

**LOCAL CONTROLLABILITY OF
BIOLOGICAL NETWORKS**

LUO CHANG

(B.Sc., NKU, China)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
2015**

Declaration

I thereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Luo Chang

Luo Chang

31 March 2015

Acknowledgements

I would like to express my sincere thanks to my parents, for their continuous support for my PhD career. They have been very patient with me, encouraged me all the way of the journey of my pursuing the PhD degree.

My most sincere gratitude goes to my supervisor, Associate Professor Zhang Louxin, for his consistent guidance in the research project. His consultation helped me a lot when I need help in research. His research mentality and hardworking spirit has inspired me throughout all these years and will be an invaluable part of my future career.

I would also like to thank all the members of our research lab: Associate Professor Choi Kwok Pui, Zheng Yu, Tian Dechao and David Chew. I have benefited a lot from discussion with them. Our weekly meetings have broadened my knowledge.

This list is by no means complete. I also give thanks to Franco Lim for helping me with some proofreading of the thesis, and Zhou Feng for discussion of the thesis writing style.

Contents

| | |
|--|------------|
| Declaration | iii |
| Acknowledgements | v |
| Summary | xi |
| 1 Introduction | 1 |
| 1.1 Networks and linear systems | 1 |
| 1.2 Contributions of the thesis | 3 |
| 1.3 Organization of the thesis | 4 |
| 2 Biological networks and their topological properties | 5 |
| 2.1 Basic concepts and notations | 6 |
| 2.2 Biological networks | 8 |
| 2.2.1 Brain networks | 9 |
| 2.2.2 Structures and functions of mouse brain subdivisions | 11 |
| 2.2.3 Transcriptional regulatory networks | 14 |
| 2.3 Network metrics | 17 |

| | | |
|----------|---|-----------|
| 2.3.1 | Clustering coefficient and transitivity | 18 |
| 2.3.2 | Betweenness | 19 |
| 2.3.3 | Closeness | 20 |
| 2.3.4 | Eigenvector centrality | 21 |
| 2.3.5 | Katz centrality | 22 |
| 2.3.6 | PageRank centrality | 22 |
| 2.4 | Network structural characteristics | 23 |
| 2.4.1 | Small-world property | 23 |
| 2.4.2 | Network motif | 25 |
| 2.4.3 | Degree correlation | 27 |
| 2.4.4 | Community structure | 30 |
| 2.5 | Topological properties of biological networks | 31 |
| 2.5.1 | Topological properties of brain networks | 31 |
| 2.5.2 | Topological properties of transcriptional regulatory networks | 34 |
| 3 | Linear systems, controllability and structural controllability | 39 |
| 3.1 | Dynamical system | 40 |
| 3.2 | Linear time-varying system | 42 |
| 3.2.1 | Structure of the solution | 42 |
| 3.2.2 | Controllability | 44 |
| 3.3 | Linear time-invariant system | 46 |
| 3.3.1 | Structure of the solution | 47 |
| 3.3.2 | Controllability | 48 |
| 3.4 | Structural controllability | 49 |
| 3.4.1 | Lin's theorem | 49 |
| 3.4.2 | Algebraic characterization | 64 |

| | | |
|----------|--|------------|
| 4 | Controllability of biological networks | 67 |
| 4.1 | Definition of network controllability | 68 |
| 4.2 | A theorem on network controllability | 70 |
| 4.3 | Controllability of brain networks | 75 |
| 4.4 | Controllability of transcriptional regulatory networks | 79 |
| 5 | Local controllability of biological networks | 81 |
| 5.1 | Definition of local controllability | 82 |
| 5.2 | Properties and bounds for local controllability | 83 |
| 5.3 | A formula for local controllability | 86 |
| 5.4 | The algorithm for local controllability | 92 |
| 5.4.1 | The Hungarian method | 92 |
| 5.4.2 | The algorithm for local controllability | 96 |
| 5.5 | Local controllability of brain networks | 97 |
| 5.6 | Local controllability of transcriptional regulatory networks | 107 |
| 6 | Local controllability of model networks | 121 |
| 6.1 | Model networks | 121 |
| 6.1.1 | Erdős–Rényi random network | 122 |
| 6.1.2 | Scale-free network | 124 |
| 6.2 | Local controllability v.s. degree | 128 |
| 6.2.1 | ER random networks | 129 |
| 6.2.2 | Scale-free networks | 134 |
| 6.3 | Local controllability v.s. betweenness | 138 |
| 6.4 | Local controllability v.s. link/node removal | 143 |
| 6.5 | Remarks | 148 |
| 7 | Future work | 151 |

| | |
|--|-----|
| Bibliography | 153 |
| A Local controllability of protein complexes | 165 |
| B The p -values for protein complexes | 191 |

Summary

This thesis focuses on local controllability of biological networks. Our main contribution is that we have proposed a network control framework, called **local controllability**, which is applicable to any directed network. Local controllability is an extension of the previous network control framework, called **network controllability** [56]. Local controllability concerns about the minimum number of inputs required to control a subset of nodes in a directed network.

For any directed network G and any nonempty subset S of nodes in G , we show that the local controllability $lc(G, S)$ can be calculated by an algorithm in time $O(n^3)$. The algorithm has been implemented in Matlab for our numerical studies. We then applied local controllability to two types of biological networks: brain networks and transcriptional regulatory networks.

We also performed numerical experiments to study local controllability of model networks. We studied how the local controllability $lc(G, S)$ depends on the choice of S (when G is fixed), and on the choice of G (when S is fixed), in ER random networks and scale-free networks. Specifically, we examined the relationships between local controllability and various centrality measures, and also investigated the robustness of local controllability against link and node removal.

List of Tables

| | | |
|-----|--|----|
| 2.1 | The network sizes of the human TF regulatory networks of 41 cell types. | 16 |
| 2.2 | Degree correlation based on degree correlation coefficient and degree correlation function. | 29 |
| 2.3 | Some network metrics of the mouse inter-region brain networks in both ipsilateral and contralateral hemispheres, corresponding to P-value cutoffs 0.05, 0.02, 0.01, 0.005 and 0.001. | 32 |
| 2.4 | The top 10 nodes with the highest centrality measures in G_r | 33 |
| 2.5 | Some network metrics of the TF regulatory network of hESC. | 36 |
| 2.6 | The top 10 nodes with the highest centrality measures in the TF regulatory network of hESC. | 36 |
| 4.1 | The network controllability of real-world networks [56]. For each network, the network type and name is given, as well as the number of nodes (N), the number of edges (L), the ratio of the minimum number of inputs to the number of nodes in the network (n_D). | 74 |

| | | |
|-----|---|-----|
| 4.2 | The network controllability of the mouse inter-region brain networks (indicated by numbers in red) and the network controllability of the subnetworks of each mouse inter-region brain network induced by 12 subdivisions. The parentheses after each subdivision indicate the number of regions belonging to that subdivision. | 77 |
| 4.3 | The network controllability of the human TF regulatory networks of 41 cell types. For each network, the cell type is given, as well as the number of nodes (N), the number of edges (L), the minimum number of inputs (N_D) and the fraction of inputs (n_D). | 80 |
| 5.1 | The local controllability of 12 subdivisions in the mouse inter-region brain networks in both ipsilateral and contralateral hemispheres, corresponding to P-value cutoffs 0.05, 0.02, 0.01, 0.005 and 0.001. The parentheses after each subdivision indicate the number of regions belonging to that subdivision. | 99 |
| 5.2 | The mean and standard deviation of the local controllability of random subsets with a given size in the mouse inter-region brain networks. The parentheses below each mean indicate the corresponding standard deviation. | 103 |
| 5.3 | The p -value for the alternative hypothesis that the local controllability of a subdivision is smaller than the local controllability of random subsets with the same size as the subdivision, in the mouse inter-region brain networks. The significant p -values ($p_1 < 0.05$) are marked in bold. | 106 |
| 5.4 | The p -value for the alternative hypothesis that the local controllability of a subdivision is larger than the local controllability of random subsets with the same size as the subdivision, in the mouse inter-region brain networks. The significant p -values ($p_2 < 0.05$) are marked in bold. | 106 |

-
- 5.5 The mean of the local controllability of protein complexes with a given size between 5 and 19 in the human TF regulatory networks of 8 cell types. The parentheses after each value of the complex size indicate the total number of protein complexes with size equal to that value. AE stands for Amniotic Epi., HSC for Hemat. Stem Cell, ADB for Adult Dermal Blood, Neu for Neuroblastoma, SM for Skeletal Myoblast, FB for Fetal Brain and SF for Skin Fib. 112
- 5.6 The mean and standard deviation of the local controllability of random subsets with a given size between 5 and 19 in the human TF regulatory networks of 8 cell types. The parentheses below each mean indicate the corresponding standard deviation. AE stands for Amniotic Epi., HSC for Hemat. Stem Cell, ADB for Adult Dermal Blood, Neu for Neuroblastoma, SM for Skeletal Myoblast, FB for Fetal Brain and SF for Skin Fib. 113
- 5.7 The mean of the local controllability of protein complexes with a given size between 5 and 19 in the TF regulatory network of hESC, calculated by ignoring ‘ZNF354C’. The parentheses after each value of the complex size indicate the total number of protein complexes with size equal to that value. 114
- 5.8 The p -value for the alternative hypothesis that the median of the local controllability of protein complexes is smaller than the median of the local controllability of random subsets with the same size, in the human TF regulatory networks of 8 cell types. We used the Wilcoxon rank sum (one-tailed) test to calculate the p -value. 116
- 5.9 The p -value for the alternative hypothesis that the median of the local controllability of protein complexes is larger than the median of the local controllability of random subsets with the same size, in the human TF regulatory networks of 8 cell types. We used the Wilcoxon rank sum (one-tailed) test to calculate the p -value. 116

| | | |
|-----|--|-----|
| 6.1 | Some properties of ER random networks. | 124 |
| 6.2 | Some properties of BA networks. | 128 |
| A.1 | The local controllability of 712 protein complexes in the TF regulatory networks of 8 cell types. For each protein complex, a numbered index (No.), the complex identifier (ID) and its size are also given. | 189 |
| B.1 | The p -values p_1 and p_2 in the TF regulatory network of hESC. | 192 |
| B.2 | The p -values p_1 and p_2 in the TF regulatory network of Amniotic Epi. | 193 |
| B.3 | The p -values p_1 and p_2 in the TF regulatory network of Hemat. Stem Cell. | 194 |
| B.4 | The p -values p_1 and p_2 in the TF regulatory network of Adult Dermal Blood. | 195 |
| B.5 | The p -values p_1 and p_2 in the TF regulatory network of Neuroblastoma. | 196 |
| B.6 | The p -values p_1 and p_2 in the TF regulatory network of Skeletal Myoblast. | 197 |
| B.7 | The p -values p_1 and p_2 in the TF regulatory network of Fetal Brain. | 198 |
| B.8 | The p -values p_1 and p_2 in the TF regulatory network of Skin Fib. | 199 |

List of Figures

- 2.1 A visualization of the mouse inter-region brain network in the ipsi-lateral hemishpere, corresponding to P-value cutoff 0.05. Each sub-division is marked with a different color and is drawn in the network visualization based on its relative location in the brain. 10
- 2.2 A visualization of the TF regulatory network of hESC. The network has a three-layer hierarchical structure. The top layer is marked in red, the middle layer is marked in cyan, and the bottom layer is marked in green. 17
- 2.3 Some simple network motifs: (a) feed-forward loop, (b) feed-back loop, (c) bi-fan. 26
- 3.1 The graph of a single-input structured pair (A, b) 50
- 3.2 The graph structure of a cacti. The cacti contains two (node-disjoint) cactus. The left cactus contains one stem and three buds. The right cactus is just a stem. 52
- 3.3 The graph of a multi-input structured pair (A, B) 63
- 4.1 A control configuration for a directed network. 68

| | | |
|-----|---|-----|
| 4.2 | A maximum matching in a directed network. The set of arcs in red form a maximum matching in the directed network. | 70 |
| 4.3 | An example of constructing a spanning cacti based on a maximum matching in a directed network. In (a), the set of arcs in red form a maximum matching. In (b), the nodes in blue are the input nodes of the cacti. The nodes in green are the unmatched nodes with respect to the maximum matching. | 72 |
| 4.4 | A visualization of the subnetworks induced by 12 subdivisions in the ipsilateral network corresponding to P-value cutoff 0.05. | 78 |
| 5.1 | An example of controlling a subset S of nodes in a directed network G . In (b), the nodes in blue are the input nodes of the cacti. The subset S is the set of nodes in red. | 82 |
| 5.2 | An example of a strongly connected network G with the nodes in S marked in red. | 84 |
| 5.3 | The correspondence between a path-cycle cover of S in G , $C_G(S)$, and a cycle cover of S in G^* , $C_{G^*}(S)$. The subset S is the set of all the nodes in red. | 89 |
| 5.4 | An example to show that $lc(G, S)$ can be calculated by finding a minimum-weight perfect matching in G^* (marked by the red arcs). The subset S is the set of all the nodes in red. | 91 |
| 5.5 | An illustration of the Hungarian method based on cost matrix. | 95 |
| 5.6 | A visualization of the controllable subnetworks of 12 subdivisions in the ipsilateral network corresponding to P-value cutoff 0.05. | 101 |
| 5.7 | The frequency of protein complexes with a given size and with a given local controllability lc , in the TF regulatory networks of cell types hESC, Amniotic Epi., Hemat. Stem Cell and Adult Dermal Blood. | 109 |

-
- 5.8 The frequency of protein complexes with a given size and with a given local controllability lc , in the TF regulatory networks of cell types Neuroblastoma, Skeletal Myoblast, Fetal Brain and Skin Fib. . 110
- 6.1 Poisson degree distributions for ER random networks with different parameters $\lambda = \langle k \rangle$. The figure is from http://en.wikipedia.org/wiki/Poisson_distribution. 123
- 6.2 A power-law degree distribution in a log-log plot. The figure is from http://mathinsight.org/image/power_law_degree_distribution_scatter. 125
- 6.3 The local controllability of low-degree, medium-degree and high-degree subsets in ER random networks. The error bar indicates one standard deviation away from the mean. 129
- 6.4 The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 3$ 131
- 6.5 The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 4$ 131
- 6.6 The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 5$ 132
- 6.7 The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 6$ 132
- 6.8 The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 7$ 133

| | | |
|------|---|-----|
| 6.9 | The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 8$ | 133 |
| 6.10 | The plot of local controllability versus average degree of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 2.2, 2.5, 3$ and 4 | 135 |
| 6.11 | The plot of local controllability versus average degree of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 6$ and $\gamma = 2.2, 2.5, 3$ and 4 | 136 |
| 6.12 | The plot of local controllability versus average degree of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 8$ and $\gamma = 2.2, 2.5, 3$ and 4 | 137 |
| 6.13 | The plot of local controllability versus average betweenness of 10 subsets in ER random networks with varying parameters $\langle k \rangle = 3, \dots, 8$. For each parameter $\langle k \rangle$, 10 ER random networks with $n = 1000$ nodes were generated. | 138 |
| 6.14 | The plot of local controllability versus average betweenness of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 2.2, 2.5, 3$ and 4 | 139 |
| 6.15 | The plot of local controllability versus average betweenness of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 6$ and $\gamma = 2.2, 2.5, 3$ and 4 | 140 |
| 6.16 | The plot of local controllability versus average betweenness of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 8$ and $\gamma = 2.2, 2.5, 3$ and 4 | 141 |
| 6.17 | The plot of local controllability versus the number of outgoing/incoming links deleted for an ER random network G_0 with $n = 1000$ and $\langle k \rangle = 4$. The dashed line $N_D = 58$ indicates the network controllability of $G_0[S]$ | 144 |

-
- 6.18 The plot of local controllability versus the number of outgoing/incoming links deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 2.5$. $N_D = 16$ indicates the network controllability of $G_0[S]$. . 145
- 6.19 The plot of local controllability versus the number of outgoing/incoming links deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 4$. The dashed line $N_D = 39$ indicates the network controllability of $G_0[S]$ 145
- 6.20 The plot of local controllability versus the number of out-neighbors/in-neighbors deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 3$ and $\gamma = 2.5$. The dashed line $N_D = 23$ indicates the network controllability of $G_0[S]$ 147
- 6.21 The plot of local controllability versus the number of out-neighbors/in-neighbors deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 3$ and $\gamma = 4$. $N_D = 50$ indicates the network controllability of $G_0[S]$. . . 147
- 6.22 The plot of local controllability versus the number of out-neighbors/in-neighbors deleted for an ER network G_0 with $n = 1000$ and $\langle k \rangle = 3$. The dashed line $N_D = 61$ indicates the network controllability of $G_0[S]$. 148

Introduction

1.1 Networks and linear systems

Complex networks [10, 69], which are usually studied in network theory, have been commonly used to model diverse natural and artificial systems. These systems range from biological systems such as DNA transcriptional regulation in the cell nucleus and neural interaction in the brain, to social, economic systems, and to technological systems such as the World Wide Web and the Internet. Complex networks arising in these systems display substantial nontrivial topological features, with patterns of connection between their elements neither purely regular nor purely random.

With the ever-increasing amount of data arising from complex systems in various fields, complex networks have been shown to be an effective modelling technique for exploiting this complexity and studying the large-scale properties of the complex systems. In the last decade, the study of complex networks has gained extensive interest from scientists in diverse research disciplines: mathematics, physics, biology, computer science, sociology and others. As a result of multidisciplinary efforts, a large amount of literature that has studied various aspects of complex networks is now available. Among these studies, there are methods that were proposed to characterize the roles of individual nodes or to uncover meaningful small subnetworks

within a network [16, 29, 65]; development of network models to explain the existing structural patterns in real networks [8, 100]; both theoretical and numerical studies on the dynamics that take place on networks, including diffusion [77], evolution [75], percolation [14], synchronizability [89], and control [56, 67].

Linear systems have long been studied in linear system theory and control theory [44, 88]. An important property of linear systems is controllability, which is the ability to drive a system, starting from any initial state, towards any final state within finite time by imposing a suitable subset of external controllers. With additional assumptions, structural controllability was first defined by Lin [55]. Recently, *network controllability* was proposed by Liu et al. [56], and it has been pursued as a promising framework for understanding the control of complex networks as well as for designing efficient control schemes over complex networks.

Network controllability, which studies how to gain control over complex networks, is deeply rooted in both network theory and linear system theory. Two factors have complicated the understanding of controllability of complex systems: (1) the complex network architecture encapsulating interactions between system components, (2) the dynamical rule that governs the time-dependent dynamics of the system. Understanding how network structure affects our ability to control complex networks becomes important for designing optimal control schemes to tame the network dynamics.

Despite a thorough investigation of controllability on the scale of a whole network, it is still unclear whether controllability can be defined for any subset of nodes within a network (on a local scale). In certain situations, one desires to only control a subset of elements within a networked system. This is the motivation for introducing the concept of *local controllability* in the thesis. In transcriptional regulatory networks, there are times when only a specific subset of genes need to be controlled for their expressions, to perform desirable biological functions. For example, It was found [93] that the introduction of a subset of four transcription factors (Oct3/4, Sox2, c-Myc, and Klf4) is sufficient to convert mouse embryonic or adult fibroblasts

(mature cells) to pluripotent stem cells.

1.2 Contributions of the thesis

Local controllability is defined for any subset of nodes in a directed network, as a measure of the minimum number of external input controllers needed in order for nodes in this subset to be all controlled, regardless of whether the rest nodes being controlled or not. Apart from theoretical interest in its own right, the application of local controllability to real-world networks could help us to understand how to control a specific group of nodes in a networked system, and design efficient control strategies to satisfy our practical needs to control target nodes on a local scale.

In this thesis, we propose a network control framework, called ‘local controllability’. We have studied the local controllability of complex networks as a theoretical problem, and also studied its applications to biological networks. The main results include:

1. Local controllability of (directed) complex networks is introduced.
2. A cubic-time algorithm is developed to calculate the local controllability.
3. The local controllability of brain subdivisions is studied in the mouse inter-region brain networks.
4. The local controllability of protein complexes is studied in human transcription factor regulatory networks of 41 cell types.
5. Simulations are carried out to investigate how the local controllability depends on both the subset and the network under consideration.

1.3 Organization of the thesis

The thesis is organized as follows. In chapter 2, we discuss brain networks and transcriptional regulatory networks and study the topological properties of these two types of biological networks.

In Chapter 3, we introduce controllability of linear time-invariant systems, which is later generalized to the notion of structural controllability (of linear structured systems). We shall give detailed exposition of structural controllability and Lin's structural controllability theorem, which are fundamental to network controllability and local controllability.

In Chapter 4, we first formally define network controllability and then give a known theorem on network controllability. We then compute the network controllability of mouse inter-region brain networks and human cell-type-specific transcription factor regulatory networks.

In Chapter 5, we first formally define the local controllability of a subset of nodes in a directed network, and then present several theorems on local controllability. Based on these theorems, we present a polynomial-time algorithm for computing the local controllability. We then apply local controllability to biological networks. We study the local controllability of brain subdivisions in mouse inter-region brain networks and the local controllability of protein complexes in human cell-type-specific transcription factor regulatory networks.

In Chapter 6, we study how the local controllability $lc(G, S)$ depends on both the network structure G and the subset S in model networks. We first examine the relationships between local controllability and various centrality measures in ER random networks and scale-free networks, and then study how the removal of links and nodes can affect the local controllability of a subset of nodes in ER random networks and scale-free networks.

In Chapter 7, we provide possible directions for future research.

Biological networks and their topological properties

The study of complex networks has advanced rapidly over the last decade, largely following the availability of large-scale real network datasets across many disciplines [69]. The widespread use of networks to model the connectivity patterns in various systems further boosts interest for researchers in studying various network properties, ultimately driving the emergence and development of a new research field called *network science* (or *network theory*).

Network science mainly deals with characterizing, modelling, and analyzing the structures and dynamics of complex networks. The core objects that are studied in network science are thus networks, whether arising in real-world situations or synthesized by network models. Although networks can also be called graphs, it should be stressed that the more recent network science is very different from the traditional field of *graph theory* in mathematics, in terms of both the research scopes and the research objectives.

Despite the heterogeneity of real networks that have been studied, a discovery that pushed the explosion of network science is that the structure and evolution of different networks can be characterized by a common set of universal organizing

principles and reproducible mechanisms. Thus, a mathematical framework to study complex networks is possible. Network science is able to produce such a framework in studying complex networks by providing a large number of network metrics, structural characteristics, network models and network dynamics. In this chapter, we will only present some selective material from network science that can be usually used to study real networks. This chapter is organized as follows.

In Section 2.1, we give some fundamental terminologies about networks. In Section 2.2, we introduce two types of biological networks: brain networks and transcriptional regulatory networks. In Section 2.3, we review some network metrics that have been used to characterize individual nodes or global network structures. In Section 2.4, several structural properties that characterize real-world networks are discussed. In Section 2.5, we discuss the topological properties of brain networks and transcriptional regulatory networks.

2.1 Basic concepts and notations

A *network* is a collection of *nodes* and *edges* (links) that connect pairs of nodes. Mathematically, we can define a network G to be $G = (V(G), E(G))$, where $V(G)$ is the set of nodes and $E(G)$ is the set of edges in which each edge is represented by an unordered pair of nodes $\{u, v\}$. Then, for each edge $\{u, v\}$, $\{u, v\}$ is said to be *incident* to the end nodes u and v . For any two nodes u and v , they are *adjacent* if there is an edge $\{u, v\} \in E(G)$. A *path* between nodes u and v is a sequence of pairwise-disjoint nodes $\{w_0, w_1, \dots, w_k\}$ for some k , such that $u = w_0, v = w_k$, and $\{w_{i-1}, w_i\} \in E(G), i = 1, \dots, k$, and such a path is said to have length k . A network $H = (V(H), E(H))$ is a *subnetwork* of a network $G = (V(G), E(G))$, if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. In this chapter, the term ‘networks’ refers to undirected networks, unless otherwise mentioned.

If there is inherent directionality in the interactions between pairs of nodes, then networks are called directed networks. A *directed network* (or digraph) is a collection

of nodes and directed edges (or arcs) that link from one node to another node. More precisely, a directed network G is defined to be $G = (V(G), A(G))$ where $V(G)$ is the set of nodes and $A(G)$ is the set of arcs in which each arc is represented by an ordered pair of nodes (u, v) (to show the directionality, an arc (u, v) can also be denoted by $u \rightarrow v$). For a directed edge $u \rightarrow v$, u is called the *initial* node, and v is called the *terminal* node. A *directed path* from node u to node v is a sequence of pairwise-disjoint nodes $\{w_0 \rightarrow w_1 \rightarrow \dots \rightarrow w_k\}$ for some k , such that $u = w_0, v = w_k$, and $(w_{i-1} \rightarrow w_i) \in A(G), i = 1, \dots, k$, and such a directed path is said to have length k . A directed network $H = (V(H), A(H))$ is a *subnetwork* of a directed network $G = (V(G), A(G))$, if $V(H) \subseteq V(G)$ and $A(H) \subseteq A(G)$.

In a network, the *degree* of a node is the number of edges that are incident to it. In a directed network, three types of degrees can be defined. For a directed network $G = (V(G), A(G))$ and a node u , the *in-degree* $k_{in}(u)$ of u is the number of nodes v such that $(v, u) \in A(G)$; the *out-degree* $k_{out}(u)$ of u is the number of nodes v such that $(u, v) \in A(G)$; and the (total) degree $k(u)$ of u is defined to be the sum of in-degree and out-degree: $k(u) = k_{in}(u) + k_{out}(u)$.

For a directed (unweighted) network G with n nodes, its *adjacency matrix* is $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ where $a_{ij} = 1$ if there is a directed edge from node i to j and $a_{ij} = 0$ otherwise. For any node i , its in-degree and out-degree can be computed by

$$k_{in}(i) = \sum_j a_{ji},$$

$$k_{out}(i) = \sum_j a_{ij},$$

and the degree of node i is thus

$$k(i) = k_{in}(i) + k_{out}(i) = \sum_j (a_{ij} + a_{ji}).$$

For a network G , the *degree distribution* is defined to be the probability distribution $P(k), k = 0, 1, \dots$, in which $P(k)$ is equal to the fraction of nodes with degree k . Therefore, it satisfies that $\sum_{k \geq 0} P(k) = 1$. In a directed network, the in-degree, out-degree and degree distributions can be similarly defined.

In a directed network, the *distance* from a node u to a node v , denoted by $d(u, v)$, is the minimum length of a directed path from u to v . Similarly, the distance $d(u, v)$ between two nodes u and v in a network is the minimum length of a path between u and v . If there is no directed path from u to v in a directed network, or no path between u and v in a network, then define $d(u, v) = \infty$.

The *average path length* \bar{d} of a network (or directed network) is defined to be the average distance between any two nodes. That is,

$$\bar{d} = \frac{1}{n(n-1)} \sum_{u,v:u \neq v} d(u, v),$$

where n is the number of nodes in the network (or directed network).

A directed network is said to be *strongly connected* if for any two nodes u and v , they are mutually reachable from each other, that is, $d(u, v) < \infty$ and $d(v, u) < \infty$.

In a directed network G , a *strongly connected component* is a subnetwork that is strongly connected, and is maximal with this property: no additional nodes or directed edges in G can be included in the subnetwork without breaking the property of being strongly connected.

For a directed network $G = (V(G), A(G))$, and a subset S of nodes in G , the *subnetwork of G induced by S* , denoted by $G[S]$, is the subnetwork defined by $V(G[S]) = S$, $A(G[S]) = \{(u, v) \in A(G) : u \in S, v \in S\}$.

2.2 Biological networks

A large class of real complex networks are biological networks. Biological systems can usually be represented by networks, and analysis of the topological properties of large-scale biological networks have been very important to understanding organization principles of interactions between network components and basic mechanisms underlying cellular processes [9, 107]. In this section, we discuss two types of biological networks: brain networks and transcriptional regulatory networks.

2.2.1 Brain networks

It has been widely recognized that the entire collection of neural interconnections in a brain forms a complex network (called ‘structural brain network’) [33, 36], and a system-level network analysis is fundamental for understanding the anatomical structure of the brain [84]. Moreover, the anatomical connectivity pattern of neural elements also provides an important constraint on the possible repertoire of dynamics and functions that emerge from the underlying brain anatomical structure [38, 79].

Brain network studies have been very active in neuroscience research [13, 84]. The ongoing effort to map the whole set of neuronal connections (called ‘connectome’) of the entire brain in diverse species will improve our understanding of the network topology and function of the brains [90]. For our discussion, brain networks refer to structural brain networks, in which nodes are neural elements (brain regions) and edges are anatomical connections (axonal projections).

The extreme complexity of neuronal connections renders it practically impossible to map the connectome at a microscale (synaptic connections between individual neurons) in humans and other animals. Recently, a mesoscale connectome of the mouse brain was obtained experimentally, by an image processing pipeline [74]. By using a linear connectivity model, Oh et al. [74] constructed an inter-region anatomical connectivity matrix for the right (ipsilateral) hemisphere, and also an inter-region anatomical connectivity matrix for the left (contralateral) hemisphere.

There are totally 213 brain regions in each anatomical connectivity matrix, represented by 213 nodes in each of the resulting networks. In the original datasets [74], each entry in the connectivity matrix (corresponding to a directed edge in the resulting network) has a P-value indicating how significant is the presence of this connection. We used five different P-value cutoffs (0.05, 0.02, 0.01, 0.005 and 0.001) to obtain the mouse inter-region brain networks. Generally, when a lower P-value cutoff is chosen, more edges will be filtered out, and the resulting network (with 213 nodes) will be sparser.

Therefore, we obtained 10 directed inter-region brain networks in total: for each P-value cutoff, there are two inter-region brain networks, one in the ipsilateral hemisphere, the other in the contralateral hemisphere. All the 213 regions can be grouped into 12 major brain subdivisions, according to the locations of these regions in the mouse brain atlas [74]. The 12 subdivisions are: **Isocortex**, **Olfactory Areas**, **Hippocampus**, **Cortical Subplate**, **Striatum**, **Pallidum**, **Thalamus**, **Hypothalamus**, **Midbrain**, **Pons**, **Medulla** and **Cerebellum**.

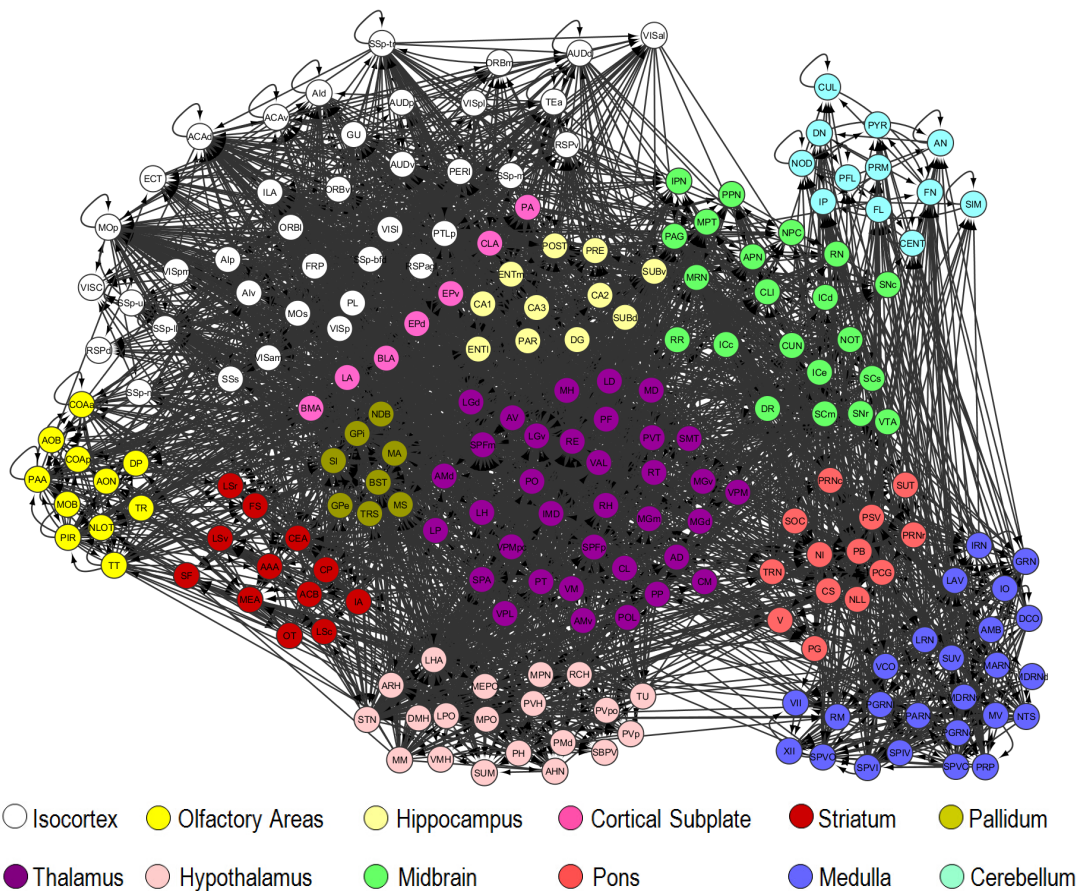


Figure 2.1: A visualization of the mouse inter-region brain network in the ipsilateral hemisphere, corresponding to P-value cutoff 0.05. Each subdivision is marked with a different color and is drawn in the network visualization based on its relative location in the brain.

We give a visualization of the mouse inter-region brain network in the ipsilateral

hemisphere (one of the 10 inter-region brain networks), obtained by choosing a P-value cutoff 0.05, in Figure 2.1. This network is a directed network with 213 nodes (brain regions) and 3123 directed edges.

2.2.2 Structures and functions of mouse brain subdivisions

In this subsection, we give an introduction to the structures and functions of the 12 subdivisions in the mouse brain.

The isocortex constitutes most of the cerebral cortex and is the most recent part of the cerebral cortex phylogenetically. In mice and primates, the isocortex is a layered structure, which consists of glia and radially extending neurons with massive interconnections [48]. The isocortex plays the important role in the central nervous system (CNS) as an associative and analytical region, providing sophisticated control and integration, based on an analytical overview of the in- and ex-teroceptive environment. The isocortex integrates simpler functions in the rest of the nervous system, and is characterized by its handling of high-level information, by receiving pre-processed sensory information via thalamic relays, and by modulating the activity of other CNS structures, and thus controls behaviour. Therefore, the cortex is more abstract, with key functions to elaborate, integrate and analyze sensory information, and to plan and oversee proper responses [26]. It is suggested that the uniform structure of the neocortex might be essential to facilitate the abstraction process.

The olfactory area of the mouse brain, also called the olfactory bulb, is located directly above the roof of the nasal cavity and can easily be distinguished by the deep transverse groove [46]. The olfactory bulb is a complex structure with several types of cells and fibers that form layers and laminae. As a neural circuit, the olfactory bulb functions as a filter that has one source of sensory input (axons from olfactory receptor neurons of the olfactory epithelium), and one output (mitral cell axons).

The hippocampus is a prominent component of the mammalian nervous system [103]. It is located under the cerebral cortex. The hippocampal complex is a rather heterogenous structure with respect to cell types, cell layers and nerve terminals [76]. The principal neuronal cell type of the hippocampus is the pyramidal cell, as well as a heterogeneous population of basket cells of various sizes and shapes, making up the various divisions of the hippocampal region [42]. The hippocampus has been studied extensively as part of a brain system responsible for spatial memory and navigation.

The cortical subplate, with its deep anatomical location below the cortical layer and above the intermediate zone, is a largely transient zone containing precocious neurons involved in several key steps of cortical development. The majority of subplate neurons form a compact layer in mouse, and contains a variety of cell types with different developmental origins, survival, connectivities, and structural and functional characteristics [98]. In rodent, subplate neurons are among the earliest born neocortical cells [99], suggesting that the evolutionary expansion of the neocortex and the expansion of the subplate zone are linked.

The striatum is a massive nucleus in the basal forebrain that plays a pivotal role in modulating motor activity and higher cognitive function. Approximately 90% of all neurons in the striatum belong to an unusual type of inhibitory projection cell referred to as medium spiny neurons, numbers of which are critical variables that influence motor performance and aspects of cognition.

The pallidum, often called globus pallidus (GP), together with the claustrum, caudate and putamen, form the basal ganglia, a set of subcortical nuclei that play an important role in facilitating voluntary movement in health and disease. The external segment of the globus pallidus (GPe) is a central nucleus in the motor-suppressing indirect pathway, which receives inputs from the striatum, subthalamic nucleus, and parafascicular nucleus of the thalamus and has been strongly implicated in the onset and maintenance of motor dysfunction in movement disorders [61]. Anatomical studies have shown that GPe neurons project to a number of brain

areas and it contains a heterogeneous population of neurons that likely contribute differentially to motor function in healthy and diseased individuals.

The thalamus is the largest part of the diencephalon in mammals [82] and its two halves are fused partly across the median third ventricle line. The nuclei of the thalamus project to the cerebral cortex in a highly organized fashion. Parts of the thalamus process all forms of sensory inputs (except olfactory), and after some modification, transmit them to a specific area of sensory cortex, while other parts are concerned with emotion, pain, memory and instinctive behaviour [46]. However, many thalamic nuclei do not receive primary sensory information, but instead are part of higher-order associative cortico-thalamic circuitry.

The hypothalamus refers to the forebrain territory that controls homeostasis. It occupies a large region between the preoptic area and the midbrain, ventral to the thalamus and the subthalamus, and there is no clear-cut caudal limit of the hypothalamus [81]. The hypothalamus is concerned with the autonomic control of cardiovascular activity, respiratory and alimentary functions, and with regulating hormone levels, as well as playing a role in eating behaviour and autonomic emotional responses (through the limbic system).

Cerebellum, pons and medulla oblongata, together with that of the midbrain, are collectively known as the brain stem and they control such essential functions as the rate of the heartbeat, breathing and involuntary gastrointestinal properties [46].

The midbrain, the smallest region of the primitive brain, links the diencephalon of the forebrain with the pons of the hindbrain (its most rostral part). Its role seems limited to processing signals from the oculomotor, trochlear and trigeminal nerves.

The pons is the rostral part of the hindbrain. It sits between the midbrain and the medulla oblongata and consists of two major parts [76]. The mature pons contains the nuclei of the abducens, facial and vestibulocochlear nerves. It also attaches to and links with the cerebellar hemispheres through the middle cerebellar peduncles, and these links are important in maximising the efficiency of motor activity.

The medulla, or the medulla oblongata, is the most caudal portion of the vertebrate hindbrain. The medulla oblongata links the pons with the spinal cord, and is larger rostrally than caudally [46]. It contains major fiber tracts as well as several motor and interneuron populations, including neural centres regulating the visceral functions and the maintenance of bodily homeostasis. Another important function of the medulla oblongata is to provide cerebrospinal fluid (CSF), which works as a buffer for the cortex of the brain, providing a basic and essential mechanical and immunological protection to the brain inside the skull, as well as acts in cerebral autoregulation of cerebral blood flow. Therefore, medulla oblongata serves a role as an important source of CSF. Medulla also functions as a tract for ascending fibres carrying sensory information principally from the body surface to the cerebral cortex, the thalamus or cerebellum, and for descending fibres from the midbrain and particularly the cerebral cortex to peripheral structures to control fine movement. The medulla oblongata has typical motor and sensory columnar anatomic structures, which were recently proved to be segmentally organized [94].

The cerebellum is connected to the midbrain and pons by the superior and inferior cerebellar peduncles, as well as the middle one. Although cerebellum has a substantial sensory input, it is principally a motor part of the brain, involved in the maintenance of equilibrium and coordination of muscle action [46].

2.2.3 Transcriptional regulatory networks

Regulation of gene expression is primarily mediated by proteins called *transcription factors* (TFs), which bind to the promoter regions of target genes (TGs) and influence the expression levels of target genes. If a TF can regulate the expression level of a TG, then there is a regulatory interaction between the TF and the TG (denoted by a directed edge $TF \rightarrow TG$). In general, a TF can either up-regulate (called ‘activation’) or down-regulate (called ‘inhibition’) a target gene, therefore, there are two types of regulatory interactions. For our study, we ignore the information of

the type of each regulatory interaction and only consider the connection patterns between TFs and TGs, since the activation and inhibition information for the regulatory interactions are unavailable for most of transcriptional regulatory networks. Thus, the set of all regulatory interactions in an organism can be simply modelled as a directed network, called the *transcriptional regulatory network*: the nodes are TFs and TGs, the directed edges are regulatory interactions that link TFs to TGs. The modelling of regulatory interactions between TFs and their target genes as a directed network could provide a general framework to identify general principles that govern the regulation of transcription at a genome scale.

In humans, a complete mapping of all the regulatory interactions is challenging, given the large number of transcription factors and genes, and the complexity of the patterns of the regulatory interactions. The regulatory interactions between TFs form a *transcription factor (TF) regulatory network*, in which nodes are TFs and a directed edge represents a regulatory relationship between two TFs (one TF binds to the regulatory region of the gene of the other TF).

There has been effort to map the human TF regulatory networks of 41 cell types, using the DNase I footprinting technique [66]. Neph et al. [66] reported a resource of 41 human cell-type-specific TF regulatory networks, and we used these network datasets for our study.

All the 41 cell types can be grouped into 8 classes based on their developmental and functional properties: **Embryonic Stem Cells, Epithelia, Blood, Endothelia, Cancer, Visceral Cells, Fetal Tissues** and **Stromal Cells**.

In Table 2.1, we show the class, the cell type, the number of nodes (N) and the number of edges (L) for the human TF regulatory network of each of the 41 cell types.

| Class | Cell type | N | L |
|-----------------------------|----------------------------|-----|-------|
| Embryonic Stem Cells | hESC | 533 | 16424 |
| Epithelia | Renal Cortical Epi. | 525 | 9597 |
| | Choroid Plexus Epi. | 527 | 13903 |
| | Small Airway Epi. | 522 | 9886 |
| | Amniotic Epi. | 526 | 13286 |
| | Esophageal Epi. | 528 | 14577 |
| | Iris Pigment Epi. | 527 | 12511 |
| Blood | Hemat. Stem Cell | 526 | 16461 |
| | Promyelocytic Leuk. | 525 | 18906 |
| | Erythroid | 493 | 9099 |
| | T-Lymphocyte | 518 | 12812 |
| | B-Lymphocyte | 515 | 16723 |
| | B-Lymphoblastoid (GM06990) | 501 | 12994 |
| | B-Lymphoblastoid (GM12865) | 513 | 15202 |
| Endothelia | Adult Dermal Blood | 520 | 13510 |
| | Neonatal Dermal Blood | 526 | 16761 |
| | Lung Lymphatic | 520 | 15435 |
| | Neonatal Dermal Lymph. | 526 | 15582 |
| Cancer | Neuroblastoma | 508 | 12761 |
| | Hepatoblastoma | 493 | 12863 |
| Visceral Cells | Hippocampal Astrocyte | 531 | 16391 |
| | Skeletal Myoblast | 523 | 13806 |
| | Skeletal Muscle | 529 | 17320 |
| | Astrocyte | 516 | 9296 |
| Fetal Tissues | Fetal Brain | 519 | 11698 |
| | Fetal Heart | 516 | 14295 |
| | Fetal Lung | 532 | 17823 |
| Stromal Cells | Aortic Fibroblast | 529 | 14795 |
| | Pulmonary Fib. | 527 | 14588 |
| | Fetal Lung Fib. | 519 | 11274 |
| | Lung Fib. | 527 | 14700 |
| | Adult Dermal Fib. | 529 | 13644 |
| | Neonatal Dermal Fib. | 521 | 15565 |
| | Cardiac Fib. | 527 | 15115 |
| | Cardiac Fib. | 522 | 14492 |
| | Pulmonary Artery Fib. | 531 | 13501 |
| | Skin Fib. | 521 | 12482 |
| | Mesenchymal Fib. | 526 | 15135 |
| | Mammary Fib. | 526 | 13961 |
| | Periodontal Fib. | 521 | 12822 |
| | Foreskin Fib. | 513 | 12126 |

Table 2.1: The network sizes of the human TF regulatory networks of 41 cell types.

In Figure 2.2, we give a visualization of the TF regulatory network of human embryonic stem cells (hESC).

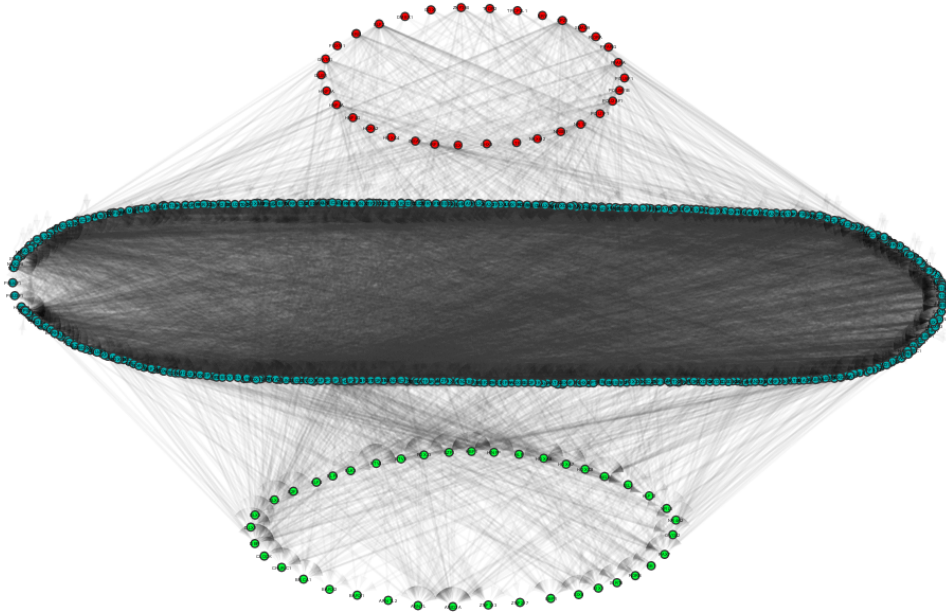


Figure 2.2: A visualization of the TF regulatory network of hESC. The network has a three-layer hierarchical structure. The top layer is marked in red, the middle layer is marked in cyan, and the bottom layer is marked in green.

2.3 Network metrics

A network metric can be thought of as a numerical measurement defined for either a node or a whole network. To characterize the structure of complex networks, there have been many network metrics (measures) proposed. The node degree is perhaps the simplest amongst all these, it helps to identify the most connected nodes (called ‘hubs’) that might have a significant role in maintaining the structural integrity of a network [3]. Hubs may also have significant functional implications, for example, in the yeast protein-protein interaction networks, the highly connected proteins are more likely to be essential for cell survival than proteins with a small number of interactions [41].

In this section, we will review several network metrics that have been widely used to study the topological properties of complex networks.

2.3.1 Clustering coefficient and transitivity

In an (undirected) network, the clustering coefficient is a local measure defined for individual nodes [100]: the (local) *clustering coefficient* of node i is the ratio of the number of connections between neighbors of i to the number of all possible connections,

$$C_i = \frac{E_i}{\frac{1}{2}k_i(k_i - 1)}, \quad (2.1)$$

where E_i is the number of edges between neighbors of node i and k_i is the degree of node i .

Then, the *global clustering coefficient* is defined to be the average of local clustering coefficients of all nodes,

$$C = \frac{1}{n} \sum_i C_i,$$

where n is the number of nodes in the network.

The local clustering coefficient of each node i , C_i , lies in the range $0 \leq C_i \leq 1$, and thus the global clustering coefficient C also lie in between 0 and 1, $0 \leq C \leq 1$.

A similar definition, usually used in sociology literature, is called transitivity. Unlike global clustering coefficient that is defined as the mean of ratios, transitivity is defined to be the ratio of means. For an (undirected) network, the *transitivity* is a global network metric defined as [69]

$$C' = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}}.$$

Here, a connected triple is a subgraph of a single node connecting to two neighbors by two edges.

The transitivity C' measures the fraction of connected triples that have their third edge filled in to complete the triangle.. Thus, this network metric C' lies in the range $0 \leq C' \leq 1$.

The definition of clustering coefficient and transitivity can be generalized to directed networks [25]. Define $k^{\leftrightarrow}(i)$ to be the number of reciprocal edges between i and its neighbors (the reciprocity is defined for a pair of nodes i and j for which $i \rightarrow j$ and $j \rightarrow i$ both exist). Then

$$k^{\leftrightarrow}(i) = \sum_{j \neq i} a_{ij} a_{ji}.$$

The clustering coefficient of node i is defined to be the ratio of the number of all directed triangles formed by i to the number of all possible triangles that i can form [25],

$$C_i = \frac{\frac{1}{2} \sum_{j \neq i} \sum_{h \neq i, j} (a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj})}{k(i)(k(i) - 1) - 2k^{\leftrightarrow}(i)}. \quad (2.2)$$

It is easy to see that $0 \leq C_i \leq 1$, and the global clustering coefficient is defined to be [25]

$$C = \frac{1}{n} \sum_i C_i.$$

Based on the same idea, transitivity can also be defined for directed networks. Using the same notation as above, transitivity can be computed by

$$C' = \frac{\sum_i \frac{1}{2} \sum_{j \neq i} \sum_{h \neq i, j} (a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj})}{\sum_i (k(i)(k(i) - 1) - 2k^{\leftrightarrow}(i))}. \quad (2.3)$$

2.3.2 Betweenness

Betweenness [29, 101] is a centrality metric that measures the extent to which a node lies on shortest (geodesic) paths between other nodes. Nodes with higher betweenness may have higher influence within a network in terms of their control over information passing between others. In other words, a node with very high betweenness tends to be a ‘bottleneck’ through which a large number of shortest paths between others pass.

In a directed network, the *betweenness* of node i is defined by [29]

$$BC(i) = \frac{1}{(n-1)(n-2)} \sum_{j, k: j \neq k \neq i} \frac{g_{jk}(i)}{g_{jk}}. \quad (2.4)$$

Here, g_{jk} is the number of shortest paths from node j to k , and $g_{jk}(i)$ is the number of shortest paths from node j to k that pass through node i . If both g_{jk} and $g_{jk}(i)$ are zero, define $\frac{g_{jk}(i)}{g_{jk}} = 0$. Note that, the normalization by $(n-1)(n-2)$ makes sure that $0 \leq BC(i) \leq 1$ for any node i . The betweenness is similarly defined for undirected networks.

One interesting feature of betweenness is that the betweenness values are typically distributed over a wide range. For example, in a star network in which a single central node is connected to $n-1$ other nodes and there are no other edges, the central node attains a maximum betweenness value 1 and for the $n-1$ other nodes each has betweenness value 0. In real networks, the range can not be $[0, 1]$ as in the extreme case of a star network, but it is nonetheless large and typically getting larger as n increases.

2.3.3 Closeness

Closeness is a centrality metric that measures the average distance of a node to other nodes in a network. It can be understood as the efficiency of each node in spreading information to other nodes.

In literature, there are various definitions for closeness, here we will use one of them that can be generally applied to all directed networks.

In a directed network, the *closeness* of node i is defined by [72]

$$CC(i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d(i, j)}. \quad (2.5)$$

It is easy to see that $0 \leq CC(i) \leq 1$ for any node i . Based on the definition, nodes with higher closeness centrality values are ‘closer’ to the other nodes of the network. The closeness can be computed based on the distance matrix, in which the distance between each pair of nodes can be computed by Dijkstra’s algorithm in time $O(|V|^2)$. The closeness is similarly defined for undirected networks.

2.3.4 Eigenvector centrality

In a directed network, the eigenvector centrality of each node is defined to be proportional to the sum of eigenvector centralities of its in-neighbors. Let $x = (x_1, x_2, \dots, x_n)$ and x_i be the eigenvector centrality of node i , $i = 1, \dots, n$. Then, x is defined to be the eigenvector of the largest eigenvalue of A^T [12], that is,

$$A^T x = \lambda_1 x.$$

Here, $A = (a_{ij})$ is the adjacency matrix of the directed network, λ_1 is the largest eigenvalue of A^T in absolute value.

Therefore, the vector x of eigenvector centralities is proportional to the leading (right) eigenvector of A^T (corresponding to the largest eigenvalue λ_1). By normalizing x , we define the *eigenvector centrality* $EC(i)$ of node i to be

$$EC(i) = n \cdot \frac{x_i}{\sum_{j=1}^n x_j}. \quad (2.6)$$

Note that, this normalization does not produce the fact that $EC(i)$ lies between 0 and 1, but it ensures that the average eigenvector centrality stays at the constant 1 as n varies.

The definition of eigenvector centrality for undirected networks is the same as that for directed networks. Note that, for an undirected network, its adjacency matrix is symmetric and has real eigenvalues.

Some difficulty may arise in the definition of eigenvector centrality for directed networks, since the leading eigenvector of A^T might not be unique. But one situation is special (due to Perron–Frobenius theorem [47]): if a directed network G is strongly connected (equivalently, its adjacency matrix A is irreducible), then there is a positive real eigenvalue λ_1 which has the largest absolute value and the leading eigenvector of A^T corresponding to λ_1 is unique and positive.

There are two generalizations of eigenvector centrality: Katz centrality and PageRank centrality.

2.3.5 Katz centrality

To define Katz centrality, one assumes that each node gains some inherent centrality, so that the Katz centrality of each node is not zero. In a directed network, the Katz centrality of each node is contributed by two parts: the sum of centralities of its in-neighbors and the inherent centrality. Introducing a parameter $\alpha > 0$ and a vector parameter $\beta = (\beta_1, \dots, \beta_n)$ with $\beta_i > 0, i = 1, \dots, n$, we have

$$x_i = \alpha \sum_j a_{ji} x_j + \beta_i.$$

In matrix form, the Katz centrality vector $x = (x_1, x_2, \dots, x_n)$ can thus be expressed as [45]

$$x = (I_n - \alpha A^T)^{-1} \beta.$$

By normalization, we define the *Katz centrality* of node i to be

$$KC(i) = n \cdot \frac{x_i}{\sum_{j=1}^n x_j}. \quad (2.7)$$

2.3.6 PageRank centrality

It is reasonable to think that the centrality one node gains from its in-neighbor is diluted: the centrality of a node is evenly distributed into its out-neighbors. Using this idea, one can further extend Katz centrality to PageRank centrality, which is known to form the underlying algorithm for ranking webpages used by Google. Mathematically, let $x = (x_1, x_2, \dots, x_n)$ be the PageRank centrality vector, then

$$x_i = \alpha \sum_j a_{ji} \frac{x_j}{k_{out}(j)} + \beta_i.$$

Here, if a node j has out-degree zero, then set $k_{out}(j) = 1$.

In matrix form, the PageRank centrality vector x can be expressed as [72]

$$x = D(D - \alpha A^T)^{-1} \beta,$$

where $D = (d_{ij})$ is a diagonal matrix in which each diagonal element d_{ii} is taken to be $d_{ii} = \max \{k_{out}(i), 1\}$.

By normalization, we define the *PageRank centrality* of node i to be

$$PC(i) = n \cdot \frac{x_i}{\sum_{j=1}^n x_j}. \quad (2.8)$$

2.4 Network structural characteristics

There are several structural properties that are relevant for real networks. These structural properties not only distinguish real networks from random networks, but also provide explanations for the differences between various behaviors of real networks and random networks.

2.4.1 Small-world property

Social networks display the ‘small-world property’: the networks have small average path lengths \bar{d} and high clustering coefficients C . In other words, in social networks, it only takes a small number of steps to get from any person to any other person, and two friends of a person are likely to be friends themselves. In popular parlance, the small-world property is also known as ‘six degrees of separation’. Perhaps, the most well-known experimental study [63] to demonstrate the small-world phenomenon was taken in 1967 by Stanley Milgram, who showed that it takes on average about six steps for a letter to reach from any individual to a target person.

A network is *random*, if there is no rule that is biased towards connecting certain nodes, that is, the probability of connecting any two nodes is equal [24]. A network is called *regular*, if every node has the same degree, it is further called *k-regular*, if every node has degree k .

To explain the origin of the small-world property, Watts and Strogatz [100] introduced a small-world network model, called *WS model*, to interpolate between regular networks and random networks by varying a rewiring probability p . At

$p = 0$, the network is a regular ring network, which is highly clustered ($C(0) \sim \frac{3}{4}$) and has a large average path length ($\bar{d}(0) \sim \frac{n}{2k}$) with $k \ll n$. At $p = 1$, the network becomes a random network, which has a small clustering coefficient ($C(1) \sim \frac{k}{n}$) and a small average path length ($\bar{d}(1) \sim \frac{\ln n}{\ln k}$).

At intermediate values of p , it was observed [100] that there is a wide interval over which the networks generated by WS model at p show comparable (small) average path length to random networks (at $p = 1$), and comparable (high) clustering coefficient to regular ring networks (at $p = 0$). Networks interpolated at p within this interval are regarded to show small-world property: both small average path length and high clustering coefficient (in comparison to random networks).

Therefore, for a regular network, a small fraction of long-range connections (shortcuts) introduced to the network can significantly reduce the average path length, while keeping a relatively high clustering coefficient.

To test whether a network is ‘small-world’, Humphries [40] proposed a quantitative measure S called ‘small-worldness’. A network is deemed small-world, if $S > 1$ is true, which can be tested statistically.

For a network G with N nodes and L edges, denote the average path length by \bar{d} and the global clustering coefficient by C . Let \bar{d}_{rand} (C_{rand}) be the average path length (global clustering coefficient) of the randomized network of G (random network with N nodes and L edges). That is, \bar{d}_{rand} (C_{rand}) is calculated as the mean of average path lengths (global clustering coefficients) of an ensemble of randomized networks of G .

Let α be the ratio of the global clustering coefficients

$$\alpha = \frac{C}{C_{rand}},$$

and β be the ratio of the average path lengths

$$\beta = \frac{\bar{d}}{\bar{d}_{rand}},$$

define the measure of *small-worldness* as

$$S = \frac{\alpha}{\beta}.$$

Then, the network G is called a *small-world* network, if $S > 1$ [40]. The small-worldness can be similarly defined for directed networks that are strongly connected.

Using the measure of small-worldness, many real networks can be quantified, and have been observed to be small-world. It has also been found [40] that the small-worldness S (positively) scales linearly with the network size N across a large number of real networks ($S \sim N$).

The small-world property can have nontrivial effects on the dynamics of complex networks. For example, in a small-world network structure, the disease (or information) can spread more easily and quickly in the network [53, 100].

2.4.2 Network motif

In many real networks, there are certain small connection patterns that are recurring at high frequency. Such small recurring connection patterns (subnetworks) are called ‘network motifs’ [65]. Network motifs are usually defined as small directed and connected subnetworks within a directed network, even though network motifs can be similarly defined in undirected networks.

Generally, there are multiple possible connection patterns for a n -node (directed and connected) subnetwork. For example, there are totally 13 types of 3-node subnetworks, including directed triangle, feed-forward loop, feed-back loop and others.

For a directed network G , to uncover a network motif as a n -node subnetwork ($n = 3, 4, \dots$), a set of randomized networks should be generated as the null hypothesis, and statistical analysis is applied. Specifically, an ensemble of randomized networks are generated, each of which has the same $(n - 1)$ -node subnetwork count (for each $(n - 1)$ -node subnetwork) as in network G .

Then, a n -node subnetwork is defined to be a (n -node) *motif*, if the number of times it occurs in G is significantly higher than those in randomized networks [65]. The criteria for defining ‘significance’ here are as follows [65]: (i) The probability

that the n -node subnetwork appears in a randomized network an equal or greater number of times than in network G is smaller than $P = 0.01$; (ii) The number of times it appears in network G with distinct sets of nodes is at least $U = 4$; (iii) The number of occurrence of the n -node subnetwork in network G , N_{real} , satisfies the relation: $N_{real} - N_{rand} > 0.1N_{rand}$, where N_{rand} is the average number of times the n -node subnetwork appears in the randomized networks.

In Figure 2.3, we show some network motifs ($n = 3, 4$), including feed-forward loop, feed-back loop and bi-fan, that can be found in real networks.

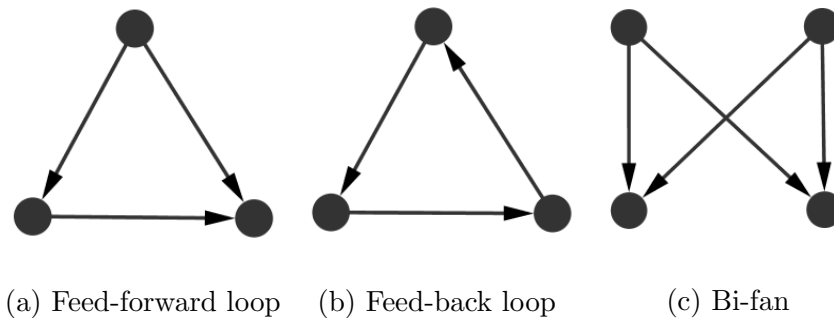


Figure 2.3: Some simple network motifs: (a) feed-forward loop, (b) feed-back loop, (c) bi-fan.

Network motifs are local substructures that may have particular functional roles in real networks [4]. For instance, the feed-forward loop, observed to be a motif in both the transcriptional regulatory network of *E. coli* and the neural network of the nematode *Caenorhabditis elegans* [65], may carry out the function of information processing, important in both types of networks.

The similarity of local substructures in real networks can help to classify real networks of various kinds into distinct superfamilies, based on the significance profile of 3-node subnetworks [64].

In structural analysis of biological networks, network motifs (functional substructures) can be used to develop new centrality measures in order to rank the network elements, or to find key elements in biological networks [51]. For instance, the motif-based centrality analysis can help to identify key regulators in the gene

regulatory network of *E. coli*. The motif-based centrality measures can result in interesting different rankings of elements in the gene regulatory network of *E. coli*, and outperform other commonly-used centrality measures [50].

2.4.3 Degree correlation

In real networks, there are two interesting features: the tendency for hubs to connect to hubs (assortativity), and the tendency for hubs to connect to nodes with small degree (disassortativity). For example, in social networks, celebrities (hubs) usually know each other; while in protein-protein interaction networks, high-degree proteins tend to be linked to proteins with low degree. These phenomena observed in real networks have consequences on the network topology and other properties.

Therefore, the connection between two nodes (based on their degrees) is not random in real networks, in other words, the degrees of the two end nodes of an edge are correlated. We say that there are usually degree correlations in real networks.

To quantify the degree correlation in a network $G = (V(G), E(G))$, there were two methods proposed: one based on degree correlation coefficient [68], the other based on degree correlation function [78].

Denote by q_k the probability that there is a node with degree k at the end of one randomly selected edge. Then, q_k is given by

$$q_k = \frac{kp_k}{\langle k \rangle},$$

where p_k is the probability that a node has degree k in the network, and $\langle k \rangle$ is the average degree of the network defined by $\langle k \rangle = \frac{1}{|V(G)|} \sum_{u \in V(G)} k(u)$.

Let e_{ij} be the probability that there is an edge that connects a node with degree i to a node with degree j . By definition, we have $\sum_{i,j} e_{ij} = 1$ and $\sum_j e_{ij} = q_i$. If the network is neutral, that is, there is no degree correlation, then we expect $e_{ij} = q_i q_j$. If there is degree correlation in the network, then e_{ij} will deviate from $q_i q_j$.

The *degree correlation coefficient* is defined to be [68]

$$r = \frac{1}{\sigma_q^2} \sum_{j,k} jk(e_{jk} - q_j q_k), \quad (2.9)$$

where the variance σ_q^2 is given by

$$\sigma_q^2 = \sum_k k^2 q_k - (\sum_k k q_k)^2.$$

The degree correlation coefficient r is a Pearson correlation coefficient, and thus lies in the range $-1 \leq r \leq 1$. A network is called *assortative*, if $r > 0$; it is called *disassortative*, if $r < 0$; it is called *neutral*, if $r = 0$.

The degree correlation coefficient r is defined for undirected networks, and can be readily generalized to directed networks [28]. In a directed network, four types of degree correlation coefficients exist: $r_{in,in}$, $r_{in,out}$, $r_{out,in}$ and $r_{out,out}$.

In assortative networks, nodes tend to connect to other nodes with comparable degrees, thus, hubs tend to connect to hubs; in disassortative networks, nodes with dissimilar degrees tend to connect to each other. Based on the degree correlation coefficient r , social networks were found to be assortative networks, whereas technological networks and biological networks are mostly disassortative [68].

Another way to quantify the degree correlation in a network is by considering the average degree of the neighbors of each node i , which is given by

$$k_{nn}(k_i) = \frac{1}{k_i} \sum_j a_{ij} k_j,$$

where k_i denotes the degree of node i , and $A = (a_{ij})$ is the adjacency matrix of the network.

If $k_{nn}(k_i)$ is evaluated for all nodes with the same degree k , there is a relationship between $k_{nn}(k)$ and k that can be thought of as a function. Here, $k_{nn}(k)$ is the average degree of neighbors of all degree- k nodes, and is a function of the degree k . This function is called the *degree correlation function*, as defined by [78]

$$k_{nn}(k) = \sum_{k'} k' P(k'|k), \quad (2.10)$$

where $P(k'|k)$ is the conditional probability of having a degree- k' neighbor for a degree- k node.

In a neutral network, we have

$$P(k'|k) = \frac{e_{kk'}}{\sum_{k'} e_{kk'}} = \frac{e_{kk'}}{q_k} = \frac{q_k q_{k'}}{q_k} = q_{k'}.$$

Thus, $k_{nn}(k)$ can be expressed as

$$k_{nn}(k) = \sum_{k'} k' q_{k'} = \sum_{k'} k' \frac{k' p_{k'}}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle}.$$

Therefore, in a neutral network (no degree correlation), $k_{nn}(k)$ is independent of k , that is, a horizontal line at $k_{nn}(k) = \frac{\langle k^2 \rangle}{\langle k \rangle}$ will be observed when $k_{nn}(k)$ is plotted versus k .

The scaling of $k_{nn}(k)$ as a function of k can be approximated by

$$k_{nn}(k) = ak^\mu,$$

where μ is called the *degree correlation exponent*.

In an assortative network, it is expected that $k_{nn}(k)$ increases with k , that is, $\mu > 0$. In a disassortative network, $k_{nn}(k)$ decreases with k , that is, $\mu < 0$.

Therefore, both single parameters r and μ can be used to characterize the degree correlation in a network, as shown in Table 2.2.

| | Degree correlation coefficient $r = \frac{1}{\sigma_q^2} \sum_{j,k} j k (e_{jk} - q_j q_k)$ | Degree correlation coefficient $k_{nn}(k) = \sum_{k'} k' P(k' k) = ak^\mu$ |
|-------------------------|---|--|
| Assortative networks | $r > 0$ | $\mu > 0$ |
| Disassortative networks | $r < 0$ | $\mu < 0$ |
| Neutral networks | $r = 0$ | $\mu = 0$ |

Table 2.2: Degree correlation based on degree correlation coefficient and degree correlation function.

The degree correlation also has impact on other network properties. For instance, assortative networks have lower phase transition points for the emergence of a giant

component and are more robust to removal of hubs [68]. Degree correlations can also impact our ability to control a directed network [80].

2.4.4 Community structure

Many real networks are found to divide naturally into close-knit clusters, called *communities* or *modules*. The community structure (or modular structure) of a network refers to a partition of all the nodes into groups such that within each group, the nodes are densely connected to each other, while there are only sparse connections between two groups [73]. Detecting such community structures can help us better understand the topology and function of a large-scale network.

Therefore, to find a community structure in a large-scale network that is ‘optimal’ in some sense is of practical significance. For such purposes, many algorithms or methods have been proposed [27]. A set of approaches to the community-detection problem are based on a measure called ‘modularity’: it is high for ‘good’ divisions of the network (or community structures with good quality) and low for ‘bad’ ones [16, 70, 71].

There is also a method for detecting community structure in directed networks, based on maximizing the generalized modularity by incorporating the edge direction information [54]. In directed networks, the edge direction cannot be ignored for community detection, since this method can find more meaningful community structure than community-detection methods that work for the corresponding undirected networks by ignoring the edge direction [54].

The detection of community structures could provide insight into understanding the network organization and function in real networks. In social networks, the communities uncovered by the community-detection algorithms might represent groups of people, and within each group the people may share common interest, belief, profession, and other social ties. Biological networks are often characterized by a modular organization, reflecting functional associations between network

components. For example, it has been observed that in brain networks, there are remarkable modular structures with each module responsible for specific functions [35, 95].

2.5 Topological properties of biological networks

In this section, we will discuss the topological properties of brain networks and transcriptional regulatory networks, based on the tools developed in the previous sections.

2.5.1 Topological properties of brain networks

Brain networks have been shown to have small-world properties across several species, ranging from *C. elegans* [96] to cats and monkeys [92], and humans [37]. It has also been found that brain networks have a modular structure [15], that is, the brain regions can be grouped into modules. The small-world property and modularity of brain networks could deliver both segregated and integrated information processing, in favor of both local specialization and global efficiency of information transfer [91]. In human brain networks, there is a ‘rich-club’ organization: high-degree nodes (structural hubs) tend to be more densely connected among themselves than nodes with lower degree [95].

Next, we calculated several network metrics of the 10 (directed) mouse inter-region brain networks we obtained (see Subsection 2.2.1). The results are given in Table 2.3. Note that, for each network, the average path length and small-worldness were calculated for the largest strongly connected component of this network.

Table 2.3 shows that the values for global clustering coefficient, transitivity, average betweenness and average closeness in ipsilateral networks are higher than those in contralateral networks. This might indicate that ipsilateral networks are more clustered and more efficient in information passing along the network edges.

Moreover, we found that the largest strongly connected component of each mouse inter-region brain network is a small-world network.

| Network metrics | | | | | | | | | | |
|-------------------------------|---------------------|--------|--------|--------|--------|-----------------------|--------|--------|--------|--------|
| | ipsilateral network | | | | | contralateral network | | | | |
| P-value cutoff | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| Number of edges | 3123 | 2721 | 2487 | 2278 | 1947 | 2451 | 2160 | 1979 | 1830 | 1587 |
| Global clustering coefficient | 0.2965 | 0.2825 | 0.2732 | 0.2725 | 0.2642 | 0.2489 | 0.2355 | 0.2342 | 0.2284 | 0.2191 |
| Transitivity | 0.2302 | 0.2165 | 0.2087 | 0.2054 | 0.1953 | 0.1868 | 0.1743 | 0.1679 | 0.1608 | 0.1516 |
| Average betweenness | 0.0083 | 0.0087 | 0.0093 | 0.0097 | 0.0108 | 0.0065 | 0.0068 | 0.0070 | 0.0074 | 0.0076 |
| Average closeness | 0.3878 | 0.3589 | 0.3394 | 0.3234 | 0.2886 | 0.2990 | 0.2788 | 0.2665 | 0.2542 | 0.2287 |
| Average path length | 2.8668 | 3.0245 | 3.1886 | 3.3142 | 3.6810 | 2.8632 | 3.0087 | 3.1129 | 3.2513 | 3.4855 |
| Small-worldness | 3.2563 | 3.4582 | 3.5324 | 3.8000 | 3.8778 | 3.0101 | 3.1271 | 3.2569 | 3.3412 | 3.7090 |

Table 2.3: Some network metrics of the mouse inter-region brain networks in both ipsilateral and contralateral hemispheres, corresponding to P-value cutoffs 0.05, 0.02, 0.01, 0.005 and 0.001.

In the following, we further analyze the ipsilateral network, obtained by choosing a P-value cutoff 0.05. We denote this network by G_r , and a network visualization of G_r is given in Figure 2.1.

First, G_r has a largest strongly connected component with size 201, connecting to the other 12 nodes which are all sink nodes (nodes with out-degree zero). These 12 sink nodes are from subdivisions: Hippocampus (1 region), Striatum (3 regions), Thalamus (4 regions), Pons (3 regions) and Medulla (1 region).

Next, we found a substantial amount of reciprocal (bidirectional) links and self-loops in G_r : there are totally 479 reciprocal links (2-loops) and self-loops. Moreover, the reciprocal links are more prevalent in the subdivision Isocortex than in the other subdivisions, suggesting that the cortex may be characterized by different connection rules [74].

Furthermore, we take the undirected version of the directed network G_r (ignoring self-loops), which is an undirected network that contains 213 nodes and 2644 undirected edges. We then calculated the small-worldness for this undirected network: the average path length $\bar{d} = 2.1906$, the global clustering coefficient $C = 0.4472$, and the small-worldness is $S = 3.3667$ ($\bar{d}_{rand} = 1.9306$, $C_{rand} = 0.1171$, for 1000 randomized networks). Therefore, G_r is a small-world network.

We further calculated various centrality measures for all the 213 nodes (brain regions) in G_r , and identified the nodes with the highest values with respect to each centrality measure. These results are given in Table 2.4. Note that, in Table 2.4, for each centrality measure, the top 10 centrality hubs are sorted in decreasing order of the centrality values.

| Centrality measure | The top 10 centrality hubs |
|------------------------|---|
| In-degree | RT, LHA, DR, PCG, VTA, SPA, PB, PP, CP, SPFP |
| Out-degree | MOp, ENTl, LHA, STN, MOs, PERI, PTLp, SUBd, NDB, ACAd |
| Degree | LHA, ENTl, MOs, MOp, STN, NDB, PERI, PTLp, ACAd, SCm |
| Betweenness | LHA, NDB, PCG, ENTl, PRNc, CLI, PAG, NI, PP, STN |
| Closeness | STN, MOp, ENTl, LHA, MOs, PERI, ACAd, PTLp, SSp-bfd, SUBd |
| Eigenvector centrality | DR, LHA, PCG, VTA, PH, PAG, SPA, LPO, SPFm, MPO |
| Katz centrality | DR, LHA, PCG, VTA, PH, PAG, SPA, LPO, SPFm, MPO |
| PageRank centrality | FL, FN, DCO, CUL, CLI, NOD, LAV, CENT, IP, PYR |

Table 2.4: The top 10 nodes with the highest centrality measures in G_r .

Table 2.4 shows that the in-degree hubs are different from the out-degree hubs (except for LHA), and the degree hubs are the same as the out-degree hubs (except for SCm). This can be explained by the fact that the highest out-degrees are considerably larger than the highest in-degrees in G_r . In the 10 degree hubs, MOs, MOp, PERI, PTLp and ACAd are brain regions in Isocortex, suggesting that Isocortex is characterized by dense connections. In the 10 closeness hubs, there are six regions (MOp, MOs, PERI, ACAd, PTLp and SSp-bfd) that are in Isocortex, further suggesting that Isocortex plays a central role in passing information to other subdivisions in the mouse brain.

The largest out-degree hub is found to be MOp (primary motor area) in Isocortex. The largest in-degree hub is found to be RT (reticular nucleus of the thalamus) in Thalamus. This might be consistent with that the thalamus is highly heterogeneous, receiving and relaying diverse sensory, motor, behavioural state and cognitive information in parallel pathways from the isocortex [74].

Both the largest degree hub and the largest betweenness hub are found to be

LHA (lateral hypothalamic area) in Hypothalamus, suggesting its central role in sending signals to other regions and receiving signals from other regions as well. Stimulation in the lateral hypothalamic area not only causes thirst and hunger, but also increases the general level of activity of an animal including a mouse [23].

2.5.2 Topological properties of transcriptional regulatory networks

Transcriptional regulatory networks have a complex structure and can be investigated at both local and global levels. At a local level, the transcriptional regulatory networks have been shown to contain small recurrent patterns of interconnections, or network motifs [85], that may perform specific information processing tasks [4]. For example, the feed-forward loop (FFL) is a network motif in the transcriptional regulatory networks of several organisms, and different structural types of FFLs (based on the ‘activation’ or ‘repression’ of each regulatory interaction) may carry out different functions [60]. At a global level, the transcriptional regulatory networks are characterized by a power-law distribution for the out-degree: there are a few TFs, called ‘global regulators’, that regulate a strikingly large number of TGs and a vast majority of TFs (referred to as ‘fine tuners’) that regulate a small number of TGs; on the other hand, the in-degree (or the number of regulating proteins per regulated gene) follows an exponential distribution [34].

Furthermore, the transcriptional regulatory networks display extensive combinatorial regulation, i.e., regulation of a gene by two or more transcription factors simultaneously or under different conditions [7]. The overall regulation of a large number of genes by a relatively small set of TFs in a combinatorial fashion can set up notably complex spatial and temporal patterns of gene expression. The transcriptional regulatory networks also possess a multi-layer hierarchical structure [104]. It has also been found that TFs have static and dynamic properties that are similar

within a layer and different across layers, in the three-layer hierarchical organization of the Yeast transcriptional regulatory network [43]. The network properties mentioned above have also been shown to be existent in the human transcriptional regulatory network derived from the ENCODE data [30].

In the following, we will analyze the topological properties of the TF regulatory network of hESC. The TF regulatory network of hESC contains 533 nodes (TFs) and 16424 directed edges and is visualized in Figure 2.2. It has a largest strongly connected component (SCC) with size 459, and each of the rest 74 TFs forms a SCC by itself. The largest SCC contains 459 TFs and 14483 directed edges.

The TF regulatory network of hESC has a well-defined three-layer hierarchical structure: the top layer contains 33 TFs, the middle layer contains 459 TFs (that form the largest SCC), and the bottom layer contains 41 TFs. There are no directed edges from nodes at a lower layer back to nodes at an upper layer, in the three-layer hierarchical structure. Furthermore, there is no intra-connectivity among TFs in the top layer, i.e., each TF in the top layer has in-degree zero and only regulates downstream TFs; and there is no intra-connectivity among TFs in the bottom layer, i.e., each TF in the bottom layer has out-degree zero and is only regulated by upstream TFs.

Recently, study of the 41 human cell-type-specific TF regulatory networks [106] has shown that TF regulatory networks of different cell types demonstrate similar global three-layer hierarchical structures and the TF regulatory network of hESC is dense with an enriched number of hESC-specific regulatory interactions.

Next, we calculated several network metrics of the TF regulatory network of hESC. The results are summarized in Table 2.5. Note that, the average path length and small-worldness were calculated for the largest SCC of the TF regulatory network of hESC.

Table 2.5 shows that the largest SCC of the TF regulatory network of hESC is a small-world network with a small average path length ~ 2.4 , which is consistent with that the largest SCC is densely connected.

| | |
|-------------------------------|--------|
| Global clustering coefficient | 0.2079 |
| Transitivity | 0.1571 |
| Average betweenness | 0.0024 |
| Average closeness | 0.3925 |
| Average path length | 2.4363 |
| Small-worldness | 2.7666 |

Table 2.5: Some network metrics of the TF regulatory network of hESC.

We further calculated various centrality measures for all the 533 nodes (TFs) in the TF regulatory network of hESC, and identified the nodes with the highest values with respect to each centrality measure. The results are given in Table 2.6. Note that, in Table 2.6, for each centrality measure, the top 10 centrality hubs are sorted in decreasing order of the centrality values.

| Centrality measure | The top 10 centrality hubs |
|------------------------|---|
| In-degree | MEF2D, NR6A1, HES1, ZBTB7B, FOXD3, BHLHE40, DDIT3, FOXO3, TFAP4, E2F3 |
| Out-degree | SP1, SP3, SP4, SP2, ZBTB7B, TFAP2A, EGR3, EGR2, EGR1, TFAP2C |
| Degree | SP1, SP3, ZBTB7B, SP4, SP2, TFAP2A, EGR2, TFAP2C, EGR1, EGR3 |
| Betweenness | ZBTB7B, TFAP2C, SP1, SOX2, HES1, EGR3, REST, TFAP2A, POU5F1, EGR2 |
| Closeness | SP1, SP3, SP4, SP2, ZBTB7B, TFAP2A, EGR3, EGR2, EGR1, TFAP2C |
| Eigenvector centrality | MEF2D, HES1, ZBTB7B, NR6A1, DDIT3, BHLHE40, FOXD3, E2F3, TFE3, JUN |
| Katz centrality | MEF2D, HES1, ZBTB7B, NR6A1, DDIT3, BHLHE40, FOXD3, E2F3, TFE3, JUN |
| PageRank centrality | SOX2, PAX3, FOXD3, HBP1, SMAD3, DLX5, TFAP2C, HES1, SOX6, TGIF1 |

Table 2.6: The top 10 nodes with the highest centrality measures in the TF regulatory network of hESC.

We can see from Table 2.6 that the in-degree hubs are different from the out-degree hubs (except for ZBTB7B), and the degree hubs are the same as the out-degree hubs. This is because that the highest out-degrees are considerably larger than the highest in-degrees in the TF regulatory network of hESC, consistent with the previous finding that a TF can regulate a large number of targets, while the number of TFs that can regulate a single target is constrained [34]. Moreover, the

degree hubs contain TFs that are within the same protein family, for example, SP1, SP2, SP3 and SP4, suggesting that TFs within the same protein family might have similar connectivity. We can also see that the centrality hubs for out-degree and closeness are exactly the same, so are the centrality hubs for eigenvector centrality and Katz centrality.

In the betweenness hubs, we found three TFs (TFAP2C, SOX2, POU5F1) that are known to be specific to hESC [5], that is, they are highly expressed in hESC but not in other cell types. Given the high betweenness values of these three TFs, they might be TFs that are specific to hESC and play some central roles in the TF regulatory network of hESC. In the PageRank centrality hubs, we also found a transcription factor (FOXD3) that is specific to hESC [5].

Chapter 3

Linear systems, controllability and structural controllability

Dynamical systems [58] are mathematical objects used to model physical phenomena whose state evolves over time. A dynamical system consists of two ingredients: a *state vector* describing the state of the system at particular time and condition, a *function* that characterizes the evolution of system states over time. It is common to represent a dynamical system with a differential equation in continuous time, or a difference equation in discrete time. Generally, the equation is not linear. In this work, we focus on linear systems that are characterized by linear differential equations [59]. Depending on whether the system parameters vary with time, linear systems include two types: linear time-varying systems and linear time-invariant systems. In this chapter, we shall focus on linear time-invariant systems, and discuss the system-specific properties as *controllability* and *structural controllability*.

3.1 Dynamical system

Generally, we consider the following dynamical system, which incorporates an additional variable, the (external) input vector $u(t)$ that is injected into the system:

$$\dot{x}(t) = f(x(t), u(t), t), \quad (3.1)$$

with

$$t \in \mathbb{R}, x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m,$$

$$f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^n.$$

In order to find a solution to equation (3.1), the input vector and initial condition should be specified. Therefore, the standard problem is:

- Given:

$$f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^n, \text{ (dynamics)}$$

$$(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n, \text{ (initial condition)}$$

$$u : \mathbb{R} \rightarrow \mathbb{R}^m, \text{ (input trajectory)}$$

- Find:

$$x : \mathbb{R} \rightarrow \mathbb{R}^n, \text{ (state trajectory)}$$

- Subject to:

$$x(t_0) = x_0,$$

$$\dot{x}(t) = f(x(t), u(t), t), \forall t \in \mathbb{R}.$$

What properties should these functions possess, in order to make sure that, for example, there is always a solution? First, the input function $u : \mathbb{R} \rightarrow \mathbb{R}^m$ should

not be too ‘wild’, and in practice, we would expect u to be somehow continuous. To be continuous, on the other hand, seems to be too restrictive for u , since many interesting input functions are not continuous. Here, we assume that u is a piecewise continuous function.

Definition 3.1. A function $u : \mathbb{R} \rightarrow \mathbb{R}^m$ is *piecewise continuous* if it is continuous at all time point $t \in \mathbb{R}$ except for those t in a discontinuity set $D \subseteq \mathbb{R}$ such that:

1. For any $\tau \in \mathbb{R}$, left and right limits of u exist when t goes to τ , that is, $\lim_{t \rightarrow \tau^-} u(t)$ and $\lim_{t \rightarrow \tau^+} u(t)$ exist.
2. For any $t_0, t_1 \in \mathbb{R}$ with $t_0 < t_1$, $D \cap [t_0, t_1]$ contains a finite number of points.

We will use $PC([t_0, t_1], \mathbb{R}^m)$ to denote the linear space of all piecewise continuous functions $u : [t_0, t_1] \rightarrow \mathbb{R}^m$.

Definition 3.2. A function $p(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ is *globally Lipschitz* in its first argument $x \in \mathbb{R}^n$, if there exists a piecewise continuous function $k : \mathbb{R} \rightarrow \mathbb{R}_+$ such that

$$\forall x, x' \in \mathbb{R}^n, \forall t \in \mathbb{R}, \|p(x, t) - p(x', t)\| \leq k(t) \|x - x'\|.$$

Here, $k(t)$ is called the *Lipschitz constant* of p at $t \in \mathbb{R}$.

Next, define $p(x(t), t) = f(x(t), u(t), t)$, and we require $p(x(t), t)$ to be piecewise continuous in its second argument t , and globally Lipschitz in its first argument $x(t)$. Under these conditions, the existence and uniqueness of the solution to $\dot{x}(t) = p(x(t), t)$ with $x(t_0) = x_0$ can be established in Theorem 3.1, which is also well known as Picard’s existence theorem.

Theorem 3.1. (Existence and Uniqueness) [17] Assume $p(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ is piecewise continuous with respect to its second argument (with the discontinuity set D) and globally Lipschitz with respect to its first argument. Then for all $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$, there exists a unique continuous function $\phi : \mathbb{R} \rightarrow \mathbb{R}^n$ such that:

1. $\phi(t_0) = x_0$.
2. $\forall t \in \mathbb{R} \setminus D, \dot{\phi}(t) = p(\phi(t), t)$.

This theorem provides the existence and uniqueness of solutions to dynamical systems with some ‘well-behaved’ properties. Because of this theorem, it is now safe to say ‘the solution’ of a general differential equation provided that the conditions given in Theorem 3.1 are satisfied.

3.2 Linear time-varying system

It should be noted that, for a general dynamical system, there is no explicit form or formula for the solution. However, for a linear time-varying system, the solution can be written in an explicit form.

Linear time-varying systems naturally arise in modelling many real physical systems, they also help to study non-linear processes by some techniques called ‘linearization’ [87]. In this section, we will first study the structure and property of the solution to a linear time-varying system (equation (3.2)), and then characterize controllability of linear time-varying systems, algebraically.

3.2.1 Structure of the solution

The differential equation representing a linear time-varying system has the following form [44]:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{3.2}$$

where

$$t \in \mathbb{R}, x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m,$$

and

$$A : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}, B : \mathbb{R} \rightarrow \mathbb{R}^{n \times m}.$$

In the rest of this section, we assume that $A(t), B(t)$ and $u(t)$ are piecewise continuous. Note that, A, B are matrix-valued functions with a single argument (the time t). Based on Theorem 3.1, the uniqueness of the solution to (3.2) can be obtained.

Theorem 3.2. [59] *Assume that $A(t), B(t)$ and $u(t)$ are piecewise continuous, given $u : \mathbb{R} \rightarrow \mathbb{R}^m$ and $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$, there is a unique solution $x : \mathbb{R} \rightarrow \mathbb{R}^n$ to the system (3.2) that satisfies $x(t_0) = x_0$.*

The solution function to (3.2), given by

$$x(t) = s(t, t_0, x_0, u),$$

is called the *state transition map*. This function maps the *input* $u : \mathbb{R} \rightarrow \mathbb{R}^m$ and the initial condition (t_0, x_0) to the *state* x .

Theorem 3.3. [59] *Let D be the union of the discontinuity sets of $A(t), B(t)$ and $u(t)$. Then*

1. *For all $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n, u \in PC(\mathbb{R}, \mathbb{R}^m)$, $s(\cdot, t_0, x_0, u) : \mathbb{R} \rightarrow \mathbb{R}^n$ is continuous and differentiable for all $t \in \mathbb{R} \setminus D$.*
2. *For all $t, t_0 \in \mathbb{R}, u \in PC(\mathbb{R}, \mathbb{R}^m)$, $s(t, t_0, \cdot, u) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuous.*
3. *For all $t, t_0 \in \mathbb{R}, x_{01}, x_{02} \in \mathbb{R}^n, u_1, u_2 \in PC(\mathbb{R}, \mathbb{R}^m), a_1, a_2 \in \mathbb{R}$, we have*

$$s(t, t_0, a_1 x_{01} + a_2 x_{02}, a_1 u_1 + a_2 u_2) = a_1 s(t, t_0, x_{01}, u_1) + a_2 s(t, t_0, x_{02}, u_2).$$

4. *For all $t, t_0 \in \mathbb{R}, x_0 \in \mathbb{R}^n, u \in PC(\mathbb{R}, \mathbb{R}^m)$, we have*

$$s(t, t_0, x_0, u) = s(t, t_0, x_0, 0_U) + s(t, t_0, 0_{\mathbb{R}^n}, u),$$

where $0_{\mathbb{R}^n} = (0, \dots, 0)$ denotes the zero element of \mathbb{R}^n and $0_U : \mathbb{R} \rightarrow \mathbb{R}^m$ denotes the zero input function, i.e., $0_U(t) = (0, \dots, 0)$ for all $t \in \mathbb{R}$.

According to part 4 of Theorem 3.3, the solution can be decomposed into two distinct components:

$$s(t, t_0, x_0, u) = s(t, t_0, x_0, 0_U) + s(t, t_0, 0_{\mathbb{R}^n}, u).$$

In the above, $s(t, t_0, x_0, 0_U)$, $s(t, t_0, 0_{\mathbb{R}^n}, u)$ are called the *zero input transition map* and the *zero state transition map*, respectively.

Part 3 of Theorem 3.3 shows that $s(t, t_0, \cdot, 0_U) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear transformation. Therefore, by fixing a set of basis vectors for \mathbb{R}^n , there exists a *state transition matrix* $\Phi(t, t_0) \in \mathbb{R}^{n \times n}$ such that

$$s(t, t_0, x_0, 0_U) = \Phi(t, t_0)x_0.$$

Therefore, the state transition matrix $\Phi(t, t_0)$ completely characterizes the zero input transition map $s(t, t_0, x_0, 0_U)$. Now we can write out the explicit formula for the solution to the linear time-varying system (3.2), as given in the following theorem.

Theorem 3.4. [59] For all $t, t_0 \in \mathbb{R}, x_0 \in \mathbb{R}^n, u \in PC(\mathbb{R}, \mathbb{R}^m)$,

$$s(t, t_0, x_0, u) = \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, \tau)B(\tau)u(\tau)d\tau,$$

(state transition = zero input transition + zero state transition)

Theorem 3.4 gives the final explicit formula for the unique solution to a linear time-varying system (3.2). First, the solution is completely characterized by the state transition matrix $\Phi(t, t_0)$. Second, the structure of the solution also indicates that the zero input transition map is linear in the initial state x_0 and the zero state transition map is linear in the input u .

3.2.2 Controllability

Controllability concerns about the following question:

Can the input be used to steer the state of the system to an arbitrary value?

Now we investigate controllability of the linear time-varying system (3.2). In the rest of our discussion, we assume that the input $u : \mathbb{R} \rightarrow \mathbb{R}^m$ lies in a larger space, the space of all square integrable functions. A vector-valued function $u : \mathbb{R} \rightarrow \mathbb{R}^m$ is a square integrable function on $[t_0, t_1]$ if

$$\int_{t_0}^{t_1} u(t)^T u(t) dt < \infty.$$

Then the set of all square integrable functions on $[t_0, t_1]$ is a Hilbert space, denoted by $L^2([t_0, t_1], \mathbb{R}^m)$, with inner product defined by

$$\langle u, \hat{u} \rangle = \int_{t_0}^{t_1} u(t)^T \hat{u}(t) dt,$$

for $u, \hat{u} \in L^2([t_0, t_1], \mathbb{R}^m)$.

Since controllability of the linear time-varying system (3.2) only depends on $A(\cdot)$ and $B(\cdot)$, we simply use the pair $(A(\cdot), B(\cdot))$ to represent the linear time-varying system (3.2) and give the following definition for controllability.

Definition 3.3. The pair $(A(\cdot), B(\cdot))$ is *controllable* on $[t_0, t_1]$ if and only if for any $x_0, x_1 \in \mathbb{R}^n$, there exists $u \in L^2([t_0, t_1], \mathbb{R}^m)$ that steers the system from (x_0, t_0) to (x_1, t_1) , that is,

$$x_1 = \Phi(t_1, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau.$$

Theorem 3.5. [59] *The following statements are equivalent:*

1. $(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$.
2. For any $x_0 \in \mathbb{R}^n$, there exists $u \in L^2([t_0, t_1], \mathbb{R}^m)$ that steers the system from (x_0, t_0) to $(0, t_1)$.
3. For any $x_1 \in \mathbb{R}^n$, there exists $u \in L^2([t_0, t_1], \mathbb{R}^m)$ that steers the system from $(0, t_0)$ to (x_1, t_1) .

Theorem 3.5 states that for linear time-varying systems, controllability is equivalent to ‘controllability to zero’ and ‘reachability from zero’. These equivalent statements can simplify the analysis. First, we give the following definition.

Definition 3.4. For the pair $(A(\cdot), B(\cdot))$, a state $x_1 \in \mathbb{R}^n$ is *reachable* on $[t_0, t_1]$ if and only if there exists $u \in L^2([t_0, t_1], \mathbb{R}^m)$ that steers the system from $(0, t_0)$ to (x_1, t_1) . The *reachability map* on $[t_0, t_1]$ is the map

$$\begin{aligned} \mathbf{L}_r &: L^2([t_0, t_1], \mathbb{R}^m) \rightarrow \mathbb{R}^n, \\ u &\mapsto \int_{t_0}^{t_1} \Phi(t_1, \tau) B(\tau) u(\tau) d\tau. \end{aligned}$$

The set of reachable states under the reachability map \mathbf{L}_r is denoted by $\text{Range}(\mathbf{L}_r)$. Moreover, \mathbf{L}_r is linear and continuous. Next, the notion of controllability gramian is given in the following definition.

Definition 3.5. The *controllability gramian* of the pair $(A(\cdot), B(\cdot))$ on $[t_0, t_1]$ is the matrix

$$W_r(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_1, \tau) B(\tau) B(\tau)^T \Phi(t_1, \tau)^T d\tau.$$

The controllability gramian is a Gramian matrix, and is commonly used to check for controllability.

Theorem 3.6. [59] *The following statements are equivalent:*

1. $(A(\cdot), B(\cdot))$ is controllable on $[t_0, t_1]$.
2. $\text{Range}(\mathbf{L}_r) = \mathbb{R}^n$.
3. $W_r(t_0, t_1)$ is an invertible matrix.

The above theorem shows that for any $x_0, x_1 \in \mathbb{R}^n$, there exists $u : [t_0, t_1] \rightarrow \mathbb{R}^m$ that steers the system from (x_0, t_0) to (x_1, t_1) if and only if the matrix $W_r(t_0, t_1) \in \mathbb{R}^{n \times n}$ is invertible.

3.3 Linear time-invariant system

In this section, we turn to linear time-invariant systems, for which most results are directly based on those in Section 3.2. A linear time-invariant system is a special

case of a linear time-varying system in which the matrices $A(\cdot)$ and $B(\cdot)$ are both time-independent constant matrices.

A linear time-invariant system is usually represented by the differential equation below:

$$\dot{x}(t) = Ax(t) + Bu(t), \quad (3.3)$$

where

$$t \in \mathbb{R}, x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m,$$

and

$$A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}.$$

In the linear time-invariant system (3.3), A is called the *state matrix*, and B is called the *input matrix*.

3.3.1 Structure of the solution

For the linear time-invariant system (3.3), the solution can be given in explicit formula.

Theorem 3.7. [59] *For any given $u \in PC(\mathbb{R}, \mathbb{R}^m)$ and $(x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}$, there exists a unique solution $x : \mathbb{R} \rightarrow \mathbb{R}^n$ to the system (3.3) that satisfies $x(t_0) = x_0$. Moreover, the solution gives a function:*

$$x(t) = s(t, t_0, x_0, u) = e^{A(t-t_0)}x_0 + \int_{t_0}^t e^{A(t-\tau)}Bu(\tau)d\tau.$$

The state transition matrix for the linear time-invariant system (3.3) is given by $\Phi(t, t_0) = e^{A(t-t_0)}$. For the zero input case, the system has the ‘time-memoryless’ property: $\Phi(t, t_0) = \Phi(t - t_0, 0), \forall t, t_0 \in \mathbb{R}$. Here, the exponential of a matrix $A \in \mathbb{R}^{n \times n}$ is defined by

$$e^A = I + A + \frac{A^2}{2!} + \dots + \frac{A^k}{k!} + \dots$$

The method for computing e^A involves matrix diagonalization and canonicalization, which are mostly discussed in linear algebra.

3.3.2 Controllability

Controllability of a linear time-invariant system can be characterized by the controllability matrix of the system. We simply use the pair (A, B) to represent the linear time-invariant system (3.3).

Definition 3.6. The *controllability matrix* $P \in \mathbb{R}^{n \times nm}$ of the pair (A, B) is defined to be

$$P = [B \quad AB \quad \dots \quad A^{n-1}B].$$

Here, the notation means that P is the augmented matrix obtained by putting together the matrices $B, AB, \dots, A^{n-1}B$.

Theorem 3.8. [59] *The following statements are equivalent:*

1. *The pair (A, B) is controllable on $[t_0, t_1]$.*
2. **(Kalman's rank condition)** *The controllability matrix has full rank, that is, $\text{Rank}(P) = n$.*
3. **(Popov-Belevich-Hautus rank condition)** $\forall \lambda \in \mathbb{C}, \text{Rank}([\lambda I - A \quad B]) = n$.
4. *The controllability gramian $W_r(t_0, t_1)$ is nonsingular, where*

$$W_r(t_0, t_1) = \int_{t_0}^{t_1} e^{A(t_1-\tau)} B B^T e^{A^T(t_1-\tau)} d\tau.$$

5. *For any $v \in \mathbb{R}^n$, if $v^T B = 0$ and $v^T A = \lambda v^T$, then $v = 0$.*

Kalman's rank condition is commonly used to check for controllability of a linear time-invariant system [55, 56]. Recently, Popov-Belevich-Hautus rank condition was used by Yuan et al., as the starting point to derive the framework of exact controllability [105].

Theorem 3.8 also indicates that controllability of a linear time-invariant system is a property that is independent of the time interval chosen, moreover, it only

depends on A and B . Thus, the pair (A, B) is controllable on $[t_0, t_1]$ if and only if it is controllable on $[t_0, t_2]$ for any $t_2 \geq t_1$. Therefore, we can just talk about controllability of a linear time-invariant system, without specifying the time interval.

3.4 Structural controllability

Structural controllability was first introduced in a seminal paper by Lin [55]. Since then it has become an important topic in the research community of control theory. Recently it has found application to complex networks [56, 67].

Structural controllability makes use of the fact that the weights of the links (interactions) in real-world systems cannot be exactly known. For example, in gene (transcriptional) regulatory networks, which are directed and weighted, there is no standard method to measure the weights. Moreover, structural controllability is important because the task of verifying Kalman's rank condition in order to determine controllability is computationally intensive, especially when the size of the system is very large.

3.4.1 Lin's theorem

To discuss Lin's theorem, we consider a single-input linear time-invariant system

$$\dot{x}(t) = Ax(t) + bu(t). \quad (3.4)$$

Here, the input matrix is a column vector $b = (b_i) \in \mathbb{R}^{n \times 1}$ and the input $u(t)$ is a scalar function. A general (multi-input) linear time-invariant system is given in (3.3).

In a linear time-invariant system (3.3), entries in the matrices A and B are taken to be known constants. However, in practical situations, the entries in A and B are parameters describing the system that can not be exactly known, they are only known within approximation by some error of measurement. Only zero entries are

known with 100 percent precision, since they usually indicate a lack of interaction between specific components of the system. For each nonzero entry, we assume that it can vary freely, it is a ‘free parameter’. For each zero entry, it is a ‘fixed zero’. Thus, we have the following definition.

Definition 3.7. M is called a *structured matrix*, if its entries are either fixed zeros or independent free parameters. \tilde{M} is called an *admissible matrix* of M if it can be obtained by fixing the free parameters of M at some particular real values.

Note that, for each admissible matrix, it is a constant matrix since all entries are now specified. The definition of structural controllability requires us to assume that the pair (A, B) is structured, that is, both A and B are structured matrices.

Definition 3.8. Let A, B be structured matrices. Then the pair (A, B) is called *structurally controllable* if there exists an admissible pair (\tilde{A}, \tilde{B}) that is controllable.

If the pair (A, B) is structured, we call the linear time-invariant system (3.3) a *linear structured system*. Therefore, structural controllability is a property of a linear structured system. From now on, we assume that both A and B are structured matrices.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ 0 \\ 0 \end{bmatrix}$$

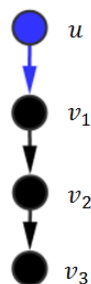


Figure 3.1: The graph of a single-input structured pair (A, b) .

For a single-input structured pair (A, b) as in (3.4), we first define the graph of (A, b) , and denote this graph by $G(A, b)$. Given $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^{n \times 1}$, $G(A, b)$ is

constructed in this way: the node set of $G(A, b)$ contains $n + 1$ nodes, v_1, v_2, \dots, v_n, u , $\{v_1, v_2, \dots, v_n\}$ are called the *state nodes*, u is called the *input node*; the edge set of $G(A, b)$ is obtained by the rule that, there is a directed edge from state node v_j to state node v_i if and only if the corresponding entry a_{ij} in A is a free parameter and there is a directed edge from the input node u to state node v_i if and only if the corresponding entry b_i in b is a free parameter. We call u the *origin* of the graph $G(A, b)$. We give some illustration for the graph of (A, b) in Figure 3.1.

Definition 3.9. A *stem* in $G(A, b)$ is a directed path starting from the input node. The initial (terminal) node of a stem is called the root (top) of the stem.

Definition 3.10. A *bud* is a directed cycle with an extra edge (called *distinguished edge*) that ends, but not begins, in a node in the cycle. The initial node of the distinguished edge is called the *origin* of the bud.

Definition 3.11. A state node v_i is called *non-accessible* if there is no stem that contains v_i . Otherwise, it is called accessible.

Definition 3.12. The graph $G(A, b)$ contains a *dilation* if there is a subset S of state nodes such that $|S| > |T(S)|$. $T(S)$ is defined to be the set of nodes w_j (can be either a state node or input node) with the property that there is a directed edge from w_j to a node in S .

Definition 3.13. A *cactus* is a digraph defined recursively as follows. First, a stem is a cactus. Given a stem S_0 and a bud B_1 , then $S_0 \cup B_1$ is a cactus if the origin of B_1 is the initial node of a directed edge in S_0 and it is the only node that belongs to both S_0 and B_1 . If $S_0 \cup B_1 \cup \dots \cup B_{k-1}$ is a cactus for some positive integer k , then $S_0 \cup B_1 \cup \dots \cup B_{k-1} \cup B_k$ is a cactus provided that B_k is a bud, the origin of B_k is the initial node of a directed edge in $S_0 \cup B_1 \cup \dots \cup B_{k-1}$ and it is the only node that belongs to both B_k and $S_0 \cup B_1 \cup \dots \cup B_{k-1}$. Moreover, a set of node-disjoint cactus is called a *acti*.

In Figure 3.2, we give an illustration of a *acti*.

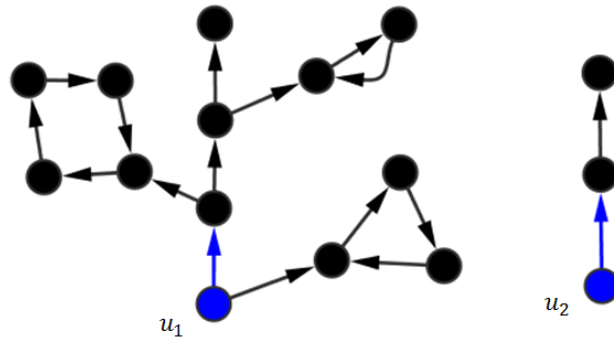


Figure 3.2: The graph structure of a cacti. The cacti contains two (node-disjoint) cactus. The left cactus contains one stem and three buds. The right cactus is just a stem.

Lin's theorem (Theorem 3.9) gives a graph-theoretic characterization for structural controllability of a single-input linear structured system (3.4) and it will be extended to general (multi-input) case later.

Definition 3.14. Let H be a subgraph of a graph G , then G is *spanned by* H , if H has the same node set as G . If G is spanned by H , then H is called a *spanning subgraph* of G .

Theorem 3.9. (Lin's theorem [55]) For a linear structured system (3.4), the following statements are equivalent:

1. The pair (A, b) is structurally controllable.
2. The graph of (A, b) contains no non-accessible nodes and no dilation.
3. The graph of (A, b) is spanned by a cactus.

Proof. The proof given below is adapted from Lin's paper on structural controllability [55]. We first prove Lemma 1 and then show $3 \Rightarrow 1$.

Lemma 1: Suppose G is the graph of a structurally controllable pair. Let Γ be a bud with the origin e , and suppose e is the only node that belongs to both the node

set of G and the node set of Γ . Then $G \cup \Gamma$ is the graph of a structurally controllable pair.

Proof. (1) First consider the case in which G and Γ have their origins as the same node (i.e., the input node). Suppose G is the graph of a structurally controllable pair (A, b) , and Γ is a bud of a structured pair (A_1, b_1) . Then, with some proper ordering of the nodes, $G \cup \Gamma$ is the graph of the pair

$$\left(\begin{bmatrix} A & 0 \\ 0 & A_1 \end{bmatrix}, \begin{bmatrix} b \\ b_1 \end{bmatrix} \right). \quad (3.5)$$

Since Γ is a bud, A_1 and b_1 have the form:

$$A_1 = \begin{pmatrix} 0 & \alpha_1 & 0 & \cdot & 0 \\ 0 & 0 & \alpha_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \alpha_{n-1} \\ \alpha_{n+1} & 0 & 0 & \cdot & 0 \end{pmatrix}, \quad b_1 = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ 0 \\ \alpha_n \end{pmatrix}. \quad (3.6)$$

It is easy to verify that the pair (A_1, b_1) is structurally controllable, based on Kalman's rank condition.

Since (A, b) is structurally controllable, there exists an admissible pair (\tilde{A}, \tilde{b}) that is controllable in the sense that Kalman's rank condition is satisfied. It is possible to find an admissible pair of (A_1, b_1) , $(\tilde{A}_1, \tilde{b}_1)$, such that \tilde{A}_1 and \tilde{A} have no common eigenvalue.

It only remains for us to prove that the pair

$$\left[\begin{array}{c|c} \tilde{A} & 0 \\ \hline 0 & \tilde{A}_1 \end{array} \right], \quad \begin{bmatrix} \tilde{b} \\ \tilde{b}_1 \end{bmatrix} \quad (3.7)$$

is controllable.

Recall that a pair (A, b) is controllable if and if the relation $c^T A = \alpha c^T$ (α is a complex number, $c \neq 0$) always implies $c^T b \neq 0$ (see part 6 of Theorem 3.8). Thus,

suppose (3.7) is not controllable, then there exist c_1 and c_2 , not both zeros, and α , such that

$$[c_1^T \quad c_2^T] \begin{bmatrix} \tilde{A} & 0 \\ 0 & \tilde{A}_1 \end{bmatrix} = \alpha [c_1^T \quad c_2^T] \quad \text{and} \quad [c_1^T \quad c_2^T] \begin{bmatrix} \tilde{b} \\ \tilde{b}_1 \end{bmatrix} = 0.$$

Then, we have

$$c_1^T \tilde{A} = \alpha c_1^T, \quad c_2^T \tilde{A}_1 = \alpha c_2^T.$$

Since \tilde{A} and \tilde{A}_1 have no common eigenvalue, it then follows that either $c_1 = 0$, or $c_2 = 0$. Otherwise, c_1, c_2 would be eigenvectors of \tilde{A} and \tilde{A}_1 with respect to the same eigenvalue α , respectively.

Suppose $c_1 \neq 0$ and $c_2 = 0$, then $c_1^T \tilde{A} = \alpha c_1^T, c_1^T \tilde{b} = 0$, this is a contradiction to (\tilde{A}, \tilde{b}) being controllable. Suppose $c_1 = 0$ and $c_2 \neq 0$, similarly, this is another contradiction. Therefore, the structured pair (3.5) is structurally controllable.

(2) Consider the case in which the origin of Γ is not the origin of G . Then $G \cup \Gamma$ is the graph of the pair

$$\begin{bmatrix} A & 0 \\ b_1 q^T & A_1 \end{bmatrix}, \quad \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{3.8}$$

where q^T is a nonzero vector such that $b_1 q^T$ has one and only one nonzero entry, and the column of the nonzero entry corresponds to the only node that belongs to both the node set of G and the node set of Γ .

We can choose an admissible pair of $(A, b), (\tilde{A}, \tilde{b})$. Furthermore, choose an admissible matrix of A_1, \tilde{A}_1 , such that \tilde{A}_1 and \tilde{A} have no common eigenvalue.

It remains to prove that the pair

$$\begin{bmatrix} \tilde{A} & 0 \\ b_1 q^T & \tilde{A}_1 \end{bmatrix}, \quad \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \tag{3.9}$$

is controllable.

If (3.9) is not controllable, then there exist c_1 and c_2 , not both zeros, and α , such that

$$c_1^T \tilde{A} + c_2^T b_1 q^T = \alpha c_1^T, \quad c_2^T \tilde{A}_1 = \alpha c_2^T, \quad c_1^T \tilde{b} = 0.$$

Moreover, we have $c_2^T b_1 \neq 0$ and $\det(\alpha I - \tilde{A}) \neq 0$.

Indeed, suppose $c_2^T b_1 = 0$, then it follows that $c_1^T \tilde{A} = \alpha c_1^T$, $c_1^T \tilde{b} = 0$. This indicates $c_1 = 0$, since (\tilde{A}, \tilde{b}) is controllable. Similarly, $c_2^T A_1 = \alpha c_2^T$, $c_2^T b_1 = 0$ implies $c_2 = 0$. A contradiction.

Next, suppose $\det(\alpha I - \tilde{A}) = 0$, then α is an eigenvalue of \tilde{A} , but $c_2^T \tilde{A}_1 = \alpha c_2^T$ implies that α is also an eigenvalue of \tilde{A}_1 , a contradiction.

From $c_1^T \tilde{A} + c_2^T b_1 q^T = \alpha c_1^T$, we have

$$c_1^T (\alpha I - \tilde{A}) = c_2^T b_1 q^T.$$

Right multiplying both sides of the equation by $(\alpha I - \tilde{A})^{-1} \tilde{b}$, we get

$$c_1^T \tilde{b} = c_2^T b_1 q^T (\alpha I - \tilde{A})^{-1} \tilde{b} = 0 \quad \Rightarrow \quad q^T (\alpha I - \tilde{A})^{-1} \tilde{b} = 0.$$

Additionally, observe that q^T must have the form

$$q^T = (0, \dots, \delta, \dots, 0), \quad \delta \neq 0.$$

Assume that the only nonzero entry δ is at the i th position. Now, if necessary, modify the nonzero entries of (\tilde{A}, \tilde{b}) such that $(\alpha I - \tilde{A})^{-1} \tilde{b}$ is a column vector with its i th entry nonzero. Then it leads to $\delta = 0$, i.e., $q^T = 0$, a contradiction. Therefore, the structured pair (3.8) is structurally controllable.

Thus, we complete the proof for Lemma 1. □

3 \Rightarrow 1:

Having proved Lemma 1, we continue to prove Lin's theorem. According to Lemma 1, and by the construction process of a cactus, it is easy to know that a cactus has a structurally controllable pair. Therefore, if a graph is spanned by a cactus, then its corresponding pair is structurally controllable.

1 \Rightarrow 2:

Suppose first the graph has a non-accessible node, by a simple reordering of the nodes in the graph, the corresponding pair will be of the form

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ b_2 \end{bmatrix}$$

where $A_{11} \in \mathbb{R}^{k \times k}$, $A_{21} \in \mathbb{R}^{(n-k) \times k}$, $A_{22} \in \mathbb{R}^{(n-k) \times (n-k)}$, $b_2 \in \mathbb{R}^{(n-k) \times 1}$.

It is easy to verify that the controllability matrix $[b \quad Ab \quad \dots \quad A^{n-1}b]$ has generic rank less than n , therefore, the pair (A, b) is not structurally controllable.

Now suppose the graph contains a dilation. By a simple reordering of the nodes, it is possible to make the corresponding pair (A, b) satisfy

$$[A \quad b] = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

where $[A \quad b]$ is the $n \times (n + 1)$ (augmented) matrix formed by putting together A and b , $P_2 \in \mathbb{R}^{(n-k) \times (n+1)}$, and P_1 is a $k \times (n + 1)$ matrix with no more than $k - 1$ nonzero columns. In this case, the generic rank of $[A \quad b]$ is less than n , it then follows that $[b \quad Ab \quad \dots \quad A^{n-1}b]$ is not of full rank generically, thus, the pair (A, b) is not structurally controllable.

Therefore, if the structured pair (A, b) is structurally controllable, then the graph $G(A, b)$ contains no non-accessible nodes and no dilation.

2 \Rightarrow 3:

First, assume that G is the graph of a pair (A, b) which satisfies the following conditions:

- (a) There are no non-accessible nodes in G .
- (b) There is no dilation in G .
- (c) G is minimal.

We will show that G must be a cactus. First, we prove several lemmas.

Lemma 2: Every node in G is accessible from the origin along one and only one simple path.

Proof. Otherwise, suppose a node α can be reached from the origin along two distinct paths p_1 and p_2 , let β and $\tilde{\beta}$ be the last nodes met before α on p_1 and p_2 , respectively. One may assume $\beta \neq \tilde{\beta}$, since the two paths are different. After deleting the edge (β, α) , we obtain a new graph G_1 . By condition (c), G_1 must have either a non-accessible node or a dilation. But it is impossible for G_1 to have a non-accessible node, since α can be reached along an alternate path p_2 while all the other nodes are still accessible. Therefore, G_1 must contain a dilation.

So, there exists a set S of k nodes such that $T_1(S)$ (the set of nodes in G_1 that form an edge pointing to nodes in S) contains no more than $k - 1$ nodes. Let $S = \{\alpha_1, \alpha_2, \dots, \alpha_k\}$ and $T_1(S) = \{\beta_1, \beta_2, \dots, \beta_l\}$ with $l \leq k - 1$. Then it is easy to see that $\alpha \in S$, $\tilde{\beta} \in T_1(S)$, since there is only one edge (β, α) being deleted (affected) in G_1 , and G itself has no dilation. We may take $\alpha = \alpha_1$, moreover, $\beta \notin T_1(S)$, since if $\beta \in T_1(S)$, then $T_1(S) = T(S)$ ($T(S)$ is the set of nodes in G that form an edge pointing to nodes in S) and as a result, G will have a dilation, a contradiction.

Similarly, in the graph G_2 obtained by deleting the edge $(\tilde{\beta}, \alpha)$, there exists a set \tilde{S} of \tilde{k} nodes such that $T_2(\tilde{S})$ (the set of nodes in G_2 that form an edge pointing to nodes in \tilde{S}) contains no more than $\tilde{k} - 1$ nodes. Let $\tilde{S} = \{\tilde{\alpha}_1, \tilde{\alpha}_2, \dots, \tilde{\alpha}_{\tilde{k}}\}$ and $T_2(\tilde{S}) = \{\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_{\tilde{l}}\}$ with $\tilde{l} \leq \tilde{k} - 1$. Then as before, we have $\alpha \in \tilde{S}$, $\beta \in T_2(\tilde{S})$, $\tilde{\beta} \notin T_2(\tilde{S})$, with $\alpha = \tilde{\alpha}_1$.

Define $\hat{S} = S \cup \tilde{S}$, $T(\hat{S})$ is the set of nodes in G that form an edge pointing to nodes in \hat{S} . Then $T(\hat{S}) = (T_1(S) \cup \{\beta\}) \cup (T_2(\tilde{S}) \cup \{\tilde{\beta}\}) = T_1(S) \cup T_2(\tilde{S})$. If M is a set of nodes, denote by $N(M)$ the number of distinct nodes in M .

Suppose S and \tilde{S} have only one node in common, i.e. $\alpha = \alpha_1 = \tilde{\alpha}_1$, then

$N(\hat{S}) = N(S) + N(\tilde{S}) - 1 = k + \tilde{k} - 1$. And we have

$$\begin{aligned}
N(T(\hat{S})) &= N(T_1(S) \cup T_2(\tilde{S})) \\
&\leq N(T_1(S)) + N(T_2(\tilde{S})) \\
&= l + \tilde{l} \\
&\leq (k - 1) + (\tilde{k} - 1) \\
&= k + \tilde{k} - 2.
\end{aligned} \tag{3.10}$$

This indicates that G has a dilation, which is a contradiction.

Now consider the case in which S and \tilde{S} have other common nodes other than α . Suppose there are j common nodes, and $\alpha_i = \tilde{\alpha}_i$, for $i = 1, 2, \dots, j$, $1 \leq j \leq \min(k, \tilde{k})$. Define $S_0 = \{\alpha_2, \alpha_3, \dots, \alpha_j\}$ and consider the corresponding $T(S_0)$ in G .

Then we claim:

- (i) $\beta \notin T(S_0)$
- (ii) $\tilde{\beta} \notin T(S_0)$
- (iii) $T(S_0) \subseteq (T_1(S) \cap T_2(\tilde{S}))$
- (iv) $N(T(S_0)) \geq j - 1$

To prove (i), suppose $\beta \in T(S_0)$, since (β, α_1) has been deleted in G_1 , $\alpha_1 \notin S_0$, we have $\beta \in T_1(S)$, and this is impossible. The proof for (ii) is similar to that for (i) as above. To prove (iii), we can see that for each $v \in T(S_0)$, it is obvious that $v \in T_1(S) \cup \{\beta\} = T(S)$, and similarly, $v \in T_2(\tilde{S}) \cup \{\tilde{\beta}\} = T(\tilde{S})$. But $v \neq \beta$ and $v \neq \tilde{\beta}$, so $v \in T_1(S)$ and $v \in T_2(\tilde{S})$. To prove (iv), suppose, otherwise, $N(T(S_0)) < j - 1$, since $N(S_0) = j - 1$, then there will be a dilation in G , contradiction.

Now, using the above results, we get

$$\begin{aligned}
N(T(\hat{S})) &= N(T_1(S) \cup T_2(\tilde{S})) \\
&= N(T_1(S)) + N(T_2(\tilde{S})) - N(T_1(S) \cap T_2(\tilde{S})) \\
&\leq N(T_1(S)) + N(T_2(\tilde{S})) - N(T(S_0)) \\
&\leq l + \tilde{l} - (j - 1) \\
&\leq (k - 1) + (\tilde{k} - 1) - (j - 1) \\
&= k + \tilde{k} - j - 1.
\end{aligned} \tag{3.11}$$

However, we have

$$N(\hat{S}) = j + (k - j) + (\tilde{k} - j) = k + \tilde{k} - j.$$

Thus, $|\hat{S}| > |T(\hat{S})|$ and G has a dilation, a contradiction.

The proof for Lemma 2 is complete. □

Next, denote the origin of the graph G by e , let $\tau_1, \tau_2, \dots, \tau_r$ be directed edges in G that start from e . Define V_i to be the set of all nodes which can be reached from e by passing through the edge τ_i , $1 \leq i \leq r$. Based on Lemma 2, all these sets V_i are (pairwise) disjoint. It follows that $V = V_1 \cup \dots \cup V_r \cup \{e\}$.

Now denote by G_i the subgraph of G whose node set is $V_i \cup \{e\}$ and whose edge set contains all the edges (α, β) from G with $\alpha \in V_i \cup \{e\}$, $\beta \in V_i$. We call G_i , defined as above, *bunches*. A subgraph G_i is called a *terminal bunch*, if there exists a subset $S \subset V_i$ such that $N(T(S)) = N(S)$ and $T(S)$ contains the origin of G .

Lemma 3: In a terminal bunch, The set S contains one and only one final node v such that $v \in S$, and $v \notin T(S)$.

Proof. Otherwise, first suppose there is no final node, then $S \subset T(S)$, but $T(S)$ contains e which doesn't belong to S , so, $N(T(S)) > N(S)$, a contradiction. Next suppose there are more than one final nodes, let v_1, v_2, \dots, v_k be the final nodes, $k > 1$. It follows that $N(T(S)) = N(S) + 1 - k < N(S)$, another contradiction.

Therefore, Lemma 3 is proved. □

Lemma 4: There exists at most one terminal bunch in G .

Proof. Otherwise, suppose there are more than one terminal bunches. Assume that G_1 and G_2 are two terminal bunches in G . Then there exists $S_i \subset V_i$ such that $N(T(S_i)) = N(S_i)$ and $T(S_i)$ contains the origin of G , for $i = 1, 2$.

Define $S = S_1 \cup S_2$, one obtains

$$\begin{aligned}
 N(T(S)) &= N(T(S_1)) + N(T(S_2)) - 1 \\
 &= N(S_1) + N(S_2) - 1 \\
 &= N(S) - 1.
 \end{aligned}
 \tag{3.12}$$

This implies that there is a dilation in G , a contradiction.

Thus, we have shown Lemma 4. □

Now, We give a definition for pre-cactus. A graph \bar{H} is called a *pre-cactus*, if one can write $\bar{H} = B_1 \cup B_2 \cup \dots \cup B_r$, where each B_i is a bud, such that for every $j = 2, 3, \dots, r$, the origin e_j of B_j is the initial node of one directed edge in the graph of $B_1 \cup \dots \cup B_{j-1}$. Moreover, e_j is the only node that belongs to both B_j and $B_1 \cup \dots \cup B_{j-1}$.

The following lemma gives the relation between a pre-cactus and a cactus.

Lemma 5: Every pre-cactus becomes a cactus after eliminating one or more suitable edges.

Proof. By definition, a pre-cactus \bar{H} can be written as $\bar{H} = B_1 \cup B_2 \cup \dots \cup B_r$. Let e_1 be the origin of B_1 and (e_1, β_1) be the distinguished edge of B_1 . Since B_1 is a bud, one can find another node e_2 such that the edge (e_2, β_1) exists in B_1 . Remove the edge (e_2, β_1) , one obtains a stem S_1 , and $B_1 = S_1 \cup (e_2, \beta_1)$. If e_2 is not the origin of any other bud, the lemma is immediate, since at this step we already had a cactus $S_1 \cup B_2 \cup \dots \cup B_r$. If not, e_2 is the origin of another bud, say B_2 , then the same procedure produces a new node e_3 and the relation $B_1 \cup B_2 = S_2 \cup (e_2, \beta_1) \cup (e_3, \beta_2)$, where S_2 is a stem.

After applying the same procedure t (finite) times, one will eventually obtain a node e_{t+1} , $t \leq r$, such that e_{t+1} is not the origin of any bud from $\{B_{t+1}, \dots, B_r\}$. Then we have $B_1 \cup \dots \cup B_t = S_t \cup (e_2, \beta_1) \cup \dots \cup (e_{t+1}, \beta_t)$, where S_t is a stem. Let $P = S_t \cup B_{t+1} \cup \dots \cup B_r$, then P is a cactus.

This completes the proof for Lemma 5.

□

Lemma 6: Any non-terminal bunch is spanned by a pre-cactus.

Proof. Assume G_1 is a non-terminal bunch, let β_1 be the terminal node of the edge τ_1 , which is defined as the edge whose initial node is the origin e of G . Then there must be another edge entering β_1 from a node ε_1 , otherwise, take $S = \{\beta_1\}$, then $T(S) = \{e\}$. This shows G_1 is a terminal bunch, contradictory to the assumption. One might have $\varepsilon_1 = \beta_1$, but this is a trivial case. Let us assume that $\varepsilon_1 \neq \beta_1$. Since there is only one simple path reaching ε_1 from e , but this path must go through β_1 , one is then able to find a cycle containing both β_1 and ε_1 . This cycle together with the distinguished edge (e, β_1) forms a bud B_1 .

If the node set of B_1 is the same as the node set of G_1 , then G_1 is spanned by a pre-cactus B_1 . If not, we take a new node q from G_1 , then there is a unique simple path from e to q . Let e_2 be the last node in the node set of B_1 which is met along this path. Let β_2 be the first node which is met after e_2 along this path. The same procedure results in a new bud B_2 , moreover, e_2 is the origin of B_2 and is the only node that belongs to both the node set of B_1 and the node set of B_2 . This procedure can only go on finite times until we can find buds B_1, \dots, B_r , such that the union of node sets of all the buds covers all the nodes in G_1 . So, $\bar{H} = B_1 \cup B_2 \cup \dots \cup B_r$ is a pre-cactus that spans G_1 .

The proof for Lemma 6 is complete.

□

Lemma 7: There always exists a terminal bunch in G .

Proof. Suppose there are no terminal bunches in G . Since all the non-terminal bunches in G are pairwise edge-disjoint, with only one common node e (the origin), by Lemma 6, G itself is spanned by a pre-cactus. But based on Lemma 5, G doesn't satisfy the minimality condition (condition (c)), a contradiction.

Therefore, Lemma 7 is true. □

Lemma 8: Any terminal bunch is spanned by a cactus.

Proof. Within the terminal bunch G_1 , there is a final node v . Then there exists a unique simple path from the origin to v , which is a stem denoted by S_0 . If all the nodes of G_1 belong to S_0 , the lemma is true. Otherwise, choose a new node q in G_1 . Note that q is not a final node, because there is only one final node in a terminal bunch. Now there is one unique simple path from the origin to q , let e_1 be the last node in the node set of S_0 met along this path, let β_1 be the first node after e_1 met along this path. Using the same arguments as in Lemma 6, one can obtain a finite set of buds, B_1, B_2, \dots, B_r , and let $P = S_0 \cup B_1 \cup \dots \cup B_r$, then P is a cactus that spans G_1 .

Therefore, the proof for Lemma 8 is complete. □

Based on these lemmas above, we return to the 3 assumptions of G and observe that G satisfying conditions (a), (b) and (c) must have the following properties:

- (1) The graph G has one and only one terminal bunch, which is itself a cactus.
- (2) If G has a non-terminal bunch, then each non-terminal bunch must be a pre-cactus.

A graph with properties (1) and (2) must be a cactus. Therefore, G is a cactus.

Finally, a graph that has no non-accessible nodes and no dilation (without the minimality condition (c)) must be spanned by a cactus. This concludes the proof for the part 2 \Rightarrow 3.

Therefore, the proof for Lin's theorem is complete. □

Lin's theorem [55] was originally stated as in Theorem 3.9. It only gives a graph-theoretic characterization for structural controllability of a single-input linear

structured system (3.4), while the extension of Lin's theorem to the general multi-input case (3.3) is also true, which is called Lin's structural controllability theorem.

Generally, for a multi-input linear structured system (A, B) , we define the graph of (A, B) , denote this graph by $G(A, B)$. Given $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$, $G(A, B)$ is constructed in this way: the node set of $G(A, B)$ contains $n + m$ nodes, $v_1, v_2, \dots, v_n, u_1, u_2, \dots, u_m$, $\{v_1, v_2, \dots, v_n\}$ are the state nodes, $\{u_1, u_2, \dots, u_m\}$ are the input nodes; the edge set of $G(A, B)$ is obtained by the rule that, there is a directed edge from state node v_j to state node v_i if and only if the corresponding entry a_{ij} in A is a free parameter and there is a directed edge from input node u_j to state node v_i if and only if the corresponding entry b_{ij} in B is a free parameter. In Figure 3.3, we give an example of the graph of (A, B) .

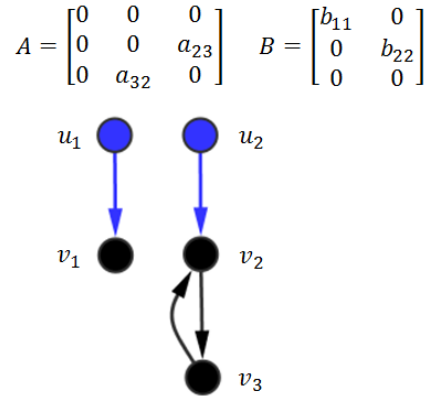


Figure 3.3: The graph of a multi-input structured pair (A, B) .

Theorem 3.10. (Lin's structural controllability theorem) [62] For a linear structured system (3.3), the following statements are equivalent:

1. The pair (A, B) is structurally controllable.
2. The graph of (A, B) contains no non-accessible nodes and no dilation.
3. The graph of (A, B) is spanned by a cacti.

Lin's structural controllability theorem (Theorem 3.10) shows that a cacti is the minimal network structure for which the corresponding linear structured system is structurally controllable.

3.4.2 Algebraic characterization

Lin's structural controllability theorem (Theorem 3.10) provides a graph-theoretic characterization for structural controllability, and the graph-theoretic conditions of accessibility and dilation can be efficiently examined by using tools in graph theory. In this subsection, we give other equivalent algebraic conditions for structural controllability, based on generic analysis of the structured pair (A, B) .

Structural controllability is a generic property [21, 86]. In other words, for a structured pair (A, B) that is structurally controllable, almost all admissible pairs will be controllable except for those that are in a proper algebraic variety in the parameter space. These concepts can be explained as follows.

Let A, B be structured matrices of dimension $n \times n$ and $n \times m$, respectively. Let N be the total number of free parameters in (A, B) , one can use a vector $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_N) \in \mathbb{R}^N$ to represent the N independent free parameters in (A, B) , each free parameter being a variable λ_i that takes value in \mathbb{R} . Then, each admissible pair of (A, B) can be identified with a point $c = (c_1, c_2, \dots, c_N)$ in the parameter space \mathbb{R}^N , after choosing a specific real-valued vector c for λ . Denote the ring of polynomials in N variables $\lambda_1, \lambda_2, \dots, \lambda_N$ by $R[\lambda]$. Then A, B can be viewed as matrices with entries in $R[\lambda]$.

A *property* Π is a function from the parameter space \mathbb{R}^N to $\{0, 1\}$, where $\Pi(c) = 1$ when the property is true for $c \in \mathbb{R}^N$ and $\Pi(c) = 0$ when the property is false for $c \in \mathbb{R}^N$. Therefore, structural controllability is a property Π stated about the structured pair (A, B) : for any point $c \in \mathbb{R}^N$, $\Pi(c) = 1$ if and only if the admissible pair of (A, B) identified with c is controllable. Consider a finite number of polynomials $p_i \in R[\lambda]$, $i = 1, 2, \dots, k$, a *variety* $V \subseteq \mathbb{R}^N$ is the set of common

zeros of polynomials p_1, p_2, \dots, p_k in the polynomial ring $R[\lambda]$. A *proper variety* V is a variety that is a proper subset of \mathbb{R}^N . A property Π is *generic* if $\text{Ker}(\Pi) \subseteq V$ for some proper variety $V \subset \mathbb{R}^N$. Note that, a proper variety is some hypersurface in \mathbb{R}^N and will be of Lebesgue measure zero, that is, a proper variety is a somewhat ‘negligible’ subset.

As an example, the maximal rank of a structured matrix M (i.e., the maximal rank that any admissible matrix of M can attain) is a generic property. Specifically, almost all admissible matrices of M have a maximal rank r , those admissible matrices that have rank less than r are ‘negligible’, lying within a proper variety. The maximal rank of a structured matrix M is called the *generic rank* of M , denoted by $\text{Rank}_g(M)$.

Theorem 3.11. [86] *Let A, B be structured matrices of dimension $n \times n$ and $n \times m$, respectively. (A, B) is a structured pair associated with a parameter space \mathbb{R}^N , where N is the total number of independent free parameters in (A, B) . Then (A, B) is structurally controllable if and only if all admissible pairs of (A, B) that are uncontrollable lie on a proper variety $V \subset \mathbb{R}^N$.*

Definition 3.15. A structured pair (A, B) is said to be in *form I* (or reducible), if there exists a permutation matrix P such that

$$PAP^{-1} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

and

$$PB = \begin{bmatrix} 0 \\ B_2 \end{bmatrix},$$

where A_{11} is of order $q \times q$, B_2 of order $(n - q) \times m$.

Definition 3.16. A structured pair (A, B) is said to be in *form II*, if the $n \times (n + m)$ structured matrix $[A \ B]$ is not of generic full rank, that is,

$$\text{Rank}_g([A \ B]) < n.$$

Lin's structural controllability theorem (Theorem 3.10) has an equivalent algebraic statement, as in Theorem 3.12. This equivalency can be seen, by considering that the graph-theoretic interpretation of (A, B) in form I is that $G(A, B)$ has non-accessible nodes, whereas (A, B) in form II amounts to that $G(A, B)$ contains a dilation [62].

Theorem 3.12. [31] *The structured pair (A, B) is structurally controllable if and only if (A, B) is neither in form I nor in form II.*

Controllability of biological networks

Despite the considerable advancement on the study of the topological characteristics and dynamics of complex networks during the past decades, the ultimate proof of our understanding of natural or technological systems is reflected in our ability to control them. Given the complexity of the structure and time-dependent dynamics of complex networks, our understanding of how to control a network is still limited.

Structural controllability of linear structured systems could serve as a good starting point in order to define a framework for network control [56]. This usefulness could be rationalized based on: (1) complex networks can be seen as structured, since only the network topology is most relevant and edge weights are not accessible in many situations; (2) nonlinear dynamics have similar structural controllability as the linearized dynamics [87], thus understanding network control under linear dynamics is essential for addressing control of networks with non-linear dynamics.

In this chapter, we will mainly introduce network controllability, and study the network controllability of biological networks.

4.1 Definition of network controllability

To define network controllability, we start with a directed network $G = (V(G), A(G))$. Our goal is to (fully) control the whole network G by introducing a set of inputs (or input nodes) into the network. The way of connecting the inputs to the nodes in G is called a *control configuration*. Given a control configuration, the network G corresponds to a linear structured system defined as follows.

Suppose $V(G) = \{v_1, v_2, \dots, v_n\}$, and the set of inputs is denoted by $U = \{u_1, u_2, \dots, u_m\}$. Then, given a control configuration, we obtain a linear structured system, as in (3.3), where $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ is a structured matrix defined as

$$a_{ij} = \begin{cases} \text{a free paramter,} & \text{if } (v_j \rightarrow v_i) \in A(G), \\ \text{zero,} & \text{if } (v_j \rightarrow v_i) \notin A(G). \end{cases}$$

$B = (b_{ij}) \in \mathbb{R}^{n \times m}$ is a structured matrix defined as

$$b_{ij} = \begin{cases} \text{a free paramter,} & \text{if } u_j \text{ is connected to } v_i, \\ \text{zero,} & \text{if } u_j \text{ is not connected to } v_i. \end{cases}$$

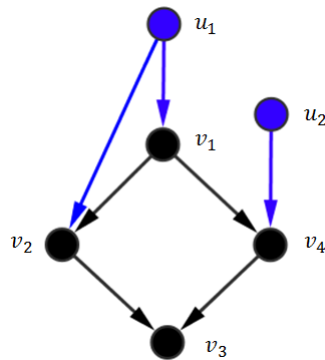


Figure 4.1: A control configuration for a directed network.

We give an example of a control configuration for a directed network in Figure 4.1. For this control configuration, the network corresponds to a linear structured system:

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ 0 & a_{32} & 0 & a_{34} \\ a_{41} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{pmatrix} + \begin{pmatrix} b_{11} & 0 \\ b_{21} & 0 \\ 0 & 0 \\ 0 & b_{42} \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}.$$

Network controllability concerns about the minimum number of inputs required to control the network G , equivalently, a control configuration with the minimum number of inputs such that the corresponding linear structured system (3.3) is structurally controllable.

The problem of finding the minimum number of inputs to control a directed network G is an optimization problem, which can be formulated as follows.

Suppose $V(G) = \{v_1, v_2, \dots, v_n\}$. Let $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$ be a column vector of n state functions. Let $u(t) = (u_1(t), u_2(t), \dots, u_n(t))^T$ be a column vector of n input functions. Let $E = (e_{ij}) \in \mathbb{R}^{n \times n}$ be a binary diagonal matrix, that is, $e_{ii} \in \{0, 1\}$, $i = 1, 2, \dots, n$.

Let $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ be a structured matrix in which each entry is a free parameter. Let $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ be a structured matrix defined as

$$a_{ij} = \begin{cases} \text{a free paramter,} & \text{if } (v_j \rightarrow v_i) \in A(G), \\ \text{zero,} & \text{if } (v_j \rightarrow v_i) \notin A(G). \end{cases}$$

Define the structured matrix $B = CE \in \mathbb{R}^{n \times n}$. The pair (A, B) represents a linear structured system of the form (3.3).

Then, *the network controllability of G* is defined as

$$N_D = \min_E \left\{ \sum_{i=1}^n e_{ii} : (A, B) \text{ is structurally controllable} \right\}. \quad (4.1)$$

The network controllability of G , denoted by N_D , is equal to the minimum number of inputs required to control the whole network G .

4.2 A theorem on network controllability

The network controllability of a directed network G only depends on the network topology of G . The minimum number of inputs can be determined by a maximum matching in the directed network G . Let us first give the definition for matching and maximum matching in a directed network. In Figure 4.2, we give an illustration of a maximum matching in a directed network.

Definition 4.1. For a directed graph, a *matching* M is a set of directed edges in which any two edges have different initial nodes and different terminal nodes. A node is *matched* if it is the terminal node of an edge in M , otherwise, it is *unmatched* (exposed). A *maximum matching* M^* is a matching with maximum cardinality. A *perfect matching* is a matching in which each node is matched.

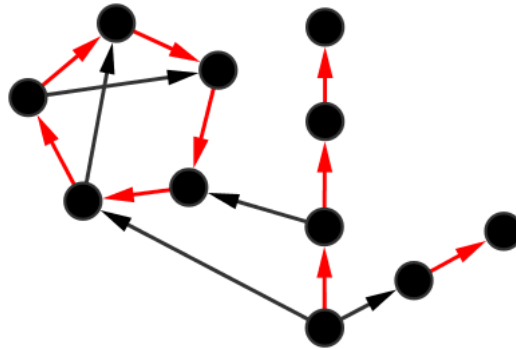


Figure 4.2: A maximum matching in a directed network. The set of arcs in red form a maximum matching in the directed network.

Theorem 4.1. (Minimum input theorem) [56] *The minimum number of inputs, N_D , is one if there is a perfect matching in a directed network G . Otherwise, it equals the number of unmatched nodes with respect to any maximum matching M^* , and to fully control G , each unmatched node needs to be directly connected to a different input node. Therefore,*

$$N_D = \max \{1, n - |M^*|\}. \quad (4.2)$$

Proof. Denote by M^* a maximum matching in G .

If G has a perfect matching, i.e., $|M^*| = n$, where n is the number of nodes in G , then G can be spanned by a set of disjoint (independent) cycles, then we just need one input to control G . Denote a single input node by u , and the set of independent cycles that spans G by $C_1 \cup \dots \cup C_k$, for each independent cycle C_i , we connect u to a node in C_i . These operations result in a new graph, which is spanned by a cacti. By Lin's structural controllability theorem (Theorem 3.10), the corresponding linear structured system of this graph is structurally controllable. So, $N_D = 1$.

If there is no perfect matching in G , i.e., $|M^*| < n$, then within a particular maximum matching, the matching edges will form simple paths and cycles, we call them *matching paths* and *matching cycles*. Given the maximum matching M^* , there are $n - |M^*|$ unmatched nodes. We can connect a different input node to each unmatched node, to form $n - |M^*|$ stems, thus, $n - |M^*|$ input nodes are needed. All the other nodes in G are covered by matching cycles. For each given matching cycle C , we connect an existing input node to a node in the matching cycle C , to form a bud. Therefore, we don't need extra input nodes to control the matching cycles. The above operations will turn G to a graph spanned by a cacti. By Lin's structural controllability theorem, the corresponding linear structured system of this graph is structurally controllable. Thus, it is able to control the network G with $n - |M^*|$ inputs, i.e., $N_D \leq n - |M^*|$. To construct a spanning cacti based on a maximum matching is illustrated in Figure 4.3.

On the other hand, suppose we can control the network G by introducing r inputs into the network. Then by Lin's structural controllability theorem, we have a spanning cacti with r inputs (or r stems). Based on this cacti, we can obtain a matching M in the network G as follows. For each stem in the cacti, take all the edges except the first edge in the stem to be in M . For each bud in the cacti, take all the edges in the cycle of this bud to be in M . It is easy to see that the matching M has size $n - r$. Since the maximum size of a matching is $|M^*|$, we have $n - r \leq |M^*|$, hence, $r \geq n - |M^*|$. Thus, the minimum number of inputs is at least $n - |M^*|$, i.e.,

$$N_D \geq n - |M^*|.$$

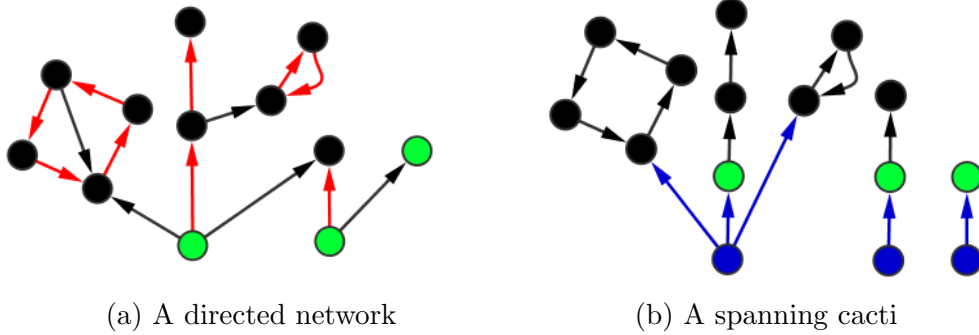


Figure 4.3: An example of constructing a spanning cacti based on a maximum matching in a directed network. In (a), the set of arcs in red form a maximum matching. In (b), the nodes in blue are the input nodes of the cacti. The nodes in green are the unmatched nodes with respect to the maximum matching.

Therefore, $N_D = n - |M^*|$, if $|M^*| < n$. The proof is thus complete. \square

The minimum input theorem provides a way to find a control configuration with the minimum number of inputs to control the network G . Indeed, based on this theorem, a minimum spanning cacti that delivers the full control of a directed network can be constructed by identifying a set of matching paths and matching cycles [2].

To find a maximum matching in a directed network G , one can find a maximum matching in an (undirected) bipartite representation H of G . Suppose $G = G(V, A)$ has its node set $V = \{v_1, v_2, \dots, v_n\}$, then $H = H(V^+ \cup V^-, \Gamma)$ is defined in this way: its node set contains two copies of V , $V^+ = \{v_1^+, \dots, v_n^+\}$, $V^- = \{v_1^-, \dots, v_n^-\}$ and its edge set Γ contains an (undirected) edge $\{v_i^+, v_j^-\}$ if and only if the directed edge $v_i \rightarrow v_j$ is in G .

For the (undirected) bipartite graph H , a maximum matching can be found efficiently by using the well-known Hopcroft-Karp algorithm, which runs in $O(\sqrt{|V|} |A|)$ time [39].

The network controllability of several real-world networks were calculated in [56] (see Table 4.1). In Table 4.1, for each network, n_D is the ratio of the minimum number of inputs to the number of nodes in the network (or the fraction of inputs), that is, $n_D = \frac{N_D}{N}$.

Observing the network controllability of real networks in Table 4.1, some interesting trends could be seen that might contradict with our expectation: gene regulatory networks display high n_D (~ 0.8), whereas we expect that the gene regulatory networks have been evolved to efficiently control the gene expressions for cellular development or cell identity, thus in favor of controllability; social networks, in which the population are commonly perceived to be resistant to control, turn out to be easily controllable given the relatively small n_D .

Liu et al. [57] argued that the unexpected high fraction of inputs in gene regulatory networks might be due to three factors: (1) currently the gene regulatory network maps are largely incomplete, (2) taking gene auto-regulation (self loops) into account will further lower the fraction of inputs, (3) n_D could be further reduced by the nonlinearities in the dynamics, which potentially help the system to explore its state space.

| Type | Name | N | L | n_D |
|----------------------|---------------------------------|--------|---------|-------|
| Regulatory | TRN-Yeast-1 | 4441 | 12873 | 0.965 |
| | TRN-Yeast-2 | 688 | 1079 | 0.821 |
| | TRN-EC-1 | 1550 | 3340 | 0.891 |
| | TRN-EC-2 | 418 | 519 | 0.751 |
| | Ownership-USCorp | 7253 | 6726 | 0.820 |
| Trust | College student | 32 | 96 | 0.188 |
| | Prison inmate | 67 | 182 | 0.134 |
| | WikiVote | 7115 | 103689 | 0.666 |
| | Epinions | 75888 | 508837 | 0.549 |
| Food web | Ythan | 135 | 601 | 0.511 |
| | Little Rock | 183 | 2494 | 0.541 |
| | Grassland | 88 | 137 | 0.523 |
| Power grid | Texas | 4889 | 5855 | 0.325 |
| Metabolic | <i>Escherichia coli</i> | 2275 | 5763 | 0.382 |
| | <i>Saccharomyces cerevisiae</i> | 1511 | 3833 | 0.329 |
| | <i>Caenorhabditis elegans</i> | 1173 | 2864 | 0.302 |
| Electronic circuits | s838 | 512 | 819 | 0.232 |
| Neuronal | <i>Caenorhabditis elegans</i> | 297 | 2345 | 0.165 |
| Citation | ArXiv-HepTh | 27770 | 352807 | 0.216 |
| | ArXiv-HepPh | 34546 | 421578 | 0.232 |
| World Wide Web | nd.edu | 325729 | 1497134 | 0.677 |
| | Political blogs | 1224 | 19025 | 0.356 |
| Internet | p2p-1 | 10876 | 39994 | 0.552 |
| | p2p-2 | 8846 | 31839 | 0.578 |
| Social communication | UCIonline | 1899 | 20296 | 0.323 |
| | Email-epoch | 3188 | 39256 | 0.426 |
| | Cellphone | 36595 | 91826 | 0.204 |
| Intra-organization | Freemans-1 | 34 | 695 | 0.029 |
| | Manufacturing | 77 | 2228 | 0.013 |
| | Consulting | 46 | 879 | 0.043 |

Table 4.1: The network controllability of real-world networks [56]. For each network, the network type and name is given, as well as the number of nodes (N), the number of edges (L), the ratio of the minimum number of inputs to the number of nodes in the network (n_D).

4.3 Controllability of brain networks

We calculated the network controllability of the mouse inter-region brain networks (see Subsection 2.2.1) and the network controllability of the subnetworks induced by the 12 subdivisions in each mouse inter-region brain network, and the results are given in Table 4.2.

As can be seen from Table 4.2, for each P-value cutoff, the corresponding ipsilateral network contains more edges than the corresponding contralateral network. Table 4.2 also shows that the network controllability of each ipsilateral network is smaller than the network controllability of each contralateral network, suggesting that ipsilateral networks need smaller numbers of inputs to control. Considering that the contralateral networks corresponding to P-value cutoffs 0.05, 0.02 and 0.01 each contains even more edges than the ipsilateral network corresponding to P-value cutoff 0.001, the large values of the network controllability of contralateral networks suggest that contralateral networks have very different structures from ipsilateral networks.

To further examine this structural difference, we found that in the contralateral network corresponding to P-value cutoff 0.05, there are 46 sink nodes (much more than 12 sink nodes in the ipsilateral network corresponding to P-value cutoff 0.05). Since each sink node needs a different input to control, thus, we provide an explanation for the observation that contralateral networks are more difficult to control than ipsilateral networks.

This structural difference also indicates that in contralateral networks, there are fewer axonal projections from the downstream brain regions back to those in the upstream, which might have effect on the functions and dynamics of the contralateral connections. For example, with such structures, the contralateral networks are less efficient in global signal processing and integration [84].

The subnetworks induced by Striatum and Thalamus in each of the 10 networks both need a large fraction of inputs to control, suggesting that in mouse inter-region

brain networks, there are only few interconnections among brain regions in Striatum or Thalamus. For example, in the contralateral networks corresponding to P-value cutoffs 0.005 and 0.001, there are no interconnections among the 12 brain regions in Striatum. We found that in the contralateral network corresponding to P-value cutoff 0.05, there are 8 sink nodes that are in Striatum, and 19 sink nodes that are in Thalamus, indicating that Striatum and Thalamus lie mostly at the downstream ends of the contralateral networks, receiving signals from upstream brain regions.

For Isocortex, we found that its induced subnetworks in contralateral networks contain less edges than its induced subnetworks in ipsilateral networks. However, the network controllability of the subnetworks induced by Isocortex in contralateral networks are smaller than those in ipsilateral networks. This may indicate that the brain regions in Isocortex have different structures of interconnections in contralateral networks than in ipsilateral networks.

Moreover, in the ipsilateral network corresponding to each P-value cutoff, each of the subdivisions Hippocampus, Cortical Subplate and Cerebellum spans a subnetwork that needs only one input to control.

Next, we visualize the subnetworks induced by the 12 subdivisions in the ipsilateral network corresponding to P-value cutoff 0.05, in Figure 4.4. In this figure, each node is marked in a color according to the subdivision to which it belongs (see Figure 2.1).

We can see from Figure 4.4, that the subnetworks induced by the subdivisions Striatum and Thalamus are sparse. Thus, a relatively large number of inputs is required to control each of these two subnetworks. There seem to be relatively dense interconnections among brain regions in each of the subdivisions Isocortex, Hippocampus, Hypothalamus and Medulla.

| Network controllability | | | | | | | | | | |
|------------------------------|---------------------|------|------|-------|-------|-----------------------|------|------|-------|-------|
| | ipsilateral network | | | | | contralateral network | | | | |
| P-value cutoff | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| Number of edges | 3123 | 2721 | 2487 | 2278 | 1947 | 2451 | 2160 | 1979 | 1830 | 1587 |
| | 12 | 17 | 18 | 21 | 28 | 46 | 47 | 49 | 51 | 61 |
| Isocortex (38) | 1 | 3 | 6 | 6 | 9 | 1 | 1 | 2 | 3 | 3 |
| Olfactory Areas (11) | 2 | 2 | 3 | 3 | 3 | 6 | 6 | 6 | 6 | 7 |
| Hippocampus (11) | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 4 | 4 | 4 |
| Cortical Subplate (7) | 1 | 1 | 1 | 1 | 1 | 6 | 6 | 6 | 6 | 6 |
| Striatum (12) | 9 | 10 | 10 | 11 | 11 | 10 | 11 | 11 | 12 | 12 |
| Pallidum (8) | 1 | 4 | 4 | 5 | 5 | 4 | 4 | 4 | 4 | 4 |
| Thalamus (35) | 25 | 26 | 27 | 29 | 30 | 25 | 26 | 26 | 26 | 27 |
| Hypothalamus (20) | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 3 | 3 | 3 |
| Midbrain (21) | 2 | 3 | 5 | 5 | 6 | 3 | 3 | 3 | 3 | 4 |
| Pons (13) | 5 | 5 | 5 | 5 | 7 | 3 | 3 | 3 | 3 | 4 |
| Medulla (25) | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 5 |
| Cerebellum (12) | 1 | 1 | 1 | 1 | 1 | 7 | 7 | 7 | 7 | 7 |

Table 4.2: The network controllability of the mouse inter-region brain networks (indicated by numbers in red) and the network controllability of the subnetworks of each mouse inter-region brain network induced by 12 subdivisions. The parentheses after each subdivision indicate the number of regions belonging to that subdivision.

4.4 Controllability of transcriptional regulatory networks

We calculated the network controllability of the 41 human cell-type-specific TF regulatory networks (see Subsection 2.2.3), and the results are given in Table 4.3. In Table 4.3, n_D is the fraction of inputs, that is, $n_D = \frac{N_D}{N}$.

In Table 4.3, we observed a relatively small n_D (< 0.35) for each TF regulatory network. This observation seems to be inconsistent with the previous finding that gene regulatory networks are hard to control, as indicated in Table 4.1. We think that the relative small values of n_D in the human TF regulatory networks might be due to two factors: (1) there are self-loops for TFs in the networks, thus decreasing n_D , and (2) in the TF regulatory networks, there are only a small number of downstream targets (TFs with out-degree zero), whereas in gene regulatory networks, there are a large number of downstream target genes.

Table 4.3 shows that among the 41 networks, the TF regulatory network of hESC needs the smallest number of inputs N_D (and also the smallest fraction of inputs n_D) to control. This may suggest that the transcription factor regulatory network of human embryonic stem cells is structurally rewired to be amenable to control, reflecting the cell type specificity of pluripotency and differentiation for hESC. On one hand, the TF regulatory networks of cell types Hemat. Stem Cell, Hippocampal Astrocyte and Fetal Lung are also easy to control, compared to other cell types. On the other hand, the TF regulatory networks of cell types Renal Cortical Epi., Small Airway Epi., Erythroid, B-Lymphoblastoid (GM06990) and Astrocyte are most difficult to control.

We can also see from Table 4.3, that the TF regulatory networks have quite different N , L and N_D , even for cell types in the same class, suggesting that the human TF regulatory network structures are cell-type-specific.

| Class | Cell type | N | L | N_D | n_D |
|-----------------------------|----------------------------|-------|-------|-------|-------|
| Embryonic Stem Cells | hESC | 533 | 16424 | 71 | 0.133 |
| Epithelia | Renal Cortical Epi. | 525 | 9597 | 176 | 0.335 |
| | Choroid Plexus Epi. | 527 | 13903 | 130 | 0.247 |
| | Small Airway Epi. | 522 | 9886 | 168 | 0.322 |
| | Amniotic Epi. | 526 | 13286 | 145 | 0.276 |
| | Esophageal Epi. | 528 | 14577 | 109 | 0.206 |
| | Iris Pigment Epi. | 527 | 12511 | 137 | 0.260 |
| Blood | Hemat. Stem Cell | 526 | 16461 | 74 | 0.141 |
| | Promyelocytic Leuk. | 525 | 18906 | 107 | 0.204 |
| | Erythroid | 493 | 9099 | 169 | 0.343 |
| | T-Lymphocyte | 518 | 12812 | 156 | 0.301 |
| | B-Lymphocyte | 515 | 16723 | 120 | 0.233 |
| | B-Lymphoblastoid (GM06990) | 501 | 12994 | 158 | 0.315 |
| B-Lymphoblastoid (GM12865) | 513 | 15202 | 127 | 0.248 | |
| Endothelia | Adult Dermal Blood | 520 | 13510 | 139 | 0.267 |
| | Neonatal Dermal Blood | 526 | 16761 | 107 | 0.203 |
| | Lung Lymphatic | 520 | 15435 | 115 | 0.221 |
| | Neonatal Dermal Lymph. | 526 | 15582 | 122 | 0.232 |
| Cancer | Neuroblastoma | 508 | 12761 | 131 | 0.258 |
| | Hepatoblastoma | 493 | 12863 | 136 | 0.276 |
| Visceral Cells | Hippocampal Astrocyte | 531 | 16391 | 85 | 0.160 |
| | Skeletal Myoblast | 523 | 13806 | 105 | 0.201 |
| | Skeletal Muscle | 529 | 17320 | 98 | 0.185 |
| | Astrocyte | 516 | 9296 | 162 | 0.314 |
| Fetal Tissues | Fetal Brain | 519 | 11698 | 122 | 0.235 |
| | Fetal Heart | 516 | 14295 | 114 | 0.221 |
| | Fetal Lung | 532 | 17823 | 74 | 0.139 |
| Stromal Cells | Aortic Fibroblast | 529 | 14795 | 121 | 0.229 |
| | Pulmonary Fib. | 527 | 14588 | 135 | 0.256 |
| | Fetal Lung Fib. | 519 | 11274 | 152 | 0.293 |
| | Lung Fib. | 527 | 14700 | 117 | 0.222 |
| | Adult Dermal Fib. | 529 | 13644 | 128 | 0.242 |
| | Neonatal Dermal Fib. | 521 | 15565 | 116 | 0.223 |
| | Cardiac Fib. | 527 | 15115 | 110 | 0.209 |
| | Cardiac Fib. | 522 | 14492 | 114 | 0.218 |
| | Pulmonary Artery Fib. | 531 | 13501 | 131 | 0.247 |
| | Skin Fib. | 521 | 12482 | 128 | 0.246 |
| | Mesenchymal Fib. | 526 | 15135 | 131 | 0.249 |
| | Mammary Fib. | 526 | 13961 | 116 | 0.221 |
| | Periodontal Fib. | 521 | 12822 | 142 | 0.273 |
| | Foreskin Fib. | 513 | 12126 | 112 | 0.218 |

Table 4.3: The network controllability of the human TF regulatory networks of 41 cell types. For each network, the cell type is given, as well as the number of nodes (N), the number of edges (L), the minimum number of inputs (N_D) and the fraction of inputs (n_D).

Chapter 5

Local controllability of biological networks

In this chapter, we propose a new network control framework called *local controllability* and provide theoretical results about local controllability. Local controllability concerns about the minimum number of inputs required to control a subset of nodes in a directed network, by appropriately connecting the inputs into the network (control configuration). It is important to study local controllability, not only due to its theoretic interest, but also because of its potential applications in the control of real networks considering that there are certain situations when one prefers to control just a subset of nodes in a network. For example, it is sometimes critical to control just a few target molecules (biomarkers) in the metabolic networks, as a therapeutic intervention in biomedical research [19].

It should be stressed that local controllability is an extension of network controllability [56], onto a local scale. Therefore, like the minimum input theorem (Theorem 4.1) for network controllability, all the results for local controllability derived in this chapter are also based on the structural controllability theory.

In this chapter, we will give some theoretical results for local controllability and then study the local controllability of biological networks.

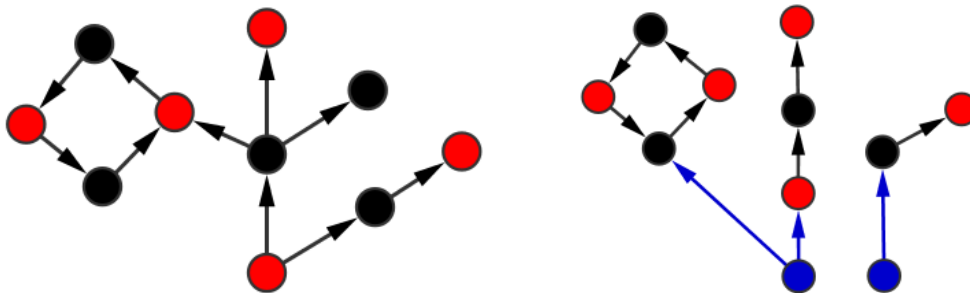
5.1 Definition of local controllability

Local controllability concerns about the minimum number of inputs required to control a subset of nodes in a directed network. The problem we study is formally stated as follows:

Given any nonempty subset S of nodes in any directed network $G = (V(G), A(G))$, find a minimum number of different inputs $\{u_1, \dots, u_l\}$ such that when the inputs are connected to the nodes in G appropriately, the subset S can be controlled (regardless of whether nodes in $V(G) \setminus S$ being controlled or not).

Here, the statement that a subset S is controlled in a directed network G , is equivalent to that there is a subsystem that is controllable and contains S after introducing inputs to the network G , also equivalent to that there is a cacti containing S after introducing inputs to the network G .

The problem above is well-defined, since we know that the whole network system becomes controllable by introducing N_D external inputs. In Figure 5.1, we give an illustration of controlling a subset of nodes in a directed network.



(a) A directed network G and a subset S

(b) A cacti containing S

Figure 5.1: An example of controlling a subset S of nodes in a directed network G . In (b), the nodes in blue are the input nodes of the cacti. The subset S is the set of nodes in red.

The problem of finding the minimum number of inputs to control S in a directed network G is an optimization problem, which can be formulated as follows.

Suppose $V(G) = \{v_1, v_2, \dots, v_n\}$ and $S = \{v_{i_1}, v_{i_2}, \dots, v_{i_r}\}$ with $1 \leq i_1 < i_2 < \dots < i_r \leq n$. Let $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$ be a column vector of n state functions. Let $u(t) = (u_1(t), u_2(t), \dots, u_n(t))^T$ be a column vector of n input functions. Let $E = (e_{ij}) \in \mathbb{R}^{n \times n}$ be a binary diagonal matrix, that is, $e_{ii} \in \{0, 1\}, i = 1, 2, \dots, n$. Let $F = (f_{ij}) \in \mathbb{R}^{n \times n}$ be a binary diagonal matrix such that $f_{i_k i_k} = 1, k = 1, 2, \dots, r$.

Let $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ be a structured matrix in which each entry is a free parameter. Let $D = (d_{ij}) \in \mathbb{R}^{n \times n}$ be a structured matrix defined as

$$d_{ij} = \begin{cases} \text{a free parameter,} & \text{if } (v_j \rightarrow v_i) \in A(G), \\ \text{zero,} & \text{if } (v_j \rightarrow v_i) \notin A(G). \end{cases}$$

Define the structured matrix $A = FDF \in \mathbb{R}^{n \times n}$, and define the structured matrix $B = FCE \in \mathbb{R}^{n \times n}$. The pair (A, B) represents a linear structured system of the form (3.3).

Then, *the local controllability of S in G* is defined as

$$lc(G, S) = \min_{E, F} \left\{ \sum_{i=1}^n e_{ii} : (A, B) \text{ is structurally controllable} \right\}. \quad (5.1)$$

The local controllability of S in G , denoted by $lc(G, S)$, is equal to the minimum number of inputs required to control S in a directed network G . As the definition suggests, $lc(G, S)$ depends on both G and S .

5.2 Properties and bounds for local controllability

First, in a directed network, a smaller subset of nodes is easier to control.

Theorem 5.1. *In a directed network $G = (V(G), A(G))$, S and S' are any two subsets of nodes in G , that is, $S \subseteq V(G)$ and $S' \subseteq V(G)$. Suppose S is a subset of S' , that is, $S \subseteq S'$. Then*

$$lc(G, S) \leq lc(G, S').$$

In particular,

$$lc(G, V(G)) = N_D(G).$$

where $N_D(G)$ is the network controllability of G .

Lin's structural controllability theorem (Theorem 3.10) shows that a subset S of nodes in a directed network G is controlled if there is a cacti that contains every node in S , after introducing inputs to the network G . The minimum input theorem (Theorem 4.1) further shows that each matching cycle in G doesn't need extra inputs to control and each matching path needs a different input.

If there exists a path in network G containing all the nodes in S , then one input turns the path into a stem and all the nodes in S are controlled, thus, $lc(G, S) = 1$. If there is a subset of independent cycles C_1, \dots, C_k such that S is contained in $C_1 \cup \dots \cup C_k$, then given one input node, by connecting this input node to a node in C_i for each cycle C_i , all the nodes in S will be controlled, thus, $lc(G, S) = 1$.

It might be expected that S is easy to control if the nodes in S tend to be locally clustered. For example, if all the nodes in S are within a strongly connected component (SCC) of the network G , $lc(G, S)$ may have a small value with respect to $|S|$. But as it turns out, this is not generally true.

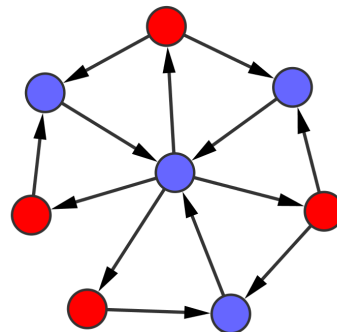


Figure 5.2: An example of a strongly connected network G with the nodes in S marked in red.

Consider an example of a strongly connected network G where we have an extreme case that $lc(G, S) = |S| - 1$, as shown in Figure 5.2. In this extreme case, $lc(G, S)$ is not small relative to $|S|$, as $|S| = 4$ and $lc(G, S) = 3$.

Examining the network G and the subset S in Figure 5.2 reveals that the high value of $lc(G, S)$ relative to $|S|$ is due to the fact that the central node lies on every path between any two nodes in S , thus blocking independent control along paths through nodes in S .

Since every directed network has a SCC decomposition, we can generalize this example to establish some bound for $lc(G, S)$.

Theorem 5.2. *Let G be a directed network and S be a subset of nodes in G . Suppose $|S| = t$ and all the nodes in S belong to totally p SCCs of G , i.e., there are p disjoint SCCs of G , G_1, \dots, G_p , and there is a partition of S , $S = S_1 \cup \dots \cup S_p$, such that S_i belongs to G_i for $i = 1, \dots, p$. Furthermore, assume each SCC is nontrivial, that is, the number of nodes in G_i is no less than two, $|V(G_i)| > 1$. Then*

$$lc(G, S) \leq \max\{1, t - p\}.$$

Proof. Consider any SCC G_i and S_i that belongs to G_i . If $|S_i| > 1$, consider any two nodes u and v from S_i , there is a path from u to v in G_i since G_i is a SCC. Then by introducing an input to u and a different input to every node w that is from $S_i - \{u, v\}$, all the nodes in S_i will be controlled. This means that it requires at most $|S_i| - 1$ different inputs to control S_i . If $|S_i| = 1$, consider this single node u from S_i , there is a cycle that goes through u in G_i since G_i is a SCC and $|V(G_i)| > 1$. Since a cycle doesn't need extra inputs, no different inputs are required to control S_i .

Therefore, these arguments imply that S can be controlled if a total number of

$$\sum_{i=1}^p (|S_i| - 1) = \sum_{i=1}^p |S_i| - p = t - p$$

different inputs are introduced to the network G .

Last, note that at least one input is required to control S in network G , by the minimum input theorem. Therefore, we have the following bound

$$lc(G, S) \leq \max \{1, t - p\}.$$

□

The example given in Figure 5.2 indicates that the equality can be attained in the bound: $lc(G, S) \leq \max \{1, t - p\}$.

Now, we define the concept of dilation in a directed network. In a directed network $G = (V(G), A(G))$, there is a *dilation* if a subset W of nodes in G can be found such that $|W| > |T(W)|$, where $T(W)$ is defined by:

$$T(W) = \{u \in V(G) : \text{there exists a directed edge from } u \text{ to some node in } W\}.$$

In the case that S forms a dilation in a directed network G , i.e., $|S| > |T(S)|$, then, at least $|S| - |T(S)|$ independent matching paths are needed to go through all the nodes in S . In other words, the number of inputs required to control S in G is at least $|S| - |T(S)|$.

Theorem 5.3. *Let G be a directed network, and S be a subset of nodes in G . If S forms a dilation, i.e., $|S| > |T(S)|$, then we have*

$$lc(G, S) \geq |S| - |T(S)|.$$

These bounds for $lc(G, S)$ are sometimes useful and easy to obtain, especially when $|S|$ is small.

5.3 A formula for local controllability

In this section, we show that the local controllability of S in a directed network G , $lc(G, S)$, can be calculated by transforming it to the problem of finding a minimum-weight perfect matching in a weighted directed network G^* , where G^* is constructed from the network G .

Let $G = (V(G), A(G))$ be a directed network with $|V(G)| = n$, and S be a subset of nodes in G , i.e., $S \subseteq V(G)$. Now, extend G to the complete directed network G^* : $G^* = (V(G^*), A(G^*))$ with $V(G^*) = V(G)$ and $A(G^*) = V(G) \times V(G)$, where $V(G) \times V(G)$ denotes the Cartesian product of the set $V(G)$ with itself, that is, for any node $u \in V(G^*)$ and any node $v \in V(G^*)$, there is a directed edge $(u, v) \in A(G^*)$.

Next, we assign a weight function w onto the edges in G^* as follows. Let $V(G) = \{v_1, v_2, \dots, v_n\}$. Suppose $S = \{v_{i_1}, v_{i_2}, \dots, v_{i_t}\}$ with an indexing set $I_S = \{i_1, i_2, \dots, i_t\}$ and $V(G) \setminus S = \{v_{j_1}, v_{j_2}, \dots, v_{j_{n-t}}\}$. Denote the set of self-loops for each node in $V(G) \setminus S$ by Γ , that is,

$$\Gamma = \{(v_{j_1}, v_{j_1}), (v_{j_2}, v_{j_2}), \dots, (v_{j_{n-t}}, v_{j_{n-t}})\}.$$

Then, the weight function w is a mapping, $w : A(G^*) \rightarrow \mathbb{R}$, defined as

$$w((u, v)) = \begin{cases} 0, & (u, v) \in \Gamma, \\ 1, & (u, v) \in A(G) \setminus \Gamma, \\ n, & (u, v) \in A(G^*) \setminus (A(G) \cup \Gamma). \end{cases} \quad (5.2)$$

Since G^* is a complete directed network, a cycle cover of S in G^* always exists. The definition of cycle cover is given as follows.

Definition 5.1. Let G be any directed network and S be a subset of nodes in G . A *cycle cover* of S in G is a set of disjoint (independent) cycles in G that contain all the nodes in S , that is, $C = C_1 \cup \dots \cup C_k$ is a cycle cover of S in G if all the cycles C_1, \dots, C_k are disjoint and each node in S is contained within some cycle C_j .

Definition 5.2. For any directed network $G = (V(G), A(G))$ and a subset S of nodes in G , let $C = C_1 \cup \dots \cup C_k$ be any cycle cover of S in G and let $w : A(G) \rightarrow \mathbb{R}$ be a weight function. The *weight of the cycle cover* C , $w(C)$, is then defined as

$$w(C) = \sum_{i=1}^k w(C_i),$$

where the weight of a cycle, $w(C_i)$, is given by

$$w(C_i) = \sum_{(u,v) \in C_i} w((u,v)).$$

Moreover, if \overline{C} is a cycle cover such that

$$w(\overline{C}) = \min \{w(C) : C \text{ is a cycle cover of } S \text{ in } G\},$$

then \overline{C} is called a *minimum-weight cycle cover* of S in G .

Our first result is that the local controllability $lc(G, S)$ can be calculated by finding a minimum-weight cycle cover of S in G^* .

Theorem 5.4. *Let $G = (V(G), A(G))$ be a directed network with $|V(G)| = n$, and S be a subset of nodes in G . Let the weight function $w : A(G^*) \rightarrow \mathbb{R}$ be defined as in (5.2). Let \overline{C} be a minimum-weight cycle cover of S in G^* , then we have*

$$lc(G, S) = \max \left\{ 1, \left\lfloor \frac{w(\overline{C})}{n} \right\rfloor \right\},$$

where $\lfloor \cdot \rfloor$ is the floor function.

Proof. The proof is due to the correspondence between the set of path-cycle covers of S in G and the set of cycle covers of S in G^* .

First, define a *path-cycle cover* of S in G to be a set of disjoint paths and cycles in G that contain S : $C_G(S) = P_1 \cup \dots \cup P_l \cup C_1 \cup \dots \cup C_k$ is a path-cycle cover of S in G , if P_1, \dots, P_l are directed paths in G , C_1, \dots, C_k are directed cycles in G (all the paths and cycles are disjoint), and each node in S is contained in $C_G(S)$.

For each path-cycle cover of S in G , denoted by $C_G(S) = P_1 \cup \dots \cup P_l \cup C_1 \cup \dots \cup C_k$, it corresponds to a cycle cover of S in G^* , denoted by $C_{G^*}(S)$, by adding a directed edge from the last node to the first node in each directed path P_j , $j = 1, \dots, l$. After the edge addition, each path thus becomes a (closed) cycle. Note that, in the resulting cycle cover $C_{G^*}(S)$, each cycle contains at most one directed edge that is not in G .

Conversely, for each cycle cover of S in G^* , denoted by $C_{G^*}(S)$, it corresponds to a path-cycle cover of S in G , denoted by $C_G(S)$, by removing from $C_{G^*}(S)$ all

the directed edges that are not contained in G . After the edge removal, some cycles are broken into paths.

The correspondence between a path-cycle cover of S in G and a cycle cover of S in G^* is illustrated in Figure 5.3. In this figure, $C_G(S)$ is given by the thin red arcs in the left, $C_{G^*}(S)$ is given by all the red arcs (both thin and thick) in the right. The thick red arcs are the edges added to $C_G(S)$ to form the cycle cover $C_{G^*}(S)$.

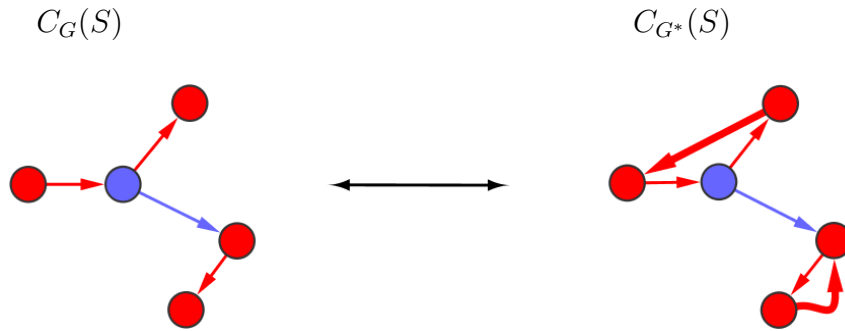


Figure 5.3: The correspondence between a path-cycle cover of S in G , $C_G(S)$, and a cycle cover of S in G^* , $C_{G^*}(S)$. The subset S is the set of all the nodes in red.

We know that $lc(G, S)$ (the minimum number of inputs required to control S in G) is given by the minimum number of paths in a path-cycle cover of S in G , since each independent path requires a different input to control and each independent cycle requires no additional input, according to the minimum input theorem (Theorem 4.1). Note that, in the case that there is actually a cycle cover of S in G (a cycle cover is also a path-cycle cover, without any path), then one input is sufficient for control, and $lc(G, S) = 1$.

For example, in Figure 5.3, let G be the network shown in the left panel, then the red arcs form a path-cycle cover $C_G(S)$ in G which contains a minimum number of two paths. Therefore, to control S (the nodes in red) in G , at least two inputs are required, that is, $lc(G, S) = 2$.

Consider a path-cycle cover of S in G , $C_G(S) = P_1 \cup \dots \cup P_l \cup C_1 \cup \dots \cup C_k$, where $l \neq 0$. Furthermore, assume that $C_G(S)$ has the minimum number of paths among all path-cycle covers of S in G . Then, the corresponding cycle cover of S in

G^* , $C_{G^*}(S)$, has weight $w(C_{G^*}(S)) = l \cdot n + q$, where q is the number of edges in $C_G(S)$ and $q < n$. Thus, $lc(G, S) = l = \left\lfloor \frac{w(C_{G^*}(S))}{n} \right\rfloor$.

Therefore, if $\bar{C} = \bar{C}_{G^*}(S)$ is a minimum-weight cycle cover of S in G^* , it corresponds to a path-cycle cover of S in G , $\bar{C}_G(S)$, which should contain exactly l paths. Otherwise, $\bar{C}_G(S)$ would contain more than l paths, and consequently, $\bar{C}_{G^*}(S)$ would have weight at least $(l + 1) \cdot n$. This is a contradiction with the assumption that $\bar{C} = \bar{C}_{G^*}(S)$ is a minimum-weight cycle cover of S in G^* .

These arguments show that if $l \neq 0$, then the local controllability $lc(G, S)$ is given by

$$lc(G, S) = l = \left\lfloor \frac{w(\bar{C})}{n} \right\rfloor.$$

Lastly, note that $lc(G, S)$ is at least one, thus, the proof is complete. □

To calculate $lc(G, S)$ efficiently, we can find a minimum-weight perfect matching in G^* .

Theorem 5.5. *Let $G = (V(G), A(G))$ be a directed network with $|V(G)| = n$, and S be a subset of nodes in G . Let the weight function $w : A(G^*) \rightarrow \mathbb{R}$ be defined as in (5.2). Let \bar{M} be a minimum-weight perfect matching in G^* , then we have*

$$lc(G, S) = \max \left\{ 1, \left\lfloor \frac{w(\bar{M})}{n} \right\rfloor \right\}.$$

Proof. To show this theorem, we note that each cycle cover of S in G^* corresponds to a perfect matching in G^* as follows. Suppose $C_{G^*}(S) = C_1 \cup \dots \cup C_k$ is a cycle cover of S in G^* , denote the set of nodes contained in $C_{G^*}(S)$ by $V(C_{G^*}(S))$, then $S \subseteq V(C_{G^*}(S))$. Let $V(G^*) \setminus V(C_{G^*}(S)) = \{w_1, w_2, \dots, w_r\}$, then by definition, $\{w_1, w_2, \dots, w_r\}$ is a subset of $V(G) \setminus S = \{v_1, \dots, v_{n-t}\}$. Form a self-loop for each node w_i and denote it by L_i , that is, $L_i = (w_i, w_i)$ for $i = 1, 2, \dots, r$. Then $C_{G^*}(S)$ corresponds to a perfect matching M_{G^*} in G^* defined as

$$M_{G^*} = C_{G^*}(S) \cup L_1 \cup L_2 \cup \dots \cup L_r.$$

It is easy to see that M_{G^*} is a cycle decomposition in G^* , thus a perfect matching in G^* . Furthermore, since $w(L_i) = 0$ for $i = 1, 2, \dots, r$, we have $w(M_{G^*}) = w(C_{G^*}(S))$.

Conversely, a perfect matching in G^* easily corresponds to a cycle cover of S in G^* as follows. Given a perfect matching M_{G^*} , since a perfect matching is a cycle decomposition, we can denote $M_{G^*} = \tilde{C}_1 \cup \tilde{C}_2 \cup \dots \cup \tilde{C}_p$ for some positive integer p , where each \tilde{C}_i is a cycle (including self-loop) in G^* . Identify all the cycles \tilde{C}_i which contains at least one node in S , and denote the set of all the identified cycles by $\{\tilde{C}_{j_1}, \tilde{C}_{j_2}, \dots, \tilde{C}_{j_s}\}$. Then, M_{G^*} corresponds to a cycle cover $C_{G^*}(S)$ of S in G^* defined as

$$C_{G^*}(S) = \tilde{C}_{j_1} \cup \tilde{C}_{j_2} \cup \dots \cup \tilde{C}_{j_s}.$$

Moreover, $w(C_{G^*}(S)) \leq w(M_{G^*})$.

Therefore, a minimum-weight perfect matching in G^* corresponds to a minimum-weight cycle cover of S in G^* and they have equal weight. Thus, this theorem follows from Theorem 5.4. □

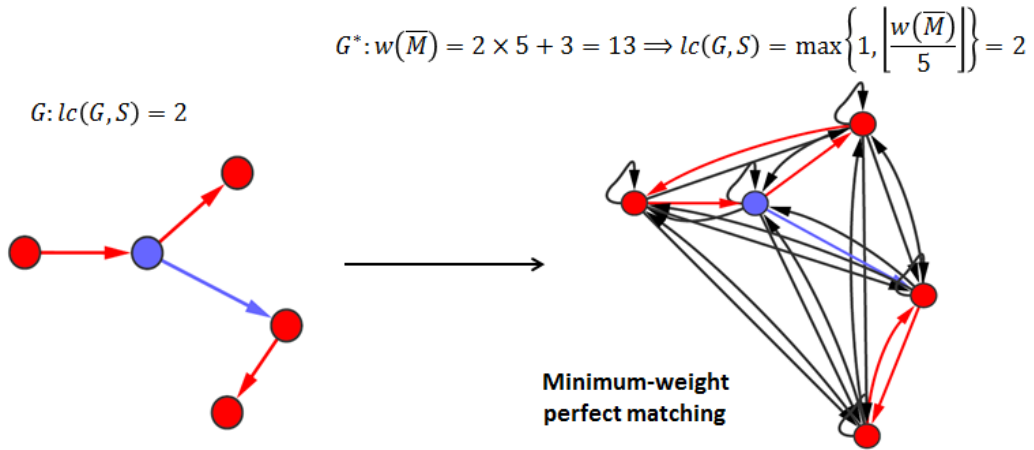


Figure 5.4: An example to show that $lc(G, S)$ can be calculated by finding a minimum-weight perfect matching in G^* (marked by the red arcs). The subset S is the set of all the nodes in red.

Here, we illustrate Theorem 5.5 (minimum-weight perfect matching) by an example, shown in Figure 5.4. In Figure 5.4, we have drawn the initial network G (in the left) and the complete directed network G^* (in the right). The red arcs in G^* form a minimum-weight perfect matching with weight $2 \times 5 + 3 = 13$, thus, $lc(G, S) = \max\{1, \lfloor 13/5 \rfloor\} = 2$. Generally, to find a minimum-weight perfect matching in G^* , we used the Hungarian method for assignment problem [52].

5.4 The algorithm for local controllability

In this section, we present a cubic-time algorithm (the algorithm for local controllability) to calculate $lc(G, S)$. Before that, we present the Hungarian method based on cost matrix.

5.4.1 The Hungarian method

The Hungarian method is an algorithm to find an optimal assignment in the general assignment problem [52]. It can be formulated based on cost matrix. In the following, we will present the Hungarian method based on cost matrix.

The Hungarian method based on cost matrix

Input: a $n \times n$ cost matrix $C = (c_{ij})$.

Output: an optimal assignment, i.e., a permutation σ of the set $\{1, \dots, n\}$ such that $\sum_{i=1}^n c_{i\sigma(i)}$ is minimum.

Step 1: For each row of the cost matrix, find the smallest element in this row, and subtract it from every element in this row.

Step 2: Find a zero element (denoted by Z) in the resulting matrix. If there is no starred zero in its row or column, star Z (by starring Z , we mean labelling Z by a star *, and Z is called ‘starred’). Repeat for each zero element in the matrix.

Step 3: Cover each column that contains a starred zero (by covering a column,

we mean covering the elements in the column by a straight line, and the column together with the elements in the column are called ‘covered’, a column that is not covered is called ‘uncovered’). If n columns are covered, the starred zeros describe an optimal assignment, stop and output the assignment indicated by a permutation σ defined as follows: denote the set of n starred zero entries by $\{c_{1k_1}, \dots, c_{nk_n}\}$, then $\sigma(i) = k_i$, for $i = 1, \dots, n$. If less than n columns are covered, go to Step 4.

Step 4: Find an uncovered zero (a zero entry that is not covered by a straight line) and underline it (label it by an underline, and it is called ‘underlined’), If there is no starred zero in the row containing this underlined zero, go to Step 5. Otherwise, cover this row and uncover the column containing the starred zero which is in the row of this underlined zero (by uncovering the column, we mean removing the straight line that covers the column). Continue in this manner until there are no uncovered zeros left. Save the smallest uncovered value and go to Step 6.

Step 5: Construct a series of alternating underlined and starred zeros as follows. Let Z_0 denote the uncovered underlined zero found in Step 4. Let Z_1 denote the starred zero in the column of Z_0 (if any). Let Z_2 denote the underlined zero in the row of Z_1 (there will always be one). Continue until the series terminates at an underlined zero that has no starred zero in its column. Unstar each starred zero of the series (by unstarring a starred zero, we mean removing the star * that labels this starred zero), star each underlined zero of the series, erase all underlines and uncover every line (a line is a row or a column) in the matrix. Return to Step 3.

Step 6: Add the value found in Step 4 to every element of each covered row, and subtract it from every element of each uncovered column. Return to Step 4 without altering any stars, underlines, or covered lines.

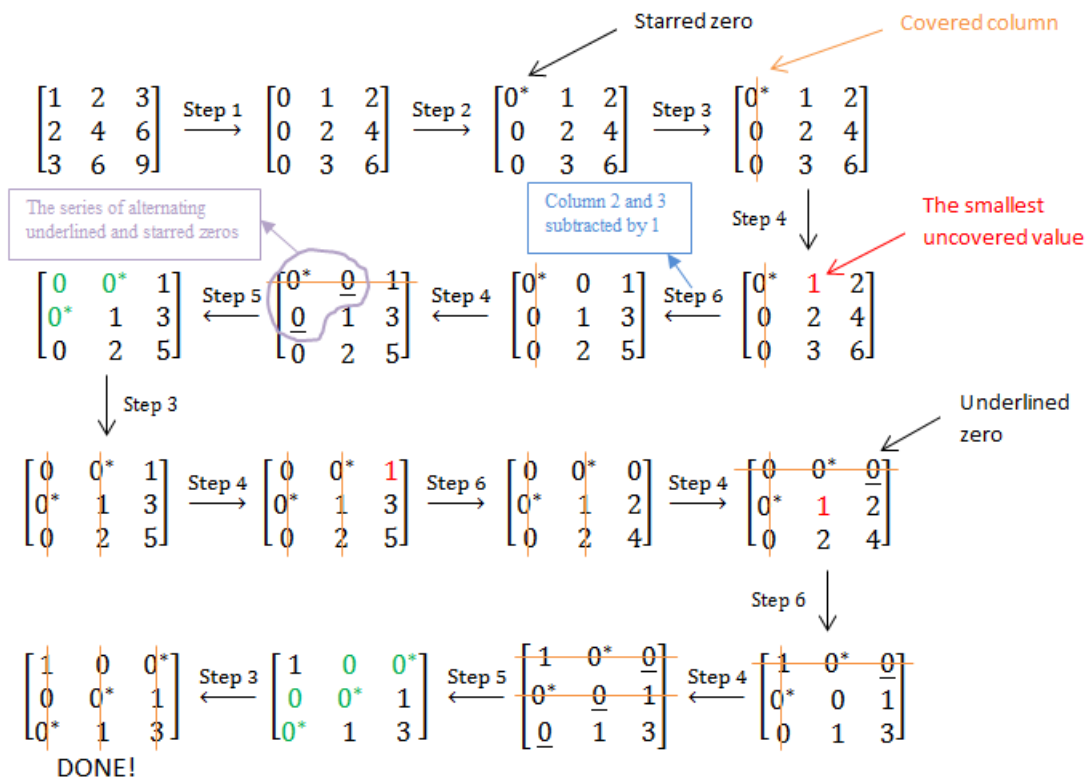
The above algorithm has time complexity $O(n^3)$, since there are at most n times of iterations of Step 5 (after Step 5, the number of starred zeros increases), and in Step 5 it takes time $O(n)$ to identify a series of alternating underlined and starred zeros, and there are at most n times of adjusting the values in the cost matrix in

Step 6 before a series of alternating underlined and starred zeros can be found in Step 5.

In Figure 5.5, we give an example to illustrate the Hungrain method based on cost matrix, step by step. In this example, we start with the input, a cost matrix, and show the result after every step of the algorithm until an optimal assignment σ is found (DONE!), and output σ . This example is in perfect agreement with the algorithm (the Hungrain method based on cost matrix) presented above.

Input: a cost matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$



Output: an optimal assignment σ with $\sigma(1) = 3, \sigma(2) = 2, \sigma(3) = 1$.

Figure 5.5: An illustration of the Hungarian method based on cost matrix.

5.4.2 The algorithm for local controllability

The algorithm for local controllability uses the cost matrix-based Hungarian method as its core ingredient.

The algorithm for local controllability

Input: a directed network $G = (V(G), A(G))$ with $|V(G)| = n$ and a subset S of nodes in G .

Output: the minimum number of inputs required to control S in G , i.e., $lc(G, S)$.

Step 1: Denote $V(G) = \{v_1, v_2, \dots, v_n\}$ and assume $S = \{v_{i_1}, v_{i_2}, \dots, v_{i_t}\}$ with an indexing set $I_S = \{i_1, i_2, \dots, i_t\}$. Denote the complement of S in $V(G)$ by \bar{S} , i.e., $\bar{S} = V(G) \setminus S$, and assume that there is an indexing set $I_{\bar{S}} = \{j_1, j_2, \dots, j_{n-t}\}$ such that $\bar{S} = \{v_{j_1}, v_{j_2}, \dots, v_{j_{n-t}}\}$.

Step 2: Extend G to the complete directed network G^* , which is defined as

$$G^* = (V(G^*) = V(G), A(G^*) = V(G) \times V(G)).$$

Assign a weight function w onto the directed edges in G^* , i.e., $w : A(G^*) \rightarrow \mathbb{R}$, defined as in (5.2).

Step 3: Construct a $n \times n$ cost matrix C to be the weighted adjacency matrix of G^* . The cost matrix $C = (c_{ij})$ is therefore defined as

$$c_{ij} = w((v_i, v_j)).$$

Step 4: Apply the cost matrix-based Hungarian method to C to find an optimal assignment σ . Output the local controllability to be

$$lc(G, S) = \max \left\{ 1, \left\lfloor \frac{\sum_{i=1}^n c_{i\sigma(i)}}{n} \right\rfloor \right\}.$$

5.5 Local controllability of brain networks

With the algorithm for local controllability developed as above, in this section, the local controllability of brain subdivisions in mouse inter-region brain networks (see Subsection 2.2.1) is studied.

For each of the 10 mouse inter-region brain networks obtained, denoted by G , we calculated the local controllability $lc(G, S)$ of each subdivision S in the network G . These results are summarized in Table 5.1.

As can be seen from Table 5.1, for each subdivision, except for Pons and Medulla, the local controllability of the subdivision in each ipsilateral network is smaller than or equal to the local controllability of the subdivision in each contralateral network. This indicates that, except for Pons and Medulla, each subdivision is easier to control in ipsilateral networks than in contralateral networks.

For Pons, less inputs are needed to control it in contralateral networks than in ipsilateral networks, as can be seen from Table 5.1. This might be due to that the subnetworks induced by Pons in contralateral networks are more dense than the subnetworks induced by Pons in ipsilateral networks. For example, the total numbers of edges of the subnetworks induced by Pons in the 5 ipsilateral networks are 27, 23, 20, 19 and 12; whereas those in the 5 contralateral networks are 42, 34, 34, 30 and 23.

For Medulla, less inputs are needed to control it in contralateral networks than in ipsilateral networks, corresponding to P-value cutoffs 0.05, 0.02, 0.01 and 0.005. However, the subnetworks induced by Medulla in contralateral networks have smaller numbers of edges than the subnetworks induced by Medulla in ipsilateral networks. This may indicate that Medulla has different connection structures in contralateral networks that make it easy to control in these networks. We will further discuss this observation by comparing the local controllability of Medulla with the local controllability of random subsets later in this section (Section 5.5).

For Striatum and Thalamus, a relatively large number of inputs is required to

control each of them in contralateral networks due to their positions in those networks. As observed before, there are 8 and 19 sink nodes in Striatum and Thalamus respectively in the contralateral network corresponding to P-value cutoff 0.05. Since each sink node needs a different input to control, the local controllability of Striatum and Thalamus in contralateral networks are relatively large.

Isocortex is the easiest to control; it needs only one input to control in all networks. This finding might be corroborated by the features of the isocortex that it has massive interconnections, and also that the isocortex, as a sophisticated centre for processing high-level information, has an extraordinary regularity within the mouse brain [48]. This will be discussed further in the later comparative study in this section (Section 5.5).

| Local controllability | | | | | | | | | | |
|------------------------------|---------------------|------|------|-------|-------|-----------------------|------|------|-------|-------|
| | ipsilateral network | | | | | contralateral network | | | | |
| P-value cutoff | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| Number of edges | 3123 | 2721 | 2487 | 2278 | 1947 | 2451 | 2160 | 1979 | 1830 | 1587 |
| Isocortex (38) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Olfactory Areas (11) | 1 | 1 | 2 | 2 | 3 | 5 | 5 | 5 | 5 | 7 |
| Hippocampus (11) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 |
| Cortical Subplate (7) | 1 | 1 | 1 | 1 | 1 | 4 | 4 | 4 | 4 | 5 |
| Striatum (12) | 3 | 3 | 3 | 3 | 4 | 9 | 9 | 9 | 10 | 10 |
| Pallidum (8) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| Thalamus (35) | 4 | 6 | 6 | 9 | 11 | 19 | 20 | 20 | 20 | 21 |
| Hypothalamus (20) | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 |
| Midbrain (21) | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 |
| Pons (13) | 3 | 3 | 4 | 4 | 4 | 2 | 2 | 2 | 2 | 3 |
| Medulla (25) | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 3 |
| Cerebellum (12) | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 |

Table 5.1: The local controllability of 12 subdivisions in the mouse inter-region brain networks in both ipsilateral and contralateral hemispheres, corresponding to P-value cutoffs 0.05, 0.02, 0.01, 0.005 and 0.001. The parentheses after each subdivision indicate the number of regions belonging to that subdivision.

Now we consider the ipsilateral network corresponding to P-value cutoff 0.05. For each of the 12 subdivisions, we found a controllable subnetwork (that delivers control of the subdivision with the minimum number of inputs) in the ipsilateral network corresponding to P-value cutoff 0.05. The controllable subnetworks are shown in Figure 5.6. The controllable subnetwork of each subdivision is drawn as a path-cycle cover of the subdivision in the ipsilateral network corresponding to P-value cutoff 0.05 (see Theorem 5.4).

In Figure 5.6, each node is marked in a color according to the subdivision to which it belongs (see Figure 2.1). Note that, in the controllable subnetwork of a subdivision, there could be nodes that are in other subdivisions.

For the subdivisions Striatum, Thalamus, Pons and Medulla, whose local controllability in the ipsilateral network corresponding to P-value cutoff 0.05 are larger than 1, they contain sink nodes (nodes with out-degree zero). Figure 5.6 shows that each sink node is the last node of a directed path in the controllable subnetwork of each subdivision. This is not unexpected since each sink node always needs a different input to control in a directed network. Figure 5.6 also shows that for each of the subdivisions Isocortex, Hippocampus and Medulla, the controllable subnetwork only contains nodes in that subdivision.

Next, we tried to compare the local controllability of each subdivision with the local controllability of random subsets with the same size as the subdivision. Since the sizes of the 12 subdivisions (in the order from Isocortex to Cerebellum as in Table 5.1) are 38, 11, 11, 7, 12, 8, 35, 20, 21, 13, 25 and 12, we thus generated random subsets with sizes equal to the possible sizes of the 12 subdivisions. Specifically, for any given (subdivision) size s ($s = 38, 11, 7, 12, 8, 35, 20, 21, 13, 25$), in each of the 10 inter-region brain networks, we selected 5000 random subsets of nodes (brain regions) with size s . For each random subset S , we calculated the local controllability $lc(G, S)$ of S in the inter-region brain network G . Then, the mean and standard deviation of $lc(G, S)$ over 5000 random subsets were obtained. The results are summarized in Table 5.2.

Table 5.2 shows that for a given size s , the average local controllability of random subsets with size s in each contralateral network is much higher than that in each ipsilateral network. However, the contralateral networks corresponding to P-value cutoffs 0.05, 0.02 and 0.01 each contains more edges than the ipsilateral network corresponding to P-value cutoff 0.001. This might indicate that contralateral networks are structurally rewired, so that it is more difficult to control a random subset in a contralateral network than in any ipsilateral network. Given the previous finding that contralateral networks are much more difficult to control than ipsilateral networks (see Section 4.3), we confirm that they are more difficult to control both globally and locally than ipsilateral networks. As explained in Section 4.3, this behavior could be due to the structural difference in contralateral networks: there are many more sink nodes (nodes with out-degree zero) in contralateral networks than in ipsilateral networks.

| The mean (std) of $lc(G, S)$ | | | | | | | | | | |
|------------------------------|---------------------|-------------------|-------------------|-------------------|-------------------|-----------------------|-------------------|-------------------|-------------------|--------------------|
| | ipsilateral network | | | | | contralateral network | | | | |
| P-value cutoff | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| 38 | 2.2448 (1.176) | 3.0754 (1.470) | 3.32 (1.549) | 3.9062 (1.659) | 5.1416 (1.877) | 8.843 (2.285) | 9.1536 (2.314) | 9.7 (2.383) | 9.7436 (2.402) | 11.3478 (2.503) |
| 11 | 1.145 (0.413) | 1.2812 (0.591) | 1.332 (0.634) | 1.4352 (0.720) | 1.6926 (0.898) | 2.6574 (1.315) | 2.7814 (1.311) | 2.8886 (1.366) | 2.8698 (1.348) | 3.3538 (1.441) |
| 7 | 1.055 (0.242) | 1.1178 (0.364) | 1.1386 (0.385) | 1.1882 (0.467) | 1.286 (0.583) | 1.7916 (0.913) | 1.859 (0.948) | 1.9262 (0.990) | 1.9754 (1.009) | 2.1874 (1.077) |
| 12 | 1.175 (0.451) | 1.3116 (0.611) | 1.3696 (0.664) | 1.5048 (0.771) | 1.7914 (0.963) | 2.8972 (1.350) | 2.9766 (1.387) | 3.1486 (1.437) | 3.1528 (1.417) | 3.6526 (1.516) |
| 8 | 1.0688 (0.269) | 1.1466 (0.406) | 1.1768 (0.463) | 1.2336 (0.521) | 1.3934 (0.668) | 2.0078 (1.037) | 2.0408 (1.033) | 2.1842 (1.098) | 2.1804 (1.085) | 2.4894 (1.181) |
| 35 | 2.0988 (1.113) | 2.8278 (1.398) | 3.0694 (1.456) | 3.6002 (1.589) | 4.7468 (1.834) | 8.207 (2.215) | 8.4918 (2.285) | 8.9458 (2.311) | 9.0118 (2.346) | 10.4932 (2.437) |
| 20 | 1.4188 (0.698) | 1.7846 (0.974) | 1.8904 (1.022) | 2.1758 (1.165) | 2.7776 (1.407) | 4.7436 (1.818) | 4.8978 (1.797) | 5.1706 (1.838) | 5.1834 (1.823) | 6.0684 (1.935) |
| 21 | 1.4756 (0.754) | 1.8428 (1.001) | 1.9674 (1.057) | 2.2466 (1.189) | 2.8784 (1.428) | 4.965 (1.858) | 5.1294 (1.823) | 5.4684 (1.871) | 5.4632 (1.876) | 6.233 (1.991) |
| 13 | 1.1834 (0.457) | 1.3642 (0.656) | 1.4386 (0.719) | 1.5642 (0.831) | 1.8872 (1.017) | 3.0862 (1.422) | 3.207 (1.434) | 3.4072 (1.502) | 3.4098 (1.484) | 3.9508 (1.569) |
| 25 | 1.6536 (0.888) | 2.0892 (1.136) | 2.3036 (1.208) | 2.6286 (1.340) | 3.393 (1.557) | 5.8584 (1.965) | 6.1016 (1.979) | 6.4618 (2.012) | 6.4634 (2.006) | 7.5136 (2.151) |

Table 5.2: The mean and standard deviation of the local controllability of random subsets with a given size in the mouse inter-region brain networks. The parentheses below each mean indicate the corresponding standard deviation.

Furthermore, for each subdivision, in each of the 10 inter-region brain networks, we compared the local controllability of the subdivision with the local controllability of 5000 random subsets with the same size as the subdivision, and calculated the p -value. Denoting the local controllability of the subdivision by lc_0 , and the local controllability of 5000 random subsets by a vector $d = (lc_1, lc_2, \dots, lc_{5000})$, we then calculated two (one-tailed) p -values: p_1 and p_2 . The p -value p_1 (p_2) is given by the fraction of elements lc_i in the vector d such that $lc_i \leq lc_0$ ($lc_i \geq lc_0$). The p -value p_1 (p_2) is used to test how significantly smaller (larger) is the local controllability of the subdivision than that of random subsets with the same size as the subdivision. The results for the two types of p -values are given in two tables: Table 5.3 and Table 5.4. In these two tables, the significant p -values (< 0.05) are marked in bold.

In Table 5.3, we can see that the subdivisions Isocortex and Medulla are significantly easier to control than the random subsets in contralateral networks ($p_1 < 0.05$). This behavior is also true for Midbrain, though less significant ($p_1 < 0.09$). Moreover, Isocortex is significantly easier to control than the random subsets in the ipsilateral network corresponding to P -value cutoff 0.001.

The low local controllability of Isocortex (significantly smaller than the local controllability of the random subsets in contralateral networks) might be due to a high level of functional regularity of the isocortex. As mentioned in Section 2.2.2, one striking feature of the isocortex is its relatively uniform structure, which implies that its connectivity with other cortical regions and subcortical structures is established, maintained and altered by sophisticated regulatory mechanisms [48]. It would be interesting to investigate the relationship between the low local controllability of Isocortex and the high cytoarchitectonic homogeneity and high functional regularity of the isocortex.

It was recently suggested that the medulla oblongata of the mouse is organized in a segmental pattern, and structures of the medullary systems such as the nucleus of the solitary tract or the reticular formation appear subdivided into successive

segment-like units [94]. Furthermore, the neurogenetic pattern of the medulla oblongata is rather homogeneous along the rostrocaudal axis, without apparent discontinuity with the spinal cord [94]. Therefore, the relationship between the low local controllability of Medulla and its highly organized and homogeneous structure would need to be explored further.

In Table 5.4, we can see that the subdivisions Striatum and Thalamus are significantly more difficult to control than the random subsets in contralateral networks ($p_2 < 0.05$). This is because most of the brain regions in Striatum and Thalamus are sink nodes in contralateral networks. Table 5.4 also shows that in the ipsilateral network corresponding to P-value cutoff 0.05, Striatum and Pons are significantly more difficult to control than the random subsets.

Recent studies have revealed that heterogeneity was observed in the neuroanatomy, microcircuits, synaptology, electrophysiology, or functional organization of the striatum and the thalamus [1, 83]. Understanding how the heterogeneity of Striatum and Thalamus contributes to their high local controllability (significantly larger than the local controllability of the random subsets in contralateral networks) might facilitate our understanding of the structural basis for the local controllability of a particular subdivision in the mouse inter-region brain networks.

In summary, the homogeneous subdivisions Isocortex and Medulla are (significantly) easy to control, whereas the heterogeneous subdivisions Striatum and Thalamus are difficult to control. In the future, it might be interesting to investigate whether there is structural basis for the local controllability of a brain subdivision in the mouse inter-region brain networks, such as the homogeneity or heterogeneity of the subdivision.

| The p -value p_1 | | | | | | | | | | |
|-----------------------|---------------------|--------|--------|--------|---------------|-----------------------|---------------|---------------|---------------|---------------|
| | ipsilateral network | | | | | contralateral network | | | | |
| P-value cutoff | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| Isocortex (38) | 0.3254 | 0.1534 | 0.123 | 0.0686 | 0.0188 | 0.0002 | 0 | 0.0004 | 0.0002 | 0 |
| Olfactory Areas (11) | 0.8756 | 0.7816 | 0.94 | 0.9082 | 0.9532 | 0.9744 | 0.974 | 0.965 | 0.9658 | 0.9966 |
| Hippocampus (11) | 0.8756 | 0.7816 | 0.742 | 0.6776 | 0.5382 | 0.2194 | 0.1896 | 0.4202 | 0.4198 | 0.2938 |
| Cortical Subplate (7) | 0.948 | 0.8956 | 0.8754 | 0.8414 | 0.7722 | 0.9922 | 0.9894 | 0.9874 | 0.983 | 0.996 |
| Striatum (12) | 0.9974 | 0.9926 | 0.988 | 0.9744 | 0.9866 | 1 | 1 | 1 | 1 | 1 |
| Pallidum (8) | 0.9354 | 0.8716 | 0.8542 | 0.8072 | 0.6966 | 0.391 | 0.3744 | 0.3248 | 0.3202 | 0.8028 |
| Thalamus (35) | 0.97 | 0.9896 | 0.982 | 1 | 0.9998 | 1 | 1 | 1 | 1 | 1 |
| Hypothalamus (20) | 0.684 | 0.5046 | 0.4572 | 0.3538 | 0.2124 | 0.2594 | 0.2218 | 0.1826 | 0.1796 | 0.0836 |
| Midbrain (21) | 0.6588 | 0.4792 | 0.4224 | 0.332 | 0.1874 | 0.0868 | 0.068 | 0.0478 | 0.0522 | 0.0862 |
| Pons (13) | 0.9974 | 0.9874 | 0.9974 | 0.9936 | 0.9824 | 0.3672 | 0.3354 | 0.2966 | 0.2868 | 0.4012 |
| Medulla (25) | 0.8366 | 0.6838 | 0.6134 | 0.512 | 0.3058 | 0.0062 | 0.0064 | 0.0036 | 0.0028 | 0.0234 |
| Cerebellum (12) | 0.8514 | 0.758 | 0.7196 | 0.6374 | 0.493 | 0.6878 | 0.67 | 0.6162 | 0.612 | 0.4732 |

Table 5.3: The p -value for the alternative hypothesis that the local controllability of a subdivision is smaller than the local controllability of random subsets with the same size as the subdivision, in the mouse inter-region brain networks. The significant p -values ($p_1 < 0.05$) are marked in bold.

| The p -value p_2 | | | | | | | | | | |
|-----------------------|---------------------|---------------|---------------|---------------|--------------|-----------------------|----------|---------------|----------|---------------|
| | ipsilateral network | | | | | contralateral network | | | | |
| P-value cutoff | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| Isocortex (38) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Olfactory Areas (11) | 1 | 1 | 0.258 | 0.3224 | 0.175 | 0.0936 | 0.1032 | 0.124 | 0.1192 | 0.0178 |
| Hippocampus (11) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.8262 | 0.8224 | 0.9014 |
| Cortical Subplate (7) | 1 | 1 | 1 | 1 | 1 | 0.049 | 0.061 | 0.075 | 0.0806 | 0.0248 |
| Striatum (12) | 0.0236 | 0.061 | 0.076 | 0.1144 | 0.0624 | 0 | 0 | 0.0004 | 0 | 0.0002 |
| Pallidum (8) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.4638 |
| Thalamus (35) | 0.1176 | 0.0394 | 0.0576 | 0.0016 | 0.002 | 0 | 0 | 0 | 0 | 0 |
| Hypothalamus (20) | 1 | 1 | 1 | 1 | 1 | 0.8976 | 0.9176 | 0.9296 | 0.9346 | 0.9746 |
| Midbrain (21) | 1 | 1 | 1 | 1 | 1 | 0.9776 | 0.9844 | 0.9916 | 0.9888 | 0.9752 |
| Pons (13) | 0.0242 | 0.0724 | 0.0172 | 0.0346 | 0.0786 | 0.8572 | 0.8788 | 0.9012 | 0.9026 | 0.812 |
| Medulla (25) | 0.4362 | 0.6114 | 0.689 | 0.7688 | 0.8896 | 1 | 1 | 1 | 1 | 0.9948 |
| Cerebellum (12) | 1 | 1 | 1 | 1 | 1 | 0.5848 | 0.6046 | 0.644 | 0.6552 | 0.7676 |

Table 5.4: The p -value for the alternative hypothesis that the local controllability of a subdivision is larger than the local controllability of random subsets with the same size as the subdivision, in the mouse inter-region brain networks. The significant p -values ($p_2 < 0.05$) are marked in bold.

5.6 Local controllability of transcriptional regulatory networks

In this section, the local controllability of transcription factor protein complexes in human cell-type-specific transcription factor regulatory networks (see Subsection 2.2.3) is studied.

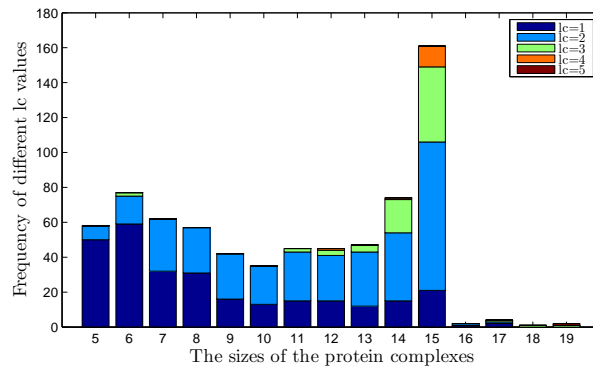
Sequence-specific transcription factors are the key effectors of eukaryotic gene control. In humans, the mutual interactions among TFs can determine cellular identity and shape complex cellular functions [20]. Typically, transcription factors do not work in isolation, they form protein complexes with other TFs to combinatorially regulate the gene expression in a concerted manner [102].

Recently, a comprehensive human protein complex resource was generated [97]. From this resource, we obtained a list of 834 protein complexes that contains transcription factors in the human TF regulatory networks introduced in Subsection 2.2.3. Since each protein complex can be considered as a subset of TFs, we calculated the local controllability of these 834 protein complexes in the human TF regulatory networks.

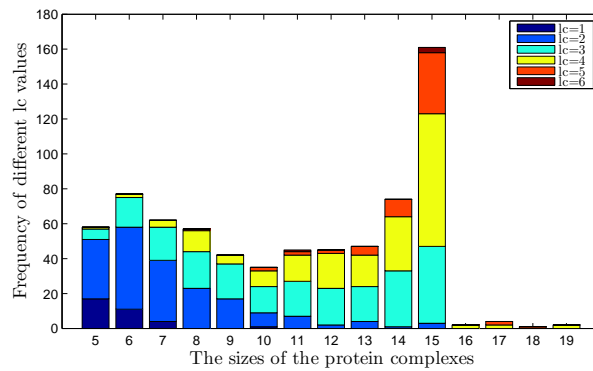
For each of the 834 protein complexes, the number of TFs it contains is in the range from 2 to 19. We calculated the local controllability of those protein complexes with sizes from 5 to 19 in the TF regulatory networks of 8 cell types, where the size of a protein complex is the number of TFs contained in the complex. The full results are presented in Table A.1, in Appendix A.

To study how the local controllability of a protein complex varies with the size of the complex, we represent the results with bar plots, shown in Figure 5.7 and Figure 5.8. In each of the bar plots in Figure 5.7 and Figure 5.8, for a given protein complex size s ($s = 5, 6, \dots, 19$), the number of protein complexes with size s and with local controllability lc ($lc = 1, 2, \dots$) is shown. Each bar plot represents the TF regulatory network of one cell type.

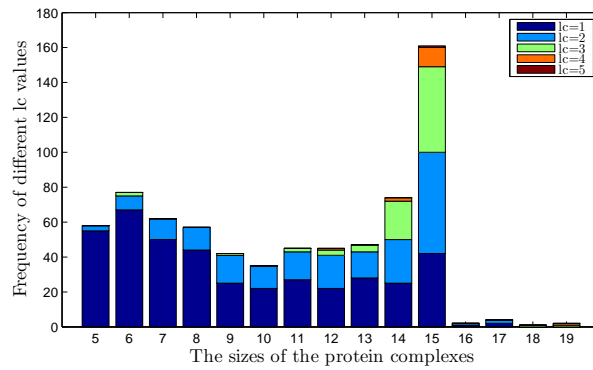
In Figure 5.7 and Figure 5.8, for protein complexes with a relatively large size s ($s \geq 9$), in the TF regulatory networks of cell types Amniotic Epi., Neuroblastoma, Skeletal Myoblast and Fetal Brain, there are very few complexes with local controllability equal to 1. Moreover, for protein complexes with a relatively small size s ($s < 9$), in the TF regulatory networks of cell types hESC, Hemat. Stem Cell, Adult Dermal Blood and Skin Fib., there are a large fraction of complexes with local controllability equal to 1.



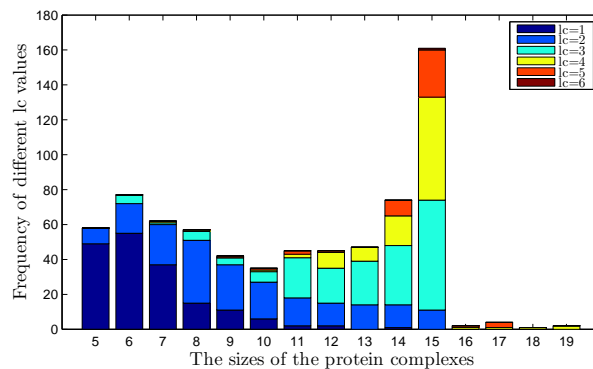
(a) hESC



(b) Amniotic Epi.

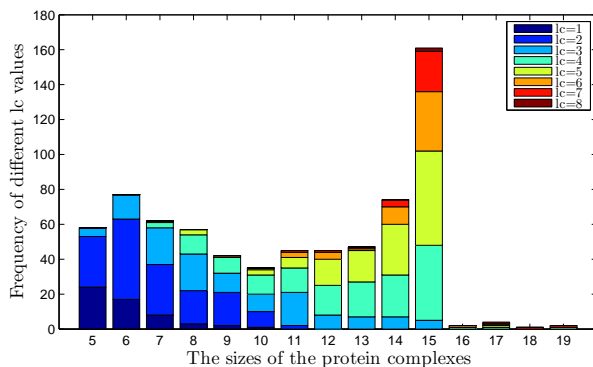


(c) Hemat. Stem Cell

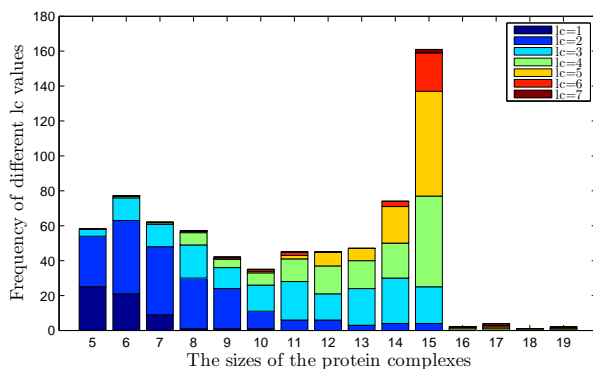


(d) Adult Dermal Blood

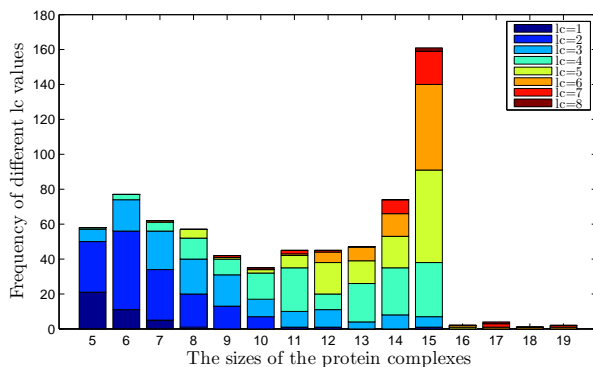
Figure 5.7: The frequency of protein complexes with a given size and with a given local controllability lc , in the TF regulatory networks of cell types hESC, Amniotic Epi., Hemat. Stem Cell and Adult Dermal Blood.



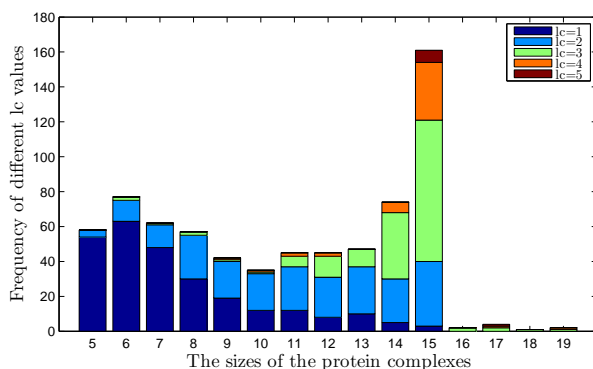
(a) Neuroblastoma



(b) Skeletal Myoblast



(c) Fetal Brain



(d) Skin Fib.

Figure 5.8: The frequency of protein complexes with a given size and with a given local controllability lc , in the TF regulatory networks of cell types Neuroblastoma, Skeletal Myoblast, Fetal Brain and Skin Fib.

Next, based on Table A.1 in Appendix A, we calculated the average local controllability of protein complexes with a given size in the TF regulatory network of each of the 8 cell types. The results are summarized in Table 5.5.

Table 5.5 shows that the average local controllability of protein complexes with a given size in the TF regulatory networks of cell types hESC, Hemat. Stem Cell, Adult Dermal Blood and Skin Fib., are smaller than those in the TF regulatory networks of the 4 other cell types.

Furthermore, to compare the local controllability of protein complexes with the local controllability of random subsets with the same size, we generated random subsets of TFs in the TF regulatory network of each of the 8 cell types and calculated the local controllability of each random subset in the network.

For any given protein complex size s ($s = 5, 6, \dots, 19$), in the TF regulatory network of each of these 8 cell types, 5000 random subsets of TFs with size s were selected. For each random subset S , the local controllability $lc(G, S)$ of S in the TF regulatory network of that cell type G was then calculated. The mean and standard deviation of $lc(G, S)$ over 5000 random subsets are summarized in Table 5.6.

| Complex size | The mean of $lc(G, S)$ | | | | | | | |
|--------------|------------------------|-------|-------|-------|-------|-------|-------|-------|
| | hESC | AE | HSC | ADB | Neu | SM | FB | SF |
| 5 (58) | 1.138 | 1.845 | 1.052 | 1.155 | 1.672 | 1.638 | 1.793 | 1.069 |
| 6 (77) | 1.26 | 2.13 | 1.156 | 1.351 | 1.961 | 1.922 | 2.169 | 1.208 |
| 7 (62) | 1.484 | 2.371 | 1.193 | 1.452 | 2.355 | 2.113 | 2.5 | 1.242 |
| 8 (57) | 1.456 | 2.842 | 1.228 | 1.86 | 2.86 | 2.614 | 3.017 | 1.509 |
| 9 (42) | 1.619 | 2.714 | 1.428 | 1.905 | 2.738 | 2.595 | 3.048 | 1.619 |
| 10 (35) | 1.628 | 3.086 | 1.371 | 2.143 | 3.257 | 3 | 3.457 | 1.743 |
| 11 (45) | 1.711 | 3.333 | 1.444 | 2.689 | 3.822 | 3.378 | 4.089 | 1.955 |
| 12 (45) | 1.778 | 3.489 | 1.622 | 2.867 | 4.4 | 3.578 | 4.467 | 2.178 |
| 13 (47) | 1.83 | 3.511 | 1.489 | 2.872 | 4.34 | 3.574 | 4.532 | 2 |
| 14 (74) | 2.081 | 3.676 | 2.013 | 3.27 | 4.73 | 3.905 | 4.811 | 2.608 |
| 15 (161) | 2.286 | 3.944 | 2.199 | 3.652 | 5.205 | 4.503 | 5.292 | 3.025 |
| 16 (2) | 1.5 | 4 | 1.5 | 4.5 | 5 | 4.5 | 5.5 | 3 |
| 17 (4) | 1.75 | 4.5 | 1.5 | 4.75 | 5.75 | 5.5 | 7 | 3.75 |
| 18 (1) | 3 | 5 | 3 | 4 | 7 | 5 | 6 | 3 |
| 19 (2) | 4 | 4 | 3.5 | 4 | 5.5 | 5 | 6.5 | 3.5 |

Table 5.5: The mean of the local controllability of protein complexes with a given size between 5 and 19 in the human TF regulatory networks of 8 cell types. The parentheses after each value of the complex size indicate the total number of protein complexes with size equal to that value. AE stands for Amniotic Epi., HSC for Hemat. Stem Cell, ADB for Adult Dermal Blood, Neu for Neuroblastoma, SM for Skeletal Myoblast, FB for Fetal Brain and SF for Skin Fib.

| Subset size | The mean (std) of $lc(G, S)$ | | | | | | | |
|-------------|------------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | hESC | AE | HSC | ADB | Neu | SM | FB | SF |
| 5 | 1.0938 (0.311) | 1.6028 (0.769) | 1.2118 (0.462) | 1.6286 (0.758) | 1.6352 (0.763) | 1.4048 (0.630) | 1.5088 (0.695) | 1.5206 (0.724) |
| 6 | 1.1302 (0.368) | 1.8382 (0.902) | 1.306 (0.556) | 1.8818 (0.895) | 1.8676 (0.877) | 1.5452 (0.743) | 1.7108 (0.825) | 1.7184 (0.829) |
| 7 | 1.1688 (0.430) | 2.0724 (1.011) | 1.4104 (0.649) | 2.1092 (0.987) | 2.1152 (0.980) | 1.6932 (0.845) | 1.9268 (0.909) | 1.9246 (0.932) |
| 8 | 1.2298 (0.497) | 2.3534 (1.112) | 1.5306 (0.727) | 2.3586 (1.089) | 2.3284 (1.053) | 1.882 (0.931) | 2.1338 (1.015) | 2.174 (1.042) |
| 9 | 1.285 (0.546) | 2.5778 (1.190) | 1.6274 (0.771) | 2.5686 (1.148) | 2.6042 (1.137) | 2.0818 (1.007) | 2.3672 (1.103) | 2.3904 (1.126) |
| 10 | 1.3448 (0.609) | 2.859 (1.266) | 1.742 (0.837) | 2.909 (1.259) | 2.8538 (1.179) | 2.2722 (1.089) | 2.5744 (1.156) | 2.6492 (1.190) |
| 11 | 1.4024 (0.658) | 3.1016 (1.309) | 1.871 (0.896) | 3.1234 (1.313) | 3.0772 (1.259) | 2.428 (1.148) | 2.8612 (1.244) | 2.8702 (1.253) |
| 12 | 1.5006 (0.720) | 3.3564 (1.394) | 2.0098 (0.952) | 3.4094 (1.389) | 3.343 (1.354) | 2.658 (1.226) | 3.0296 (1.305) | 3.1208 (1.335) |
| 13 | 1.5596 (0.773) | 3.6344 (1.476) | 2.1408 (1.006) | 3.66 (1.427) | 3.5962 (1.390) | 2.8148 (1.300) | 3.291 (1.357) | 3.3418 (1.405) |
| 14 | 1.611 (0.782) | 3.8934 (1.532) | 2.2666 (1.047) | 3.916 (1.480) | 3.8418 (1.436) | 3.0046 (1.305) | 3.5328 (1.420) | 3.5762 (1.437) |
| 15 | 1.7096 (0.849) | 4.1964 (1.615) | 2.3814 (1.082) | 4.2214 (1.604) | 4.1424 (1.505) | 3.2046 (1.373) | 3.7726 (1.473) | 3.7842 (1.507) |
| 16 | 1.7618 (0.871) | 4.4584 (1.651) | 2.5428 (1.138) | 4.439 (1.623) | 4.3416 (1.531) | 3.398 (1.401) | 3.9896 (1.515) | 4.0732 (1.584) |
| 17 | 1.833 (0.900) | 4.7398 (1.719) | 2.6732 (1.164) | 4.7182 (1.681) | 4.616 (1.564) | 3.5966 (1.477) | 4.212 (1.578) | 4.3138 (1.639) |
| 18 | 1.9602 (0.953) | 4.962 (1.755) | 2.806 (1.206) | 4.9528 (1.700) | 4.8446 (1.637) | 3.7714 (1.503) | 4.4592 (1.597) | 4.5396 (1.673) |
| 19 | 2.0254 (0.974) | 5.2308 (1.831) | 2.9218 (1.230) | 5.2366 (1.778) | 5.113 (1.697) | 4.0028 (1.561) | 4.6688 (1.662) | 4.798 (1.712) |

Table 5.6: The mean and standard deviation of the local controllability of random subsets with a given size between 5 and 19 in the human TF regulatory networks of 8 cell types. The parentheses below each mean indicate the corresponding standard deviation. AE stands for Amniotic Epi., HSC for Hemat. Stem Cell, ADB for Adult Dermal Blood, Neu for Neuroblastoma, SM for Skeletal Myoblast, FB for Fetal Brain and SF for Skin Fib.

We noticed that the transcription factor ‘ZNF354C’ is not in the node set of the TF regulatory network of hESC. Thus, a different way was also used to calculate the local controllability of a protein complex that contains ‘ZNF354C’ in the TF regulatory network of hESC: the local controllability of this protein complex was calculated by ignoring the TF ‘ZNF354C’ (see Appendix A). Based on the local controllability calculated in the different way (by ignoring ‘ZNF354C’), the average local controllability of protein complexes with a given size in the TF regulatory network of hESC was obtained, and the results are summarized in Table 5.7.

| Complex size | The mean of $lc(G, S)$ |
|--------------|------------------------|
| 5 (58) | 1 |
| 6 (77) | 1.026 |
| 7 (62) | 1.032 |
| 8 (57) | 1.018 |
| 9 (42) | 1.119 |
| 10 (35) | 1.029 |
| 11 (45) | 1.111 |
| 12 (45) | 1.156 |
| 13 (47) | 1.128 |
| 14 (74) | 1.324 |
| 15 (161) | 1.46 |
| 16 (2) | 1.5 |
| 17 (4) | 1.5 |
| 18 (1) | 2 |
| 19 (2) | 3 |

Table 5.7: The mean of the local controllability of protein complexes with a given size between 5 and 19 in the TF regulatory network of hESC, calculated by ignoring ‘ZNF354C’. The parentheses after each value of the complex size indicate the total number of protein complexes with size equal to that value.

Since the TF ‘ZNF354C’ is not in the node set of the TF regulatory network of hESC, the local controllability of each protein complex in the TF regulatory network of hESC calculated in the different way (by ignoring ‘ZNF354C’) will be used for our study. Thus, for the TF regulatory network of hESC, the results given in Table 5.7 will be used.

Comparison between Table 5.7 and the column of hESC in Table 5.6 indicates

that in the TF regulatory network of hESC, the mean of the local controllability of protein complexes is smaller than that of random subsets with the same size.

For the 7 other cell types, from Table 5.5 and Table 5.6, we find that in the TF regulatory networks of HSC, ADB, and SF, the mean of the local controllability of protein complexes is smaller than that of random subsets with the same size; while in the TF regulatory networks of Neu, SM and FB, the mean of the local controllability of protein complexes is larger than that of random subsets with the same size. Note that, in these comparisons, we only considered protein complex sizes $s = 5, 6, \dots, 15$, since there are only very few protein complexes (in the list) with each of the sizes $s = 16, 17, 18, 19$. In the TF regulatory network of AE, there is a mixed behavior: for sizes $s = 5, 6, 7, 8, 9, 10, 11, 12$, the mean of the local controllability of protein complexes is larger than that of random subsets with the same size; for sizes $s = 13, 14, 15$, the former is smaller than the latter.

Apart from direct comparison between the mean of the local controllability of protein complexes and that of random subsets with the same size, we also calculated the p -value. For any given size s ($s = 5, 6, \dots, 15$), in the TF regulatory network of each of the 8 cell types, we used the Wilcoxon rank-sum (one-tailed) test to calculate the p -value. The p -value is used to test how significantly smaller or significantly larger is the median of the local controllability of protein complexes with size s than that of random subsets with the same size s . The results for the two types of p -values are given in Table 5.8 and Table 5.9.

As shown in Table 5.8 and Table 5.9, in the TF regulatory networks of hESC, HSC, ADB and SF, the local controllability of protein complexes tend to be significantly smaller than the local controllability of random subsets with the same size ($p < 0.05$); while in the TF regulatory networks of Neu, SM and FB, the local controllability of protein complexes tend to be significantly larger than those of random subsets with the same size ($p < 0.05$). Note that, there are two p -values (0.168 and 0.199) in the column of Neu in Table 5.9 that are insignificant.

| Complex size | The p -value | | | | | | | |
|--------------|----------------|-------|----------|----------|-------|----|----|----------|
| | hESC | AE | HSC | ADB | Neu | SM | FB | SF |
| 5 | 8.98E-03 | 0.999 | 3.60E-03 | 3.24E-07 | 0.832 | 1 | 1 | 1.75E-07 |
| 6 | 5.80E-03 | 1 | 5.12E-03 | 1.77E-08 | 0.965 | 1 | 1 | 3.41E-09 |
| 7 | 5.20E-03 | 0.999 | 6.32E-03 | 1.38E-08 | 0.992 | 1 | 1 | 3.02E-10 |
| 8 | 3.14E-04 | 1 | 1.01E-03 | 2.85E-04 | 1 | 1 | 1 | 3.12E-07 |
| 9 | 2.76E-02 | 0.914 | 8.42E-02 | 2.45E-05 | 0.801 | 1 | 1 | 1.10E-06 |
| 10 | 4.47E-04 | 0.935 | 6.20E-03 | 5.06E-05 | 0.984 | 1 | 1 | 6.26E-07 |
| 11 | 1.06E-03 | 0.953 | 7.27E-04 | 1.37E-02 | 1 | 1 | 1 | 9.44E-08 |
| 12 | 2.74E-04 | 0.885 | 3.10E-03 | 4.50E-03 | 1 | 1 | 1 | 2.36E-07 |
| 13 | 2.06E-05 | 0.380 | 1.48E-06 | 1.59E-05 | 1 | 1 | 1 | 7.78E-13 |
| 14 | 1.19E-03 | 0.193 | 3.89E-02 | 5.70E-05 | 1 | 1 | 1 | 3.09E-10 |
| 15 | 1.47E-04 | 0.057 | 4.32E-02 | 2.90E-06 | 1 | 1 | 1 | 1.57E-11 |

Table 5.8: The p -value for the alternative hypothesis that the median of the local controllability of protein complexes is smaller than the median of the local controllability of random subsets with the same size, in the human TF regulatory networks of 8 cell types. We used the Wilcoxon rank sum (one-tailed) test to calculate the p -value.

| Complex size | The p -value | | | | | | | |
|--------------|----------------|----------|-------|-------|----------|----------|----------|----|
| | hESC | AE | HSC | ADB | Neu | SM | FB | SF |
| 5 | 0.991 | 6.90E-04 | 0.996 | 1 | 0.168 | 2.37E-04 | 2.67E-04 | 1 |
| 6 | 0.994 | 5.31E-05 | 0.995 | 1 | 3.50E-02 | 8.44E-08 | 7.62E-09 | 1 |
| 7 | 0.995 | 7.05E-04 | 0.994 | 1 | 7.61E-03 | 4.12E-07 | 1.47E-07 | 1 |
| 8 | 1 | 4.76E-05 | 0.999 | 1 | 3.61E-05 | 1.11E-10 | 1.76E-10 | 1 |
| 9 | 0.972 | 8.57E-02 | 0.916 | 1 | 0.199 | 1.67E-04 | 1.41E-05 | 1 |
| 10 | 1 | 6.51E-02 | 0.994 | 1 | 1.56E-02 | 1.81E-05 | 2.20E-06 | 1 |
| 11 | 0.999 | 4.74E-02 | 0.999 | 0.986 | 2.37E-05 | 2.51E-09 | 5.04E-12 | 1 |
| 12 | 1 | 0.115 | 0.997 | 0.995 | 4.68E-09 | 2.28E-08 | 1.44E-12 | 1 |
| 13 | 1 | 0.620 | 1 | 1 | 3.98E-06 | 8.27E-07 | 8.27E-12 | 1 |
| 14 | 0.999 | 0.807 | 0.961 | 1 | 5.14E-10 | 8.33E-11 | 5.13E-15 | 1 |
| 15 | 1 | 0.943 | 0.957 | 1 | 1.00E-21 | 7.98E-35 | 1.19E-38 | 1 |

Table 5.9: The p -value for the alternative hypothesis that the median of the local controllability of protein complexes is larger than the median of the local controllability of random subsets with the same size, in the human TF regulatory networks of 8 cell types. We used the Wilcoxon rank sum (one-tailed) test to calculate the p -value.

Table 5.8 and Table 5.9 also show that in the TF regulatory network of AE, there seems to be no significant difference between the local controllability of protein complexes and those of random subsets with the same size, except for protein complex sizes $s = 5, 6, 7, 8$.

Therefore, in the TF regulatory networks of hESC, HSC, ADB and SF, protein complexes are easier to control than random subsets with the same size (with statistical significance); while in the TF regulatory networks of Neu, SM and FB, protein complexes are more difficult to control than random subsets with the same size. In the TF regulatory network of AE, we do not have a definite conclusion.

The above analysis only compared the local controllability of protein complexes with the local controllability of random subsets at a population (distribution) level, i.e., for a given complex size s , we compared a set of protein complexes with size s with a set of random subsets with size s .

Next, we compared the local controllability of each (individual) protein complex with the local controllability of 5000 random subsets with the same size as the protein complex in the TF regulatory network of each of the 8 cell types, and calculated the p -value. Denoting the local controllability of the protein complex by lc_0 , and the local controllability of 5000 random subsets by a vector $d = (lc_1, lc_2, \dots, lc_{5000})$, we then calculated two (one-tailed) p -values: p_1 and p_2 . The p -value p_1 (p_2) is given by the fraction of elements lc_i in the vector d such that $lc_i \leq lc_0$ ($lc_i \geq lc_0$). The p -value p_1 (p_2) is used to test how significantly smaller (larger) is the local controllability of the protein complex than those of random subsets with the same size. The results of the two types of p -values (p_1 and p_2) for protein complexes in the TF regulatory networks of 8 cell types are shown in Table B.1-B.8 in Appendix B.

We have identified all the protein complexes that have significant p -values ($p_1 < 0.05$ or $p_2 < 0.05$), based on Table B.1-B.8 in Appendix B. We found that there is only one protein complex that is significantly easier to control in the TF regulatory network of Adult Dermal Blood, and there are three protein complexes that are significantly easier to control in the TF regulatory network of Skin Fib.

On the other hand, the protein complexes that are significantly more difficult to control include: 1 protein complex in hESC, 3 protein complexes in Amniotic Epi., 3 protein complexes in Hemat. Stem Cell, 19 protein complexes in Neuroblastoma, 16 protein complexes in Skeletal Myoblast and 53 protein complexes in Fetal Brain.

Therefore, there are only very few protein complexes that are significantly easier to control than random subsets ($p_1 < 0.05$), and there are some protein complexes that are significantly more difficult to control than random subsets ($p_2 < 0.05$). This may suggest that overall, protein complexes are difficult to control in the human TF regulatory networks.

In the following, we discuss some specific protein complexes that are significantly easier or more difficult to control. For each of these protein complexes, we discuss the functions of this protein complex in relation to the functions of the corresponding cell type in which this protein complex has significant p -value.

The protein complex that is significantly easier to control in the TF regulatory network of Adult Dermal Blood is HC8898 with size 14 and with local controllability 1. The three protein complexes that are significantly easier to control in the TF regulatory network of Skin Fib. are HC4783, HC9284 and HC8674 with size 15 and with local controllability 1 (see Table A.1 in Appendix A for the IDs of protein complexes). These four protein complexes all have general functions in regulating protein modification [97]. These general roles of regulation might suggest that these protein complexes need to be easily activated (easy to control) in order for cells to rapidly adapt and respond to external signals.

For the protein complexes that are significantly more difficult to control in the TF regulatory network of a given cell type, such as Hemat. Stem Cell, Fetal brain and Skeletal Myoblast, we found that some of the protein complexes have functions related to the functions of the corresponding cell type.

Hematopoietic stem cells (HSC) are multipotent, self-renewing progenitor cells that can give rise to all types of blood cells. Multi-potency is the ability to differentiate into all functional blood cells. Self-renewal is the ability to give rise to HSC

itself without differentiation. The protein complex HC6967 is significantly more difficult to control in the TF regulatory network of Hemat. Stem Cell, and one of the functions of this protein complex is response to wounding [97] which is related to hematopoietic stem cells. We hypothesize that this protein complex might be robust (insensitive) to be controlled in hematopoietic stem cells. In other words, in normal situations, a small number of arbitrary regulatory signals is not able to turn on this protein complex. Only in the situation of wounding is this protein complex triggered to regulate the hematopoietic stem cell differentiation for blood cell supply.

Fetal brain (FB) cells play an essential role in the rapid growth of the brain in the prenatal period. They are responsible for production of new neurons, and migrate to different brain areas and differentiate to perform specialised functions. Among the protein complexes which are significantly more difficult to control in the TF regulatory network of Fetal Brain, we have identified several protein complexes with particular functions in the proliferation and differentiation of the neurons. HC1009, for example, has a role in regulating axon guidance and cell morphogenesis involved in neuron differentiation [97]. HC9168, on the other hand, regulates cell division [97]. These two protein complexes are not easily activated in fetal brain cells, and thus might be robust to be controlled to perform the functions related to proliferation and differentiation of the neurons.

Skeletal myoblasts (SM) are embryonic progenitor cells that differentiate to give rise to muscle cells and therefore, have been an important source of donor cells for transplantation, to augment myocardial function. HC3943 is a protein complex that is significantly more difficult to control in the TF regulatory network of Skeletal Myoblast. It has functions in muscle organ development, muscle structure development and actin filament-based process, which are important for the differentiation of skeletal myoblasts [97]. This might indicate that the protein complex HC3943 is robust to be controlled in skeletal myoblasts to perform the cell type-related functions.

In summary, there are few protein complexes that are significantly easier to control, and there are some protein complexes that are significantly more difficult

to control in given cell types. For the protein complexes that are significantly easier to control, they carry out some general functions of regulation which might not be directly related to the corresponding cell types. For some protein complexes that are significantly more difficult to control, their functions are related to the functions of the corresponding cell types. A further detailed biological investigation into the roles of these protein complexes in specific cellular events in their corresponding cell types, might yield insights into the relationship between the local controllability and the cell-type-specific functionality of protein complexes.

Chapter 6

Local controllability of model networks

In this chapter, we will introduce model networks and study the application of local controllability to model networks. We study local controllability of two types of model networks: Erdős–Rényi random networks and scale-free networks. All the model networks we generated are directed, since local controllability is only applicable to directed networks.

6.1 Model networks

In this section, we describe two types of model networks: Erdős–Rényi random networks and scale-free networks. These two types of model networks are commonly used as benchmark networks in reference to real networks, in other words, real networks can be compared with them to see whether real networks fall into these two types, or show significantly different behavior. We will discuss network models to generate these two types of model networks and review some network properties of the model networks.

6.1.1 Erdős–Rényi random network

An *Erdős–Rényi (ER) random network* is a network in which any two nodes are connected randomly, with the same probability [24]. Similarly, a *directed ER random network* is a directed network in which there is a directed edge from each node to each other node with the same probability. We can also simply call ER random networks by random networks.

There are two closely related ways to generate ER random networks: $G(N, L)$ model and $G(N, p)$ model, both called *Erdős–Rényi (ER) models*.

$G(N, L)$ model: a graph is constructed with N labelled nodes and L edges that are connected uniformly at random.

$G(N, p)$ model: a graph is constructed with N labelled nodes and any two nodes are connected with probability p .

ER models can be used to generate directed ER random networks as well, with nodes being connected by directed edges (uniformly at random) instead.

Note that, random networks generated by $G(N, L)$ model have a fixed number of edges, while the number of edges in random networks generated by $G(N, p)$ model is a variable. We will only discuss properties of random networks generated by $G(N, p)$ model in the following.

The expected number of edges in a random network is $\langle L \rangle = p \cdot \frac{N(N-1)}{2}$. The expected value of the average degree in a random network is $\langle k \rangle = p(N-1)$.

The degree distribution of a random network follows a binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \quad (6.1)$$

where p_k is the probability that a node has degree k .

In the large N limit with $\langle k \rangle \ll N$, the binomial distribution (6.1) is well approximated by a Poisson distribution

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \quad (6.2)$$

Due to analytical simplicity, the Poisson distribution (6.2) is preferred to describe the degree distribution of a random network with $\langle k \rangle \ll N$. The condition $\langle k \rangle \ll N$ is easily satisfied for real networks as most real networks are sparse. Nevertheless, the degree distribution of real networks usually deviates significantly from a Poisson distribution. The graphs of Poisson distributions are shown in Figure 6.1.

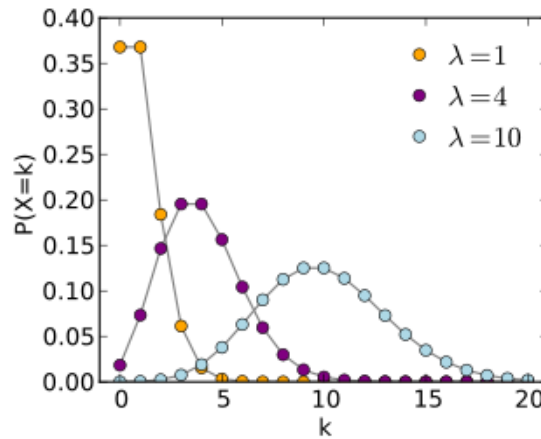


Figure 6.1: Poisson degree distributions for ER random networks with different parameters $\lambda = \langle k \rangle$. The figure is from http://en.wikipedia.org/wiki/Poisson_distribution.

Random networks lack hubs (nodes with very high degree), as the probability p_k decreases exponentially when k increases. Since the standard deviation of the Poisson distribution (6.2) is $\sigma = \langle k \rangle^{\frac{1}{2}}$, the degrees of most nodes in a large random network are within a small vicinity of $\langle k \rangle$, that is, ER random networks are homogeneous.

The average path length \bar{d} of a random network is approximately given by [11]

$$\bar{d} \propto \frac{\ln N}{\ln \langle k \rangle}.$$

The clustering coefficient C_i of each node i is the same as the global clustering coefficient C in a random network, which is given by

$$C_i = C = p = \frac{\langle k \rangle}{N-1}.$$

Therefore, in the case $\langle k \rangle \ll N$, ER random networks have small average path length, and small clustering coefficient (poorly clustered) [100]. Since most

real networks are highly clustered, ER random network models can not explain the structure of real networks. The properties of ER random networks are summarized in Table 6.1.

| | |
|---|---|
| Degree distribution | $p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$ |
| Average degree | $\langle k \rangle = p(N - 1)$ |
| Standard deviation of degree distribution | $\sigma = \langle k \rangle^{\frac{1}{2}}$ |
| Average length path | $\bar{d} \propto \frac{\ln N}{\ln \langle k \rangle}$ |
| Clustering coefficient | $C_i = C = p = \frac{\langle k \rangle}{N-1}$ |

Table 6.1: Some properties of ER random networks.

6.1.2 Scale-free network

If real networks can not be characterized by ER random network models, what should the structure of real networks look like? It has been found that many real networks are characterized by a fat-tail degree distribution, which can be described by a power-law degree distribution.

Definition 6.1. A *power-law* degree distribution is given by

$$p_k \sim k^{-\gamma}, k = 1, 2, \dots, \quad (6.3)$$

where p_k is the fraction of nodes with degree k , and γ is called the *degree exponent*.

If we take logarithm on both sides of (6.3), we obtain

$$\log p_k \sim -\gamma \log k.$$

Therefore, if a power-law degree distribution is plotted on a log-log scale, the plot should be observed as an approximate straight line with slope $-\gamma$, since $\log p_k$ depends linearly on $\log k$. The plot for a power-law degree distribution is shown in Figure 6.2.

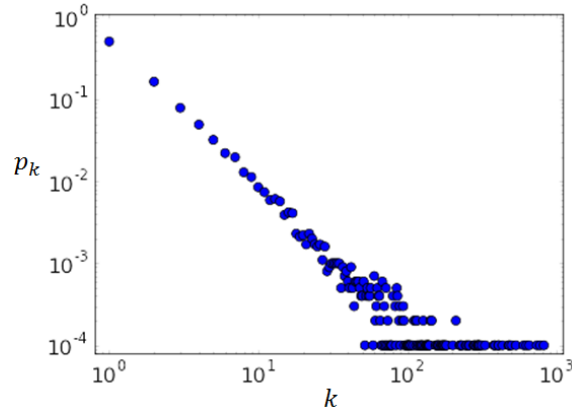


Figure 6.2: A power-law degree distribution in a log-log plot. The figure is from http://mathinsight.org/image/power_law_degree_distribution_scatter.

Definition 6.2. A network is called a *scale-free* network, if its degree distribution follows a power-law distribution.

Many real networks have been found to be scale-free, with degree exponents $2 < \gamma < 3$. In a scale-free network, there are a few nodes with very high degree (hubs) and also a large number of nodes with small degree.

The n -th moment of the degree distribution of a scale-free network is

$$\langle k^n \rangle = \int_{k_{min}}^{k_{max}} k^n p(k) dk = \frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n - \gamma + 1}. \quad (6.4)$$

According to (6.4), in the limit $k_{max} \rightarrow \infty$, $\langle k^n \rangle$ is finite if $n < \gamma - 1$; $\langle k^n \rangle$ is infinite if $n > \gamma - 1$. Therefore, for a scale-free network with $\gamma < 3$, in the limit $N \rightarrow \infty$, the first moment $\langle k \rangle$ is finite, but the second moment $\langle k^2 \rangle$ is infinite, thus, the standard deviation $\sigma = \sqrt{\langle k^2 \rangle - \langle k \rangle^2}$ is infinite.

Even though real networks can not be of infinite size, it is expected that σ is considerably larger than $\langle k \rangle$ in a large scale-free network with $\gamma < 3$. In other words, a scale-free network has a heterogeneous (wide) degree distribution, whereas a random network has a homogeneous degree distribution as most nodes have degrees within a narrow vicinity of $\langle k \rangle$ ($\sigma = \langle k \rangle^{\frac{1}{2}}$). Moreover, the smaller is γ , the more heterogeneous is the degree distribution of a scale-free network.

The presence of hubs in scale-free networks can contribute to the small-world

property: the average path length in a scale-free network is either smaller than or equal to that in a random network with the same number of nodes and the same number of edges as the scale-free network. The dependence of average path length \bar{d} on network size N and degree exponent γ can be captured by the following expressions [18]

$$\bar{d} \sim \begin{cases} \frac{\ln \ln N}{\ln(\gamma-1)} & 2 < \gamma < 3, \\ \frac{\ln N}{\ln \ln N} & \gamma = 3, \\ \ln N & \gamma > 3. \end{cases} \quad (6.5)$$

Scale-free networks with $2 < \gamma < 3$ are called ‘ultra-small world’, as the presence of hubs can radically reduce the average path length. Generally, in a scale-free network, the smaller is γ , the smaller is the average path length \bar{d} .

In the following, we will describe two network models to generate scale-free networks: the Barabási–Albert (BA) model [8] and the static model [32].

Barabási–Albert model: Start with m_0 nodes, the links between which are arbitrary as long as each node has at least one link (suppose there are l_0 links in this initial network). At each time step, we add a new node with m ($\leq m_0$) links that connect the new node to m nodes that already exist. The probability $\Pi(k_i)$ that the new node connects to node i depends linearly on the degree k_i of node i as

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \quad (6.6)$$

After t time steps, the network has $m_0 + t$ nodes and $l_0 + mt$ links.

Static model: Start with N isolated nodes, which are labelled by $1, 2, \dots, N$. We assign the weight $p_i = i^{-\alpha}$ to each node i , where α is a parameter in $[0, 1)$. Next, we randomly select two different nodes i and j with probabilities equal to the normalized weights, $\frac{p_i}{\sum_k p_k}$ and $\frac{p_j}{\sum_k p_k}$, respectively. Add an edge between nodes i and j unless there is already an edge between them. This process is repeated zN times until zN edges are present in the network.

We mention here that these two models can be modified to generate directed scale-free networks. For the directed version of BA model, the only difference is that in the network growth, directed arcs rather than undirected links are added from the new node to the existing nodes at each time step. The directed version of the static model can be described as follows [32].

Directed static model: Start with N isolated nodes, which are labelled by $1, 2, \dots, N$. Assign two weights $p_i = i^{-\alpha_{out}}$ and $q_i = i^{-\alpha_{in}}$ to each node i , where the parameters α_{out} and α_{in} are both in $[0, 1)$. Then, randomly select two different nodes i and j with probabilities $\frac{p_i}{\sum_k p_k}$ and $\frac{q_j}{\sum_k q_k}$, respectively. A directed edge is connected from node i to j if this directed edge $i \rightarrow j$ is not available. Repeat the process zN times until zN directed edges are present in the network.

For the static model, the resulting network is scale-free, with a degree distribution $p_k \sim k^{-\gamma}$, where the degree exponent γ is given by $\gamma = 1 + \frac{1}{\alpha}$. By adjusting the parameter α in $[0, 1)$, we can obtain values for γ such that $2 < \gamma < \infty$. Similarly, in the directed static model, the resulting directed network has a power law distribution for both out-degree and in-degree, the degree exponents of which are $\gamma_{out} = 1 + \frac{1}{\alpha_{out}}$ and $\gamma_{in} = 1 + \frac{1}{\alpha_{in}}$, respectively.

The BA network model has two important ingredients: *growth* and *preferential attachment*. In the BA model, the network grows as a new node is added in each time step, and the new node is preferentially attached to nodes that already have high degree as shown in (6.6). Neither of these two ingredients can be neglected in the BA model in order for the network to evolve to a scale-free state [8].

In the BA model, the degree dynamics of each node i follows $k_i(t) = m(\frac{t}{t_i})^{0.5}$, where t_i the time at which node i is added to the network, and $k_i(t)$ is the degree of node i at time t . Thus, the earlier node i is added, the higher is the degree $k_i(t)$. Therefore, older nodes have advantage over younger nodes in acquiring links, eventually turning into hubs.

Networks generated by the BA model are called *BA networks*, which have a

power-law degree distribution. In the BA model, over time, the network evolves into a stationary degree distribution (independent of time t) that can be exactly given by [22]

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}.$$

Therefore, the degree exponent of the power-law degree distribution of BA networks is $\gamma = 3$, independent of the choice of m and m_0 . According to (6.5), the average path length of BA networks scales as

$$\bar{d} \sim \frac{\ln N}{\ln \ln N}.$$

The global clustering coefficient of BA networks is given by [49]

$$C = \frac{m-1}{8} \frac{(\ln N)^2}{N}.$$

The properties of BA networks are summarized in Table 6.2.

| | |
|------------------------|---|
| Degree distribution | $p_k = \frac{2m(m+1)}{k(k+1)(k+2)}$ |
| Average length path | $\bar{d} \sim \frac{\ln N}{\ln \ln N}$ |
| Clustering coefficient | $C = \frac{m-1}{8} \frac{(\ln N)^2}{N}$ |

Table 6.2: Some properties of BA networks.

6.2 Local controllability v.s. degree

We know that the local controllability $lc(G, S)$ generally depends on both the directed network G and the subset S . Thus, we first want to study how the local controllability $lc(G, S)$ depends on the subset S when G is fixed.

For such purposes, we studied how local controllability is related to various network centrality measures in model networks. Specifically, we generate a model network G , and then make different choices of subset S based on centrality measures, in order to find the relationship between $lc(G, S)$ and S .

In this section, we study how local controllability is related to the degree in both ER random networks and scale-free networks.

6.2.1 ER random networks

We first generated 100 directed ER random networks with $n = 900$ nodes and each of the parameters for average degree $\langle k \rangle = 2, 3, \dots, 8$. For each network G generated, all the nodes were divided into three equal-sized subsets based on the order of the node degrees: low-degree, medium-degree, and high-degree. For each of the three subsets, S , we calculated the local controllability $lc(G, S)$.

Then, $lc(G, S)$ was averaged over all 100 network realizations for low-degree, medium-degree and high-degree subsets. We also calculated the standard deviation of $lc(G, S)$ in the 100 network realizations. The results are summarized in a bar plot, shown in Figure 6.3.

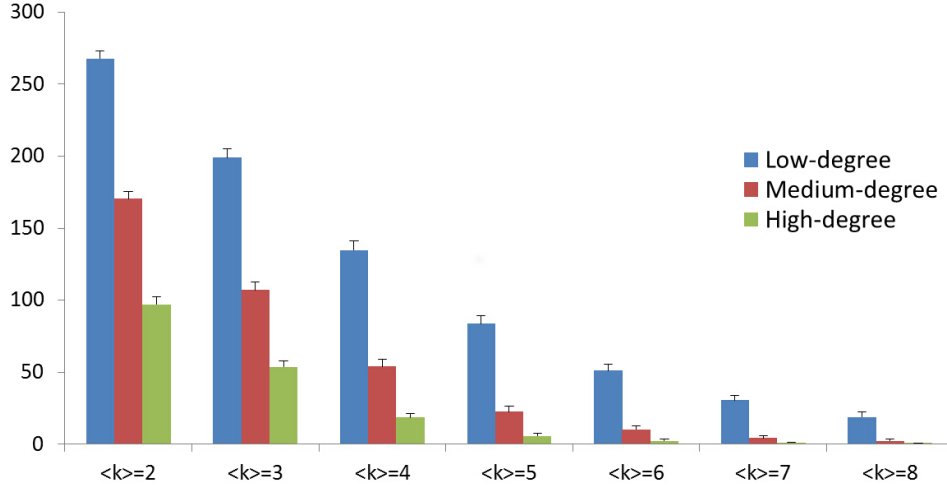


Figure 6.3: The local controllability of low-degree, medium-degree and high-degree subsets in ER random networks. The error bar indicates one standard deviation away from the mean.

Figure 6.3 shows that the local controllability $lc(G, S)$ is smallest for high-degree subsets, and largest for low-degree subsets.

To further verify this observation, we generated directed ER random networks

with $n = 1000$ nodes and each of the parameters for average degree $\langle k \rangle = 3, 4, \dots, 8$. For each parameter $\langle k \rangle$, 10 ER random networks were generated. For each network G , all the nodes were divided into 10 equal-sized subsets based on the order of the node degrees/in-degrees/out-degrees. For each of these 10 subsets, S , we calculated $lc(G, S)$ and the average degree/in-degree/out-degree of S . Then, we plotted the average $lc(G, S)$ versus the mean of the average degree/in-degree/out-degree of S over all 10 network realizations. These plots are given in Figure 6.4-6.9.

One can observe a decreasing trend for $lc(G, S)$ as a function of the average degree of S in each of these plots, that is, a smaller number of inputs are required to control a subset of nodes with higher degree. This further shows that nodes with higher degree are easier to control in ER random networks.

On the other hand, there is no detectable monotone trend for the plot of local controllability versus average in-degree or out-degree of 10 subsets. Therefore, we can not conclude that local controllability is related to in-degree or out-degree, in ER random networks.

A possible explanation might be that in directed ER random networks, nodes with higher degree have greater tendency to be connected to each other through a larger number of different paths, and thus less inputs are needed to control them (as there are a smaller number of paths in a path-cycle cover of the higher-degree nodes).

This behavior may not be true for in-degree or out-degree alone. In directed ER random networks, there is no correlation between the in-degree and the out-degree of a node, and a node with large in-degree (out-degree) may only have small out-degree (in-degree), therefore, nodes with high in-degree (or high out-degree) might be prevented from being connected to each other through a large number of different paths.

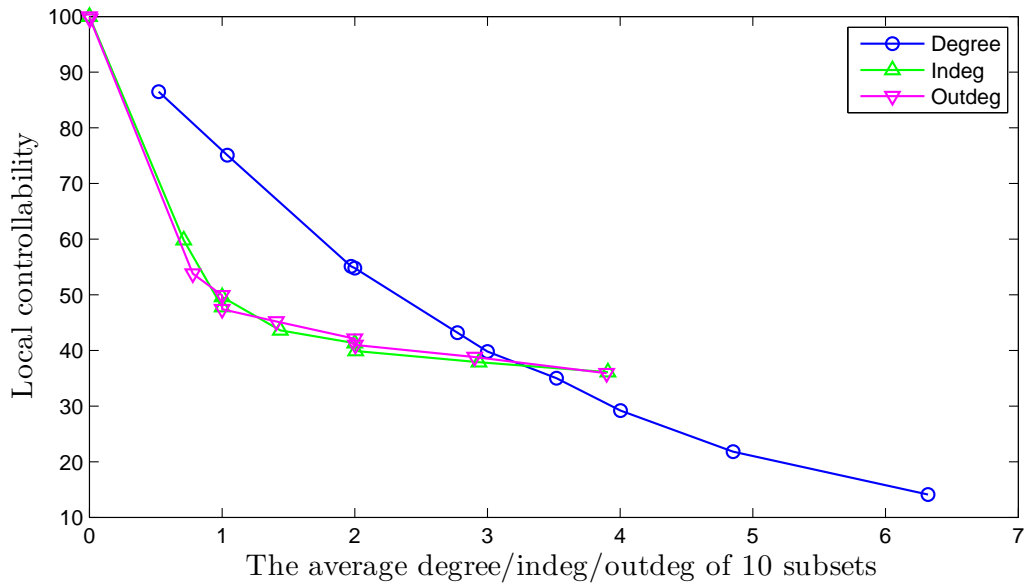


Figure 6.4: The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 3$.

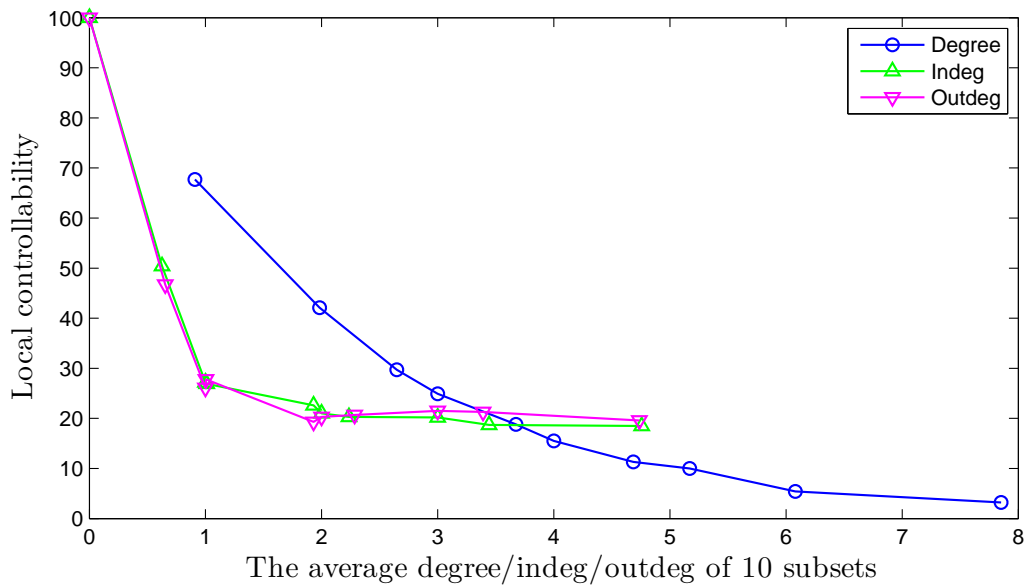


Figure 6.5: The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 4$.

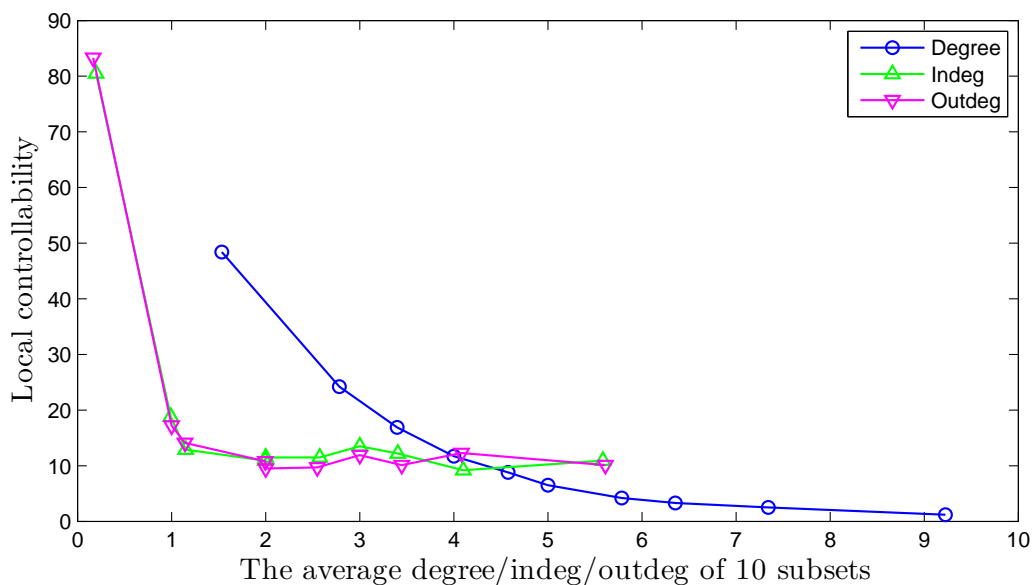


Figure 6.6: The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 5$.

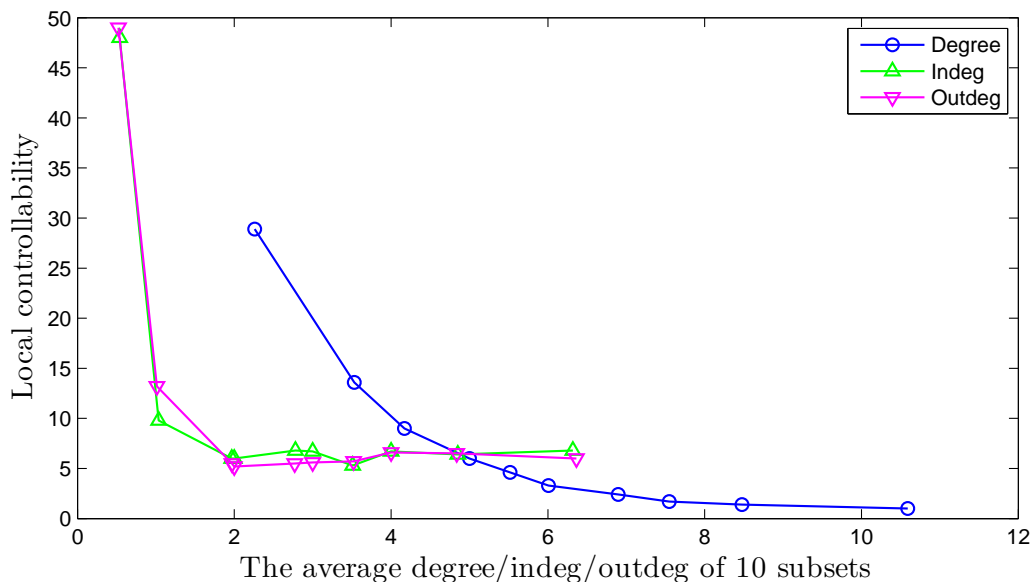


Figure 6.7: The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 6$.

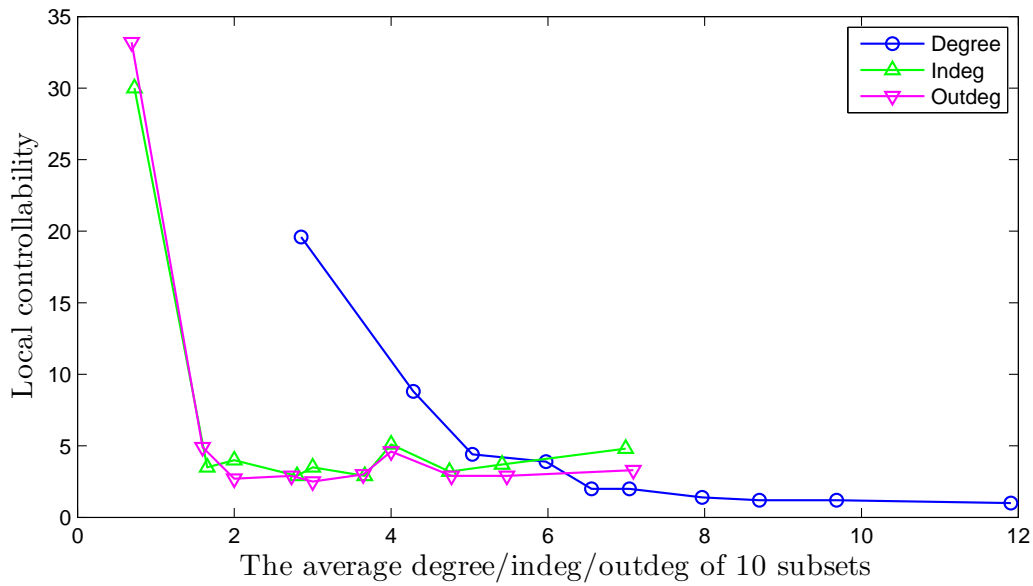


Figure 6.8: The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 7$.

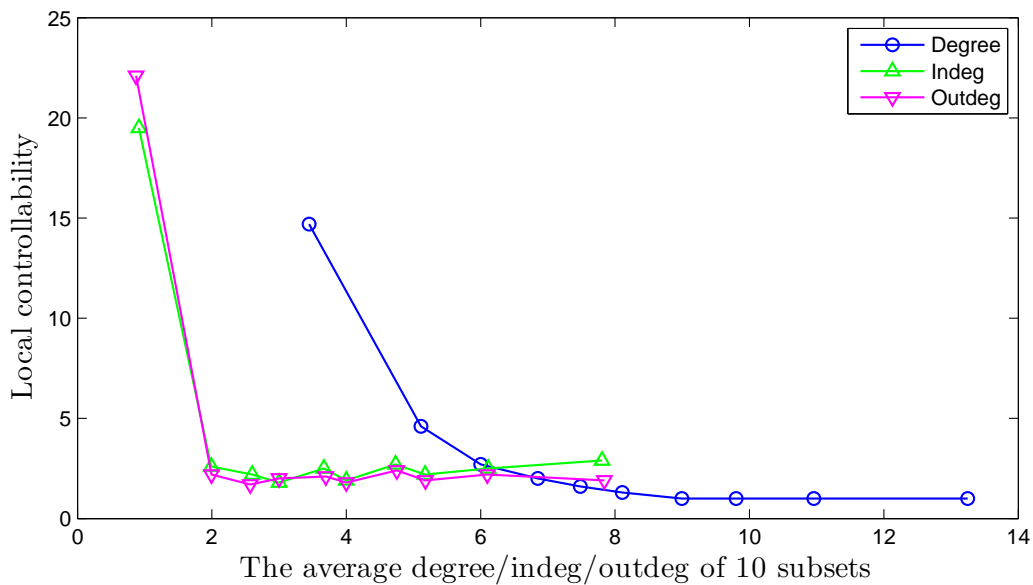


Figure 6.9: The plot of local controllability versus average degree/in-degree/out-degree of 10 subsets in ER random networks with $n = 1000$ and $\langle k \rangle = 8$.

6.2.2 Scale-free networks

We also performed similar numerical experiments in directed scale-free networks. We generated scale-free networks with $n = 1000$ nodes and varying parameters for average degree $\langle k \rangle$ and degree exponent γ . For each combination of parameters $\langle k \rangle$ and γ , we take the in-degree exponent γ_{in} and the out-degree exponent γ_{out} to be both equal to γ , that is, $\gamma_{in} = \gamma_{out} = \gamma$. Then, using the parameters $\langle k \rangle$, γ_{in} and γ_{out} , we generated 10 directed scale-free networks based on the directed static model (see Subsection 6.1.2).

For each scale-free network G generated, all the nodes were divided into 10 equal-sized subsets based on the order of the node degrees. For each of these 10 subsets, S , the local controllability $lc(G, S)$ was calculated together with the average degree of S . Then, $lc(G, S)$ and average degree of S were averaged over all 10 network realizations, and the relationships between $lc(G, S)$ and average degree of S were plotted, as shown in Figure 6.10-6.12.

These plots all show a decreasing trend (also true for plots for scale-free networks with other parameters, not shown here), indicating that higher-degree nodes tend to be easier to control in scale-free networks. In scale-free networks generated by the directed static model, it is expected that nodes with higher degree tend to be more interconnected (as two higher-degree nodes are more likely to be connected by a directed edge) than nodes with smaller degree, thus less inputs might be required to control them.

It can also be found in these figures, that when the degree exponent γ becomes larger (hence scale-free networks become less heterogeneous) with $\langle k \rangle$ unchanged, the local controllability of the subset of nodes with the smallest degrees in scale-free networks becomes smaller.

In summary, we did numerical experiments to examine the relationship between local controllability and the degree measure, and found that higher-degree nodes tend to be easier to control in both ER random networks and scale-free networks.

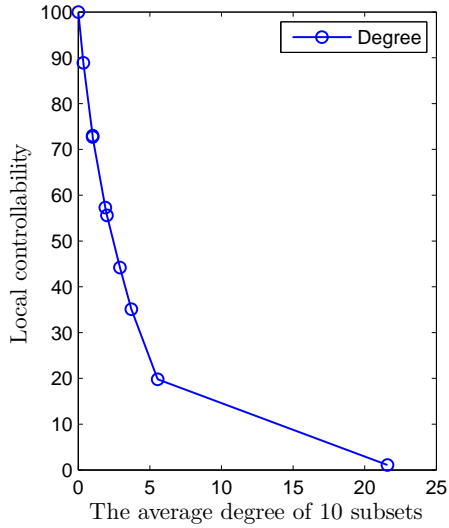
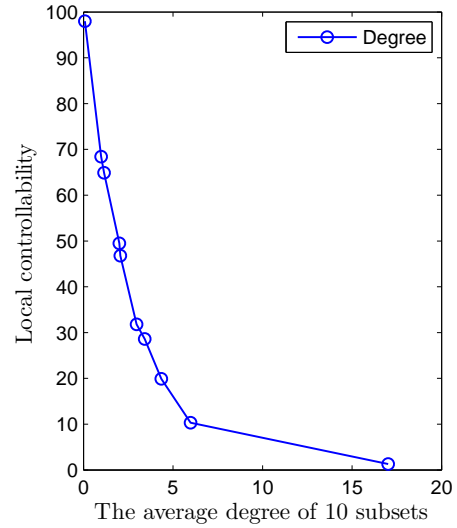
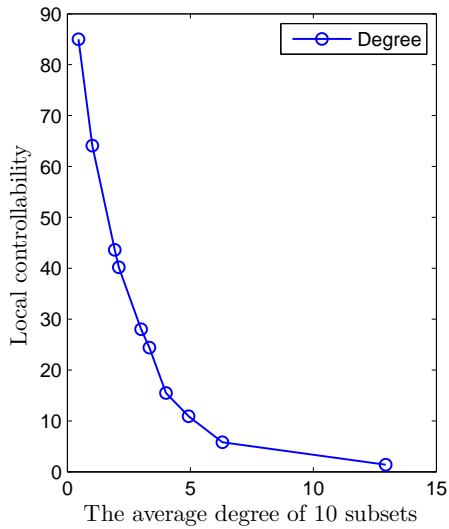
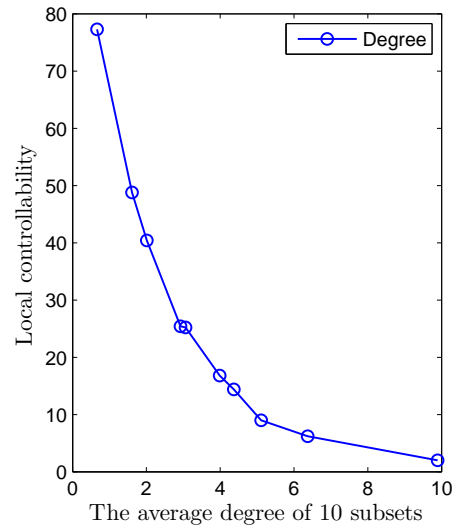
(a) $\gamma = 2.2$ (b) $\gamma = 2.5$ (c) $\gamma = 3$ (d) $\gamma = 4$

Figure 6.10: The plot of local controllability versus average degree of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 2.2, 2.5, 3$ and 4 .

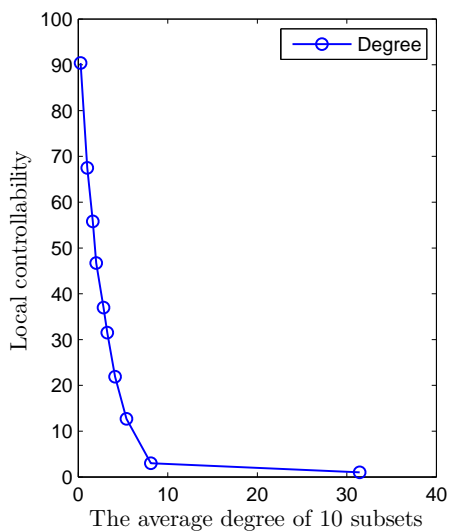
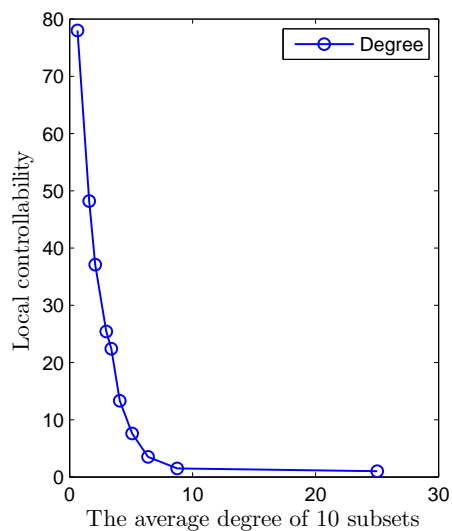
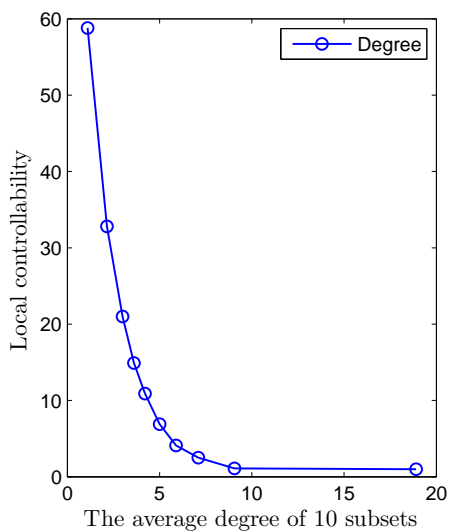
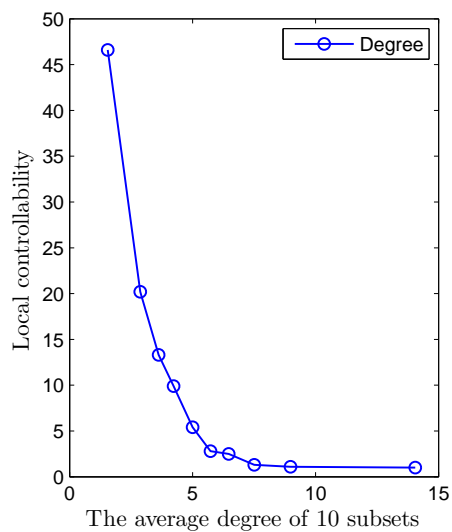
(a) $\gamma = 2.2$ (b) $\gamma = 2.5$ (c) $\gamma = 3$ (d) $\gamma = 4$

Figure 6.11: The plot of local controllability versus average degree of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 6$ and $\gamma = 2.2, 2.5, 3$ and 4 .

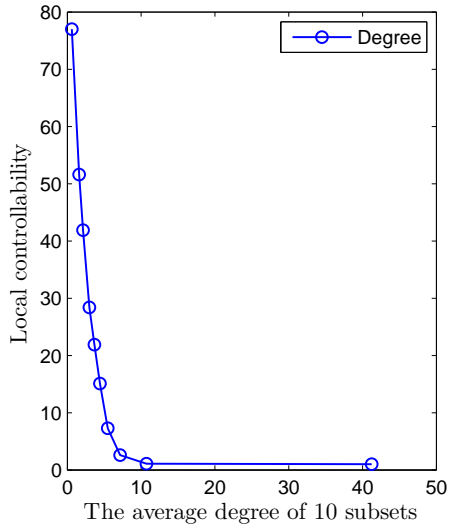
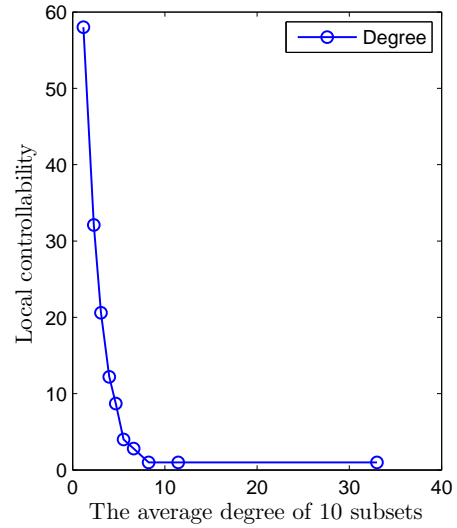
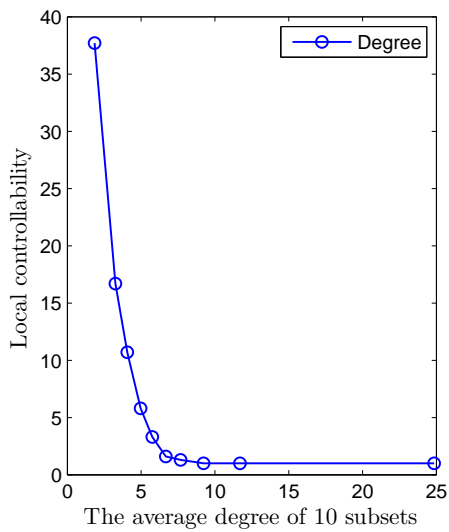
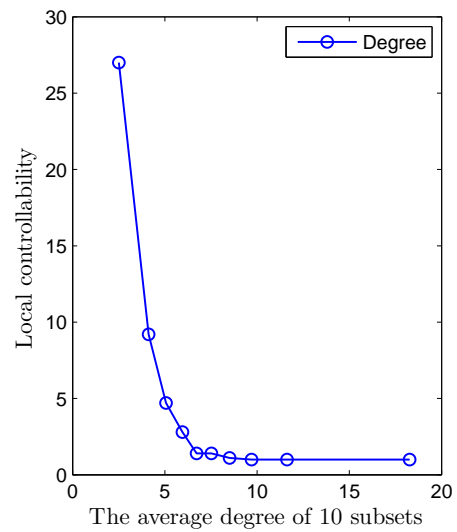
(a) $\gamma = 2.2$ (b) $\gamma = 2.5$ (c) $\gamma = 3$ (d) $\gamma = 4$

Figure 6.12: The plot of local controllability versus average degree of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 8$ and $\gamma = 2.2, 2.5, 3$ and 4 .

6.3 Local controllability v.s. betweenness

In this section, we study the relationship between local controllability and the betweenness (centrality measure) in both ER random networks and scale-free networks.

For ER random networks, the (overall) plot of local controllability versus average betweenness of 10 subsets is shown in Figure 6.13. In this plot, the relationship between local controllability $lc(G, S)$ and average betweenness of S for 10 subsets of nodes were averaged over 10 realizations of ER random networks with parameters $n = 1000$ and $\langle k \rangle = 3, 4, \dots, 8$. In each network realization, the 10 subsets are equal-sized: all the nodes were divided into 10 subsets of equal size based on the order of the betweenness values of all the nodes.

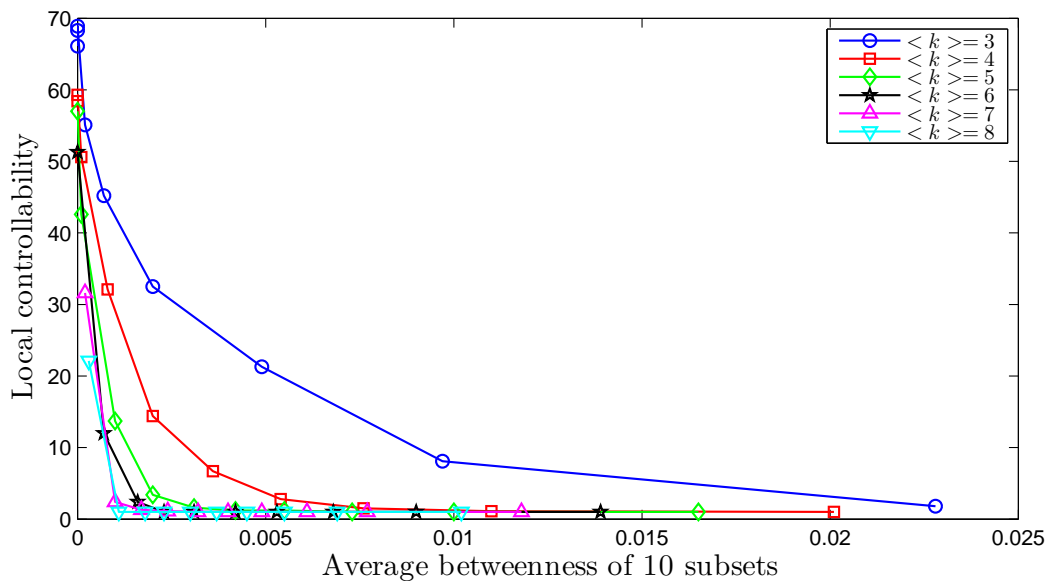


Figure 6.13: The plot of local controllability versus average betweenness of 10 subsets in ER random networks with varying parameters $\langle k \rangle = 3, \dots, 8$. For each parameter $\langle k \rangle$, 10 ER random networks with $n = 1000$ nodes were generated.

As can be seen from Figure 6.13, a subset of nodes with higher betweenness needs less inputs to control (or has lower local controllability), indicating that nodes with higher betweenness are easier to control in ER random networks.

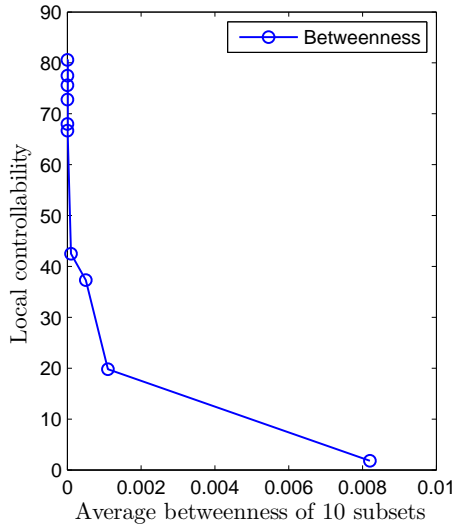
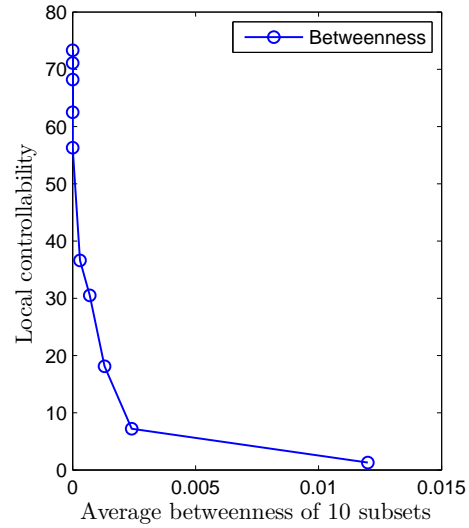
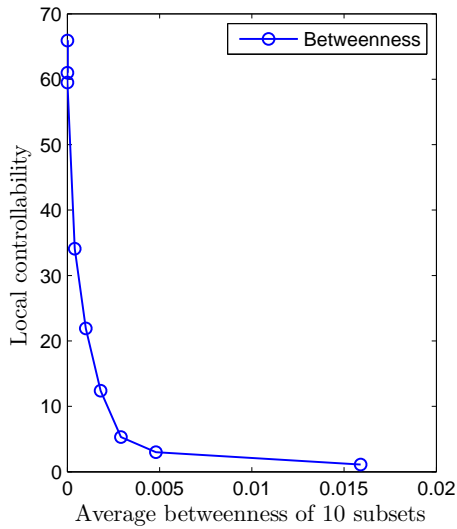
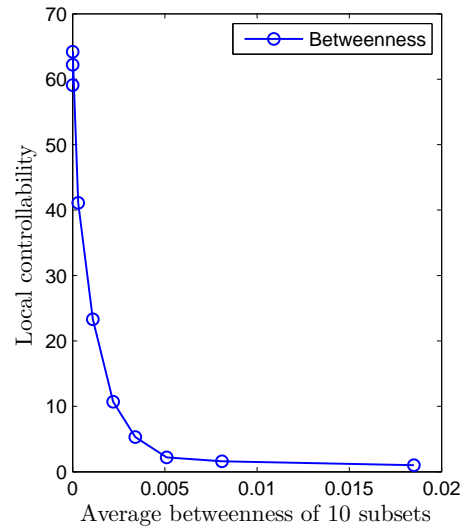
(a) $\gamma = 2.2$ (b) $\gamma = 2.5$ (c) $\gamma = 3$ (d) $\gamma = 4$

Figure 6.14: The plot of local controllability versus average betweenness of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 2.2, 2.5, 3$ and 4 .

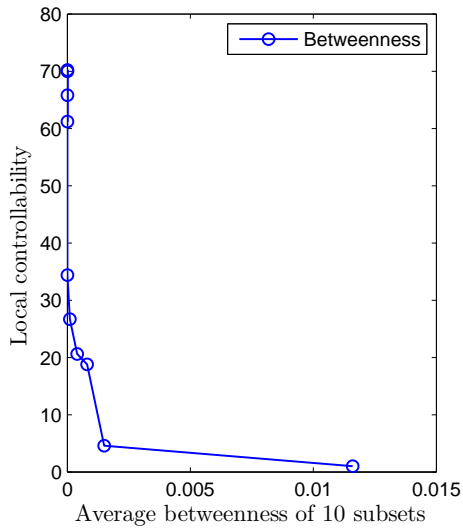
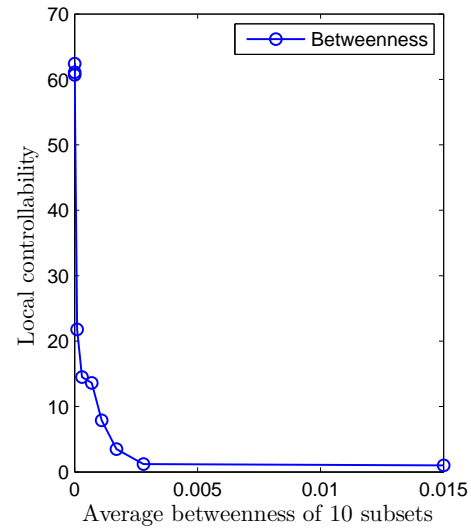
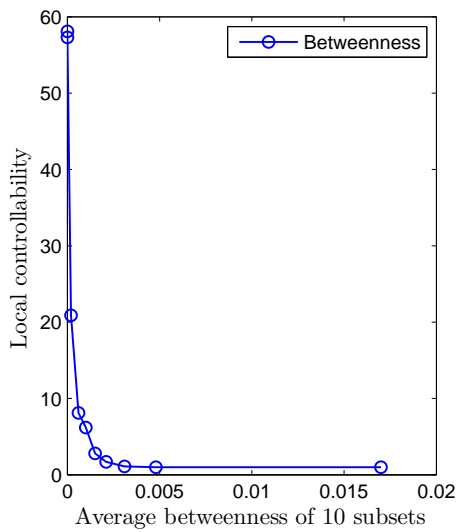
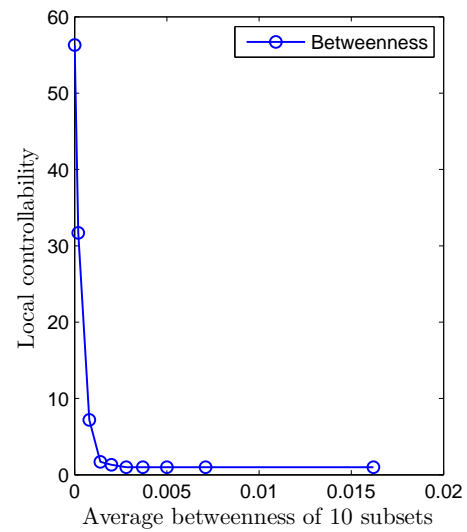
(a) $\gamma = 2.2$ (b) $\gamma = 2.5$ (c) $\gamma = 3$ (d) $\gamma = 4$

Figure 6.15: The plot of local controllability versus average betweenness of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 6$ and $\gamma = 2.2, 2.5, 3$ and 4 .

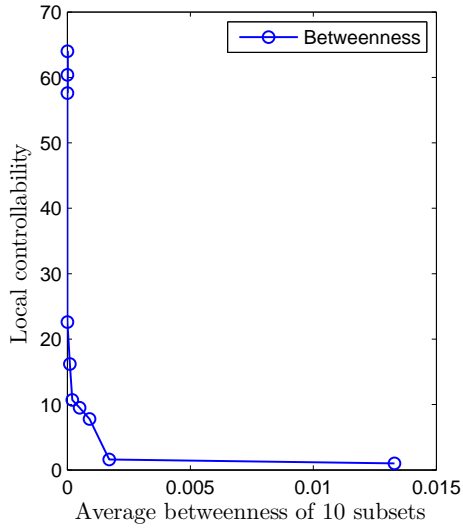
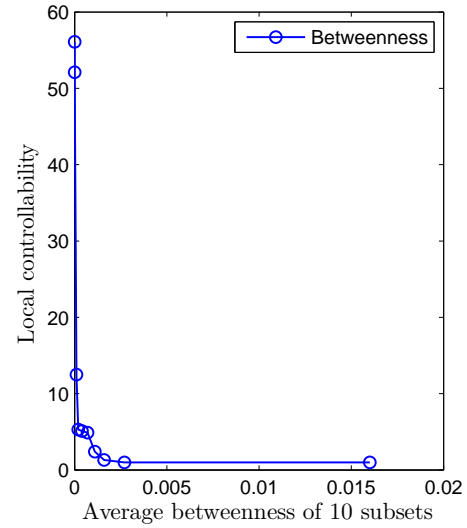
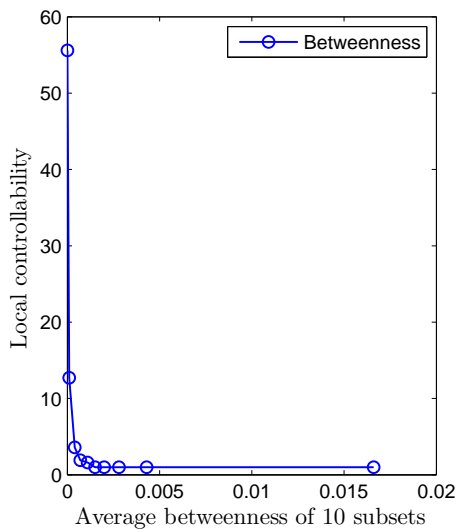
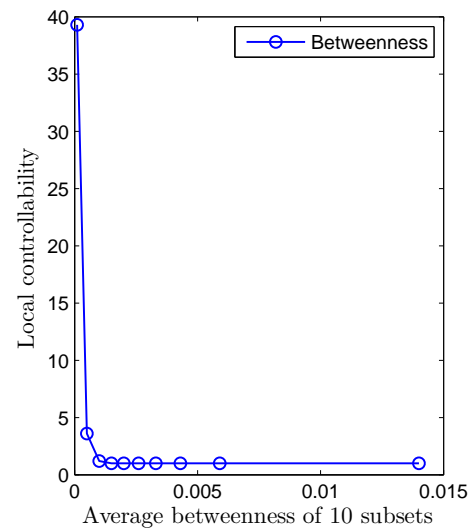
(a) $\gamma = 2.2$ (b) $\gamma = 2.5$ (c) $\gamma = 3$ (d) $\gamma = 4$

Figure 6.16: The plot of local controllability versus average betweenness of 10 subsets in scale-free networks with $n = 1000$, $\langle k \rangle = 8$ and $\gamma = 2.2, 2.5, 3$ and 4 .

We also examined the relationship between local controllability and the betweenness measure in directed scale-free networks. We generated scale-free networks with $n = 1000$ nodes and varying parameters for average degree $\langle k \rangle$ and degree exponent γ . For each combination of parameters $\langle k \rangle$ and $\gamma (= \gamma_{in} = \gamma_{out})$, we generated 10 scale-free networks, based on the directed static model.

For each scale-free network G generated, all the nodes were divided into 10 equal-sized subsets based on the order of the betweenness values of all the nodes. For each of these 10 subsets, S , we calculated $lc(G, S)$ and the average betweenness of S , which were then averaged over all 10 network realizations. The plots for the relationships between $lc(G, S)$ and average betweenness of S are given in Figure 6.14-6.16.

All these plots reveal a decreasing trend, suggesting that nodes with higher betweenness are easier to control in scale-free networks. Moreover, when the degree exponent γ gets larger, the scale-free network becomes less heterogeneous, and the plot for scale-free networks with given $\langle k \rangle$ and γ becomes more similar to the plot for ER random networks with the same $\langle k \rangle$.

In summary, our numerical experiments show that nodes with higher betweenness tend to be easier to control in both ER random networks and scale-free networks.

We have performed similar numerical experiments to examine the relationships between local controllability and other centrality measures, including clustering coefficient, closeness, eigenvector centrality, Katz centrality and PageRank centrality (see Section 2.3 in Chapter 2). Similar plots of local controllability versus average value of these centrality measures of 10 subsets have been obtained.

In all these plots (not shown here), we didn't observe a clear monotone trend, therefore, we can not conclude that local controllability is related to these centrality measures in both ER random networks and scale-free networks.

6.4 Local controllability v.s. link/node removal

It is expected that $lc(G_1, S)$ will be different from $lc(G_2, S)$, when S is a subset of nodes in both directed networks G_1 and G_2 that have different network structures. Thus, next we are interested in studying how the local controllability $lc(G, S)$ depends on the network topology G when S is fixed.

For such purposes, we did numerical experiments to investigate the robustness of local controllability against link (edge) and node removal in model networks: the effect of deleting links or nodes in a directed network G on the local controllability $lc(G, S)$ of a subset S .

First, we studied the robustness of local controllability against link removal. Specifically, we generated a directed ER random network with $n = 1000$ nodes and $\langle k \rangle = 4$, and denote the network by $G_0 = (V(G_0), A(G_0))$. Then, we identified the top 100 hubs in G_0 , the subset of which is denoted by S . The *out-neighborhood* of S in G_0 is defined to be

$$N_{G_0}^+(S) = \{v \in V(G_0) - S : (u, v) \in A(G_0) \text{ for some node } u \in S\}.$$

The *in-neighborhood* of S in G_0 is defined to be

$$N_{G_0}^-(S) = \{v \in V(G_0) - S : (v, u) \in A(G_0) \text{ for some node } u \in S\}.$$

The set of *outgoing links* of S in G_0 is thus the set of arcs $(u, v) \in A(G_0)$ such that $u \in S$ and $v \in N_{G_0}^+(S)$. Similarly, the set of *incoming links* of S in G_0 is the set of arcs $(u, v) \in A(G_0)$ such that $u \in N_{G_0}^-(S)$ and $v \in S$.

After we identified S , we deleted the set of outgoing links of S in G_0 step by step: starting with the network G_0 , we calculated $lc(G_0, S)$; in each step, we randomly deleted an outgoing link of S in the current network, and recalculated the local controllability of S in the updated network (after the link removal in this step), until all outgoing links of S in G_0 have been deleted.

The same procedure was also performed to delete the set of incoming links of S in G_0 , starting again with the same initial network G_0 . The local controllability

was then plotted against the time steps of the procedure of outgoing/incoming link deletion (or the number of outgoing/incoming links deleted), and the plot is shown in Figure 6.17.

Similar plots have been obtained for scale-free networks with $n = 1000$ nodes, $\langle k \rangle = 4$ and $\gamma (= \gamma_{in} = \gamma_{out}) = 2.5, 4$. These two plots are given in Figure 6.18-6.19.

In all these plots, we observed that the removal of outgoing links and the removal of incoming links both make the local controllability of S converge to its upper bound $N_D(G_0[S])$, which is the network controllability of the subnetwork $G_0[S]$ of G_0 induced by S .

For the scale-free network with small degree exponent $\gamma = 2.5$ (see Figure 6.18), the hubs are very densely interconnected, and we found that the local controllability of S (the subset of top 100 hubs) is robust to the removal of both outgoing links and incoming links: only after a large number of outgoing/incoming links have been deleted can the local controllability of S start to increase.

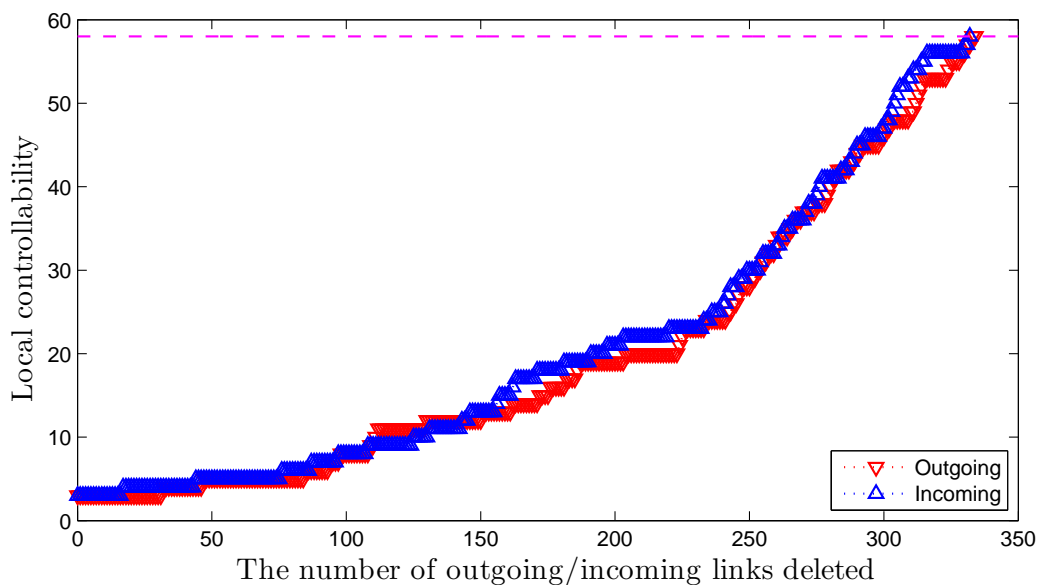


Figure 6.17: The plot of local controllability versus the number of outgoing/incoming links deleted for an ER random network G_0 with $n = 1000$ and $\langle k \rangle = 4$. The dashed line $N_D = 58$ indicates the network controllability of $G_0[S]$.

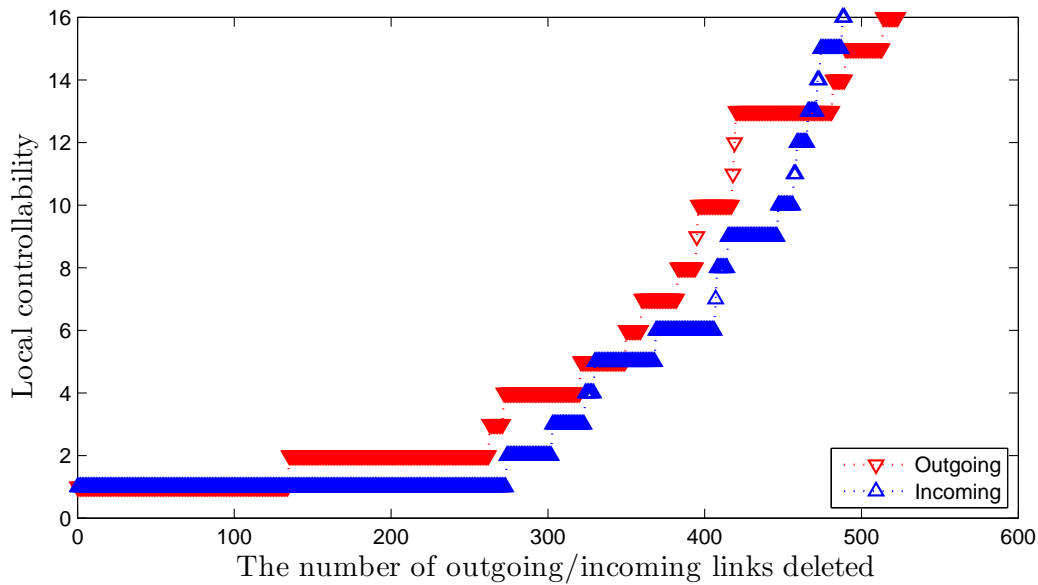


Figure 6.18: The plot of local controllability versus the number of outgoing/incoming links deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 2.5$. $N_D = 16$ indicates the network controllability of $G_0[S]$.

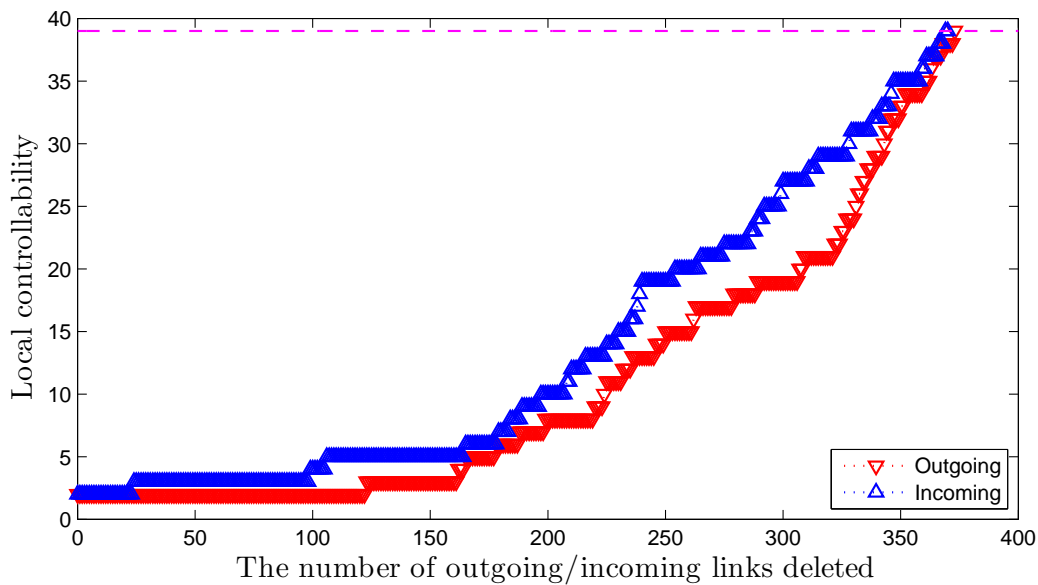


Figure 6.19: The plot of local controllability versus the number of outgoing/incoming links deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 4$ and $\gamma = 4$. The dashed line $N_D = 39$ indicates the network controllability of $G_0[S]$.

For the ER random network and the scale-free network with large degree exponent $\gamma = 4$ (see Figure 6.17 and 6.19), the local controllability $lc(G_0, S)$ varies greatly with the network controllability $N_D(G_0[S])$, and the removal of outgoing links and incoming links can increase the local controllability of S steadily, especially in the later time steps. Moreover, in these networks, we can see that local controllability is less robust to incoming link removal: the removal of a few incoming links (early time steps in the plots) seems to have a greater impact on the local controllability of S than the removal of the same number of outgoing links.

Next, we also studied the robustness of local controllability against node removal. We generated a scale-free network with $n = 1000$ nodes, $\langle k \rangle = 3$ and $\gamma (= \gamma_{in} = \gamma_{out}) = 2.5$, and denoted the network by G_0 . The subset of the top 100 hubs was identified, and is denoted by S . We considered the removal of both out-neighbors and in-neighbors of S , each by two means: random removal and hub attack. Starting with the initial network G_0 , we define four node removal strategies:

Out-neighbor (in-neighbor) random removal: in each step, randomly delete an out-neighbor (in-neighbor) of S , until all out-neighbors (in-neighbors) of S in G_0 have been deleted.

Out-neighbor (in-neighbor) hub attack: in each step, delete the out-neighbor (in-neighbor) of S that currently has the largest degree, until all out-neighbors (in-neighbors) of S in G_0 have been deleted.

The plot of local controllability versus the number of nodes deleted based on these four strategies is then given in Figure 6.20.

A similar plot for a scale-free network with $n = 1000$ nodes, $\langle k \rangle = 3$ and $\gamma = 4$ is given in Figure 6.21. A plot for an ER random network with $n = 1000$ nodes and $\langle k \rangle = 3$ is also given, as in Figure 6.22.

As can be seen from Figure 6.20-6.22, hub attack can increase (damage) the local controllability more effectively than random removal, for both out-neighbors and in-neighbors. The in-neighbor hub attack seems to be more effective than the out-neighbor hub attack, but the difference is not that obvious.

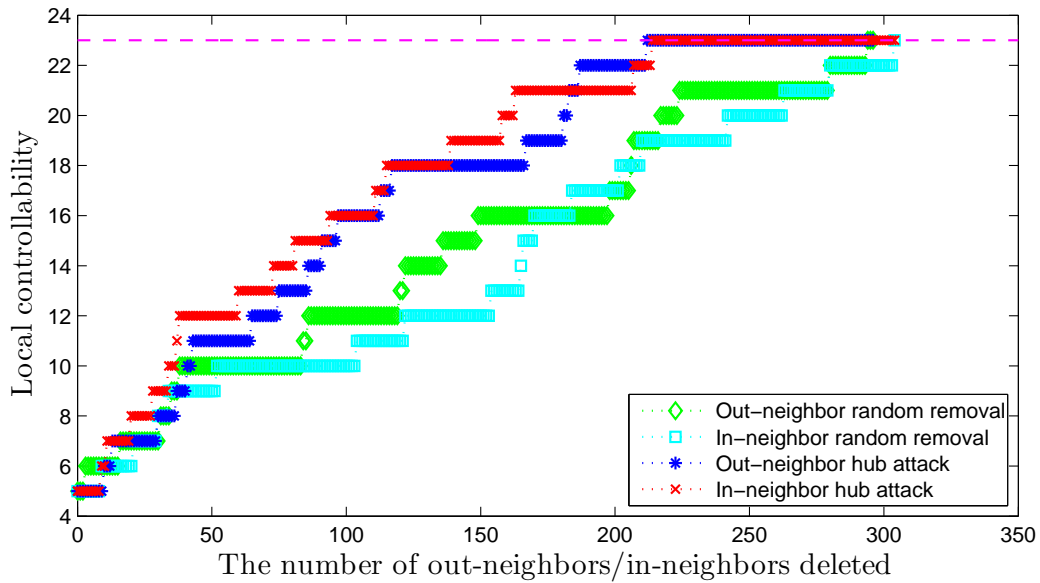


Figure 6.20: The plot of local controllability versus the number of out-neighbors/in-neighbors deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 3$ and $\gamma = 2.5$. The dashed line $N_D = 23$ indicates the network controllability of $G_0[S]$.

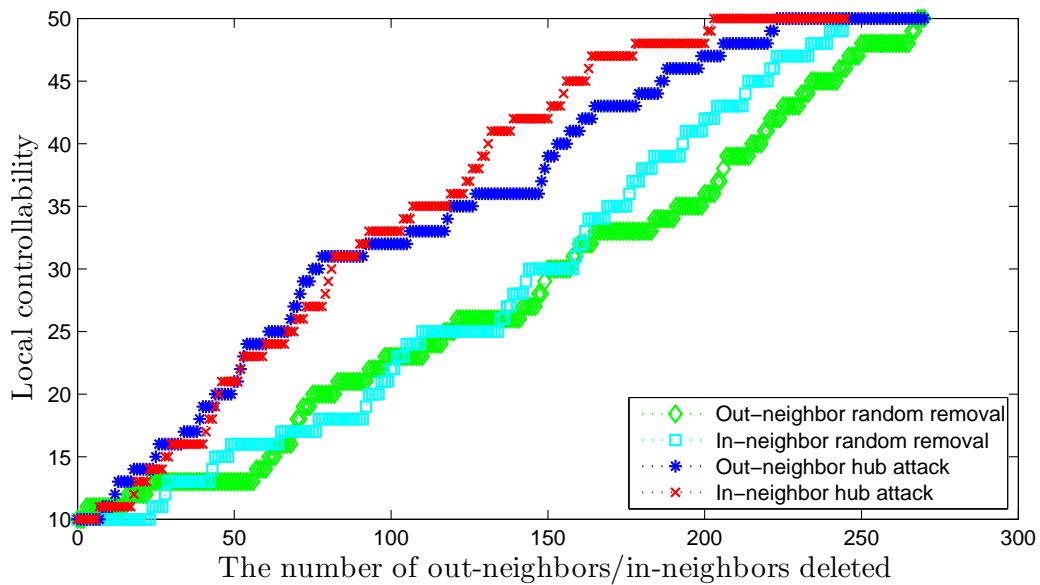


Figure 6.21: The plot of local controllability versus the number of out-neighbors/in-neighbors deleted for a scale-free network G_0 with $n = 1000$, $\langle k \rangle = 3$ and $\gamma = 4$. $N_D = 50$ indicates the network controllability of $G_0[S]$.

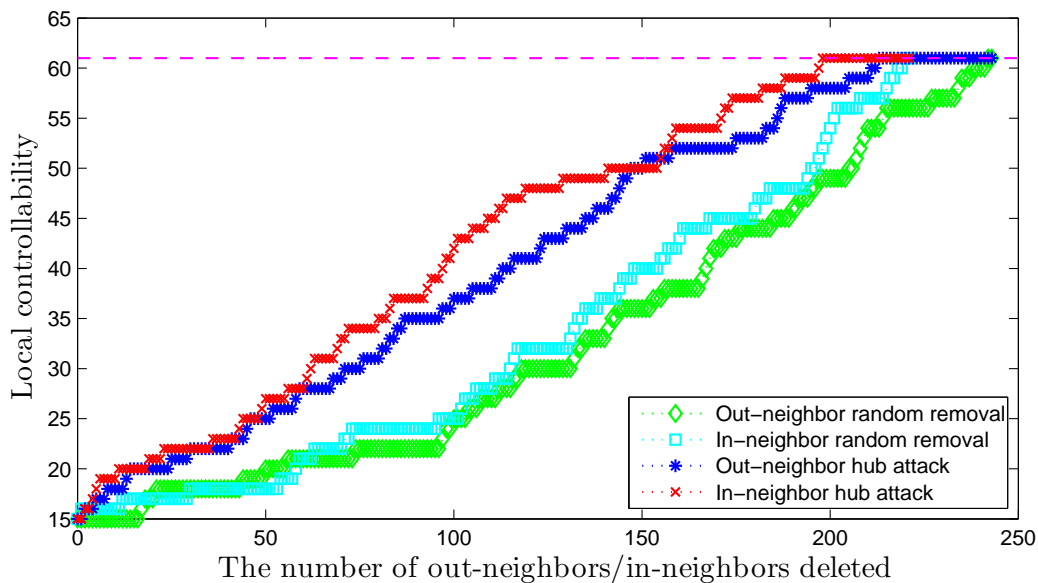


Figure 6.22: The plot of local controllability versus the number of out-neighbors/in-neighbors deleted for an ER network G_0 with $n = 1000$ and $\langle k \rangle = 3$. The dashed line $N_D = 61$ indicates the network controllability of $G_0[S]$.

In summary, local controllability is affected by both link removal and node removal in a directed network. The removal of incoming links of S and the removal of in-neighbors (out-neighbors) of S with the highest degree, seem to be effective at increasing the local controllability of S in a directed network.

6.5 Remarks

The two types of model networks, ER random networks and scale-free networks show some similar behaviors in terms of local controllability. First, we found that in both ER random networks and scale-free networks, nodes with higher degree are easier to control than nodes with lower degree, and nodes with higher betweenness are easier to control than nodes with lower betweenness. Second, for a subset S , the removal of in-neighbors (out-neighbors) of S as well as the removal of outgoing (incoming) links of S in both ER random networks and scale-free networks will increase the local

controllability of S towards the network controllability of the subnetwork induced by S . Third, in both ER random networks and scale-free networks, the removal of in-neighbors (out-neighbors) of S with the highest degree (hub attack) causes more damage to the local controllability of S than the random removal of in-neighbors (out-neighbors) of S .

ER random networks and scale-free networks also have some different behaviors. First, it is easier to control nodes with the highest degrees in scale-free networks than in ER random networks, while it is harder to control nodes with the lowest degrees in scale-free networks than in ER random networks. Second, the local controllability of a subset S of hubs is more robust to the removal of outgoing and incoming links of S in scale-free networks than in ER random networks. These differences could be explained by the different structures of ER random networks and scale-free networks: ER random networks are homogeneous, while scale-free networks are heterogeneous. As the degree exponent γ becomes larger, scale-free networks become less heterogeneous, and their behaviors become more similar to those of ER random networks.

Future work

In this thesis, we have introduced a new network control framework, called ‘local controllability’. We have applied local controllability to biological networks, especially the brain networks and the transcriptional regulatory networks. The results suggest that in the mouse inter-region brain networks, there are some subdivisions that are easy to control and also some subdivisions that are difficult to control, which might be related to the structural basis of these subdivisions. The results also suggest that in the human TF regulatory networks, protein complexes tend not to be easy to control.

Local controllability is defined for any directed network, and it concerns about controlling a subset of nodes in a directed network with the minimum number of inputs. Like the framework of network controllability [56], our framework of local controllability can only be applied to directed networks.

Our research focus is to develop new tools as well as incorporating existing methods, to study the control of complex networks. In the following, we will list several possible problems for future study.

Firstly, we will seek to generalize the framework of local controllability so that it can also be defined for undirected networks (and possibly weighted networks). One limitation of local controllability is that it is only applicable to directed networks.

Thus, a future avenue of research is to generalize local controllability to undirected networks. Since the existing framework of exact controllability [105] is defined by using a set of algebraic formulations (for example, the Popov-Belevich-Hautus rank condition), local controllability might be able to be defined in undirected networks by incorporating algebraic conditions.

Secondly, with the general framework of exact controllability [105], we will study the exact controllability of undirected networks, including some important types of biological networks. For example, the human (structural) brain networks are mostly undirected, as directionality in the anatomical connections between brain regions is difficult to be detected, by using the common mapping techniques such as diffusion imaging and tractography [13, 35, 95]. In the future, we might investigate how the exact controllability of the human brain networks is related to their structures, such as the small-world property [37], and the rich-club organization [95]. We can also compare the exact controllability of healthy and diseased brain networks, and see how the abnormal brain network structure can affect the controllability of human brain networks. It is also possible to use exact controllability to study the relationship between the brain networks of different species.

Lastly, with our framework of local controllability, we can also study the local controllability of other types of directed biological networks, such as metabolic networks and cell signaling networks. For example, we can use local controllability to study how to control a biomarker set for a cancer disease, which might be important in finding a possible cure.

Bibliography

- [1] Q. A. Acton. *Advances in Corpus Striatum Research and Application: 2012 Edition*, Scholarly Editions, Atlanta, USA, 2012.
- [2] N. T. T. Ahn. *Spanning Cacti for Structurally Controllable Networks*, master thesis, National University of Singapore, 2012.
- [3] R. Albert, H. Jeong, and A. L. Barabási. Error and attack tolerance of complex networks, *Nature*, 406 (2000), 378-382.
- [4] U. Alon. Network motifs: theory and experimental approaches, *Nature*, 8 (2007), 450-461.
- [5] S. Assou, T. L. Carrou, et al. A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas, *Stem Cells*, 25 (2007), 961-973.
- [6] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks, *Current Opinion in Structural Biology*, 14 (2004), 283-291.

-
- [7] S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of Yeast, *J. Mol. Biol.*, 360 (2006), 213-227.
- [8] A. L. Barabási and R. Albert. Emergence of scaling in random networks, *Science*, 286 (1999), 509-512.
- [9] A. L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization, *Nature Reviews Genetics*, 5 (2004), 101-113.
- [10] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: structure and dynamics, *Physics Reports*, 424 (2006), 175-308.
- [11] B. Bollobás. *Random Graphs*, Cambridge University Press, Cambridge, UK, 2001.
- [12] P. Bonacich. Power and centrality: a family of measures, *American Journal of Sociology*, 92 (1987), 1170-1182.
- [13] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems, *Nature Reviews Neuroscience*, 10 (2009), 186-198.
- [14] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: percolation on random graphs, *Phys. Rev. Lett.*, 85 (2000), 5468-5471.
- [15] Z. J. Chen, Y. He, P. Rosa-Neto, J. Germann, and A. C. Evans. Revealing modular architecture of human brain structural networks by using cortical thickness from MRI, *Cerebral Cortex*, 18 (2008), 2374-2381.
- [16] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks, *Physical Review E*, 70, 066111 (2004).

-
- [17] E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, USA, 1955.
- [18] R. Cohen and S. Havlin. Scale-free networks are ultra-small, *Phys. Rev. Lett.*, 90, 058701 (2003).
- [19] P. Csermely, T. Korcsmaros, H. J. M. Kiss, G. London, and R. Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery, a comprehensive review, *Pharmacology and Therapeutics*, 138 (2013), 333-408.
- [20] E. H. Davidson. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*, Academic Press, San Diego, USA, 2010.
- [21] J. M. Dion, C. Commault, and J. van der Woude. Generic properties and control of linear structured systems: a survey, *Automatica*, 39 (2003), 1125-1144.
- [22] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking, *Phys. Rev. Lett.*, 85 (2000), 4633-4636.
- [23] J. K. Elmquist and C. F. Elias. From lesions to leptin: hypothalamic control of food intake and body weight, *Neuron*, 22 (1999), 221-232.
- [24] P. Erdős and A. Rényi. On the evolution of random graphs, *Publications of the Mathematical Institute of the Hungarian Academy of Science*, 5 (1960), 17-61.
- [25] G. Fagiolo. Clustering in complex directed networks, *Physical Review E*, 76, 026107 (2007).
- [26] I. Ferezou, F. Haiss, et al. Spatiotemporal dynamics of cortical sensorimotor integration in behaving mice, *Neuron*, 56 (2007), 907-923.
- [27] S. Fortunato. Community detection in graphs, *Physics Reports*, 486 (2010), 75-174.
- [28] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski. Edge direction and the structure of networks, *Proc. Natl. Acad. Sci.*, 107 (2010), 10815-10820.

-
- [29] L. C. Freeman. Centrality in social networks conceptual clarification, *Social Networks*, 1 (1979), 215-239.
- [30] M. B. Gerstein, A. Kundaje, et al. Architecture of the human regulatory network derived from ENCODE data, *Nature*, 489 (2012), 91-100.
- [31] K. Glover and L. M. Silverman. Characterization of structural controllability, *IEEE Trans. Automat. Contr.*, 21 (1976), 534-537.
- [32] K. I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks, *Phys. Rev. Lett.*, 87, 278701 (2001).
- [33] G. L. Gong, Y. He, L. Concha, C. Lebel, D. W. Gross, A. C. Evans, and C. Beaulieu. Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography, *Cerebral Cortex*, 19 (2009), 524-536.
- [34] N. Guelzim, S. Bottani, P. Bourguin, and F. Képès. Topological and causal structure of the yeast transcriptional regulatory network, *Nature Genetics*, 31 (2002), 60-63.
- [35] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex, *PLoS Biol*, 6 (2008), 1479-1493.
- [36] P. Hagmann, M. Kurant, X. Gigandet, P. Thiran, V. J. Wedeen, R. Meuli, and J. P. Thiran. Mapping human whole-brain structural networks with diffusion MRI, *PLoS One*, 2, e597 (2007).
- [37] Y. He, Z. J. Chen, and A. C. Evans. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI, *Cerebral Cortex*, 17 (2007), 2407-2419.

-
- [38] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural connectivity, *Proc. Natl. Acad. Sci.*, 106 (2009), 2035-2040.
- [39] J. E. Hopcroft and R. M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs, *SIAM Journal on Computing*, 2 (1973), 225-231.
- [40] M. D. Humphries and K. Gurney. Network 'small-worldness': a quantitative method for determining canonical network equivalence, *PLoS One*, 3, e0002051 (2008).
- [41] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks, *Nature*, 411 (2001), 41-42.
- [42] S. Jinno and T. Kosaka. Cellular architecture of the mouse hippocampus: a quantitative aspect of chemically defined GABAergic neurons with stereology, *Neurosci Res*, 56 (2006), 229-245.
- [43] R. Jothi, S. Balaji, et al. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture, *Molecular Systems Biology*, 5, 294 (2009).
- [44] R. E. Kalman. Mathematical description of linear dynamical systems, *J. Soc. Indus. Appl. Math. Ser. A*, 1 (1963), 152-192.
- [45] L. Katz. A new status index derived from sociometric analysis, *Psychometrika*, 18 (1953), 39-43.
- [46] M. H. Kaufman and J. B. L. Bard. *The Anatomical Basis of Mouse Development*, Academic Press, San Diego, USA, 1999.
- [47] J. P. Keener. The Perron-Frobenius theorem and the ranking of football teams, *SIAM Review*, 35 (1993), 80-93.

-
- [48] M. T. K. Kirkcaldie. Neocortex, In *The Mouse Nervous System*, pp. 52-111, Elsevier Inc., Netherlands, 2012.
- [49] K. Klemm and V. M. Eguíluz. Growing scale-free networks with small-world behavior, *Physical Review E*, 65, 057102 (2002).
- [50] D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks, *Gene Regulation and System Biology*, 2 (2008), 193-201.
- [51] D. Koschützki, H. Schwöbbermeyer, and F. Schreiber. Ranking of network elements based on functional substructures, *Journal of Theoretical Biology*, 248 (2007), 471-479.
- [52] H. W. Kuhn. The Hungarian method for the assignment problem, *Naval Research Logistics Quarterly*, 2 (1955), 83-97.
- [53] L. Y. Lu, D. B. Chen, and T. Zhou. The small world yields the most effective information spreading, *New Journal of Physics*, 13, 123005 (2011).
- [54] E. A. Leicht and M. E. J. Newman. Community structure in directed networks, *Phys. Rev. Lett.*, 100, 118703 (2008).
- [55] C. T. Lin. Structural controllability, *IEEE Trans. Automat. Contr.*, 19 (1974), 201-208.
- [56] Y. Y. Liu, J. J. Slotine, and A. L. Barabási. Controllability of complex networks, *Nature*, 473 (2011), 167-173.
- [57] Y. Y. Liu, J. J. Slotine, and A. L. Barabási. Few inputs can reprogram biological networks, *Nature*, 478 (2011), E4-E5.
- [58] D. G. Luenberger. *Introduction to Dynamical Systems: Theory, Models and Applications*, Wiley, USA, 1979.

-
- [59] J. Lygeros and F. Ramponi. *Lecture Notes on Linear System Theory*, ETH, Zurich, Switzerland, 2010.
- [60] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif, *Proc. Natl. Acad. Sci.*, 100 (2003), 11980-11985.
- [61] K. J. Mastro, R. S. Bouchard, H. A. K. Holt, and A. H. Gittis. Transgenic mouse lines subdivide external segment of the globus pallidus (GPe) neurons and reveal distinct GPe output pathways, *The Journal of Neuroscience*, 34 (2014), 2087-2099.
- [62] H. Mayeda. On structural controllability theorem, *IEEE Trans. Automat. Contr.*, 26 (1981), 795-798.
- [63] S. Milgram. The small world problem, *Psychology Today*, 1 (1967), 61-67.
- [64] R. Milo, S. Itzkovitz, et al. Superfamilies of evolved and designed networks, *Science*, 303 (2003), 1538-1542.
- [65] R. Milo, S. Shen-Orr, et al. Network motifs: simple building blocks of complex networks, *Science*, 298 (2002), 824-827.
- [66] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks, *Cell*, 150 (2012), 1274-1286.
- [67] T. Nepusz and T. Vicsek. Controlling edge dynamics in complex networks, *Nature Physics*, 8 (2012), 568-573.
- [68] M. E. J. Newman. Assortative mixing in networks, *Phys. Rev. Lett.*, 89, 208701 (2002).
- [69] M. E. J. Newman. The structure and function of complex networks, *SIAM Review*, 45 (2003), 167-256.

-
- [70] M. E. J. Newman. Fast algorithm for detecting community structure in networks, *Physical Review E*, 69, 066133 (2004).
- [71] M. E. J. Newman. Modularity and community structure in networks, *Proc. Natl. Acad. Sci.*, 103 (2006), 8577-8582.
- [72] M. E. J. Newman. *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.
- [73] M. E. J. Newman. Communities, modules and large-scale structure in networks, *Nature Physics*, 8 (2012), 25-31.
- [74] S. Oh, J. A. Harris, et al. A mesoscale connectome of the mouse brain, *Nature*, 508 (2014), 207-214.
- [75] P. Oikonomou and P. Cluzel. Effects of topology on network evolution, *Nature Physics*, 2 (2006), 532-536.
- [76] M. Palkovits and M. J. Brownstein. *Maps and Guide to Microdissection of the Rat Brain*, Elsevier, Netherlands, 1988.
- [77] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks, *Phys. Rev. Lett.*, 86 (2001), 3200-3203.
- [78] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet, *Phys. Rev. Lett.*, 87, 258701 (2001).
- [79] S. C. Ponten, A. Daffertshofer, A. Hillebrand, and C. J. Stam. The relationship between structural and functional connectivity: graph theoretical analysis of an EEG neural mass model, *NeuroImage*, 52 (2010), 985-994.
- [80] M. Pósfai, Y. Y. Liu, J. J. Slotine, and A. L. Barabási. Effect of correlations on network controllability, *Science Reports*, 3, 1067 (2013).

-
- [81] L. Puelles, M. M. de-la-Torre, S. Bardet, and J. L. R. Rubenstein. Hypothalamus, In *The Mouse Nervous System*, pp. 221-312, Elsevier Inc., Netherlands, 2012.
- [82] L. Puelles, M. M. de-la-Torre, J. L. Ferran, and C. Watson. Diencephalon, In *The Mouse Nervous System*, pp. 313-336, Elsevier Inc., Netherlands, 2012.
- [83] R. W. Rieck, M. S. Ansari, et al. Distribution of dopamine D2-like receptors in the human thalamus: Autoradiographic and PET studies, *Neuropsychopharmacology*, 29 (2004), 362-372.
- [84] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations, *NeuroImage*, 52 (2010), 1059-1069.
- [85] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli, *Nature Genetics*, 31 (2002), 64-68.
- [86] R. W. Shields and J. B. Pearson. Structural controllability of multi-input linear systems, *IEEE Trans. Automat. Contr.*, 21 (1976), 203-212.
- [87] J. J. Slotine and W. Li. *Applied Nonlinear Control*, Prentice Hall, New Jersey, USA, 1991.
- [88] E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems, Second Edition*, Springer, New York, USA, 1998.
- [89] F. Sorrentino, M. di Bernardo, F. Garofalo, and G. R. Chen. Controllability of complex networks via pinning, *Physical Review E*, 75, 046103 (2007).
- [90] O. Sporns. The human connectome: a complex network, *Ann. N.Y. Acad. Sci.*, 1224 (2011), 109-125.
- [91] O. Sporns. Network attributes for segregation and integration in the human brain, *Current Opinion in Neurobiology*, 23 (2013), 162-171.

- [92] O. Sporns and J. D. Zwi. The small world of the cerebral cortex, *Neuroinformatics*, 2 (2004), 145-162.
- [93] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors, *Cell*, 126 (2006), 663-676.
- [94] L. Tomás-Roca, R. Corral-San-Miguel, P. Aroca, L. Puelles, and F. Marín. Crypto-rhombomeres of the mouse medulla oblongata, defined by molecular and morphological features, *Brain Structure and Function*, doi: 10.1007/s00429-014-0938-y (2014), 1-24.
- [95] M. P. van den Heuvel and O. Sporns. Rich-club organization of the human connectome, *Journal of Neuroscience*, 31 (2011), 15775-15786.
- [96] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural properties of the caenorhabditis elegans neuronal network, *PLoS Computational Biology*, 7, e1001066 (2011).
- [97] A. Vinayagam, Y. H. Hu, M. Kulkarni, C. Roesel, R. Sopko, S. E. Mohr, and N. Perrimon. Protein complex-based analysis framework for high-throughput data sets, *Science Signaling*, 6, rs5 (2013).
- [98] W. Z. Wang, F. M. Oeschger, et al. Comparative aspects of subplate zone studied with gene expression in sauropsids and mammals, *Cerebral Cortex*, 21 (2011), 2187-2203.
- [99] W. Z. Wang and A. H. Suabedissen. Subplate in the developing cortex of mouse and human, *Journal of Anatomy*, 217 (2010), 368-380.
- [100] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks, *Nature*, 393 (1998), 440-442.
- [101] D. R. White and S. P. Borgatti. Betweenness centrality measures for directed graphs, *Social Networks*, 16 (1994), 335-346.

-
- [102] T. Will and V. Helms. Identifying transcription factor complexes and their roles, *Bioinformatics*, 30 (2014), i415-i421.
- [103] M. Witter. Hippocampus, In *The Mouse Nervous System*, pp. 112-139, Elsevier Inc., Netherlands, 2012.
- [104] H. Y. Yu and M. Gerstein. Genomic analysis of the hierarchical structure of regulatory networks, *Proc. Natl. Acad. Sci.*, 103 (2006), 14724-14731.
- [105] Z. Z. Yuan, C. Zhao, Z. R. Di, W. X. Wang, and Y. C. Lai. Exact controllability of complex networks, *Nature Communications*, 4, 2447 (2013).
- [106] S. H. Zhang, D. C. Tian, N. H. Tran, K. P. Choi, and L. X. Zhang. Profiling the transcription factor regulatory networks of human cell types, *Nucleic Acid Research*, 42 (2014), 12380-12387.
- [107] X. W. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks, *Genes Dev.*, 21 (2007), 1010-1024.

Local controllability of protein complexes

In this appendix, we give the results for the local controllability of 712 protein complexes (with size from 5 to 19) in the human transcription factor regulatory networks of 8 cell types: **hESC**, **Amniotic Epi.**, **Hemat. Stem Cell**, **Adult Dermal Blood**, **Neuroblastoma**, **Skeletal Myoblast**, **Fetal Brain**, and **Skin Fib.** Each of these 8 cell types is a representative from its corresponding class, see Table 2.1.

These results are presented in a long table, Table A.1, which is shown in the subsequent pages. We give a numbered index (No.) and the complex identifier (ID) for each protein complex, each entry in the column of a cell type indicates the local controllability of the corresponding protein complex in the human TF regulatory network of that cell type.

In Table A.1, we use acronyms for the names of 7 cell types in following way: AE stands for **Amniotic Epi.**, HSC for **Hemat. Stem Cell**, ADB for **Adult Dermal Blood**, Neu for **Neuroblastoma**, SM for **Skeletal Myoblast**, FB for **Fetal Brain** and SF for **Skin Fib.**

Note that, the transcription factor ‘ZNF354C’ is not in the node set of the TF regulatory network of hESC, thus, the local controllability of a protein complex that contains ‘ZNF354C’ in the TF regulatory network of hESC was calculated by

regarding ‘ZNF354C’ as an isolated node in the network, as shown in Table A.1 (or the column of hESC) in this appendix.

In Section 5.6 of the main text, we also calculated the local controllability of the 712 protein complexes in the TF regulatory network of hESC in a different way. For a protein complex that contains ‘ZNF354C’, we calculated the local controllability of this protein complex in the TF regulatory network of hESC by ignoring the TF ‘ZNF354C’, in other words, it is equal to the local controllability of the subset of TFs in this protein complex except ‘ZNF354C’. Thus, the values of the local controllability of a protein complex that contains ‘ZNF354C’ in the TF regulatory network of hESC, calculated in these two different ways, differ by 1.

| No. | ID | Protein complex | Size | hESC | AE | HSC | ADB | Neu | SM | FB | SF |
|-----|--------|------------------------------------|------|------|----|-----|-----|-----|----|----|----|
| 1 | HC6801 | ETS1;FOXC1;FOXL1;GATA2;ZEB1 | 5 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 2 | HC1040 | ETS1;FOXC1;GATA2;MZF1;ZNF354C | 5 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | HC1027 | ETS1;FOXC1;GATA2;YY1;ZEB1 | 5 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | HC3201 | ETS1;GATA2;SP1;TFAP2A;YY1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 5 | HC3204 | ETS1;FOXC1;GATA2;TFAP2A;YY1 | 5 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 6 | HC6706 | ETS1;GATA2;KLF4;SPI1;TFAP2A | 5 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 7 | HC6759 | BRCA1;ETS1;GATA2;MZF1;PDX1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 8 | HC2673 | ETS1;FOXC1;GATA2;GATA3;YY1 | 5 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 9 | HC4715 | BRCA1;GATA2;GATA3;MAFB;YY1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 10 | HC8852 | ETS1;FOXC1;GATA2;PDX1;YY1 | 5 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 11 | HC5 | ETS1;FOXC1;FOXL1;GATA2;SPIB | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 3 | 1 |
| 12 | HC6678 | ETS1;FOXL1;GATA2;YY1;ZEB1 | 5 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 |
| 13 | HC1001 | BRCA1;FOXC1;GATA2;GATA3;SPIB | 5 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| 14 | HC6601 | ETS1;FOXC1;GATA2;SPIB;ZEB1 | 5 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 15 | HC3165 | FOXC1;GATA2;PDX1;TFAP2A;YY1 | 5 | 1 | 4 | 1 | 2 | 2 | 3 | 2 | 1 |
| 16 | HC4608 | ETS1;FOXC1;GATA2;YY1;ZNF354C | 5 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 17 | HC1365 | BRCA1;ETS1;GATA2;PDX1;ZNF354C | 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 18 | HC4551 | ARNT;ETS1;GATA2;TFAP2A;YY1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 19 | HC2244 | ARID3A;ETS1;GATA2;GATA3;ZNF354C | 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 20 | HC1219 | ETS1;FOXC1;GATA2;PRRX2;TFAP2A | 5 | 1 | 3 | 1 | 2 | 2 | 2 | 3 | 1 |
| 21 | HC1245 | E2F5;EP300;ETS1;GATA2;ZEB1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 22 | HC2310 | ETS1;GATA2;PRRX2;TFAP2A;YY1 | 5 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 1 |
| 23 | HC615 | ARID3A;ELK1;ETS1;GATA2;GATA3 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 24 | HC677 | BRCA1;ETS1;GATA2;YY1;ZEB1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 25 | HC9173 | ETS1;FOXC1;FOXL1;GATA2;GATA3 | 5 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 26 | HC6270 | ETS1;GATA2;GATA3;SPIB;TFAP2A | 5 | 1 | 2 | 1 | 2 | 2 | 3 | 3 | 1 |
| 27 | HC7762 | ETS1;GATA2;GATA3;ZEB1;ZNF354C | 5 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 28 | HC7695 | ETS1;GATA2;GATA3;MZF1;TFAP2A | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 29 | HC6560 | ETS1;GATA2;GATA3;TFAP2A;ZEB1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 30 | HC3808 | ETS1;GATA2;GATA3;SP1;ZEB1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 31 | HC916 | ELK1;ETS1;GATA2;PDX1;TP53 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 32 | HC993 | ELF5;FOXC1;GATA2;KLF4;YY1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| 33 | HC6460 | CDC5L;ETS1;GATA2;GATA3;ZEB1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 34 | HC5160 | ARID3A;ETS1;FOXL1;GATA2;ZEB1 | 5 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 |
| 35 | HC1981 | E2F4;HC5160FOXL1;GATA2;TFAP2A;ZEB1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 36 | HC7444 | CDC5L;ETS1;GATA2;GATA3;YY1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 37 | HC1961 | ARID3A;ETS1;GATA2;ZEB1;ZNF354C | 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 38 | HC9606 | BRCA1;ETS1;GATA2;TFAP2A;ZEB1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 39 | HC1815 | ETS1;GATA2;GATA3;MZF1;PRRX2 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| 40 | HC7304 | ARID3A;ETS1;FOXC1;FOXL1;GATA3 | 5 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 |
| 41 | HC1804 | ETS1;GATA2;MZF1;SP1;ZEB1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 42 | HC8289 | ELK1;ETS1;GATA2;GATA3;MZF1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 43 | HC8288 | FOXC1;GATA2;TFAP2A;YY1;ZNF354C | 5 | 2 | 3 | 1 | 1 | 1 | 2 | 1 | 1 |
| 44 | HC8270 | ETS1;FOXC1;GATA2;MZF1;TFAP2A | 5 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 45 | HC8291 | ETS1;FOXC1;GATA2;MZF1;YY1 | 5 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 46 | HC5561 | BRCA1;ETS1;FOXC1;GATA2;YY1 | 5 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 47 | HC1558 | ETS1;FOXC1;GATA2;MZF1;SP1 | 5 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 48 | HC7065 | ETS1;GATA2;GATA3;SPIB;ZEB1 | 5 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 |
| 49 | HC7024 | ETS1;FOXC1;GATA2;GATA3;ZEB1 | 5 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 50 | HC5811 | ETS1;FOXC1;GATA2;SPIB;ZNF354C | 5 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |

| | | | | | | | | | | | |
|-----|--------|--------------------------------------|---|---|---|---|---|---|---|---|---|
| 51 | HC1412 | ETS1;FOXL1;GATA2;SPIB;TFAP2A | 5 | 1 | 2 | 1 | 2 | 3 | 3 | 4 | 1 |
| 52 | HC5657 | ETS1;FOXC1;FOXL1;GATA2;YY1 | 5 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 53 | HC8464 | ETS1;GATA2;MZF1;SOX10;YY1 | 5 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| 54 | HC2913 | ETS1;GATA2;MZF1;TFAP2A;YY1 | 5 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 55 | HC5942 | ETS1;FOX1;GATA2;PDX1;YY1 | 5 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 2 |
| 56 | HC5386 | ETS1;FOXL1;GATA1;GATA2;YY1 | 5 | 1 | 2 | 1 | 1 | 3 | 2 | 2 | 1 |
| 57 | HC2790 | ETS1;FOXL1;GATA2;GATA3;PDX1 | 5 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 58 | HC1663 | ETS1;GATA2;SOX10;YY1;ZEB1 | 5 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| 59 | HC6806 | ETS1;FOXC1;GATA2;PDX1;YY1;ZEB1 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 60 | HC4860 | ETS1;FOXC1;GATA2;GATA3;REL;ZNF354C | 6 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 61 | HC8803 | ARID3A;ETS1;FOXL1;GATA2;MZF1;SOX10 | 6 | 1 | 1 | 1 | 1 | 3 | 2 | 3 | 1 |
| 62 | HC4757 | ETS1;GATA2;MZF1;SPIB;YY1;ZNF354C | 6 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 |
| 63 | HC8878 | EP300;ETS1;FOXC1;GATA2;GATA3;PRRX2 | 6 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 |
| 64 | HC3139 | ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 65 | HC3119 | ARID3A;ETS1;FOXC1;GATA2;YY1;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 66 | HC1009 | BRCA1;ETS1;GATA2;SOX10;SPIB;TFAP2A | 6 | 1 | 2 | 1 | 3 | 3 | 4 | 4 | 2 |
| 67 | HC6605 | BRCA1;ETS1;GATA2;SOX10;YY1;ZEB1 | 6 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| 68 | HC6603 | ETS1;FOXC1;GATA2;PRRX2;SOX10;ZNF354C | 6 | 2 | 2 | 1 | 2 | 2 | 2 | 3 | 1 |
| 69 | HC4689 | ETS1;GATA2;PRRX2;SPIB;TFAP2A;YY1 | 6 | 1 | 2 | 1 | 3 | 2 | 3 | 4 | 1 |
| 70 | HC2525 | ETS1;FOXC1;GATA2;GATA3;SP1;SPIB | 6 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 71 | HC4617 | ETS1;FOXC1;GATA2;GATA3;PAX2;PRRX2 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 72 | HC1293 | ETS1;FOXC1;GATA2;SPIB;YY1;ZNF354C | 6 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 73 | HC1357 | ETS1;FOXC1;GATA2;NKX3-1;YY1;ZEB1 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 74 | HC6967 | ARID3A;BRCA1;ETS1;GATA2;SRY;ZNF354C | 6 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 75 | HC4561 | ETS1;GATA2;NR4A2;SPIB;ZEB1;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 76 | HC4556 | ETS1;FOXC1;FOXL1;GATA2;SOX10;YY1 | 6 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 77 | HC6904 | BRCA1;ETS1;FOXC1;GATA2;PAX2;ZEB1 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| 78 | HC1138 | ARID3A;ETS1;GATA2;GATA3;SPIB;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 79 | HC4492 | ARID3A;ETS1;FOXL1;GATA2;GATA3;ZEB1 | 6 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 |
| 80 | HC7987 | ETS1;FOXC1;GATA2;GATA3;SOX17;TFAP2A | 6 | 1 | 4 | 1 | 1 | 3 | 3 | 3 | 2 |
| 81 | HC4453 | ETS1;FOXC1;FOXL1;GATA2;SP1;ZEB1 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 82 | HC4439 | ETS1;FOXC1;GATA2;MAFB;MZF1;SOX10 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 83 | HC4434 | BRCA1;ETS1;GATA2;GATA3;PRRX2;SOX10 | 6 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 1 |
| 84 | HC3796 | ETS1;FOXC1;GATA2;TFAP2A;YY1;ZEB1 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 85 | HC3784 | ETS1;FOXC1;GATA2;GATA3;MZF1;TFAP2A | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 86 | HC7903 | ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2 | 6 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 |
| 87 | HC7899 | ETS1;FOXC1;GATA2;MZF1;PDX1;SP1 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 88 | HC670 | ETS1;GATA2;MZF1;NFATC2;NKX2-5;ZEB1 | 6 | 1 | 3 | 1 | 1 | 2 | 3 | 2 | 1 |
| 89 | HC675 | ETS1;FOXC1;FOXL1;GATA2;GATA3;SP1 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 90 | HC660 | ETS1;FOXC1;GATA2;MZF1;YY1;ZNF354C | 6 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 91 | HC7862 | ARNT;ETS1;GATA2;MZF1;TFAP2A;YY1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 92 | HC6387 | ETS1;FOXC1;GATA2;GATA3;TFAP2A;YY1 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 93 | HC2062 | ETS1;FOXC1;GATA2;MZF1;PDX1;TFAP2A | 6 | 1 | 4 | 1 | 2 | 3 | 3 | 3 | 1 |
| 94 | HC7774 | ETS1;FOXC1;GATA2;MZF1;SPIB;TFAP2A | 6 | 1 | 3 | 1 | 2 | 2 | 3 | 3 | 1 |
| 95 | HC7725 | CIZ1;ELK1;ETS1;GATA2;MZF1;ZEB1 | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 96 | HC9298 | ETS1;FOXC1;GATA2;GATA3;MAFB;TFAP2A | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 97 | HC6268 | ETS1;GATA2;GATA3;TFAP2A;YY1;ZEB1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 98 | HC3587 | ARID3A;BRCA1;ETS1;GATA2;ZEB1;ZNF354C | 6 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 99 | HC6507 | ETS1;FOXC1;GATA2;NFKB1;PDX1;TP53 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 100 | HC7692 | ETS1;FOXC1;GATA2;SP1;ZEB1;ZNF354C | 6 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | | | | | |
|-----|--------|--|---|---|---|---|---|---|---|---|---|
| 101 | HC3803 | BRCA1;EP300;ETS1;GATA2;TFAP2A;TP53 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 102 | HC9090 | BRCA1;ETS1;FOXC1;GATA2;TFAP2A;ZNF354C | 6 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 103 | HC7517 | ARID3A;ETS1;FOXC1;GATA2;MZF1;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 104 | HC8126 | ARID3A;ETS1;GATA2;HOXA5;SP1;TFAP2A | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 105 | HC1993 | BRCA1;ETS1;FOXC1;GATA2;ZEB1;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 106 | HC8175 | ETS1;FOXC1;GATA2;NR4A2;SP1;USF1 | 6 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 107 | HC7433 | ETS1;FOXC1;FOXL1;GATA2;GATA3;TFAP2A | 6 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 108 | HC9682 | ETS1;FOXC1;GATA2;GATA3;PDX1;SPIB | 6 | 1 | 3 | 1 | 2 | 2 | 3 | 3 | 2 |
| 109 | HC352 | ETS1;FOXL1;GATA2;GATA3;SPIB;YY1 | 6 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 1 |
| 110 | HC8010 | ETS1;FOXC1;GATA2;GATA3;MZF1;SPIB | 6 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 111 | HC55 | ETS1;GATA2;MZF1;SP1;YY1;ZEB1 | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 112 | HC25 | ETS1;GATA2;SOX10;SP1;TFAP2A;YY1 | 6 | 1 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 113 | HC6122 | BRCA1;ETS1;FOXL1;GATA2;GATA3;TFAP2A | 6 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 114 | HC487 | ETS1;FOXC1;GATA2;GATA3;MZF1;ZNF354C | 6 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 115 | HC9440 | ARID3A;ETS1;FOXC1;GATA2;PRRX2;YY1 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 116 | HC8211 | ETS1;GATA2;GATA3;HOXA5;SP1;ZNF354C | 6 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 117 | HC5062 | ETS1;FOXC1;GATA2;MZF1;REL;SP1 | 6 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 118 | HC4036 | ETS1;GATA2;GATA3;HOXA5;TFAP2A;ZEB1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 119 | HC4074 | BRCA1;ETS1;FOXC1;GATA2;TFAP2A;YY1 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 120 | HC4052 | ETS1;GATA2;PAX2;SP1;SPIB;TFAP2A | 6 | 1 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 121 | HC8502 | FOXC1;GATA2;GATA3;SOX10;SPIB;YY1 | 6 | 1 | 2 | 1 | 2 | 1 | 3 | 2 | 2 |
| 122 | HC5894 | ETS1;GATA2;MZF1;TFAP2A;YY1;ZEB1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 123 | HC4128 | BRCA1;ETS1;FOXC1;GATA2;SOX10;YY1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 124 | HC5927 | BRCA1;ETS1;FOXC1;GATA2;YY1;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 125 | HC5328 | CDC5L;ETS1;GATA2;GATA3;YY1;ZEB1 | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 126 | HC5301 | ETS1;FOXL1;GATA2;TFAP2A;YY1;ZEB1 | 6 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 127 | HC9879 | BRCA1;ETS1;FOXC1;GATA2;PDX1;ZEB1 | 6 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 128 | HC9785 | ETS1;FOXC1;FOXL1;GATA2;GATA3;SOX10 | 6 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 129 | HC1609 | BRCA1;ETS1;GATA2;YY1;ZEB1;ZNF354C | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 130 | HC5428 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 131 | HC5459 | ETS1;GATA2;GATA3;PDX1;PRRX2;TFAP2A | 6 | 1 | 3 | 1 | 3 | 3 | 3 | 4 | 1 |
| 132 | HC5469 | ETS1;FOXC1;GATA2;SOX10;SP1;YY1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 133 | HC4332 | ETS1;GATA2;GATA3;MZF1;SOX10;YY1 | 6 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| 134 | HC1655 | ETS1;FOXC1;GATA2;MZF1;SOX10;YY1 | 6 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 135 | HC8664 | ETS1;GATA2;MZF1;SP1;SPIB;TFAP2A | 6 | 1 | 2 | 1 | 2 | 2 | 3 | 3 | 1 |
| 136 | HC4837 | ETS1;FOXL1;GATA2;MAFB;SPIB;TFAP2A;YY1 | 7 | 1 | 2 | 1 | 2 | 3 | 3 | 4 | 1 |
| 137 | HC4810 | ETS1;FOXC1;GATA2;GATA3;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 138 | HC4806 | ARID3A;ELK1;ETS1;FOXC1;FOXL1;GATA2;NKX2-5 | 7 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 139 | HC4883 | ETS1;GATA2;MZF1;SPIB;TFAP2A;YY1;ZNF354C | 7 | 2 | 2 | 1 | 2 | 2 | 3 | 3 | 1 |
| 140 | HC6819 | BRCA1;ETS1;GATA2;MAFB;SPIB;TFAP2A;YY1 | 7 | 1 | 2 | 1 | 2 | 2 | 3 | 3 | 1 |
| 141 | HC3308 | ETS1;FOXC1;GATA2;GATA3;MAFB;PDX1;YY1 | 7 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 142 | HC8991 | ARID3A;ETS1;FOXC1;GATA2;SP1;TFAP2A;ZNF354C | 7 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 143 | HC6731 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;TFAP2A | 7 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 144 | HC6754 | ETS1;FOXC1;FOXL1;GATA2;GATA3;PAX2;PDX1 | 7 | 1 | 3 | 1 | 1 | 4 | 2 | 3 | 1 |
| 145 | HC2549 | ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;YY1 | 7 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 146 | HC1274 | BRCA1;ETS1;GATA2;SP1;SPIB;TFAP2A;ZEB1 | 7 | 1 | 2 | 1 | 2 | 2 | 3 | 3 | 1 |
| 147 | HC3481 | ETS1;FOXC1;GATA2;GATA3;SP1;YY1;ZEB1 | 7 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 148 | HC4549 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 149 | HC6984 | ETS1;FOXC1;GATA2;GATA3;PDX1;ZEB1;ZNF354C | 7 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 150 | HC6945 | ETS1;FOXC1;GATA2;PRRX2;TFAP2A;ZEB1;ZNF354C | 7 | 2 | 3 | 1 | 2 | 2 | 2 | 3 | 1 |

| | | | | | | | | | | | |
|-----|--------|--|---|---|---|---|---|---|---|---|---|
| 151 | HC2283 | ETS1;FOXC1;GATA2;GATA3;TFAP2A;YY1;ZNF354C | 7 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 152 | HC3774 | ARID3A;ETS1;FOXC1;GATA2;SPIB;YY1;ZEB1 | 7 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 153 | HC2142 | ARID3A;ETS1;GATA2;GATA3;TFAP2A;ZEB1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 154 | HC2136 | BRCA1;ETS1;FOXC1;GATA2;PDX1;TFAP2A;ZEB1 | 7 | 1 | 4 | 1 | 2 | 3 | 3 | 3 | 1 |
| 155 | HC7840 | BRCA1;ETS1;GATA2;SP1;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 156 | HC6325 | ETS1;FOXL1;GATA2;GATA3;TP63;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 3 | 1 | 3 | 1 |
| 157 | HC6367 | ETS1;FOXC1;GATA2;MZF1;TFAP2A;YY1;ZNF354C | 7 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 158 | HC6364 | ETS1;GATA2;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 159 | HC2073 | ARID3A;ELF5;ETS1;GATA2;SP1;TFAP2A;USF1 | 7 | 2 | 2 | 1 | 1 | 3 | 3 | 3 | 1 |
| 160 | HC2054 | ETS1;GATA2;GATA3;PDX1;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 161 | HC7764 | ARID3A;ELF5;ETS1;FOXL1;GATA2;TFAP2A;ZEB1 | 7 | 2 | 2 | 1 | 1 | 4 | 3 | 4 | 1 |
| 162 | HC3589 | ARID3A;ETS1;GATA2;GATA3;MZF1;TFAP2A;ZEB1 | 7 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 163 | HC838 | ARID3A;ETS1;FOXL1;GATA2;KLF4;PDX1;SP1 | 7 | 1 | 2 | 1 | 1 | 4 | 2 | 3 | 1 |
| 164 | HC3912 | ETS1;GATA2;MZF1;SP1;USF1;YY1;ZNF354C | 7 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 165 | HC7628 | ETS1;FOXC1;FOXL1;GATA2;GATA3;SOX10;SPIB | 7 | 1 | 2 | 1 | 2 | 3 | 3 | 4 | 2 |
| 166 | HC7577 | ETS1;FOXC1;GATA2;PDX1;TFAP2A;YY1;ZEB1 | 7 | 1 | 4 | 1 | 2 | 3 | 3 | 3 | 1 |
| 167 | HC9065 | BRCA1;ETS1;FOXC1;GATA2;KLF4;SOX10;ZNF354C | 7 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 |
| 168 | HC8124 | ETS1;GATA2;GATA3;NF1;TFAP2A;YY1;ZEB1 | 7 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 169 | HC5165 | ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX2-5;REL | 7 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 170 | HC5140 | ETS1;FOXL1;GATA2;NKX2-5;PDX1;SPIB;TFAP2A | 7 | 1 | 4 | 1 | 4 | 5 | 5 | 6 | 3 |
| 171 | HC8164 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;NKX2-5;ZNF354C | 7 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 |
| 172 | HC8163 | BRCA1;ETS1;FOXL1;GATA2;PRRX2;SOX10;ZNF354C | 7 | 2 | 2 | 2 | 2 | 3 | 2 | 4 | 2 |
| 173 | HC1984 | BRCA1;ETS1;FOXC1;GATA2;MAFB;TFAP2A;ZEB1 | 7 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 174 | HC8023 | ETS1;FOXC1;GATA2;MZF1;TFAP2A;YY1;ZEB1 | 7 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 175 | HC1885 | BRCA1;ETS1;FOXC1;GATA2;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 176 | HC8091 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;SOX10 | 7 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 177 | HC3064 | ETS1;FOXL1;GATA2;TFAP2A;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 178 | HC9407 | ETS1;FOXC1;FOXL1;GATA2;GATA3;TFAP2A;ZNF354C | 7 | 2 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 179 | HC8290 | BRCA1;CREB1;ETS1;FOXL1;GATA2;HOXA5;SP1 | 7 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 |
| 180 | HC8200 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;MZF1;YY1 | 7 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 181 | HC8206 | ETS1;GATA2;GATA3;PRRX2;SP1;YY1;ZEB1 | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| 182 | HC6036 | ARID3A;ETS1;GATA2;TFAP2A;YY1;ZEB1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 183 | HC5577 | ETS1;FOXC1;GATA2;MZF1;SP1;YY1;ZEB1 | 7 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 184 | HC4056 | BRCA1;ETS1;GATA2;SPIB;TFAP2A;YY1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 |
| 185 | HC1578 | ARID3A;EP300;ETS1;FOXC1;GATA2;SMAD1;TFAP2A | 7 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 186 | HC7077 | ETS1;GATA2;MZF1;SOX10;TFAP2A;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 187 | HC5821 | BRCA1;ETS1;FOXL1;GATA2;GATA3;PRRX2;SOX10 | 7 | 1 | 1 | 1 | 2 | 3 | 2 | 4 | 1 |
| 188 | HC5852 | BRCA1;ETS1;FOXC1;GATA2;GATA3;SP1;ZNF354C | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 189 | HC7059 | ETS1;FOXC1;GATA2;GATA3;PDX1;YY1;ZEB1 | 7 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 190 | HC4146 | ETS1;GATA2;MAFB;MZF1;TFAP2A;ZEB1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 191 | HC4181 | ARID3A;ETS1;FOXC1;GATA2;NKX3-1;TFAP2A;ZNF354C | 7 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 2 |
| 192 | HC9802 | ETS1;FOXC1;GATA2;GATA3;SOX10;USF1;ZEB1 | 7 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 193 | HC9796 | BRCA1;ETS1;FOXC1;GATA2;GATA3;SPIB;YY1 | 7 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 194 | HC4350 | ETS1;FOXC1;GATA2;MZF1;PDX1;TFAP2A;YY1 | 7 | 1 | 4 | 1 | 2 | 3 | 3 | 3 | 1 |
| 195 | HC4369 | ETS1;FOXL1;GATA2;PDX1;USF1;YY1;ZNF354C | 7 | 2 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 196 | HC9734 | ETS1;FOX11;GATA2;MAFB;TFAP2A;YY1;ZNF354C | 7 | 2 | 3 | 1 | 2 | 3 | 3 | 3 | 1 |
| 197 | HC9764 | ARID3A;ETS1;FOXC1;GATA2;KLF4;MZF1;TFAP2A | 7 | 1 | 3 | 1 | 1 | 3 | 2 | 2 | 1 |
| 198 | HC4830 | ETS1;FOXC1;GATA2;MZF1;SOX17;SPIB;TFAP2A;USF1 | 8 | 1 | 4 | 1 | 2 | 3 | 4 | 4 | 2 |
| 199 | HC8943 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2;ZNF354C | 8 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 |
| 200 | HC8984 | ETS1;FOXC1;FOXL1;GATA2;GATA3;SOX17;SP1;TFAP2A | 8 | 1 | 4 | 1 | 1 | 4 | 3 | 4 | 2 |

| | | | | | | | | | | | |
|-----|--------|--|---|---|---|---|---|---|---|---|---|
| 201 | HC1055 | ETS1;FOXC1;GATA2;GATA3;NKX3-1;SOX10;ZEB1;ZNF354C | 8 | 2 | 2 | 1 | 2 | 3 | 2 | 3 | 1 |
| 202 | HC2480 | BPTF;BRCA1;ETS1;FOXC1;GATA2;SPIB;TFAP2A;ZEB1 | 8 | 1 | 3 | 1 | 2 | 2 | 3 | 3 | 1 |
| 203 | HC4702 | ETS1;GATA2;MZF1;NF1;SOX10;TFAP2A;YY1;ZEB1 | 8 | 1 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 204 | HC8811 | ETS1;GATA2;MAX;MZF1;SOX10;TFAP2A;ZEB1;ZNF354C | 8 | 2 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 205 | HC8853 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;PRRX2;TFAP2A;ZEB1 | 8 | 1 | 3 | 1 | 2 | 3 | 2 | 4 | 1 |
| 206 | HC3113 | ETS1;FOXC1;GATA2;NKX3-1;PDX1;SOX10;TFAP2A;ZEB1 | 8 | 1 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 207 | HC6618 | ARID3A;ETS1;FOXC1;GATA2;NFATC2;PRRX2;YY1;ZNF354C | 8 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 2 |
| 208 | HC4693 | ARID3A;ETS1;FOXC1;GATA2;HOXA5;MZF1;TFAP2A;YY1 | 8 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 209 | HC4429 | ARID3A;ETS1;GATA2;HOXA5;MZF1;SP1;TFAP2A;ZEB1 | 8 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 210 | HC6399 | CDC5L;ETS1;FOXC1;GATA2;TFAP2A;YY1;ZEB1;ZNF354C | 8 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 211 | HC646 | ETS1;GATA2;MECP2;MYB;PDX1;SOX10;SP1;ZNF354C | 8 | 2 | 2 | 1 | 2 | 3 | 3 | 4 | 3 |
| 212 | HC6303 | ARID3A;ETS1;FOXL1;GATA2;GATA3;NKX2-5;TFAP2A;ZEB1 | 8 | 1 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 213 | HC2082 | BRCA1;ETS1;FOXC1;GATA1;GATA2;GATA3;ZEB1;ZNF354C | 8 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 214 | HC780 | ETS1;FOXC1;FOXL1;GATA2;NKX3-2;TFAP2A;YY1;ZNF354C | 8 | 2 | 3 | 1 | 2 | 3 | 2 | 3 | 1 |
| 215 | HC6225 | ETS1;FOXC1;GATA2;PDX1;SOX10;SPIB;TFAP2A;YY1 | 8 | 1 | 4 | 1 | 4 | 4 | 5 | 5 | 3 |
| 216 | HC7717 | ETS1;FOXC1;FOXL1;GATA2;PDX1;SOX10;YY1;ZNF354C | 8 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 2 |
| 217 | HC7677 | BRCA1;ETS1;GATA2;GATA3;MZF1;PRRX2;ZEB1;ZNF354C | 8 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 |
| 218 | HC7690 | ETS1;FOXC1;GATA2;GATA3;SOX10;TFAP2A;ZEB1;ZNF354C | 8 | 2 | 3 | 1 | 2 | 3 | 3 | 3 | 1 |
| 219 | HC6591 | ETS1;FOXC1;FOXL1;GATA2;PDX1;SP1;YY1;ZNF354C | 8 | 2 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 220 | HC7607 | ARID3A;ETS1;FOXC1;GATA2;GATA3;SPIB;YY1;ZNF354C | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 221 | HC3807 | BRCA1;ETS1;FOXC1;GATA2;MZF1;SOX10;SP1;YY1 | 8 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 222 | HC4953 | ETS1;FOXC1;GATA2;GATA3;MZF1;SP1;YY1;ZEB1 | 8 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 223 | HC6412 | ETS1;FOXC1;GATA2;GATA3;HOXA5;TFAP2A;YY1;ZNF354C | 8 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 224 | HC372 | ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;SPIB;YY1 | 8 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 225 | HC1898 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;NKX2-5;PDX1;ZNF354C | 8 | 2 | 4 | 2 | 2 | 4 | 3 | 4 | 2 |
| 226 | HC20 | BRCA1;ELK1;ETS1;FOXC1;GATA2;MZF1;SPIB;YY1 | 8 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 227 | HC8384 | ETS1;FOXC1;GATA2;GATA3;MAFB;PRRX2;SOX10;YY1 | 8 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 1 |
| 228 | HC7260 | ETS1;FOXC1;GATA2;MYCN;SPIB;TFAP2A;YY1;ZNF354C | 8 | 2 | 4 | 1 | 2 | 2 | 3 | 3 | 2 |
| 229 | HC8303 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;SPIB;YY1 | 8 | 1 | 2 | 1 | 1 | 2 | 2 | 3 | 1 |
| 230 | HC6008 | ETS1;GATA2;KLF4;MAFB;MZF1;TFAP2A;ZEB1;ZNF354C | 8 | 2 | 2 | 1 | 1 | 3 | 2 | 2 | 1 |
| 231 | HC9389 | ETS1;FOXC1;GATA2;GATA3;HOXA5;SOX10;SPIB;TFAP2A | 8 | 1 | 3 | 1 | 3 | 3 | 4 | 4 | 2 |
| 232 | HC6024 | ETS1;FOXC1;FOXL1;NKX2-5;PDX1;SP1;YY1;ZNF354C | 8 | 2 | 3 | 1 | 2 | 4 | 2 | 4 | 2 |
| 233 | HC6045 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;MZF1;YY1;ZEB1 | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 234 | HC5553 | ETS1;FOXC1;GATA2;MZF1;PDX1;TFAP2A;YY1;ZEB1 | 8 | 1 | 4 | 1 | 2 | 3 | 3 | 3 | 1 |
| 235 | HC5534 | ETS1;FOXC1;FOXL1;GATA2;PAX2;PDX1;SPIB;TFAP2A | 8 | 1 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 236 | HC8573 | ETS1;GATA2;MZF1;NF1;PDX1;SOX10;TFAP2A;YY1 | 8 | 1 | 3 | 1 | 3 | 4 | 4 | 4 | 2 |
| 237 | HC8583 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;PDX1;YY1;ZNF354C | 8 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 |
| 238 | HC5594 | ETS1;FOXC1;GATA2;KLF4;SOX17;TFAP2A;YY1;ZNF354C | 8 | 2 | 4 | 2 | 2 | 4 | 3 | 3 | 2 |
| 239 | HC8507 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;MZF1;TFAP2A;ZEB1 | 8 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 240 | HC5661 | ETS1;FOXC1;GATA2;GATA3;MAFB;PDX1;TFAP2A;YY1 | 8 | 1 | 4 | 1 | 2 | 3 | 3 | 3 | 1 |
| 241 | HC5695 | ETS1;FOXC1;FOXL1;GATA2;GATA3;PDX1;TFAP2A;ZEB1 | 8 | 1 | 4 | 1 | 2 | 4 | 3 | 4 | 1 |
| 242 | HC5631 | ETS1;FOXC1;FOXL1;GATA2;NKX2-5;SPIB;TFAP2A;ZEB1 | 8 | 1 | 4 | 1 | 3 | 4 | 4 | 5 | 2 |
| 243 | HC8467 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;SOX10;TFAP2A;YY1 | 8 | 1 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 244 | HC2905 | ARID3A;ETS1;GATA2;GATA3;TFAP2A;YY1;ZEB1;ZNF354C | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 245 | HC1398 | ETS1;GATA2;GATA3;MZF1;SOX10;SP1;TFAP2A;YY1 | 8 | 1 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 246 | HC4215 | ETS1;FOXC1;GATA2;GATA3;PDX1;RELA;SPIB;ZNF354C | 8 | 2 | 3 | 1 | 2 | 2 | 3 | 3 | 2 |
| 247 | HC4224 | ETS1;FOXC1;FOXL1;GATA2;MAFB;PDX1;SOX17;TFAP2A | 8 | 1 | 5 | 1 | 2 | 5 | 4 | 5 | 2 |
| 248 | HC2751 | ETS1;GATA2;MAFB;MZF1;SP1;TFAP2A;YY1;ZEB1 | 8 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 249 | HC1613 | ETS1;GATA2;GATA3;MAFB;MZF1;SOX10;SP1;TFAP2A | 8 | 1 | 2 | 1 | 2 | 3 | 3 | 3 | 1 |
| 250 | HC176 | ARID3A;ETS1;FOXL1;GATA2;MZF1;TFAP2A;YY1;ZNF354C | 8 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 251 | HC4342 | BRCA1;ETS1;FOXC1;GATA2;GATA3;SPIB;TFAP2A;YY1 | 8 | 1 | 3 | 1 | 2 | 2 | 3 | 3 | 1 |
| 252 | HC4365 | ARID3A;ETS1;FOXC1;GATA2;HOXA5;MZF1;NKX2-5;ZEB1 | 8 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 253 | HC8691 | ETS1;FOXC1;GATA2;MAFB;SOX10;YY1;ZEB1;ZNF354C | 8 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 254 | HC9772 | ARID3A;ETS1;FOXL1;GATA2;HOXA5;MZF1;ZEB1;ZNF354C | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 255 | HC1068 | ETS1;FOXC1;GATA2;MAFB;PAX2;SOX10;TFAP2A;ZEB1;ZNF354C | 9 | 2 | 3 | 1 | 2 | 4 | 3 | 3 | 1 |
| 256 | HC1108 | ELF5;ETS1;FOXC1;GATA2;GATA3;SPIB;TFAP2A;YY1;ZNF354C | 9 | 2 | 3 | 1 | 2 | 3 | 4 | 4 | 2 |
| 257 | HC4713 | ETS1;FOXC1;GATA2;GATA3;MZF1;SP1;TFAP2A;ZEB1;ZNF354C | 9 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 258 | HC6703 | ARID3A;ETS1;FOXC1;FOXO1;GATA2;NKX2-5;REL;SP1;YY1 | 9 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 259 | HC6693 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;YY1;ZNF354C | 9 | 2 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 260 | HC4616 | BRCA1;ETS1;FOXC1;GATA2;GATA3;PRRX2;SOX10;ZEB1;ZNF354C | 9 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |
| 261 | HC2228 | ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;SOX10;SP1;TFAP2A | 9 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 1 |
| 262 | HC3764 | BRCA1;ETS1;FOXC1;GATA2;MAX;MZF1;NR2E3;TP53;YY1 | 9 | 1 | 3 | 1 | 1 | 1 | 1 | 2 | 2 |
| 263 | HC2132 | ETS1;FOXC1;FOXL1;GATA2;NKX3-2;SPIB;TFAP2A;YY1;ZNF354C | 9 | 2 | 3 | 1 | 2 | 3 | 3 | 4 | 1 |
| 264 | HC2125 | ETS1;FOXC1;GATA2;MAFB;MZF1;PAX2;SOX10;TFAP2A;ZNF354C | 9 | 2 | 3 | 1 | 2 | 4 | 3 | 3 | 1 |
| 265 | HC640 | ETS1;FOXC1;GATA2;HOXA5;MZF1;PRRX2;SOX10;YY1;ZEB1 | 9 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 1 |
| 266 | HC6301 | BRCA1;ETS1;FOXC1;GATA2;GATA3;HOXA5;MAFB;TFAP2A;ZEB1 | 9 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 267 | HC9166 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;SOX17;SPIB;TFAP2A | 9 | 2 | 4 | 3 | 3 | 4 | 4 | 5 | 3 |
| 268 | HC9149 | ARID3A;ETS1;FOXC1;GATA2;HOXA5;MZF1;SOX10;ZEB1;ZNF354C | 9 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 269 | HC7868 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PRRX2;ZNF354C | 9 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |
| 270 | HC7736 | ETS1;FOXL1;GATA2;GATA3;MAFB;PRRX2;SP1;TFAP2A;ZEB1 | 9 | 1 | 2 | 1 | 2 | 3 | 2 | 4 | 1 |
| 271 | HC7730 | ARID3A;ETS1;FOXL1;GATA2;GATA3;NKX2-5;TFAP2A; ZEB1;ZNF354C | 9 | 2 | 3 | 2 | 2 | 4 | 3 | 4 | 2 |
| 272 | HC6478 | ETS1;FOXL1;GATA2;MZF1;NKX2-5;PDX1;SOX10;SPIB;TFAP2A | 9 | 1 | 4 | 1 | 5 | 6 | 6 | 7 | 4 |
| 273 | HC8100 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;SP1 | 9 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 274 | HC8147 | ETS1;GATA2;GATA3;MZF1;NKX2-5;REL;SPIB;TFAP2A;YY1 | 9 | 1 | 3 | 1 | 3 | 3 | 4 | 4 | 2 |
| 275 | HC9619 | ARID3A;ETS1;FOXC1;GABPA;GATA2;PRRX2;TFAP2A;YY1;ZNF354C | 9 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 |
| 276 | HC7490 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;SP1;SPIB;ZNF354C | 9 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 |
| 277 | HC5279 | ARID3A;ETS1;FOXC1;GATA2;MZF1;NKX2-5;SOX10; TFAP2A;ZNF354C | 9 | 2 | 4 | 2 | 3 | 4 | 4 | 4 | 2 |
| 278 | HC5226 | ARID3A;ETS1;GATA2;HOXA5;MZF1;SP1;SPIB;TFAP2A;ZEB1 | 9 | 1 | 2 | 1 | 2 | 2 | 3 | 3 | 1 |
| 279 | HC8012 | BRCA1;ETS1;GATA2;GATA3;MZF1;PRRX2;SP1;ZEB1;ZNF354C | 9 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 |
| 280 | HC1828 | ARID3A;BPTF;BRCA1;ETS1;FOXC1;GATA2;SPIB;TFAP2A;ZEB1 | 9 | 2 | 3 | 2 | 2 | 2 | 3 | 3 | 2 |
| 281 | HC41 | BRCA1;ETS1;GATA2;GATA3;MAFB;MZF1;PDX1;SOX10;ZEB1 | 9 | 1 | 2 | 1 | 2 | 3 | 3 | 3 | 2 |
| 282 | HC9520 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;SP1;TFAP2A;YY1 | 9 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 283 | HC8398 | BRCA1;ELK1;ETS1;FOXC1;GATA2;MZF1;PDX1;YY1;ZEB1 | 9 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 284 | HC5029 | ETS1;FOXC1;FOXL1;GATA2;GATA3;SP1;SPIB;ZEB1;ZNF354C | 9 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 1 |
| 285 | HC7115 | ETS1;FOXC1;FOXL1;GATA2;GATA3;PAX2;PDX1;ZEB1;ZNF354C | 9 | 2 | 3 | 1 | 1 | 4 | 2 | 3 | 1 |
| 286 | HC9373 | ETS1;FOXC1;FOXL1;GATA2;GATA3;NR4A2;SPIB;YY1;ZNF354C | 9 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 |
| 287 | HC5506 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;NF1;PDX1;TFAP2A;ZNF354C | 9 | 2 | 4 | 2 | 2 | 4 | 3 | 4 | 2 |
| 288 | HC7082 | ARID3A;BRCA1;ETS1;GATA2;MAX;MZF1;TFAP2A;TP53;ZEB1 | 9 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 289 | HC7084 | CDC5L;ETS1;FOXC1;GATA2;GATA3;SPIB;YY1;ZEB1;ZFX | 9 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 290 | HC4139 | BRCA1;ETS1;FOXC1;GATA2;GATA3;SOX10;SPIB;ZEB1;ZNF354C | 9 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 |
| 291 | HC5635 | BRCA1;ETS1;FOXC1;GATA2;GATA3;PDX1;SPIB;YY1;ZEB1 | 9 | 1 | 3 | 1 | 2 | 2 | 3 | 3 | 2 |
| 292 | HC5860 | ETS1;FOXC1;FOXL1;GATA2;PDX1;TFAP2A;YY1;ZEB1;ZNF354C | 9 | 2 | 4 | 1 | 2 | 4 | 3 | 4 | 1 |
| 293 | HC1442 | CREB1;ETS1;FOXC1;GATA2;GATA3;PRRX2;SOX10;YY1;ZEB1 | 9 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 1 |
| 294 | HC9790 | BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;PDX1;YY1;ZEB1 | 9 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 295 | HC4385 | ARID3A;BRCA1;ETS1;FOXL1;GATA2;MAX;MZF1;TP53;YY1 | 9 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 296 | HC4390 | ARID3A;ETS1;GATA2;GATA3;MZF1;NKX2-5;SOX10; TFAP2A;ZNF354C | 9 | 2 | 3 | 2 | 3 | 4 | 4 | 4 | 2 |
| 297 | HC4738 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;SOX10;TFAP2A; ZEB1;ZNF354C | 10 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 298 | HC3123 | ETS1;FOXC1;GATA2;MAFB;PAX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 10 | 2 | 3 | 1 | 2 | 4 | 3 | 3 | 1 |
| 299 | HC6684 | ELF5;ETS1;FOXC1;FOXL1;GATA2;MZF1;TFAP2A;YY1;ZEB1;ZNF354C | 10 | 2 | 3 | 1 | 1 | 4 | 3 | 4 | 1 |
| 300 | HC6669 | ETS1;FOXC1;FOXL1;GATA2;MZF1;PRRX2;SOX10;TFAP2A;YY1;ZEB1 | 10 | 1 | 3 | 1 | 3 | 4 | 3 | 5 | 1 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 301 | HC4644 | ARID3A;CEBPA;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;SOX10 | 10 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 2 |
| 302 | HC3509 | ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;SPIB;TFAP2A;YY1;ZNF354C | 10 | 2 | 4 | 1 | 3 | 3 | 4 | 4 | 2 |
| 303 | HC4580 | CDC5L;ETS1;FOXC1;FOXL1;GATA2;HOXA5;MZF1;SOX10;TFAP2A;ZNF354C | 10 | 2 | 3 | 2 | 2 | 4 | 3 | 4 | 2 |
| 304 | HC1218 | ETS1;FOXL1;GATA2;GATA3;MAFB;MZF1;SP1;TFAP2A;YY1;ZEB1 | 10 | 1 | 2 | 1 | 1 | 3 | 2 | 3 | 1 |
| 305 | HC4410 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;SOX10;SP1;TFAP2A;YY1 | 10 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 1 |
| 306 | HC3760 | CDC5L;ETS1;FOXC1;GATA2;GATA3;MZF1;SPIB;TFAP2A;YY1;ZNF354C | 10 | 2 | 3 | 2 | 2 | 2 | 3 | 3 | 2 |
| 307 | HC6280 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;SOX10;SP1;TFAP2A;ZEB1 | 10 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 1 |
| 308 | HC3633 | ARID3A;CREB1;ETS1;GATA2;GATA3;MYB;MZF1;SOX10;SP1;YY1 | 10 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 309 | HC6588 | ETS1;FOXC1;GATA2;GATA3;MZF1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 10 | 2 | 2 | 1 | 2 | 2 | 3 | 3 | 2 |
| 310 | HC6589 | ARID3A;ARNT;BRCA1;ETS1;FOXL1;GATA2;PRRX2;REL;TFAP2A;YY1 | 10 | 2 | 2 | 2 | 2 | 3 | 2 | 4 | 2 |
| 311 | HC3839 | ETS1;FOXC1;GATA2;KLF4;MZF1;PDX1;SPIB;TFAP2A;YY1;ZEB1 | 10 | 1 | 4 | 1 | 3 | 4 | 4 | 4 | 2 |
| 312 | HC6464 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;NFATC2;NKX2-5;TFAP2A;YY1;ZNF354C | 10 | 2 | 5 | 2 | 2 | 4 | 4 | 4 | 2 |
| 313 | HC9051 | ETS1;GATA2;MZF1;NKX2-5;REL;SP1;SPIB;TFAP2A;YY1;ZEB1 | 10 | 1 | 3 | 1 | 3 | 3 | 4 | 4 | 2 |
| 314 | HC1937 | ETS1;FOXC1;GATA2;GATA3;NKX2-5;PDX1;SP1;TBP;TFAP2A;ZNF354C | 10 | 2 | 5 | 1 | 3 | 4 | 4 | 4 | 2 |
| 315 | HC5284 | ETS1;FOXC1;GATA2;GATA3;MZF1;REL;SPIB;YY1;ZEB1;ZNF354C | 10 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 316 | HC384 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;TFAP2A;YY1;ZEB1;ZNF354C | 10 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 317 | HC333 | BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;YY1;ZNF354C | 10 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| 318 | HC7370 | ELK1;ETS1;FOXC1;GATA2;HSF2;MZF1;PDX1;SPIB;TFAP2A;YY1 | 10 | 1 | 4 | 1 | 3 | 3 | 4 | 4 | 2 |
| 319 | HC7251 | ETS1;FOXC1;FOXL1;GATA2;GATA3;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 10 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 320 | HC9384 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;NF1;NKX2-5;TFAP2A;ZNF354C | 10 | 2 | 4 | 2 | 2 | 5 | 3 | 4 | 2 |
| 321 | HC8259 | ARID3A;ETS1;GATA2;GATA3;MZF1;NKX2-5;PDX1;SOX10;TFAP2A;ZNF354C | 10 | 2 | 4 | 2 | 4 | 5 | 5 | 5 | 3 |
| 322 | HC587 | ARID3A;ETS1;FOXC1;GATA2;GATA3;HOXA5;PRRX2;SOX10;YY1;ZNF354C | 10 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |
| 323 | HC5048 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;SOX10;SP1;YY1;ZNF354C | 10 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 324 | HC5099 | ARID3A;BCL6;ETS1;FOXC1;FOXL1;GATA2;JUND;MZF1;YY1;ZEB1 | 10 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| 325 | HC5669 | ETS1;FOXL1;GATA2;MZF1;NKX2-5;PDX1;SOX10;SPIB;TFAP2A;ZNF354C | 10 | 2 | 4 | 1 | 5 | 6 | 6 | 7 | 4 |
| 326 | HC4166 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;NF1;PDX1;TFAP2A | 10 | 1 | 4 | 1 | 2 | 4 | 3 | 4 | 1 |
| 327 | HC5878 | BRCA1;ETS1;FOXC1;GATA2;MZF1;PDX1;SPIB;YY1;ZEB1;ZNF354C | 10 | 2 | 3 | 2 | 2 | 2 | 3 | 3 | 2 |
| 328 | HC2942 | ARID3A;ETS1;FOXC1;GATA2;MAFB;MZF1;SP1;TFAP2A;YY1;ZEB1 | 10 | 1 | 3 | 1 | 1 | 2 | 2 | 2 | 1 |
| 329 | HC2869 | ELF5;ETS1;FOXC1;GATA2;GATA3;KLF4;PDX1;TBP;TFAP2A;YY1 | 10 | 1 | 4 | 1 | 2 | 5 | 4 | 4 | 2 |
| 330 | HC2844 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2;SOX10;ZEB1;ZNF354C | 10 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 331 | HC9763 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;MZF1;PDX1; TFAP2A;ZEB1;ZNF354C | 10 | 2 | 4 | 2 | 2 | 4 | 3 | 4 | 2 |
| 332 | HC4840 | ARID3A;ETS1;FOXL1;GATA2;HOXA5;MZF1;SOX10;SP1; TFAP2A;ZEB1;ZNF354C | 11 | 2 | 2 | 2 | 2 | 4 | 3 | 4 | 2 |
| 333 | HC8945 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;PDX1; PRRX2;TFAP2A;YY1 | 11 | 1 | 4 | 1 | 3 | 3 | 3 | 4 | 2 |
| 334 | HC3104 | ELK1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;PDX1; YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 1 | 1 | 4 | 2 | 3 | 1 |
| 335 | HC2534 | ARID3A;BRCA1;ETS1;FOXL1;GATA2;GATA3;MZF1;SOX10; SP1;TFAP2A;YY1 | 11 | 2 | 2 | 2 | 2 | 4 | 3 | 4 | 2 |
| 336 | HC4638 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;PRRX2; REL;SOX10;ZNF354C | 11 | 2 | 2 | 1 | 2 | 3 | 2 | 4 | 1 |
| 337 | HC2357 | ELF5;ETS1;FOXC1;GATA2;GATA3;KLF4;PDX1;SOX10; TBP;TFAP2A;YY1 | 11 | 1 | 4 | 1 | 3 | 6 | 5 | 5 | 3 |
| 338 | HC3477 | ETS1;FOXC1;FOX11;FOXL1;GATA2;MZF1;PDX1;SOX10; SPIB;TFAP2A;ZNF354C | 11 | 2 | 5 | 1 | 5 | 6 | 6 | 7 | 4 |
| 339 | HC4560 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;PRRX2; SOX10;YY1;ZNF354C | 11 | 2 | 3 | 2 | 3 | 3 | 3 | 4 | 2 |
| 340 | HC4554 | ETS1;FOXC1;GATA2;GATA3;MZF1;NKX2-5;PRRX2;SOX10; TFAP2A;YY1;ZNF354C | 11 | 2 | 4 | 1 | 4 | 4 | 4 | 5 | 2 |
| 341 | HC2279 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MAFB;MZF1; SOX10;YY1;ZNF354C | 11 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 |
| 342 | HC4515 | BRCA1;ETS1;FOXC1;GATA2;GATA3;PDX1;SP1;SPIB; TFAP2A;YY1;ZEB1 | 11 | 1 | 4 | 1 | 3 | 3 | 4 | 4 | 2 |
| 343 | HC4470 | BRCA1;ETS1;FOXC1;GATA2;GATA3;NKX3-1;SOX10; TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 2 |
| 344 | HC3736 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;MZF1;PDX1;SOX10; TFAP2A;YY1;ZEB1 | 11 | 1 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 345 | HC2149 | ARID3A;ETS1;GATA2;HOXA5;MZF1;NKX2-5;PRRX2;SOX10; YY1;ZEB1;ZNF354C | 11 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 2 |
| 346 | HC6305 | ETS1;FOXC1;GATA2;GATA3;HOXA5;NKX3-1;SOX10;TFAP2A; YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 347 | HC7837 | BRCA1;ETS1;FOXA1;FOXC1;FOXL1;GATA2;GATA3;SOX10; TFAP2A;ZEB1;ZNF354C | 11 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 348 | HC9174 | ARID3A;ETS1;FOXC1;GATA2;KLF4;MZF1;PDX1;SPIB;TFAP2A; YY1;ZNF354C | 11 | 2 | 4 | 2 | 3 | 4 | 4 | 4 | 2 |
| 349 | HC3943 | ETS1;FOXC1;FOX11;FOXL1;GATA2;KLF4;MZF1;NKX2- 5;PDX1; TFAP2A;YY1 | 11 | 1 | 6 | 1 | 4 | 7 | 5 | 6 | 3 |
| 350 | HC7671 | EBF1;ELK1;ETS1;FOXC1;FOXL1;GATA2;NKX2- 5;SPIB;TFAP2A; ZEB1;ZNF354C | 11 | 2 | 4 | 1 | 3 | 4 | 4 | 5 | 2 |
| 351 | HC6583 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;NKX3- 1;PDX1; TFAP2A;YY1 | 11 | 2 | 4 | 2 | 3 | 5 | 3 | 5 | 2 |
| 352 | HC4979 | ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX2- 5;PAX2;SP1;TFAP2A; USF1;YY1 | 11 | 1 | 4 | 1 | 2 | 5 | 3 | 4 | 1 |
| 353 | HC3829 | EP300;ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX2- 5;SOX10;TFAP2A; ZEB1;ZNF354C | 11 | 2 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 354 | HC4889 | ARNT;ETS1;GATA2;GATA3;PDX1;REL;SPIB;TFAP2A;YY1; ZEB1;ZNF354C | 11 | 2 | 3 | 1 | 3 | 3 | 4 | 4 | 2 |
| 355 | HC9023 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;PAX2;SOX10;TFAP2A; TP53;YY1;ZEB1 | 11 | 1 | 3 | 1 | 2 | 5 | 3 | 4 | 1 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 356 | HC259 | ETS1;FOXC1;GATA2;GATA3;NR4A2;PDX1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 2 | 3 | 3 | 4 | 4 | 3 |
| 357 | HC8158 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2;SPIB;YY1;ZEB1;ZNF354C | 11 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |
| 358 | HC7450 | ETS1;FOXC1;FOXJ1;FOXJ1;GATA2;GATA3;PDX1;SOX10;SPIB;TFAP2A;YY1 | 11 | 1 | 5 | 1 | 5 | 6 | 6 | 7 | 4 |
| 359 | HC9684 | BRCA1;ETS1;FOXC1;FOXJ1;GATA2;GATA3;MZF1;SP1;TFAP2A;YY1;ZNF354C | 11 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 |
| 360 | HC9662 | ARID3A;ETS1;FOXC1;GATA2;GATA3;NKX2-5;NR4A2;PDX1;SOX10;ZEB1;ZNF354C | 11 | 3 | 4 | 3 | 3 | 4 | 4 | 4 | 3 |
| 361 | HC8067 | ETS1;FOXC1;GATA2;GATA3;MZF1;NFATC2;PDX1;PRRX2;SOX10;YY1;ZNF354C | 11 | 2 | 4 | 1 | 3 | 3 | 4 | 4 | 2 |
| 362 | HC7230 | BRCA1;CDC5L;ETS1;FOXC1;FOXJ1;GATA2;GATA3;SOX10;SPIB;YY1;ZEB1 | 11 | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 2 |
| 363 | HC7237 | E2F1;ETS1;FOXC1;GATA2;GATA3;MZF1;PAX2;PDX1;SPIB;YY1;ZEB1 | 11 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 2 |
| 364 | HC6007 | ARID3A;ETS1;FOXC1;GATA2;GATA3;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 2 | 3 | 3 | 4 | 4 | 2 |
| 365 | HC9382 | E2F1;ETS1;FOXC1;GATA2;GATA3;NKX2-5;PAX2;SP1;SPIB;YY1;ZEB1 | 11 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 2 |
| 366 | HC8248 | ARID3A;ETS1;FOXC1;FOXJ1;GATA2;GATA3;MAX;MZF1;PDX1;SP1;ZEB1 | 11 | 1 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |
| 367 | HC9343 | ETS1;FOXC1;GATA2;GATA3;HSF1;HSF2;MAFB;NKX2-5;PRRX2;TFAP2A;YY1 | 11 | 1 | 4 | 1 | 3 | 3 | 3 | 4 | 1 |
| 368 | HC6047 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 369 | HC4131 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;MYB;MZF1;PDX1;SOX10;TFAP2A;YY1 | 11 | 2 | 4 | 2 | 3 | 4 | 4 | 4 | 3 |
| 370 | HC4135 | ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 1 | 3 | 2 | 3 | 4 | 1 |
| 371 | HC8418 | ETS1;FOXC1;FOXJ1;GATA2;MZF1;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 4 | 1 | 3 | 4 | 4 | 5 | 2 |
| 372 | HC5989 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2;SP1;SPIB;TFAP2A;ZEB1 | 11 | 1 | 3 | 1 | 3 | 2 | 3 | 4 | 1 |
| 373 | HC1775 | BRCA1;ETS1;FOXC1;GATA2;MZF1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 2 | 3 | 3 | 3 | 4 | 2 |
| 374 | HC9792 | ETS1;FOXC1;GATA2;GATA3;MZF1;NKX3-1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 375 | HC1635 | ETS1;FOXC1;GATA2;MAFB;MZF1;PAX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 11 | 2 | 3 | 1 | 2 | 4 | 3 | 3 | 1 |
| 376 | HC4330 | ETS1;FOXC1;GATA2;MAFB;PAX2;REL;SOX10;TFAP2A;TP53;YY1;ZEB1 | 11 | 1 | 3 | 1 | 2 | 4 | 3 | 3 | 1 |
| 377 | HC2617 | ARID3A;EN1;ETS1;FOXC1;FOXJ1;GATA2;MZF1;PRRX2;TFAP2A;USF1;YY1;ZEB1 | 12 | 1 | 4 | 2 | 3 | 4 | 2 | 5 | 1 |
| 378 | HC4770 | ARID3A;BRCA1;ETS1;FOXC1;FOXJ1;GATA2;NFYA;PRRX2;RORA;TFAP2A;YY1;ZEB1 | 12 | 2 | 3 | 2 | 3 | 4 | 2 | 5 | 2 |
| 379 | HC3294 | ELF5;ETS1;FOXC1;FOXJ1;GATA2;GATA3;KLF4;PDX1;SOX10;TBP;TFAP2A;YY1 | 12 | 1 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 380 | HC6698 | ETS1;FOXC1;FOXJ1;GATA2;GATA3;MZF1;SP1;TFAP2A;USF1;YY1;ZEB1;ZNF354C | 12 | 2 | 3 | 1 | 1 | 3 | 2 | 3 | 1 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 381 | HC4688 | ELK1;ETS1;FOXC1;GATA2;GATA3;IKZF1;PAX2;PRRX2; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 3 | 1 | 3 | 4 | 3 | 4 | 1 |
| 382 | HC4630 | ETS1;FOXC1;FOXL1;GATA2;GATA3;NFYA;PDX1;SOX10; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 383 | HC4614 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1; PRRX2;SOX10;TFAP2A;YY1 | 12 | 1 | 3 | 1 | 3 | 5 | 3 | 5 | 1 |
| 384 | HC4583 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;SOX10; SP1;SPIB;ZEB1;ZNF354C | 12 | 2 | 3 | 1 | 3 | 4 | 4 | 5 | 3 |
| 385 | HC9117 | ARID3A;BRCA1;ETS1;GATA2;GATA3;NKX2-5;NR4A2;PDX1; SOX10;SP1;ZEB1;ZNF354C | 12 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 386 | HC9146 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MAFB;MZF1;SP1; TFAP2A;YY1;ZEB1 | 12 | 1 | 3 | 1 | 1 | 3 | 2 | 2 | 1 |
| 387 | HC6353 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1; SPIB;YY1;ZEB1;ZNF354C | 12 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 388 | HC3651 | CREB1;ETS1;FOXC1;GATA2;GATA3;MZF1;NKX2-5;PRRX2; SOX10;TFAP2A;YY1;ZEB1 | 12 | 1 | 4 | 1 | 4 | 4 | 4 | 5 | 2 |
| 389 | HC6289 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;SOX10; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 3 | 2 | 2 | 4 | 3 | 4 | 2 |
| 390 | HC721 | ETS1;FOXC1;GATA2;MAFB;MZF1;PAX2;PDX1;SOX10;TFAP2A; YY1;ZEB1;ZNF354C | 12 | 2 | 4 | 1 | 3 | 5 | 4 | 4 | 2 |
| 391 | HC7688 | ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1;SOX10; SPIB;TFAP2A;YY1;ZEB1 | 12 | 1 | 3 | 1 | 3 | 5 | 4 | 5 | 2 |
| 392 | HC3834 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX3-1;SOX10; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 3 | 2 | 3 | 5 | 3 | 5 | 2 |
| 393 | HC3871 | ETS1;FOXL1;GATA2;KLF4;MYB;MZF1;NKX3-1;SP1; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 2 | 1 | 2 | 5 | 2 | 4 | 1 |
| 394 | HC9005 | BRCA1;ETS1;GATA2;GATA3;KLF4;MAFB;MZF1;PDX1; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 3 | 2 | 2 | 4 | 3 | 3 | 2 |
| 395 | HC6452 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;PDX1; TBP;TFAP2A;YY1;ZEB1 | 12 | 1 | 4 | 1 | 2 | 6 | 4 | 5 | 2 |
| 396 | HC6443 | ARID3A;CDC5L;ETS1;FOXC1;FOXO3;GATA2;GATA3;MYB; SOX10;TFAP2A;YY1;ZEB1 | 12 | 1 | 3 | 2 | 2 | 3 | 3 | 3 | 2 |
| 397 | HC8130 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NKX2-5; PDX1;TFAP2A;YY1;ZNF354C | 12 | 2 | 5 | 2 | 3 | 5 | 4 | 5 | 2 |
| 398 | HC194 | ELF5;ETS1;FOXC1;GATA2;GATA3;KLF4;PDX1;PRRX2; SOX10;TBP;TFAP2A;YY1 | 12 | 1 | 4 | 1 | 4 | 6 | 5 | 6 | 3 |
| 399 | HC5196 | EN1;ETS1;FOXC1;GATA2;GATA3;KLF4;MAFB;SOX10; SPIB;TFAP2A;ZEB1;ZNF354C | 12 | 2 | 4 | 2 | 4 | 5 | 4 | 5 | 2 |
| 400 | HC8385 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2; SOX10;SP1;SPIB;TFAP2A | 12 | 1 | 4 | 1 | 5 | 5 | 5 | 7 | 3 |
| 401 | HC9478 | BRCA1;CREB1;ETS1;FOXL1;GATA2;GATA3;HOXA5;MZF1; PDX1;SOX10;SOX17;TFAP2A | 12 | 1 | 4 | 2 | 3 | 6 | 5 | 6 | 3 |
| 402 | HC7289 | ETS1;FOXC1;FOXL1;GATA2;PAX2;PDX1;SP1;SPIB; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 403 | HC6199 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1; PDX1;PRRX2;ZEB1;ZNF354C | 12 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 |
| 404 | HC5036 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;MZF1;PDX1;SPIB; TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 4 | 2 | 3 | 4 | 4 | 5 | 2 |
| 405 | HC524 | ARID3A;ELK1;ETS1;FOXC1;GATA2;GATA3;KLF4;MZF1; PDX1;SPIB;YY1;ZEB1 | 12 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 2 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 406 | HC7108 | EGR1;ETS1;FOXC1;GATA2;GATA3;MZF1;NR4A2;PDX1;TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 4 | 2 | 2 | 3 | 3 | 3 | 2 |
| 407 | HC5084 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MAFB;MZF1;PDX1;SP1;TFAP2A;YY1 | 12 | 1 | 4 | 1 | 2 | 4 | 3 | 3 | 1 |
| 408 | HC4080 | ARID3A;ETS1;FOXL1;GATA2;GATA3;MZF1;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 12 | 2 | 2 | 2 | 2 | 4 | 3 | 4 | 2 |
| 409 | HC5592 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;HOXA5;MAFB;SOX10;SP1;TFAP2A;ZEB1;ZNF354C | 12 | 2 | 3 | 2 | 2 | 4 | 3 | 4 | 2 |
| 410 | HC8521 | ELK1;ETS1;FOXL1;GATA2;GATA3;PAX2;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZNF354C | 12 | 2 | 3 | 1 | 4 | 6 | 4 | 6 | 2 |
| 411 | HC5600 | ELF5;ETS1;FOXC1;GATA2;GATA3;MYB;NKX2-5;SOX10;SP1;TFAP2A;ZEB1;ZNF354C | 12 | 2 | 4 | 1 | 3 | 5 | 5 | 5 | 4 |
| 412 | HC4094 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;SOX10;TFAP2A;USF1;YY1;ZNF354C | 12 | 2 | 3 | 2 | 2 | 3 | 3 | 3 | 2 |
| 413 | HC5628 | ETS1;FOXC1;GATA2;GATA3;MZF1;NFATC2;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZNF354C | 12 | 2 | 5 | 1 | 4 | 4 | 5 | 5 | 2 |
| 414 | HC8430 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;PAX2;PDX1;PRRX2;TFAP2A;YY1;ZNF354C | 12 | 2 | 4 | 1 | 3 | 5 | 3 | 5 | 1 |
| 415 | HC5718 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAX;MZF1;PRRX2;TFAP2A;YY1 | 12 | 2 | 3 | 2 | 2 | 3 | 2 | 4 | 2 |
| 416 | HC1738 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;NKX2-5;SOX10;SPIB;TFAP2A;ZEB1 | 12 | 1 | 4 | 1 | 4 | 5 | 5 | 6 | 3 |
| 417 | HC8710 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;SOX5;YY1;ZNF354C | 12 | 2 | 3 | 3 | 4 | 4 | 4 | 4 | 3 |
| 418 | HC5444 | BRCA1;ELF5;ETS1;FOXC1;GATA2;GATA3;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 12 | 3 | 3 | 2 | 3 | 4 | 5 | 5 | 3 |
| 419 | HC2710 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;MZF1;PAX2;PDX1;PRRX2;SOX10;SPIB;ZNF354C | 12 | 2 | 3 | 2 | 4 | 5 | 4 | 6 | 3 |
| 420 | HC8630 | ETS1;FOXC1;FOXL1;FOXL1;GATA2;GATA3;HOXA5;MZF1;SOX10;SOX17;ZEB1;ZNF354C | 12 | 2 | 4 | 2 | 2 | 5 | 4 | 5 | 3 |
| 421 | HC9738 | BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;NKX2-5;SOX5;SP1;TFAP2A;YY1;ZEB1 | 12 | 1 | 4 | 2 | 3 | 4 | 4 | 3 | 2 |
| 422 | HC4742 | EBF1;ELK1;ETS1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2;SPIB;TFAP2A;ZEB1;ZNF354C | 13 | 2 | 3 | 1 | 4 | 4 | 4 | 6 | 2 |
| 423 | HC4660 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;HOXA5;MZF1;SOX10;SOX17;SPIB;TFAP2A;ZEB1;ZNF354C | 13 | 2 | 4 | 3 | 3 | 5 | 5 | 6 | 3 |
| 424 | HC4522 | ELK1;ETS1;FOXC1;FOXL1;GATA2;MAFB;MAX;NR4A2;PAX2;SOX10;TFAP2A;YY1;ZEB1 | 13 | 1 | 3 | 1 | 2 | 5 | 3 | 4 | 1 |
| 425 | HC7954 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAX;MZF1;PDX1;PRRX2;TFAP2A;YY1 | 13 | 1 | 4 | 1 | 3 | 4 | 3 | 5 | 1 |
| 426 | HC6873 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 427 | HC3799 | BRCA1;E2F1;ETS1;FOXC1;GATA2;GATA3;NFATC2;PAX2;SOX10;SPIB;TFAP2A;YY1;ZEB1 | 13 | 1 | 4 | 1 | 3 | 4 | 5 | 4 | 2 |
| 428 | HC3753 | ETS1;FOXL1;GATA2;GATA3;MZF1;NR4A2;PAX2;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 2 | 2 | 3 | 5 | 3 | 5 | 2 |
| 429 | HC639 | ARID3A;CREB1;ETS1;FOXO3;GATA2;GATA3;MYB;MZF1;NR4A2;SOX10;SP1;TFAP2A;YY1 | 13 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 2 |
| 430 | HC7879 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;MYB;MZF1;PDX1;PRRX2;YY1;ZNF354C | 13 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 431 | HC3666 | EP300;ETS1;FOXC1;GATA2;HOXA5;MAFB;MZF1;PDX1;SOX10; TFAP2A;USF1;ZEB1;ZNF354C | 13 | 2 | 4 | 1 | 3 | 4 | 4 | 4 | 2 |
| 432 | HC747 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;NR4A2;PARP1; PRRX2;SOX10;SP1;TFAP2A;YY1 | 13 | 2 | 3 | 1 | 3 | 5 | 4 | 6 | 2 |
| 433 | HC9228 | ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;NKX2-5;PRRX2; SOX10;SP1;TFAP2A;YY1;ZNF354C | 13 | 2 | 4 | 1 | 4 | 4 | 4 | 5 | 3 |
| 434 | HC6264 | ARID3A;CNOT3;ETS1;FOXC1;GATA2;MAFB;MZF1;PAX2; PDX1;PRRX2;TFAP2A;YY1;ZEB1 | 13 | 1 | 4 | 1 | 3 | 4 | 3 | 4 | 1 |
| 435 | HC3582 | BRCA1;CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1; NFYA;PDX1;SOX10;TFAP2A;ZEB1 | 13 | 1 | 4 | 1 | 3 | 5 | 4 | 5 | 2 |
| 436 | HC7643 | ARID3A;ETS1;FOXC1;FOXL1;FOXP3;GATA2;GATA3;MEF2C; MZF1;PRRX2;SOX10;SP1;YY1 | 13 | 1 | 2 | 1 | 2 | 3 | 2 | 4 | 1 |
| 437 | HC6590 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;KLF4; PRRX2;SOX5;TFAP2A;YY1;ZNF354C | 13 | 3 | 3 | 3 | 3 | 5 | 3 | 4 | 3 |
| 438 | HC3847 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MYB;MZF1; PAX2;PDX1;PRRX2;YY1 | 13 | 1 | 3 | 1 | 2 | 4 | 2 | 4 | 2 |
| 439 | HC9012 | BCL6;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;SOX10; SPIB;TFAP2A;YY1;ZEB1 | 13 | 1 | 4 | 1 | 4 | 5 | 5 | 6 | 3 |
| 440 | HC5110 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NKX3-2;REL; SOX10;SPIB;TFAP2A;YY1;ZNF354C | 13 | 2 | 3 | 1 | 3 | 4 | 4 | 5 | 2 |
| 441 | HC5117 | ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;NFATC2;PDX1; PRRX2;SOX10;TFAP2A;YY1;ZNF354C | 13 | 2 | 5 | 1 | 4 | 4 | 5 | 5 | 2 |
| 442 | HC9579 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MAX;MZF1;PDX1;SOX10; TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 5 | 1 | 4 | 5 | 5 | 5 | 3 |
| 443 | HC7208 | ETS1;FOXC1;FOXL1;GATA2;GATA3;NFYA;PDX1;PRRX2; SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 4 | 1 | 4 | 5 | 4 | 6 | 2 |
| 444 | HC6196 | EP300;ETS1;FOXC1;FOXL1;GATA2;ING4;MAFB;NKX2-5; PDX1;PRRX2;TFAP2A;ZEB1;ZNF354C | 13 | 2 | 5 | 2 | 4 | 5 | 4 | 6 | 2 |
| 445 | HC6157 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;SPIB; TFAP2A;TP53;TP73;ZEB1;ZNF354C | 13 | 2 | 3 | 2 | 2 | 3 | 3 | 4 | 2 |
| 446 | HC5000 | ETS1;FOXC1;FOXL1;GATA2;KLF4;NR4A2;PDX1;REL;SPIB; TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 4 | 2 | 3 | 5 | 4 | 5 | 2 |
| 447 | HC5031 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1; NKX2-5;PAX2;TFAP2A;YY1;ZNF354C | 13 | 2 | 4 | 2 | 2 | 5 | 3 | 4 | 2 |
| 448 | HC9317 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MAFB;PRRX2; SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 2 |
| 449 | HC6033 | ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;NKX2-5;PAX2;SP1; SPIB;TFAP2A;USF1;YY1 | 13 | 1 | 4 | 1 | 3 | 6 | 4 | 5 | 2 |
| 450 | HC6059 | ELK1;ETS1;FOXC1;GATA2;GATA3;NKX3-1;PAX2;PRRX2; SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 1 | 3 | 5 | 3 | 5 | 1 |
| 451 | HC6067 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NFYA;SP1;SPIB; TFAP2A;USF1;YY1;ZNF354C | 13 | 2 | 3 | 1 | 2 | 3 | 3 | 4 | 1 |
| 452 | HC5576 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MAFB;MYB; SOX10;TFAP2A;YY1;ZNF354C | 13 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 453 | HC1525 | ARNT;ETS1;FOXC1;GATA2;MAFB;MZF1;PAX2;PDX1;SOX10; TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 4 | 1 | 3 | 5 | 4 | 4 | 2 |
| 454 | HC5524 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1; SMAD1;SOX10;TFAP2A;YY1;ZEB1 | 13 | 1 | 3 | 1 | 2 | 4 | 3 | 4 | 1 |
| 455 | HC7058 | BRCA1;ELK1;ETS1;FOXL1;GATA2;GATA3;MYB;MZF1;PAX2; PDX1;PRRX2;SP1;YY1 | 13 | 1 | 2 | 1 | 2 | 4 | 2 | 4 | 2 |

| | | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|---|
| 456 | HC5861 | ARID3A;BRCA1;ELK4;ETS1;FOXC1;GATA2;MAFB;MAX;MZF1;PDX1;YY1;ZEB1;ZNF354C | 13 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 457 | HC8410 | BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;SOX10;SP1;TFAP2A;YY1;ZNF354C | 13 | 2 | 4 | 2 | 3 | 4 | 4 | 4 | 4 | 2 |
| 458 | HC5997 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;PDX1;PRRX2;SOX10;SPIB;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 1 | 4 | 4 | 4 | 6 | 3 | |
| 459 | HC4279 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;SOX10;SP1;TFAP2A;USF1;YY1;ZEB1 | 13 | 1 | 4 | 1 | 3 | 5 | 4 | 5 | 2 | |
| 460 | HC5324 | BRCA1;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NFYA;NKX2-5;SOX10;TFAP2A;ZNF354C | 13 | 3 | 4 | 2 | 3 | 7 | 5 | 6 | 3 | |
| 461 | HC4270 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PAX2;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 1 | 3 | 5 | 3 | 5 | 1 | |
| 462 | HC5304 | BRCA1;ETS1;FOXC1;FOX11;FOX11;FOXO3;GATA2;GATA3;PDX1;SOX10;YY1;ZEB1;ZNF354C | 13 | 2 | 4 | 2 | 3 | 5 | 4 | 5 | 3 | |
| 463 | HC5342 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MYB;PRRX2;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 2 | 3 | 3 | 3 | 4 | 2 | |
| 464 | HC8735 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;SOX10;SP1;TFAP2A;USF1;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 1 | 2 | 4 | 3 | 4 | 1 | |
| 465 | HC8718 | ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;NFATC2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 5 | 1 | 3 | 4 | 5 | 4 | 2 | |
| 466 | HC9786 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;PRRX2;SOX5;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 3 | 2 | 3 | 3 | 3 | 3 | 2 | |
| 467 | HC5448 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;HOXA5;MAFB;MZF1;NFATC2;PDX1;TFAP2A;ZEB1;ZNF354C | 13 | 2 | 5 | 2 | 2 | 4 | 4 | 4 | 2 | |
| 468 | HC2765 | BRCA1;ETS1;FOXC1;FOXL1;GATA1;GATA2;GATA3;MZF1;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 13 | 2 | 4 | 2 | 2 | 4 | 3 | 3 | 2 | |
| 469 | HC4801 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 4 | 3 | 4 | 6 | 4 | 6 | 3 | |
| 470 | HC4868 | BRCA1;ETS1;FOXC1;GATA2;JUNB;MAFB;MZF1;PAX2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 2 | 3 | 5 | 4 | 4 | 2 | |
| 471 | HC8982 | ARID3A;CDC5L;ETS1;FOXC1;GATA2;GATA3;HOXA5;NKX2-5;NR4A2;PDX1;SOX10;YY1;ZEB1;ZNF354C | 14 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | |
| 472 | HC4780 | ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MAFB;PDX1;PRRX2;SOX10;SPIB;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 1 | 4 | 4 | 4 | 6 | 3 | |
| 473 | HC4700 | BRCA1;E2F1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NFATC2;PAX2;SOX10;SPIB;TFAP2A;YY1;ZEB1 | 14 | 1 | 4 | 1 | 3 | 5 | 5 | 5 | 2 | |
| 474 | HC4707 | ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX3-2;PDX1;PRRX2;REL;SOX10;SPIB;TFAP2A;YY1;ZNF354C | 14 | 2 | 4 | 1 | 5 | 5 | 5 | 7 | 3 | |
| 475 | HC8806 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;SP1;SPIB;TFAP2A;YY1;ZNF354C | 14 | 2 | 4 | 1 | 5 | 5 | 5 | 7 | 3 | |
| 476 | HC8849 | ARID3A;CREB1;ETS1;FOX11;FOX11;GATA2;GATA3;HOXA5;KLF4;MZF1;SOX10;SOX17;TFAP2A;ZEB1 | 14 | 1 | 4 | 2 | 3 | 7 | 5 | 6 | 3 | |
| 477 | HC8891 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NR4A2;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 4 | 4 | 5 | 3 | |
| 478 | HC8898 | BPTF;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;PAX2;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 1 | 1 | 4 | 2 | 3 | 1 | |
| 479 | HC3136 | ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MZF1;NKX2-5;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 1 | 4 | 5 | 5 | 6 | 3 | |
| 480 | HC6690 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;MYB;PRRX2;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 | |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 481 | HC4684 | ELK1;ETS1;FOXC1;FOXL1;GATA2;MAFB;MAX;NR4A2;PAX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 482 | HC2530 | BRCA1;EBF1;ETS1;FOXC1;GATA1;GATA2;GATA3;MAFB;MZF1;SOX10;SP1;TFAP2A;YY1;ZEB1 | 14 | 1 | 4 | 1 | 3 | 4 | 4 | 3 | 2 |
| 483 | HC4512 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5;PAX2;PRRX2;SOX10;SPIB;TFAP2A;ZEB1 | 14 | 1 | 4 | 1 | 5 | 7 | 5 | 7 | 3 |
| 484 | HC4523 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MAFB;PDX1;SOX10;TBP;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 2 | 4 | 3 | 4 | 2 |
| 485 | HC7938 | ELK1;ETS1;FOXC1;FOXL1;GATA2;MAFB;MAX;MZF1;NR4A2;PAX2;SOX10;TFAP2A;YY1;ZEB1 | 14 | 1 | 3 | 1 | 2 | 5 | 3 | 4 | 1 |
| 486 | HC7944 | ETS1;FOXC1;FOXL1;GATA2;GATA3;NFYA;NKX2-5;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 5 | 1 | 5 | 6 | 5 | 7 | 3 |
| 487 | HC7931 | ARID3A;CDC5L;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MAFB;MZF1;PDX1;TFAP2A;YY1;ZNF354C | 14 | 2 | 4 | 3 | 3 | 5 | 3 | 4 | 3 |
| 488 | HC7919 | ARID3A;ELK1;ETS1;FOXC1;GATA2;GATA3;MYB;PRRX2;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 3 | 3 | 3 | 4 | 2 |
| 489 | HC7917 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NR4A2;PDX1;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 4 | 3 | 4 | 5 | 5 | 6 | 3 |
| 490 | HC4458 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NR4A2;PDX1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 4 | 4 | 5 | 3 |
| 491 | HC3780 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 2 | 4 | 5 | 4 | 6 | 2 |
| 492 | HC3719 | ELK1;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;PAX2;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 1 | 3 | 4 | 3 | 4 | 2 |
| 493 | HC3713 | ARID3A;CDC5L;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;PDX1;SOX10;TFAP2A;YY1;ZNF354C | 14 | 2 | 4 | 3 | 3 | 5 | 4 | 5 | 3 |
| 494 | HC9133 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;PRRX2;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 3 | 3 | 3 | 4 | 2 |
| 495 | HC9168 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;SPIB;TFAP2A;USF1;YY1;ZNF354C | 14 | 2 | 4 | 1 | 5 | 5 | 5 | 7 | 3 |
| 496 | HC6380 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;SOX5;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 3 | 5 | 5 | 5 | 5 | 3 |
| 497 | HC3658 | ARID3A;CREB1;ETS1;FOXC1;FOX11;FOXL1;GATA2;GATA3;HOXA5;MZF1;SOX10;SOX17;TFAP2A;ZEB1 | 14 | 1 | 5 | 2 | 3 | 6 | 5 | 6 | 3 |
| 498 | HC707 | ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;PAX2;PDX1;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 1 | 3 | 5 | 4 | 4 | 2 |
| 499 | HC9211 | BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;PAX2;PRRX2;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 |
| 500 | HC7740 | ELK1;ETS1;FOXC1;FOX11;FOXL1;GATA2;GATA3;HOXA5;MZF1;SOX10;SOX17;SP1;TFAP2A;ZEB1 | 14 | 1 | 5 | 1 | 3 | 6 | 5 | 6 | 3 |
| 501 | HC7713 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX2-5;PDX1;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 5 | 2 | 5 | 6 | 6 | 7 | 4 |
| 502 | HC3978 | BRCA1;ESRRB;ETS1;FOXL1;GATA2;GATA3;MAFB;MZF1;PAX2;PDX1;SPIB;TFAP2A;ZEB1;ZNF354C | 14 | 2 | 4 | 2 | 4 | 6 | 5 | 5 | 3 |
| 503 | HC3972 | ETS1;FOXC1;FOX11;FOXL1;GATA2;GATA3;HOXA5;MZF1;SOX10;SOX17;SP1;TFAP2A;YY1;ZEB1 | 14 | 1 | 5 | 1 | 3 | 6 | 5 | 6 | 3 |
| 504 | HC3944 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;PDX1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 4 | 3 | 3 | 4 | 3 | 4 | 3 |
| 505 | HC6597 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MYB;PAX2;PRRX2;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 2 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 506 | HC3888 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;PDX1;PRRX2;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 4 | 3 | 3 | 4 | 3 | 4 | 3 |
| 507 | HC7548 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NR4A2;PDX1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 14 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 |
| 508 | HC9016 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MYB;MZF1;PAX2;PDX1;PRRX2;YY1;ZEB1 | 14 | 1 | 3 | 1 | 2 | 4 | 2 | 4 | 2 |
| 509 | HC7597 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;SOX10;SOX5;SPIB;USF1;ZEB1;ZNF354C | 14 | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 4 |
| 510 | HC9084 | ARID3A;BRCA1;ESRRB;ETS1;GATA2;GATA3;MAFB;PAX2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 4 | 3 | 4 | 5 | 5 | 4 | 3 |
| 511 | HC6446 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;NFYA;NKX2-5;PDX1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 1 | 4 | 5 | 5 | 6 | 4 |
| 512 | HC7467 | ARID3A;CEBPA;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5;PDX1;SOX10;SP1;TFAP2A;YY1 | 14 | 1 | 5 | 1 | 4 | 7 | 5 | 6 | 3 |
| 513 | HC7402 | ARID3A;ETS1;FOXL1;GATA2;GATA3;MAFB;MZF1;SOX10;SP1;TFAP2A;USF1;YY1;ZEB1;ZNF354C | 14 | 2 | 2 | 2 | 2 | 4 | 3 | 4 | 2 |
| 514 | HC9618 | BRCA1;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PAX2;PDX1;SPIB;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 2 | 2 | 5 | 4 | 5 | 3 |
| 515 | HC367 | ARID3A;ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;NKX3-2;PAX2;SOX10;TFAP2A;TRIM28;YY1;ZNF354C | 14 | 2 | 3 | 2 | 3 | 4 | 3 | 3 | 2 |
| 516 | HC5254 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MZF1;SOX10;SOX17;SP1;TFAP2A;YY1;ZEB1 | 14 | 1 | 4 | 2 | 2 | 5 | 4 | 5 | 2 |
| 517 | HC7368 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5;NR4A2;SOX10;SP1;USF1;YY1;ZNF354C | 14 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 518 | HC7381 | ARID3A;BRCA1;ETS1;GATA2;GATA3;MAFB;MZF1;NKX2-5;PDX1;SP1;SPIB;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 3 |
| 519 | HC6117 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;MAFB;PDX1;PRRX2;SIRT6;SP1;SPIB;YY1;ZEB1 | 14 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| 520 | HC9340 | BRCA1;ETS1;FOXC1;GATA2;KLF4;MZF1;NKX2-5;REL;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 2 | 3 | 5 | 4 | 4 | 2 |
| 521 | HC7166 | ETS1;FOXC1;FOXL1;GATA2;MAFB;NFYA;NKX2-5;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 5 | 1 | 5 | 6 | 5 | 7 | 3 |
| 522 | HC5558 | BRCA1;ETS1;FOXC1;GATA2;MZF1;NKX2-5;PAX2;PDX1;SOX10;SP1;TFAP2A;USF1;YY1;ZNF354C | 14 | 2 | 5 | 2 | 4 | 6 | 5 | 5 | 3 |
| 523 | HC5768 | BRCA1;ETS1;FOXC1;GATA2;GATA3;HOXA5;MAFB;MYB;PRRX2;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 |
| 524 | HC8585 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PRRX2;SP1;SPIB;TFAP2A;USF1;YY1;ZEB1;ZNF354C | 14 | 2 | 3 | 1 | 3 | 3 | 3 | 5 | 1 |
| 525 | HC5586 | ARID3A;BRCA1;ELK1;ETS1;GATA2;GATA3;MYB;MZF1;PAX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 |
| 526 | HC8534 | ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1;NKX2-5;NR4A2;SOX10;SP1;USF1;YY1;ZNF354C | 14 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 527 | HC5814 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;NKX2-5;NR4A2;PRRX2;SOX10;SP1;USF1;YY1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 5 | 3 | 5 | 3 |
| 528 | HC5855 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;NFATC2;NKX2-5;PRRX2;SOX10;SPIB;TFAP2A;YY1 | 14 | 1 | 5 | 1 | 5 | 4 | 6 | 6 | 3 |
| 529 | HC4151 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;PAX2;PDX1;PRRX2;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| 530 | HC5885 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;PAX2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 4 | 2 | 3 | 5 | 4 | 4 | 2 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 531 | HC5906 | BRCA1;ESRRB;ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX2-5; PAX2;PRRX2;TFAP2A;YY1;ZEB1;ZNF354C | 14 | 2 | 5 | 2 | 4 | 6 | 4 | 5 | 2 |
| 532 | HC5313 | EP300;ETS1;FOXC1;GATA2;HOXA5;KLF4;MAFB;MZF1; NFYA;PDX1;TFAP2A;USF1;ZEB1;ZNF354C | 14 | 2 | 4 | 1 | 2 | 4 | 3 | 3 | 1 |
| 533 | HC5312 | ARID3A;BRCA1;CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3; HOXA5;MZF1;SOX10;SOX17;TFAP2A;ZEB1 | 14 | 2 | 4 | 3 | 3 | 5 | 4 | 5 | 3 |
| 534 | HC4210 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;NR4A2; PRRX2;SOX10;SPIB;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 4 | 3 | 5 | 3 |
| 535 | HC5390 | ARID3A;BRCA1;ESRRB;ETS1;GATA2;GATA3;MAFB;MZF1; PAX2;PDX1;SPIB;TFAP2A;YY1;ZNF354C | 14 | 3 | 4 | 3 | 4 | 5 | 5 | 4 | 3 |
| 536 | HC9803 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MZF1;PDX1; PRRX2;SOX10;SP1;TFAP2A;YY1;ZNF354C | 14 | 2 | 4 | 2 | 4 | 5 | 4 | 5 | 2 |
| 537 | HC9829 | ARNT;EP300;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1; NKX2-5;PRRX2;SOX10;SPI1;YY1;ZEB1 | 14 | 1 | 3 | 1 | 3 | 4 | 3 | 6 | 2 |
| 538 | HC8703 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5; NR4A2;SOX10;SP1;USF1;YY1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 5 | 3 | 4 | 3 |
| 539 | HC5466 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;PDX1; SOX10;SPIB;TBP;TFAP2A;YY1;ZEB1 | 14 | 1 | 4 | 1 | 4 | 7 | 6 | 7 | 4 |
| 540 | HC4313 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NR4A2; SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1 | 14 | 2 | 3 | 2 | 3 | 4 | 4 | 5 | 2 |
| 541 | HC8656 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1; NKX2-5;SP1;SPIB;YY1;ZEB1;ZNF354C | 14 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 542 | HC9740 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MAX; NR4A2;PAX2;SOX10;TFAP2A;YY1;ZEB1 | 14 | 1 | 3 | 1 | 2 | 5 | 3 | 4 | 1 |
| 543 | HC8906 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;NKX2-5;PDX1; PRRX2;SP1;SPIB;USF1;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 3 | 4 | 5 | 3 |
| 544 | HC8941 | ARID3A;BRCA1;ESRRB;ETS1;GATA2;GATA3;KLF4;MAFB; MZF1;PAX2;PDX1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 3 | 6 | 4 | 3 | 3 |
| 545 | HC8956 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;NKX3-1;PAX2;PDX1; SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 4 | 2 | 4 | 7 | 5 | 7 | 3 |
| 546 | HC4783 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MZF1; PRRX2;REL;SOX10;SP1;TFAP2A;YY1;ZNF354C | 15 | 2 | 3 | 1 | 3 | 4 | 3 | 5 | 1 |
| 547 | HC4717 | BRCA1;CREB1;ELK1;ETS1;FOXC1;GATA2;GATA3;MAFB; MZF1;PDX1;SOX10;TFAP2A;USF1;YY1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 4 | 4 | 4 | 2 |
| 548 | HC4752 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5; NR4A2;SOX10;SP1;USF1;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 549 | HC4756 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1; SOX10;SOX17;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 4 | 4 | 5 | 5 | 6 | 4 |
| 550 | HC8877 | ARID3A;BRCA1;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3; MYB;MZF1;PARP1;SOX10;SP1;YY1;ZNF354C | 15 | 4 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| 551 | HC6665 | BRCA1;ELF5;ELK1;ETS1;FOXC1;GATA2;GATA3;HOXA2; MZF1;PDX1;SOX10;SOX5;TFAP2A;YY1;ZNF354C | 15 | 3 | 4 | 3 | 4 | 7 | 6 | 6 | 4 |
| 552 | HC6637 | BRCA1;ETS1;FOXC1;GATA2;KLF4;MZF1;NKX2-5;PAX2; REL;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1 | 15 | 1 | 4 | 1 | 4 | 6 | 5 | 5 | 3 |
| 553 | HC4645 | ARID3A;BRCA1;ESRRB;ETS1;GATA2;GATA3;MAFB;MZF1; PAX2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 4 | 5 | 5 | 4 | 3 |
| 554 | HC4605 | ARID3A;BRCA1;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3; MZF1;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 4 | 4 | 3 | 3 | 5 | 5 | 6 | 3 |
| 555 | HC4623 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;GTF2A1;HSF1; KLF4;PDX1;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 2 | 6 | 4 | 5 | 2 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 556 | HC4585 | ARID3A;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4; PDX1;POU3F2;SOX10;SP1;TBP;TFAP2A;YY1 | 15 | 2 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 557 | HC6961 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3; MYB;MZF1;PAX2;PDX1;PRRX2;SP1;ZEB1 | 15 | 2 | 3 | 2 | 2 | 4 | 2 | 4 | 2 |
| 558 | HC6972 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MZF1; PDX1;PRRX2;REL;SOX10;SP1;TFAP2A;ZNF354C | 15 | 2 | 4 | 1 | 4 | 5 | 4 | 6 | 2 |
| 559 | HC6929 | BRCA1;ELK1;ETS1;FOXC1;GATA2;MAFB;NKX2-5;PDX1; SOX10;SOX17;SOX5;SP1;TFAP2A;ZEB1;ZNF354C | 15 | 2 | 6 | 4 | 5 | 7 | 7 | 6 | 5 |
| 560 | HC6936 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MAFB;MZF1; NFATC2;NKX2-5;PRRX2;SP1;SPIB;TFAP2A;YY1 | 15 | 1 | 5 | 1 | 4 | 4 | 5 | 5 | 2 |
| 561 | HC6935 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;PAX6; PDX1;SOX10;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 5 | 1 | 3 | 8 | 6 | 6 | 4 |
| 562 | HC4521 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX3-1; SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 3 | 2 | 3 | 5 | 4 | 6 | 2 |
| 563 | HC7918 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;NR4A2; SOX10;SOX17;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 564 | HC7961 | ARID3A;BRCA1;BRF1;ETS1;FOXC1;GATA2;KLF4;MZF1; NR4A2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 565 | HC4463 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;HSF1;KLF4; PDX1;SOX10;TBP;TFAP2A;YY1;ZEB1;ZFX | 15 | 1 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 566 | HC4475 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5; NR4A2;SOX10;SP1;TFAP2A;USF1;YY1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 6 | 4 | 5 | 2 |
| 567 | HC4406 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3; MZF1;NKX2-5;NR4A2;PDX1;SOX10;SPIB;TFAP2A;ZEB1 | 15 | 3 | 5 | 3 | 5 | 6 | 6 | 7 | 4 |
| 568 | HC6390 | ARID3A;BRCA1;CREB1;ETS1;FOXC1;FOXD3;GATA2;GATA3; MAFB;NKX2-5;NR4A2;SOX10;SPIB;ZEB1;ZNF354C | 15 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 4 |
| 569 | HC3723 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;MZF1;NR4A2; PRRX2;SOX10;SOX17;SPIB;TFAP2A;YY1;ZEB1 | 15 | 4 | 4 | 4 | 4 | 5 | 5 | 6 | 4 |
| 570 | HC3721 | ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;PAX2;PDX1;REL; SOX10;SP1;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 4 | 7 | 5 | 6 | 3 |
| 571 | HC3722 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;ING4;MAFB;MYB; PRRX2;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| 572 | HC7887 | ARID3A;E2F1;ETS1;FOXC1;FOX11;FOXL1;GATA2;GATA3; PDX1;PRRX2;SOX10;SPIB;YY1;ZNF148;ZNF354C | 15 | 2 | 4 | 2 | 5 | 5 | 5 | 7 | 4 |
| 573 | HC9131 | ARID3A;CDC5L;ETS1;FOXC1;FOXO3;GATA2;GATA3;MYB; MZF1;NKX2-5;SOX5;USF1;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 3 | 3 | 3 | 3 | 2 | 3 |
| 574 | HC9132 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MAX; NR4A2;PAX2;PDX1;REL;SOX10;YY1;ZEB1 | 15 | 1 | 3 | 1 | 2 | 5 | 3 | 4 | 2 |
| 575 | HC9135 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5; PDX1;SOX10;SPIB;TBP;TFAP2A;YY1;ZEB1 | 15 | 1 | 5 | 1 | 5 | 8 | 7 | 8 | 5 |
| 576 | HC9137 | BRCA1;CUX1;ETS1;FOXC1;GATA2;GATA3;KLF4;MZF1; NR4A2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 3 | 4 | 4 | 4 | 3 |
| 577 | HC9105 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1; NFATC2;NKX2-5;PRRX2;SOX10;SPIB;TFAP2A;YY1 | 15 | 1 | 5 | 1 | 5 | 5 | 6 | 7 | 3 |
| 578 | HC6304 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1; NFATC2;NKX2-5;PRRX2;SPIB;TFAP2A;YY1;ZEB1 | 15 | 1 | 5 | 1 | 4 | 4 | 5 | 6 | 2 |
| 579 | HC7843 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;KLF4;NFYA;NKX2-5; NR4A2;PDX1;SOX10;SP1;TFAP2A;YY1;ZNF354C | 15 | 3 | 5 | 3 | 4 | 7 | 5 | 6 | 3 |
| 580 | HC7865 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;NKX3-1;PDX1; SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 4 | 2 | 4 | 6 | 5 | 7 | 3 |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 581 | HC9152 | ARID3A;CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX2-5;NR4A2;PDX1;SOX10;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 5 | 3 | 5 | 6 | 6 | 7 | 4 |
| 582 | HC7871 | ELF5;ETS1;FOXC1;FOXL1;FOXO3;GATA2;GATA3;GTF2A1;KLF4;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 583 | HC7823 | CEBPA;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1;NKX2-5;PDX1;SOX10;SP1;SPIB;TFAP2A;YY1 | 15 | 1 | 5 | 1 | 5 | 7 | 6 | 7 | 4 |
| 584 | HC3693 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;PDX1;PRRX2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 5 | 5 | 5 | 7 | 3 |
| 585 | HC6292 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;SOX10;SOX5;SPIB;USF1;ZEB1;ZNF354C | 15 | 2 | 2 | 3 | 3 | 4 | 4 | 4 | 3 |
| 586 | HC9244 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NKX2-5;PDX1;PRRX2;SP1;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 5 | 5 | 5 | 7 | 3 |
| 587 | HC7758 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1;SOX5;TFAP2A;USF1;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 3 |
| 588 | HC7751 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;NKX2-5;PDX1;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 5 | 2 | 5 | 6 | 6 | 7 | 4 |
| 589 | HC9284 | BRCA1;ETS1;FOXC1;FOXL1;FOXM1;GATA2;GATA3;MZF1;PRRX2;SOX10;SP1;TFAP2A;USF1;YY1;ZEB1 | 15 | 1 | 3 | 1 | 3 | 4 | 3 | 5 | 1 |
| 590 | HC7737 | CEBPA;CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;PDX1;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 3 | 6 | 4 | 5 | 2 |
| 591 | HC9293 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;PRRX2;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 4 | 5 | 4 | 6 | 3 |
| 592 | HC9292 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX2-5;PAX2;PDX1;SOX10;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 5 | 3 | 5 | 7 | 6 | 7 | 4 |
| 593 | HC7705 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NR4A2;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 4 | 4 | 5 | 3 |
| 594 | HC3987 | BRCA1;ELK1;ETS1;FOXC1;FOXD1;FOXL1;GATA2;MZF1;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 3 | 2 | 3 | 4 | 4 | 5 | 2 |
| 595 | HC845 | ETS1;FOXC1;GATA2;GATA3;IKZF1;IKZF2;MZF1;PAX2;PDX1;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 4 | 6 | 5 | 5 | 3 |
| 596 | HC3939 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NR4A2;PDX1;SOX10;SP1;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 4 | 5 | 5 | 6 | 3 |
| 597 | HC6526 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 5 | 5 | 5 | 7 | 3 |
| 598 | HC6518 | ARID3A;BRCA1;ETS1;FOXL1;GATA2;GATA3;MZF1;PDX1;SOX10;SP1;TFAP2A;USF1;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 5 | 4 | 5 | 3 |
| 599 | HC6533 | ARID3A;ETS1;FOXC1;FOXD3;FOXL1;GATA2;GATA3;MAFB;NR4A2;PDX1;SOX10;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 5 | 3 | 4 | 5 | 6 | 6 | 4 |
| 600 | HC6537 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MYB;MZF1;PAX2;PDX1;PRRX2;TFAP2A;YY1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 5 | 3 | 5 | 2 |
| 601 | HC7682 | ELK1;ETS1;FOXC1;GATA2;MYB;MZF1;NKX2-5;PAX2;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 4 | 5 | 5 | 5 | 4 |
| 602 | HC7698 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;PDX1;POU3F2;SOX10;SP1;TBP;TFAP2A;YY1;ZEB1 | 15 | 1 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 603 | HC7614 | ARID3A;CEBPA;ETS1;FOXC1;FOXL1;GATA2;MZF1;NKX2-5;PDX1;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 6 | 5 | 6 | 3 |
| 604 | HC6575 | ARID3A;BRCA1;ELF5;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;NKX2-5;PARP1;SOX10;SP1;YY1;ZNF354C | 15 | 4 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| 605 | HC7627 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;JUN;KLF4;MZF1;PDX1;SOX10;TBP;TFAP2A;YY1;ZEB1 | 15 | 1 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 606 | HC3849 | ARID5B;ETS1;FOXC1;FOXL1;GATA2;HOXA5;MAFB;PAX2;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 4 | 6 | 4 | 6 | 2 |
| 607 | HC3846 | ARID3A;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;GTF2A1;KLF4;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 2 | 3 | 7 | 5 | 6 | 3 |
| 608 | HC3866 | E2F1;ETS1;FOXC1;FOX11;GATA2;GATA3;MZF1;NFATC2;PAX2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 6 | 1 | 4 | 5 | 6 | 5 | 3 |
| 609 | HC9002 | BRCA1;ESRRB;ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;PAX2;PDX1;SP1;SPIB;TFAP2A;YY1;ZEB1 | 15 | 1 | 5 | 1 | 4 | 5 | 5 | 4 | 3 |
| 610 | HC9080 | BRCA1;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PAX2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 2 | 3 | 6 | 5 | 6 | 3 |
| 611 | HC9095 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MAFB;MYB;NFE2L2;NKX2-5;NR2E3;SOX10;SPIB;TFAP2A;YY1;ZEB1 | 15 | 1 | 5 | 2 | 5 | 4 | 5 | 6 | 5 |
| 612 | HC6454 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NKX2-5;PAX2;PDX1;SP1;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 6 | 5 | 6 | 3 |
| 613 | HC7512 | CREB1;ETS1;FOXC1;FOX11;FOXO3;GATA2;GATA3;MAFB;NKX2-5;NR4A2;SOX10;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 4 | 6 | 5 | 4 |
| 614 | HC8125 | ARID3A;ETS1;FOXC1;GATA2;MAFB;MAX;MZF1;PAX2;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 5 | 4 | 5 | 2 |
| 615 | HC8115 | ARID3A;CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NKX2-5;NR4A2;SOX10;SP1;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 3 | 6 | 4 | 5 | 3 |
| 616 | HC8143 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MYB;MZF1;PAX2;PDX1;PRRX2;SP1;YY1;ZEB1 | 15 | 1 | 3 | 1 | 2 | 4 | 2 | 4 | 2 |
| 617 | HC5109 | E2F1;ETS1;FOXC1;FOX11;FOXL1;GATA2;GATA3;MZF1;NFATC2;PAX2;PDX1;SPIB;YY1;ZEB1;ZNF354C | 15 | 2 | 5 | 1 | 3 | 5 | 5 | 5 | 3 |
| 618 | HC8187 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;REL;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 2 | 2 | 4 | 3 | 4 | 2 |
| 619 | HC7435 | ARID3A;CREB1;ETS1;FOXC1;FOX11;GATA2;GATA3;MAFB;NKX2-5;NR4A2;PDX1;SOX10;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 6 | 3 | 4 | 5 | 6 | 5 | 4 |
| 620 | HC9648 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MZF1;NKX2-5;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 5 | 5 | 6 | 3 |
| 621 | HC9637 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NKX3-2;PAX2;PDX1;SPIB;TFAP2A;TP53;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 5 | 4 | 5 | 2 |
| 622 | HC9605 | BRCA1;ELK1;ETS1;FOXC1;GATA2;MAFB;NKX2-5;PDX1;SOX10;SOX5;SP1;SPIB;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 3 | 5 | 5 | 6 | 5 | 5 |
| 623 | HC9617 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;PDX1;PRRX2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 5 | 5 | 5 | 7 | 3 |
| 624 | HC7484 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;NKX2-5;SOX10;SOX5;SP1;SPIB;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 4 |
| 625 | HC5191 | BRCA1;ETS1;FOXC1;FOX11;FOXL1;GATA2;GATA3;KLF4;MZF1;PAX2;PDX1;SPIB;USF1;YY1;ZEB1 | 15 | 1 | 4 | 1 | 3 | 6 | 4 | 5 | 3 |
| 626 | HC5195 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1;PRRX2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 5 | 5 | 5 | 7 | 3 |
| 627 | HC5298 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PRRX2;SOX10;SOX5;SPIB;TFAP2A;USF1;ZEB1;ZNF354C | 15 | 2 | 3 | 3 | 5 | 5 | 5 | 6 | 3 |
| 628 | HC5287 | ARID3A;CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1;NR4A2;SOX10;SP1;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 5 | 3 | 4 | 3 |
| 629 | HC8022 | BRCA1;CREB1;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1;PDX1;SOX10;SRY;TFAP2A;USF1;YY1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 5 | 5 | 5 | 3 |
| 630 | HC8024 | BRCA1;ETS1;FOXC1;GATA2;KLF4;MZF1;NKX2-5;PAX2;REL;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 6 | 4 | 4 | 2 |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 631 | HC8048 | ELK1;ETS1;FOXC1;FOXL1;GATA2;MAFB;MAX;NR4A2; PAX2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 6 | 4 | 5 | 2 |
| 632 | HC9589 | ARID3A;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3; KLF4;PDX1;SOX10;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 2 | 3 | 7 | 5 | 6 | 3 |
| 633 | HC5217 | ARID3A;ETS1;FOXC1;FOXD3;GATA2;GATA3;MAFB;MZF1; NR4A2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 5 | 3 | 3 | 4 | 5 | 4 | 3 |
| 634 | HC9543 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MAFB; NR4A2;PAX2;PRRX2;SOX10;YY1;ZEB1;ZNF354C | 15 | 4 | 4 | 4 | 4 | 5 | 4 | 4 | 4 |
| 635 | HC8397 | ALX4;ELF5;ETS1;FOXA3;FOXC1;FOXL1;GATA2;GATA3; KLF4;MZF1;PDX1;SOX10;TBP;YY1;ZNF354C | 15 | 2 | 4 | 1 | 3 | 7 | 5 | 6 | 4 |
| 636 | HC6130 | ARID3A;CDC5L;ELK1;ETS1;FOXC1;FOXL1;GATA2;MAX; MZF1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 3 | 3 | 4 | 3 | 5 | 3 |
| 637 | HC8366 | BRCA1;ESRRB;ETS1;GATA2;GATA3;KLF4;MAFB;MZF1; PAX2;PDX1;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 6 | 4 | 3 | 2 |
| 638 | HC8348 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1; NKX2-5;PDX1;SOX10;SP1;TFAP2A;YY1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 6 | 5 | 6 | 3 |
| 639 | HC9424 | BRCA1;CDC5L;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3; MZF1;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 3 | 5 | 5 | 6 | 3 |
| 640 | HC7233 | ARNT;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MYB; MZF1;PRRX2;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 2 | 2 | 3 | 2 | 4 | 2 |
| 641 | HC7223 | ELK1;ETS1;FOXC1;FOXL1;GATA2;MAFB;MAX;MZF1; NR4A2;PAX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 642 | HC7270 | BRCA1;ETS1;FOXC1;GATA2;KLF4;MZF1;NKX2-5;PAX2; PDX1;REL;SOX10;SP1;TFAP2A;YY1;ZEB1 | 15 | 1 | 5 | 1 | 4 | 7 | 5 | 5 | 3 |
| 643 | HC7291 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;GTF2A1; KLF4;PDX1;SOX10;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 644 | HC8300 | ARID3A;CREB1;ETS1;FOXC1;FOXD3;GATA2;GATA3; MAFB;NR4A2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 5 | 3 | 3 | 4 | 5 | 4 | 3 |
| 645 | HC6159 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3; MZF1;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 4 | 4 | 5 | 3 |
| 646 | HC8220 | EN1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MYCN; PRRX2;SOX10;SPIB;TFAP2A;USF1;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 5 | 6 | 4 | 7 | 3 |
| 647 | HC5041 | ARID3A;CREB1;ETS1;FOXC1;FOXD3;GATA2;GATA3; MAFB;NKX2-5;NR4A2;SOX10;SPIB;TFAP2A;ZEB1;ZNF354C | 15 | 3 | 5 | 3 | 4 | 4 | 6 | 5 | 4 |
| 648 | HC9335 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1; PAX2;PDX1;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 5 | 4 | 5 | 2 |
| 649 | HC9330 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NFYA;NKX2-5; SOX10;SOX5;SP1;SPIB;TFAP2A;YY1;ZNF354C | 15 | 2 | 4 | 3 | 5 | 6 | 6 | 6 | 4 |
| 650 | HC7119 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4; NR4A2;SOX10;SPIB;TBP;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 4 | 4 | 4 | 4 | 5 | 4 | 5 | 4 |
| 651 | HC7112 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4; MAFB;PDX1;PRRX2;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 4 | 5 | 4 | 5 | 3 |
| 652 | HC7154 | BRCA1;ESRRB;ETS1;GATA2;GATA3;KLF4;MAFB;MZF1; PAX2;PDX1;SOX10;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 7 | 5 | 4 | 3 |
| 653 | HC5076 | CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;PDX1; PRRX2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 5 | 6 | 5 | 7 | 3 |
| 654 | HC6099 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1; NKX2-5;PAX2;REL;SOX10;TFAP2A;YY1;ZEB1 | 15 | 2 | 4 | 2 | 3 | 7 | 4 | 5 | 2 |
| 655 | HC6095 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PDX1; PRRX2;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 1 | 4 | 4 | 4 | 6 | 2 |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 656 | HC5781 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;GATA2;MAFB;MZF1; NKX2-5;PDX1;SOX10;SOX5;SP1;ZEB1;ZNF354C | 15 | 3 | 4 | 4 | 4 | 5 | 5 | 4 | 4 |
| 657 | HC4044 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MYB; MZF1;PAX2;PDX1;PRRX2;SOX10;YY1;ZEB1 | 15 | 1 | 3 | 1 | 3 | 5 | 3 | 5 | 3 |
| 658 | HC5737 | EP300;ETS1;FOXC1;GATA2;GATA3;GATA5;GATA6;KLF4; MZF1;SP1;SREBF1;SREBF2;TFAP2A;YY1;ZNF354C | 15 | 3 | 4 | 2 | 2 | 4 | 3 | 3 | 3 |
| 659 | HC5738 | ARID3A;ETS1;GATA2;GATA3;KLF4;MAFB;MYB;MZF1; PAX2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 2 | 3 | 5 | 4 | 4 | 3 |
| 660 | HC5508 | BRCA1;ETS1;FOXC1;GATA2;KLF4;MZF1;NKX2-5;PAX2; REL;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 6 | 5 | 5 | 3 |
| 661 | HC4063 | ARID3A;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX2-5; PDX1;SP1;SP3;SPIB;TFAP2A;YY1;ZNF354C | 15 | 3 | 5 | 2 | 4 | 6 | 6 | 7 | 4 |
| 662 | HC5583 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB; MZF1;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 3 | 2 | 3 | 4 | 4 | 5 | 2 |
| 663 | HC5846 | ARID3A;BPTF;ETS1;FOXC1;FOXL1;GATA2;GATA3; MAFB;PDX1;SOX10;SOX5;SPIB;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 4 |
| 664 | HC4093 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB; MZF1;NR4A2;PDX1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 4 | 4 | 5 | 3 |
| 665 | HC5845 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;MAFB;NKX2-5; PDX1;SOX10;SOX5;SP1;SPIB;ZEB1;ZNF354C | 15 | 2 | 4 | 3 | 5 | 6 | 6 | 6 | 5 |
| 666 | HC5842 | ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB; MZF1;PDX1;PRRX2;SOX10;SPIB;TFAP2A;YY1;ZNF354C | 15 | 2 | 4 | 1 | 5 | 5 | 5 | 7 | 3 |
| 667 | HC7036 | CREB1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3; KLF4;NKX2-5;NR4A2;SOX10;SP1;USF1;YY1;ZNF354C | 15 | 2 | 3 | 2 | 2 | 5 | 3 | 4 | 2 |
| 668 | HC7030 | ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1;NKX2-5; NR4A2;SOX10;SP1;TFAP2A;USF1;YY1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 6 | 4 | 5 | 2 |
| 669 | HC5668 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2; PRRX2;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 3 | 3 | 3 | 4 | 4 | 4 | 6 | 3 |
| 670 | HC5665 | BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4;MZF1;NKX2-5; PAX2;PDX1;REL;SOX10;TFAP2A;YY1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 7 | 5 | 5 | 3 |
| 671 | HC5691 | ARID3A;BRCA1;CDC5L;ELK1;ETS1;FOXC1;GATA2; GATA3;MAFB;MYB;MZF1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 4 | 4 | 4 | 4 | 3 | 4 |
| 672 | HC8474 | ARID3A;ELF5;ETS1;FOXC1;GATA2;GATA3;KLF4; NFYA;NKX2-5;PDX1;SP1;TBP;TFAP2A;YY1;ZEB1 | 15 | 2 | 5 | 1 | 3 | 6 | 5 | 5 | 3 |
| 673 | HC8409 | ARID3A;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5; MZF1;NFATC2;PDX1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 5 | 2 | 3 | 4 | 5 | 5 | 2 |
| 674 | HC8431 | BRCA1;ETS1;FOXC1;GATA2;KLF4;MZF1;NKX2-5; PDX1;REL;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1 | 15 | 1 | 5 | 1 | 5 | 6 | 6 | 6 | 4 |
| 675 | HC5973 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NR4A2; PDX1;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 3 | 4 | 5 | 5 | 6 | 3 |
| 676 | HC5724 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB; NR4A2;PDX1;PRRX2;SOX10;SPIB;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 4 | 4 | 4 | 6 | 3 |
| 677 | HC5928 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX2-5; SOX10;SOX5;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 3 | 5 | 6 | 6 | 6 | 4 |
| 678 | HC5988 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2; GATA3;MYB;MZF1;PAX2;PDX1;PRRX2;YY1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| 679 | HC5995 | BRCA1;CREB1;ELK1;ETS1;FOXC1;FOXL1;GATA2; GATA3;MZF1;PDX1;SOX10;TFAP2A;USF1;YY1;ZNF354C | 15 | 2 | 4 | 2 | 3 | 5 | 4 | 5 | 2 |
| 680 | HC5363 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;KLF4;MAFB; MZF1;NR4A2;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |

| | | | | | | | | | | | |
|-----|--------|--|----|---|---|---|---|---|---|---|---|
| 681 | HC5362 | BRCA1;CREB1;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1; PDX1;PRRX2;SOX10;TFAP2A;USF1;YY1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 4 | 4 | 5 | 2 |
| 682 | HC5369 | ARID3A;BRCA1;ELF5;ETS1;FOXC1;GATA2;GATA3;HOXB7; MYB;MZF1;PARP1;SOX10;SP1;YY1;ZNF354C | 15 | 4 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| 683 | HC5375 | ARID3A;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;PDX1; PRRX2;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 3 | 4 | 2 | 5 | 6 | 6 | 8 | 4 |
| 684 | HC5358 | ARID3A;ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4; PDX1;SMAD1;SOX10;TBP;TFAP2A;YY1;ZEB1 | 15 | 2 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 685 | HC5357 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NKX3-1; SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 3 | 2 | 3 | 5 | 4 | 6 | 2 |
| 686 | HC4203 | CREB1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;MZF1; PAX2;PDX1;SOX10;SPIB;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 1 | 3 | 5 | 4 | 5 | 3 |
| 687 | HC8794 | BRCA1;ETS1;FOXC1;FOXL1;FOXL1;GATA2;GATA3;NKX2-5; PRRX2;SOX10;SOX5;SP1;SPIB;YY1;ZNF354C | 15 | 2 | 4 | 3 | 6 | 6 | 6 | 7 | 5 |
| 688 | HC8777 | ARID3A;BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4; MZF1;NR4A2;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 689 | HC9813 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;NKX3-1; SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 3 | 3 | 3 | 4 | 5 | 4 | 6 | 3 |
| 690 | HC9855 | ARID3A;CEBPA;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4; NKX2-5;PDX1;SOX10;SP1;TFAP2A;YY1;ZNF354C | 15 | 2 | 5 | 2 | 4 | 7 | 5 | 6 | 3 |
| 691 | HC9844 | BRCA1;ETS1;FOXC1;GATA2;GATA3;MAFB;MZF1;NFATC2; NKX2-5;PDX1;PRRX2;SOX10;SPIB;YY1;ZEB1 | 15 | 1 | 5 | 1 | 5 | 4 | 6 | 6 | 4 |
| 692 | HC9878 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;GATA2;GATA3;MZF1; PDX1;PRRX2;SOX10;SPIB;YY1;ZEB1;ZNF354C | 15 | 3 | 3 | 3 | 4 | 4 | 4 | 5 | 3 |
| 693 | HC4329 | ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;MZF1; PDX1;SOX10;SP2;TBP;TFAP2A;YY1;ZEB1 | 15 | 1 | 4 | 1 | 3 | 7 | 5 | 6 | 3 |
| 694 | HC5468 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;PAX2; PDX1;SOX10;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 6 | 5 | 6 | 3 |
| 695 | HC4339 | ARID3A;BRCA1;ELK1;ETS1;FOXC1;FOXL1;FOXL1; FOXO3;GATA2;GATA3;HOXA5;REL;SP1;YY1;ZNF354C | 15 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 696 | HC4349 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1; NR4A2;PRRX2;SP1;SPIB;TBP;TFAP2A;YY1;ZEB1 | 15 | 2 | 3 | 2 | 3 | 4 | 3 | 5 | 2 |
| 697 | HC5487 | ARID3A;ETS1;FOXC1;GATA2;GATA3;KLF4;MYB;PDX1; PRRX2;SOX10;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 4 | 2 | 4 | 5 | 4 | 5 | 3 |
| 698 | HC8674 | BPTF;ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1; PRRX2;SOX10;USF1;USF2;YY1;ZEB1;ZNF354C | 15 | 2 | 2 | 1 | 2 | 3 | 2 | 4 | 1 |
| 699 | HC8625 | ARID3A;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3; MAX;MZF1;NKX3-2;PRRX2;SOX10;SPIB;YY1;ZEB1 | 15 | 1 | 2 | 1 | 3 | 4 | 3 | 5 | 2 |
| 700 | HC8637 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NKX3-1; SIRT6;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 3 | 3 | 3 | 4 | 5 | 4 | 6 | 3 |
| 701 | HC9709 | BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;NKX2-5; NR4A2;PRRX2;SOX10;SP1;TFAP2A;USF1;YY1;ZNF354C | 15 | 3 | 4 | 3 | 4 | 6 | 4 | 6 | 3 |
| 702 | HC9737 | ETS1;FOXC1;FOXL1;GATA2;GATA3;MAFB;MZF1; PRRX2;SOX10;SOX5;SP1;TFAP2A;YY1;ZEB1;ZNF354C | 15 | 2 | 3 | 2 | 4 | 5 | 4 | 5 | 2 |
| 703 | HC9743 | BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3; MZF1;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C | 15 | 2 | 3 | 2 | 3 | 4 | 4 | 5 | 2 |
| 704 | HC3554 | ARID3A;ELK1;ETS1;FOXC1;FOXL1;FOXL1;GATA2; GATA3;MZF1;NKX3-2;PRRX2;SP1;SPIB;TFAP2A;THRA;ZEB1 | 16 | 1 | 4 | 1 | 5 | 4 | 4 | 6 | 3 |
| 705 | HC232 | ARID3A;BRCA1;ETS1;FOXC1;GATA2;GATA3;KLF4; MZF1;NKX2-5;PAX2;REL;SOX10;SPIB;TFAP2A;YY1;ZEB1 | 16 | 2 | 4 | 2 | 4 | 6 | 5 | 5 | 3 |

| | | | | | | | | | | | |
|-----|--------|---|----|---|---|---|---|---|---|---|---|
| 706 | HC2656 | ARID3A;ELK1;ETS1;FOXC1;FOXJ1;FOXL1;GATA2;GATA3; MZF1;NKX3-2;PRRX2;SP1;SPIB;TFAP2A;THRA;YY1;ZEB1 | 17 | 1 | 4 | 1 | 5 | 4 | 4 | 6 | 3 |
| 707 | HC2456 | ARID3A;BRCA1;ELF5;ELK1;ETS1;FOXC1;FOXL1;GATA2; GATA3;HOXA5;MAFB;PDX1;SOX10;SPIB;TFAP2A;YY1;ZEB1 | 17 | 3 | 4 | 2 | 4 | 6 | 6 | 7 | 4 |
| 708 | HC700 | ARID3A;ELK1;ETS1;FOXC1;FOXJ1;FOXL1;GATA2;GATA3; MZF1;NKX2-5;PRRX2;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 17 | 2 | 5 | 2 | 5 | 5 | 5 | 7 | 3 |
| 709 | HC2815 | ELF5;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5; MZF1;NKX2-5;PAX2;PDX1;SOX10;SPIB;TFAP2A;YY1;ZEB1 | 17 | 1 | 5 | 1 | 5 | 8 | 7 | 8 | 5 |
| 710 | HC1787 | ARID3A;BRCA1;CEBPA;ETS1;FOXC1;FOXL1;GATA2; MAFB;MZF1;NKX2-5;PAX2;PDX1;REL;SOX10;TFAP2A; YY1;ZEB1;ZNF354C | 18 | 3 | 5 | 3 | 4 | 7 | 5 | 6 | 3 |
| 711 | HC682 | ARID3A;BRCA1;CREB1;ELF5;ETS1;FOXC1;FOXJ1; FOXL1;GATA2;GATA3;KLF4;NR4A2;SOX10; SP1;SPIB;TFAP2A;YY1; ZEB1;ZNF354C | 19 | 5 | 4 | 4 | 4 | 7 | 6 | 7 | 4 |
| 712 | HC2931 | ARID3A;BPTF;BRCA1;ELK1;ETS1;FOXC1;FOXL1; GATA2;GATA3;MAFB;MZF1;NKX2-5;PRRX2;SP1; SPIB;TFAP2A;YY1;ZEB1;ZNF354C | 19 | 3 | 4 | 3 | 4 | 4 | 4 | 6 | 3 |

Table A.1: The local controllability of 712 protein complexes in the TF regulatory networks of 8 cell types. For each protein complex, a numbered index (No.), the complex identifier (ID) and its size are also given.

The p -values for protein complexes

In this appendix, we show the two types of p -values for protein complexes: p_1 and p_2 . For a given complex size s ($s = 5, 6, \dots, 19$), we calculated p_1 and p_2 for each complex with size s .

Denote the local controllability of the complex (with size s) by lc_0 , and the local controllability of 5000 random subsets with size s by a vector $d = (lc_1, lc_2, \dots, lc_{5000})$. Then, the p -value p_1 (p_2) is given by the fraction of elements lc_i in the vector d such that $lc_i \leq lc_0$ ($lc_i \geq lc_0$). The p -value p_1 (p_2) is used to test how significantly smaller (larger) is the local controllability of the complex than the local controllability of the random subsets.

Since there might be multiple protein complexes with size s and with the same local controllability (lc value), we thus show the results of p -values (p_1 and p_2) for each specific lc value of protein complexes. The results are shown for 8 cell types, as in Table B.1-B.8. In these tables, the significant p -values ($p_1 < 0.05$ or $p_2 < 0.05$) are marked in bold.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------|---------------|
| 5 | 1 | 0.9118 | 1 |
| 6 | 1 | 0.8808 | 1 |
| | 2 | 0.9892 | 0.1192 |
| 7 | 1 | 0.8526 | 1 |
| | 2 | 0.9796 | 0.1474 |
| 8 | 1 | 0.8014 | 1 |
| | 2 | 0.972 | 0.1986 |
| 9 | 1 | 0.7578 | 1 |
| | 2 | 0.9616 | 0.2422 |
| 10 | 1 | 0.7176 | 1 |
| | 2 | 0.9468 | 0.2824 |
| 11 | 1 | 0.6794 | 1 |
| | 2 | 0.9304 | 0.3206 |
| 12 | 1 | 0.614 | 1 |
| | 2 | 0.9034 | 0.386 |
| | 3 | 0.983 | 0.0966 |
| 13 | 1 | 0.582 | 1 |
| | 2 | 0.8866 | 0.418 |
| 14 | 1 | 0.5456 | 1 |
| | 2 | 0.8712 | 0.4544 |
| | 3 | 0.9746 | 0.1288 |
| 15 | 1 | 0.497 | 1 |
| | 2 | 0.8356 | 0.503 |
| | 3 | 0.9652 | 0.1644 |
| | 4 | 0.9934 | 0.0348 |
| 16 | 1 | 0.4668 | 1 |
| | 2 | 0.8212 | 0.5332 |
| 17 | 1 | 0.4294 | 1 |
| | 3 | 0.9488 | 0.2012 |
| 18 | 2 | 0.75 | 0.6262 |
| 19 | 2 | 0.7244 | 0.6542 |
| | 4 | 0.9842 | 0.0774 |

Table B.1: The p -values p_1 and p_2 in the TF regulatory network of hESC.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------|---------------|
| 5 | 1 | 0.55 | 1 |
| | 2 | 0.869 | 0.45 |
| | 3 | 0.9796 | 0.131 |
| | 4 | 0.9986 | 0.0204 |
| 6 | 1 | 0.4344 | 1 |
| | 2 | 0.785 | 0.5656 |
| | 3 | 0.95 | 0.215 |
| | 4 | 0.9932 | 0.05 |
| 7 | 1 | 0.3406 | 1 |
| | 2 | 0.6996 | 0.6594 |
| | 3 | 0.909 | 0.3004 |
| | 4 | 0.9808 | 0.091 |
| 8 | 2 | 0.5972 | 0.7508 |
| | 3 | 0.8482 | 0.4028 |
| | 4 | 0.9588 | 0.1518 |
| | 5 | 0.9936 | 0.0412 |
| 9 | 2 | 0.522 | 0.8062 |
| | 3 | 0.7876 | 0.478 |
| | 4 | 0.9344 | 0.2124 |
| 10 | 1 | 0.1352 | 1 |
| | 2 | 0.4308 | 0.8648 |
| | 3 | 0.7134 | 0.5692 |
| | 4 | 0.8948 | 0.2866 |
| | 5 | 0.974 | 0.1052 |
| 11 | 2 | 0.3482 | 0.903 |
| | 3 | 0.6506 | 0.6518 |
| | 4 | 0.8596 | 0.3494 |
| | 5 | 0.9556 | 0.1404 |
| | 6 | 0.9888 | 0.0444 |
| 12 | 2 | 0.2926 | 0.9232 |
| | 3 | 0.5658 | 0.7074 |
| | 4 | 0.8006 | 0.4342 |
| | 5 | 0.9296 | 0.1994 |
| 13 | 2 | 0.235 | 0.9428 |
| | 3 | 0.4852 | 0.765 |
| | 4 | 0.74 | 0.5148 |
| | 5 | 0.8972 | 0.26 |
| 14 | 2 | 0.1816 | 0.9636 |
| | 3 | 0.4364 | 0.8184 |
| | 4 | 0.6758 | 0.5636 |
| | 5 | 0.8516 | 0.3242 |
| 15 | 2 | 0.1442 | 0.9734 |
| | 3 | 0.3586 | 0.8558 |
| | 4 | 0.5996 | 0.6414 |
| | 5 | 0.797 | 0.4004 |
| | 6 | 0.9166 | 0.203 |
| 16 | 4 | 0.5374 | 0.7006 |
| 17 | 4 | 0.4706 | 0.7532 |
| | 5 | 0.6884 | 0.5294 |
| 18 | 5 | 0.6452 | 0.5744 |
| 19 | 4 | 0.3684 | 0.8224 |

Table B.2: The p -values p_1 and p_2 in the TF regulatory network of Amniotic Epi.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------|---------------|
| 5 | 1 | 0.8104 | 1 |
| | 2 | 0.9788 | 0.1896 |
| 6 | 1 | 0.7388 | 1 |
| | 2 | 0.9588 | 0.2612 |
| | 3 | 0.9966 | 0.0412 |
| 7 | 1 | 0.6676 | 1 |
| | 2 | 0.9326 | 0.3324 |
| 8 | 1 | 0.5866 | 1 |
| | 2 | 0.9022 | 0.4134 |
| 9 | 1 | 0.5258 | 1 |
| | 2 | 0.8706 | 0.4742 |
| | 3 | 0.979 | 0.1294 |
| 10 | 1 | 0.4654 | 1 |
| | 2 | 0.8326 | 0.5346 |
| 11 | 1 | 0.4056 | 1 |
| | 2 | 0.782 | 0.5944 |
| | 3 | 0.95 | 0.218 |
| 12 | 1 | 0.346 | 1 |
| | 2 | 0.729 | 0.654 |
| | 3 | 0.93 | 0.271 |
| | 4 | 0.9872 | 0.07 |
| 13 | 1 | 0.2966 | 1 |
| | 2 | 0.6806 | 0.7034 |
| | 3 | 0.908 | 0.3194 |
| 14 | 1 | 0.2538 | 1 |
| | 2 | 0.6342 | 0.7462 |
| | 3 | 0.8804 | 0.3658 |
| | 4 | 0.9712 | 0.1196 |
| 15 | 1 | 0.2224 | 1 |
| | 2 | 0.5842 | 0.7776 |
| | 3 | 0.8574 | 0.4158 |
| | 4 | 0.9648 | 0.1426 |
| | 5 | 0.9912 | 0.0352 |
| 16 | 1 | 0.1822 | 1 |
| | 2 | 0.5266 | 0.8178 |
| 17 | 1 | 0.1568 | 1 |
| | 2 | 0.4766 | 0.8432 |
| 18 | 3 | 0.7454 | 0.5654 |
| 19 | 3 | 0.7082 | 0.5992 |
| | 4 | 0.893 | 0.2918 |

Table B.3: The p -values p_1 and p_2 in the TF regulatory network of Hemat. Stem Cell.

| Complex size | l_c value of complex | p_1 | p_2 |
|--------------|------------------------|---------------|--------|
| 5 | 1 | 0.5236 | 1 |
| | 2 | 0.8656 | 0.4764 |
| 6 | 1 | 0.4008 | 1 |
| | 2 | 0.774 | 0.5992 |
| | 3 | 0.952 | 0.226 |
| 7 | 1 | 0.3074 | 1 |
| | 2 | 0.6918 | 0.6926 |
| | 3 | 0.9122 | 0.3082 |
| | 4 | 0.9816 | 0.0878 |
| 8 | 1 | 0.2418 | 1 |
| | 2 | 0.5868 | 0.7582 |
| | 3 | 0.8544 | 0.4132 |
| | 4 | 0.9648 | 0.1456 |
| 9 | 1 | 0.1778 | 1 |
| | 2 | 0.5186 | 0.8222 |
| | 3 | 0.811 | 0.4814 |
| | 5 | 0.9868 | 0.0608 |
| 10 | 1 | 0.1184 | 1 |
| | 2 | 0.4146 | 0.8816 |
| | 3 | 0.7046 | 0.5854 |
| | 4 | 0.8896 | 0.2954 |
| | 5 | 0.9708 | 0.1104 |
| 11 | 1 | 0.0912 | 1 |
| | 2 | 0.3472 | 0.9088 |
| | 3 | 0.647 | 0.6528 |
| | 4 | 0.848 | 0.353 |
| | 5 | 0.9544 | 0.152 |
| 12 | 1 | 0.064 | 1 |
| | 2 | 0.2748 | 0.936 |
| | 3 | 0.559 | 0.7252 |
| | 4 | 0.7918 | 0.441 |
| | 5 | 0.928 | 0.2082 |
| 13 | 2 | 0.2114 | 0.954 |
| | 3 | 0.4852 | 0.7886 |
| | 4 | 0.7434 | 0.5148 |
| 14 | 1 | 0.0278 | 1 |
| | 2 | 0.169 | 0.9722 |
| | 3 | 0.42 | 0.831 |
| | 4 | 0.6784 | 0.58 |
| | 5 | 0.8578 | 0.3216 |
| 15 | 2 | 0.1312 | 0.975 |
| | 3 | 0.3522 | 0.8688 |
| | 4 | 0.6024 | 0.6478 |
| | 5 | 0.7938 | 0.3976 |
| | 6 | 0.9128 | 0.2062 |
| | 16 | 4 | 0.5448 |
| 16 | 5 | 0.7518 | 0.4552 |
| 17 | 4 | 0.4848 | 0.755 |
| | 5 | 0.6932 | 0.5152 |
| 18 | 4 | 0.4228 | 0.8014 |
| 19 | 4 | 0.369 | 0.8318 |

Table B.4: The p -values p_1 and p_2 in the TF regulatory network of Adult Dermal Blood.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------|---------------|
| 5 | 1 | 0.5176 | 1 |
| | 2 | 0.8678 | 0.4824 |
| | 3 | 0.9812 | 0.1322 |
| 6 | 1 | 0.3992 | 1 |
| | 2 | 0.7852 | 0.6008 |
| | 3 | 0.9562 | 0.2148 |
| 7 | 1 | 0.2998 | 1 |
| | 2 | 0.6926 | 0.7002 |
| | 3 | 0.9122 | 0.3074 |
| | 4 | 0.9824 | 0.0878 |
| | 5 | 0.9978 | 0.0176 |
| 8 | 1 | 0.2348 | 1 |
| | 2 | 0.6068 | 0.7652 |
| | 3 | 0.8648 | 0.3932 |
| | 4 | 0.9692 | 0.1352 |
| | 5 | 0.996 | 0.0308 |
| 9 | 1 | 0.1636 | 1 |
| | 2 | 0.5048 | 0.8364 |
| | 3 | 0.8012 | 0.4952 |
| | 4 | 0.9402 | 0.1988 |
| | 6 | 0.998 | 0.0118 |
| | 10 | 1 | 0.107 |
| 2 | | 0.4196 | 0.893 |
| 3 | | 0.733 | 0.5804 |
| 4 | | 0.9128 | 0.267 |
| 5 | | 0.9788 | 0.0872 |
| 6 | | 0.9958 | 0.0212 |
| 11 | 2 | 0.3498 | 0.912 |
| | 3 | 0.6594 | 0.6502 |
| | 4 | 0.8726 | 0.3406 |
| | 5 | 0.961 | 0.1274 |
| | 6 | 0.9934 | 0.039 |
| | 7 | 0.9988 | 0.0066 |
| | 12 | 3 | 0.5878 |
| 4 | | 0.8192 | 0.4122 |
| 5 | | 0.9328 | 0.1808 |
| 6 | | 0.981 | 0.0672 |
| 7 | | 0.9952 | 0.019 |
| 13 | 3 | 0.508 | 0.779 |
| | 4 | 0.7592 | 0.492 |
| | 5 | 0.9092 | 0.2408 |
| | 6 | 0.9714 | 0.0908 |
| | 7 | 0.9934 | 0.0286 |
| 14 | 3 | 0.4438 | 0.8338 |
| | 4 | 0.7072 | 0.5562 |
| | 5 | 0.8736 | 0.2928 |
| | 6 | 0.9566 | 0.1264 |
| | 7 | 0.9876 | 0.0434 |
| 15 | 3 | 0.3608 | 0.875 |
| | 4 | 0.6192 | 0.6392 |
| | 5 | 0.8264 | 0.3808 |
| | 6 | 0.9306 | 0.1736 |
| | 7 | 0.9804 | 0.0694 |
| | 8 | 0.9952 | 0.0196 |
| 16 | 4 | 0.5742 | 0.6838 |
| | 6 | 0.909 | 0.2126 |
| 17 | 4 | 0.5038 | 0.7508 |
| | 5 | 0.7328 | 0.4962 |
| | 6 | 0.88 | 0.2672 |
| | 8 | 0.9872 | 0.0406 |
| 18 | 7 | 0.9358 | 0.1578 |
| 19 | 4 | 0.382 | 0.8328 |
| | 7 | 0.911 | 0.1952 |

Table B.5: The p -values p_1 and p_2 in the TF regulatory network of Neuroblastoma.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------|---------------|
| 5 | 1 | 0.6666 | 1 |
| | 2 | 0.9346 | 0.3334 |
| | 3 | 0.9944 | 0.0654 |
| 6 | 1 | 0.5862 | 1 |
| | 2 | 0.8874 | 0.4138 |
| | 3 | 0.983 | 0.1126 |
| | 4 | 0.9982 | 0.017 |
| 7 | 1 | 0.511 | 1 |
| | 2 | 0.837 | 0.489 |
| | 3 | 0.9634 | 0.163 |
| | 4 | 0.9998 | 0.0044 |
| | 5 | | |
| 8 | 1 | 0.4128 | 1 |
| | 2 | 0.7782 | 0.5872 |
| | 3 | 0.939 | 0.2218 |
| | 4 | 0.9896 | 0.061 |
| | 5 | 0.9984 | 0.0104 |
| 9 | 1 | 0.3312 | 1 |
| | 2 | 0.7008 | 0.6688 |
| | 3 | 0.9082 | 0.2992 |
| | 4 | 0.9804 | 0.0918 |
| | 5 | 0.9998 | 0.0022 |
| | 6 | | |
| 10 | 1 | 0.2708 | 1 |
| | 2 | 0.6272 | 0.7292 |
| | 3 | 0.874 | 0.3728 |
| | 4 | 0.9628 | 0.126 |
| | 5 | 0.9944 | 0.0372 |
| | 6 | 0.9986 | 0.0056 |
| 11 | 2 | 0.5744 | 0.7742 |
| | 3 | 0.8352 | 0.4256 |
| | 4 | 0.953 | 0.1648 |
| | 5 | 0.9864 | 0.047 |
| | 6 | 0.9974 | 0.0136 |
| | | | |
| 12 | 2 | 0.4986 | 0.8292 |
| | 3 | 0.7774 | 0.5014 |
| | 4 | 0.9232 | 0.2226 |
| | 5 | 0.9778 | 0.0768 |
| | | | |
| 13 | 2 | 0.4472 | 0.8452 |
| | 3 | 0.7278 | 0.5528 |
| | 4 | 0.8962 | 0.2722 |
| | 5 | 0.9684 | 0.1038 |
| | | | |
| 14 | 2 | 0.385 | 0.891 |
| | 3 | 0.683 | 0.615 |
| | 4 | 0.8702 | 0.317 |
| | 5 | 0.9588 | 0.1298 |
| | 6 | 0.9912 | 0.0412 |
| | | | |
| 15 | 2 | 0.326 | 0.9058 |
| | 3 | 0.6196 | 0.674 |
| | 4 | 0.834 | 0.3804 |
| | 5 | 0.9418 | 0.166 |
| | 6 | 0.9834 | 0.0582 |
| | 7 | 0.9966 | 0.0166 |
| | | | |
| 16 | 4 | 0.798 | 0.4344 |
| | 5 | 0.924 | 0.202 |
| 17 | 4 | 0.7528 | 0.4838 |
| | 5 | 0.8924 | 0.2472 |
| | 6 | 0.9628 | 0.1076 |
| | 7 | 0.9904 | 0.0372 |
| | | | |
| 18 | 5 | 0.8756 | 0.2862 |
| 19 | 4 | 0.6536 | 0.5968 |
| | 6 | 0.9336 | 0.167 |

Table B.6: The p -values p_1 and p_2 in the TF regulatory network of Skeletal Myoblast.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------|---------------|
| 5 | 1 | 0.5958 | 1 |
| | 2 | 0.907 | 0.4042 |
| | 3 | 0.989 | 0.093 |
| | 4 | 0.9994 | 0.011 |
| 6 | 1 | 0.4842 | 1 |
| | 2 | 0.8424 | 0.5158 |
| | 3 | 0.9676 | 0.1576 |
| | 4 | 0.9952 | 0.0324 |
| 7 | 1 | 0.3772 | 1 |
| | 2 | 0.7602 | 0.6228 |
| | 3 | 0.9446 | 0.2398 |
| | 4 | 0.993 | 0.0554 |
| | 6 | 0.9998 | 0.0016 |
| 8 | 1 | 0.3068 | 1 |
| | 2 | 0.6812 | 0.6932 |
| | 3 | 0.9026 | 0.3188 |
| | 4 | 0.979 | 0.0974 |
| | 5 | 0.997 | 0.021 |
| 9 | 2 | 0.5952 | 0.7652 |
| | 3 | 0.853 | 0.4048 |
| | 4 | 0.9582 | 0.147 |
| | 5 | 0.9926 | 0.0418 |
| | 7 | 1 | 0.001 |
| 10 | 2 | 0.5158 | 0.8208 |
| | 3 | 0.807 | 0.4842 |
| | 4 | 0.9416 | 0.193 |
| | 5 | 0.9844 | 0.0584 |
| | 7 | 0.9998 | 0.0022 |
| 11 | 2 | 0.4222 | 0.8752 |
| | 3 | 0.726 | 0.5778 |
| | 4 | 0.902 | 0.274 |
| | 5 | 0.9728 | 0.098 |
| | 6 | 0.9924 | 0.0272 |
| 12 | 7 | 0.9986 | 0.0076 |
| | 2 | 0.3766 | 0.8946 |
| | 3 | 0.6712 | 0.6234 |
| | 4 | 0.872 | 0.3288 |
| | 5 | 0.9556 | 0.128 |
| 13 | 6 | 0.9912 | 0.0444 |
| | 7 | 0.9984 | 0.0088 |
| | 3 | 0.602 | 0.703 |
| | 4 | 0.8216 | 0.398 |
| | 5 | 0.9342 | 0.1784 |
| 14 | 6 | 0.9834 | 0.0658 |
| | 3 | 0.528 | 0.7576 |
| | 4 | 0.7732 | 0.472 |
| | 5 | 0.9112 | 0.2268 |
| | 6 | 0.9708 | 0.0888 |
| 15 | 7 | 0.9916 | 0.0292 |
| | 2 | 0.1938 | 0.9614 |
| | 3 | 0.468 | 0.8062 |
| | 4 | 0.7062 | 0.532 |
| | 5 | 0.8762 | 0.2938 |
| 16 | 6 | 0.9582 | 0.1238 |
| | 7 | 0.99 | 0.0418 |
| | 8 | 0.9974 | 0.01 |
| | 5 | 0.843 | 0.339 |
| | 6 | 0.944 | 0.157 |
| 17 | 6 | 0.9166 | 0.1954 |
| | 7 | 0.9714 | 0.0834 |
| | 8 | 0.9934 | 0.0286 |
| 18 | 6 | 0.8942 | 0.243 |
| 19 | 6 | 0.8628 | 0.2848 |
| | 7 | 0.9424 | 0.1372 |

Table B.7: The p -values p_1 and p_2 in the TF regulatory network of Fetal Brain.

| Complex size | lc value of complex | p_1 | p_2 |
|--------------|-----------------------|--------------|--------|
| 5 | 1 | 0.6012 | 1 |
| | 2 | 0.893 | 0.3988 |
| 6 | 1 | 0.484 | 1 |
| | 2 | 0.8324 | 0.516 |
| | 3 | 0.9698 | 0.1676 |
| 7 | 1 | 0.3908 | 1 |
| | 2 | 0.7576 | 0.6092 |
| | 3 | 0.9374 | 0.2424 |
| 8 | 1 | 0.2968 | 1 |
| | 2 | 0.6684 | 0.7032 |
| | 3 | 0.892 | 0.3316 |
| 9 | 1 | 0.2326 | 1 |
| | 2 | 0.5888 | 0.7674 |
| | 3 | 0.8454 | 0.4112 |
| | 4 | 0.9548 | 0.1546 |
| 10 | 1 | 0.1726 | 1 |
| | 2 | 0.4886 | 0.8274 |
| | 3 | 0.779 | 0.5114 |
| | 4 | 0.9294 | 0.221 |
| 11 | 1 | 0.1274 | 1 |
| | 2 | 0.422 | 0.8726 |
| | 3 | 0.7198 | 0.578 |
| | 4 | 0.8966 | 0.2802 |
| 12 | 1 | 0.0956 | 1 |
| | 2 | 0.3478 | 0.9044 |
| | 3 | 0.6494 | 0.6522 |
| | 4 | 0.8538 | 0.3506 |
| 13 | 1 | 0.0776 | 1 |
| | 2 | 0.2938 | 0.9224 |
| | 3 | 0.5818 | 0.7062 |
| 14 | 1 | 0.0512 | 1 |
| | 2 | 0.239 | 0.9488 |
| | 3 | 0.5142 | 0.761 |
| | 4 | 0.7578 | 0.4858 |
| 15 | 1 | 0.039 | 1 |
| | 2 | 0.202 | 0.961 |
| | 3 | 0.4644 | 0.798 |
| | 4 | 0.7058 | 0.5356 |
| | 5 | 0.8672 | 0.2942 |
| 16 | 3 | 0.3976 | 0.8478 |
| 17 | 3 | 0.3354 | 0.88 |
| | 4 | 0.5864 | 0.6646 |
| | 5 | 0.7762 | 0.4136 |
| 18 | 3 | 0.2866 | 0.9046 |
| 19 | 3 | 0.2316 | 0.9236 |
| | 4 | 0.4618 | 0.7684 |

Table B.8: The p -values p_1 and p_2 in the TF regulatory network of Skin Fib.

**LOCAL CONTROLLABILITY OF
BIOLOGICAL NETWORKS**

LUO CHANG

NATIONAL UNIVERSITY OF SINGAPORE

2015