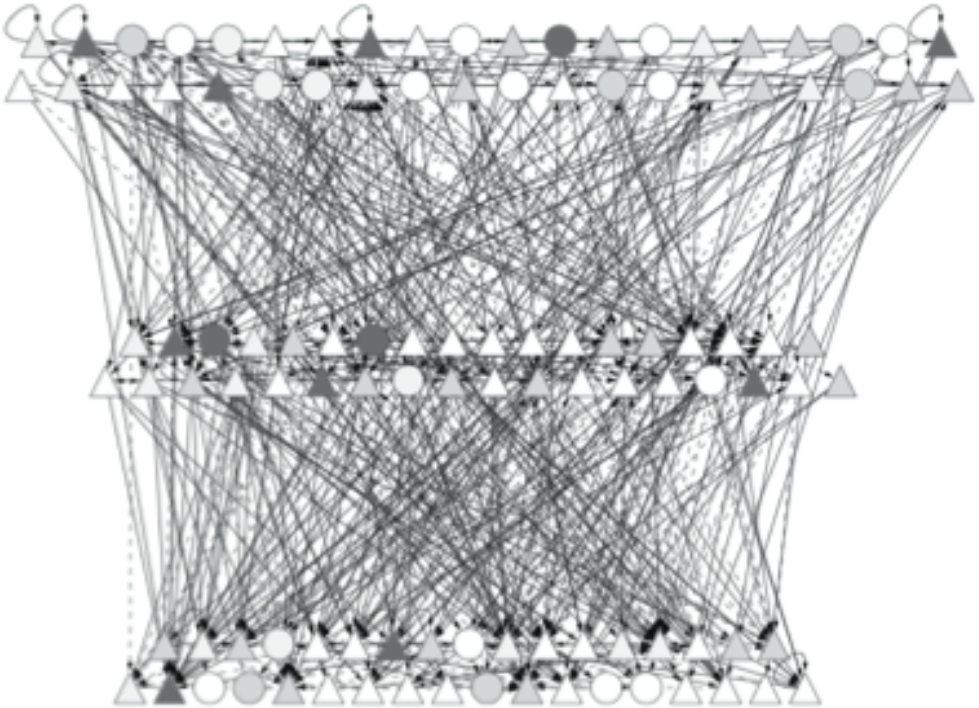
Deciphering Functions and Regulatory Programs of Coding Genes and ncRNAs using ENCODE Data

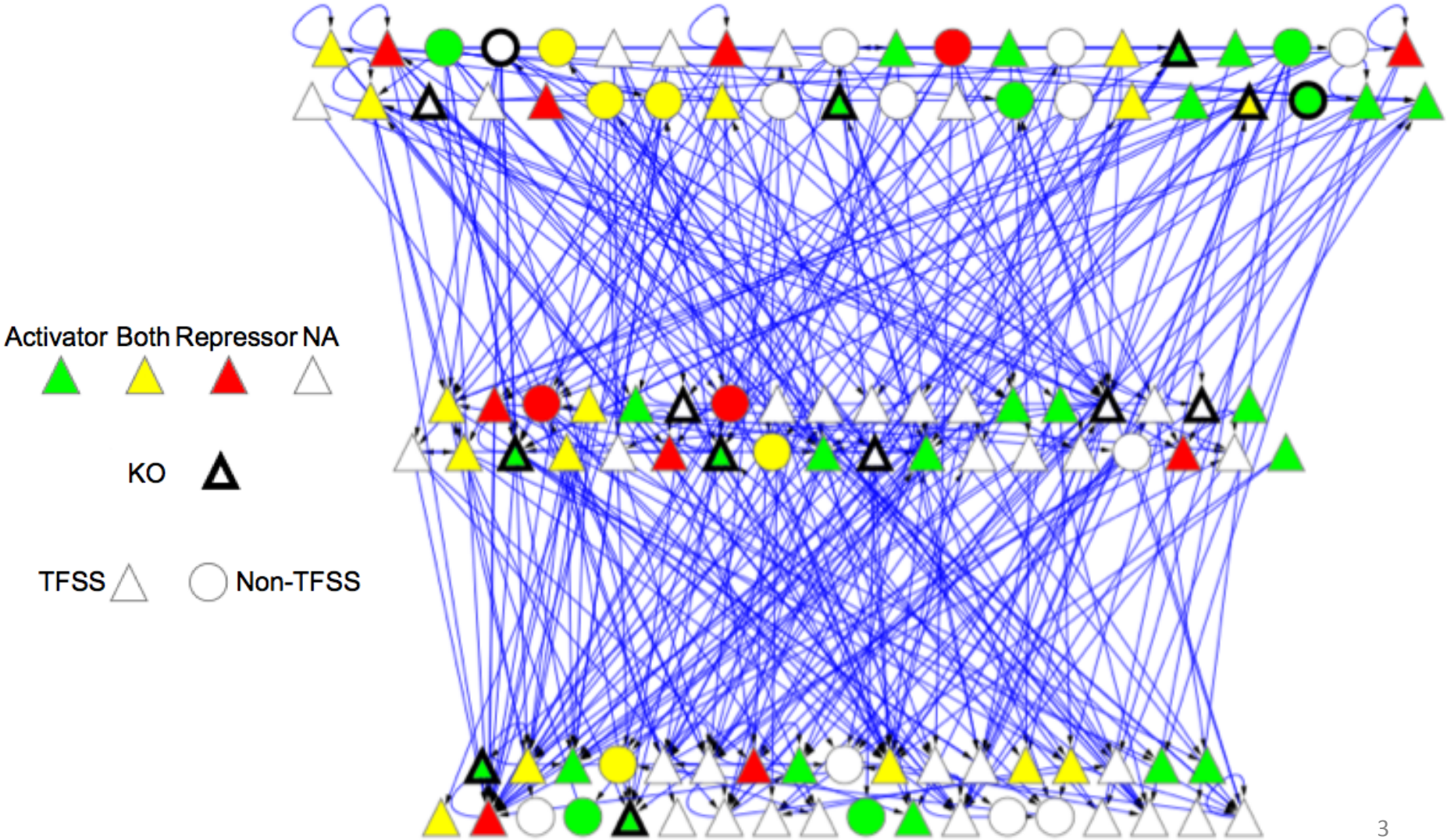
Renqiang (Martin) Min
In collaboration with
Chao Cheng and Mark Gerstein
Program in Computational Biology and Bioinformatics
Yale University
Dec 8, 2011

Building hierarchy (Chao Cheng and Koon-Kiu Yan)

- To find the hidden intrinsic direction in the presence of cycles
- Simple-minded method:
  - Divide TFs into 3-layers (Top, Middle, Bottom)
  - Divide TFs into layers based on whether they are regulating or being regulated by TFs at different levels
  - Minimize bottom-top directional edges



Functional Support (multi-tasker argument) (Thanks Pedro!)



Chromatin Organization and Chromatin Modification TFs Enriched in Top Level

- ID Gene Name Species
- CTCF CCCTC-binding factor (zinc finger protein) Homo sapiens
- CTCFL CCCTC-binding factor (zinc finger protein)-like Homo sapiens
- EP300 E1A binding protein p300 Homo sapiens
- KAT2A K(lysine) acetyltransferase 2A Homo sapiens
- SETDB1 SET domain, bifurcated 1 Homo sapiens
- SMARCA4 SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 Homo sapiens
- SMARCB1 SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1 Homo sapiens
- SMARCC1 SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1 Homo sapiens
- SMARCC2 SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2 Homo sapiens
- CHD2 chromodomain helicase DNA binding protein 2 Homo sapiens
- HDAC2 histone deacetylase 2 Homo sapiens
- IRF4 interferon regulatory factor 4 Homo sapiens
- NFE2 nuclear factor (erythroid-derived 2), 45kDa Homo sapiens
- NR3C1 nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor) Homo sapiens
- SUZ12 suppressor of zeste 12 homolog (Drosophila) Homo sapiens

Functional Analysis of TF Targets

Categories of TF Coding Gene Targets	Most Significant Functional Categories	P-Value
Targets of only top level TFs	regulation of transcription	4.60E-19
	transcription	1.50E-14
	positive regulation of transcription	9.10E-6
	limb development	7.00E-7
	cell morphogenesis involved in differentiation	2.60E-5
	cellular component morphogenesis	1.20E-3
	regulation of nervous system development	1.60E-2
	regulation of cell development	6.50E-2
	regulation of RNA metabolic process	4.00E-30
	regulation of mRNA metabolic process	4.00E-30
	positive regulation of mRNA metabolic process	3.70E-5
	positive regulation of biosynthetic process	3.70E-5
	positive regulation of macromolecule metabolic process	9.40E-3
	cell migration	2.00E-3
	cell motility	4.00E-3
	intrinsic to membrane	4.10E-5
	integral to membrane	6.00E-4
Targets of only middle level TFs	**Non-Significant**	Non-Significant
Targets of only bottom level TFs	integral to membrane	7.40E-2
	intrinsic to membrane	1.50E-1

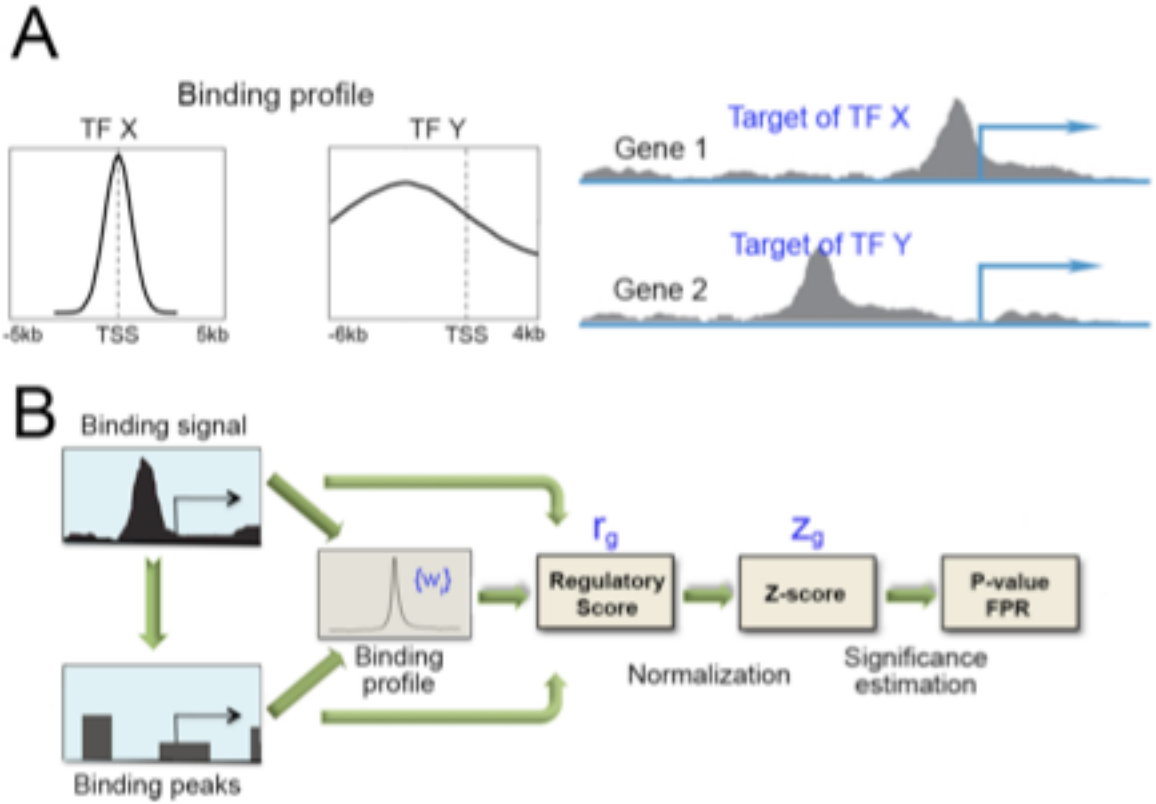
Functional Analysis of TF Targets (2)

Common targets of top and middle level TFs	translational elongation	7.50E-8
	cytosolic ribosome	7.70E-8
	ribosome	3.40E-7
	organelle lumen	2.50E-7
	membrane-enclosed lumen	2.90E-7
	intracellular organelle lumen	9.10E-7
	nuclear lumen	5.50E-6
	nucleosome	4.60E-7
	nucleosome organization	7.30E-4
	nucleosome assembly	3.10E-3
	chromatin assembly or disassembly	1.20E-2
	DNA packaging	1.50E-2
Common targets of	DNA packaging	7.50E-6
	nucleosome	7.50E-6
	nucleosome assembly	1.70E-4
	nucleosome organization	2.00E-4
middle and bottom level TFs	chromatin assembly or disassembly	3.10E-4
	ribosome	2.70E-2
	nucleotide binding	4.80E-2
	membrane-enclosed lumen	2.50E-2
	intracellular organelle lumen	3.90E-2
	organelle lumen	4.30E-2
Common targets of top and bottom level TFs	cytosolic ribosome	1.60E-2
	ribosome	6.10E-2
	translational elongation	6.20E-2
	nucleosome assembly	2.70E-4
	chromatin assembly or disassembly	3.80E-4
	nucleosome organization	4.60E-4
	nucleosome	1.10E-3
	DNA packaging	8.80E-3
	nucleotide binding	1.70E-2
Common targets of top, middle, and bottom level TFs	nucleosome	4.10E-7
	nucleosome organization	1.40E-5
	nucleosome assembly	1.90E-5
	chromatin assembly or disassembly	2.10E-5
	cytosolic ribosome	2.30E-3
	ribosome	7.30E-3
	translational elongation	7.60E-3
	membrane-enclosed lumen	1.60E-2
	organelle lumen	2.00E-2
	intracellular organelle lumen	2.20E-2
	nuclear lumen	5.40E-2

Too many peaks and targets!

- We want to know whether or not TFs tend to collaborate each other during transcriptional regulation
- Can we observe any obvious dependent binding patterns from ChIP-Seq data?
- Kevin Yip did peak overlapping studies, but I found that almost every TF overlaps with any other TF, so they even used a z-score of 100 to eliminate noise (challenging)!
- Better ideas? Filtering!

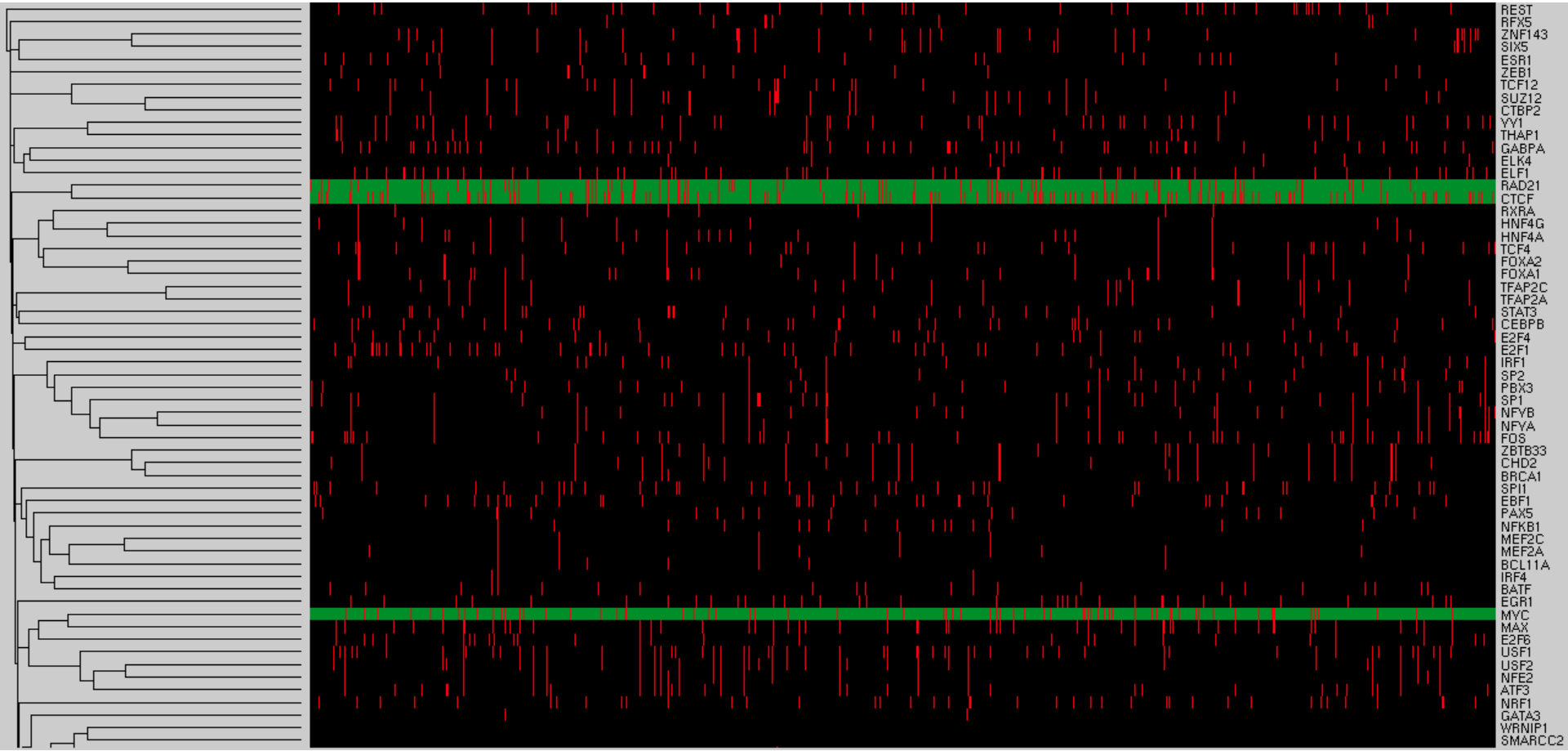
TIP (Cheng, Min, Gerstein, 2011)



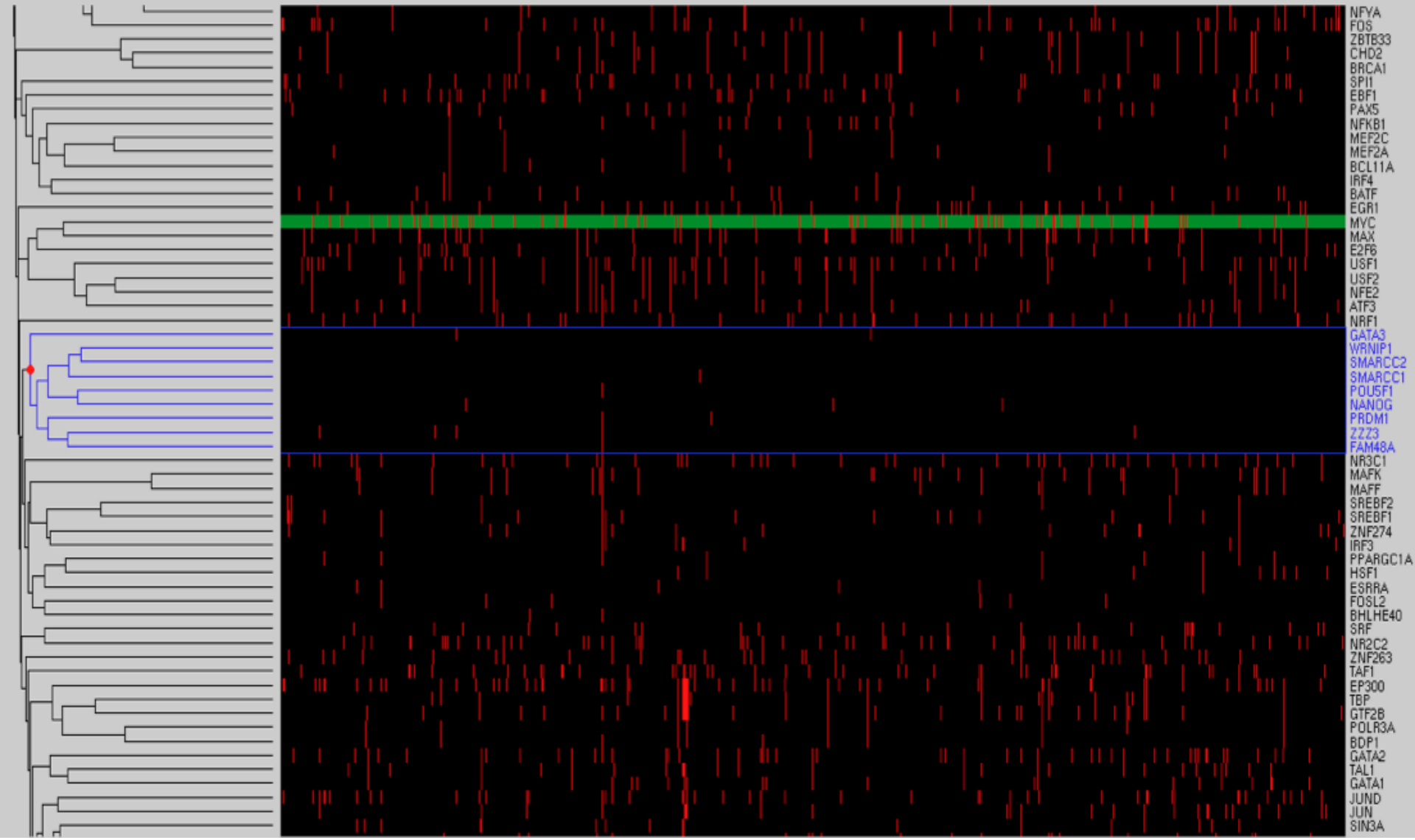
$$p(T(t,g)=1 \mid \vec{s}(t,g)) = C \sum_i w_i(t) s_i(t,g)$$

$$p(T(t,g)=1 \mid \vec{s}(t,g)) = \frac{p(T(t,g)=1, \vec{s}(t,g))}{p(\vec{s}(t,g))}$$
$$= C_1 p(T(t,g)=1, s_i(t,g) \vec{s}_{-i}(t,g))$$
$$= C_1 \Sigma_i p(T_i(t,g)=1) p(s_i(t,g) \mid T_i(t,g)=1) p(\vec{s}_{-i}(t,g))$$
$$\approx C_2 \Sigma_i p(T_i(t,g)=1) p(s_i(t,g) \mid T_i(t,g)=1)$$
$$= C_2 \Sigma_i w_i(t,g) f(s_i(t,g))$$
$$= C \Sigma_i w_i(t) s_i(t,g)$$

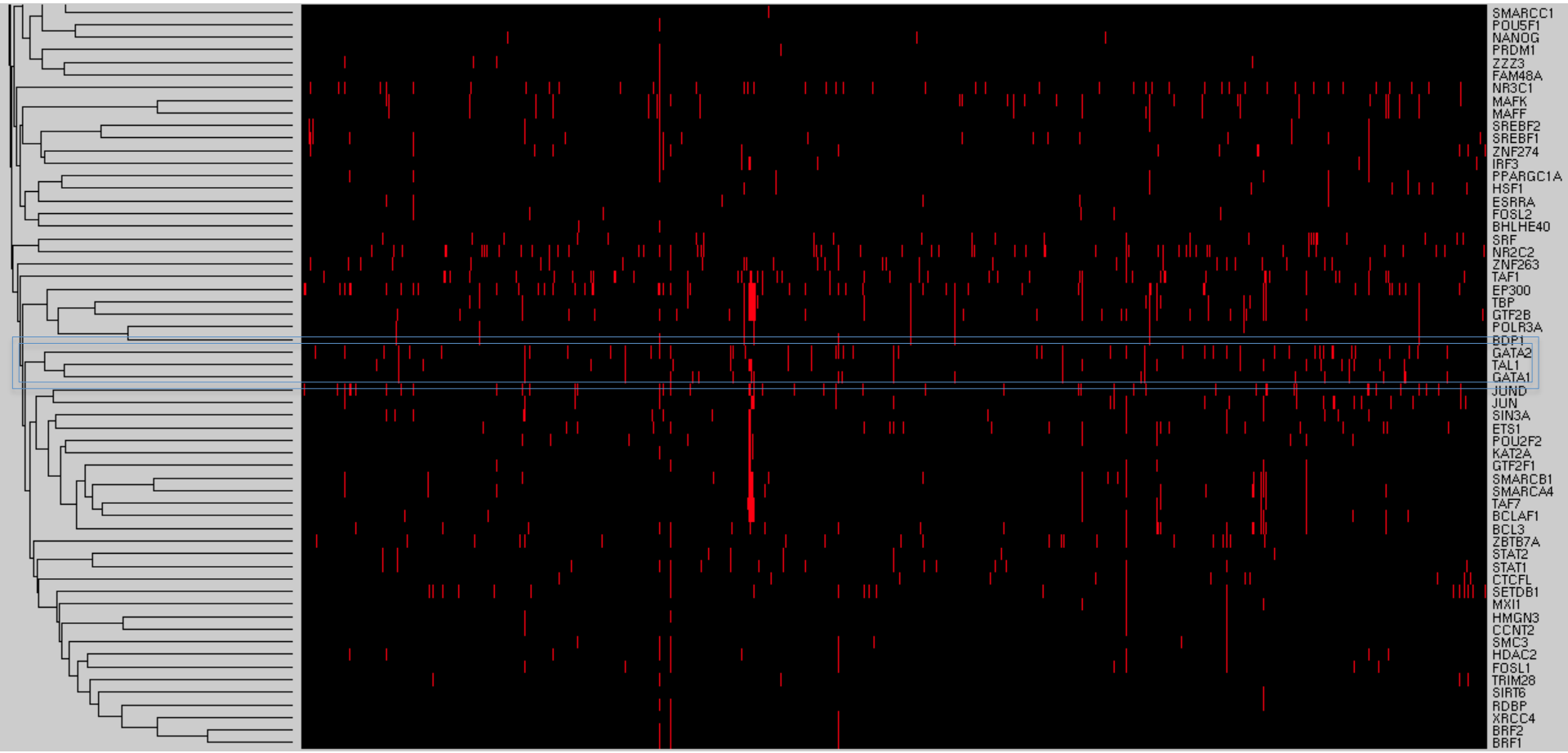
Co-association: clustering TFs (1)



Co-association: clustering TFs (2)



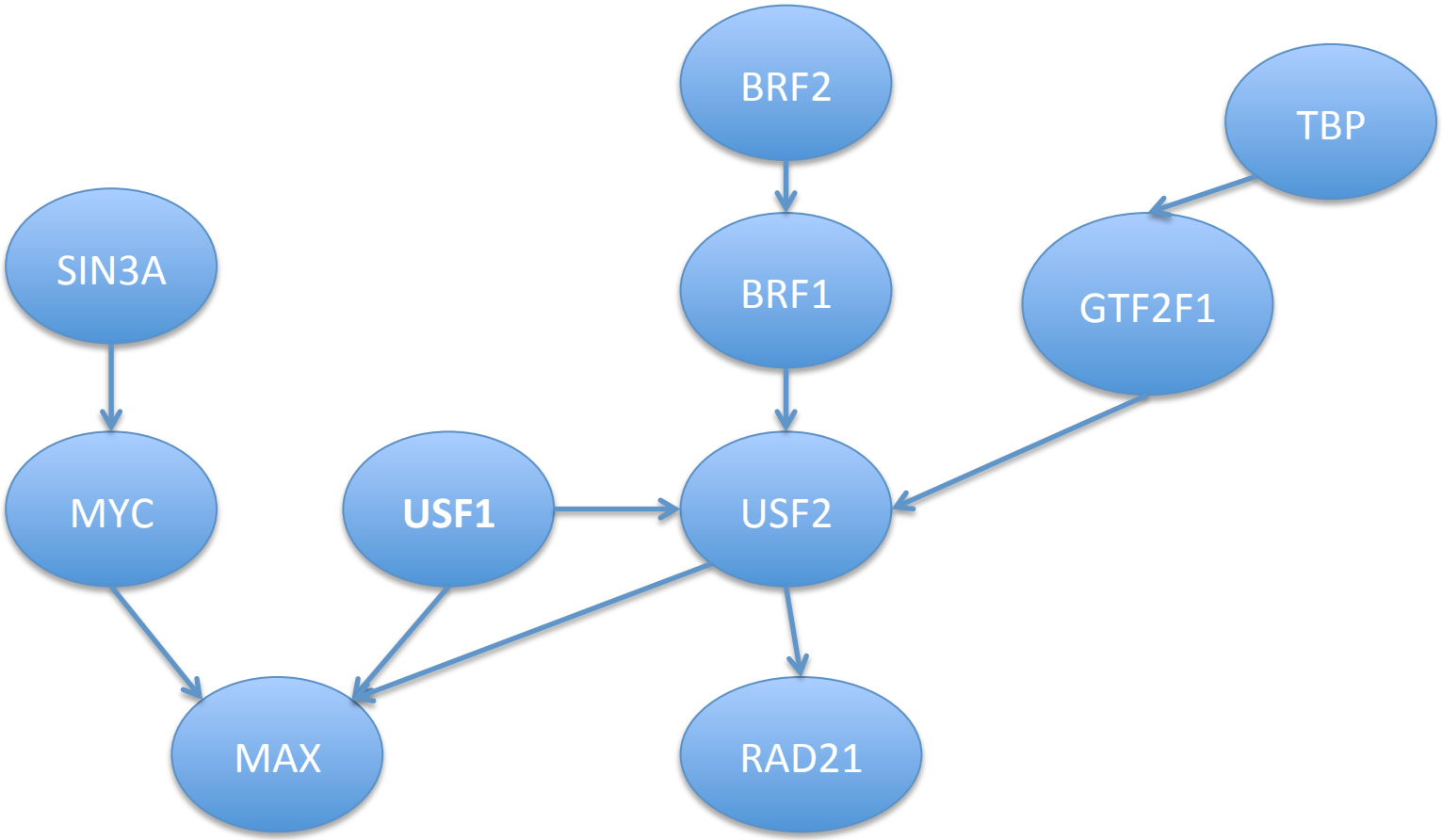
Co-association: clustering TFs (3)



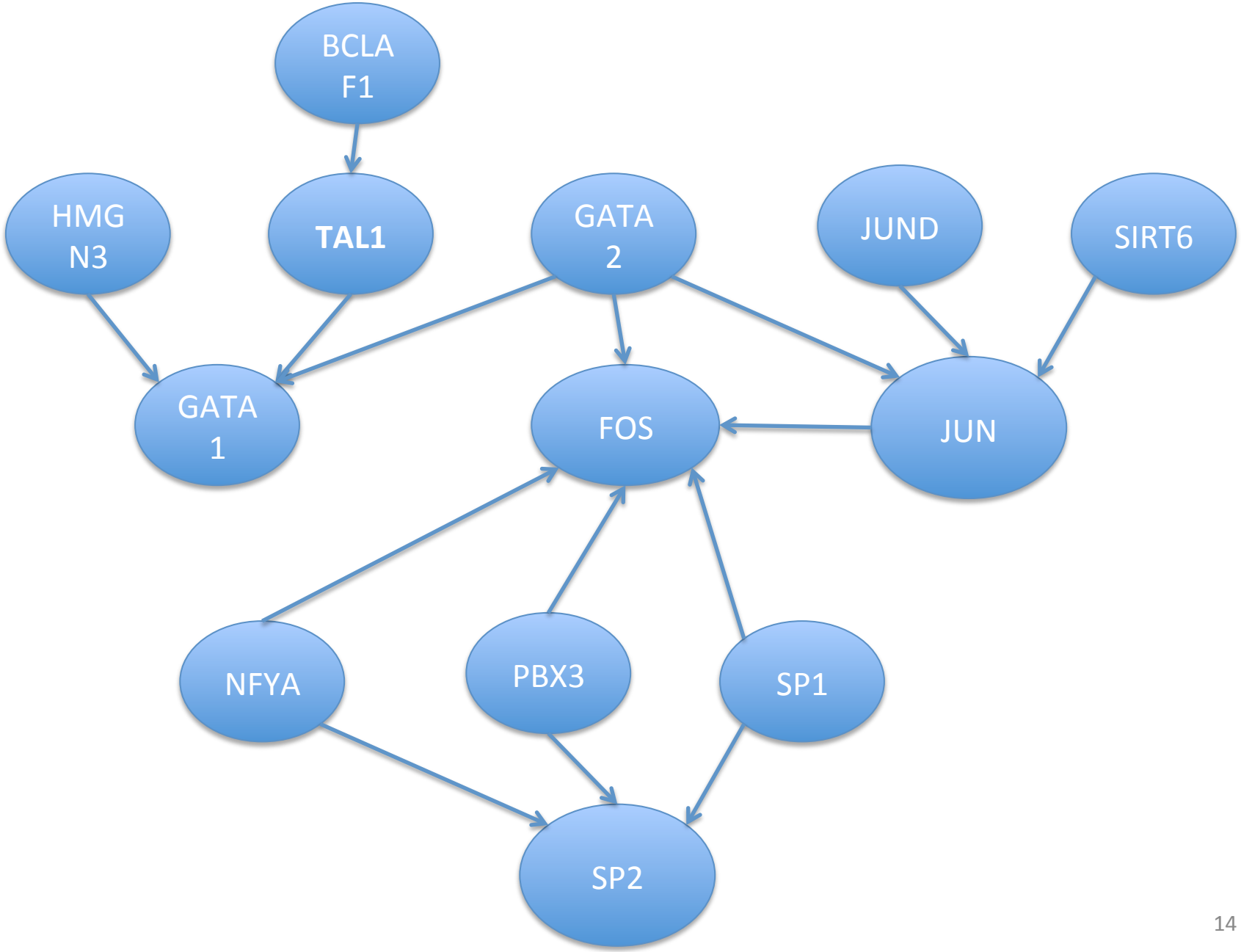
Dependencies between TF Bindings: Bayesian Networks!

- Using a graph to represent dependencies between variables, in which nodes correspond to variables and edges correspond to dependencies between variables
- Lack of edges represent conditional independency
- Main theorem: p(V) = Product_c p(V_c| parents of V_c); Conditioned on its parents, each variable is independent of all the variables except its children and descendants

Bayesian Network for TF Binding Profiles from ENCODE



Bayesian Network (2)



TFs with the largest TF Out-degrees in Bayesian Net

- SP1
- BCL11A
- BCLAF1
- EP300
- SMARCA4
- USF2
- GATA2
- HMGN3
- HNF4G
- JUND

- RXRA->SP1->[SP2 FOS IRF1 IRF3 NFYA SREBF1]
- BCL11A->[BATF NFKB1 PAX5 RXRA TCF12]
- BCLAF1->[MEF2A POU2F2 TAF7 TAL1]
- EP300 ->[JUND RXRA TCF4 ZNF143]
- SMARCA4->[BCL3 KAT2A SMARCB1 TAF7]
- [BRF1 GTF2F1 USF1]->USF2->[ATF3 MAX NFE2 RAD21]

TFs with the largest TF In-degrees in Bayesian Net

- FOS
- JUN
- MAX
- RXRA
- SP2
- TBP
- TCF12
- TCF4
- USF2

- [GATA2 JUN NFYA PBX3 SP1]->FOS
- [GATA2 JUND SIRT6]->JUN
- [MYC USF1 USF2]->MAX
- [BCL11A EP300 HNF4G]->RXRA
- [NFYA PBX3 SP1]->SP2
- [POLR3A TAF7]->TBP

Common Top-Layer TFs comparing BN to Hierarchy

- EP300
- HDAC2
- HMGN3
- HNF4G
- PBX3
- PPARGC1A
- SETDB1
- SMARCA4
- SRF
- STAT3
- SUZ12
- TFAP2C
- ZBTB33

Why TF-ncRNAs?

- Mark asked me to work on intersecting ncRNAs with indels, snps, and deletions
- I want to contribute to supporting the hierarchical network constructed and ENCODE-nets by looking at TF-ncRNA interactions, hoping to find novel results ☺

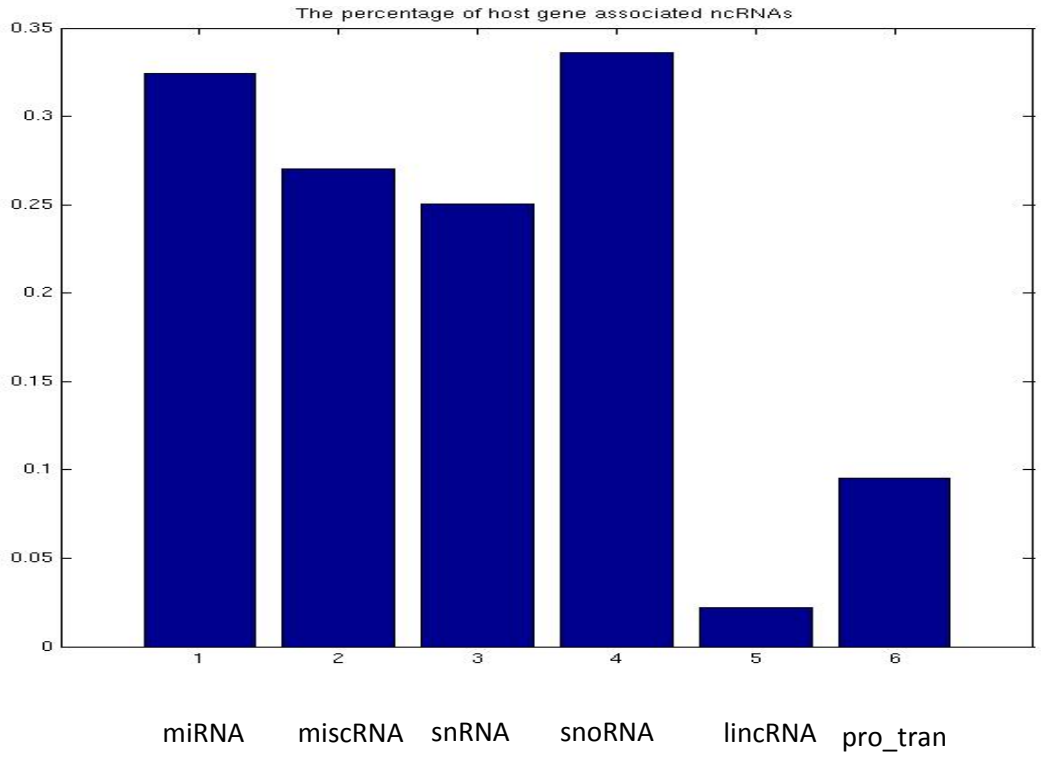
Overview of ncRNAs in ENCODE Data

- We consider the following RNAs from Gencode v7 annotation as non-coding RNAs:

	number	avg. length
-microRNA:	1756	92
-misc-RNA:	1187	153
-snRNA:	1944	107
-snoRNA:	1521	110
-lincRNA:	1239	43970 (11828, median)
-processed transcript:	8401	26346 (6372, median)

Host-Gene Associated ncRNAs

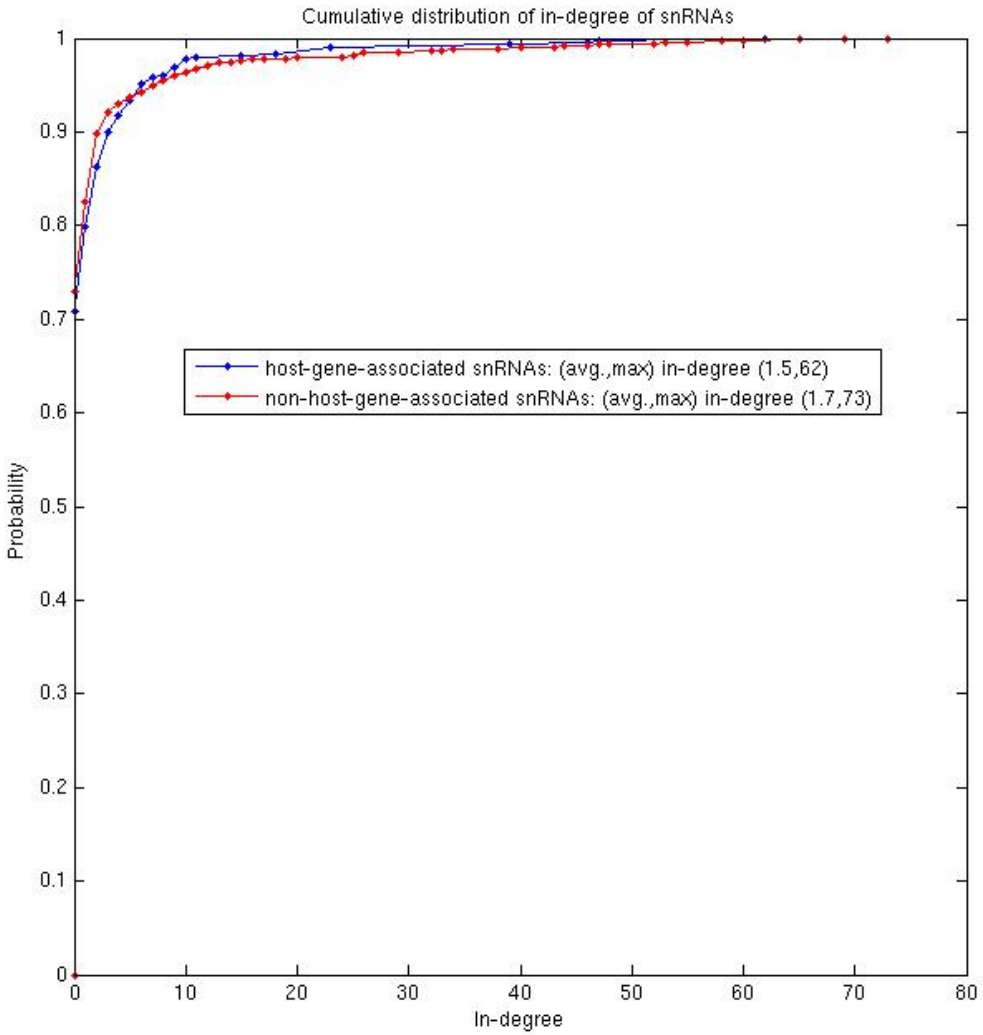
- The ncRNAs that lie in the protein coding regions



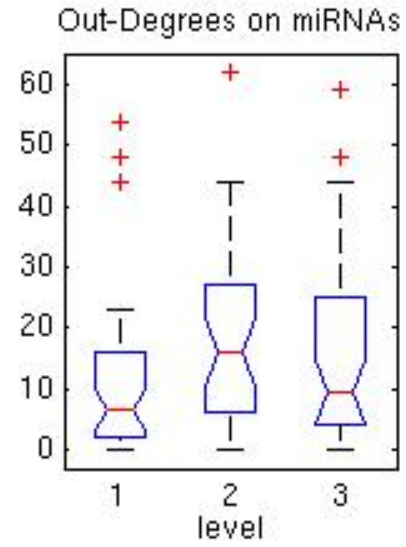
ENCODE ChIP-Seq Data for ncRNAs

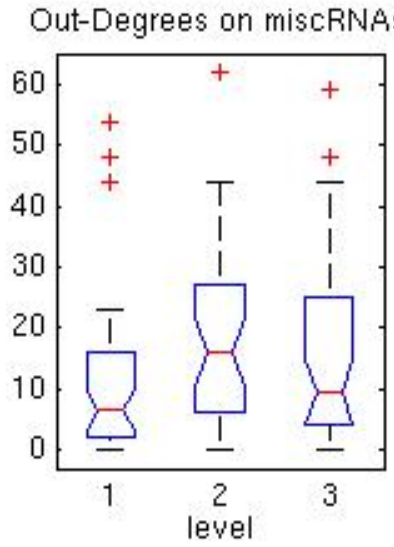
- There are around 500 ChIP-Seq experiments for about 120 unique TFs
- To identify ncRNA targets of TFs, I used 1.5 KB upstream region of the starting position as promoters of miRNAs, misc_RNAs, snRNAs, and snoRNAs
- Because lincRNAs and processed transcripts are much longer, which are comparable or even longer than coding genes, I used 1.5KB upstream and 500B downstream of the starting position as promoters of lincRNAs and processed_transcript.

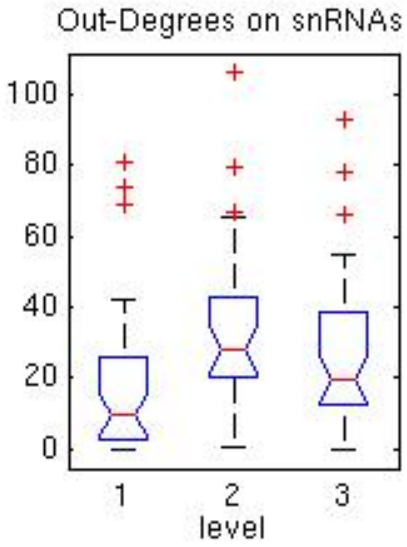
In-degree of snRNAs

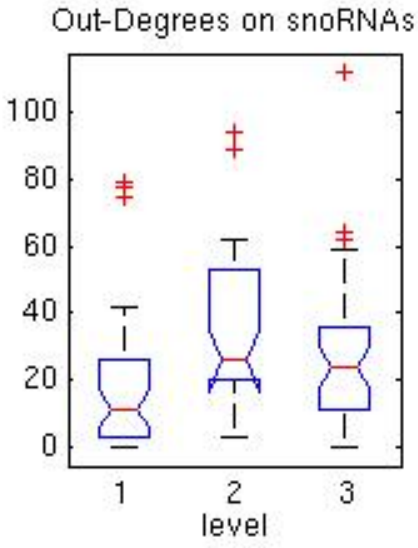


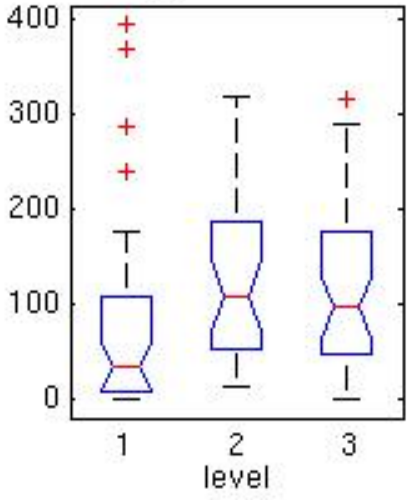
Box Plots of Out-Degrees of TFs in different levels on ncRNAs

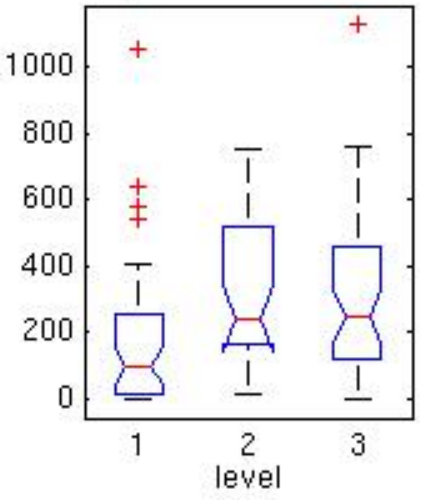




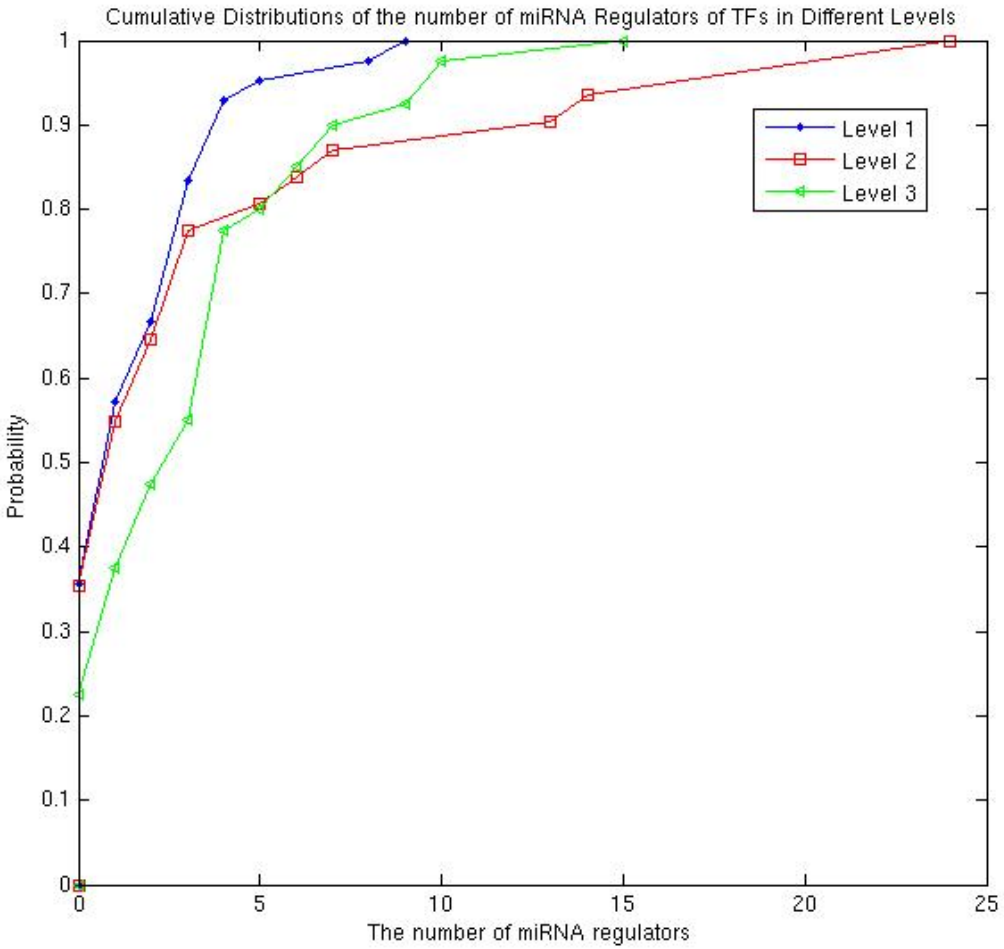




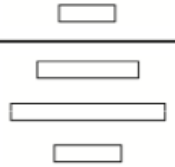


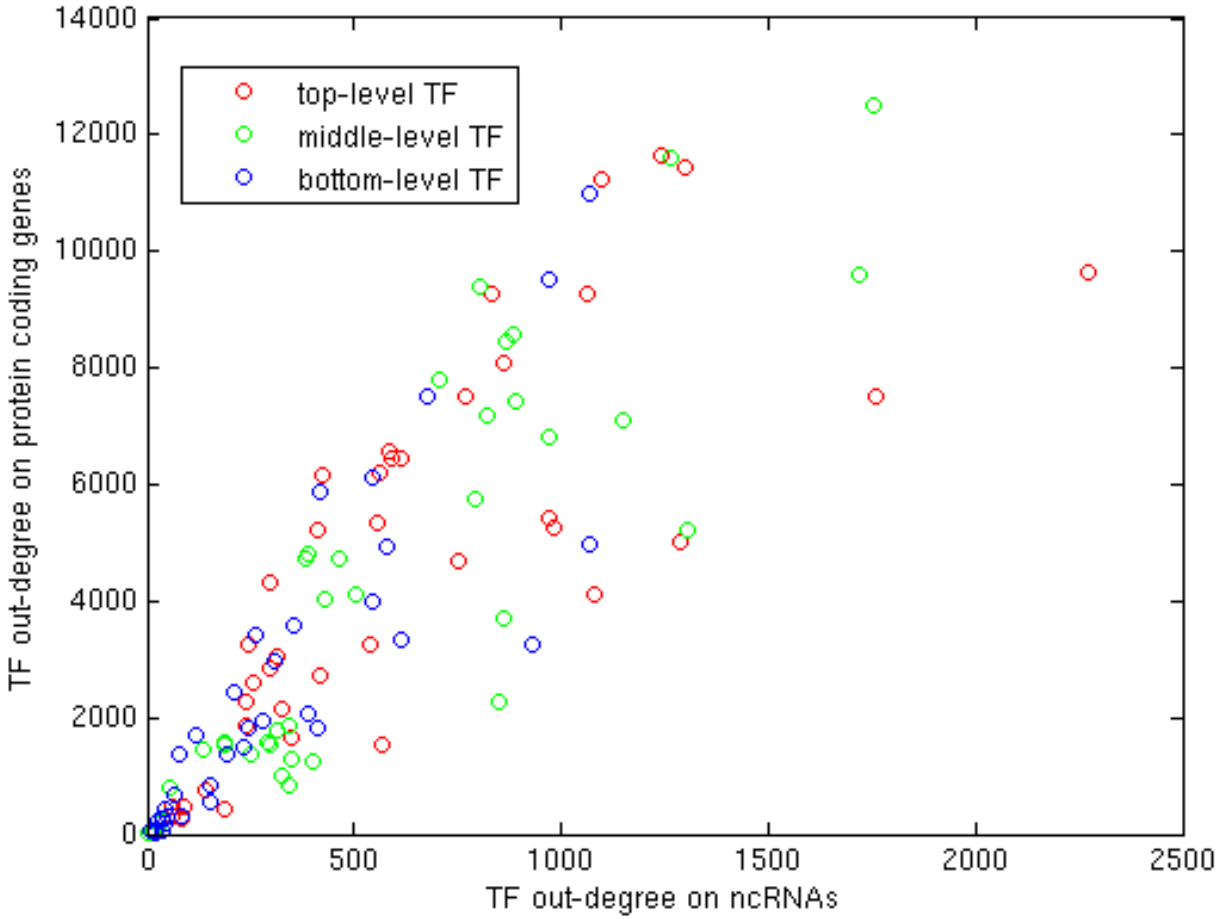


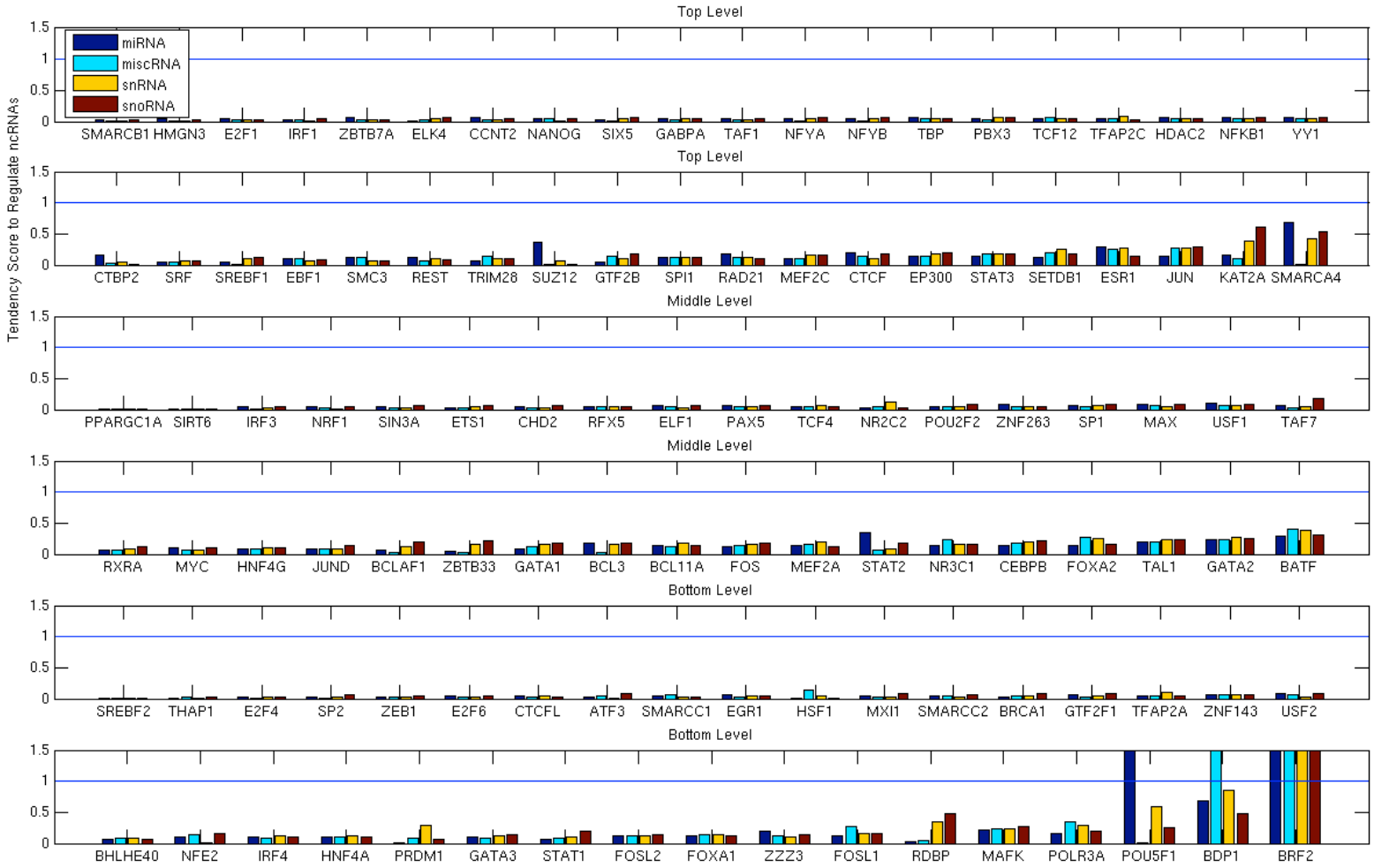
No. of miRNA Regulators

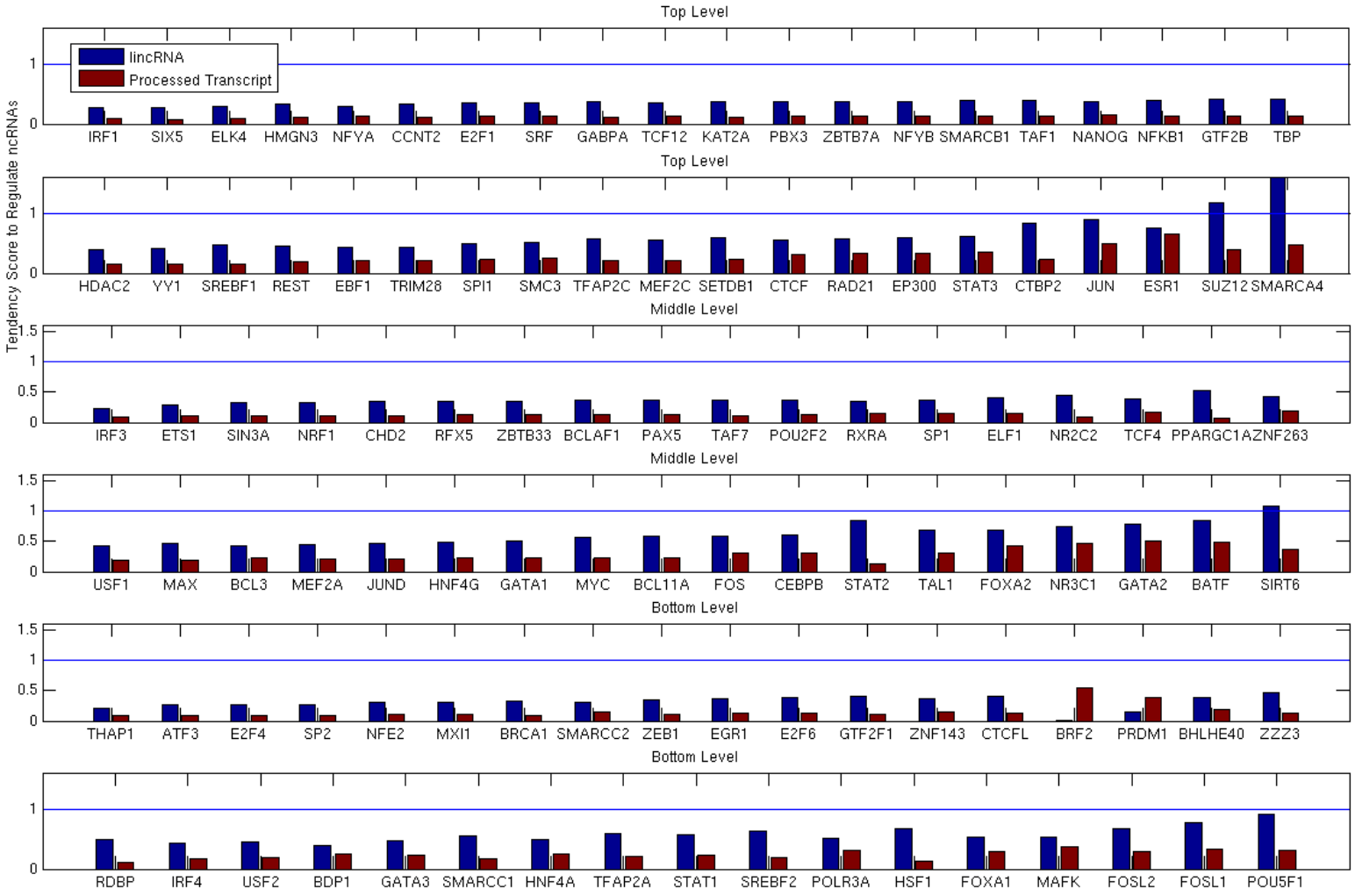


Thanks Nitin!

Topology			
	# of miRNA regulators		+0.32 
	# of ncRNA targets		+0.48 


top-level TF
middle-level TF
bottom-level TF
TF out-degree on protein coding genes
TF out-degree on ncRNAs
0
2000
4000
6000
8000
10000
12000
14000
0
500
1000
1500
2000
2500

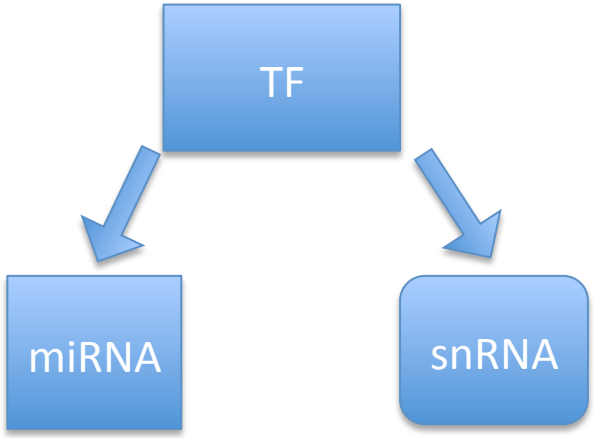

Top Level
miRNA
miscRNA
snRNA
snoRNA
SMARCB1 HMGN3 E2F1 IRF1 ZBTB7A ELK4 CCNT2 NANOG SIX5 GABPA TAF1 NFYA NFYB TBP PBX3 TCF12 TFAP2C HDAC2 NFKB1 YY1
Top Level
CTBP2 SRF SREBF1 EBF1 SMC3 REST TRIM28 SUZ12 GTF2B SPI1 RAD21 MEF2C CTCF EP300 STAT3 SETDB1 ESR1 JUN KAT2A SMARCA4
Middle Level
PPARGC1A SIRT6 IRF3 NRF1 SIN3A ETS1 CHD2 RFX5 ELF1 PAX5 TCF4 NR2C2 POU2F2 ZNF263 SP1 MAX USF1 TAF7
Middle Level
RXRA MYC HNF4G JUND BCLAF1 ZBTB33 GATA1 BCL3 BCL11A FOS MEF2A STAT2 NR3C1 CEBPB FOXA2 TAL1 GATA2 BATF
Bottom Level
SREBF2 THAP1 E2F4 SP2 ZEB1 E2F6 CTCFL ATF3 SMARCC1 EGR1 HSF1 MXI1 SMARCC2 BRCA1 GTF2F1 TFAP2A ZNF143 USF2
Bottom Level
BHLHE40 NFE2 IRF4 HNF4A PRDM1 GATA3 STAT1 FOSL2 FOXA1 ZZZ3 FOSL1 RDBP MAFK POLR3A POU5F1 BDP1 BRF2
Tendency Score to Regulate ncRNAs
0
0.5
1
1.5

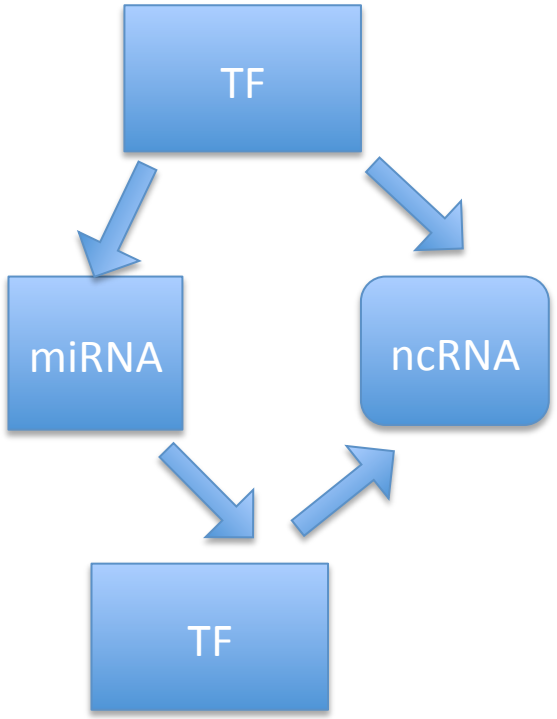

Top Level
lincRNA
Processed Transcript
Tendency Score to Regulate ncRNAs
IRF1 SIX5 ELK4 HMGN3 NFYA CCNT2 E2F1 SRF GABPA TCF12 KAT2A PBX3 ZBTB7A NFYB SMARCB1 TAF1 NANOG NFKB1 GTF2B TBP
Top Level
HDAC2 YY1 SREBF1 REST EBF1 TRIM28 SPI1 SMC3 TFAP2C MEF2C SETDB1 CTCF RAD21 EP300 STAT3 CTBP2 JUN ESR1 SUZ12 SMARCA4
Middle Level
IRF3 ETS1 SIN3A NRF1 CHD2 RFX5 ZBTB33 BCLAF1 PAX5 TAF7 POU2F2 RXRA SP1 ELF1 NR2C2 TCF4 PPARGC1A ZNF263
Middle Level
USF1 MAX BCL3 MEF2A JUND HNF4G GATA1 MYC BCL11A FOS CEBPB STAT2 TAL1 FOXA2 NR3C1 GATA2 BATF SIRT6
Bottom Level
THAP1 ATF3 E2F4 SP2 NFE2 MXI1 BRCA1 SMARCC2 ZEB1 EGR1 E2F6 GTF2F1 ZNF143 CTCFL BRF2 PRDM1 BHLHE40 ZZZ3
Bottom Level
RDBP IRF4 USF2 BDP1 GATA3 SMARCC1 HNF4A TFAP2A STAT1 SREBF2 POLR3A HSF1 FOXA1 MAFK FOSL2 FOSL1 POU5F1

Significant TFs Focusing on Regulating ncRNAs Found

- Comparing TF-ncRNA regulatory network to TF-protein coding gene regulatory network reveals known TFs and novel TFs that tend to focus on regulating different sub-classes of ncRNAs (including miRNAs).
- BDP1, BRF1, BRF2, POU5F1, ZNF274 ($P<0.067$)
- All these TFs that tend to focus on regulating short ncRNAs (including miRNAs) are from the bottom level of the hierarchy. (consistent with Pedro's beautiful figure)

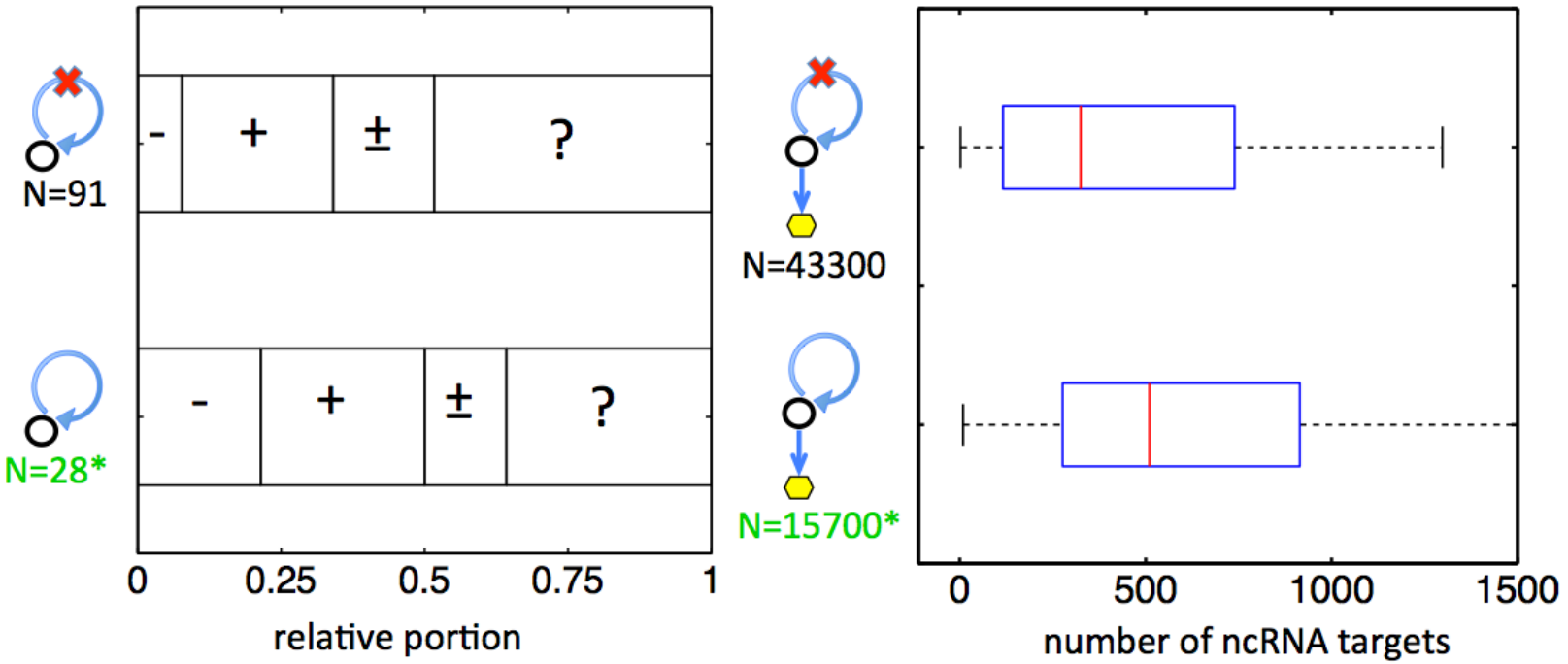
ncRNA Regulatory Network Motifs

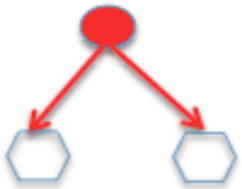




Auto-regulating TFs regulating ncRNAs are enriched

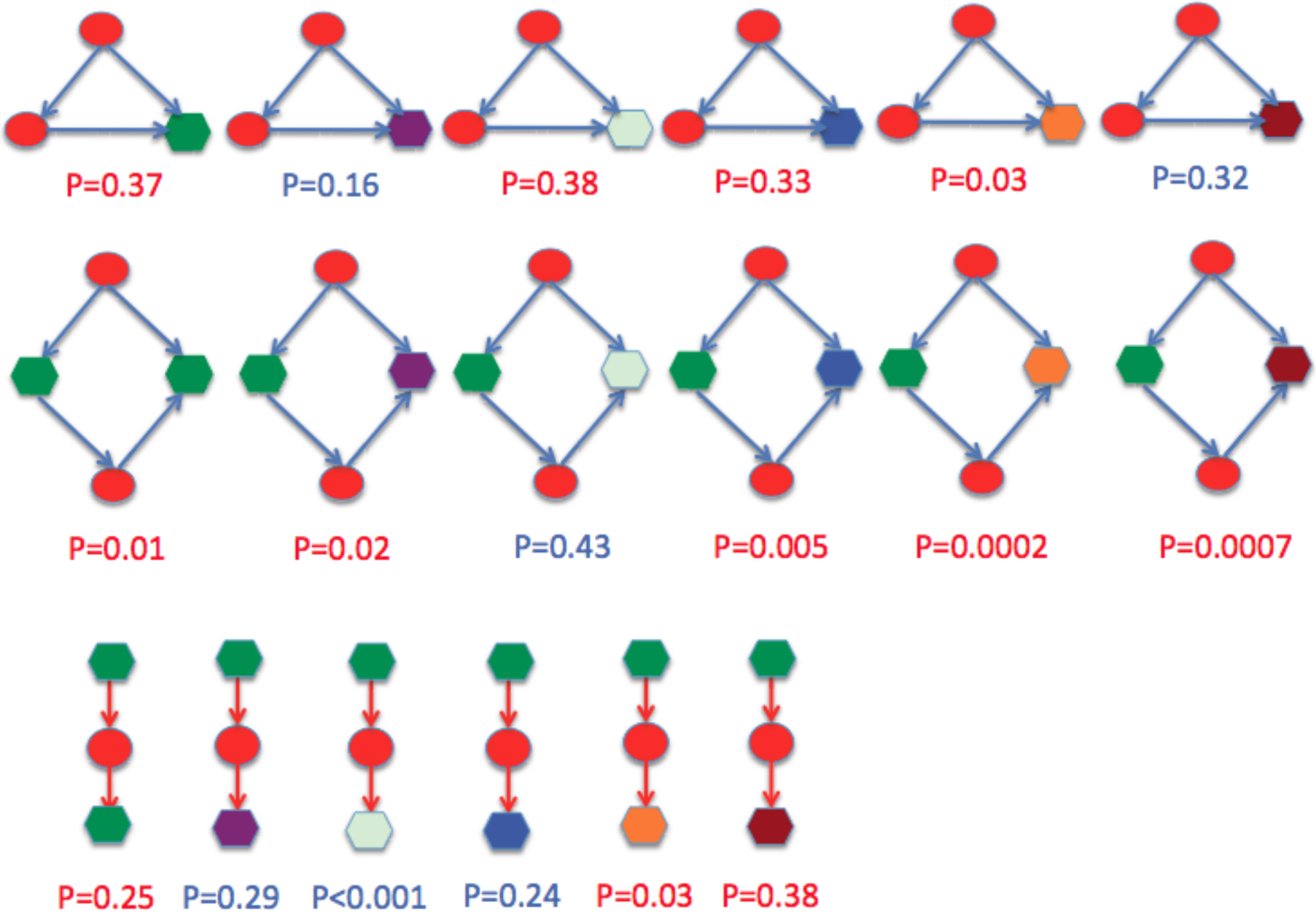
Koon-Kiu also showed his enriched auto-regulating TFs for coding genes





0.24	0.1	P<1e-8	P<1e-5	0.46	0.18
	0.24	P<0.01	P<1e-5	P<1e-14	P<0.001
		P=0.02	P<1e-8	0	0
			0.48	1e-10	P<1e-9
				0	P<1e-14
					P<1e-7

TFs
ncRNAs
miRNAs
miscRNA
snRNAs
snoRNA
lincRNAs
processed transcripts


P=0.37
P=0.16
P=0.38
P=0.33
P=0.03
P=0.32
P=0.01
P=0.02
P=0.43
P=0.005
P=0.0002
P=0.0007
P=0.25
P=0.29
P<0.001
P=0.24
P=0.03
P=0.38

Acknowledgement

Mark Gerstein
Chao Cheng
Joel Rozowsky
Koon-Kiu Yan
Pedro Alves
Roger Alexander
Baikang Pei
Arif Harmanci
Jing Leng
Xinmeng Mu

Thank You!

Questions?