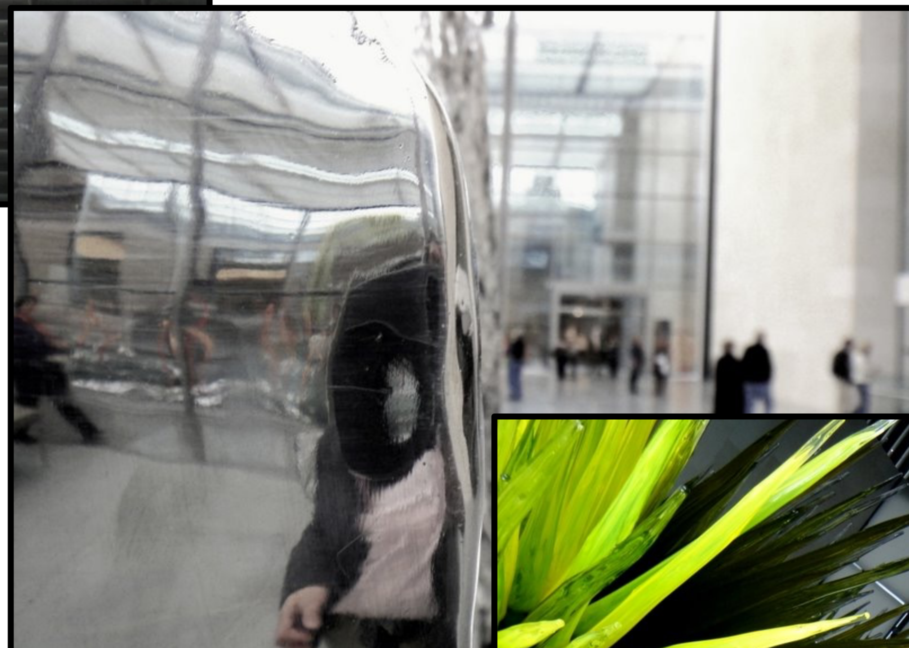
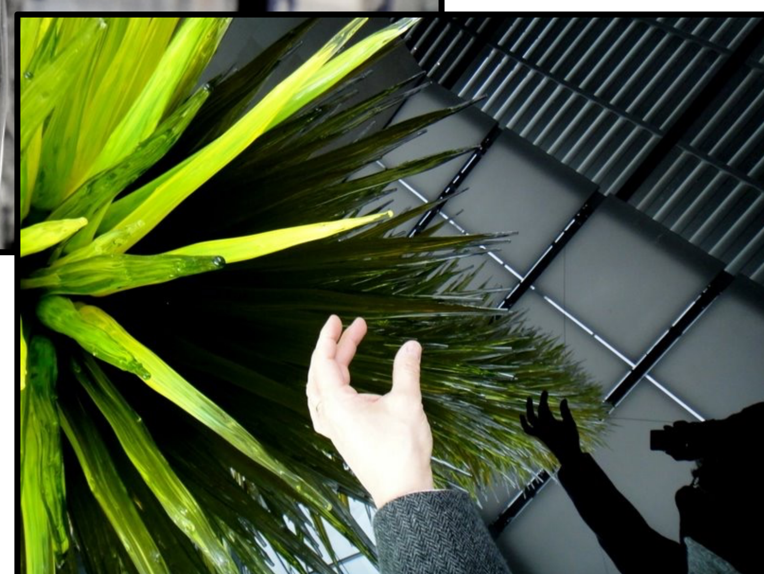
An Overview of the Data & Some of the Key Analyses
of the ENCODE & modENCODE Consortia:

Interpreting the Transcriptome in terms of the Regulome



M Gerstein, Yale

See last slide for
references & more info.



Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable” (via @markgerstein).

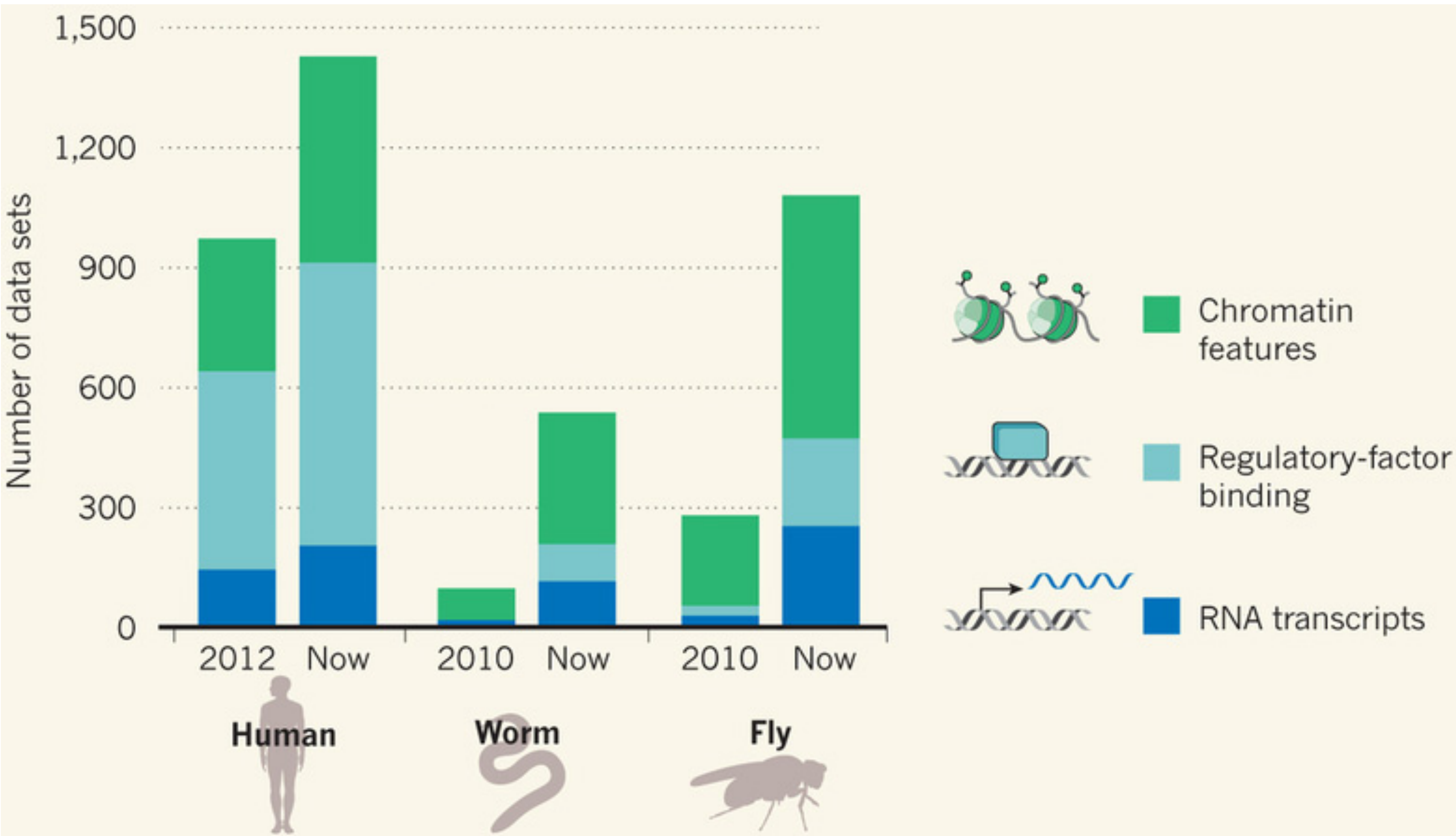
Trying to interpret RNA-seq in terms of Regulation

- Large amount of transcriptome (RNA-seq) data generated on diverse systems & in diverse conditions
- Less but still considerable amount of regulatory network & chromatin structure data available, mostly on canonical human & model organism systems
- One goal is to interpret the RNA-seq in light of frameworks provided by the regulatory data

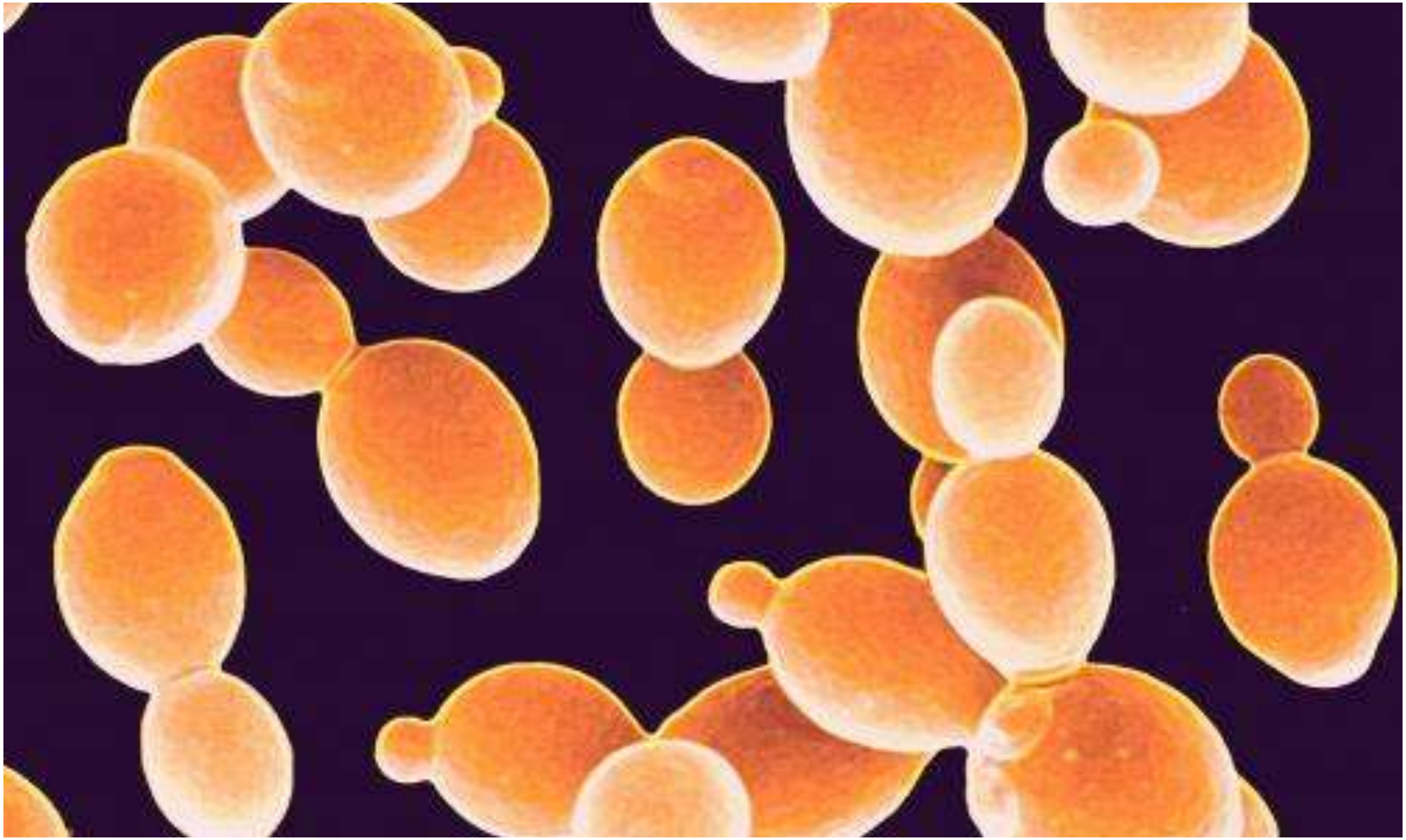
Comparative ENCODE Functional Genomics Resource

(EncodeProject.org/modENCODE.org)

- Broad sampling of conditions across transcriptomes & regulomes for human, worm & fly
  - embryo & ES cells
  - developmental time course (worm-fly)
- In total: ~3000 datasets (~130B reads)



Also Large Amount of Yeast Functional Genomics Data



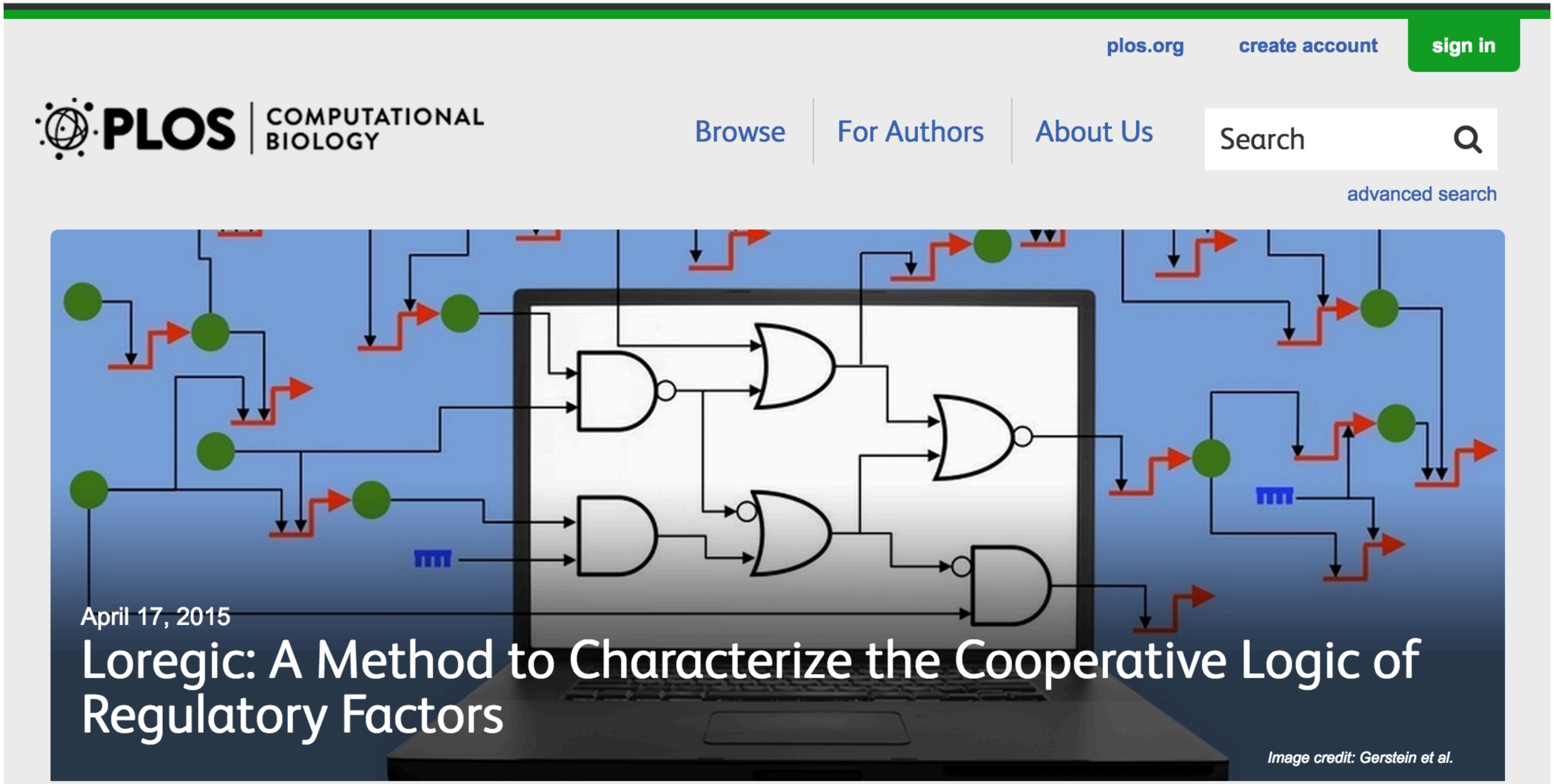
Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models** [if time]
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

Loregic: A method to characterize the cooperative logic of regulatory factors



Wang D, Yan K-K, Sisu C, Cheng C, Rozowsky J, Meyerson W, Gerstein M (2015), PLoS Comput Biol 11(4): e1004132. doi: 10.1371/journal.pcbi.1004132

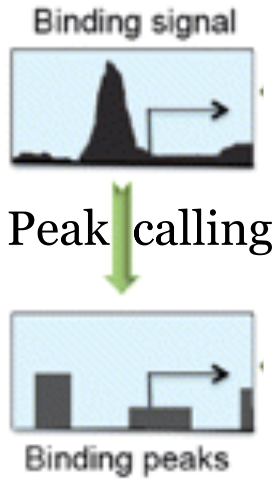
General-purpose tool, R package: github.com/gersteinlab/loregic

A gene can be regulated by multiple gene regulatory factors

Next generation sequencing techniques (e.g., ChIP-seq, CLIP-seq) predict **gene regulatory factors (RFs)** and their target genes

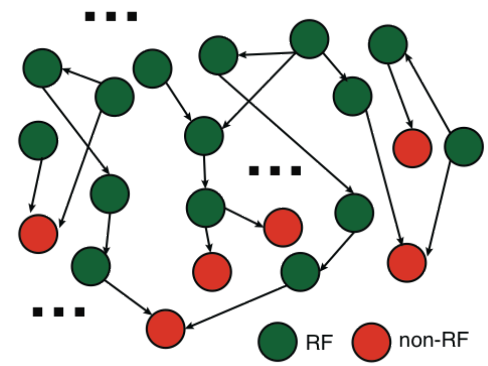
- transcription factors (TFs)
- micro-RNAs

…



Gene regulatory network

Regulatory Factor (RF)	Target (T)
TF 1	Gene 1
TF 2	Gene 1
TF 3	Gene 2
miRNA 1	Gene 1
miRNA 2	Gene 3
miRNA 3	Gene 2
…	…



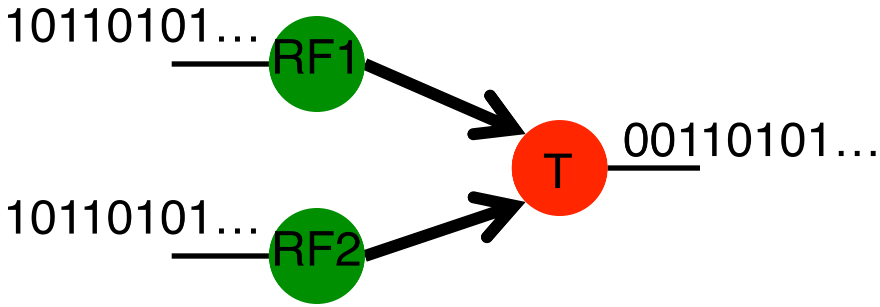
Many genes are regulated by multiple RFs.
How RFs coordinate to regulate target gene expression?

- cooperative?
- competitive?
- independent?

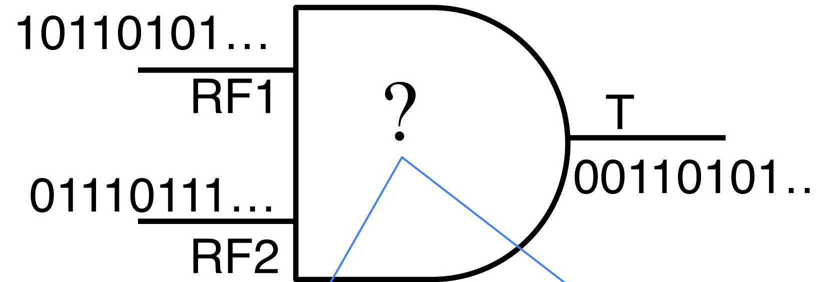
…

Modeling cooperativity between RFs to target gene using logic gates

A regulatory triplet



2-input-1-output logic gate

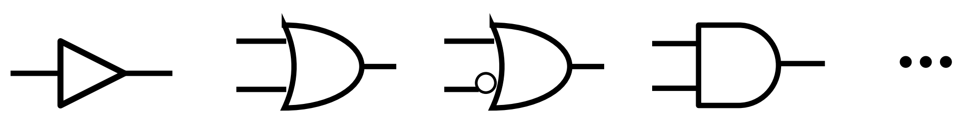


0 – gene off
1 – gene on
after binarizing gene expression data*

Input type (RF1, RF2)	RF1	0	0	1	1
	RF2	0	1	0	1
Output	T	*X*	*X*	*X*	*X*

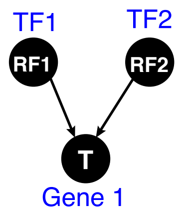
Binarized expression

X can be 0 or 1, so there are $2^4=16$ possible output combinations, each of which corresponds to a unique 2-input-1-output logic gate

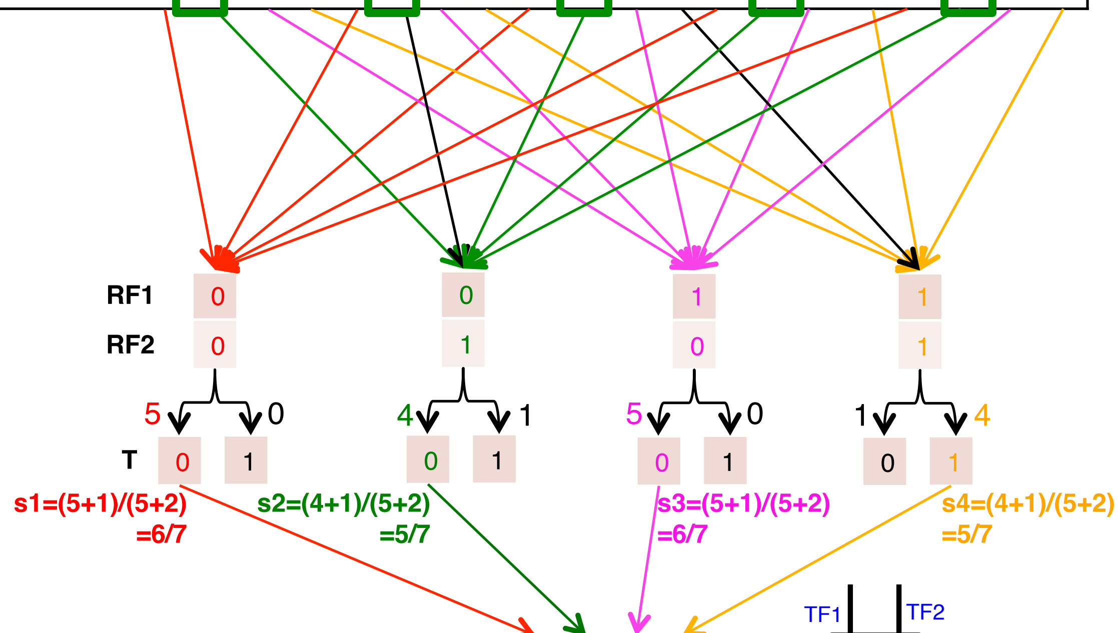


*BoolNet, R package

An example: selection of the best-matched logic gate



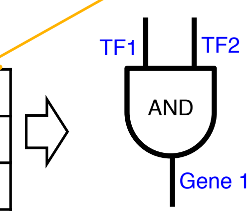
Gene	20 samples																			
RF1=TF 1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
RF2=TF 2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
T=Gene 1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1



Laplace's rule of succession
s=(# of selected output state for the input type + 1)
/(# of input type + 2)

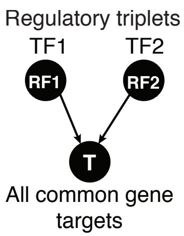
Consistency score:
6/7*5/7*6/7*5/7=0.37

RF1=TF 1	0	0	1	1
RF2=TF 2	0	1	0	1
T=Gene 1	0	0	0	1



Wang, et al., *PLoS Computational Biology*, 2015

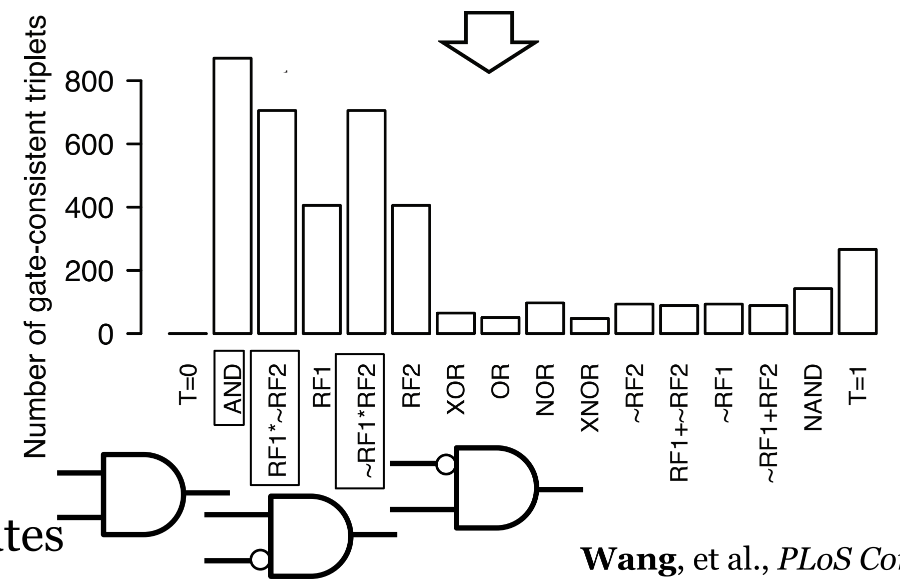
Application 1 – transcription factor cooperativity in Yeast cell cycle



Target gene	2464
TF	176
Triplet	39,011
Time point	59

Yeast Cell Cycle

Triplet ID	RF1	RF2	Common Target Gene (T)	Matched logic gate
1	YHR084W	YBR083W	YBR082C	AND
2	YKL112W	YIL131C	YMR198W	OR
…	…	…	…	…
39011	YOR113W	YBL103C	YDR042C	XOR



AND-like gates

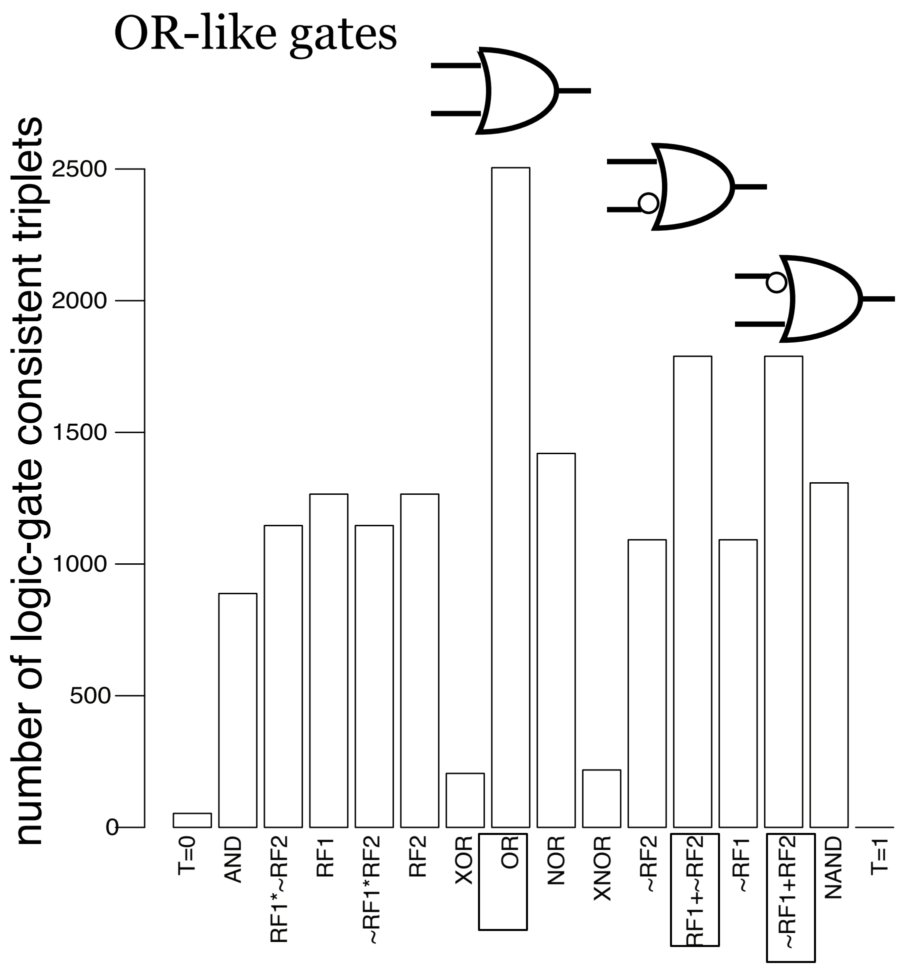
Wang, et al., *PLoS Computational Biology*, 2015

Application 2 – transcription factor cooperativity in Acute Myeloid Leukemia (AML)

Target gene	1824	ENCODE Data (K562, ChIP-seq) http://encodenets.gersteinlab.org/
TF	70	
Regulatory triplet	50,865	TCGA Data (AML, level 3, RNA-seq) https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp
Patient sample	197	

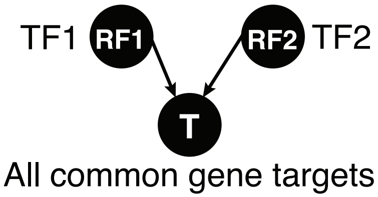
Wang, et al., *PLoS Computational Biology*, 2015

Application 2 – transcription factor cooperativity in Acute Myeloid Leukemia (AML)



Human TF-TF-target

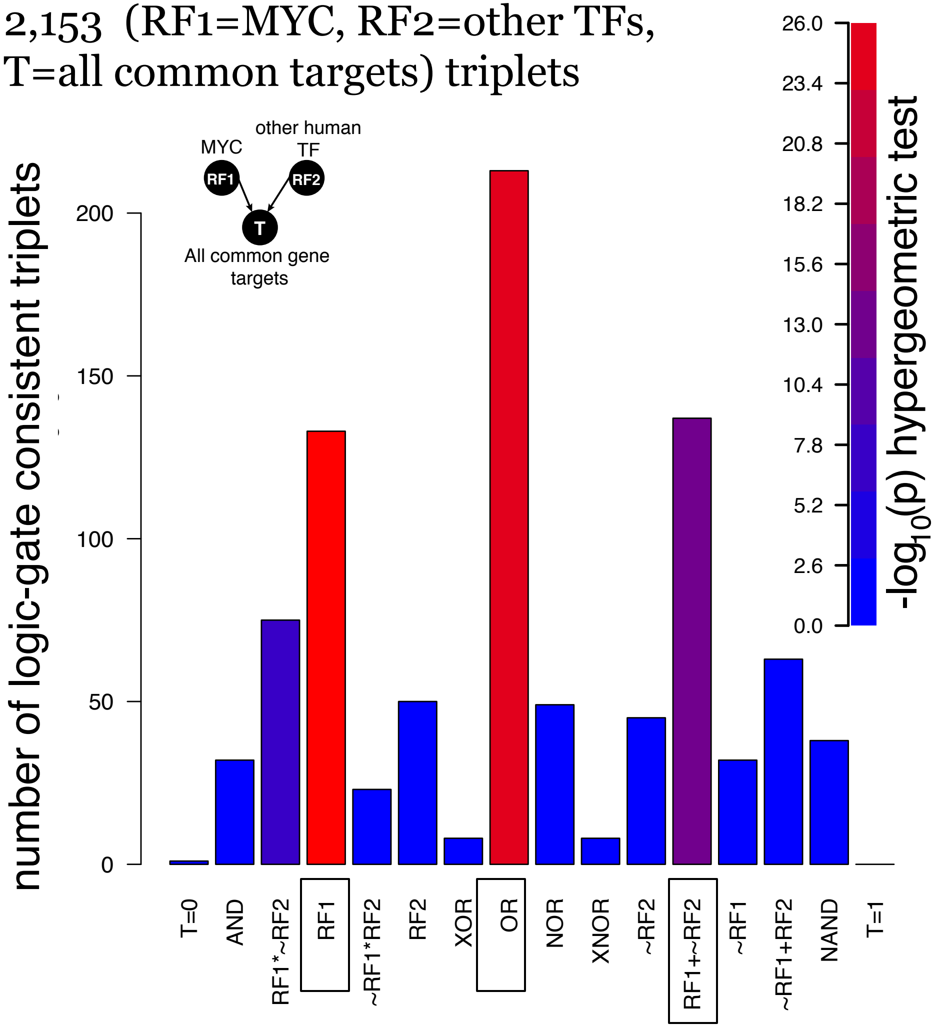
RF1	RF2	Common Target Gene (T)	Matched logic gate
ATF3	BDP1	YPEL1	AND
MYC	BCL3	BCR	T=RF1
ATF3	BRF2	AIF1L	AND
…	…	…	…

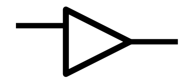
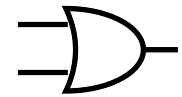
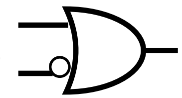


Wang, et al., *PLoS Computational Biology*, 2015

Cancer-related TF, MYC universally amplifies target expression

2,153 (RF1=MYC, RF2=other TFs, T=all common targets) triplets



- RF1 
- **OR**(RF1, RF2) 
- **OR**(RF1, **NOT** RF2) 

High expression of MYC is sufficient for high target gene expression

c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells

Cell

Zuqin Nie,[1,6] Gangqing Hu,[2,6] Gang Wei,[2] Kairong Cui,[2] Arito Yamane,[3] Wolfgang Resch,[3] Ruoning Wang,[4] Douglas R. Green,[4] Lino Tessarollo,[5] Rafael Casellas,[3] Keji Zhao,[2,*] and David Levens[1,*]

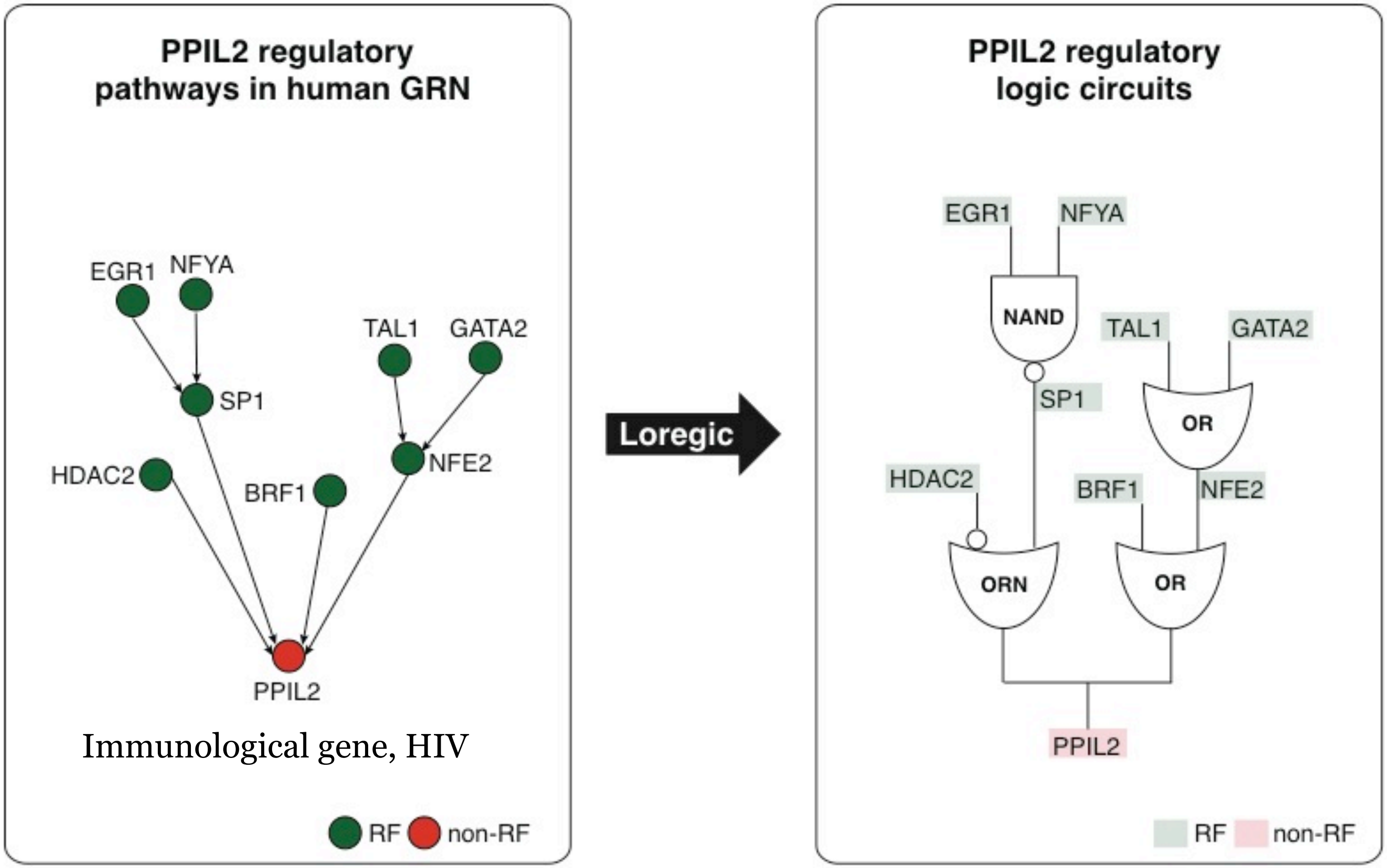
Wang, et al., *PLoS Computational Biology*, 2015

An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:

Interpreting the Transcriptome in terms of the Regulome

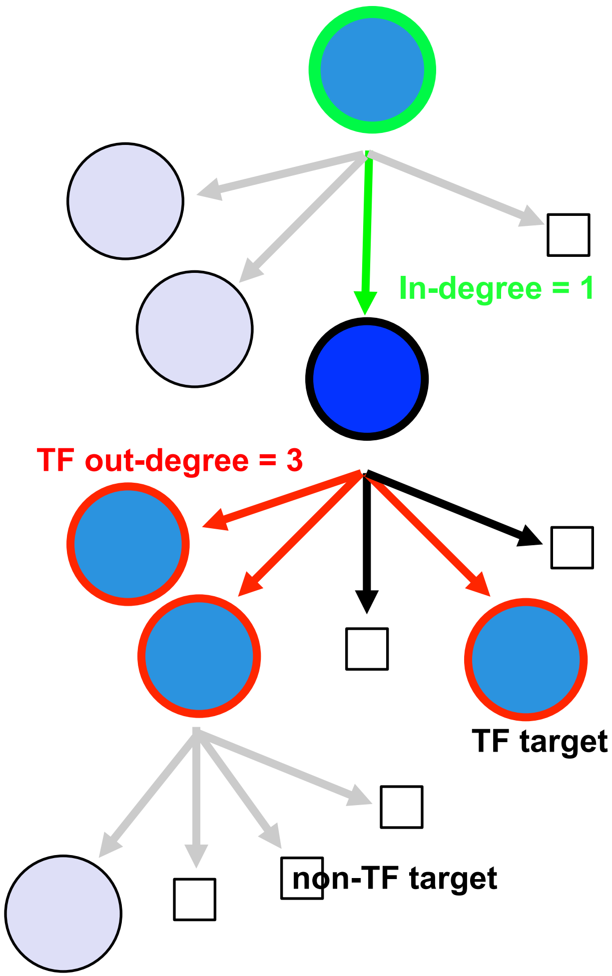
- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

Gene regulatory pathways have logic-circuit behaviors

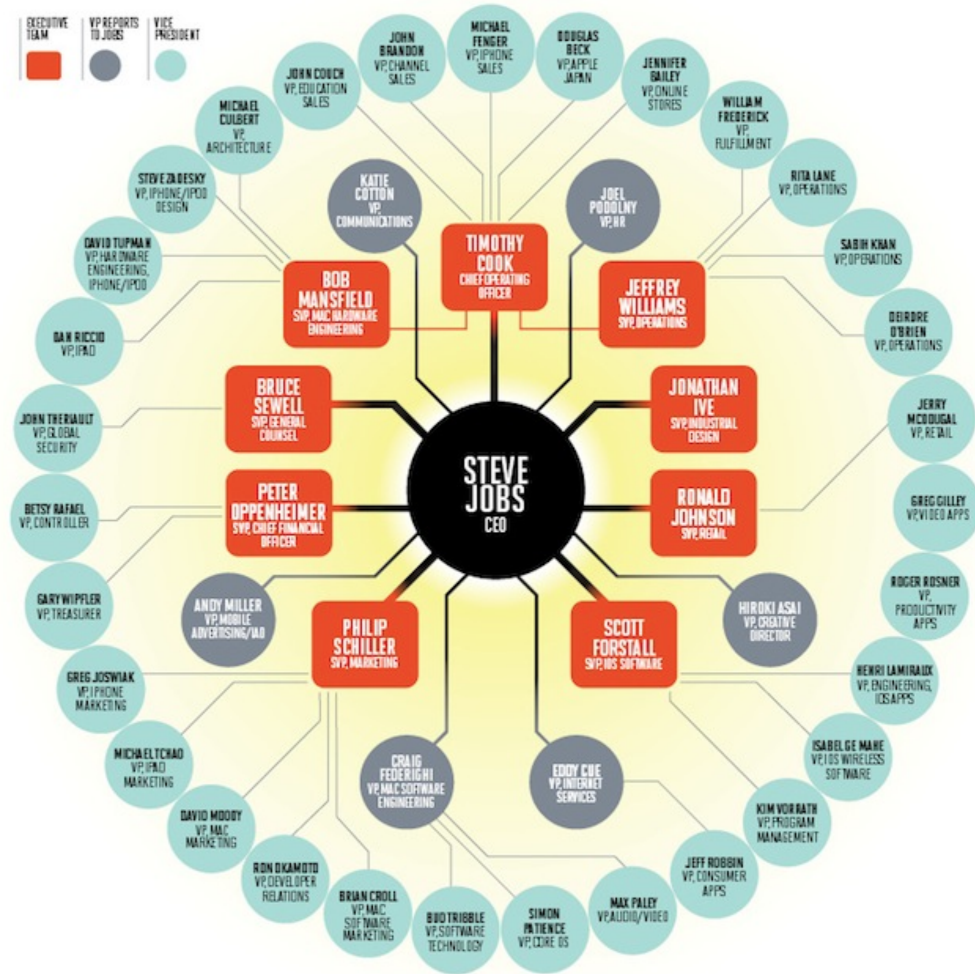


Wang, et al., *PLoS Computational Biology*, 2015

Global Hierarchical Structure



Hierarchy Height Statistic = (normalized TF Out deg. – In deg.)

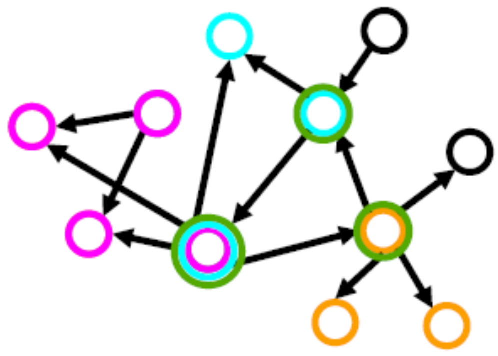


Algorithms for Defining Hierarchical Structure

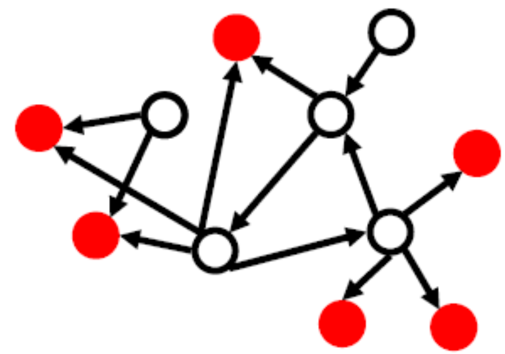
- Breadth-first Search
- Globally minimize upward edges
- Globally maximize hierarchy "score"

Breadth-first Search (Locally Optimal)

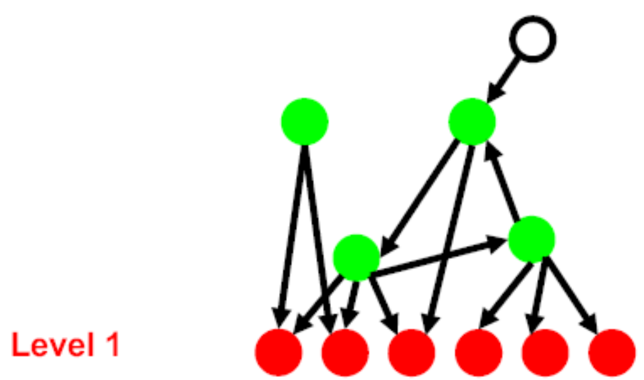
I. Example network with all 4 motifs



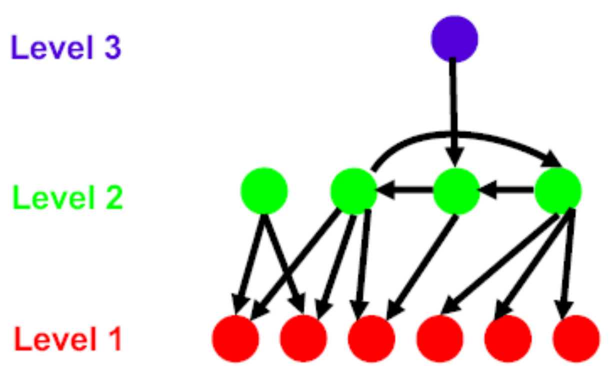
II. Finding terminal nodes (Red)



III. Finding mid-level nodes (Green)

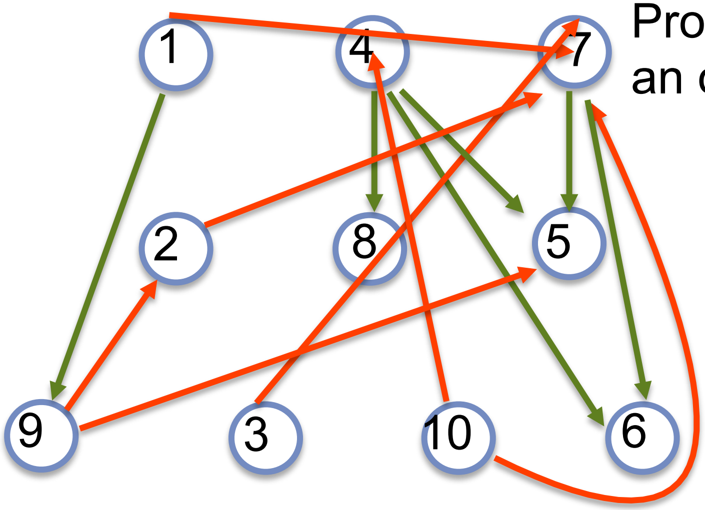


IV. Finding top-most nodes (Blue)



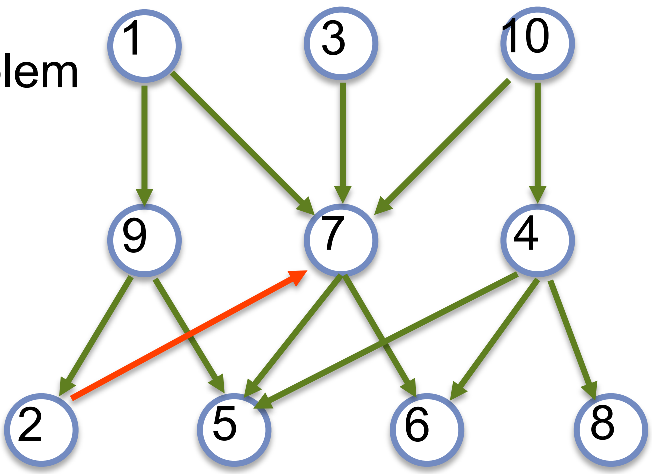
[Yu et al., PNAS (2006)]

Using Simulated Annealing to Globally Minimize the Number of Upward Pointing Edges



Probing direction is an optimization problem

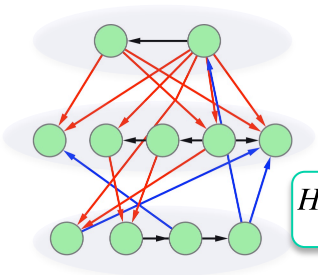
relocating nodes



[Gerstein et al. Nature ('12)]

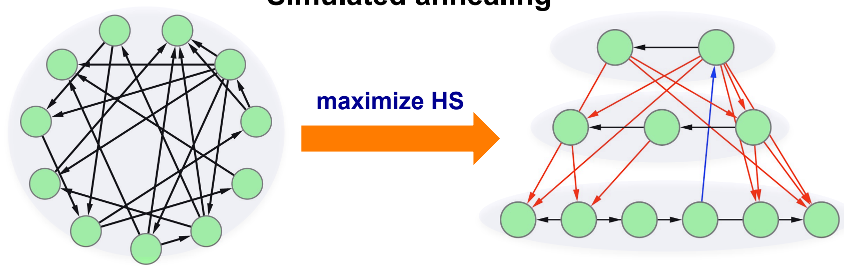
Hierarchy Score Maximization Algorithm

Definition

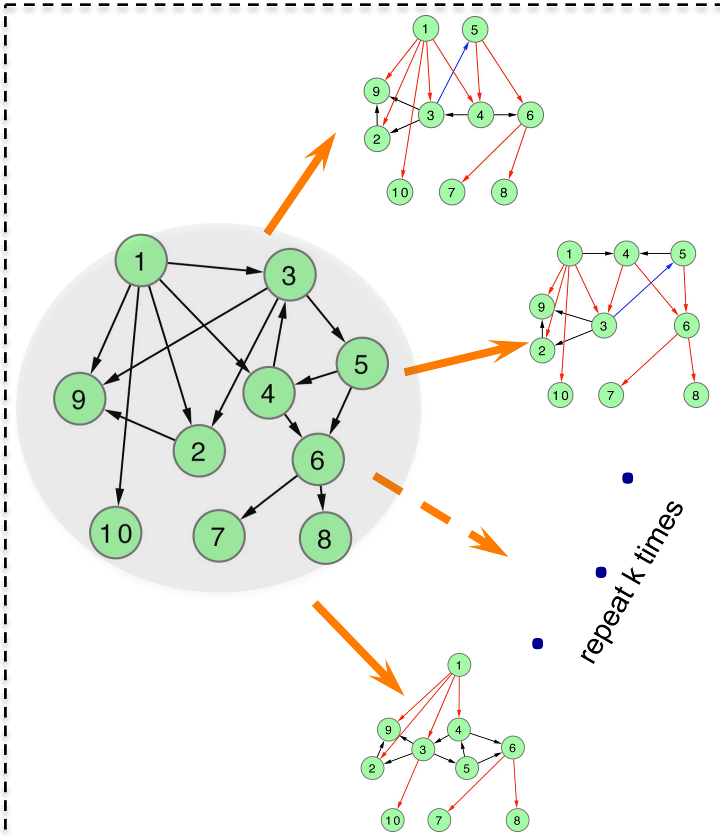


$$HS = \frac{N_d + N_h}{N_u + N_h}$$

Simulated annealing

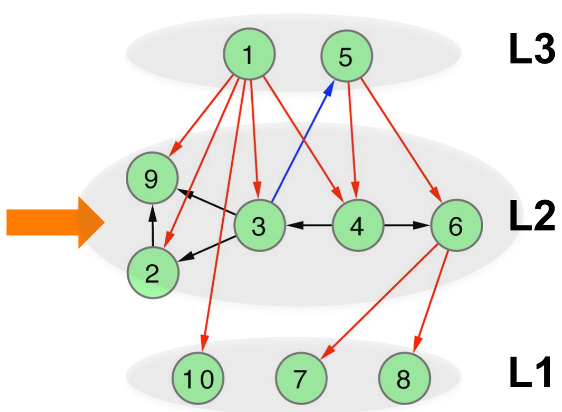


HSM algorithm



	L1	L2	L3
N1	0	0	1
N2	0.4	0.6	0
N3	0	0.84	0.16
N4	0	0.56	0.44
N5	0	0	1
N6	0	1	0
N7	0.46	0.54	0
N8	1	0	0
N9	1	0	0
N10	1	0	0

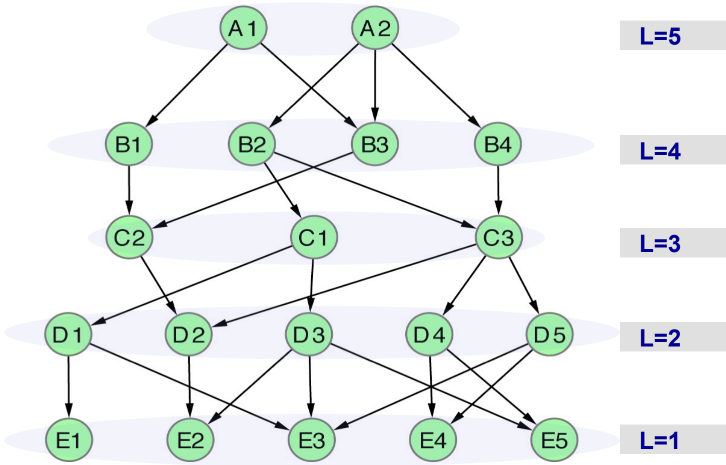
Probabilistic hierarchy network

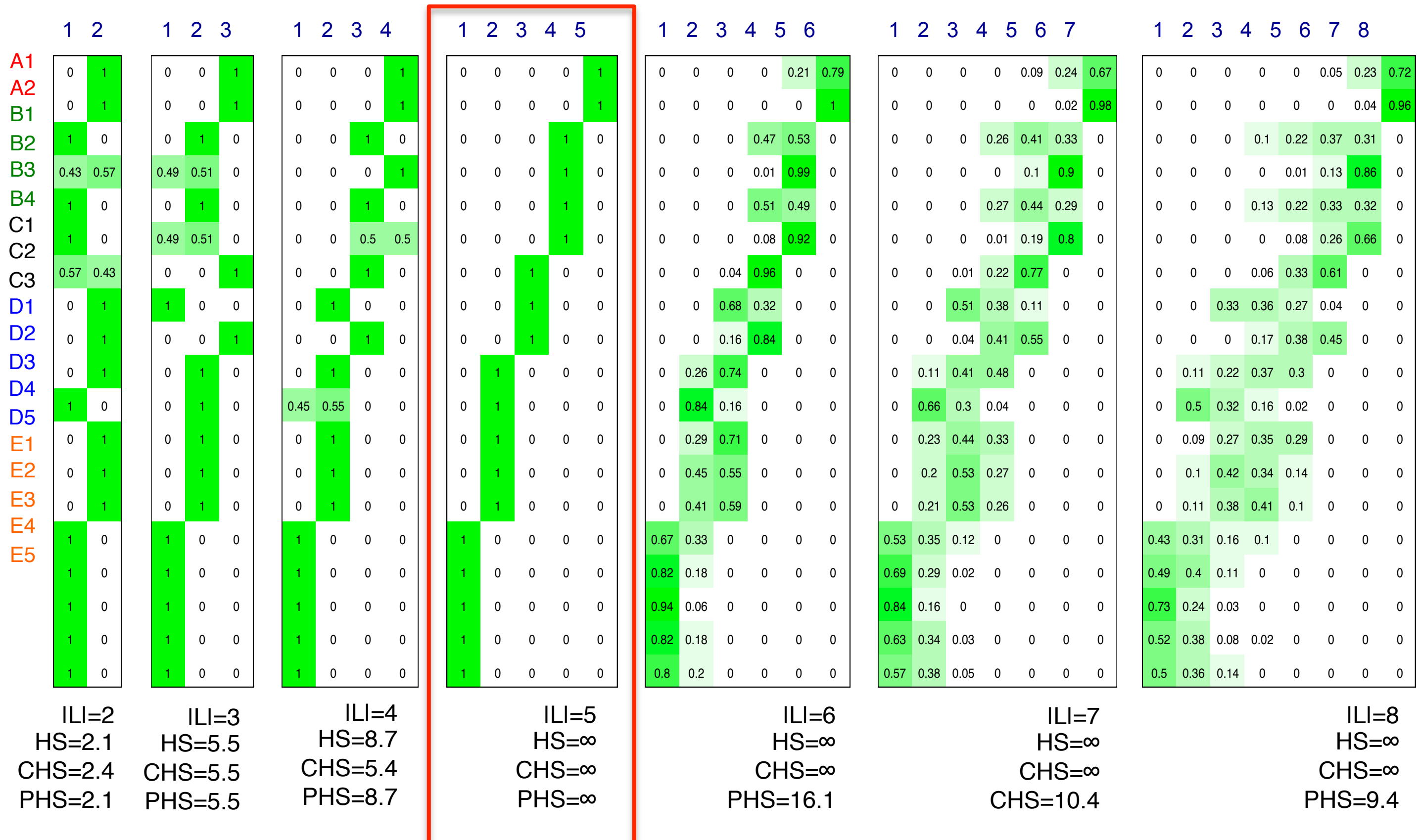


Discretized hierarchy network

[Cheng et al. Genome Biol. ('15)]

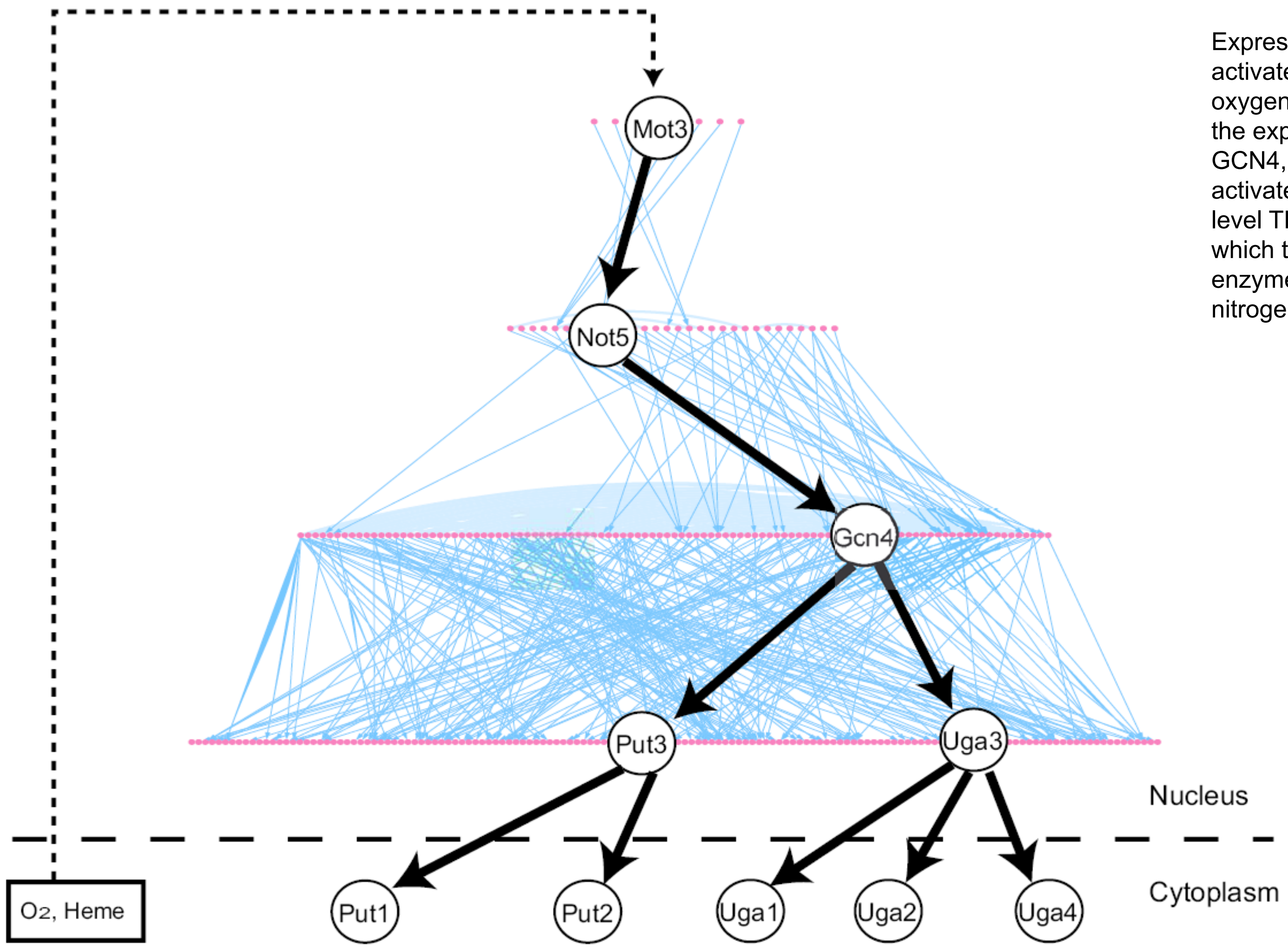
Apply HSM to a toy example





[Cheng et al. GenomeBiol. (in press, '15)]

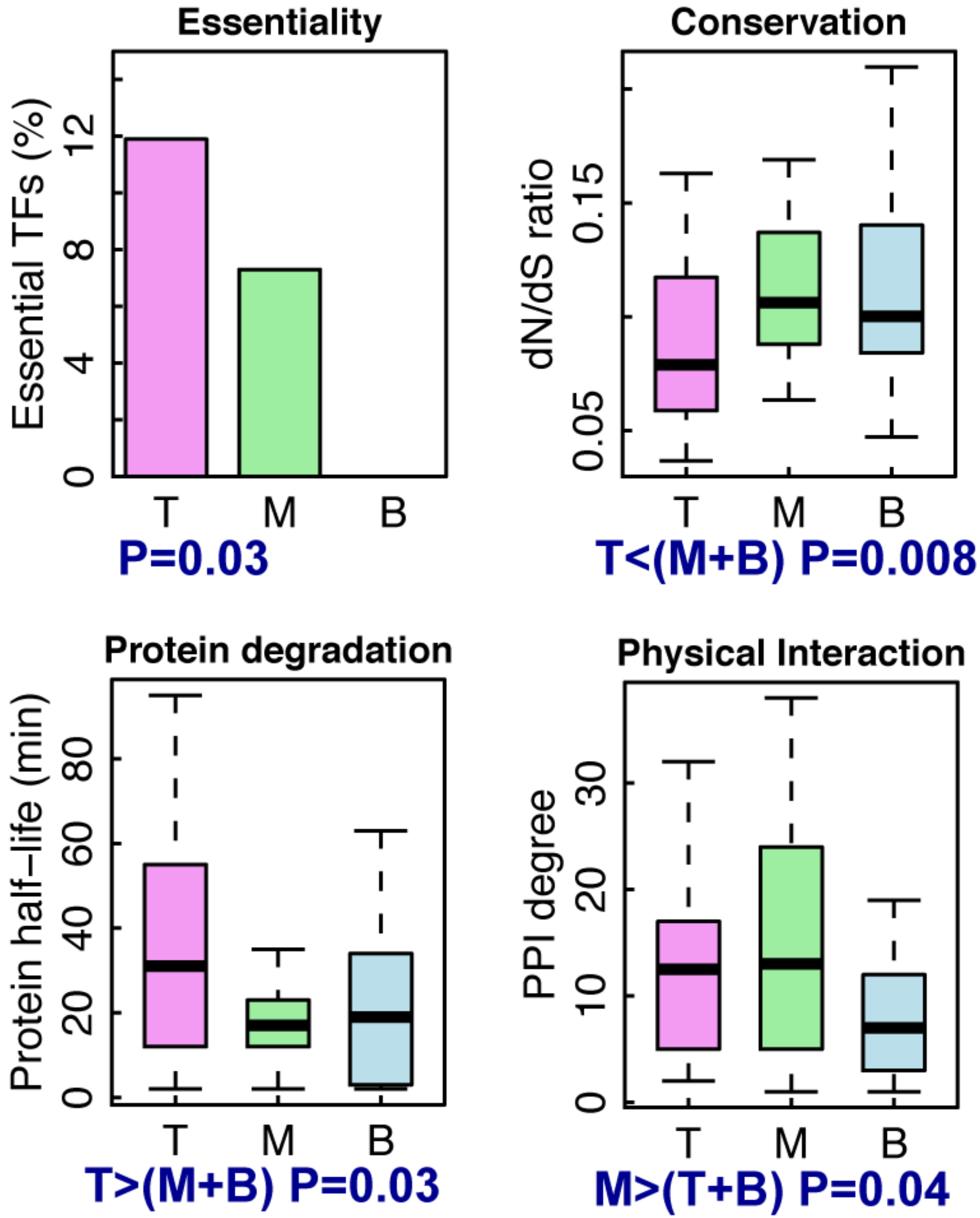
Example of Path Through Regulatory Network



Expression of MOT3 is activated by heme and oxygen. Mot3 in turn activates the expression of NOT5 and GCN4, mid-level hubs. GCN4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization.

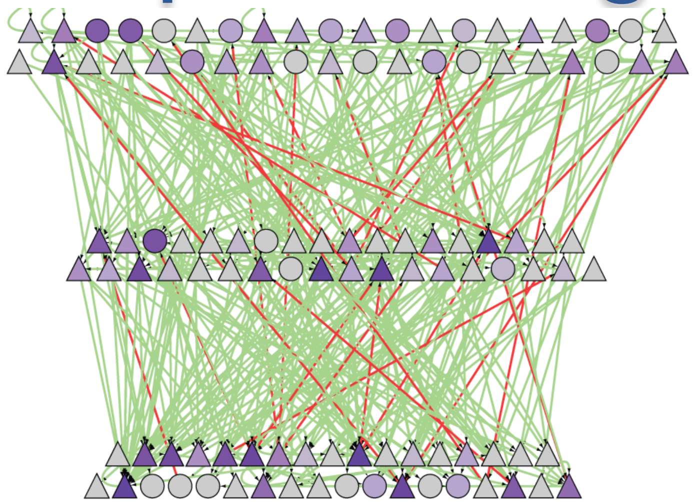
[Yu et al., PNAS (2006)]

Biological Insights from Hierarchy in Yeast TF Regulatory Network



[Cheng et al. GenomeBiol. (in press, '15)]

Hierarchical organization of human transcriptional regulatory network



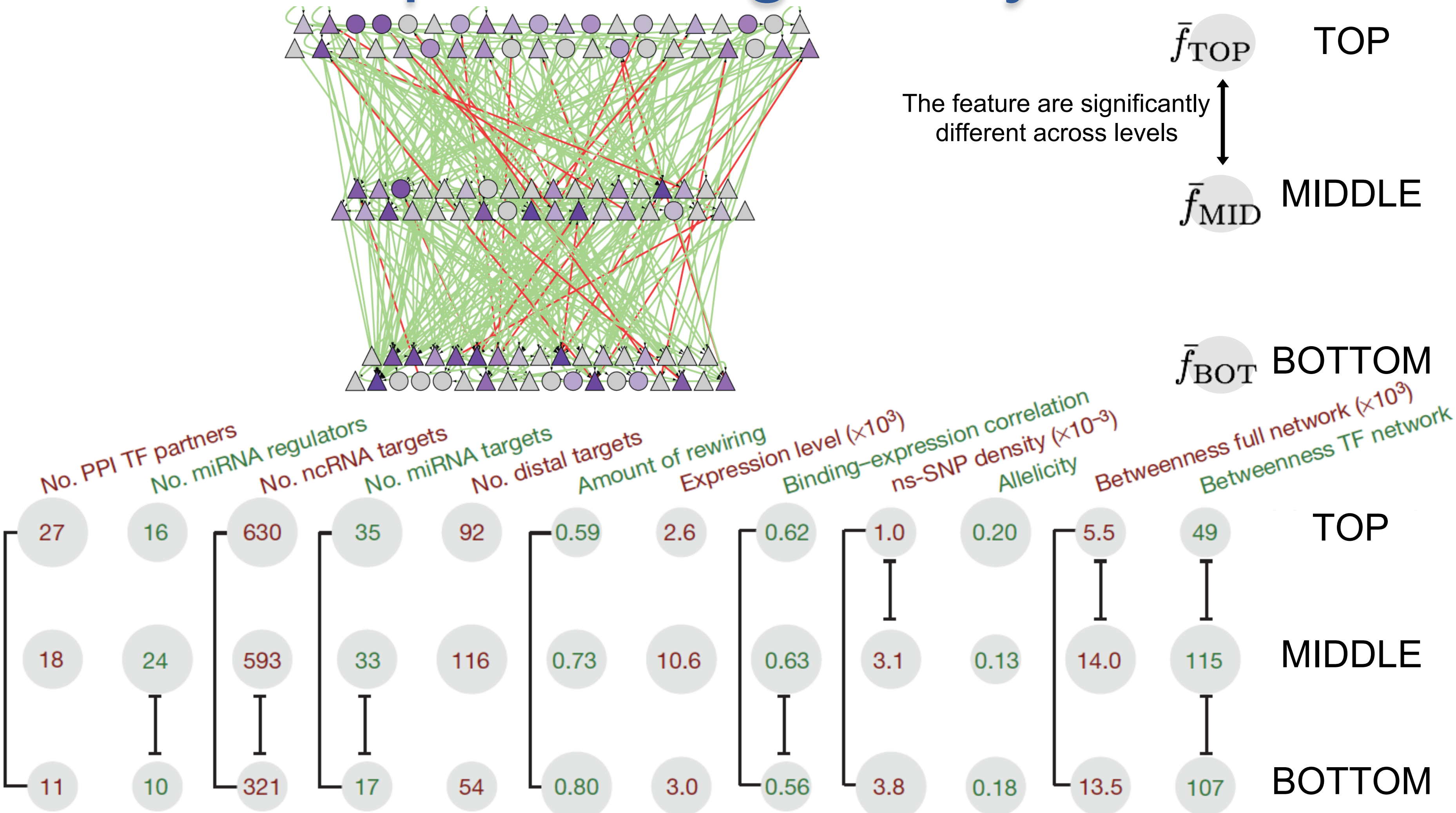
TOP

MIDDLE

BOTTOM

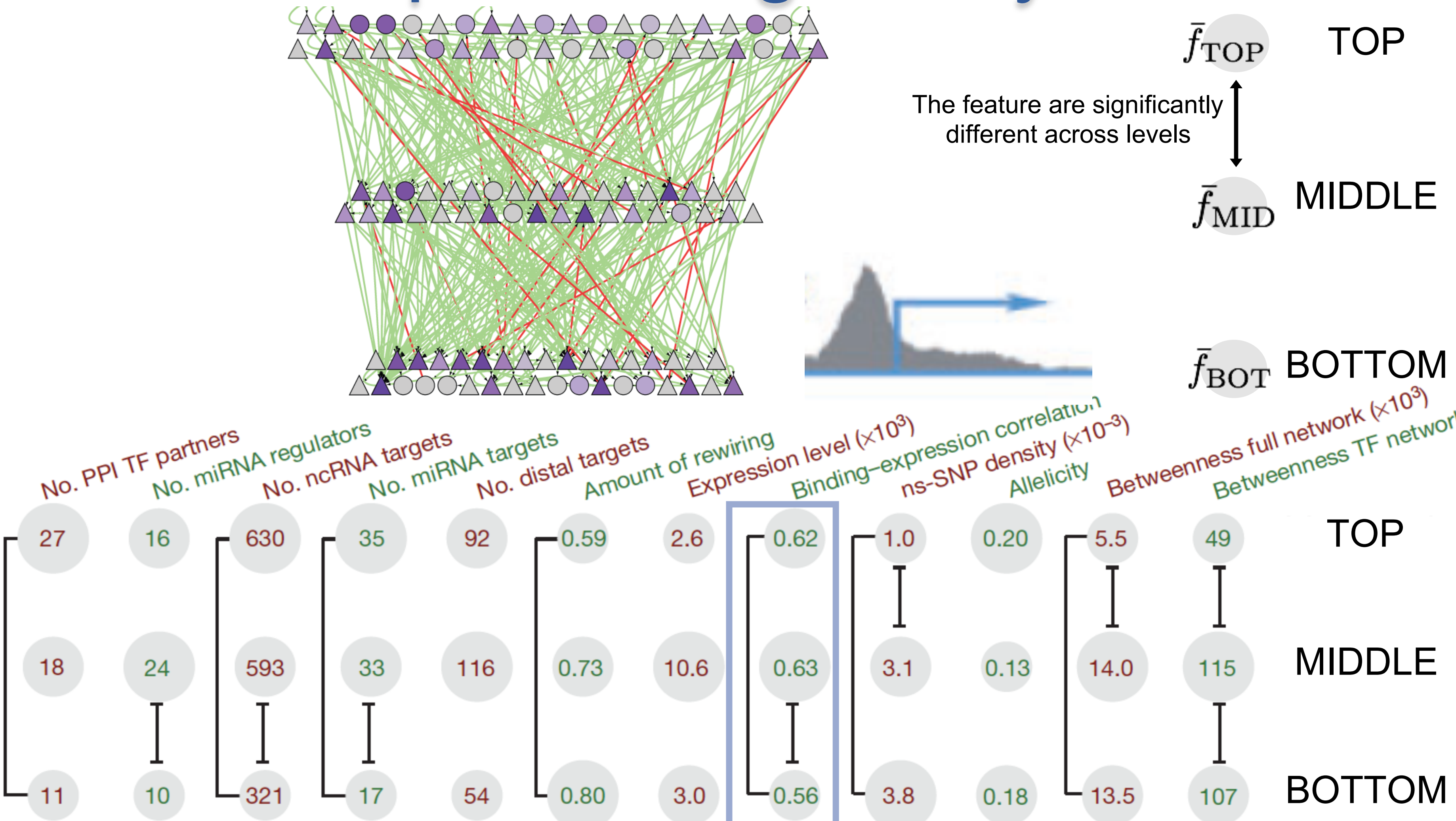
[Nature 489: 91]

Hierarchical organization of human transcriptional regulatory network



[Nature 489: 91]

Hierarchical organization of human transcriptional regulatory network

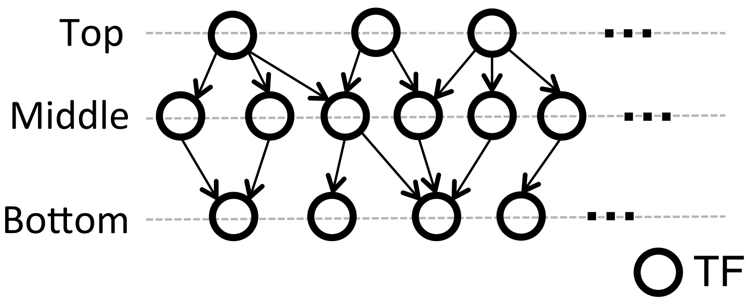


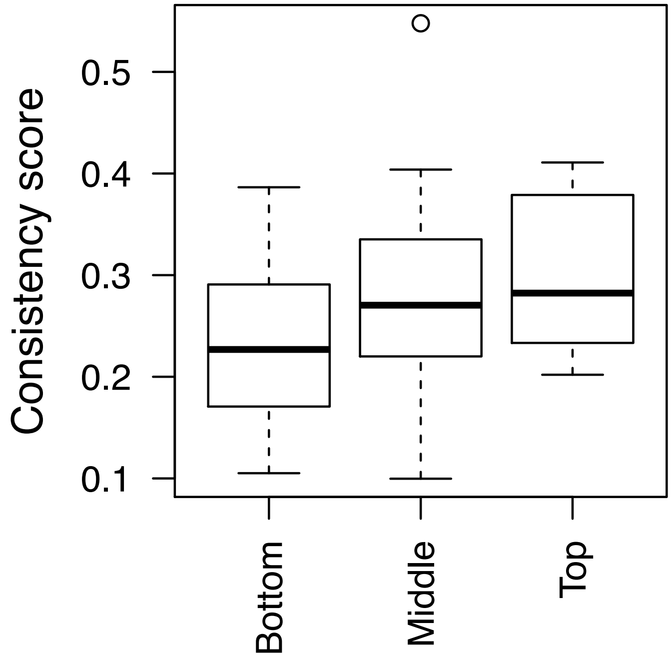
more influential TFs at the top

[Nature 489: 91]

Logical cooperativity across hierarchical layers in gene regulatory network

Hierarchical gene regulatory network in yeast





The regulations of middle and top TFs more likely follow logical operations than the bottom TFs.

Wang, et al., *PLoS Computational Biology*, 2015

Putting the regulatory hierarchy in perspective: Kinase network is more hierarchical than the TF reg. network

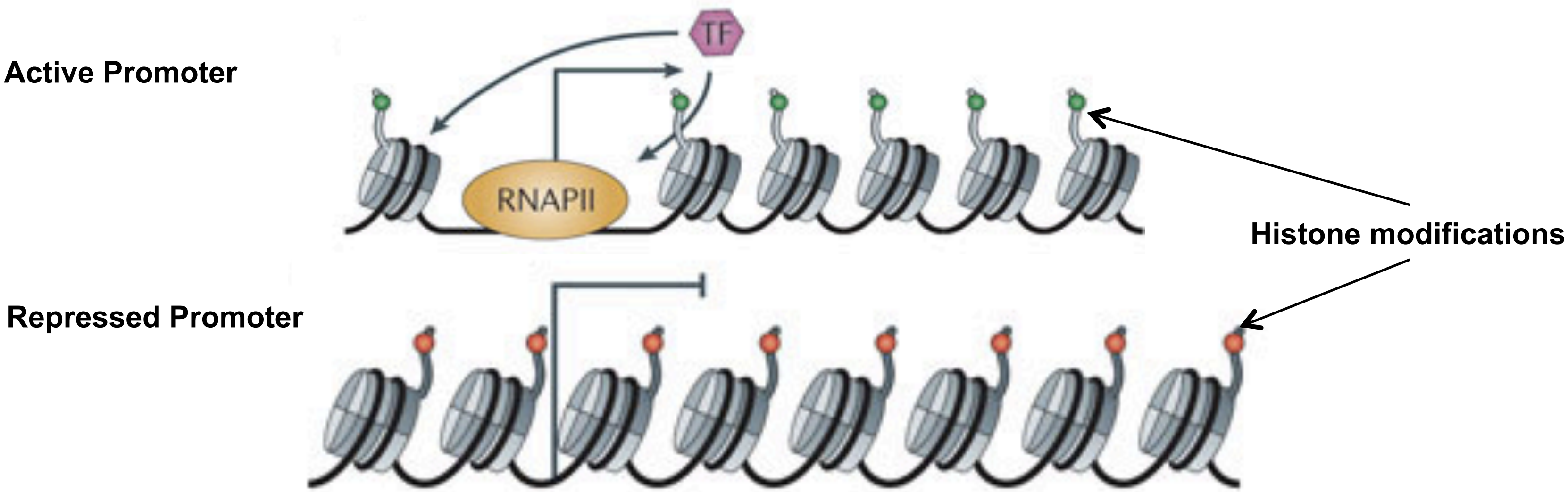
	CHS
Worm neural	2.3
Political blogs	3.1
Yeast TF	3.8
Human TF	5.6
P2P file sharing	5.8
Foodweb	6.4
Human Kinase	13.3
Yeast Kinase	13.9

[Cheng et al. GenomeBiol. (in press, '15)]

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

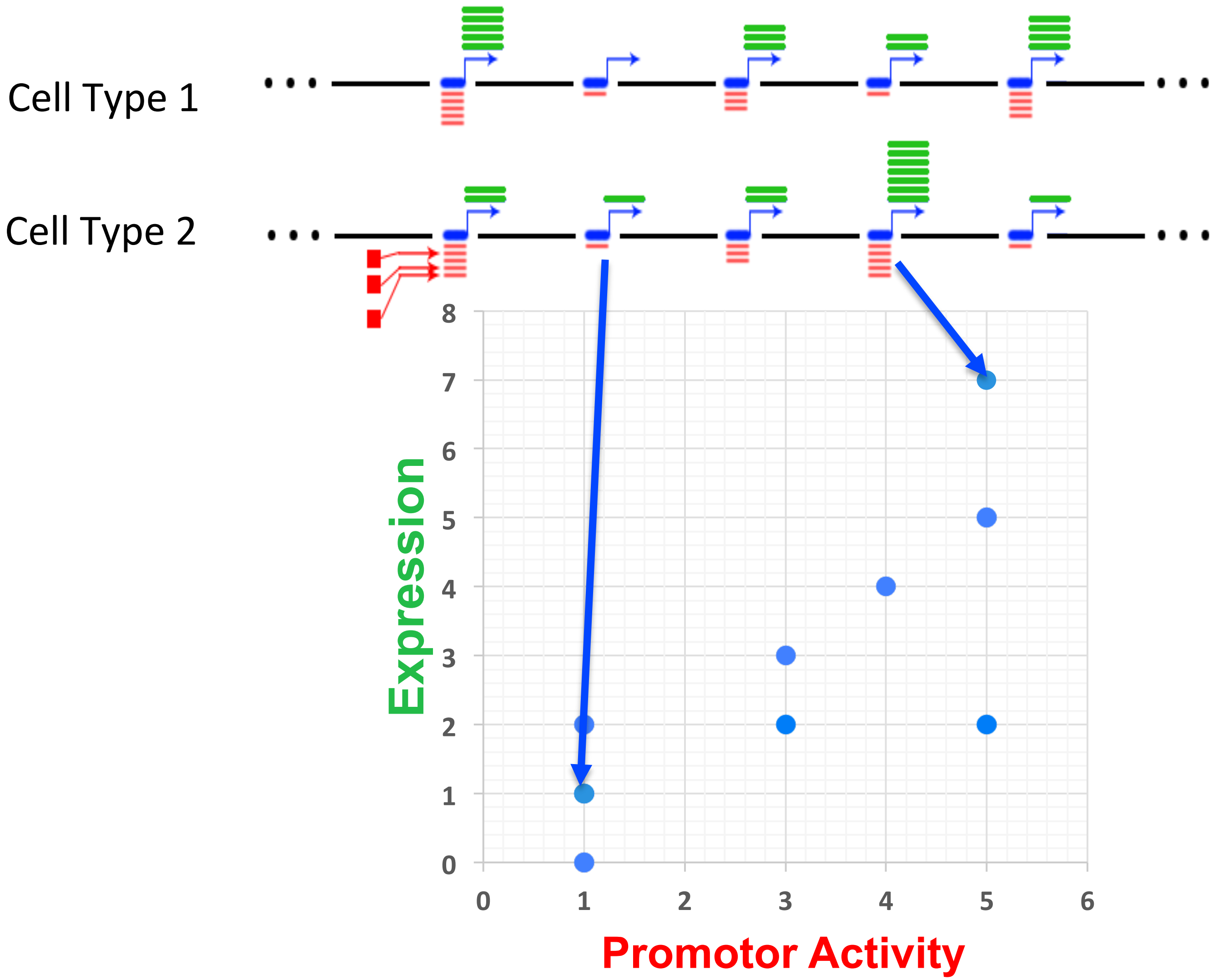
Focus on Promoters

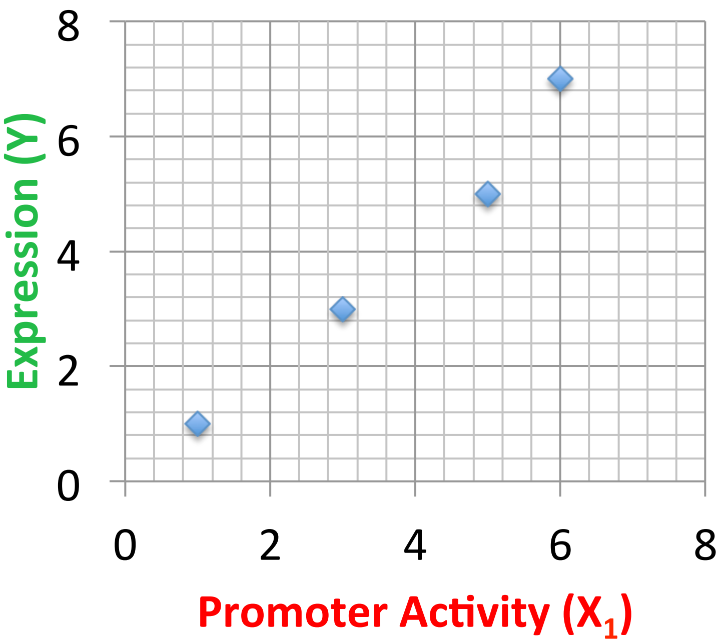


- Key Questions
  - How do we define the active regions of promoter?
  - For an active promoter, how do we relate it bound TFs, its epigenetic marks & its chromatin state to the level of transcription?
  - Are these definitions & relationships conserved between very different species?

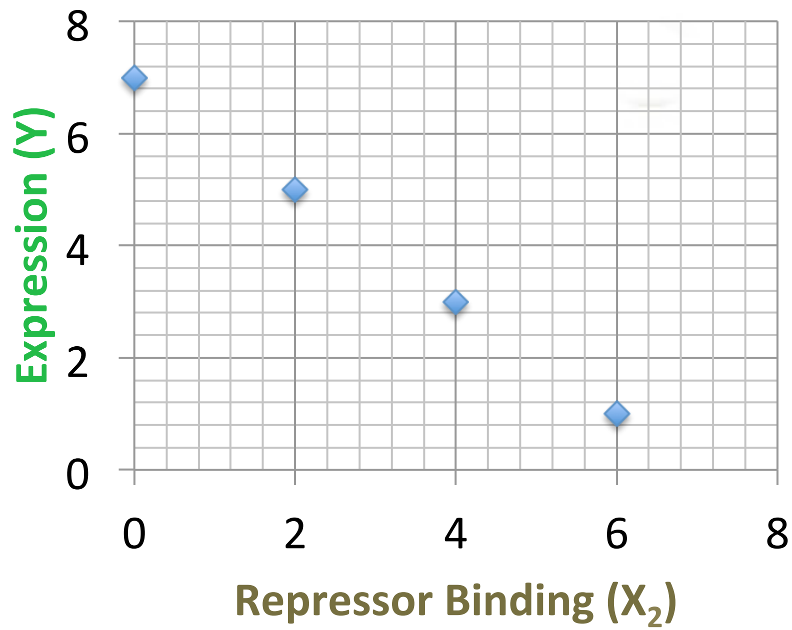
Zhou *et al. Nat. Rev. Genetics* 2011

Relating Genomic Inputs to Outputs





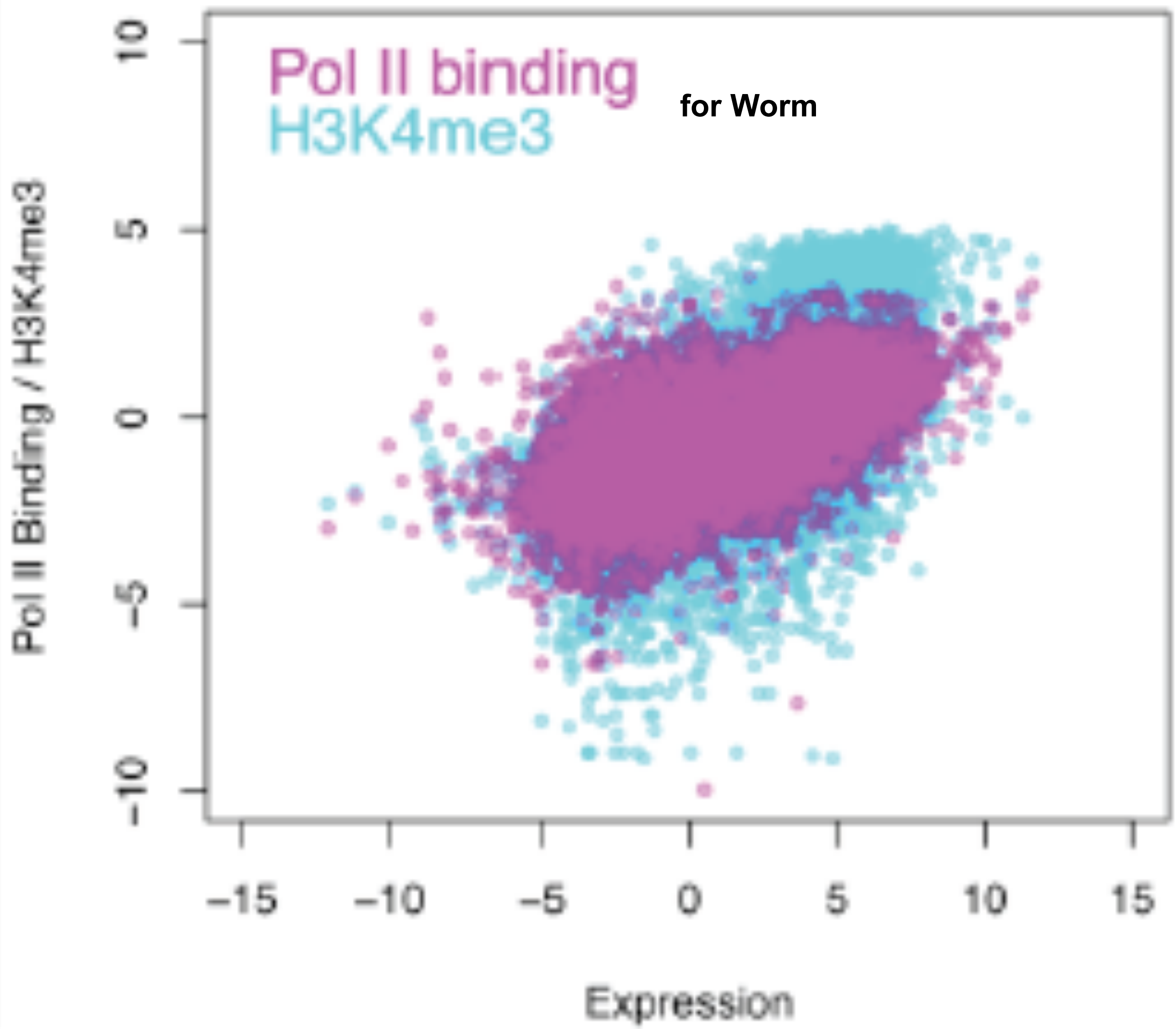
Expression (Y)
0
2
4
6
8
Promoter Activity (X_1)



Expression (Y)
0
2
4
6
8
Repressor Binding (X_2)

$$Y = aX_1 + bX_2 + c$$

Inputs v Outputs: Upstream Binding/Modification v Expression

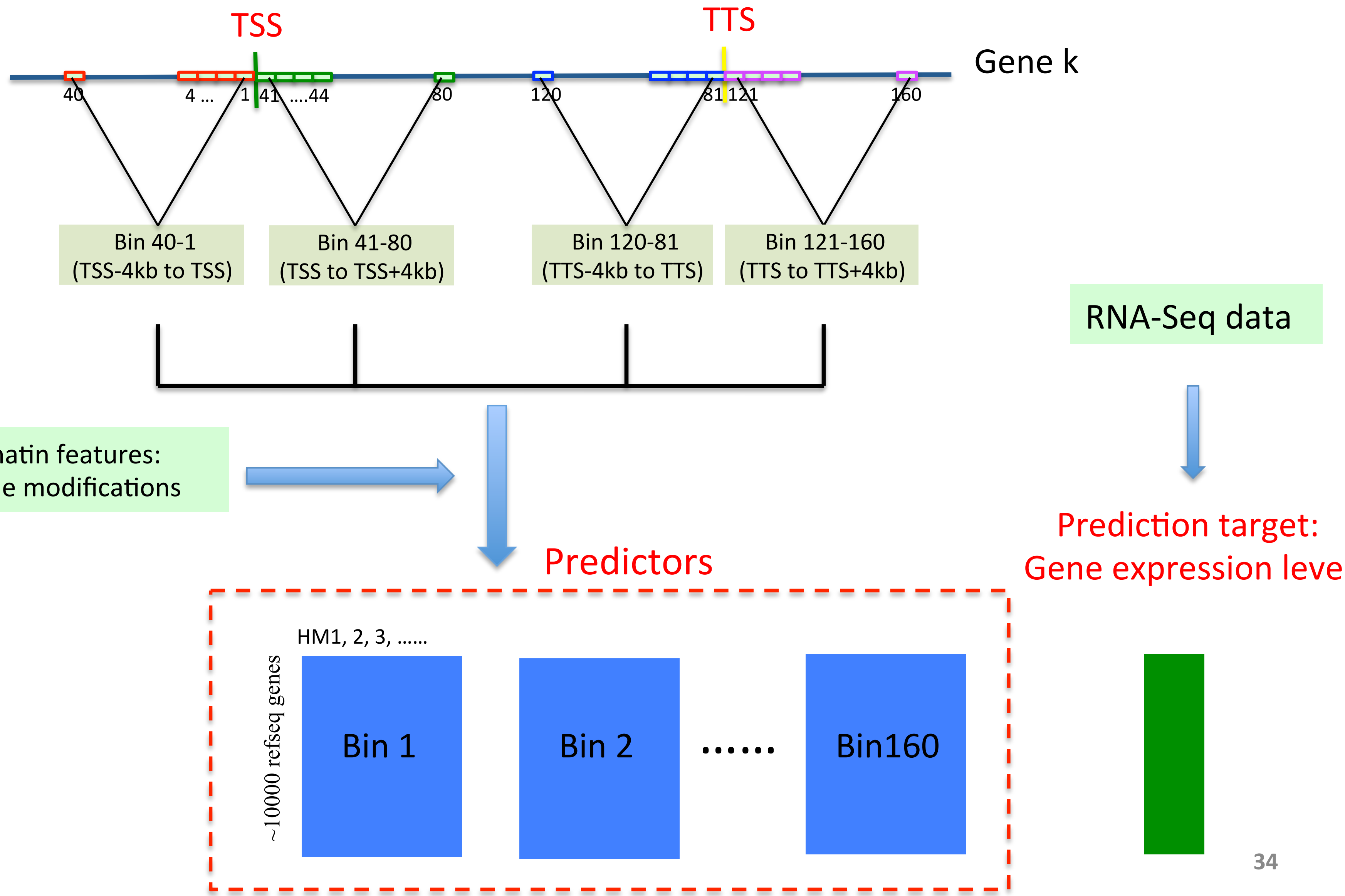


PCC: Pol II, **0.33;** H3K4me3, 0.28

[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Histone Modification (HM) model

[Cheng et al. ('11) *Genome Biol.* 12: R15]



His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion

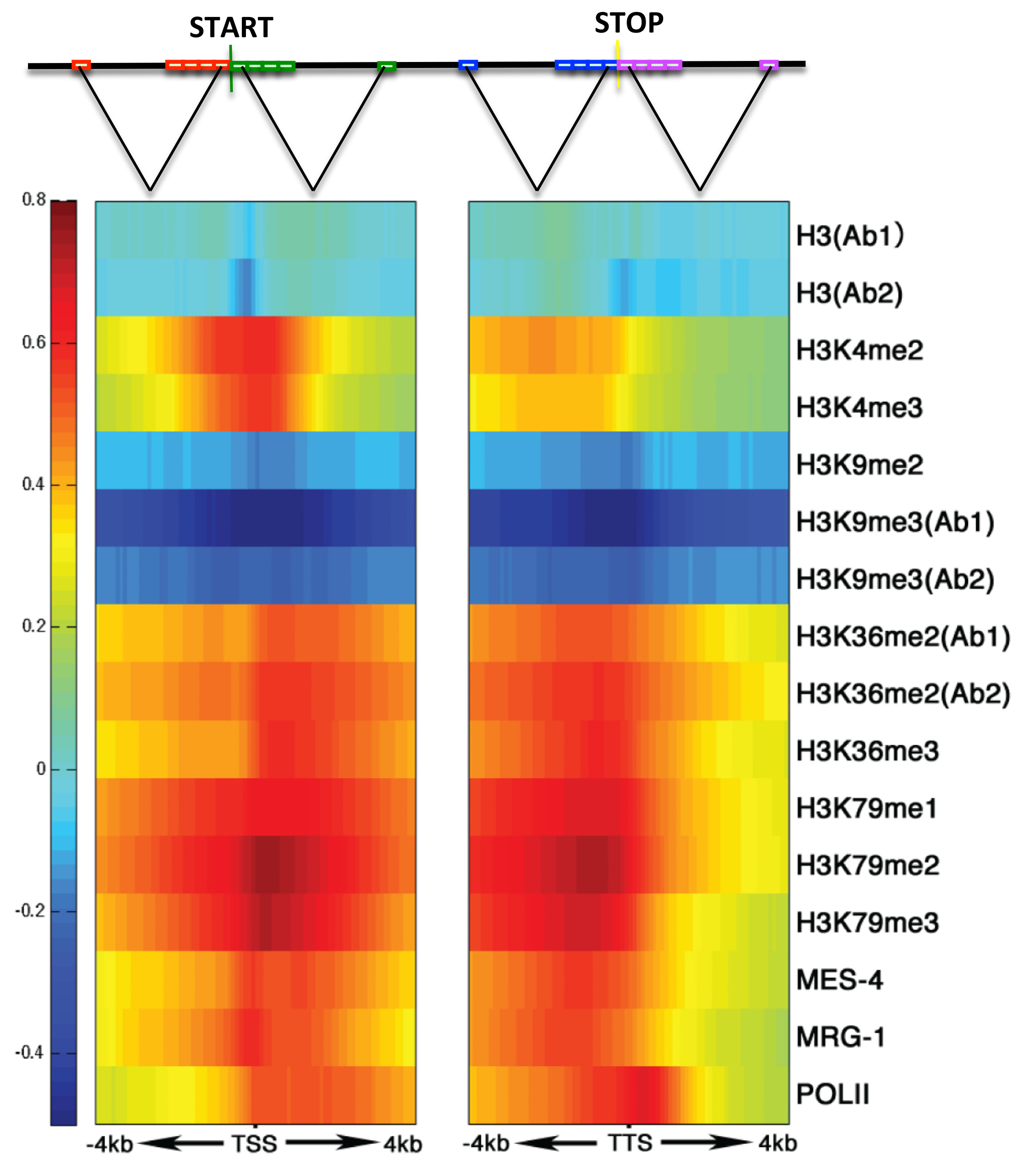
Early work in '09/'10

Science 330:6012 [here]

Also:

Ouyang, Zhou, Wong ('09) *PNAS*;

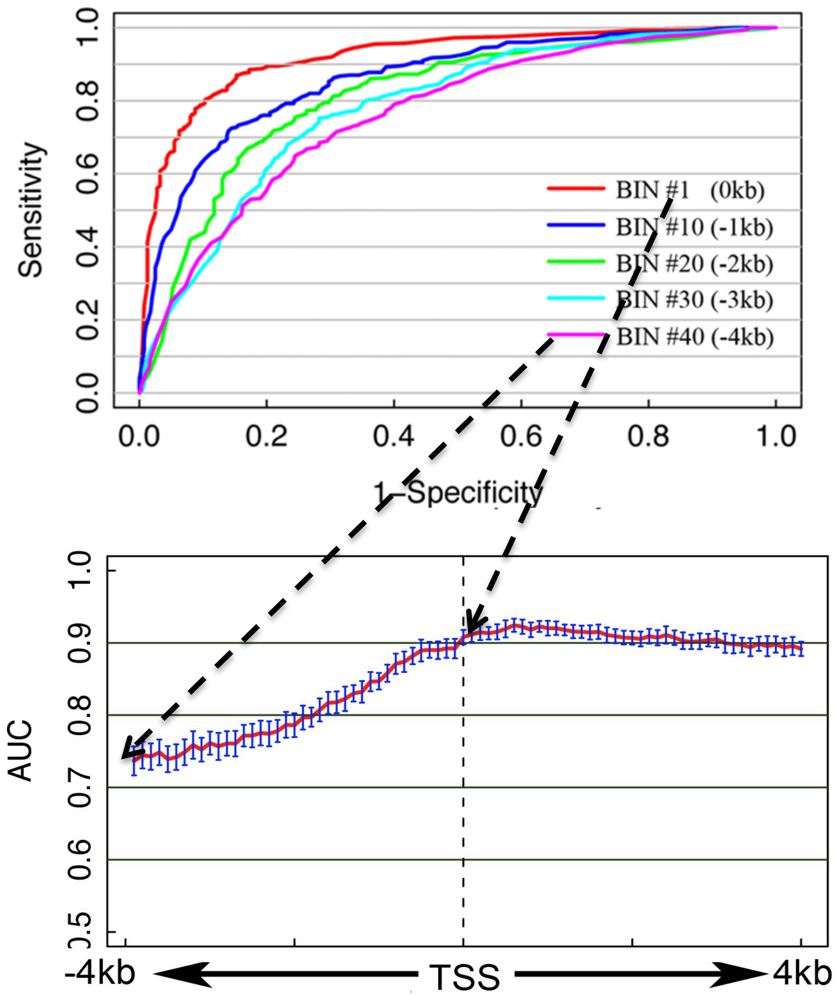
Karlic et al. & Vingron ('10) *PNAS*



Integrate all histone modifications to predict gene expression levels

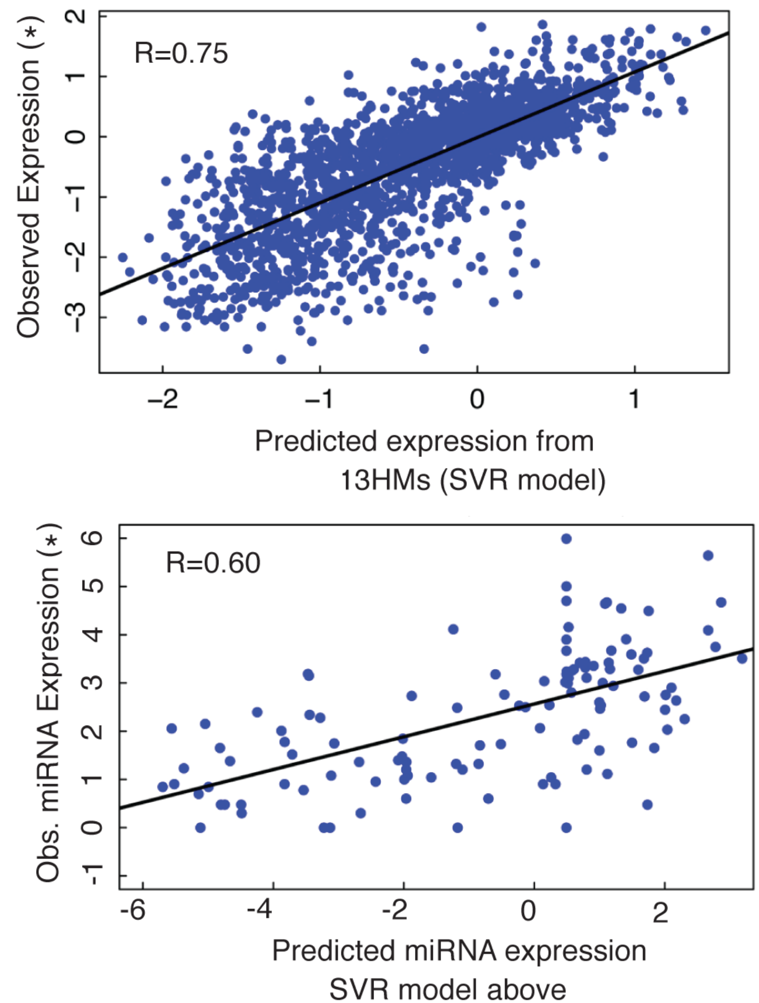
[Cheng et al. ('11) *Genome Biol.* 12: R15]

Classify H/L genes (SVM)



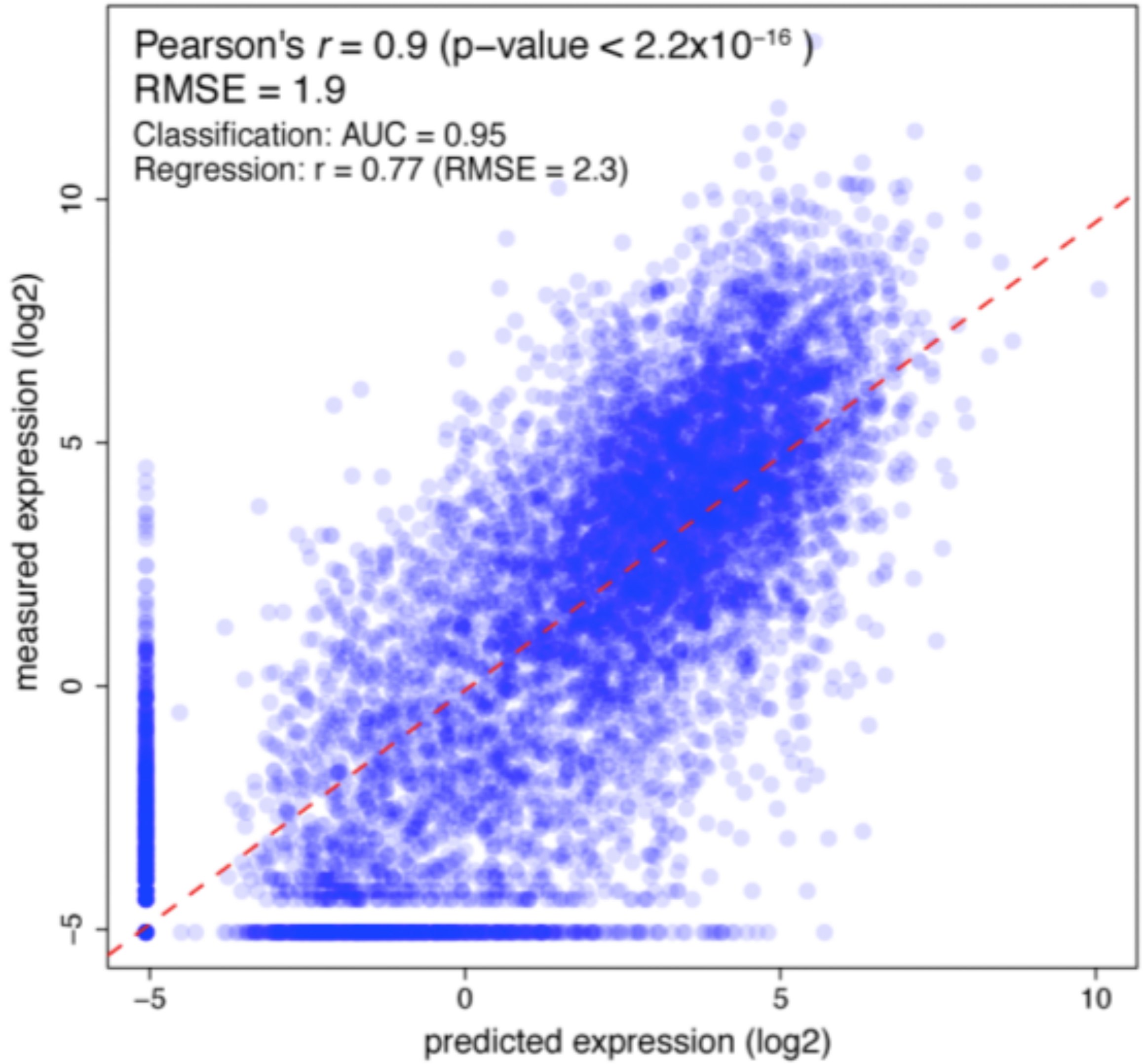
Magnitude of Prediction from a "bin" around the TSS

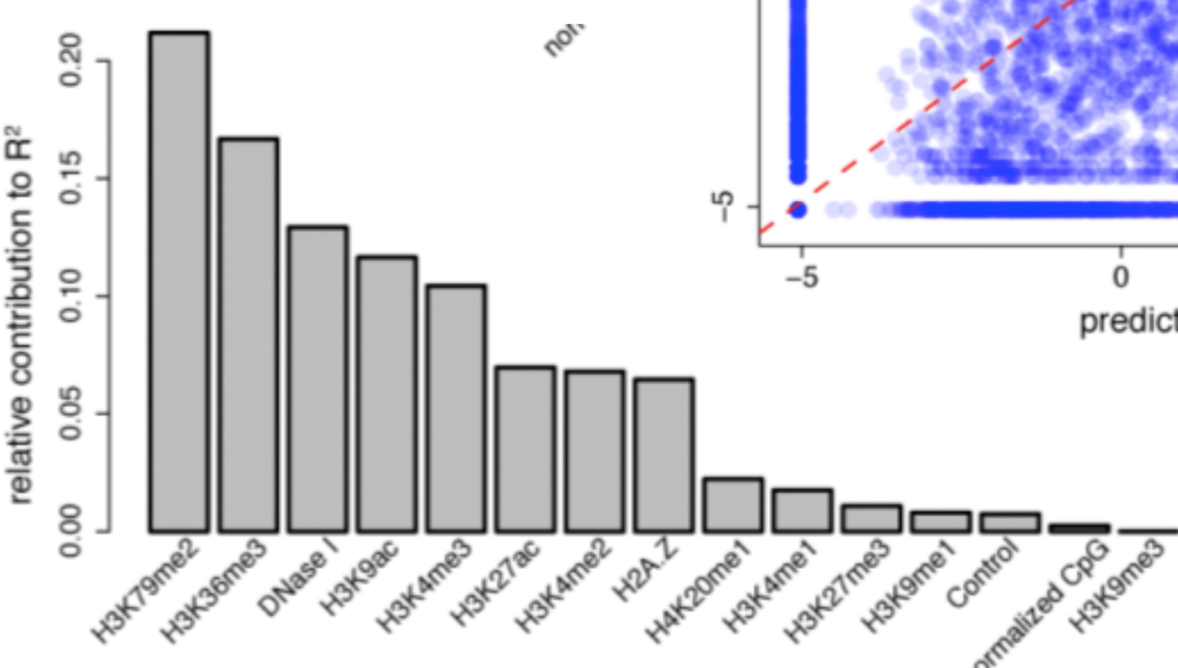
Predict expression values



$* = LOG_{10}RPKM$

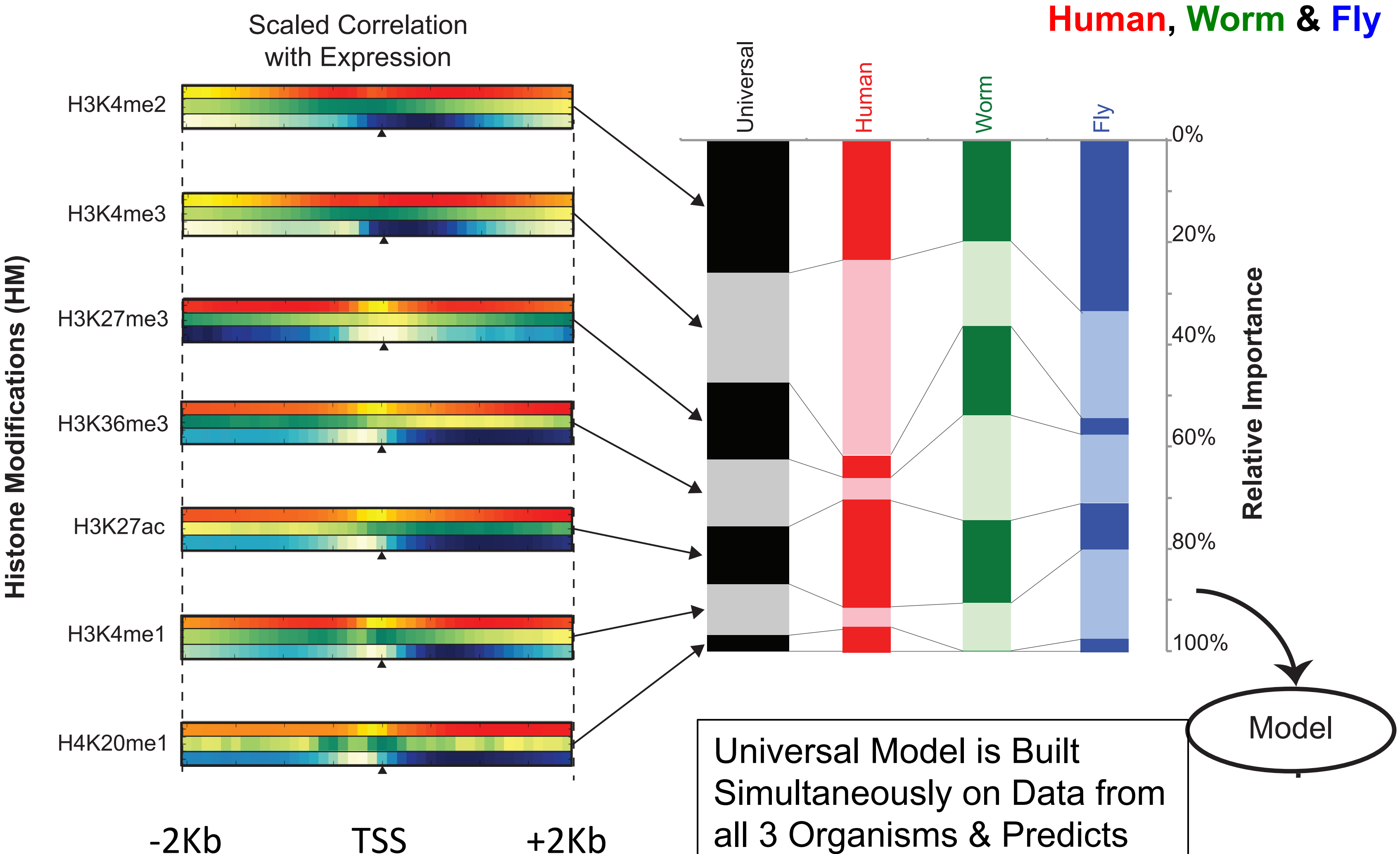
Human ENCODE Results





[Encode Consortium, *Nature* ('12)]

Comparison of Models for Gene Expression, Building a Universal Model



Universal Model is Built Simultaneously on Data from all 3 Organisms & Predicts on all 3 with a Single Set of Parameters

[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Performance of Universal, cross-organism Model

- works almost as well as species specific models
- works for both mRNAs and ncRNAs

Prediction Accuracy for Protein-coding Genes

Model Trained in	Human	Worm	Fly
Human	.82	.66	.69
Worm	.66	.74	.70
Fly	.69	.68	.84

Prediction Accuracy of Universal Model

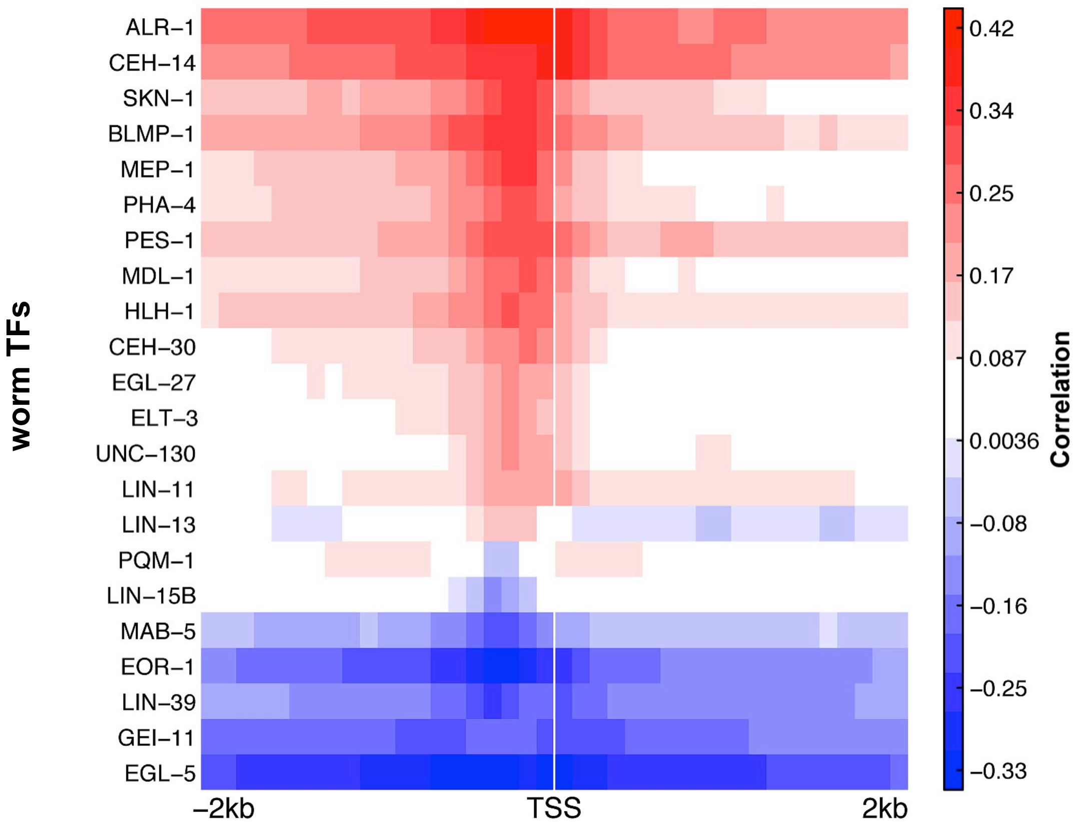
Protein coding	.80	.73	.83
ncRNA	.69	.51	.60

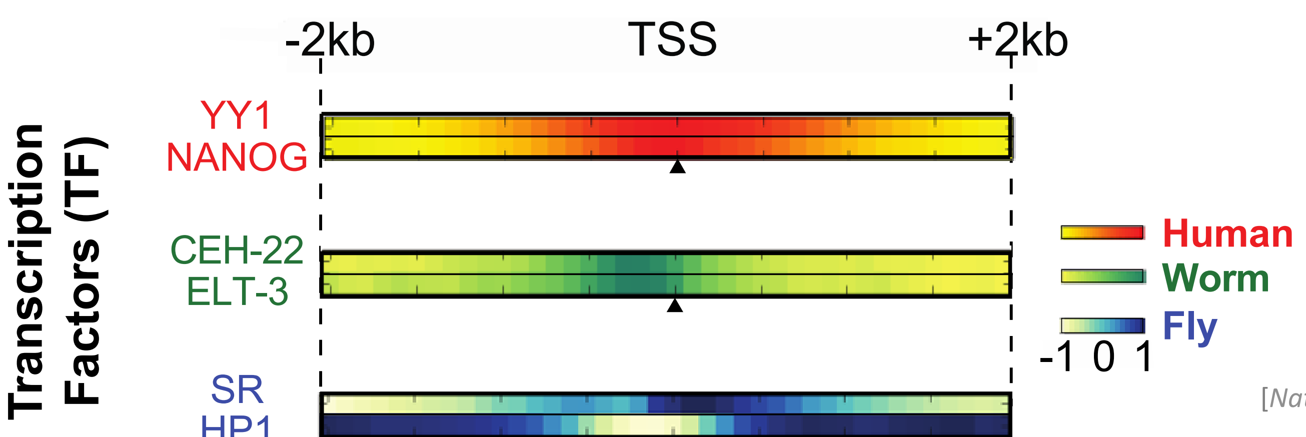
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for ‘prediction'
  - Surprisingly, a few TFs are quite predictive

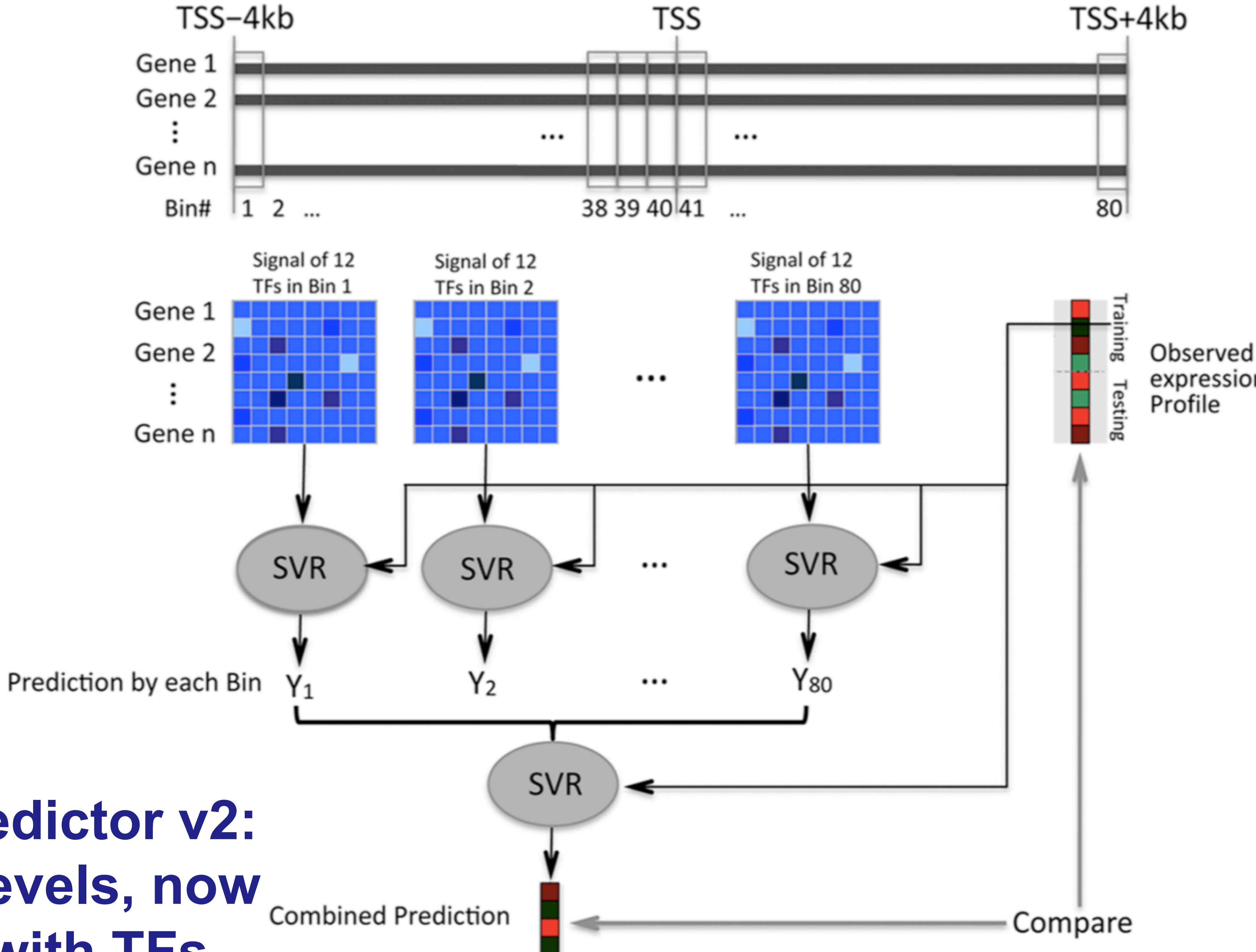
Doing a Model with TFs: Positive and negative regulators from correlating TF signal at TSS with gene expression





[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

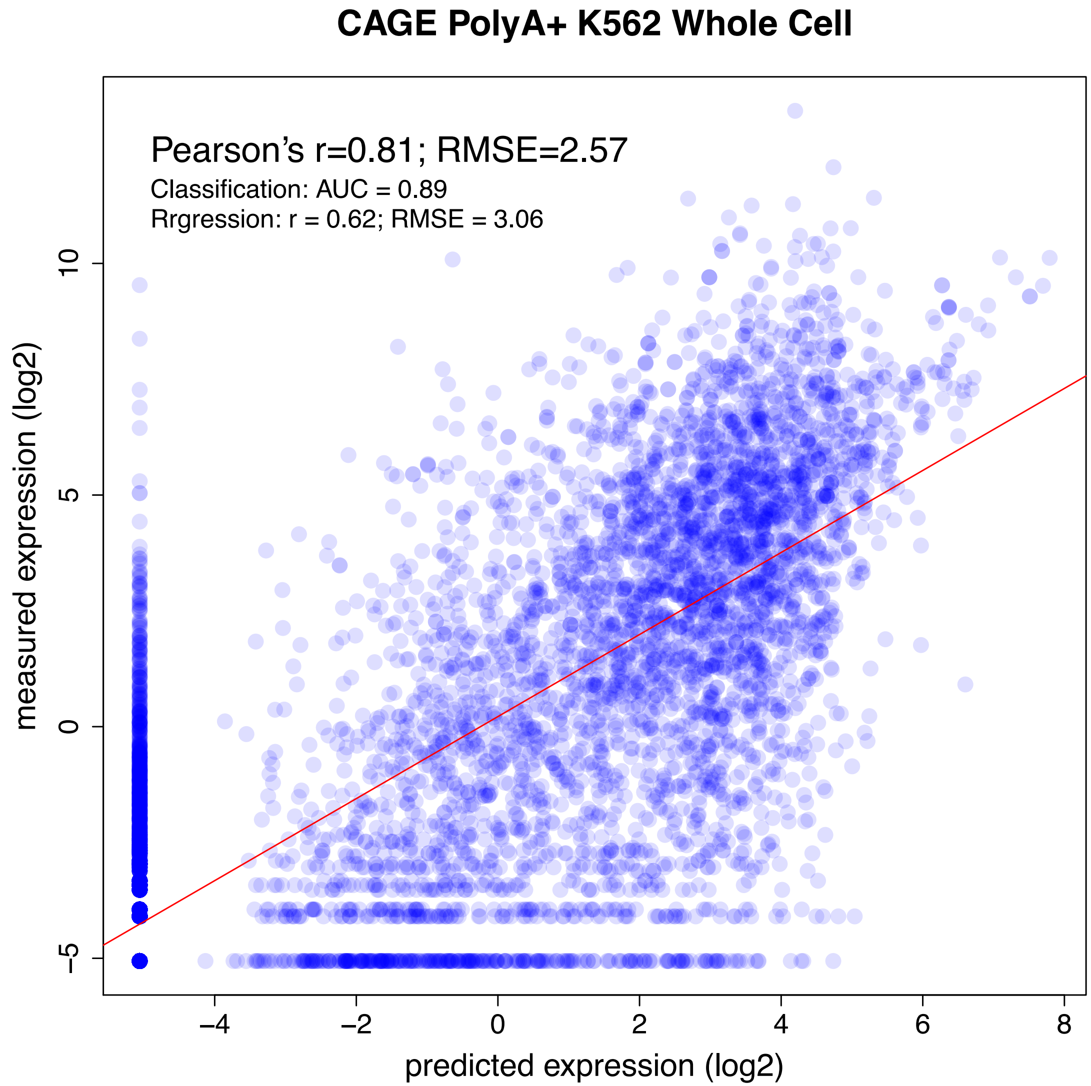
[Cheng et al. ('11) *PLOS CB*]

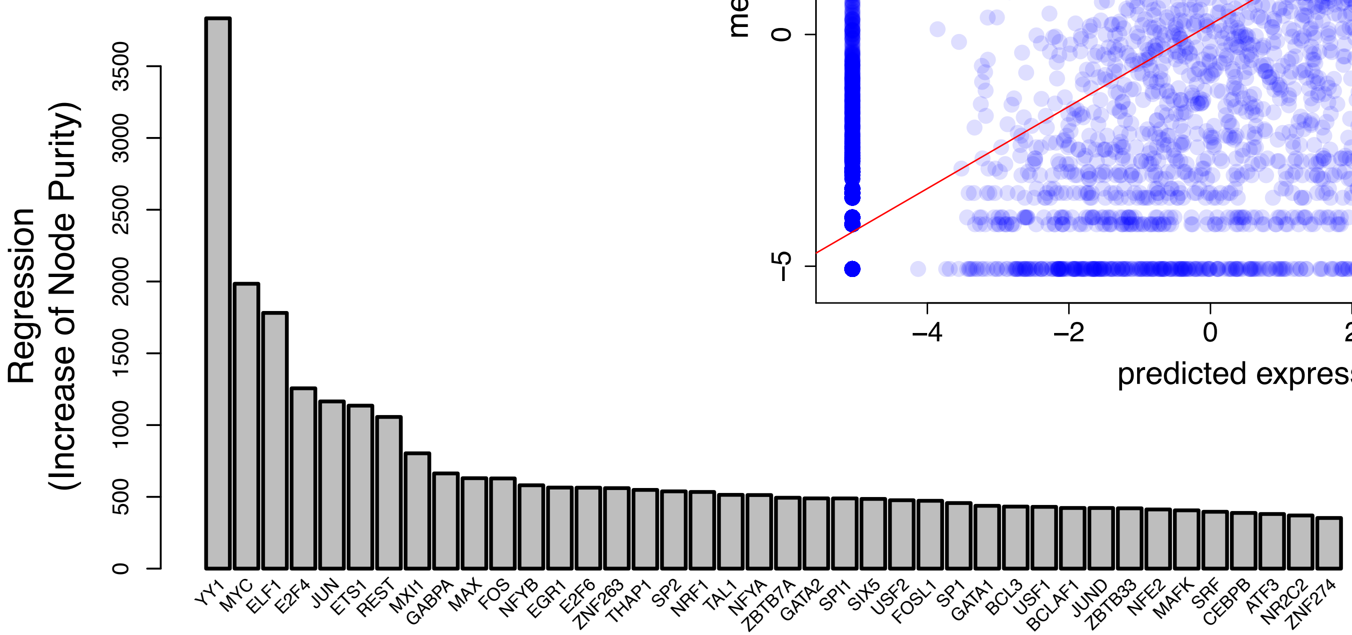


Predictor v2: 2-levels, now with TFs

[Cheng et al. *NAR* ('11)]

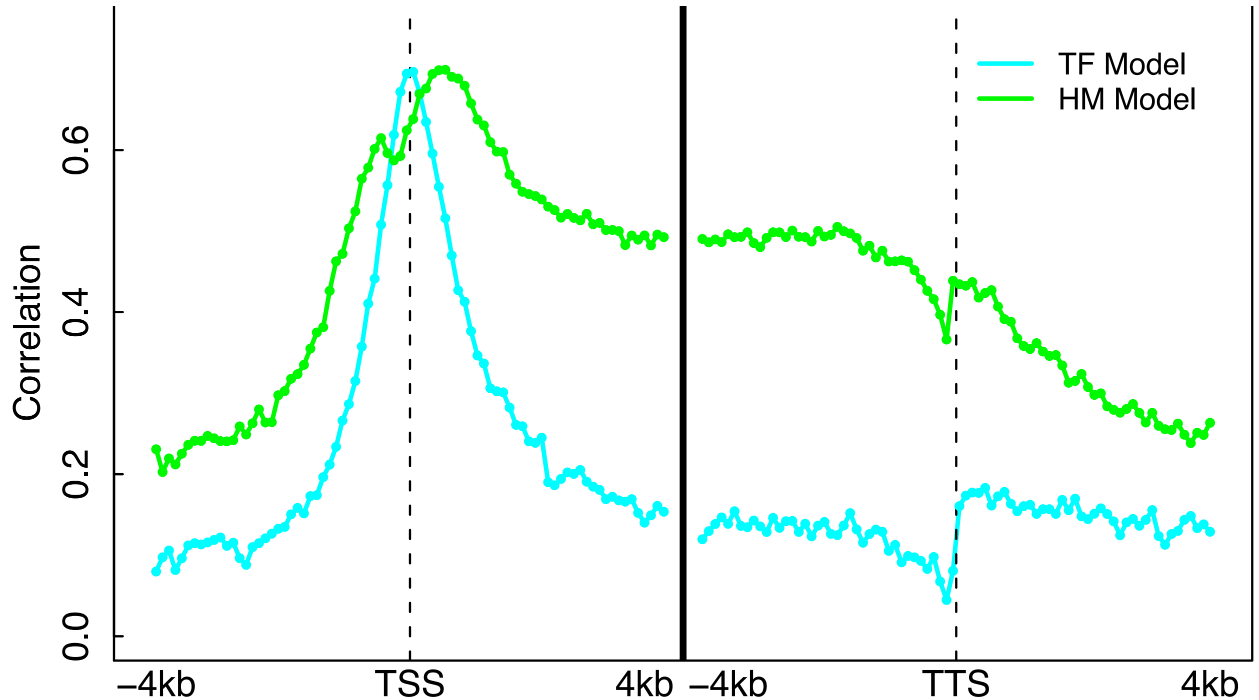
Human Results





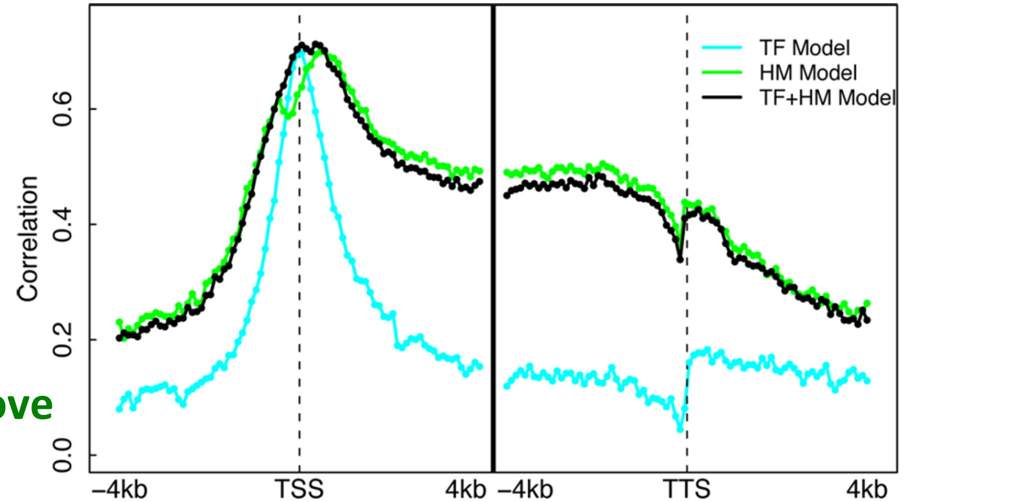
[Encode Consortium, *Nature* ('12)]

Models Illuminates Different Regions of Influence for TFs vs HMs



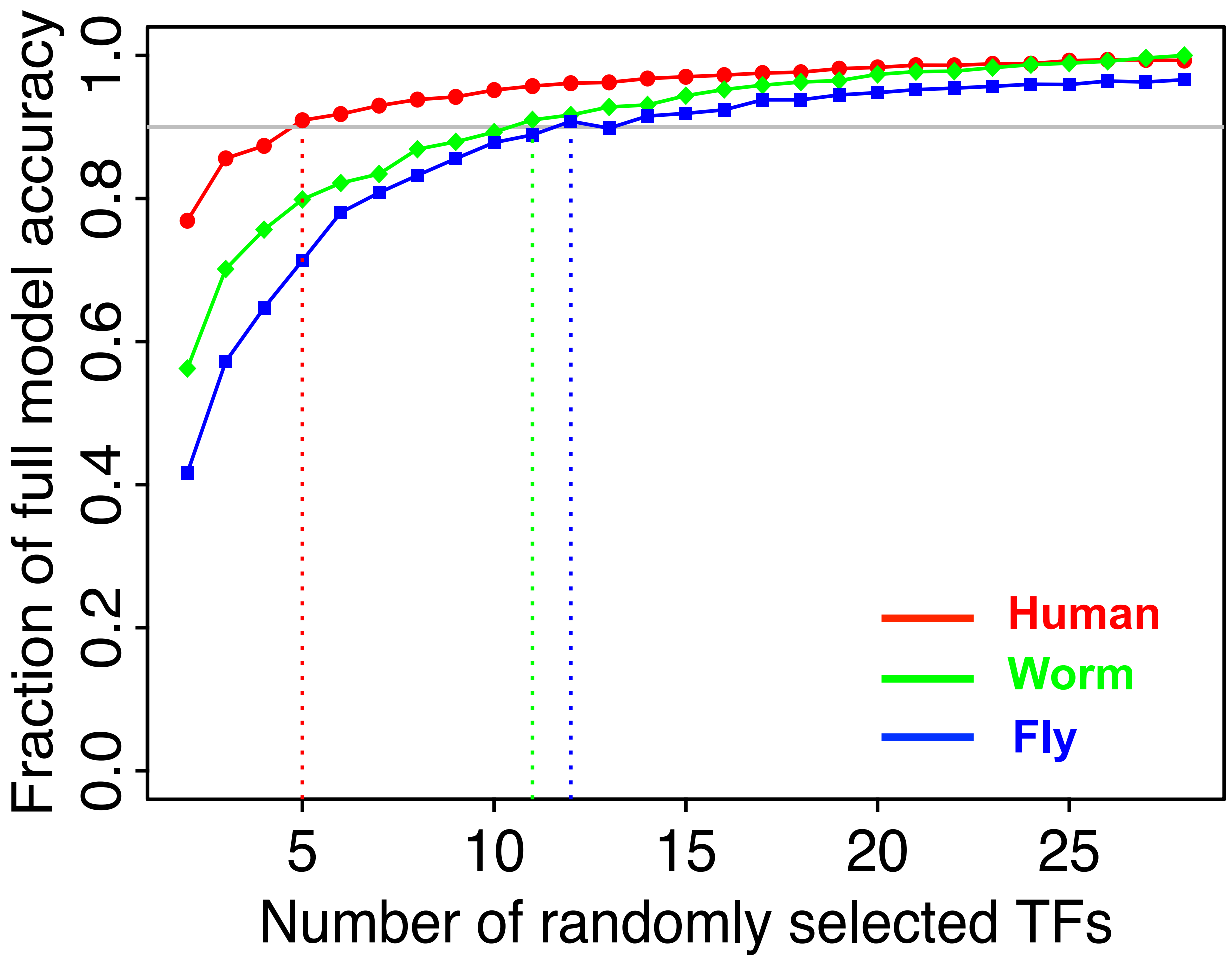
- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.'08; Mikkelsen et al. '07)
  - RNA-Seq (Cloonan et al. 2008)

A TF+HM model that combine TF and HM features does NOT improve accuracy!



Cheng et al. 2011, Nucleic Acids Res.

TF model accuracy only needs a small number of TFs for high accuracy (>90%)



[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

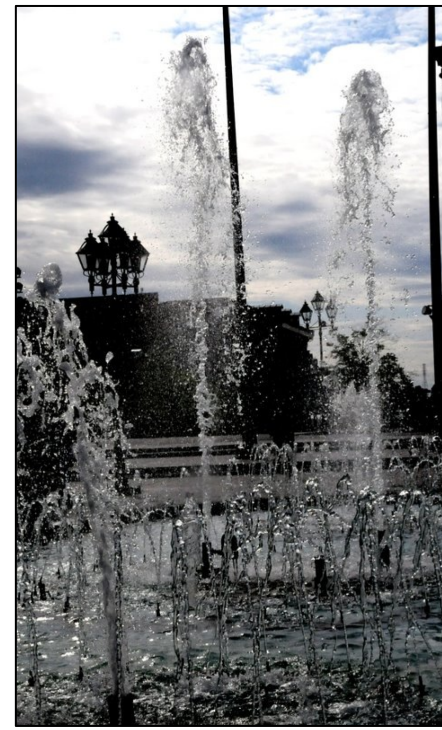
Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)
- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy
- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **Similarly constructed TF Models** [if time]
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

Acknowledgements

modENCODE/ENCODE Transcriptome subgroup

Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B. Brown, Carrie A. Davis, LaDeana Hillier, Cristina Sisu, **Jingyi Jessica Li,** Baikang Pei, Arif O. Harmanci, Michael O. Duff, Sarah Djebali, Roger P. Alexander, Burak H. Alver, Raymond K. Auerbach, Kimberly Bell, Peter J. Bickel, Max E. Boeck, Nathan P. Boley, Benjamin W. Booth, Lucy Cherbas, Peter Cherbas, Chao Di, Alex Dobin, Jorg Drenkow, Brent Ewing, Gang Fang, Megan Fastuca, Elise A. Feingold, Adam Frankish, Guanjun Gao, Peter J. Good, Phil Green, Roderic Guigó, Ann Hammonds, Jen Harrow, Roger A. Hoskins, Cédric Howald, Long Hu, Haiyan Huang, Tim J. P. Hubbard, Chau Huynh, Sonali Jha, Dionna Kasper, Masaomi Kato, Thomas C. Kaufman, Rob Kitchen, Erik Ladewig, Julien Lagarde, Eric Lai, Jing Leng, **Zhi Lu,** Michael MacCoss, Gemma May, Rebecca McWhirter, Gennifer Merrihew, David M. Miller, Ali Mortazavi, Rabi Murad, Brian Oliver, Sara Olson, Peter Park, Michael J. Pazin, Norbert Perrimon, Dmitri Pervouchine, Valerie Reinke, Alexandre Reymond, Garrett Robinson, Anastasia Samsonova, Gary I. Saunders, Felix Schlesinger, Anurag Sethi, Frank J. Slack, William C. Spencer, Marcus H. Stoiber, Pnina Strasbourger, Andrea Tanzer, Owen A. Thompson, Kenneth H. Wan, Guilin Wang, Huaien Wang, Kathie L. Watkins, Jiayu Wen, Kejia Wen, Chenghai Xue, Li Yang, Kevin Yip, Chris Zaleski, Yan Zhang, Henry Zheng, **Steven E. Brenner, Brenton R. Graveley, Susan E. Celniker, Thomas R Gingeras, Robert Waterston**



Hiring Postdocs. See gersteinlab.org/jobs !

ENCODE

(11 Main Projects, ~50 labs, >700 substantial contributors + NHGRI)

Network Acknowledgements

Networks/Elements (~60 participants):

A Kundaje, M Hariharan, S Landt, K Yan, C Cheng, X Mu, E Khurana, J Rozowsky, R Alexander, R Min, P Alves, A Abyzov, N Addleman, N Bhardwaj, A Boyle, P Cayting, A Charos, D Chen, Y Cheng, **D Clarke**, C Eastman, G Euskirchen, S Frietze, **Y Fu**, J Gertz, F Grubert, A Harmanci, P Jain, M Kasowski, P Lacroute, **J Leng**, J Lian, H Monahan, H O'Geen, Z Ouyang, E Partridge, D Patacsil, F Pauli, D Raha, L Ramirez, T Reddy, B Reed, M Shi, T Slifer, J Wang, L Wu, X Yang, **K Yip**, G Zilberman-Schapira, S Batzoglou, A Sidow, P Farnham, **R Myers**, S Weissman, **M Snyder**

Hiring Postdocs. See gersteinlab.org/jobs !

Acknowledgements

- Hierarchy Construction –

  Chao **Cheng**, Erik Andrews, Koon-Kiu Yan, Matthew Ung, Daifeng Wang

  [papers.gersteinlab.org/papers/hinet coming soon]

- Loregic –

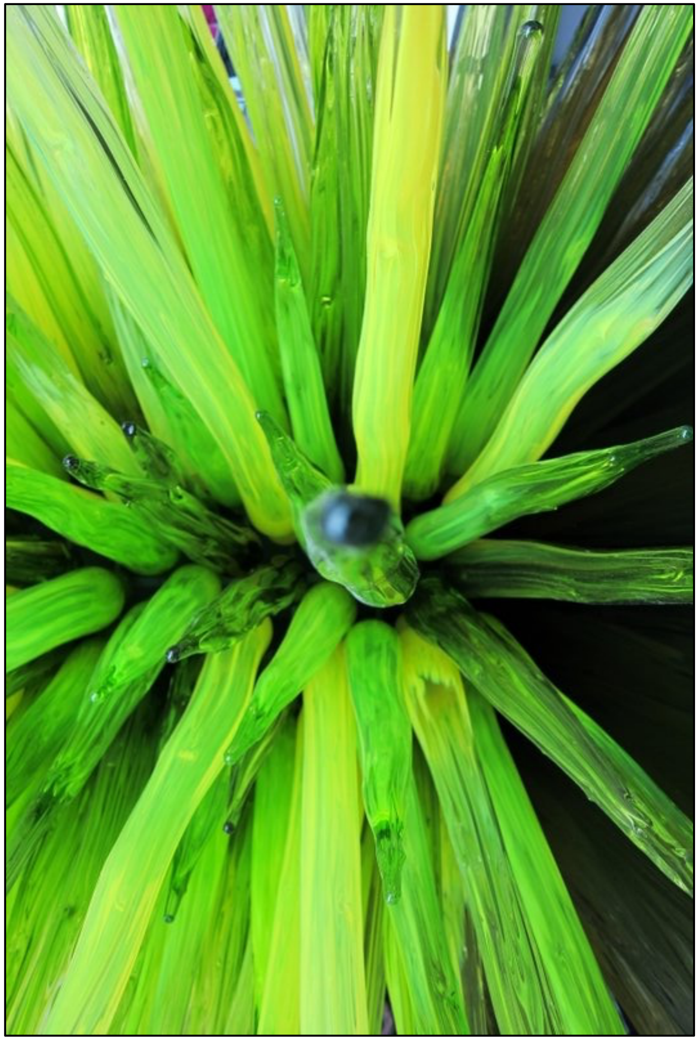
  Daifeng **Wang**, Koon-Kiu Yan, Cristina Sisu, Chao Cheng, Joel Rozowsky, William Meyerson

  [papers.gersteinlab.org/papers/loregic coming soon]

TF-v-expr:
Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M

worm-HM:
Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C

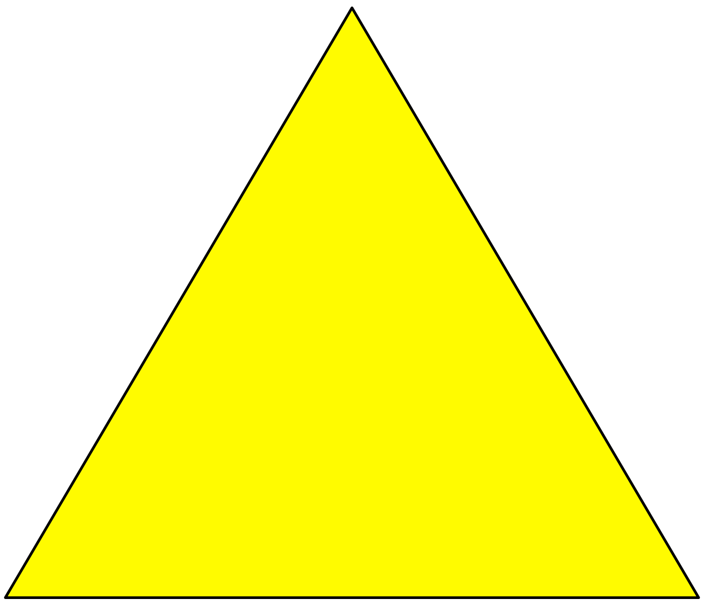
ENCODE:
Chao Cheng, Roger Alexander, **Renqiang Min**, Kevin Y. Yip, Jing Leng, Joel Rozowsky, Koon-kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie A Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, Roderic Guigó Serra, Ewan Birney, Zhiping Weng, Michael Snyder



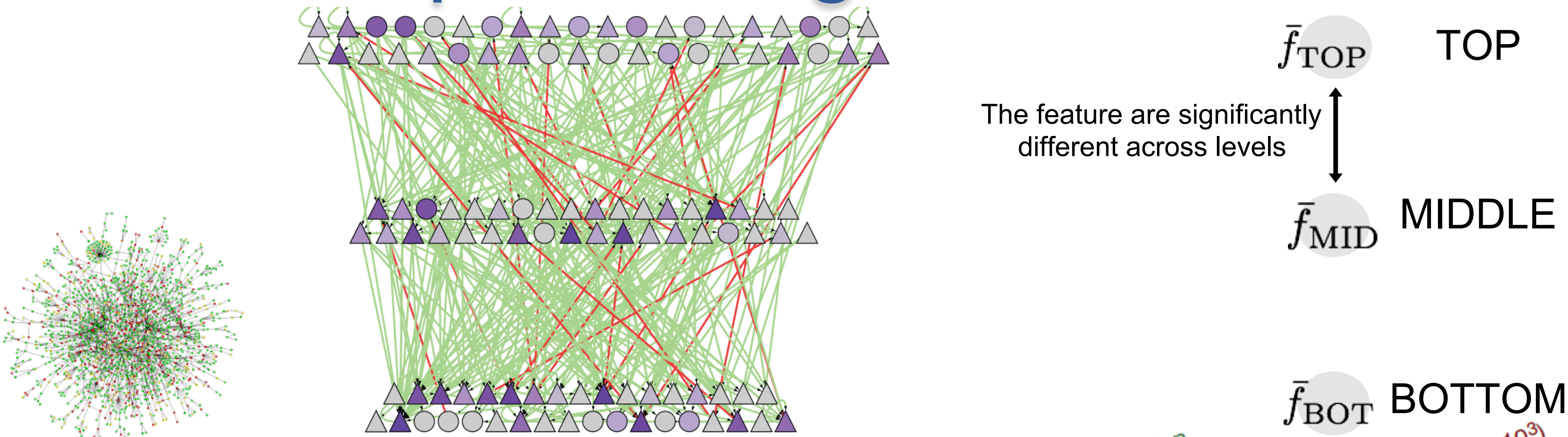
Hiring Postdocs. See gersteinlab.org/jobs !

Default Theme

- Default Outline Level 1
  - Level 2

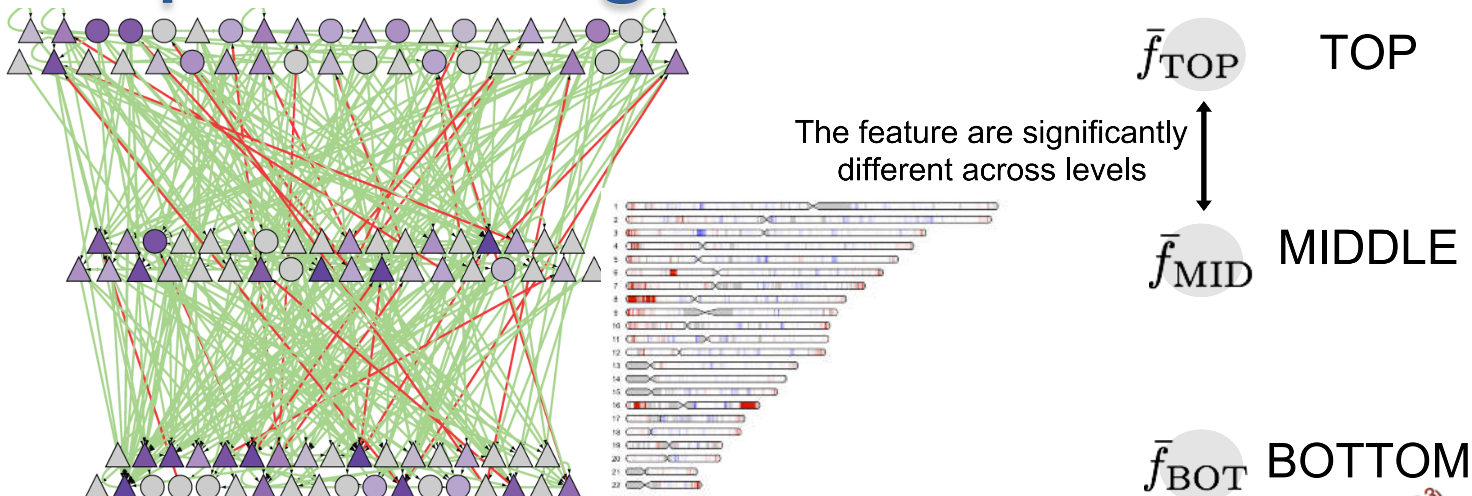


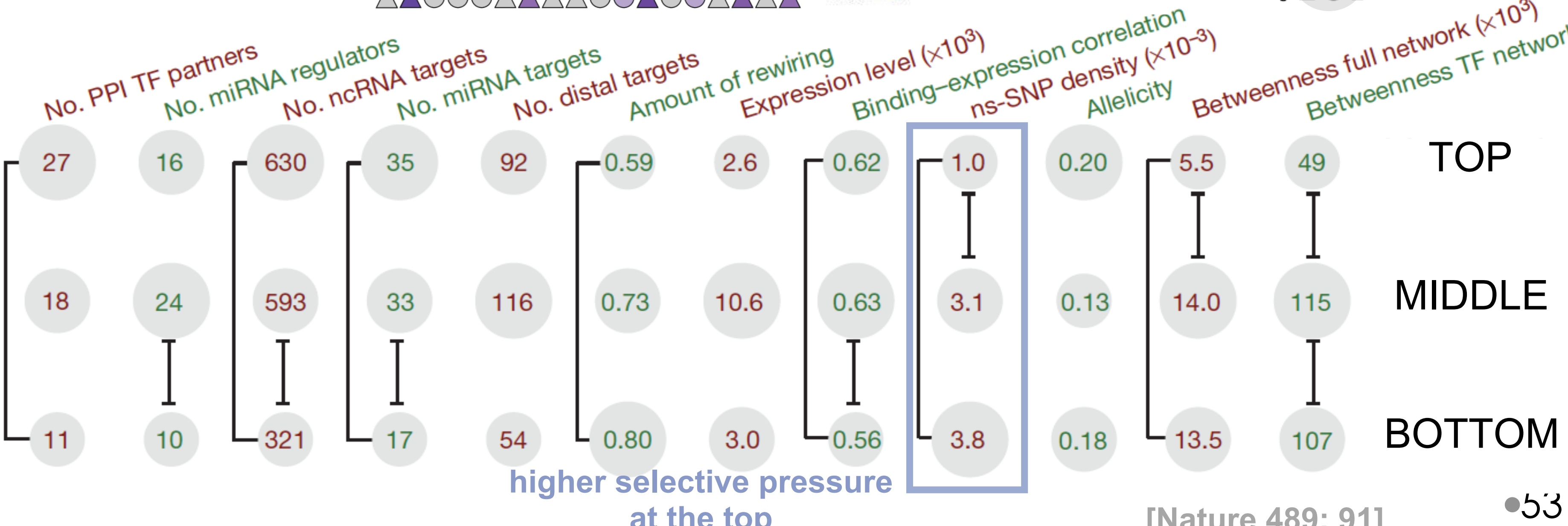
Hierarchical organization of human transcriptional regulatory network



[Nature 489: 91]

Hierarchical organization of human transcriptional regulatory network





[Nature 489: 91]

Info about content in this slide pack

- PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2015 (and beyond). Please read statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to appropriate place on gersteinlab.org).
- Paper references in the talk were mostly from Papers.GersteinLab.org.
- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt