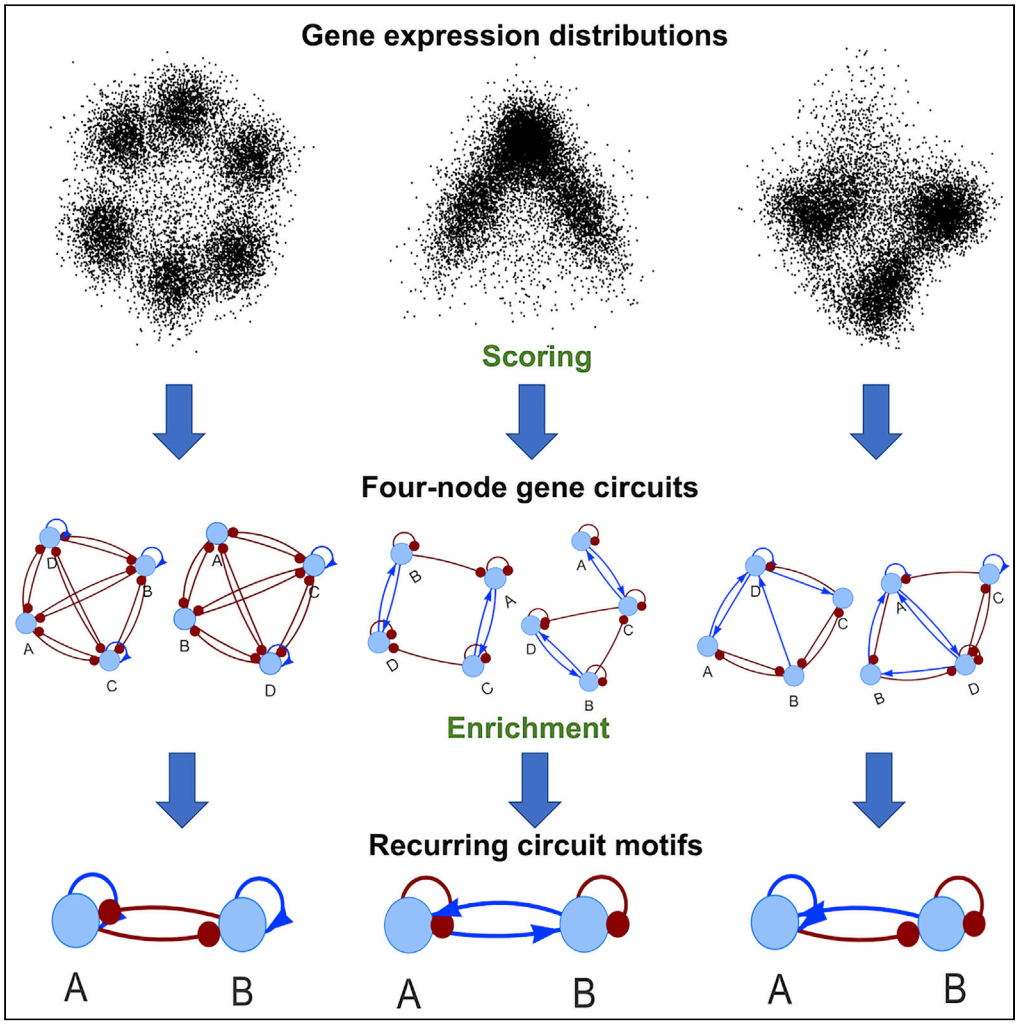
iScience

CellPress
OPEN ACCESS

Article

A quantitative evaluation of topological motifs and their coupling in gene circuit state distributions



Benjamin Clauss, Mingyang Lu

m.lu@northeastern.edu

Highlights

A novel quantitative circuit motif analysis based on circuit functions

An enrichment analysis to identify important circuit motifs and their coupling

Quantitative classification of circuit motifs and patterns of motif coupling

Circuit motif analysis directly from single cell gene expression distribution

Clauss & Lu, iScience 26, 106029
February 17, 2023
https://doi.org/10.1016/j.isci.2023.106029

iScience

CellPress
OPEN ACCESS

Article

A quantitative evaluation of topological motifs and their coupling in gene circuit state distributions

Benjamin Clauss[2,3,4] and Mingyang Lu[1,2,3,4,5,*]

SUMMARY

One of the major challenges in biology is to understand how gene interactions collaborate to determine overall functions of biological systems. Here, we present a new computational framework that enables systematic, high-throughput, and quantitative evaluation of how small transcriptional regulatory circuit motifs, and their coupling, contribute to functions of a dynamical biological system. We illustrate how this approach can be applied to identify four-node gene circuits, circuit motifs, and motif coupling responsible for various gene expression state distributions, including those derived from single-cell RNA sequencing data. We also identify seven major classes of four-node circuits from clustering analysis of state distributions. The method is applied to establish phenomenological models of gene circuits driving human neuron differentiation, revealing important biologically relevant regulatory interactions. Our study will shed light on a better understanding of gene regulatory mechanisms in creating and maintaining cellular states.

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.[1,2] Over the years, researchers have focused on studying gene circuit motifs, defined as reoccurring small circuit topologies within larger biological gene regulatory networks.[3] It has been shown, by approaches in synthetic biology,[4] computational systems-biology modeling,[5,6] and experimental systems biology,[7] 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[8,9]; a two-node toggle switch circuit can generate bistability[10,11]; and an incoherent feed forward loop can achieve adaptation.[12] 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. Owing 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 lager networks in general.

Gene circuit motifs were classically identified by searching the topology of a large biological network, such those from *Escherichia coli* and yeast, for the presence of smaller circuit motifs.[2,3,13] Motifs are important when they are over-represented in biological networks compared to similarly generated random networks. This approach usually only considers the frequency of circuit motifs' appearance, but not their functionality, for initial identification. To address this issue, recent studies[5,6,14,15] 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[16–18] and multiple stable steady states.[6,19,20] Ye et al.[6] 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.[6] Schaerli et al.[14] 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

[1]Department of Bioengineering, Northeastern University, Boston, MA 02115, USA

[2]Center for Theoretical Biological Physics, Northeastern University, Boston, MA 02115, USA

[3]Genetics Program, Graduate School of Biomedical Sciences, Tufts University, Boston, MA 02111, USA

[4]The Jackson Laboratory, Bar Harbor, ME 04609, USA

[5]Lead contact

*Correspondence: m.lu@northeastern.edu

https://doi.org/10.1016/j.isci.2023.106029

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 is no 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.[21,22] In particular, the role of circuit coupling may depend on the proportion of shared nodes between the two coupled circuit motifs.[5] Another recent study[23] 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. In this study, we performed the first-ever comprehensive analysis on all non-redundant four-node transcriptional regulatory circuits. Compared to previous studies on three-node circuits,[6,14,24] our analysis has the following advantages. First, there are around 60,000 non-redundant four-node circuits (see STAR Methods), 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. This 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),[25] 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,[10,21,26] gene networks regulating stem cell differentiation,[27] cell cycle,[28] B-cell development,[29] and epithelial-mesenchymal transitions[30,31] (EMTs). 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 first give an overview of the computational framework. We then illustrate the methodology with two applications to identify all four-node gene circuits allowing a triangular three-state distribution and a linear three-state distribution. From an enrichment analysis, we can identify the enriched two-node circuit motifs and the patterns of their coupling. Next, we demonstrate how this framework can be applied to identify (1) clusters of circuits with distinct gene expression behaviors and (2) circuits with similar state distributions to any other starting circuit. Finally, we demonstrate how our method can be applied to identify circuits motifs and their coupling responsible for experimentally observed single-cell gene expression state distributions.

RESULTS

A quantitative method to identify circuits with defined functionalities

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 Figure 1.

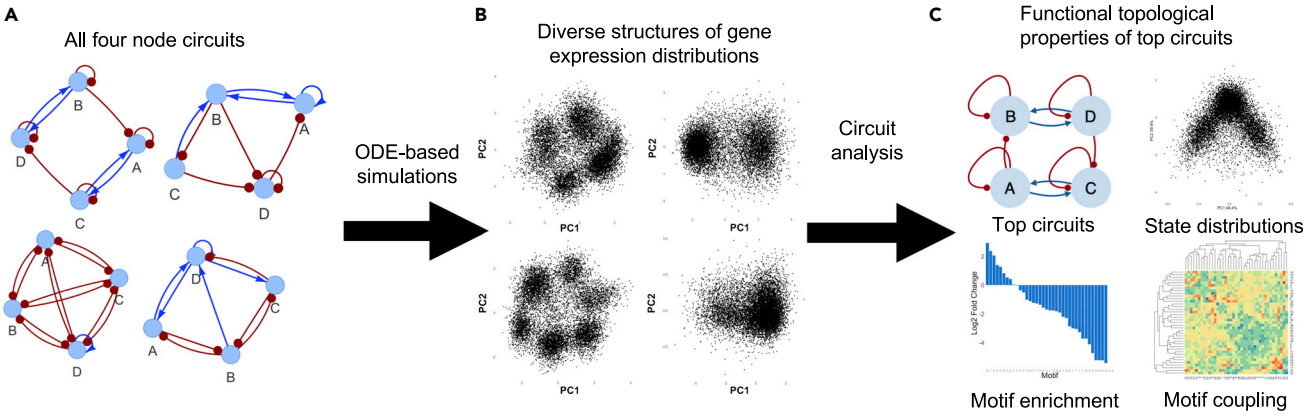


Figure 1. A schematic overview of circuit motif analysis

(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 (Figure 1A) (see STAR Methods 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[25] to generate an ensemble of 10,000 ODE models with randomly generated kinetic parameters (see STAR Methods 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 (Figure 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.[25] 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 (Figure 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.

Characterizing circuits of three states with a triangular or linear state distribution

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 Figure 2A. 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.[32,33] 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 STAR Methods 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 Figure 2B. As demonstrated in the PCA projections of the simulated gene expression data of the corresponding circuits (Figure 2B, bottom row), these circuits create gene expression state distributions of three states arranged in a triangular shape (see Figures S9–S13 for the contour maps and the outcomes for RACIPE simulations with random parameters derived from Gaussian distributions). Of interest, the topologies of top ranked circuits are remarkably

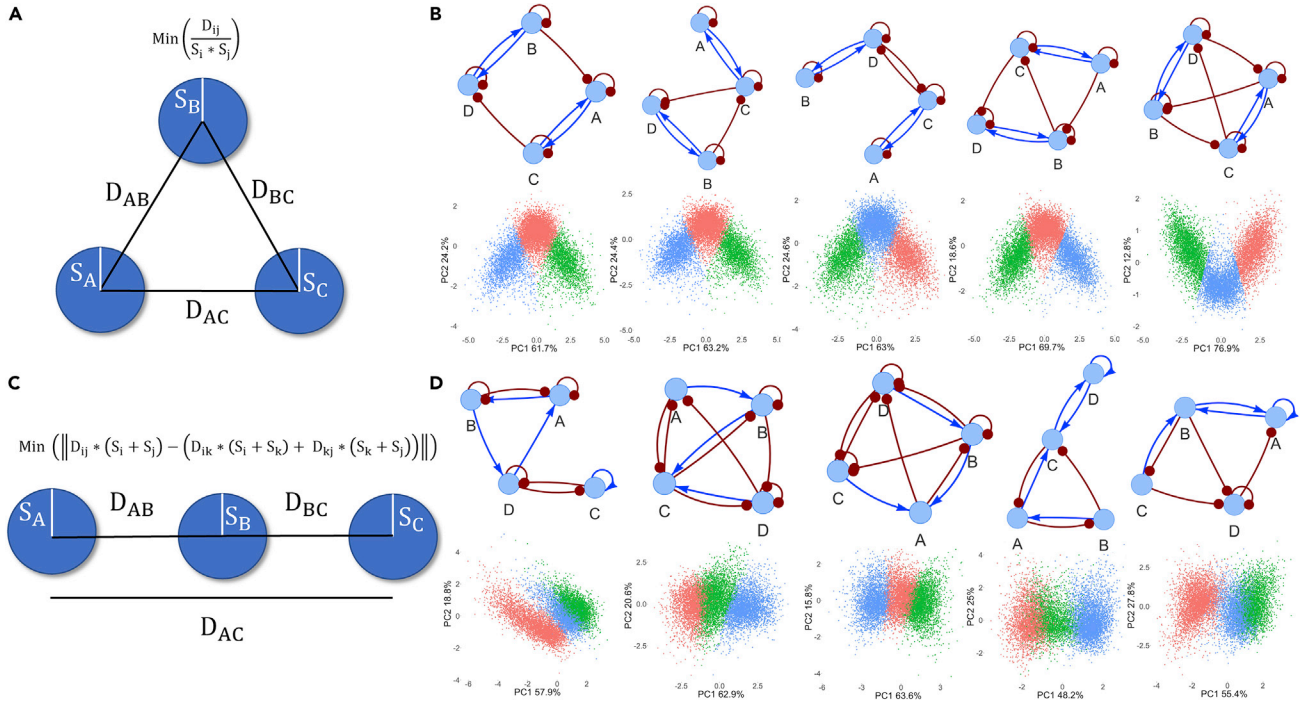


Figure 2. Identifying four-node circuits with triangular and linear state distributions

(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). The kinetic parameters of these models were randomized using uniform distributions, and the outcomes using Gaussian distributions are shown in Figures S9–S13. See Figures S14–S18 and S24–28 for scatterplots of gene pairs. 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.

similar with clear patterns of two-node circuit motifs, such as motif 25 (Figure S1 for the list of motifs and their indices) appearing twice in each network without sharing a node and motif 25 and 16 co-occurring while always sharing a node (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 Figure 2C (see FiguresS19–S23 for the contour maps and the outcomes for RACIPE simulations with random parameters derived from Gaussian distributions). 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.[34–36] 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 the STAR Methods section. The top five ranked circuits of linear state distribution are illustrated in Figure 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. Although 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.

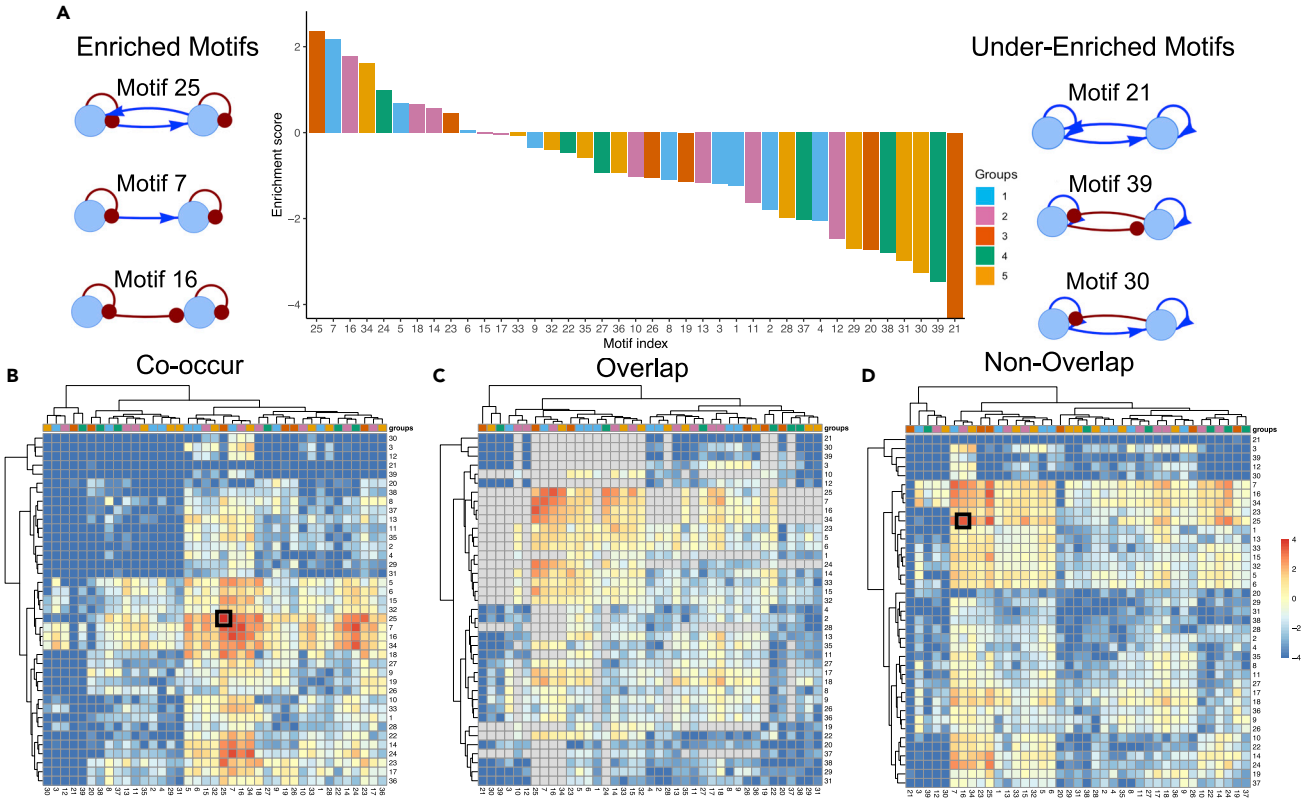


Figure 3. Motif enrichment analysis using the triangularity score

(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 (see STAR Methods section for details).

Enrichment analysis identifies circuit motifs and motif 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 (see Figure S1) and identified their occurrence in each four-node circuit. We evaluated the enrichment of each two-node motif in circuits with top triangularity scores (600 circuits, see STAR Methods for details), as shown in Figure 3A. 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). Of interest, 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. Of interest, 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,[3] and we observe that removing the negative autoregulation in the top ranked circuits would generate an additional state cluster (FiguresS29–S38). 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 (Figure 3B), for co-occurrence of two motifs with a shared node (motifs with overlapping,

Figure 3C), and for co-occurrence of two motifs without any overlapping node (Figure 3D). Of interest, the topmost enriched motif coupling patterns are (1) between two motifs of #25 and (2) between motifs 25 and 16. Furthermore, coupling between two motifs of #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 (Figures S2 and S3). 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 (see Figure S41 for simulations of motif 25 by itself and Figures S43 and S44 for two motif 25s are present but uncoupled). In addition, we observed that motifs 14, 17, 18, and 24 all have relatively higher enrichment for coupling with motifs 7 and 25 (see Figure 3B), despite having relatively lower enrichment of as standalone motif (Figure 3A). For motif coupling without a shared node (Figure 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 STAR Methods 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 (Figures S39and S40).

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 was identified, as shown in Figure S4. 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 (see Figure S1). 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; whereas 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[3] may contribute to the ability of the identified motifs to generate separate states. Figure S42 shows that motif 34 by itself is not able to generate multiple states; whereas from the top five circuits for the linear state distribution (Figure 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,[10,11,37,38] 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 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 versus 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 Figures S4D–S4F.

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 (Figures 3 and S4). 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.

Circuit motifs of linear and triangular state distributions are frequently observed in biological networks

Next, we searched for the occurrence of the top enriched motifs in both the linear and triangular scores in PluriNetWork,[39] 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. Among these motifs, Sall4 and Oct4[40] 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.[41] Oct4 is a well-known pan-pluripotency transcription factor, and Sall4 is important in ESC proliferation and differentiation.[40,42,43] 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,

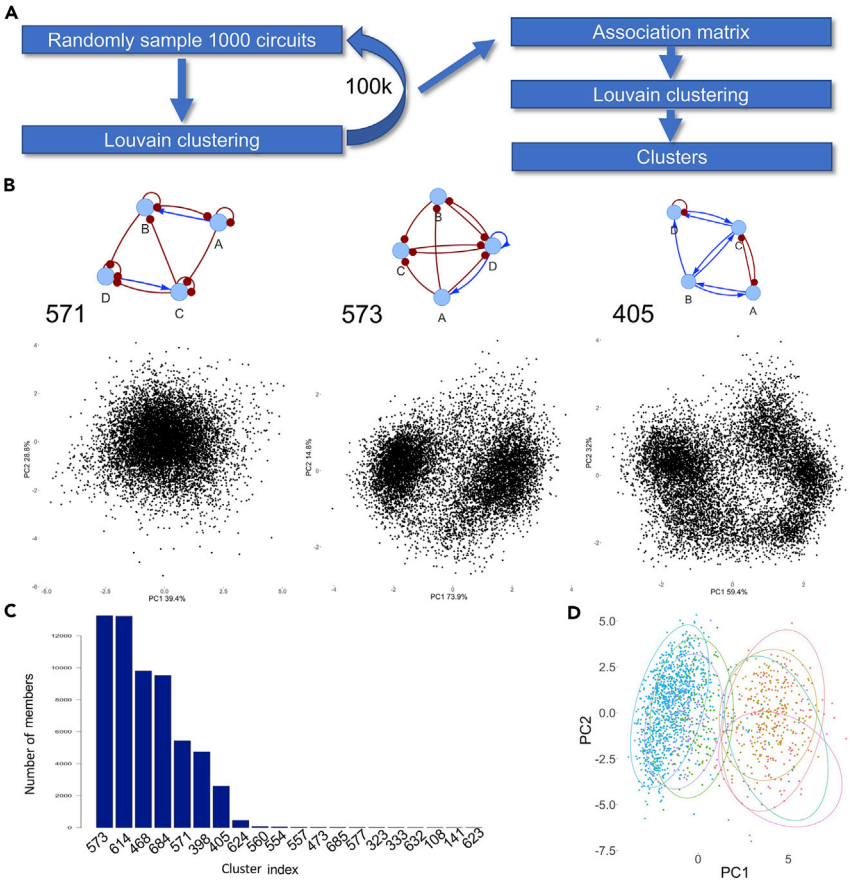


Figure 4. Clustering of all non-redundant four-node gene circuits by the similarity of state distributions

(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.

demonstrating its importance in generating three states.[44] 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.

Classifying types of state distributions of all four-node circuits

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 STAR Methods). We then devised a subsampling approach of Louvain clustering (schematic overview in Figure 4A, details in STAR 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. For example, Figure 4B 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 (Figure 4C), we observe seven major circuit clusters. Of interest, circuit clusters with more members tend to have simpler state distributions (*i.e.*, distributions with one or two gene expression clusters), whereas 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 principal component analysis (PCA). To do so, we constructed a vector of statistics summarizing the expression of each circuit (details in the STAR Methods section), for the most representative circuits in the largest seven clusters and projected the data to the first two principal components (Figure 4D). 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, whereas 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, whereas 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.

Identifying related 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 Figure 5. In the first case, we started with a circuit with a state distribution of a ring of six states (Figure 5A). We show the top five circuits with the closest state distributions, based on the described distance function (Figure 5B). PCA (second row in Figure 5B) and UMAP (third row in Figure 5B) 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 STAR 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. Of interest, 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 (Figure 5D). This state distribution was previously described by the triangularity score defined earlier (Figure 2). With our current analysis, we successfully identified circuits with the most similar state distributions, without using the triangularity score (Figure 5E). 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 (Figure 5B). 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 Figure 5F, are the same motifs as identified from our earlier analysis using the triangularity score (Figure 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

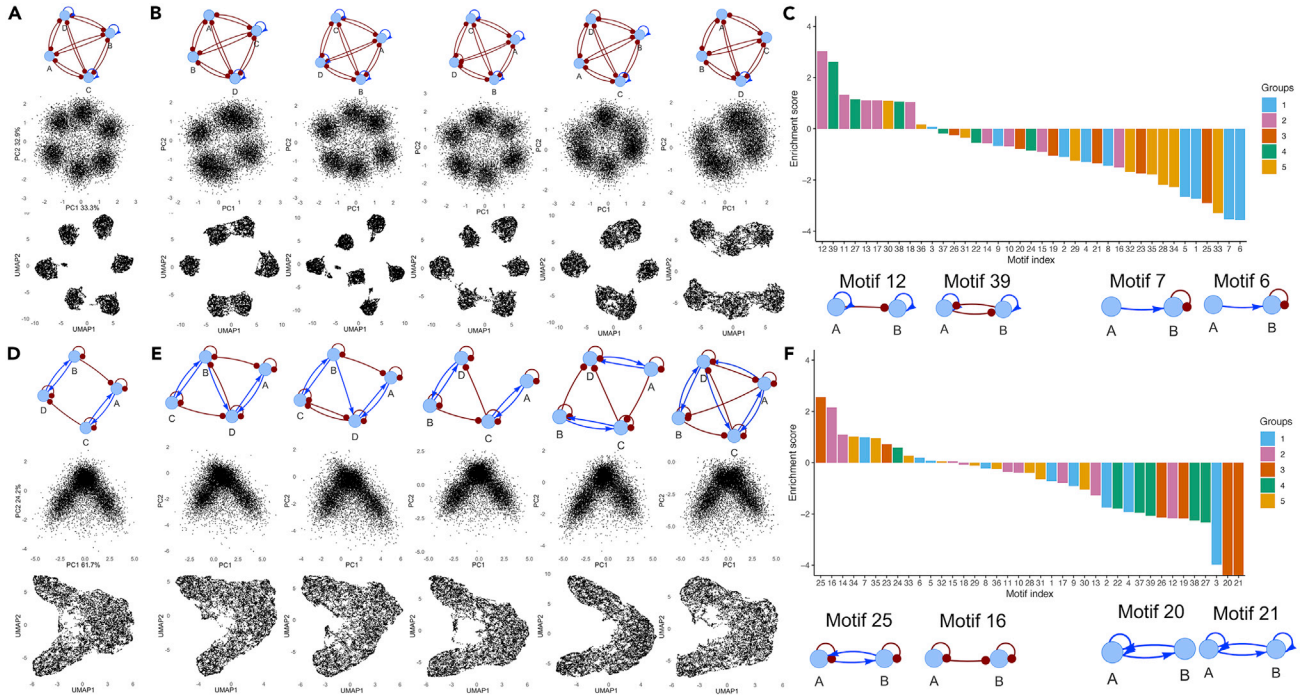


Figure 5. Identifying circuits with similar state distributions

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. The enrichment results for motif coupling are shown in Figures S6 and S7, respectively. 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 (see STAR Methods).

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.

Identifying small core regulatory circuits from single-cell gene expression

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 because of the potential modular structure and redundancy observed in large gene networks.[45,46] 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 STAR Methods).

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.[47] Figure 6A shows the PCA projection of the expression of 1448 genes to the first two principal components. Although 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.[47] From the circuit analysis we

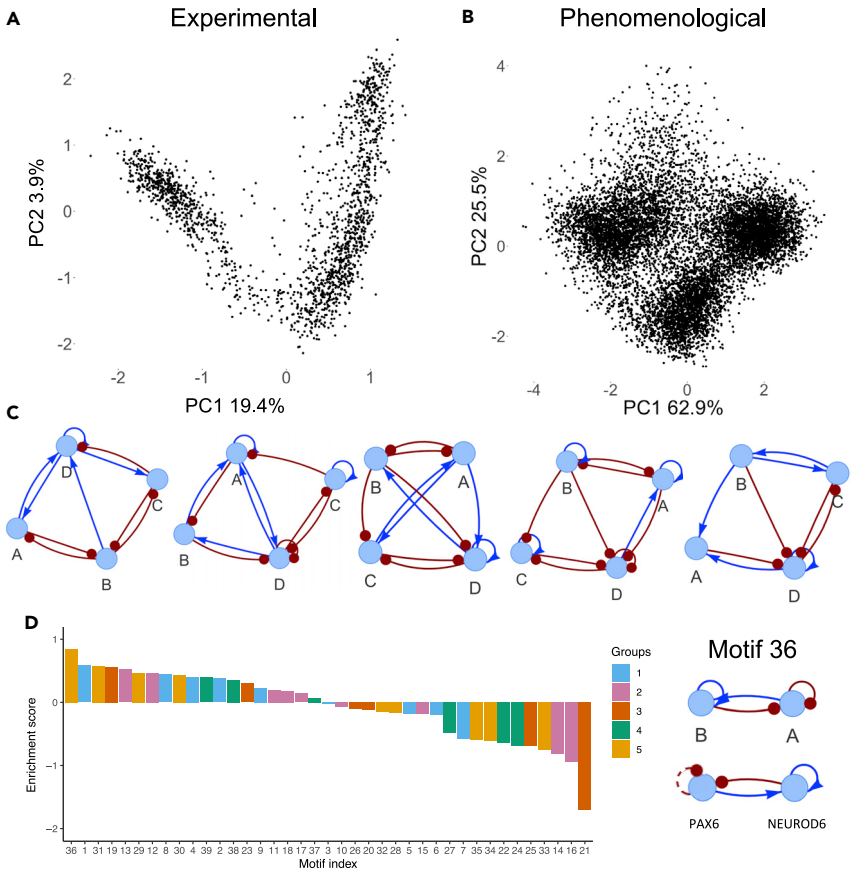


Figure 6. Application of the circuit analysis to scRNA-seq data of human glutamatergic neuron differentiation

(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. See Figure S45 for the scatterplot of gene pairs.

(C) The diagrams of the top five ranked circuits, with ranks shown below circuits. Additional top ranked circuits are shown in Figure S5.

(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. The outcomes of motif coupling analysis are shown in Figure S8. Colors in the column plots represent different groups of motif types (see STAR Methods).

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 Figure 6C and the state distribution of the topmost circuit in Figure 6B (see STAR Methods for the statistical evaluation and Figure S46 for the histogram of circuit distances). The top 20 four-node circuits are shown in Figure S5. 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. However, the clusters from the circuit simulations appear more spherical whereas 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 (Figure 6D). Of interest, 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 (Figure S8). 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[48] 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.[49] Furthermore, it has been shown that decreasing Pax6 expression is required to turn off the neural stem-cell self-renewal program.[50] 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.[51,52]

Remarkably, the top identified motif (#36) is consistent with the regulatory interactions responsible for neuronal cell differentiation.[48–52] In this circuit motif (bottom right circuit diagram in Figure 6D), one node (network involving *PAX6*) decreases its own expression and activates another node (network involving *NEUROD6*), which activates itself and represses the other node. Of interest, 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.

DISCUSSION

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 over-represented 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. Although 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[6,14] also utilized this motif identification strategy, however the current study provides a more quantitative and generalized methodology. Although 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.[27,28,30,31,53] 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.[54,55] We have shown its application to study glutamatergic neuron differentiation. We expect that this approach is invaluable to elucidate the regulatory mechanism for systems with more complex structures of cellular states.

Limitations of the study

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., feed forward loops with AND or OR logics.[56] 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,[57] 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. Third, our current analysis is illustrated with 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. 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.[58,59] A potential future direction would be the functional motif analysis for circuits more than the transcriptional regulation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Generation of all four-node gene circuits
  - Simulation of gene circuits by RACIPE
  - Metrics to quantify circuits with a triangular or linear structure of three states
  - Circuit motif enrichment
  - Grouping scheme of two-node circuit motifs
  - Classification of topological configurations for four-node circuits
  - Quantifying the differences between two gene expression distributions
  - Identifying circuits with similar state distributions
  - Associating gene circuits with single cell gene expression data
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical tests for circuit motif enrichment
  - Statistical tests for selecting top-ranked circuits associated with a gene expression state distribution

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106029.

ACKNOWLEDGMENTS

The study is supported by startup funds from The Jackson Laboratory and Northeastern University, and by the National Institute of General Medical Sciences of the National Institutes of Health under Award

Number R35GM128717. Special thanks Christopher Baker, Gregory Carter, and Amy Yee, whom together form the Thesis Advisory committee of Benjamin Clauss. We also thank Danial A. Ramirez for reproducing the pre-processing of the scRNA-seq data of human glutamatergic neuron differentiation, as described in reference (La Manno et al. 2018).

AUTHOR CONTRIBUTIONS

Conceptualization, M.L. and B.C.; Methodology, B.C. and M.L.; Software, B.C. and M.L.; Investigation, B.C.; Formal Analysis, B.C.; Writing – Original Draft, B.C. and M.L.; Writing – Review and Editing, B.C. and M.L.; Funding Acquisition, M.L; Supervision, M.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper received support from a program designed to increase minority representation in their field of research.

Received: August 11, 2022
Revised: December 19, 2022
Accepted: January 17, 2023
Published: January 23, 2023

REFERENCES

1. Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. Annu. Rev. Cell Dev. Biol. *26*, 721–744. https://doi.org/10.1146/annurev-cellbio-100109-104122.

2. Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet. *31*, 64–68. https://doi.org/10.1038/ng881.

3. Alon, U. (2007). Network motifs: theory and experimental approaches. Nat. Rev. Genet. *8*, 450–461. https://doi.org/10.1038/nrg2102.

4. Healy, C.P., and Deans, T.L. (2019). Genetic circuits to engineer tissues with alternative functions. J. Biol. Eng. *13*, 39. https://doi.org/10.1186/s13036-019-0170-7.

5. Jiménez, A., Cotterell, J., Munteanu, A., and Sharpe, J. (2017). A spectrum of modularity in multi-functional gene circuits. Mol. Syst. Biol. *13*, 925. https://doi.org/10.15252/msb.20167347.

6. Ye, Y., Kang, X., Bailey, J., Li, C., and Hong, T. (2019). An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. PLoS Comput. Biol. *15*, e1006855. https://doi.org/10.1371/journal.pcbi.1006855.

7. Gorochowski, T.E., Espah Borujeni, A., Park, Y., Nielsen, A.A., Zhang, J., Der, B.S., Gordon, D.B., and Voigt, C.A. (2017). Genetic circuit characterization and debugging using RNA -seq. Mol. Syst. Biol. *13*, 952. https://doi.org/10.15252/msb.20167461.

8. Becskei, A., and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. Nature *405*, 590–593. https://doi.org/10.1038/35014651.

9. Gardner, T.S., and Collins, J.J. (2000). Neutralizing noise in gene networks. Nature *405*, 520–521. https://doi.org/10.1038/35014708.

10. Huang, B., Lu, M., Jia, D., Ben-Jacob, E., Levine, H., and Onuchic, J.N. (2017). Interrogating the topological robustness of gene regulatory circuits by randomization. PLoS Comput. Biol. *13*, e1005456. https://doi.org/10.1371/journal.pcbi.1005456.

11. Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature *403*, 339–342. https://doi.org/10.1038/35002131.

12. Hong, J., Brandt, N., Abdul-Rahman, F., Yang, A., Hughes, T., and Gresham, D. (2018). An incoherent feedforward loop facilitates adaptive tuning of gene expression. Elife *7*, e32323. https://doi.org/10.7554/eLife.32323.

13. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. Science *298*, 824–827. https://doi.org/10.1126/science.298.5594.824.

14. Schaerli, Y., Munteanu, A., Gili, M., Cotterell, J., Sharpe, J., and Isalan, M. (2014). A unified design space of synthetic stripe-forming networks. Nat. Commun. *5*, 4905. https://doi.org/10.1038/ncomms5905.

15. Nordick, B., and Hong, T. (2021). Identification, visualization, statistical analysis and mathematical modeling of high-feedback loops in gene regulatory networks. BMC Bioinf. *22*, 481. https://doi.org/10.1186/s12859-021-04405-z.

16. Panovska-Griffiths, J., Page, K.M., and Briscoe, J. (2013). A gene regulatory motif that generates oscillatory or multiway switch outputs. J. R. Soc. Interface *10*, 20120826. https://doi.org/10.1098/rsif.2012.0826.

17. Nordick, B., Yu, P.Y., Liao, G., and Hong, T. (2022). Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression. Nucleic Acids Res. *50*, 3693–3708. https://doi.org/10.1093/nar/gkac217.

18. van Dorp, M., Lannoo, B., and Carlon, E. (2013). Generation of oscillating gene regulatory network motifs. Phys. Rev. E - Stat. Nonlinear Soft Matter Phys. *88*, 012722. https://doi.org/10.1103/PhysRevE.88.012722.

19. Thomas, P., Popović, N., and Grima, R. (2014). Phenotypic switching in gene regulatory networks. Proc. Natl. Acad. Sci. USA *111*, 6994–6999. https://doi.org/10.1073/pnas.1400049111.

20. Hortsch, S.K., and Kremling, A. (2018). Characterization of noise in multistable genetic circuits reveals ways to modulate heterogeneity. PLoS One *13*, e0194779. https://doi.org/10.1371/journal.pone.0194779.

21. Hari, K., Harlapur, P., Gopalan, A., Ullanat, V., Duddu, A.S., and Jolly, M.K. (2021). Emergent properties of coupled bistable switches. Syst. Biol. https://doi.org/10.1101/2021.06.15.448553.

22. Jolly, M.K., Jia, D., Boareto, M., Mani, S.A., Pienta, K.J., Ben-Jacob, E., and Levine, H. (2015). Coupling the modules of EMT and stemness: a tunable 'stemness window'

model. Oncotarget *6*, 25161–25174. https://doi.org/10.18632/oncotarget.4629.

23. Adler, M., and Medzhitov, R. (2022). Emergence of dynamic properties in network hypermotifs. Proc. Natl. Acad. Sci. USA *119*. e2204967119. https://doi.org/10.1073/pnas.2204967119.

24. Adler, M., Szekely, P., Mayo, A., and Alon, U. (2017). Optimal regulatory circuit topologies for fold-change detection. Cell Syst. *4*, 171–181.e8. https://doi.org/10.1016/j.cels.2016.12.009.

25. Kohar, V., and Lu, M. (2018). Role of noise and parametric variation in the dynamics of gene regulatory circuits. NPJ Syst. Biol. Appl. *4*, 40. https://doi.org/10.1038/s41540-018-0076-x.

26. Sabuwala, B., Hari, K., Abhishek, S.V., and Jolly, M.K. (2022). Coupled Mutual Inhibition and Mutual Activation Motifs as Tools for Cell-Fate Control. Syst. Biol. https://doi.org/10.1101/2022.05.27.493756.

27. Huang, B., Lu, M., Galbraith, M., Levine, H., Onuchic, J.N., and Jia, D. (2020). Decoding the mechanisms underlying cell-fate decision-making during stem cell differentiation by random circuit perturbation. J. R. Soc. Interface *17*, 20200500. https://doi.org/10.1098/rsif.2020.0500.

28. Katebi, A., Kohar, V., and Lu, M. (2020). Random parametric perturbations of gene regulatory circuit uncover state transitions in cell cycle. iScience *23*, 101150. https://doi.org/10.1016/j.isci.2020.101150.

29. Huang, B., Jia, D., Feng, J., Levine, H., Onuchic, J.N., and Lu, M. (2018). RACIPE: a computational tool for modeling gene regulatory circuits using randomization. BMC Syst. Biol. *12*, 74. https://doi.org/10.1186/s12918-018-0594-6.

30. Katebi, A., Ramirez, D., and Lu, M. (2021). Computational systems-biology approaches for modeling gene networks driving epithelial–mesenchymal transitions. Comput. Syst. Oncol. *1*, e1021. https://doi.org/10.1002/cso2.1021.

31. Ramirez, D., Kohar, V., and Lu, M. (2020). Toward modeling context-specific EMT regulatory networks using temporal single cell RNA-seq data. Front. Mol. Biosci. *7*, 54. https://doi.org/10.3389/fmolb.2020.00054.

32. Watcham, S., Kucinski, I., and Gottgens, B. (2019). New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. Blood *133*, 1415–1426. https://doi.org/10.1182/blood-2018-08-835355.

33. Pellin, D., Loperfido, M., Baricordi, C., Wolock, S.L., Montepeloso, A., Weinberg, O.K., Biffi, A., Klein, A.M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat. Commun. *10*, 2395. https://doi.org/10.1038/s41467-019-10291-0.

34. Xu, J., Lamouille, S., and Derynck, R. (2009). TGF-β-induced epithelial to mesenchymal transition. Cell Res. *19*, 156–172. https://doi.org/10.1038/cr.2009.5.

35. Lu, M., Jolly, M.K., Levine, H., Onuchic, J.N., and Ben-Jacob, E. (2013). MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. Proc. Natl. Acad. Sci. USA *110*, 18144–18149. https://doi.org/10.1073/pnas.1318192110.

36. Zhang, J., Tian, X.-J., Zhang, H., Teng, Y., Li, R., Bai, F., Elankumaran, S., and Xing, J. (2014). TGF-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. Sci. Signal. *7*, ra91. https://doi.org/10.1126/scisignal.2005304.

37. Lu, M., Jolly, M.K., Gomoto, R., Huang, B., Onuchic, J., and Ben-Jacob, E. (2013). Tristability in cancer-associated MicroRNA-TF chimera toggle switch. J. Phys. Chem. B *117*, 13164–13174. https://doi.org/10.1021/jp403156m.

38. Jia, D., Jolly, M.K., Harrison, W., Boareto, M., Ben-Jacob, E., and Levine, H. (2017). Operating principles of tristable circuits regulating cellular differentiation. Phys. Biol. *14*, 035007. https://doi.org/10.1088/1478-3975/aa6f90.

39. Som, A., Harder, C., Greber, B., Siatkowski, M., Paudel, Y., Warsow, G., Cap, C., Schöler, H., and Fuellen, G. (2010). The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. PLoS One *5*, e15165. https://doi.org/10.1371/journal.pone.0015165.

40. Yang, J., Gao, C., Chai, L., and Ma, Y. (2010). A novel SALL4/OCT4 transcriptional feedback network for pluripotency of embryonic stem cells. PLoS One *5*, e10766. https://doi.org/10.1371/journal.pone.0010766.

41. Han, H., Shim, H., Shin, D., Shim, J.E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., et al. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. Sci. Rep. *5*, 11432. https://doi.org/10.1038/srep11432.

42. Tatetsu, H., Kong, N.R., Chong, G., Amabile, G., Tenen, D.G., and Chai, L. (2016). SALL4, the missing link between stem cells, development and cancer. Gene *584*, 111–119. https://doi.org/10.1016/j.gene.2016.02.019.

43. Shi, G., and Jin, Y. (2010). Role of Oct4 in maintaining and regaining stem cell pluripotency. Stem Cell Res. Ther. *1*, 39. https://doi.org/10.1186/scrt39.

44. Moreira, S., Polena, E., Gordon, V., Abdulla, S., Mahendram, S., Cao, J., Blais, A., Wood, G.A., Dvorkin-Gheva, A., and Doble, B.W. (2017). A single TCF transcription factor, regardless of its activation capacity, is sufficient for effective trilineage differentiation of ESCs. Cell Rep. *20*, 2424–2438. https://doi.org/10.1016/j.celrep.2017.08.043.

45. Kafri, R., Levy, M., and Pilpel, Y. (2006). The regulatory utilization of genetic redundancy through responsive backup circuits. Proc. Natl. Acad. Sci. USA *103*, 11653–11658. https://doi.org/10.1073/pnas.0604883103.

46. Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy. Nature *388*, 167–171. https://doi.org/10.1038/40618.

47. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.

48. Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular identity of human outer radial glia during cortical development. Cell *163*, 55–67. https://doi.org/10.1016/j.cell.2015.09.004.

49. Thakurela, S., Tiwari, N., Schick, S., Garding, A., Ivanek, R., Berninger, B., and Tiwari, V.K. (2016). Mapping gene regulatory circuitry of Pax6 during neurogenesis. Cell Discov. *2*, 15045. https://doi.org/10.1038/celldisc.2015.45.

50. Sansom, S.N., Griffiths, D.S., Faedo, A., Kleinjan, D.-J., Ruan, Y., Smith, J., van Heyningen, V., Rubenstein, J.L., and Livesey, F.J. (2009). 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. https://doi.org/10.1371/journal.pgen.1000511.

51. Uittenbogaard, M., and Chiaramello, A. (2002). 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. https://doi.org/10.1002/jnr.10119.

52. Uittenbogaard, M., Baxter, K.K., and Chiaramello, A. (2010). NeuroD6 genomic signature bridging neuronal differentiation to survival via the molecular chaperone network. J. Neurosci. Res. *88*, 33–54. https://doi.org/10.1002/jnr.22182.

53. Su, K., Katebi, A., Kohar, V., Clauss, B., Gordin, D., Qin, Z.S., Karuturi, R.K.M., Li, S., and Lu, M. (2022). NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. Syst. Biol. https://doi.org/10.1101/2022.05.06.487898.

54. Hari, K., Ullanat, V., Balasubramanian, A., Gopalan, A., and Jolly, M.K. (2021). Landscape of Epithelial Mesenchymal Plasticity as an emergent property of coordinated teams in regulatory networks. Syst. Biol. https://doi.org/10.1101/2021.12.12.472090.

55. Campbell, C., Shea, K., Yang, S., and Albert, R. (2016). Motif profile dynamics and transient species in a Boolean model of mutualistic ecological communities. J. Complex Netw. *4*,

127–139. https://doi.org/10.1093/comnet/cnv008.

56. Alon, U. (2020). An Introduction to Systems Biology : Design Principles of Biological Circuits, Second Edition (CRC Press).

57. Bennett, M.R., Volfson, D., Tsimring, L., and Hasty, J. (2007). Transient dynamics of genetic regulatory networks. Biophys. J. *92*, 3501–3512. https://doi.org/10.1529/biophysj.106.095638.

58. Markevich, N.I., Hoek, J.B., and Kholodenko, B.N. (2004). Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. J. Cell Biol. *164*, 353–359. https://doi.org/10.1083/jcb.200308060.

59. Li, C.J., Liau, E.S., Lee, Y.H., Huang, Y.Z., Liu, Z., Willems, A., Garside, V., McGlinn, E., Chen, J.A., and Hong, T. (2021). MicroRNA governs bistable cell differentiation and lineage segregation via a noncanonical feedback. Mol. Syst. Biol. *17*, e9945. https://doi.org/10.15252/msb.20209945.

60. Csardi, G., and Nepusz, T. (2006). The Igraph Software Package for Complex Network Research (InterJournal), p. 1695.

61. Kolmogorov–Smirnov Test (2008). In The Concise Encyclopedia of Statistics (New York: Springer), pp. 283–287. https://doi.org/10.1007/978-0-387-32833-1_214.

62. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

63. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Motif4node (R package)	This paper	https://github.com/lusystemsbio/motif4node
Analysis R code from this study	This paper	https://doi.org/10.5281/zenodo.7534936
sRACIPE (R package)	(Kohar and Lu, 2018)	https://www.bioconductor.org/packages/release/bioc/html/sRACIPE.html
igraph (R package)	Csardi G, Nepusz T (2006)	https://igraph.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Mingyang Lu (m.lu@northeastern.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- R code for the RACIPE simulations, state distribution scoring, the construction of all two-node circuit motifs, enrichment analysis of non-redundant four-node circuits, and data files for the topologies of all non-redundant four-node gene circuits have been deposited at Zenodo and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- The circuit motif analysis tools is also available as an R package *motif4node* at GitHub with details in key resources table.
- Any additional information required to replicate the analysis reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Generation of all four-node gene 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 analyze 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.

Simulation of gene circuits by RACIPE

To explore the dynamical behavior of a gene circuit, we performed *r*andom*c*ircuit *p*erturbation (RACIPE) simulations using the sRACIPE R package.[25] RACIPE takes a network as an input and generates an ensemble of ordinary differential equation (ODE) based models, each with a unique set of randomized kinetic parameter values. RACIPE randomly samples from a wide range of parameter values, allowing for the robust exploration of circuit behavior, contrary to the conventional approaches that only explore the behavior of circuits for narrow parameter ranges.

To simulate all gene circuits in this study, we used the sRACIPE R implementation of RACIPE.[25] RACIPE takes a topology as an input and generates a system of ODEs that describe the rate of change of any gene j that is transcriptionally regulated by one or more regulators i. The equation is as follows:

$$\frac{dx_j}{dt} = \frac{G_j}{\prod_i \lambda_{ij}^+} \prod_i H^S\left(x_i, x_{ij0}, n_{ij}, \lambda_{ij}\right) - k_j x_j, \qquad \text{(Equation 1)}$$

where G_j is the expression levels of gene j, x_i or x_j is the 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:

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

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 λ_{ij}^+ for an excitatory interaction and takes a value larger than 1. In this case, H^S ranges from (1, λ_{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 λ_{ij}^+ from (1, 100), for each parameter. For an inhibitory interaction, λ_{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.[10] 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 $\frac{G_j}{k_j}\left(\frac{\prod_i \lambda_{ij}{}^-}{\prod_i \lambda_{ij}^+}\right)$.

In this study, as we focused on characterizing state distributions systematically for a very large number of gene circuits, we chose to sample one initial condition for each of the 10,000 models and output the numeric value of gene expression from the last snapshot. This allows us to finish simulations of all non-redundant four node gene circuits in two days using 60 CPU nodes (E5-2680v4@2.40GHz) on the discovery HPC at Northeastern University. As each model was simulated for 50 unit time, we expect that most of steady-state models should have converged (as the degradation rates were randomly sampled from 0.1 to 1 (unit time)$^{-1}$), and, for those who have not, they belong to fluctuating models.

Here, each circuit was simulated with sRACIPE for 10,000 models. The simulated data were processed with log transformation and standardization for further analysis.

Metrics to quantify circuits with a triangular or linear structure of three states

With the simulation data, we aimed to characterize the gene expression state distribution allowed by each gene circuit topology. We first performed k-means clustering (k = 3) to the simulated gene expression data, and defined a score for triangular structure, Q_1, as:

$$Q_1 = \min_{i,j} \frac{D_{ij}}{S_i S_j}, \quad \text{(Equation 3)}$$

where D_{ij} is the Euclidean distance between the centers of clusters i and, 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 overall 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 (see Figure 2A for a geometric illustration of Q_1). 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:

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

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 (see Figure 2C for a geometric illustration of Q_2). 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.

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 (the diagrams of motifs and their indices are listed in Figure S1) 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

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

where a = 1 *or* 2 for the two scores, $H^-(x, x_0, n) := 1/\left(1 + \left(\frac{x}{x_0}\right)^n\right)$ 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 overall 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 overall 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.

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. 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 (Figures 3, 5, and 6) and the heatmaps for two-motif enrichment analysis (Figures 3, S4, and S6–S8).

Classification of topological configurations for four-node circuits

We classified all 4-node topologies as belonging to one of five classes (ship, chain, box, cross, double-cross), as shown in Figure S39. We started by turning each adjacency matrix into an undirected graph using the graph_from_adjacency_matrix() and as.undirected() functions from igraph.[60] The circuit classification was then performed on the basis of the total number of edges and the number interactions for each node (i.e., box has four total edges, where each node has two interactions; ship has four total edges, where the nodes have one, two, or three interactions). With these classifications we evaluated the patterns of coupling between two two-node circuit motifs in the top 600 circuits capable of producing a triangular state distribution and the top 600 circuits capable of producing a linear state distribution. As shown in the left panel of Figure S40 for the triangular distribution, the line and ship configurations are enriched, while the double cross configuration is under-enriched. As shown in the right panel of Figure S40 for the linear distribution, while no specific configuration is enriched, the line configuration is under-enriched. Our results indicate that specific topological configurations of circuits play important roles in determining the circuit's state distribution.

Quantifying the differences between two gene expression distributions

To quantify the differences between the 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:

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

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[61] 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.

Identifying circuits with similar state distributions

Using the above-defined distance function, we constructed a matrix of pairwise distances for all the 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. A p value for the fit of two circuits was calculated via the *Z* score of distances. 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[60]) 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).

Associating gene circuits with single cell gene expression data

The scRNA-seq data of human glutamatergic neuron differentiation was processed using the velocyto pipeline, as described in the original paper.[47] 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 non-redundant 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests for circuit motif enrichment

The significance of the enrichment was determined by a permutation test, similar to some previous approaches.[62] A null distribution for each enrichment was created by shuffling the ranking indices of circuits for each score and applying the enrichment test. Thisstep 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[63] for multiple hypothesis testing. The number of hypotheses is the total number of two node motifs 39. Statistical details can be found in the captions of Figures 3, 5, and 6.

Statistical tests for selecting top-ranked circuits associated with a gene expression state distribution

We estimated a p value for the fit of the top ranked circuits to either an experimental gene expression state distribution, or to another simulated gene expression state distribution by calculating the Z score (using base R) of the distance. When comparing simulated gene expression state distributions, the Z score was calculated from a distance distribution generated by comparing a reference simulation to all other simulation results (see STAR Methods section "identifying circuits with similar state distributions" for details on the distance metric). When comparing to an experimental gene expression state distribution, the Z score was calculated from a distance distribution generated by comparing an experimental gene expression state distribution to all other simulation results (see STAR Methods section "associating gene circuits with single cell gene expression data" for details on the distance metric). Here, we regarded the distribution of the KS distances of all four-node circuits as the null distribution, which was found to be approximately Gaussian (Figure S46). A p value ≤ 0.05 is reached when the Z score ≤ -1.64. Statistical details can be found in the caption of Figure 6.