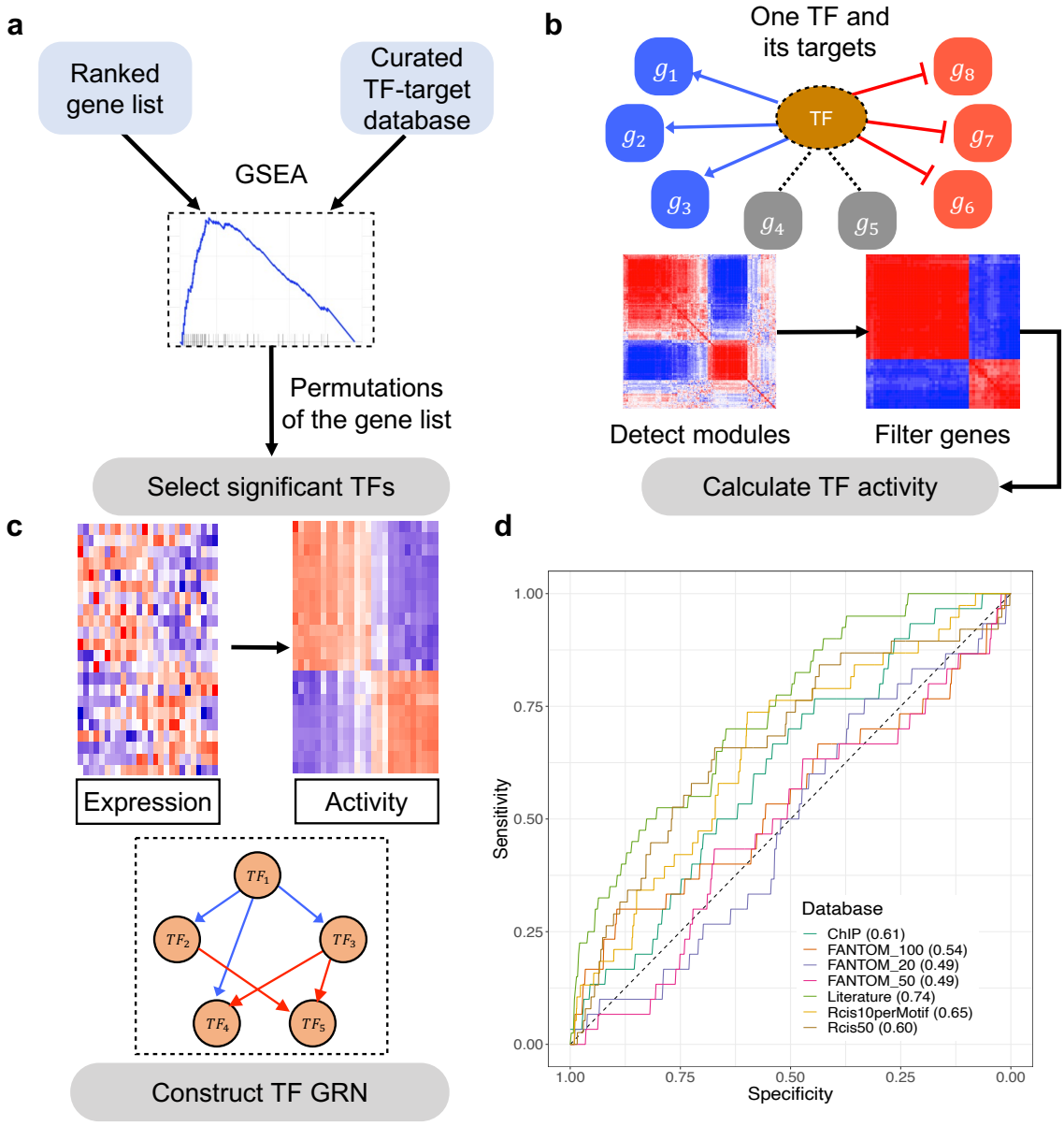

Figure 4.1. Schematics of NetAct. **a** First, key TFs are identified using GSEA with a literature-based TF-target database. **b** Second, the TF activity of an individual sample is inferred from the expression of target genes. From the co-expression and modularity analysis of target genes, we find target genes that are either activated (blue), inhibited (red), or not strongly related to the TF (gray). The activity is defined as the weighted average of target genes activated by the TF minus the weighted average of target genes inhibited by the TF. **c** Lastly, a TF regulatory network is constructed according to the mutual information of inferred TF activity and literature-based regulatory interactions. **d** Performance of GSEA for various TF-target gene set databases. The plot shows the sensitivity and specificity with different *q*-value cutoffs. The gene set databases in the benchmark include the combined literature-based database (D1); FANTOM5-based databases (D2) with 20, 50, and 100 target genes per TF; the combined experimental-based database (D3, ChIP); and RcisTarget databases (D4), one with 10 targets per TF binding motif and another with 50 total number of targets per TF

My personal contribution to this work is mainly the development of a method to correct low confidence edge assignments and the application of NetAct to two Biological settings, EMT and Macrophage activation. The method is presented in full to understand the application.

4.2.1 Database for TF-target interactions

To establish a comprehensive gene set database containing TF-target relationships, we considered data from different sources. They are (D1) a literature-based database, consisting of data from TRRUST[175], RegNetwork[176], TFactS[177], and TRED[178]; (D2) a gene regulatory network database FANTOM5[179], whose interactions are extracted from networks constructed using RNA expression data from 394 individual tissues; (D3) a database derived from resources of putative TF binding targets, including ChEA[180], TRANSFAC[181], JASPAR[182], and ENCODE[183]; and (D4) a database derived from motif-enrichment analysis, RcisTarget[184]. These databases have been frequently used to study transcriptional regulations and have already been utilized for network construction[67,185].

We evaluated the performance of these databases by GSEA on a benchmark dataset. GSEA is a popular statistical method that can be used to evaluate significant overlapping between a set of genes and differentially expressed genes between two experimental conditions. Using various types of TF-target databases, our goal is to find the best version of the database, so that GSEA can detect the target gene sets of the relevant TFs to be statistically significant. This benchmark dataset, denoted as *set B*, consists of a compilation of 12 microarray and 32 RNA-Seq gene expression data. Each of these datasets contains at least three samples under the normal condition (control) and three samples under the treatment condition in which a single specific TF is treated by knockdown (KD) leading to 44 different TF knockouts in either mouse or human. TFs knocked out include EPAS1, TCFL72, FOXD3, PAX2, AR, TCF21, ASCC3, RUNX1, EZH2, AHR, GATA1, ETS1, PARP1, BRCA1, PAX2, RELA, NFKB1, VHL, JUND, SREBF1, SREBF2, GATA6, EGR1, ATF3, SOX10, FLI1, ZEB2, CTCF, PTBP1, SFRS1, HOXA1, BCL6, FOXM1, MYB, ATM, RELA, P53, NANOG, SOX2, OCT4, PPARG, DNMT1, and CTNNB1. We applied GSEA (with slight modifications, details in Chapter 2) on the set *B* to evaluate whether the enrichment analysis can detect the perturbed TFs. The underlying assumption is that, with a better TF-target gene set database, GSEA will be more likely to detect the corresponding perturbed TFs. For each TF-target database and each gene expression data in set *B*, we calculated the q-values of all the TFs in the database by GSEA to determine whether the target genes of the perturbed TF are enriched in the differentially expressed genes. We found that more significant q-values are usually associated with relatively larger number of targets for

each TF; however, too many (*e.g.*, greater than 2000) targets will result in non-significant q-values. Furthermore, these corresponding q-values from all the gene expression data are converted to specificity and sensitivity values (see Chapter 2), and different databases are compared based on the area under the sensitivity-specificity curves (**Fig. 4.1D**). We found that the literature-based database, consisting of combined interactions from TRRUST[175], RegNetwork[176], TFactS[177], and TRED[178] has the best overall performance, thus we used this database for further analyses. Our results are in line with a previous benchmark study[186] that literature-based TF-target database outperforms others in capturing transcriptional regulation.

4.2.2 Inferring TF activity levels

NetAct can accurately infer TF activity for an individual sample directly from the expression of genes targeted by the TF (see Chapter 2). In the following, we will illustrate how NetAct infers TF activity on two cases of microarray KD experiments -- one case for shRNA KD of FOXM1 and shRNA KD of MYB in lymphoma cells (GEO: GSE17172[187]), and another case for KD of BCL6 on both OCI-Ly7 and Pfeiffer GCB-DLBCL cell lines (GEO: GSE45838[168]). NetAct first successfully identified the TFs that undergo knockdown in each case, *i.e.*, FOXM1, MYB and BCL6 respectively, by applying GSEA on the optimized TF-target database (q value < 0.15)

Next, for each identified TF, NetAct calculates its activity using the mRNA expression of the direct targets of the TF. We first constructed a Spearman correlation matrix from the expression of the targeted genes. As shown in Fig. **4.2A**, the correlation matrix after

hierarchical clustering analysis typically consists of two red diagonal blocks, two blue off-diagonal blocks, and the remaining elements with low correlations which will be filtered out subsequently (details in Chapter 2). Within the red blocks, the expression of any column gene is positively correlated with that of any row gene; while within the blue blocks, the expression of any column gene is negatively correlated with that of any row gene. This indicates that the genes in the two red blocks are anti-correlated in gene expression with each other. However, if the correlation matrix is constructed from 100 or 200 randomly selected genes (Fig. **4.2BC**), such a clear pattern disappears. Thus, our observation suggests that genes from one of the red blocks are activated by the TF, whereas genes from the other block are inhibited by the TF. Moreover, filtered genes are not likely to be directly targeted by the TF in this context, or they are regulated by multiple factors simultaneously and are thus likely not a good indicator for the TF activity.

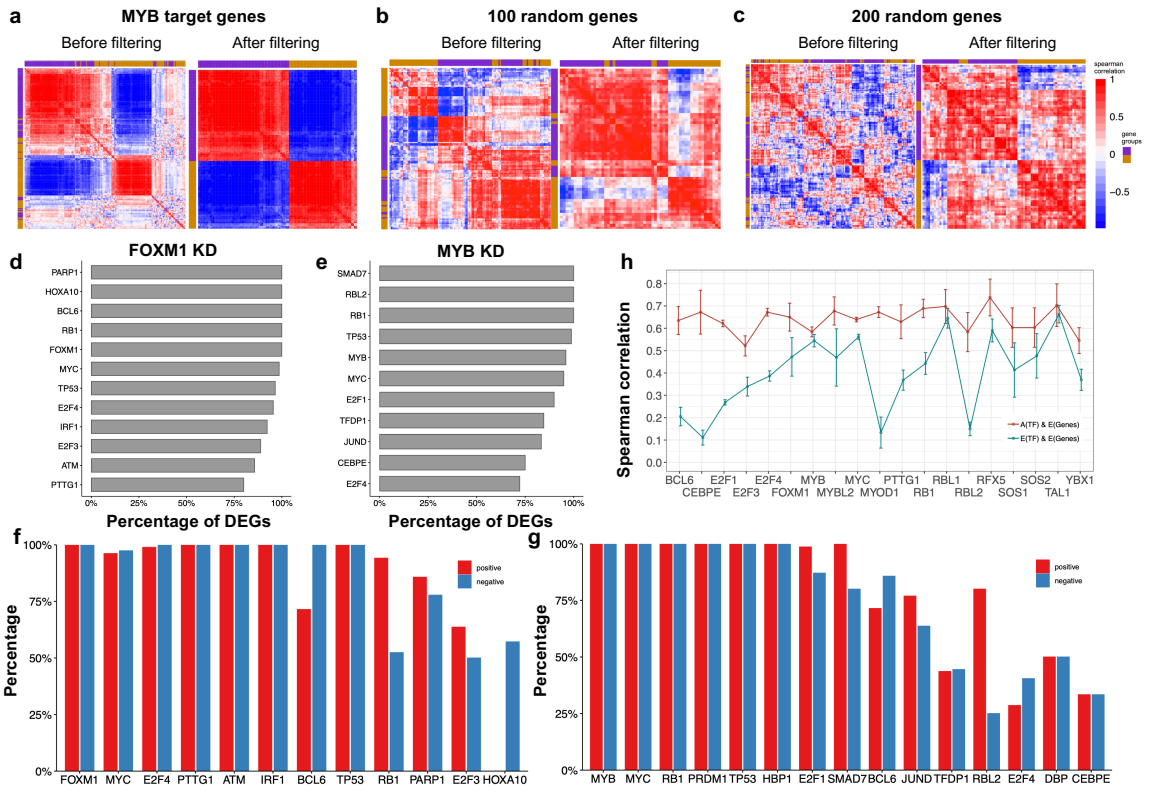

Figure 4.2 Illustration of the grouping scheme for target genes of a transcription factor. a The co-expression matrix of MYB target genes in shRNA knockdown of MYB lymphoma cells by hierarchical clustering analysis (Pearson correlation and complete linkage). **b**, **c** The poor clustering results from the co-expression of randomly selected 100 (**b**) and 200 genes (**c**). In panels **a**–**c**, the left subplots show the outcomes of all tested genes, and the right subplots show the outcomes of genes after the filtering step. Compared to the random cases, MYB target genes have a clear pattern of red and blue diagonal blocks from their co-expression. **d**, **e** The percentage of differentially expressed genes remained after the filtering step in the case of FOXM1 and MYB knockdown, respectively. **f**, **g** The proportion of genes from the activation group that are positively correlated with the TF expression (red bars) and the proportion of genes from the inhibition group that are negatively correlated with the TF expression (blue bars). **h** Spearman correlation (average and standard deviation) between TF activity and target expression (red) and between TF expression and target expression (blue)

We further evaluated how the filtering step removes noise and retains the important genes in the analysis. We found that, after the filtering step, most of the DE genes are retained, as evidenced by Fig. **4.2D**. Here, DE genes from each comparison were retrieved by using *limma* with a cutoff for the adjusted p-values at 0.05 and a cutoff for the log2 fold changes at 2. Subsequently, for DE TFs we evaluated the Spearman correlations between

the TFs and the corresponding targeted genes. In traditional approaches (such as ARACNe[156], WGCNA[188], and BEST[189]), the co-expression between a TF and its targeted genes are commonly used to identify its association and assign the sign (activation or inhibition) of the regulation. We found that, for each TF, most of the genes in a block either positively correlate with the TF expression (Fig. **4.2FG**, blue bars), or they negatively correlate with the TF expression (Fig. **4.2FG**, red bars). The tests demonstrate that, without directly using TF expression, NetAct can successfully identify two groups of important target genes – genes in each group are either activated or inhibited by the TF. These two groups of genes are further used to infer TF activity by a weighted average of their gene expression (Equation 10 in Chapter 2). Additionally, we found that the correlations between inferred TF activity and target expression are usually higher than the correlations between TF expression and target expression (Fig. **4.2H**).

4.2.3 Benchmarking NetAct

To evaluate the accuracy and robustness of inferred TF activity, we performed extensive benchmark tests to compare NetAct with other existing methods. We first performed the benchmark tests on simulated data because TF activity is usually not directly measurable. The activity of a TF can be related to its protein level or the level of a particular posttranslational modification, such as phosphorylation. Therefore, it is very difficult to obtain the ground truth of TF activity from an experimental data set. Thus, in this

benchmark test, we rely on mathematical modeling to simulate both the expression and activity of each TF from a synthetic TF-target network. With this simulated data, we benchmark NetAct against other methods.

To establish the simulated benchmark data set, we first constructed a synthetic TF-target network with a total of 30 TFs. Each TF has 20 target genes randomly selected with replacement from a pool of 1000 genes. In addition, each TF also regulates two (randomly selected) of the 30 TFs. This synthetic network has a hierarchical structure, where a target gene may be co-regulated by multiple TFs. The type of each TF-to-TF regulation is either excitatory, inhibitory, or signaling, with a chance of 25%, 25%, and 50%, respectively; the type of each TF-to-target regulation is either excitatory or inhibitory with a 50% chance for each. Here, the signaling regulation changes the activity of a TF without changing its expression; whereas the excitatory or inhibitory interactions changes both the activity and expression. From one realization of the synthetic network generation, the final synthetic network contains a total of 477 genes (30 TFs, 447 targeted genes) and 660 regulatory links (**Fig. 4.2A**).

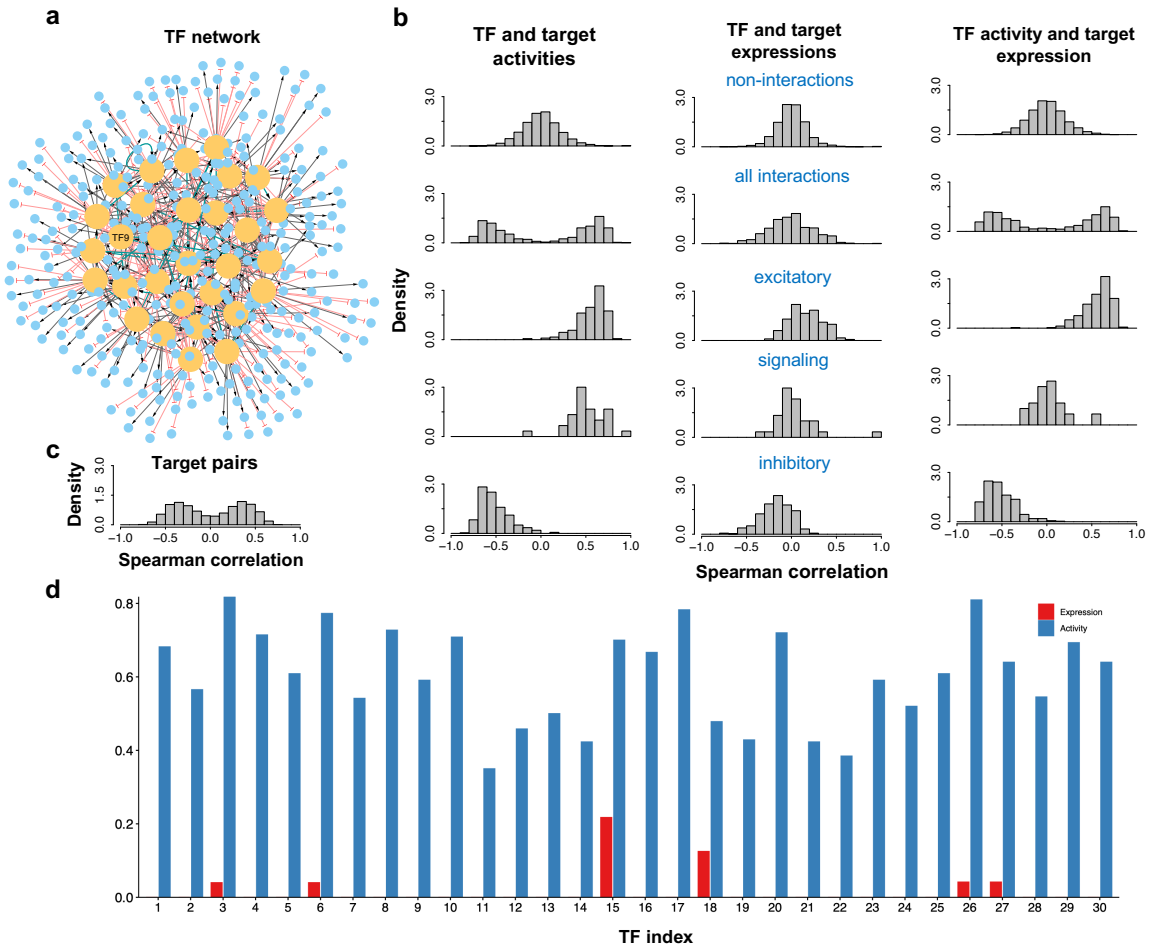

Figure 4.3. Simulation of both gene expression and activity of a synthetic GRN. **a** The synthetic GRN consisting of 30 TFs and 447 target genes. An edge of transcriptional activation is shown as black line with an arrowhead; an edge of transcriptional inhibition as red line with a blunt head; an edge of signaling interaction as green line with an arrowhead. Transcription factor labeled as TF9 was selected for knockdown simulations. **b** The summary of the correlation analyses of the simulated expression and activity. The left, middle, and right columns represent the outcomes for TF and target activities, TF and target expressions, and TF activities and target expressions, respectively. For each category, the histograms of Spearman correlations are shown for non-interacting gene pairs (first row), interacting gene pairs (second row), gene pairs of excitatory transcriptional regulation (third row), gene pairs of excitatory signaling regulation (fourth row), and gene pairs of inhibitory transcriptional regulation (fifth row). Here, the target activity is set to be the same as the target expression for non-TF genes. **c** The histograms of Spearman correlations for gene pairs of target genes from the same TF. **d** Jaccard indices between the ground truth regulons of the synthetic GRN and the regulons inferred by ARACNe using either the simulated expression (red) or activity data (blue).

To simulate the gene expression of the TF-target network, we applied a generalized version of the mathematical modeling algorithm, RACIPE[171]. Using the network topology as the only input, RACIPE can generate an ensemble of random models, each corresponds to a set of randomly sampled parameters. Here, we used RACIPE to generate simulated data including gene expression and TF activity for benchmark. Some previous studies have also adopted a similar modeling approach for benchmarking[190,191]. To consider the effects of a signaling regulatory link, we generalized RACIPE to simulate both expression and activity for each TF.

In the benchmark test, we used RACIPE to simulate 100 models with randomly generated kinetic parameters. From these 100 models we obtained 83 stable steady-state gene expression and activity profiles for the 477 genes. As expected, TF activity and target activity from a regulatory link are correlated (1st column, 2nd row in **Fig. 4.3B**); TF activity and target expression (3rd column, 2nd row in **Fig. 4.3B**) are correlated; and the expression of two target genes (**Fig. 4.3C**) are correlated. However, there is no strong correlation between TF expression and target expression (2nd column, 2nd row in **Fig. 4.3B**) and, for a signaling regulatory link, between TF activity and target expression (3rd column, 4th row in **Fig. 4.3B**). Next, we applied ARACNe to predict the regulon (*i.e.*, the list of targeted genes by a specific TF) using either the simulated expression profiles or the simulated activity profiles. We found that the regulons predicted from the activity profiles are substantially more similar to the predefined regulons (measured by the Jaccard similarity[192]) than those predicted from the expression profiles (**Fig. 4.3D**). The

results indicate the need of using the TF activity, instead of TF expression, to identify TF-target relationships.

Next, we compared the performance of NetAct with several related algorithms, NCA, VIPER, and AUCell, in inferring TF activity using both the simulated expression profiles from the 83 models and a predefined regulon (*i.e.*, the association of each TF with its target genes). The predicted activity was then compared with the simulated activity (ground truth) to evaluate the performance. To mimic the real-life scenario where the target information may not be complete and accurate, we consider more challenging tests where the regulon data is randomly perturbed. Here, for a specific perturbation level, we generated 100 sets of regulon data by replacing a certain number of target genes for each TF with non-interacting genes. The numbers of replaced genes are 0 (0% level of perturbation), 5 (25%), 10 (50%) and 15 (75%), respectively, in different tests. We then evaluated the performance of NetAct, NCA, and VIPER. AUCell protocol advises to include the target genes with only positive interactions in the regulons. To satisfy this criterion, we updated the regulons for both unperturbed and perturbed regulons. For the unperturbed regulons, we retained only the positive interactions; for the perturbed regulons, we retained the positive target genes that were not replaced and a random half of the replaced target genes (assuming that half of the genes are positively regulated by the TF). We then evaluated AUCell performance using these updated regulons (denoted AUCell 1) and non-updated regulons (denoted AUCell 2). As shown in **Fig. 4.4A**, NetAct significantly outperforms each of the other methods in reproducing the simulated activity profiles at each perturbation level. As expected, the performance of NetAct is

decreased by increasing the perturbation levels of the regulon data; however, NetAct still performs reasonably well even when only 25% of the actual target genes are kept in the regulon data. The results indicate that NetAct can robustly and accurately infer TF activity even with a noisy TF-target database.

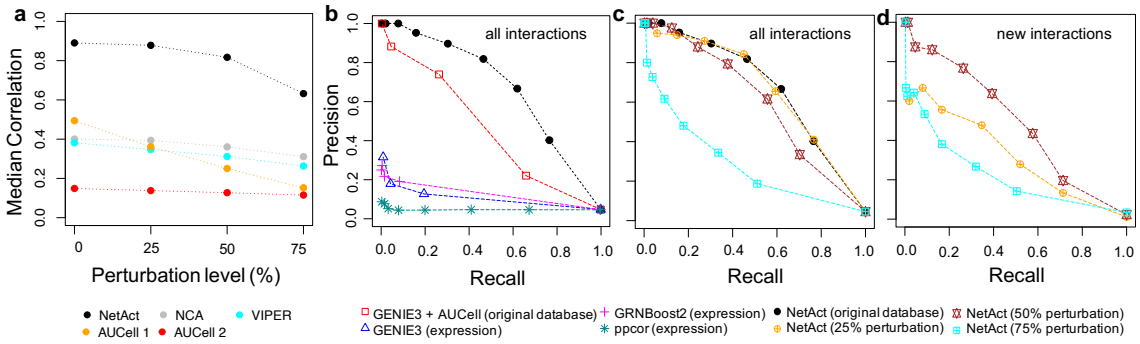

Figure 4.4 The performance of activity and network inference from a simulation benchmark. a TF activity inference. TF activity was inferred by several methods using the gene expression data simulated from the synthetic TF-target GRN and the corresponding regulons. For each TF, we computed Spearman correlations between the inferred activity and simulated activity (ground truth) for all the simulated models. Then, we calculated the average correlation values over all TFs. The plots show the median of average correlations for the cases where we used the original regulons defined by the TF-target network (0% perturbation), and the regulons where 5 (25% perturbation), 10 (50% perturbation), and 15 (75% perturbation) target genes are randomly replaced with non-interacting genes. The median values were computed over 100 repeats of random replacement for each perturbation level, and the values of the average correlations are reported for the case of zero perturbation. Shown are the results for NetAct (black), NCA (gray), VIPER (cyan), AUCELL 1 where regulons contain only positively associated target genes (orange), and AUCELL 2 where regulons contain all target genes (red). **b–d** Network inference. The panels show the performance of network inference algorithms from the simulation benchmark by the precision and recall for different link selection thresholds. **b** Network inference performance against all ground truth regulatory interactions. Tested methods are GENIE3, GRNBoost2, and PPCOR, using TF expression; GENIE3 using TF activity inferred by AUCell; NetAct using its inferred TF activity. For the latter two methods, original (unperturbed) regulons obtained from the regulatory network were used. **c** Network inference performance of NetAct against all ground truth regulatory interactions using the regulons with 0% (the original), 25%, 50%, and 75% target perturbations. **d** Network inference performance of NetAct in discovering new regulatory interactions not existing in the regulons. NetAct was applied using the regulons at different perturbation levels (25%, 50%, and 75%). The benchmark results shown here are for the case of the untreated simulation.

Furthermore, we tested another scenario where the test data contains simulated data from two experimental conditions, *e.g.*, one representing an unperturbed condition and the other representing a perturbed condition. Here, we used the same synthetic network but compiled 40 expression and activity data from the above-mentioned simulation (unperturbed condition), together with 43 expression and activity data from the simulations in which a specific TF (TF9) is knocked down (perturbed condition). We then performed a similar test as above and found that NetAct outperformed each of the other. The notable performance gain of NetAct mainly emanates from the removal of incoherent (or noisy) targets of a TF before the activity calculation in NetAct (see Chapter 2).

In addition, we performed a network construction benchmark of NetAct and a few other network construction algorithms using the in-silico simulation data set, as shown in **Fig. 4.4BCD** NetAct, using the TF activity inferred from the original regulon database, outperforms not only network construction methods using gene expression, such as GENIE3[193], GRNBoost2[46], and ppcor[25,194], but also GENIE3 using the TF activity inferred by AUCell (**Fig. 4.4B**). The last approach was presented to mimic a popular method SCENIC. Moreover, we evaluated the performance of NetAct when using a perturbed regulon database. We found that NetAct continues to performing well when the perturbation level is as large as 50%, when evaluated by all the ground-truth interactions (**Fig. 4.4C**) and by those not presented in regulon database (**Fig. 4.4D**). The latter case was designed to evaluate the capability of NetAct in predicting novel interactions. We observed similar outcomes for the case of the second scenario of the simulation data from

two conditions. In summary, our in-silico benchmark test demonstrates the high performance of NetAct over existing state-of-the-art methods in both inferring TF activity and gene regulatory networks.

4.2.4 Modeling cell state transitions

In the previous sections, we demonstrated the capability of NetAct in identifying the key TFs and predicting TF activity. With these data, NetAct further constructs a TF-based GRN using the mutual information (MI) of the activity from the identified TFs (details in Chapter 2). We then applied RACIPE to the constructed network to check whether the simulated network dynamics are consistent with experimental observations. In the following, we show the utility of NetAct with two biological examples: EMT and macrophage polarization.

4.2.4.1 EMT

In the first case (EMT), we analyzed a set of time-series microarray data on A549 epithelial cells undergoing TGF-ß induced EMT (GEO: GSE17708)[195]. According to the overall structure of the transcriptomics profiles, we arranged samples from different time points into three groups – early stage (time points 0h, 0.5h and 1h), middle stage (time points 2h, 4h, and 8h) and late stage (time points 16h, 24h, and 72h). We then performed three-way GSEA with our human literature-based TF-target database to identify enriched TFs that are active between either early-middle, early-late and middle-late timepoints.

Forty-one TFs (q-value cutoff 0.01) were identified including many major transcriptional master regulators, such as BRCA1, CTNNB1, MYC, TWIST1, TWIST2 and ZEB1, and factors that are directly associated with TGF- ß signaling pathway, such as SMAD3[196], FOS and JUN[197]. The hierarchical clustering analysis (HCA) of the expression and activity profiles for these TFs is shown in Fig. 19a. While the expression profiles are quite noisy, the activities show a clear gradual transition from the epithelial to mesenchymal (M) state. Note that the signs of the activity of a few non-DE TFs were flipped according to experimental evidence of protein-protein interactions and the nature of transcriptional regulation (see Chapter 2 for detailed procedures).

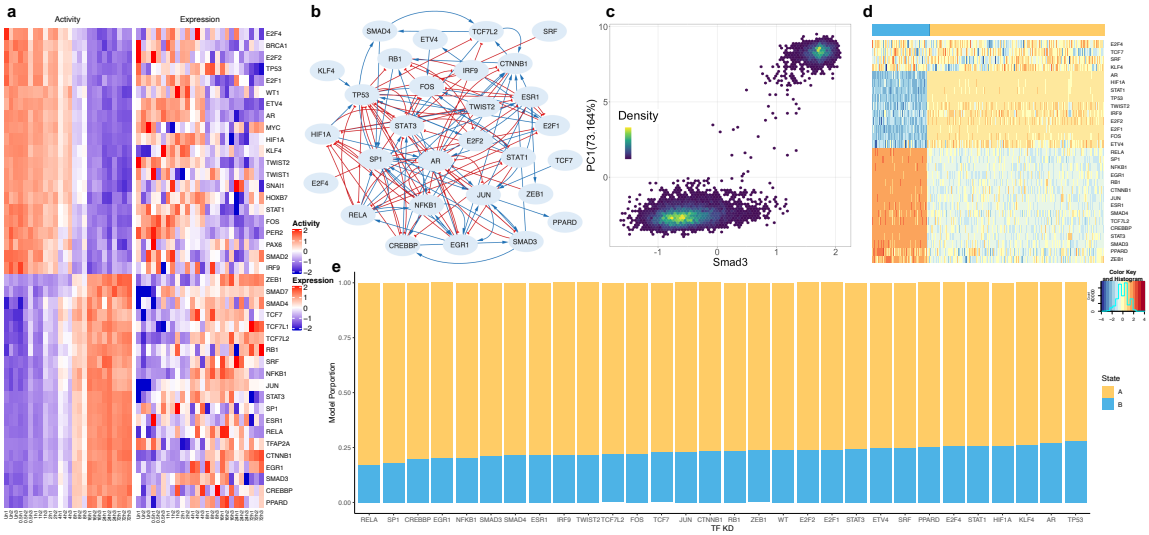

Figure 4.5 Network modeling of TGF-β-induced EMT. Application of NetAct to an EMT in human cell lines using time-series microarray data. **a** Experimental expression and activity of enriched transcription factors. **b** Inferred TF regulatory network. Blue lines and arrowheads represent the gene activation; red lines and blunt heads represent gene inhibition. **c** The relationship between SMAD3 gene activity and the first principal component of the activity of all network genes from RACIPE simulations. **d** Hierarchical clustering analysis of simulated gene activity (with Pearson correlation as the distance function and Ward.D2 linkage method). Colors at the top indicate the two clusters from the simulated gene activity. The blue cluster represents the mesenchymal state, and the yellow cluster represents the epithelial state. The color legend for the heatmap is at the bottom right. **e** Knockdown simulations of the TF regulatory network. The bar plot shows the proportion of RACIPE models in each state (epithelial or mesenchymal) for the conditions of the knockdown of every TF

We then constructed a TF regulatory network (**Fig. 4.5B**) and performed mathematical modeling to simulate the dynamical behavior of the network using RACIPE (**Fig. 4.5CD**). We found that, consistent with the expression and activity profiles (**Fig.4.5A**), the network clearly allows two distinct transcriptional clusters that can be associated with E (the yellow cluster in **Fig. 4.5D**) and M states (the blue cluster in **Fig. 4.5D**). To assess the role of TGF- ß signaling in inducing EMT, we performed a global bifurcation analysis[67] in which the SMAD3 level is used as the control parameter (**Fig. 4.5C**). Here, SMAD3 was selected as it is the direct target of TGF- ß signaling[196]. As shown in (**Fig. 4.5C**), when SMAD3 level is either very low or high, the cells reside in E or M states. However, when SMAD3 is at the intermediate level, the cells could be driven into some rare hybrid phenotypes. These results are consistent with our previous studies on the hybrid states of EMT[198,199]. Using RACIPE, we systematically performed perturbation analyses by knocking down every TF in the network. Our simulation results (**Fig. 4.5E**) suggest that knocking down TFs, such as RELA, SP1, EGR1, and CREBBP, *etc.*, has major effects in driving M to E transition (MET), while knocking down TFs, such as TP53, AR, and KLF4, *etc.*, has major effects in driving E to M transition (EMT). These predictions are all consistent with existing experimental evidence[200–206].

Compared to a previous model of the EMT network based on an extensive literature survey[207], the GRN constructed by NetAct identified some of the same regulators induced by the TGF-β pathway, such as SMAD3/4, TWIST2, ZEB1, CTNNB1, NFKB1, RELA, FOS and EGR1. Because of the lack of microRNAs and protein-protein interactions in

the database, NetAct didn't identify factors like miR200 and signaling molecules like PI3K. Interestingly, the NetAct model identifies STAT1/3, which was connected to other signaling pathways, such as HGF, PDGF, IGF1and FGR, but not TGF-β in the previous network model. In addition, the NetAct model identified regulators in other important pathways in TGF-β-induced EMT in cancer cells, *e.g.*, cell cycle pathway (RB1 and E2F1) and DNA damage pathway (P53).

4.2.4.2 Macrophage activation

In the second case, we studied the macrophage polarization program in mouse bone-marrow-derived macrophage cells using time series RNA-seq data (GEO: GSE84517)[208]. In this experiment, macrophage progenitor cells (denoted as UT condition) were treated with (1) IFNγ to induce a transition to the M1 state; (2) IL4 to induce a transition to the M2 state; (3) both IFNγ and IL4 to induce a transition to a hybrid M state. Here, we reprocessed the raw counts of RNA-seq with a standard protocol. From PCA on the whole transcriptomics (**Fig. 4.6B**), we found that the gene expression undergoes distinct trajectories when macrophage cells were treated with either IFN γ (M1 state) or IL4 (M2 state). When both IFN γ and IL4 were administered, the gene expression trajectories are in the middle of the previous two trajectories, suggesting that cells are in a hybrid state (hybrid M state). We aim to use NetAct to elucidate the crosstalk in transcriptional regulation downstream of cytokine-induced signaling pathways during macrophage polarization.

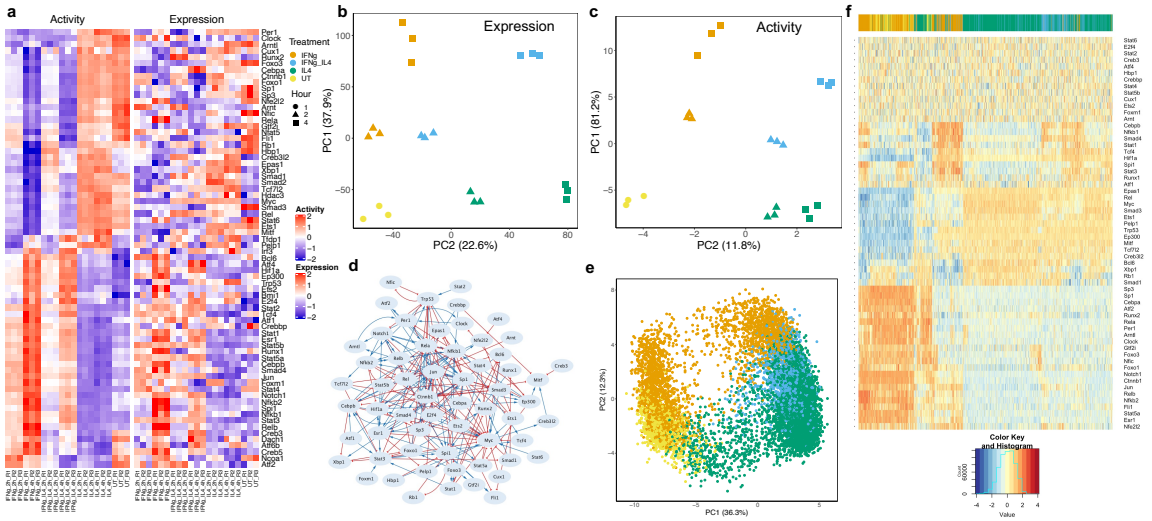

Figure 4.6 Network modeling of macrophage polarization. Application of NetAct to induced macrophage polarization via drug treatment in mice using RNA-seq data. **a** Experimental expression and activity of enriched TFs. **b** PCA projection of genome-wide gene expression profiles. Different point shapes indicate the time after treatment, and colors indicate treatment types **c** PCA projection of gene activity of enriched TFs. **d** Inferred TF regulatory network. Blue lines and arrowheads represent the gene activation; red lines and blunt heads represent the gene inhibition. **e** PCA projection of simulated gene activity of inferred network colored by mapping each model back to experimental data. **f** Hierarchical clustering analysis of simulated gene activity (with Pearson correlation as the distance function and Ward.D2 linkage method). Colors at the top indicate the mapped experimental conditions. The color legend of the heatmap is at the bottom

Here, we applied GSEA on six comparisons – untreated versus IFNγ treated samples (one comparison between the untreated and the treated after two hours, another between the untreated and the treated after four hours, same for the other comparisons), untreated versus IL4 treated samples, and untreated versus IFNγ +IL4 treated samples. Using our mouse literature-based TF-target database, we identified 79 TFs (q-value cutoff 0.05 for UT vs IL4-2h and 0.01 for all others). The expression and activity profiles of these TFs (**Fig. 4.6ABC**) captures the essential dynamics of transcriptional state transitions during macrophage polarization as follows. NetAct successfully identified important TFs in these processes, including Stat1, the major target of IFNγ, Stat2, Stat6, Cebpb, Nfkb

family members, Hif1a and Myc[209–211]. Myc is known to be induced by IL-4 at later phases of M2 activation and required for early phases of M1 activation[210]. Interestingly, we find Myc has high expression in both IL4 stimulation and its co-stimulation with IFNγ but its activity is high only in IL4 stimulation. We then constructed a TF regulatory network that connects 60 TFs (**Fig. 4.6D**) and simulated the network with RACIPE, from which we found that simulated gene expression (**Fig. 4.6F**) matches well with experimental gene expression data (**Fig. 4.6A**). RACIPE simulations display disparate trajectories from UT to IL4 or IFNγ activation and stimulation with both IL4 and IFNγ. Strikingly, we found in the simulation that there is a spectrum of hybrid M states between M1 and M2 (**Fig. 4.6E**), which is consistent with experimental observations of macrophage polarization[209]. Moreover, we also predict from our GRN modeling that the transition from UT to hybrid M is likely to first undergo a transition to either M1 or M2 before a second transition to hybrid M (**Fig. 4.6E**). This is because of our observation from the simulation data that there are fewer models connecting UT and hybrid M than any of the other two routes (*i.e.*, UT to M1, and UT to M2). Taken together we showed that the NetAct-constructed GRN model captures the multiple cellular state transitions during macrophage polarization.

In conclusion, we show that NetAct can identify the core TF-based GRN using both the literature-based TF-target database and the gene expression data. We also demonstrate how RACIPE-based mathematical modeling complements NetAct-based GRN inference in elucidating the dynamical behaviors of the inferred GRNs. Together these two

methods can be applied to infer biologically relevant regulatory interactions and the dynamical behavior of biological processes.

4.3 Discussion for NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity

In this study, we have developed NetAct—a computational platform for constructing and modeling core TF-based regulatory networks. NetAct takes a data-driven approach to establish GRN models directly from transcriptomics data and takes a mathematical modeling approach to characterize cellular state transitions driven by the inferred GRN. The method specifically integrates both literature-based TF-target databases and transcriptomics data of multiple experimental conditions to accurately infer TF transcriptional activity based on the expression of their target genes. Using the inferred TF activity, NetAct further constructs a TF-based GRN, whose dynamics can then be evaluated and explored by mathematical modeling. Our approach in combining top-down and bottom-up systems biology approaches will contribute to a better understanding of the gene regulatory mechanism of cellular decision-making.

One of the key components of NetAct is a pre-compiled TF-target gene set database. Here, we have evaluated different types of TF-target databases in identifying knocked-down TFs using publicly available transcriptomics datasets. In this test, we have considered databases derived from the literature, gene co-expression, cis-motif

prediction, and TF-binding motif data. Our benchmark tests suggest that the literature-based database clearly outperformed the other databases.

NetAct also has a unique approach to infer the TF activity from the gene expression of the target genes with the consideration of activation/inhibition nature. From our in silico benchmark tests, we found that NetAct outperforms major activity inference methods, owing to the design of the filtering step and the use of a high-quality TF-target database. NetAct is also robust against some inaccuracy in the TF-target database and noises in gene expression data, because of its capability of filtering out irrelevant targets as well as remaining key targets.

One potential issue is the assignment of the sign of TF activity, as it is algorithmically assigned according to the correlation with TF expression. In the case where the TF expression is very noisy or the expression is completely unrelated to TF activity, the sign assignment might be inaccurate. To deal with this issue, we have devised a semi-manual approach that identifies the sign of TF activity according to the sign of other interacting TFs. Another potential issue is that some TFs from the same family may have very similar target genes; therefore, NetAct will have difficulty in identifying exactly which TF from the family is most relevant. Additional data resources, such as epigenomics[212], TF-binding data[48], and Hi-C data[213], will be helpful to address this problem. One of the future directions is to design methods to integrate these data resources.

Lastly, instead of constructing a global transcriptional regulatory network, NetAct focuses on modeling a core regulatory network with only interactions between key TFs. The underlying hypothesis is that these TFs and the associated regulatory interactions play major roles in controlling the gene expression of different cellular states and the patterns of state transitions. With the core network identified using NetAct, we can further perform simulations with mathematical modeling algorithms, such as RACIPE, to analyze the control mechanism of the core network. These simulations allow us to generate new hypotheses, which can be further tested experimentally. The validation data can further help to improve the model. Ideally, this needs to be an iterative process to refine a core network model, which is indeed another interesting future direction.

5. Discussion

5.1 Summary

The novel work presented in this thesis addresses key issues in the field of network modeling, mainly focusing on developing novel computational methods to identify both the factors and regulatory interactions that drive cellular state transitions.

In chapter three, we develop a method capable of identifying circuits, motifs, and motif coupling defining any type of state distribution or other quantitative score by analyzing the simulated gene expression distributions of all possible four-node circuits. This study is the first ever comprehensive analysis of four node circuits and does not suffer from issues of sampling seen in other motif identification methods, which rely on statistical comparisons to a limited number of similar networks. We also explore the behavior of gene circuits and motifs in a much more robust manner, through the use of RACIPE, which generates models with parameters that cover the breadth of biological contexts. Due to the direct analysis of all possible four node networks and the robust exploration of both circuit and motif behavior, our method provides a more rigorous approach to the identification of circuit motifs, the core information processing units of biological networks. We apply this method to identify circuit motifs responsible for two types of commonly observed state distributions, triangular state distributions, which are similar to bifurcations during development, and linear state distributions, which are often observed when a cell state transitions through an intermediate state, such as in EMT. We also show

that starting with any state distribution we can identify circuits capable of generating similar state distributions, the causal motifs, and the coupling of the causal motifs that define that specific type of state distribution. We conclude the first part of chapter three by showing how our method can be applied to identify phenomenological circuits, circuit motifs, and motif coupling for a single-cell RNA-seq dataset of human glutamatergic neuron differentiation. The interactions we identify as causative for the neuron differentiation state distribution are back up by literature information for well-known TFs involved in the process. The development of this method is highly valuable for informing network construction; The identification of causal circuits, motifs, and motif coupling from our method can be used as prior information for the construction of more accurate networks. The analysis of four-node circuits could potentially be extended to build larger transcriptional regulatory networks by combining the identified circuits, motifs, and coupling interactions into larger networks or by ensuring that inferred networks contain the enriched motifs and coupling interactions. Furthermore, future studies could also focus on implementing different logic rules for each circuit and evaluating the resulting behavior.

In the second section of chapter three, we extend our analysis to identify circuits, motifs, and coupling interactions responsible for key aspects of functioning regulatory networks, such as the ability to create multiple states and to respond to environmental cues. Circuits, motifs, and coupling interactions were identified for both functions. The behavior of the identified motifs was explored in networks of varying types and sizes leading to the identification of intermediate sized networks, around 30 nodes, as having

the highest scores for both the ability to create multiple states and respond to environmental cues. Both the motifs identified and the knowledge of intermediate sized networks preforming best are both significant findings. First, the motifs identified should be commonly shared interactions for many different networks involved in a plethora of biological contexts. Second, having some estimate of network size is extremely valuable for network inference since different approaches can generate networks of vastly different sizes, ranging from networks so large they are computationally intractable to networks so small that it is hard to believe their impact.

There are a few limitations of our current approach worth investigating in the future. First, current RACIPE modeling assumes AND logics to model regulation of multiple regulators to the same target gene. But, it is well known that circuits with different types of multivariate regulation can exhibit distinct behaviors, *e.g.*, feedforward loops with AND or OR logics [214]. An extensive analysis on this aspect would improve our understanding of the roles of logical rules in gene circuits. Second, we focused on steady-state gene expression distributions in this study, but temporal gene expression dynamics (both deterministic and stochastic) are crucial to evaluate dynamical properties of circuits [215], such as oscillations, excitability, and chaotic dynamics. It is possible that different types of circuits exhibit similar steady-state distributions but drastically different dynamics. Investigating both the steady-state and deterministic/stochastic dynamical behaviors of gene circuits could improve our understanding of gene regulatory circuit motifs in by providing additional dynamical information as to how circuit motifs generate specific attractor states. This is important because different motifs could potentially

generate similar state distributions but achieve this through disparate dynamics. This would help to classify motifs more accurately and could aid to infer the correct motifs governing cellular state transitions. Furthermore exploration of motif behavior with stochastic modeling could potentially revealing additional functions/behaviors of motifs that arise that were overlooked when only using deterministic modeling. Third, our current analysis for motif identification utilizes circuits of only four nodes. The approach can be generalized to analyze gene circuits or networks of larger sizes. It would be interesting to discover more complex circuit motifs and patterns of motif coupling that are not observed from the analysis of four-node circuits. It is worth noting however that our particular method would not scale well and evaluating all circuits of five or more nodes would probably require some type of subsampling approach due to the vast number of potential circuits. Fourth, we have observed that multiplicity gets saturated for large networks. Indeed, biological networks usually exhibit a limited number of cellular states, thus limiting the level of multiplicity. It is worth some further studies to elucidate the saturation of cellular states in biological networks. Finally, the formalism of our modeling limits our approach to identifying motifs of transcriptional gene regulation and does not capture important biological phenomena such as post-transcriptional and post-translational regulation, which are known to generate important biologically relevant circuits [216,217]. A potential future direction would be the functional motif analysis for circuits more than the transcriptional regulation.

In chapter four, we present a method for the inference of transcriptional networks from the activity of transcription factors. Briefly, this method identifies transcription factors

that are active using GSEA, infers the activity of the TFs using a weighted sum method, constructs networks from the MI and spearman's correlation of the activity of the TFs, and applies RACIPE to validate networks and identify TFs driving cell state transitions. We apply this method to model both TGF-B driven EMT and macrophage activation. The method is an example of top-down data driven approach's combined with more classical systems biology methods that utilize literature information and mathematical modeling. The significance of this method arises not from its ability to implicate new transcription factors in specific cell state transitions, but rather from its ability to create a mechanistic depiction of the regulatory interactions of known TFs and the prediction of specific cell state outcomes from direct TF perturbations.

A few limitations exist for NetAct. The literature-based database usually contains a small static number of curated regulatory interactions depending on the TF (~ 30), but these data have direct experimental evidence, therefore being more reliable than those from the other sources. However, the literature-based database has missing regulatory interactions, therefore maybe limiting the overall performance of NetAct. One way to address this issue is to further update the literature-based database, once new information is available. Another potential approach is to compile a database by combining different types of databases together. However, this might be quite challenging as different databases have data of very different sizes (the number of target genes) and quality. Future investigations on this direction can help to expand our knowledge of transcriptional regulation and meanwhile improve the performance of the algorithm.

5.2 Outlook

Overall, the work presented in this thesis represents novel ways to combine classical systems biology methodologies with top-down data driven approaches to infer mechanistic depictions of critical regulatory interactions. The use of mathematical modeling and the aid of literature databases represent significant steps forward for bioinformatic methods since they guarantee experimentally validated regulators are included and simultaneously ensuring the inferred networks accurately represent the biological context. On the other hand, these methods represent a major step forward for classical systems biology modeling by taking a data driven approach to inference of their networks certifying that the inferred networks are context specific. The method in chapter 3 demonstrates how we can identify the critical motifs driving different state distributions. In section 3.2.7 we show how this can be applied to neuron differentiation to understand how PAX6 and NeuroD6 gene expression programs interact to allow for the transition from radial glial progenitor cells, through intermediate progenitors, and finally to differentiated neurons.

The method in chapter 3 highlights types of interactions between two programs that support the dynamical behavior required for neuron differentiation (i.e. the correct type of state distribution). Further computational analysis can now be performed to predict how strengthening or weakening interactions in the identified motif can alter the state distribution, leading to hypothesis on how to enhance neuron differentiation that can be tested by experimental perturbations. Such experimental perturbations could include CRISPRi or CRISPRa. Furthermore, our method is easily applied to other single-cell

RNA-seq datasets for which it could also identify important regulatory motifs. Another key area where the method in chapter 3 could be applied is to identify how gene expression programs interact in disease progression. The approach in section 3.2.7 could potentially be used to identify the types of interactions that drive the transition from a healthy cellular state to a diseased one. By identifying both the regulatory programs and interactions between them that are allowing for a disease state, research efforts could be focused on disrupting those interactions to prevent disease.

This approach represents a novel method on how to prioritize the types of interactions causing any cellular state transition of interest and does so in a purely data driven way, addressing key issues with systems biology methods. This method demonstrates how classical systems biology approaches can be applied to infer context specific regulation without the use of any regulatory interaction database. Our method also relies entirely on comparisons to dynamical behaviors for inferring regulatory interactions, thereby ensure it can reproduce the cellular state transition being studies. Furthermore, the method is entirely deterministic so repeated use on the same dataset will result in the same information.

The method in Chapter 4 gives much more information on the regulatory interactions that control cellular state distributions. NetAct provides detailed information on the key TFs and how they specifically interact in a network to drive a specific state transition. NetAct also addresses key issues in the field by deriving interactions from the data, but also utilizes literature databases and mathematical modeling to ensure the inferred network is

accurate. The implementation of the TF GSEA in NetAct can identify context specific activities of different TFs. This could potentially be used to identify mechanisms that underlie the impact of genetic or environmental variation on differences in cellular state transitions. Performing TF GSEA on the same cellular state transitions in varied conditions could highlight the activities of TFs that are driving the observed differences. The network that NetAct infers also reflects interactions observed from the data since edges are constructed between TFs if their activities are above a mutual information threshold. While NetAct would be able to infer context specific interactions from the data it would not be able to identify novel interactions that are not in its database since inferred edges are required to agree with the literature database.

NetAct is modeled with RACIPE in order to ensure the behavior of the network matches experimental observations and also to predict how different TFs specifically impact the state distribution. The knockout analysis we present in Fig 4.5 shows direct predictions on how specific perturbations to TFs can shift cell state distributions in different ways. We show how this can be used to see the impact of knockdowns, but this method can also easily be applied to predict over-expression also. The resulting shift in cellular state distributions resulting from TF perturbations can also be easily converted into a score, for each TF, that could be used for target prioritization efforts in both research and industry. Furthermore, as stated above, the entire output of the TF GSEA could be used to both stratify diseases, and identify key TF targets for disease indication. This method could be invaluable for identifying drivers of any cellular state transition and how they should be perturbed. Further studies using NetAct to identify drivers of cellular state distributions

followed with direct experimental validation of the TF perturbation scores would help improve the impact for discovery.

Both methods provide information on the regulatory interactions that drive cellular state distributions however the situations in which they can be used and the information they provide are different. The method in Chapter 3 can be used to identify key motifs that should be overrepresented in networks and can provide phenomenological circuits capable of producing the same dynamics of the network being investigated. While this does not provide a mechanistic understanding of cellular state transitions, it can be used to prioritize avenues of research that would then lead to mechanistic understanding. The method presented in Chapter 3 also takes single-cell RNA-seq as an input. A major benefit to this method is that it does not require any prior information to infer regulatory motifs and circuits. This makes this method amenable to studying organisms or processes where the main factors and their interactions are unknown. There are relatively few organisms that have extensively studied regulatory mechanisms, presenting a perfect opportunity for the application of the method presented in Chapter 3.

Alternatively, NetAct takes bulk RNA-seq as an input and provides a more detailed description of the regulatory interactions that govern cellular state transitions. NetAct also provides direct predictions for how TFs can be experimentally perturbed to control the transition. This method is ideal for gaining a mechanistic understanding of the regulatory interactions that govern a well-studied process. The requirement for a database

of known interactions limits the application in areas where the important regulatory factors are not well known.

Together, these methods represent novel ways to model GRNs that address key issues such as the lack of context specificity from systems biology methods and the lack of dynamical consideration in bioinformatics. While they are designed for disparate situations, there are potential ways in which they could be combined. One such way is that motifs that are identified as over-represented could be used to inform network construction from NetAct. When generating networks with NetAct, network inference parameters could potentially be sampled to obtain multiple networks. The presence of the overrepresented motif in the inferred networks could then serve as one of the metrics for evaluating which network to use.

Regardless of how the methods are combined they show that bioinformatic and systems biology methods can be jointly implemented to with the goal of improving GRN inference. Improving GRN inference will greatly enhance our ability to treat disease, enhance differentiation protocols, and gain a mechanistic understanding of what is driving environmental or genetic differences in cellular state transitions. Through the implementation of methods that similarly combine Bioinformatics and Systems Biology, direct hypothesis on how to manipulate transcription factors to shift cellular state distributions toward a desired outcome can be formulated and tested. Altering cellular state transitions in a predicted manner is the key to developing treatments to any disease and ensuring robust differentiation protocols with minimal unintended cell types.

6. Appendix

This section contains unfinished work not to be included in the body of the thesis. The unfinished project is presented to document my contribution.

6.1 Modeling multi-state transitions

Another significant issue in the field of network modeling is the limited representation of multi-state networks, with network models often focusing on a single state or a simple cell-state transition between two states. However, cell-state transitions in many biological processes exhibit more complex dynamics. For instance, during embryonic development, transitions can start from a single state and lead to multiple final states. Similarly, the transition from naïve to primed pluripotency involves intermediate cell states. This highlights the necessity for network models that can effectively capture the activity of three or more states.

Moreover, networks inferred from data are often not simulated to ensure their ability to capture the steady-state behavior of the represented system. This issue becomes particularly crucial when modeling bifurcations, where a progenitor cell differentiates into two or more distinct cell types, in contrast to linear transitions involving hybrid states. Consequently, there is a pressing need for a novel method that not only infers a network encompassing more than two states but also optimizes the network's topology to ensure accurate representation of the steady-state behavior across multiple states.

In order to tackle these challenges, we acquired a publicly available single-cell RNA-seq dataset of in-vitro myelopoiesis[218]. This dataset serves as the foundation for developing a method that addresses the aforementioned issues. The dataset specifically captures the induced differentiation of human induced pluripotent stem cells (hiPSCs) into macrophages, wherein a substantial number of cells undergo fibroblast differentiation instead. Importantly, the dataset encompasses both bifurcation events and linear transitions, providing valuable examples for analysis. The dataset comprises 135,000 cells obtained from three donors, and it includes 20 single-cell RNA-seq measurements that were collected over a period of 38 days **Figure 6.1**.

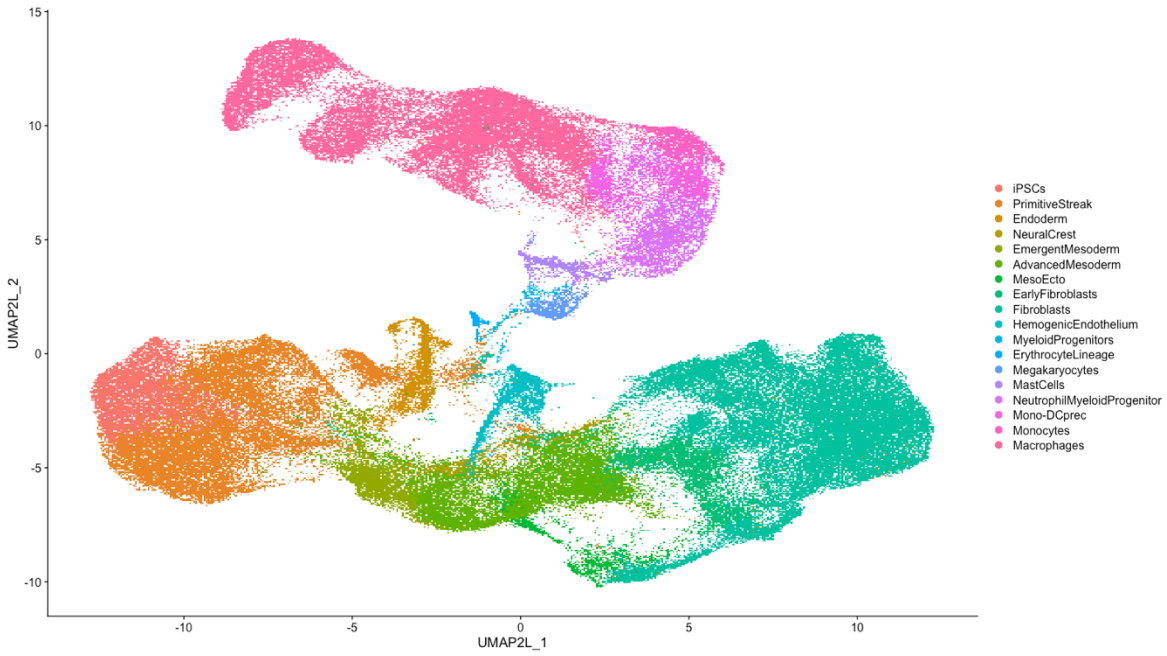

Figure 6.1. UMAP projection of in vitro myelopoiesis. Cells are colored by cell type. Dataset consists of 135,000 cells from three donors captured over a period of 38 days

For the initial development of our method, we focused on the bifurcation of primitive streak cells into either emergent mesoderm or endoderm lineages. Initially, we subset the data to remove other cell types, and then down-sampled it to ensure an equal

representation of each state (**Figure 6.2**). This equal representation was necessary to prevent bias in the inferred regulatory interactions due to states with a larger number of cells. Next, the down-sampled dataset was divided into two lineages: one consisting of primitive streak and endoderm cells, and the other consisting of primitive streak and emergent mesoderm cells.

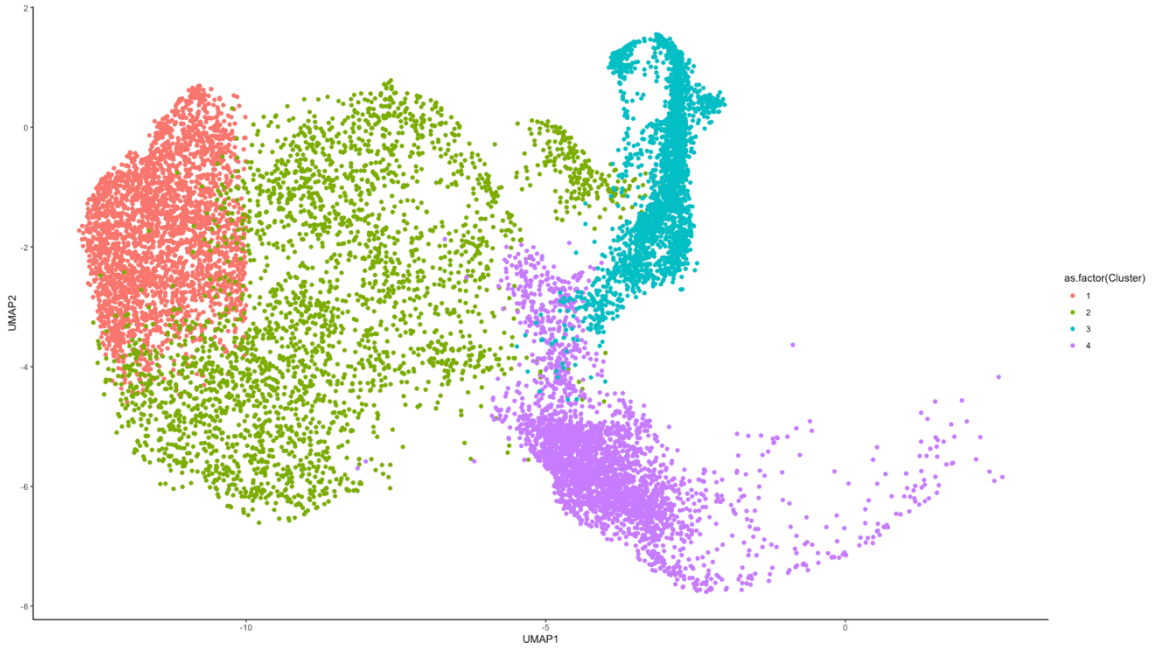

Figure 6.2 Umap projection of down-sampled and subset dataset for network construction. Original dataset was subset to include only first bifurcation and down-sampled to ensure even representation of cell states . Cells are colored by cell type.

To identify the active TFs in our dataset and quantify their activity, we applied Scenic to both lineages individually. Lineages were defined by cell type, with one lineage containing iPSC, primitive streak, and emergent mesoderm cells while the other lineage consisting of iPSC, primitive streak and endoderm cells. This analysis generated activity matrices for each lineage, visualized with UMAP projections in **Figure 6.3**, which captured the activity levels of each TF in every cell for each lineage. Instead of using

counts, we used the activity matrices with Seurat to perform differential activity analysis. By adjusting both the p-value and fold-change thresholds, we identified multiple sets of differentially active TFs for each lineage.

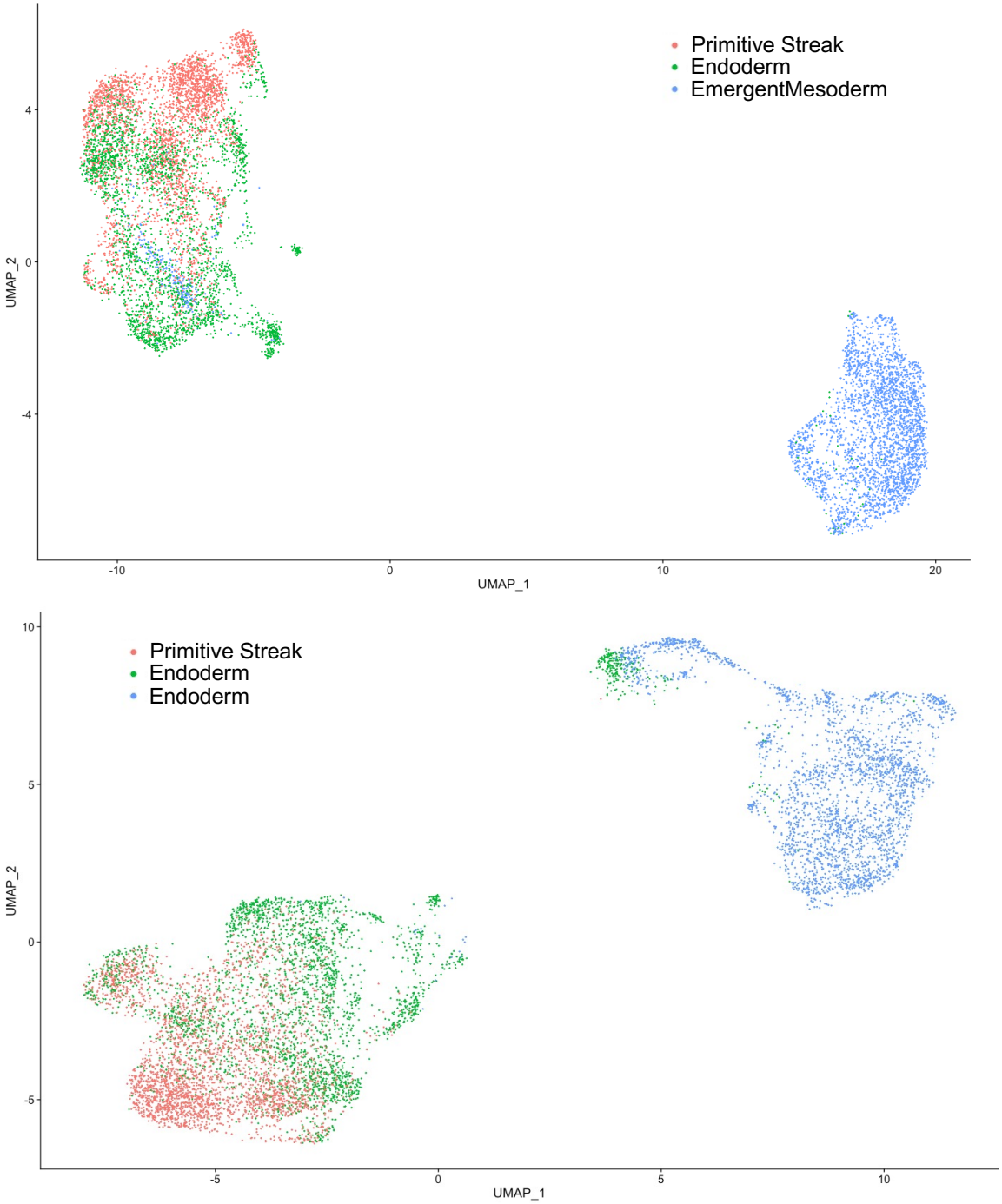

Figure 6.3 Umap projection of activity of identified regulons for each lineage from SCENIC. Colored by Cell type

Next, we constructed networks based on the differentially active TFs. To do this, we calculated the mutual information (MI) of the TF activity separately for each lineage and

considered an edge between two TFs if the MI exceeded a certain threshold. The sign of the edge (activating or inhibitory) was inferred from the Spearman correlation between TFs. We constructed 100 different networks for each lineage by varying parameters such as fold change, p-value, and the MI cutoff. Six representative networks are shown for each lineage in **Figure 6.4**

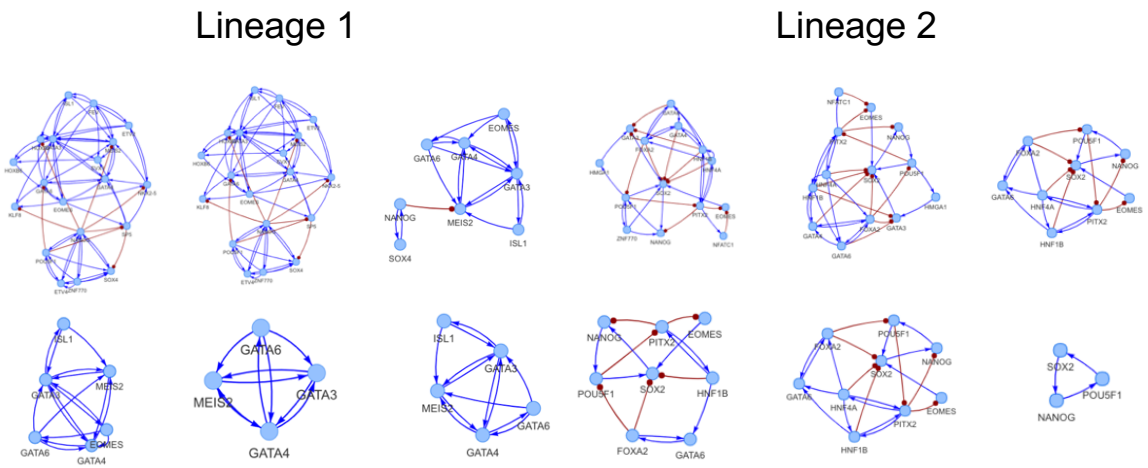

Figure 6.4 Example of different networks generated for each lineage. Varied parameters for network construction such as MI of activity or p-value and LFC of differential activity analysis.

To determine the network that best represents each lineage, we simulated all the inferred networks using RACIPE. We then aimed to develop a scoring metric that compares the simulated data to the experimental data using Gaussian Mixture Models (GMMs). We fit a GMM to each lineage of the experimental data and utilized it to assign the cell type to each model simulated by RACIPE. Specifically, GMMs were fitted to the UMAP projections of the experimental data and used to classify the UMAP projections of the simulated data. Currently, we are using the log likelihood of the GMM to assess how well each RACIPE simulation fits the data, but further validation is required to confirm its

effectiveness as a score. An example of three RACIPE simulations and their log likelihood score are shown in **Figure 6.5**

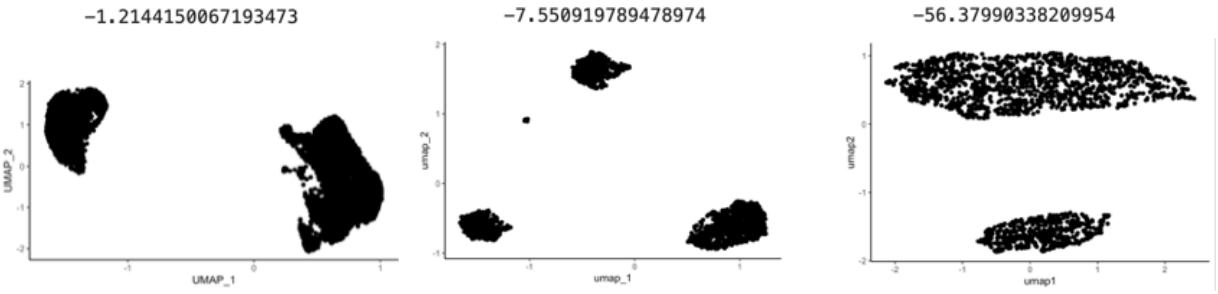

Figure 6.5 UMAP projections of RACIPE simulations of three inferred networks. Top is log likelihood score which indicates the similarity of the distribution to the original data the GMM was fit to.

Once the scoring criteria for comparing simulated and experimental data has been established, along with the selection of networks that best represent each lineage, our next step is to integrate these lineage-specific networks into a unified network while preserving the behavior of each individual network. To achieve this, we will identify overlapping interactions between the networks and sample different interactions to combine them. By generating a large number of interconnected networks with varying arrangements, we can apply the same scoring metric used for the lineage-specific networks to identify the unified network that best recapitulates the experimentally observed bifurcation. Once we have constructed the unified network, we can extend our method to analyze other multi-state transitions present in the dataset, including linear transitions.

The development of a method specifically tailored for inferring networks in multistate transitions, optimized to closely align with the biological context, holds immense value in

the field of network modeling. It paves the way for the identification of networks that can better capture complex processes such as embryonic development and EMT. Unified networks will significantly contribute to the optimization of differentiation protocols by enabling the identification of specific perturbations that can shift cell state distributions towards a desired lineage. Moreover, this framework holds the potential to unravel how specific genetic variants can lead to altered developmental trajectories and increased susceptibility to various diseases.

7 Bibliography

1. Miroshnikova, Y. A., Shahbazi, M. N., Negrete, J., Chalut, K. J. & Smith, A. Cell state transitions: catch them if you can. *Development* **150**, dev201139 (2023).
2. Casey, M. J., Stumpf, P. S. & MacArthur, B. D. Theory of cell fate. *WIREs Syst. Biol. Med.* **12**, (2020).
3. Mulas, C., Chaigne, A., Smith, A. & Chalut, K. J. Cell state transitions: definitions and challenges. *Development* **148**, dev199950 (2021).
4. Pera, M. F. & Rossant, J. The exploration of pluripotency space: Charting cell state transitions in peri-implantation development. *Cell Stem Cell* **28**, 1896–1906 (2021).
5. Blanpain, C. & Fuchs, E. Epidermal Stem Cells of the Skin. *Annu. Rev. Cell Dev. Biol.* **22**, 339–373 (2006).
6. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
7. Yeh, C.-Y., Huang, W.-H., Chen, H.-C. & Meir, Y.-J. J. Capturing Pluripotency and Beyond. *Cells* **10**, 3558 (2021).
8. Ribatti, D., Tamma, R. & Annese, T. Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Transl. Oncol.* **13**, 100773 (2020).
9. Deyneko, I. V. *et al.* Modeling and cleaning RNA-seq data significantly improve detection of differentially expressed genes. *BMC Bioinformatics* **23**, 488 (2022).
10. D'Agostino, N., Li, W. & Wang, D. High-throughput transcriptomics. *Sci. Rep.* **12**, 20313, s41598-022-23985–1 (2022).
11. Corbett, A. H. Post-transcriptional regulation of gene expression and human disease. *Curr. Opin. Cell Biol.* **52**, 96–104 (2018).
12. Velázquez-Cruz, A., Baños-Jaime, B., Díaz-Quintana, A., De La Rosa, M. A. & Díaz-Moreno, I. Post-translational Control of RNA-Binding Proteins and Disease-Related Dysregulation. *Front. Mol. Biosci.* **8**, 658852 (2021).
13. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
14. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biol.* **14**, e2000640 (2016).
15. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
16. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
17. Morris, S. A. The evolving concept of cell identity in the single cell era. *Development* **146**, dev169748 (2019).
18. Chaigne, A. *et al.* Abscission Couples Cell Division to Embryonic Stem Cell Fate. *Dev. Cell* **55**, 195-208.e5 (2020).
19. Corominas-Murtra, B. *et al.* Stem cell lineage survival as a noisy competition for niche access. *Proc. Natl. Acad. Sci.* **117**, 16969–16975 (2020).
20. Maki, K. *et al.* Hydrostatic pressure prevents chondrocyte differentiation through heterochromatin remodeling. *J. Cell Sci.* **134**, jcs247643 (2021).

21. Almeida, N. *et al.* Employing core regulatory circuits to define cell identity. *EMBO J.* **40**, e106785 (2021).
22. Zhang, S., Tian, D., Tran, N. H., Choi, K. P. & Zhang, L. Profiling the transcription factor regulatory networks of human cell types. *Nucleic Acids Res.* **42**, 12380–12387 (2014).
23. He, B. & Tan, K. Understanding transcriptional regulatory networks using computational models. *Curr. Opin. Genet. Dev.* **37**, 101–108 (2016).
24. Ament, S. A. *et al.* Transcriptional regulatory networks underlying gene expression changes in Huntington's disease. *Mol. Syst. Biol.* **14**, e7435 (2018).
25. 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. *Nat. Methods* **17**, 147–154 (2020).
26. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
27. Ay, A. & Arnosti, D. N. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit. Rev. Biochem. Mol. Biol.* **46**, 137–151 (2011).
28. Katebi, A., Ramirez, D. & Lu, M. Computational systems-biology approaches for modeling gene networks driving epithelial–mesenchymal transitions. *Comput. Syst. Oncol.* **1**, (2021).
29. Shahzad, K. & J. Loor, J. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Curr. Genomics* **13**, 379–394 (2012).
30. Nguyen, H., Tran, D., Tran, B., Pehlivan, B. & Nguyen, T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief. Bioinform.* **22**, bbaa190 (2021).
31. Kang, Y., Thieffry, D. & Cantini, L. Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. *Front. Genet.* **12**, 617282 (2021).
32. Butte, A. The use and analysis of microarray data. *Nat. Rev. Drug Discov.* **1**, 951–960 (2002).
33. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
34. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).
35. Johnson, K. A. & Krishnan, A. Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. *Genome Biol.* **23**, 1 (2022).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
38. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

39. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
40. Schlitt, T. & Brazma, A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics* **8**, S9 (2007).
41. Greenfield, A., Madar, A., Ostrer, H. & Bonneau, R. DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. *PLoS ONE* **5**, e13397 (2010).
42. The DREAM5 Consortium *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
43. Walker, A. M. *et al.* Evaluating the performance of random forest and iterative random forest based methods when applied to gene expression data. *Comput. Struct. Biotechnol. J.* **20**, 3372–3386 (2022).
44. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* **5**, e12776 (2010).
45. Talekar, B. A Detailed Review on Decision Tree and Random Forest. *Biosci. Biotechnol. Res. Commun.* **13**, 245–248 (2020).
46. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
47. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ. - Comput. Inf. Sci.* **35**, 757–774 (2023).
48. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
49. Chen, S. & Mar, J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232 (2018).
50. Schwab, J. D., Kühlwein, S. D., Ikonomi, N., Kühl, M. & Kestler, H. A. Concepts in Boolean network modeling: What do they all mean? *Comput. Struct. Biotechnol. J.* **18**, 571–582 (2020).
51. Ni, Y., Müller, P., Wei, L. & Ji, Y. Bayesian graphical models for computational network biology. *BMC Bioinformatics* **19**, 63 (2018).
52. Städter, P., Schälte, Y., Schmiester, L., Hasenauer, J. & Stapor, P. L. Benchmarking of numerical integration methods for ODE models of biological systems. *Sci. Rep.* **11**, 2696 (2021).
53. Huang, S. Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. *Cancer Metastasis Rev.* **32**, 423–448 (2013).
54. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
55. Healy, C. P. & Deans, T. L. Genetic circuits to engineer tissues with alternative functions. *J. Biol. Eng.* **13**, (2019).
56. Jiménez, A., Cotterell, J., Munteanu, A. & Sharpe, J. A spectrum of modularity in multi-functional gene circuits. *Mol. Syst. Biol.* **13**, 925 (2017).

57. Ye, Y., Kang, X., Bailey, J., Li, C. & Hong, T. An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. *PLOS Comput. Biol.* **15**, e1006855 (2019).
58. Gorochowski, T. E. *et al.* Genetic circuit characterization and debugging using RNA -seq. *Mol. Syst. Biol.* **13**, 952 (2017).
59. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).
60. Milo, R. *et al.* Network Motifs: Simple Building Blocks of Complex Networks. *Science* **298**, 824–827 (2002).
61. Schaerli, Y. *et al.* A unified design space of synthetic stripe-forming networks. *Nat. Commun.* **5**, (2014).
62. Nordick, B. & Hong, T. Identification, visualization, statistical analysis and mathematical modeling of high-feedback loops in gene regulatory networks. *BMC Bioinformatics* **22**, 481 (2021).
63. Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S. & Smith, A. G. Defining an essential transcription factor program for naïve pluripotency. *Science* **344**, 1156–1160 (2014).
64. Dunn, S., Li, M. A., Carbognin, E., Smith, A. & Martello, G. A common molecular logic determines embryonic stem cell self-renewal and reprogramming. *EMBO J.* **38**, e100003 (2019).
65. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 (2015).
66. Sha, Y., Wang, S., Bocci, F., Zhou, P. & Nie, Q. Inference of Intercellular Communications and Multilayer Gene-Regulations of Epithelial–Mesenchymal Transition From Single-Cell Transcriptomic Data. *Front. Genet.* **11**, 604585 (2021).
67. Kohar, V. & Lu, M. Role of noise and parametric variation in the dynamics of gene regulatory circuits. *Npj Syst. Biol. Appl.* **4**, 1–11 (2018).
68. Kolmogorov–Smirnov Test. in *The Concise Encyclopedia of Statistics* 283–287 (Springer New York, 2008). doi:10.1007/978-0-387-32833-1_214.
69. Csardi, G. & Nepusz, T. the igraph software package for complex network research. *InterJournal* 1695 (2006).
70. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
71. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
72. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
73. Tripathi, S., Kessler, D. A. & Levine, H. Biological Networks Regulating Cell Fate Choice Are Minimally Frustrated. *Phys. Rev. Lett.* **125**, 088101 (2020).
74. Huang, B. *et al.* Decoding the mechanisms underlying cell-fate decision-making during stem cell differentiation by random circuit perturbation. *J. R. Soc. Interface* **17**, 20200500 (2020).
75. Lord, W. M., Sun, J. & Bollt, E. M. Geometric k-nearest neighbor estimation of entropy and mutual information. *Chaos Interdiscip. J. Nonlinear Sci.* **28**, 033114 (2018).

76. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **46**, 175–185 (1992).
77. Dodge, Y. *The Concise Encyclopedia of Statistics*. (Springer Science & Business Media, 2008).
78. Gabor Csardi, T. N. The igraph software package for complex network research. *InterJournal* 1695 (2006).
79. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**, 016110 (2006).
80. Newman, M. E. J. Analysis of weighted networks. *Phys. Rev. E* **70**, 056131 (2004).
81. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
82. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**, 8577–8582 (2006).
83. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**, 016110 (2006).
84. Chuang, H.-Y., Hofree, M. & Ideker, T. A Decade of Systems Biology. *Annu. Rev. Cell Dev. Biol.* **26**, 721–744 (2010).
85. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
86. Gardner, T. S. & Collins, J. J. Neutralizing noise in gene networks. *Nature* **405**, 520–521 (2000).
87. Huang, B. *et al.* Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Comput. Biol.* **13**, e1005456 (2017).
88. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).
89. Hong, J. *et al.* An incoherent feedforward loop facilitates adaptive tuning of gene expression. *eLife* **7**, e32323 (2018).
90. Panovska-Griffiths, J., Page, K. M. & Briscoe, J. A gene regulatory motif that generates oscillatory or multiway switch outputs. *J. R. Soc. Interface* **10**, 20120826 (2013).
91. Nordick, B., Yu, P. Y., Liao, G. & Hong, T. Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression. *Nucleic Acids Res.* **50**, 3693–3708 (2022).
92. van Dorp, M., Lannoo, B. & Carlon, E. Generation of oscillating gene regulatory network motifs. *Phys. Rev. E* **88**, (2013).
93. Thomas, P., Popović, N. & Grima, R. Phenotypic switching in gene regulatory networks. *Proc. Natl. Acad. Sci.* **111**, 6994–6999 (2014).
94. Hortsch, S. K. & Kremling, A. Characterization of noise in multistable genetic circuits reveals ways to modulate heterogeneity. *PLOS ONE* **13**, e0194779 (2018).
95. Hari, K. *et al. Emergent properties of coupled bistable switches*. http://biorxiv.org/lookup/doi/10.1101/2021.06.15.448553 (2021) doi:10.1101/2021.06.15.448553.
96. Jolly, M. K. *et al.* Coupling the modules of EMT and stemness: A tunable 'stemness window' model. *Oncotarget* **6**, 25161–25174 (2015).

97. Adler, M. & Medzhitov, R. Emergence of dynamic properties in network hypermotifs. *Proc. Natl. Acad. Sci.* **119**, e2204967119 (2022).
98. Adler, M., Szekely, P., Mayo, A. & Alon, U. Optimal Regulatory Circuit Topologies for Fold-Change Detection. *Cell Syst.* **4**, 171-181.e8 (2017).
99. Kohar, V. & Lu, M. Role of noise and parametric variation in the dynamics of gene regulatory circuits. *Npj Syst. Biol. Appl.* **4**, (2018).
100. Sabuwala, B., Hari, K., Abhishek, S. V. & Jolly, M. K. *Coupled Mutual Inhibition and Mutual activation motifs as tools for cell-fate control*. http://biorxiv.org/lookup/doi/10.1101/2022.05.27.493756 (2022) doi:10.1101/2022.05.27.493756.
101. Huang, B. *et al.* Decoding the mechanisms underlying cell-fate decision-making during stem cell differentiation by random circuit perturbation. *J. R. Soc. Interface* **17**, 20200500 (2020).
102. Katebi, A., Kohar, V. & Lu, M. Random Parametric Perturbations of Gene Regulatory Circuit Uncover State Transitions in Cell Cycle. *iScience* **23**, 101150 (2020).
103. Huang, B. *et al.* RACIPE: a computational tool for modeling gene regulatory circuits using randomization. *BMC Syst. Biol.* **12**, 74 (2018).
104. Ramirez, D., Kohar, V. & Lu, M. Toward Modeling Context-Specific EMT Regulatory Networks Using Temporal Single Cell RNA-Seq Data. *Front. Mol. Biosci.* **7**, 54 (2020).
105. Ye, Y., Kang, X., Bailey, J., Li, C. & Hong, T. An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. *PLoS Comput. Biol.* **15**, e1006855 (2019).
106. Duddu, A. S., Sahoo, S., Hati, S., Jhunjhunwala, S. & Jolly, M. K. Multi-stability in cellular differentiation enabled by a network of three mutually repressing master regulators. *J. R. Soc. Interface* **17**, 20200631 (2020).
107. Laurent, M. & Kellershohn, N. Multistability: a major means of differentiation and evolution in biological systems. *Trends Biochem. Sci.* **24**, 418–422 (1999).
108. Guantes, R. & Poyatos, J. F. Multistable Decision Switches for Flexible Control of Epigenetic Differentiation. *PLOS Comput. Biol.* **4**, e1000235 (2008).
109. Bhalla, U. S. & Iyengar, R. Emergent Properties of Networks of Biological Signaling Pathways. *Science* **283**, 381–387 (1999).
110. Li, M., Gao, H., Wang, J. & Wu, F.-X. Control principles for complex biological networks. *Brief. Bioinform.* **20**, 2253–2266 (2019).
111. Zañudo, J. G. T., Yang, G. & Albert, R. Structure-based control of complex networks with nonlinear dynamics. *Proc. Natl. Acad. Sci.* **114**, 7234–7239 (2017).
112. Liu, Y.-Y. & Barabási, A.-L. Control principles of complex systems. *Rev. Mod. Phys.* **88**, 035006 (2016).
113. Zhang, W. & Liu, H. T. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* **12**, 9–18 (2002).
114. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).
115. Bao, Y. *et al.* Analysis of Critical and Redundant Vertices in Controlling Directed Complex Networks Using Feedback Vertex Sets. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **25**, 1071–1090 (2018).

116. Bhattacharya, P., Raman, K. & Tangirala, A. K. Discovering adaptation-capable biological network structures using control-theoretic approaches. *PLOS Comput. Biol.* **18**, e1009769 (2022).
117. Watcham, S., Kucinski, I. & Gottgens, B. New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. *Blood* **133**, 1415–1426 (2019).
118. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019).
119. Xu, J., Lamouille, S. & Derynck, R. TGF-β-induced epithelial to mesenchymal transition. *Cell Res.* **19**, 156–172 (2009).
120. Lu, M., Jolly, M. K., Levine, H., Onuchic, J. N. & Ben-Jacob, E. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Natl. Acad. Sci.* **110**, 18144–18149 (2013).
121. Zhang, J. *et al.* TGF-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci. Signal.* **7**, (2014).
122. Lu, M. *et al.* Tristability in Cancer-Associated MicroRNA-TF Chimera Toggle Switch. *J. Phys. Chem. B* **117**, 13164–13174 (2013).
123. Jia, D. *et al.* Operating principles of tristable circuits regulating cellular differentiation. *Phys. Biol.* **14**, 035007 (2017).
124. Som, A. *et al.* The PluriNetWork: An Electronic Representation of the Network Underlying Pluripotency in Mouse, and Its Applications. *PLoS ONE* **5**, e15165 (2010).
125. Yang, J., Gao, C., Chai, L. & Ma, Y. A Novel SALL4/OCT4 Transcriptional Feedback Network for Pluripotency of Embryonic Stem Cells. *PLoS ONE* **5**, e10766 (2010).
126. Han, H. *et al.* TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* **5**, 11432 (2015).
127. Mehravar, M., Ghaemimanesh, F. & Poursani, E. M. An Overview on the Complexity of OCT4: at the Level of DNA, RNA and Protein. *Stem Cell Rev. Rep.* **17**, 1121–1136 (2021).
128. Han, D. *et al.* A balanced Oct4 interactome is crucial for maintaining pluripotency. *Sci. Adv.* **8**, eabe4375 (2022).
129. Tatetsu, H. *et al.* SALL4, the missing link between stem cells, development and cancer. *Gene* **584**, 111–119 (2016).
130. Shi, G. & Jin, Y. Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Res. Ther.* **1**, 39 (2010).
131. Moreira, S. *et al.* A Single TCF Transcription Factor, Regardless of Its Activation Capacity, Is Sufficient for Effective Trilineage Differentiation of ESCs. *Cell Rep.* **20**, 2424–2438 (2017).
132. Zhou, Q., Chipperfield, H., Melton, D. A. & Wong, W. H. A gene regulatory network in mouse embryonic stem cells. *Proc. Natl. Acad. Sci.* **104**, 16438–16443 (2007).
133. Kafri, R., Levy, M. & Pilpel, Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl. Acad. Sci.* **103**, 11653–11658 (2006).
134. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature* **388**, 167–171 (1997).

135. Pollen, A. A. *et al.* Molecular Identity of Human Outer Radial Glia during Cortical Development. *Cell* **163**, 55–67 (2015).
136. Thakurela, S. *et al.* Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discov.* **2**, 15045 (2016).
137. Sansom, S. N. *et al.* The Level of the Transcription Factor Pax6 Is Essential for Controlling the Balance between Neural Stem Cell Self-Renewal and Neurogenesis. *PLoS Genet.* **5**, e1000511 (2009).
138. Uittenbogaard, M. & Chiaramello, A. Constitutive overexpression of the basic helix-loop-helix Nex1/MATH-2 transcription factor promotes neuronal differentiation of PC12 cells and neurite regeneration. *J. Neurosci. Res.* **67**, 235–245 (2002).
139. Uittenbogaard, M., Baxter, K. K. & Chiaramello, A. NeuroD6 genomic signature bridging neuronal differentiation to survival via the molecular chaperone network. *J. Neurosci. Res.* **88**, 33–54 (2010).
140. Ochi, S., Manabe, S., Kikkawa, T. & Osumi, N. Thirty Years' History since the Discovery of Pax6: From Central Nervous System Development to Neurodevelopmental Disorders. *Int. J. Mol. Sci.* **23**, 6115 (2022).
141. Tutukova, S., Tarabykin, V. & Hernandez-Miranda, L. R. The Role of Neurod Genes in Brain Development, Function, and Disease. *Front. Mol. Neurosci.* **14**, 662774 (2021).
142. Uittenbogaard, M., Baxter, K. K. & Chiaramello, A. NeuroD6 genomic signature bridging neuronal differentiation to survival via the molecular chaperone network. *J. Neurosci. Res.* **88**, 33–54 (2010).
143. Bartholomä, A. & Nave, K.-A. NEX-1: a novel brain-specific helix-loop-helix protein with autoregulation and sustained expression in mature cortical neurons. *Mech. Dev.* **48**, 217–228 (1994).
144. Su, K. *et al. NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity*. http://biorxiv.org/lookup/doi/10.1101/2022.05.06.487898 (2022) doi:10.1101/2022.05.06.487898.
145. Hari, K., Ullanat, V., Balasubramanian, A., Gopalan, A. & Jolly, M. K. *Landscape of Epithelial Mesenchymal Plasticity as an emergent property of coordinated teams in regulatory networks*. http://biorxiv.org/lookup/doi/10.1101/2021.12.12.472090 (2021) doi:10.1101/2021.12.12.472090.
146. Campbell, C., Shea, K., Yang, S. & Albert, R. Motif profile dynamics and transient species in a Boolean model of mutualistic ecological communities. *J. Complex Netw.* **4**, 127–139 (2016).
147. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).
148. Clauss, B. & Lu, M. A Quantitative Evaluation of Topological Motifs and Their Coupling in Gene Circuit State Distributions. 2022.07.19.500691 Preprint at https://doi.org/10.1101/2022.07.19.500691 (2022).
149. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
150. Huang, B. *et al.* Interrogating the topological robustness of gene regulatory circuits by randomization. *PLoS Comput. Biol.* **13**, e1005456 (2017).

151. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).
152. Tian, X.-J., Zhang, X.-P., Liu, F. & Wang, W. Interlinking positive and negative feedback loops creates a tunable motif in gene regulatory networks. *Phys. Rev. E* **80**, 011926 (2009).
153. Github repository of this study. https://github.com/huanglijiaU201614513/circuitanalysis.
154. Kafri, R., Levy, M. & Pilpel, Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl. Acad. Sci.* **103**, 11653–11658 (2006).
155. Cooke, J., Nowak, M. A., Boerlijst, M. & Maynard-Smith, J. Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* **13**, 360–364 (1997).
156. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**, S7 (2006).
157. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838 (2016).
158. Ament, S. A. *et al.* Transcriptional regulatory networks underlying gene expression changes in Huntington's disease. *Mol. Syst. Biol.* **14**, e7435 (2018).
159. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst.* **5**, 251-267.e3 (2017).
160. Carré, C., Mas, A. & Krouk, G. Reverse engineering highlights potential principles of large gene regulatory network design and learning. *Npj Syst. Biol. Appl.* **3**, 17 (2017).
161. Fiers, M. W. E. J. *et al.* Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* doi:10.1093/bfgp/elx046.
162. Gérard, C. & Goldbeter, A. Temporal self-organization of the cyclin/Cdk network driving the mammalian cell cycle. *Proc. Natl. Acad. Sci.* **106**, 21643–21648 (2009).
163. Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M. & Shapiro, L. Global Analysis of the Genetic Network Controlling a Bacterial Cell Cycle. *Science* **290**, 2144–2148 (2000).
164. Li, F., Long, T., Lu, Y., Ouyang, Q. & Tang, C. The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci.* **101**, 4781–4786 (2004).
165. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
166. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. *Cell* **132**, 1049–1061 (2008).
167. Loh, Y.-H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431 (2006).
168. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838 (2016).
169. Liao, J. C. *et al.* Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci.* **100**, 15522–15527 (2003).
170. Huang, B. *et al.* Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Comput. Biol.* **13**, e1005456 (2017).

171. Katebi, A., Kohar, V. & Lu, M. Random Parametric Perturbations of Gene Regulatory Circuit Uncover State Transitions in Cell Cycle. *iScience* **23**, 101150 (2020).
172. Ramirez, D., Kohar, V. & Lu, M. Toward Modeling Context-Specific EMT Regulatory Networks Using Temporal Single Cell RNA-Seq Data. *Front. Mol. Biosci.* **7**, 54 (2020).
173. Huang, B. *et al.* Decoding the mechanisms underlying cell-fate decision-making during stem cell differentiation by random circuit perturbation. *J. R. Soc. Interface* **17**, 20200500 (2020).
174. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
175. Han, H. *et al.* TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* **5**, 11432 (2015).
176. Liu, Z.-P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, (2015).
177. Essaghir, A. & Demoulin, J.-B. A Minimal Connected Network of Transcription Factors Regulated in Human Tumors and Its Application to the Quest for Universal Cancer Biomarkers. *PLOS ONE* **7**, e39666 (2012).
178. Jiang, C., Xuan, Z., Zhao, F. & Zhang, M. Q. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.* **35**, D137–D140 (2007).
179. Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database J. Biol. Databases Curation* **2016**, (2016).
180. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
181. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. *Nucleic Acids Res.* **24**, 238–241 (1996).
182. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
183. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
184. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
185. Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database J. Biol. Databases Curation* **2016**, baw105 (2016).
186. Garcia-Alonso, L., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *bioRxiv* 337915 (2018) doi:10.1101/337915.
187. Alvarez, M. J., Sumazin, P., Rajbhandari, P. & Califano, A. Correlating measurements across samples improves accuracy of large-scale expression profile experiments. *Genome Biol.* **10**, R143 (2009).
188. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

189. Hu, M. & Qin, Z. S. Query Large Scale Microarray Compendium Datasets Using a Model-Based Bayesian Approach with Variable Selection. *PLOS ONE* **4**, e4495 (2009).
190. Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **27**, 2263–2270 (2011).
191. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**, S7 (2006).
192. Levandowsky, M. & Winter, D. Distance between Sets. *Nature* **234**, 34 (1971).
193. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).
194. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun. Stat. Appl. Methods* **22**, 665–674 (2015).
195. Sartor, M. A. *et al.* ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* **26**, 456–463 (2010).
196. Schiffer, M., Von Gersdorff, G., Bitzer, M., Susztak, K. & Böttinger, E. P. Smad proteins and transforming growth factor-β signaling. *Kidney Int.* **58**, S45–S52 (2000).
197. Zhang, Y., Feng, X.-H. & Derynck, R. Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-β-induced transcription. *Nature* **394**, 909–913 (1998).
198. Lu, M., Jolly, M. K., Levine, H., Onuchic, J. N. & Ben-Jacob, E. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Natl. Acad. Sci.* **110**, 18144–18149 (2013).
199. Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front. Oncol.* **5**, (2015).
200. Subbalakshmi, A. R. *et al.* KLF4 Induces Mesenchymal–Epithelial Transition (MET) by Suppressing Multiple EMT-Inducing Transcription Factors. *Cancers* **13**, 5135 (2021).
201. Zhu, M. & Kyprianou, N. Role of androgens and the androgen receptor in epithelial-mesenchymal transition and invasion of prostate cancer cells. *FASEB J.* **24**, 769–777 (2010).
202. Chang, C.-J. *et al.* p53 regulates epithelial–mesenchymal transition and stem cell properties through modulating miRNAs. *Nat. Cell Biol.* **13**, 317–323 (2011).
203. Dai, X. *et al.* CBP-mediated Slug acetylation stabilizes Slug and promotes EMT and migration of breast cancer cells. *Sci. China Life Sci.* **64**, 563–574 (2021).
204. Wang, Y. *et al.* EGR1 induces EMT in pancreatic cancer via a P300/SNAI2 pathway. *J. Transl. Med.* **21**, 201 (2023).
205. Kim, I. *et al.* Specific protein 1(SP1) regulates the epithelial-mesenchymal transition via lysyl oxidase-like 2(LOXL2) in pancreatic ductal adenocarcinoma. *Sci. Rep.* **9**, 5933 (2019).
206. Tian, B. *et al.* The NFκB subunit RELA is a master transcriptional regulator of the committed epithelial-mesenchymal transition in airway epithelial cells. *J. Biol. Chem.* **293**, 16528–16545 (2018).
207. Steinway, S. N. *et al.* Network Modeling of TGFβ Signaling in Hepatocellular Carcinoma Epithelial-to-Mesenchymal Transition Reveals Joint Sonic Hedgehog and Wnt Pathway Activation. *Cancer Res.* **74**, 5963–5977 (2014).

208. Piccolo, V. *et al.* Opposing macrophage polarization programs show extensive epigenomic and transcriptional cross-talk. *Nat. Immunol.* **18**, 530–540 (2017).
209. Mosser, D. M. & Edwards, J. P. Exploring the full spectrum of macrophage activation. *Nat. Rev. Immunol.* **8**, 958–969 (2008).
210. Bae, S. *et al.* MYC-mediated early glycolysis negatively regulates proinflammatory responses by controlling IRF4 in inflammatory macrophages. *Cell Rep.* **35**, 109264 (2021).
211. Hu, X. & Ivashkiv, L. B. Cross-regulation of Signaling Pathways by Interferon-γ: Implications for Immune Responses and Autoimmune Diseases. *Immunity* **31**, 539–550 (2009).
212. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).
213. Malysheva, V., Mendoza-Parra, M. A., Saleem, M.-A. M. & Gronemeyer, H. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. *Genome Med.* **8**, 57 (2016).
214. Alon, U. *An introduction to systems biology : design principles of biological circuits*. (CRC Press, c2020).
215. Bennett, M. R., Volfson, D., Tsimring, L. & Hasty, J. Transient Dynamics of Genetic Regulatory Networks. *Biophys. J.* **92**, 3501–3512 (2007).
216. Markevich, N. I., Hoek, J. B. & Kholodenko, B. N. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J. Cell Biol.* **164**, 353–359 (2004).
217. Li, C. *et al.* MicroRNA governs bistable cell differentiation and lineage segregation via a noncanonical feedback. *Mol. Syst. Biol.* **17**, (2021).
218. Alsinet, C. *et al.* Robust temporal map of human in vitro myelopoiesis using single-cell genomics. *Nat. Commun.* **13**, 2885 (2022).