Comparing transcription factor regulatory networks of human cell types

The Protein Network Workshop
June 8–12, 2015

KWOK-PUI CHOI

Dept of Statistics & Applied Probability,
Dept of Mathematics,
NUS

Outline

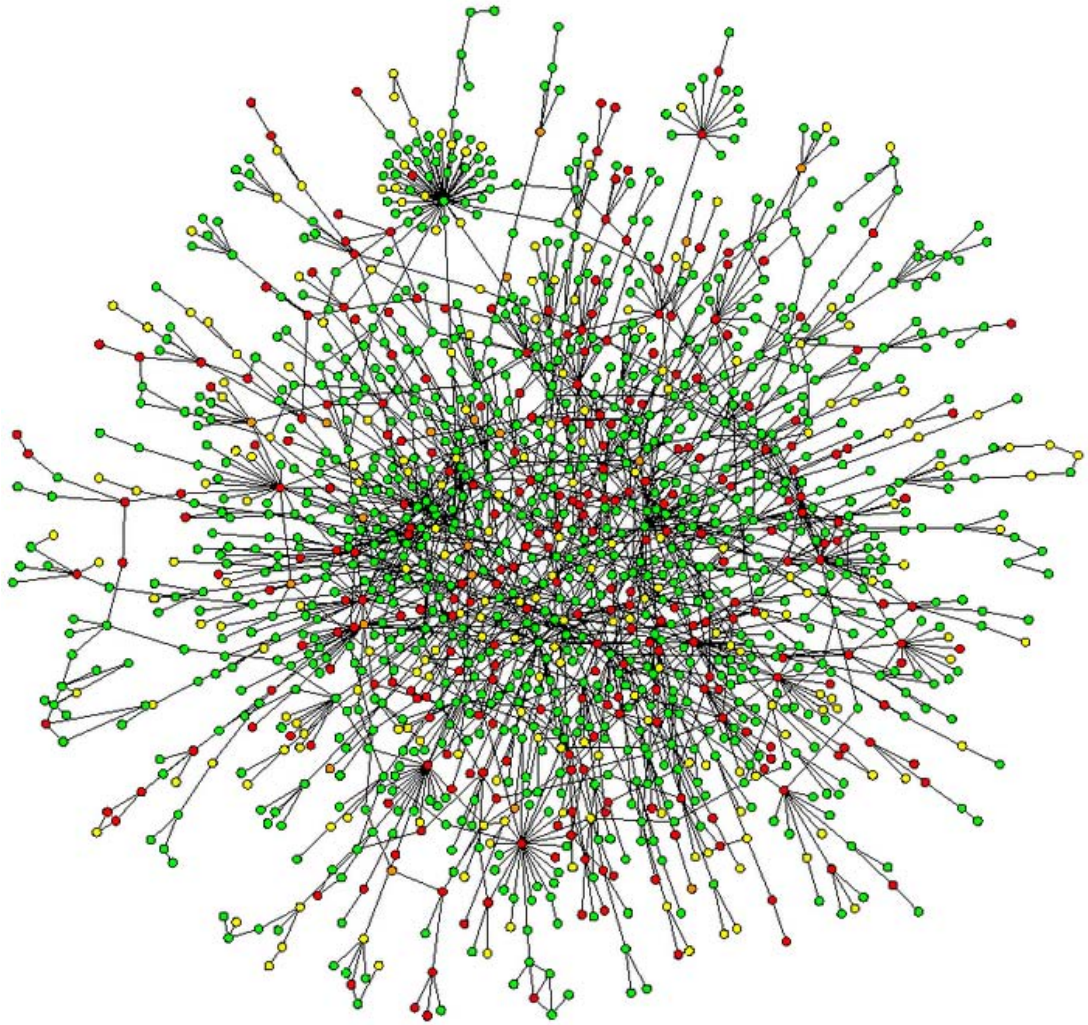
Part I: Summary statistics for networks

1. Introduction
2. Wiener-type indices: To normalize or not to normalize
3. Simulation studies

Part II: Comparing transcriptional factor regulatory networks of human cell types

1. Introduction
2. Global/local features across networks
3. Work in progress

1. Introduction



Yeast protein-protein interaction

http://compbio.pbworks.com/w/page/16252899/Mass-Spectrophotometry-and-Protein-Interaction-Networks

1. Introduction

- Terminologies
  - Nodes: biological units
  - Edges/links/interactions: interactions between 2 units
  - Networks: directed/undirected
- Network structure enhances understanding about the structure–function of the units in the system
  - Ma and Gao (2012). Biological network analysis: insights into structure and functions. *Breifings in Functional Genomics*, **2**.
  - Ali et al. (2014). Alignment-free protein network comparison. *Bioinformatics*, **30**

    Introduced a network topology based measure, Netdis; Topology of PPI networks contain information about evolutionary process.

1. Introduction

- Scale-free, small world (small graph diameter)
- Pržulj and her coworkers introduced Graphlet Degree Distribution Agreement (GDDA) to measure the extent of agreement between two networks. *Bioinformatics* **23**

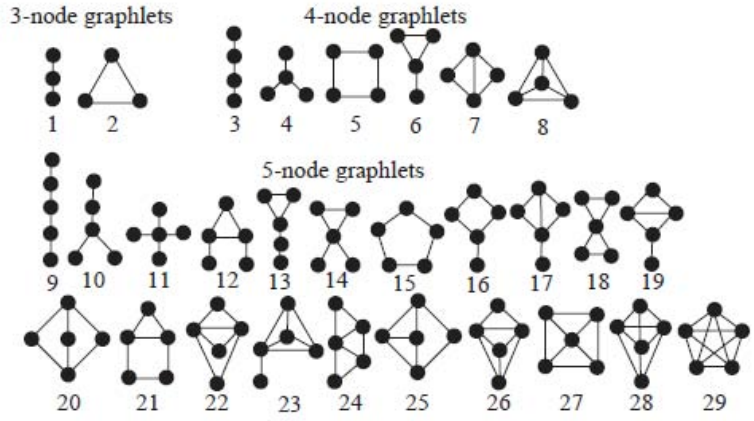


Fig. 1. All 3-node, 4-node and 5-node connected networks (graphlets), ordered within groups from the least to the most dense with respect to the number of edges when compared to the maximum possible number of edges in the graphlet; they are numbered from 1 to 29.

- Rito et al. (2010) pointed out GDDA score depends on the number of edges and nodes in the network *Bioinformatics* **26**

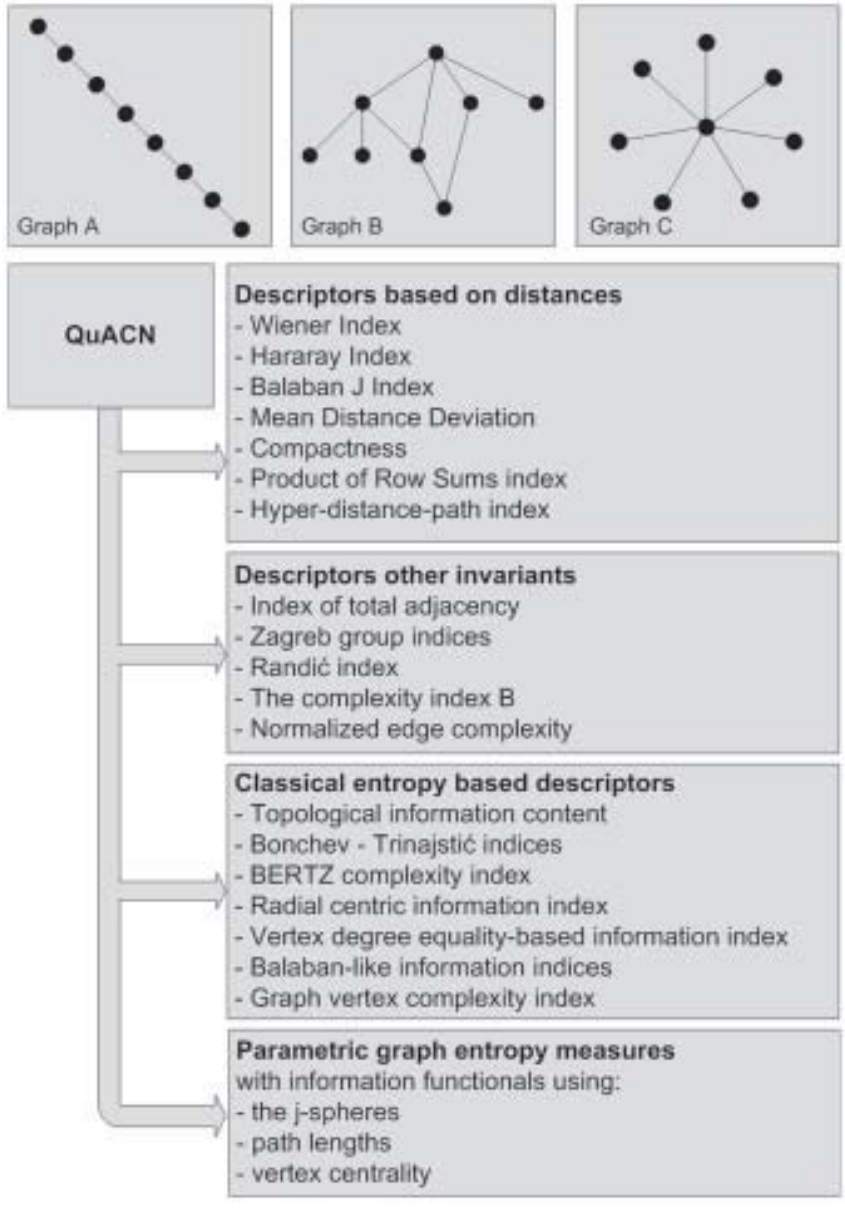
1. Introduction

- Mueller et al. (2011). QuACN: An R Package for Analyzing Complex Biological networks quantitatively. *Bioinformatics* **27**
- QuACN stands for (**Qu**antitative **A**nalysis of **C**omplex **N**etwork)
- In chemometrics, topological features are used to characterize chemical compounds for identifying potential drug targets.
- Todeschini et al (2002) lists a large number of descriptors for possible use to analyze molecular networks.

  > $\cdots$ *a mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number* $\cdots$

- Extended to provide summary statistics of a complex network.

1. Introduction



Graphical overview of QuACN (Fig 1 in Mueller et al. (2011) *Bioinformatics*)

2. Wiener-type indices: To normalize or not to normalize

- Harry Wiener (1947) introduced an index to predict physical properties of isomers:

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

- Many other indices are introduced. To name a few related to this presentation:
  - Harary index: $H(G) = \sum_{1 \le i < j \le n} \frac{1}{d(i,j)}$;
  - Hyper-Wiener index:

$$WW(G) = \frac{1}{2} \sum_{1 \le i < j \le n} d(i,j)^2 + \frac{1}{2} \sum_{1 \le i < j \le n} d(i,j);$$

  - Generalized Wiener index, Q–index (Hosoya polynomial or Wiener polynomial

$$\sum_{1 \le i < j \le n} \lambda^{d(i,j)}.$$

2. Wiener-type indices: To normalize or not to normalize

- All these Wiener type indices depend on the number of nodes besides the number of edges and shape
- Example:
  - P_4 (P_5) a straight line graph of 4 nodes (5 nodes);
  - K_4 a complete graph of 4 nodes; and
  - S_5 a star with 5 nodes
  - Values of the Wiener index are:

$$W(K_4) = 6, \quad W(P_4) = 10, \quad W(S_5) = 16, \quad W(P_5) = 20.$$

- Given the number of nodes, if we know the maximum of $W(G)$ and the minimum of $W(G)$, then we can introduce a normalized Wiener index:

$$W^*(G) = \frac{\max - W(G)}{\max - \min}.$$

  So $W^*(G) \in [0, 1]$.
- Better still: if we can identify for what networks the maximum and the minimum are attained.

- Tian + C introduced f-Wiener index of a network G:

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

where f is a monotone function.

- They proved that for f increasing

$$\begin{aligned} \max &= W_f(P_n) = \sum_{k=1}^{n-1} (n-k) f(k); \\ \min &= W_f(K_n) = \frac{n(n-1)}{2} f(1). \end{aligned}$$

- For f strictly increasing, maximum is attained if and only if G is a path of n nodes.
- For f strictly increasing, minimum is attained if and only if G is a complete graph of n nodes.
- Similar results hold for f decreasing.
- Extends these results to G being a tree, to G being a network (or tree) with given maximum degree m.

3. A SIMULATION STUDY

(1) Hierarchical clustering of random networks

(2) Random networks used:

10 Erdös-Renyi random graphs $ER(n, 0.05)$:
n: number of nodes are $500, 550, \ldots, 950$

10 Scale-free networks with the same numbers of nodes as above

3-dim Geometric networks with the same numbers of nodes as above

(3) Functions f used

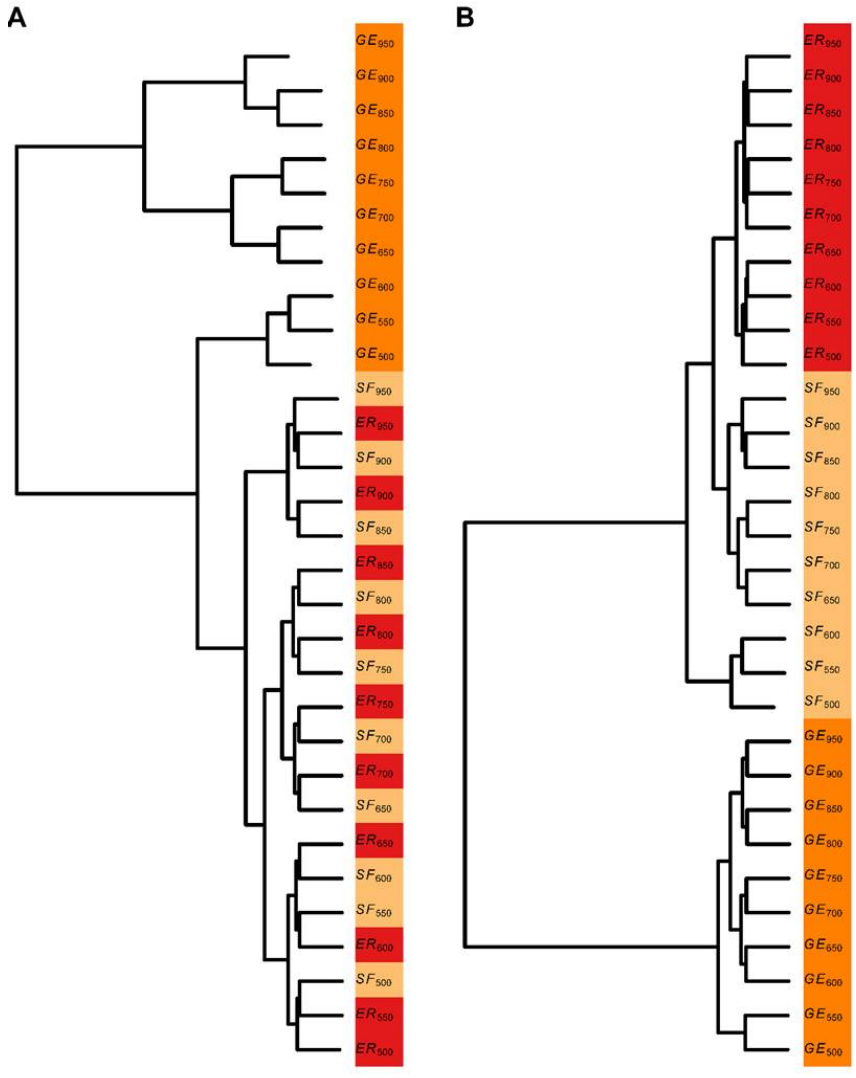
$$f_1(k) = \sqrt{k}, \quad f_2(k) = k, \quad f_3(k) = (k^2 + k)/2, \quad f_4(k) = \frac{4k}{n(n-1)}$$

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

(4) Each network is summarized by 7 statistics: $v_G = (W_{f_1}(G), \ldots, W_{f_7}(G))$
Cluster G's in (2) using v_G's.

(5) Each network is also summarized by 7 statistics:
$G \Rightarrow v_G^* = (W_{f_1}^*(G), \ldots, W_{f_7}^*(G))$
Cluster G's in (2) using v_G^*'s.

3. A SIMULATION STUDY

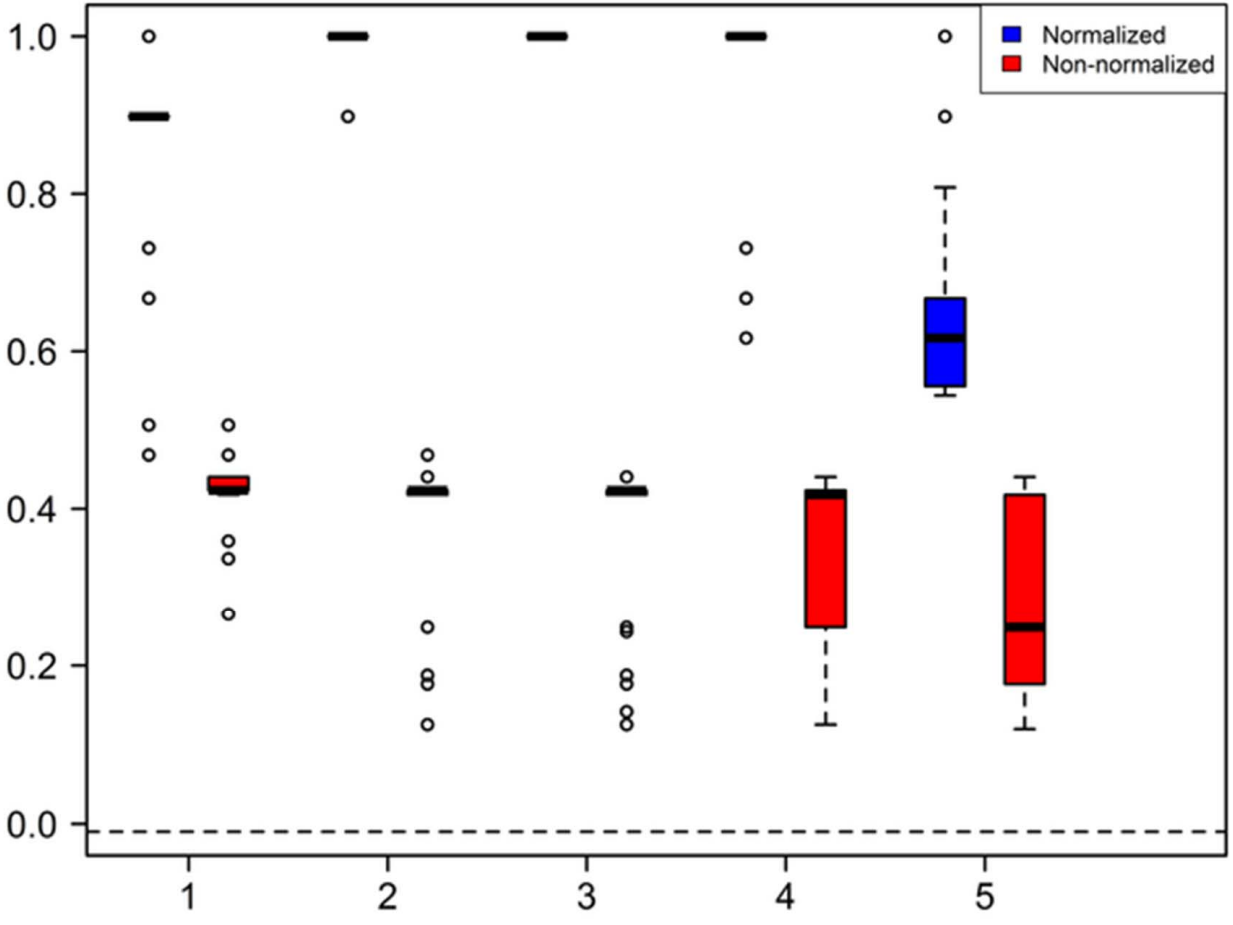


Adjusted Rand Index using non-normalized indices = 0.24

Adjusted Rand Index using normalized indices = 0.67

3. A SIMULATION STUDY

Boxplots of adjusted rand index for measuring the extent of agreement of clustering of the random networks using non-normalized



Tian D, Choi KP (2013) Sharp Bounds and Normalization of Wiener-Type Indices. PLoS ONE 8(11): e78448. doi:10.1371/journal.pone.0078448
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0078448

Summary for Part I

- Measures for graphical structures useful for biological network comparison and analysis
- Effect of number of nodes in summary statistics needs to be accounted for.
- Future work: How about other indices used in QuACN: Need normalization? And how?

Part II COMPARING 41 TF REGULATORY NETWORKS OF HUMAN CELL TYPES

1. Introduction

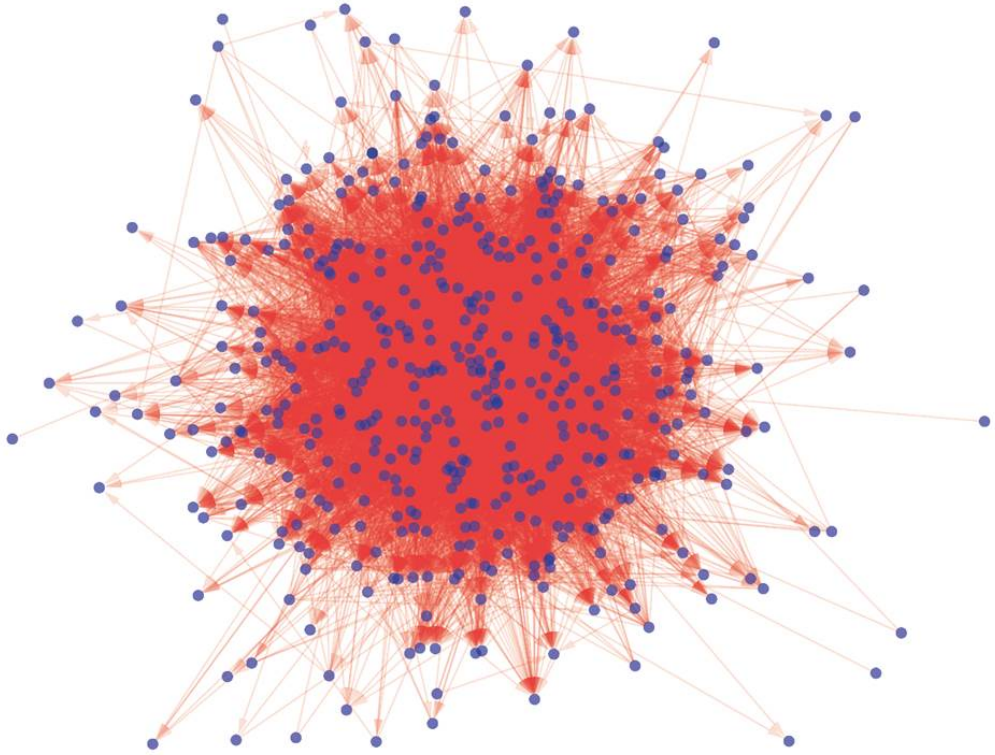
Dataset

- Taken from Neph at al. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**(6), p1274–1286.
- TF regulatory networks of 41 human cell types
- Constructed using the DNaseI footprinting technology

Cell type	# of networks
Blood	7
Cancer	2
Endothelia	4
Epithelia	6
Embryonic SC	1
Fetal	3
Stroma	14
Viscera	4

- About 475 TFs (altogether there are 538 TFs)
- About 11,200 interactions
- Graph density: (# of interactions)$/[n(n-1)] \approx 5\%$

1. INTRODUCTION



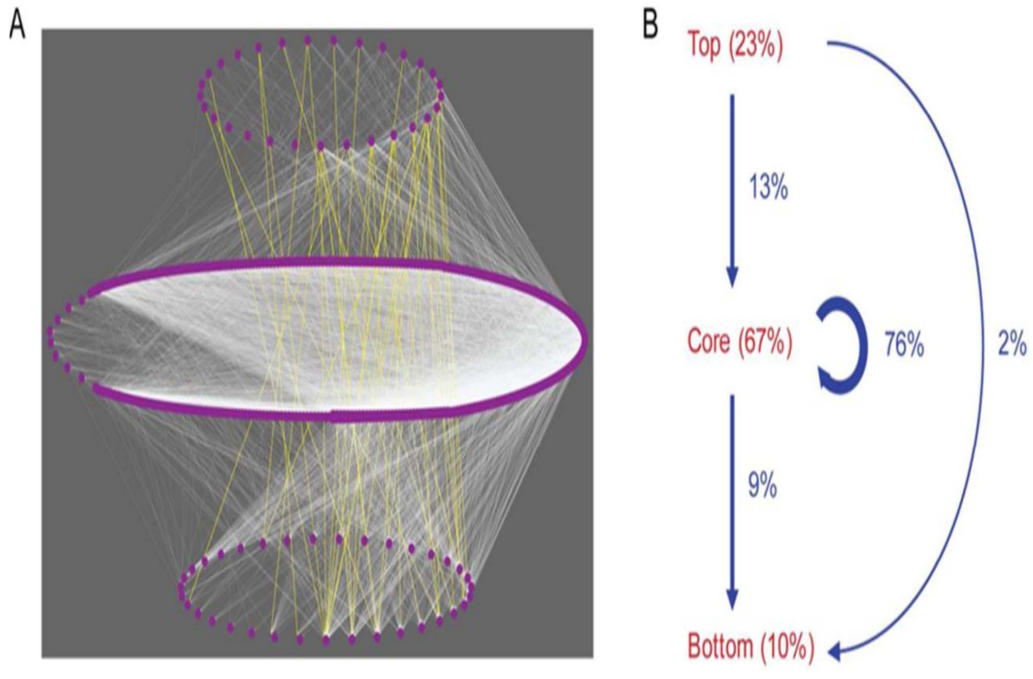
ESC TF regulatory network

2. GLOBAL/LOCAL FEATURES ACROSS NETWORKS

2.1 HIERARCHICAL STRUCTURES OF THE REGULATORY NETWORKS

- Using Vertex Sort Algorithm (Jothi et al. Mol. Syst. Biol. (2009)), all 41 networks share a 3-layer structure on each network.

A schematic view of the three-layer hierarchical structure of the hESC TF regulatory network.



Shihua Zhang et al. Nucl. Acids Res. 2014;42:12380-12387

Nucleic Acids Research

2.1 HIERARCHICAL STRUCTURES OF THE REGULATORY NETWORKS

- hESC TF regulatory network stands out:

  (1) In terms of TF distribution in each layer as compared against all networks

Layer	% TF	SD from mean (rank)
Top	6.2%	3 SD below 23% (lowest)
Core	85.3%	3.1 SD above 67% (highest)
Bottom	8.5%	1 SD below 10% (5th lowest)

  (2) In terms of interactions from one layer to another

Layer A to layer B	% TF	SD from mean (rank)
Top $\rightarrow$ Core	3.8%	2.6 SD below 13% (lowest)
Core $\rightarrow$ Core	87.6%	2.5 SD above 76% (highest)
Top $\rightarrow$ Bottom	0.3%	2.8 SD below 2% (12/13 th)

2.1 HIERARCHICAL STRUCTURES OF THE REGULATORY NETWORKS

- We apply global reach centrality (GRC) to measure the extent of hierarchy in these 3-layer structures.
- GRC: intended to quantify the concept of flow hierarchy
- Introduced by Mones et al. (2012). Hierarchy measure for complex networks. *PLoS One*, **7**
  - Local reach centrality of node i, $C_R(i)$: fraction of nodes in the network that can be reached by node i

$$GRC = \frac{1}{n-1} \sum_{i=1}^{n} \left(C_R^{\max} - C_R(i) \right).$$

- Almost all networks (with 2 exceptions):
  LRC(any top node) $>$ LRC(any core node) $>$ LRC(any bottom node)
- 5-number summary of GRC

Min	1st Qu	Median	3rd Qu	Max
0.065	0.078	0.083	0.089	0.121

- hESC has the highest LRC among 41 networks for top (0.94) and core (0.94) layers. Both are 3SD above their means.

2.2 Housekeeping regulatory interactions

- Analogous to the concept of housekeeping genes, housekeeping interactions (HK interactions) are interactions found in all cell types.
- Neph et al. (2012) identified 2041 interaction.
- Dataset used only 41 networks. We did leave-k-out validation if these are indeed HK interactions.
- For $k = 1$, no more than 1.5% increase in the number of HK interactions. For $k = 2$, average increase is 1.5%, 3rd quartile around 2%, max 5%.
- Enrichment analysis:
  Proportions of HK interactions in "Core $\rightarrow$ Core" and "Core $\rightarrow$ Bottom" are comparable and higher than the other combinations ("Top $\rightarrow$ Core" and "Top $\rightarrow$ Bottom").

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

- Neph et al. (2012) used global connectivity of these 41 networks to classify the cell types.

  They used normalized node degree vectors, a global feature.

- We explored local connectivity.
  - Pick k random TFs (k small). Call this set A.
  - For a network with n nodes, construct $(x_1, \ldots, x_n)$ where

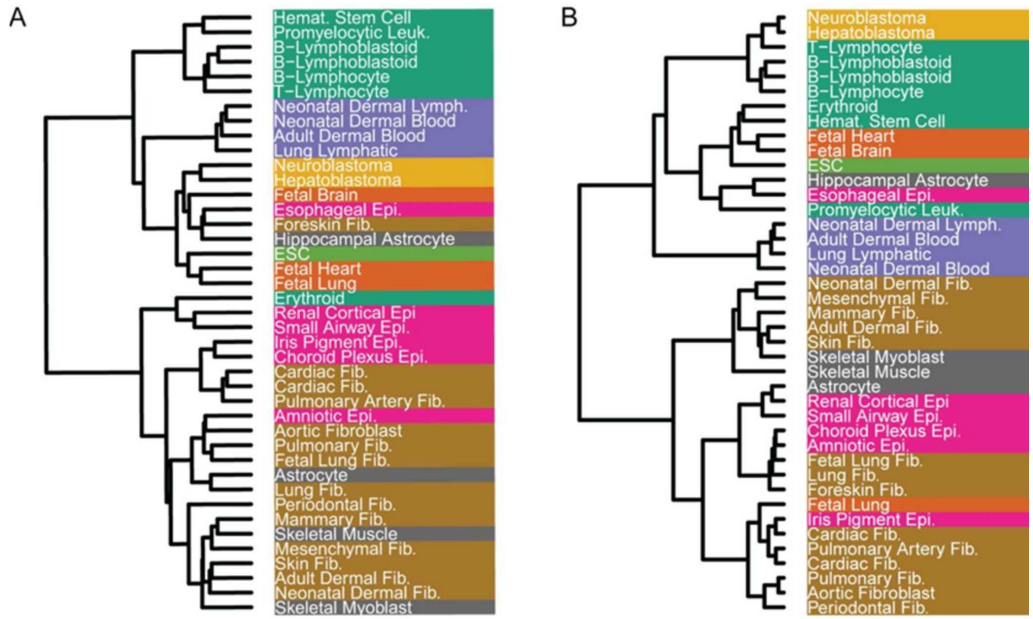
$$x_i = \begin{cases} 1 & \text{if node } i \text{ is a target of a TF in } A \\ 0 & \text{otherwise} \end{cases}.$$

  - Apply PCA for dimension reduction, and use the first seven principal components.
  - Ward clustering to classify the cell types.

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

- Take 7 STAT (signal transducer and activator) TFs, and use their local connectivity for clustering
- Left panel is from Neph et al. (2012); right panel based on 7 STAT TFs.

Hierarchical clustering of 41 cell types (colour indicates which class it belongs)



Zhang et al. Nucl. Acids Res. 2014;42:12380-12387

Rand Index $= 0.801$ Rand Index $= 0.856$

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

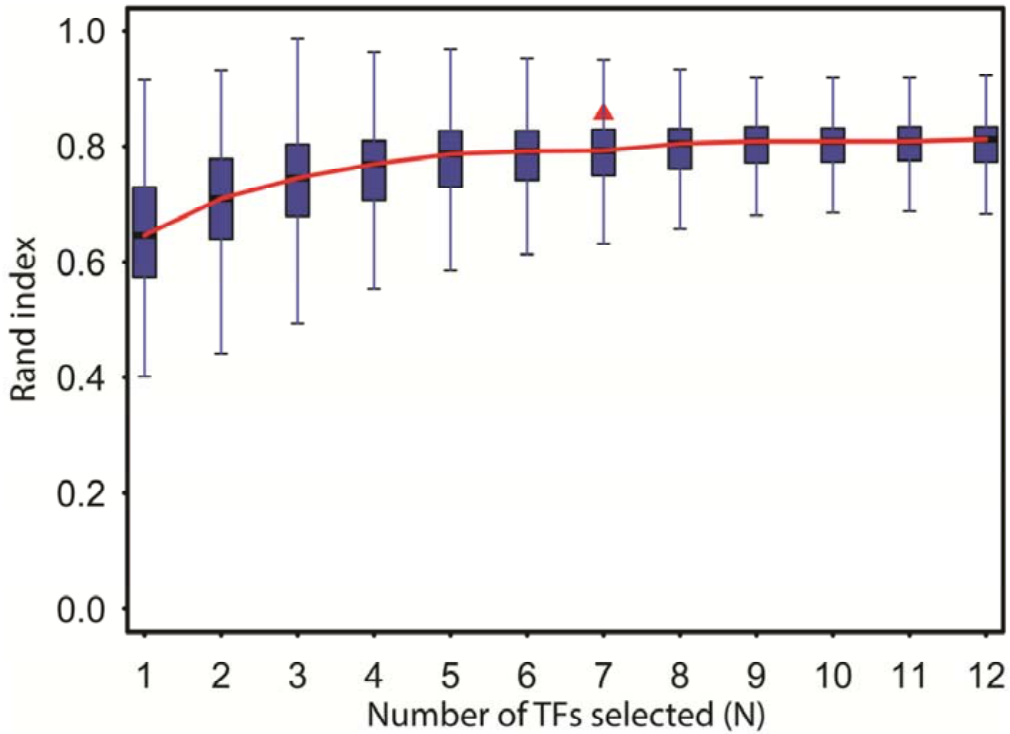
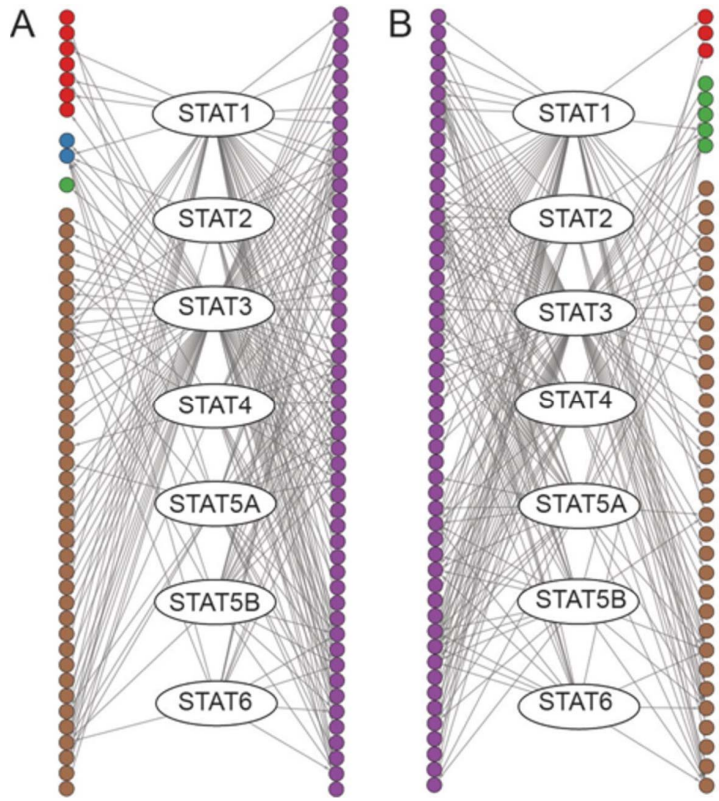


Figure S1 The boxplots of the Rand Index values for the classifications of the 41 cell types using N randomly selected TFs. The number of repetitions taken was 1000 for each N. See the Methods section for details.

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

- Downstream targets of STAT TFs in hESC & HSC
- GO term: cell fate commitment process (GO: 0045165), in red dots, is enriched in hESC

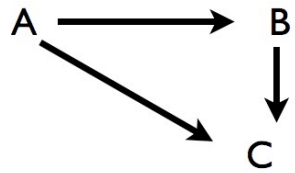
The STATs and their downstream regulatory targets in hESCs (A) and HSCs (B).



Shihua Zhang et al. Nucl. Acids Res. 2014;42:12380-12387

3. Work in progress

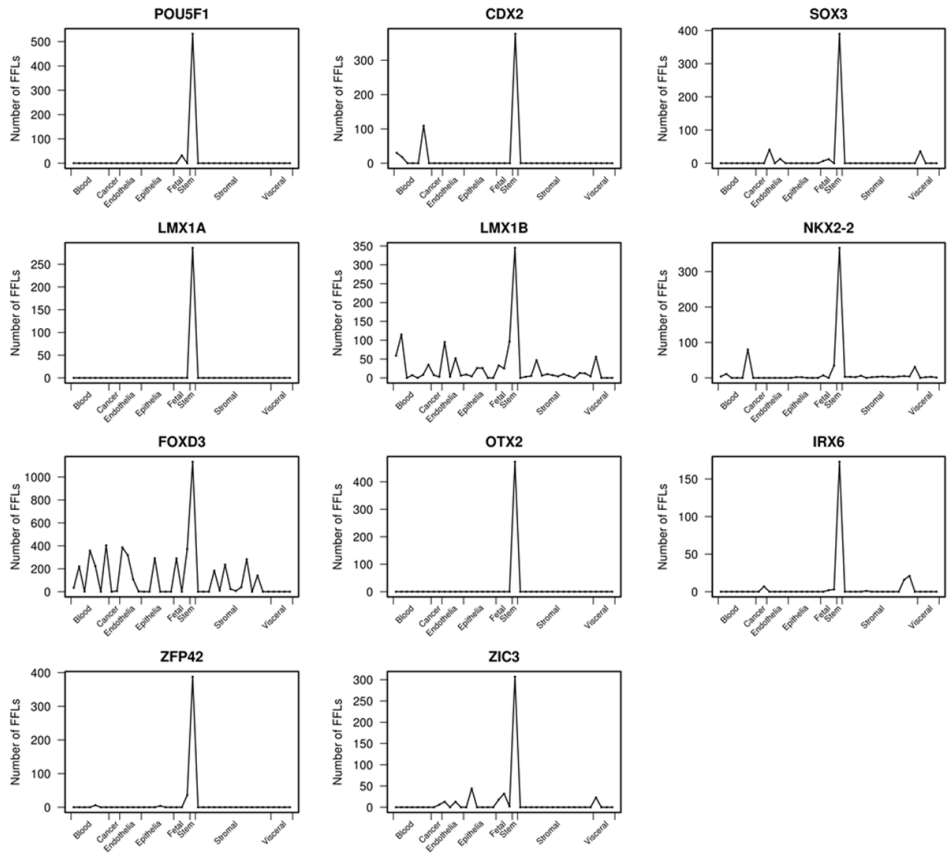
- Feed forward loops
- Examine targets, C, of feed forward loops (FFL)



- Study TFs that are extensively regulated by FFLs
- Construct a count matrix: $M = [m_{ij}]$ where i is the ith TF, and j is the jth network.

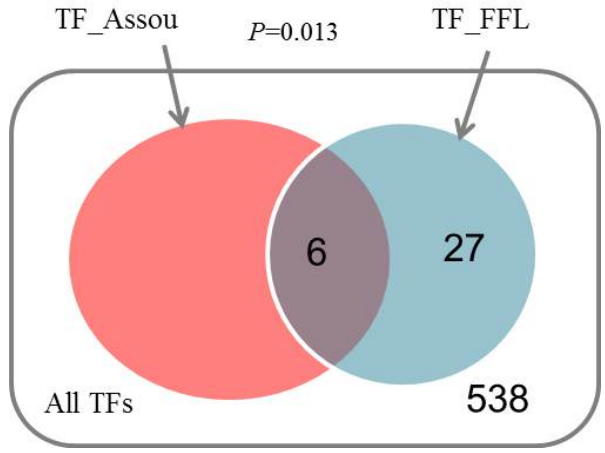
	hESC	Blood 1	Blood 2	$\cdots$	Cancer 1	$\cdots$
OCT4	532	0	0	$\cdots$	0	$\cdots$
SOX2	513	67	58	$\cdots$	101	$\cdots$
NANOG	37	0	0	$\cdots$	3	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	

TFs extensively regulated by FFLs in hESC



3. WORK IN PROGRESS

- 33 TFs extensively regulated by FFLs in hESC network
- Meta-analysis study by Assou et al (2007) identified 1076 up-regulated genes. 34 TFs are found in the hESC network.



- Common TFs found are FOXD3, OCT4, OTX2, SOX3, ZFP42, ZIC3
- TFs extensively regulated by FFLs in hESC are not found in the down-regulated gene list in Assou et al (2007).
- Our question: how best to refine our approach?

Summary for Part II

- 41 human cell type TF regulatory networks provide opportunity to compare and contrast their organizational architecture.
- Global & local connectivity in the networks are different for different cell types.
- In many ways, the organization architecture of hESC stands out from the rest.
- Identify housekeeping interactions.

Thank You

Acknowledgement

Based on joint work with

- Dechao Tian
- Ngoc Hieu Tran
- Louxin Zhang, NUS
- Shihua Zhang, Chinese Academy of Sciences, China