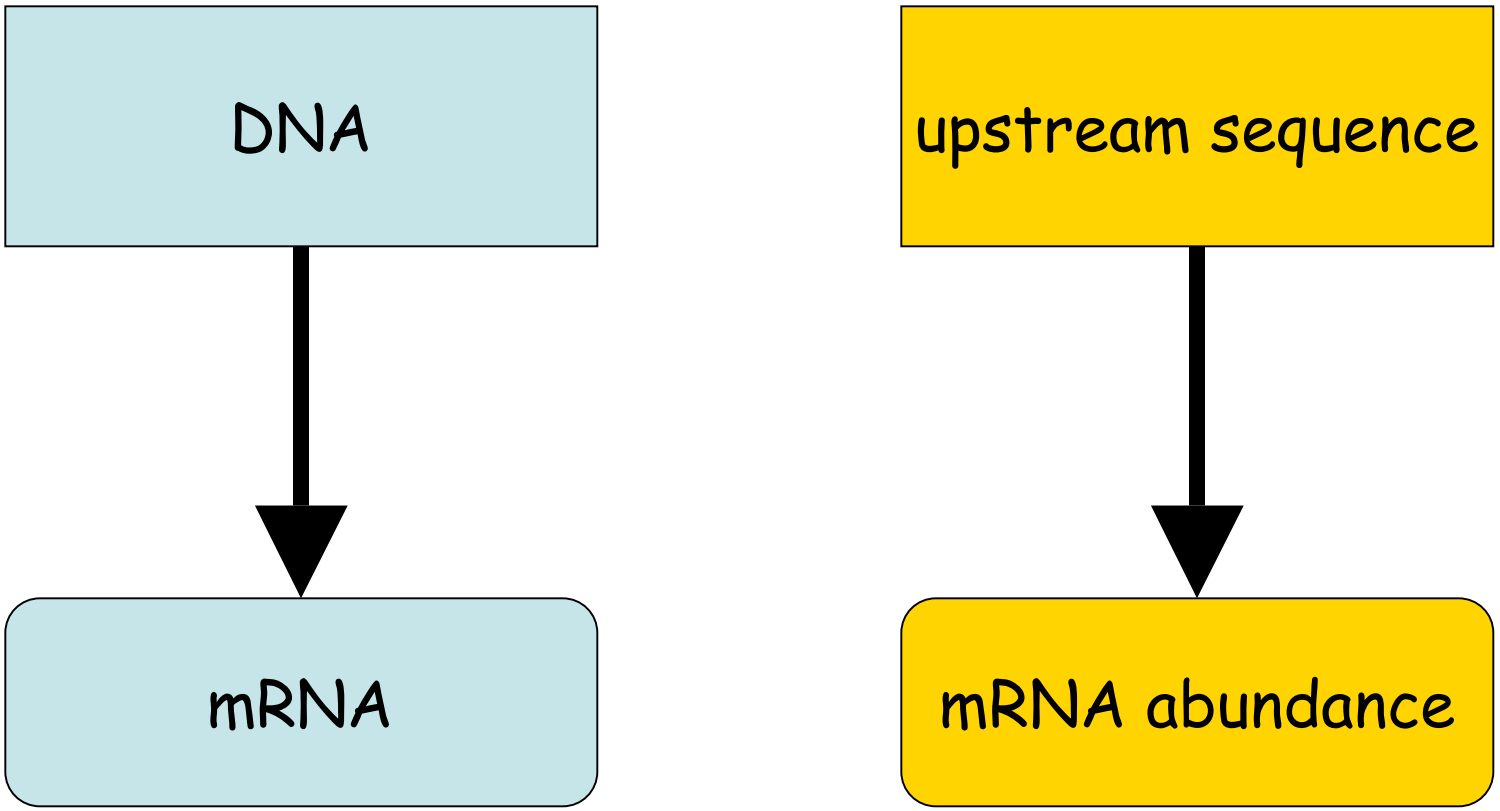
Model-based analysis of microarray data: From Central Dogma to "Omes Law"

Harmen J. Bussemaker

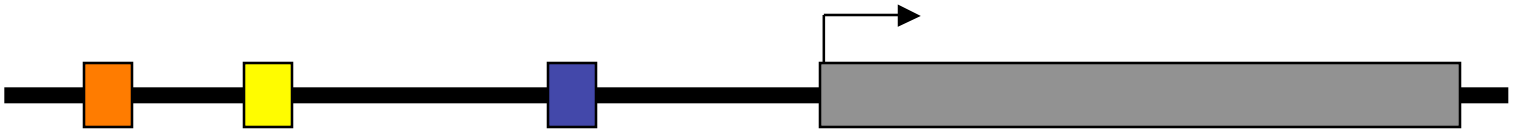
Department of Biological Sciences
Center for Comp. Biology and Bioinformatics
Columbia University

The Cell as a Molecular Computer

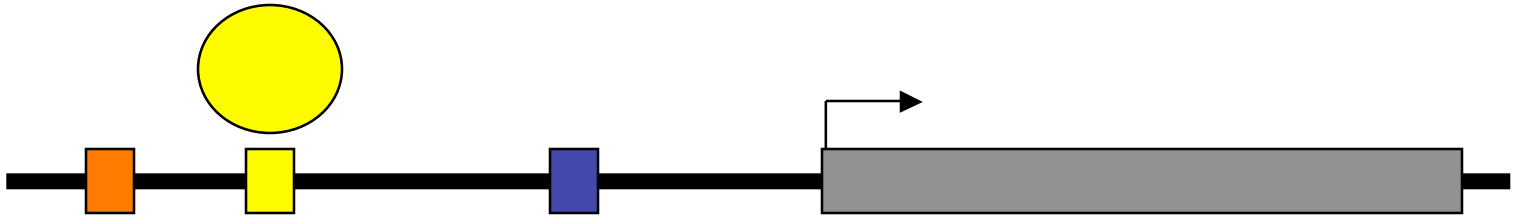


Transcribed Region

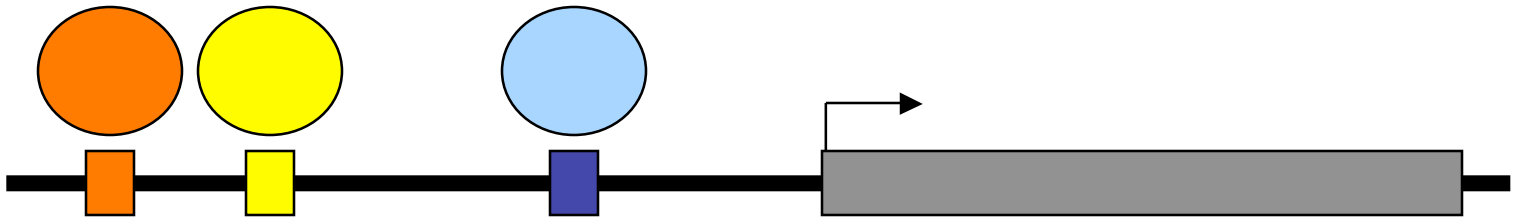
Cis-regulatory Elements

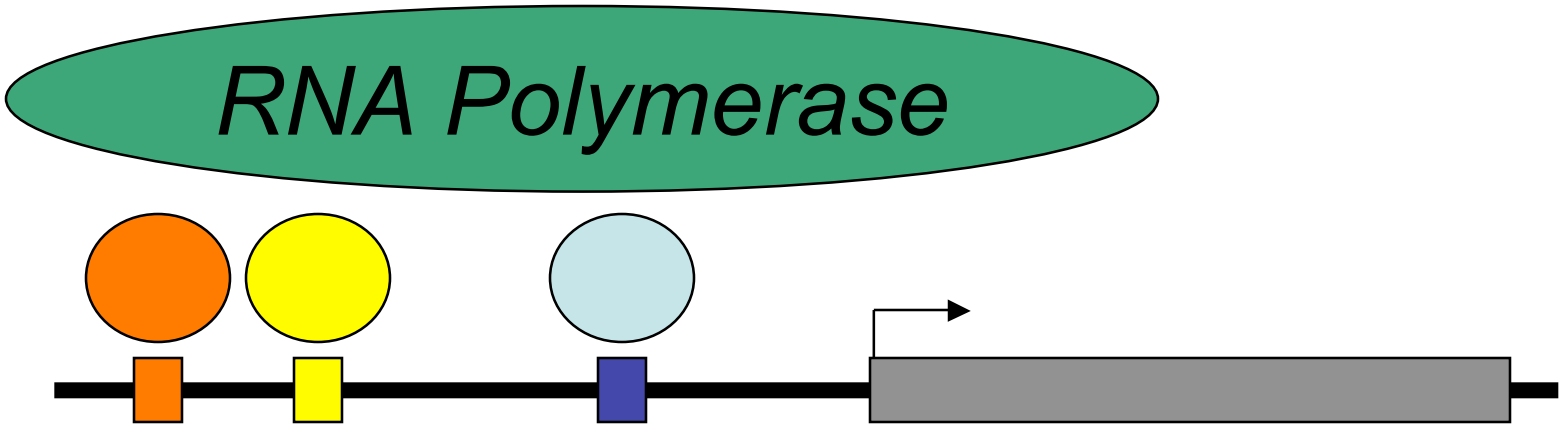


Transcription Factor

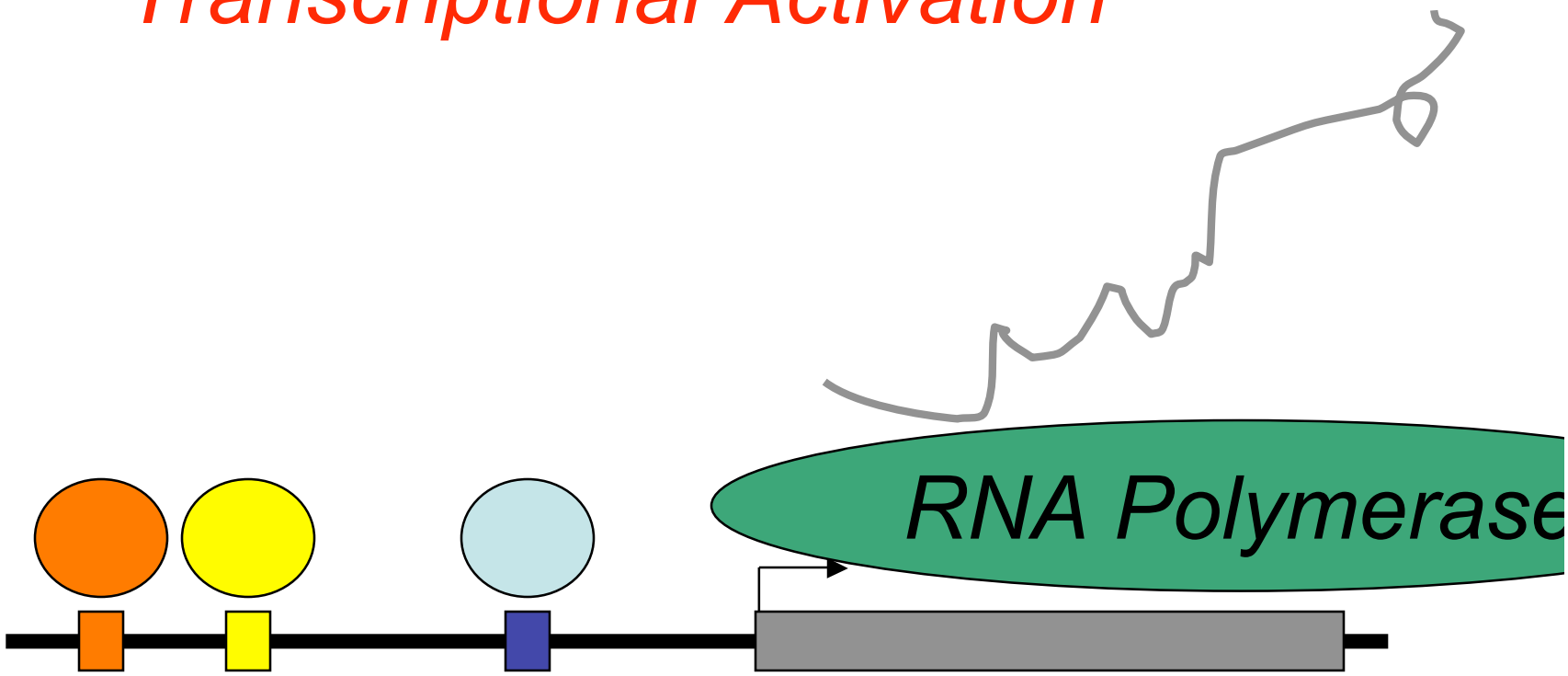


Occupancy Level

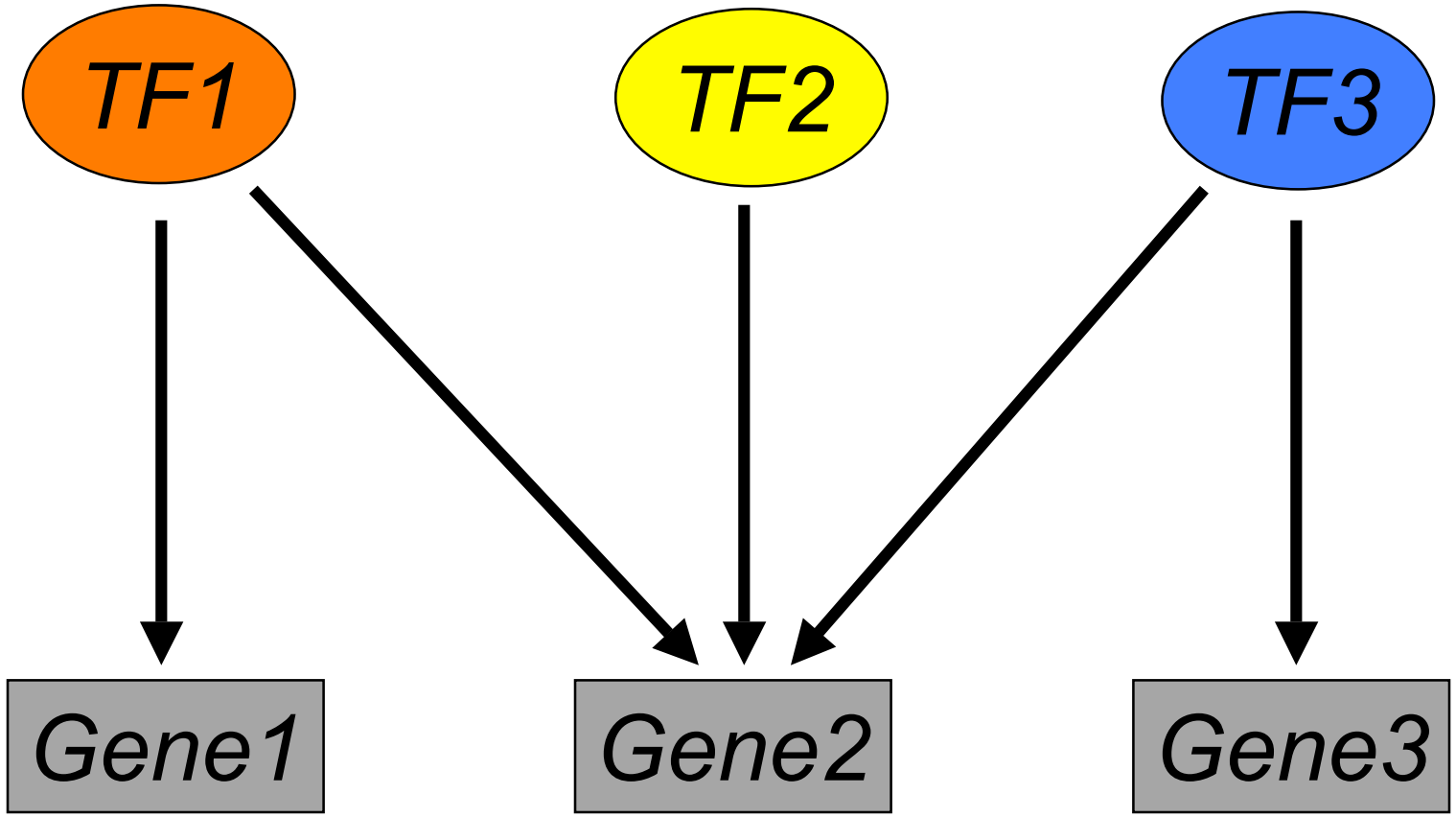



RNA Polymerase

Transcriptional Activation

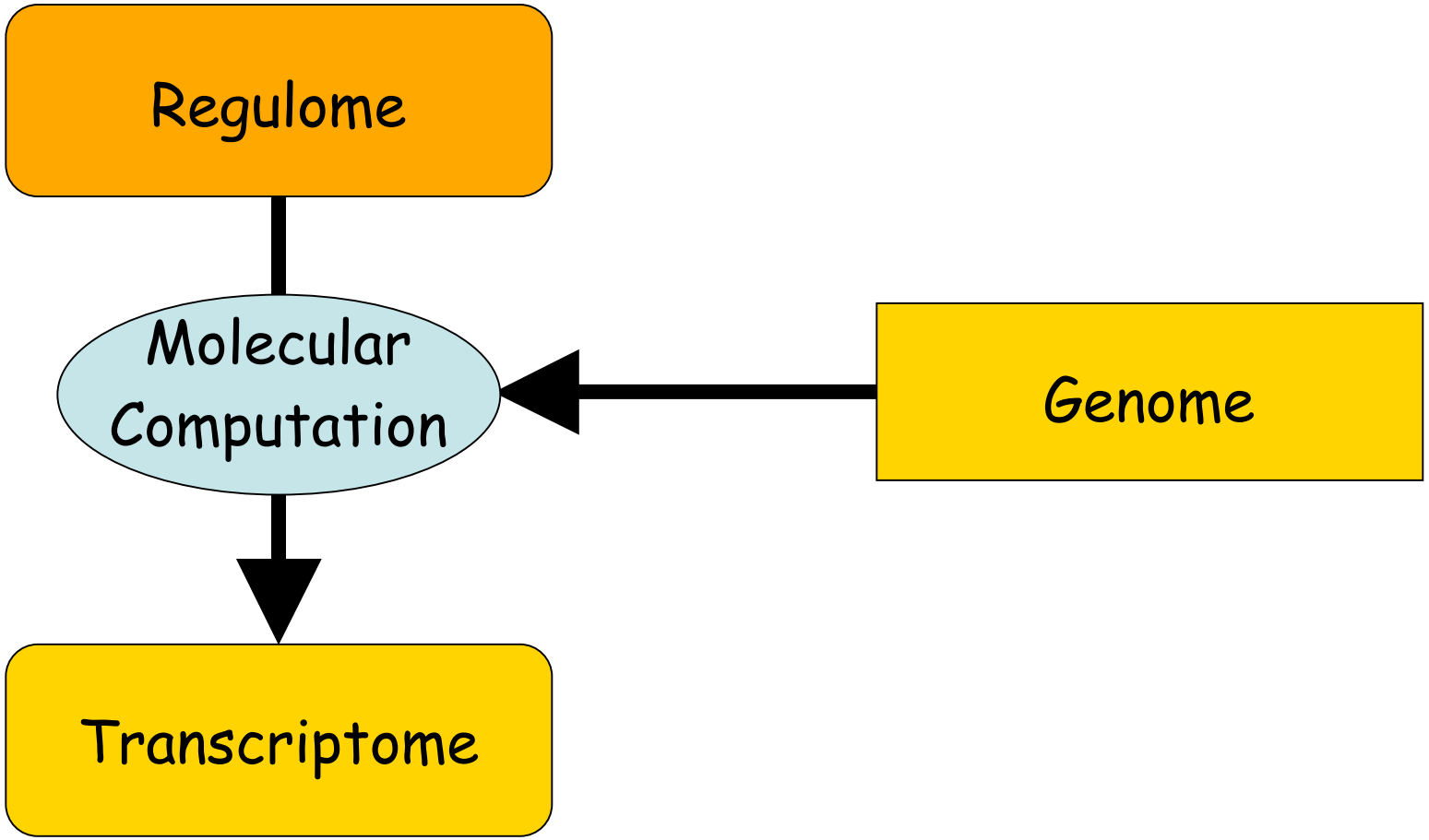


Protein, not mRNA!

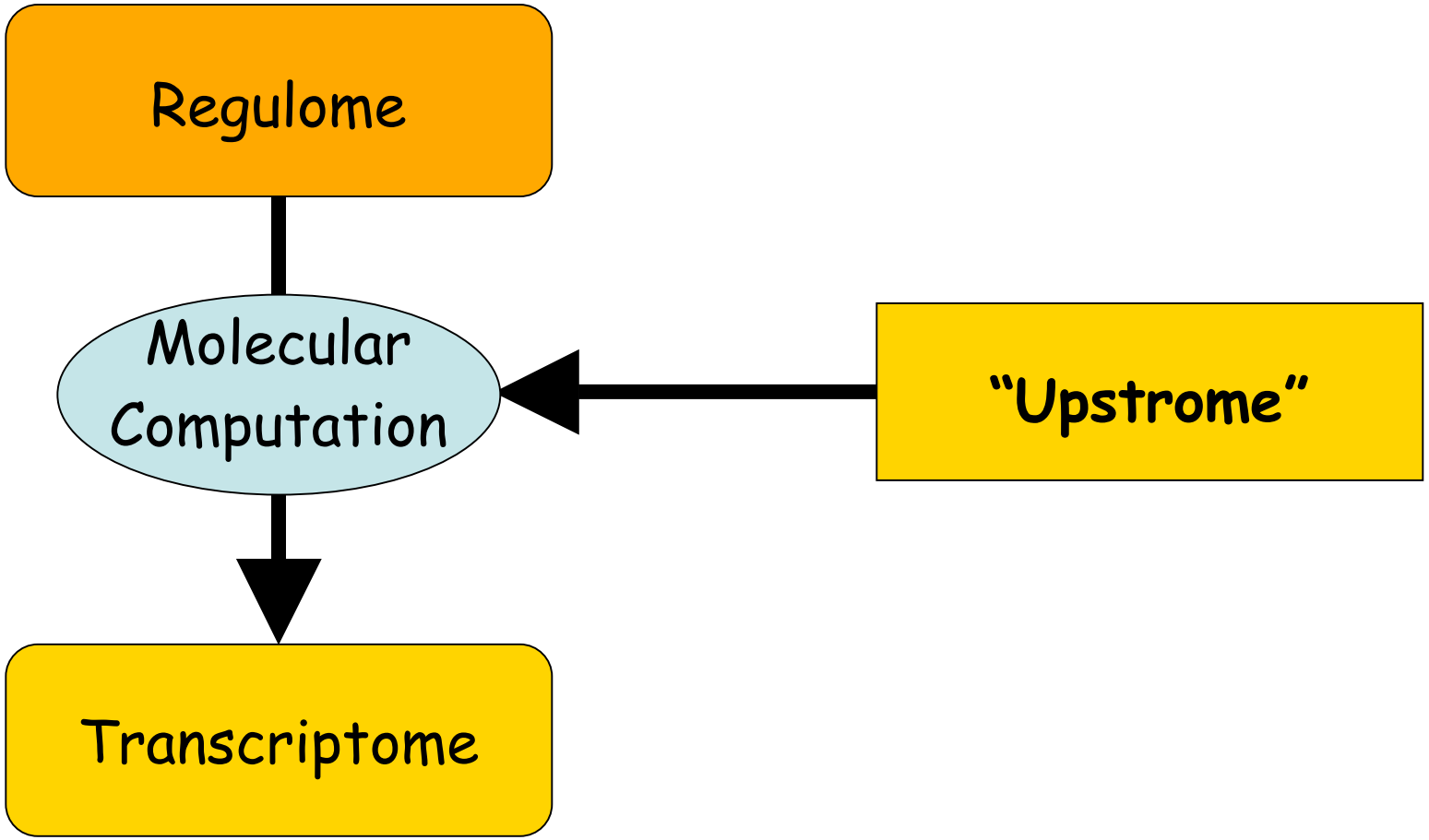


mRNA

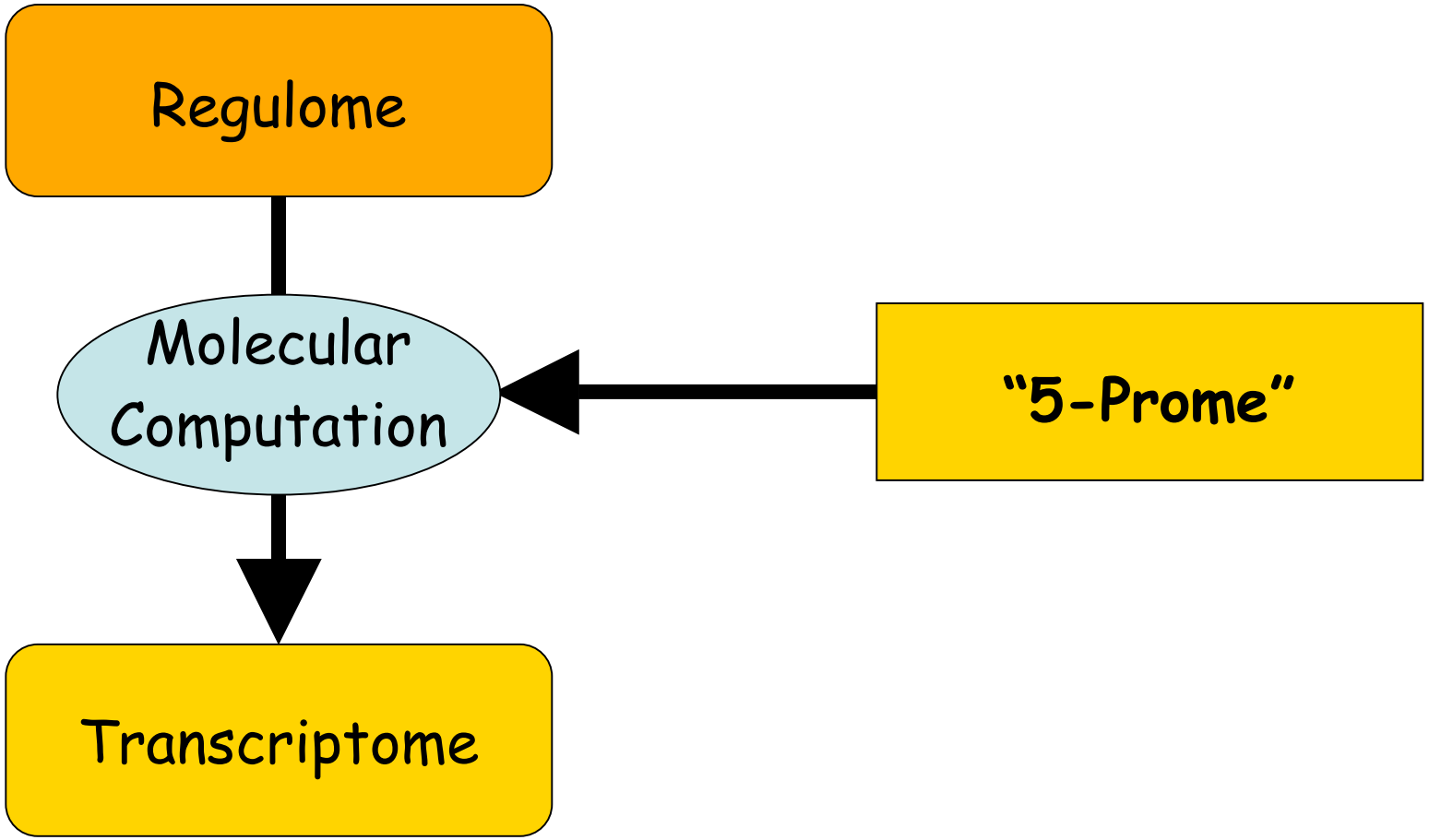
"Regulome" = concentration of all TF proteins in nucleus



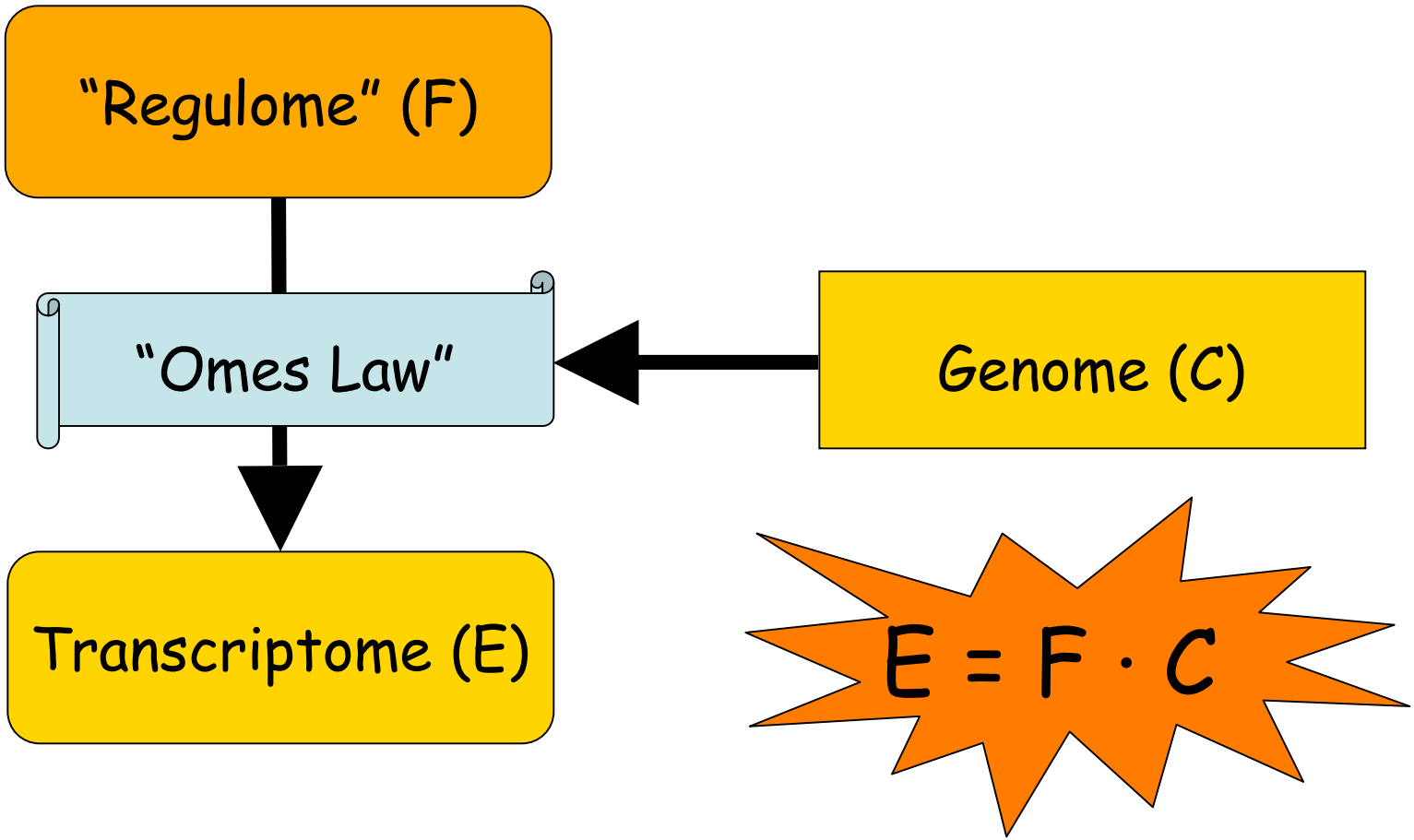
"Regulome" = concentration of all TF proteins in nucleus

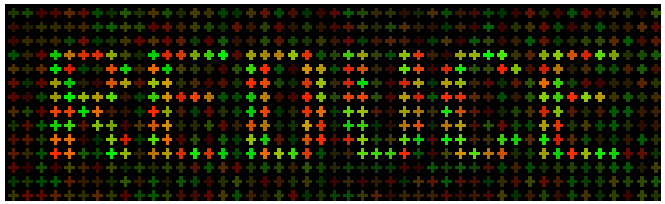


"Regulome" = concentration of all TF proteins in nucleus

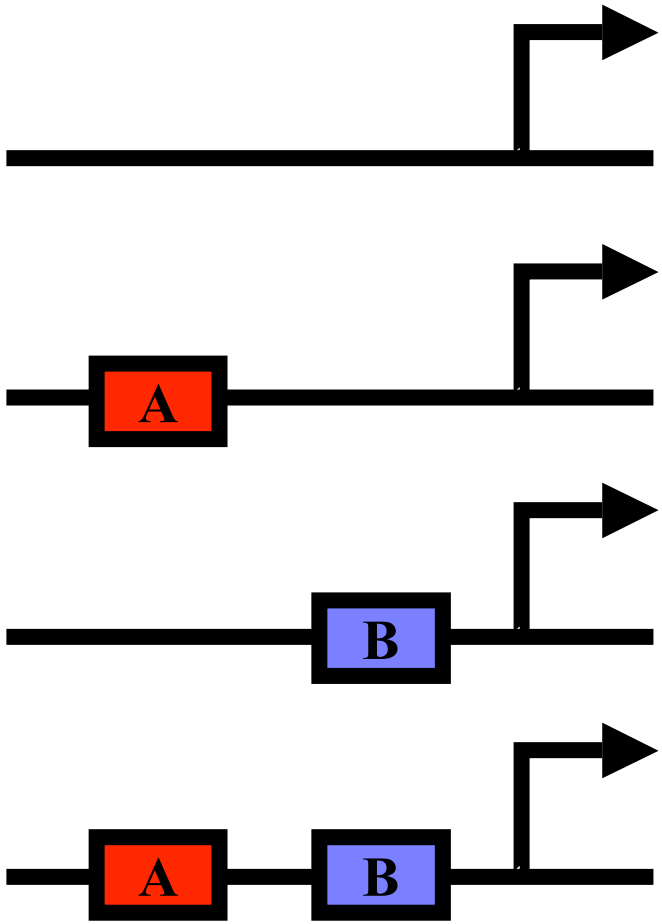


Linear Response Theory for Cells: "Omes Law"





$$E_{gt} = E_0 + \sum_{\mu} F_{\mu t} N_{\mu g}$$



$E = E_0$

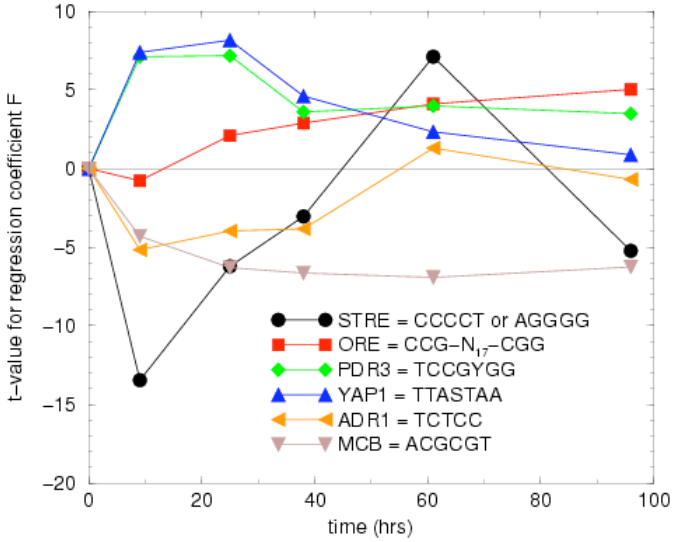
$E = E_0 + F_A$

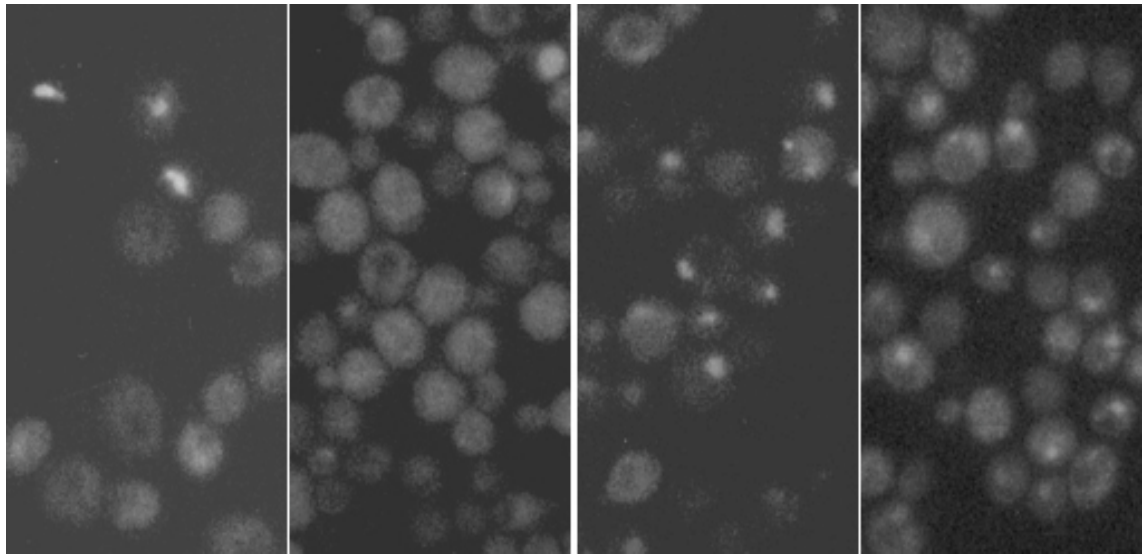
$E = E_0 + F_B$

$E = E_0 + F_A + F_B$

HJ Bussemaker, H Li, & ED Siggia, Nature Genet. (2001)

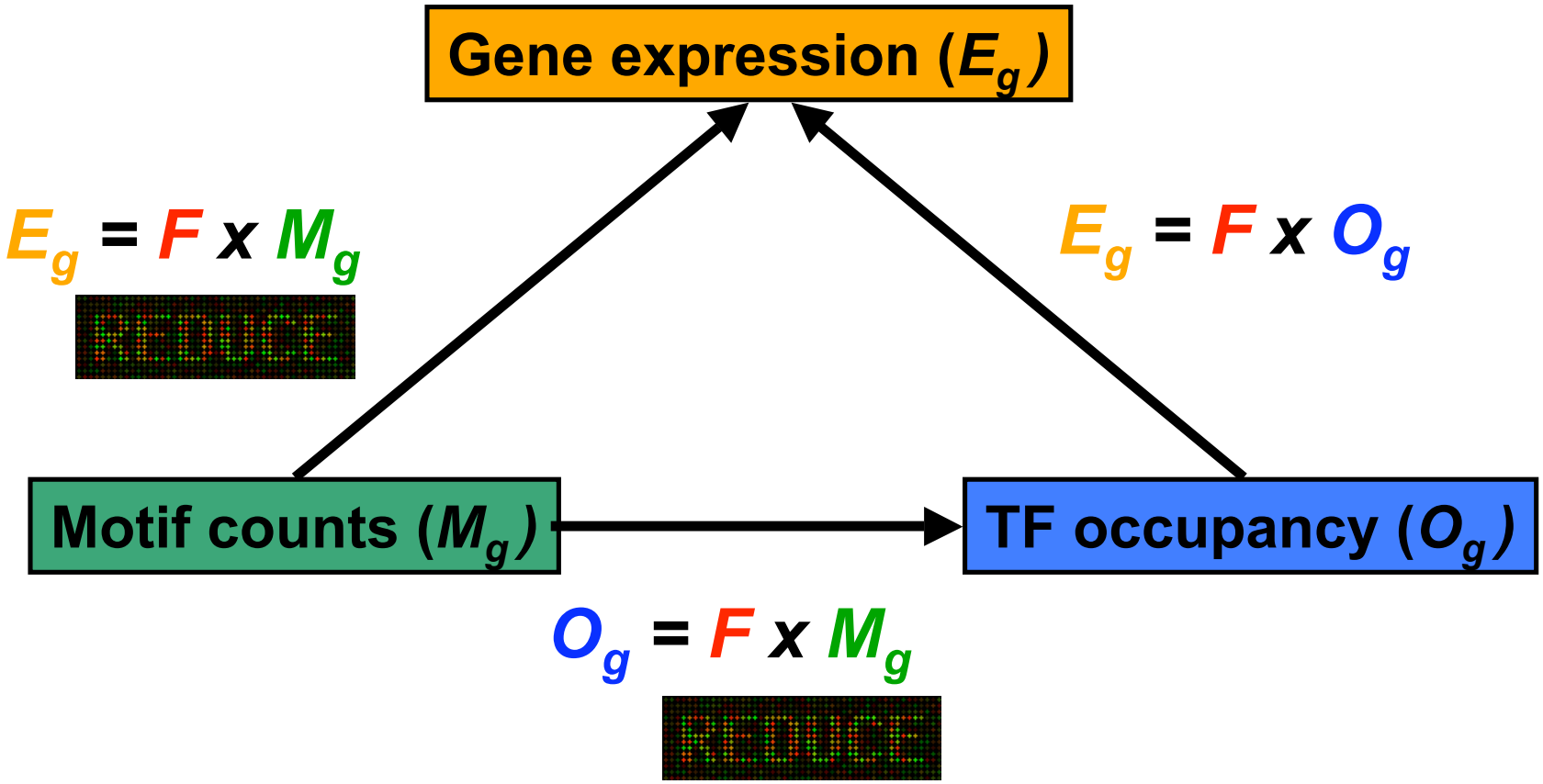
Experimental validation of TF activity profile inferred using REDUCE



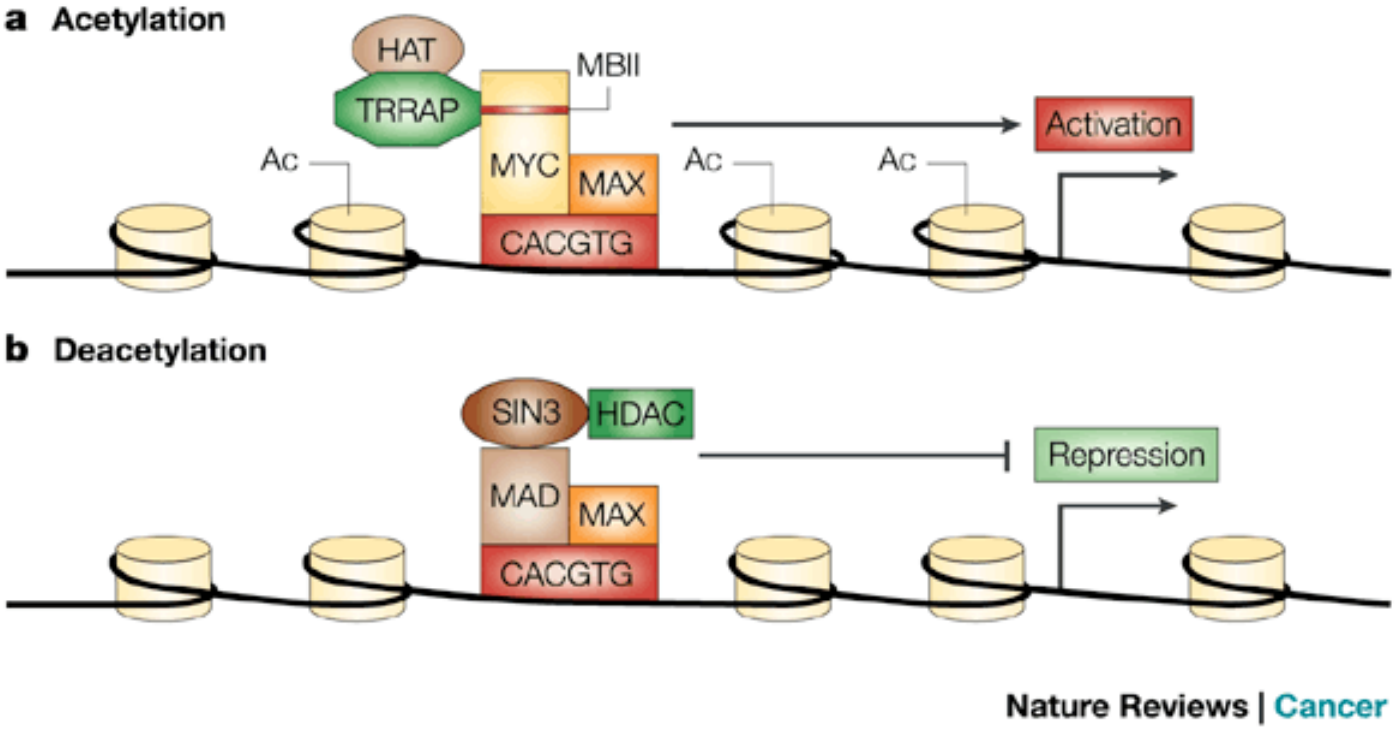


MG Koerkamp et al., Mol. Biol. Cell (2002)

Regression Analysis as a Paradigm for Microarray Data Analysis



The Myc/Mad/Max network in *D. melanogaster*



A. A. Orian, B. van Steensel, J. Delrow, H.J. Bussemaker, L. Li, T. Sawado, E. Williams, L.M. Loo, S.M. Cowley, C. Yost, S. Pierce, B.A. Edgar, S.M. Parkhurst, and R.N. Eisenman. Genes & Development (2003)

DamID: *B van Steensel, J Delrow, S Henikoff, Nature Genet. 2001*

Motif	R^2	P-value	F	Matches	Loci	Consensus
dMnt						
cgcg	0.051	0.0E+00	0.01	66,217	4,366	cg-repeat
gcgc	0.048	0.0E+00	0.01	97,925	4,367	cg-repeat
cgcgc	0.047	0.0E+00	0.02	16,661	4,121	cg-repeat
gcgcg	0.042	0.0E+00	0.02	17,392	4,129	cg-repeat
tatcgata	0.026	0.0E+00	0.06	1,618	1,258	tatcgata
atcgata	0.024	0.0E+00	0.04	3,863	2,367	tatcgata
tcgata	0.015	1.0E-12	0.02	8,262	3,541	tatcgata
tatcgat	0.015	5.0E-12	0.03	3,765	2,369	tatcgata
ggtcacac	0.024	0.0E+00	0.09	788	706	gtcacac
gtcacact	0.017	0.0E+00	0.08	691	633	gtcacac
cacgtg	0.019	0.0E+00	0.03	4,214	2,558	cacgtg
gcacgtg	0.016	0.0E+00	0.05	1,370	1,139	cacgtg
gcacgtgt	0.012	9.8E-09	0.10	319	301	cacgtg

Motif	R^2	P-value	F	Matches	Loci	Consensus
dMax						
cgcgc	0.025	0.0E+00	0.02	16,401	4,058	cg-repeat
cgcg	0.022	0.0E+00	0.01	65,098	4,301	cg-repeat
gcgcg	0.020	0.0E+00	0.02	17,086	4,069	cg-repeat
gcgc	0.014	2.0E-12	0.00	96,364	4,302	cg-repeat
tatcgata	0.024	0.0E+00	0.07	1,600	1,247	tatcgata
atcgata	0.018	0.0E+00	0.04	3,797	2,330	tatcgata
tatcgat	0.016	2.0E-12	0.04	3,705	2,334	tatcgata
ggtcacac	0.013	2.0E-09	0.08	776	693	gtcacac
gtcacact	0.009	1.1E-05	0.07	681	624	gtcacac

Motif	R^2	P-value	F	Matches	Loci	Consensus
dMyc (low dMax)						
aa	0.066	0.0E+00	0.00	2,763,243	4,332	at-rich
a	0.065	0.0E+00	0.00	8,493,006	4,332	at-rich
t	0.063	0.0E+00	0.00	8,153,995	4,332	at-rich
aat	0.063	0.0E+00	0.00	729,002	4,332	at-rich
tt	0.062	0.0E+00	0.00	2,601,186	4,332	at-rich

Motif	R^2	P-value	F	Matches	Loci	Consensus
dMyc (high dMax)						
cacgtg	0.032	0.0E+00	0.03	4,203	2,555	cacgtg
acgtg	0.024	0.0E+00	0.01	18,323	4,228	cacgtg
cacgt	0.022	0.0E+00	0.01	17,082	4,221	cacgtg
gcacgtg	0.021	0.0E+00	0.05	1,365	1,134	cacgtg
atcgata	0.022	0.0E+00	0.03	3,853	2,362	tatcgata
tcgata	0.022	0.0E+00	0.02	8,255	3,535	tatcgata
tatcgata	0.020	0.0E+00	0.04	1,616	1,256	tatcgata
atcgat	0.014	2.2E-11	0.01	13,336	4,037	tatcgata
tatcgat	0.013	2.1E-10	0.02	3,751	2,362	tatcgata
cgcgc	0.027	0.0E+00	0.01	16,646	4,116	cg-repeat
cgcg	0.026	0.0E+00	0.00	66,118	4,361	cg-repeat
gcgc	0.024	0.0E+00	0.00	97,752	4,362	cg-repeat
gcgcg	0.019	0.0E+00	0.01	17,348	4,124	cg-repeat

Known: CACGTG (E-box)

Unexpected: TATCGATA (DRE)

Novel: GTCACAC (???)

Both from DamID & mRNA data!

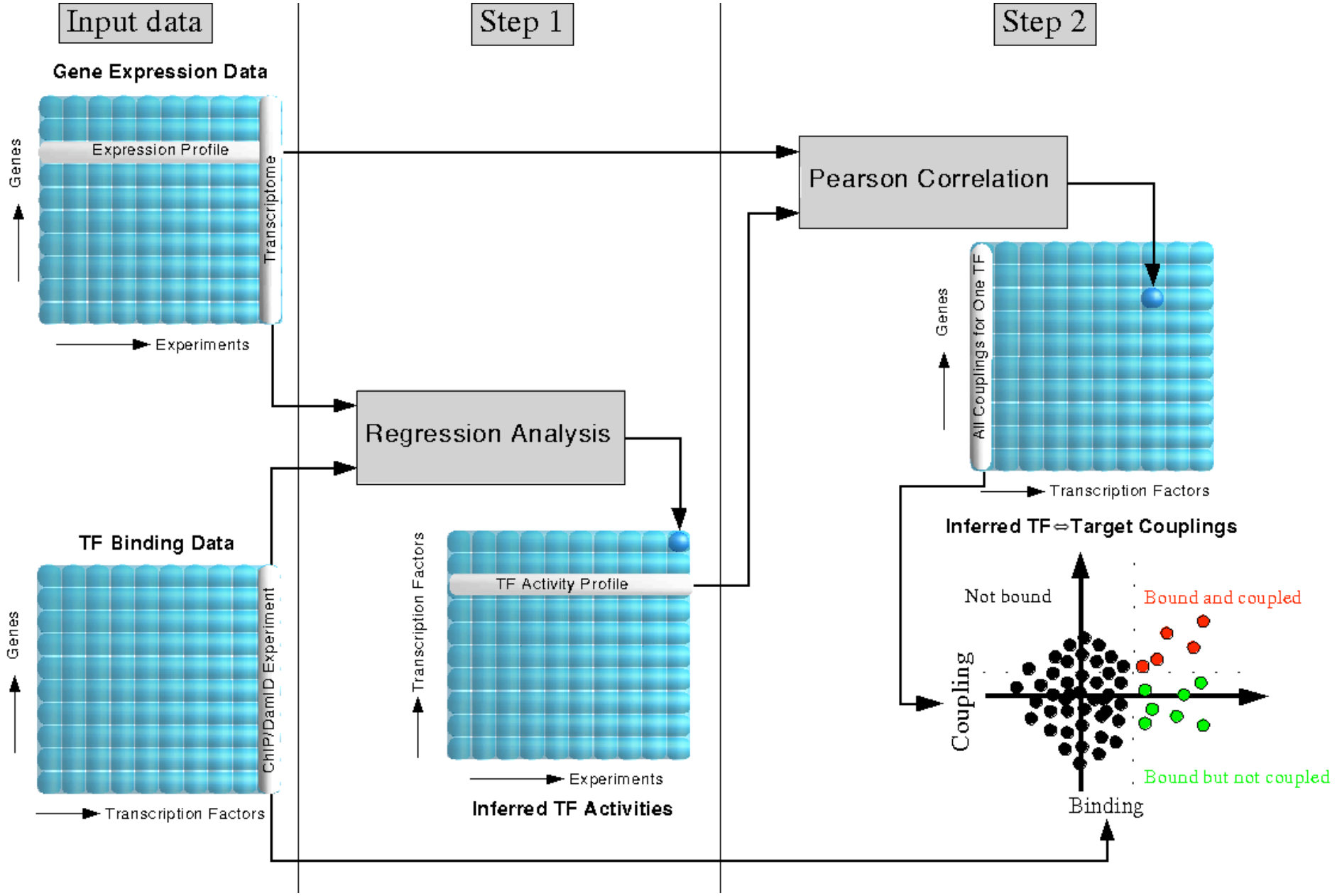
"MA-Networker": Integrating mRNA expression and ChIP data

F Gao, BC Foat, and HJ Bussemaker, BMC Bioinformatics (2004)

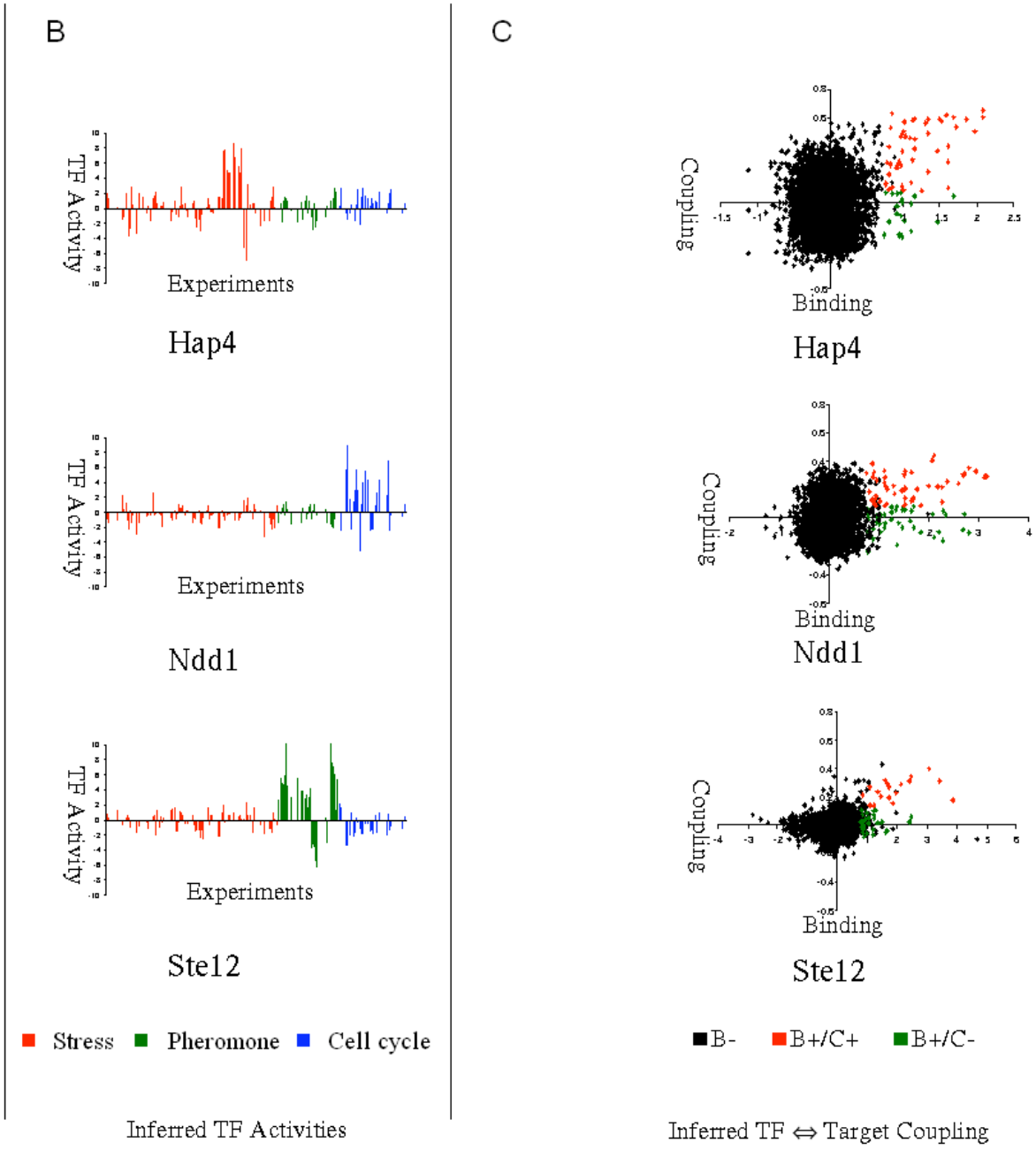
- ChIP data for >100 TFs (Lee et al., *Science* (2002))
- mRNA expression data for ~800 conditions

- TF activity profiles from regression (A~O)
- Response of individual genes to TF activity profile

- Increase specificity of TF target prediction
- Overcome context dependence of TF deletion experiments related to combinatorial control



F Gao, BC Foat, and HJ Bussemaker, BMC Bioinformatics (2004)


B
TF Activity
Experiments
Hap4
TF Activity
Experiments
Ndd1
TF Activity
Experiments
Ste12
Stress
Pheromone
Cell cycle
Inferred TF Activities
C
Coupling
Binding
Hap4
Coupling
Binding
Ndd1
Coupling
Binding
Ste12
B-
B+/C+
B+/C-
Inferred TF ⇔ Target Coupling

MA-Networker results

- Only 37 out of 113 transcription factors are statistically significant predictor of gene expression in one or more conditions
- On average only 58% of significantly bound genes are **functional** targets

F Gao, BC Foat, and HJ Bussemaker, BMC Bioinformatics (2004)

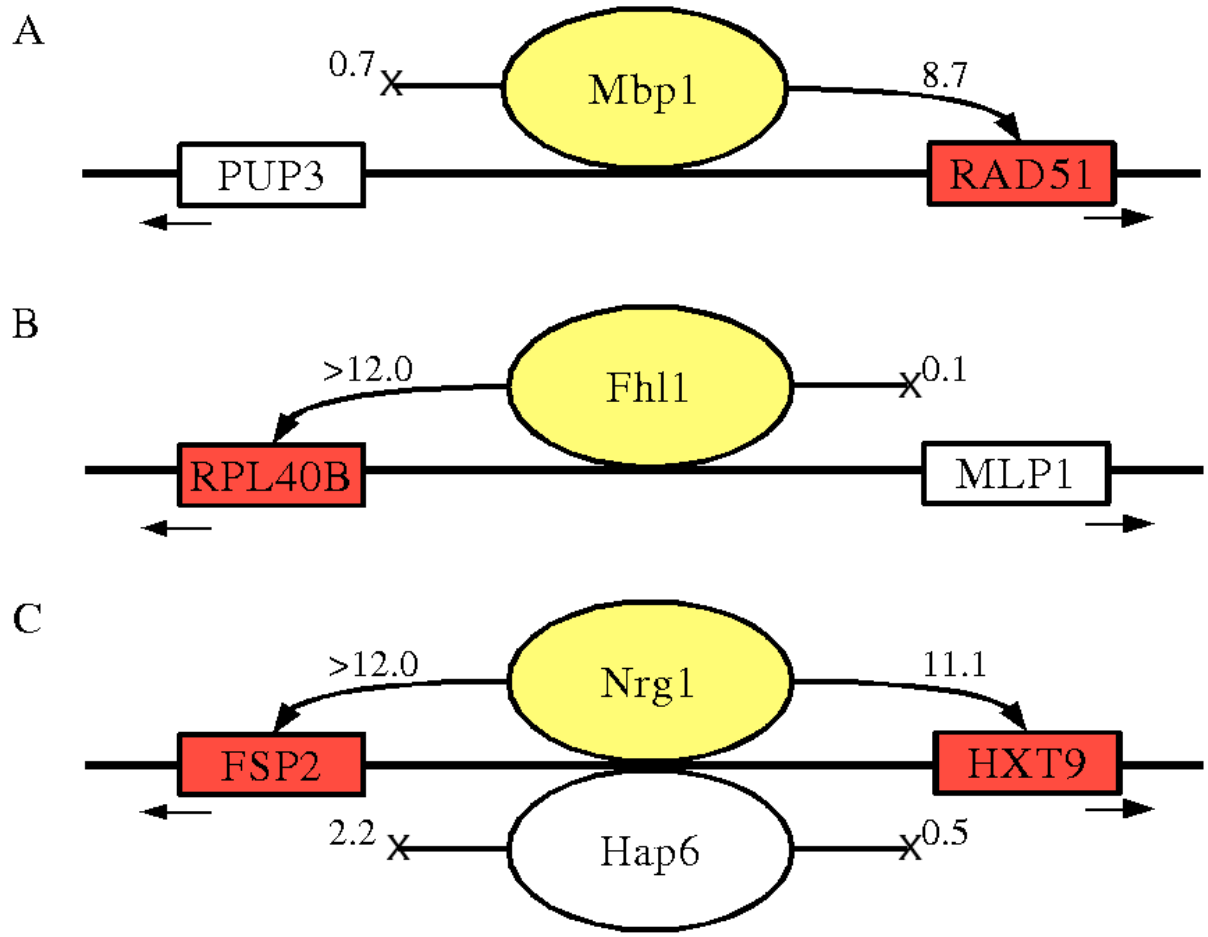
MA-Networker validation

Among **non-functional** TF targets, there is:

- **No** enrichment for GO categories
- **No** transcriptional response to TF deletion
- **Less** DNA motif over-representation

F Gao, BC Foat, and HJ Bussemaker, BMC Bioinformatics (2004)

Divergently transcribed gene pairs



F Gao, BC Foat, and HJ Bussemaker, BMC Bioinformatics (2004)

Identification of TF interactions

$$E_g = E_0 + F_1 O_{1g} + F_2 O_{2g} + F_{1:2} O_{1g} O_{2g}$$

Proof of principle:

Yeast cell cycle: Mbp1, Swi4, Swi6

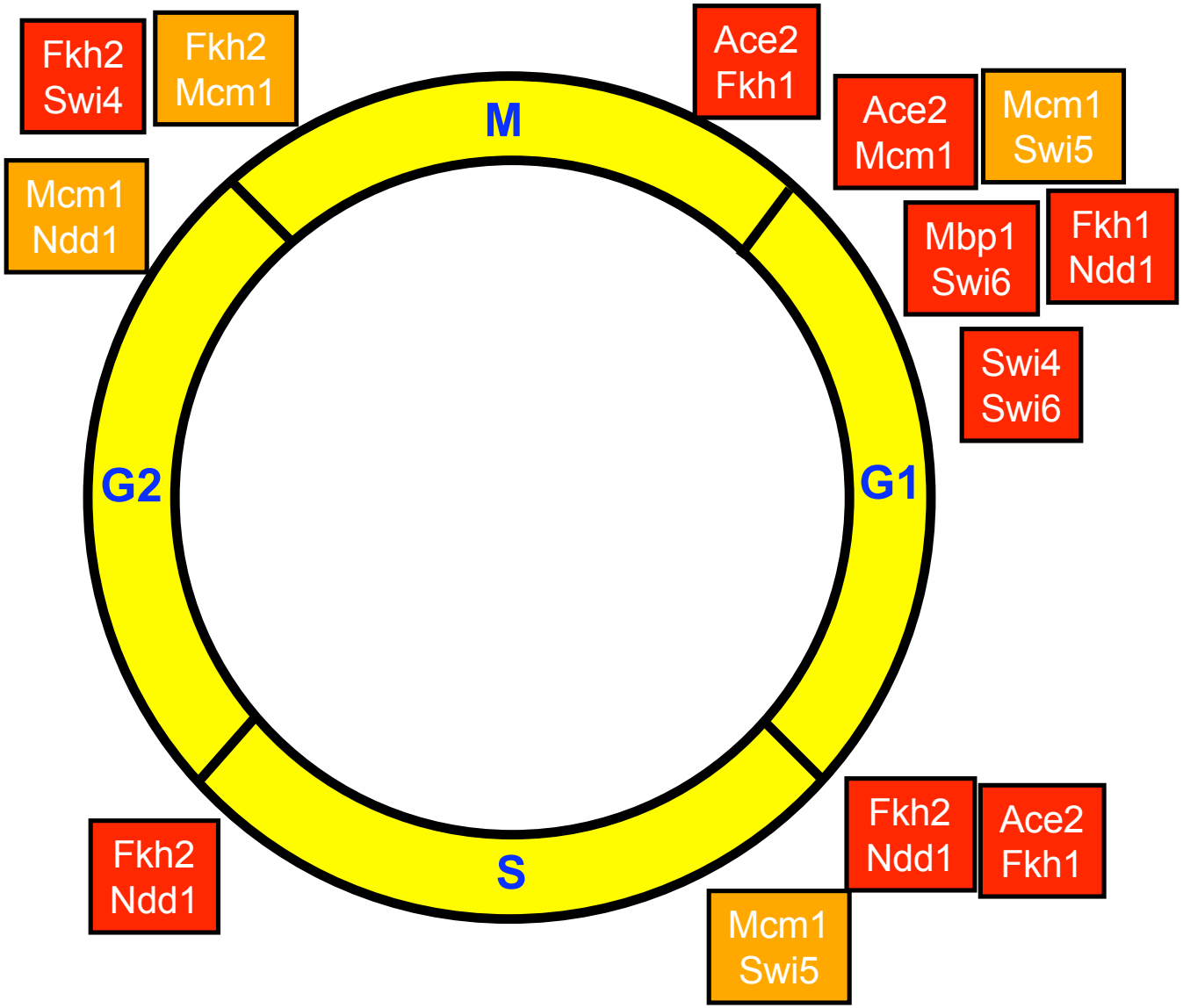
MBF = Mbp1::Swi6

SBF = Swi4::Swi6

Mpb1-Swi6 interaction inferred from ChIP & Mbp1 deletion data

Potential elements	t value	Pr (>|t|)
Mbp1	0.52	0.6010
Swi4	1.82	0.0694
Swi6	-0.48	0.6289
Mbp1+Swi4	-1.15	0.2490
Mbp1+Swi6	3.56	0.0004
Swi4+Swi6	-0.04	0.9665
Mbp1+Swi4+Swi6	0.54	0.5895

Detect Context-Dependent Interaction



A Gene Expression Map for the Euchromatic Genome of *Drosophila melanogaster*

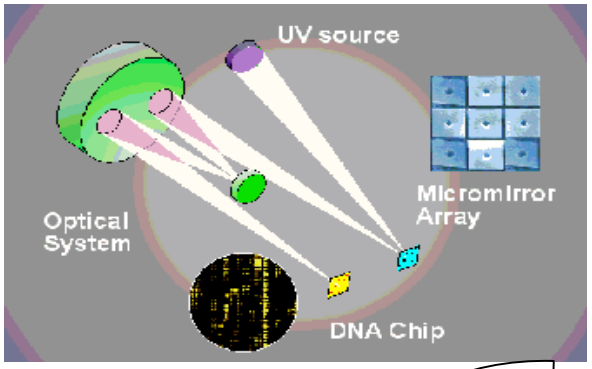
Viktor Stolc[1,2], Zareen Gauhar[2,3], Christopher Mason[3], Gabor Halasz[4,5], Marinus F. van Batenburg[4,9], Scott A Rifkin[3,6], Sujun Hua[3], Tine Herreman[3], Waraporn Tongprasit[7], Paolo Barbano[3,8], Harmen J. Bussemaker[4,10], and Kevin P White[3,6]

1. Genome Research Facility, NASA Ames Research Center
2. Department of Molecular, Cellular, and Developmental Biology, Yale University
3.Department of Genetics, Yale University School of Medicine
4.Department of Biological Sciences, Columbia University
5.Integrated Program in Cellular, Molecular and Biophysical Studies, Columbia University
6. Department of Ecology and Evolutionary Biology, Yale University
7. Eloret Corporation, at NASA Ames Research Center
8. Department of Mathematics, Yale University
9. Bioinformatics laboratory, Academic Medical Center, University of Amsterdam
10. Center for Computational Biology and Bioinformatics, Columbia University

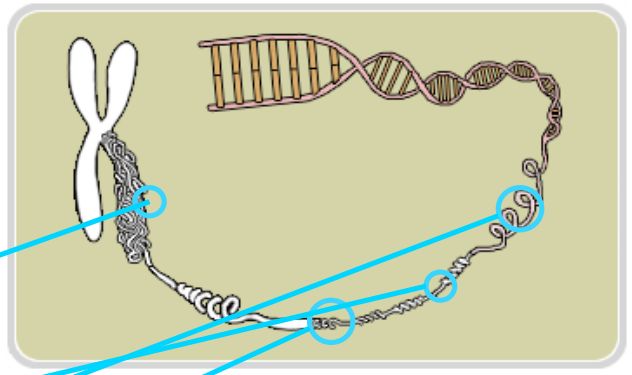
Science, Oct 22, 2004

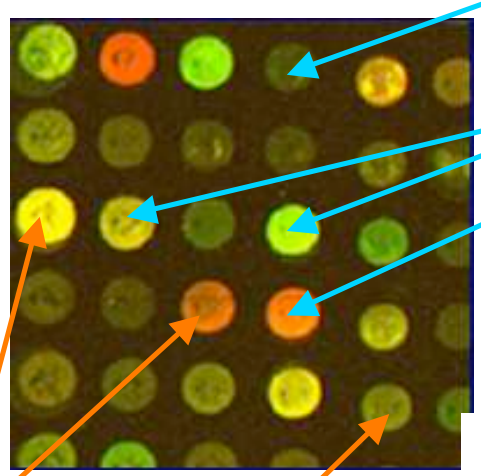
Experimental Setup

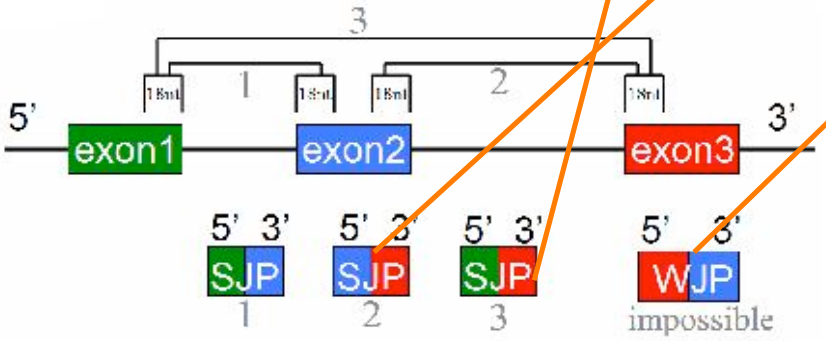
MAS

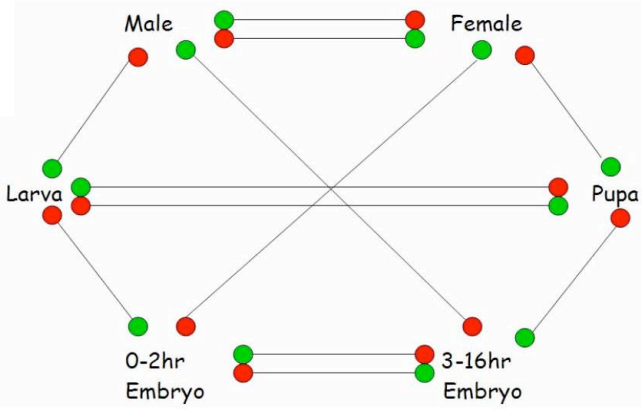


Tiling Probes





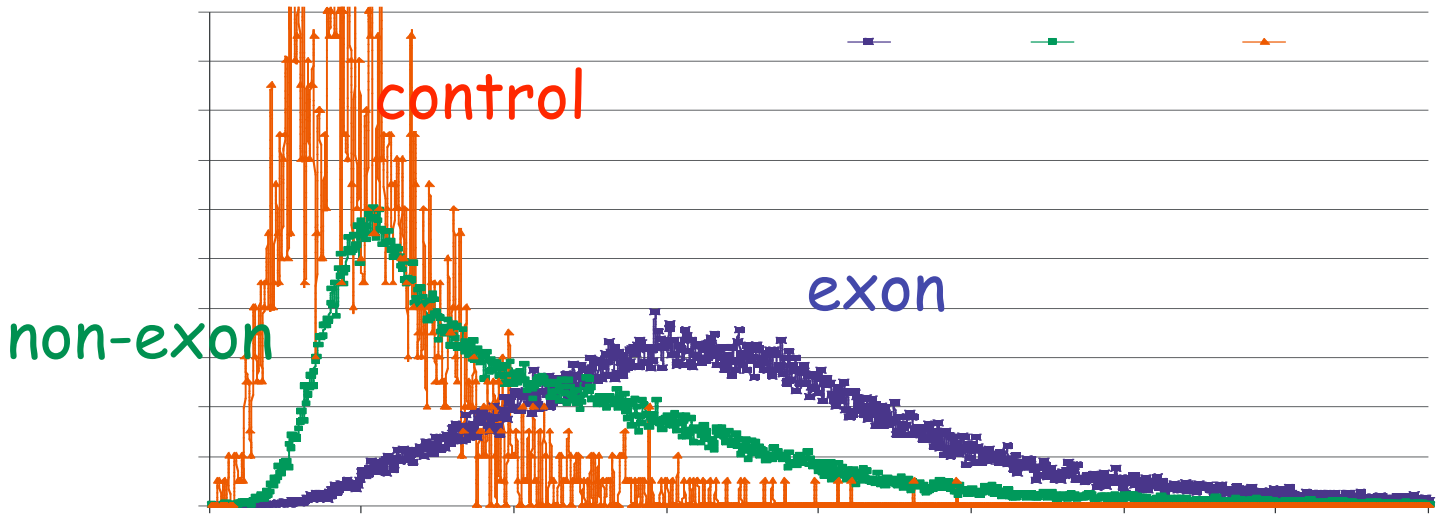




Splice Forms

Developmental Stages

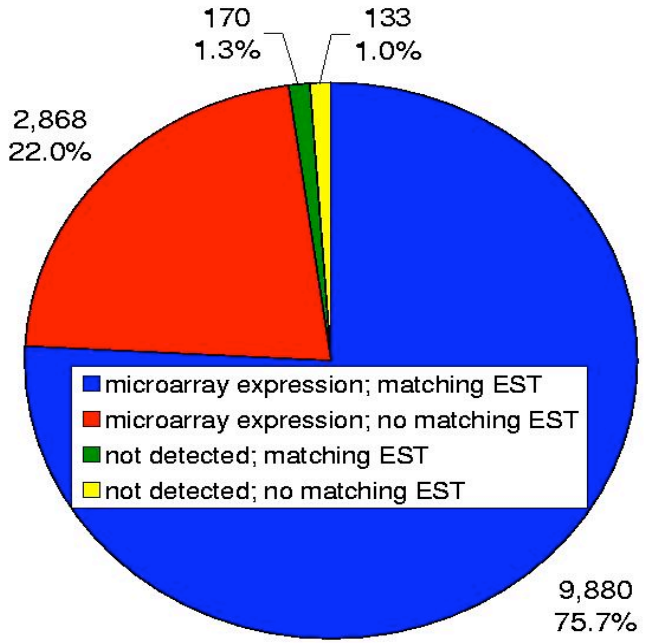
Detecting Probes Expressed Above Background



- Model-based correction for probe sequence bias
- Convert probe intensities to P-values separately for each channel
- Combine P-values for all 24 channels into single P-value
- FDR procedure

In fruitfly, 41 % of the non-coding genome is being transcribed

	Total Probes	Significantly Expressed	% *Expressed*	Differentially Expressed	% Differentially Expressed
Exon	61371	47419	*77*	21176	31
Tiling	87814	35985	*41*	5508	15
Junction	30788	8732	*28*		

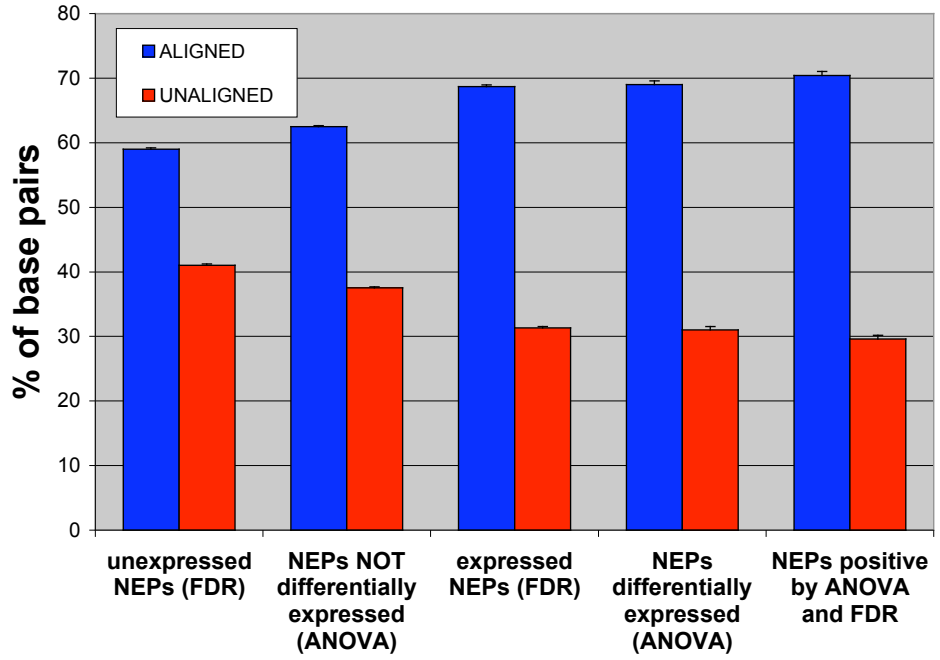


Exon expressio: Comparison with EST databases from the Berkeley Drosophila Genome Project

Over a quarter of splice junctions show expression, including 5,440 previously undetected splice variants.

15% of expressed non-coding transcripts ios developmentally regulated

Expressed non-coding regions are evolutionarily constrained



$P=10^{-15}$ (t-test)

Based on alignment between *D. melanogaster* and *D. pseudoobscura*

S. Richards et al., Genome Research (in press)

Acknowledgements

Bussemaker Lab:

Barrett Foat

Gabor Halasz

Ron Tepper

Marcel van Batenburg

Junbai Wang

Xiang-Jun Lu

Feng Gao

Crispin Roven

Collaborators:

Kevin White (Yale)

Bas van Steensel (NKI)

Robert Eisenman (FHCRC)

Frank Holstege (Utrecht)

Henk Tabak (A'dam/Utrecht)

Antoine van Kampen (AMC)

Rogier Versteeg (AMC)

Funding: NIH, HFSP, NWO

Looking for Postdoc (CS/Physics background) !