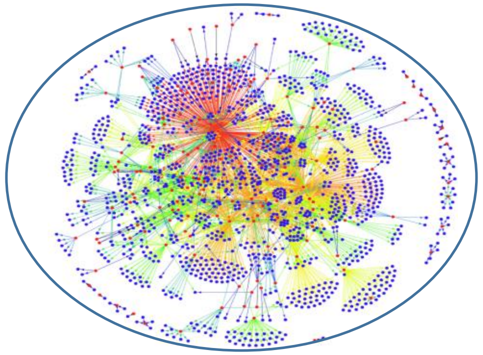
Profiling Human Cell/Tissue Specific Gene Regulation Networks

Louxin Zhang
Department of Mathematics
National University of Singapore
matzlx@nus.edu.sg

Network Biology



$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$$

$$< y_1\ (t), y_2(t), \cdots, y_k(t) >$$

subject to:

$y_j(t) = \varphi_j(y_1, \ldots, y_k, I_1, \ldots, I_m)(t),$

$\varphi_1, \varphi_2, \ldots, \varphi_k$ are governing functions

I_1(t), I_2(t), …, $I_m(t)$ are input functions

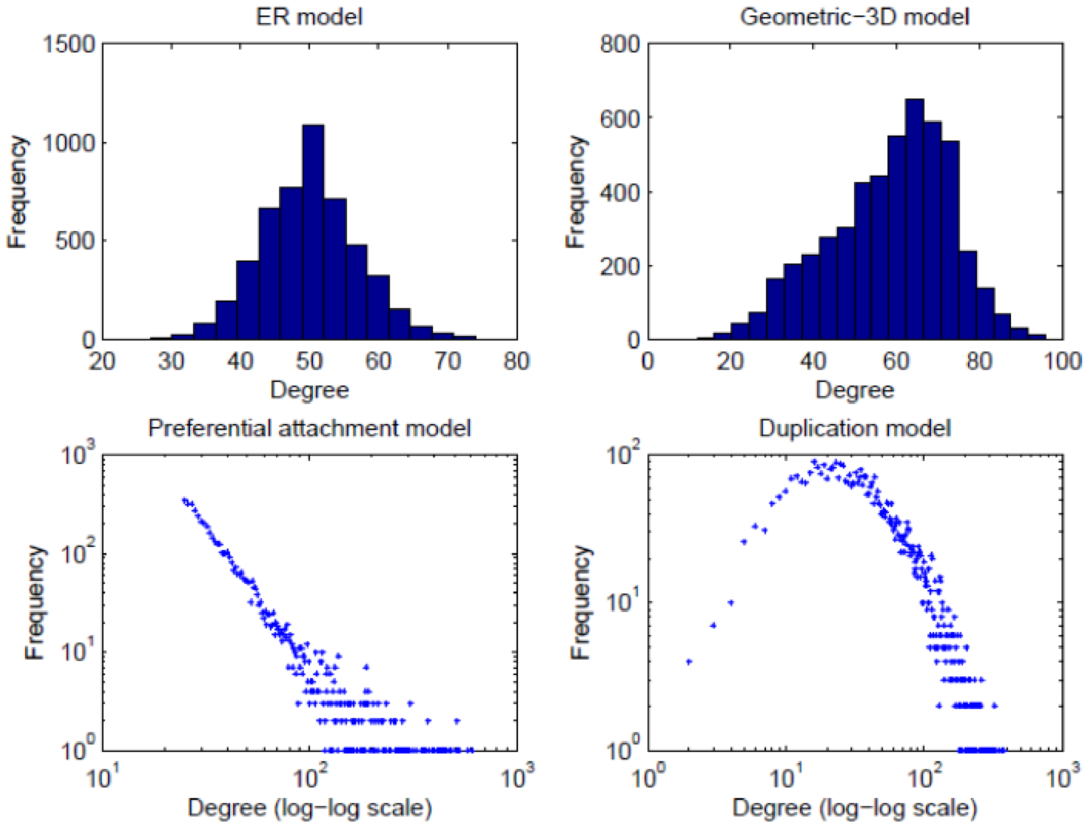
Cell

Cell State

Cell Type

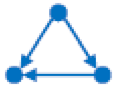
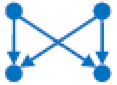
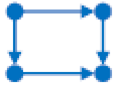
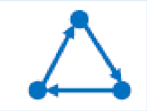
Topological Features of Cellular Protein Networks

- Degree distribution: scale-free or not



➢ Over (or under)-represented network motifs

Motif: A graph pattern that appears much more frequently than would be expected in randomized networks

 Feed-Forward Loop  Bi-Fan  Bi-Parallel	Over-represented in transcriptional regulation networks
 Feedback Loop	Under-represented in transcriptional regulation networks

Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii, Alon, ***Science***, 2002
Albergante, Blow, Newman, ***eLife***, 2014

Today’s Talk

- Difference in regulatory interaction between human cells/tissues
- Hierarchical organization of the TF regulatory networks of human cells/tissues
- Motif counts for human cell/tissue specific TF reg. networks

Tran, Choi, Zhang, ***Nature Communications****, 2013*
Zhang, Tian, Tran, Choi, Zhang, ***Nucleic Acids Res***., 2014

Dataset

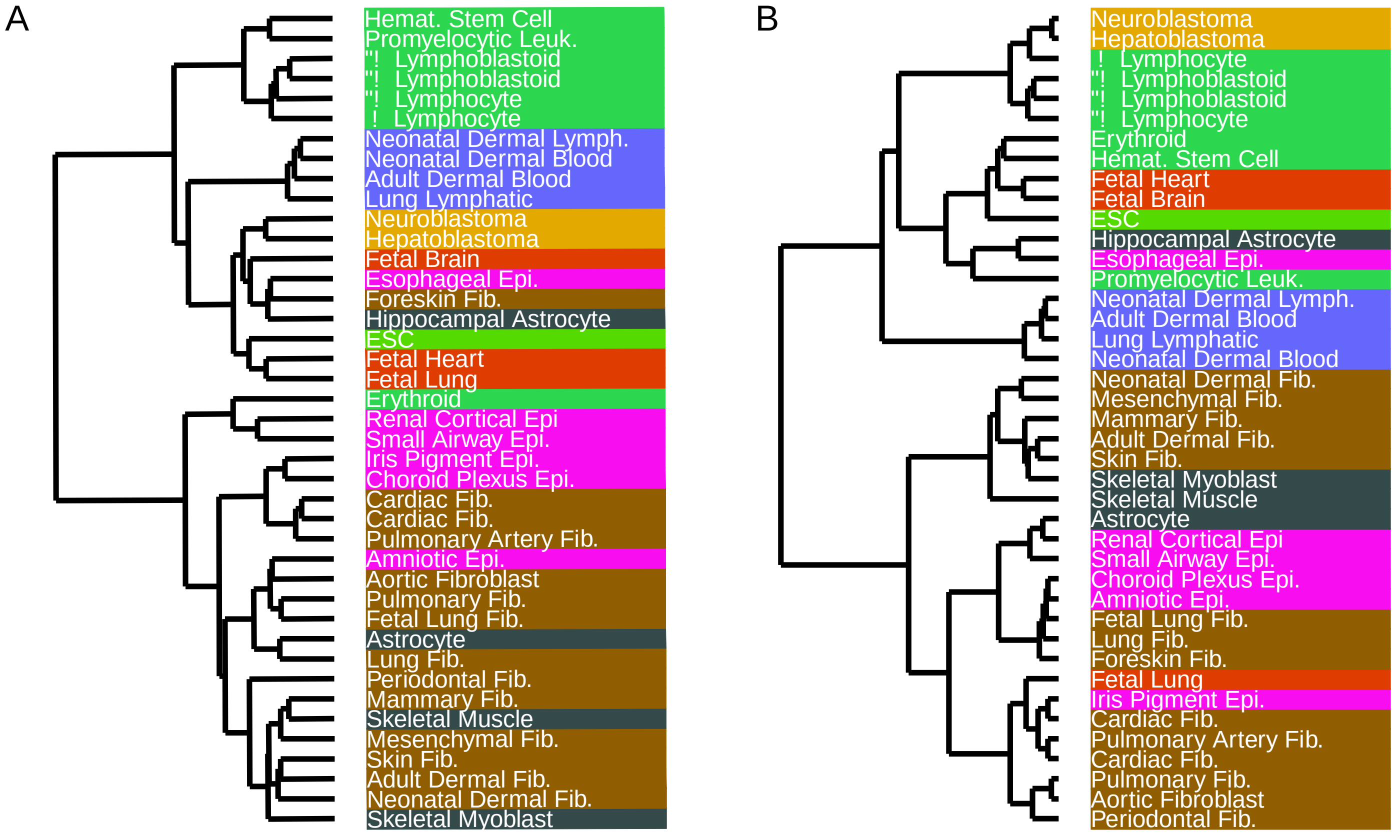
- 41 cell/tissue specific TF regulatory networks for humans

  Blood (7)
  Cancer (2)
  Endothelia (4)
  Epithelia (6)
  Embryonic stem cell (ESC) (1)
  Fetal (3)
  Stroma (14)
  Viscera (4)

- Each has about 475 TFs and 11,200 regulatory interactions
- Derived from DNaseI footprint data and predicted TRANSFAC binding-site motifs.

Neph et al., ***Cell***, 2012

Part I Difference in Reg. Interaction between Human Cell Types

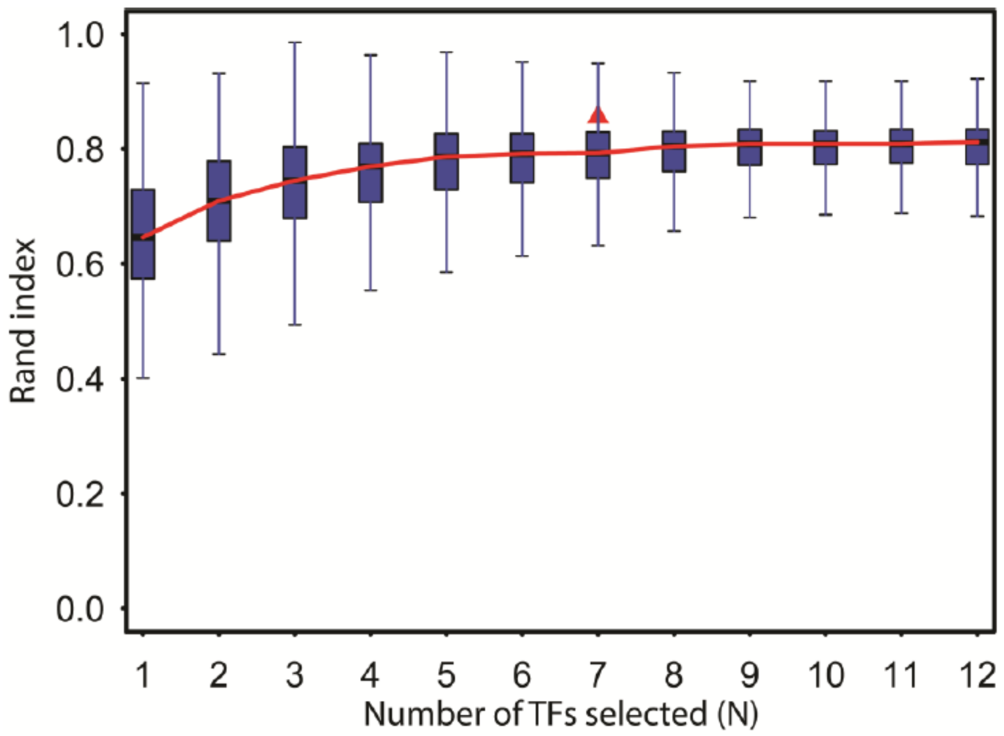


- Left: Using the node-degree vectors of the entire networks (Neph et al., 2012)
- Right: Using the node-degree vectors of a subnetwork around six STATs

	Rand Index
The classification A:	0.801
The classification B:	0.856

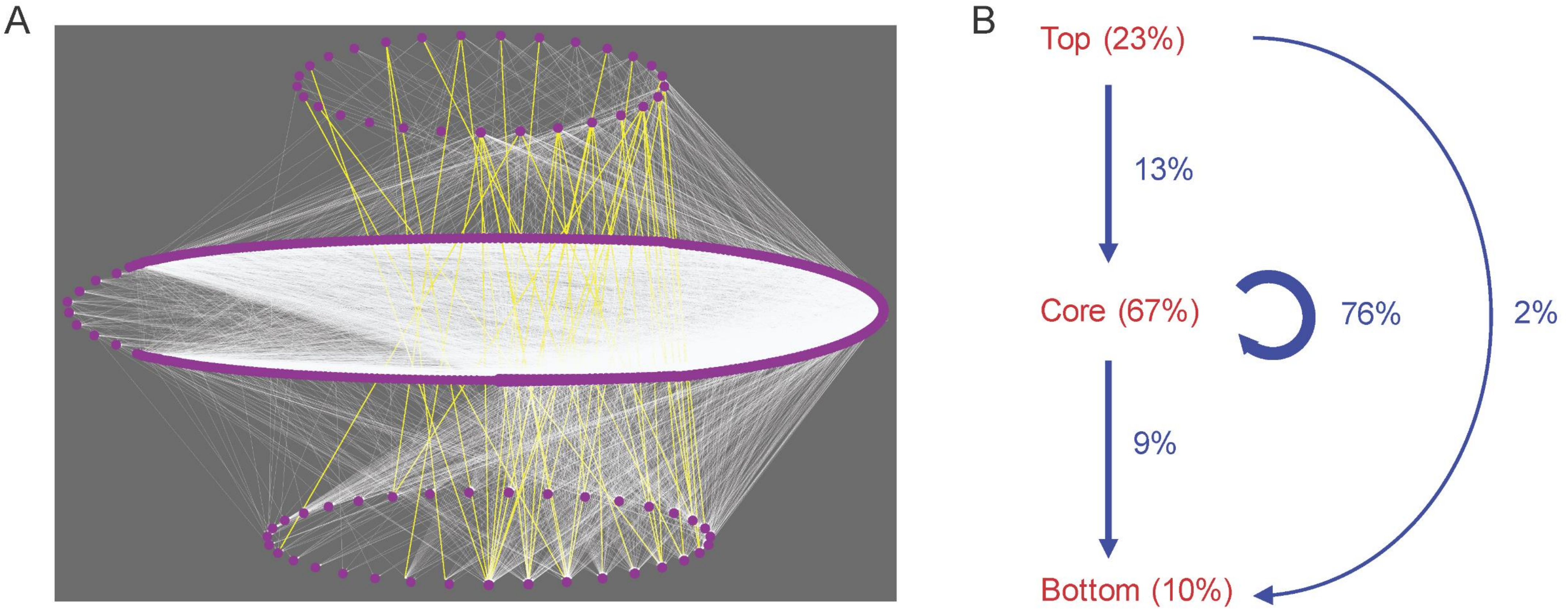
Finding:

Wiring around a few TFs can distinguish cell identity well

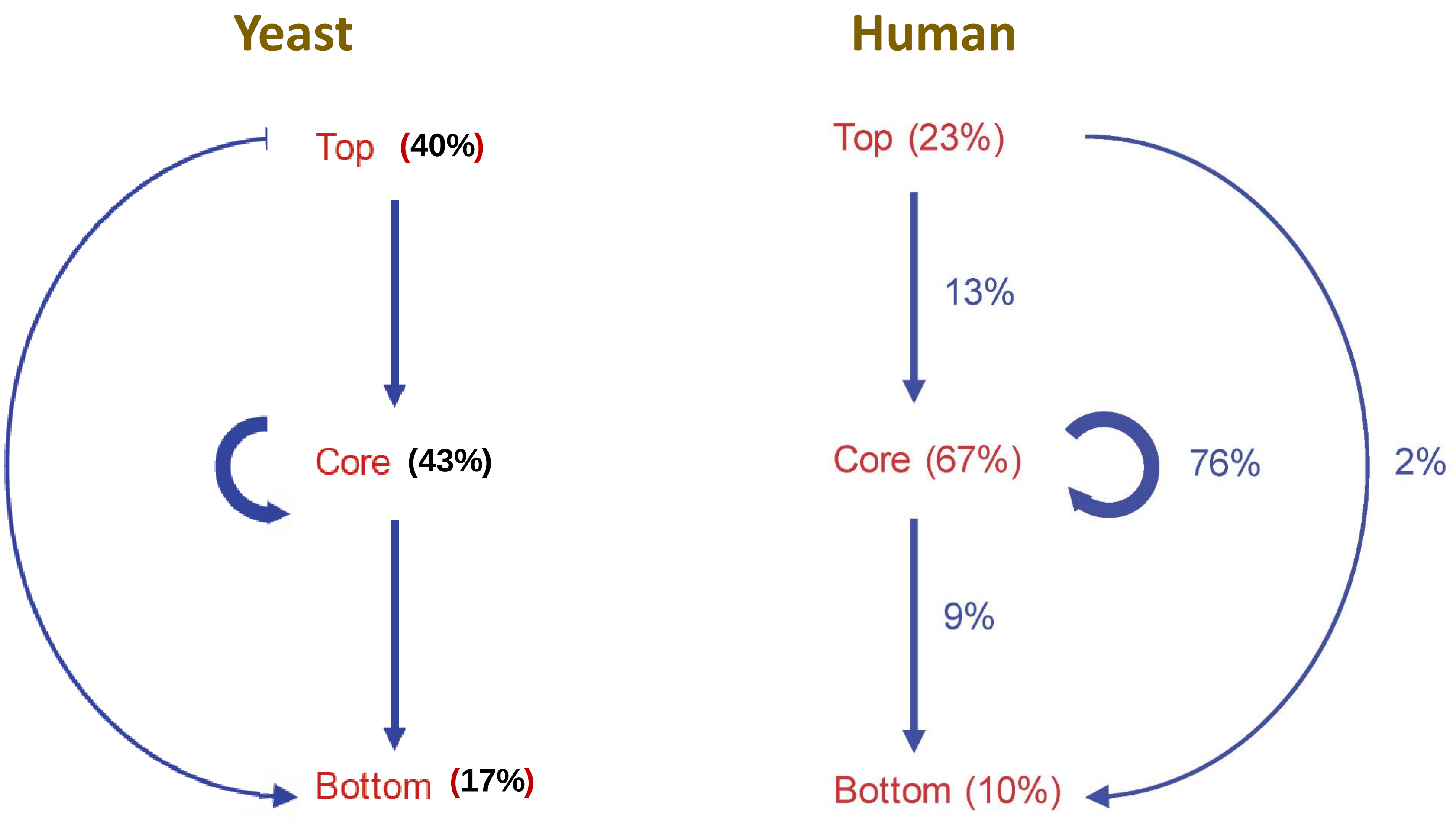


- ➢ Both Neph et al (2012) and our studies suggest that 41 cell specific human TF regulatory networks are different globally as well as locally.

Part II Hierarchical Organization of human TF Networks



- ➢ The hierarchical structures obtained using a vertex-sort algorithm (Jothi et al., 2009)
- ➢ hESC has much less TFs in the top layer (6%), much more TFs in the core layer (85%) than other cell types.



- A difference in the topological organization between human and yeast TF networks.

The Enrichment (+) and Depletion (–) of Hub, Essential and Housekeeping (HK) TFs

	Hub TFs*			Essential TFs			HK TFs		
	Top	Core	Bottom	Top	Core	Bottom	Top	Core	Bottom
Blood (7)	–	+	–			–	–	+	
Cancer (2)	–	+	–				–	+	
Endothelia (4)	–	+	–			–	–	+	
Epithelia (6)	–	+	–				–	+	
ESC (1)		+	–		+	–			
Fetal (3)	–	+	–		+	–	–	+	
Stroma (14)		+	–				–	+	
Viscera (4)	–	+	–			–	–	+	

* Hubs: top 20% TFs with the largest degrees

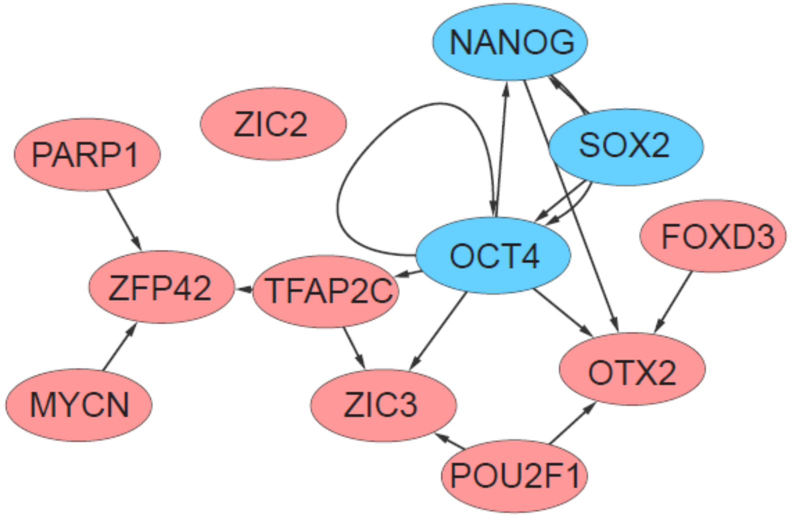
Regulatory Interactions Specific to hESCs

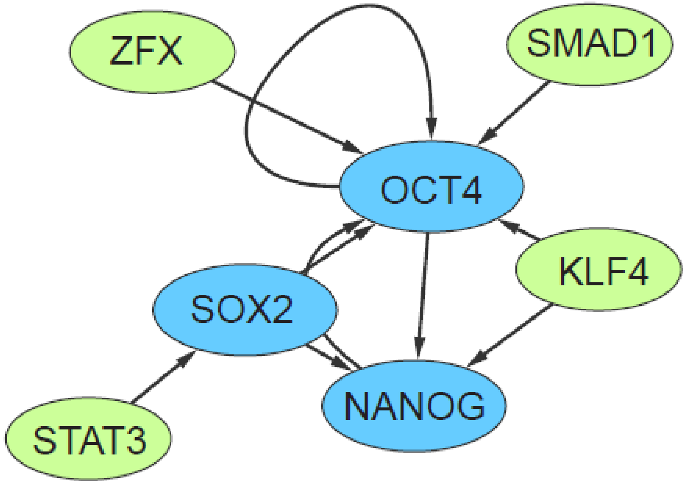
- 1509 interactions are specific to hESCs, involving 411 TFs
- The subnetwork induced by these specific interactions has 82 hubs (top 20%, degree >11), only 35 are the hubs in the original network of ESCs.

	Hubs only in the specific network				Hubs also in the original networks			
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	SPI1	ZFP42
	BARHL2	FOXJ1	NR5A2	TBX22	MYCN	OCT4	SREBF2	ZNF148
	CDX2	GFI1	OTP	VAX1	NF1	PURA	STAT3	ZNF219
	CRX	HOXB13	OTX2	ZIC1	NFKB2	REST	TCF3	ZNF216
	DMRT1	LHX4	PARP1	ZIC2	NR2F1	RXRA		
	DMRT3	LMX1A	PAX2	ZIC3				
	ETV7							
Bottom	HBP1	OVOL2	PAX7	SIX6	Red TFs are encoded by a gene with a super-enhancer domain			

- The 82 hubs are enriched with the TFs encoded by the 1076 genes that are overexpressed in hESCs (p-value < 1.6e-3) (Assou et al, 2007)

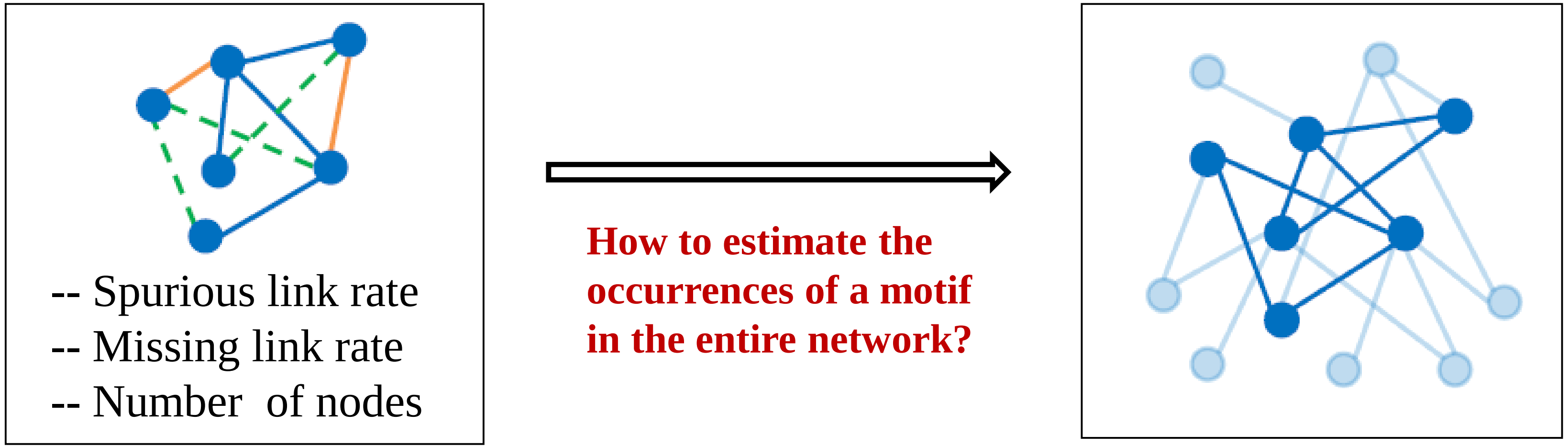
- The core transcriptional regulatory subnetwork for ESC reported by Chen et al. (2008) is enriched with hESC specific regulatory interactions.





Part III Counting Motifs for TF Regulatory Networks

- Currently, the existing regulatory interaction data are incomplete and noise
- What is the size of the entire TF regulatory network *N* of a cell type?
- How many times does the feed-forward loop appear in the entire network *N* ?



-- Spurious link rate
-- Missing link rate
-- Number of nodes

How to estimate the occurrences of a motif in the entire network?

Incomplete, noisy data

The entire network

Estimators for Error-Free Subnetwork Data

M: A graphlet with m nodes

$\mathcal{G}$: An entire network with n nodes

$\mathcal{G}^{\text{obs}}$: An observed, error-free sub-network (of $\mathcal{G}$) with n^{obs} nodes, in which M occurs $N_M{}^{\text{obs}}$ times

Estimator:

$$\hat{N}_M = \frac{\binom{n}{m}}{\binom{n^{obs}}{m}} N_M{}^{\mathbf{obs}}$$

$$\hat{N}_{\mathrm{E}} = \frac{\binom{n}{2}}{\binom{n^{obs}}{2}} N_{\mathrm{E}}{}^{\mathbf{obs}}$$

Stumpf *et al.*, PNAS, 2008

Theorem Assume $\mathcal{G}^{\text{obs}}$ is obtained by uniformly selecting nodes each with probability 0<p<1 in $\mathcal{G}$. Then,

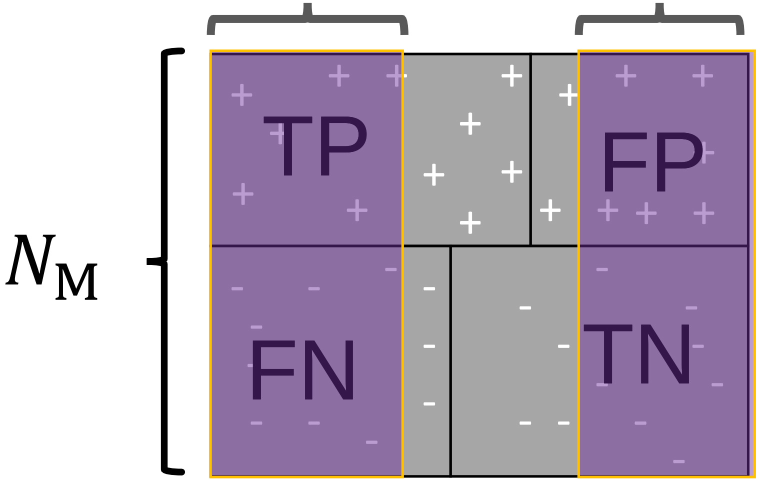
$$E\left(\frac{\hat{N}_M}{N_M}\right) = 1 - \sum_{0 \le j < m} \binom{n}{j} p^j (1-p)^{n-j},$$

converging to 1 fast as n goes to infinity.

Estimator $\widetilde{N}_M$ from Incomplete and Noise Subnetwork Data

$$\widehat{N}_M = \frac{\binom{n}{m}}{\binom{n^{obs}}{m}} N_M{}^{\mathbf{obs}}$$

$$N_M{}^{obs} = \text{TP} + \text{FP}$$



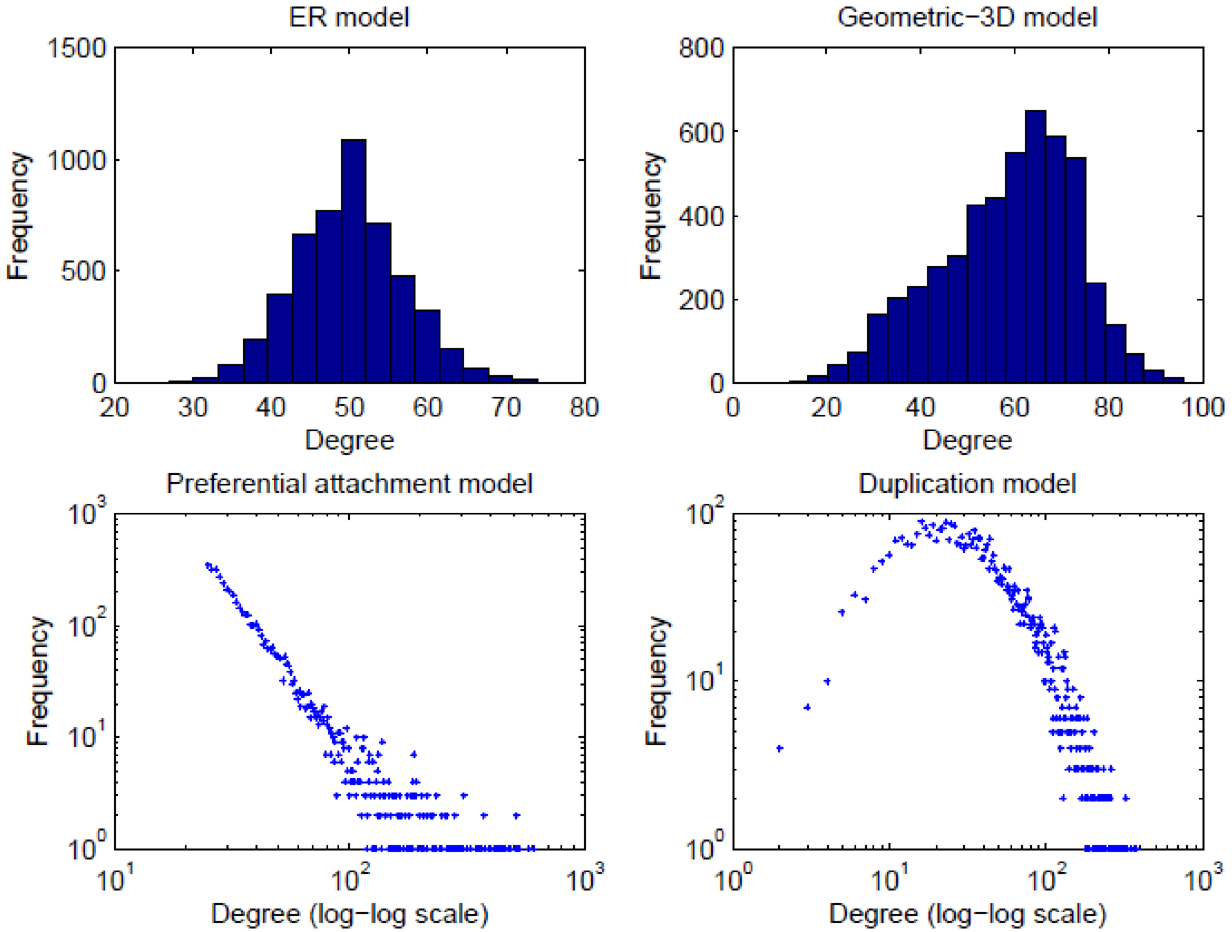
Missing link rate $r_- = \dfrac{\text{FN}}{\text{TP+F}N}$

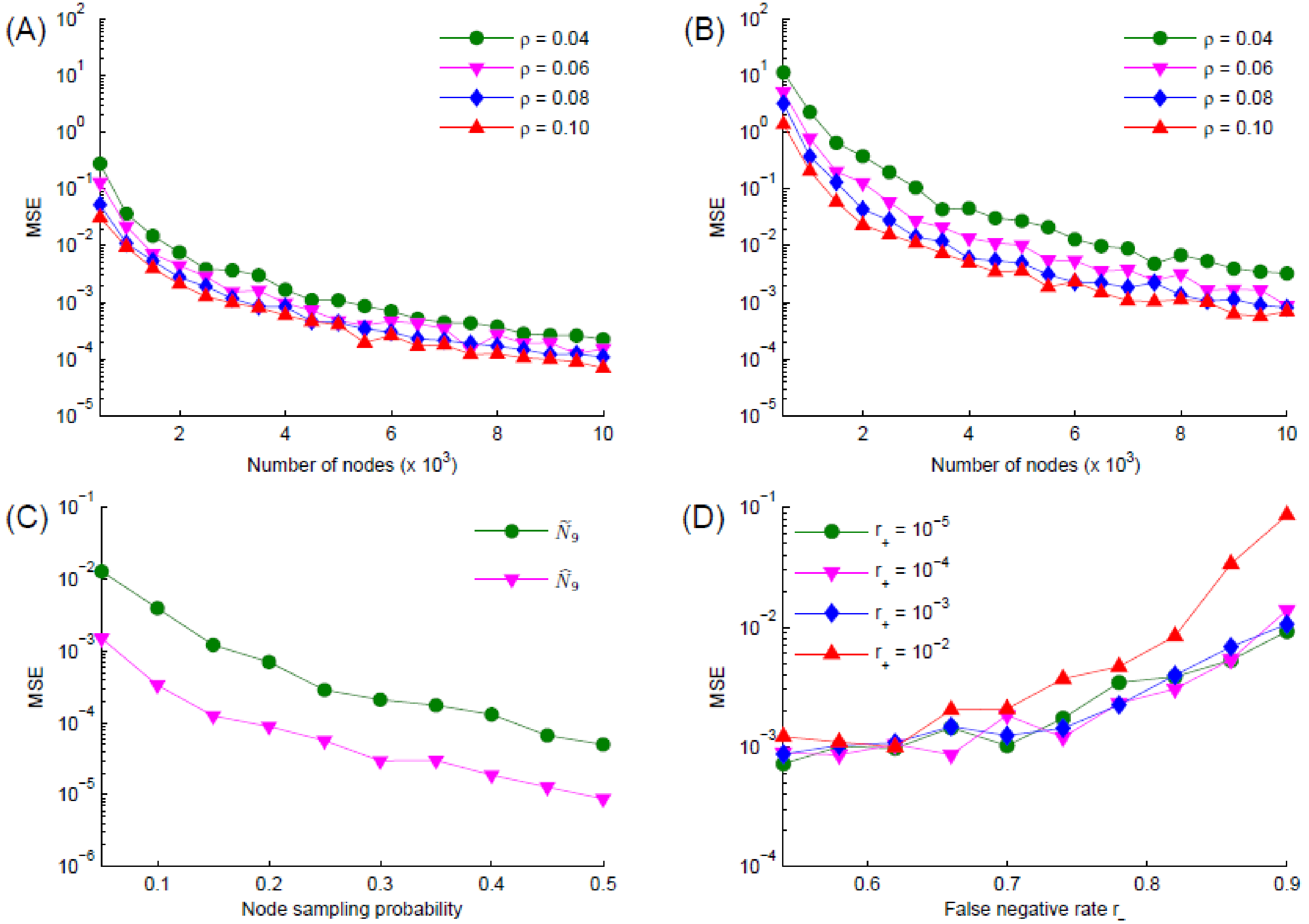
Spurious link rate $r_+ = \dfrac{\text{FP}}{\text{FP+}TN}$

$$E\left(\widehat{N}_M\right) = \left(1 - \sum_{0 \le j < m} \binom{n}{j} p^j (1-p)^{n-j}\right)\left[(1 - r_+ - r_-)^{|M|} N_M + \boldsymbol{W_M(n, r_+, r_-, N_{M'})}\right]$$

$$\widetilde{N}_M = \frac{1}{(1 - r_+ - r_-)^{|M|}}\left(\widehat{N}_M - \widetilde{W}_M(n, r_+, r_-, N_{M'})\right)$$

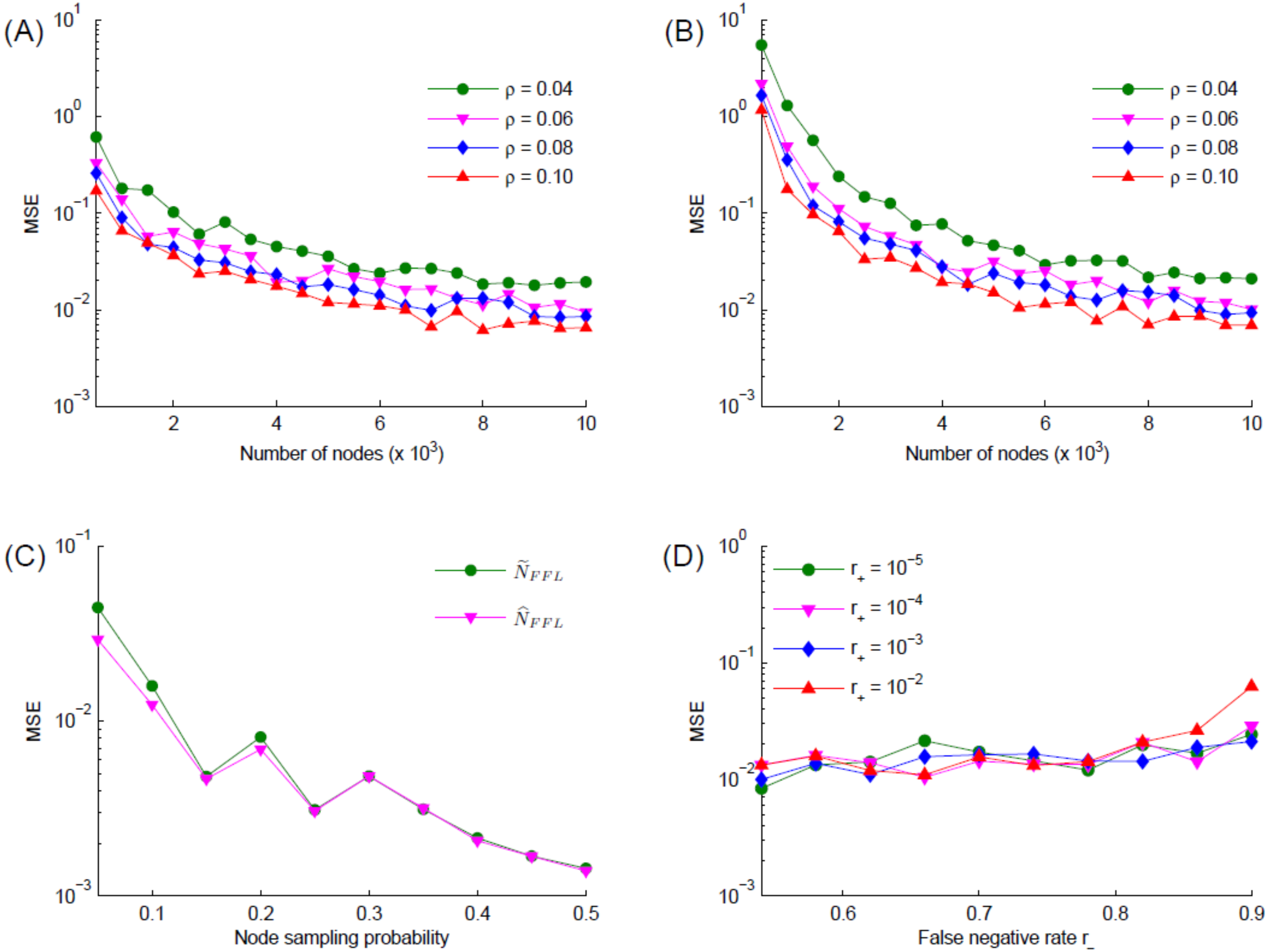
Validation for Unbiasedness and Consistence





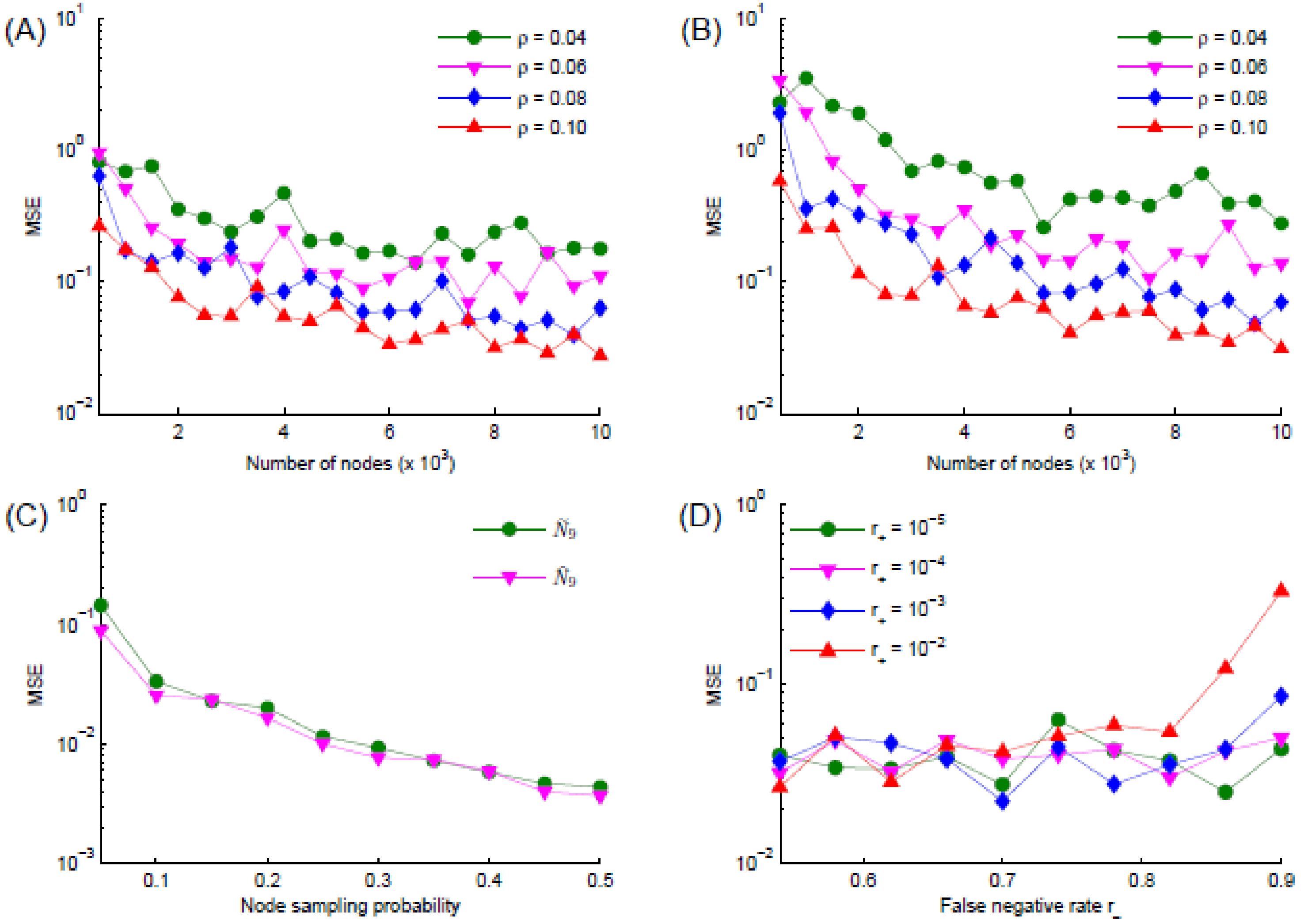
Motif: Feed Forward Loop

Model: ER



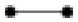
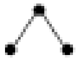
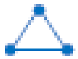
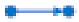
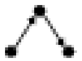
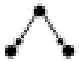
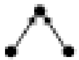
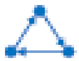
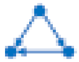
Motif: Feed Forward Loop

Model: Preferential Attachment

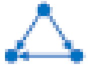
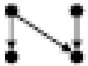
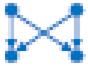
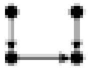
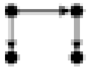
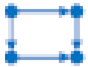


Motif: Feed Forward Loop

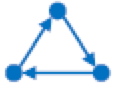
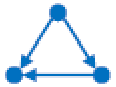
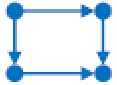
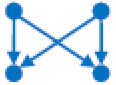
Model: Duplication

Motif		Bias-corrected Estimator
1		$\widetilde{N}_1 = \frac{1}{r}\left[\widehat{N}_1 - \binom{n}{2}r_+\right]$
2		$\widetilde{N}_2 = \frac{1}{r^2}\left[\widehat{N}_2 - 2(n-2)r_+r\widetilde{N}_1 - 3\binom{n}{3}r_+^2\right]$
3		$\widetilde{N}_3 = \frac{1}{r^3}\left[\widehat{N}_3 - r_+r^2\widetilde{N}_2 - (n-2)r_+^2r\widetilde{N}_1 - \binom{n}{3}r_+^3\right]$
4		$\widetilde{N}_4 = \frac{1}{r}\left[\widehat{N}_4 - 2\binom{n}{2}r_+\right]$
5		$\widetilde{N}_5 = \frac{1}{r^2}\left[\widehat{N}_5 - 2(n-2)r_+r\widetilde{N}_4 - 6\binom{n}{3}r_+^2\right]$
6		$\widetilde{N}_6 = \frac{1}{r^2}\left[\widehat{N}_6 - (n-2)r_+r\widetilde{N}_4 - 3\binom{n}{3}r_+^2\right]$
7		$\widetilde{N}_7 = \frac{1}{r^2}\left[\widehat{N}_7 - (n-2)r_+r\widetilde{N}_4 - 3\binom{n}{3}r_+^2\right]$
8		$\widetilde{N}_8 = \frac{1}{r^3}\left[\widehat{N}_8 - r_+r^2\widetilde{N}_5 - (n-2)r_+^2r\widetilde{N}_4 - 2\binom{n}{3}r_+^3\right]$
9		$\widetilde{N}_9 = \frac{1}{r^3}\left[\widehat{N}_9 - r_+r^2(\widetilde{N}_5 + 2\widetilde{N}_6 + 2\widetilde{N}_7) - 3(n-2)r_+^2r\widetilde{N}_4 - 6\binom{n}{3}r_+^3\right]$

$$\widehat{N}_M = \frac{\binom{n}{|M|}}{\binom{n^{\mathrm{obs}}}{|M|}} N_M{}^{\mathrm{obs}}, \quad r{=}1 - r_+ - r_-$$

	Motif	Bias-corrected Estimator
9		$\widetilde{N}_9 = \frac{1}{r^3}\left[\widehat{N}_9 - r_+r^2(\widetilde{N}_5 + 2\widetilde{N}_6 + 2\widetilde{N}_7) - 3(n-2)r_+^2 r\widetilde{N}_4 - 6\binom{n}{3}r_+^3\right]$
10		$\widetilde{N}_{10} = \frac{1}{r^3}\left\{\widehat{N}_{10} - 2r_+r^2\left[\binom{\bar{N}_4}{2} + (n-3)(\widetilde{N}_6 + \widetilde{N}_7)\right] - 6\binom{n-2}{2}r_+^2 r\widetilde{N}_4 - 24\binom{n}{4}r_+^3\right\}$
11		$\widetilde{N}_{11} = \frac{1}{r^4}\left\{\widehat{N}_{11} - r_+r^3\widetilde{N}_{10} - r_+^2r^2\left[\binom{\bar{N}_4}{2} + (n-3)(\widetilde{N}_6 + \widetilde{N}_7)\right] - 2\binom{n-2}{2}r_+^3 r\widetilde{N}_4 - 6\binom{n}{4}r_+^4\right\}$
12		$\widetilde{N}_{12} = \frac{1}{r^3}\left\{\widehat{N}_{12} - r_+r^2\left[2\binom{\bar{N}_4}{2} + (n-3)(\widetilde{N}_5 + 2\widetilde{N}_7)\right] - 6\binom{n-2}{2}r_+^2 r\widetilde{N}_4 - 24\binom{n}{4}r_+^3\right\}$
13		$\widetilde{N}_{13} = \frac{1}{r^3}\left\{\widehat{N}_{13} - r_+r^2\left[2\binom{\bar{N}_4}{2} + (n-3)(\widetilde{N}_5 + 2\widetilde{N}_6)\right] - 6\binom{n-2}{2}r_+^2 r\widetilde{N}_4 - 24\binom{n}{4}r_+^3\right\}$
14		$\widetilde{N}_{14} = \frac{1}{r^4}\Big\{\widehat{N}_{14} - r_+r^3(\widetilde{N}_{12} + \widetilde{N}_{13}) - r_+^2r^2\left[2\binom{\bar{N}_4}{2} + (n-3)(\widetilde{N}_5 + \widetilde{N}_6 + \widetilde{N}_7)\right]$ $-4\binom{n-2}{2}r_+^3 r\widetilde{N}_4 - 12\binom{n}{4}r_+^4\Big\}$

Estimation for Four Motifs in the Human Cell-Specific TF Networks

				
Blood cells	3,687	37,884	4,379,527	7,359,970
Cancer cells	2,738	30,122	2,862,215	6,267,987
Endothelia cells	3,160	35,314	3,844,161	6,877,606
Epithelia cells	1,896	19,901	1,858,957	3,238,587
Fetal cells	3,088	33,782	3,660,840	6,498,027
Stroma cells	2,727	29,155	3,052,803	5,094,576
ES cells	2,766	32,400	3,282,473	6,436,708

- Embryonic stem cell has the smallest motif count relative to its network size
- Promyelocytic leukemia blood cell has the largest motif count
- The feed-forward loop count is about 10 times that of the feed-back loop.

Motif Counts for PPI Networks

	S. cerevisiae[1]	*C. elegans*[2]	*H. sapiens*[3]	*A. Thaliana*[4]
Total no. of proteins	6,000	20,065	22,500	27,029
No. of proteins examined	3,676	9,906	7,194	7,108
No. of interactions detected	967	1,816	2,754	4,890
Precision*	0.9400	0.8600	0.7940	0.8030
Sensitivity*	0.1700	0.0496	0.0950	0.1570
Missing link rate r_-	0.8300	0.9500	0.9050	0.8430
Spurious link rate r_+	0.000008	0.000005	0.000002	0.000003

[1] Yu et al. 2008

[2] Simonis et al. 2009

[3] Rual et al 2005; Venkatesan et al. 2009

[4] Arabbidopsis Interactome Mapping Consortium, 2011

* Reported in 1-4

Predicted sizes of four model interactomes

	S. cerevisiae	*C. elegans*	*H. sapiens*	*A. Thaliana*
CCSB estimate	18,000±4,500	116,000±26,400	130,000±32,600	299,000±79,200
Our estimate	15,000±2,700	122,000±16,600	214,000±32,200	376,000±45,600
Hart et al. estimate	37,800-75,500		154,000-369,000	

Hart, Ramani and Marcotte, 2006

Triangle counts for four model interactomes

	S. cerevisiae	*C. elegans*	*H. sapiens*	*A. Thaliana*
Our estimate	53,000	6,263,000	10,270,000	10,697,000
Triangle density	1×10^{-6}	5×10^{-6}	5×10^{-6}	5×10^{-6}

The number of triangles in the human PPI network is 194 times that of the yeasts,
3 times as large as expected.

Summary

- Topological structures of 41 cell/tissue specific TF networks
- Human ESC specific regulatory interactions
- Motif counting for the TF networks.

Acknowledgements

Co-authors and collaborators

Kwok Pui Choi,	National University of Singapore
Dechao Tian,	A Postdoc at Carnegie Mellon University
Ngoc Hieu Tran	A Postdoc at Nanyang Tech. University
Shihua Zhang	Academy of Math and Systems Sci., China

Funding

Singapore Ministry of Education Tier-2 Research Grant