Genetic regulatory networks

Eukaryotic transcriptional regulation

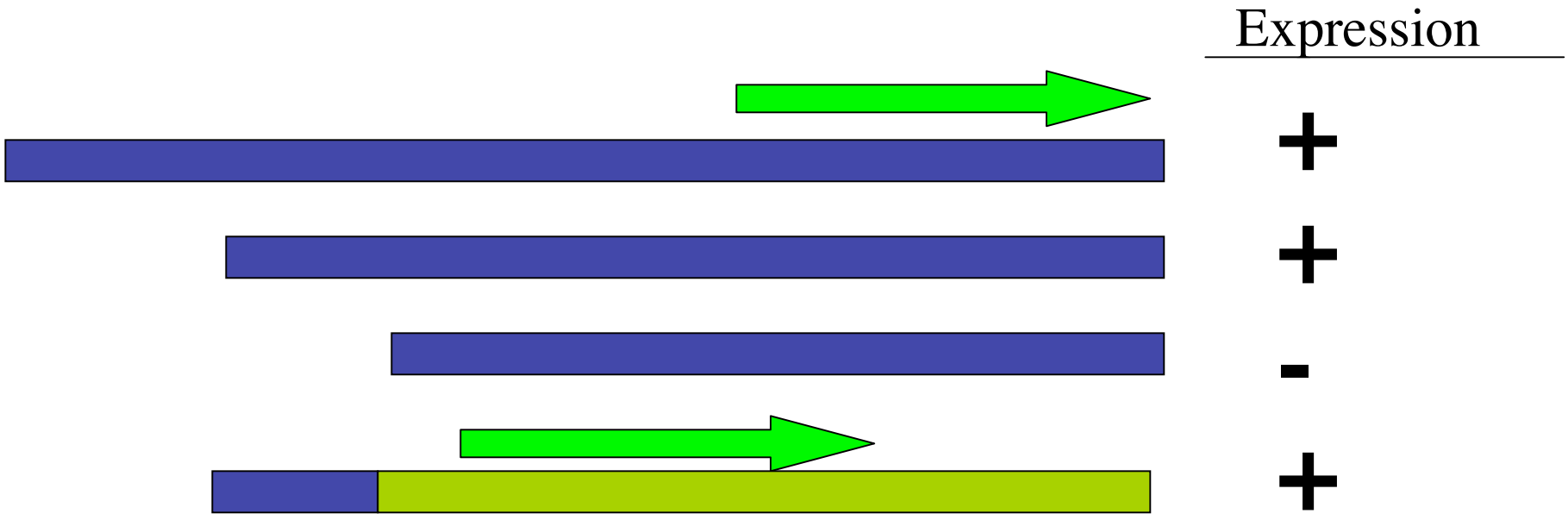
Cis-regulatory analysis

Finding trans-regulators (genetic perturbation)

Finding cis-regulatory sites

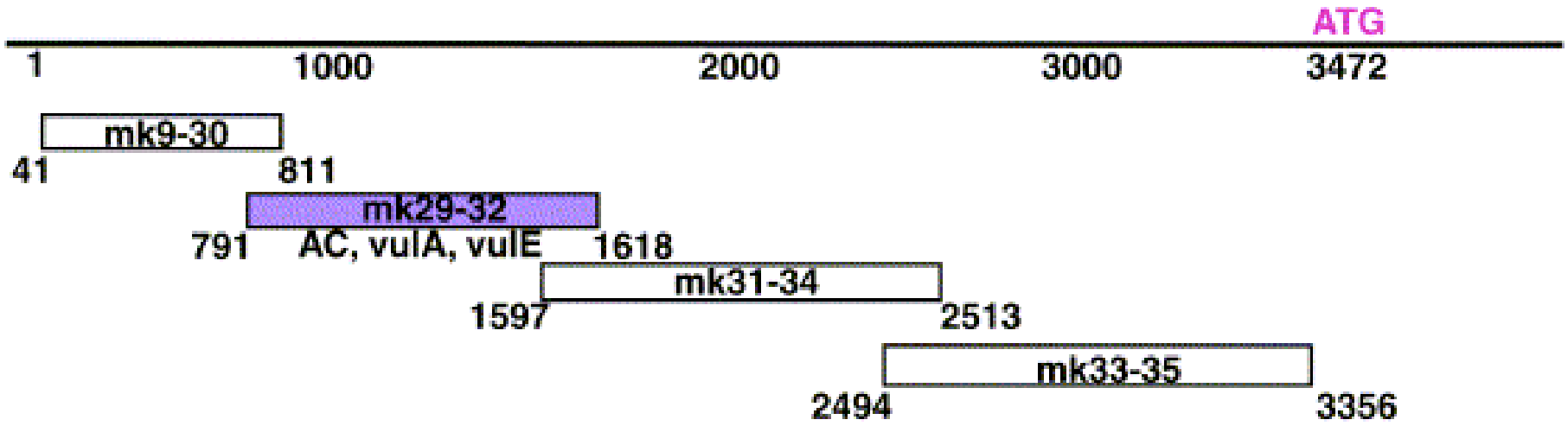
- **computational (conservation, motifs)**
- **experimental (SELEX, deletion/mutation)**

Cis-regulatory analysis

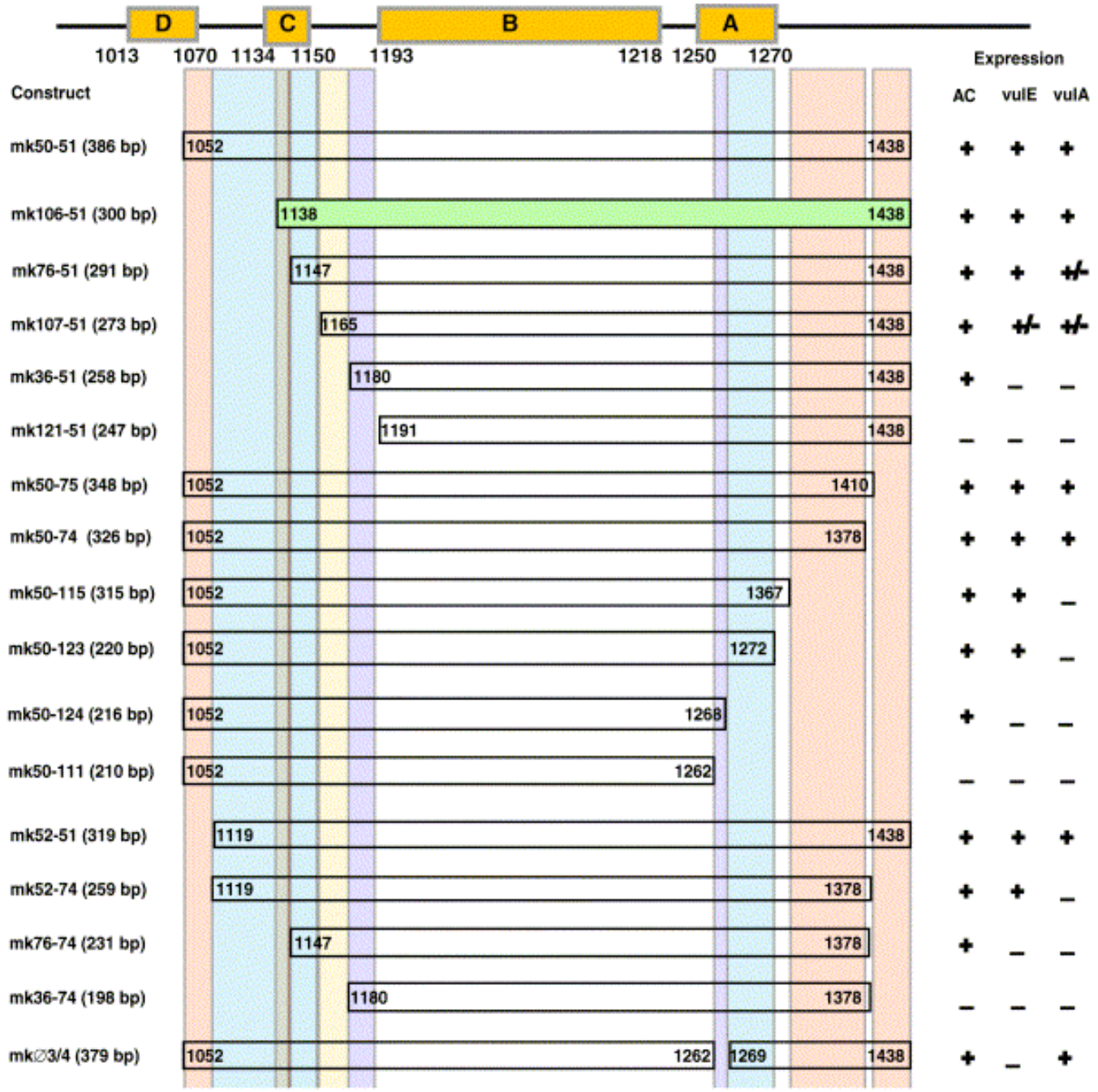


Deletion analysis: necessity
Enhancer assay: sufficiency

A pJB100 *zmp-1* 3472 bp upstream sequence



B Multiple sub-regions of mk29-32 direct *zmp-1* expression



Multicellular organisms: the binding site problem

2×10^4 genes

10^9 bp DNA

Average gene size is thus $10^9/(2 \times 10^4) = 5 \times 10^4$

A hexamer with random GC content occurs once per ~4000 nucleotides (4^6)

IUPAC Ambiguity Symbols

IUPAC Symbol	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
X/N	G or A or T or C	X/N
./-	gap character	./-

Binding sites (simple example)

consensus G C R A C

Position weight matrix (PWM)

Pos	G	C	A	T
1	1	0	0	0
2	0	1	0	0
3	0.5	0	0.5	0
4	0	0	1	0
5	0	1	0	0

sequence logo

http://weblogo.berkeley.edu/logo.cgi

CTCGTA
CACGTG
CAGGTC
CACGTG
CAGGTG
CACGTG
CAGGTG
CACGTG

G C R A C

Pos	G	C	A	T
1	1	0	0	0
2	0	1	0	0
3	0.5	0	0.5	0
4	0	0	1	0
5	0	1	0	0

sequence
G C A T C

Pos	1	2	3	4	5
G	1	0	0	0	0
C	0	1	0	0	1
A	0	0	1	0	0
T	0	0	0	1	0

(5 x4) • (4 x 5) = (5 x 5)

trace (5x5)

```
1 . .   . .
. 1 .   . .
. . 0.5 . .
. . .   0 .
. . .   . 1
```

trace of the matrix product) = 3.5

normalize?

Regulator	Distance[1]	Discovered	Literature
Abf1	0.143	rTCAytnnnnAcg	rTCAyTnnnnACGw
Ace2	0.18	tGCTGGT	GCTGGT
Aft2	0.15	rCACCC	ATCTTCAAAAGTGCACCCATTTGCAGGTGC
Azf1	0.203	YwTTkcKkTyyckgykky	TTTTTCTT
Bas1	0.045	TGACTC	TGACTC
Cad1	0.089	mTTAsTmAkC	TTACTAA
Cbf1	0.105	tCACGTG	rTCACrTGA
Cin5	0.324	TTAcrTAA	TTACTAA
Fkh1	0.123	gtAAAcAA	GGTAAACAA
Fkh2	0.212	GTAAACA	GGTAAACAA
Gal4	0.11	CGGnnnnnnnnnnncCg	CGGnnnnnnnnnnCCG
Gat1	0.004	aGATAAG	GATAA
Gcn4	0.123	TGAsTCa	ArTGACTCw
Gln3	0.148	GATAAGa	GATAAGATAAG
Hap1	0.191	GGnnaTAnCGs	CGGnnnTAnCGG
Hap4	0.146	gnCcAAtcA	YCNNCCAATNANM
Hsf1	0.198	TTCynnnnnnTTC	TTCTAGAAnnTTCT
Ino2	0.236	CAcaTGc	ATTTCACATC
Ino4	0.163	CATGTGaa	CATGTGAAAT
Leu3	0.131	cCGgtacCGG	yGCCGGTACCGGyk
Mbp1	0.073	ACGCGt	ACGCGT
Mcm1	0.181	CCnrAtnngg	wTTCCyAAwnnGGTAA
Msn2	0.308	mAGGGGsgg	mAGGGG
Nrg1	0.042	GGaCCCT	CCCT
Pdr1	0.301	ccGCCgRAwr	CCGCGG
Pho4	0.096	CACGTGs	cacgtkng
Rap1	0.181	cayCCrtrCa	wrmACCCATACAyy
Rcs1	0.184	ggGTGcant	AAmTGGGTGCAkT
Reb1	0.055	TTACCCG	TTACCCGG
Rpn4	0.049	GGTGGCAAA	GGTGGCAAA
Sip4	0.184	CGGnynAATGGrr	yCGGAyrrAwGG
Skn7	0.228	GnCnnGsCs	ATTTGGCyGGsCC
Stb5	0.058	CGGnstTAta	CGG
Ste12	0.087	tgAAAC	ATGAAAC
Sum1	0.221	gyGwCAswaaw	AGyGwCACAAAAk
Sut1	0.295	gcsGsgnnsG	CGCG
Swi4	0.122	CgCsAAA	CnCGAAA
Swi6	0.214	CGCgaaa	CnCGAAA
Tec1	0.064	CATTCyy	CATTCy
Tye7	0.193	tCACGTGa	CAnnTG
Ume6	0.16	taGCCGCCsa	wGCCGCCGw
Yap1	0.124	TTaGTmAGc	TTAsTmA
Yap7	0.15	mTkAsTmA	TTACTAA
Zap1	0.085	ACCCTmAAGGTyrT	ACCCTAAAGGT

Name **Sp1**

Description stimulating protein 1

Factors T00754; Sp1; Species: rat, Rattus norvegicus.
T00752; Sp1; Species: mouse, Mus musculus.
T00759; Sp1; Species: human, Homo sapiens.

Matrix

	Info	N	A	C	G	T	Consensus
01	0.043	108.00	**32**	21	**35**	20	N
02	0.298	108.00	24	20	**56**	8	G
03	0.418	108.00	14	10	**65**	19	G
04	1.225	108.00	17	1	**89**	1	G
05	2.000	108.00	0	0	**108**	0	G
06	1.867	108.00	0	2	**106**	0	G
07	0.940	108.00	19	**80**	0	9	C
08	1.467	108.00	2	5	**99**	2	G
09	1.544	108.00	0	1	**99**	8	G
10	0.747	108.00	21	5	**76**	6	G
11	0.574	108.00	17	10	**72**	9	G
12	0.392	108.00	3	**55**	21	29	Y
13	0.151	108.00	9	**40**	**32**	27	N

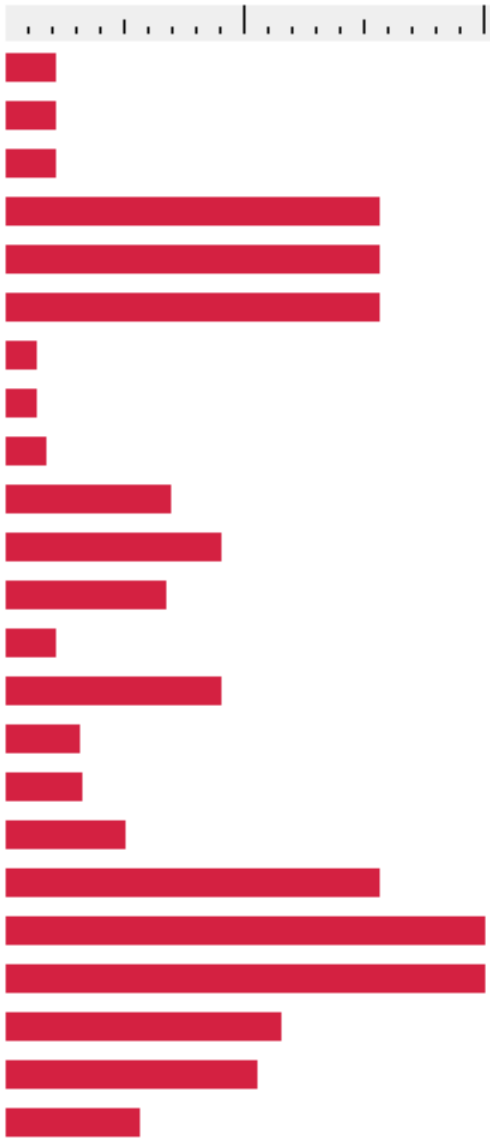
Basis 108 compiled sequences

Comments TRANSFAC Sites of quality <= 6

Description GAL4

Factors T00302; GAL4; Species: yeast, Saccharomyces cerevisiae.

Matrix



	Info	N	A	C	G	T	Consensu
01	0.210	11.00	1	5	3	2	N
02	0.210	11.00	5	2	1	3	N
03	0.210	11.00	3	2	1	5	N
04	1.561	11.00	1	10	0	0	C
05	1.561	11.00	0	0	10	1	G
06	1.561	11.00	0	1	10	0	G
07	0.132	11.00	4	3	3	1	N
08	0.132	11.00	1	3	4	3	N
09	0.177	11.00	2	4	4	1	N
10	0.691	11.00	7	0	2	2	A
11	0.904	11.00	1	8	2	0	C
12	0.678	11.00	4	1	0	6	W
13	0.210	11.00	1	3	5	2	N
14	0.904	11.00	0	2	1	8	T
15	0.314	11.00	1	6	2	2	C
16	0.323	11.00	1	5	4	1	S
17	0.509	11.00	2	1	1	7	T
18	1.561	11.00	0	10	1	0	C
19	2.000	11.00	0	11	0	0	C
20	2.000	11.00	0	0	11	0	G
21	1.155	11.00	8	0	0	3	A
22	1.054	11.00	7	0	4	0	R
23	0.565	11.00	2	6	3	0	S

Basis 11 genomic binding sites from 6 genes

Gene Regulatory Networks

Bi188

cis-regulation

E. coli **and phage regulatory circuits provided a powerful conceptual framework for understanding gene regulation in multicellular organisms.**

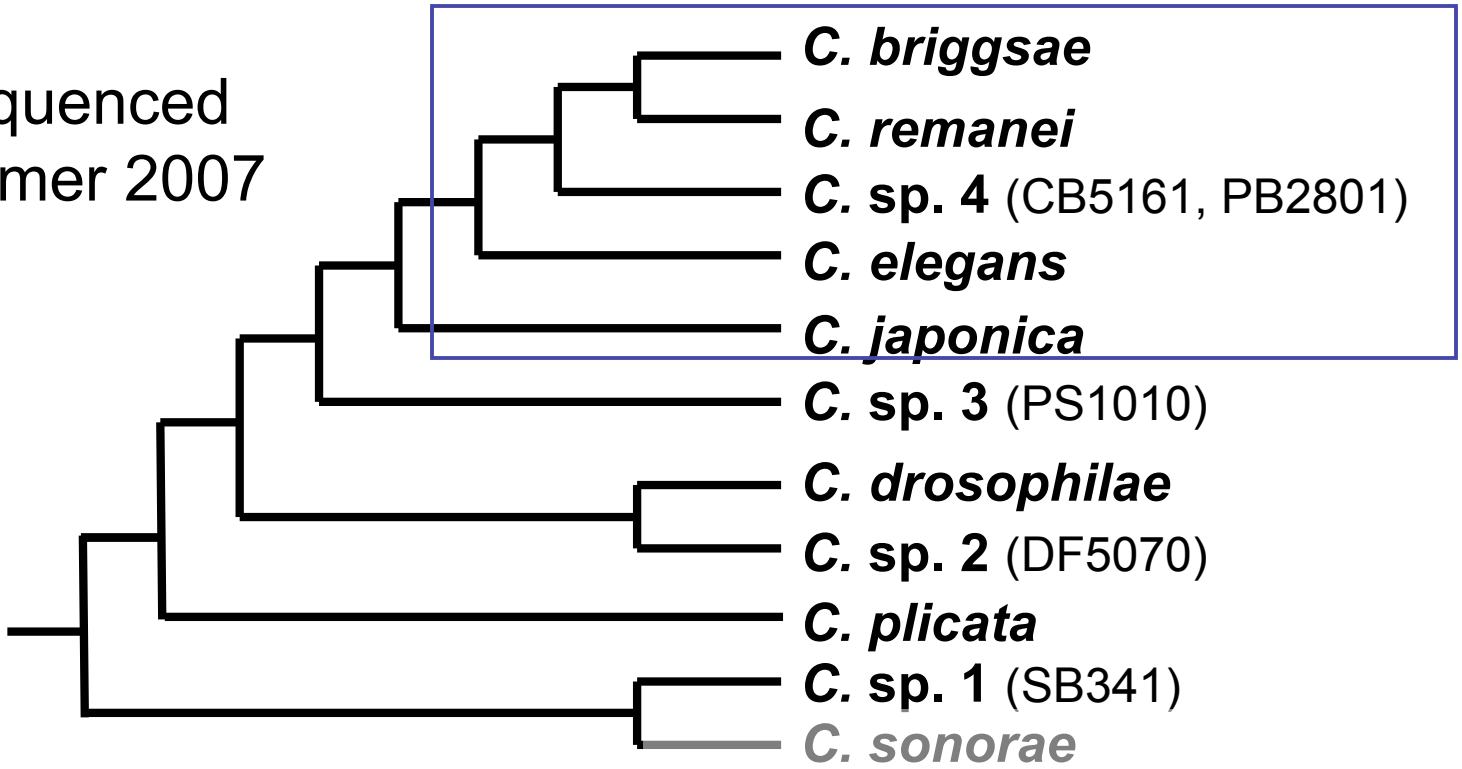
cis regulatory elements

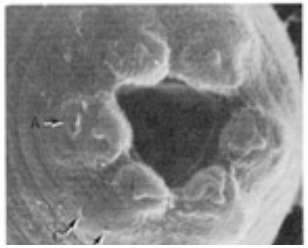
conservation to find regions

sets of genes to find motifs

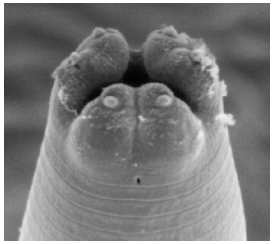
Caenorhabditis species

Five sequenced
by Summer 2007

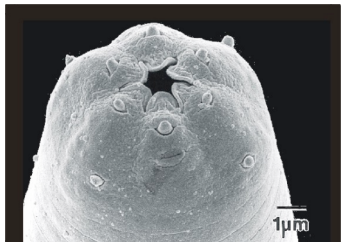




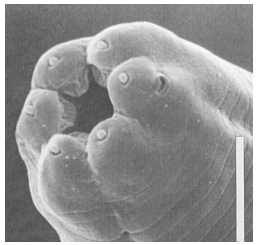
C. elegans



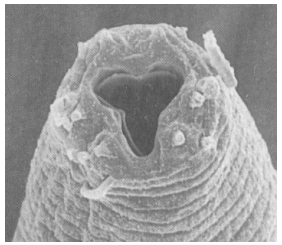
C. japonica



C. sp. 3 PS1010



C. drosophilae



C. sonorae

Reference: Kiontke, K. and David H.A. Fitch. (2005). *www.wormbook.org*.

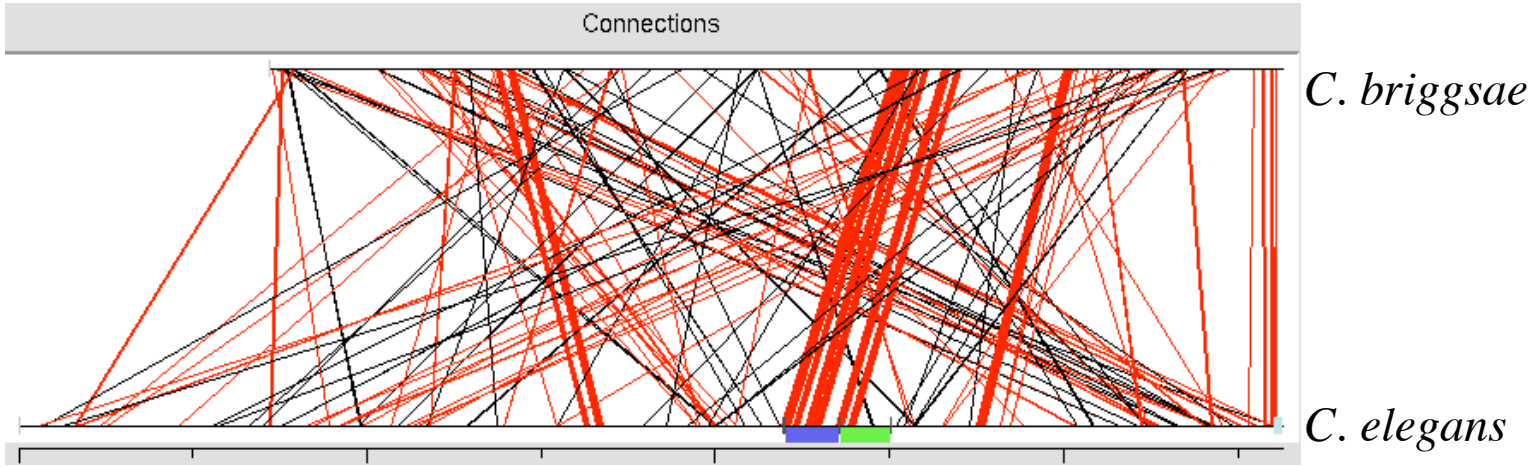
lin-3 Anchor Cell Enhancer

```
              POU  E-box    POU                          Ftz-F1                E-box

elegans   TATTCAATGCACCTGTGTATTTTATGCTGGTTT--T-TTCTTGTGACCCTGAAAACTGTACACACAGGTGTTCTT

briggsae  TAGTTGGAACACCTGCAATTTATGCTGCCATACAGGATTTGTGTGACCCTGAT---------CACAGGTGTTCTC
```

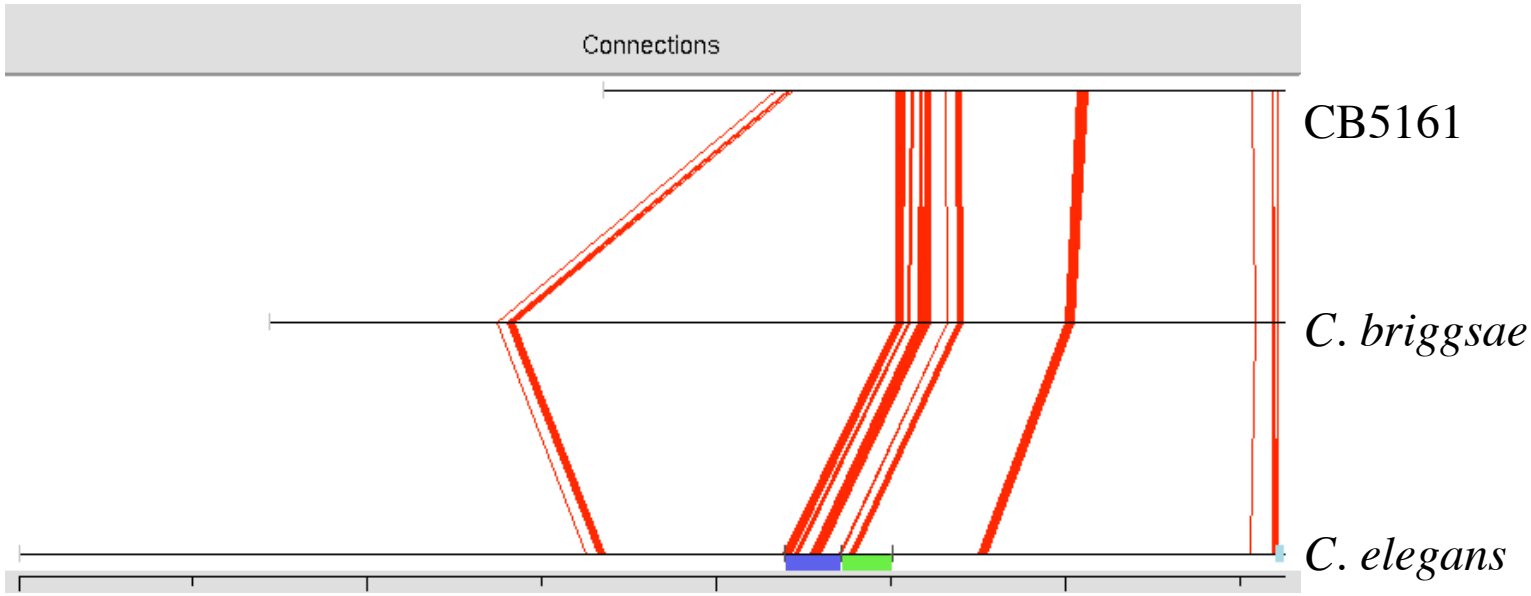
Byung Hwang [Devel. 2004]
John DeModena, Erich Schwarz

Ungapped blocks =~ regulatory sites in *lin-11*

21/30:



21/30:

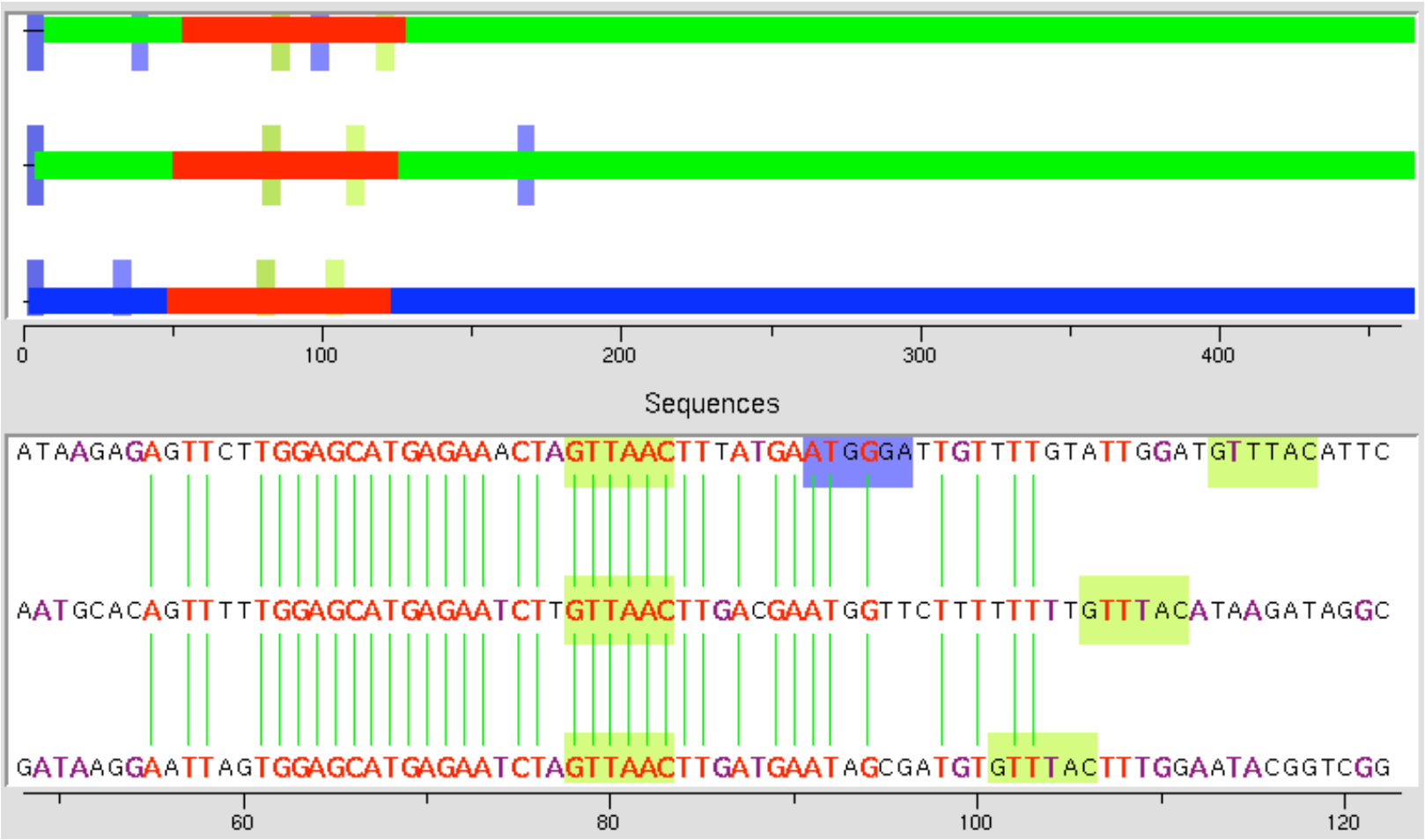


0 2000 4000 6000 nt

Erich Schwarz

Vulval and uterine elements: Gupta and Sternberg (2002), Dev Biol. *247*, 102-115; unpub. res.

Blocks only *partly* overlap smaller sites (*lin-11*)



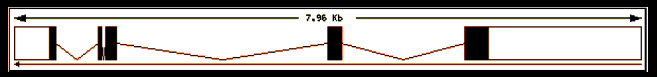
'ATGGGA' and 'GTTWAC' identified by YMF/Explanators

Reference: Sinha and Tompa (2003), Nucleic Acids Res. *31*, 3586-3688.

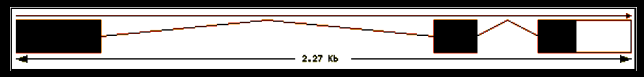
Erich Schwarz

C. elegans vs. C. briggsae vs. CB5161: lin-39 and ceh-13

- Window: 30 Threshold: 25

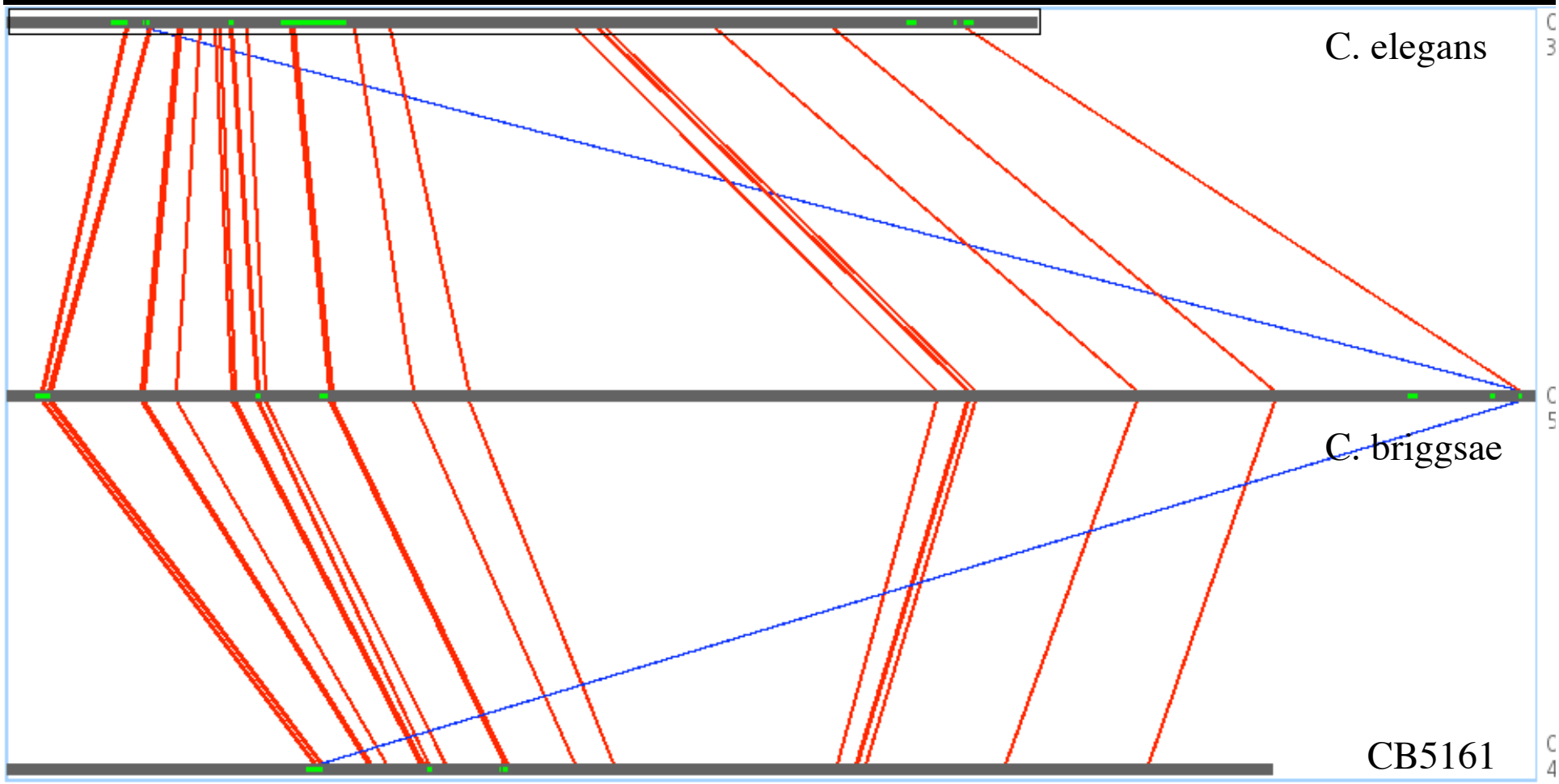


Lin-39 (- strand)

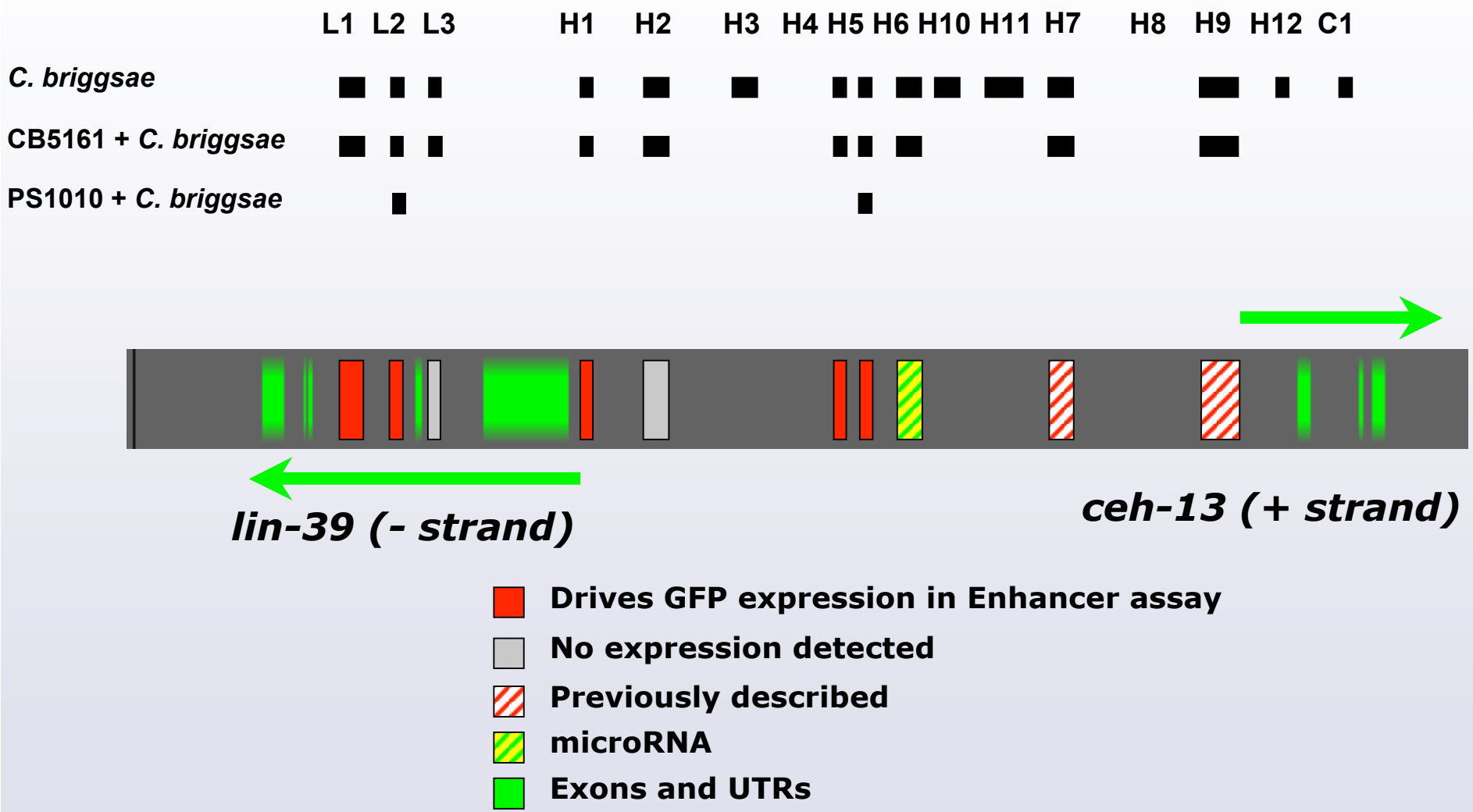


Ceh-13 (+ strand)

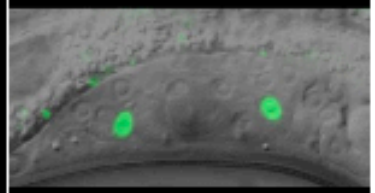
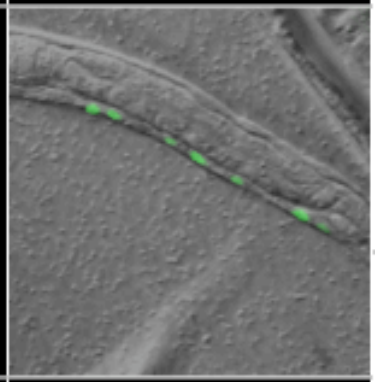
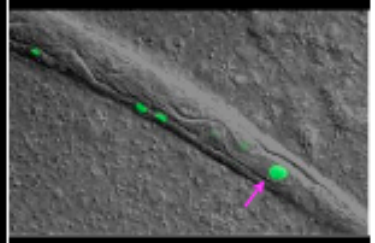
H1 H2 H4 H5 H7 H9

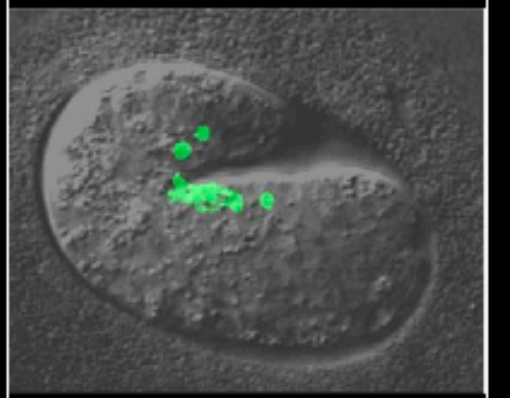
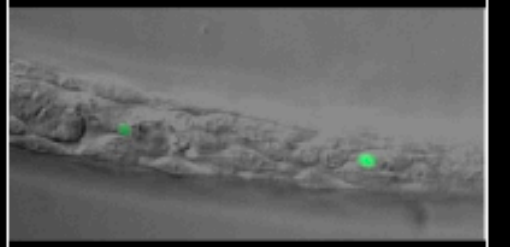


Ungapped blocks = ~2% Hox DNA



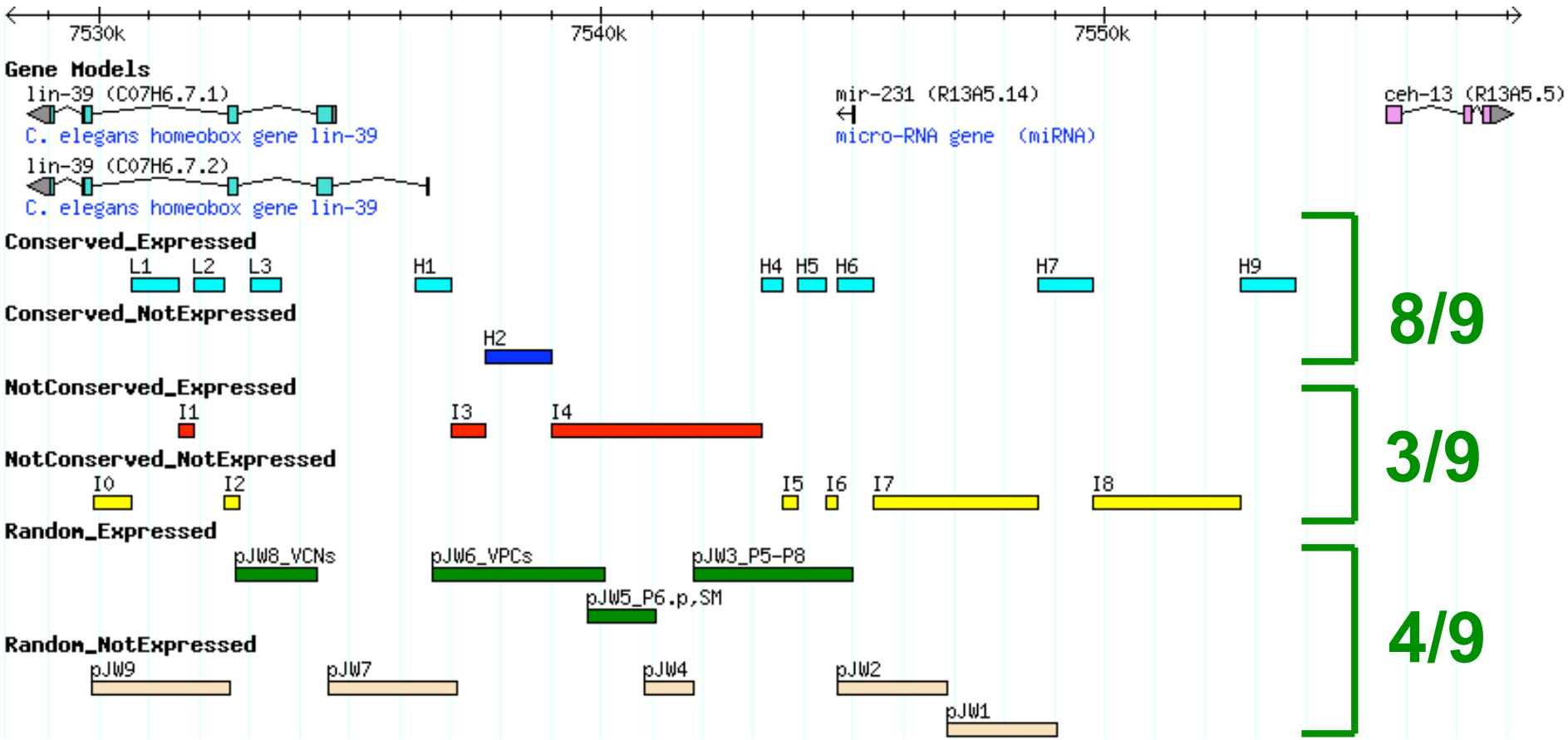
Steven Kuntz

Insert	Cells	Image
L1	vulval muscle	
L2	ventral cord neurons	
L2	Q cells	

Insert	Cells	Image
H1	MS lineage or Capa and Da lineages	
H1	V6 cells	

conservation is a good indicator of function

lin-39–ceh-23 hox cluster on chr III



Steven Kuntz (+Barbara Wold)

Gleason/Eisenmann Dev Biol. 2006

Tests of methods: Prediction of *dpy-8* element

Cuticle collagen family (*dpy-7*, *dpy-8* etc.)

algorithm

Paircomp

Binomial

WABA

P value:

e^-10

0.05^2

0.05

4.6e^-5

0.05^2

0.05

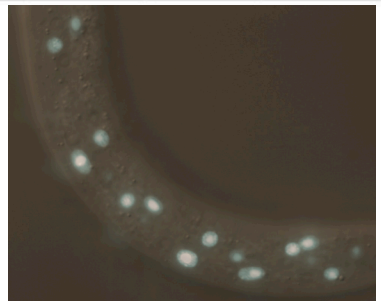
A13.4

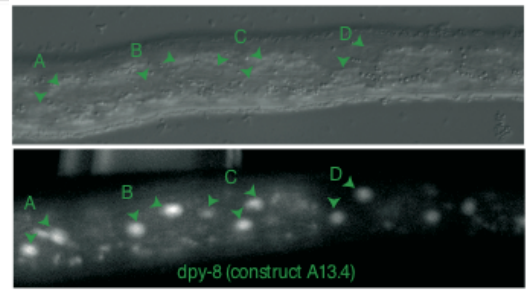
A8.3

dpy-8 exons

Experimentally Tested

0 800 1600 2400 3200 4000 4800

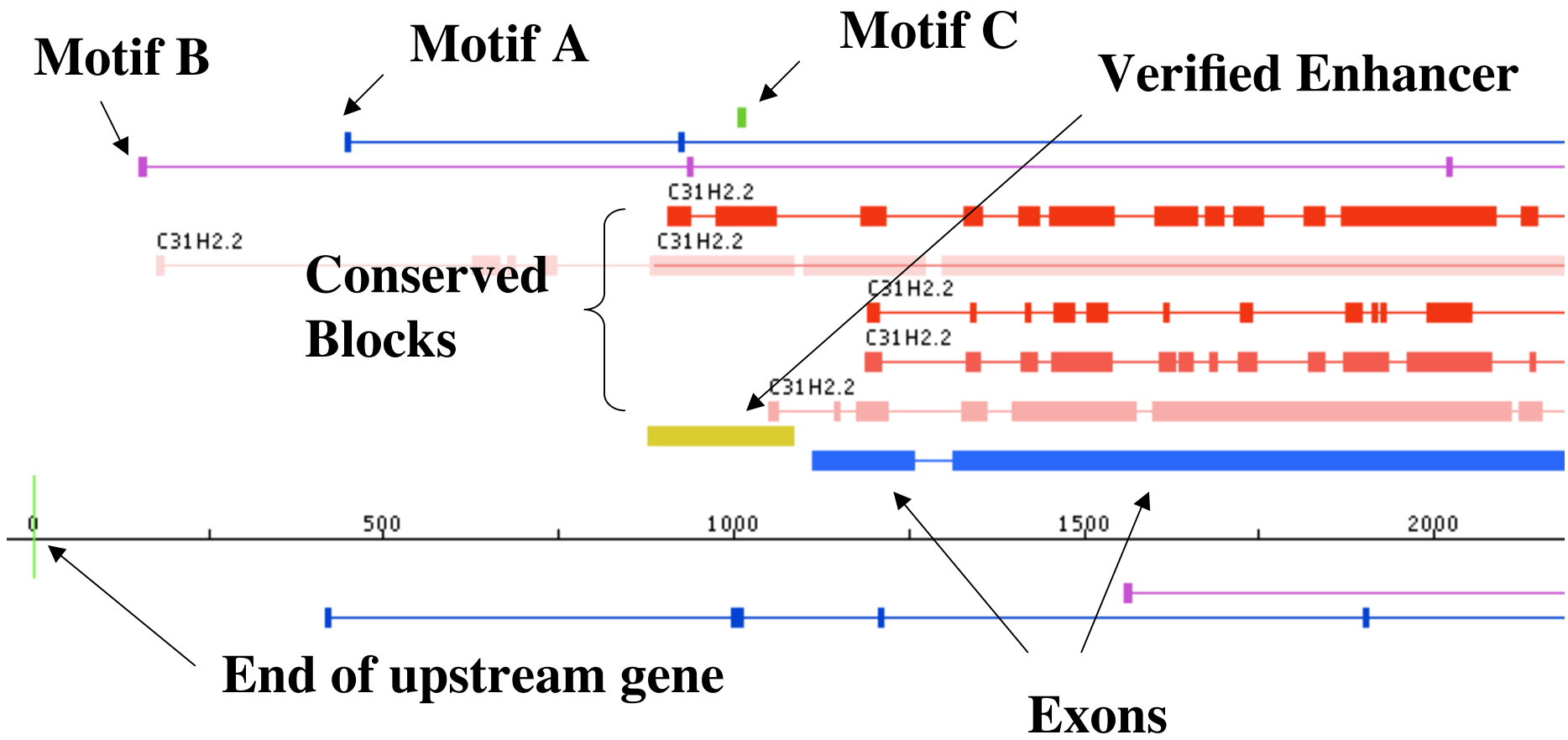




Alok Saldanha

Integrating motifs with conservation & enhancers

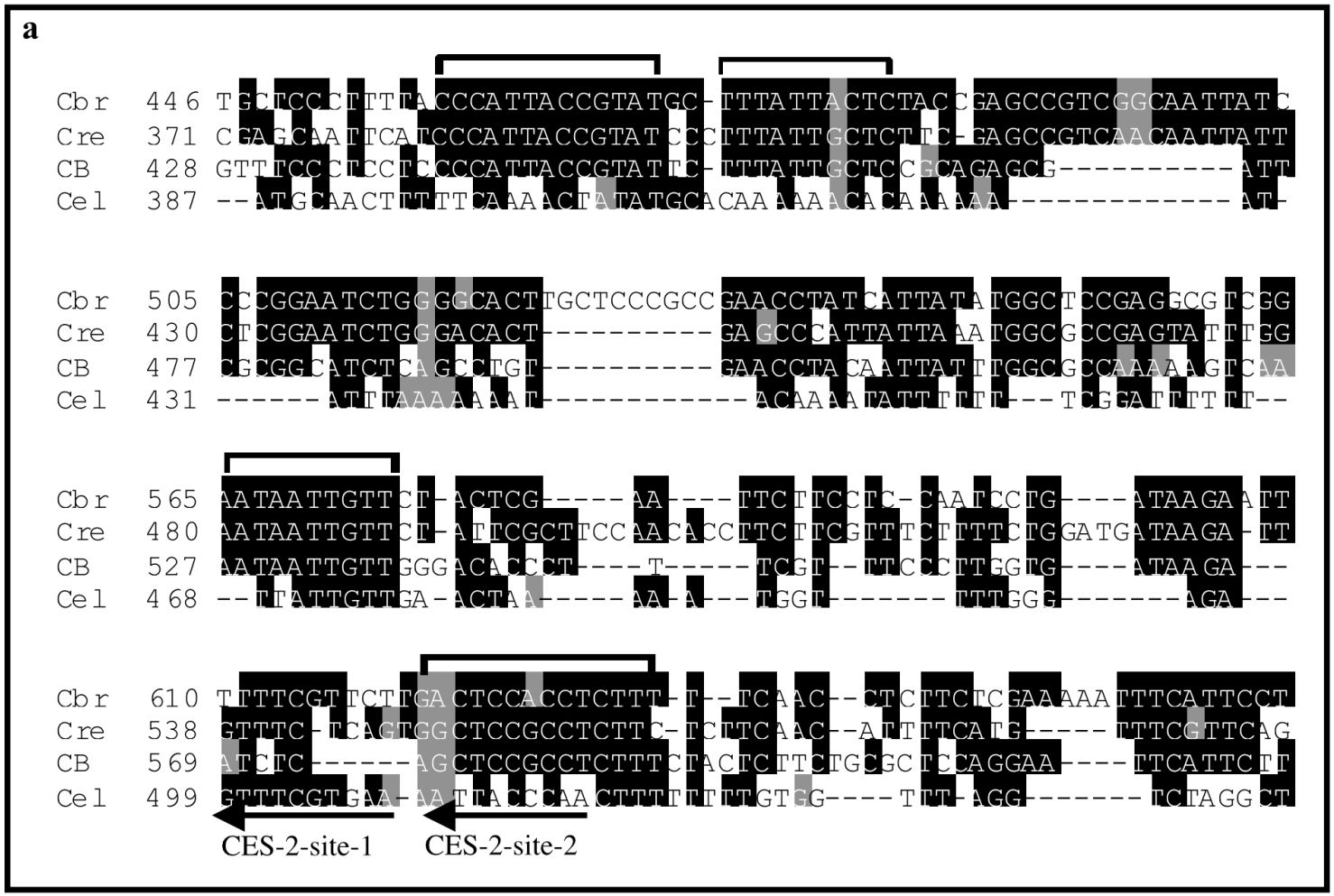
(*dpy-8* upstream)



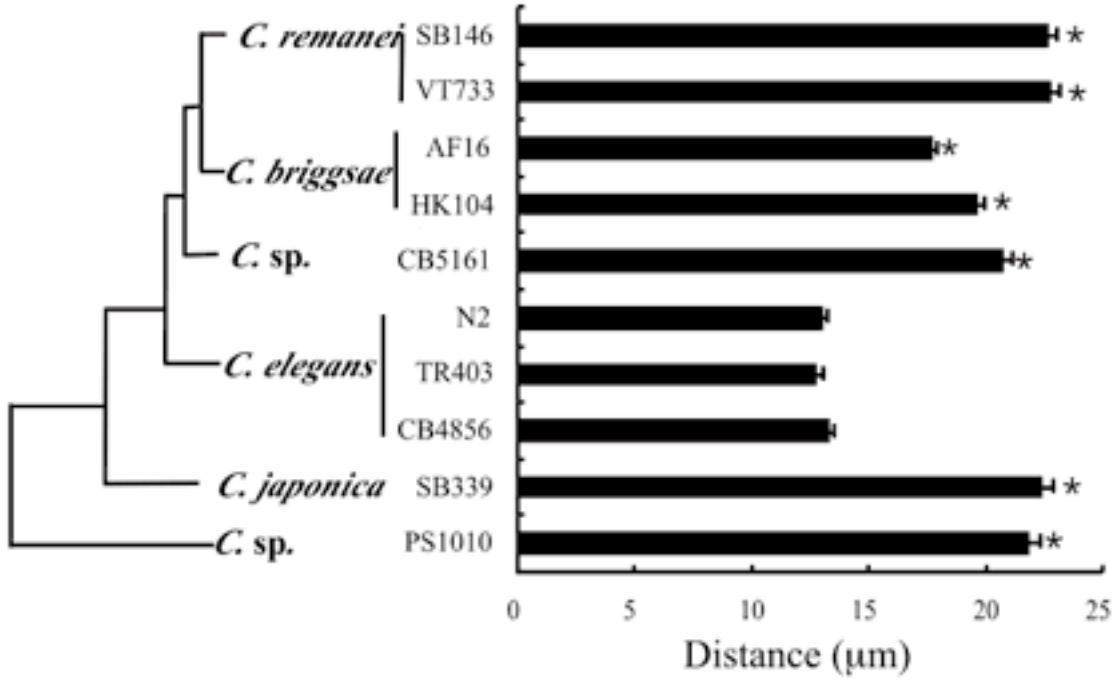
Alok Saldanha; Ali Mortazavi

Display: Apollo (moving to Gbrowse)

The proximal *lin-48* sequences in *C. elegans* are distinct from those in other *Caenorhabditis* species



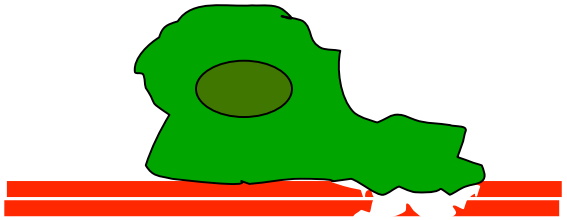
Wang & Chamberlin 2004 Nat Genet


C. remanei
SB146
VT733
C. briggsae
AF16
HK104
C. sp.
CB5161
C. elegans
N2
TR403
CB4856
C. japonica
SB339
C. sp.
PS1010
0
5
10
15
20
25
Distance (μm)

Two distinct transcriptional regulatory pathways in the *C. elegans* anchor cell

regulator | **Target genes** | **Biological Function**

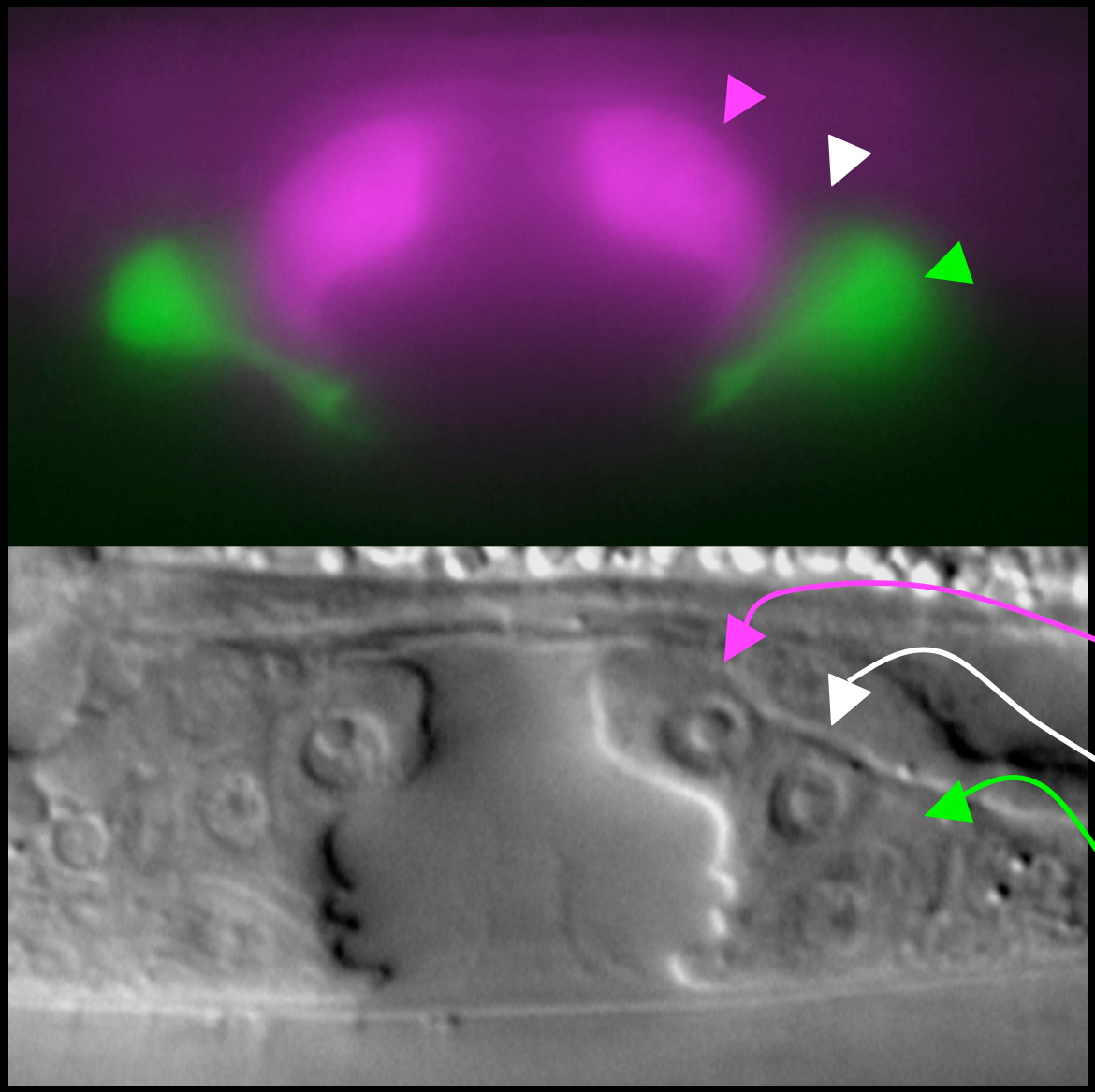
FOS-1 → **ZMP-1**, **HIM-4**, **CDH-3**, **etc.** → **INVASION**



HLH-2
AND
some NHR
(Ftz-F1 type) → **LIN-3** → **Vulval INDUCTION**

B. Hwang Dev. 2004; M. Kirouac Dev. Biol. 2003 ; D. Sherwood Cell 2005

mature vulval cell-type specific genes



cdh-3::cfp
ceh-2::yfp

Takao Inoue
MoD 2002
PNAS 2005

Enhancer assays

~12 genes with vulval cell type specific gene expression

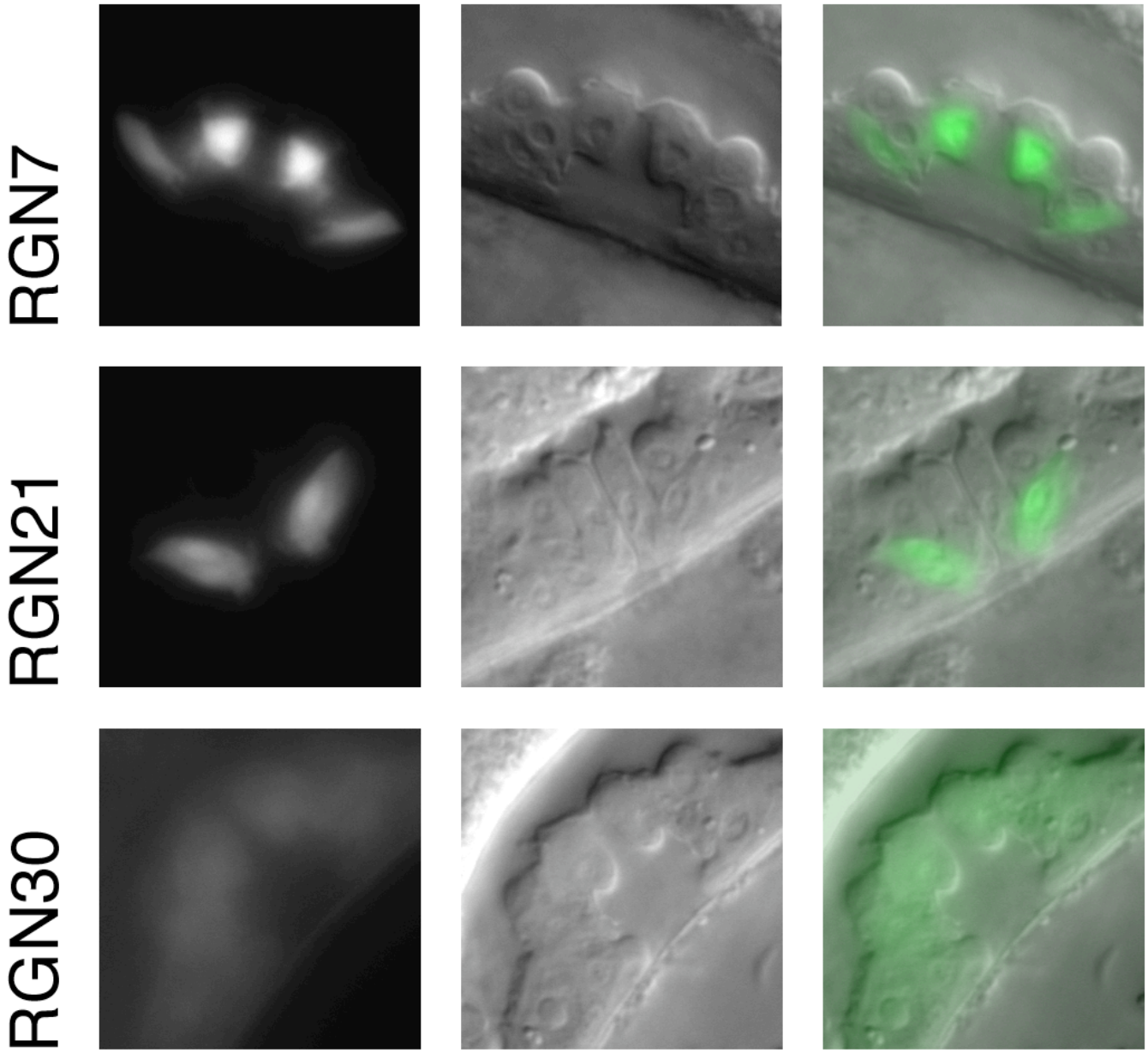
Defined 48 conserved (*elegans-briggsae*) regions (RGN)

~200 bp: 9/32 had vulval expression

then, sub-regions ~80 bp

grl-4	**RGN7**	**A, B, D**
grl-4	**RGN7a**	**A, B, D**
pax-2	**RGN12**	**C, D**
col-48	**RGN14**	**B, C, D**
col-48	**RGN14b**	**B, C, D**
F48B9.5	**RGN16**	**A, B, C**
sqv-4	**RGN17**	**C, D**
sqv-4	**RGN28**	**E, F**
sqv-4	**RGN28a**	**E, F**
sqv-4	**RGN30**	**vulval cells**
daf-6	**RGN44**	**vulval cells**

Takao Inoue, Shahla Gharib

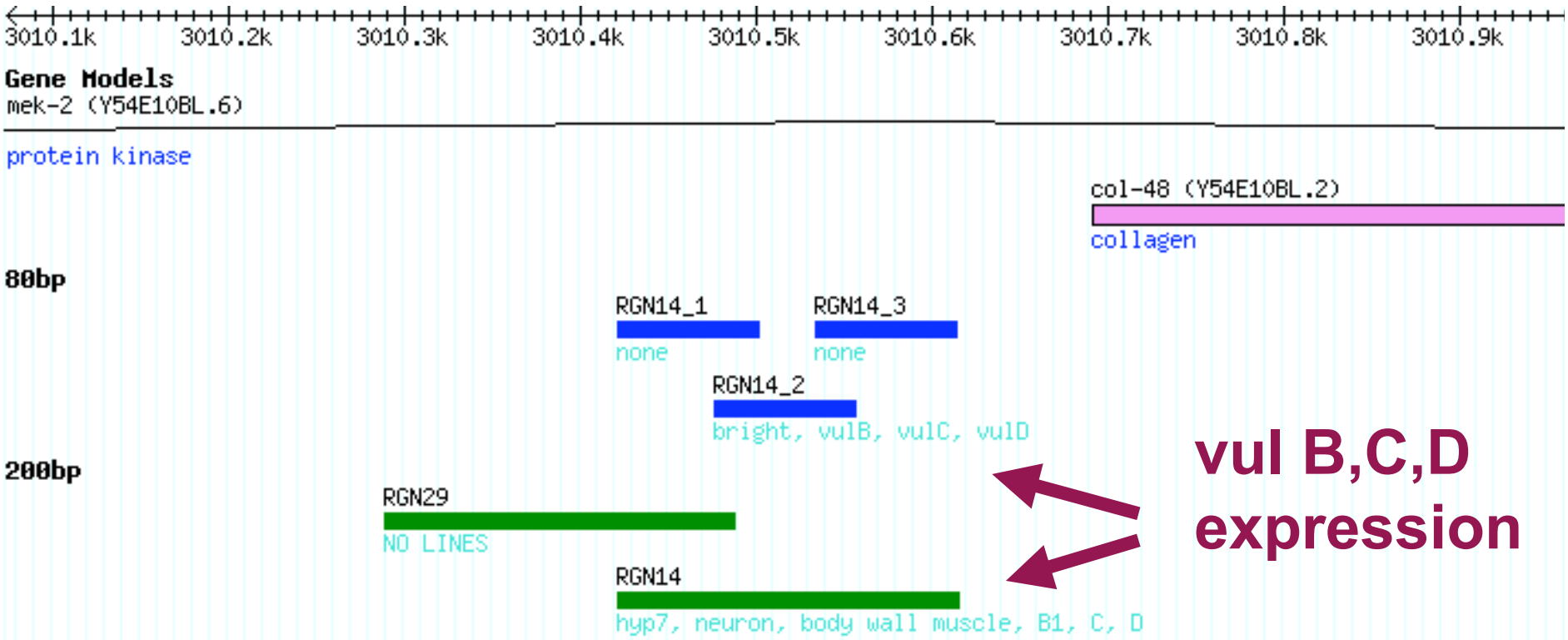


A, B, D

vulval cells

Takao Inoue, Shahla Gharib

col-48: 1 of 1 region tested expressed; 1 of 3 subregions expressed



Takao Inoue, Shahla Gharib

Enhancer assays
~12 genes with vulval cell type specific gene expression
Defined 48 conserved (*elegans-briggsae*) regions (RGN)
~200 bp: 9/32 had vulval expression
then, sub-regions ~80 bp

grl-4	**RGN7**	**A, B, D**
grl-4	**RGN7a**	**A, B, D**
pax-2	**RGN12**	**C, D**
col-48	**RGN14**	**B, C, D**
col-48	**RGN14b**	**B, C, D**
F48B9.5	**RGN16**	**A, B, C**
sqv-4	**RGN17**	**C, D**
sqv-4	**RGN28**	**E, F**
sqv-4	**RGN28a**	**E, F**
sqv-4	**RGN30**	**vulval cells**
daf-6	**RGN44**	**vulval cells**

yeast 1-hybrid with Walhout lab

Txn Factors

Takao Inoue, Shahla Gharib

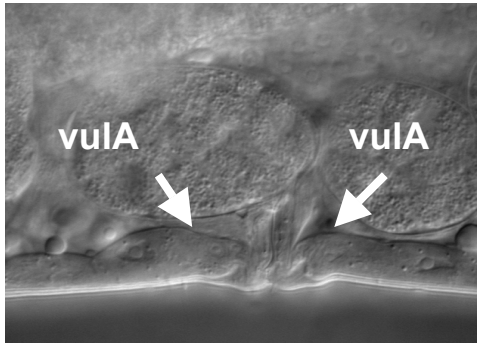
cis regulatory elements

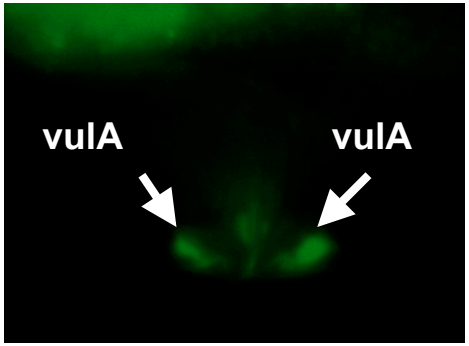
conservation to find regions

sets of genes to find motifs

cis-Regulatory Analysis of *zmp-1 (an MMP)*

3-species comparison of the 386 bp zmp-1 vulA-E enhancer





Three motifs with MEME

Construct			Expression: % adults vulA (n)	Expression: % late L4 vulE (n)
mk50-51	1052	1438	86	100
103/4	1135	1150	24 (41)	0 (23)
105/6	1206	1219	65 (48)	48 (25)
107/8	1235	1250	27 (15)	13 (16)

substitute 15 bp and test in transgenic worms

Ted Ririe

No Motif Discovery Method is Perfect (Yet)

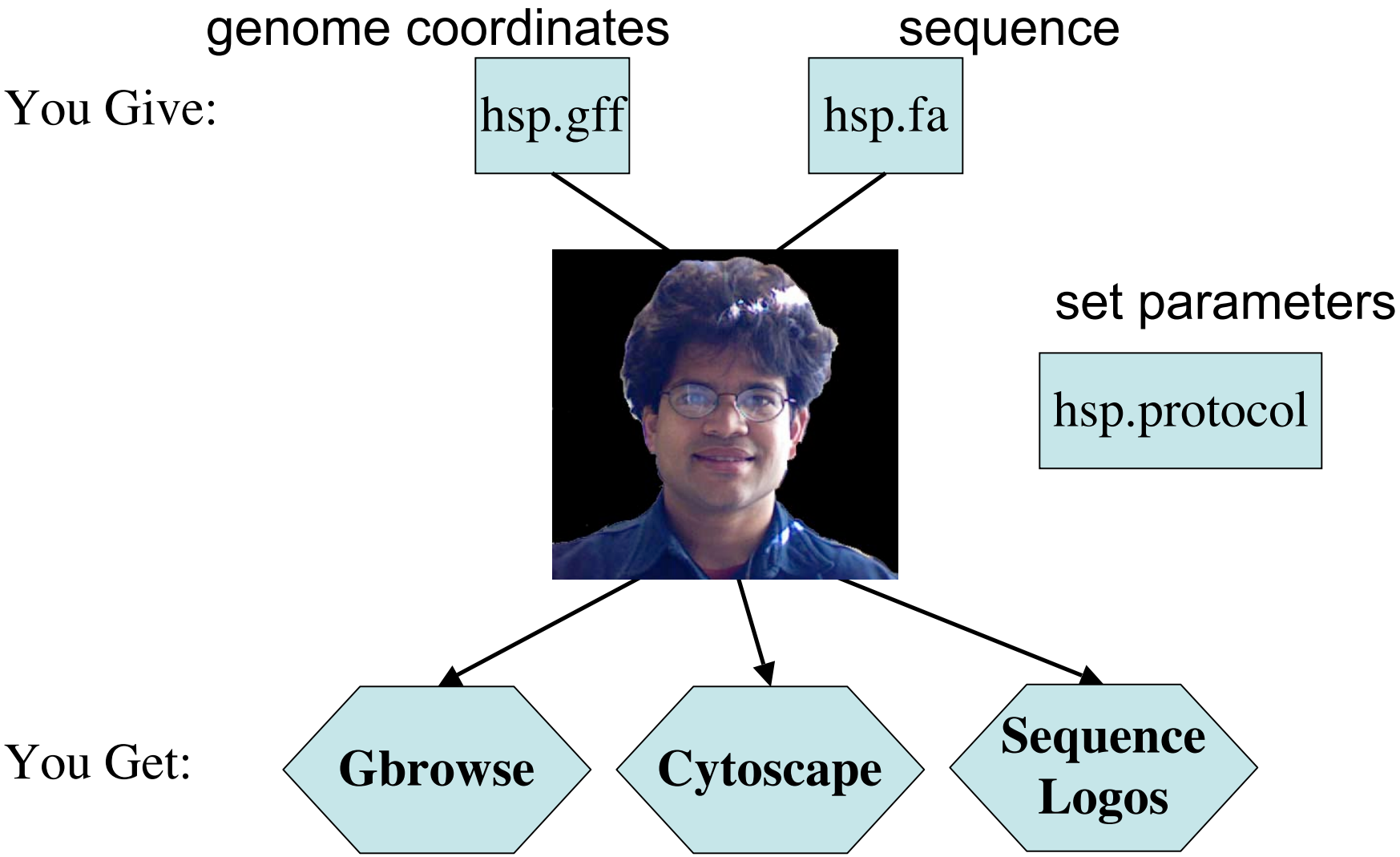
Different papers use different motif finders, often referring to an *ad hoc* search to find one that worked

Tompa et al. (Nat Biotech 2005) compared 13 different motif finders, and concluded: "*Biologists would be well advised to use a few complementary tools in combination*."

We devised a reproducible method of combination:

Maximal Clique Motif Reduction (MCMR)

MCMR conceptual plan



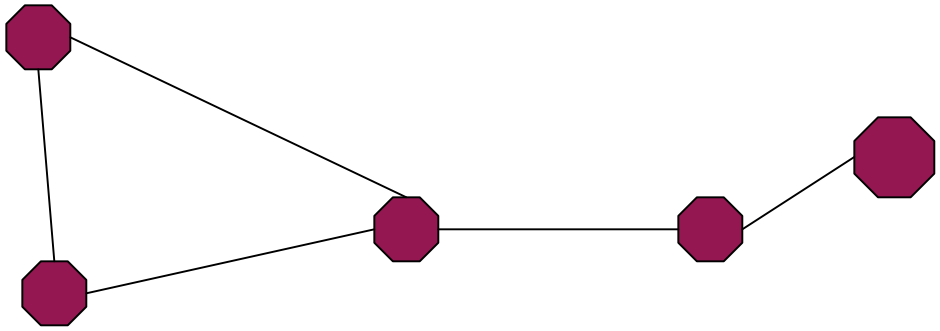
Maximal Clique Motif Reduction

Given a set of input motifs:

- **motif 1**
- **motif 2**
- **motif 3**
- **motif 4**
- **motif 5**

Alok Saldanha

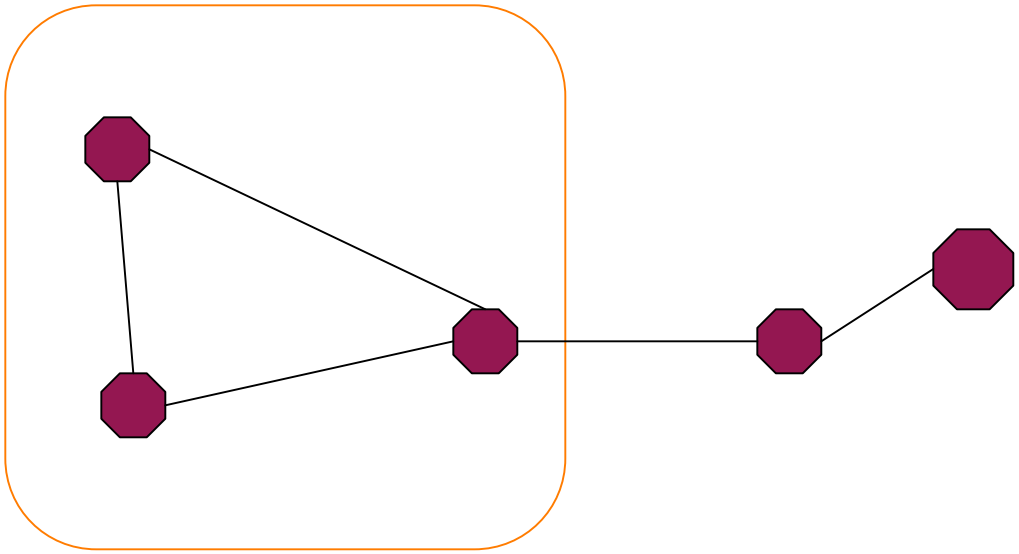
Maximal Clique Motif Reduction



Calculate and threshold pairwise similarity

Alok Saldanha

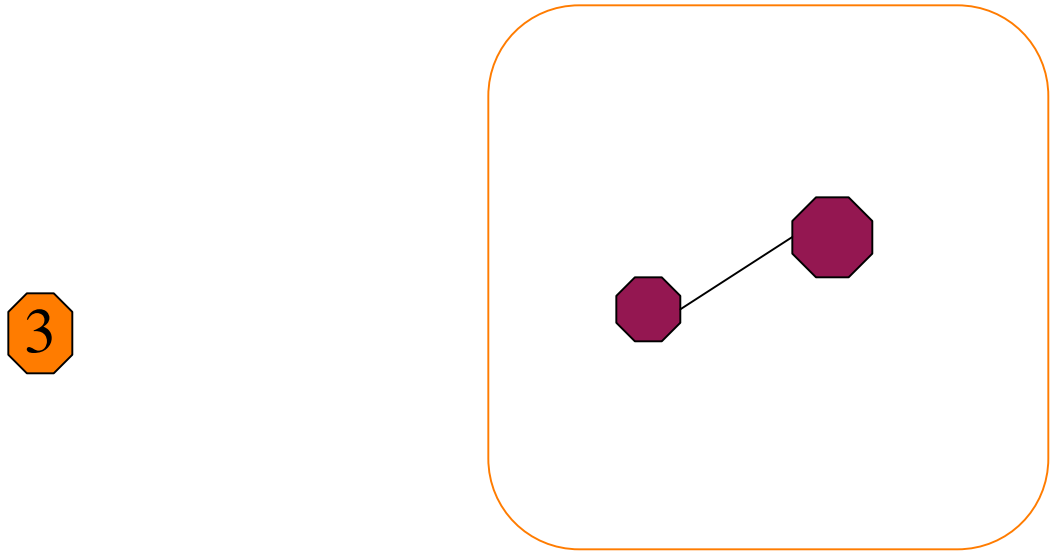
Maximal Clique Motif Reduction



Find maximal clique of largest size

Alok Saldanha

Maximal Clique Motif Reduction



Replace clique with "reduced" version, disconnect from graph and repeat

Alok Saldanha

Maximal Clique Motif Reduction

Alok Saldanha

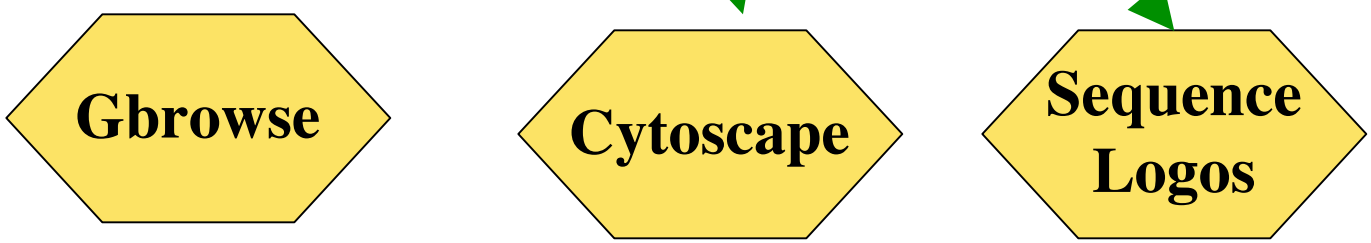
Example of MCMR Output

Maximal Clique Motif Reduction

1. Browse Sequence Logos
2. Cytoscape download
3. Gbrowse download

Input: Top 25 genes from GuhaThakurta et al Genome Research 2002:

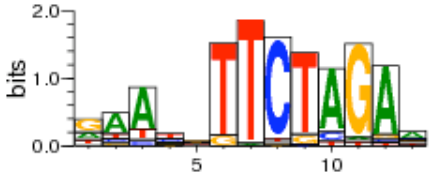
F55A12.9a
T27A3.4
C12C8.1
W02D9.10
F33H12.6
F41C3.2
ZK1290.5
M05D6.1
D2013.9
R03D7.2
F44E5.5
F08G2.5
C30C11.4
T28H11.7
C50F7.5
H14N18.1a
T27E4.2
T27F2.4
F58E10.4
C25D7.1
C25F9.2
Y38H6C.7
F53A9.2
R07B1.4
F09B9.1

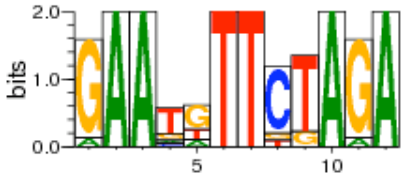
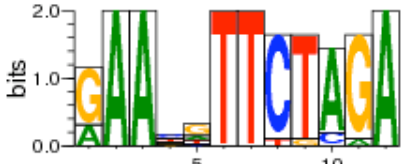
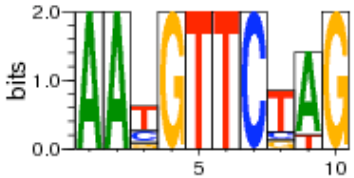
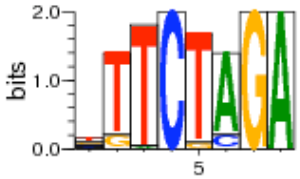


Sequence logo of top clique from hsp_guha

clique 0 had 6 motifs (list hsp_guha_2_19.fa 25)

Core:

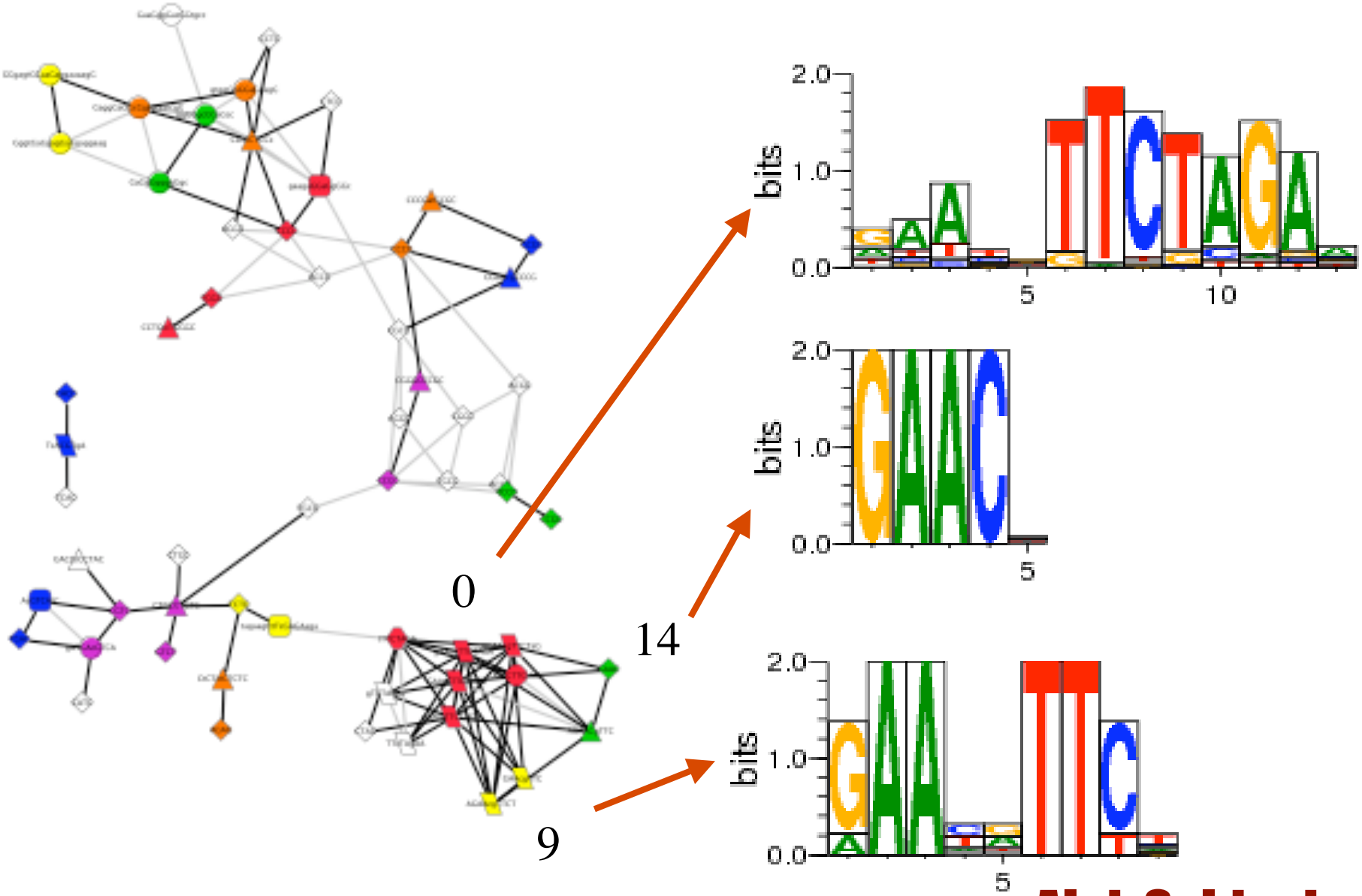
gaatgTTCTAGAa  49 matches in 19 regions 

GAAtgTTCTAGA		12 matches in 8 regions	motif 3 from WEEDER run 12209 (ID 23)	hsp_guha_2_19.fa 
gAAcgTTCTAGA		15 matches in 12 regions	motif 5 from WEEDER run 12209 (ID 25)	hsp_guha_2_19.fa 
AAtGTTCtAG		7 matches in 5 regions	motif 4 from NMICA run 12212 (ID 53)	hsp_guha_2_19.fa 
tTTCTAGA		34 matches in 19 regions	motif 4 from MOTIFSAMPLER run 12214 (ID 145)	hsp_guha_2_19.fa 

alignment **logo** **matches** **id** **coverage**

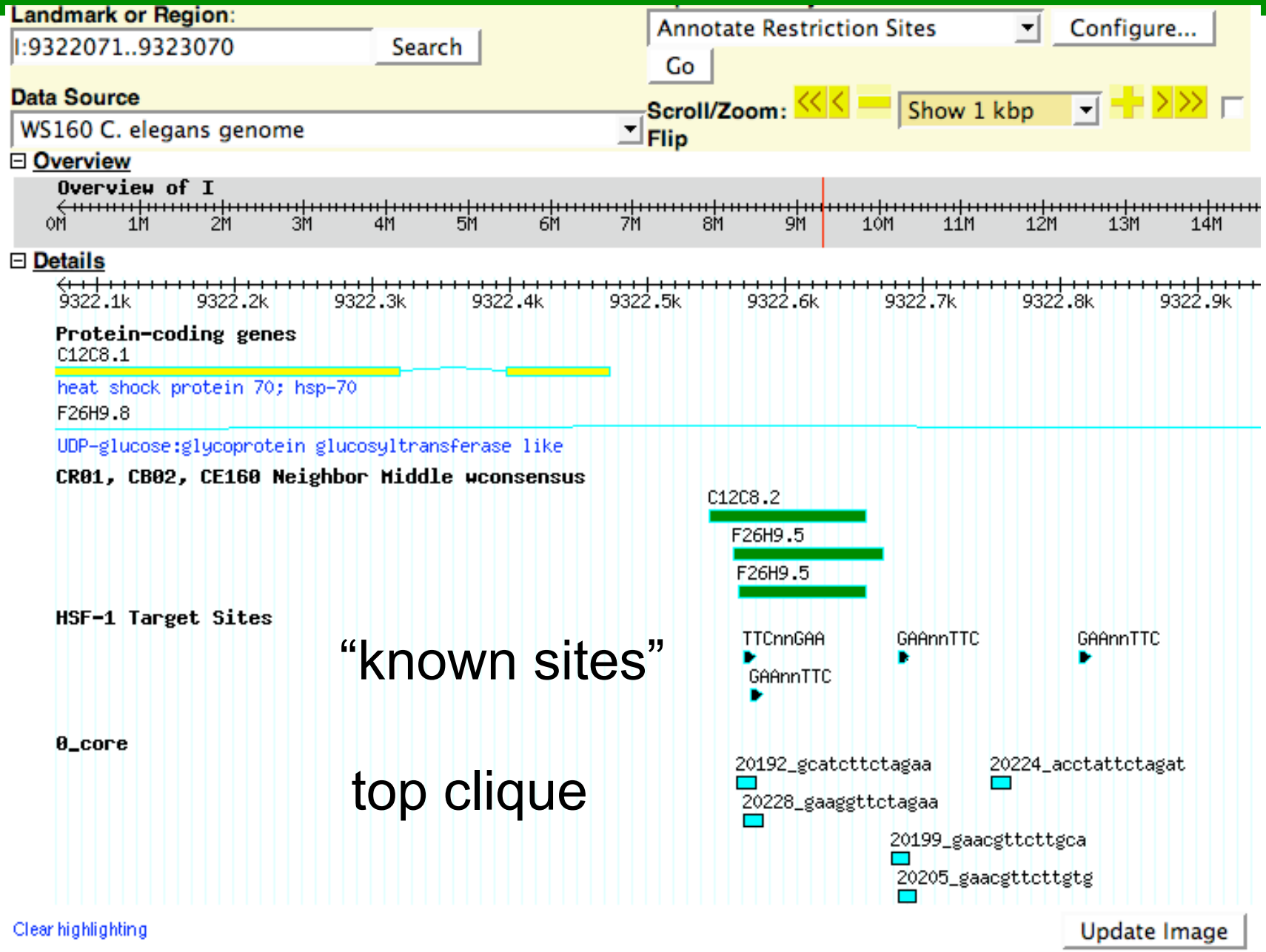
19/23 have HSE

Cytoscape graph of all motifs from hsp_guha



Alok Saldanha

Gbrowse mapping of top clique from hsp_guha



"known sites"

top clique

individual motif finder

site	AlignACE	Improbizer	MEME	Mobydick	N-MICA	Weeder	YMF	MCMR
mec	11	-	6	4	-	1	-	1
HSE	21	Y	2	-	Y	1	-	1
HSAS	8	Y	4	7	Y	-	1	2
AIY	2	Y	-	-	-	4	-	2

= rank: Yes, present; -, absent

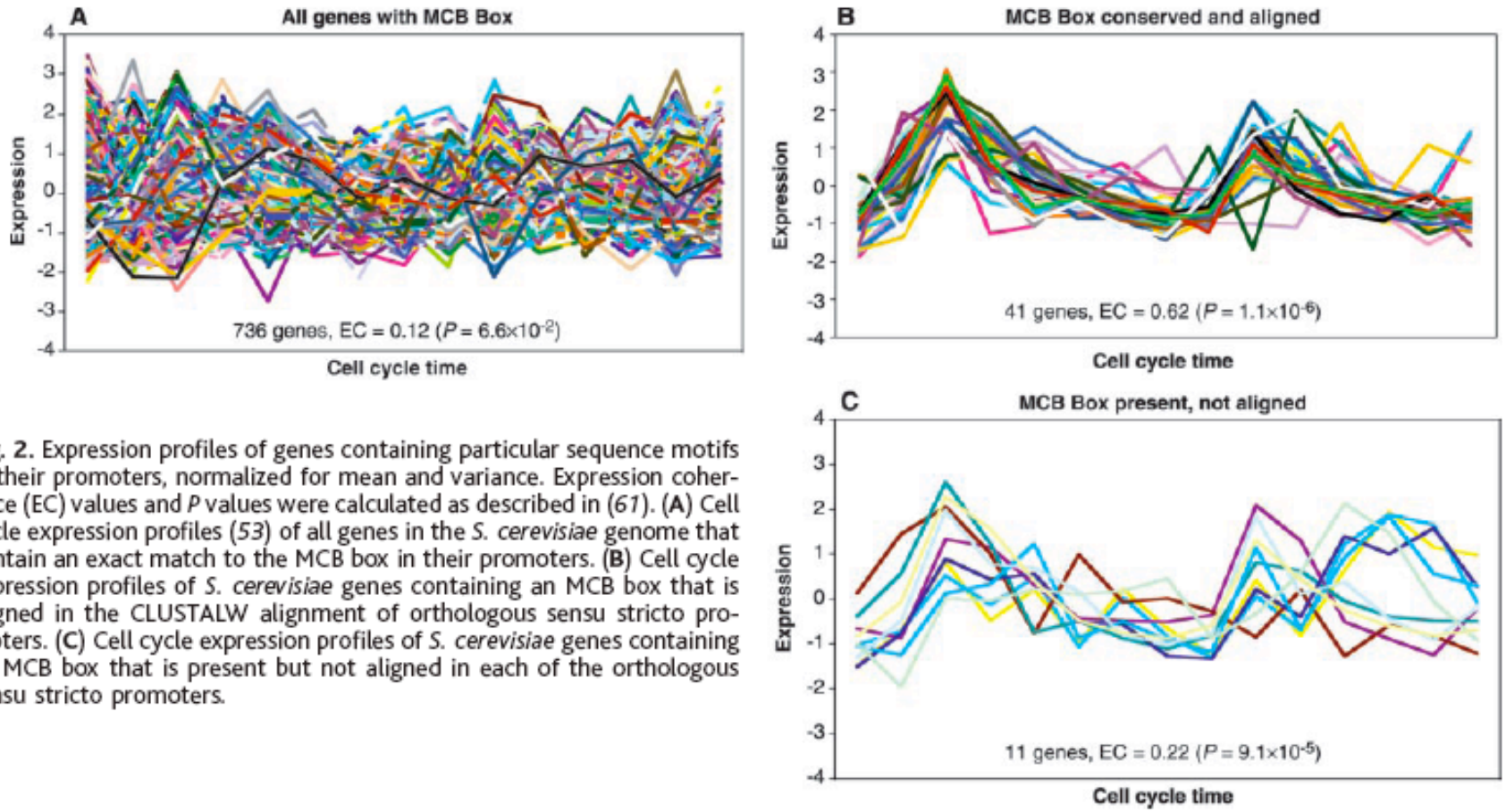


Fig. 2. Expression profiles of genes containing particular sequence motifs in their promoters, normalized for mean and variance. Expression coherence (EC) values and *P* values were calculated as described in (*61*). (**A**) Cell cycle expression profiles (*53*) of all genes in the *S. cerevisiae* genome that contain an exact match to the MCB box in their promoters. (**B**) Cell cycle expression profiles of *S. cerevisiae* genes containing an MCB box that is aligned in the CLUSTALW alignment of orthologous sensu stricto promoters. (**C**) Cell cycle expression profiles of *S. cerevisiae* genes containing an MCB box that is present but not aligned in each of the orthologous sensu stricto promoters.

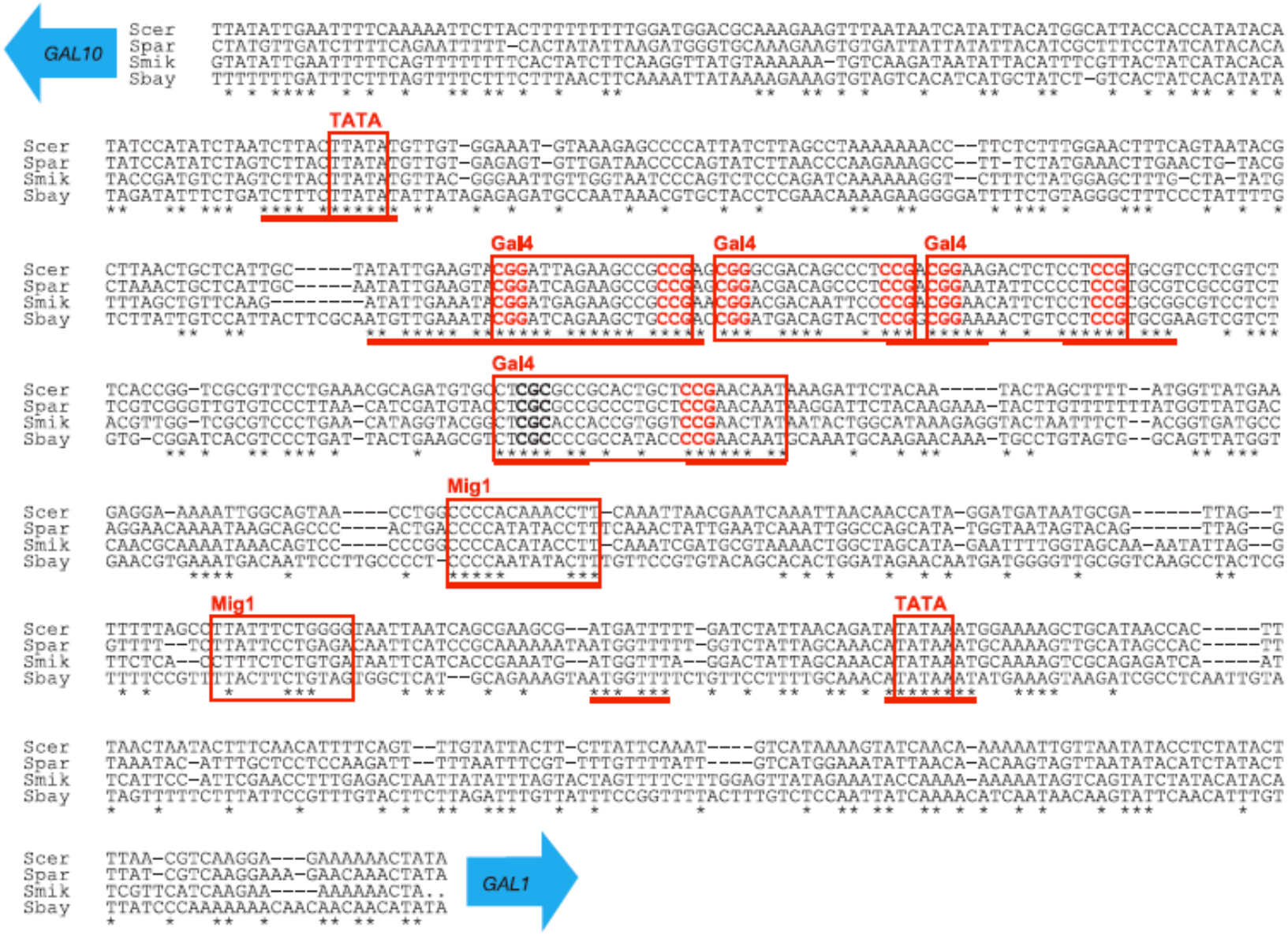


Figure 6 Conservation in the *GAL1–GAL10* intergenic region. Multiple alignment of the four species shows a strong overlap between functional nucleotides and stretches of conservation. Asterisks denote conserved positions in the multiple alignment. Blue arrows denote the start and transcriptional orientation of the flanking ORFs. Experimentally validated factor-binding footprints are boxed and labelled according to the bound factor. Stretches of conserved nucleotides are underlined. Nucleotides matching the published Gal4 motif are shown in red. The fourth experimentally validated site differs: it shows a longer footprint and a non-standard consensus motif (bold). This variant motif is also conserved across all four species. Scer, *S. cerevisiae*; Spar, *S. paradoxus*; Smik, *S. mikatae*; Sbay, *S. bayanus*.