BIOLOGICAL NETWORK ANALYSIS AND COMPARISON

TIAN DECHAO

(MASTER OF SCIENCE, NORTHEAST NORMAL UNIVERSITY, CHINA)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby 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.

Tian Dechao

5th February, 2015

Thesis Supervisor

Choi Kwok Pui Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546

Papers and Manuscripts

Papers

Tian, D., and Choi, K. P. (2013). Sharp Bounds and Normalization of Wiener-Type Indices. *PLoS ONE*, 8(11), e78448.

Zhang, S.*, **Tian, D.***, Tran, N. H., Choi, K. P., and Zhang, L.X (2014). Profiling the Transcription Factor Regulatory Networks of Human Cell Types. *Nucleic Acids Research*, 42(20), 12380–12387.

Manuscript

Tian, D., Choi, K. P., and Zhang, L.X. Profiling Human Embryonic Stem Cell via Feed-Forward Loops in Transcription Factor Regulatory Network.

* co-first authorship

Acknowledgements

I would like first to express my gratitude and appreciation to my supervisor, Prof. Choi Kwok Pui for his complete trust, endless patience, and expert guidance in my research. He helps me to build up my confidence in further pursuit of my academic dream and provides detailed recommendations for my future plan. He also takes care of my daily life and always hope the best for me.

I would like to show my gratitude for Prof. Zhang Louxin. His style of thinking, analyzing, and writing help me a lot in research. My thanks also goes to the other Network Biology group members for helpful discussion and warm friendship.

I want to take this opportunity to thank Prof. Bai Zhidong for his support in my PhD application, encouragement in my research and care for my daily life. I would like to express special thanks to other faculty members and support staff. I am grateful to National University of Singapore for awarding me the Graduate Research Scholarship to pursue research in my area of interest.

I would also like to express my sincere thanks to my friends Dr. Li Xiang and Dr. Huang Zhipeng for their friendship and encouragement in the journey. I would like to thank seniors Dr. Hu Jiang, Dr. Li Hua and Dr. Xia Ningning for their generous guidance and kind help in many aspects.

Also I would like to thank other PhD students in Department of Statistics and Applied Probability who helped me in one way or another. All my friends whom I have forgotten to mention here are also greatly appreciated for their assistance and encouragement.

Finally, I am grateful to my family for their unconditional support and encouragement.

Contents

Summary

Complex networks abound in physical, biological and social sciences. Quantifying a network's topological structure facilitates network exploration and analysis, and network comparison, clustering and classification. A number of Wiener type indices have recently been incorporated as distance-based descriptors of complex networks, such as the R package QuACN. Wiener type indices are known to depend both on the network's number of nodes and topology. To apply these indices to measure similarity of networks of different numbers of nodes, normalization of these indices is needed to correct the effect of the number of nodes in a network. Chapter 2 aims to fill this gap. Moreover, we introduce an f-Wiener index of network G, denoted by $W_f(G)$. This notion generalizes the Wiener index to a very wide class of Wiener type indices including all known Wiener type indices. We identify the maximum and minimum of $W_f(G)$ over a set of networks with n nodes. We then introduce our normalized-version of f-Wiener index. The normalized f-Wiener indices were demonstrated, in a number of experiments, to improve significantly the hierarchical clustering over the non-normalized counterparts.

Neph et al. (2012a) reported the transcription factor (TF) regulatory networks of 41 human cell types using the DNaseI footprinting technique. This provides a valuable resource for uncovering regulation principles in

different human cells. In chapter 3, the architectures of the 41 regulatory networks and the distributions of housekeeping and specific regulatory interactions are investigated. The TF regulatory networks of different human cell types demonstrate similar global three-layer (top, core, and bottom) hierarchical architectures, which are greatly different from the yeast TF regulatory network. However, they have distinguishable local organizations, as suggested by the fact that wiring patterns of only a few TFs are enough to distinguish cell identities. The TF regulatory network of human embryonic stem cells (hESCs) is dense and enriched with interactions that are unseen in the networks of other cell types. The examination of specific regulatory interactions suggests that specific interactions play important roles in hESCs.

An Feed-Forward Loop (FFL) consists of 3 nodes A, B and C in which A regulates B, and both A and B regulate C. In chapter 4, we compared local regulatory landscapes on each TF in terms of FFLs in regulatory network of hESC with those in other 40 differentiated cell types reported by Neph et al. (2012a). Firstly we found that distributional properties of FFL regulating each TF can reproduce embryonic origin and known cell-lineage relationship well. The clustering is comparable with clusterings based on distance matrices produced by *netdis* (Ali et al., 2014). Secondly we identified 28 TFs extensively regulated by FFLs in hESC only. Among them 13 TFs perform hESC related functions. While remaining 15 TFs are master TFs in various differentiated cell types. Thirdly, our proposed scores perform better in identifying hESC related TFs than FFL-based centrality measures in Koschützki et al. (2007).

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Complex biological networks

Living cells' characteristics are maintained by complex biological systems which contain numerous components such as DNA, RNA, proteins, and their interactions. Each of these components has been extensively studied to investigate its functions in maintaining cell states and decipher complex cellular systems. It is increasingly clear that biological functions can rarely be attributed to an individual component. Instead, recently more and more evidence demonstrates that important functions are played by interactions between components in maintaining cellular functions (Barabasi and Oltvai, 2004). These discoveries highlight the need to study complex biological systems as a whole. A key challenge is to study structure and dynamics of complex biological systems across conditions, e.g. cell stages, cell types or species, etc. To this direction, complex biological systems are represented by biological networks. A network can be metabolic network, protein-protein interaction (PPI) network, regulatory network, etc. Metabolic networks are classic examples for using a network to represent metabolic pathways. Two

metabolic substrates, denoted as a and b, are connected by a directed interaction if a known metabolic reaction exists that acts on a and produces b. PPI networks symbolize physical interactions between proteins. Gene regulatory networks (GRNs) depict gene expression regulation, where a gene's expression is regulated by their regulators (Figure 1.1). Transcription factor (TF) regulatory networks represent regulation of a TF by other TFs (Figures 1.2 and 1.5). Interactions between cellular components rewire at different conditions, for example, stages of a cell, different cell types or across species. Thus these networks could be time-specific, cell type-specific or species-specific, etc. Moreover, these networks are associated with each other and form a "network of networks" that control cell behaviours.

1.2 High-throughput technologies to map networks

1.2.1 High-throughput technologies

Currently two high-throughput technologies are widely used to map PPI networks, namely Yeast two-hybrid (Y2H) assays (Chen et al., 2010) and affinity purification followed by mass spectrometry (AP-MS) assays (Gingras et al., 2007). Y2H assays can detect direct physical interactions between proteins whereas AP-MS assays can detect protein complexes and indirect association between proteins.

To map regulatory networks, technologies are Yeast one-hybrid (Y1H) assays (Deplancke et al., 2004), Chromatin Immunoprecipitation (ChIP) experiments (Lee et al., 2002) and DNaseI footprinting (Boyle et al., 2011; Galas and Schmitz, 1978; Gusmao et al., 2014; Neph et al., 2012b) are widely applied high-throughput technologies. In Y1H assays, a specific reg-

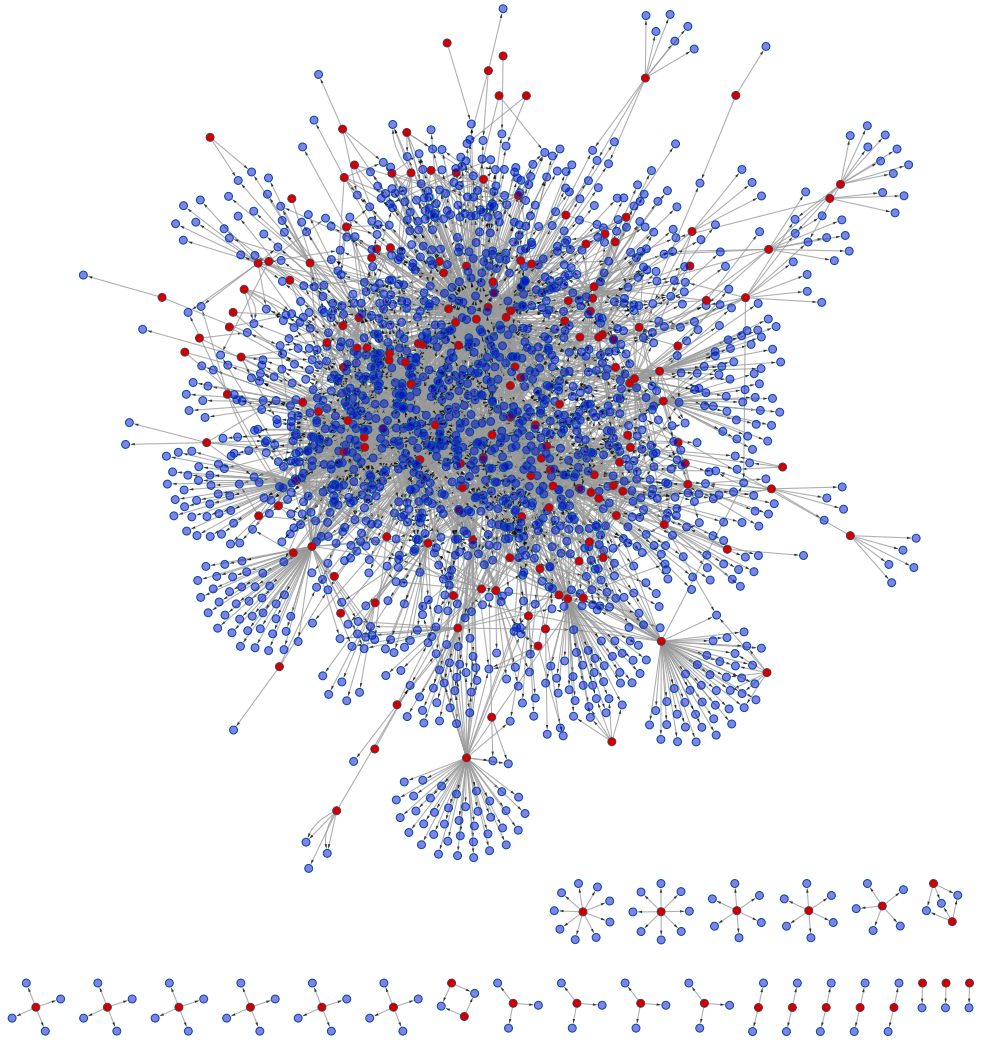


Figure 1.1. Gene regulatory network of E. coli. There are 197 TFs (red circle), 1745 target genes (blue circle) and 1942 directed interactions. Data from RegulonDB (version 8.0, Salgado et al. (2013)). Network visualization: Cytoscape (version 3.1.0, Kohl et al. (2011)).

ulatory DNA sequence of interest, named as promotor, is used as a bait to identify all putative TFs (preys) that bind to this sequence. On the other hand, Chip experiments are applied to delineate all potentially associated DNA binding sites for a DNA-binding protein of interest.

TF regulatory networks studied in chapters 3 and 4 of this thesis are produced by this approach. DNaseI footprinting is developed by Galas and

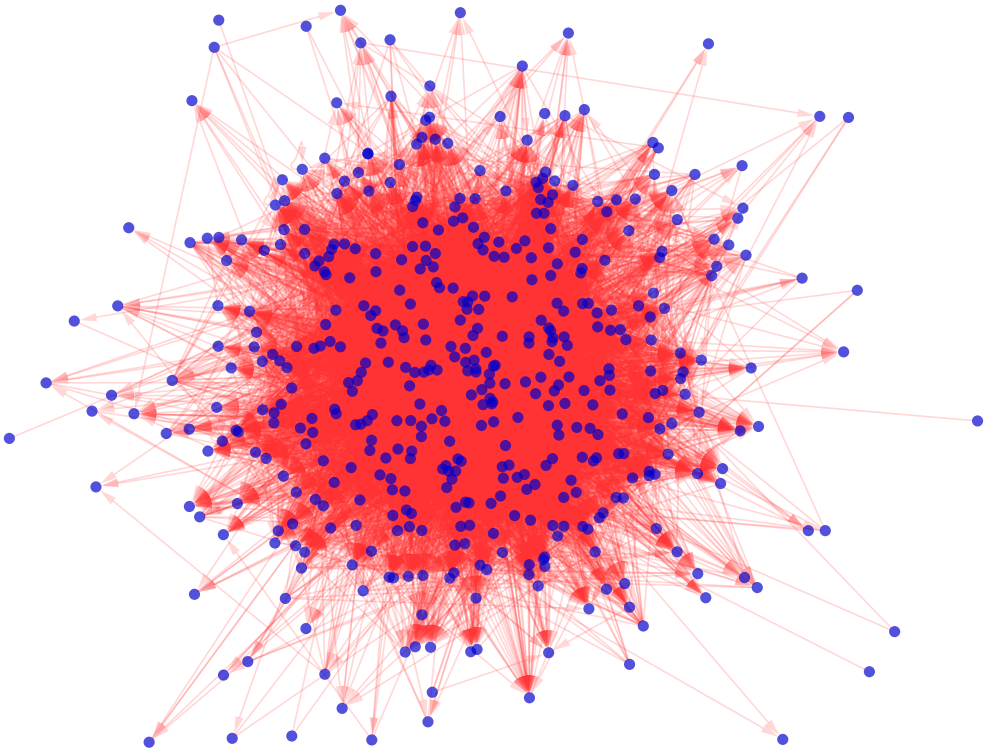


Figure 1.2. Transcription factor (TF) regulatory network in human embryonic stem cell. There are 470 TFs and 13176 interactions. Data from Neph et al. (2012a). Network visualization: Cytoscape (version 3.1.0, Kohl et al. (2011)).

Schmitz (1978) to analyze regulatory sequences in diverse organisms. DNaseI footprinting is a well-established approach for identifying direct regulatory interactions and provides a powerful genetic approach for assaying the occupancy of specific sequence elements which can regulate downstream genes. It is successfully applied to discover the first human sequence-specific TFs (Dynan and Tjian, 1983). DNaseI footprinting technology first binds nuclear chromatin with a protein of interest. Then the chromatin sequence is cleaved by certain enzyme. The protein will protect the binding region from being cleaved thus leaving "footprints" which indicate binding of the protein to the chromatin. The workflow is illustrated in Figure 1.3. Armed

with next-generation sequencing, upgraded DNaseI footprinting approaches are able to identify DNA footprints on a genome-wide scale (Boyle et al., 2011; Gusmao et al., 2014; Neph et al., 2012b). Neph et al. (2012b) improves this approach by integrating DNaseI footprints and the predicted TRANSFAC motif-binding sites (Ravasi et al., 2010). This approach can accurately and quantitatively recapitulate Chip-seq data for individual TFs, while simultaneously interrogating the genomic occupancy of potentially all expressed DNA-binding factors in a single experiment.

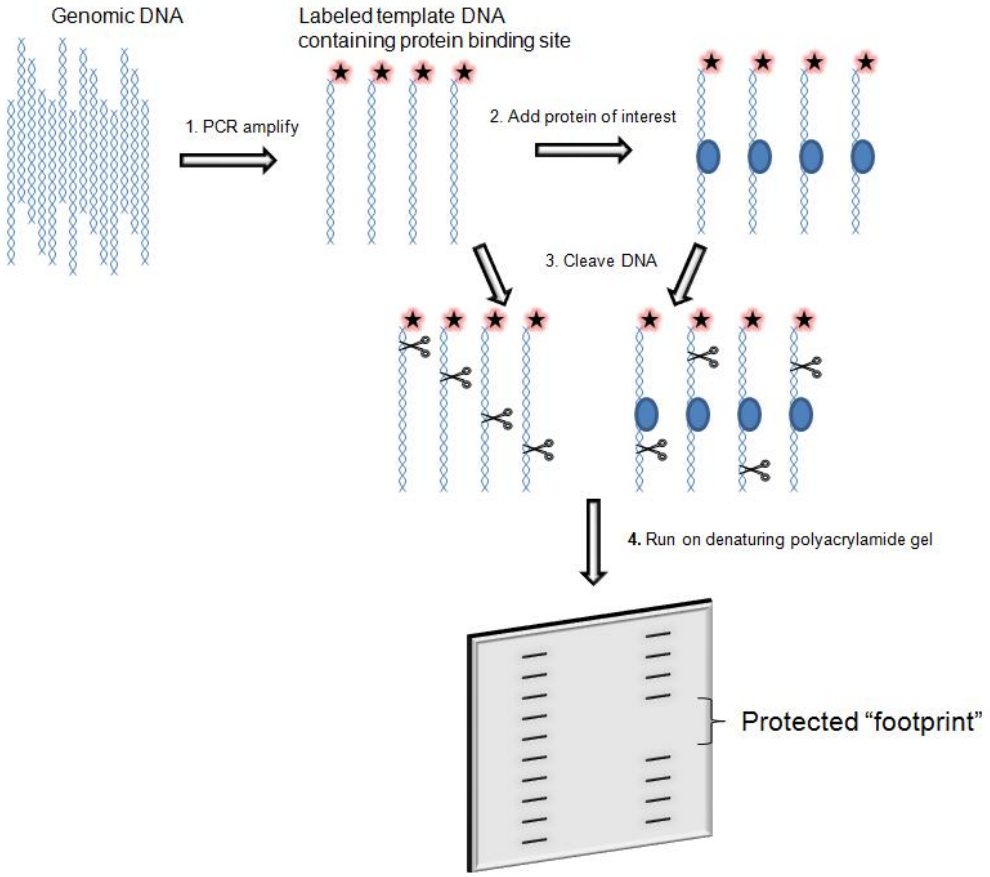


Figure 1.3. Illustration of DNaseI footprinting workflow. Figure is downloaded from Wikipedia (http://en.wikipedia.org/wiki/DNA_footprinting#cite_note-PMID22955618-14).

1.2.2 Errors in the observed biological networks

An observed network is a network detected by experiments to infer the respective real but unknown network. In an observed network, false positives (FP) refer to interactions that do not exist in the real network but are detected by experiments. False negatives (FN) refer to interactions in a real network but are not detected by experiments. True positives (TP) refer to interactions in a real network and are correctly detected by experiments. True negatives (TN) refer to interactions that do not exist in a real network and are not falsely detected by experiments. A high sensitivity, $TP/(TP + FN)$, indicates that a large proportion of interactions in the real network are identified in the observed network, and a high precision, $TP/(TP + FP)$, indicates that a high proportion of interactions in the observed network are actually exist in the real network.

Observed networks are prone to inaccuracy and low coverage due to limitations in high-throughput technologies and the complex nature of corresponding systems. For example, the precision for the human PPI dataset CCSB-HI1 was estimated at $\sim 79.4\%$, which corresponds to a false discovery rate $\sim 20.6\%$ (Venkatesan et al., 2008). The precision for a new high-quality PPI dataset of S. cerevisiae, CCSH-YI1, was estimated at $\sim 94\%$ by Yu et al. (2008). Although new Y2H assays achieve very high precision, the sensitivity is quite low, where the best sensitivity is at $\sim 17\%$ for S. cerevisiae.

1.2.3 Network resources and databases

Recent years witness exponential growth in biological network data thanks to the rapid development of high-throughput technologies. A number of

open-access databases have been established to bank numerous network data sets. To name a few well-known databases, PPI networks of multiple species are available in DIP (Xenarios et al., 2002), BioGRID (Chatr-aryamontri et al., 2013), STRING (Chatr-aryamontri et al., 2013), etc. Although the quality of results from Y2H studies, which supply the core of DIP database, is debated (Von Mering et al., 2002), the manually curated DIP database represents currently most reliable yeast protein interactions and provides sufficient data for their unambiguous statistical analyses (Wuchty et al., 2003). GRN can be downloaded from TRANSFAC (Matys et al., 2003), RegulonDB (Salgado et al., 2013), AtRegNet (Palaniswamy et al., 2006), etc. KEGG (Kanehisa and Goto, 2000), perhaps, is the most comprehensive database for metabolic networks and pathways. Some other useful resources include MIPS (Pagel et al., 2005), BIND (Bader et al., 2003), BioCyc (Caspi et al., 2008), Reactome (Croft et al., 2010), etc. A brief summary of over 300 resources related to biological networks and pathways can be found in the meta-database Pathguide (http://www.pathguide.org).

Besides open-access database, some publications also provide valuable network data. One example is the 41 human TF regulatory networks produced by Neph et al. (2012a). The authors combined DNaseI footprinting technology and TRANSFAC motif-binding sites (Ravasi et al., 2010) to map regulatory networks across 41 diverse human cell and tissue types. Each network contains about 475 sequence-specific TFs and 11,200 interactions. This data provide a good opportunity to study structural organizations and dynamics of human TF regulatory networks across cell and tissue types.

1.3 Mathematical formulation

1.3.1 Mathematical representation

A complex biological network is mathematically represented by a graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ is the set of nodes in the graph and $E \subseteq V \times V$ is the set of edges connecting nodes in V. A node stands for a functional component in the network. An edge (i, j) stands for a certain kind of relationship between nodes i and j in the network, depending on the nature of interactions in the network, (i, j) could be directed or undirected. For example, in a PPI network, nodes are proteins in the network and (i, j) denotes the physical interaction or functional association between proteins i and j and is an undirected edge. In a GRN, $i \in V$ denotes a gene and (i, j) denotes regulation of expression level of gene j by gene i hence a directed edge. For a TF regulatory network, i is a TF and (i, j) represents regulation of TF j by TF i, thus a directed edge.

1.3.2 Definitions and notations

Let $G = (V, E)$ be a connected directed or undirected graph. Denote by $N(G)$ the number of nodes in G. Size of G also refers to the number of nodes in G. The degree of a node, when G is an undirected graph, is the number of edges incident to this node. When G is a directed graph, the out-degree of a node is the number of edges originated from this node, the in-degree is the number of edges ended with this node, and the total-degree is the sum of in-degree and out-degree of this node. Degree distribution $P(k)$, $k = 0, 1, 2, \ldots$, is the probability that a randomly selected node has degree k. Similarly for out-degree and in-degree distribution. Hubs are nodes with

high degrees. Throughout this thesis, hubs refer to those nodes with top 20% degrees (Jothi et al., 2009). Out-degree hubs, in-degree hubs, and total-degree hubs are similarly defined.

Let $G = (V, E)$ be a simple (that is, no self-loops nor multiple edges) undirected and connected graph. Let $\mathcal{G}_n$ denote the set of all simple, connected graphs with n nodes. A graph having no cycles is called a tree, and we let $\mathcal{T}_n$ denote the set of all connected trees with n nodes. The distance $d(i, j)$ between any pair of nodes, i and j, in G is the number of edges in a shortest path from i to j. Let $D(G) = [d(i, j)]_{1 \leq i,j \leq n}$ be the distance matrix. We denote the maximum degree of G by $\Delta(G)$.

1.3.3 Biological network analysis and comparison

Complex biological networks are modelled as graphs. So that one can apply a wide-repertoire of results in graph theory to quantify network topological structures with the aim to find associations between significant topological structures and functional properties in biological networks. Many such associations have been reported in literature. For example, biological networks have many topological properties which are different from those in random networks. It is believed that these differences are attributed to functional constraints imposed on biological networks. To name a few of these associations, degree distributions in many biological networks are scale free, in other words, it follows a power law distribution, $P(k) \sim k^{-\lambda}$ (Barabasi and Albert, 1999), implying that a few nodes are hubs and connect to many others while the majority of nodes have a few connections. Biological networks are noted for having high clustering coefficients and small diameters and thus are small-world (Amaral et al., 2000). Good summaries can be

found in the review papers by Barabasi and Oltvai (2004) and Ma'ayan (2011). Chapter 2 will focus on Wiener type indices and their applications in network comparison.

A second area is to discover structural organization principles from complex biological networks, with the hope to universally interpret and model complex biological systems. A few global and local principles have been discovered across various biological networks. Global organization principle includes bow-tie structure organization and hierarchical structure organization. Local organization principle includes network motifs, which are viewed as basic building blocks of complex biological networks. The global and local organization principles nest together in biological networks and perform multiple functions as the network backbone. They can be starting points to model biological networks and decipher complex biological systems.

A third area is to study network dynamics. Examples are study on housekeeping (HK) interactions and cell type-specific interactions. These two types of interactions provide further understanding on cell dynamic organizations.

In the following paragraphs, I will first introduce Wiener type indices and a relevant R package QuACN, then bow-tie, hierarchical structures, network motifs, HK and cell type-specific interactions. This sub-section ends with a brief introduction to network analysis and comparison tools.

Wiener type indices

The use of Wiener index and related type of indices dates back to the seminal work of Wiener in 1947 (Wiener, 1947a,b). Wiener introduced his

celebrated index to predict the physical properties, such as boiling point, heats of isomerization and differences in heats of vaporization, of isomers of paraffin by their chemical structures. Viewing the chemical structure of an isomer as a connected graph, the Wiener index is defined as $\sum_{i,j} d(i,j)$ where i, j represent nodes in the graph, $d(i,j)$ the distance between nodes i and j, and the sum is over all pairs of nodes in the graph. Wiener index has since inspired many distance-based descriptors in Chemometrics. These include Harary index (Plavšić et al., 1993), hyper Wiener index (Randić, 1993), q-analog of Wiener index (Zhang et al., 2012b), Wiener polynomial (Hosoya, 1988), Q-index (Brückler et al., 2011), Balaban J index (Balaban, 1982), and information indices (Dehmer, 2008; Dehmer and Mowshowitz, 2011; Dehmer et al., 2009). These indices, or commonly called descriptors, play significant roles in quantitative structure-activity relationship/quantitative structure-property relationship (QSAR/QSPR) models (Todeschini and Consonni, 2009). The definitions of these indices are detailed in chapter 2.

QuACN: an R package for calculating network indices

Mueller et al. (2011a) introduced the R package QuACN, which facilitates the systematic calculation of network indices in a network. QuACN computes the values of different categories of indices in a network. There are 4 categories in this package. (1) Descriptors based on distances in a graph: this class consists of measures using distances to describe the networks structure (e.g. Wiener index, Harary index, etc.). (2) Descriptors based on other graph invariants: the descriptors in this class use other graph invariants other than distances (e.g. degree, number of nodes, number of edges,

etc.). (3) Partition-based graph entropy descriptors: these measures use an arbitrary graph invariant and an equivalence criteria to induce partitions. A probability value is then calculated for each partition to determine the entropy. (4) Parametric graph entropy measures: to determine the entropy measures of this class (Dehmer et al., 2009), by assigning a probability value to each vertex of the network, using the so-called information functionals.

Mueller et al. (2011b) applied a set of indices in QuACN to quantify metabolic networks from three domains of life. Each network is represented by a numeric vector whose elements are the calculated indices. Then three domains of life are classified based on their numeric vectors from their metabolic networks. Their classification results show that these selected indices capture domain-specific structural characteristics of metabolic networks.

Bow-tie structure organization in biological network

A Bow-tie network has a conserved core which interacts with numerous input and output components (Figure 1.4). The three components are connected and each component has global and local feedback regulations. As a result, there are multiple flows of information from input to output through the core (Kitano, 2004).

Bow-tie structure organization is shown to be a common but fundamental organizing principle evidenced by large amount of accumulated biological data (Csete and Doyle, 2004; Li et al., 2012; Nelson et al., 2011). Early evidences for bow-tie structures in biological networks can be found in the review paper by Csete and Doyle (2004). Recently bow-tie structure organization was first found in GRN governing male tail tip morphogene-

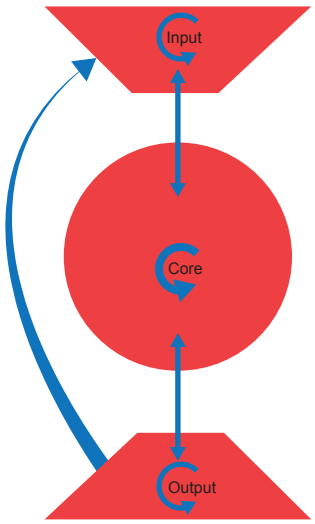


Figure 1.4. A framework for bow-tie structure organization. Red objects stands for input, core and output components. Blue arrows stands for regulation within or between components.

sis in C. elegans (Nelson et al., 2011). Li et al. (2012) also found bow-tie organization with diverse patterns in GRNs of 8 human tissues.

In metabolic networks, the bow-tie structure design is robust. It facilitates control, accommodating perturbations and fluctuations on many timescales and spatial scales. Bow-tie structure has inherent fragilities. A chief source of fragility is that the universal common currencies responsible for robustness can be easily hijacked by parasites or used to amplify pathological processes. Bow-tie structure is also capable to maintain evolvability over multiple timescales. Thus it can be viewed as a starting point for modeling complex biological systems (Csete and Doyle, 2004).

Hierarchical structure of regulatory networks

Hierarchical structure as shown in Figure 1.5 is pervasive in complex systems and is believed to be attributed by functional constraints in GRNs (Corominas-Murtra et al., 2013). Hierarchical structure classifies nodes in

a network into N-layers ($N \geq 3$), i.e., top layer, intermediate layers, and bottom layer. We call the intermediate layer core in a 3-layer hierarchical structure. The regulatory networks for E. coli (Yu and Gerstein, 2006), S. cerevisiae (Jothi et al., 2009; Yu and Gerstein, 2006), worm (Boyle et al., 2014), fly (Boyle et al., 2014), mouse (Bookout et al., 2006) and human (Boyle et al., 2014; Gerstein et al., 2012) exhibit hierarchical organizations. The hierarchical organization of complex networks has been shown to increase adaptabilities and avoid conflicting constraints compared with non-hierarchical networks (Kauffman, 1993).

Most importantly, these hierarchical organizations are associated with TF dynamics (Gerstein et al., 2012; Jothi et al., 2009; Yu and Gerstein, 2006). More specifically, TFs from different layers in one regulatory network exhibit distinct properties. For example, in human regulatory network, Gerstein et al. (2012) showed that the core layer TFs have the highest betweeness and tend to have the most regulatory collaboration among the TFs. Conversely, top layer TFs have more partners in a protein-protein interaction network and a phosphorylome. In yeast regulatory network, Jothi et al. (2009) showed that (1) top and bottom layers are depleted in hubs while the core is enriched with hubs. (2) The percentage of essential proteins in the top layer ($\sim 12\%$) is higher than in the core layer ($\sim 6\%$) and in the bottom layer ($\sim 3\%$). Essential proteins are necessary for performing basic developmental functions. If they are disrupted, they will cause pre- or neonatal lethality (Georgi et al., 2013). (3) Top layer TFs are relatively abundant, had a much longer half-life, and are noisy compared with the core and bottom layer TFs. Here noise of a TF was calculated as the ratio of the standard deviation to its mean abundance. In E. coli and S. cere-

visiae regulatory networks, Yu and Gerstein (2006) showed that (1) TFs in top layers are close to all proteins in a protein-protein interaction network, and they receive most of the input for the whole regulatory hierarchy through protein interactions. Moreover, they have maximal influence over other genes, in terms of affecting expression-level changes. (2) TFs at the lower levels of both networks have a much higher tendency to be essential.

Moreover, regulatory networks across species exhibit different hierarchical structures. Boyle et al. (2014) showed that regulatory network from human have more TFs in top layer than those from worm and fly. Zhang et al. (2014) showed that hierarchical structures of regulatory networks from 41 human cell types are different from that of yeast regulatory network, in terms of distribution of TFs, enrichment of essential proteins, etc.

Furthermore, regulatory networks across human cell types exhibit different hierarchical structures. Zhang et al. (2014) revealed that the hESC TF regulatory network has a topological structure that is different from the rest of the 40 non-hESC networks. (1) It has significantly small top and bottom layers and therefore a large core layer. (2) Its top layer is neither enriched with nor depleted of hub, essential and housekeeping TFs, in contrast to the TF regulatory networks of the 40 differentiated cells.

To classify nodes from a directed network into N layers, a number of deterministic and probabilistic algorithms have been developed by Boyle et al. (2014); Jothi et al. (2009); Yu and Gerstein (2006) and Gerstein et al. (2012). This thesis extensively applied vertex-sort algorithm developed by Jothi et al. (2009) to construct a 3-layer structure for each of these 41 human regulatory networks.

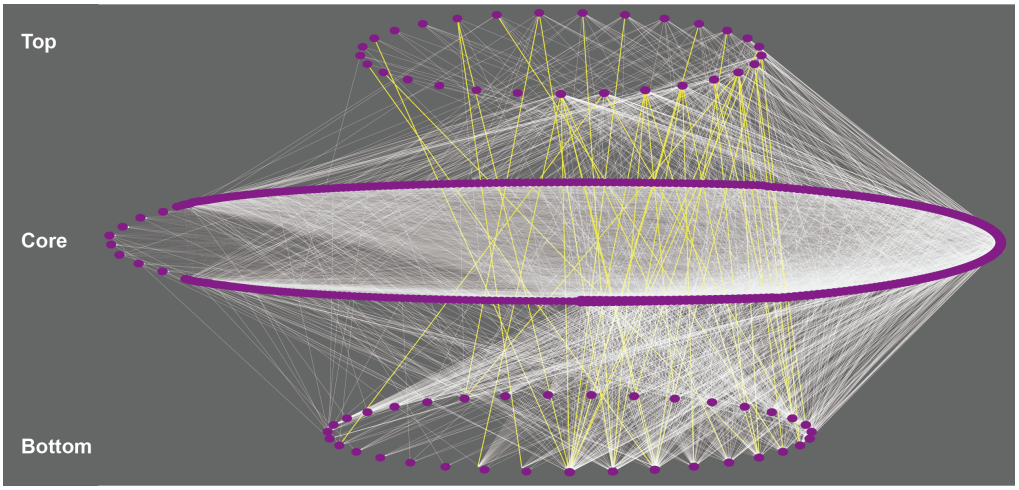


Figure 1.5. A schematic view of three-layer hierarchical structure of the hESC TF regulatory network produced by the vertex-sort algorithm. The TFs are colored red. The links between the top and bottom layers are colored yellow. The other links are in white color. Network data is from Neph et al. (2012a). Network visualization: Cytoscape (version 3.1.0, Kohl et al. (2011)).

Global and local reaching centrality

As mentioned above, hierarchical structure exists in various biological networks across different species, cell types, or cell stages, etc. To quantitatively characterize the level of network hierarchy, Mones et al. (2012) proposed global reaching centrality (GRC) and local reaching centrality (LRC). Given an unweighted directed graph G, LRC of node i, denoted as $LRC(i)$, is defined as proportion of all nodes in the network that can be reached from node i via outgoing interactions. The authors defined GRC based on a heterogeneous distribution of the LRC, where $GRC = \sum_{v \in V(G)} (\max(LRC) - LRC(v))/(N(G) - 1)$. The upper bound of GRC is 1 and is attained when the network is a star. GRC is demonstrated to neatly capture the degree of hierarchy in random networks and real networks. Also it is strongly associated with controllability from real networks in (Mones

et al., 2012).

Network motifs

Network motifs are over-represented connected sub-graph patterns in a real network as compared to a random network (Figure 1.6). For example, to test significance of over-represented connected n-node sub-graphs in a real network, Milo et al. (2002) generated random networks satisfy following two conditions. Each node in randomized networks has the same in-degree and out-degree as the corresponding node in the real network. The number of all $(n-1)$-node sub-graphs are the same as in the real network. Theoretical and experimental evidences demonstrate that network motifs perform dynamic and specific functions in context of the respective networks. Alon (2007) and Shoval and Alon (2010) are two excellent reviews on various motifs and their functions discovered in GRNs from bacteria to plant to human and other types of biological network. Discovery of conserved motifs like Feed-Forward Loop (FFL) in GRNs indicates that motifs are the manifestation of evolutionary design principles favored by selection (Artzy-Randrup et al., 2004). Network motifs are building blocks of complex networks thus are one design principle of complex networks and can control behaviours and states of the corresponding complex systems (Alon, 2007). Figure 1.6 illustrates a number of motifs with 2 to 4 nodes found in various biological networks.

One of the important and extensively studied network motifs is FFL. An FFL, as illustrated in Figure 1.6B, consists of 3 nodes A, B and C in which A regulates B, and both A and B regulate C. FFL in regulatory networks can speed-up the response time of the target gene expression or act as sign-sensitivity delays. FFL can generate pulse of gene expression.

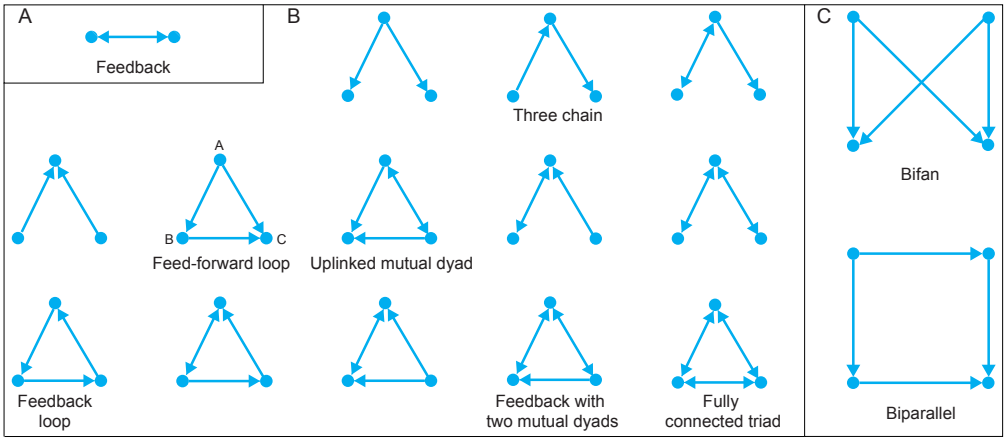


Figure 1.6. Network motifs with 2, 3, and 4 nodes. (A) feedback motif. (B) All 13 types of three-node connected subgraph. (C) Bifan and Biparallel motifs.

FFL can cooperatively enhance induction of gene C by inducers of TF A. Here inducers of A are small molecules, protein partners, or covalent modifications that activate or inhibit the transcription activities of A (Alon, 2007; Shoval and Alon, 2010). Early studies revealed that FFL is over-represented in the regulatory networks of organisms from bacteria and yeast to plants and animals (Alon, 2007). Recent studies show that FFL as a motif is also found in regulatory networks of worm (Boyle et al., 2014), fly (Boyle et al., 2014), human (Boyle et al., 2014; Gerstein et al., 2012; Neph et al., 2012a). Regarding to regulatory networks from embryonic stem cells (ESCs), FFLs is enriched in hESC (Neph et al., 2012a).

Bearing in mind important and dynamic functions played by FFLs and other motifs in various biological networks, some network centrality measures based on network motifs have been proposed to quantify the importance of nodes in directed networks (Harriger et al., 2012; Koschützki and Schreiber, 2008; Koschützki et al., 2007; Sporns et al., 2007; Sporns and Kötter, 2004; Wang et al., 2014). The underlying idea of these centrality

measures is when a node is involved in more motifs, this node is more likely to be functionally important. These centrality measures are named as motif centrality in general and can identify different sets of important nodes in networks partially because they can integrate structural information between local and global information.

Given a directed network G, Koschützki et al. (2007) proposed a few centrality measures. First the authors quantified centrality of a node by the number of FFLs this node is involved in. Then this centrality measure is generalized to as node A (or node B, node C) in an FFL. Furthermore, a path tree is used instead of FFL to define new centrality measures. The proposed centrality measures are comparable with other centrality measures like PageRank, in-degree, and out-degree. They can identify new important nodes partially due to the fact that they are not strongly correlated with previous centrality measures (Koschützki and Schreiber, 2008).

Sporns and Kötter (2004) proposed "network fingerprint" to characterize areas (nodes) in one brain network from Macaque Visual Cortex. Network fingerprint for a node is a vector with length 13, each element is the number of times this node is involved in a particular 3-node connected subgraph. Network fingerprint identified five areas which show highly similar patterns of network fingerprints. In their following up work, Sporns et al. (2007) identified and classified putative hub regions in brain networks based on network fingerprints and other centrality measures. Rich club regions are hub regions that are densely connected than expected based on their degree alone. Harriger et al. (2012) discovered that rich club regions in brain networks tend to form star-like configurations based on application of network motif analysis, which indicate that hubs regions embed within

sets of nodes.

Wang et al. (2014) generalized motif centrality measures by taking into account of 2 to 4 node motifs. For a given network with n nodes, the authors first calculated $B = [b_{ij}]_{1 \leq i \leq n, 1 \leq j \leq m}$, where m is number of types of motifs in the network, $b_{ij} = u_{ij} \times w_j$, u_{ij} equals to the number of $j\text{-}th$ motif involved by node i, $w_j = c_j / \sum_{k=1}^{m} c_k$ where c_k denotes total occurrences of $k\text{-}th$ motif in the network. Then centrality for node i is the $i\text{-}th$ element of first principal component derived from B. The proposed centrality measure can robustly identify functionally important nodes in five biological networks.

Housekeeping and cell type-specific interactions

It is believed that biological systems undergo differential change depending on the environments, tissue types, disease states, development or speciation while part of a system will remain unaffected. Rapid development in technology and experimental designs enables large-scale differential network mapping. Some interactions are observed to appear or disappear dynamically, and many others are observed irrespective of conditions. The latter group of interactions are considered housekeeping interactions. Housekeeping interactions and condition-specific (differential) interactions are proposed to model the two types of interactions and they offer deep biological insights into complex systems (Bolouri, 2014; Ideker and Krogan, 2012; Mitra et al., 2013; Srivas et al., 2013). To name one example, in the study of DNA damage-induced genetic networks in yeast, identified housekeeping interactions in untreated and treated networks are enriched for housekeeping functions, like transcription, translation, chromatin and other cellular housekeeping machinery, whereas identified differential interactions effec-

tively capture DNA damage response genes (Ideker and Krogan, 2012). Neph et al. (2012a) observed that $\sim$ 5% of all interactions are common across the 41 cell types and interactions unique to one cell form a well-connected subnetwork, highlighting regulatory diversity within humans. The 41 networks enable us to examine the concepts of housekeeping interactions and cell type-specific interactions in human TF regulatory networks, in terms of deep topological and functional analysis.

1.3.4 Network analysis and comparison toolsets

Multiple softwares and platforms are developed to provide comprehensive analysis and comparison on real network data sets. To name a few, Bioconductor (http://bioconductor.org), Cytoscape (http://cytoscape.org/, Kohl et al. (2011)), Galaxy (http://galaxyproject.org/, Goecks et al. (2010)), GenePattern (http://www.broadinstitute.org/cancer/software/genepattern), and GenomeSpace (http://www.genomespace.org/). For brief introduction to and comparison on these toolsets, refer to a recent review paper by Bolouri (2014).

1.4 Thesis organization

This thesis is organized as follows. In chapter 2, we first introduce f-Wiener index for a given network G, denoted as $W_f(G)$, which generalizes all existing Wiener type indices. Then we propose a normalized version of $W_f(G)$. In section 2.3, we state our main Theorems 1 to 4, which give sharp bounds of $W_f(G)$ in different classes of networks and trees. We also give a brief description of related works in section 2.4. Then, we consider special cases of f in $W_f(G)$ to provide explicit expressions of the maximum and the mini-

mum of Wiener, Harary, hyper Wiener, generalized Wiener indices. In the experiment section, we report the performance of hierarchical clustering based on the usual Wiener type indices and the normalized version of these in our experiments. Followed is conclusions section. We end chapter 2 with the proofs of Theorems 1 to 4.

In chapter 3, we first present materials and methods in section 3.2. In results section, we classify cell types based on local bipartite patterns constructed by a number TFs and their targets in the 41 TF regulatory networks reported by Neph et al. (2012a) in section 3.3.1. Next, we investigate hierarchical structures of 41 human regulatory networks in section 3.3.2. Then, we report dynamic structures of human regulatory networks in terms of HK interactions and hESC specific interactions in sections 3.3.3 and 3.3.4 respectively. In conclusion section, we summarize our contributions and discussed limitations of the study.

In chapter 4, we present materials and methods in section 4.2. In results section, we first classify cell types based on distributions of FFLs regulating each TF 41 in the TF regulatory networks reported by Neph et al. (2012a) in section 4.3.1. Next we study functions of TFs which are extensively regulated by FFLs in hESC only in section 4.3.3. Then we compare our proposed scores with motif centrality measures in identifying hESC related TFs in section 4.3.5. In conclusion section, we summarize our contributions and discuss limitations. Besides we discuss potential generalization of TFs extensively regulated by FFLs in hESC only.

We end this thesis with Chapter 5 for further discussions and conclusions.

Chapter 2

Sharp Bounds and Normalization of Wiener-type Indices

2.1 Introduction

Recent years witness exponential growth of available biological network data. Thanks to past decades' breakthrough in biotechnology, researchers now are able to interrogate molecular interactions at systems level. It has since been observed that topological properties of these networks provide important insight into the functions of proteins, and their relationship with one another (Barabasi et al., 2011; Delprato, 2012; Hu et al., 2011; Junker and Schreiber, 2008; Milenković et al., 2011; Newman, 2002; Resendis-Antonio et al., 2012; Vidal et al., 2011). For examples, degree distribution, average clustering coefficient, diameter, centrality, lethality and graphlet distribution have been extensively studied. Hopefully, based on a carefully chosen list of network topological properties and methods in quantifying them, a complex network is adequately summarized in the form of a numerical d-dimensional vector where d is the number of topological properties in consideration. This representation enables us to take full advantage of

a host of classification and clustering techniques to compare complex networks.

A significant step towards this direction is facilitated by the introduction of the R package QuACN by Mueller et al. (2011a). QuACN computes the values of different categories of descriptors in a network. One such category is the distance-based descriptors which include Wiener index, Harary index, etc. The use of Wiener index and related type of indices dates back to the seminal work of Wiener in 1947 (Wiener, 1947a,b). Wiener introduced his celebrated index to predict the physical properties, such as boiling point, heats of isomerization and differences in heats of vaporization, of isomers of paraffin by their chemical structures. Viewing the chemical structure of an isomer as a connected graph, the Wiener index is defined as $\sum_{i,j} d(i,j)$ where i, j represent nodes in the graph, $d(i,j)$ the distance between nodes i and j which is defined as the length of a shortest path between them, and the sum is over all pairs of nodes in the graph. Wiener index has since inspired many distance-based descriptors in Chemometrics. These include Harary index (Plavšić et al., 1993), hyper Wiener index (Randić, 1993), q-analog of Wiener index (Zhang et al., 2012b), Wiener polynomial (Hosoya, 1988), Q-index (Brückler et al., 2011), Balaban J index (Balaban, 1982), and information indices (Dehmer, 2008; Dehmer and Mowshowitz, 2011; Dehmer et al., 2009). These indices, or commonly called descriptors, play significant roles in quantitative structure-activity relationship/quantitative structure-property relationship (QSAR/QSPR) models (Todeschini and Consonni, 2009).

It is known that the Wiener type indices depend both on a network's number of nodes and its topology. When the numbers of nodes in the net-

works are equal, as in the applications to isomers, these indices provide informative measures of the branching property of the networks and hence a fair comparison among them. However, when they are used to measure similarities of networks with different numbers of nodes, the intended measure of topological structures will be masked by the sizes of the networks. Normalization of a Wiener type index expectedly minimizes the effect of the network's number of nodes and hence brings forth its topological structure better. Furthermore, it is also desirable for the normalized index to take value in an absolute scale for better understanding and interpretation. This chapter seeks to fill this gap. The normalization introduced in definition 2 below fulfils this purpose. This definition will be of limited practical value if the sharp upper and lower bounds of the index on a graph cannot be found explicitly. The objective of this chapter is three-fold. First, introduce a very general Wiener type index. We call it f-Wiener index, and denote it by $W_f(G)$ for a graph G. This definition includes all known Wiener type indices as special cases. Second, identify the maximum and minimum values of $W_f(G)$ over a class of connected networks G or a class of connected trees G. We are able to derive explicit formulas for these optimal values. Third, propose a normalized version, $W_f^*(G)$ which takes value in $[0, 1]$ for better interpretation and network comparison.

This chapter is organized as follows. We first introduce some standard graph-theoretic notations and recall some special graphs. We then introduce the functional analog of Wiener index, $W_f(G)$, and our proposed normalized versions of this functional Wiener index in section 2.2.1. In section 2.3, we provide our main results Theorems 1 to 4. Theorem 1 gives the maximum and the minimum of $W_f(G)$ over the set of connected graphs

of n nodes, and characterization of graphs achieving the maximum or the minimum. Theorem 2 gives a parallel result when the maximum and minimum are taken over the set of connected trees of n nodes. Theorem 3, (respectively Theorem 4) identifies the maximum of $W_f(G)$ over the set of connected graphs (respectively connected trees) of n nodes with specified maximum degree. We also give a brief description of related works in next section. Then, we consider special cases of f in $W_f(G)$ to provide explicit expressions of the maximum and the minimum of Wiener, Harary, hyper Wiener, generalized Wiener indices. In the experiment section, we report the performance of hierarchical clustering based on the usual Wiener type indices and the normalized version of these in our experiments. Followed with conclusions section. We end with the proofs of Theorems 1 to 4 of this chapter.

2.2 Methods

2.2.1 Definitions and terminologies

Let $G = (V, E)$ be a simple (that is, no self-loops nor multiple edges) connected graph on n nodes where $V = \{1, \ldots, n\}$ and $E \subseteq V \times V$. Denote by $N(G)$ as the number of nodes in G. Let $\mathcal{G}_n$ denote the set of all simple, connected graphs with n nodes. A graph having no cycles is called a tree, and we let $\mathcal{T}_n$ denote the set of all connected trees with n nodes. The distance $d(i, j)$ between any pair of nodes, i and j, in G is the number of edges in a shortest path from i to j. Let $D(G) = [d(i, j)]_{1 \leq i,j \leq n}$ be the distance matrix. We denote the maximum degree of G by $\Delta(G)$.

Figure 2.1 shows some special graphs we frequently refer to in this

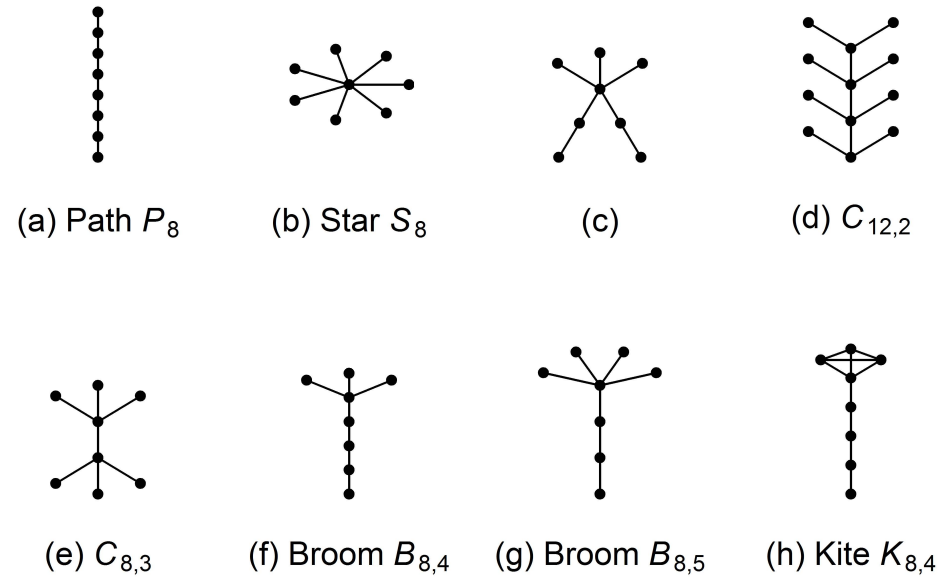


Figure 2.1. Some special graphs. Figure 2.1 (a) to (g) are trees.

chapter. A path graph, P_n, is a graph that can be drawn so that all of its vertices and edges lie on a straight line. Figure 2.1(a) shows P_8. A star, S_n, is a tree with one internal node and $n-1$ leaves. S_8 is shown in Figure 2.1(b). A complete graph, K_n, is a graph with n nodes in which every pair of distinct nodes is connected by an edge. A caterpillar, $C_{n,k}$, is a tree with a central path with number of nodes $\in [n/(k+1),\ (n+k)/(k+1)]$ where at most one end node of the central path has less than k leaves, each of the other nodes in the central path has k leaves. Figures 2.1(d) and 2.1 (e) show caterpillars $C_{12,2}$ and $C_{8,3}$ respectively. A broom $B_{n,k}$ is a tree joining a star S_{k+1} and a path P_{n-k-1} by attaching a pendant node (or leaf) in P_{n-k-1} to a pendant node of S_{k+1}. For examples, brooms $B_{8,4}$ and $B_{8,5}$ are shown in Figures 2.1(f) and 2.1(g) respectively. A kite $K_{n,\ell}$ is a graph obtained from connecting two end nodes one from a complete graph K_ℓ

and one from a path $P_{n-\ell}$. Figure 2.1(h) shows a kite $K_{8,4}$.

Throughout this chapter, f denotes a monotone function defined on nonnegative integers. We define a functional-analog Wiener index below. Our definition contains the Wiener index, Harary index, hyper Wiener index, compactness, average efficiency, generalized Wiener index, Wiener polynomial, Q-index, q-analogy of Wiener index as special cases. For detail, see section 2.4.1. We abbreviate it as f-Wiener index. This definition has also been independently introduced by Schmuck et al. (2012).

Definition 1. *The f-Wiener index of $G \in \mathcal{G}_n$ is defined by*

$$W_f(G) = \sum_{1 \le i < j \le n} f(d(i,j)).$$

Here $d(i,j)$ denotes the shortest distance between nodes i and j.

The number of nodes of G has a very strong effect on Wiener type indices (Section 2.2.2). In order to apply f-Wiener index for comparing networks, which often differ in the numbers of nodes, we are led to propose a normalized version for graphs and a normalized version for trees for better interpretation of the index.

Definition 2. *(a) The normalized f-Wiener index for a graph $G \in \mathcal{G}_n$ is defined as*

$$W_f^\star(G) = \frac{M_f - W_f(G)}{M_f - m_f}.$$

Here $M_f = \max_{H \in \mathcal{G}_n}\{W_f(H)\}$ and $m_f = \min_{H \in \mathcal{G}_n}\{W_f(H)\}$.

(b) The normalized f-Wiener index for a tree $T \in \mathcal{T}_n$ is similarly defined where the maximum M_f and the minimum m_f are taken over $\mathcal{T}_n$ instead.

These normalized versions will be of limited practical value if one cannot compute M_f nor m_f. Our main results, stated in Theorems 1 and 2, show that these optimal upper and lower bounds can be easily computed. Moreover, they characterize those graphs which attain the maximum or the minimum.

By definition, $W_f^*(G)$ takes values in $[0, 1]$. When f is a non-decreasing function, Theorem 1 below shows that $W_f^*(G) = 0$ if and only if G is a path graph, and $W_f^*(G) = 1$ if and only if G is a complete graph. So $W_f^*(G) \approx 0$ (respectively, $W_f^*(G) \approx 1$) suggests G looks like a path graph (respectively, a complete graph). And hence the numerical value of $W_f^*(G)$ provides an indication how G is like.

2.2.2 Effect of number of nodes on Wiener type indices

It is known that the Wiener index for a connected graph with n nodes ranges from $n(n-1)/2$ to $n(n-1)(n+1)/6$ (see Corollary 5 below or Dobrynin et al. (2001); Soltés (1991), and Gutman et al. (1997)). This wide range can be undesirable if it is used for comparing similarity of graphs with different number of nodes. For example, consider two path graphs, P_4 and P_5, with 4 nodes and 5 nodes respectively, and a star graph with 5 nodes, S_5. Values of the Wiener index for P_4, P_5 and S_5 are respectively 10, 20 and 16, giving the false impression that P_5 and S_5 are more similar than that of P_4 and P_5. However, values of the normalized Wiener index are 0 for P_4 and P_5, and 1 for S_5. This example is far from being an isolated case, it can be shown that if the number of nodes of a path graph is at least 26% more than the number of nodes in another path graph, there exists a star graph whose Wiener index is closer to that of the path graph with smaller

number of nodes.

The normalized Wiener index of S_n, star with n nodes, is $1 - 3/n$, suggesting stars of sufficiently large n, based on the normalized Wiener index, S_n is very similar to a complete graph. This is concordant with the fact that a K_n is the line graph of S_{n+1} (Resendis-Antonio et al., 1931).

2.2.3 Main idea

A key ingredient in our proofs is a matrix majorization (see section 2.8 for definition) argument. Given a connected graph G, we can transform it to another graph G' such that the distance matrix of G, $D(G) = [d(i,j)]_{1\leq i,j\leq n}$ majorizes the corresponding distance matrix of G'. Since the Wiener index of G, or its generalization f-Wiener index for increasing function f, is the sum of the upper diagonal entries in the distance matrix, it follows that $W_f(G) \geq W_f(G')$. The construction of G' is fairly straightforward as can be seen in the proofs. Similarly, we can transform G to another graph G'' such that $D(G)$ is majorized by $D(G'')$. And thus $W_f(G) \leq W_f(G'')$. The construction of G'' requires delicate and judicious pruning and regrafting. However, the essential idea remains the same. Technical details of proofs are given in section 2.8.

2.3 Results

We provide explicit expressions for the maximum and minimum of $W_f(G)$ over $\mathcal{G}_n$, and over $\mathcal{T}_n$ in Theorems 1 and 2 below. We also characterize those graphs or trees attaining the extremum. Theorems 3 and 4 concern trees or graphs with a specified maximum degree. For simplicity of presentations, we shall only state our results for non-decreasing function f. Analogous

results for non-increasing f can be deduced easily by replacing f by $-f$.

Theorem 1. *Let f be a non-decreasing function on nonnegative integers, and $G \in \mathcal{G}_n$, then*

$$\frac{n(n-1)}{2} f(1) \leq W_f(G) \leq \sum_{i=1}^{n-1} (n-i) f(i).$$

The lower bound is attained if and only if G is K_n. The upper bound is attained if and only if G is P_n.

Theorem 2. *Let f be a non-decreasing function on nonnegative integers, and $T \in \mathcal{T}_n$, then*

$$\frac{(n-1)((n-2)f(2)+2f(1))}{2} \leq W_f(T) \leq \sum_{i=1}^{n-1} (n-i) f(i).$$

The lower bound is attained if and only if T is S_n. The upper bound is attained if and only if T is P_n.

Theorem 3. *Let f be a non-decreasing function on nonnegative integers. Then, for any $T \in \mathcal{T}_n$ with $\Delta(T) = k$, we have*

$$W_f(T) \leq W_f(B_{n,k+1}).$$

The upper bound is attained if and only if T is a broom $B_{n,k+1}$.

Theorem 4. *Let f be a non-decreasing function on nonnegative integers. Then, for any $G \in \mathcal{G}_n$ with $\Delta(G) = k$, we have*

$$W_f(G) \leq W_f(B_{n,k+1}).$$

Moreover,

$$W_f(B_{n,k+1}) = \sum_{j=1}^{n-k+1} (n-j)f(j) + \frac{(k-1)(k-2)}{2} f(2).$$

Equality holds if and only if G *is* $B_{n,k+1}$.

2.4 Related work

The proofs of Theorems 1 to 4 will be given in section 2.8. Theorem 2 has also been independently obtained by Wagner et al. (see Theorem 2.7 and Corollary 4.1 in Wagner et al. (2013)). Special cases of Theorems 1 to 4 for particular Wiener type index are known in the literature. For examples, the complete graph (respectively, the path graph) is shown to be the minimizer (respectively, maximizer) of the Wiener index among simple connected graphs with the same number of nodes in Dobrynin et al. (2001); Soltés (1991), and Gutman et al. (1997). Similar conclusions are proved to hold for the hyper Wiener index in Gutman et al. (1997), and the Harary index in Gutman (1997). The results in Theorems 1 to 4 in its full generality as f-Wiener index are novel to the best of our knowledge. Moreover, we have provided a unifying methodology for the proofs.

2.4.1 Important special cases

Since its introduction, Wiener index has inspired many variants and thoroughly studied in a sizeable literature (Todeschini and Consonni, 2009). By choosing appropriate functions f, the f-Wiener index can be reduced to a number of commonly used descriptors as follows.

If we take $f(k) = k$, $W_f(G)$ written as $W(G)$ is the well-studied de-

scriptor introduced by Wiener in 1947 (Wiener, 1947a,b).

Taking $f(k) = 1/k$, the f-Wiener index is the Harary index (Plavšić et al., 1993), denoted by $H(G)$ which is shown to be more discriminating than the Wiener index (Plavšić et al., 1993). Watts and Strogatz (1998) used a scaled version of the Harary index (more precisely, $f(k) = \frac{2}{n(n-1)k}$) to measure a network's efficiency in information exchange.

Taking $f(k) = k^\alpha$, where α can be positive or negative, the f-Wiener index is called generalized Wiener index, denoted by $W_\alpha(G)$ (Gutman et al., 1998).

If $f(k) = (k^2 + k)/2$, the f-Wiener index is known as the hyper Wiener index (Randić, 1993), denoted by $WW(G)$.

Taking $f(k) = \lambda^k$, where λ is regarded as a parameter, the f-Wiener index is called the Hosoya polynomial or Wiener polynomial (Hosoya, 1988). With an additional factor 2, the Hosoya polynomial is called Q-index and denoted by $Q(\lambda)$ in Brückler et al. (2011).

The q-analog of the Wiener index, introduced by Zhang et al. (2012c) is simply the f-Wiener index by choosing $f(k) = (1 - q^k)/(1 - q) = \sum_{t=0}^{k-1} q^t$.

2.5 Applications

By specializing f to various forms in Theorems 1 and 2, we provide below explicit sharp upper bounds and sharp lower bounds for the Wiener index $W(G)$, the Harary index $H(G)$, the hyper Wiener index $WW(G)$, and the generalized Wiener index $W_\alpha(G)$ for $\alpha > 0$ and $\alpha < 0$.

Corollary 5. *Let G be a simple, connected graph with n nodes (that is,*

$G \in \mathcal{G}_n$), we have

$$\frac{n(n-1)}{2} \leq W(G) \leq \frac{n(n-1)(n+1)}{6},$$
$$n\sum_{i=2}^{n-1}\frac{1}{i} + 1 \leq H(G) \leq \frac{n(n-1)}{2},$$
$$\frac{n(n-1)}{2} \leq WW(G) \leq \frac{n(n-1)(n+1)(n+2)}{24},$$

when $\alpha < 0$,

$$n\sum_{i=1}^{n-1} i^{\alpha} - \sum_{i=1}^{n-1} i^{\alpha+1} \leq W_{\alpha}(G) \leq \frac{n(n-1)}{2},$$

when $\alpha > 0$,

$$\frac{n(n-1)}{2} \leq W_{\alpha}(G) \leq n\sum_{i=1}^{n-1} i^{\alpha} - \sum_{i=1}^{n-1} i^{\alpha+1}.$$

Corollary 6. *Let T be a tree with n nodes (that is, $T \in \mathcal{T}_n$), we have*

$$(n-1)^2 \leq W(T) \leq \frac{n(n-1)(n+1)}{6},$$
$$n\sum_{i=2}^{n-1}\frac{1}{i} + 1 \leq H(T) \leq \frac{(n-1)(n+2)}{4},$$
$$\frac{(n-1)(3n-4)}{2} \leq WW(T) \leq \frac{n(n-1)(n+1)(n+2)}{24},$$

when $\alpha < 0$,

$$n\sum_{i=1}^{n-1} i^{\alpha} - \sum_{i=1}^{n-1} i^{\alpha+1} \leq W_{\alpha}(T) \leq ((n-2)2^{\alpha-1}+1)(n-1),$$

when $\alpha > 0$,

$$((n-2)2^{\alpha-1}+1)(n-1) \leq W_{\alpha}(T) \leq n\sum_{i=1}^{n-1} i^{\alpha} - \sum_{i=1}^{n-1} i^{\alpha+1}.$$

2.6 Experiments

We describe below three experiments to compare the hierarchical clustering using normalized f-Wiener indices with the hierarchical clustering us-

ing non-normalized f-Wiener indices. Each experiments consists of 3 main steps.

Step 1: A collection of networks (or graphs) or trees, $\mathcal{C}$, are chosen to be clustered. The collection is detailed in each experiment below.

Step 2: Seven functions are chosen to form the f-Wiener indices. In all our experiments, we choose

$$f_1(k) = \sqrt{k},\ f_2(k) = k,\ f_3(k) = \frac{k + k^2}{2},$$

and

$$f_4(k) = \frac{4k}{N(G)(N(G) - 1)},$$

$$f_5(k) = k^{-1/2},\ f_6(k) = k^{-1},\ f_7(k) = k^{-2}.$$

The first four functions chosen are increasing and the f-Wiener indices correspond to the usual $W_{1/2}$ index, Wiener index, the hyper Wiener index and the compactness index. The remaining 3 functions chosen are decreasing and correspond to the $W_{-1/2}$ index, the Harary index and the W_{-2} index. Hopefully these indices collectively capture some essential characters of networks and useful for clustering. For $G \in \mathcal{C}$, we construct two characteristic vectors,

$$v_G = (W_{f_1}(G), \ldots, W_{f_7}(G)),$$

$$v_G^\star = (W_{f_1}^\star(G), \ldots, W_{f_7}^\star(G)).$$

Step 3: We adopt a clustering algorithm to cluster $\mathcal{C}$ using v_G and then produce a dendrogram. We do the same using v_G^*. Minimum variance method algorithm due to Ward (Ward, 1963) which is made available in R

base package, was used in all the experiments. The computed the Adjusted Rand Index (ARI) in all the experiments are summarized in Table 2.1 below.

2.6.1 Experiment 1: Hierarchical clustering of random networks

The collection of networks chosen for this experiment is the networks generated by some commonly used random network models, namely, Erdos-Renyi (ER) model (Erdős and Rényi, 1959, 1960), scale-free (SF) network model (Barabasi and Albert, 1999) and 3-D geometric model (GE) (Pržulj et al., 2004). Each of these random network models is applied to generate 10 random networks with the number of nodes ranging from 500 to 950 with step of increment 50. Experiment 1 consists of 5 small, but similar, experiments. We enumerate these 5 small experiments as 1.1,. . . , 1.5. The subsection after experiments provides more details on how to generate these random networks. We then apply Steps 2 and 3 above to form two dendrograms: one using f-Wiener indices without normalization (Figure 2.2A) and the other dendrogram using normalized f-Wiener indices (Figure 2.2B). To quantify the classification of the two methods: with and without normalization, we adopt the commonly used Adjusted Rand Index (ARI) (Rand, 1971) for classification validation. ARI measures the accuracy of classification, and takes values between -1 and 1. The larger the ARI is, the better is the classification. The ARI for Figures 2.2A and 2.2B are respectively 0.18 and 0.58 for Experiment 1.5. Using normalized f-Wiener indices lead to a substantial improvement in the classification. We repeat Experiments 1.1 to 1.5 1000 times each. The boxplots of the ARI are shown in Figure 2.3. The means and standard deviations for these experiments are given in

Table 2.1. They clearly demonstrate the superiority of classification using normalized f-Wiener indices.

Table 2.1. Adjusted Rand Index (ARI) for clustering (or classification) of networks in our three experiments. For experiments 1.1 to 1.5, we report the mean and the standard deviation (number in parenthesis) of ARI. Mean and standard deviation of ARI for experiments 1.1 to 1.5 under random clustering are 0 and 0.05 respectively.

	Non-normalized	Normalized
Experiment 1.1	0.44 (0.02)	0.88 (0.07)
Experiment 1.2	0.41 (0.06)	1.00 (0.01)
Experiment 1.3	0.38 (0.10)	1.00 (0.00)
Experiment 1.4	0.36 (0.11)	0.97 (0.10)
Experiment 1.5	0.30 (0.12)	0.62 (0.07)
Experiment 2	0.10	1.00
Experiment 3	0.04	0.86

2.6.2 Experiment 2: Hierarchical clustering of trees

The collection of trees to be classified consists of 10 paths (P_n), 10 stars (S_n), 10 brooms ($B_{n,\frac{n}{2}}$), 20 caterpillars ($C_{n,2}$ which is like a path, and $C_{n,\frac{n-10}{10}}$ which is like a star), and for n ranging from 500 to 950 with step of increment 50.

Figure 2.4 shows the two dendrograms. The ARI for Figures 2.4A and 2.4B are respectively 0.10 and 1.00. This demonstrates that using normalized f-Wiener indices provides much better accuracy for classification purposes. The result in this experiment is consistent with that of experiment 1.

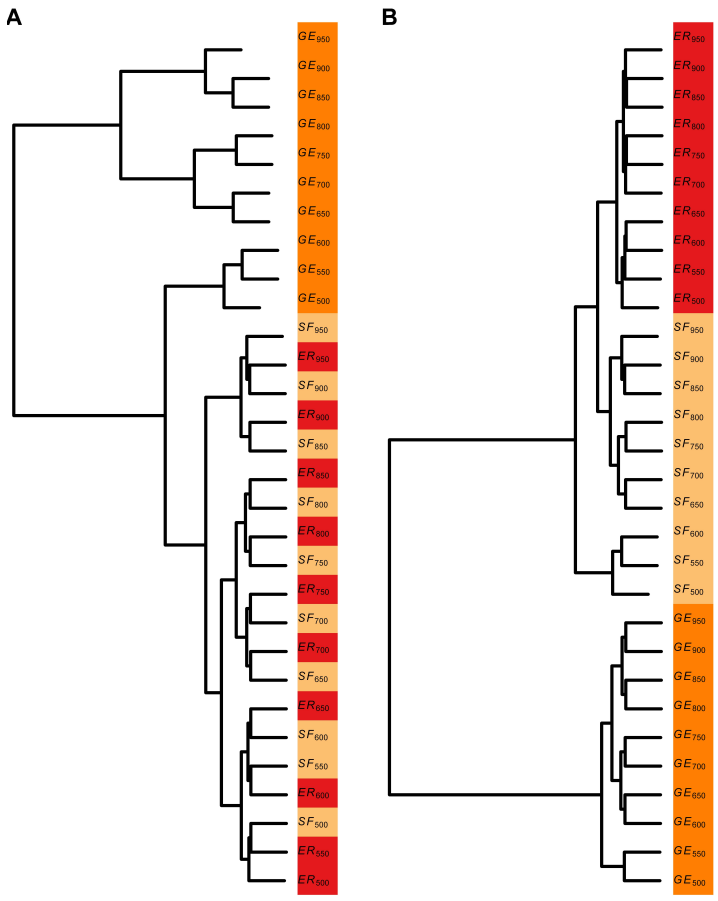


Figure 2.2. Hierarchical clustering of random networks. 30 networks with 10 each generated by the Erdos-Renyi (ER), scale-free (SF) and geometric (GE) random network models. Panel (A) shows the hierarchical clustering based on the f-Wiener indices (see Step 1 on page 35 for functions used). The adjusted rand index (ARI) for this clustering is 0.24. Panel (B) is the hierarchical clustering based on the normalized versions of the same f-Wiener indices. The ARI of this clustering is 0.67. Number of nodes chosen are 500, 550, ... , 950, and p is 0.05 in the Erdos-Renyi model. A scale-free network with 500 nodes is denoted by SF_{500}. The others are denoted in a similar way.

2.6.3 Experiment 3: Hierarchical clustering of random networks and trees

The collection of networks consists of (i) networks generated by three random network models, namely, ER model, SF Model and 3-D geometric

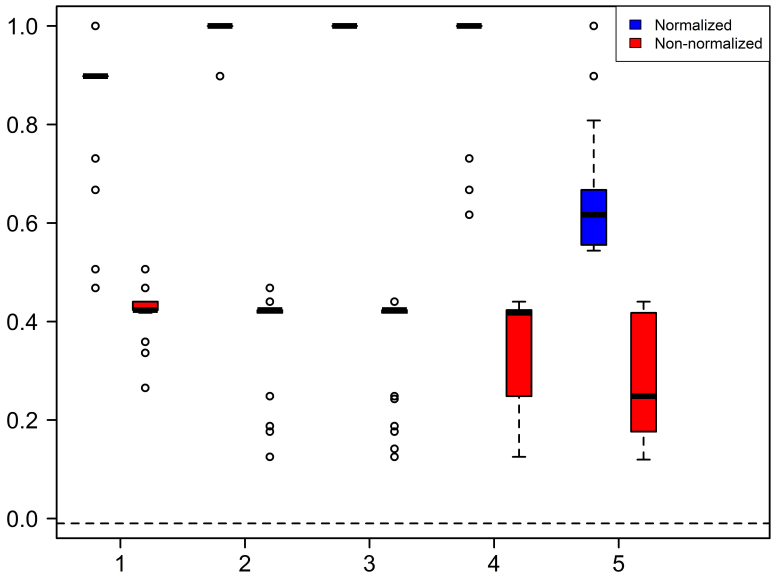


Figure 2.3. Boxplots of Adjusted Rand Index for measuring the extent of agreement of clustering of the random networks using non-normalized f-Wiener indices versus normalized f-Wiener indices.

model; (ii) some trees such as paths, brooms, caterpillars, stars. Figure 2.5 shows the two dendrograms formed. And the ARI for Figures 2.5A and 2.5B are respectively 0.04 and 0.86.

2.6.4 Details on generating random networks

We describe here in details on how to choose the networks generated by the three random network models in experiments 1 and 3.

ER model

There are two parameters in the ER model, namely, n, the number of nodes, and p, the probability that an edge is formed between a pair of nodes. All edges are formed independently of each other. In Experiment

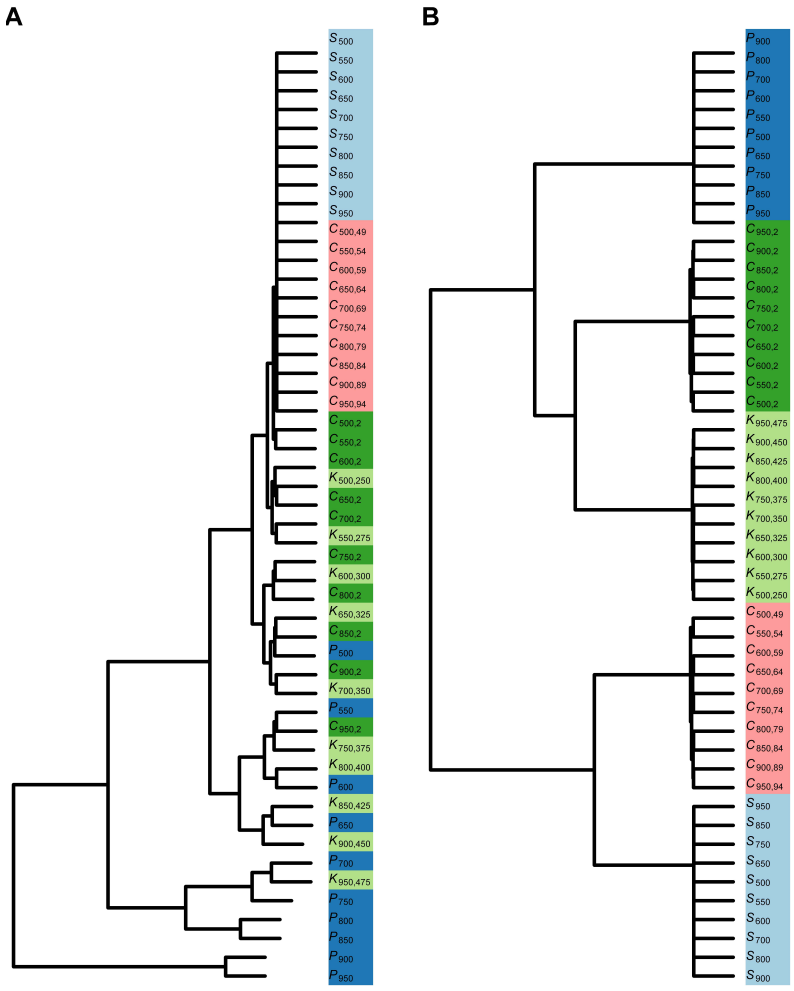


Figure 2.4. Hierarchical clustering of trees. Panel (A) shows the hierarchical clustering based on the f-Wiener indices (see Step 1 on page 6 for functions used). The Adjusted Rand Index (ARI) is 0.1. Panel (B) shows the hierarchical clustering based on normalized f-Wiener indices. The ARI is 1. Trees used in the clustering consist of paths (P_n), stars (S_n), caterpillar-like trees ($C_{n,k}$), kites ($K_{n,k}$). Number of nodes $n = 500, 550, ..., 950$.

1.5, where $p = 0.05$, we choose n ranging from 500 to 950 with step of increment 50. We generate an ER network using the 'erdos.renyi.game' function available in the R package igraph (Csardi and Nepusz, 2006). If the network is connected, we keep it in $\mathcal{C}$ and denote it as ER_{500}. If not, then we repeat the function 'erdos.renyi.game' until a connected network

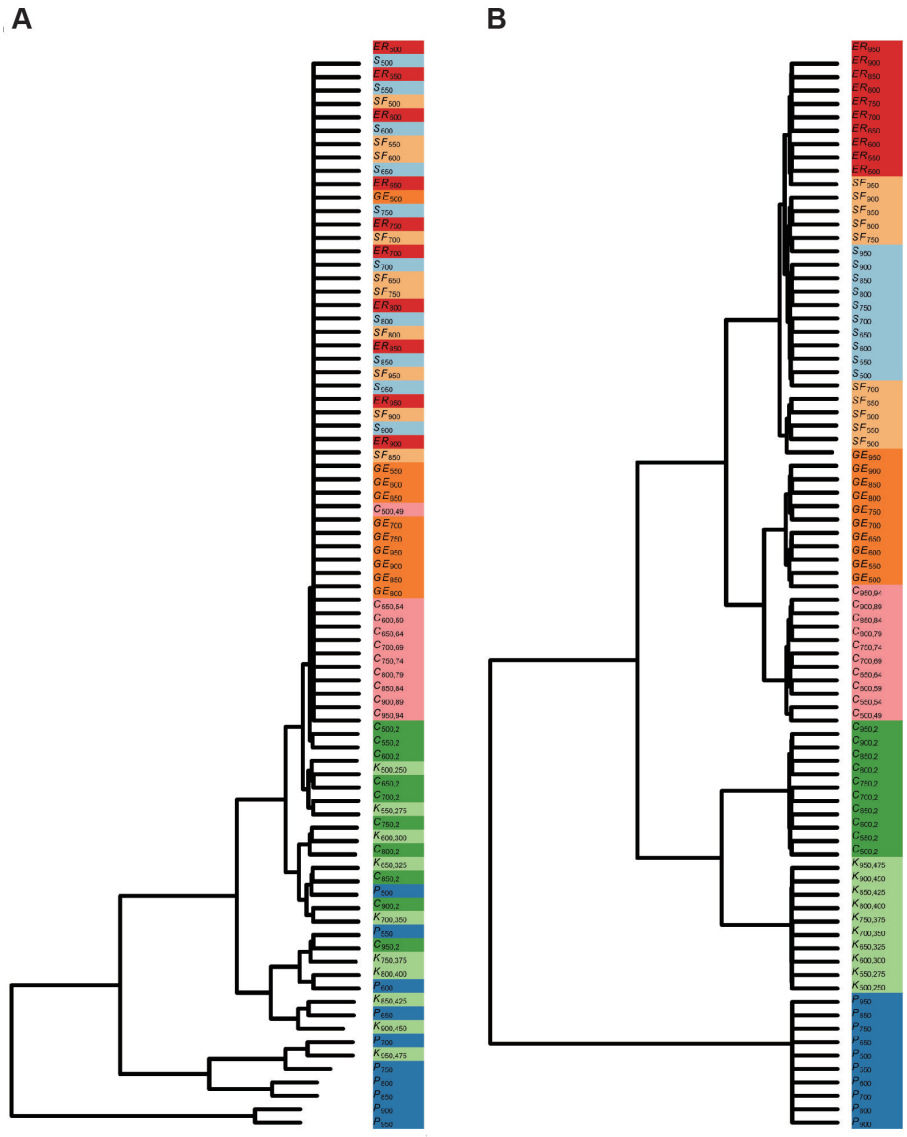


Figure 2.5. Hierarchical clusters of trees and graphs. Panel (A) shows the hierarchical clustering based on the f-Wiener indices (see Step 1 on page 6 for functions used). The Adjusted Rand Index (ARI) is 0.04. Panel (B) shows the hierarchical clustering based on normalized f-Wiener indices, and ARI = 0.86. Trees used are paths (P_n), stars (S_n), caterpillar-like trees ($C_{n,k}$), kites ($K_{n,k}$). Graphs are generated by Erdos-Renyi (ER_n), scale-free (SF_n) and geometric (GE_n) random network models. The parameter, p, in the Erods-Renyi random graph equals to 0.05, number of nodes n = 500,550,...,950.

is obtained. Similarly, $ER_{550}, \ldots, ER_{950}$ are generated.

SF model

We also construct ten SF networks by the function 'barabasi.game' available in the R igraph package. We shall describe how to grow a SF network with 500 nodes for a given p, say $p = 0.05$. The other 9 SF networks with $550, \ldots, 950$ nodes are constructed in a similar manner. In 'barabasi.game' function, we set number of vertices 500, number of edges to be added in each time step $500 \times 0.05/2$ rounded to the nearest integer, and the option to create a directed graph false.

Geometric model

We generate ten 3-D geometric networks with $500, 550, \ldots, 950$ nodes. We shall describe how to construct one with 500 nodes as follows. The rest are constructed similarly. We first place 500 nodes in a unit cube uniformly and independently, then we compute all the $\binom{500}{2}$ pairwise distances and rank these distances in ascending order. We choose the top $100p\%$ of these pairwise distances and connect their corresponding nodes. If this network is connected, then we keep it in $\mathcal{C}$ and denote it by GE_{500}. Otherwise, we discard it, and repeat the above procedure until we get a connected network. The other networks $GE_{550}, \ldots, GE_{950}$ are constructed similarly.

2.7 Conclusions

Wiener index and other Wiener type indices have been commonly applied in Chemometrics to associate structures and physicochemical properties of molecules. Recently, these indices are incorporated in quantifying complex

networks as in QuACN (Mueller et al., 2011a) and NetCAD (Ren and Liu, 2013). In this chapter, we first generalize Wiener index to a general functional form, called f-Wiener index. This f-Wiener index contains all well-known Wiener type indices as special cases such as Wiener index, Harary index, hyper Wiener index, compactness, and average efficiency. We provide a unifying method to identify the maximum and minimum over the set of simple connected graphs with n nodes, or the set of simple connected trees with n nodes (Theorems 1 and 2). Explicit sharp upper and lower bounds for Wiener index, Harary index, hyper Wiener index and the generalized index are deduced over networks (Corollary 5) and over trees (Corollary 6). Moreover, the maximizer and minimizer are characterized in Theorems 1 and 2. We believe these results are general and of independent interests.

Armed with these maximum and minimum values, we propose a normalized version of f-Wiener index over networks, and a similar version over trees. These normalized versions provide better interpretation of indices over networks of varying number of nodes than the non-normalized one. We conduct a number of experiments to compare the clustering performance using normalized f-Wiener indices with that of the non-normalized f-Wiener indices. The results of these experiments consistently demonstrate that using normalized versions improved clustering substantially. The normalized versions capture similar topological structures among networks with different number of nodes better. Our method of optimizing $W_f(G)$ can be easily extended to index of the form $\Phi(W_f(G))$ where Φ and f are monotone functions. For example, taking $\Phi(x) = 1/x$ and $f(k) = \frac{2}{n(n-1)k}$ leads to $\Phi(W_f(G)) = \frac{n(n-1)}{2\sum_{i<j} 1/d(i,j)}$ which measures small-world behvaior of

network G (Newman, 2002). For other descriptors, it is of interest to study whether normalization is needed; if so, how best to normalize them; and to what extent normalization improve network comparison.

Observe that $W_f(G) = \sum_{r=1}^{\infty} f(r)n_r(G) = \sum_{r=0}^{\infty}[f(r+1) - f(r)]N_r(G)$ where we assume $f(0) = 0$, $n_r(G)$ denotes the number of pairs of nodes in G with distance equals r, and $N_r(G)$ the number of pairs of nodes in G with distance greater than r. Since in most biological networks the number of nodes is large, one may normalize a scaled-version of $W_f(G)$ in terms of the asymptotic distribution of the N_r's under the assumption that the observed network G is generated by a given random network model $\mathcal{M}$. This will enable us to determine the likelihood that the observed network is generated by $\mathcal{M}$. Currently a fair amount of information about shortest paths in some network models is available in Barbour and Reinert (2011) and Fronczak et al. (2004). How to make use of these results seems like a worthwhile future project.

2.8 Proofs for Theorems 1-4

We describe here detailed proofs of Theorems 1-4. We start with some definitions and three Lemmas.

A matrix $A = [a_{ij}]_{1\le i,j\le n}$ is majorized by matrix $B = [b_{ij}]_{1\le i,j\le n}$, denoted by $A \preccurlyeq B$ or $B \succcurlyeq A$ if and only if

$$a_{(i)} \le b_{(i)} \quad \text{for } 1 \le i \le n \times n,$$

where $a_{(i)}$ and $b_{(i)}$ are the i-th smallest elements in A and B. A is strictly

majorized by B, denoted by $A \prec B$ or $B \succ A$ if and only if

$$a_{(i)} \leq b_{(i)} \quad \text{for } 1 \leq i \leq n \times n$$

and

$$a_{(i)} < b_{(i)} \quad \text{for some } i.$$

Matrices A and B are said to be equivalent, denoted by $A \equiv B$ if and only if

$$a_{(i)} = b_{(i)} \quad \text{for } 1 \leq i \leq n \times n.$$

Majorization, strict majorization, and equivalent between two vectors $A = (a_i)_{1 \leq i \leq n}$ and $B = (b_i)_{1 \leq i \leq n}$ are defined similarly.

Let G be a graph, define $V(G)$ as the set of nodes in G, and $E(G)$ as the set of edges in G. Let $\deg_G(u)$ denote the degree of node u in graph G. When there is no risk of ambiguity which graph G we are considering, we abbreviate $\deg_G(u)$ to $\deg(u)$. Define $ne(u) = \{v \in V(G) : (u, v) \in E(G)\}$ and call it neighborhood of node u. A node of degree 1 is called a pendant node or a leaf. A node which is not a pendant node is called an internal node. It is known that a path tree is the only tree on n nodes with maximal degree 2. Only tree on n nodes with maximal degree $n - 1$ is a star tree.

A tree is called a starlike tree if it has exactly one node of degree greater than two. Figures 2.1(c), (f), and (g) show 8-node starlike trees with maximum degree equal to 5, 4, and 5 respectively.

Lemma 1. *Let T be a connected tree, u_1 a pendant node and u_2 an internal node. Suppose all nodes, if there is any, in the shortest path connecting u_1*

and u_2 are of degree 2. Then

$$(d(u_2, v))_{v \in V(T)} \prec (d(u_1, v))_{v \in V(T)}.$$

Proof. Let P_{u_1,u_2} denote the path connecting u_1 with u_2. For $v \in V(T) \backslash V(P_{u_1,u_2})$

$$\begin{aligned} d(u_1, v) &= d(u_1, u_2) + d(u_2, v) \\ &> d(u_2, v). \end{aligned}$$

And

$$(d(u_1, v))_{v \in V(P_{u_1,u_2})} \equiv (d(u_2, v))_{v \in V(P_{u_1,u_2})}.$$

Thus

$$(d(u_2, v))_{v \in V(T)} \prec (d(u_1, v))_{v \in V(T)}.$$

□

Lemma 2. *Consider two distinct trees T_1 and T_2. Let $u_1, u_2 \in V(T_1)$ with u_1 of degree at least 2 and u_2 a pendant node satisfying the property that any node, if there is any, on the shortest path connecting u_1 and u_2 is of degree 2. Let $u_3 \in V(T_2)$. A new tree T is constructed by connecting u_1 and u_3, and T' is constructed by connecting u_2 and u_3. Then,*

$$D(T) \prec D(T').$$

Proof. Observe that

$$\begin{aligned} (d(v_1, v_2))_{v_1,v_2 \in V(T_1)} &\equiv (d'(v_1, v_2))_{v_1,v_2 \in V(T_1)}, \\ (d(v_1, v_2))_{v_1,v_2 \in V(T_2)} &\equiv (d'(v_1, v_2))_{v_1,v_2 \in V(T_2)}. \end{aligned}$$

For $v_1 \in V(T_2)$, we have

$$\begin{aligned} & (d'(v_1, v_2))_{v_2 \in V(T_1)} \\ \equiv \; & d'(v_1, u_3) + 1 + (d'(u_2, v_2))_{v_2 \in V(T_1)} \\ \equiv \; & d(v_1, u_3) + 1 + (d(u_2, v_2))_{v_2 \in V(T_1)} \end{aligned}$$

and

$$\begin{aligned} & (d(v_1, v_2))_{v_2 \in V(T_1)} \\ \equiv \; & d(v_1, u_3) + 1 + (d(u_1, v_2))_{v_2 \in V(T_1)} \\ \prec \; & (d'(v_1, v_2))_{v_2 \in V(T_1)}. \end{aligned}$$

Thus $D(T) \prec D(T')$. □

Manipulations in Lemma 2 are illustrated in Figure 2.6.

Starting from a tree T with m number of nodes with maximum degree $\Delta(T)$. If $m \geq 2$, Lemma 2 can be iteratively applied to construct a tree T' such that the maximum degree is equal to that of T but the number of nodes in T' with the maximum degree is reduced by 1. If $m = 1$, then Lemma 2 can also be iteratively applied to construct a tree T' with maximum degree $\Delta(T') = \Delta(T) - 1$.

Lemma 3. *Given $i + j = k + \ell = n$, $1 \leq \ell < i \leq j < k$, T is created by connecting internal node u_1 of S_i and internal node u_2 of S_j. T' is created by connecting internal node u_3 of S_k and internal node u_4 of S_ℓ. Then*

$$\begin{aligned} (d'(u_3, v))_{v \in V(T')} & \prec (d(u_1, v))_{v \in V(T)}, \\ D(T') & \prec D(T). \end{aligned}$$

Proof. Note that $|V(T)| = |V(T')| = n$.

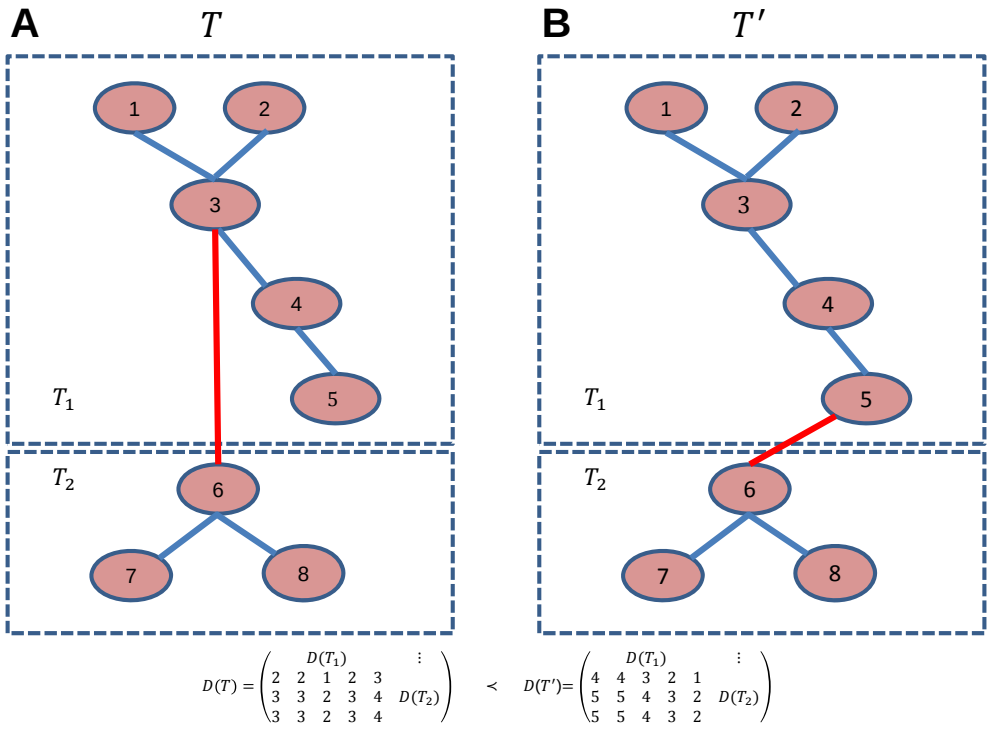


Figure 2.6. Illustrating the choices of u_1, u_2 and u_3 in Lemma 2. Here T_1 has 5 nodes, T_2 3 nodes. We choose $u_1 = 3, u_2 = 5$ and $u_3 = 6$. Tree T is constructed by joining u_1 and u_3 while T' by joining u_2 and u_3. $D(T)$ and $D(T')$ are 8×8 matrices where the first 5 columns correspondent to the 5 nodes in T_1, and the last 3 rows correspondent to the 3 nodes in T_2.

Note also that $(d(u_1, v))_{v \in V(T)}$ has 1 entry equals to 0, i entries equal to 1 and $j-1$ entries equal to 2. Similarly $(d'(u_3, v))_{v \in V(T')}$ has 1 entry equals to 0, k entries equal to 1 and $\ell - 1$ entries equal to 2. Thus $(d(u_3, v))_{v \in V(T')} \prec (d'(u_1, v))_{v \in V(T)}$ proving the first majorization.

Both $D(T)$ and $D(T')$ have n entries equal to 0, $2(n-1)$ entries equal to 1. $D(T)$ has $2(i-1)(j-1)$ entries equal to 3 and the rest of entries 2, $D(T')$ has $2(k-1)(\ell-1)$ entries equal to 3 and the rest of entries 2. Since $(k-1)(l-1) < (i-1)(j-1)$, thus $D(T') \prec D(T)$ proving the second majorization, and hence the proof of Lemma 3. □

Manipulations in Lemma 3 are illustrated in Figure 2.7, where $n = 10, i = j = 5, \ell = 3, k = 7$.

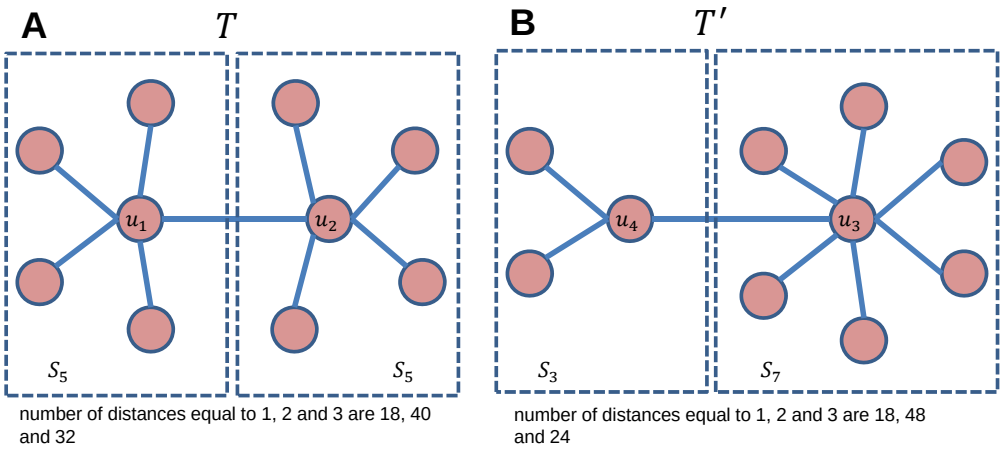


Figure 2.7. Illustration of Lemma 3. Here $n = 10, i = j = 5, \ell = 3, k = 7$. From the counts of the distances above, it is clear that $(d'(u_3, v))_{v \in V(T')} \prec (d(u_1, v))_{v \in V(T)}$ and $D(T') \prec D(T)$.

2.8.1 Proof of Theorem 2

In this section we will find upper and lower bounds of $W_f(T)$ for $T \in \mathcal{T}_n$. Lemmas 4 and 5 are dedicated to investigate the relationship between a tree's distance matrix and its maximum degree.

Consider the following subtree pruning and regrafting (SPR) algorithm:

Input $T \in \mathcal{T}_n$ with $\Delta(T) \geq 3$:

1. Choose a pendant node u_1, and an internal node u_2 with $\deg(u_2) \geq 3$ satisfying the condition that all nodes lying on the shortest path connecting u_1 and u_2, if any, are of degree 2.

2. Choose $u_3 \in ne(u_2)$ such that u_3 does not lie on the shortest path connecting u_1 and u_2.

3. A new tree $T^0 \in \mathcal{T}_n$ is constructed by first deleting (u_2, u_3) and then connecting u_3 to u_1.

This algorithm outputs a tree T^0 with these properties: (i) $D(T) \prec D(T^0)$; (ii) $\Delta(T) - 1 \leq \Delta(T^0) \leq \Delta(T)$; and (iii) number of pendant nodes is one less than that of T.

To see this, let P_{u_1,u_2} denote the path connecting u_1 with u_2. Observe that

$$\begin{aligned} & (d(v_1, v_2))_{v_1,v_2 \in V(T) \backslash V(P_{u_1,u_2})} \\ \equiv \; & (d^0(v_1, v_2))_{v_1,v_2 \in V(T) \backslash V(P_{u_1,u_2})} \end{aligned}$$

and

$$\begin{aligned} & (d(v_1, v_2))_{v_1,v_2 \in V(P_{u_1,u_2})} \\ \equiv \; & (d^0(v_1, v_2))_{v_1,v_2 \in V(P_{u_1,u_2})}. \end{aligned}$$

For $v_1 \in V(T) \backslash V(P_{u_1,u_2})$, we have

$$\begin{aligned} & (d(v_1, v_2))_{v_2 \in P_{u_1,u_2}} \\ \equiv \; & d(v_1, u_3) + 1 + (d(u_2, v_2))_{v_2 \in P_{u_1,u_2}} \end{aligned}$$

and

$$\begin{aligned} & (d^0(v_1, v_2))_{v_2 \in P_{u_1,u_2}} \\ \equiv \; & d^0(v_1, u_3) + 1 + (d^0(u_1, v_2))_{v_2 \in P_{u_1,u_2}} \\ \equiv \; & d(v_1, u_3) + 1 + (d(u_1, v_2))_{v_2 \in P_{u_1,u_2}} \\ \succ \; & (d(v_1, v_2))_{v_2 \in P_{u_1,u_2}} \quad \text{by Lemma 2.} \end{aligned}$$

Thus $D(T) \prec D(T^0)$ and property (i) follows. Since $\deg_{T^0}(u_2) = \deg_T(u_2) - 1$, $\deg_{T^0}(u_1) = 2$, $\deg_{T^0}(u) = \deg_T(u)$ for $u \neq u_1, u_2$. Then properties (ii)

and (iii) follow.

Manipulations of SPR algorithms are illustrated in Figure 2.8.

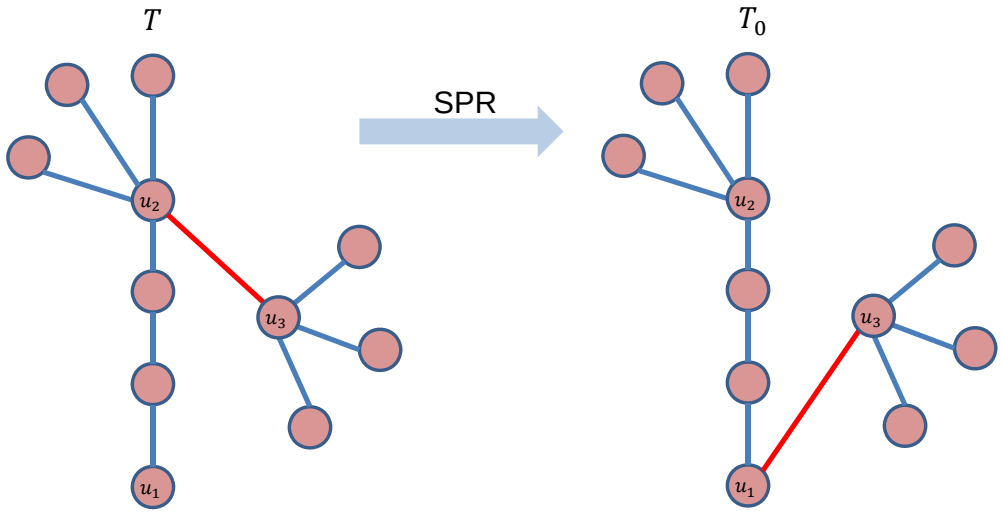


Figure 2.8. Illustration of the subtree pruning and regrafting algorithm. Here T_0 is obtained from T first by deleting the edge (u_2, u_3) and then connecting u_1 and u_3. T_0 is proved to satisfy these properties: (i) $D(T) \prec D(T^0)$; (ii) $\Delta(T) - 1 \leq \Delta(T^0) \leq \Delta(T)$; and (iii) number of pendant nodes is one less than that of T.

Lemma 4. *Let $T \in \mathcal{T}_n$ with $\Delta(T) \geq 3$. There exists $T' \in \mathcal{T}_n$ such that $\Delta(T') = \Delta(T) - 1$ and*

$$D(T) \prec D(T').$$

Proof. Let ℓ be the number of pendant nodes in T. Apply SPR algorithm to T to obtain T^0. If $\Delta(T^0) = \Delta(T) - 1$, then we stop and take $T' = T^0$. Otherwise let $T = T^0$ and apply SPR algorithm again. We repeat this algorithm until we obtain the desired tree T'. Note that this algorithm will be repeated at most $\ell - 2$ times to get the desired tree. Because each application of SPR algorithm reduces number of pendant nodes by 1. There

are at least 2 pendant nodes in a tree. □

Lemma 5. *Let $T \in \mathcal{T}_n$ with $2 \leq \Delta(T) < n-1$. There exists $T' \in \mathcal{T}_n$ such that $\Delta(T') = \Delta(T)+1$ and*

$$D(T') \prec D(T).$$

Proof. We write $\Delta(T) = k$. Choose $u \in V(T)$ with degree m, $m \geq 2$, in such a way that all its neighbors except one are pendant nodes. Write $ne(u) = \{u_1, \ldots, u_{m-1}, u_m\}$ where u_m is the only internal node in T. We consider two cases: 1: $m-1+\deg_T(u_m) < k+1$ and 2: $m-1+\deg_T(u_m) \geq k+1$.

1. A new tree T^0 is constructed by deleting edge (u, u_j), and then connecting u_j to u_m for $1 \leq j \leq m-1$. We claim that T^0 satisfies that $\Delta(T^0) = k$ and $D(T^0) \prec D(T)$. Since $\deg_{T^0}(v) = \deg_T(v), v \in V(T)\backslash\{u, u_m\}$, $\deg_{T^0}(u) = 1$, $\deg_{T^0}(u_m) = \deg_T(u_m) + m - 1 \leq k$, so $\Delta(T^0) = k$. Let $V = V(T) = V(T^0)$, $V_1 = ne(u_m)\backslash u$, $V_2 = V_1 \bigcup \{u, u_1, \ldots, u_m\}$ and $V_3 = V\backslash V_2$.

$$(d(i,j))_{i,j\in V_3} \equiv (d(i,j))^0_{i,j\in V_3}.$$

$$(d(i,j))_{i,j\in V_2} \succ (d(i,j))^0_{i,j\in V_2} \text{ by Lemma 3.}$$

$$(d(i,j))_{i\in V_2, j\in V_3} \succ (d(i,j))^0_{i\in V_2, j\in V_3}.$$

Thus $D(T^0) \prec D(T)$. Let $T = T^0$ and repeat this procedure again. Note that the number of pendant nodes in T increases by 1 for each application of this procedure.

2. A new tree T^0 is constructed by deleting edge (u, u_j), and connecting u_j to u_m for $1 \leq j \leq k - \deg_T(u_m) + 1$. As in case 1, T^0 satisfies $D(T^0) \prec D(T)$. Since $\deg_{T^0}(v) = \deg_T(v), v \in V(T)\backslash\{u_1, u_m\}$, $\deg_{T^0}(u) = \deg_T(u) - (k+1-\deg_T(u_m)) < k$, $\deg_{T^0}(u_m) = k+1$, so $\Delta(T^0) = k+1$. Let $T' = T^0$ and T' satisfies conditions in Lemma 5.

As we claimed in case 1, each time case 1 occurs, the number of pendant nodes in T decreases by 1, and $(deg(i))_{i\in V(T)} \prec (deg(i))_{i\in V(T^0)}$. Thus eventually only case 2 remains and produces a tree as required in Lemma 5. □

In the proof of Lemma 5, we can easily choose $u \in V(T)$ such that its degree equals to m, $m \geq 2$, and all its neighbors except one are pendant nodes. We write $\Delta(T) = k$. So T has at least one node with degree k. We choose one such node and denote it as v. Let $v_1, \ldots, v_{n_1}$ be the pendant nodes in T and satisfy $d(v, v_1) \leq d(v, v_2) \leq \ldots \leq d(v, v_{n_1})$. Denote the path connecting v and v_{n_1} by $v \to w_1 \to w_2 \to \cdots \to w_p \to v_{n_1}$. Then w_p is one such node u. Otherwise, $d(v, v_1) \leq d(v, v_2) \leq \ldots \leq d(v, v_{n_1})$ does not hold.

Since the star graph has the largest maximum degree, and the path graph has the smallest maximum degree among trees in $\mathcal{T}_n$, by Lemmas 4 and 5, we obtain the following corollary.

Corollary 7. *Let $T \in \mathcal{T}_n$ with $2 < \Delta(T) < n-1$. Then*

$$D(S_n) \prec D(T) \prec D(P_n).$$

Proof of Theorem 2. Applying Corollary 7 and the fact that f is increasing can prove Theorem 2. □

2.8.2 Proof of Theorem 1

Define $\mathcal{G}_n(m)$ as a set of connected graphs with the number of nodes n and the number of edges m, $n-1 \leq m \leq \frac{n(n-1)}{2}$.

First we will show that maximum value of $W_f(G)$ over $G \in \mathcal{G}_n(m)$ is a monotone function of the number of edges, m, of G.

Lemma 6. *Let $G \in \mathcal{G}_n$. Then $\max_{G\in\mathcal{G}_n(m)} W_f(G)$ and $\min_{G\in\mathcal{G}_n(m)} W_f(G)$ are decreasing functions in m.*

Proof. For any $G \in \mathcal{G}_n(m)$ with $D(G) = (d(i,j))_{1\leq i,j\leq n}$. When $m \geq n$, G cannot be a tree and hence contains a cycle. Choose an edge in a cycle in G and delete it to form G'. Let's say the deleted edge is $(1,2)$. Note that $G' \in \mathcal{G}_n(m-1)$. Write $D(G') = (d'(i,j))_{1\leq i,j\leq n}$. Since $E(G') \subsetneqq E(G)$, $d(i,j) \leq d'(i,j)$, $1 \leq i < j \leq n$, $W_f(G) \leq W_f(G')$. So $\max_{G\in\mathcal{G}_n(m)} W_f(G) \leq \max_{G\in\mathcal{G}_n(m-1)} W_f(G)$, for $m \geq n$.

Consider $n \leq m \leq \frac{n(n-1)}{2}$. For any $G \in \mathcal{G}_n(m-1)$, we connect two nodes with distance greater than 1 in G and call the resulting graph G''. Now $G'' \in \mathcal{G}_n(m)$ with $D(G'') = (d''(i,j))_{1\leq i\leq j\leq n}$. Since $E(G) \subset E(G'')$, $d''(i,j) \leq d(i,j)$, $1 \leq i < j \leq n$, thus $W_f(G'') \leq W_f(G)$. So $\min_{G\in\mathcal{G}_n(k)} W_f(G) \leq \min_{G\in\mathcal{G}_n(m-1)} W_f(G)$ for $m \geq n$. □

Proof of Theorem 1. From Lemma 6 we have

$$W_f(K_n) \leq W_f(G) \leq \max\{W_f(T) : T \in \mathcal{T}_n\}.$$

From Theorem 2

$$W_f(P_n) = \max\{W_f(T) : T \in \mathcal{T}_n\}.$$

Thus Theorem 1 follows. □

2.8.3 Proof of Theorem 3

In this section, we consider trees with a given maximum degree. The relationship between the distance matrix and the number of nodes with degree equal to maximum degree is investigated.

Lemma 7. *Let $T \in \mathcal{T}_n$ with n_1 nodes with degree equal to $\Delta(T)$. Suppose $n_1 \geq 2$ and $\Delta(T) \geq 3$. There exists $T' \in \mathcal{T}_n$ with $\Delta(T') = \Delta(T)$ and $n_1 - 1$ nodes with degree equal to $\Delta(T)$. Moreover, we have*

$$D(T) \prec D(T').$$

Proof. Let ℓ be the number of pendant nodes in T. Apply SPR algorithm to T to obtain T^0. If T^0 has $n_1 - 1$ nodes with degree equal to $\Delta(T)$, then we stop and take $T' = T^0$. Otherwise let $T = T^0$ and apply SPR algorithm again. We repeat this algorithm until we obtain desired tree T'. Note that this algorithm will be repeated at most $\ell - 2$ times to obtain desired tree. Because each application of SPR algorithm reduces number of pendant nodes by 1. There are at least 2 pendant nodes in a tree. □

Corollary 8. *Let $T \in \mathcal{T}_n$ with $2 < \Delta(T) < n - 1$. There exists a starlike tree T' with $\Delta(T) = \Delta(T')$ such that*

$$D(T) \prec D(T').$$

Corollary 8 states that among trees with equal maximum degree, distance matrix of a tree with more than one node with maximum degree is

strictly majorized by a distance matrix of a starlike tree. Next to find a tree whose distance matrix majorizes all starlike trees.

Lemma 8. *Let T be a starlike tree with $\Delta(T) = k \geq 3$. Then*

$$D(T) \preceq D(B_{n,k+1}),$$

with equality holds if and only if T is $B_{n,k+1}$.

Proof. Assume T is non-isomorphic to $B_{n,k+1}$. T is a starlike tree thus T has k pendant nodes by definition. Denote by u the node with maximum degree k, by $u_1, \ldots, u_k$ pendant nodes in T and satisfy $d(u, u_1) \leq d(u, u_2) \leq \ldots \leq d(u, u_k)$, and by V_i set of nodes in the shortest path connecting node u and u_i, $1 \leq i \leq k$. Next a new tree T^0 is constructed by deleting edge $(u_{k-1}, ne(u_{k-1}))$ and connecting u_{k-1} to u_k.

For $i, j \in V \backslash \{u_{k-1}\}$,

$$d(i, j) = d^0(i, j).$$

For $i \in V \backslash (V_{k-1} \cup V_k)$

$$\begin{aligned} d(i, u_{k-1}) &= d(i, u) + d(u, u_{k-1}) \\ d^0(i, u_{k-1}) &= d^0(i, u) + d^0(u, u_{k-1}) \\ &= d(i, u) + d(u, u_k) + 1 \end{aligned}$$

thus

$$d(i, u_{k-1}) < d^0(i, u_{k-1}).$$

And

$$(d(i,u_{k-1}))_{V_{k-1}\cup V_k} \equiv (d^0(i,u_{k-1}))_{V_{k-1}\cup V_k},$$

since both vectors are distances of a pendant node to other nodes in one path with length $d(u_{k-1},u_k)$. Thus $D(T) \prec D(T^0)$. If T^0 satisfies $d^0(u,u_1) = \cdots = d^0(u,u_{k-1}) = 1$, then we stop and T^0 is $B_{n,k+1}$. Otherwise let $T = T^0$ and we repeat this process until get tree $B_{n,k+1}$. Note that this algorithm will be repeated $n-k-d(u,u_k)$ times. Because each repetition will increase $d(u,u_k)$ by 1. And maximum of $d(u,u_K)$ is $n-k$ and is attained when T is $B_{n,k+1}$. □

Lemma 9. *For $k \geq 3$,*

$$D(B_{n,k+1}) \prec D(B_{n,k})$$

Proof. Lemma 9 follows directly from Lemmas 4 and 8. □

Proof of Theorem 3. Applying Lemma 8 and the fact that f is increasing. □

Remark It has been proven in corollary 3.5 of Schmuck et al. (2012) that

$$W_f(T_n(k)) = \min\{W_f(T) : T \in \mathcal{T}_n, \quad \Delta(T) = k\} \qquad (\star)$$

where $T_n(k)$ is a k-ary tree, also called Volkmann tree (Fischermann et al., 2002). It remains open whether

$$D(T_n(k)) \preceq D(T) \quad \text{for } T \in \mathcal{T}_n,\ \Delta(T) = k \qquad (\star\star)$$

holds for all k, n and $k \leq n$. We have verified that ($\star\star$) holds for $6 \leq n \leq 9$ and $k = 3$. If ($\star\star$) is true for all n and k, it provides an alternative proof of

$$W_f(T_n(k)) \leq W_f(T)$$

for $T \in \mathcal{T}_n$, $\Delta(T) = k$, and f monotonically increasing.

2.8.4 Proof of Theorem 4

Proof. Let T be a spanning tree of G satisfying $\Delta(T) = k$. Similar to the proof of Theorem 1, one can prove that $D(G) \preceq D(T)$. By Theorem 3, $D(T) \preceq D(B_{n,k+1})$. Thus $W_f(G) \leq W_f(B_{n,k+1})$. □

Chapter 3

Profiling the Transcription Factor Regulatory Networks of Human Cell Types

3.1 Introduction

Living cells are the products of transcription programs involving thousands of genes. Sequence-specific transcription factor (TF) proteins regulate target genes by binding to promoter regions adjacent to the DNA sequences of the genes. There are less than 2,000 TFs in the human genome (Babu et al., 2004; Ravasi et al., 2010; Vaquerizas et al., 2009; Zhang et al., 2012a). They work cooperatively to enhance or inhibit their target genes to achieve high specificity, and thus to precisely control the condition-dependent expression of the genes to respond to extracellular stimuli. Hence, the mutual interactions among TFs determine cellular identity and shape complex cellular functions (Csermely et al., 2014; Davidson, 2010). This makes the study of human TFs on a system-wide scale of vital importantce (Csermely et al., 2013). In systems biology, regulatory interactions among TFs are modeled

as a TF regulatory network in which the nodes are the TFs and the links represent the regulatory relationship among TFs.

Over the past decade, a great deal of information on the organization of regulatory interactions has been obtained particularly for E. coli and S. cerevisiae (Balazsi et al., 2005; Banerjee and Zhang, 2003; Gerstein et al., 2012; Ma et al., 2004; Yu et al., 2006). However, comprehensive generation of cell-type regulatory interactions for humans has been a challenge. First, there are a large number of human TFs as mentioned above, but the data collected from individual experiments often target one cell type and only a few TFs in a particular condition (Davidson et al., 2002; Gerstein et al., 2010; Kim et al., 2008). Second, correlation-based analyses of microarray gene expression data often do not capture the direction of transcriptional regulations, a necessity for deep analyses of regulatory interactions (Basso et al., 2005; Carro et al., 2009). Fortunately, the genome-wide DNaseI footprinting technique has recently been adopted to determine the regulatory interactions of sequence-specific TFs in the 41 human cell types (Neph et al., 2012a). This provides a valuable resource for deciphering regulatory mechanisms in different human cells.

The TF regulatory networks for E. coli (Yu and Gerstein, 2006), S. cerevisiae (Jothi et al., 2009; Yu and Gerstein, 2006), mouse (Bookout et al., 2006) and humans (Gerstein et al., 2012) exhibit hierarchical organizations. Most importantly, these organizations are associated with TF dynamics (Jothi et al., 2009; Yu and Gerstein, 2006). In the present thesis, we investigate the structural organizations and dynamics of the 41 human cell-type TF regulatory networks reported in Neph et al. (2012a) using the vertex-sort algorithm developed in Jothi et al. (2009). Our findings are

interpreted to indicate three insightful conclusions. First, the human cell-type TF regulatory networks share similar global three-layer (top, core, and bottom) hierarchical architectures, which are markedly different from that of the yeast TF regulatory network. On the other hand, there are significant differences in the TF regulatory interactions among cell types, as suggested by our finding that wirings around a few TFs can distinguish cell identities well. Second, the TF regulatory network of the human embryonic stem cell (hESC) is dense and has different topological properties from all the other networks. Finally, there are more specific regulatory interactions than thought in the hESCs. These cell-type regulatory interactions and the TFs involved may play unique roles in maintaining pluripotency.

3.2 Materials and Methods

3.2.1 Network data

The TFs regulatory networks of 41 human cell types have been taken from the recent work by Neph et al. (2012a). These networks were derived from DNaseI footprinting data and the predicted TRANSFAC motif-binding sites (Ravasi et al., 2010). Each network contains about 475 TFs and 11,200 interactions.

According to the physiological and functional properties, Neph et al. (2012a) divided the 41 cell types into eight classes: blood (seven cell types), cancer (two cell types), endothelia (four cell types), epithelia (six cell types), ESCs (one cell type), fetal (three cell types), stroma (14 cell types), and viscera (four cell types).

3.2.2 Discovery of the hierarchical structures of the regulatory networks

We used the vertex-sort algorithm (Jothi et al., 2009) to identify the hierarchical structure of a regulatory network. The vertex-sort algorithm first collapses strongly connected components into super-nodes to form a directed acyclic graph, and then constructs its transposed graph by reversing the directions of the edges. A strongly connected component is a subnetwork in which, for each pair of nodes u and v in the subnetwork, these exists a directed path from u to v and from v to u. Next, it uses the topological structures of both the directed acyclic graph and its transposed graph to classify the original nodes into the top, core and bottom layers.

3.2.3 Classifying cell types based on TF regulatory networks

Neph et al. (2012a) made use of the connectivity of the TF regulatory networks to classify the 41 human cell types. Specifically, they computed all the pairwise Euclidean distances between the normalized node-degree (NND) vectors of the networks, and then applied the Ward clustering method (Ward, 1963) to cluster the cell types.

Instead, we used local connectivity, defined by a subset of nodes in the networks, to classify the cell types. Given a small set of TFs, A, we define the feature vector of each cell type to be $(x_1, \ldots, x_n)$, where n is the number of TFs in the corresponding network and where $x_i = 1$ if the $i\text{-}th$ TF is a target of some TFs in A and 0 otherwise. Principal component analysis was then applied to the feature vectors to reduce the dimension and the noise of feature vector data. We computed the pairwise Euclidean distances based on the first seven principal components of the 41 feature vectors and

then applied Ward clustering to classify the cell types.

To answer one question that how well local topological features of randomly selected TF group distinguish the cell identities, we apply the described strategy for 1000 randomly selected TF groups with n TFs ($n = 1, \ldots, 12$)

3.2.4 Measuring the accuracy of the classifications of cell types

The Rand Index (RI) (Rand, 1971) was used to assess the quality of cell type classifications. To this end, the 41 cell types are partitioned into four categories: (i) stromal and epithelial, (ii) blood, (iii) endothelial, and (iv) cancer, ESC, and fetal tissues.

3.2.5 Detection of regulatory complex-target modules in hESCs

The hESC specific interactions are interactions that are only found in the regulatory network of hESCs. A total of 1,509 interactions were identified (Table A.1).

We used these interactions to identify regulatory complex-target modules that are specific to hESCs. For a protein complex, C, and a set of TFs, B, we say that C and B form a regulatory complex-target module if C contains two or more TFs such that all TFs in B are regulated by every TF (in C) only in the hESCs. We detected 55 regulatory complex-target modules (Table A.2 using the protein complexes reported in Vinayagam et al. (2013).

3.2.6 Comparing two distributions

The Wilcoxon rank-sum test was used to determine whether the RI was significantly higher when grouping the 41 cell types based on the targets of a few TFs compared to random grouping.

The gene expression data of 79 human tissues (Su et al., 2004) was used to investigate whether a TF gene was stably expressed across tissues. The deviation of an expression level from being a constant is measured in terms of its relative entropy (also known as Kullback-Leibler divergence). In our context, for a gene, it is computed as $\log_2 79 + \sum_j f_j \log_2(f_j)$, where $f_j = e_j/(\sum_{k=1}^{79} e_k)$ and e_j is the expression level of the gene in tissue j (Ravasi et al., 2010). The entropy equals 0 if the gene expression levels are identical in all 79 tissues. The Wilcoxon rank-sum test was also used to test whether the TFs involved in housekeeping (HK) interactions were more stably expressed than the other TFs.

Wilcoxon rank-sum test require below 3 assumptions. (1) Data are paired and come from the same population. (2) Each pair is chosen randomly and independently. (3) The data are measured at least on an ordinal scale (cannot be nominal). In our applications, the first two assumptions may not be entirely satisfied. For example the independence assumption may not hold considering potential bias in TFs detection.

3.3 Results

3.3.1 Wirings around a few TFs are enough to distinguish cell identities

Neph et al. (2012a) made use of the global connectivity of the TF regu-

latory networks to classify the 41 human cell types (Section 3.2.3). The resulting grouping (redrawn in Figure 3.1A) strikingly groups the anatomical and functional cell-type groups into clearly pre-annotated classes with RI=0.801. Surprisingly, the local connection patterns involving five to nine arbitrarily selected TFs are also good enough to obtain comparable classifications with the RI being in the range from 0.7 to 0.9 on average (Section 3.2.4, Figure 3.2).

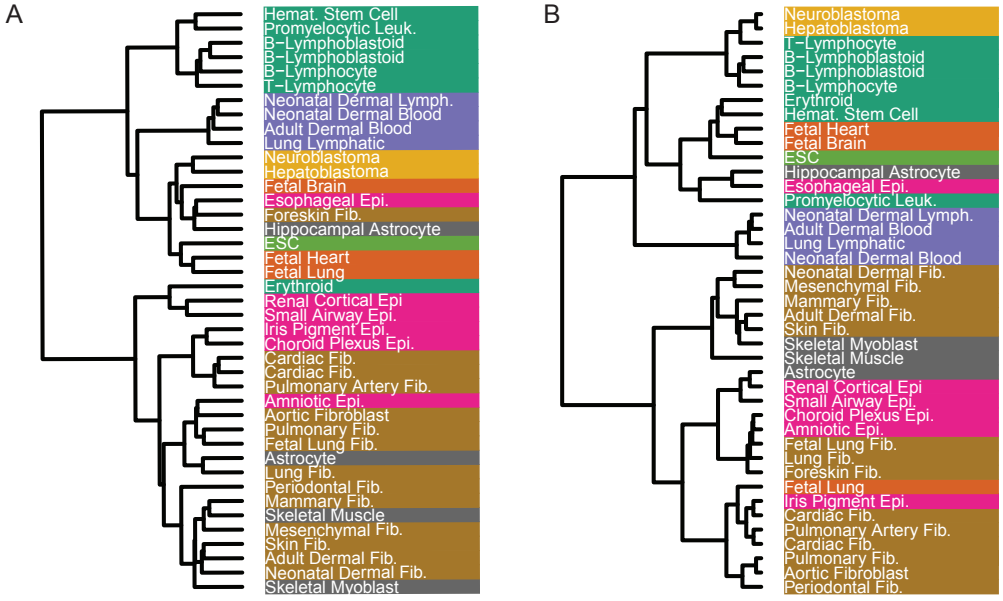


Figure 3.1. The hierarchical clustering of 41 cell types, where the color indicates which classes they belong to (Section 3.2.1). (A) The clustering reported in Neph et al. (2012a) and redrawn for the purpose of comparison, which is based on the pairwise Euclidean distances between the NND vectors of the corresponding TF regulatory networks, has RI=0.801. (B) Our clustering, which is based on the distribution of the downstream targets of the seven STATs, has RI=0.856.

Let us consider the seven mammalian signal transducer and activator of transcription (STAT) proteins. The activation of STATs by the Janus kinase proteins serves as an alternative to the second messenger system, transmitting extracellular signals from a wide spectrum of cytokines,

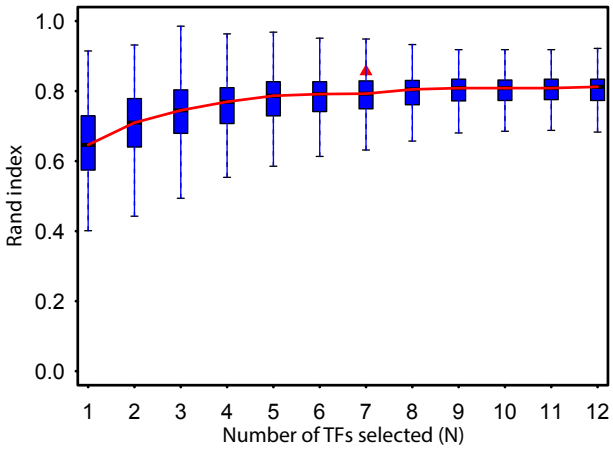


Figure 3.2. The evaluation of how the clustering results of limited number of TFs reflect the original cell/tissue groups. The red triangle marks the RI value of the STAT family.

growth factors and other polypeptide ligands to the nuclei (Horvath, 2000; Levy and Darnell, 2002). A close examination finds that the TFs regulated by the STATs are annotated with different gene ontology (GO) terms in different regulatory networks. For example, as illustrated in Figure 3.3, TFs that are regulated by STATs in hESCs but not in hematopoietic stem cells (HSCs) are enriched in GO:0045165 (cell fate commitment, Benjamini corrected p-value =2.72e-7). By contrast, TFs that are regulated by STATs in HSCs but not in hESCs are enriched in GO:0048534 (hemopoietic or lymphoid organ development, Benjamini corrected p-value =0.03).

The diversity of the downstream TFs of the STATs might indicate their strong distinguishability for the classification of human cell types. Indeed, using the information on how the STAT proteins connect with their targets to classify the cell types, we obtained a grouping with RI=0.856 (Figure 3.1B), which is even higher than the RI of the grouping of Neph et al. (2012a) mentioned above.

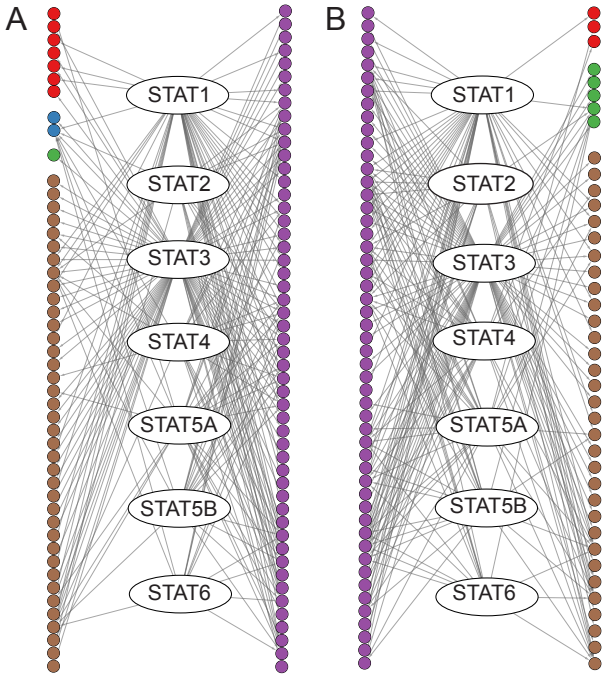


Figure 3.3. The STATs and their downstream regulatory targets in hESCs (A) and HSCs (B). Purple TFs are those regulated by some STATs in both cell types. The cell fate commitment process (GO:0045165) is enriched in the targets of STATs in hESCs (Benjamini corrected p-value =2.72e-7). Dark red and blue targets are the TFs annotated with the GO term. The hemopoietic or lymphoid organ development process (GO:0048534) is enriched in the targets of STATs in HSCs (Benjamini corrected p-value =0.03). Green and blue targets are the TFs annotated with this GO term. Brown targets are other targets whose GO annotations are not given.

3.3.2 The hierarchical structures of 41 cell-type regulatory networks

The E. coli, yeast, rat, mouse, and human regulatory networks all exhibit hierarchical organization (Bookout et al., 2006; Gerstein et al., 2012; Jothi et al., 2009; Yu and Gerstein, 2006). We investigate the hierarchical organization of the 41 human cell type networks using the vertex-sort algorithm (Jothi et al., 2009).

For each network, the vertex-sort algorithm partitioned its nodes into

the top, core and bottom layers (Figure 3.4A) (Section 3.2.2). The percentages of TFs in the three layers of the 41 regulatory networks are reported in Table A.3. On average, 23% of TFs are classified into the top layer, 67% into the core layer, and the lowest amount of TFs (10%) into the bottom layer (Figure 3.4B). The top, core and bottom layers of the 41 networks have 1 (that is HNF4G), 141 and 15 TFs in common, respectively.

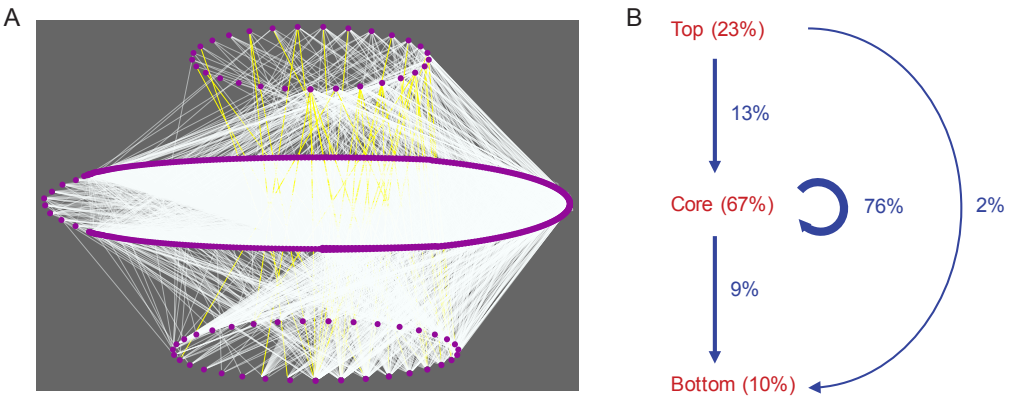


Figure 3.4. (A) A schematic view of the three-layer hierarchical structure of the hESC TF regulatory network. The links between the top and bottom layers are colored yellow. (B) A summary of average percentages of nodes (dark red) in the three layers and of links (blue) within and across the top, core and bottom layers in a human cell-type TF regulatory network.

When compared to the regulatory networks of other cell types, the hESC TF regulatory network has a significantly low number of TFs in the top layer (6%, p-value < 0.01, one-tailed test) and its core layer contains a significantly high number of TFs (85%, p-value < 0.01, one-tailed test). However, its bottom layer has a size (9%) similar to those of the other cell type networks (Table A.3).

To measure the degree of hierarchy in the three-layer structures obtained above, we calculated the local reaching centrality (LRC) of TFs in each of the 41 networks (Mones et al., 2012). As expected, the LRC of each

TF in a layer is always greater than that of each TF in the layers below it in all except two stromal (HCF and HCM) networks. In the HCF network, only HOXC9 and NKX2-1 in the top layer have an extremely low LRC, smaller than the LRC of the TFs in the core layer. In the HCM network, only HOXC9 and NKX6-1 in the top layer have smaller LRC than that of TFs in the core layer. The mean values of the LRC of the TFs in a layer in the 41 regulatory networks are given in Table A.4. The global reaching centrality (GRC) of the 41 regulatory networks ranges from 0.065 to 0.125. Low GRC for each network is due to (1) there are only three hierarchal layers, (2) the core layer is much larger than the top layers (67% vs 23% on average), and (3) the LRC of a TF is slightly smaller in the core layer than in the top layer. These facts leads to the distribution of LRCs skew to the maximum LRC resulting in small GRC.

Distributions of network links. Seventy-six percent of links are distributed within the core layer (Table A.3 and Figure 3.4B). Both the size of the core layers and the links within them reveal the complex regulatory relationships among TFs in different human cells. The remaining links are distributed as follows: top $\rightarrow$ core (13%), top $\rightarrow$ bottom (2%), and core $\rightarrow$ bottom (9%), suggesting that TFs in the top layer mainly regulate TFs in the core layer.

Distributions of hubs. TFs with high out-degrees are crucial in that they have a large numbers of downstream targets. Following Jothi et al. (2009), the top 20% TFs with the largest out-degree are defined as hubs in a regulatory network. There are 96 to 98 hubs that regulate at least 21 TFs in each of the 41 cell-type regulatory networks. The core layers of the networks are all enriched in hubs (all p-values ≤ 0.005, hypergeometric test,

Figure 3.5A). All the top layers are depleted in hubs (all p-values ≤ 0.05, hypergeometric test) except in the networks of hESCs, HSCs, hippocampal astrocytes and mammary fibroblasts (Figure 3.5A). These results on hub enrichment are concordant with those of the yeast transcription network (Jothi et al., 2009).

Distributions of essential TFs. Essential proteins are necessary for performing basic developmental functions. If they are disrupted, they will cause pre- or neonatal lethality (Georgi et al., 2013). There are 280 essential TFs in each of the 41 networks. For each network, the percentages of essential proteins in the top and core layers are about the same (average difference 1%) (Figure 3.5B). By contrast, the percentage of essential proteins in the top layer (12%) is higher than in the core layer (6%) and in the bottom layer (3%) in the yeast transcription network (Jothi et al., 2009).

Distributions of HK TFs. Here TFs encoded by HK genes (Eisenberg and Levanon, 2013) are called HK TFs. There are 63 HK TFs in each of the 41 networks. There are 2, 54 and 7 HK TFs respectively in the top, core and bottom layers of the hESC TF regulatory network. In the remaining 40 networks, all the core layers are enriched, whereas all the top layers are depleted in HK TFs (Figure 3.5C).

3.3.3 HK and specific regulatory interactions

In analogy to genes, some regulatory interactions appear in only certain cell types, whereas many others are found in all cell types. Regulatory interactions that are only found in one cell type are called specific interactions; those that are found in all cell types are called HK interactions. Identifying the regulatory interactions belonging to the classes provides important bi-

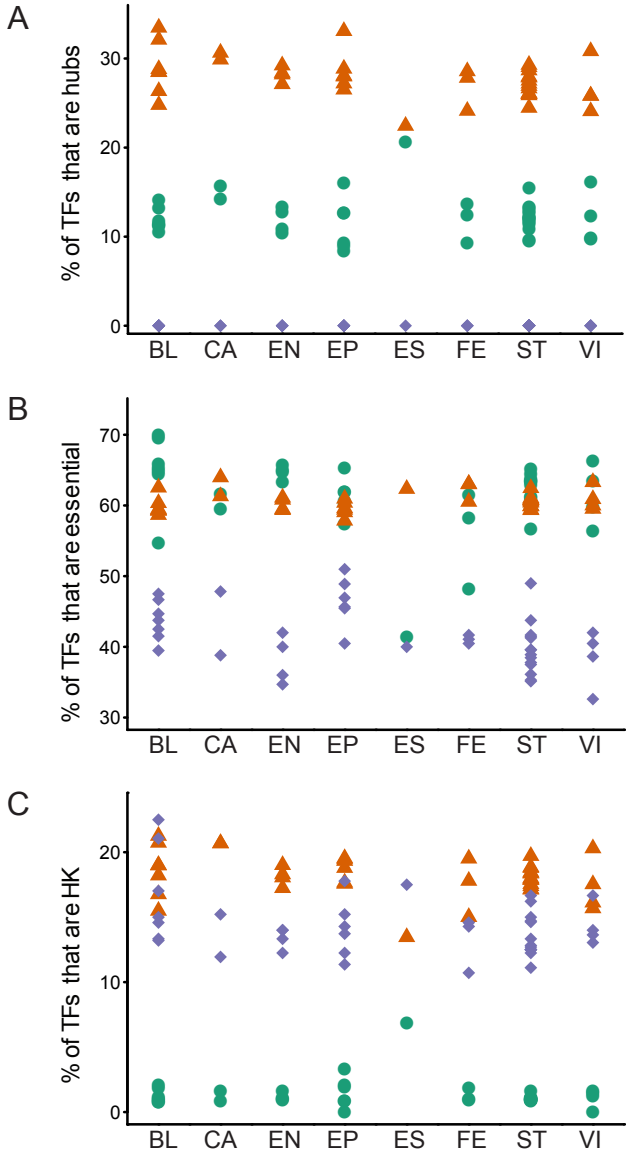


Figure 3.5. Percentages of TFs that are hubs (A), essential (B) and HK (C) in the top (green circle), core (brown triangle) and bottom (blue diamond) layers in 41 human cell-type TF regulatory networks, grouped according to cell class. Abbreviations: BL, blood; CA, cancer; EN, endothelia; EP, epithelia; ES, ESC; FE, fetal; ST, stromal cells; VI, visceral cells.

ological insights into complex biological systems (Bolouri, 2014; Ideker and Krogan, 2012; Mitra et al., 2013; Srivas et al., 2013).

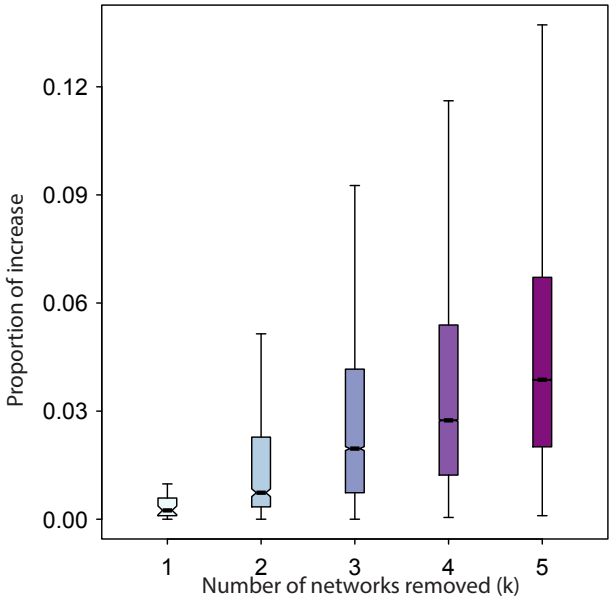


Figure 3.6. Proportion of increase in number of HK interactions in all potential 41-k TF regulatory networks. Where for each k, we enumerate all possible percentage of increase in number of common interactions in 41-k TF regulatory networks.

Neph et al. (2012a) remarked that 5% of all interactions (i.e. 2041 interactions) (Table A.5) are common across the 41 cell types. Encouraging fact is that HK interactions are remarkable robust with median increase from 0.24% ($k = 1$) to 3.87% ($k = 5$) (Figure 3.6). We therefore take these 2041 interactions as HK regulatory interactions. Enrichment analyses show that the proportions of HK links within the core layer and between the core and bottom layers are comparable and higher than those between the top and core layers and between the top and bottom layers (Figure 3.7C).

There are 296 TFs involved in HK interactions (Figure 3.7A). These TFs are not necessarily encoded by HK genes. But, as expected, they are enriched with TFs encoded by the HK genes listed in Eisenberg and Levanon (2013) (p-value =1.27e-10; hypergeometric test). Additionally, the expressions of genes encoding them are much stabler than other TF genes

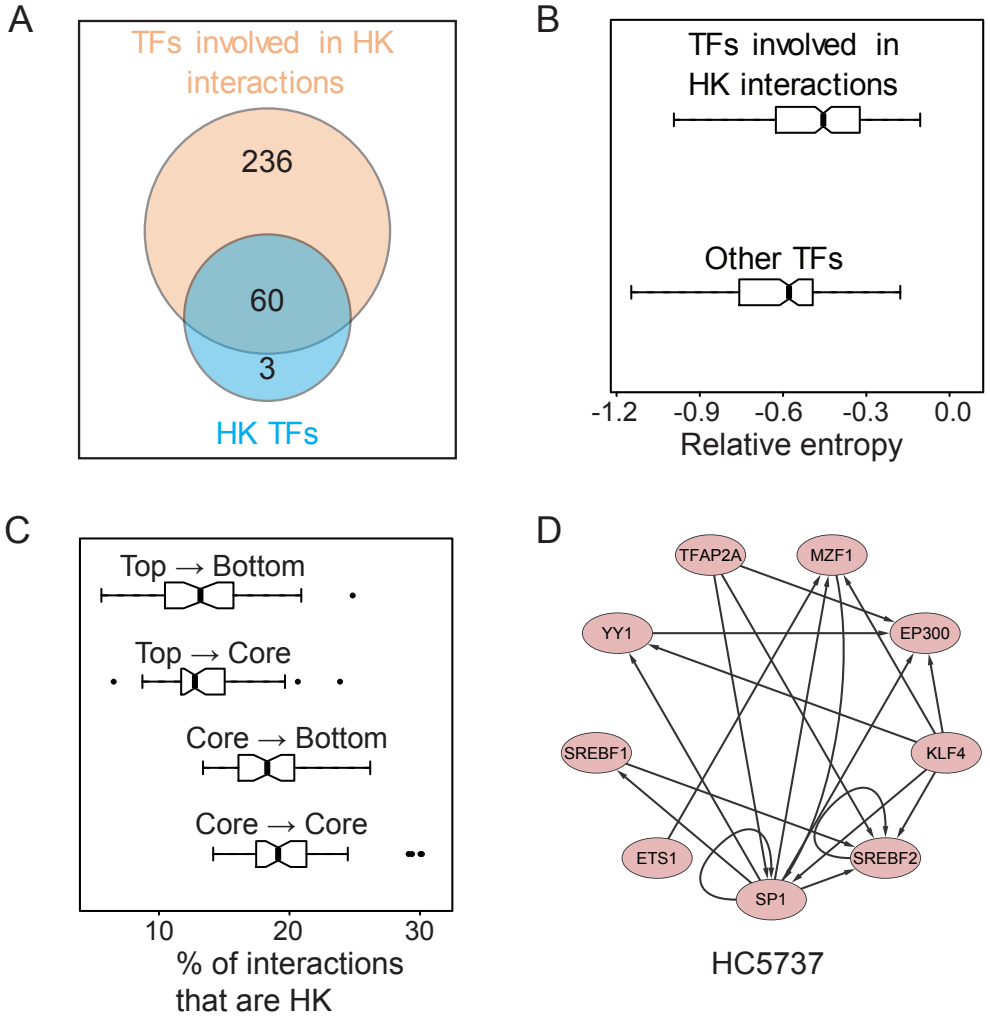


Figure 3.7. A) The intersection of the subset of TFs that are involved in HK interactions and the subset of TFs that are encoded by HK genes. (B) The box plots of the relative entropy of the expression values of the genes encoding TFs involved in HK interactions (above) and other TFs (below). (C) The box plots of the proportions of HK interactions within the core layer and among the top, core, and bottom layers in the 41 human cell-type TF regulatory networks. (D) TFs and HK interactions among them in a protein complex (id: HC5737) (Vinayagam et al., 2013)

across 79 human tissues (p-value =4.32e-10) based on the entropy analysis of the gene expression data reported in Su et al. (2004) (Figure 3.7B). Similar results hold for the HK gene list obtained from combining the lists in Eisenberg and Levanon (2003); She et al. (2009), and Chang et al. (2011)

(Figure 3.8).

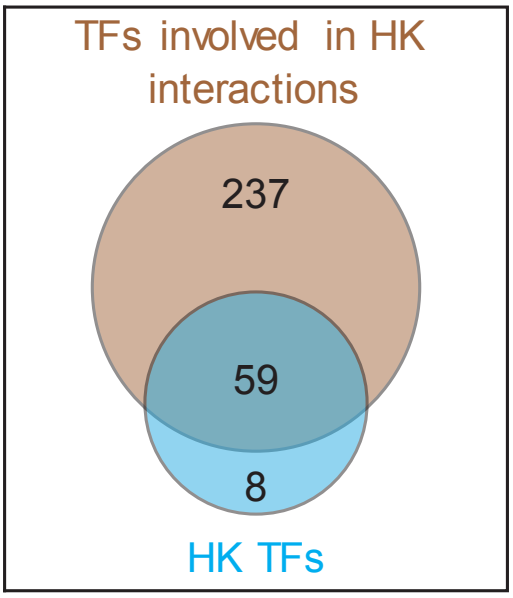


Figure 3.8. The TFs involved in HK interactions that appeared in all of the 41 TF regulatory networks are significantly (p value=5.62e-07) enriched in HK TFs list obtained by combining the lists in Eisenberg and Levanon (2003); She et al. (2009), and Chang et al. (2011).

3.3.4 Regulatory interactions specific to hESCs

ESCs are derived from the inner cell mass of an early-stage embryo. Although OCT4, NANOG and other markers of hESCs have been identified, the whole picture of how TFs cooperate with each other in hESCs is largely unclear (Chen et al., 2008; Liu et al., 2009; Young, 2011). There are 1509 regulatory interactions specific to hESCs, involving 411 TFs. The network induced by specific interactions over these TFs is referred to as the hESC specific network (ESCSN). There are 82 hubs (the top 20% of the TFs with the largest total degree) (Table 3.1). Among the 82 hubs, only 35 are the hub TFs in the original hESC TF regulatory network. The remaining 47 hubs, including popular NANOG, seem to play unique roles in hESCs.

Table 3.1. There are 82 hub TFs in the ESCSN. Forty-seven of them, include NANOG, are not hubs in the original hESC TF regulatory network. TFs encoded by hESC-specific genes with super-enhance are colored red.

	Hubs only in the specific netowrk				Hubs also in the original network			
Top	HNF4A	PPARA			SPZ1			
Core	ALX1	FOXA1	LMX1B	PAX6	ETS1	NR2F2	SOX2	TFAP2B
	ALX3	FOXA2	MNX1	POU2F3	FOXD3	NR2F6	SP1	TFAP2C
	ALX4	FOXC1	MSX2	POU4F3	GTF2I	PAX4	SP2	VDR
	ARX	FOXH1	NANOG	SIX3	IKZF1	PAX5	SP3	ZBTB7B
	ATOH1	FOXI1	NKX2−2	SMAD4	MAZ	POU2F1	SP11	ZFP42
	BARHL2	FOXJ1	NR5A2	TBX22	MYCN	OCT4	SREBF2	ZNF148
	CDX2	GFI1	OTP	VAX1	NF1	PURA	STAT3	ZNF216
	CRX	HOXB13	OTX2	ZIC1	NFKB2	REST	TCF3	ZNF219
	DMRT1	LHX4	PARP1	ZIC2	NR2F1	RXTA		
	DMRT3	LMX1A	PAX2	ZIC3				
	ETV7							
Bottom	HBP1	OVOL2	PAX7	SIX6				

Super-enhancers are large collections of transcriptional enhancers. Genes with super-enhancer domain play important roles in the control of cell identity and diseases (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). In mouse and human ESCs, master transcription factors OCT4, SOX2 and NANOG are each encoded by a gene with super-enhancer. They also have DNA binding motifs that are often found in super-enhancer domains (Whyte et al., 2013). Most interestingly, nine hub TFs (colored red in Table 3.1) are each encoded by hESC-specific genes with super-enhancer (p-value = 0.03; hypergeometric test) based on super-enhancers reported in Hnisz et al. (2013). They are FOXD3, GTF2I, NANOG, NR2F6, OCT4, SIX3, SOX2, ZBTB7B, and ZIC3.

Assou et al. (2007) in a meta-analysis compiled a list of 1076 genes that are overexpressed in hESCs. In the ESCSN the hubs are significantly enriched with the TFs encoded by the overexpressed genes in this list (p-value =1.61e-3; hypergeometric test, Figure 3.9A). More interestingly, 12 of the hubs that are encoded by the genes in the list are well connected, except for ZIC2 (Figure 3.9B). Interestingly, NANOG, OTX2, PARP1, ZIC2 and ZIC3 are not hubs in the original hESC TF regulatory network.

ESCs self-renew indefinitely while maintaining pluripotency. Activin A is a member of the transforming growth factor beta superfamily. It is found to play a central role in maintaining stemness (James et al., 2005; Xiao et al., 2006). Activin A initially binds to type II Activin A receptors and then recruits the Activin A receptor, type IB (ALK4). ALK4 further phosphorylates SMAD2/3. Upon activation by phosphorylation and association with SMAD4, SMAD2/3 translocates to the nucleus and up-regulates the expression of other TF genes, such as Oct4, Nanog, Modal, Wnt3, and Fgf8, and down-regulates Bmp7 (James et al., 2005). In hESCs, SMAD3 tends to co-occupy DNA binding sites with OCT4, SOX2 and NANOG in responses to transforming growth factor beta signaling (Mullen et al., 2011). The Nadal/Activin A signaling pathway is also enriched (False discovery rate=9.86e-5) with the hubs in the ESCSN.

In addition, a core transcriptional regulatory network of hESCs (Chen et al., 2008) is enriched in hESC specific interactions (p-value =6.92e-6; hypergeometric test, Figure 3.9C), as shown in Figure 3.9D.

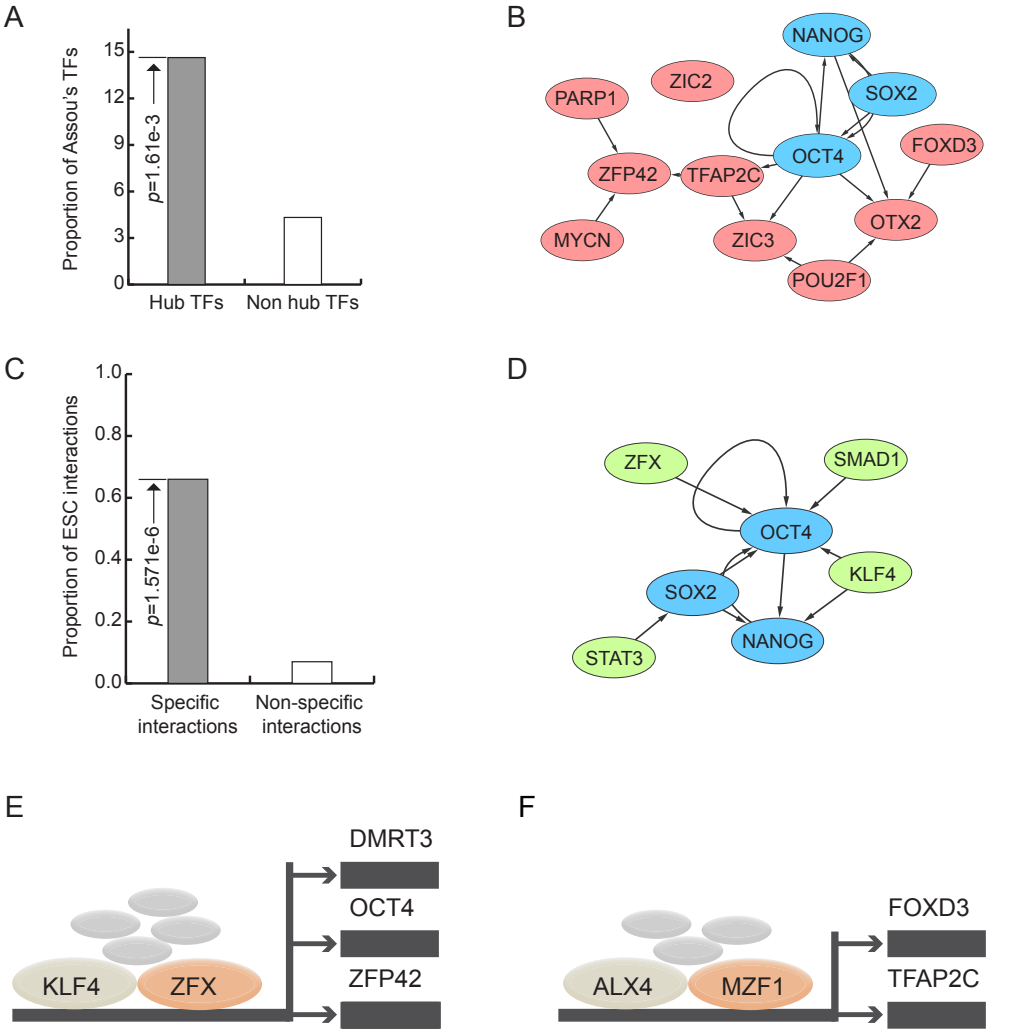


Figure 3.9. (A) Proportions of hub TFs that are in Assou et al. (2007) and the significance of their enrichment in the ESCSN. (B) The subnetwork induced by the hub TFs in the Assou et al.s list in the ESCSN. (C) Proportions of known hESC interactions (38) and the significance of their enrichment in the ESCSN. (D) The hESC specific regulatory interactions appearing in a reported core transcription network for hESCs (Chen et al., 2008). (E) and (F) Two specific regulatory complex-target modules in the hESCs.

3.4 Discussion

We have studied the organizational architectures of the 41 human cell-type TF regulatory networks that were reported by Neph et al. (2012a). First, we

have showed that the wiring around five to seven TFs in the networks can be used to classify all the 41 cell types well. Both Neph et al. (2012a) and our studies indicate that the human TF regulatory networks are different globally as well as locally.

Human regulatory networks exhibit hierarchical and modular structure (Rodriguez-Caso et al., 2005). We have examined the three-layer hierarchical organizations of the human cell-type TF regulatory networks. The networks are each partitioned into the top, core and bottom layers, containing 23%, 67% and 10% of TFs on average (Figure 3.4B, Table A.3), respectively. The large size and well-connectedness of the core layers are probably due to (1) master cell-type specific TFs have a large number of target genes and (2) their encoding genes have a super-enhancer domain (Hnisz et al., 2013; Whyte et al., 2013). For example, in the core layer of the hESC TF regulatory network, 326 TFs (81.3%) out of 401 are either the regulators or regulated by nine TFs each encoded by a gene with super-enhancer domain, forming a large bow-tie subnetwork (Csete and Doyle, 2004).

The same hierarchical analysis (Jothi et al., 2009) indicates that in the yeast TF regulatory networks both the core and bottom layers have similar sizes (43% vs 40%) whereas the top layer contains only 13% of the TFs. Taken together, these two facts together imply a difference in the topological organizations between the human and yeast TF regulatory networks.

Enrichment analyses (Table 3.2) indicate that for each TF regulatory network of the 40 non-ESC cell types, (a) the top layer is lacking in both hub and HK TFs, (b) the core layer is enriched with both hubs and HK

TFs and (c) the bottom layer is only enriched with hub TFs. However, essential TFs seem to be distributed evenly in the top and core layers, but, by and large, sparsely in the bottom layers.

Table 3.2. The summary of the enrichments of hubs, essential and HK TFs in the top, core and bottom layers of the 41 cell-type TF regulatory networks. For clarity, the cell types are divided into eight classes, listed (together with the numbers of cell types) in the first column. The symbols + and represent the enrichment and depletion of TFs of a type in a hierarchical layer in all the networks of a class.

	Hub TFs			**Essential TFs**			**Housekeeping TFs**		
	Top	Core	Bottom	Top	Core	Bottom	Top	Core	Bottom
Blood (7)	−	+	−			−	−	+	
Cancer (2)	−	+	−		$+^{c}$	$-^{c}$	−	+	
Endothelia (4)	−	+	−			−	−	+	
Epithelia (6)	−	+	−			$-^{b}$	−	+	
ESC (1)		+	−		+	−			
Fetal (3)	−	+	−		+	−	−	+	
Stroma (14)	$-^{a}$	+	−			$-^{a}$	−	+	
Viscera (4)	−	+	−			−	−	+	

[a] 13 out of 14 are poor in hubs or essential TFs;

[b] 3 out of 6 are poor in essential TFs;

[c] 1 out of 2 are enriched with or poor in essential TFs.

Interestingly, the hESC TF regulatory network has a topological structure that is different from the rest. It has significantly small top and bottom layers and therefore a large core layer. Indeed, seven STATs and 15 key TFs (appearing in Figures 3.9B and 3.9D) are all found in the core layer. Moreover, 87.6% of links are within the core layer, whereas there are only 40 links (0.3%) between the top and bottom layers. These two facts together suggest that hESCs have a highly dense and well-connected TF regulatory network. And our analyses indicate that master TFs and super-enhancers associated TFs are in the kernel of the core layer. Its top layer is neither enriched with nor depleted of hub, essential and HK TFs, in contrast to the TF regulatory networks of the other cell types.

We have also studied the dynamic properties of the human cell-type TF regulatory networks. The HK interactions are related to basic life support

such as bio-molecular synthesis and transcription mechanisms. One of our findings is that most HK interactions are within the core layer or between the core and bottom layers. Using the identified HK interactions to investigate the protein complex database, we identified 23 protein complexes in which the proteins are highly connected with HK links (Table A.6). One of these complexes is given in Figure 3.7D. Most of the identified protein complexes are as predicted and hence it would be interesting to investigate their biological functions.

The ESCSN, the subnetwork induced by specific links in the hESC TF regulatory network, has also been investigated. The 82 hub TFs in the ESCSN (Table 3.1) seem to play important roles in hESCs due to the following facts: (i) their genes are overexpressed, (ii) they are enriched in the Activin A/Nodal signaling pathway, and (iii) specific interactions are enriched in a core transcriptional regulatory network of the hESCs reported in Chen et al. (2008). In general, specific regulatory interactions are difficult to detect because the network of each cell type is based on independent data, leading to a high false negative rate. Since the number of specific interactions in hESCs is much higher than that in other cell types, our results should not be greatly affected by the limitations of the data chosen.

Cell type specificity is believed to be the outcome of the interplay of the DNA sequence binding specificity of TFs, co-factors and epigenetics (Boyer et al., 2005; Chen et al., 2008). Through the integration of a database of protein complexes (Vinayagam et al., 2013) and the ESCSN, we identified 55 hESC- specific regulatory complex-target modules (Section 3.2.5, Table A.2). One of these modules is illustrated in Figure 3.9E: in a complex (id #: HC4463), both KLF4 and ZFX have three common downstream

targets: FOXD3, OCT4 and ZFP42. As expected, KLF4, ZFX and their targets are important in the maintenance of pluripotency, self-renewal and development processes in ESCs (Boyer et al., 2005; Chan et al., 2009; Chen et al., 2008; Galan-Caridad et al., 2007; Jiang et al., 2008; Ramalho-Santos et al., 2002; Rogers et al., 1991). Another is given in Figure 3.9F, in which both ALX4 and MZF1 regulate FOXD3 and TFAP2C. Notably, FOXD3 has recently been demonstrated to be responsible in directing pluripotency and paraxial mesoderm fates in hESCs (Arduini and Brivanlou, 2012). All these facts together suggest that specific regulatory interactions may play important roles in hESCs.

Chapter 4

Profiling Human Embryonic Stem Cell via Feed-Forward Loops in Transcription Factor Regulatory Network

4.1 Introduction

Embryonic stem cells (ESCs) are derived from the inner cell mass of an early-stage embryo. ESCs are capable to maintain self-renewal and pluripotency simultaneously. Self-renewal is the process that ESCs divide to produce more ESCs. Pluripotency is the ability that ESCs differentiate into endoderm, mesoderm, or ectoderm germ layer, then into all human cell types. Deciphering molecular mechanisms which control ESC self-renewal and pluripotency is key to understanding development. It may also help to discover new therapies for diseases resulted from defects in development.

Living human cells are the products of transcription programs involving approximately 21,000 protein-coding genes (Pennisi, 2012). TF proteins

regulate target genes by binding to either promoter or enhancer regions adjacent to the DNA sequences of the genes. There are less than 2,000 TFs in the human genome (Babu et al., 2004; Ravasi et al., 2010; Vaquerizas et al., 2009; Zhang et al., 2012a). Pluripotency of ESCs are largely controlled by TFs OCT4, SOX2, and NANOG. OCT4 and NANOG are essential for establishing or maintaining a robust pluripotent state. SOX2 functions as a heterodimer with OCT4 in ESCs. Expression of SOX2 is generally required for reprogramming somatic cells into induced pluripotent cells (Young, 2011). Super-enhancers are large collections of transcriptional enhancers. Genes with super-enhancer domain play important roles in the control of cell identity and diseases (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). In mouse and human ESCs, OCT4, SOX2 and NANOG are each encoded by a gene with super-enhancer. Their DNA binding motifs are found in super-enhancer domains (Whyte et al., 2013). Hnisz et al. (2013) reported 60 TFs which are encoded by hESC-specific genes with super-enhancers. In another study, Assou et al. (2007) in a meta-analysis compiled a list of 1076 genes that are overexpressed in hESCs.

TFs work cooperatively to enhance or inhibit their target genes to achieve high specificity, and thus to precisely control the condition-dependent expression of the genes to respond to extracellular stimuli. Hence, the mutual interactions among TFs determine cellular identity and shape complex cellular functions (Csermely et al., 2014; Davidson, 2010). This makes the study of human TFs on a system-wide scale of vital importance (Csermely et al., 2013). In systems biology, regulatory interactions among TFs are modeled as a TF regulatory network in which the nodes are the TFs and the links represent the regulatory relationship among TFs.

Over the past decade, a great deal of information on the organization of regulatory interactions has been obtained particularly for E. coli and S. cerevisiae (Balazsi et al., 2005; Banerjee and Zhang, 2003; Gerstein et al., 2012; Ma et al., 2004; Yu et al., 2006). However, comprehensive generation of cell-type regulatory interactions for humans has been a challenge. First, there are a large number of human TFs, but the data collected from individual experiments often target one cell type and only a few TFs in a particular condition (Davidson et al., 2002; Gerstein et al., 2010; Kim et al., 2008). Second, correlation-based analyses of microarray gene expression data often do not capture the direction of transcriptional regulations, a necessity for deep analyses of regulatory interactions (Basso et al., 2005; Carro et al., 2009). Fortunately, the genome-wide DNaseI footprinting technique has recently been widely adopted to determine the regulatory interactions of sequence-specific TFs in the 41 human cell types including hESC (Neph et al., 2012a). This provides a valuable resource for deciphering local regulatory mechanisms on ESC related TFs by comparing local structures of regulatory networks in hESC with those in the other 40 differentiated cell types.

Network motifs are connected sub-graph patterns which are over-represented in the observed network as compared against a network model. One of the most important and extensively studied network motifs is Feed-Forward Loop (FFL) (Alon, 2007). An FFL, as illustrated in Figure 1.6B, consists of 3 nodes A, B and C in which A regulates B, and both A and B regulate C. FFL in regulatory networks can speed-up the response time of the target gene expression or act as sign-sensitivity delays. FFL can generate pulse of gene expression. FFL can also cooperatively enhance induction of gene

C by inducers of TF A. Here inducers of A are small molecules, protein partners, or covalent modifications that activate or inhibit the transcription activities of A (Alon, 2007; Mangan and Alon, 2003; Shoval and Alon, 2010). Early studies revealed that FFL is over-represented in the regulatory networks of organisms ranging from bacteria and yeast to plants and animals (Alon, 2007). Recently FFL as a motif is also found in regulatory networks of worm (Boyle et al., 2014), fly (Boyle et al., 2014), human (Boyle et al., 2014; Gerstein et al., 2012; Neph et al., 2012a). Core TFs in hESC regulatory network form an FFL where OCT4/SOX2 can be viewed as node A, NANOG as node B, and ESC related genes as node C (Boyer et al., 2005). The number of FFLs varies according to the developmental stages in worm and in fly, with L1 stage in worm and late-embryo stage in fly showing the highest number of FFLs, suggesting increased filtering fluctuations and accelerating responses in these stages (Boyle et al., 2014).

Recognising that FFLs play important and dynamic functions in various biological networks, some network centrality measures based on network motifs have been proposed to quantify the importance of nodes in directed networks (Harriger et al., 2012; Koschützki and Schreiber, 2008; Koschützki et al., 2007; Sporns et al., 2007; Sporns and Kötter, 2004; Wang et al., 2014). The underlying idea of these centrality measures is that the more motifs a node is involved in the network, the more important the node could be. These centrality measures are called motif centrality in general and can identify different sets of important nodes in networks partially because they can integrate structural information between local and global information. However these centrality measures only take into account of the structural information, e.g. FFLs, in a single network. They fail to capture dynamic

organization principles of regulatory network across cell types.

Our objectives in this chapter are two folds: (1) to study whether the distributional properties of FFL are distinctive in hESC network as compared with those in the differentiated tissue/cell types; and (2) to identify TFs that are extensively regulated by FFL in the hESC network only.

In this chapter, we compare local regulatory landscape on each TF in terms of FFLs in regulatory network of hESC with those in the other 40 differentiated cell types reported by Neph et al. (2012a). Firstly we find that distributional properties of FFL regulating each TF can recapture embryonic origin and classify known cell-lineage relationship well. Secondly, we identify 28 TFs extensively regulated by FFLs in hESC only. Among them 13 TFs perform hESC related functions, and the remaining 15 TFs are master TFs in various differentiated cell types. Thirdly, our proposed scores perform better in identifying hESC related TFs than FFL-based centrality measures in Koschützki et al. (2007).

4.2 Materials and Methods

4.2.1 FFL count matrices

We constructed an FFL count matrix $MC = [mc_{i,j}]_{1 \le i \le 475, 1 \le j \le 41}$: the ij-th element $mc_{i,j}$ is the number of times TF i is regulated by FFLs in network j. In other words, the ij-th element represents the number of times TF i taking position C in FFLs in network j. Seven TFs, GCM1, HNF4G, POU1F1, PROP1, SPZ1, SPY and TFDP2, are not regulated at all by any FFL in the 41 networks. As a result, the rows in MC for these TFs have constant value 0, and then we removed constant rows of 0's in matrix MC

before further analysis. Without loss of generality, we label the network of hESC as network 1 and the other 40 networks as network j ($2 \leq j \leq 41$). Table 4.1 illustrates a portion of FFL count matrix MC.

Table 4.1. A portion of FFL count matrix MC. Values are numbers of FFLs regulating each of 475 TFs in the 41 networks. Abbreviation: H7, h7-ESC; BL1, B-Lymphocyte; HEM, hematopoietic stem cell; BL2, B-Lymphoblastoid; ERY, erythroid; PRO, promyelocytic leukemia; TLY, T-Lymphocyte; HEP, hepatoblastoma; NEU, neuroblastoma.

	hESC	Blood							Cancer		···
	H7	BL1	HEM	BL2	BL2	ERY	PRO	TLY	HEP	NEU	···
OTX2	431	0	0	0	0	0	0	0	0	0	···
POU5F1	495	0	0	0	0	0	0	0	0	0	···
ZFP42	309	0	0	0	0	6	0	0	0	0	···
ZIC3	263	0	0	0	0	0	0	0	0	3	···
FOXD3	920	27	154	1	224	157	0	277	0	6	···
SIX3	592	146	2	0	123	0	452	4	0	191	···
SOX2	457	67	58	170	1	0	1	57	79	67	···
NANOG	24	0	0	0	0	0	53	0	2	0	···
KLF4	82	234	282	75	145	16	152	187	127	0	···
STAT3	243	249	232	262	411	194	206	292	213	226	···
PAX4	69	0	0	0	0	0	0	0	0	0	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Furthermore, we introduced normalized versions of MC, denoted by MC^c and MC^r. Since total occurrences of FFLs in the 41 networks has an extremely wide range: from 27264 in epithelia cell HRCEpiC to 122646 in blood cell NB4. MC^c is derived from normalizing MC in a way that each column is divided by the total of this column. Each column in MC^c sums to 1 and is a distribution of the relative frequencies of FFLs regulating 475 TFs in the corresponding network. MC^c will be used as the input for hierarchical clustering on cell types.

Then MC^r is derived from standardizing MC^c in a way that each row is subtracted from its empirical mean and divided by its empirical standard deviation. Thus each row of MC^r is z-score of corresponding row in MC^c

and has mean of 0 and standard deviation of 1. A higher z-score of a TF in network j_1 than in network j_2 means a higher number of FFLs regulating this TF in network j_1 than in network j_2. MC^r will be used as the input for principal component analysis on cell types.

Similarly we can construct FFL count matrices MA, MB, and $MSum$, where the $ij\text{-}th$ element is the number of times TF i taking position A, taking position B, and involved in FFLs in network j, respectively. Here $MSum = MA + MB + MC$. We define their normalized versions in a similar fashion.

4.2.2 TFs extensively regulated by FFLs in hESC only

We introduced a score, denoted by RC, to quantify to what extent a TF is regulated by FFLs in hESC only. For a TF i, the score

$$RC_i = mc_{i,1} / \max_{2 \le j \le 41} \{mc_{i,j}\},$$

that is, RC_i is the ratio of the number of FFLs regulating TF i in hESC network to the maximum number of FFLs regulating TF i in the other 40 tissue/cell-type networks. TFs with scores exceeding a threshold, which is to be chosen suitably, are defined as TFs extensively regulated by FFLs in hESC only. Following a 2-fold gene expression analysis practice, we chose threshold of 2. To determine the significance of TFs extensively regulated by FFLs in hESC only at threshold 2, we found that the distribution of number of FFLs regulating each TF in the 41 networks can be approximated by a lognormal distribution (Figure 4.1). Let X_j be the number of FFLs regulating a TF in network j for $1 \le j \le 41$. Let Φ and ϕ be the distribution and density functions of a standard normal distribution. We

assumed that X_j independently follows a lognormal distribution. Let $t > 0$ and represent $X_j = e^{\mu_j + \sigma_j Z_j}$ where Z_j's are independent standard normal random variables. Write $s = \log t$. We have

$$\begin{aligned}
& P(X_1 / \max_{2 \le j \le 41} \{X_j\} > t) \\
= \; & P(X_1 > t \max_{2 \le j \le 41} \{X_j\}) \\
= \; & P(X_1 > tX2, \ldots, X_1 > tX_{41}) \\
= \; & P(\mu_1 + \sigma_1 Z_1 > s + \mu_2 + \sigma_2 Z_2, \ldots, \mu_1 + \sigma_1 Z_1 > s + \mu_{41} + \sigma_{41} Z_{41}) \\
= \; & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \prod_{j=2}^{41} \Phi\Big(\frac{\mu_1 - \mu_j + \sigma_1 z - \log t}{\sigma_j}\Big) dz
\end{aligned}$$

Parameters $\hat{\mu}_j$, $\hat{\sigma}_j$ were estimated from MC by equation (4.1)

$$\hat{\mu}_j = \ln\left(\frac{m_j^2}{\sqrt{v_j + m_j^2}}\right) \text{ and } \hat{\sigma}_j^2 = \ln\Big(1 + \frac{v_j}{m_j^2}\Big), \tag{4.1}$$

where m_j and v_j are the mean and variance of the j-th column in MC. Plugging in $t = 2$ and the estimated μ_j and σ_j, we have $P(X_1/\max_{2 \le j \le 41}\{X_j\} > 2) = 4.6e\text{-}18$. An extremely small probability indicates that TFs extensively regulated by FFLs in hESC only at threshold 2 are most likely not caused by chance. This phenomenon may be attributed to organization principles and dynamic properties of regulatory networks that maintain self-renewal and pluripotency of hESC. Considering that the 41 cells belong to 8 cell types, the independence assumption between $X_i, 1 \le i \le 41$, may not be satisfied.

Similarly we introduced RA, RB and $RSum$ based on matrices MA, MB, and $MSum$ respectively. Then we can quantify TFs extensively taking position A in FFLs in hESC only, TFs extensively taking position B in

FFLs in hESC only, and TFs extensively involved in FFLs in hESC only.

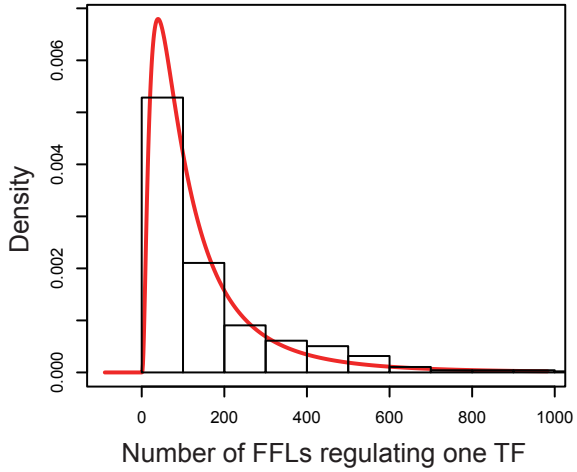


Figure 4.1. Histogram and fitted log-normal density curve of number of FFLs regulating each TF in the regulatory network of hESC.

4.2.3 hESC specific TF lists

In a meta-analysis, Assou et al. (2007) compiled a list of 1076 genes that are overexpressed in hESCs by at least three studies. Among them 29 are found in 475 TFs of the 41 networks. We labeled the list of these 29 TFs as "Assou TFs"

Super-enhancers are large collections of transcriptional enhancers. Genes with super-enhancer domain play important roles in the control of cell identity and diseases (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). In mouse and human ESCs, master TFs OCT4, SOX2, NANOG are each encoded by a gene with super-enhancer and also have DNA binding motifs that are often found in super-enhancer domains (Whyte et al., 2013). We used "Master TFs" to label 24 TFs out of 475 TFs which are encoded by hESC-specific genes with super-enhancer based on super-enhancers reported in Hnisz et al. (2013).

"Duplicated TFs" denotes a list of 8 TFs which belong to both the "Assou TFs" and "Master TFs". "Combined TFs" denotes a list of 45 TFs which is the union of the two lists "Assou TFs" and "Master TFs".

4.3 Results

4.3.1 FFLs in regulatory networks globally distinguish hESC from the other 40 differentiated cell types

Considering that FFLs play multiple important functions in regulatory networks and hESC represents a common developmental ancestor to the other differentiated tissue/cell-types, it is interesting to investigate whether hESC can be distinguished from the other 40 differentiated tissue/cell-types by FFLs in regulatory networks, and whether FFLs can recover known cell-lineage relationships between the 41 tissue/cell-types. To answer these questions, we first constructed an FFL count matrix MC whose $ij\text{-}th$ element is number of FFLs regulating TF i in network j. Then we calculated distances between cell types by the Manhattan distance of MC^c which is the normalized MC (Section 4.2.1). Next hierarchical clustering was carried out with complete linkage method. Hierarchical clustering has RI=0.69 and produced a dendrogram that reproduced known cell-lineage relationship with remarkable detail, as well as broader features of embryonic origin. On a gross level, hESC was the root of the dendrogram, and functional or anatomical related cells were in one major cluster group, e.g. blood cells, cancer cells, endothelia cells, fetal tissues and stromal cells (Figure 4.2A). hESC was also the root in dendrograms produced by hierarchical clustering with a number of linkage methods (Figure 4.3).

To confirm these observations, we applied principal component analysis on MC^r (Section 4.2.1). The first two principal components (PC1 and PC2) together explain 21.4% of total variance of MC^r. The scatterplot of PC1 and PC2 clearly shows the distinctiveness of these major cluster groups (Figure 4.2C). The scatterplot also reveals that hESC is far away from the other 40 tissue/cell-types.

These results together suggest that regulatory networks of functional or anatomical related cells share similar local organization principles. So local structures could shed light on cell type related TFs.

4.3.2 *Netdis* and FFL based measure produce comparable cell type classification

Ali et al. (2014) proposed an alignment-free distance measure $netdis \in [0, 1]$ to compare two simple undirected networks. Given two query networks and a gold-standard network, the authors first counted occurrences of all k-node induced subgraphs, $k = 3$ or 4, in the two-step ego graph of each node. Two-step ego graph of a node is the subgraph induced by this node and its neighbours within two edges. *Netdis* of the two query networks is constructed after k-node induced subgraphs counts are normalized by those in the gold-standard network. It is demonstrated to be able to reconstruct phylogenetic tree of species and separate different random network models.

To apply *netdis* to the 41 TF regulatory networks, we first converted them to simple undirected networks by removing self-loops and duplicated edges. Then we iteratively chose a network as a gold-standard network and classified the other 40 networks by hierarchical clustering with ward method (Ward, 1963) on a distance matrix built from *netdis* between re-

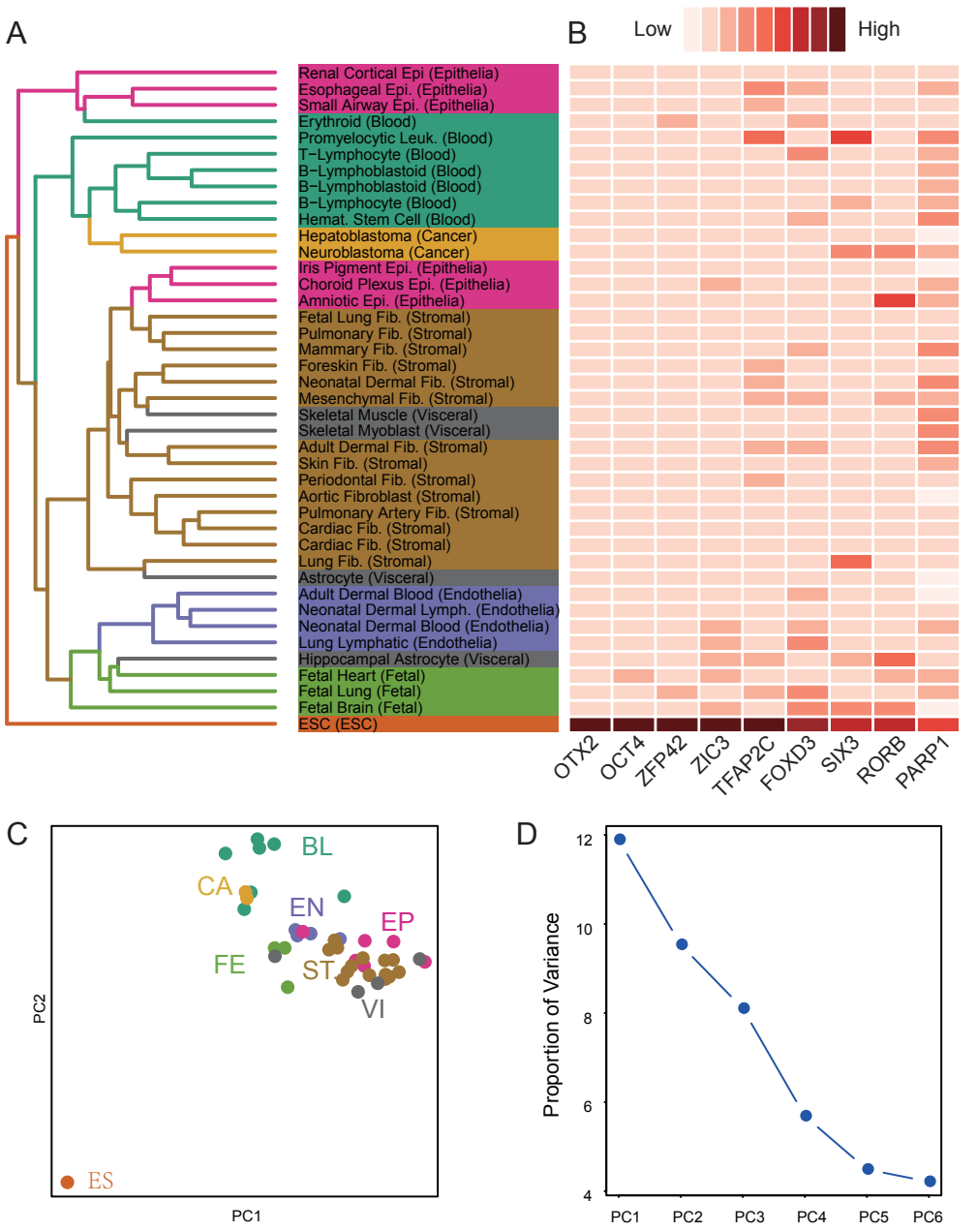


Figure 4.2. (A) Hierarchical clustering of the 41 cell types based on MC^c. It has RI=0.69. (B) z-score of number of FFLs regulating master TFs in the 41 networks. For a given TF and cell type, high z-score (dark color) indicates this TF is regulated by large number of FFLs in that cell type. For example, pluripotent marker OCT4 is regulated by most FFLs in hESC than in the other 40 cell types. (C) Scatterplot of first 2 principal components (PC1 and PC2) from MC^r. (D) Proportion of variance explained by the first 6 PCs. PC1 and PC2 explained 21.4% of total variance. Abbreviations: BL, blood; CA, cancer; EN, endothelia; EP, epithelia; ES, ESC; FE, fetal; ST, stromal cells; VI, visceral cells.

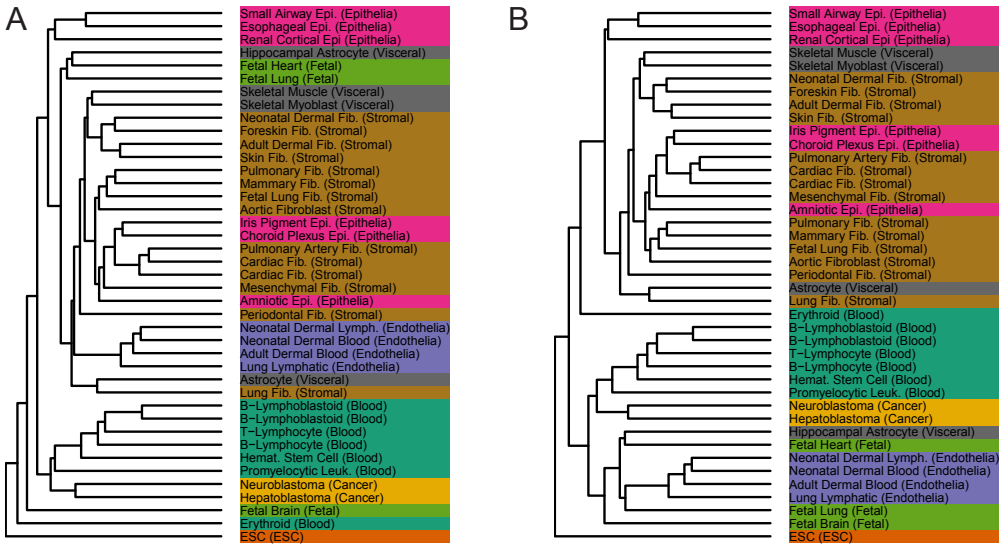


Figure 4.3. Dendrograms produced by hierarchical clustering with linkage *Method="average"* (A) and *Method="mcquitty"* (B) in *hclust* function in R. The classifications have RI=0.49 (A) and RI=0.85 (B).

spective networks. The clusterings produced by *netdis* are comparable with the results based on FFL count. Because there is no significant difference between RIs based on *netdis* and RI based on FFL count. Table 4.2 reported the five-number summary of RI based on *netdis*. The median RI is 0.603 and 0.565 for k=3 and 4, respectively. The maximum RI=0.74 is obtained when $k = 4$ and the network of fetal brain is the gold-standard network (Figure 4.4).

4.3.3 TFs extensively regulated by FFLs in hESC only carry out important hESC specific functions

We next investigated differences on local regulatory landscapes between hESC and the other 40 tissue/cell-types. Among TFs in "Combined TFs" list (Section 4.2.3), OTX2, OCT4, ZFP42, ZIC3, TFAP2C, FOXD3, SIX3, RORB, and PARP1 are extensively regulated by FFLs in hESC when com-

Table 4.2. Five-number summary of RIs from hierarchical clusterings based on distance matrices produced by *netdis*. We iteratively chose one out of the 41 networks as a gold-standard network and constructed pair-wise *netdis* with k=3 or 4 for remaining 40 networks. Then we performed hierarchical clustering with Ward method and computed RI for resulting clustering.

	Minimum	Lower quartile	Median	Upper quartile	Maximum
$k = 3$	0.415	0.517	0.603	0.659	0.732
$k = 4$	0.413	0.5	0.565	0.621	0.74

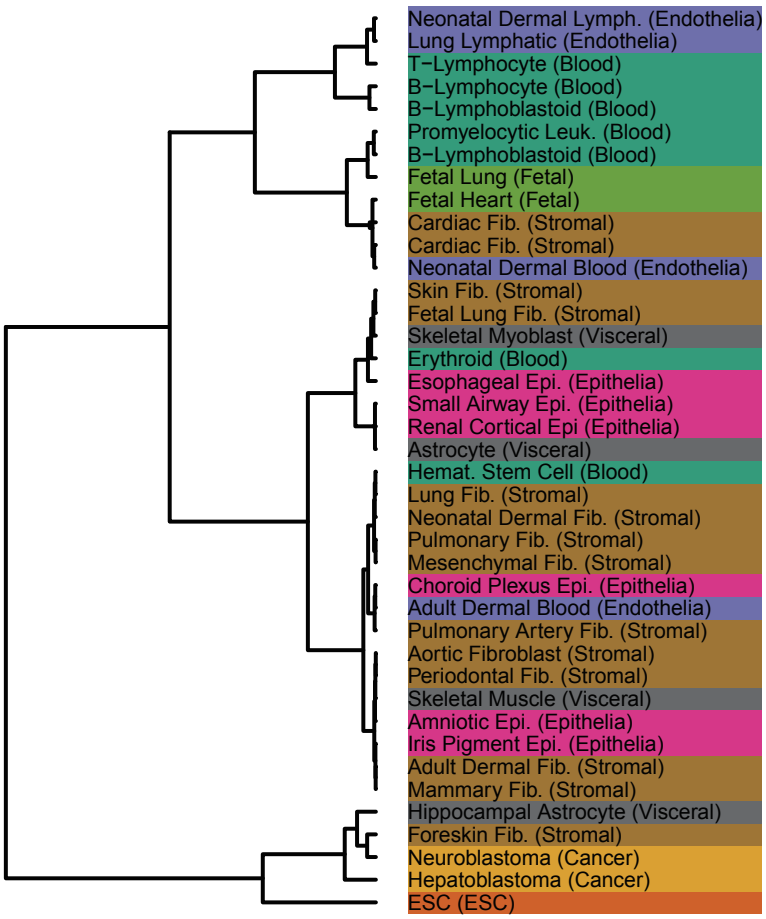


Figure 4.4. Dendrogram produced by hierarchical clustering based on a distance matrix produced by *netdis* (Ali et al., 2014). The network of fetal brain is used as the gold-standard network for *netdis*. The clustering has RI=0.74. The classification is comparable with the result (Section 4.3.1) produced by the distributional properties of FFL (RI=0.69).

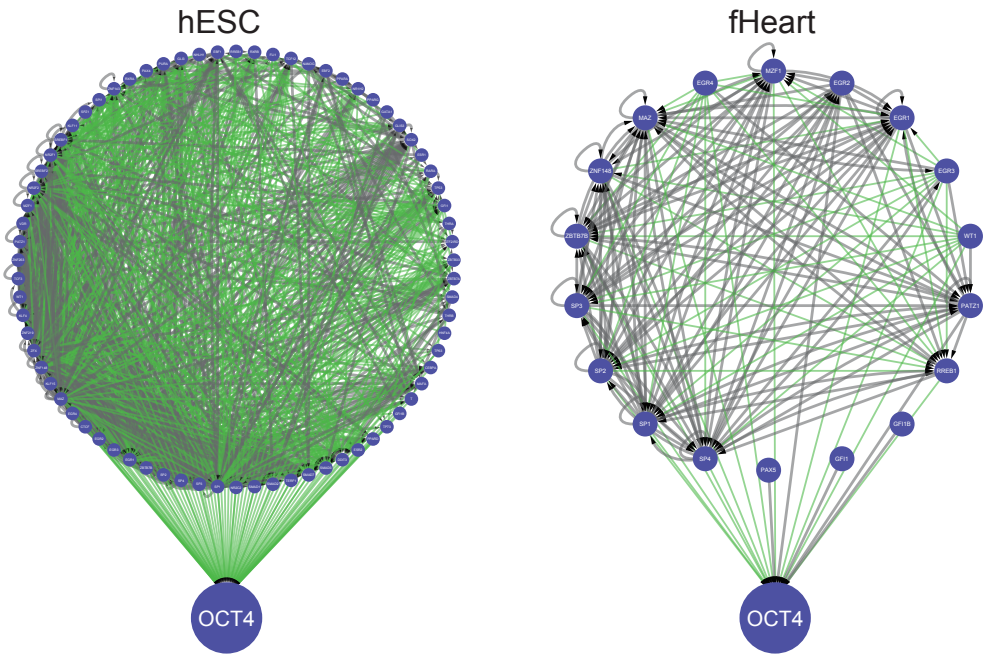


Figure 4.5. (A) Subgraph induced by OCT4 and its upstream neighbours (76) in the regulatory network of hESC. There are 495 FFLs regulating OCT4 in this subnetwork. (B) Subgraph induced by OCT4 and its upstream neighbours (18) in the network of fetal heart (fHeart). There are 32 FFLs regulating OCT4 in this subnetwork. Interactions involving in FFLs are colored in green.

pared against other cell types (Figure 4.2B). For example, OCT4 was regulated by 495 FFLs in hESC, while it were regulated by only 32 FFLs in fetal heart, and it was not even regulated by FFLs in the other 39 tissue/cell-types. Induced subgraphs by OCT4 and its upstream neighbours in hESC (Figure 4.5A) and in fetal heart (Figure 4.5B) clearly show that OCT4 was extensively regulated by FFLs in hESC only, indicating that the local regulatory landscape of OCT4 in hESC is very different from that of the other 40 cell types. Given the fact that OCT4 is a master TF for pluripotency in hESC (Young, 2011), we further asked whether other TFs extensively regulated by FFLs in regulatory network of hESC only also perform important functions in hESC.

To answer this question we used the score, RC, defined earlier for each TF. We defined TFs with score not less than 2 as TFs extensively regulated by FFLs in hESC only (Section 4.2.2). Totally 28 TFs are identified. We denoted the list of these 28 TFs by TFC.

We first did enrichment analysis of TFC in TFs in hESC specific TFs lists (Section 4.2.3). Overall TFC is significantly enriched in hESC specific TFs. Detailed results are listed below.

1. TFC is significantly (p-value=0.039) enriched in hESC specific TFs list "Combined TFs". There are 6 TFs common to TFC and "Combined TFs". They are FOXD3, POU5F1, TFAP2C, ZFP42, ZIC3 and OTX2.

2. TFC is significantly (p-value=0.005) enriched in hESC specific TFs list "Assou TFs". There are 6 TFs common to TFC and "Assou TFs". They are FOXD3, POU5F1, TFAP2C, ZFP42, ZIC3 and OTX2.

3. TFC is not significantly (p-value=0.161) enriched in hESC specific TFs list "Master TFs". There are only 3 TFs common to TFC and "Master TFs". They are FOXD3, OCT4 and ZIC3.

4. TFC is significantly (p-value=0.008) enriched in hESC specific TFs list "Duplicated TFs". There are 3 TFs common to TFC and "Duplicated TFs". They are FOXD3, OCT4 and ZIC3.

The hESC specific TF lists do not necessarily include all TFs that play some functions related to hESC. Thus we searched functions of TFs in TFC by Google Scholar. Totally 13 TFs in TFC have been reported in literature to perform hESC related functions. These TFs are ALX1, CDX2, DMRT1, FOXD3, HOXB13, LMX1A, LMX1B, NKX2-2, OTX2, OCT4,

PAX4, ZFP42, and ZIC3. Tables 4.3 and 4.4 list important functions played by these 28 TFs. Table 4.4 lists TFs that are uniquely regulated by FFLs in hESC, no FFL regulates these TFs in the other tissue/cell-types.

Table 4.3. TFs extensively regulated by Feed-Forward Loops (FFLs) in hESC regulatory network only.

	TF	Score	Function	Reference
Development	PAX7	64.8	Plays a central role in muscle development	Seale et al. (2000)
	OTP	33	Involved in brain development and neuronal differentiation. OTP has been identified as specifically required for development of the A11 DA group in mice	Ryu et al. (2007)
	ALX1	10	Encoded gene expressed selectively in chondrocyte lineage during embryonic development	Beverdam and Meijlink (2001)
	ZFP42	8.4	Specific to very early stages of development in hESCs	Rogers et al. (1991)
	PAX2	7.4	Relates to midbrain and eye development	Bäumer et al. (2003)
	FOXI1	5.8	In zebrafish, Foxi1 is required for cells to respond to FGF signalling in patterning the developing ear and jaws	Solomon et al. (2003)
	VSX1	3.6	Associated with development and maintenance of ocular tissues, which expressed in embryonic craniofacial	Semina et al. (2000)
	SIX6	2.9	Involved in specification and morphogenesis of the eye in the first few weeks of human development	Jean et al. (1999)
	CDX2	2.4	One of trophectoderm markers and markedly up-regulated upon POU5F1 reduction	Loh et al. (2006)
	HOXB13	2.4	In mouse ESCs, HOXB13 is involved in tail and neuronal development	John et al. (2004)
	TFAP2C	2.3	Activate genes involved in a large spectrum of important biological functions including proper eye, face, body wall, limb and neural tube development. They also suppress a number of genes including MCAM/MUC18, C/EBP alpha and MYC	Safran et al. (2010)
	ARX	2	Required for normal brain development. May be important for maintenance of specific neuronal subtypes in the cerebral cortex and axonal guidance in the floor plate	Safran et al. (2010)
Differentiation	POU4F3	10	Essential for hair cell differentiation and maintenance	Kim et al. (2002)
	NKX2-2	4.3	Markedly induced during differentiation in ESCs. It is a marker of the endocrine lineage	Xiang et al. (1997)
	POU2F3	3.3	A member of the Oct transcription factor family. It is involved in keratinocyte and epidermal differentiation	Jensen (2004)
	PAX4	3.2	Controls endocrine cell differentiation, it promotes the development of insulin-producing cells such as pancreatic cell during differentiation of mouse ESCs	Andersen et al. (1997)
	LMX1B	3	Associated with control dopaminergic differentiation, LMX1B is a key TF in directed differentiation of dopaminergic neuronal subtypes from human ESCs	Blyszczuk et al. (2003)
	ATOH1	2.3	Plays a role in the differentiation of subsets of neural cells by activating E box-dependent transcription (By similarity)	Safran et al. (2010)
Pluripotency	POU5F1	15	Crucial for ESC self-renewal and pluripotency	Young (2011)
	ZIC3	8.3	Required for maintenance of pluripotency in ESCs	Lim et al. (2007)
	FOXD3	3.3	Important in maintaining pluripotency of mouse ESCs, and is specific to the very early stages of development in hESCs	Hanna et al. (2002)

Table 4.4. TFs regulated by FFLs in regulatory network of hESC only.

TF	NO. FFLs	Function	Reference
OTX2	431	Associated with early pan-neural epithelium in day 7 embryoid bodies	Goulburn et al. (2011)
CRX	335	Belonging to homebox family and is one photoreceptor marker	Safran et al. (2010)
DMRT3	221	Plays key roles in neurogenesis	Bellefroid et al. (2013)
LMX1A	190	Plays a pivotal role in the mDA differentiation of human ESCs	Cai et al. (2009)
DMRT1	43	One sex determination gene, DMRT1 is present during embryogenesis in Sertoli and germ cells	Raymond et al. (2000)
TBX22	37	Part of one pathway to regulate mammalian palate development	Liu et al. (2008)
ESX1	6	A trophoblast-specific transcription factor, regulating placental development and fetal growth	Li and Behringer (1998)

4.3.4 Significance of TFs extensively regulated by FFLs in hESC only

We tested the significance of TFs extensively regulated by FFLs in hESC only by procedures listed below. Let $N = 10,000$ and $1 \leq i \leq 28$ stands for the 28 TFs extensively regulated by FFLs in hESC only. Firstly we generated one random network for the network of hESC by randomly rewiring its interactions while preserving the degree sequence. We counted number of FFLs regulating TF i in this random network and denoted it by x^i. Secondly we generated one random graph for each of the other 40 networks in the same way and counted numbers of FFLs regulating TF i in the resulted 40 random networks. We denoted the maximum number by y^i. We repeated the above two steps N times and constructed two vectors $X^i = (x^i_k)$ and $Y^i = (y^i_\ell)$, $1 \leq k, \ell \leq N$ for TF i. We calculated p-value, denoted by p^i, for TF i extensively regulated by FFLs in hESC only by equation (4.2).

$$p^i = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{\ell=1}^{N} \mathbb{1}(x^i_k > 2y^i_\ell). \tag{4.2}$$

P-values for TFs extensively regulated by FFLs are not greater than 0.0001, suggesting these TFs extensively regulated by FFLs are unlikely due to chance.

4.3.5 Comparison with motif centrality measures

Koschützki et al. (2007) proposed 4 centrality measures (4.3)-(4.6) based on FFL to quantify importance of each node in a directed network. A higher

value implicates greater importance of a node in the network.

$$ffl Sum = \text{Number of times a node involved in an FFL,} \tag{4.3}$$

$$ffl A = \text{Number of times a node taking position } A \text{ of an FFL,} \tag{4.4}$$

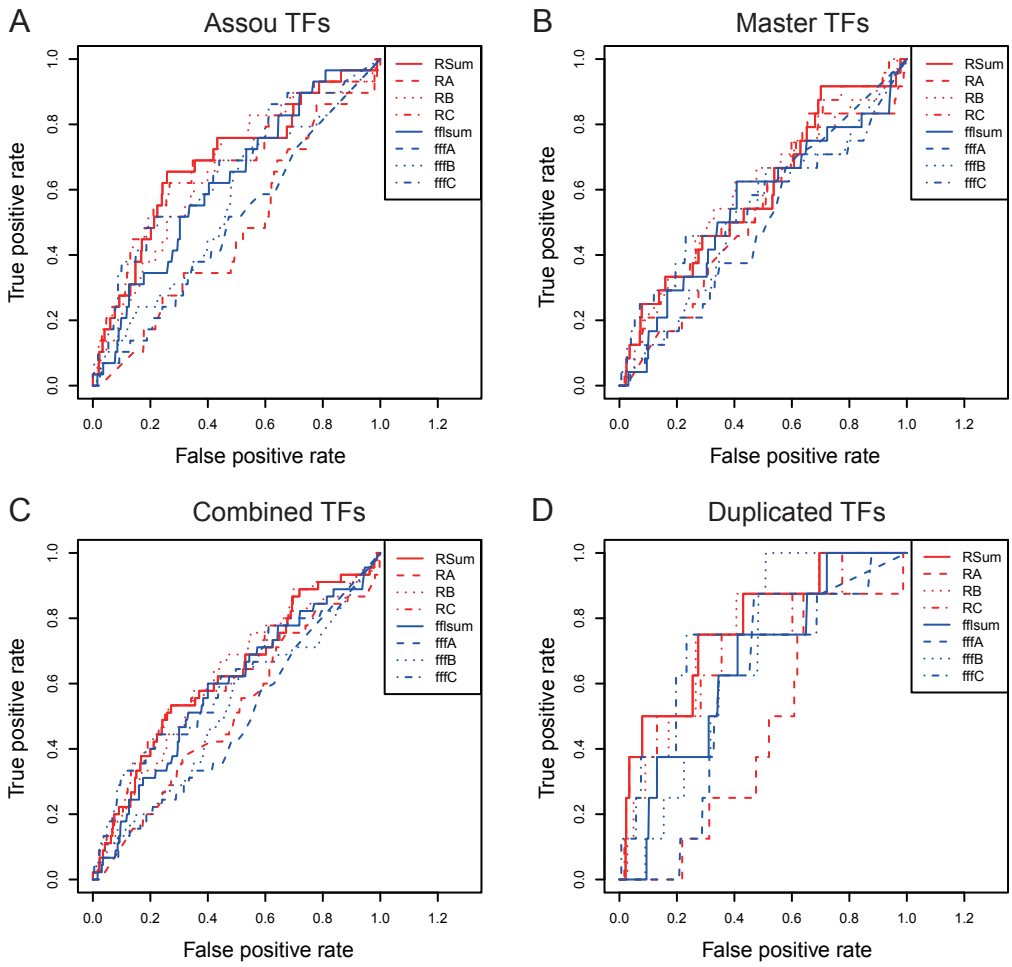
$$ffl B = \text{Number of times a node taking position } B \text{ of an FFL,} \tag{4.5}$$

$$ffl C = \text{Number of times a node taking position } C \text{ of an FFL.} \tag{4.6}$$

To identify hESC related TFs in regulatory network of hESC, Koschützki et al. (2007) proposed centrality measures is limited to TFs participating FFLs in regulatory network of hESC only. However, our proposed ratio-based scores $RSum$, RA, RB, and RC take into account TFs participating in FFLs in the network of hESC as well as in the networks of the other 40 differentiated cell types. We compared $RSum$ against $ffl Sum$, RA against $ffl A$, RB against $ffl B$, and RC against $ffl C$ in identifying hESC related TFs. Reference lists used for hESC related TF lists are "Assou TFs", "Master TFs", "Duplicated TFs ", and "Combined TFs" (Section 4.2.3). Receiver operating characteristic (ROC) curves (Figures 4.6A to D) and area under the curve (AUC) (Figure 4.6E) consistently demonstrate superiority of our proposed ratio based scores to the 4 centrality measures.

4.4 Conclusions

In this chapter, we contrasted the local regulatory landscape of each TF in terms of FFLs in regulatory network of hESC with the other 40 differentiated cell types reported by Neph et al. (2012a). We first found that the distributional properties of FFL regulating each TF can recapture embryonic origin and classify known cell-lineage relationship. These results



E

Area under the curve (AUC)

	Assou (29)	Master (24)	Combined (45)	Duplicated (8)
RSum	0.685	0.594	0.624	0.773
fflSum	0.624	0.554	0.583	0.655
RA	0.475	0.528	0.509	0.452
fflA	0.498	0.504	0.483	0.593
RB	0.657	0.599	0.611	0.754
fflB	0.555	0.526	0.517	0.675
RC	0.683	0.591	0.630	0.722
fflC	0.664	0.561	0.603	0.710

Figure 4.6. Receiver operating characteristic (ROC) curves and area under the curve (AUC). We compared $RSum$ against $fflSum$, RA against $fflA$, RB against $fflB$, RC against $fflC$ in identifying hESC related TFs in reference lists of "Assou TFs" (A), "Master TFs" (B), "Combined TFs" (C), and "Duplicated TFs"(D). (E) Area under the curve (AUC). ROC and AUC demonstrate superiority of $RSum$ to $fflSum$, RA to $fflA$, RB to $fflB$, RC to $fflC$.

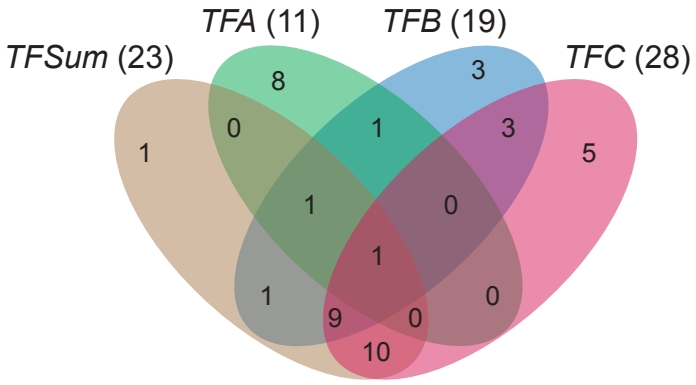


Figure 4.7. Venn diagram between TFs extensively involving in FFLs, taking positions A, B, or C in FFL in hESC only. The lists of TFs are labeled as $TFSum$, TFA, TFB, and TFC respectively. Interesting the 4 lists of TFs have many common TFs. Especially TFC and $TFSum$ have 20 common TFs, TFC and TFB have 13 common TFs. But TFC and TFA only has 1 common TF (ESX1). Total number of TFs in each list is given in parentheses.

together suggest that regulatory networks of functionally or anatomically related cells share similar local organization principles. Local structures could shed light on identifying cell type related TFs.

Next we identified 28 TFs extensively regulated by FFLs in hESC only. These 28 TFs are significantly extensively regulated by FFLs in hESC only as evidenced by simulation results. Among them ALX1, CDX2, DMRT1, FOXD3, HOXB13, LMX1A, LMX1B, NKX2-2, OTX2, OCT4, PAX4, ZFP42, and ZIC3 perform hESC related functions. Even though remaining 15 TFs are not evidenced to carry out direct functions in hESC, they are demonstrated to play multiple important roles in differentiated cell types (Tables 4.3 and 4.4). This may indicate FFLs play roles in repressing the expression of key TFs encoded genes of differentiated cell types in hESC to maintain self-renewal and pluripotency of hESC.

TFs extensively regulated by FFLs in hESC only can be generalized to TFs extensively taking positions A, B, or involved in FFLs in hESC only.

Interestingly, a remarkable number of TFs can be found in at least 2 lists, and only a small number of TFs can only be found in one list (Figure 4.7). The list of TF's extensively taking position B in FFLs in hESC only at threshold 2 has 19 TFs. Among them 13 are also extensively regulated by FFLs in hESC only, including OCT4, OTX2, ZFP42, which are well-known ESC markers The other 6 TFs are ALX3, EVX1, LHX4, MNX1, SOX17, and T. ALX3 involves in cell-type differentiation and development. EVX1 may play an important role as a transcriptional repressor during embryogenesis. LHX4 is involved in the control of differentiation and development of the pituitary gland. SOX17 involves in the regulation of embryonic development and in the determination of the cell fate. T is an embryonic nuclear TF. It effects transcription of genes required for mesoderm formation and differentiation. One master TF of hESC is SOX2 and it is extensively taking position B in FFLs in hESC only at threshold 1.

TFs extensively taking position A in FFLs in hESC only at threshold 1 has 11 TFs including GATA1, GATA2, HOXC5, and TBX5. Interestingly only one TF (ESX1) is also extensively regulated by FFLs in hESC only. GATA1 and GATA2 play an essential role in regulating transcription of genes involved in the development and proliferation of hematopoietic and endocrine cell lineages. HOXC5 plays an important role in morphogenesis in all multicellular organisms. TBX5 may play a role in heart development and specification of limb identity.

TFs extensively involved in FFLs in hESC only at threshold 2 has 23 TFs with 20 of these 23 TFs are also found in TFs extensively regulated by FFLs in hESC only. The other 3 TFs are ALX3, BARHL2, and EVX1. As discussed above, ALX3 and EVX1 play some functions related

to ESCs. Functions of these TFs presented in this section were extracted from GeneCards encyclopedia (www.genecards.org, Safran et al. (2010)).

Thirdly, we compared $RSum$ versus $fflSum$, RA versus $fflA$, RB versus $fflB$, and RC versus $fflC$ in identifying hESC related TFs. ROC and AUC consistently demonstrate superiority of our proposed ratio based scores, $RSum$, RA, RB, and RC, to the FFL-based centrality measures $fflSum$, $fflA$, $fflB$, and $fflC$ (Koschützki et al., 2007).

Advantage of *Netdis* is that it counts all k-node induced subgraphs in the two-step ego graph of a node i. Disadvantage of applying *netdis* in measuring pairwise distance among the 41 directed networks is that it is originally designed for undirected networks. Applying *netdis* requires us to remove the direction of regulation, key information contained in a TF regulatory network. In contrast, advantage of our FFL based method is that it naturally takes into account the regulation direction attribute. But the disadvantage of our FFL based method is that it only counts FFLs involving in the node i in the two-step ego graph. The advantage and disadvantage in both *Netdis* and our FFL based methods may be the reason why the two methods produce comparable classification results.

The 41 regulatory networks were constructed using DNaseI footprinting technology, which possibly contain some spurious links and missing links. TFs extensively regulated by FFLs in hESC only are defined by choosing threshold of 2. These TFs are most likely to be regulated by a higher number of FFLs in the network of hESC than in the networks of the other cell types. Also hierarchical clustering results is confirmed by a few linkage methods and by principal component analysis. Our results should not be greatly affected by the limitations of the data chosen.

The 41 regulatory networks are from 8 cell/tissue types. Studying associations between FFLs and master TFs in the other 7 cell/tissue types is an interesting future work.

Chapter 5

Conclusion and Future Work

Many biological networks have become available in the recent decade. Designing methods to compare and analyze them will enhance understanding of the biological systems at system level. Studies (Alon, 2007; Barabasi and Oltvai, 2004; Jothi et al., 2009; Neph et al., 2012a) show that the topological properties of a complex biological network often unravel its global and local organization structure, and functionally similar nodes. The three results (Chapters 2-4) represent our attempts to explore the relationship of topological structures and biological functions. Our works in Chapter 2 on f-Wiener type indices stem from our purpose to provide summarize statistics for a given network. We also underscore the need to normalize these indices for the objectives to compare networks which often have different number of nodes. In chapter 3, we compare in greater detail about the global and local organization principles of the 41 human cell type networks. We discover similar as well as distinct structures across them. In Chapter 4, based on an important network motif, FFL, we compare and contrast the distributional properties of FFL in these networks. We describe below some of our findings.

5.1 Conclusion

5.1.1 f-Wiener index

Wiener index and other Wiener type indices have been commonly applied in Chemometrics to associate structures and physicochemical properties of molecules. Recently, these indices are incorporated in quantifying complex networks as in QuACN and NetCAD. In chapter 2, we first generalized Wiener index to a general functional form, called f-Wiener index. This f-Wiener index contains all well-known Wiener type indices as special cases. We provided a unifying method to identify the maximum and minimum over the set of simple connected graphs with n nodes, or the set of simple connected trees with n nodes (Theorems 1 and 2). Explicit sharp upper and lower bounds for Wiener index, Harary index, hyper Wiener index and the generalized index were deduced over networks (Corollary 5) and over trees (Corollary 6). Moreover, the maximizer and minimizer were characterized in Theorems 1 and 2. We believed these results are general and of independent interests.

Armed with these maximum and minimum values, we proposed a normalized version of f-Wiener index over networks, and a similar version over trees. These normalized versions provide better interpretation of indices over networks of varying number of nodes than the non-normalized one. The normalized versions capture similar topological structures among networks with different number of nodes better, evidenced by significant improvement in network classification in five simulations.

Our method of optimizing $W_f(G)$ can be easily extended to index of the form $\Phi(W_f(G))$ where Φ and f are monotone functions. For example,

taking $\Phi(x) = 1/x$ and $f(k) = \frac{2}{n(n-1)k}$ leads to $\Phi(W_f(G)) = \frac{n(n-1)}{2\sum_{i<j} 1/d(i,j)}$ which measures small-world behavior of network G (Newman, 2002).

5.1.2 Profiling TF regulatory networks of human cell types

In chapter 3, we have studied the organizational architectures of the 41 human cell-type TF regulatory networks that were reported by Neph et al. (2012a). First, we have showed that the wiring around five to seven TFs in the networks can be used to classify all the 41 cell types well. Both Neph et al. (2012a) and our studies indicate that the human TF regulatory networks are different globally as well as locally.

We have examined the three-layer hierarchical organizations of the human cell-type TF regulatory networks. The networks are each partitioned into the top, core and bottom layers, containing 23%, 67% and 10% of TFs on average, respectively. The same hierarchical analysis (Jothi et al., 2009) indicates that in the yeast TF regulatory networks both the core and bottom layers have similar sizes (43% vs 40%) whereas the top layer contains only 13% of the TFs. Taken together, these two facts imply a difference in the topological organizations between the human and yeast TF regulatory networks.

Enrichment analyses indicate that for each TF regulatory network of the 40 non-ESC cell types, (a) the top layer is lacking in both hub and HK TFs, (b) the core layer is enriched with both hubs and HK TFs and (c) the bottom layer is only enriched with hub TFs. However, essential TFs seem to be distributed evenly in the top and core layers, but, sparsely in the bottom layers. Interestingly, the hESC TF regulatory network has a topological structure that is different from the rest. It has significantly

small top and bottom layers and therefore a large core layer. Its top layer is neither enriched with nor depleted in hub, essential and HK TFs.

We have also studied the dynamic properties of the human cell-type TF regulatory networks. The HK interactions are related to basic life support such as bio-molecular synthesis and transcription mechanisms. One of our findings is that most HK interactions are within the core layer or between the core and bottom layers.

The ESCSN, the subnetwork induced by specific links in the hESC TF regulatory network, has also been investigated. The 82 hub TFs in the ESCSN seem to play important roles in hESCs due to the following facts: (i) their genes are overexpressed, (ii) they are enriched in the Activin A/Nodal signaling pathway, and (iii) specific interactions are enriched in a core transcriptional regulatory network of the hESCs. In one hESC- specific regulatory complex-target module, both KLF4 and ZFX have three common downstream targets: FOXD3, OCT4 and ZFP42 in ESCSN. Notably KLF4, ZFX and their targets are important in the maintenance of pluripotency, self-renewal and development processes in ESCs. All these facts together suggest that specific regulatory interactions may play important roles in hESCs.

In general, specific regulatory interactions are difficult to detect because the network of each cell type is based on independent data, leading to a high false negative rate. Since the number of specific interactions in hESCs is much higher than that in other cell types, our results should not be greatly affected by the limitations of the data chosen.

5.1.3 Profiling Human Embryonic Stem Cell via Feed-Forward Loops in Transcription Factor Regulatory Network

In chapter 4, we compared local regulatory landscape on each TF in terms of FFLs in the regulatory network of hESC and in the other 40 differentiated cell types reported by Neph et al. (2012a). Firstly we found that distributional properties of FFL regulating each TF can reproduce embryonic origin and known cell-lineage relationship. These results together suggest that regulatory networks of functional or anatomical related cells share similar local organization principles. Local structures could shed light on identifying cell type related TFs. Moreover the hierarchical clustering of cell types by distributional properties of FFL regulating each TF is comparable with clustering based on network distances produced by *netdis* (Ali et al., 2014).

Secondly we identified 28 TFs extensively regulated by FFLs in hESC only. These 28 TFs are significantly extensively regulated by FFLs in hESC only as evidenced by simulation results. Among them ALX1, CDX2, DMRT1, FOXD3, HOXB13, LMX1A, LMX1B, NKX2-2, OTX2, OCT4, PAX4, ZFP42, and ZIC3 perform hESC related functions. The other 15 TFs are not evidenced to carry out direct functions in hESC. But they are demonstrated to play multiple important roles in differentiated cell types. This may indicate that interacting FFLs play roles in repressing expression of key TFs encoded genes of differentiated cell types in hESC to maintain self-renewal and pluripotency of hESC.

TFs extensively regulated by FFLs in hESC only can be generalized to TFs extensively taking position A, B, or involving in FFLs in hESC only. Interestingly, there are large number of TFs that are common to the 4 lists

of TFs. Only a small number of TFs are unique to each list.

Thirdly, we compared $RSum$ against $fflSum$, RA against $fflA$, RB against $fflB$, and RC against $fflC$ in identifying hESC related TFs. ROC and AUC consistently demonstrate superiority of our proposed ratio based scores, $RSum$, RA, RB, and RC, to the FFL-based centrality measures $fflSum$, $fflA$, $fflB$, and $fflC$ (Koschützki et al., 2007).

The 41 regulatory networks data set is produced based on DNaseI footprinting technology, which is believed to contain some spurious links and missing links. TFs extensively regulated by FFLs in hESC only are defined by choosing threshold at 2. These TFs are most likely to be regulated by a higher number of FFLs in hESC that in other cell types. Also hierarchical clustering results is confirmed by a few linkage methods and principal component analysis. Our results should not be greatly affected by the limitations of the data chosen.

5.2 Future work

5.2.1 f-Wiener index

Observe that $W_f(G) = \sum_{r=1}^{N(G)-1} f(r)n_r(G) = \sum_{r=0}^{N(G)-1}[f(r+1)-f(r)]N_r(G)$ where we assume $f(0) = 0$, $n_r(G)$ denotes the number of pairs of nodes in G with distance equals r, and $N_r(G)$ the number of pairs of nodes in G with distance greater than r. Since in most biological networks the number of nodes is large, one may normalize a scaled-version of $W_f(G)$ in terms of the asymptotic distribution of the N_r's under the assumption that the observed network G is generated by a given random network model $\mathcal{M}$. This will enable us to determine the likelihood that the observed network is gener-

ated by $\mathcal{M}$. Currently a fair amount of information about shortest paths in some network models is available in Barbour and Reinert (2011)and Fronczak et al. (2004). How to make use of these results seems like a worthwhile future project.

For other descriptors, it is of interest to study whether normalization is needed; if so, how best to normalize them; and to what extent normalization improves network comparison.

5.2.2 Profiling TF regulatory networks of human cell types

The 41 regulatory networks are produced based on DNaseI footprinting, which is believed prone to high false positive rate. More generally, network data produced by high-throughput technologies are prone to low coverage rate and inaccuracy. Thus how to cleanup these noisy network data is an interesting future work.

By integrating identified HK interactions and the protein complex database, we identified 23 protein complexes in which the proteins are highly connected with HK interactions. Most of the identified protein complexes are as predicted and hence it would be interesting to investigate their biological functions.

5.2.3 Profiling Human Embryonic Stem Cell via Feed-Forward Loops in Transcription Factor Regulatory Network

In this chapter we investigated network motif, FFLs, in identifying master TFs in the TF regulatory networks of hESC. There are other important networks motif, for example 4-node bi-fan. How to identify master TFs in regulatory networks based on other motifs is an interesting future work. The

41 regulatory networks are from 8 cell/tissue types. Studying associations between FFLs and master TFs in the other 7 cell/tissue types is another interesting future work.

Bibliography

Ali, W., Rito, T., Reinert, G., Sun, F., and Deane, C. M. (2014). Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17):i430–i437.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461.

Amaral, L. A. N., Scala, A., Barthelemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences, USA*, 97(21):11149–11152.

Andersen, B., Weinberg, W. C., Rennekampff, O., McEvilly, R. J., Bermingham, J., Hooshmand, F., Vasilyev, V., Hansbrough, J. F., Pittelkow, M. R., Yuspa, S. H., et al. (1997). Functions of the POU domain genes Skn-1a/i and Tst-1/Oct-6/SCIP in epidermal differentiation. *Genes & Development*, 11(14):1873–1884.

Arduini, B. L. and Brivanlou, A. H. (2012). Modulation of FOXD3 activity in human embryonic stem cells directs pluripotency and paraxial mesoderm fates. *Stem Cells*, 30(10):2188–2198.

Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004). Comment on "Network motifs: simple building blocks of complex net-

works" and "Superfamilies of evolved and designed networks". *Science*, 305(5687):1107–1107.

Assou, S., Le Carrour, T., Tondeur, S., Ström, S., Gabelle, A., Marty, S., Nadal, L., Pantesco, V., Réme, T., Hugnot, J.-P., et al. (2007). A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells*, 25(4):961–973.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291.

Bader, G. D., Betel, D., and Hogue, C. W. (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250.

Balaban, A. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89(5):399–404.

Balazsi, G., Barabási, A.-L., and Oltvai, Z. (2005). Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proceedings of the National Academy of Sciences, USA*, 102(22):7841–7846.

Banerjee, N. and Zhang, M. Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Research*, 31(23):7024–7031.

Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabasi, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Barbour, A. D. and Reinert, G. (2011). The shortest distance in random multi-type intersection graphs. *Random Structures & Algorithms*, 39(2):179–209.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390.

Bäumer, N., Marquardt, T., Stoykova, A., Spieler, D., Treichel, D., Ashery-Padan, R., and Gruss, P. (2003). Retinal pigmented epithelium determination requires the redundant activities of Pax2 and Pax6. *Development*, 130(13):2903–2915.

Bellefroid, E. J., Leclère, L., Saulnier, A., Keruzore, M., Sirakov, M., Vervoort, M., and De Clercq, S. (2013). Expanding roles for the evolutionarily conserved dmrt sex transcriptional regulators during embryogenesis. *Cellular and Molecular Life Sciences*, 70(20):3829–3845.

Beverdam, A. and Meijlink, F. (2001). Expression patterns of group-I *aristaless*-related genes during craniofacial and limb development. *Mechanisms of Development*, 107(1):163–167.

Blyszczuk, P., Czyz, J., Kania, G., Wagner, M., Roll, U., St-Onge, L., and Wobus, A. M. (2003). Expression of Pax4 in embryonic stem cells promotes differentiation of nestin-positive progenitor and insulin-producing cells. *Proceedings of the National Academy of Sciences, USA*, 100(3):998–1003.

Bolouri, H. (2014). Modeling genomic regulatory networks with big data. *TRENDS in Genetics*, 30(5):182–191.

Bookout, A. L., Jeong, Y., Downes, M., Yu, R. T., Evans, R. M., and Mangelsdorf, D. J. (2006). Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. *Cell*, 126(4):789–799.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956.

Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L. W., Janette, J., Jiang, L., et al. (2014). Comparative analysis of regulatory information and circuits across distant species. *Nature*, 512(7515):453–456.

Boyle, A. P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., and Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464.

Brückler, F., Došlić, T., Graovac, A., and Gutman, I. (2011). On a class

of distance-based molecular structure descriptors. *Chemical Physics Letters*, 503(4):336–338.

Cai, J., Donaldson, A., Yang, M., German, M. S., Enikolopov, G., and Iacovitti, L. (2009). The role of Lmx1a in the differentiation of human embryonic stem cells into midbrain dopamine neurons in culture and after transplantation into a Parkinson's disease model. *Stem Cells*, 27(1):220–229.

Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., Anne, S. L., Doetsch, F., Colman, H., et al. (2009). The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., et al. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 36(suppl 1):D623–D631.

Chan, K. K.-K., Zhang, J., Chia, N.-Y., Chan, Y.-S., Sim, H. S., Tan, K. S., Oh, S. K.-W., Ng, H.-H., and Choo, A. B.-H. (2009). KLF4 and PBX1 directly regulate NANOG expression in human embryonic stem cells. *Stem Cells*, 27(9):2114–2125.

Chang, C.-W., Cheng, W.-C., Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Huang, C.-L., and Hsu, I. C. (2011). Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE*, 6(7):e22859.

Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L., et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117.

Chen, Y.-C., Rajagopala, S. V., Stellberger, T., and Uetz, P. (2010). Exhaustive benchmarking of the yeast two-hybrid system. *Nature Methods*, 7(9):667–668.

Corominas-Murtra, B., Goñi, J., Solé, R. V., and Rodríguez-Caso, C. (2013). On the origins of hierarchy in complex networks. *Proceedings of the National Academy of Sciences, USA*, 110(33):13316–13321.

Croft, D., OKelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(S1):D691–D697.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

Csermely, P., Hódsági, J., Korcsmáros, T., Módos, D., Perez-Lopez, Á. R., Szalay, K., Veres, D. V., Lenti, K., Wu, L.-Y., and Zhang, X.-S. (2014). Cancer stem cells display extremely large evolvability: alternating plastic and rigid networks as a potential Mechanism: Network models, novel

therapeutic target strategies, and the contributions of hypoxia, inflammation and cellular senescence. In *Seminars in Cancer Biology*. Elsevier.

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & Therapeutics*, 138(3):333–408.

Csete, M. and Doyle, J. (2004). Bow ties, metabolism and disease. *TRENDS in Biotechnology*, 22(9):446–450.

Davidson, E. H. (2010). *The regulatory genome: gene regulatory networks in development and evolution*. Academic Press.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002). A genomic regulatory network for development. *Science*, 295(5560):1669–1678.

Dehmer, M. (2008). Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, 201(1):82–94.

Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Information Sciences*, 181(1):57–78.

Dehmer, M., Varmuza, K., Borgert, S., and Emmert-Streib, F. (2009). On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. *Journal of Chemical Information and Modeling*, 49(7):1655–1663.

Delprato, A. (2012). Topological and functional properties of the small GTPases protein interaction network. *PLoS ONE*, 7(9):e44882.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A. J. (2004). A gateway-compatible yeast one-hybrid system. *Genome Research*, 14(10b):2093–2101.

Dobrynin, A., Entringer, R., and Gutman, I. (2001). Wiener index of trees: Theory and applications. *Acta Applicandae Mathematicae*, 66(3):211–249.

Dynan, W. S. and Tjian, R. (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell*, 35(1):79–87.

Eisenberg, E. and Levanon, E. Y. (2003). Human housekeeping genes are compact. *TRENDS in Genetics*, 19(7):362–365.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *TRENDS in Genetics*, 29(10):569–574.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

Fischermann, M., Hoffmann, A., Rautenbach, D., Székely, L., and Volkmann, L. (2002). Wiener index versus maximum degree in trees. *Discrete Applied Mathematics*, 122(1):127–137.

Fronczak, A., Fronczak, P., and Hołyst, J. A. (2004). Average path length in random networks. *Physical Review E*, 70(5):056110.

Galan-Caridad, J. M., Harel, S., Arenzana, T. L., Hou, Z. E., Doetsch, F. K., Mirny, L. A., and Reizis, B. (2007). Zfx controls the self-renewal of embryonic and hematopoietic stem cells. *Cell*, 129(2):345–357.

Galas, D. J. and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5(9):3157–3170.

Georgi, B., Voight, B. F., and Bućan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genetics*, 9(5):e1003484.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330(6012):1775–1787.

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. *Nature Reviews Molecular Cell Biology*, 8(8):645–654.

Goecks, J., Nekrutenko, A., Taylor, J., et al. (2010). Galaxy: a comprehen-

sive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86.

Goulburn, A. L., Alden, D., Davis, R. P., Micallef, S. J., Ng, E. S., Yu, Q. C., Lim, S. M., Soh, C.-L., Elliott, D. A., Hatzistavrou, T., et al. (2011). A targeted NKX2.1 human embryonic stem cell reporter line enables identification of human basal forebrain derivatives. *Stem Cells*, 29(3):462–473.

Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151.

Gutman, I. (1997). A property of the Wiener number and its modifications. *Indian Journal of Chemistry. Sect. A: Inorganic, Physical, Theoretical & Analytical*, 36(2):128–132.

Gutman, I., Linert, W., Lukovits, I., and Dobrynin, A. (1997). Trees with extremal hyper-Wiener index: Mathematical basis and chemical applications. *Journal of Chemical Information and Computer Sciences*, 37(2):349–354.

Gutman, I., Popović, L., et al. (1998). Graph representation of organic molecules Cayley's plerograms vs. his kenograms. *Journal of the Chemical Society, Faraday Transactions*, 94(7):857–860.

Hanna, L. A., Foreman, R. K., Tarasenko, I. A., Kessler, D. S., and Labosky, P. A. (2002). Requirement for Foxd3 in maintaining pluripotent cells of the early mouse embryo. *Genes & Development*, 16(20):2650–2661.

Harriger, L., Van Den Heuvel, M. P., and Sporns, O. (2012). Rich club organization of macaque cerebral cortex and its role in network communication. *PLoS ONE*, 7(9):e46497.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947.

Horvath, C. M. (2000). STAT proteins and transcriptional responses to extracellular signals. *TRENDS in Biochemical Sciences*, 25(10):496–502.

Hosoya, H. (1988). On some counting polynomials in chemistry. *Discrete Applied Mathematics*, 19(1):239–257.

Hu, L., Huang, T., Shi, X., Lu, W.-C., Cai, Y.-D., and Chou, K.-C. (2011). Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. *PLoS ONE*, 6(1):e14556.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8:565.

James, D., Levine, A. J., Besser, D., and Hemmati-Brivanlou, A. (2005). TGFβ/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development*, 132(6):1273–1282.

Jean, D., Bernier, G., and Gruss, P. (1999). *Six6*(*Optx2*) is a novel murine *Six3*-related homeobox gene that demarcates the presumptive pituitary/hypothalamic axis and the ventral optic stalk. *Mechanisms of Development*, 84(1):31–40.

Jensen, J. (2004). Gene regulatory factors in pancreatic development. *Developmental Dynamics*, 229(1):176–200.

Jiang, J., Chan, Y.-S., Loh, Y.-H., Cai, J., Tong, G.-Q., Lim, C.-A., Robson, P., Zhong, S., and Ng, H.-H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nature Cell Biology*, 10(3):353–360.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human microrna targets. *PLoS Biology*, 2(11):e363.

Jothi, R., Balaji, S., Wuster, A., Grochow, J. A., Gsponer, J., Przytycka, T. M., Aravind, L., and Babu, M. M. (2009). Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular Systems Biology*, 5:294.

Junker, B. H. and Schreiber, F. (2008). Analysis of biological networks. volume 2, pages 31–59. Wiley-Interscience.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061.

Kim, J.-H., Auerbach, J. M., Rodríguez-Gómez, J. A., Velasco, I., Gavin, D., Lumelsky, N., Lee, S.-H., Nguyen, J., Sánchez-Pernaute, R., Bankiewicz, K., et al. (2002). Dopamine neurons derived from embryonic

stem cells function in an animal model of Parkinson's disease. *Nature*, 418(6893):50–56.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11):826–837.

Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. In *Data Mining in Proteomics*, pages 291–303. Springer.

Koschützki, D. and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*, 2:193–201.

Koschützki, D., Schwöbbermeyer, H., and Schreiber, F. (2007). Ranking of network elements based on functional substructures. *Journal of Theoretical Biology*, 248(3):471–479.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804.

Levy, D. E. and Darnell, J. (2002). STATs: transcriptional control and biological impact. *Nature Reviews Molecular Cell Biology*, 3(9):651–662.

Li, J., Hua, X., Haubrock, M., Wang, J., and Wingender, E. (2012). The architecture of the gene regulatory networks of different tissues. *Bioinformatics*, 28(18):i509–i514.

Li, Y. and Behringer, R. R. (1998). Esx1 is an X-chromosome-imprinted regulator of placental development and fetal growth. *Nature Genetics*, 20(3):309–311.

Lim, L. S., Loh, Y.-H., Zhang, W., Li, Y., Chen, X., Wang, Y., Bakre, M., Ng, H.-H., and Stanton, L. W. (2007). Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Molecular Biology of the Cell*, 18(4):1348–1358.

Liu, W., Lan, Y., Pauws, E., Meester-Smoor, M. A., Stanier, P., Zwarthoff, E. C., and Jiang, R. (2008). The Mn1 transcription factor acts upstream of Tbx22 and preferentially regulates posterior palate growth in mice. *Development*, 135(23):3959–3968.

Liu, Y., Jiang, B., and Zhang, X. (2009). Gene-set analysis identifies master transcription factors in developmental courses. *Genomics*, 94(1):1–10.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, 38(4):431–440.

Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334.

Ma, H.-W., Buer, J., and Zeng, A.-P. (2004). Hierarchical structure and modules in the escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, 5:199.

Ma'ayan, A. (2011). Introduction to network analysis in systems biology. *Science Signaling*, 4(190):tr5.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences, USA*, 100(21):11980–11985.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., et al. (2003). Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378.

Milenković, T., Memišević, V., Bonato, A., and Pržulj, N. (2011). Dominating biological networks. *PLoS ONE*, 6(8):e23016.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

Mones, E., Vicsek, L., and Vicsek, T. (2012). Hierarchy measure for complex networks. *PLoS ONE*, 7(3):e33799.

Mueller, L., Kugler, K., Dander, A., Graber, A., and Dehmer, M. (2011a). QuACN: an R package for analyzing complex biological networks quantitatively. *Bioinformatics*, 27(1):140–141.

Mueller, L. A., Kugler, K. G., Netzer, M., Graber, A., and Dehmer, M. (2011b). A network-based approach to classify the three domains of life. *Biology Direct*, 6:53.

Mullen, A. C., Orlando, D. A., Newman, J. J., Lovén, J., Kumar, R. M., Bilodeau, S., Reddy, J., Guenther, M. G., DeKoter, R. P., and Young, R. A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell*, 147(3):565–576.

Nelson, M. D., Zhou, E., Kiontke, K., Fradin, H., Maldonado, G., Martin, D., Shah, K., and Fitch, D. H. (2011). A bow-tie genetic architecture for morphogenesis suggested by a genome-wide RNAi screen in *Caenorhabditis elegans*. *PLoS Genetics*, 7(3):e1002010.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012a). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. (2012b). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90.

Newman, M. E. (2002). The structure and function of networks. *Computer Physics Communications*, 147(1):40–45.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., et al. (2005). The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834.

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006). AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiology*, 140(3):818–829.

Pennisi, E. (2012). Encode project writes eulogy for junk DNA. *Science*, 337(7):1159–1161.

Plavšić, D., Nikolić, S., Trinajstić, N., and Mihalić, Z. (1993). On the Harary index for the characterization of chemical graphs. *Journal of Mathematical Chemistry*, 12(1):235–250.

Pržulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515.

Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C., and Melton, D. A. (2002). "stemness": transcriptional profiling of embryonic and adult stem cells. *Science*, 298(5593):597–600.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Randić, M. (1993). Novel molecular descriptor for structureproperty studies. *Chemical Physics Letters*, 211(4):478–483.

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752.

Raymond, C. S., Murphy, M. W., O'Sullivan, M. G., Bardwell, V. J., and Zarkower, D. (2000). Dmrt1, a gene related to worm and fly sexual regulators, is required for mammalian testis differentiation. *Genes & Development*, 14(20):2587–2595.

Ren, G. and Liu, Z. (2013). NetCAD: a network analysis tool for coronary artery disease-associated PPI network. *Bioinformatics*, 29(2):279–280.

Resendis-Antonio, O., Hernández, M., Mora, Y., and Encarnación, S. (1931). Gráfok és mátrixok. *Matematikai és Fizikai Lapok*, 38(10):116–119.

Resendis-Antonio, O., Hernández, M., Mora, Y., and Encarnación, S. (2012). Functional modules, structural topology, and optimal activity in metabolic networks. *PLoS Computational Biology*, 8(10):e1002720.

Rodriguez-Caso, C., Medina, M. A., and Sole, R. V. (2005). Topology, tinkering and evolution of the human transcription factor network. *FEBS Journal*, 272(24):6423–6434.

Rogers, M., Hosler, B., and Gudas, L. (1991). Specific expression of a retinoic acid-regulated, zinc-finger gene, Rex-1, in preimplantation embryos, trophoblast and spermatocytes. *Development*, 113(3):815–824.

Ryu, S., Mahler, J., Acampora, D., Holzschuh, J., Erhardt, S., Omodei, D., Simeone, A., and Driever, W. (2007). Orthopedia homeodomain protein is essential for diencephalic dopaminergic neuron development. *Current Biology*, 17(10):873–880.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Stein, T. I., Shmoish, M.,

Nativ, N., Bahir, I., Doniger, T., Krug, H., et al. (2010). GeneCards Version 3: the human gene integrator. *Database*, 2010:baq020.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., et al. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(D1):D203–D213.

Schmuck, N. S., Wagner, S. G., and Wang, H. (2012). Greedy trees, caterpillars, and Wiener-type graph invariants. *MATCH Communications in Mathematical and in Computer Chemistry*, 68(1):273–292.

Seale, P., Sabourin, L. A., Girgis-Gabardo, A., Mansouri, A., Gruss, P., and Rudnicki, M. A. (2000). Pax7 is required for the specification of myogenic satellite cells. *Cell*, 102(6):777–786.

Semina, E., Mintz-Hittner, H., and Murray, J. (2000). Isolation and characterization of a novel human *paired-like* homeodomain-containing transcription factor gene, *VSX1*, expressed in ocular tissues. *Genomics*, 63(2):289–293.

She, X., Rohl, C. A., Castle, J. C., Kulkarni, A. V., Johnson, J. M., and Chen, R. (2009). Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics*, 10:269.

Shoval, O. and Alon, U. (2010). SnapShot: network motifs. *Cell*, 143(2):326–326.

Solomon, K. S., Kudoh, T., Dawid, I. B., and Fritz, A. (2003). Zebrafish

foxi1 mediates otic placode formation and jaw development. *Development*, 130(5):929–940.

Soltés, L. (1991). Transmission in graphs: a bound and vertex removing. *Mathematica Slovaca*, 41(1):11–16.

Sporns, O., Honey, C. J., and Kötter, R. (2007). Identification and classification of hubs in brain networks. *PLoS ONE*, 2(10):e1049.

Sporns, O. and Kötter, R. (2004). Motifs in brain networks. *PLoS Biology*, 2(11):e369.

Srivas, R., Costelloe, T., Carvunis, A.-R., Sarkar, S., Malta, E., Sun, S. M., Pool, M., Licon, K., van Welsem, T., van Leeuwen, F., et al. (2013). A UV-induced genetic network links the RSC complex to nucleotide excision repair and shows dose-dependent rewiring. *Cell Reports*, 5(6):1714–1724.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences, USA*, 101(16):6062–6067.

Todeschini, R. and Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. Wiley-VCH.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., et al. (2008). An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90.

Vidal, M., Cusick, M. E., and Barabasi, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–998.

Vinayagam, A., Hu, Y., Kulkarni, M., Roesel, C., Sopko, R., Mohr, S. E., and Perrimon, N. (2013). Protein complex-based analysis framework for high-throughput data sets. *Science Signaling*, 6(264):rs5.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403.

Wagner, S., Wangb, H., and Zhangc, X.-D. (2013). Distance-based graph invariants of trees and the Harary index. *Filomat*, 27(1):41–50.

Wang, P., Lü, J., and Yu, X. (2014). Identification of important nodes in directed biological networks: A network motif approach. *PLoS ONE*, 9(8):e106132.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master

transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319.

Wiener, H. (1947a). Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *Journal of the American Chemical Society*, 69(11):2636–2638.

Wiener, H. (1947b). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20.

Wuchty, S., Oltvai, Z. N., and Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305.

Xiang, M., Gan, L., Li, D., Chen, Z.-Y., Zhou, L., OMalley, B. W., Klein, W., and Nathans, J. (1997). Essential role of POU–domain factor Brn-3c in auditory and vestibular hair cell development. *Proceedings of the National Academy of Sciences, USA*, 94(17):9445–9450.

Xiao, L., Yuan, X., and Sharkis, S. J. (2006). Activin A maintains self-renewal and regulates fibroblast growth factor, Wnt, and bone morphogenic protein pathways in human embryonic stem cells. *Stem Cells*, 24(6):1476–1486.

Young, R. A. (2011). Control of the embryonic stem cell state. *Cell*, 144(6):940–954.

Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Yu, H. and Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences, USA*, 103(40):14724–14731.

Yu, X., Lin, J., Masuda, T., Esumi, N., Zack, D. J., and Qian, J. (2006). Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 34(3):917–927.

Zhang, H.-M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.-Y. (2012a). AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 40(D1):D144–D149.

Zhang, S., Tian, D., Tran, N. H., Choi, K. P., and Zhang, L. (2014). Profiling the transcription factor regulatory networks of human cell types. *Nucleic Acids Research*, 42(20):12380–12387.

Zhang, Y., Gutman, I., Liu, J., and Mu, Z. (2012b). q-analog of wiener index. *Match-Communications in Mathematical and Computer Chemistry*, 67(2):347.

Zhang, Y., Gutman, I., Liu, J., and Mu, Z. (2012c). q-analog of Wiener index. *MATCH Communications in Mathematical and in Computer Chemistry*, 67(2):347–356.

Appendix

Table A.1. 1509 ESC specific interactions which are found in hESC network, but not found in the other 40 TF regulatory networks.

N	Source	Target	N	Source	Target	N	Source	Target
1	AHR	BARHL2	2	AHR	EN2	3	AHR	IKZF1
4	AHR	SIX3	5	ALX1	HBP1	6	ALX1	NR1H4
7	ALX1	TFAP2C	8	ALX3	FOXJ1	9	ALX3	HBP1
10	ALX4	EBF2	11	ALX4	FOXD3	12	ALX4	NR1H4
13	ALX4	TFAP2C	14	AR	SIX6	15	ARID5B	DMRT1
16	ARID5B	HNF1B	17	ARID5B	NKX3-2	18	ARID5B	OVOL2
19	ARID5B	ZFP42	20	ARNT	EN2	21	ARNT	IKZF1
22	ARNT	SIX3	23	ARNT	SOX2	24	ARNT2	IKZF1
25	ARNT2	SIX3	26	ATF1	FOXC1	27	ATF2	FOXC1
28	ATF3	FOXC1	29	ATF4	DLX2	30	ATF4	FOXC1
31	ATF4	ZIC3	32	ATF5	ARX	33	ATF5	FOXC1
34	ATF5	NKX2-2	35	ATF5	PAX2	36	ATF5	SIX3
37	ATF6	FOXC1	38	ATF7	FOXC1	39	ATOH1	FGF9
40	ATOH1	HOXB13	41	ATOH1	POU3F1	42	ATOH1	REST
43	ATOH1	SIX4	44	ATOH1	STAT5A	45	ATOH1	TBX22
46	ATOH1	ZBTB7A	47	BACH1	XBP1	48	BACH2	GLI1
49	BCL6	MYF5	50	BCL6	SIX4	51	BDP1	DLX2
52	BDP1	HBP1	53	BDP1	HMGA1	54	BDP1	LHX2
55	BHLHE40	CRX	56	BHLHE40	DMRT1	57	BHLHE40	OVOL2
58	BHLHE40	ZFP42	59	BHLHE41	CRX	60	BHLHE41	DMRT1
61	BHLHE41	OVOL2	62	BHLHE41	ZFP42	63	BPTF	FOXO4
64	BPTF	WT1	65	BRF1	DLX2	66	BRF1	LHX2
67	CBFB	IRX2	68	CBFB	OTX2	69	CBFB	PARP1
70	CBFB	POU4F3	71	CDC5L	LEF1	72	CDX1	REST
73	CDX1	T	74	CDX1	VAX1	75	CDX2	GCM1
76	CDX2	ONECUT1	77	CDX2	REST	78	CDX2	VAX1
79	CEBPA	EOMES	80	CEBPA	POU5F1	81	CEBPA	SIX4
82	CEBPG	TBX22	83	CIZ1	ISL1	84	CNOT3	HOXB13
85	CNOT3	MSX2	86	CNOT3	PAX2	87	CNOT3	PAX4
88	CREB1	EN2	89	CREB1	FOXC1	90	CREM	FOXC1
91	CRX	AIRE	92	CRX	NR2E3	93	CTCF	CDX2
94	CTCF	DLX1	95	CTCF	ESX1	96	CTCF	OTP
97	CTCF	POU5F1	98	CTCF	SOX10	99	CTCF	TBX22

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
100	CTCF	VSX2	101	CTCF	ZIC3	102	CUX1	MYF5
103	DDIT3	POU5F1	104	DEAF1	ESR1	105	DEAF1	GFI1
106	DEAF1	HOXB13	107	DEAF1	POU4F3	108	DEAF1	RORB
109	DEAF1	T	110	DLX5	CRX	111	DMBX1	CRX
112	DMRT2	VAX1	113	DMRT3	HOXB13	114	E2F1	DMRT1
115	E2F1	FOXA2	116	E2F1	LMX1A	117	E2F4	LMX1A
118	E2F6	LMX1A	119	E2F7	LMX1A	120	EBF1	LHX4
121	EBF1	NF1	122	EBF1	PAX4	123	EBF1	POU5F1
124	EBF2	PAX4	125	EBF2	POU5F1	126	EGR1	ALX1
127	EGR1	CRX	128	EGR1	DMRT3	129	EGR1	OTP
130	EGR1	OTX2	131	EGR1	PAX4	132	EGR1	PAX7
133	EGR1	POU2F3	134	EGR2	ALX1	135	EGR2	CRX
136	EGR2	DMRT3	137	EGR2	OTP	138	EGR2	OTX2
139	EGR2	PAX4	140	EGR2	PAX7	141	EGR2	POU2F3
142	EGR3	ALX1	143	EGR3	CRX	144	EGR3	DMRT3
145	EGR3	OTP	146	EGR3	OTX2	147	EGR3	PAX4
148	EGR3	PAX7	149	EGR3	POU2F3	150	EGR4	CRX
151	EGR4	DMRT3	152	EGR4	NKX6-1	153	EGR4	OTX2
154	EGR4	PAX7	155	EGR4	POU2F3	156	EGR4	SIX6
157	ELF1	SMAD3	158	ELF1	ZIC1	159	ELF1	ZIC2
160	ELF2	FOXD3	161	ELF2	HMX3	162	ELF2	NKX2-2
163	ELF2	POU4F3	164	ELF2	REST	165	ELF2	SIX4
166	ELF2	ZIC1	167	ELF2	ZIC2	168	ELF3	ETV7
169	ELF3	ISL1	170	ELF3	OTX2	171	ELF3	TP63
172	ELK1	DMRT3	173	ELK1	HBP1	174	ELK1	TCF7
175	ELK1	ZIC1	176	ELK1	ZIC2	177	ELK4	TCF7
178	ELK4	ZIC1	179	ELK4	ZIC2	180	EMX2	CEBPB
181	EN1	FOXD3	182	EN2	FOXD3	183	EP300	DMRT3
184	EP300	GFI1	185	EP300	NKX2-2	186	EP300	OTX2
187	EP300	PAX7	188	EP300	RORB	189	EP300	SIX3
190	EPAS1	SOX2	191	ERF	ZIC1	192	ERF	ZIC2
193	ERG	OTX2	194	ERG	TCF7	195	ERG	ZIC1
196	ERG	ZIC2	197	ESR1	PAX2	198	ESR1	POU5F1
199	ESR2	POU5F1	200	ESRRA	HOXB13	201	ESRRB	EBF2
202	ESRRB	HOXB13	203	ESRRB	POU4F3	204	ESX1	HBP1
205	ETS1	BARHL2	206	ETS1	FOXD3	207	ETS1	HBP1
208	ETS1	MSX2	209	ETS1	TCF7	210	ETS1	ZIC1
211	ETS1	ZIC2	212	ETS1	ZIC3	213	ETS2	FOXD3
214	ETS2	HBP1	215	ETS2	ZIC1	216	ETS2	ZIC2
217	ETV7	HBP1	218	ETV7	TCF7	219	ETV7	ZIC1
220	ETV7	ZIC2	221	EVX1	SOX2	222	FGF9	FOXD3
223	FGF9	GATA3	224	FLI1	MYCN	225	FLI1	NR5A2
226	FLI1	PAX1	227	FLI1	POU5F1	228	FLI1	TCF7
229	FLI1	ZIC1	230	FLI1	ZIC2	231	FOSL1	GLI1
232	FOSL1	LMX1A	233	FOSL1	PAX6	234	FOXA1	DMRT1
235	FOXA1	ETV7	236	FOXA1	FOXD3	237	FOXA1	HOXA1
238	FOXA1	HOXA11	239	FOXA2	ETV7	240	FOXA2	HOXA1
241	FOXA2	HOXA11	242	FOXA3	ETV7	243	FOXA3	HOXA1
244	FOXA3	HOXA11	245	FOXD1	FOXD3	246	FOXD1	IRX5
247	FOXD1	MEIS1	248	FOXD3	ETV7	249	FOXD3	FOXI1

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
250	FOXD3	HOXB3	251	FOXD3	OTX2	252	FOXD3	PAX5
253	FOXD3	PAX7	254	FOXF1	ETV7	255	FOXF1	FOXI1
256	FOXF1	OTX2	257	FOXF1	PAX5	258	FOXF1	PAX7
259	FOXF2	ETV7	260	FOXF2	FOXI1	261	FOXF2	OTX2
262	FOXF2	PAX5	263	FOXF2	PAX7	264	FOXG1	DMRT1
265	FOXG1	XBP1	266	FOXH1	ETV7	267	FOXH1	FOXI1
268	FOXH1	HOXB3	269	FOXH1	OTX2	270	FOXH1	PAX5
271	FOXH1	PAX7	272	FOXI1	ETV7	273	FOXI1	FOXI1
274	FOXI1	PAX5	275	FOXI1	PAX7	276	FOXJ1	ETV7
277	FOXJ1	FOXI1	278	FOXJ1	HOXB3	279	FOXJ1	OTX2
280	FOXJ1	PAX5	281	FOXJ1	PAX7	282	FOXJ2	ETV7
283	FOXJ2	FOXI1	284	FOXJ2	HOXB3	285	FOXJ2	PAX5
286	FOXJ2	PAX7	287	FOXL1	CDX2	288	FOXL1	PAX3
289	FOXM1	ETV7	290	FOXM1	HOXA1	291	FOXM1	HOXA11
292	FOXO1	DMRT1	293	FOXO1	XBP1	294	FOXO3	FOXI1
295	FOXO3	NANOG	296	FOXO4	MAFA	297	FOXP1	BARHL2
298	FOXP1	HOXB13	299	FOXP1	OTX2	300	FOXP1	PAX2
301	FOXP3	ONECUT1	302	GABPA	BARX1	303	GABPA	NR5A2
304	GABPA	OVOL2	305	GABPA	PAX6	306	GABPB1	BARX1
307	GABPB1	NR5A2	308	GABPB1	OVOL2	309	GABPB1	PAX6
310	GATA1	OTX2	311	GATA1	PAX2	312	GATA1	SIX3
313	GATA1	STAT3	314	GATA1	STAT4	315	GATA2	FOXC1
316	GATA2	FOXJ3	317	GATA2	MSX2	318	GATA2	NR5A2
319	GATA2	SIX3	320	GATA2	STAT3	321	GATA3	FOXA2
322	GATA3	FOXC1	323	GATA3	FOXJ3	324	GATA3	MSX2
325	GATA3	NR5A2	326	GATA3	PGR	327	GATA4	CRX
328	GATA4	TFAP2B	329	GBX2	SOX2	330	GCM1	CDX2
331	GFI1	MEIS1	332	GLI1	SMAD2	333	GLI2	RORB
334	GLI2	SMAD2	335	GLI3	HMX3	336	GLI3	HOXA1
337	GLI3	POU5F1	338	GLI3	RORB	339	GLI3	VSX2
340	GLIS1	SMAD2	341	GLIS3	PAX7	342	GLIS3	POU5F1
343	GSX2	SOX2	344	GTF2A1	LHX4	345	GTF2I	ALX3
346	GTF2I	ALX4	347	GTF2I	ARNTL2	348	GTF2I	CRX
349	GTF2I	DMRT3	350	GTF2I	E2F4	351	GTF2I	EGR4
352	GTF2I	ONECUT1	353	GTF2I	OTX1	354	GTF2I	TFAP2B
355	GTF2I	VAX2	356	GTF2IRD1	CDX2	357	GTF2IRD1	FOXD3
358	GTF2IRD1	POU5F1	359	HAND1	DLX1	360	HAND1	DMRT1
361	HAND1	OVOL2	362	HAND1	SIX2	363	HAND1	SIX3
364	HAND1	ZFP42	365	HAND2	DMRT1	366	HAND2	OVOL2
367	HAND2	ZFP42	368	HES1	NEUROD1	369	HES1	PAX6
370	HIC1	BARHL2	371	HIC1	CDX2	372	HIC1	CRX
373	HIC1	DMRT1	374	HIC1	DMRT3	375	HIC1	HNF1B
376	HIC1	LMX1A	377	HIC1	NKX2-2	378	HIC1	OTP
379	HIC1	ZIC3	380	HIF1A	IKZF1	381	HIF1A	IRF3
382	HIF1A	POU3F2	383	HIF1A	SIX3	384	HIF1A	SOX2
385	HIVEP2	CRX	386	HIVEP2	OTX2	387	HMGA1	LMX1A
388	HMGA1	PARP1	389	HMGA1	PAX6	390	HMGA1	SOX21
391	HMGA2	LMX1A	392	HMGA2	PARP1	393	HMGA2	PAX6
394	HMGA2	SOX21	395	HNF1A	MNX1	396	HNF1B	DMRT2
397	HNF1B	EGR4	398	HNF1B	IRX4	399	HNF1B	NF1

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
400	HNF4A	ALX4	401	HNF4A	DMRT2	402	HNF4A	DMRT3
403	HNF4A	FOXD3	404	HNF4A	FOXH1	405	HNF4A	LHX4
406	HNF4A	LMX1A	407	HNF4A	NR5A1	408	HNF4A	OTP
409	HNF4A	POU5F1	410	HNF4A	ZNF628	411	HNF4G	DMRT3
412	HNF4G	FOXD3	413	HNF4G	FOXH1	414	HNF4G	LHX4
415	HNF4G	POU5F1	416	HNF4G	ZNF628	417	HOXA1	SOX2
418	HOXA10	REST	419	HOXA10	VAX1	420	HOXA11	LHX4
421	HOXA13	FOXD3	422	HOXA13	RAX	423	HOXA3	SOX2
424	HOXA4	TFAP2C	425	HOXA5	MYCN	426	HOXA9	LMX1B
427	HOXB3	CEBPB	428	HOXB3	SOX2	429	HOXB4	TFAP2C
430	HOXB7	BARHL2	431	HOXB8	BARHL2	432	HOXB9	GLI1
433	HOXB9	VAX1	434	HOXC10	LHX4	435	HOXC11	LHX4
436	HOXC4	NR1H4	437	HOXC4	TFAP2C	438	HOXC5	TFAP2C
439	HOXC9	REST	440	HOXD13	SOX9	441	HOXD9	MYCN
442	HSF1	FOXD3	443	HSF1	FOXJ3	444	HSF1	LEF1
445	HSF1	MECP2	446	HSF1	NKX2-1	447	HSF1	NKX2-2
448	HSF2	FOXJ3	449	HSF2	GATA3	450	HSF2	LEF1
451	HSF2	MECP2	452	HSF2	NKX2-1	453	IKZF1	CRX
454	IKZF1	ESX1	455	IKZF1	HBP1	456	IKZF1	PAX6
457	IKZF1	RUNX3	458	IKZF1	SIX4	459	IKZF1	TBX15
460	IKZF2	FOXD3	461	IRF1	ALX1	462	IRF1	EOMES
463	IRF1	GFI1	464	IRF1	SOX17	465	IRF1	ZFP42
466	IRF2	ALX1	467	IRF2	EOMES	468	IRF2	FOXD3
469	IRF2	GFI1	470	IRF2	SOX17	471	IRF2	ZFP42
472	IRF3	ALX1	473	IRF3	EOMES	474	IRF3	FOXD3
475	IRF3	GFI1	476	IRF3	PAX7	477	IRF3	SOX17
478	IRF3	ZFP42	479	IRF4	ALX1	480	IRF4	EOMES
481	IRF4	GFI1	482	IRF4	SOX17	483	IRF4	ZFP42
484	IRF6	ALX1	485	IRF6	EOMES	486	IRF6	SOX17
487	IRF6	ZFP42	488	IRF7	ALX1	489	IRF7	GFI1
490	IRF7	SOX17	491	IRF7	ZFP42	492	IRF8	ALX1
493	IRF8	EOMES	494	IRF8	GFI1	495	IRF8	SOX17
496	IRF8	ZFP42	497	IRF9	GFI1	498	IRF9	PAX2
499	IRX2	PAX3	500	IRX3	PAX3	501	IRX4	PAX3
502	IRX5	PAX3	503	ISX	FOXJ1	504	JUN	GLI1
505	JUN	LMX1A	506	JUN	PAX6	507	JUNB	GLI1
508	JUND	GLI1	509	JUND	LMX1A	510	JUND	PAX6
511	KLF11	ATF2	512	KLF11	DLX1	513	KLF11	EN2
514	KLF11	FOXH1	515	KLF11	NKX2-2	516	KLF11	POU5F1
517	KLF11	SREBF2	518	KLF15	CRX	519	KLF15	DMRT3
520	KLF15	MAFA	521	KLF15	MSX2	522	KLF15	NKX2-2
523	KLF15	OTX2	524	KLF15	PAX2	525	KLF15	PAX7
526	KLF15	POU5F1	527	KLF15	ZIC3	528	KLF4	DMRT3
529	KLF4	FOXD3	530	KLF4	MSX2	531	KLF4	NANOG
532	KLF4	PAX2	533	KLF4	POU5F1	534	KLF4	TBX15
535	KLF4	TFAP2B	536	KLF4	TP63	537	KLF4	ZFP42
538	LEF1	FOXD1	539	LEF1	HMGA1	540	LEF1	MNX1
541	LEF1	MYOG	542	LHX2	HBP1	543	LHX3	TFAP2C
544	LHX4	FOXJ1	545	LHX4	SOX2	546	LHX5	FOXI1
547	LHX5	NR1H4	548	LHX5	TFAP2C	549	LHX6	FOXJ1

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
550	LHX8	FOXJ1	551	LHX8	SOX2	552	LMO2	ATOH1
553	LMO2	FOXJ3	554	LMO2	OTX2	555	LMO2	POU3F1
556	LMO2	SOX2	557	LMX1A	FOXI1	558	LMX1A	NR1H4
559	LMX1A	TFAP2C	560	LMX1B	FOXI1	561	LMX1B	NR1H4
562	MAF	CRX	563	MAF	DLX4	564	MAF	PURA
565	MAF	TBX15	566	MAFA	DLX3	567	MAFA	PARP1
568	MAFA	POU5F1	569	MAFG	REST	570	MAX	DMRT1
571	MAX	NKX2-5	572	MAX	OVOL2	573	MAX	SATB1
574	MAX	SMAD4	575	MAX	ZFP42	576	MAX	ZNF219
577	MAZ	ALX3	578	MAZ	ALX4	579	MAZ	CDX2
580	MAZ	CRX	581	MAZ	FOXD3	582	MAZ	LMX1B
583	MAZ	MSX1	584	MAZ	MYF5	585	MAZ	NKX6-1
586	MAZ	OTX2	587	MAZ	PAX7	588	MAZ	SIX6
589	MAZ	TBX22	590	MECOM	FOXJ3	591	MECOM	MSX2
592	MECOM	OTX2	593	MECOM	PAX2	594	MECP2	PAX7
595	MEF2A	IRX2	596	MEF2A	TBX15	597	MEIS1	LMX1B
598	MEIS3	ARX	599	MEOX1	SOX2	600	MITF	DMRT1
601	MITF	OVOL2	602	MITF	ZFP42	603	MNX1	FOXD3
604	MTF1	FOXD3	605	MYB	LMX1A	606	MYB	NKX2-2
607	MYB	PARP1	608	MYB	PAX7	609	MYB	RAX
610	MYB	RUNX3	611	MYB	STAT3	612	MYB	ZIC3
613	MYC	DMRT1	614	MYC	NKX2-5	615	MYC	OVOL2
616	MYC	SATB1	617	MYC	SMAD4	618	MYC	ZFP42
619	MYC	ZNF219	620	MYCN	ATOH1	621	MYCN	DMRT1
622	MYCN	OVOL2	623	MYCN	ZFP42	624	MYF5	DMRT1
625	MYF5	HNF1B	626	MYF5	NKX3-2	627	MYF5	OVOL2
628	MYF5	ZFP42	629	MYF6	ATOH1	630	MYF6	DMRT1
631	MYF6	OTX2	632	MYF6	OVOL2	633	MYF6	T
634	MYF6	ZFP42	635	MYOD1	ATOH1	636	MYOD1	DLX1
637	MYOD1	DMRT1	638	MYOD1	HOXB13	639	MYOD1	OTX2
640	MYOD1	OVOL2	641	MYOD1	T	642	MYOD1	ZFP42
643	MYOG	ATOH1	644	MYOG	DMRT1	645	MYOG	OTX2
646	MYOG	OVOL2	647	MYOG	T	648	MYOG	ZFP42
649	MZF1	CRX	650	MZF1	FOXC1	651	MZF1	FOXD3
652	MZF1	PAX2	653	MZF1	PAX5	654	MZF1	PURA
655	MZF1	TBX22	656	MZF1	TFAP2C	657	MZF1	ZFP42
658	NANOG	ETS1	659	NANOG	LMX1A	660	NANOG	MECP2
661	NANOG	OTX2	662	NANOG	POU3F2	663	NANOG	POU5F1
664	NANOG	VAX1	665	NANOG	XBP1	666	NEUROD1	CDX2
667	NEUROD1	FOXD3	668	NEUROD1	FOXO4	669	NEUROD1	HOXB13
670	NEUROD1	MNX1	671	NEUROD1	REST	672	NF1	ALX3
673	NF1	CDX2	674	NF1	DMRT3	675	NF1	HMX3
676	NF1	MSX2	677	NF1	NANOG	678	NFATC1	FOXA1
679	NFATC1	HOXD12	680	NFATC1	ZNF219	681	NFATC2	FOXA1
682	NFATC2	ZNF219	683	NFATC3	FOXA1	684	NFATC3	ZNF219
685	NFATC4	FOXA1	686	NFATC4	POU2F3	687	NFATC4	ZNF219
688	NFE2	HMX3	689	NFE2	LMX1A	690	NFE2	PAX7
691	NFE2L1	REST	692	NFE2L2	BARX1	693	NFE2L2	LMX1A
694	NFE2L2	NR5A2	695	NFE2L2	OVOL2	696	NFE2L2	PAX6
697	NFE2L2	REST	698	NFIB	ALX3	699	NFIB	CDX2

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
700	NFIB	DMRT3	701	NFIB	MSX2	702	NFIB	NANOG
703	NFIX	ALX3	704	NFIX	CDX2	705	NFIX	DMRT3
706	NFIX	HMX3	707	NFIX	MSX2	708	NFIX	NANOG
709	NFKB1	ALX3	710	NFKB1	BARHL2	711	NFKB1	GATA3
712	NFKB1	ISL1	713	NFKB1	LMX1B	714	NFKB1	NKX2-1
715	NFKB1	OTP	716	NFKB1	PAX4	717	NFKB1	RUNX3
718	NFKB1	ZIC3	719	NFKB2	ALX3	720	NFKB2	ISL1
721	NFKB2	LMX1B	722	NFKB2	OTP	723	NFKB2	PAX4
724	NFKB2	RUNX3	725	NFKB2	ZIC3	726	NFYA	LHX4
727	NFYA	OTX2	728	NFYA	SIRT6	729	NHLH1	DMRT1
730	NHLH1	HOXB9	731	NHLH1	OVOL2	732	NHLH1	POU5F1
733	NHLH1	SMAD4	734	NHLH1	ZFP42	735	NHLH1	ZNF219
736	NKX2-1	FOXH1	737	NKX2-2	HBP1	738	NKX2-2	LMX1A
739	NKX3-1	LMX1A	740	NKX3-2	OTP	741	NKX6-2	FOXD3
742	NR0B1	BARHL1	743	NR0B1	EBF2	744	NR0B1	FOXO4
745	NR0B1	GFI1	746	NR0B1	HOXB13	747	NR0B1	PAX1
748	NR1H2	GLI1	749	NR1H2	HOXB13	750	NR1H2	IRX4
751	NR1H2	NKX2-2	752	NR1H2	POU5F1	753	NR1H4	GLI1
754	NR1H4	HOXB13	755	NR1H4	IRX4	756	NR1H4	PAX6
757	NR1I2	GFI1	758	NR1I2	HOXB13	759	NR1I3	GFI1
760	NR1I3	HOXB13	761	NR2C2	MYCN	762	NR2C2	POU5F1
763	NR2C2	ZIC1	764	NR2C2	ZIC2	765	NR2E3	MNX1
766	NR2F1	ALX1	767	NR2F1	ALX4	768	NR2F1	BARHL1
769	NR2F1	DLX1	770	NR2F1	EOMES	771	NR2F1	FOXD3
772	NR2F1	HOXB13	773	NR2F1	LHX4	774	NR2F1	LMX1A
775	NR2F1	MSX2	776	NR2F1	MYCN	777	NR2F1	NKX6-1
778	NR2F1	PARP1	779	NR2F1	PAX2	780	NR2F1	POU5F1
781	NR2F1	TBX22	782	NR2F1	ZNF628	783	NR2F2	ALX1
784	NR2F2	ALX4	785	NR2F2	BARHL1	786	NR2F2	DLX1
787	NR2F2	FOXD3	788	NR2F2	HOXB13	789	NR2F2	ISL1
790	NR2F2	LHX4	791	NR2F2	LMX1A	792	NR2F2	MSX2
793	NR2F2	MYCN	794	NR2F2	NKX6-1	795	NR2F2	ONECUT1
796	NR2F2	PARP1	797	NR2F2	PAX2	798	NR2F2	PAX4
799	NR2F2	POU5F1	800	NR2F2	TBX22	801	NR2F2	ZNF628
802	NR2F6	BARHL2	803	NR2F6	NR5A1	804	NR2F6	OTP
805	NR2F6	RAX	806	NR2F6	SIX3	807	NR2F6	SMAD4
808	NR3C1	CDX2	809	NR3C1	FOXO1	810	NR3C1	HOXB13
811	NR3C1	LEF1	812	NR3C1	MYCN	813	NR3C1	PBX1
814	NR4A1	BARHL1	815	NR4A1	EBF2	816	NR4A1	GFI1
817	NR5A1	POU4F3	818	NR5A2	NKX2-2	819	NR5A2	POU4F3
820	NR5A2	PRRX1	821	NR5A2	ZIC1	822	NR5A2	ZIC2
823	NR6A1	NKX2-2	824	NR6A1	SIX6	825	NRF1	POU4F3
826	OAZ1	FOXO4	827	OAZ1	NKX2-2	828	ONECUT1	HOXD12
829	ONECUT1	HOXD13	830	OTP	TFAP2C	831	OTX1	SIX4
832	OTX2	CRX	833	OTX2	NR2E3	834	PARP1	HOXB3
835	PARP1	ZFP42	836	PATZ1	CRX	837	PATZ1	FOXD3
838	PATZ1	NKX2-2	839	PATZ1	OTX2	840	PATZ1	PAX7
841	PATZ1	TFAP2C	842	PAX2	ALX3	843	PAX4	DMRT3
844	PAX4	NF1	845	PAX4	OTP	846	PAX4	OTX2
847	PAX4	PAX7	848	PAX4	POU5F1	849	PAX4	RORB

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
850	PAX4	SIX6	851	PAX4	TBX15	852	PAX4	ZFP42
853	PAX5	FLI1	854	PAX5	FOXH1	855	PAX5	HMX3
856	PAX5	PAX2	857	PAX5	POU2F3	858	PAX6	FOXD3
859	PAX6	SMAD3	860	PAX8	LHX5	861	PBX1	BARHL2
862	PBX1	MNX1	863	PBX1	OTX2	864	PBX1	ZIC3
865	PDX1	BARHL2	866	PDX1	FOXJ3	867	PDX1	MNX1
868	PGR	CDX2	869	PGR	MYCN	870	PGR	PBX1
871	PGR	ZNF628	872	PITX1	AIRE	873	PITX1	MEOX1
874	PITX1	NR2F6	875	PITX1	STAT4	876	PITX1	WT1
877	PITX2	ATOH1	878	PITX2	BRF1	879	PITX2	CRX
880	PITX2	NR2E3	881	PITX2	ZFP42	882	PITX3	CRX
883	PITX3	NANOG	884	PKNOX1	ARX	885	PKNOX1	BARHL2
886	PKNOX1	TP63	887	PKNOX2	ARX	888	PKNOX2	ETS1
889	POU2AF1	CDX2	890	POU2AF1	NR2F6	891	POU2AF1	OTX2
892	POU2AF1	PAX6	893	POU2AF1	SMAD4	894	POU2AF1	ZIC3
895	POU2F1	CDX2	896	POU2F1	FOXD1	897	POU2F1	FOXJ1
898	POU2F1	HBP1	899	POU2F1	IRX2	900	POU2F1	MECP2
901	POU2F1	OTX2	902	POU2F1	PAX2	903	POU2F1	PAX6
904	POU2F1	SMAD4	905	POU2F1	ZIC3	906	POU2F2	CDX2
907	POU2F2	MECP2	908	POU2F2	NR2F6	909	POU2F2	OTX2
910	POU2F2	PAX2	911	POU2F2	PAX6	912	POU2F2	SMAD4
913	POU2F2	ZIC3	914	POU2F3	CDX2	915	POU2F3	FOXD1
916	POU2F3	NR2F6	917	POU2F3	OTX2	918	POU2F3	PAX6
919	POU2F3	SMAD4	920	POU2F3	ZIC3	921	POU3F1	CDX2
922	POU3F1	IRX2	923	POU3F1	NR2F6	924	POU3F1	OTX2
925	POU3F1	PAX6	926	POU3F1	SMAD4	927	POU3F1	ZIC3
928	POU3F2	CDX2	929	POU3F2	NR2F6	930	POU3F2	OTX2
931	POU3F2	PAX6	932	POU3F2	SMAD4	933	POU3F2	ZIC3
934	POU3F3	CDX2	935	POU3F3	NR2F6	936	POU3F3	OTX2
937	POU3F3	PAX6	938	POU3F3	SMAD4	939	POU3F3	ZIC3
940	POU5F1	CDX2	941	POU5F1	ESX1	942	POU5F1	ETS1
943	POU5F1	FOXD1	944	POU5F1	HOXD12	945	POU5F1	HOXD13
946	POU5F1	ISL1	947	POU5F1	LHX5	948	POU5F1	NANOG
949	POU5F1	NR2F6	950	POU5F1	OTX2	951	POU5F1	PAX6
952	POU5F1	POU5F1	953	POU5F1	SIX3	954	POU5F1	SMAD4
955	POU5F1	TFAP2C	956	POU5F1	THRB	957	POU5F1	VSX1
958	POU5F1	ZIC3	959	POU6F1	FOXJ3	960	POU6F1	OTX2
961	PPARA	CDX2	962	PPARA	LHX4	963	PPARA	LHX8
964	PPARA	MNX1	965	PPARA	MYCN	966	PPARA	OTX2
967	PPARA	PAX6	968	PPARA	POU5F1	969	PPARA	RORB
970	PPARA	SIX3	971	PPARA	SMAD3	972	PPARD	LHX4
973	PPARD	MNX1	974	PPARD	MYCN	975	PPARD	POU5F1
976	PPARD	SIX3	977	PPARD	ZNF628	978	PPARG	LHX4
979	PPARG	MNX1	980	PPARG	MYCN	981	PPARG	POU5F1
982	PPARG	SIX3	983	PPARG	SMAD4	984	PPARG	ZNF628
985	PRDM1	ALX1	986	PRDM1	BARHL2	987	PRDM1	FOXD3
988	PRDM1	HOXB13	989	PRDM1	MYF6	990	PRDM1	NKX2-1
991	PRDM1	POU4F3	992	PRDM1	VAX1	993	PRDM1	ZIC3
994	PURA	ALX1	995	PURA	ALX4	996	PURA	FOXA2
997	PURA	IRX4	998	PURA	MSX2	999	PURA	NKX2-2

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1000	PURA	NKX3-2	1001	PURA	PITX1	1002	PURA	PITX2
1003	PURA	POU5F1	1004	PURA	RORB	1005	RARA	CRX
1006	RARA	FOXA2	1007	RARA	MYCN	1008	RARA	POU5F1
1009	RARA	REST	1010	RARA	VAX1	1011	RAX	AIRE
1012	RAX	NR2E3	1013	RBPJ	NF1	1014	REL	ISL1
1015	REL	NR5A1	1016	REL	PAX4	1017	REL	ZIC3
1018	RELA	ISL1	1019	RELA	LMX1B	1020	RELA	NR5A1
1021	RELA	PAX4	1022	RELA	RUNX3	1023	RELA	ZIC3
1024	RELB	ISL1	1025	RELB	PAX4	1026	RELB	ZIC3
1027	REST	DMRT3	1028	REST	LMX1A	1029	REST	PAX6
1030	RFX1	CRX	1031	RFX1	FOXD3	1032	RFX1	OTX2
1033	RORA	PAX6	1034	RORA	REST	1035	RREB1	ARX
1036	RREB1	EN2	1037	RREB1	FOXD3	1038	RREB1	HAND1
1039	RREB1	HMGA1	1040	RREB1	OTX2	1041	RREB1	PAX7
1042	RREB1	RAX	1043	RREB1	ZIC3	1044	RUNX1	IRX2
1045	RUNX1	OTX2	1046	RUNX1	PARP1	1047	RUNX1	POU4F3
1048	RUNX2	IRX2	1049	RUNX2	OTX2	1050	RUNX2	PARP1
1051	RUNX2	POU4F3	1052	RUNX3	IRX2	1053	RUNX3	OTX2
1054	RUNX3	PARP1	1055	RUNX3	POU4F3	1056	RXRA	CDX2
1057	RXRA	CRX	1058	RXRA	HOXB13	1059	RXRA	LHX4
1060	RXRA	MNX1	1061	RXRA	NKX2-2	1062	RXRA	OTX2
1063	RXRA	PAX6	1064	RXRA	POU5F1	1065	RXRA	REST
1066	RXRA	SIX3	1067	RXRA	VAX1	1068	RXRB	CRX
1069	RXRB	FOXA2	1070	RXRB	GFI1	1071	RXRB	POU5F1
1072	RXRB	REST	1073	RXRB	VAX1	1074	SIX4	FOXA1
1075	SMAD1	CRX	1076	SMAD1	LMX1A	1077	SMAD1	POU5F1
1078	SMAD1	XBP1	1079	SMAD2	ATF4	1080	SMAD2	CRX
1081	SMAD2	LMX1A	1082	SMAD2	POU5F1	1083	SMAD3	ATF4
1084	SMAD3	CRX	1085	SMAD3	HOXD13	1086	SMAD3	LMX1A
1087	SMAD3	POU5F1	1088	SMAD3	TBX15	1089	SMAD4	ATF4
1090	SMAD4	CRX	1091	SMAD4	FOXA2	1092	SMAD4	FOXD3
1093	SMAD4	GATA2	1094	SMAD4	LMX1A	1095	SMAD4	POU5F1
1096	SMAD4	RAX	1097	SMAD4	SIX2	1098	SMAD4	TBX15
1099	SMAD7	ATF4	1100	SMAD7	CRX	1101	SMAD7	LMX1A
1102	SMAD7	POU5F1	1103	SOX10	CDX2	1104	SOX10	NFKB2
1105	SOX17	SOX2	1106	SOX2	CDX2	1107	SOX2	ETS1
1108	SOX2	HOXD12	1109	SOX2	HOXD13	1110	SOX2	NANOG
1111	SOX2	NFKB2	1112	SOX2	POU5F1	1113	SOX2	SIX3
1114	SOX2	TGIF2	1115	SOX2	THRB	1116	SOX21	CDX2
1117	SOX21	NFKB2	1118	SOX4	CDX2	1119	SOX4	NFKB2
1120	SOX5	CDX2	1121	SOX5	NFKB2	1122	SOX9	CDX2
1123	SOX9	FOXI1	1124	SOX9	NFIB	1125	SOX9	NFKB2
1126	SOX9	NKX2-2	1127	SP1	ALX3	1128	SP1	CDX2
1129	SP1	CRX	1130	SP1	DMRT3	1131	SP1	EVX1
1132	SP1	FOXI1	1133	SP1	LMX1A	1134	SP1	NR5A2
1135	SP1	OTP	1136	SP1	OTX2	1137	SP1	PAX7
1138	SP2	ALX3	1139	SP2	CDX2	1140	SP2	CRX
1141	SP2	DMRT3	1142	SP2	EVX1	1143	SP2	FOXI1
1144	SP2	LMX1A	1145	SP2	MSX2	1146	SP2	NKX2-2
1147	SP2	OTX2	1148	SP2	PAX7	1149	SP3	ALX3

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1150	SP3	CDX2	1151	SP3	CRX	1152	SP3	DMRT3
1153	SP3	EVX1	1154	SP3	FOXI1	1155	SP3	LMX1A
1156	SP3	NKX2-2	1157	SP3	OTP	1158	SP3	OTX2
1159	SP3	PAX7	1160	SP4	ALX3	1161	SP4	CDX2
1162	SP4	CRX	1163	SP4	DMRT3	1164	SP4	EVX1
1165	SP4	FOXI1	1166	SP4	LMX1A	1167	SP4	MSX2
1168	SP4	NKX2-2	1169	SP4	OTX2	1170	SP4	PAX7
1171	SPI1	ALX1	1172	SPI1	EOMES	1173	SPI1	EVX1
1174	SPI1	FOXH1	1175	SPI1	HMX3	1176	SPI1	IRX2
1177	SPI1	LHX4	1178	SPI1	MYCN	1179	SPI1	MYOG
1180	SPI1	POU5F1	1181	SPI1	T	1182	SPI1	TFAP2B
1183	SPI1	TP63	1184	SPIB	LMX1A	1185	SPZ1	ALX1
1186	SPZ1	CDX2	1187	SPZ1	CRX	1188	SPZ1	FOXA1
1189	SPZ1	FOXA2	1190	SPZ1	HOXB3	1191	SPZ1	NKX2-2
1192	SPZ1	PAX7	1193	SPZ1	POU4F3	1194	SPZ1	POU5F1
1195	SPZ1	PURA	1196	SPZ1	RAX	1197	SPZ1	ZIC3
1198	SREBF1	ARX	1199	SREBF1	EN2	1200	SREBF1	FOXD3
1201	SREBF1	IRX4	1202	SREBF1	NKX2-2	1203	SREBF1	PAX7
1204	SREBF1	POU5F1	1205	SREBF1	TBX22	1206	SREBF1	ZNF148
1207	SREBF2	ARX	1208	SREBF2	CDX2	1209	SREBF2	DMRT3
1210	SREBF2	EN2	1211	SREBF2	FOXD3	1212	SREBF2	IRX4
1213	SREBF2	NKX2-2	1214	SREBF2	NR5A2	1215	SREBF2	PAX6
1216	SREBF2	PAX7	1217	SREBF2	POU5F1	1218	SREBF2	TBX22
1219	SREBF2	ZNF148	1220	SRF	ESX1	1221	SRF	IRX2
1222	SRY	CDX2	1223	SRY	NFKB2	1224	STAT1	FOXO4
1225	STAT1	LHX2	1226	STAT1	POU4F3	1227	STAT1	VAX1
1228	STAT3	ARX	1229	STAT3	DLX3	1230	STAT3	FOXD3
1231	STAT3	FOXO4	1232	STAT3	HMX3	1233	STAT3	LHX5
1234	STAT3	OTX2	1235	STAT3	PAX7	1236	STAT3	PKNOX2
1237	STAT3	POU4F3	1238	STAT3	SOX2	1239	STAT3	VAX1
1240	STAT4	POU2F3	1241	STAT5A	TBX22	1242	T	ATF2
1243	T	POU5F1	1244	TAL1	CDX2	1245	TAL1	DMRT1
1246	TAL1	EHF	1247	TAL1	OVOL2	1248	TAL1	TP63
1249	TAL1	ZFP42	1250	TAL1	ZNF219	1251	TBX15	FOXA1
1252	TBX15	HOXD13	1253	TBX15	TFAP2C	1254	TBX18	FOXA1
1255	TBX18	TFAP2C	1256	TBX22	TFAP2C	1257	TBX5	DMRT3
1258	TBX5	EGR2	1259	TBX5	ETV7	1260	TBX5	FOXA2
1261	TBX5	REST	1262	TBX5	SIX3	1263	TCF12	DMRT1
1264	TCF12	HNF1B	1265	TCF12	LMX1A	1266	TCF12	NKX3-2
1267	TCF12	NR2F6	1268	TCF12	OVOL2	1269	TCF12	POU5F1
1270	TCF12	ZFP42	1271	TCF3	ATOH1	1272	TCF3	DMRT1
1273	TCF3	HOXB13	1274	TCF3	HOXD12	1275	TCF3	LMX1A
1276	TCF3	OTX2	1277	TCF3	OVOL2	1278	TCF3	POU3F1
1279	TCF3	POU5F1	1280	TCF3	REST	1281	TCF3	SIX2
1282	TCF3	SIX3	1283	TCF3	T	1284	TCF3	TBX22
1285	TCF3	ZFP42	1286	TCF4	DMRT1	1287	TCF4	FOXD1
1288	TCF4	HNF1B	1289	TCF4	NKX3-2	1290	TCF4	OVOL2
1291	TCF4	ZFP42	1292	TCF7	MNX1	1293	TCF7	MYOG
1294	TEF	PAX7	1295	TEF	ZIC1	1296	TEF	ZIC2
1297	TERF1	OTX2	1298	TERF1	POU5F1	1299	TFAP2A	DMRT3

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1300	TFAP2A	HOXC12	1301	TFAP2A	LHX8	1302	TFAP2A	OVOL2
1303	TFAP2A	PAX2	1304	TFAP2A	PAX7	1305	TFAP2A	SIX6
1306	TFAP2A	TBX22	1307	TFAP2A	ZFP42	1308	TFAP2A	ZIC3
1309	TFAP2B	DMRT3	1310	TFAP2B	HOXC12	1311	TFAP2B	LHX8
1312	TFAP2B	PAX2	1313	TFAP2B	PAX7	1314	TFAP2B	RORB
1315	TFAP2B	SIX6	1316	TFAP2B	TBX22	1317	TFAP2B	ZFP42
1318	TFAP2B	ZIC3	1319	TFAP2C	DMRT3	1320	TFAP2C	HOXC12
1321	TFAP2C	LHX8	1322	TFAP2C	PAX2	1323	TFAP2C	PAX7
1324	TFAP2C	RORB	1325	TFAP2C	SIX6	1326	TFAP2C	TBX22
1327	TFAP2C	ZFP42	1328	TFAP2C	ZIC3	1329	TFAP4	ESX1
1330	TFAP4	GFI1	1331	TFAP4	OTX2	1332	TFAP4	TBX22
1333	TFAP4	ZNF628	1334	TFCP2	ARX	1335	TFCP2	EBF2
1336	TFCP2	EN2	1337	TFCP2	EOMES	1338	TFCP2	LHX5
1339	TFCP2	ZBTB7A	1340	TFCP2L1	HMGA1	1341	TFCP2L1	LHX4
1342	TFCP2L1	ZFP42	1343	TFDP1	LMX1A	1344	TFDP2	LMX1A
1345	TGIF1	ARX	1346	TGIF1	ETS1	1347	TGIF1	RXRB
1348	TGIF2	ARX	1349	THRA	CRX	1350	THRA	MYCN
1351	THRA	POU5F1	1352	THRA	REST	1353	THRA	VAX1
1354	THRB	CRX	1355	THRB	MYCN	1356	THRB	POU5F1
1357	THRB	REST	1358	THRB	VAX1	1359	TLX2	IRX4
1360	TLX2	TBX15	1361	TOPORS	FOXA2	1362	TOPORS	PARP1
1363	TOPORS	PRRX1	1364	TOPORS	TERF1	1365	TOPORS	ZFP42
1366	TP53	DMRT3	1367	TP53	FOXD1	1368	TP53	LMX1A
1369	TP53	MEOX1	1370	TP53	NKX3-2	1371	TP53	PAX2
1372	TP53	POU5F1	1373	TP63	DMRT3	1374	TP63	LMX1A
1375	TP63	NKX3-2	1376	TP63	POU5F1	1377	TP73	DMRT3
1378	TP73	LMX1A	1379	TP73	NKX3-2	1380	TP73	POU5F1
1381	TRIM28	ARX	1382	TRIM28	BARHL2	1383	TRIM28	FOXA2
1384	TRIM28	HOXA11	1385	TRIM28	LHX8	1386	TRIM28	SOX10
1387	USF1	DMRT1	1388	USF1	OVOL2	1389	USF1	ZFP42
1390	USF2	DMRT1	1391	USF2	OVOL2	1392	USF2	ZFP42
1393	VDR	ALX3	1394	VDR	ALX4	1395	VDR	CRX
1396	VDR	DMRT1	1397	VDR	EVX1	1398	VDR	LMX1B
1399	VDR	NKX2-1	1400	VDR	ONECUT1	1401	VDR	OTX2
1402	VDR	PAX2	1403	VDR	POU5F1	1404	VDR	SIX6
1405	VSX1	TFAP2C	1406	VSX2	SOX2	1407	WT1	ALX3
1408	WT1	CRX	1409	WT1	FOXD3	1410	WT1	FOXH1
1411	WT1	LMX1B	1412	WT1	PAX7	1413	WT1	PURA
1414	WT1	SIX6	1415	WT1	TFAP2C	1416	XBP1	CDX2
1417	YY1	CDX2	1418	YY1	LMX1B	1419	YY1	ZFP42
1420	ZBTB33	POU5F1	1421	ZBTB33	ZNF589	1422	ZBTB6	FOXJ3
1423	ZBTB6	LHX8	1424	ZBTB6	NANOG	1425	ZBTB6	OTX1
1426	ZBTB7A	PAX5	1427	ZBTB7A	POU5F1	1428	ZBTB7B	ALX3
1429	ZBTB7B	CDX2	1430	ZBTB7B	CRX	1431	ZBTB7B	DMRT3
1432	ZBTB7B	EVX1	1433	ZBTB7B	FOXD3	1434	ZBTB7B	MSX2
1435	ZBTB7B	NKX2-2	1436	ZBTB7B	NKX6-1	1437	ZBTB7B	OTX2
1438	ZBTB7B	PAX7	1439	ZBTB7B	SIX6	1440	ZEB1	ETV7
1441	ZEB1	HOXA11	1442	ZEB1	MNX1	1443	ZEB1	ONECUT1
1444	ZEB1	OTX1	1445	ZEB1	REST	1446	ZEB1	SOX2
1447	ZFP161	POU4F3	1448	ZFP42	CDX2	1449	ZFP42	HOXB13

Continued on next page

Table A.1 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1450	ZFP42	PAX7	1451	ZFP42	VAX1	1452	ZFX	CRX
1453	ZFX	DMRT3	1454	ZFX	EN2	1455	ZFX	OTP
1456	ZFX	OTX2	1457	ZFX	PAX7	1458	ZFX	POU4F3
1459	ZFX	POU5F1	1460	ZFX	SOX10	1461	ZFX	ZFP42
1462	ZIC2	FOXG1	1463	ZIC3	NF1	1464	ZNF143	FOXD1
1465	ZNF143	POU5F1	1466	ZNF143	TLX2	1467	ZNF148	ALX3
1468	ZNF148	CRX	1469	ZNF148	EVX1	1470	ZNF148	FOXD3
1471	ZNF148	FOXH1	1472	ZNF148	LMX1B	1473	ZNF148	NKX6-1
1474	ZNF148	OTX2	1475	ZNF148	PAX7	1476	ZNF148	PURA
1477	ZNF148	SIX6	1478	ZNF148	TFAP2C	1479	ZNF148	ZFP42
1480	ZNF148	ZIC3	1481	ZNF219	DMRT3	1482	ZNF219	NKX2-2
1483	ZNF219	OTX2	1484	ZNF219	PAX7	1485	ZNF219	POU5F1
1486	ZNF219	PURA	1487	ZNF219	SIX6	1488	ZNF219	ZIC3
1489	ZNF263	ALX4	1490	ZNF263	CRX	1491	ZNF263	DMRT3
1492	ZNF263	E2F4	1493	ZNF263	ESX1	1494	ZNF263	EVX1
1495	ZNF263	FOXC1	1496	ZNF263	FOXD3	1497	ZNF263	MEOX1
1498	ZNF263	ONECUT1	1499	ZNF263	PARP1	1500	ZNF263	PAX7
1501	ZNF263	POU5F1	1502	ZNF263	SMAD4	1503	ZNF263	TBX22
1504	ZNF350	ATOH1	1505	ZNF350	FOXC1	1506	ZNF350	NKX6-1
1507	ZNF589	PAX2	1508	ZNF589	TP63	1509	ZNF628	PAX2

Table A.2. 55 ESC regulatory complex-target modules using the ESC specific interactions and protein complexes. TFs are separated by semicolon.

N	Complex ID	TFs in complex	Common targeted TFs	Common targeted TFs in Assou's list
1	HC5737	SREBF1;SREBF2	ARX;EN2;FOXD3;IRX4;NKX2-2; PAX7;POU5F1;TBX22;ZNF148	FOXD3;POU5F1
2	HC4791	KLF4;MZF1	FOXD3;PAX2;ZFP42	FOXD3;ZFP42
3	HC4463	KLF4;ZFX	DMRT3;POU5F1;ZFP42	POU5F1;ZFP42
4	HC5737	KLF4;SREBF2	DMRT3;FOXD3;POU5F1	FOXD3;POU5F1
5	HC8397	ALX4;MZF1	FOXD3;TFAP2C	FOXD3;TFAP2C
6	HC5737	KLF4;SREBF1	FOXD3;POU5F1	FOXD3;POU5F1
7	HC5737	KLF4;SREBF1;SREBF2	FOXD3;POU5F1	FOXD3;POU5F1
8	HC7980	SP1;SP4	ALX3;CDX2;CRX;DMRT3;EVX1; FOXI1;LMX1A;OTX2;PAX7	OTX2
9	HC6813	SP1;ZFX	CRX;DMRT3;OTP;OTX2;PAX7	OTX2
10	HC4806	ELK1;ETS1	HBP1;TCF7;ZIC1;ZIC2	ZIC2
11	HC4830	MZF1;TFAP2A	PAX2;TBX22;ZFP42	ZFP42
12	HC6813	TFAP2A;ZFX	DMRT3;PAX7;ZFP42	ZFP42
13	HC6706	KLF4;TFAP2A	DMRT3;PAX2;ZFP42	ZFP42
14	HC6706	KLF4;SPI1	POU5F1;TFAP2B;TP63	POU5F1
15	HC6161	FOXD3;SP1	FOXI1;OTX2;PAX7	OTX2
16	HC7106	ELK4;ETS1	TCF7;ZIC1;ZIC2	ZIC2
17	HC7106	ELK1;ELK4	TCF7;ZIC1;ZIC2	ZIC2
18	HC7106	ELK1;ELK4;ETS1	TCF7;ZIC1;ZIC2	ZIC2
19	HC5737	EP300;SP1	DMRT3;OTX2;PAX7	OTX2
20	HC8674	USF1;USF2	DMRT1;OVOL2;ZFP42	ZFP42
21	HC4830	TFAP2A;USF1	OVOL2;ZFP42	ZFP42
22	HC4829	ETS1;KLF4	FOXD3;MSX2	FOXD3
23	HC8945	MYB;TFAP2A	PAX7;ZIC3	ZIC3
24	HC8981	MZF1;ZFX	CRX;ZFP42	ZFP42
25	HC4791	KLF4;MZF1;TFAP2A	PAX2;ZFP42	ZFP42
26	HC4750	MAX;TFAP2A	OVOL2;ZFP42	ZFP42
27	HC4463	KLF4;TFAP2A;ZFX	DMRT3;ZFP42	ZFP42
28	HC6507	ETS1;NFKB1	BARHL2;ZIC3	ZIC3
29	HC5737	MZF1;SREBF1	FOXD3;TBX22	FOXD3
30	HC5737	MZF1;SREBF2	FOXD3;TBX22	FOXD3
31	HC5737	MZF1;SREBF1;SREBF2	FOXD3;TBX22	FOXD3
32	HC4824	GATA2;GATA3	FOXC1;FOXJ3;MSX2;NR5A2	
33	HC9343	HSF1;HSF2	FOXJ3;LEF1;MECP2;NKX2-1	
34	HC5737	SP1;SREBF2	CDX2;DMRT3;NR5A2;PAX7	
35	HC5737	EP300;SREBF2	DMRT3;NKX2-2;PAX7	
36	HC5737	SREBF2;TFAP2A	DMRT3;PAX7;TBX22	
37	HC4812	FOXI1;SP1	FOXI1;PAX7	
38	HC6813	SP1;TFAP2A	DMRT3;PAX7	
39	HC6813	SP1;TFAP2A;ZFX	DMRT3;PAX7	
40	HC4771	MYB;SP1	LMX1A;PAX7	
41	HC7991	SOX10;SOX5	CDX2;NFKB2	
42	HC7837	FOXA1;ZEB1	ETV7;HOXA11	
43	HC2082	GATA1;GATA2	SIX3;STAT3	
44	HC9023	TFAP2A;TP53	DMRT3;PAX2	
45	HC6196	EP300;TFAP2A	DMRT3;PAX7	
46	HC9394	FOXM1;ZEB1	ETV7;HOXA11	
47	HC5737	EP300;SREBF1	NKX2-2;PAX7	
48	HC5737	SREBF1;TFAP2A	PAX7;TBX22	
49	HC5737	EP300;SREBF1;SREBF2	NKX2-2;PAX7	
50	HC5737	SREBF1;SREBF2;TFAP2A	PAX7;TBX22	
51	HC5737	EP300;SP1;SREBF2	DMRT3;PAX7	
52	HC5737	SP1;SREBF2;TFAP2A	DMRT3;PAX7	
53	HC5737	EP300;SREBF2;TFAP2A	DMRT3;PAX7	
54	HC5737	EP300;SP1;TFAP2A	DMRT3;PAX7	
55	HC5737	EP300;SP1;SREBF2;TFAP2A	DMRT3;PAX7	

Table A.3. The distributions of nodes and interactions among three layers: top, core, bottom in the hierarchical organization of 41 networks. The entries in red color are those significantly low/high percentages when compared to the others. Abbreviations. T-T: Top→Top. T-C: Top→Core; T-B: Top→Bottom;C-C: Core→Core; C-B: Core→Bottom; B-B: Bottom→Bottom.

	% of nodes in 3 layers			% of interactions between 3 layers					
Network	Top	Core	Bottom	T-T	T-C	T-B	C-C	C-B	B-B
01_AG10803_Skin_Fib	23.7	65.7	10.6	0	13.2	2	74.1	10.6	0
02_AoAF_Aortic_Fib	21.8	69.4	8.8	0	11.5	1.3	78.6	8.5	0
03_CD20+_B_Lymphocyte	22.6	66.8	10.6	0	14.4	1.8	74.7	9.1	0
04_CD34+_Hemat_Stem_Cell	11.5	77	11.5	0	6.7	0.9	82.6	9.8	0
05_fBrain	24.3	63.4	12.4	0.1	13.6	2.1	73	11.2	0
06_fHeart	21.2	68.2	10.6	0	11.1	1	80.6	7.3	0
07_fLung	11.5	79.5	9	0	6.5	0.8	83.7	9	0
08_GM06990_B_Lymphoblastoid	30.6	58.5	10.8	0	17	2.3	71.1	9.5	0
09_GM12865_B_Lymphoblastoid	23.9	66	10.1	0	13.5	1.5	76.7	8.3	0
10_H7-hESC_Embryonic_Stem_Cell	6.2	85.3	8.5	0	3.8	0.3	87.6	8.2	0
11_HAEpiC_Amniotic_Epi	25.9	64.4	9.7	0.1	16.4	1.9	73.9	7.8	0
12_HA-h_Hippocampal	13.3	76.8	9.9	0	9.9	0.8	81.6	7.6	0
13_HCF_Cardiac_Fib	20.9	69.3	9.8	0	10.7	1.1	79.8	8.3	0
14_HCM_Cardiac_Fib	19.4	70.3	10.3	0	10.2	1.1	80.5	8.3	0
15_HCPEpiC_Choroid_Plexus_Epi	22.5	68.3	9.2	0	12	1.1	79.9	6.9	0
16_HEEpiC_Esophageal_Epi	20.9	68.2	11	0	9.1	1.1	79.8	9.9	0
17_HepG2_Hepatoblastoma	28.3	61	10.7	0	19.3	2.9	68.8	8.9	0
18_HFF_Foreskin_Fib	20.2	65.7	14.1	0	10.6	1.4	76.4	11.5	0
19_HIPEpiC_Iris_Pigment_Epi	24.4	64.9	10.7	0	11.6	1.3	78.6	8.5	0
20_HMF_Mamary_Fib	21.1	71.1	7.8	0	12.7	1.3	78.6	7.4	0
21_HMVEC_LLy_Lung_Lymphatic	22.3	67	10.7	0	14.3	2	73.8	9.9	0
22_HMVEC-dBl-Ad_Adult_Derm_Blood	26.4	62.6	11	0	16.8	2.3	71.3	9.6	0
23_HMVEC-dBl-Neo_Derm_Blood	19.6	69.5	10.9	0	12.3	1.6	76.3	9.8	0
24_HMVEC-dLy-Neo_Derm_Lymph	22.8	67.4	9.8	0	14.2	1.6	75.6	8.6	0
25_HPAF_Pulmonary_Artery_Fib	22.5	67.2	10.4	0	11.1	1.1	79.8	8	0
26_HPdLF_Periodontal_Fib	26.4	65.5	8.1	0	14.2	1.5	76	8.3	0
27_HPF_Pulmonary_Fib	23.5	67.9	8.6	0	14.9	1.4	75.8	7.7	0
28_HRCEpiC_Renal_Cortical_Epi	32.8	57.1	10.1	0.1	23	3.3	63.8	9.8	0
29_HSMM_Skeletal_Myoblast	19.4	71	9.6	0	13.8	1.5	76.7	7.9	0
30_HVMF_Mesenchymal_Fib	23.6	67.7	8.7	0	14.4	1.3	77.1	7.1	0
31_IMR90_Fetal_Lung_Fib	27	62.3	10.7	0	15.3	2.4	71.3	10.9	0
32_K562_Erythroid	33	57.6	9.4	0	17.4	2.2	70.7	9.6	0
33_NB4_Leukemia	18.5	72.8	8.7	0	9.3	1.1	81.4	8.2	0
34_NH-A_Astrocyte	29.5	59.4	11.2	0	15.5	2.3	72.3	9.9	0
35_NHDF-Ad_Adult_Dermal_Fib	22.9	66.4	10.7	0	12.7	1.7	76.3	9.2	0
36_NHDF-neo_Neonatal_Dermal_Fib	22.6	70	7.5	0	13.7	1.1	78.8	6.3	0
37_NHLF_Lung_Fib	20.3	68	11.8	0	10.8	1.5	76.8	10.8	0
38_SAEC_Small_Airway_Epi	31.3	58.9	9.8	0	13.7	1.9	74.3	10.1	0
39_SKMC_Skeletal_Muscle	17.6	73.3	9	0	9.3	1	81.3	8.4	0
40_SK-N-SH_RA_Neuroblastoma	25.5	59.3	15.2	0	16.5	3.1	68	12.4	0
41_Th1_T_Lymphocyte	28.9	62.7	8.4	0	15.9	1.3	75.1	7.7	0
Mean(SD)	23(5.6)	67(5.9)	10(1.5)	0	13(3.6)	2(0.6)	76(4.6)	9(1.3)	0
Correlation	1	-1	0.1		1	0.9	-0.9	0.2	
	-1	1	-0.4		0.9	1	-1	0.6	
	0.1	-0.4	1		-0.9	-1	1	-0.5	
					0.2	0.6	-0.5	1	

Table A.4. Local reaching centrality (LRC) and global reaching centrality (GRC) in each of 41 networks. Here we report average LRC of TFs in Top, Core, and Bottom layers. As expected, the LRC of each TF in a layer is always greater than that of each TF in the layers below it in all except two stromal (HCF and HCM) networks from Cardiac Fibroblast.

N	Name	GRC	LRC		
			Top	Core	Bottom
1	B-Lymphocyte	0.085	0.7743	0.7721	0
2	Hemat. Stem Cell	0.108	0.8812	0.879	3e-04
3	B-Lymphoblastoid	0.085	0.6872	0.6849	2e-04
4	B-Lymphoblastoid	0.085	0.7573	0.755	0
5	Erythroid	0.068	0.6667	0.6643	2e-04
6	Promyelocytic Leuk.	0.076	0.813	0.8109	2e-04
7	T-Lymphocyte	0.066	0.7071	0.7048	1e-04
8	Hepatoblastoma	0.081	0.7167	0.7143	2e-04
9	Neuroblastoma	0.121	0.7421	0.7398	1e-04
10	Lung Lymphatic	0.089	0.7768	0.7746	1e-04
11	Adult Dermal Blood	0.086	0.7357	0.7335	1e-04
12	Neonatal Dermal Blood	0.093	0.8039	0.8017	1e-04
13	Neonatal Dermal Lymph.	0.081	0.7718	0.7696	0
14	Amniotic Epi.	0.084	0.7397	0.7374	1e-04
15	Choroid Plexus Epi.	0.079	0.7718	0.7696	2e-04
16	Esophageal Epi.	0.091	0.7931	0.7909	3e-04
17	Iris Pigment Epi.	0.087	0.7544	0.7522	1e-04
18	Renal Cortical Epi	0.078	0.6682	0.6659	1e-04
19	Small Airway Epi.	0.074	0.6856	0.6834	3e-04
20	ESC	0.082	0.9403	0.9382	0
21	Fetal Brain	0.109	0.7523	0.75	2e-04
22	Fetal Heart	0.088	0.7898	0.7876	1e-04
23	Fetal Lung	0.083	0.8868	0.8846	0
24	Skin Fib.	0.092	0.7594	0.7572	0
25	Aortic Fibroblast	0.074	0.7824	0.7802	4e-04
26	Cardiac Fib.	0.083	0.7766	0.7908	0
27	Cardiac Fib.	0.089	0.7899	0.8056	0
28	Foreskin Fib.	0.117	0.7964	0.7942	2e-04
29	Mammary Fib.	0.069	0.7892	0.787	1e-04
30	Pulmonary Artery Fib.	0.088	0.7721	0.7699	0
31	Periodontal Fib.	0.069	0.7347	0.7325	2e-04
32	Pulmonary Fib.	0.074	0.7632	0.761	3e-04
33	Mesenchymal Fib.	0.07	0.7658	0.7636	2e-04
34	Fetal Lung Fib.	0.083	0.7299	0.7277	3e-04
35	Adult Dermal Fib.	0.088	0.7713	0.7691	2e-04
36	Neonatal Dermal Fib.	0.065	0.7743	0.7719	1e-04
37	Lung Fib.	0.099	0.7974	0.7952	2e-04
38	Hippocampal Astrocyte	0.091	0.8667	0.8645	1e-04
39	Skeletal Myoblast	0.083	0.8054	0.8031	1e-04
40	Astrocyte	0.083	0.701	0.6987	1e-04
41	Skeletal Muscle	0.082	0.8237	0.8215	1e-04

Table A.5. 2041 housekeeping (HK) interactions which are found in all the 41 TF regulatory networks.

N	Source	Target	N	Source	Target	N	Source	Target
1	AHR	ATF2	2	AHR	EGR1	3	AHR	GTF2A1
4	AHR	HBP1	5	AHR	RFX1	6	AHR	YY1
7	AHR	ZNF219	8	AIRE	TCF3	9	ALX4	HMBOX1
10	ALX4	JUN	11	ALX4	RB1	12	ALX4	UBP1
13	AR	HMBOX1	14	AR	NR6A1	15	AR	RXRB
16	ARNT	BPTF	17	ARNT	DBP	18	ARNT	DLX2
19	ARNT	EGR1	20	ARNT	GTF2A1	21	ARNT	GZF1
22	ARNT	HINFP	23	ARNT	IRF9	24	ARNT	NFATC3
25	ARNT	TOPORS	26	ATF1	BACH2	27	ATF1	BDP1
28	ATF1	EGR2	29	ATF1	ING4	30	ATF1	JUND
31	ATF1	MAFF	32	ATF1	NF1	33	ATF1	NR6A1
34	ATF1	REL	35	ATF1	RELB	36	ATF1	RFX1
37	ATF1	SREBF1	38	ATF1	STAT3	39	ATF1	TFCP2
40	ATF1	TRIM28	41	ATF2	BACH2	42	ATF2	EGR2
43	ATF2	ING4	44	ATF2	JUND	45	ATF2	MAFF
46	ATF2	NF1	47	ATF2	NR6A1	48	ATF2	RELB
49	ATF2	RFX1	50	ATF2	SREBF1	51	ATF2	STAT3
52	ATF2	TFCP2	53	ATF2	TRIM28	54	ATF3	BACH2
55	ATF3	BDP1	56	ATF3	ING4	57	ATF3	JUN
58	ATF3	MAFF	59	ATF3	NF1	60	ATF3	NR6A1
61	ATF3	RELB	62	ATF3	RFX1	63	ATF3	SREBF1
64	ATF3	STAT3	65	ATF3	TFCP2	66	ATF3	TRIM28
67	ATF4	BACH2	68	ATF4	ING4	69	ATF4	MAFF
70	ATF4	NF1	71	ATF4	NR6A1	72	ATF4	RELB
73	ATF4	RFX1	74	ATF4	SREBF1	75	ATF4	SREBF2
76	ATF4	STAT3	77	ATF4	TFCP2	78	ATF4	TRIM28
79	ATF5	BACH2	80	ATF5	ING4	81	ATF5	MAFF
82	ATF5	NF1	83	ATF5	NR6A1	84	ATF5	OAZ1
85	ATF5	RELB	86	ATF5	RFX1	87	ATF5	SREBF1
88	ATF5	STAT3	89	ATF5	TFCP2	90	ATF5	TRIM28
91	ATF6	BACH2	92	ATF6	ING4	93	ATF6	MAFF
94	ATF6	NF1	95	ATF6	NR6A1	96	ATF6	RELB
97	ATF6	RFX1	98	ATF6	SREBF1	99	ATF6	STAT3
100	ATF6	TFCP2	101	ATF6	TRIM28	102	ATF7	BACH2
103	ATF7	ING4	104	ATF7	MAFF	105	ATF7	NF1
106	ATF7	NR6A1	107	ATF7	RELB	108	ATF7	RFX1
109	ATF7	SREBF1	110	ATF7	STAT3	111	ATF7	TFCP2
112	ATF7	TRIM28	113	ATOH1	CEBPE	114	ATOH1	DLX2
115	BACH1	GTF2A1	116	BACH2	GTF2A1	117	BARHL1	CNOT3
118	BARHL2	CNOT3	119	BCL6	GTF2A1	120	BHLHE41	GTF2A1
121	BHLHE41	HSF2	122	CDX1	HES1	123	CDX2	BCL6
124	CDX2	HES1	125	CEBPA	NFE2L1	126	CEBPD	SREBF2
127	CNOT3	BACH1	128	CNOT3	CTCF	129	CNOT3	EGR1
130	CNOT3	FOXN2	131	CNOT3	FOXO3	132	CNOT3	HSF2
133	CNOT3	JUND	134	CNOT3	MAFF	135	CNOT3	MAX
136	CNOT3	NF1	137	CNOT3	NFE2L2	138	CNOT3	NFYA
139	CNOT3	PITX3	140	CNOT3	PKNOX1	141	CNOT3	RB1
142	CNOT3	SMAD2	143	CNOT3	SP4	144	CNOT3	SREBF1

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
145	CNOT3	SREBF2	146	CNOT3	SRF	147	CNOT3	TFAP4
148	CNOT3	ZFP161	149	CNOT3	ZNF143	150	CNOT3	ZNF263
151	CREB1	BACH2	152	CREB1	E4F1	153	CREB1	EGR2
154	CREB1	FOXP3	155	CREB1	ING4	156	CREB1	JUND
157	CREB1	MAFF	158	CREB1	NF1	159	CREB1	NFE2L2
160	CREB1	NR6A1	161	CREB1	REL	162	CREB1	RELB
163	CREB1	RFX1	164	CREB1	SREBF1	165	CREB1	STAT3
166	CREB1	TFCP2	167	CREB1	TRIM28	168	CREM	BACH2
169	CREM	EGR2	170	CREM	ING4	171	CREM	JUND
172	CREM	MAFF	173	CREM	NF1	174	CREM	NR6A1
175	CREM	RELB	176	CREM	RFX1	177	CREM	SREBF1
178	CREM	STAT3	179	CREM	TFCP2	180	CREM	TRIM28
181	CRX	DDIT3	182	CTCF	BCL6	183	CTCF	CEBPE
184	CTCF	CTCF	185	CTCF	DLX2	186	CTCF	DLX3
187	CTCF	E4F1	188	CTCF	EGR1	189	CTCF	EP300
190	CTCF	ESRRA	191	CTCF	GLI1	192	CTCF	GZF1
193	CTCF	HOMEZ	194	CTCF	HSF2	195	CTCF	IRF2
196	CTCF	IRF9	197	CTCF	MAFA	198	CTCF	MAZ
199	CTCF	MTERF	200	CTCF	MTF1	201	CTCF	NFE2L1
202	CTCF	NFE2L2	203	CTCF	NFKB2	204	CTCF	NFYA
205	CTCF	PATZ1	206	CTCF	PKNOX1	207	CTCF	POU2F1
208	CTCF	PURA	209	CTCF	RFX2	210	CTCF	SP1
211	CTCF	SP3	212	CTCF	SRF	213	CTCF	STAT1
214	CTCF	TCF12	215	CTCF	TP53	216	CTCF	ZBTB33
217	CTCF	ZBTB7A	218	CTCF	ZBTB7B	219	DEAF1	CREB1
220	DEAF1	CTCF	221	DEAF1	DEAF1	222	DEAF1	EGR1
223	DEAF1	HMBOX1	224	DEAF1	JUNB	225	DEAF1	MAX
226	DEAF1	MZF1	227	DEAF1	SHOX2	228	DEAF1	TCF12
229	DEAF1	TP53	230	DEAF1	ZBTB33	231	DMRT1	SRF
232	DMRT2	SRF	233	E2F1	ATF2	234	E2F1	ATF4
235	E2F1	CREM	236	E2F1	DLX2	237	E2F1	E2F1
238	E2F1	JUNB	239	E2F1	MAZ	240	E2F1	PKNOX1
241	E2F1	ZFP161	242	E2F1	ZNF143	243	E2F4	ATF4
244	E2F4	E2F1	245	E2F4	MAZ	246	E2F6	ATF4
247	E2F6	E2F1	248	E2F6	MAZ	249	E2F7	ATF4
250	E2F7	E2F1	251	E2F7	MAZ	252	E4F1	RFX1
253	EBF1	EGR1	254	EBF1	HOMEZ	255	EBF1	MAX
256	EBF1	RB1	257	EBF1	ZNF143	258	EBF2	EGR1
259	EBF2	HOMEZ	260	EBF2	RB1	261	EBF2	ZNF143
262	EGR1	ATF1	263	EGR1	ATF2	264	EGR1	ATF4
265	EGR1	BHLHE40	266	EGR1	BRF1	267	EGR1	CEBPB
268	EGR1	CNOT3	269	EGR1	CREM	270	EGR1	DBP
271	EGR1	DDIT3	272	EGR1	EP300	273	EGR1	FOXH1
274	EGR1	FOXJ3	275	EGR1	FOXN2	276	EGR1	FOXO3
277	EGR1	GABPA	278	EGR1	GTF2A1	279	EGR1	GTF2I
280	EGR1	HBP1	281	EGR1	HES1	282	EGR1	HIF1A
283	EGR1	HSF2	284	EGR1	IRF1	285	EGR1	JUNB
286	EGR1	JUND	287	EGR1	MAX	288	EGR1	MAZ
289	EGR1	MZF1	290	EGR1	NFATC3	291	EGR1	NFE2L1
292	EGR1	NFE2L2	293	EGR1	NFYA	294	EGR1	NR4A1

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
295	EGR1	NR6A1	296	EGR1	OAZ1	297	EGR1	PKNOX1
298	EGR1	POU2F1	299	EGR1	RBPJ	300	EGR1	REL
301	EGR1	RELB	302	EGR1	RFX1	303	EGR1	RFX2
304	EGR1	RXRB	305	EGR1	SP1	306	EGR1	SP4
307	EGR1	SREBF1	308	EGR1	SRF	309	EGR1	STAT1
310	EGR1	STAT3	311	EGR1	TCF12	312	EGR1	TEF
313	EGR1	TOPORS	314	EGR1	TP53	315	EGR1	TRIM28
316	EGR1	UBP1	317	EGR1	YY1	318	EGR1	ZBTB7B
319	EGR1	ZNF143	320	EGR1	ZNF238	321	EGR1	ZNF333
322	EGR1	ZNF628	323	EGR2	ATF1	324	EGR2	ATF2
325	EGR2	ATF4	326	EGR2	BHLHE40	327	EGR2	BRF1
328	EGR2	CEBPB	329	EGR2	CNOT3	330	EGR2	CREM
331	EGR2	DBP	332	EGR2	DDIT3	333	EGR2	EP300
334	EGR2	FOXH1	335	EGR2	FOXJ3	336	EGR2	FOXN2
337	EGR2	FOXO3	338	EGR2	GABPA	339	EGR2	GTF2A1
340	EGR2	GTF2I	341	EGR2	HBP1	342	EGR2	HES1
343	EGR2	HIF1A	344	EGR2	HSF2	345	EGR2	IRF1
346	EGR2	JUNB	347	EGR2	JUND	348	EGR2	MAX
349	EGR2	MAZ	350	EGR2	MZF1	351	EGR2	NFATC3
352	EGR2	NFE2L1	353	EGR2	NFE2L2	354	EGR2	NFYA
355	EGR2	NR4A1	356	EGR2	NR6A1	357	EGR2	OAZ1
358	EGR2	PKNOX1	359	EGR2	POU2F1	360	EGR2	RBPJ
361	EGR2	REL	362	EGR2	RELB	363	EGR2	RFX1
364	EGR2	RFX2	365	EGR2	RXRB	366	EGR2	SP1
367	EGR2	SP4	368	EGR2	SREBF1	369	EGR2	SRF
370	EGR2	STAT1	371	EGR2	STAT3	372	EGR2	TCF12
373	EGR2	TEF	374	EGR2	TFAP4	375	EGR2	TOPORS
376	EGR2	TP53	377	EGR2	TRIM28	378	EGR2	UBP1
379	EGR2	YY1	380	EGR2	ZBTB7B	381	EGR2	ZNF143
382	EGR2	ZNF238	383	EGR2	ZNF333	384	EGR2	ZNF628
385	EGR3	ATF1	386	EGR3	ATF2	387	EGR3	ATF4
388	EGR3	BHLHE40	389	EGR3	BRF1	390	EGR3	CEBPB
391	EGR3	CNOT3	392	EGR3	CREM	393	EGR3	DBP
394	EGR3	DDIT3	395	EGR3	EP300	396	EGR3	FOXH1
397	EGR3	FOXJ3	398	EGR3	FOXN2	399	EGR3	FOXO3
400	EGR3	GABPA	401	EGR3	GTF2A1	402	EGR3	GTF2I
403	EGR3	HBP1	404	EGR3	HES1	405	EGR3	HIF1A
406	EGR3	HSF2	407	EGR3	IRF1	408	EGR3	JUNB
409	EGR3	JUND	410	EGR3	MAX	411	EGR3	MAZ
412	EGR3	MZF1	413	EGR3	NFATC3	414	EGR3	NFE2L1
415	EGR3	NFE2L2	416	EGR3	NFYA	417	EGR3	NR4A1
418	EGR3	NR6A1	419	EGR3	OAZ1	420	EGR3	PKNOX1
421	EGR3	POU2F1	422	EGR3	RBPJ	423	EGR3	REL
424	EGR3	RELB	425	EGR3	RFX1	426	EGR3	RFX2
427	EGR3	RXRB	428	EGR3	SP1	429	EGR3	SP4
430	EGR3	SREBF1	431	EGR3	SRF	432	EGR3	STAT1
433	EGR3	STAT3	434	EGR3	TCF12	435	EGR3	TEF
436	EGR3	TOPORS	437	EGR3	TP53	438	EGR3	TRIM28
439	EGR3	UBP1	440	EGR3	YY1	441	EGR3	ZBTB7B
442	EGR3	ZNF143	443	EGR3	ZNF238	444	EGR3	ZNF333

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
445	EGR3	ZNF628	446	EGR4	ATF1	447	EGR4	ATF2
448	EGR4	ATF4	449	EGR4	BHLHE40	450	EGR4	BRF1
451	EGR4	CEBPB	452	EGR4	CNOT3	453	EGR4	CREM
454	EGR4	DBP	455	EGR4	DDIT3	456	EGR4	EP300
457	EGR4	FOXH1	458	EGR4	FOXJ3	459	EGR4	FOXN2
460	EGR4	FOXO3	461	EGR4	GABPA	462	EGR4	GTF2I
463	EGR4	HBP1	464	EGR4	HES1	465	EGR4	HIF1A
466	EGR4	HSF2	467	EGR4	IRF1	468	EGR4	JUNB
469	EGR4	JUND	470	EGR4	MAZ	471	EGR4	MZF1
472	EGR4	NFATC3	473	EGR4	NFE2L1	474	EGR4	NFE2L2
475	EGR4	NFYA	476	EGR4	NR4A1	477	EGR4	PKNOX1
478	EGR4	POU2F1	479	EGR4	RBPJ	480	EGR4	REL
481	EGR4	RELB	482	EGR4	RFX1	483	EGR4	RFX2
484	EGR4	RXRB	485	EGR4	SP4	486	EGR4	SRF
487	EGR4	STAT1	488	EGR4	STAT3	489	EGR4	TCF12
490	EGR4	TEF	491	EGR4	TOPORS	492	EGR4	TP53
493	EGR4	TRIM28	494	EGR4	UBP1	495	EGR4	YY1
496	EGR4	ZBTB7B	497	EGR4	ZNF238	498	EGR4	ZNF333
499	EGR4	ZNF628	500	EHF	EGR1	501	EHF	TBP
502	ELF1	DDIT3	503	ELF2	CNOT3	504	ELF2	DDIT3
505	ELF2	GABPA	506	ELF2	GTF2I	507	ELF2	MTF1
508	ELF2	PITX3	509	ELF2	SP3	510	ELF3	ZNF143
511	ELK1	DDIT3	512	ELK1	ELK4	513	ELK1	ING4
514	ELK1	MTERF	515	ELK1	MTF1	516	ELK1	MZF1
517	ELK1	NR1H2	518	ELK1	SIRT6	519	ELK1	SP3
520	ELK1	TBP	521	ELK4	DDIT3	522	ELK4	MZF1
523	ELK4	TBP	524	EP300	MTF1	525	EP300	UBP1
526	ERF	DDIT3	527	ERG	DDIT3	528	ERG	NR1H2
529	ESR1	CTCF	530	ESR1	JUND	531	ETS1	CDC5L
532	ETS1	DDIT3	533	ETS1	EGR1	534	ETS1	GABPA
535	ETS1	HMBOX1	536	ETS1	MAZ	537	ETS1	MTERF
538	ETS1	MZF1	539	ETS1	NFYA	540	ETS1	NR1H2
541	ETS1	RXRB	542	ETS1	SIRT6	543	ETS1	SP3
544	ETS2	DDIT3	545	ETS2	EGR1	546	ETS2	GABPA
547	ETS2	MAZ	548	ETS2	SP3	549	ETV7	DDIT3
550	ETV7	ELK4	551	ETV7	ING4	552	ETV7	TBP
553	ETV7	UBP1	554	FLI1	ATF4	555	FLI1	DDIT3
556	FLI1	ELK4	557	FLI1	ING4	558	FLI1	TBP
559	FOSL1	BHLHE40	560	FOXJ2	TP53	561	FOXM1	USF1
562	FOXO3	HBP1	563	FOXO4	ZNF263	564	FOXP1	NR1I3
565	GABPA	DDIT3	566	GABPA	EGR1	567	GABPA	ING4
568	GABPA	MZF1	569	GABPA	NFYA	570	GABPA	NR1H2
571	GABPA	SIRT6	572	GABPA	SP3	573	GABPA	TBP
574	GABPB1	DDIT3	575	GABPB1	EGR1	576	GABPB1	ING4
577	GABPB1	MZF1	578	GABPB1	NFYA	579	GABPB1	NR1H2
580	GABPB1	SIRT6	581	GABPB1	SP3	582	GABPB1	TBP
583	GATA1	CTCF	584	GATA2	FOXP3	585	GBX2	NR2C2
586	GFI1	SREBF2	587	GLI3	CNOT3	588	GLI3	HINFP
589	GLIS3	NR6A1	590	GLIS3	SP1	591	GTF2A1	JUND
592	GTF2I	EGR1	593	GTF2I	FOXN2	594	GTF2I	IRF2

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
595	GTF2I	MAZ	596	GTF2I	STAT1	597	GTF2I	STAT3
598	HES1	HES1	599	HES1	MAZ	600	HIC1	ATF7
601	HIC1	BHLHE40	602	HIC1	DDIT3	603	HIC1	DEAF1
604	HIC1	FOXJ1	605	HIC1	GLI1	606	HIC1	GZF1
607	HIC1	HIF1A	608	HIC1	HOMEZ	609	HIC1	IRF1
610	HIC1	JUNB	611	HIC1	JUND	612	HIC1	MAX
613	HIC1	MAZ	614	HIC1	NFE2L2	615	HIC1	NFYA
616	HIC1	NR6A1	617	HIC1	PKNOX1	618	HIC1	RELA
619	HIC1	SP1	620	HIC1	SREBF2	621	HIC1	SRF
622	HIC1	TEF	623	HIC1	TP53	624	HIC1	XBP1
625	HIC1	ZBTB7A	626	HIC1	ZFP161	627	HIVEP2	GTF2I
628	HMX1	NFYA	629	HMX3	CNOT3	630	HNF4A	ATF7
631	HNF4A	FOXO3	632	HNF4G	ATF7	633	HNF4G	FOXO3
634	HOXA11	E4F1	635	HOXA4	TEF	636	HOXA9	PITX3
637	HOXC10	E4F1	638	HOXC11	E4F1	639	HOXC12	E4F1
640	HOXC4	TEF	641	HOXC5	TEF	642	HOXD12	E4F1
643	HSF1	CHURC1	644	HSF1	POU2F1	645	HSF1	ZNF143
646	HSF2	CHURC1	647	IKZF1	DBP	648	JUN	BHLHE40
649	JUNB	BHLHE40	650	JUND	BHLHE40	651	KLF11	DDIT3
652	KLF11	ESRRA	653	KLF11	JUND	654	KLF11	NFATC3
655	KLF11	TGIF1	656	KLF11	TP53	657	KLF11	ZNF238
658	KLF15	ATF5	659	KLF15	BHLHE40	660	KLF15	BRF1
661	KLF15	CEBPB	662	KLF15	E4F1	663	KLF15	EGR1
664	KLF15	EP300	665	KLF15	ERF	666	KLF15	ESRRA
667	KLF15	FOXN2	668	KLF15	FOXO3	669	KLF15	GTF2I
670	KLF15	HIF1A	671	KLF15	JUND	672	KLF15	MAX
673	KLF15	MAZ	674	KLF15	NFATC3	675	KLF15	NR1H2
676	KLF15	NR4A1	677	KLF15	PKNOX1	678	KLF15	POU2F1
679	KLF15	RBPJ	680	KLF15	RELB	681	KLF15	RFX1
682	KLF15	RXRB	683	KLF15	SIRT6	684	KLF15	SP1
685	KLF15	SREBF1	686	KLF15	SRF	687	KLF15	STAT6
688	KLF15	TCF12	689	KLF15	TEF	690	KLF15	TP53
691	KLF15	TRIM28	692	KLF15	ZBTB7B	693	KLF15	ZFP161
694	KLF15	ZNF143	695	KLF4	ATF5	696	KLF4	ATF7
697	KLF4	BRF1	698	KLF4	CREM	699	KLF4	CTCF
700	KLF4	DBP	701	KLF4	DDIT3	702	KLF4	DLX2
703	KLF4	E4F1	704	KLF4	EP300	705	KLF4	ESRRA
706	KLF4	FOXN2	707	KLF4	GTF2A1	708	KLF4	GTF2I
709	KLF4	GZF1	710	KLF4	HES1	711	KLF4	HINFP
712	KLF4	HSF1	713	KLF4	IRF1	714	KLF4	JUND
715	KLF4	KLF11	716	KLF4	MAZ	717	KLF4	MZF1
718	KLF4	NFATC3	719	KLF4	NFE2L2	720	KLF4	NFYA
721	KLF4	NR1H2	722	KLF4	NR6A1	723	KLF4	OAZ1
724	KLF4	PITX3	725	KLF4	PKNOX1	726	KLF4	RBPJ
727	KLF4	REL	728	KLF4	RELB	729	KLF4	RFX1
730	KLF4	RFX2	731	KLF4	RXRB	732	KLF4	SIRT6
733	KLF4	SP1	734	KLF4	SP4	735	KLF4	SREBF2
736	KLF4	SRF	737	KLF4	STAT3	738	KLF4	TBP
739	KLF4	TCF3	740	KLF4	TEF	741	KLF4	TOPORS
742	KLF4	TRIM28	743	KLF4	UBP1	744	KLF4	USF1

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
745	KLF4	YY1	746	KLF4	ZFP161	747	KLF4	ZNF238
748	KLF4	ZNF333	749	KLF4	ZNF628	750	LHX6	E2F7
751	LHX8	E2F7	752	MAF	CDC5L	753	MAF	GTF2A1
754	MAF	MTF1	755	MAF	SIRT6	756	MAFB	GTF2A1
757	MAFF	GTF2A1	758	MAFG	GTF2A1	759	MAFG	PKNOX1
760	MAX	DBP	761	MAX	FOXO3	762	MAX	GTF2A1
763	MAX	HSF1	764	MAX	IRF9	765	MAX	JUNB
766	MAX	JUND	767	MAX	MAX	768	MAX	NFE2L2
769	MAX	PITX3	770	MAX	SMAD7	771	MAX	ZFP161
772	MAZ	ATF2	773	MAZ	BHLHE40	774	MAZ	DDIT3
775	MAZ	E4F1	776	MAZ	ELK4	777	MAZ	EP300
778	MAZ	JUN	779	MAZ	MAZ	780	MAZ	MECP2
781	MAZ	NR6A1	782	MAZ	PITX3	783	MAZ	POU2F1
784	MAZ	RXRB	785	MAZ	SP2	786	MAZ	SP3
787	MAZ	STAT3	788	MAZ	TCF12	789	MAZ	TEF
790	MAZ	ZBTB7B	791	MEF2A	BDP1	792	MEF2A	JUN
793	MEF2C	JUN	794	MEIS1	PITX3	795	MTF1	RELB
796	MYB	NFYA	797	MYB	PKNOX1	798	MYB	RFX2
799	MYB	UBP1	800	MYC	DBP	801	MYC	FOXO3
802	MYC	GTF2A1	803	MYC	HSF1	804	MYC	JUNB
805	MYC	JUND	806	MYC	NFE2L2	807	MYC	NR6A1
808	MYC	PITX3	809	MYC	ZFP161	810	MYCN	MAX
811	MYCN	PITX3	812	MYF6	ATF1	813	MYF6	HES1
814	MYF6	HMBOX1	815	MYF6	TEF	816	MYF6	TERF1
817	MYF6	ZBTB7A	818	MYOD1	ATF1	819	MYOD1	HES1
820	MYOD1	HMBOX1	821	MYOD1	TEF	822	MYOD1	TERF1
823	MYOD1	ZBTB7A	824	MYOG	ATF1	825	MYOG	HES1
826	MYOG	HMBOX1	827	MYOG	TEF	828	MYOG	TERF1
829	MYOG	ZBTB7A	830	MZF1	BHLHE40	831	MZF1	DBP
832	MZF1	EGR1	833	MZF1	MAZ	834	MZF1	MECP2
835	MZF1	POU2F1	836	MZF1	SP1	837	MZF1	SP4
838	NANOG	ING4	839	NEUROD1	ARNT	840	NEUROD1	CDC5L
841	NF1	FOXP3	842	NF1	HBP1	843	NF1	MAZ
844	NF1	PATZ1	845	NF1	ZFP161	846	NFATC1	MAX
847	NFATC2	MAX	848	NFATC3	MAX	849	NFATC4	MAX
850	NFE2	FOXA3	851	NFE2	GTF2A1	852	NFE2L1	BDP1
853	NFE2L1	GTF2A1	854	NFE2L1	PKNOX1	855	NFE2L1	TBP
856	NFE2L2	DDIT3	857	NFE2L2	EGR1	858	NFE2L2	GTF2A1
859	NFE2L2	ING4	860	NFE2L2	MZF1	861	NFE2L2	NFYA
862	NFE2L2	NR1H2	863	NFE2L2	SIRT6	864	NFE2L2	SP3
865	NFE2L2	TBP	866	NFIB	MAZ	867	NFIB	ZFP161
868	NFIX	FOXP3	869	NFIX	HBP1	870	NFIX	MAZ
871	NFIX	PATZ1	872	NFIX	ZFP161	873	NFKB1	EGR1
874	NFKB1	HES1	875	NFKB1	IRF1	876	NFKB1	IRF2
877	NFKB1	JUNB	878	NFKB1	NFE2L2	879	NFKB1	NFKB2
880	NFKB1	NR4A1	881	NFKB1	RBPJ	882	NFKB1	REL
883	NFKB1	RFX5	884	NFKB2	HES1	885	NFKB2	IRF1
886	NFKB2	IRF2	887	NFKB2	NFE2L2	888	NFKB2	NFKB2
889	NFKB2	NR4A1	890	NFKB2	REL	891	NFKB2	RFX5
892	NFYA	ATF4	893	NFYA	ATF7	894	NFYA	CNOT3

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
895	NFYA	DBP	896	NFYA	DLX1	897	NFYA	E2F7
898	NFYA	FOXP3	899	NFYA	HES1	900	NFYA	HSF1
901	NFYA	NFE2L1	902	NFYA	NFYA	903	NFYA	RFX5
904	NFYA	SMAD2	905	NFYA	SP2	906	NFYA	SP3
907	NFYA	SRF	908	NFYA	STAT3	909	NFYA	TEF
910	NFYA	TERF1	911	NFYA	YY1	912	NFYA	ZBTB7B
913	NFYA	ZNF628	914	NHLH1	EGR1	915	NHLH1	RFX1
916	NKX2-1	HINFP	917	NKX2-1	TRIM28	918	NKX6-1	CNOT3
919	NR0B1	CREM	920	NR0B1	SP2	921	NR1H2	DDIT3
922	NR1H2	ING4	923	NR1H2	IRF9	924	NR1H2	SREBF1
925	NR1H4	ATF4	926	NR1I2	BHLHE40	927	NR1I2	CNOT3
928	NR1I2	ING4	929	NR1I2	IRF9	930	NR1I2	SREBF1
931	NR1I2	TBP	932	NR1I3	BHLHE40	933	NR1I3	CNOT3
934	NR1I3	ING4	935	NR1I3	IRF9	936	NR1I3	SREBF1
937	NR1I3	TBP	938	NR2F1	CNOT3	939	NR2F1	FOXO3
940	NR2F1	ING4	941	NR2F1	IRF9	942	NR2F1	MAFF
943	NR2F1	RBPJ	944	NR2F1	RELA	945	NR2F1	SREBF1
946	NR2F1	TFAP4	947	NR2F1	YY1	948	NR2F1	ZBTB7B
949	NR2F2	CNOT3	950	NR2F2	FOXO3	951	NR2F2	ING4
952	NR2F2	IRF9	953	NR2F2	MAFF	954	NR2F2	RBPJ
955	NR2F2	RELA	956	NR2F2	SREBF1	957	NR2F2	TFAP4
958	NR2F2	YY1	959	NR2F2	ZBTB7B	960	NR3C1	DDIT3
961	NR3C1	HMBOX1	962	NR3C1	NR6A1	963	NR3C1	RXRB
964	NR5A2	ARNT	965	NR6A1	GABPA	966	NRF1	ATF5
967	NRF1	CHURC1	968	NRF1	DDIT3	969	NRF1	GTF2I
970	NRF1	GZF1	971	NRF1	HBP1	972	NRF1	JUNB
973	NRF1	MAZ	974	NRF1	NF1	975	NRF1	NFYA
976	NRF1	POU2F1	977	NRF1	RXRB	978	NRF1	SIRT6
979	NRF1	SMAD4	980	NRF1	SREBF1	981	NRF1	TEF
982	NRF1	TOPORS	983	NRF1	ZFP161	984	OAZ1	JUND
985	OTX1	DDIT3	986	OTX2	DDIT3	987	OTX2	ZNF589
988	PATZ1	ATF2	989	PATZ1	DDIT3	990	PATZ1	EGR1
991	PATZ1	ELK4	992	PATZ1	EP300	993	PATZ1	ESRRA
994	PATZ1	FOXN2	995	PATZ1	HIF1A	996	PATZ1	JUN
997	PATZ1	MAZ	998	PATZ1	NFATC3	999	PATZ1	NFYA
1000	PATZ1	PITX3	1001	PATZ1	PKNOX1	1002	PATZ1	POU2F1
1003	PATZ1	SP1	1004	PATZ1	SP4	1005	PATZ1	TEF
1006	PATZ1	TGIF1	1007	PATZ1	TP53	1008	PATZ1	ZBTB7B
1009	PAX2	STAT3	1010	PAX3	DBP	1011	PAX4	ATF5
1012	PAX4	ATF7	1013	PAX4	CBFB	1014	PAX4	DBP
1015	PAX4	DDIT3	1016	PAX4	FOXN2	1017	PAX4	HIF1A
1018	PAX4	JUND	1019	PAX4	MZF1	1020	PAX4	NFE2L2
1021	PAX4	NR6A1	1022	PAX4	OAZ1	1023	PAX4	REL
1024	PAX4	RELB	1025	PAX4	SP4	1026	PAX4	SRF
1027	PAX5	BRF1	1028	PAX5	SMAD2	1029	PAX5	TCF3
1030	PAX6	TEF	1031	PDX1	ZNF143	1032	PGR	HMBOX1
1033	PGR	RXRB	1034	POU1F1	TEF	1035	POU2AF1	EGR2
1036	POU2AF1	JUND	1037	POU2AF1	SREBF2	1038	POU2F1	EGR2
1039	POU2F1	FOXN2	1040	POU2F1	JUND	1041	POU2F1	SMAD3
1042	POU2F1	SREBF2	1043	POU2F2	EGR2	1044	POU2F2	JUND

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1045	POU2F2	SREBF2	1046	POU2F3	EGR2	1047	POU2F3	JUND
1048	POU2F3	SMAD3	1049	POU2F3	SREBF2	1050	POU3F1	EGR2
1051	POU3F1	JUND	1052	POU3F1	SREBF2	1053	POU3F1	STAT3
1054	POU3F2	EGR2	1055	POU3F2	JUND	1056	POU3F2	SREBF2
1057	POU3F2	STAT3	1058	POU3F3	EGR2	1059	POU3F3	JUND
1060	POU3F3	SREBF2	1061	POU5F1	EGR2	1062	POU5F1	JUND
1063	POU5F1	SREBF2	1064	PPARA	ATF2	1065	PPARA	JUND
1066	PPARG	GABPA	1067	PPARG	SP2	1068	PURA	BHLHE40
1069	PURA	CNOT3	1070	PURA	ZBTB7B	1071	RARA	ING4
1072	RARA	IRF9	1073	RARA	SREBF1	1074	RARA	ZNF143
1075	RBPJ	RBPJ	1076	REL	HES1	1077	REL	IRF2
1078	REL	NFKB2	1079	RELA	EGR1	1080	RELA	HES1
1081	RELA	IRF1	1082	RELA	IRF2	1083	RELA	JUNB
1084	RELA	NFE2L2	1085	RELA	NFKB2	1086	RELA	RBPJ
1087	RELA	REL	1088	RELA	RFX5	1089	RELB	IRF2
1090	RELB	NFE2L2	1091	RELB	REL	1092	REST	CDC5L
1093	REST	GLI1	1094	REST	NRF1	1095	REST	ZNF219
1096	RFX1	ATF7	1097	RFX1	BDP1	1098	RFX1	BRCA1
1099	RFX1	FOXN2	1100	RFX1	GTF2I	1101	RFX1	JUNB
1102	RFX1	MZF1	1103	RFX1	NR2F2	1104	RFX1	RFX2
1105	RFX2	JUNB	1106	RFX5	JUNB	1107	RORA	DBP
1108	RREB1	NR6A1	1109	RXRA	ATF2	1110	RXRA	ATF4
1111	RXRA	BHLHE40	1112	RXRA	CNOT3	1113	RXRA	DDIT3
1114	RXRA	ING4	1115	RXRA	IRF9	1116	RXRA	JUND
1117	RXRA	SREBF1	1118	RXRA	TBP	1119	RXRA	ZNF143
1120	RXRB	BHLHE40	1121	RXRB	CNOT3	1122	RXRB	ING4
1123	RXRB	IRF9	1124	RXRB	SREBF1	1125	RXRB	TBP
1126	RXRB	ZNF143	1127	SIX4	GLI1	1128	SMAD4	NFYA
1129	SMAD4	PKNOX1	1130	SP1	ATF1	1131	SP1	ATF2
1132	SP1	ATF4	1133	SP1	ATF5	1134	SP1	ATF7
1135	SP1	BHLHE40	1136	SP1	BRF1	1137	SP1	CEBPB
1138	SP1	CHURC1	1139	SP1	CNOT3	1140	SP1	CREM
1141	SP1	CTCF	1142	SP1	DBP	1143	SP1	DDIT3
1144	SP1	DLX2	1145	SP1	E4F1	1146	SP1	EGR1
1147	SP1	ELK4	1148	SP1	EP300	1149	SP1	ERF
1150	SP1	ESRRA	1151	SP1	ETV4	1152	SP1	FOXA3
1153	SP1	FOXH1	1154	SP1	FOXJ1	1155	SP1	FOXJ3
1156	SP1	FOXN2	1157	SP1	FOXO3	1158	SP1	FOXP3
1159	SP1	GABPA	1160	SP1	GTF2A1	1161	SP1	GTF2I
1162	SP1	GZF1	1163	SP1	HBP1	1164	SP1	HES1
1165	SP1	HIF1A	1166	SP1	HINFP	1167	SP1	HOMEZ
1168	SP1	HSF1	1169	SP1	HSF2	1170	SP1	IRF1
1171	SP1	IRF2	1172	SP1	JUN	1173	SP1	JUNB
1174	SP1	JUND	1175	SP1	KLF11	1176	SP1	MAFF
1177	SP1	MAX	1178	SP1	MAZ	1179	SP1	MECP2
1180	SP1	MZF1	1181	SP1	NFATC3	1182	SP1	NFE2L1
1183	SP1	NFE2L2	1184	SP1	NFKB2	1185	SP1	NFYA
1186	SP1	NR1H2	1187	SP1	NR2C2	1188	SP1	NR3C1
1189	SP1	NR4A1	1190	SP1	NR6A1	1191	SP1	OAZ1
1192	SP1	PARP1	1193	SP1	PITX3	1194	SP1	PKNOX1

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1195	SP1	POU2F1	1196	SP1	PPARD	1197	SP1	RBPJ
1198	SP1	REL	1199	SP1	RELA	1200	SP1	RELB
1201	SP1	RFX1	1202	SP1	RFX2	1203	SP1	RXRB
1204	SP1	SIRT6	1205	SP1	SMAD2	1206	SP1	SP1
1207	SP1	SP2	1208	SP1	SP3	1209	SP1	SP4
1210	SP1	SREBF1	1211	SP1	SREBF2	1212	SP1	SRF
1213	SP1	STAT1	1214	SP1	STAT2	1215	SP1	STAT3
1216	SP1	STAT6	1217	SP1	TBP	1218	SP1	TCF12
1219	SP1	TCF3	1220	SP1	TEF	1221	SP1	TGIF1
1222	SP1	TOPORS	1223	SP1	TP53	1224	SP1	TRIM28
1225	SP1	UBP1	1226	SP1	USF1	1227	SP1	YY1
1228	SP1	ZBTB33	1229	SP1	ZBTB7A	1230	SP1	ZBTB7B
1231	SP1	ZFP161	1232	SP1	ZNF143	1233	SP1	ZNF238
1234	SP1	ZNF333	1235	SP1	ZNF628	1236	SP2	ATF1
1237	SP2	ATF2	1238	SP2	ATF4	1239	SP2	ATF5
1240	SP2	ATF7	1241	SP2	BHLHE40	1242	SP2	BRF1
1243	SP2	CEBPB	1244	SP2	CHURC1	1245	SP2	CNOT3
1246	SP2	CREM	1247	SP2	CTCF	1248	SP2	DBP
1249	SP2	DDIT3	1250	SP2	DLX2	1251	SP2	E4F1
1252	SP2	EGR1	1253	SP2	ELK4	1254	SP2	EP300
1255	SP2	ESRRA	1256	SP2	ETV4	1257	SP2	FOXH1
1258	SP2	FOXJ1	1259	SP2	FOXN2	1260	SP2	FOXO3
1261	SP2	GABPA	1262	SP2	GTF2A1	1263	SP2	GTF2I
1264	SP2	GZF1	1265	SP2	HBP1	1266	SP2	HES1
1267	SP2	HIF1A	1268	SP2	HINFP	1269	SP2	HOMEZ
1270	SP2	HSF1	1271	SP2	IRF1	1272	SP2	IRF2
1273	SP2	JUN	1274	SP2	JUNB	1275	SP2	JUND
1276	SP2	KLF11	1277	SP2	MAX	1278	SP2	MAZ
1279	SP2	MECP2	1280	SP2	MZF1	1281	SP2	NFATC3
1282	SP2	NFE2L1	1283	SP2	NFE2L2	1284	SP2	NFYA
1285	SP2	NR1H2	1286	SP2	NR2C2	1287	SP2	NR4A1
1288	SP2	NR6A1	1289	SP2	OAZ1	1290	SP2	PARP1
1291	SP2	PITX3	1292	SP2	PKNOX1	1293	SP2	POU2F1
1294	SP2	PPARD	1295	SP2	RBPJ	1296	SP2	REL
1297	SP2	RELB	1298	SP2	RFX1	1299	SP2	RFX2
1300	SP2	RXRB	1301	SP2	SIRT6	1302	SP2	SMAD2
1303	SP2	SP1	1304	SP2	SP2	1305	SP2	SP3
1306	SP2	SP4	1307	SP2	SREBF2	1308	SP2	SRF
1309	SP2	STAT2	1310	SP2	STAT3	1311	SP2	TBP
1312	SP2	TCF12	1313	SP2	TCF3	1314	SP2	TEF
1315	SP2	TGIF1	1316	SP2	TOPORS	1317	SP2	TP53
1318	SP2	TRIM28	1319	SP2	UBP1	1320	SP2	USF1
1321	SP2	YY1	1322	SP2	ZBTB33	1323	SP2	ZBTB7B
1324	SP2	ZFP161	1325	SP2	ZNF143	1326	SP2	ZNF238
1327	SP2	ZNF333	1328	SP2	ZNF628	1329	SP3	ATF1
1330	SP3	ATF2	1331	SP3	ATF4	1332	SP3	ATF5
1333	SP3	ATF7	1334	SP3	BHLHE40	1335	SP3	BRF1
1336	SP3	CEBPB	1337	SP3	CHURC1	1338	SP3	CNOT3
1339	SP3	CREM	1340	SP3	CTCF	1341	SP3	DBP
1342	SP3	DDIT3	1343	SP3	DLX2	1344	SP3	E4F1

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1345	SP3	EGR1	1346	SP3	ELK4	1347	SP3	EP300
1348	SP3	ESRRA	1349	SP3	ETV4	1350	SP3	FOXA3
1351	SP3	FOXH1	1352	SP3	FOXJ1	1353	SP3	FOXJ3
1354	SP3	FOXN2	1355	SP3	FOXO3	1356	SP3	GABPA
1357	SP3	GTF2A1	1358	SP3	GTF2I	1359	SP3	GZF1
1360	SP3	HBP1	1361	SP3	HES1	1362	SP3	HIF1A
1363	SP3	HINFP	1364	SP3	HOMEZ	1365	SP3	HSF1
1366	SP3	HSF2	1367	SP3	IRF1	1368	SP3	IRF2
1369	SP3	JUN	1370	SP3	JUNB	1371	SP3	JUND
1372	SP3	KLF11	1373	SP3	MAX	1374	SP3	MAZ
1375	SP3	MECP2	1376	SP3	MZF1	1377	SP3	NFATC3
1378	SP3	NFE2L1	1379	SP3	NFE2L2	1380	SP3	NFYA
1381	SP3	NR1H2	1382	SP3	NR2C2	1383	SP3	NR3C1
1384	SP3	NR4A1	1385	SP3	NR6A1	1386	SP3	OAZ1
1387	SP3	PARP1	1388	SP3	PITX3	1389	SP3	PKNOX1
1390	SP3	POU2F1	1391	SP3	PPARD	1392	SP3	RBPJ
1393	SP3	REL	1394	SP3	RELA	1395	SP3	RELB
1396	SP3	RFX1	1397	SP3	RFX2	1398	SP3	RXRB
1399	SP3	SIRT6	1400	SP3	SMAD2	1401	SP3	SP1
1402	SP3	SP2	1403	SP3	SP3	1404	SP3	SP4
1405	SP3	SREBF1	1406	SP3	SREBF2	1407	SP3	SRF
1408	SP3	STAT1	1409	SP3	STAT2	1410	SP3	STAT3
1411	SP3	TBP	1412	SP3	TCF12	1413	SP3	TCF3
1414	SP3	TEF	1415	SP3	TERF1	1416	SP3	TGIF1
1417	SP3	TOPORS	1418	SP3	TP53	1419	SP3	TRIM28
1420	SP3	UBP1	1421	SP3	USF1	1422	SP3	YY1
1423	SP3	ZBTB33	1424	SP3	ZBTB7A	1425	SP3	ZBTB7B
1426	SP3	ZFP161	1427	SP3	ZNF143	1428	SP3	ZNF238
1429	SP3	ZNF333	1430	SP3	ZNF628	1431	SP4	ATF1
1432	SP4	ATF2	1433	SP4	ATF4	1434	SP4	ATF5
1435	SP4	ATF7	1436	SP4	BHLHE40	1437	SP4	BRF1
1438	SP4	CEBPB	1439	SP4	CHURC1	1440	SP4	CNOT3
1441	SP4	CREM	1442	SP4	CTCF	1443	SP4	DBP
1444	SP4	DDIT3	1445	SP4	DLX2	1446	SP4	E4F1
1447	SP4	EGR1	1448	SP4	ELK4	1449	SP4	EP300
1450	SP4	ESRRA	1451	SP4	ETV4	1452	SP4	FOXH1
1453	SP4	FOXJ1	1454	SP4	FOXJ3	1455	SP4	FOXN2
1456	SP4	FOXO3	1457	SP4	GABPA	1458	SP4	GTF2A1
1459	SP4	GTF2I	1460	SP4	GZF1	1461	SP4	HBP1
1462	SP4	HES1	1463	SP4	HIF1A	1464	SP4	HINFP
1465	SP4	HOMEZ	1466	SP4	HSF1	1467	SP4	IRF1
1468	SP4	IRF2	1469	SP4	JUN	1470	SP4	JUNB
1471	SP4	JUND	1472	SP4	KLF11	1473	SP4	MAX
1474	SP4	MAZ	1475	SP4	MECP2	1476	SP4	MZF1
1477	SP4	NFATC3	1478	SP4	NFE2L1	1479	SP4	NFE2L2
1480	SP4	NFYA	1481	SP4	NR1H2	1482	SP4	NR2C2
1483	SP4	NR4A1	1484	SP4	NR6A1	1485	SP4	OAZ1
1486	SP4	PARP1	1487	SP4	PITX3	1488	SP4	PKNOX1
1489	SP4	POU2F1	1490	SP4	PPARD	1491	SP4	RBPJ
1492	SP4	REL	1493	SP4	RELA	1494	SP4	RELB

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1495	SP4	RFX1	1496	SP4	RFX2	1497	SP4	RXRB
1498	SP4	SIRT6	1499	SP4	SMAD2	1500	SP4	SP1
1501	SP4	SP2	1502	SP4	SP3	1503	SP4	SP4
1504	SP4	SREBF1	1505	SP4	SREBF2	1506	SP4	SRF
1507	SP4	STAT2	1508	SP4	STAT3	1509	SP4	TBP
1510	SP4	TCF12	1511	SP4	TCF3	1512	SP4	TEF
1513	SP4	TGIF1	1514	SP4	TOPORS	1515	SP4	TP53
1516	SP4	TRIM28	1517	SP4	UBP1	1518	SP4	USF1
1519	SP4	YY1	1520	SP4	ZBTB33	1521	SP4	ZBTB7B
1522	SP4	ZFP161	1523	SP4	ZNF143	1524	SP4	ZNF238
1525	SP4	ZNF263	1526	SP4	ZNF333	1527	SP4	ZNF628
1528	SPI1	MAZ	1529	SPI1	MTF1	1530	SPI1	NR1I3
1531	SPI1	RBPJ	1532	SPI1	RELB	1533	SPI1	RFX5
1534	SPI1	TBP	1535	SPIB	RXRB	1536	SPZ1	DBP
1537	SPZ1	EGR1	1538	SPZ1	ELK4	1539	SPZ1	GABPA
1540	SPZ1	HMBOX1	1541	SPZ1	IRF2	1542	SPZ1	NR4A1
1543	SPZ1	SREBF1	1544	SPZ1	ZBTB7B	1545	SREBF1	IRF2
1546	SREBF1	JUND	1547	SREBF1	MAZ	1548	SREBF1	NFYA
1549	SREBF1	PPARD	1550	SREBF1	RELB	1551	SREBF1	SP2
1552	SREBF1	SREBF2	1553	SREBF1	TOPORS	1554	SREBF1	USF1
1555	SREBF2	IRF2	1556	SREBF2	JUN	1557	SREBF2	JUND
1558	SREBF2	MAZ	1559	SREBF2	NFE2L1	1560	SREBF2	PPARD
1561	SREBF2	SP2	1562	SREBF2	SREBF2	1563	SREBF2	UBP1
1564	SRF	E4F1	1565	SRF	EGR1	1566	SRF	EGR2
1567	SRF	EGR3	1568	SRF	ING4	1569	SRF	JUNB
1570	SRF	NR2F2	1571	SRF	NR4A1	1572	SRF	SRF
1573	STAT1	CHURC1	1574	STAT1	FOXA3	1575	STAT1	IRF9
1576	STAT1	MAFF	1577	STAT1	MAX	1578	STAT1	SRF
1579	STAT2	MAFF	1580	STAT2	MAX	1581	STAT3	CHURC1
1582	STAT3	FOXA3	1583	STAT3	IRF9	1584	STAT3	MAFF
1585	STAT3	MAX	1586	STAT4	IRF9	1587	STAT4	MAFF
1588	STAT4	MAX	1589	STAT5A	FOXA3	1590	STAT5A	IRF9
1591	STAT5A	MAFF	1592	STAT5A	MAX	1593	STAT5B	IRF9
1594	STAT5B	MAFF	1595	STAT5B	MAX	1596	STAT6	MAFF
1597	STAT6	MAX	1598	TAL1	ATF7	1599	TAL1	GTF2A1
1600	TAL1	TOPORS	1601	TCF3	ATF1	1602	TCF3	CEBPE
1603	TCF3	DLX2	1604	TCF3	HES1	1605	TCF3	HMBOX1
1606	TCF3	ING4	1607	TCF3	SRF	1608	TCF3	TEF
1609	TCF3	TERF1	1610	TCF3	ZBTB7A	1611	TFAP2A	ATF1
1612	TFAP2A	ATF5	1613	TFAP2A	ATF7	1614	TFAP2A	BCL6
1615	TFAP2A	BHLHE40	1616	TFAP2A	CTCF	1617	TFAP2A	DEAF1
1618	TFAP2A	E2F7	1619	TFAP2A	E4F1	1620	TFAP2A	EP300
1621	TFAP2A	ESRRA	1622	TFAP2A	ETV4	1623	TFAP2A	FOXO3
1624	TFAP2A	GZF1	1625	TFAP2A	HES1	1626	TFAP2A	HMBOX1
1627	TFAP2A	HSF1	1628	TFAP2A	IRF2	1629	TFAP2A	JUN
1630	TFAP2A	JUNB	1631	TFAP2A	KLF11	1632	TFAP2A	KLF15
1633	TFAP2A	MAFA	1634	TFAP2A	MAFF	1635	TFAP2A	MAZ
1636	TFAP2A	NFE2L2	1637	TFAP2A	OAZ1	1638	TFAP2A	PARP1
1639	TFAP2A	PATZ1	1640	TFAP2A	POU2F1	1641	TFAP2A	PURA
1642	TFAP2A	RBPJ	1643	TFAP2A	REL	1644	TFAP2A	RELB

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1645	TFAP2A	SIRT6	1646	TFAP2A	SP1	1647	TFAP2A	SP4
1648	TFAP2A	SREBF2	1649	TFAP2A	SRF	1650	TFAP2A	STAT3
1651	TFAP2A	TEF	1652	TFAP2A	TFAP4	1653	TFAP2A	TGIF1
1654	TFAP2A	TOPORS	1655	TFAP2A	TP53	1656	TFAP2A	TRIM28
1657	TFAP2A	USF1	1658	TFAP2A	ZBTB7B	1659	TFAP2A	ZFP161
1660	TFAP2A	ZNF143	1661	TFAP2A	ZNF238	1662	TFAP2B	ATF1
1663	TFAP2B	BCL6	1664	TFAP2B	BHLHE40	1665	TFAP2B	CTCF
1666	TFAP2B	DEAF1	1667	TFAP2B	E2F7	1668	TFAP2B	E4F1
1669	TFAP2B	EP300	1670	TFAP2B	ESRRA	1671	TFAP2B	ETV4
1672	TFAP2B	FOXO3	1673	TFAP2B	GZF1	1674	TFAP2B	HES1
1675	TFAP2B	HMBOX1	1676	TFAP2B	HSF1	1677	TFAP2B	IRF2
1678	TFAP2B	JUN	1679	TFAP2B	JUNB	1680	TFAP2B	KLF11
1681	TFAP2B	KLF15	1682	TFAP2B	MAFF	1683	TFAP2B	MAZ
1684	TFAP2B	NFE2L2	1685	TFAP2B	OAZ1	1686	TFAP2B	PARP1
1687	TFAP2B	PATZ1	1688	TFAP2B	POU2F1	1689	TFAP2B	PURA
1690	TFAP2B	RBPJ	1691	TFAP2B	REL	1692	TFAP2B	RELB
1693	TFAP2B	SIRT6	1694	TFAP2B	SP4	1695	TFAP2B	SREBF2
1696	TFAP2B	SRF	1697	TFAP2B	STAT3	1698	TFAP2B	TFAP4
1699	TFAP2B	TGIF1	1700	TFAP2B	TOPORS	1701	TFAP2B	TP53
1702	TFAP2B	TRIM28	1703	TFAP2B	USF1	1704	TFAP2B	ZBTB7B
1705	TFAP2B	ZFP161	1706	TFAP2B	ZNF143	1707	TFAP2B	ZNF238
1708	TFAP2C	ATF1	1709	TFAP2C	ATF7	1710	TFAP2C	BCL6
1711	TFAP2C	BHLHE40	1712	TFAP2C	CTCF	1713	TFAP2C	DEAF1
1714	TFAP2C	E2F7	1715	TFAP2C	E4F1	1716	TFAP2C	EP300
1717	TFAP2C	ESRRA	1718	TFAP2C	ETV4	1719	TFAP2C	FOXO3
1720	TFAP2C	GZF1	1721	TFAP2C	HES1	1722	TFAP2C	HMBOX1
1723	TFAP2C	HSF1	1724	TFAP2C	IRF2	1725	TFAP2C	JUN
1726	TFAP2C	JUNB	1727	TFAP2C	KLF11	1728	TFAP2C	KLF15
1729	TFAP2C	MAFF	1730	TFAP2C	MAZ	1731	TFAP2C	NFE2L2
1732	TFAP2C	OAZ1	1733	TFAP2C	PARP1	1734	TFAP2C	PATZ1
1735	TFAP2C	POU2F1	1736	TFAP2C	PURA	1737	TFAP2C	RBPJ
1738	TFAP2C	REL	1739	TFAP2C	RELB	1740	TFAP2C	SIRT6
1741	TFAP2C	SP4	1742	TFAP2C	SREBF2	1743	TFAP2C	SRF
1744	TFAP2C	STAT3	1745	TFAP2C	TEF	1746	TFAP2C	TFAP4
1747	TFAP2C	TGIF1	1748	TFAP2C	TOPORS	1749	TFAP2C	TP53
1750	TFAP2C	TRIM28	1751	TFAP2C	USF1	1752	TFAP2C	ZBTB7B
1753	TFAP2C	ZFP161	1754	TFAP2C	ZNF143	1755	TFAP2C	ZNF238
1756	TFAP4	CDC5L	1757	TFAP4	E2F7	1758	TFAP4	TP53
1759	TFCP2	CTCF	1760	TFCP2	DBP	1761	TFCP2	GTF2I
1762	TFCP2	MAZ	1763	TFCP2L1	DBP	1764	TFCP2L1	TP53
1765	TFDP1	ATF4	1766	TFDP1	E2F1	1767	TFDP1	MAZ
1768	TFDP2	ATF4	1769	TFDP2	E2F1	1770	TFDP2	MAZ
1771	THRA	ZNF143	1772	THRB	ZNF143	1773	TLX2	CREM
1774	TP53	GABPB1	1775	TP53	HOMEZ	1776	TP53	RBPJ
1777	TP53	RELB	1778	TP63	RELB	1779	TP73	RELB
1780	TRIM28	CBFB	1781	TRIM28	ELK4	1782	TRIM28	ESRRA
1783	TRIM28	FOXO3	1784	TRIM28	GTF2A1	1785	TRIM28	GZF1
1786	TRIM28	MAFA	1787	TRIM28	PITX3	1788	TRIM28	SMAD2
1789	TRIM28	SREBF2	1790	TRIM28	SRF	1791	TRIM28	TGIF1
1792	TRIM28	TP53	1793	TRIM28	YY1	1794	USF1	DBP

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1795	USF1	GTF2A1	1796	USF1	HINFP	1797	USF1	IRF9
1798	USF1	MAFG	1799	USF1	PITX3	1800	USF1	TOPORS
1801	USF2	DBP	1802	USF2	GTF2A1	1803	USF2	HINFP
1804	USF2	IRF9	1805	USF2	MAFG	1806	USF2	PITX3
1807	VDR	BHLHE40	1808	VDR	CNOT3	1809	VDR	HBP1
1810	VDR	MECP2	1811	VDR	NR6A1	1812	VDR	RFX5
1813	VDR	SP1	1814	VDR	SP3	1815	VDR	SP4
1816	VDR	TBP	1817	WT1	BHLHE40	1818	WT1	DBP
1819	WT1	DDIT3	1820	WT1	ELK4	1821	WT1	EP300
1822	WT1	FOXO3	1823	WT1	GZF1	1824	WT1	JUN
1825	WT1	MAZ	1826	WT1	MECP2	1827	WT1	NFATC3
1828	WT1	PATZ1	1829	WT1	POU2F1	1830	WT1	SP3
1831	WT1	STAT3	1832	WT1	TEF	1833	WT1	TP53
1834	WT1	ZBTB7B	1835	WT1	ZFP161	1836	XBP1	RXRB
1837	YY1	ARNT	1838	YY1	DEAF1	1839	YY1	E4F1
1840	YY1	EGR2	1841	YY1	EP300	1842	YY1	ING4
1843	YY1	NFYA	1844	YY1	NR1I3	1845	YY1	RXRB
1846	YY1	USF1	1847	ZBTB6	NFYA	1848	ZBTB7B	ATF1
1849	ZBTB7B	ATF2	1850	ZBTB7B	ATF5	1851	ZBTB7B	BHLHE40
1852	ZBTB7B	BRF1	1853	ZBTB7B	CEBPB	1854	ZBTB7B	CREM
1855	ZBTB7B	DBP	1856	ZBTB7B	DDIT3	1857	ZBTB7B	DLX2
1858	ZBTB7B	E4F1	1859	ZBTB7B	ELK4	1860	ZBTB7B	EP300
1861	ZBTB7B	ESRRA	1862	ZBTB7B	FOXN2	1863	ZBTB7B	FOXO3
1864	ZBTB7B	FOXP3	1865	ZBTB7B	GTF2I	1866	ZBTB7B	GZF1
1867	ZBTB7B	HBP1	1868	ZBTB7B	HES1	1869	ZBTB7B	IRF1
1870	ZBTB7B	JUN	1871	ZBTB7B	JUND	1872	ZBTB7B	KLF11
1873	ZBTB7B	MAFF	1874	ZBTB7B	MAX	1875	ZBTB7B	MAZ
1876	ZBTB7B	MECP2	1877	ZBTB7B	MZF1	1878	ZBTB7B	NFATC3
1879	ZBTB7B	NFE2L1	1880	ZBTB7B	NFYA	1881	ZBTB7B	NR4A1
1882	ZBTB7B	NR6A1	1883	ZBTB7B	OAZ1	1884	ZBTB7B	PITX3
1885	ZBTB7B	PKNOX1	1886	ZBTB7B	POU2F1	1887	ZBTB7B	RBPJ
1888	ZBTB7B	RELB	1889	ZBTB7B	RFX1	1890	ZBTB7B	RFX5
1891	ZBTB7B	SIRT6	1892	ZBTB7B	SP1	1893	ZBTB7B	SP2
1894	ZBTB7B	SP3	1895	ZBTB7B	SP4	1896	ZBTB7B	SRF
1897	ZBTB7B	STAT3	1898	ZBTB7B	TBP	1899	ZBTB7B	TCF12
1900	ZBTB7B	TCF3	1901	ZBTB7B	TEF	1902	ZBTB7B	TOPORS
1903	ZBTB7B	TP53	1904	ZBTB7B	TRIM28	1905	ZBTB7B	YY1
1906	ZBTB7B	ZBTB7B	1907	ZBTB7B	ZNF143	1908	ZBTB7B	ZNF238
1909	ZBTB7B	ZNF628	1910	ZEB1	TERF1	1911	ZFP161	DDIT3
1912	ZFP161	ELK4	1913	ZFP161	MAZ	1914	ZFP161	SP3
1915	ZFP42	ARNT	1916	ZFP42	E4F1	1917	ZFP42	EP300
1918	ZFP42	ING4	1919	ZFP42	MAZ	1920	ZFP42	NFYA
1921	ZFX	BRF1	1922	ZFX	CREM	1923	ZFX	CTCF
1924	ZFX	DBP	1925	ZFX	DEAF1	1926	ZFX	E2F7
1927	ZFX	E4F1	1928	ZFX	EGR1	1929	ZFX	ELK4
1930	ZFX	EP300	1931	ZFX	ESRRA	1932	ZFX	FOXO3
1933	ZFX	GTF2I	1934	ZFX	GZF1	1935	ZFX	HES1
1936	ZFX	HIF1A	1937	ZFX	HSF1	1938	ZFX	IRF2
1939	ZFX	JUNB	1940	ZFX	JUND	1941	ZFX	KLF15
1942	ZFX	MAFF	1943	ZFX	NFATC3	1944	ZFX	NFE2L1

Continued on next page

Table A.5 – *Continued from previous page*

N	Source	Target	N	Source	Target	N	Source	Target
1945	ZFX	NFE2L2	1946	ZFX	NFYA	1947	ZFX	NR2C2
1948	ZFX	PATZ1	1949	ZFX	PKNOX1	1950	ZFX	POU2F1
1951	ZFX	RBPJ	1952	ZFX	RELB	1953	ZFX	RFX1
1954	ZFX	SMAD2	1955	ZFX	SMAD7	1956	ZFX	SP1
1957	ZFX	SP4	1958	ZFX	SREBF2	1959	ZFX	SRF
1960	ZFX	TCF3	1961	ZFX	TFAP4	1962	ZFX	TOPORS
1963	ZFX	TRIM28	1964	ZFX	YY1	1965	ZFX	ZBTB7A
1966	ZFX	ZNF143	1967	ZFX	ZNF263	1968	ZNF143	ATF7
1969	ZNF143	FOXA3	1970	ZNF143	GABPA	1971	ZNF143	GTF2I
1972	ZNF143	MAX	1973	ZNF143	MZF1	1974	ZNF143	NFE2L1
1975	ZNF143	NR6A1	1976	ZNF143	TP73	1977	ZNF143	ZBTB7A
1978	ZNF143	ZNF143	1979	ZNF143	ZNF219	1980	ZNF143	ZNF263
1981	ZNF143	ZNF628	1982	ZNF148	BHLHE40	1983	ZNF148	CNOT3
1984	ZNF148	CTCF	1985	ZNF148	DDIT3	1986	ZNF148	ELK4
1987	ZNF148	EP300	1988	ZNF148	ESRRA	1989	ZNF148	HES1
1990	ZNF148	HOMEZ	1991	ZNF148	JUN	1992	ZNF148	MAZ
1993	ZNF148	MECP2	1994	ZNF148	NFATC3	1995	ZNF148	POU2F1
1996	ZNF148	SP1	1997	ZNF148	SP3	1998	ZNF148	STAT3
1999	ZNF148	TEF	2000	ZNF148	ZBTB7B	2001	ZNF148	ZNF238
2002	ZNF219	BHLHE40	2003	ZNF219	BRF1	2004	ZNF219	CNOT3
2005	ZNF219	CTCF	2006	ZNF219	DBP	2007	ZNF219	DDIT3
2008	ZNF219	DLX2	2009	ZNF219	EP300	2010	ZNF219	ETV4
2011	ZNF219	FOXN2	2012	ZNF219	FOXO3	2013	ZNF219	GTF2I
2014	ZNF219	IRF2	2015	ZNF219	JUN	2016	ZNF219	JUNB
2017	ZNF219	JUND	2018	ZNF219	NFATC3	2019	ZNF219	NFYA
2020	ZNF219	NR4A1	2021	ZNF219	NR6A1	2022	ZNF219	POU2F1
2023	ZNF219	RBPJ	2024	ZNF219	SP4	2025	ZNF219	SRF
2026	ZNF219	TP53	2027	ZNF219	ZBTB7B	2028	ZNF219	ZNF238
2029	ZNF263	BHLHE40	2030	ZNF263	CNOT3	2031	ZNF263	EGR1
2032	ZNF263	MAZ	2033	ZNF263	MECP2	2034	ZNF263	NFE2L2
2035	ZNF263	SP1	2036	ZNF263	SP4	2037	ZNF263	STAT3
2038	ZNF263	ZNF238	2039	ZNF350	FOXM1	2040	ZNF350	TBP
2041	ZNF589	MECP2						

Table A.6. 23 protein complexes in which the proteins in the complex are highly connected with HK interactions. Rows without background are TFs in one complex, while rows with gray background are HK interactions connecting TFs in the complex.

Complex ID	TFs or interactions lists	N
HC5737	EP300;ETS1;FOXC1;GATA2;GATA3;GATA5;GATA6;KLF4;MZF1;SP1;SREBF1;SREBF2;TFAP2A;YY1;ZNF354C	15
	KLF4-EP300;SP1-EP300;TFAP2A-EP300;YY1-EP300;ETS1-MZF1;KLF4-MZF1; SP1-MZF1;KLF4-SP1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;SP1-SREBF1;KLF4-SREBF2; SP1-SREBF2;SREBF1-SREBF2;SREBF2-SREBF2;TFAP2A-SREBF2;KLF4-YY1;SP1-YY1	19
HC6033	ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;NKX2-5;PAX2;SP1;SPIB;TFAP2A;USF1;YY1	13
	ETS1-MZF1;KLF4-MZF1;SP1-MZF1;KLF4-SP1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;KLF4-USF1; SP1-USF1;TFAP2A-USF1;YY1-USF1;KLF4-YY1;SP1-YY1	13
HC3896	BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;NFIC;SP1;SPIB;TFAP2A;USF1;YY1;ZNF354C	14
	ETS1-MZF1;KLF4-MZF1;SP1-MZF1;KLF4-SP1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;KLF4-USF1; SP1-USF1;TFAP2A-USF1;YY1-USF1;KLF4-YY1;SP1-YY1	13
HC6644	BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;NFIC;NKX2-5;SP1;SPIB;TFAP2A;USF1;YY1;ZNF354C	15
	ETS1-MZF1;KLF4-MZF1;SP1-MZF1;KLF4-SP1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;KLF4-USF1; SP1-USF1;TFAP2A-USF1;YY1-USF1;KLF4-YY1;SP1-YY1	13
HC4454	ETS1;FOXC1;FOXL1;GATA2;GATA3;HOXA5;HSF1;MZF1;NFIC;SP1;TFAP2A;USF1;YY1;ZNF354C	14
	SP1-HSF1;TFAP2A-HSF1;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;SP1-USF1; TFAP2A-USF1;YY1-USF1;SP1-YY1	11
HC8755	CDC5L;ETS1;FOXC1;GATA2;GATA3;MZF1;NFIC;NKX2-5;REL;SP1;ZEB1	11
	ETS1-CDC5L;ETS1-MZF1;SP1-MZF1;SP1-REL;MZF1-SP1;SP1-SP1	6
HC6745	BRCA1;ETS1;FOXC1;FOXL1;GATA2;KLF4;MZF1;NFIC;NR4A2;SP1;SPIB;TBP;TFAP2A;YY1;ZEB1	15
	ETS1-MZF1;KLF4-MZF1;SP1-MZF1;KLF4-SP1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;KLF4-TBP; SP1-TBP;KLF4-YY1;SP1-YY1	11
HC4615	ETS1;FOXC1;FOXL1;GATA2;MZF1;NFIC;NKX3-2;REL;SOX10;SP1;TFAP2A;YY1;ZNF354C	13
	ETS1-MZF1;SP1-MZF1;SP1-REL;TFAP2A-REL;MZF1-SP1;SP1-SP1;TFAP2A-SP1;SP1-YY1	8
HC4912	ARNT;ELF5;ETS1;FOXC1;GATA2;GATA3;MZF1;NFIC;SOX10;SP1;SPIB;TFAP2A;YY1	13
	YY1-ARNT;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;SP1-YY1	7
HC6667	BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;MZF1;NFIC;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C	15
	ELK1-MZF1;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;ZFX-SP1;SP1-YY1;ZFX-YY1	9
HC9842	BRCA1;ELK1;ETS1;FOXC1;FOXL1;GATA2;MZF1;NFIC;SOX10;SP1;SPIB;TFAP2A;YY1;ZFX;ZNF354C	15
	ELK1-MZF1;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;ZFX-SP1;SP1-YY1;ZFX-YY1	9
HC2178	BRCA1;ETS1;FOXC1;GATA2;MZF1;NFIC;NKX2-5;REL;SOX10;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C	15
	ETS1-MZF1;SP1-MZF1;SP1-REL;TFAP2A-REL;MZF1-SP1;SP1-SP1;TFAP2A-SP1;SP1-YY1	8
HC9330	BRCA1;ETS1;FOXC1;FOXL1;GATA2;GATA3;NFYA;NKX2-5;SOX10;SOX5;SP1;SPIB;TFAP2A;YY1;ZNF354C	15
	ETS1-NFYA;NFYA-NFYA;SP1-NFYA;YY1-NFYA;SP1-SP1;TFAP2A-SP1;NFYA-YY1;SP1-YY1	8
HC4460	ARID3A;ELF5;ETS1;FOXL1;GATA2;GATA3;HSF1;NFIC;NFIL3;PARP1;SP1;SPIB;TFAP2A;YY1;ZNF354C	15
	SP1-HSF1;TFAP2A-HSF1;SP1-PARP1;TFAP2A-PARP1;SP1-SP1;TFAP2A-SP1;SP1-YY1	7
HC6205	ELF5;ETS1;FOXL1;GATA2;GATA3;HSF1;NFIC;NFIL3;PARP1;SP1;SPIB;TFAP2A;YY1;ZEB1;ZNF354C	15
	SP1-HSF1;TFAP2A-HSF1;SP1-PARP1;TFAP2A-PARP1;SP1-SP1;TFAP2A-SP1;SP1-YY1	7
HC9314	ARNT;ATF7;BRCA1;ESRRB;ETS1;FOS;FOXC1;GATA2;GATA3;MZF1;NFIC;PAX2;SP1;YY1;ZEB1	15
	YY1-ARNT;SP1-ATF7;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;SP1-YY1	7
HC7980	ELK1;ETS1;FOXC1;GATA2;GATA3;HOXA5;MAFB;NFIC;SOX10;SP1;SP4;YY1;ZEB1;ZNF354C	14
	SP1-SP1;SP4-SP1;SP1-SP4;SP4-SP4;SP1-YY1;SP4-YY1	6
HC1277	CREB1;ETS1;FOXC1;FOXO3;GATA2;GATA3;MZF1;NFIC;NFYA;PDX1;REL;SOX10;SPIB;TFAP2A; YY1;ZEB1;ZNF354C	17
	TFAP2A-FOXO3;ETS1-MZF1;ETS1-NFYA;NFYA-NFYA;YY1-NFYA;CREB1-REL;TFAP2A-REL;NFYA-YY1	8
HC7936	ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;KLF4;NFIC;NFYA;NKX2-5;PDX1;SOX10;TFAP2A;YY1;ZEB1	15
	ETS1-NFYA;KLF4-NFYA;NFYA-NFYA;YY1-NFYA;KLF4-YY1;NFYA-YY1	6
HC4463	ELF5;ETS1;FOXC1;FOXL1;GATA2;GATA3;HSF1;KLF4;PDX1;SOX10;TBP;TFAP2A;YY1;ZEB1;ZFX	15
	KLF4-HSF1;TFAP2A-HSF1;ZFX-HSF1;KLF4-TBP;KLF4-YY1;ZFX-YY1	6
HC6575	ARID3A;BRCA1;ELF5;ETS1;FOXC1;GATA2;GATA3;MYB;MZF1;NKX2-5;PARP1;SOX10;SP1;YY1;ZNF354C	15
	ETS1-MZF1;SP1-MZF1;SP1-PARP1;MZF1-SP1;SP1-SP1;SP1-YY1	6
HC2683	ETS1;FOXC1;FOXO3;GATA2;GATA3;HOXA5;MAFB;MZF1;NFIC;NKX2-5; PAX2;PDX1;SOX10;SP1;TBP;TFAP2A;YY1;ZEB1;ZNF354C	19
	SP1-FOXO3;TFAP2A-FOXO3;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;TFAP2A-SP1;SP1-TBP;SP1-YY1	9
HC6143	ARID3A;BRCA1;CDC5L;CREB1;ELK1;ETS1;FOXC1;FOXL1;GATA2;GATA3;MZF1;NFIC;PDX1; PRRX2;SOX10;SP1;YY1;ZEB1	18
	ETS1-CDC5L;ELK1-MZF1;ETS1-MZF1;SP1-MZF1;MZF1-SP1;SP1-SP1;SP1-YY1	7