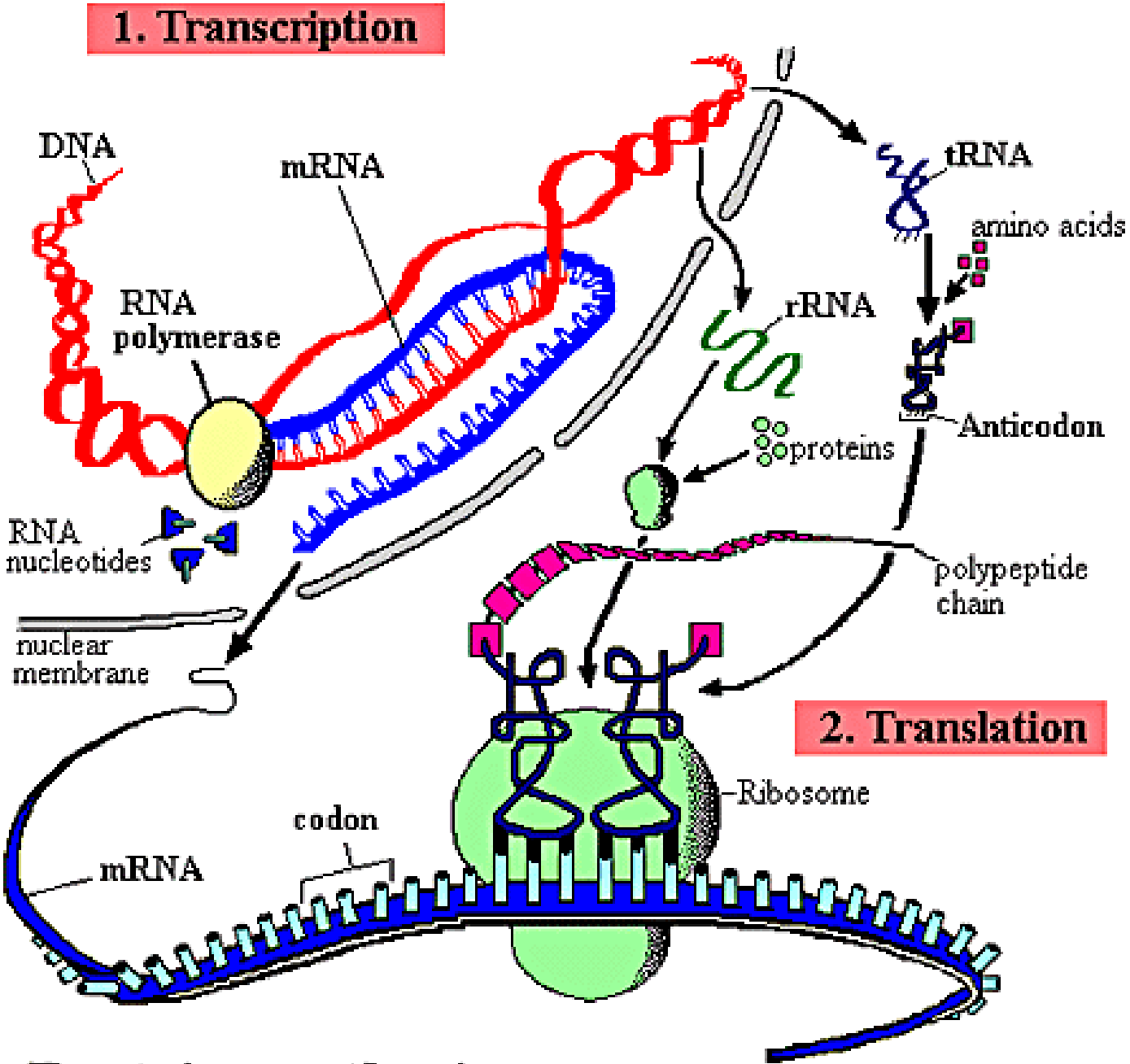
Using Regression Analysis to Infer Regulatory Circuitry from Genome-Wide Expression and Binding Data

Harmen J. Bussemaker

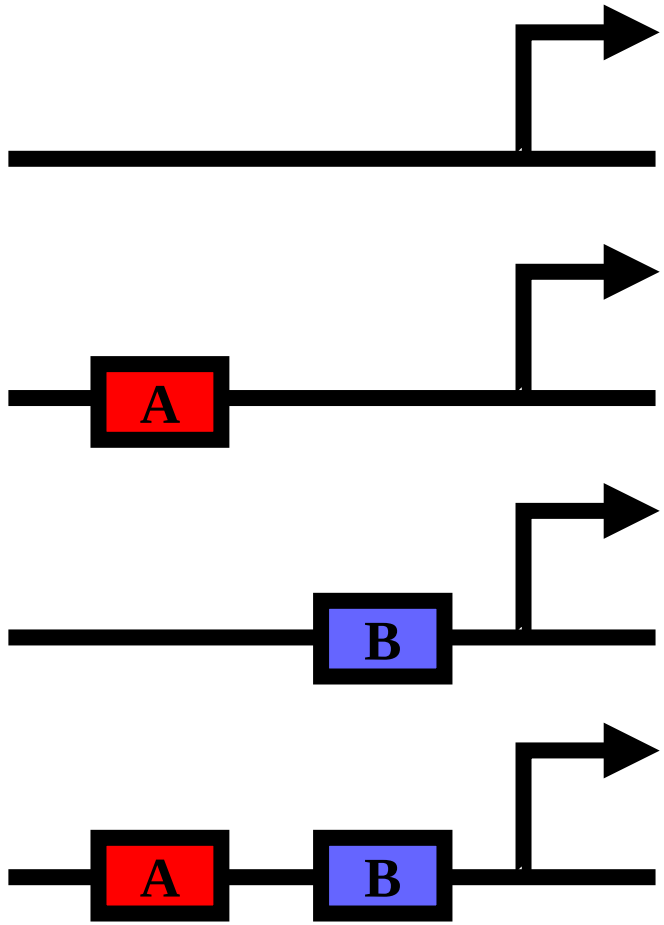
Department of Biological Sciences, and

Center for Computational Biology and Bioinformatics

Columbia University



Protein synthesis



$A = C$

$A = C + F_A$

$A = C + F_B$

$A = C + F_A + F_B$

$$A_{gt} = C + \sum_m F_{mt} N_{mg} + noise$$

Top motifs for G1 phase of cell cycle

(Spellman et al, *MBC* 1998)

AAAATTT

AAATTTT

ACGCGT

CGATGAG

GATGAGC

AAATTT

AGGGG

GATGAG

AAAATT

ACGCG

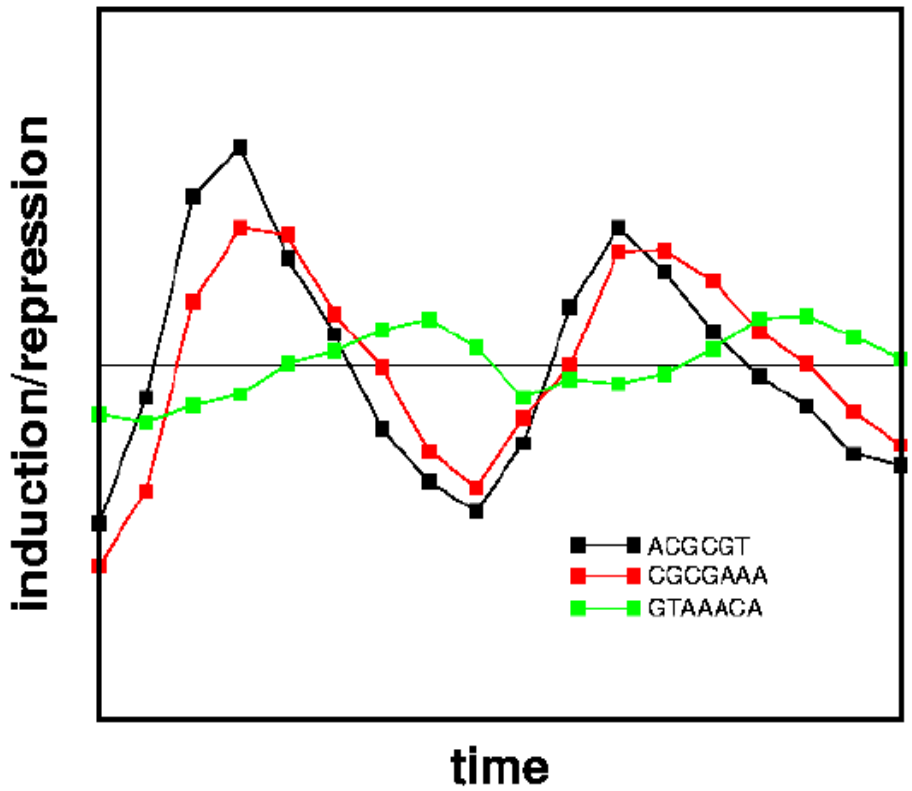
Subtract AAATTTT signal...

ACGCGT
AGGGG
AAGGGG
ACGCG
CGATGAG
GATGAGC
GGGG
CGCGT
TGACGCG
GACGCGT

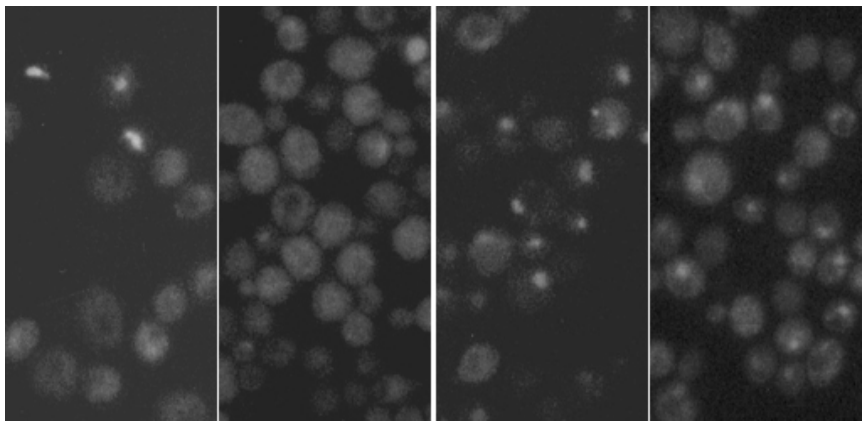
Final model resulting from iteration...

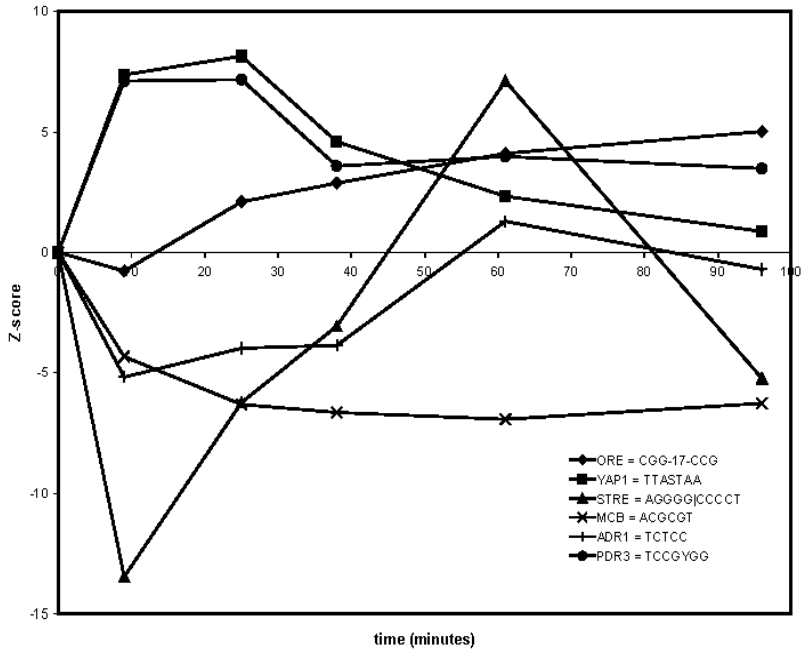
AAAATTT
ACGCGT
AGGGG
CGATGAG
CTCATCG
CCTCGAC
CCCCT
TAAACAA

Module activity time course



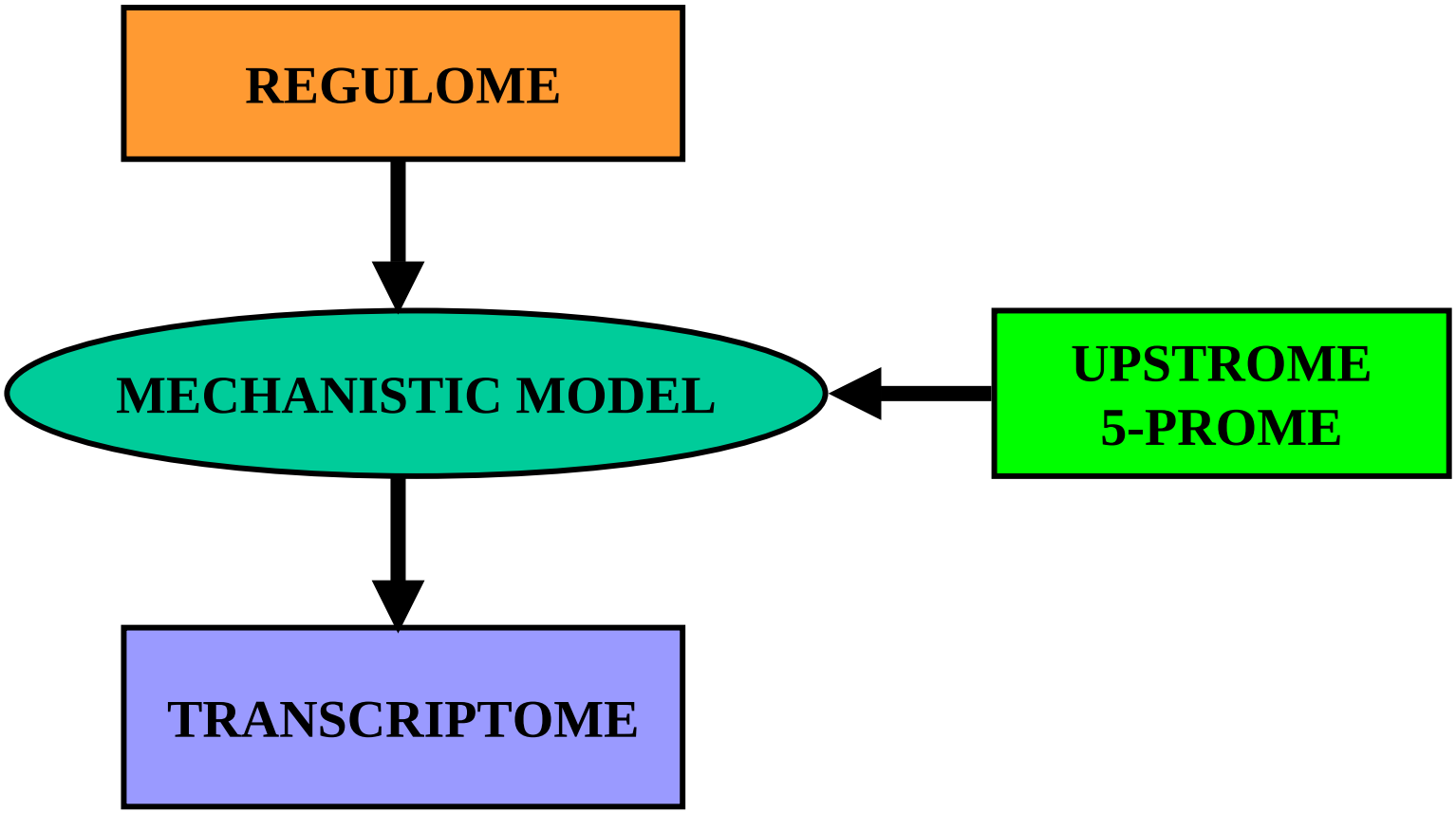
Koerkamp et al., *Mol. Biol. Cell* (2002)



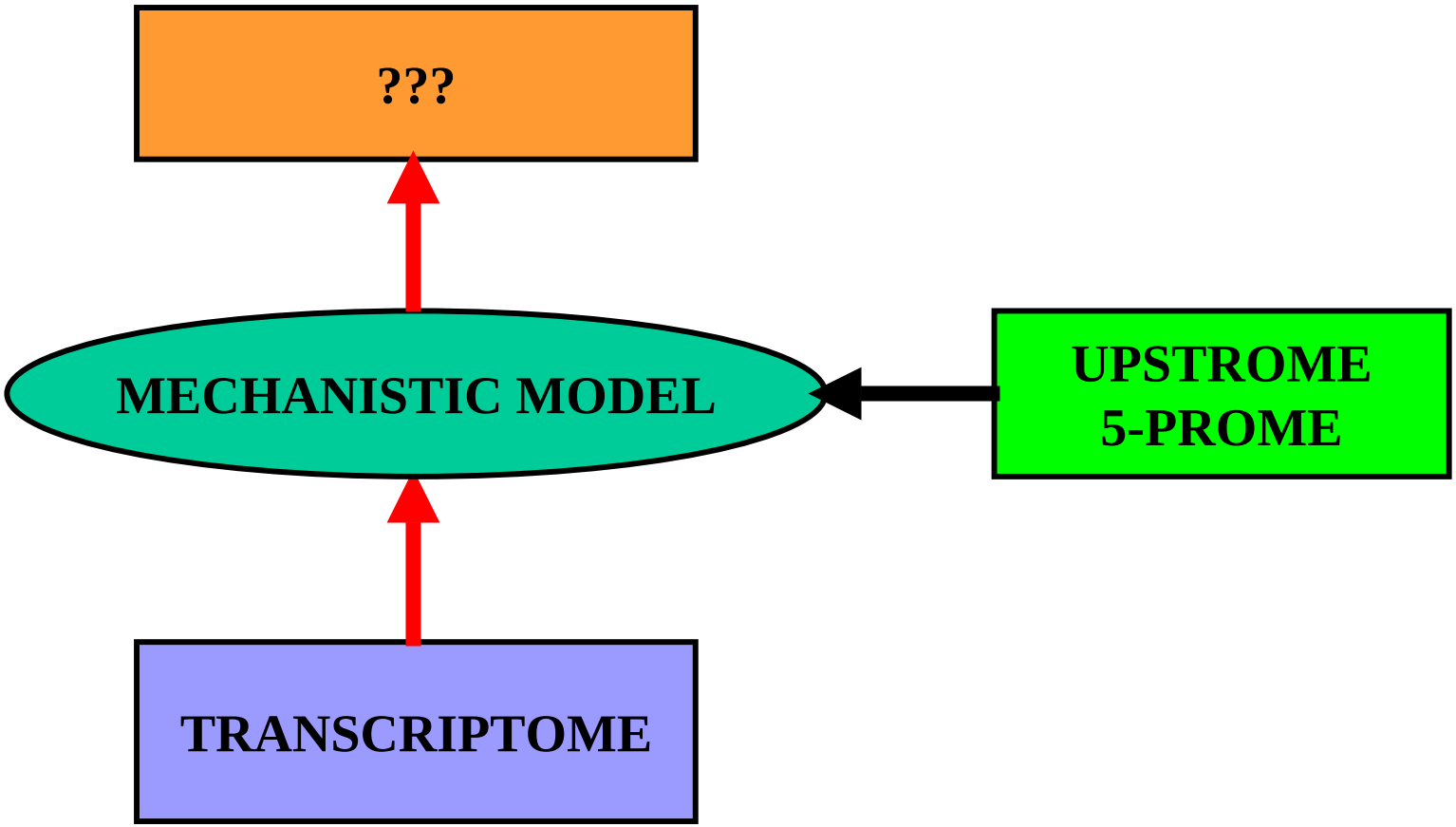


GFP fusion experiment confirms REDUCE prediction of the role of Msn2/4 during metabolic switch from glucose to oleate

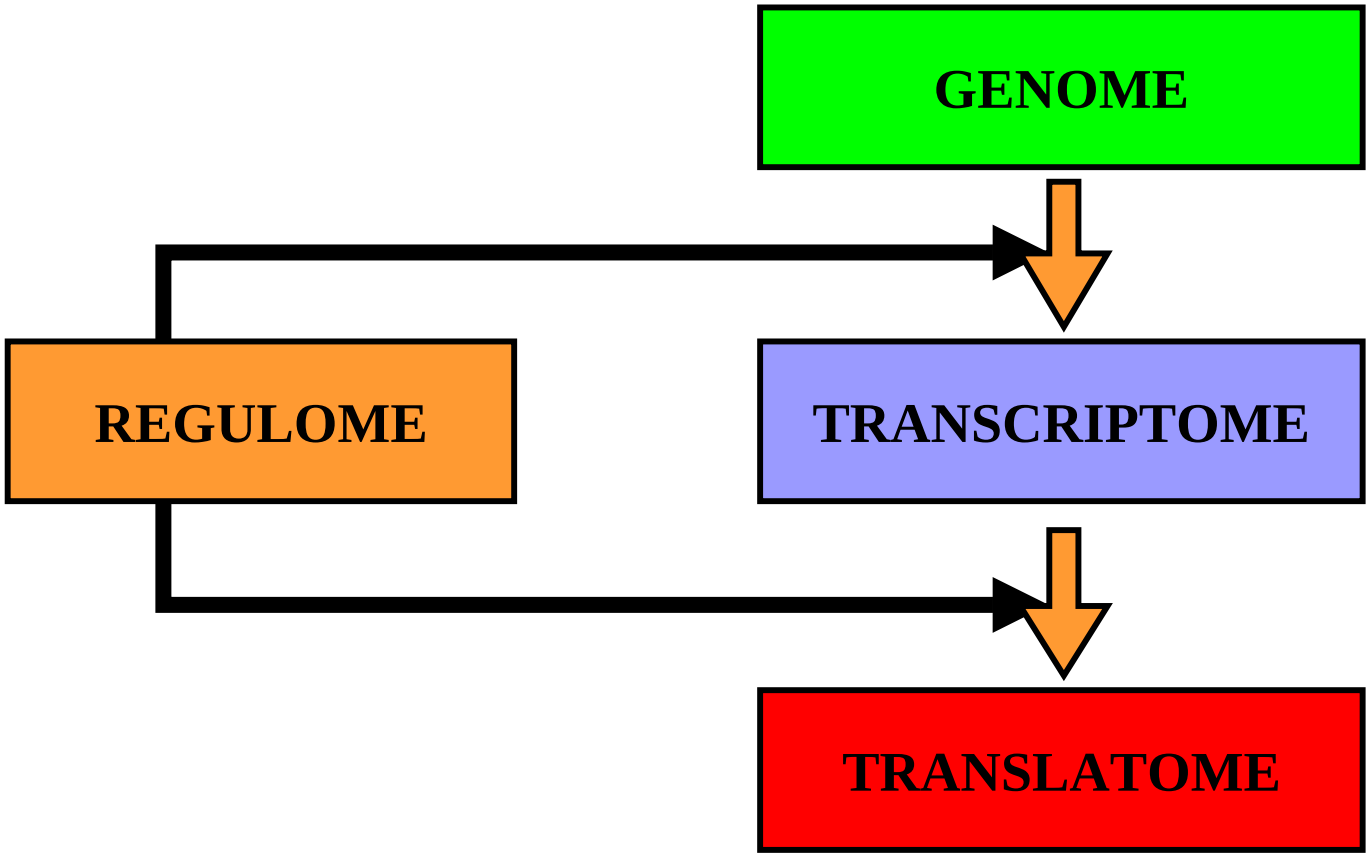
“Regulome” = the “state” of the cell



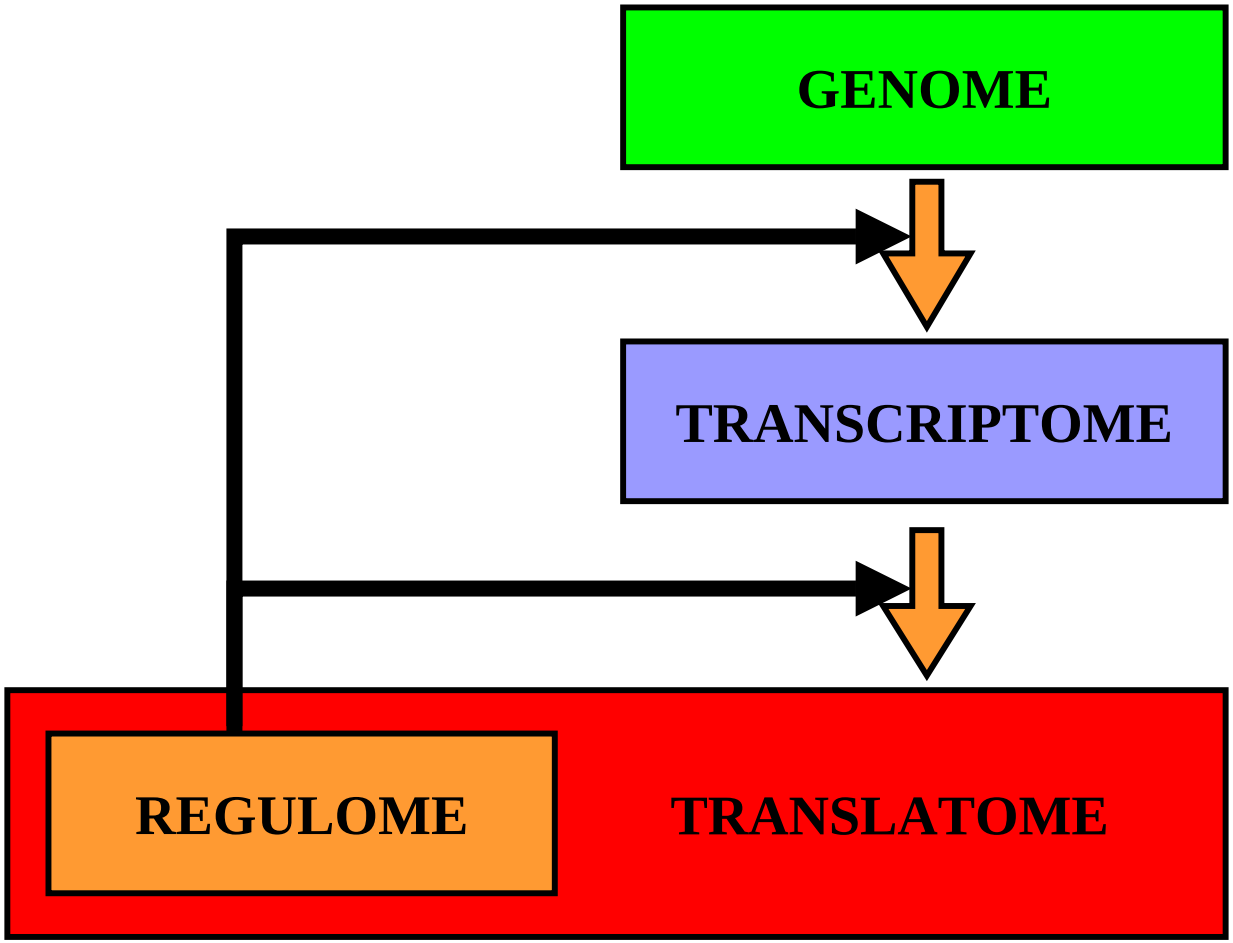
REDUCE viewed as inverse problem

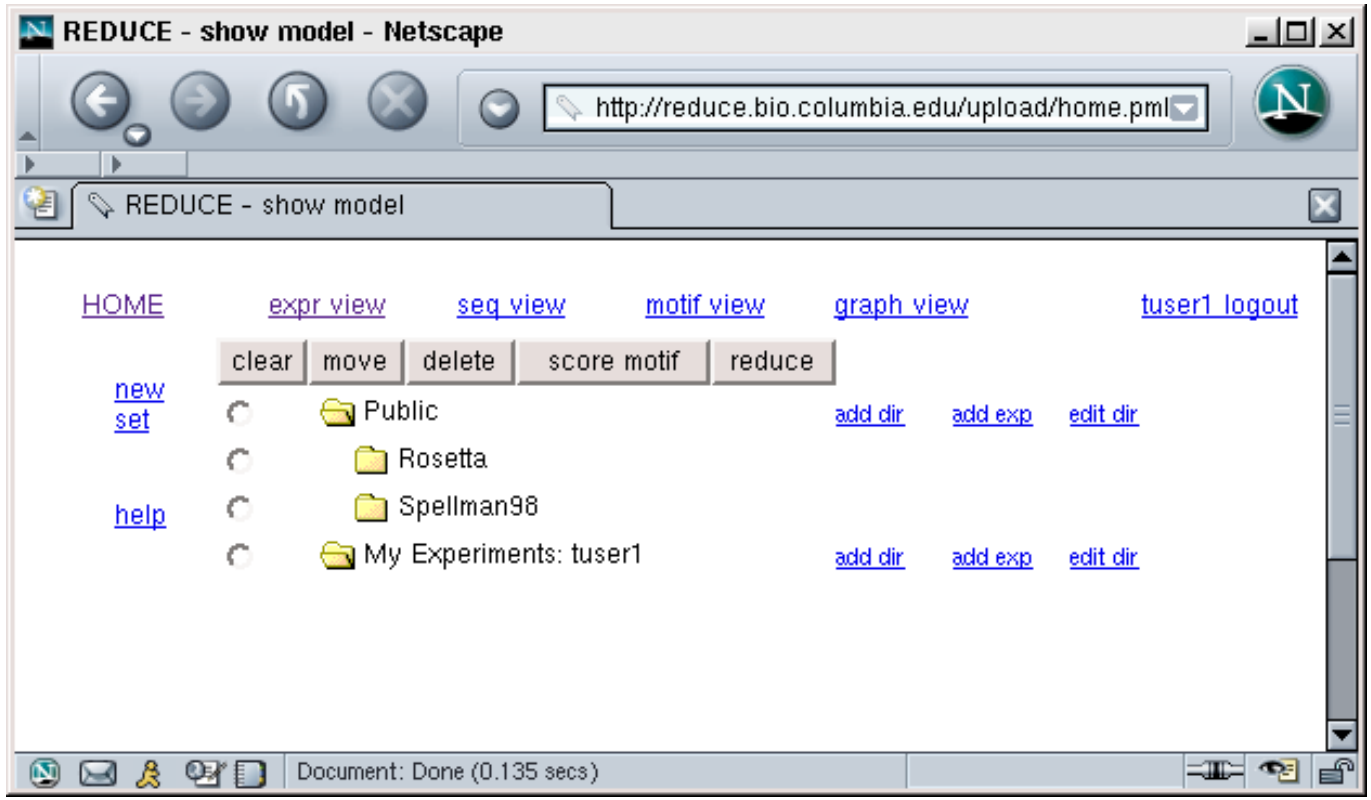


From Central Dogma to “Omes Law”

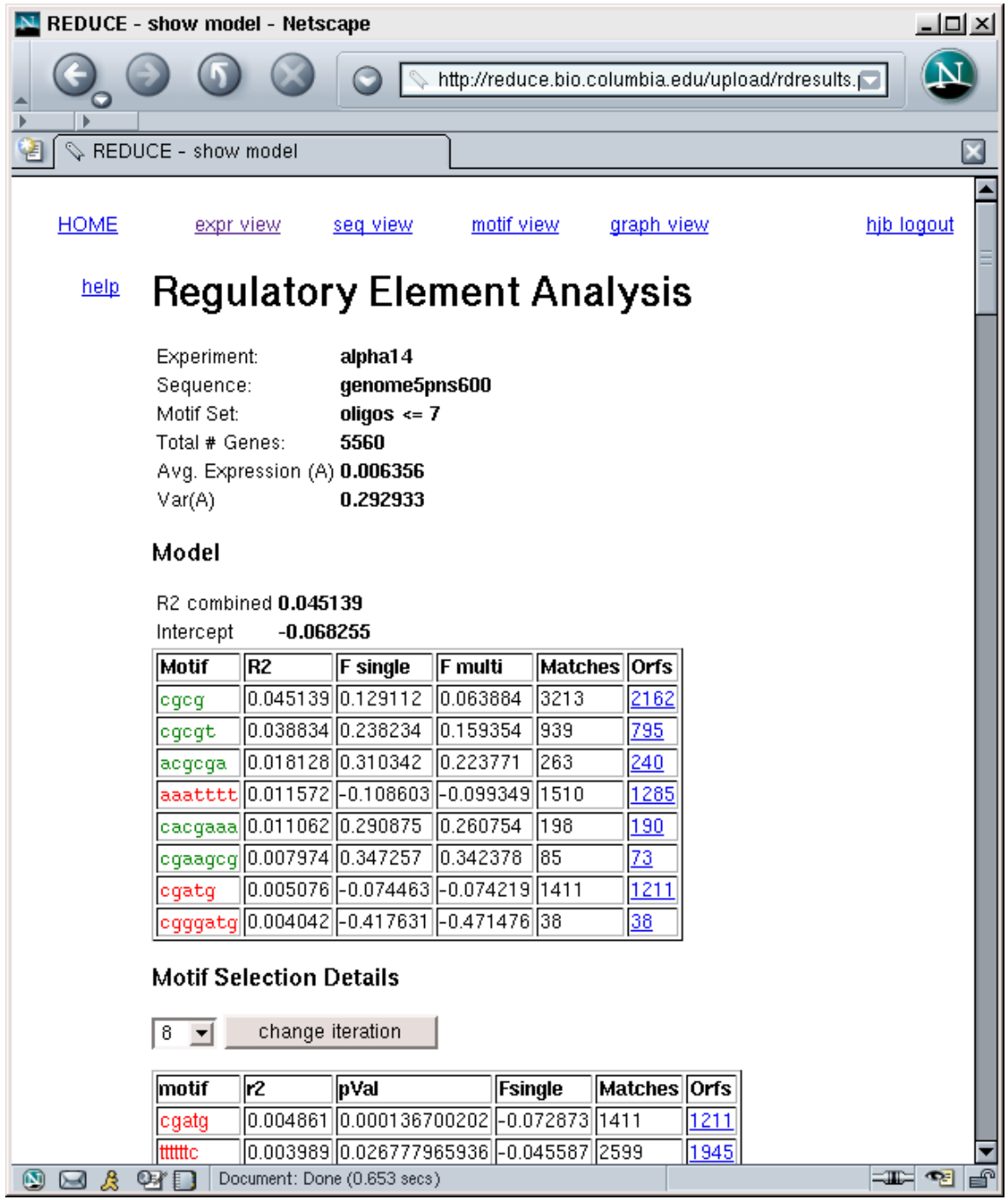


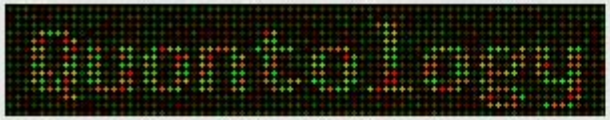
Closing the circle - the genetic network





http://bussemaker.bio.columbia.edu/reduce/


REDUCE - show model - Netscape
http://reduce.bio.columbia.edu/upload/rdresults.
REDUCE - show model
HOME
expr view
seq view
motif view
graph view
hjb logout
help
Regulatory Element Analysis
Experiment: alpha14
Sequence: genome5pns600
Motif Set: oligos <= 7
Total # Genes: 5560
Avg. Expression (A) 0.006356
Var(A) 0.292933
Model
R2 combined 0.045139
Intercept -0.068255
Motif R2 F single F multi Matches Orfs
cgcg 0.045139 0.129112 0.063884 3213 2162
cgcgt 0.038834 0.238234 0.159354 939 795
acgcga 0.018128 0.310342 0.223771 263 240
aaattttt 0.011572 -0.108603 -0.099349 1510 1285
cacgaaa 0.011062 0.290875 0.260754 198 190
cgaagcg 0.007974 0.347257 0.342378 85 73
cgatg 0.005076 -0.074463 -0.074219 1411 1211
cgggatg 0.004042 -0.417631 -0.471476 38 38
Motif Selection Details
8
change iteration
motif r2 pVal Fsingle Matches Orfs
cgatg 0.004861 0.000136700202 -0.072873 1411 1211
ttttttc 0.003989 0.026777965936 -0.045587 2599 1945
Document: Done (0.653 secs)



Gene ontology analysis

switch to regulatory element analysis

Change category | Change experiment

Selected parameters

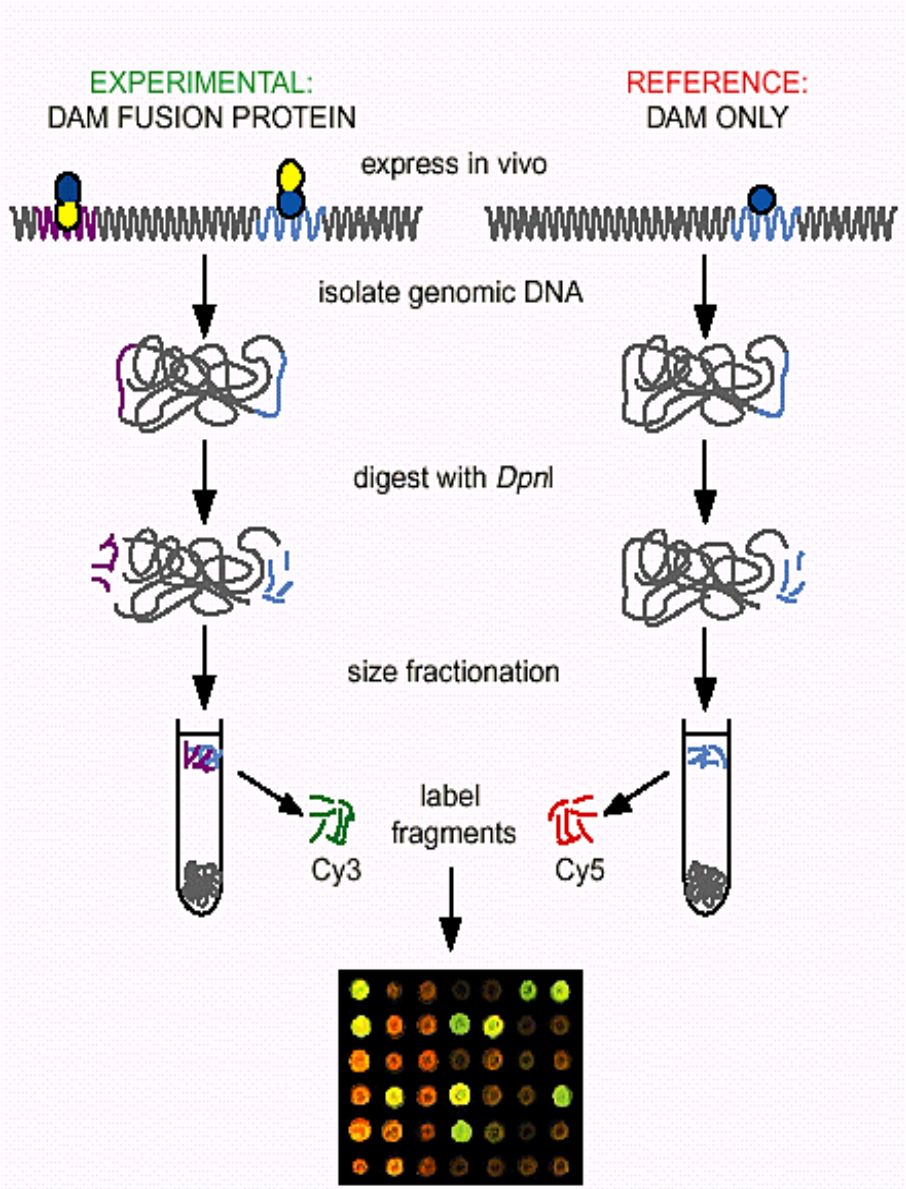
- Expression data used is Thevelein (Roosen:Ratio_1)

Significantly induced GO categories

Z score	P value	Description	Aspect	ORFs
+9.22	0.000000	heat shock protein	F	17
+8.60	0.000000	stress response	P	36
+5.65	0.000022	multicatalytic endopeptidase	F	30
+5.45	0.000068	19S proteasome regulatory particle	C	20
+5.40	0.000089	protein folding	P	35
+5.24	0.000218	glutathione transferase	F	2
+4.69	0.003772	oxidative stress response	P	16
+4.67	0.003995	tricarboxylic acid cycle	P	14
+4.59	0.005893	ubiquitin-dependent protein degradation	P	58
+4.54	0.007539	periplasmic space	C	8
+4.53	0.007977	superoxide dismutase	F	2
+4.40	0.014743	glutathione metabolism	P	3
+4.25	0.028237	asparagine catabolism to aspartate	P	4
+4.25	0.028237	nitrogen starvation response	P	4
+4.24	0.030379	malate metabolism	P	3

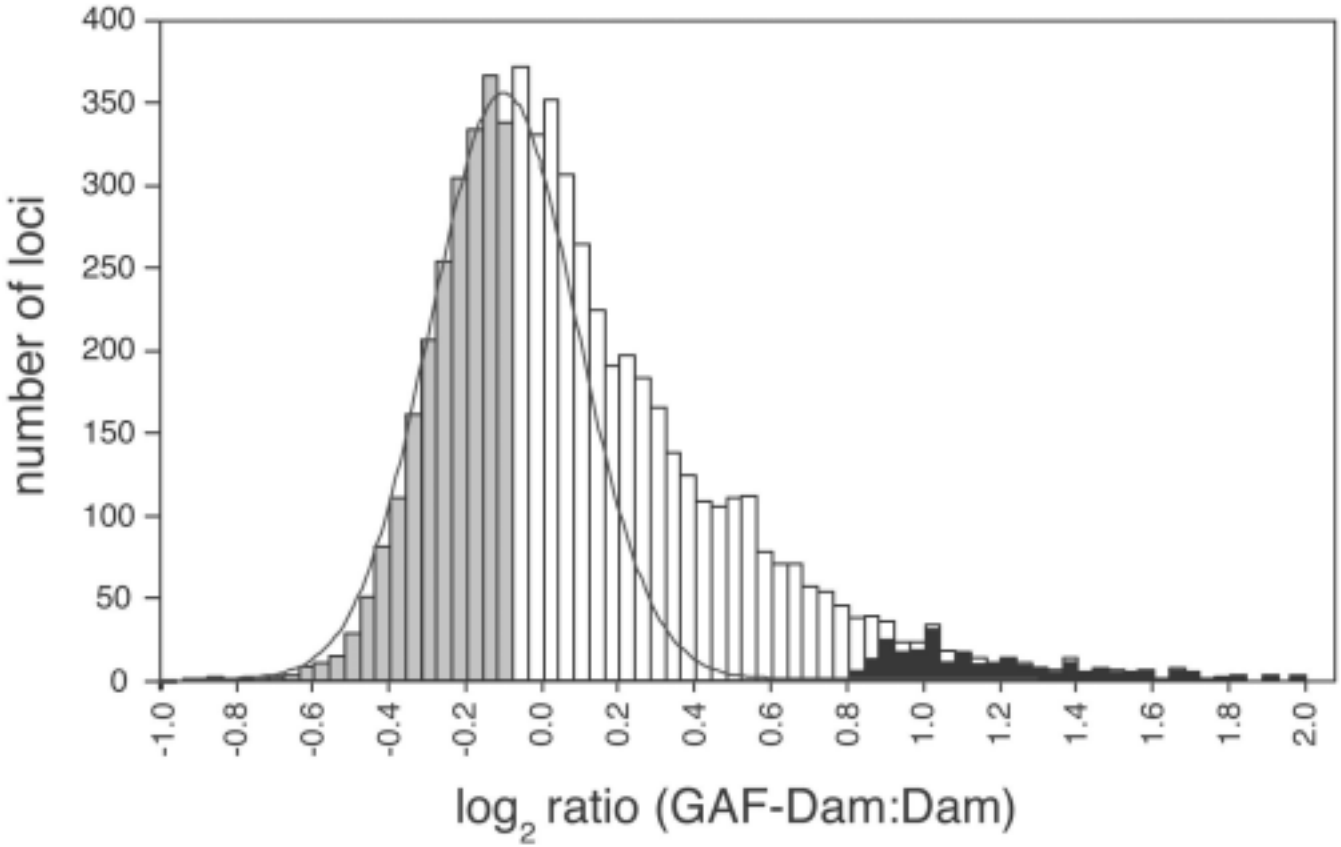
“DamID”

Bas van Steensel
NKI Amsterdam

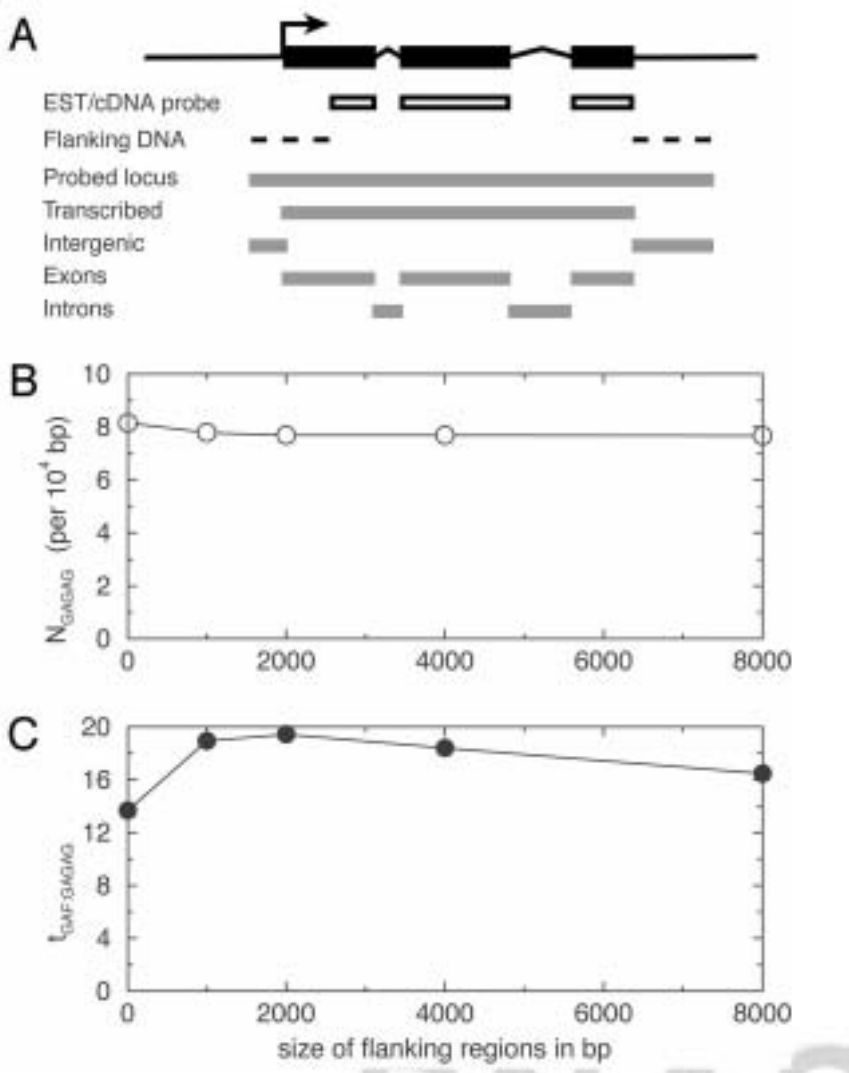


B. Van Steensel, J. Delrow, S. Henikoff, Nature Genet. (2001)

Context dependence of GAGA factor binding



B. van Steensel, J. Delrow, and H.J. Bussemaker, *PNAS* (in press)



B. van Steensel, J. Delrow, and H.J. Bussemaker, *PNAS* (in press)

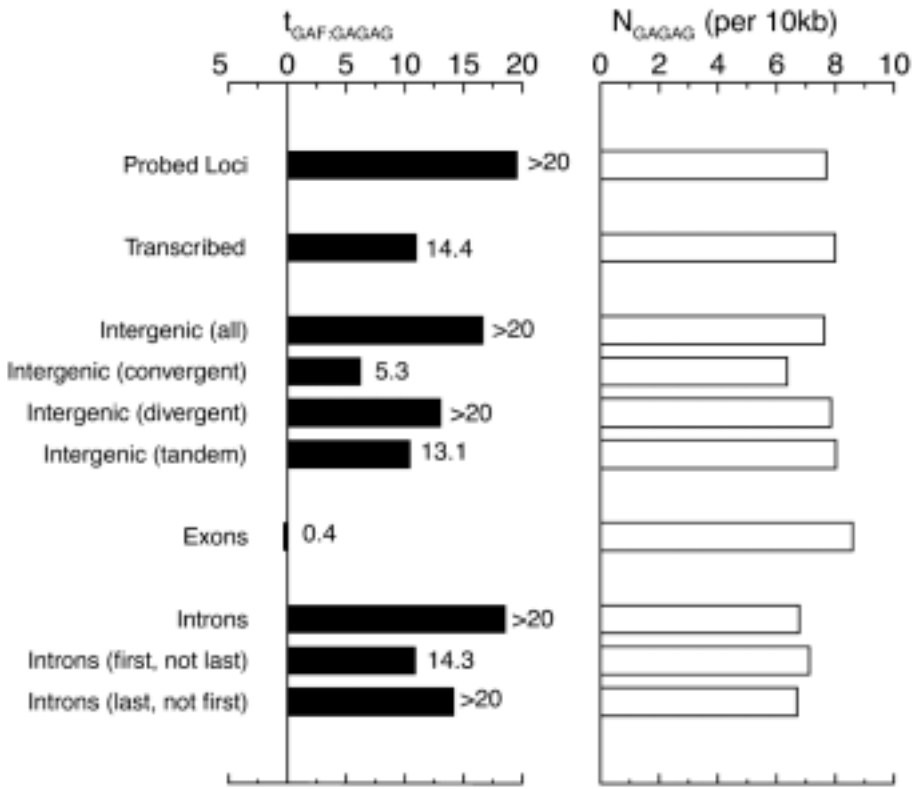


Fig. 3. Binding of GAF to GAGAG elements in subregions of target genes. (*Left*) *t* value calculated as in Fig. 2*C*; for each bar, the statistical significance is indicated by value of $-\log_{10}(P)$. (*Right*) GAGAG density. GAGAG occurrences were counted only in specific subregions of the probed loci as indicated in Fig. 2*A*.

B. van Steensel, J. Delrow, and H.J. Bussemaker, *PNAS* (in press)

Table 2. Most significant motifs found by REDUCE analysis, for probed loci with 2 kb flanking sequence

Rank	Motif	r^2	t	$\log_{10}$, P value	Matches	Loci with match
1	AGAGAG	0.08411	19.646	<−16	6,958	2,677
2	GAGAG	0.08353	19.572	<−16	22,454	4,084
3	AGAGA	0.07613	18.61	<−16	25,640	4,120
4	GAGAGA	0.07043	17.846	<−16	6,675	2,720
8	AAGAGAG	0.05598	15.788	<−16	1,940	1,384
11	GAGAGAG	0.05345	15.406	<−16	2,762	1,217
13	AGAGAGC	0.04817	14.585	<−16	1,848	1,350
14	AGAGAGA	0.04673	14.353	<−16	2,876	1,309
16	AAGAGA	0.04087	13.382	<−16	7,993	3,373
17	GAGAGAA	0.03944	13.137	<−16	1,731	1,333
22	AAAGAGA	0.03592	12.514	<−16	3,054	2,023
26	GAGAGCG	0.03245	11.872	<−16	2,125	1,528
28	AGAGAGT	0.03229	11.842	<−16	1,178	981
35	AGAGAA	0.03012	11.424	−15.5	8,221	3,418
36	GAGAGC	0.02949	11.302	−15.2	7,275	3,269
42	CGAGAGA	0.02823	11.051	−14.6	1,268	1,055
5	CTCTC	0.06802	17.514	<−16	21,401	4,062
6	TCTCT	0.0619	16.653	<−16	24,773	4,115
7	CTCTCT	0.06175	16.632	<−16	7,068	2,711
9	CTCTCTT	0.05526	15.68	<−16	1,861	1,339
10	GCTCTCT	0.05501	15.641	<−16	1,772	1,307
12	TCTCTC	0.04922	14.751	<−16	6,624	2,665
15	TCTCTT	0.04443	13.979	<−16	7,390	3,298
18	CTCTCTC	0.03852	12.976	<−16	2,742	1,150
19	CTCT	0.03762	12.818	<−16	88,116	4,205
20	GCTCTC	0.03709	12.723	<−16	6,627	3,103
23	CGCTCTC	0.03579	12.49	<−16	1,858	1,325
31	TCTCTTT	0.03069	11.537	−15.7	2,923	1,970
33	TCTCTCT	0.03017	11.435	−15.5	2,971	1,300
34	CTCTTT	0.03017	11.435	−15.5	8,940	3,540
44	TTCTCT	0.02798	11	−14.4	7,486	3,313
46	CGCTCT	0.0274	10.882	−14.2	5,904	2,986

B. van Steensel, J. Delrow, and H.J. Bussemaker, *PNAS* (in press)

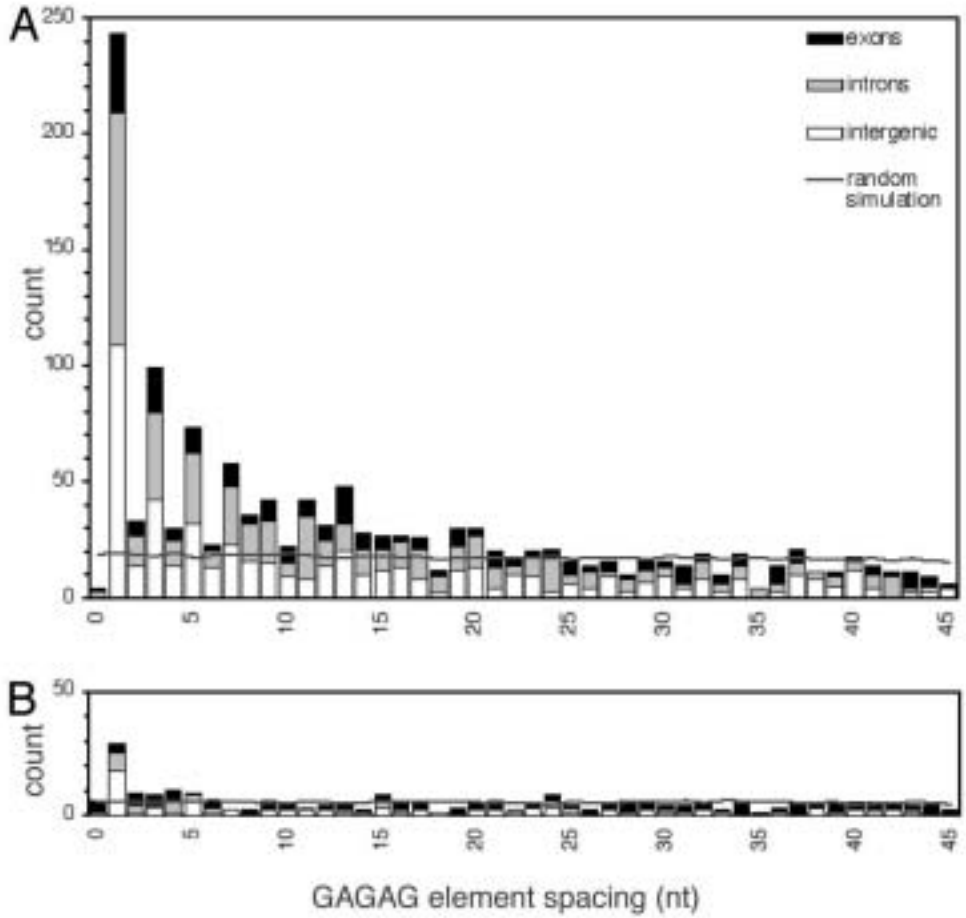
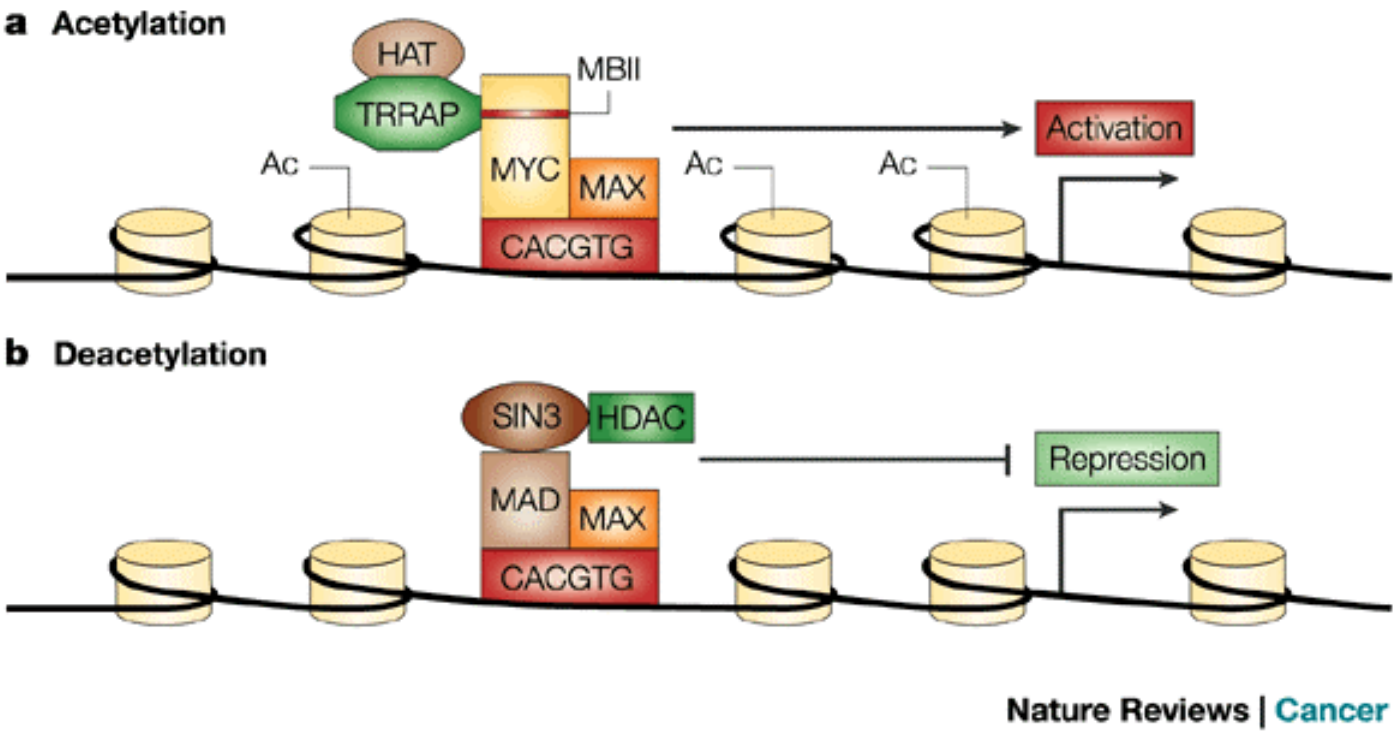


Fig. 4. Spacing distribution of GAGAG elements in 500 loci with strongest GAF binding (average methylation log ratio 0.96) (*A*) and 500 loci with weakest GAF binding (average methylation log ratio −0.44) (*B*), for intergenic regions (black), introns (gray), and exons (open bars). Lines indicate the spacing distribution after randomization of the positions of all GAGAG elements in each locus (average of >1,000 random simulations is shown).

B. van Steensel, J. Delrow, and H.J. Bussemaker, *PNAS* (in press)

The Drosophila Myc/Mad/Max Network



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 Dev.* (in press)

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

CACGTG = E-box
TATCGATA = DRE
GTCACAC = ???

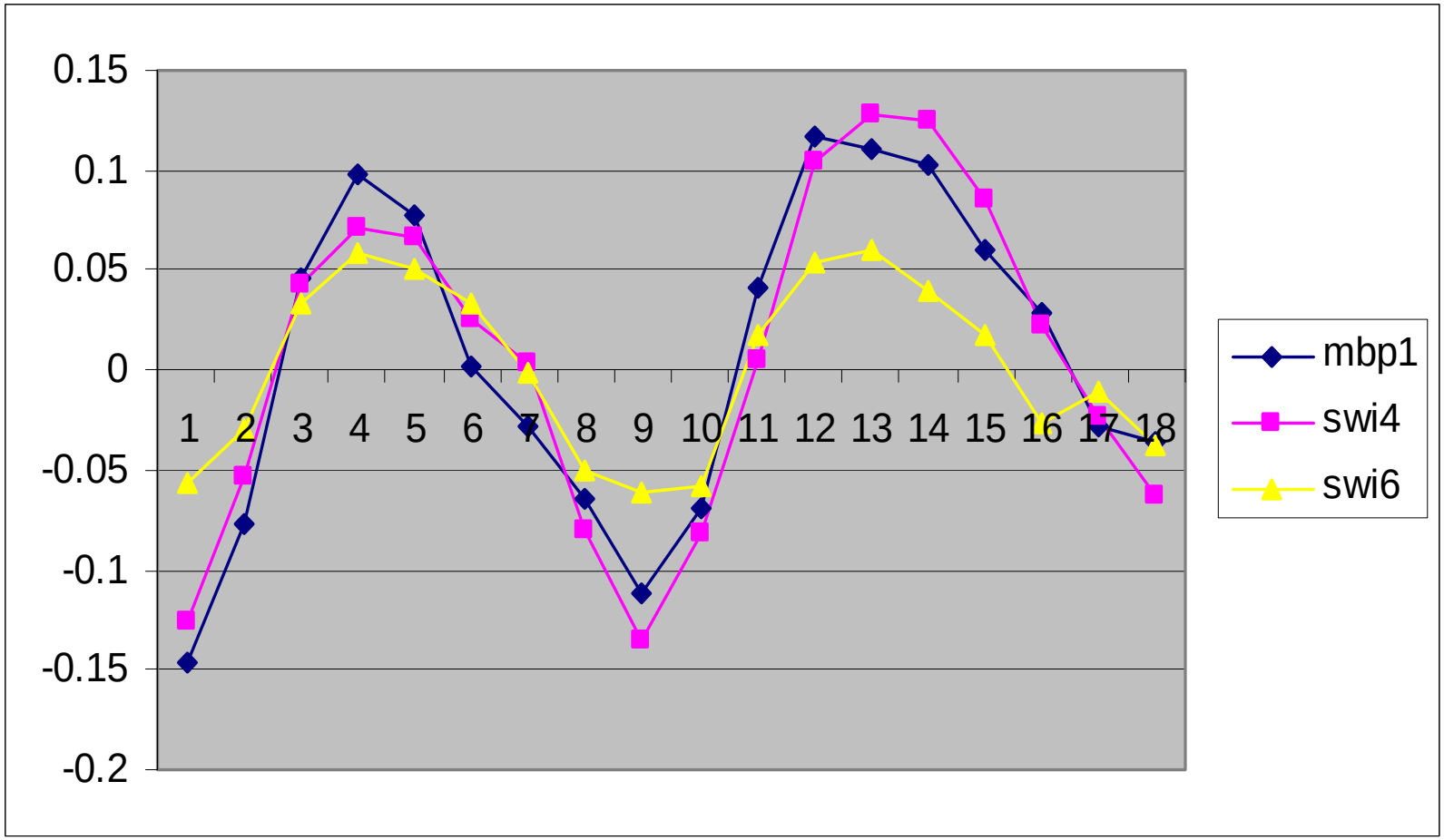
Making the most of ChIP data

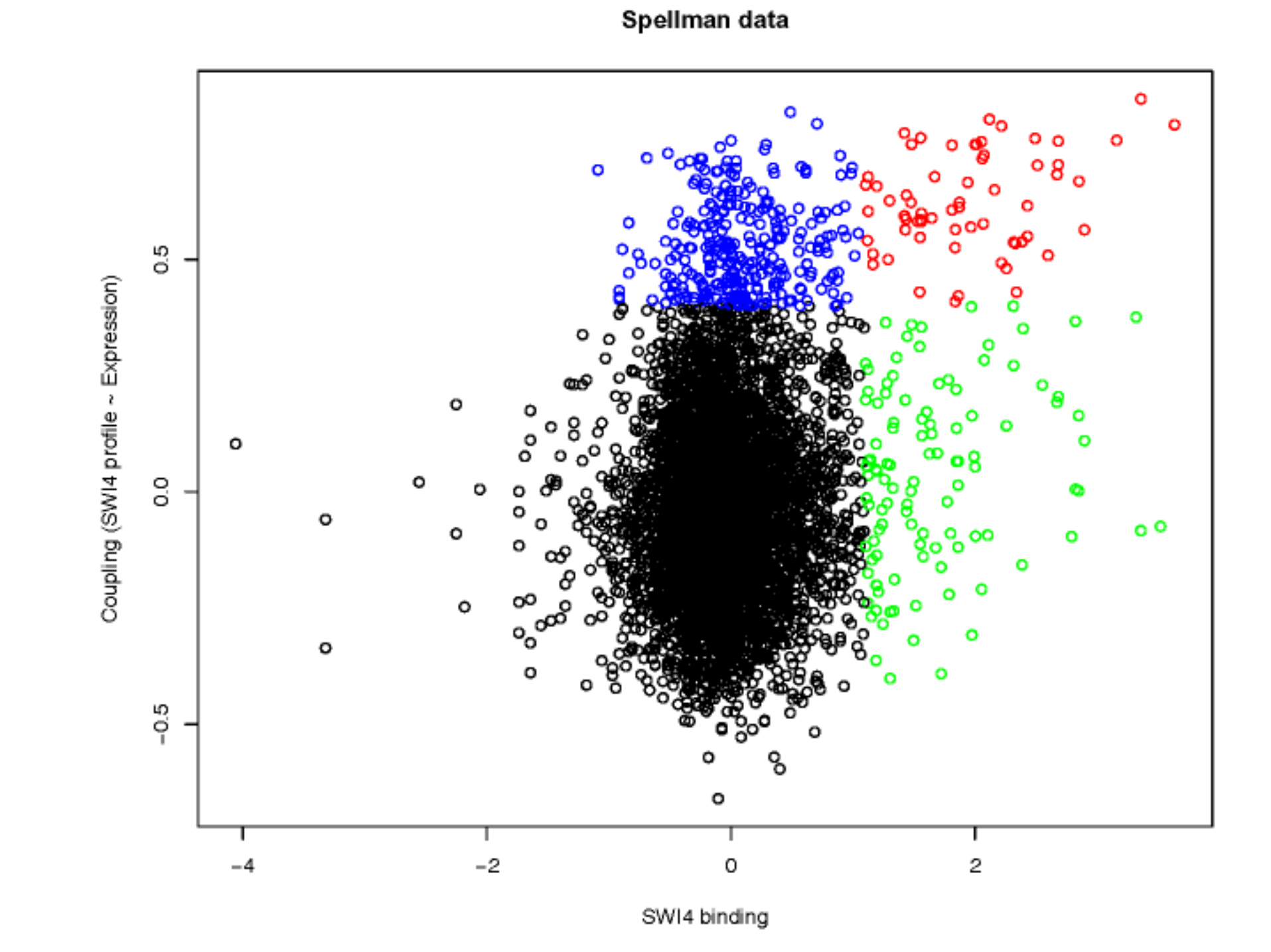
- Lee et al. - recent *Nature* paper Young Lab

- TF activity profiles from regression (E~B)
- Coupling of individual genes to TF activity

- Increase specificity of TF target prediction

TF activity inferred from ChIP + Spellman data (alpha)




Spellman data
Coupling (SWI4 profile ~ Expression)
0.5
0.0
−0.5
−4
−2
0
2
SWI4 binding

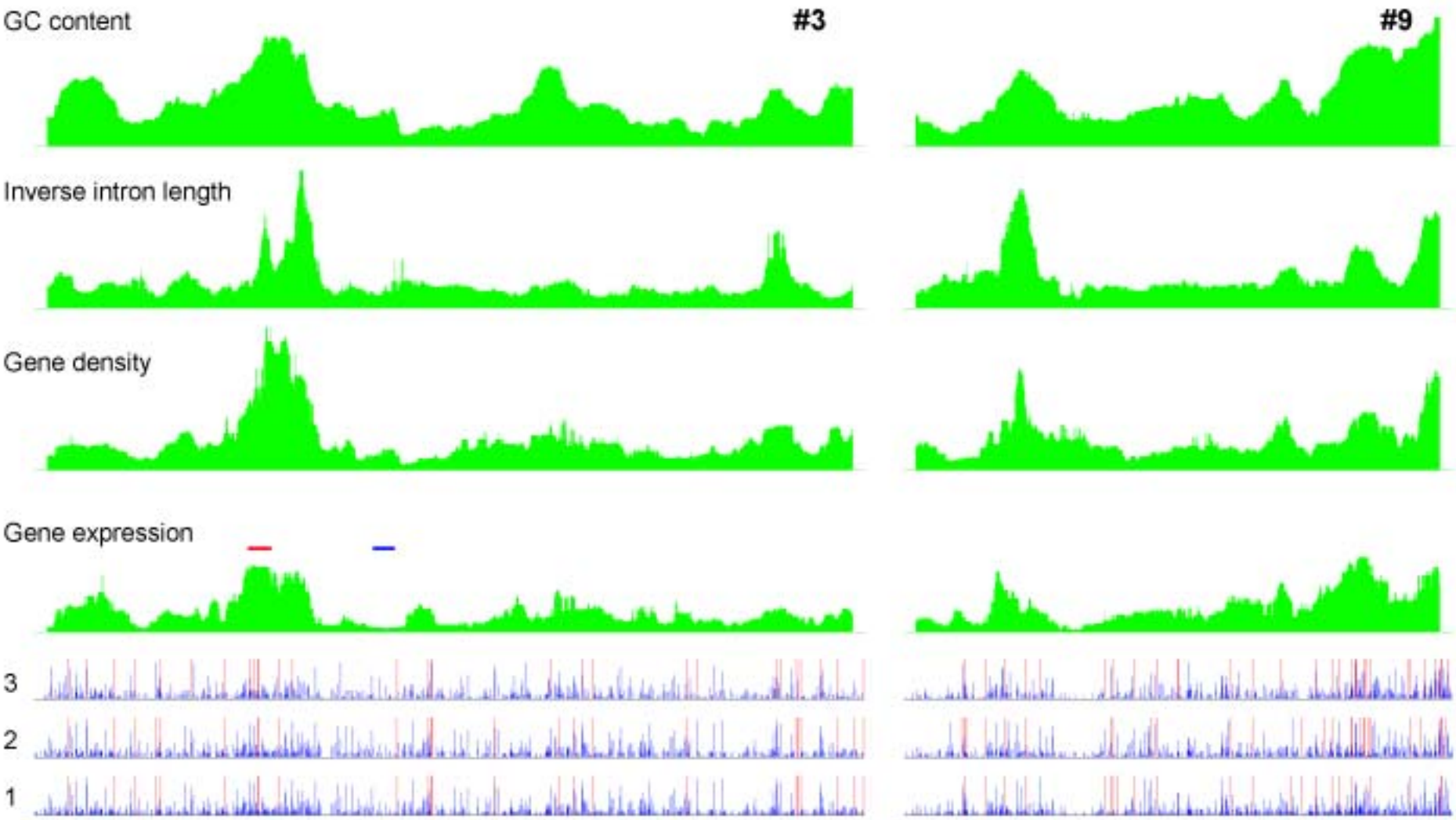
Group	P	Match	Csize	Gnumber	Category	Cname
B+/C+	3.65E-06	15	427	63	MIPS_funcat_fc14	CELL FATE
B+/C+	2.17E-05	13	382	63	MIPS_funcat_fc14_04	cell differentiation
B+/C+	2.17E-05	13	382	63	MIPS_funcat_fc14_04_03	fungal cell differentiation
B+/C+	8.82E-07	10	170	63	MIPS_funcat_fc14_04_03_01	budding, cell polarity and filament formation
B+/C+	4.6E-06	5	44	63	MIPS_funcat_fc40_10_03	chromosome
B+/C+	4.14E-05	4	39	63	MIPS_subcell_fc1	cell wall
B+/C+	2.74E-06	4	23	63	MIPS_subcell_fc27	chromosome structure
B-/C+	8.19E-05	4	11	291	MIPS_funcat_fc01_03_07	deoxyribonucleotide metabolism
B-/C+	-4.6E-12	79	628	291	MIPS_funcat_fc03	CELL CYCLE AND DNA PROCESSING
B-/C+	-8.7E-12	43	251	291	MIPS_funcat_fc03_01	DNA processing
B-/C+	1.39E-09	20	94	291	MIPS_funcat_fc03_01_03	DNA synthesis and replication
B-/C+	6.21E-11	28	153	291	MIPS_funcat_fc03_01_05	DNA recombination and DNA repair
B-/C+	2.04E-06	43	451	291	MIPS_funcat_fc03_03	cell cycle
B-/C+	3.42E-07	38	352	291	MIPS_funcat_fc03_03_01	mitotic cell cycle and cell cycle control
B-/C+	0.000173	134	2239	291	MIPS_funcat_fc40	SUBCELLULAR LOCALISATION
B-/C+	8.47E-07	9	31	291	MIPS_funcat_fc40_05	centrosome
B-/C+	1.24E-05	20	157	291	MIPS_funcat_fc40_07	endoplasmic reticulum
B-/C+	2.74E-05	9	44	291	MIPS_funcat_fc40_10_03	chromosome
B-/C+	1.24E-08	21	115	291	MIPS_subcell_fc4	cytoskeleton
B-/C+	4.6E-07	8	23	291	MIPS_subcell_fc6	tubulin cytoskeleton
B-/C+	3.93E-06	9	36	291	MIPS_subcell_fc7	spindle pole body
B-/C+	4.34E-06	21	159	291	MIPS_subcell_fc9	ER
B-/C+	3.93E-06	5	11	291	MIPS_subcell_fc15	ER-golgi transport vesicles
B-/C+	0.000123	61	836	291	MIPS_subcell_fc22	nucleus
B+/C-	No significant enrichment for any MIPS category!					

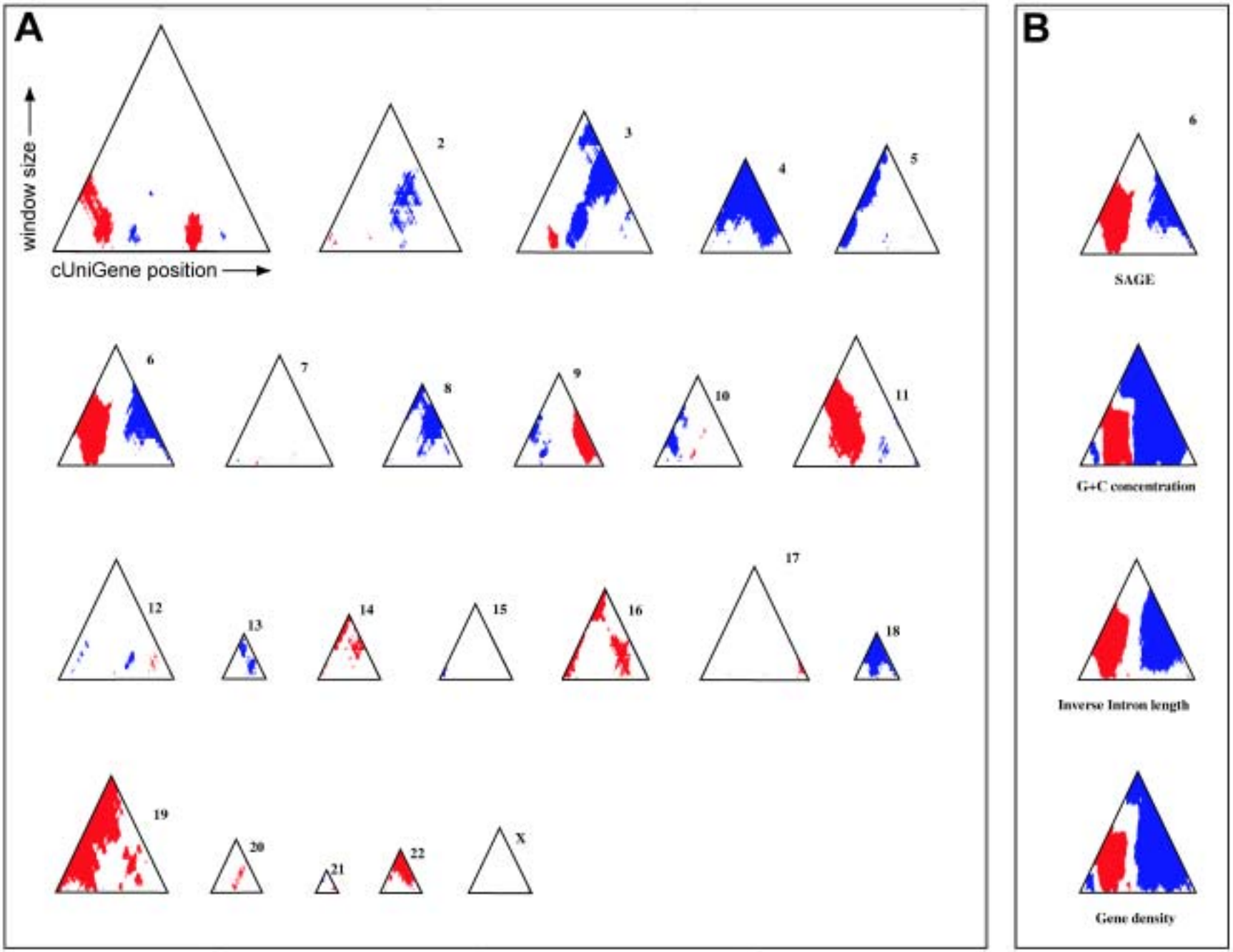
Two hypotheses:

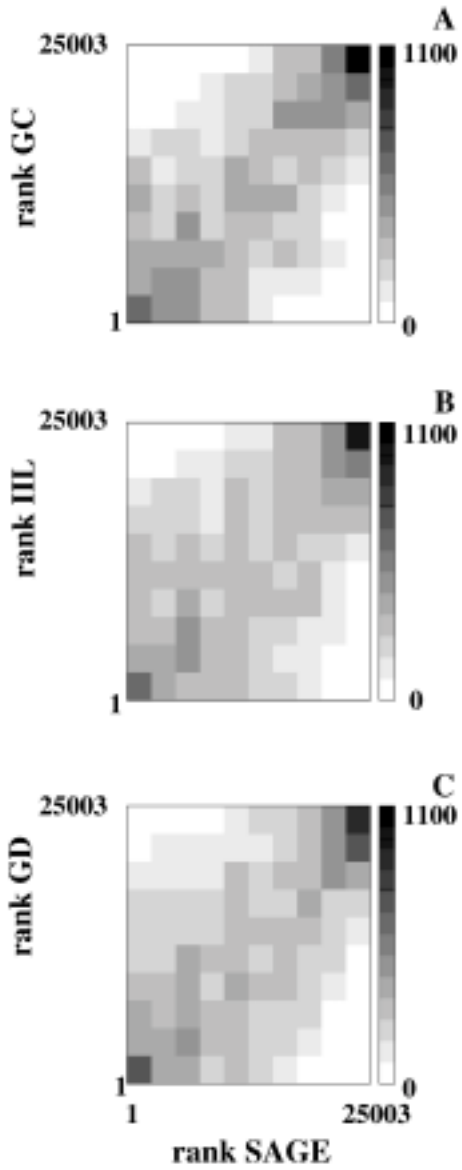
- Non-coupling promoters are ChIP false positives
- Non-coupling promoters are truly bound by TF, but they fail to recruit Pol II complex (do they have another role?)

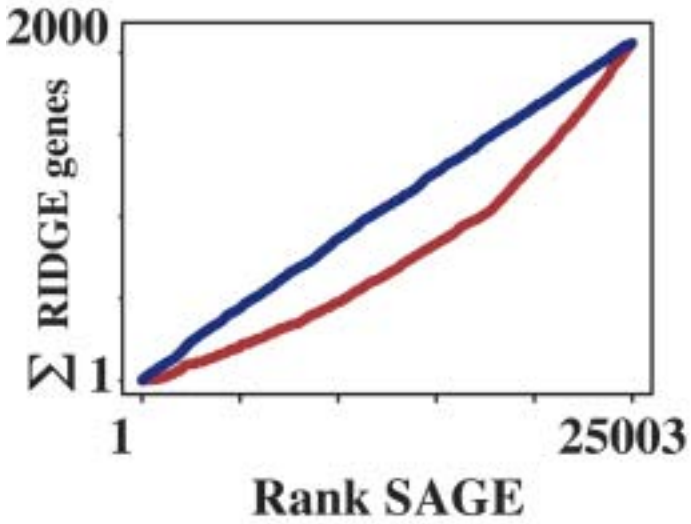
Human Transcriptome Map

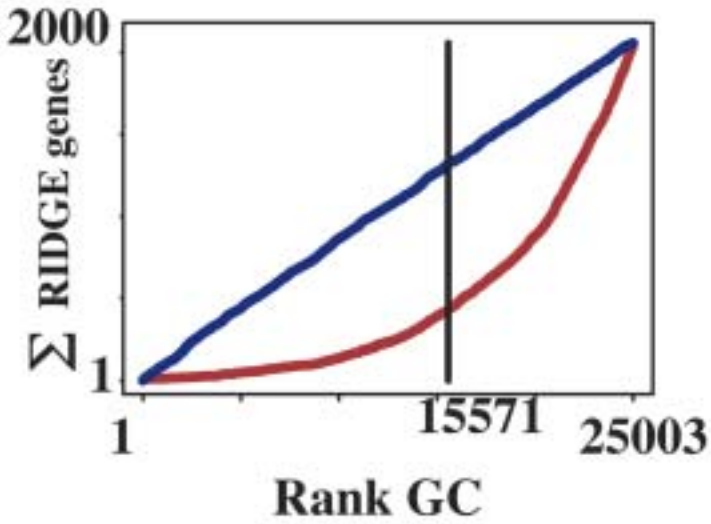
- Map SAGE data to genome assembly
- Spatial clustering of high/low expression
- Multiple length scales: 30-1000 genes
  - RIDGES / “ridgeograms”
- Correlation with, and clustering of:
  - GC-content
  - gene density
  - inverse intron length

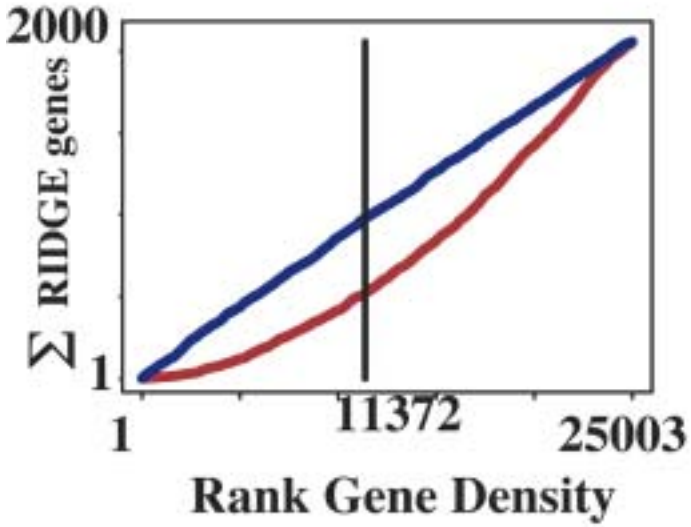

GC content
#3
#9
Inverse intron length
Gene density
Gene expression
3
2
1

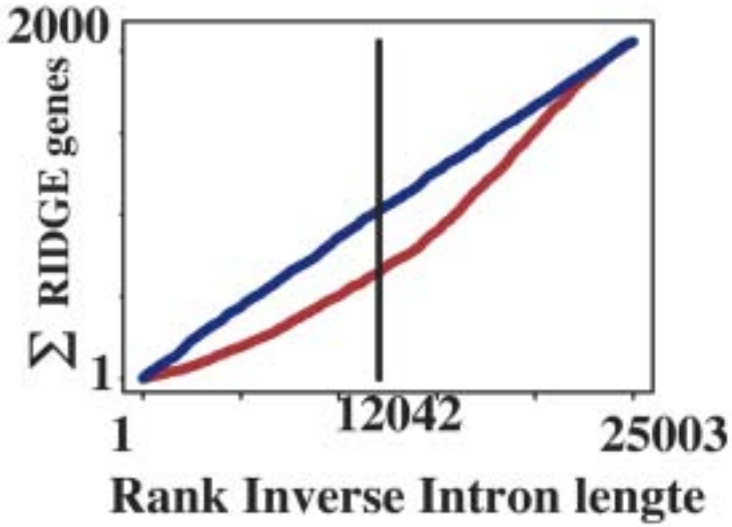

A
window size
cUniGene position
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
B
6
SAGE
G+C concentration
Inverse Intron length
Gene density


A
25003
1100
rank GC
1
0
B
25003
1100
rank IIL
1
0
C
25003
1100
rank GD
1
0
1
25003
rank SAGE


2000
Σ RIDGE genes
1
1
25003
Rank SAGE


2000
Σ RIDGE genes
1
1
15571
25003
Rank GC


2000
Σ RIDGE genes
1
1
11372
25003
Rank Gene Density


2000
Σ RIDGE genes
1
1
12042
25003
Rank Inverse Intron lengte

Two hypotheses:

- *Passive role of chromatin*: evolution favors highly expressed genes to be physically close, as they will help each other in "opening up" the chromatin
- *Active role of chromatin*: chromosomal domains are controlled by polycomb proteins, boundary elements, …?

Acknowledgements

Eric Siggia, Hao Li, Boris Shraiman (Rockefeller University)

Marian Koerkamp, Martijn Rep, Henk Tabak (AMC Amsterdam)

Rogier Versteeg, Antoine van Kampen, Marco Roos (AMC Amsterdam)

Bas van Steensel (Netherlands Cancer Institute)

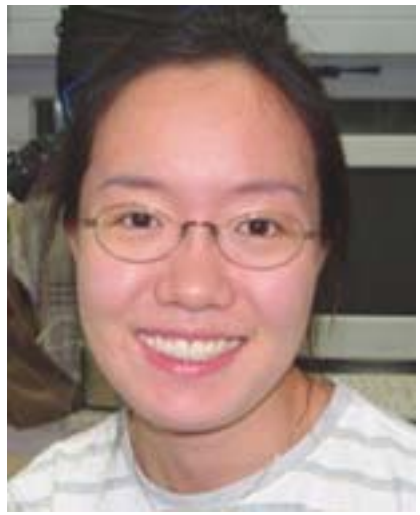
Jeff Delrow, Amir Orian, Bob Eisenman (FHCRC, Seattle)

NIH / Netherlands Organization for Scientific Research (NWO)

The Lab

Barrett Foat
(mRNA structure)

Crispin Roven
(database)



Min-sung Choi
(rotation)

Marcel van Batenburg
(Human Transcriptome Map)

Feng Gao
(ChIP binding data)