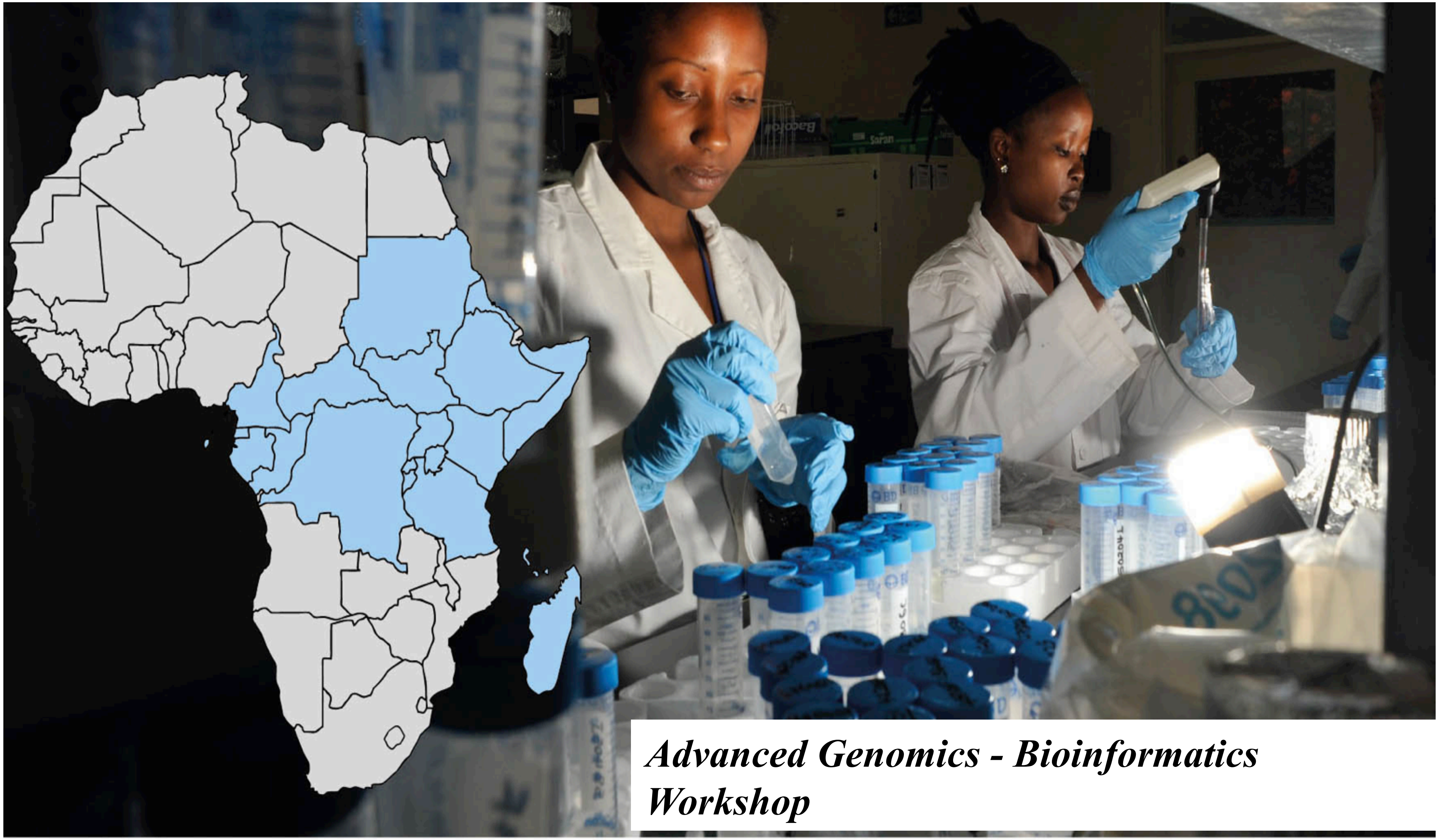


Advanced Genomics - Bioinformatics Workshop

Mark Wamalwa
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
m.wamalwa@cgiar.org

7th – 18th September 2015

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

THE NEW PARTNERSHIP FOR AFRICA'S DEVELOPMENT
NEPAD
A PROGRAMME OF THE AFRICAN UNION

biosciences eastern and central africa

Evolution and constraint on *cis*-regulatory motifs

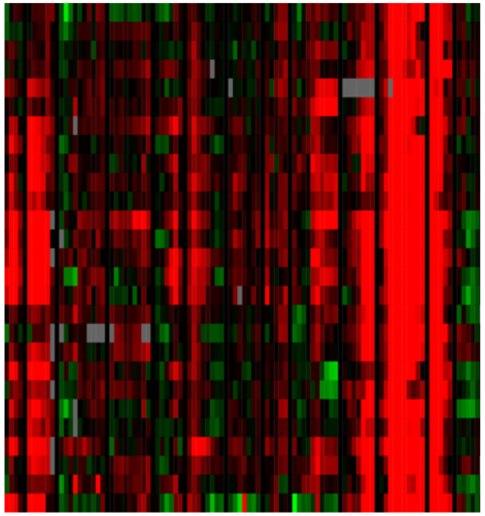
(focusing on TF binding sites)

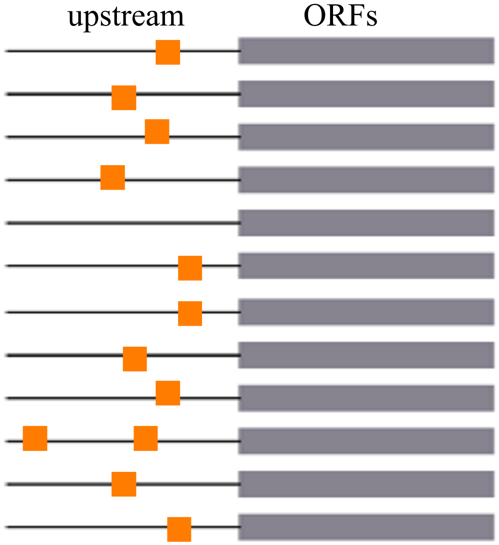
Many DNA binding proteins recognize specific (often short) DNA sequences.

Often bind ‘degenerate’ sequences, since some bases more important for contact.

Many work cooperatively with other factors to bind.

Representing the set of TF binding sites *within* a genome





Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G
IPUAC consensus:	**W**	**G**	**A**	**T**	**G**	**G**	**A**	**T**	**N**	**G**

Position-weight matrices are a better representation

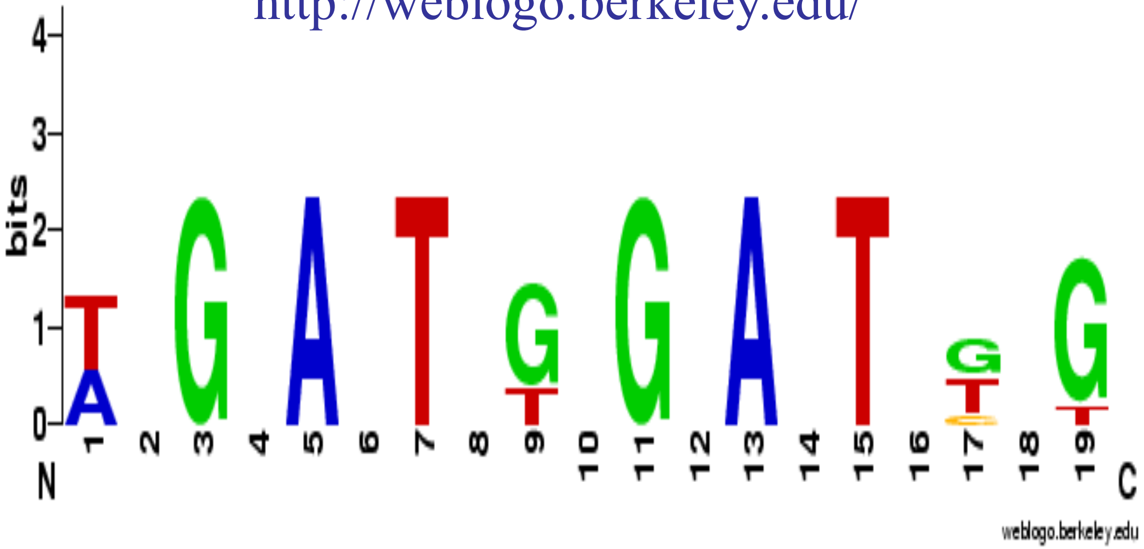
Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G

PWM represents frequencies of each base at each position in the motif

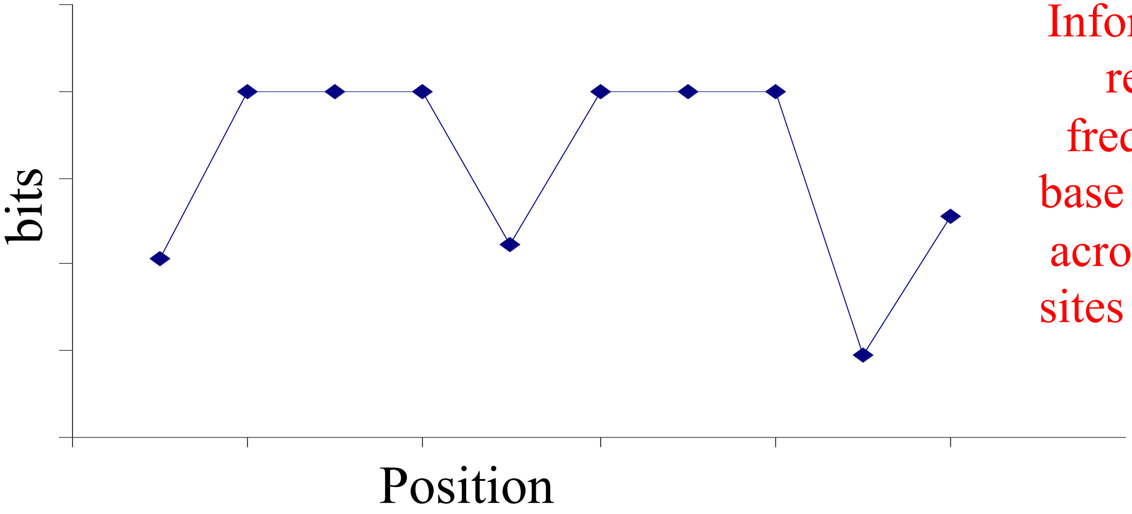
G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

Web-logo: A graphical representation of PWMs

http://weblogo.berkeley.edu/



Information Profile:



Information content represents the frequency of each base at each position across ALL binding sites in an individual

To study the evolution of cis regulatory elements, we first need to identify them in genomes

Identification of *cis*-regulatory elements

Computational predictions:

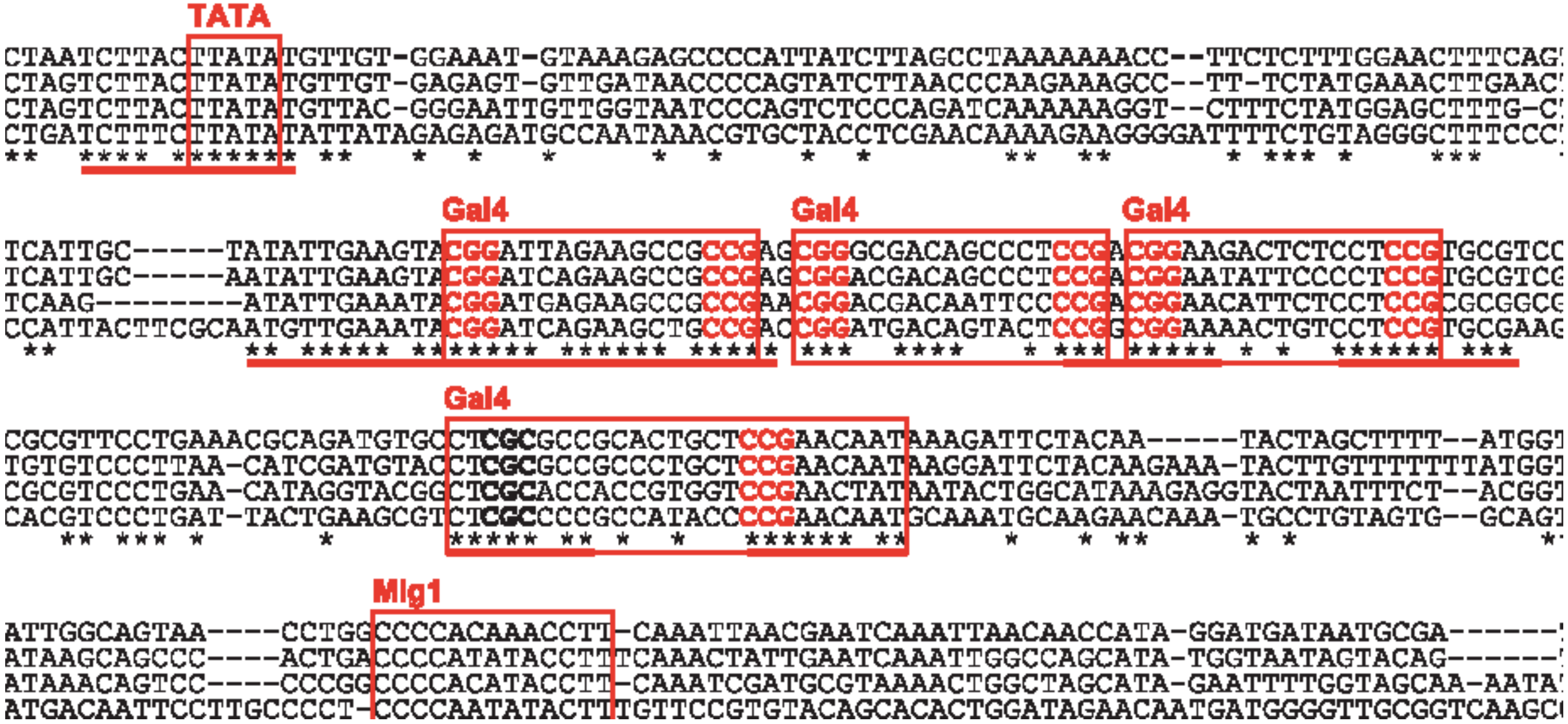
1. Scan genome for matches to known matrix/consensus

   *problem is that there are many nonfunctional in the genome - poor predictor of functio*n

2. Phylogenetic footprinting: overly-conserved sequences in multiple alignments

   Variation within element is typically lower than surrounding 'nonfunctional' DNA

Simplest case: stretches of very highly conserved sequence


TATA
CTAATCTTACTTATATGTTGT-GGAAAT-GTAAAGAGCCCCATTATCTTAGCCTAAAAAAACC--TTCTCTTTGGAACTTTCAG
CTAGTCTTACTTATATGTTGT-GAGAGT-GTTGATAACCCCAGTATCTTAACCCAAGAAAGCC--TT-TCTATGAAACTTGAAC
CTAGTCTTACTTATATGTTAC-GGGAATTGTTGGTAATCCCAGTCTCCCAGATCAAAAAAGGT--CTTTCTATGGAGCTTTG-C
CTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAGAAGGGGATTTTCTGTAGGGCTTTCCC
Gal4
Gal4
Gal4
TCATTGC-----TATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCC
TCATTGC-----AATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCG
TCAAG--------ATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCG
CCATTACTTCGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAG
Gal4
CGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAA-----TACTAGCTTTT--ATGGT
TGTGTCCCTTAA-CATCGATGTACCTCGCGCCGCCCTGCTCCGAACAATAAGGATTCTACAAGAAA-TACTTGTTTTTTTATGGT
CGCGTCCCTGAA-CATAGGTACGGCTCGCACCACCGTGGTCCGAACTATAATACTGGCATAAAGAGGTACTAATTTCT--ACGGT
CACGTCCCTGAT-TACTGAAGCGTCTCGCCCCGCCATACCCCGAACAATGCAAATGCAAGAACAAA-TGCCTGTAGTG--GCAGT
Mig1
ATTGGCAGTAA----CCTGGCCCCACAAACCTT-CAAATTAACGAATCAAATTAACAACCATA-GGATGATAATGCGA------
ATAAGCAGCCC----ACTGACCCCATATACCTTTCAAACTATTGAATCAAATTGGCCAGCATA-TGGTAATAGTACAG------
ATAAACAGTCC----CCCGGCCCCACATACCTT-CAAATCGATGCGTAAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATA
ATGACAATTCCTTGCCCCT-CCCCAATATACTTTGTTCCGTGTACAGCACACTGGATAGAACAATGATGGGGTTGCGGTCAAGC

Kellis *et al.* 2003 "Sequencing and comparison of yeast species to identify genes and regulatory elements"
Sequenced 4 closely related *Saccharomyces* genomes & identified conserved sequences in multiple alignments of orthologous sequences from the four species.

Need species close enough to get reliable DNA alignment

Position of elements has to be conserved for detection
(keep this in mind when we get to stabilizing selection at the end ...)

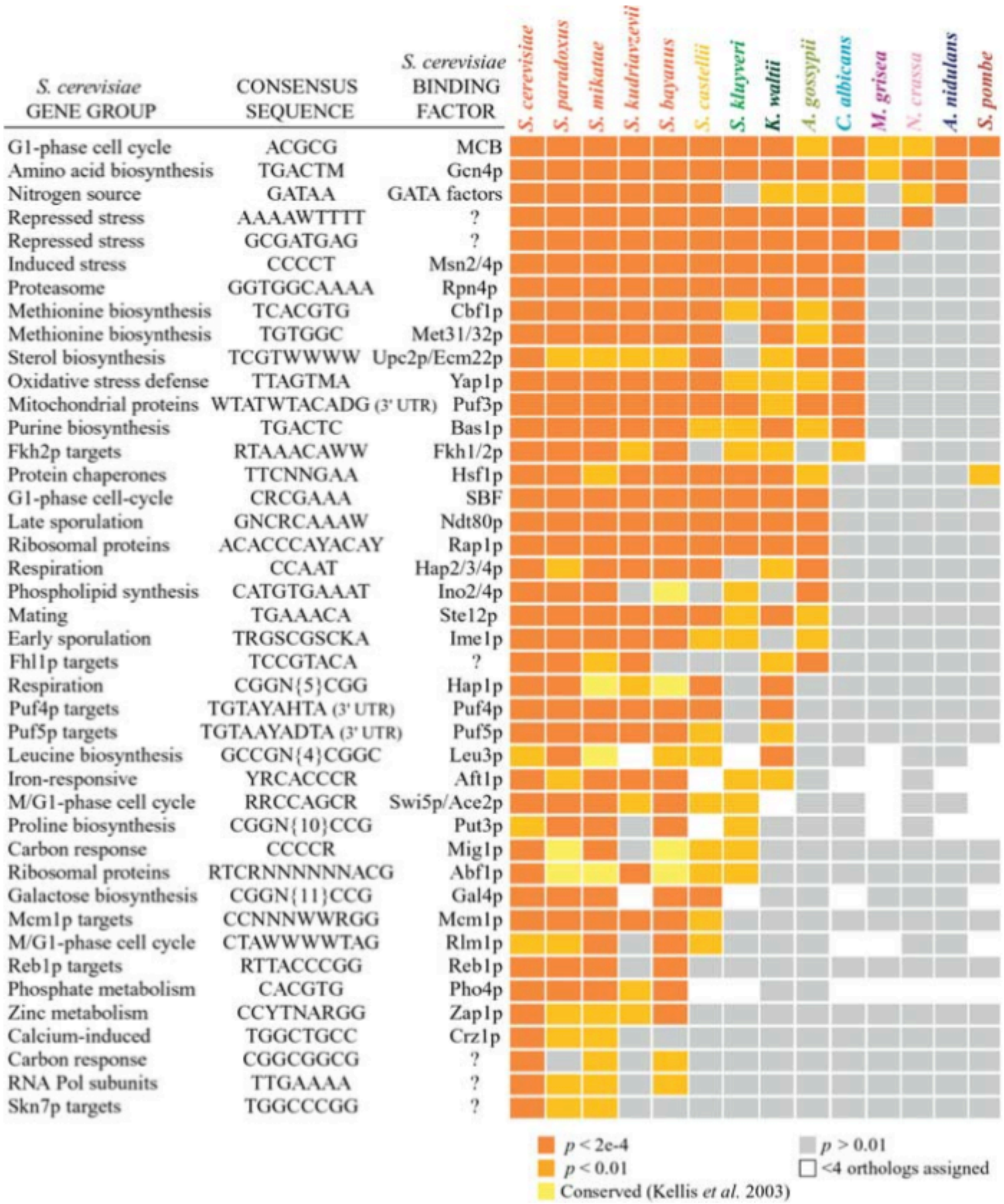
Identification of *cis*-regulatory elements

Computational predictions:

1. Scan genome for matches to known matrix/consensus
   problem is that there are many nonfunctional in the genome - poor predictor of function

2. Phylogenetic footprinting: overly-conserved sequences in multiple alignments
   Variation within element is typically lower than surrounding 'nonfunctional' DNA

3. Network/module approach: Focus on groups of co-regulated genes to increase statistical power
   Look for statistically significant enrichment of sequences in the group of upstream regions from a group of co-regulated genes



S. cerevisiae GENE GROUP	CONSENSUS SEQUENCE	S. cerevisiae BINDING FACTOR
G1-phase cell cycle	ACGCG	MCB
Amino acid biosynthesis	TGACTM	Gcn4p
Nitrogen source	GATAA	GATA factors
Repressed stress	AAAAWTTTT	?
Repressed stress	GCGATGAG	?
Induced stress	CCCCT	Msn2/4p
Proteasome	GGTGGCAAAA	Rpn4p
Methionine biosynthesis	TCACGTG	Cbf1p
Methionine biosynthesis	TGTGGC	Met31/32p
Sterol biosynthesis	TCGTWWWW	Upc2p/Ecm22p
Oxidative stress defense	TTAGTMA	Yap1p
Mitochondrial proteins	WTATWTACADG (3' UTR)	Puf3p
Purine biosynthesis	TGACTC	Bas1p
Fkh2p targets	RTAAACAWW	Fkh1/2p
Protein chaperones	TTCNNGAA	Hsf1p
G1-phase cell-cycle	CRCGAAA	SBF
Late sporulation	GNCRCAAAW	Ndt80p
Ribosomal proteins	ACACCCAYACAY	Rap1p
Respiration	CCAAT	Hap2/3/4p
Phospholipid synthesis	CATGTGAAAT	Ino2/4p
Mating	TGAAACA	Ste12p
Early sporulation	TRGSCGSCKA	Ime1p
Fhl1p targets	TCCGTACA	?
Respiration	CGGN{5}CGG	Hap1p
Puf4p targets	TGTAYAHTA (3' UTR)	Puf4p
Puf5p targets	TGTAAYADTA (3' UTR)	Puf5p
Leucine biosynthesis	GCCGN{4}CGGC	Leu3p
Iron-responsive	YRCACCCR	Aft1p
M/G1-phase cell cycle	RRCCAGCR	Swi5p/Ace2p
Proline biosynthesis	CGGN{10}CCG	Put3p
Carbon response	CCCCR	Mig1p
Ribosomal proteins	RTCRNNNNNNACG	Abf1p
Galactose biosynthesis	CGGN{11}CCG	Gal4p
Mcm1p targets	CCNNNWWRGG	Mcm1p
M/G1-phase cell cycle	CTAWWWWTAG	Rlm1p
Reb1p targets	RTTACCCGG	Reb1p
Phosphate metabolism	CACGTG	Pho4p
Zinc metabolism	CCYTNARGG	Zap1p
Calcium-induced	TGGCTGCC	Crz1p
Carbon response	CGGCGGCG	?
RNA Pol subunits	TTGAAAA	?
Skn7p targets	TGGCCCGG	?

"Conservation and evolution of *cis*-regulatory systems in ascomycete fungi"

Gasch *et al.* 2004
PLoS Biol

Results:

* Many conserved elements are connected to similar gene groups over 100's of millions of years.

* Some gene groups show show evidence of conserved co-regulation but evolved elements

* One example of co-evolved TF binding specificity and upstream sequence elements

Identification of *cis*-regulatory elements

Computational predictions:

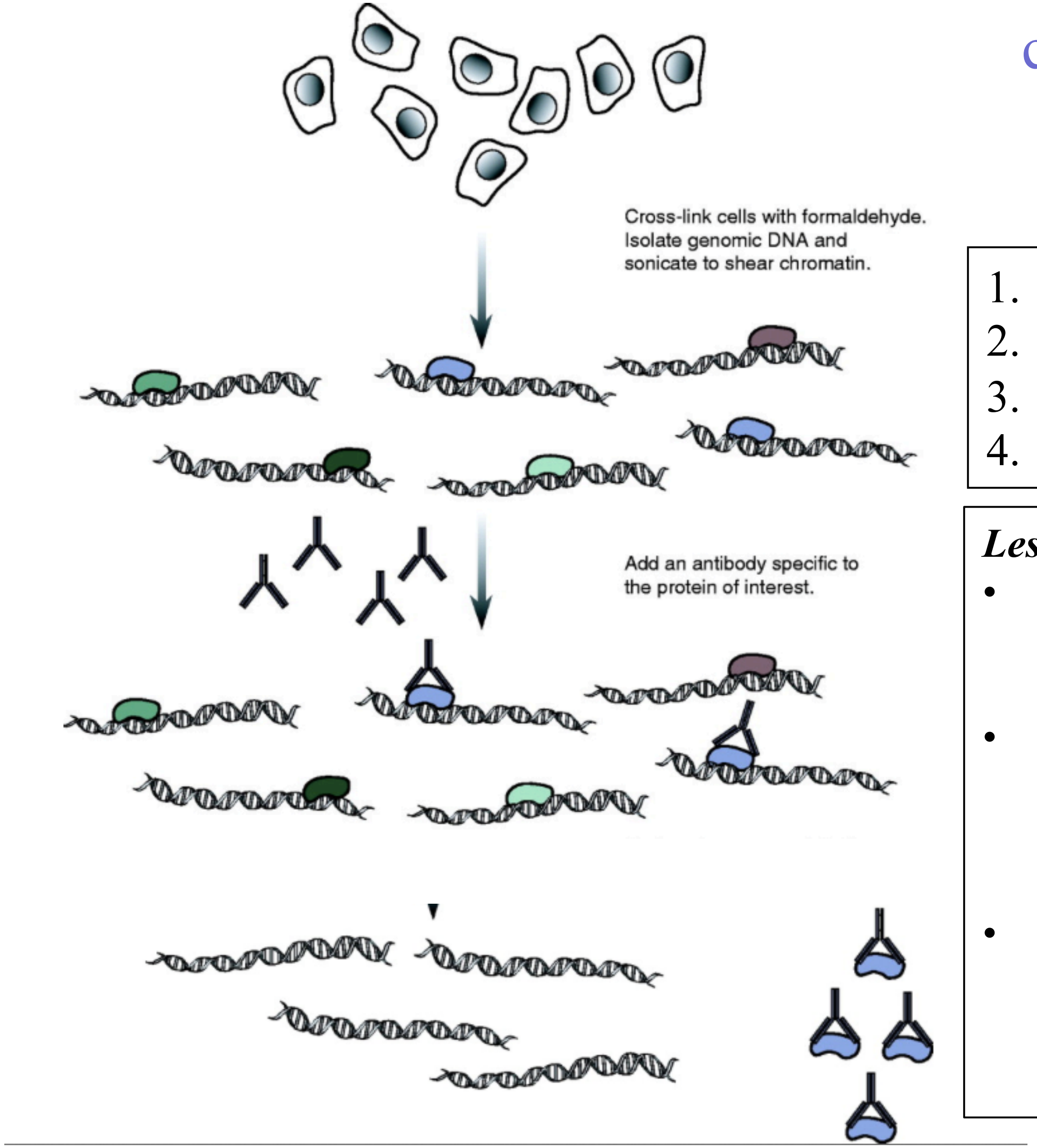
1. Scan genome for matches to known matrix/consensus
   problem is that there are many nonfunctional in the genome - poor predictor of function

2. Phylogenetic footprinting: overly-conserved sequences in multiple alignments
   Variation within element is typically lower than surrounding 'nonfunctional' DNA

3. Network/module approach: Focus on groups of co-regulated genes to increase statistical power
   Look for statistically significant enrichment of sequences in
   the group of upstream regions from a group of co-regulated genes

Experimental:

4. Chromatin immunoprecipitation (ChIP-chip or ChIP-seq) to identify binding loci genomewide
   can do ChIP analysis across species or in one species then compare computationally


Cross-link cells with formaldehyde. Isolate genomic DNA and sonicate to shear chromatin.
Add an antibody specific to the protein of interest.

Chromatin-immunoprecipitation coupled to deep sequencing:

ChIP-Seq:

1. Add crosslinker to cells
2. Lyse & shear DNA
3. IP protein of interest with antibody
4. Process recovered DNA & sequence

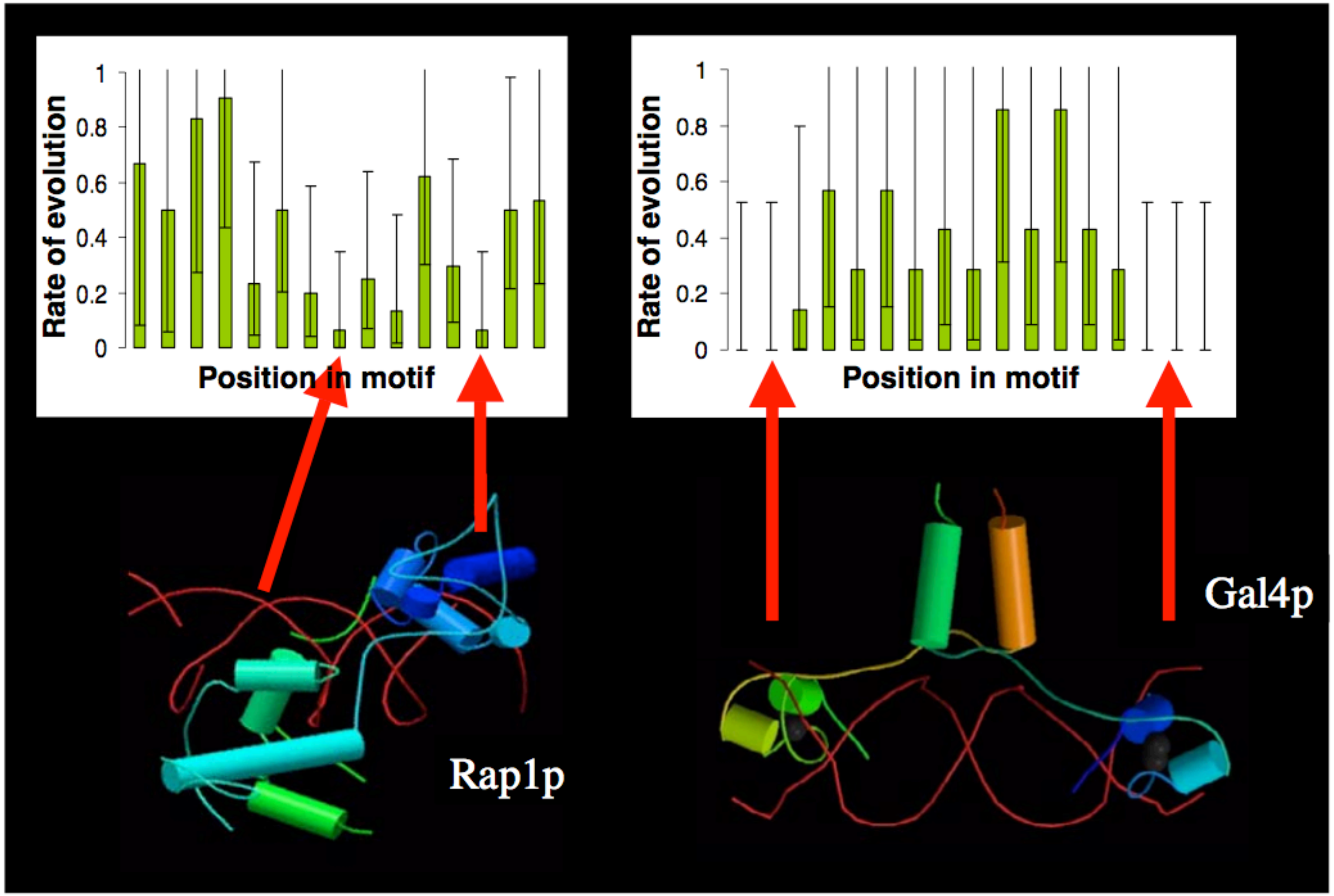
Lessons from ChIP

- Best/most DNA recovery usually means highest TF-DNA affinity
- Often TFs bind DNA despite no recognizable 'binding site' in the region (note ChIP identifies a region bound, not a site)
- Many "low-occupancy" (e.g. weakly recovered) sites may be real binding that is non-functional

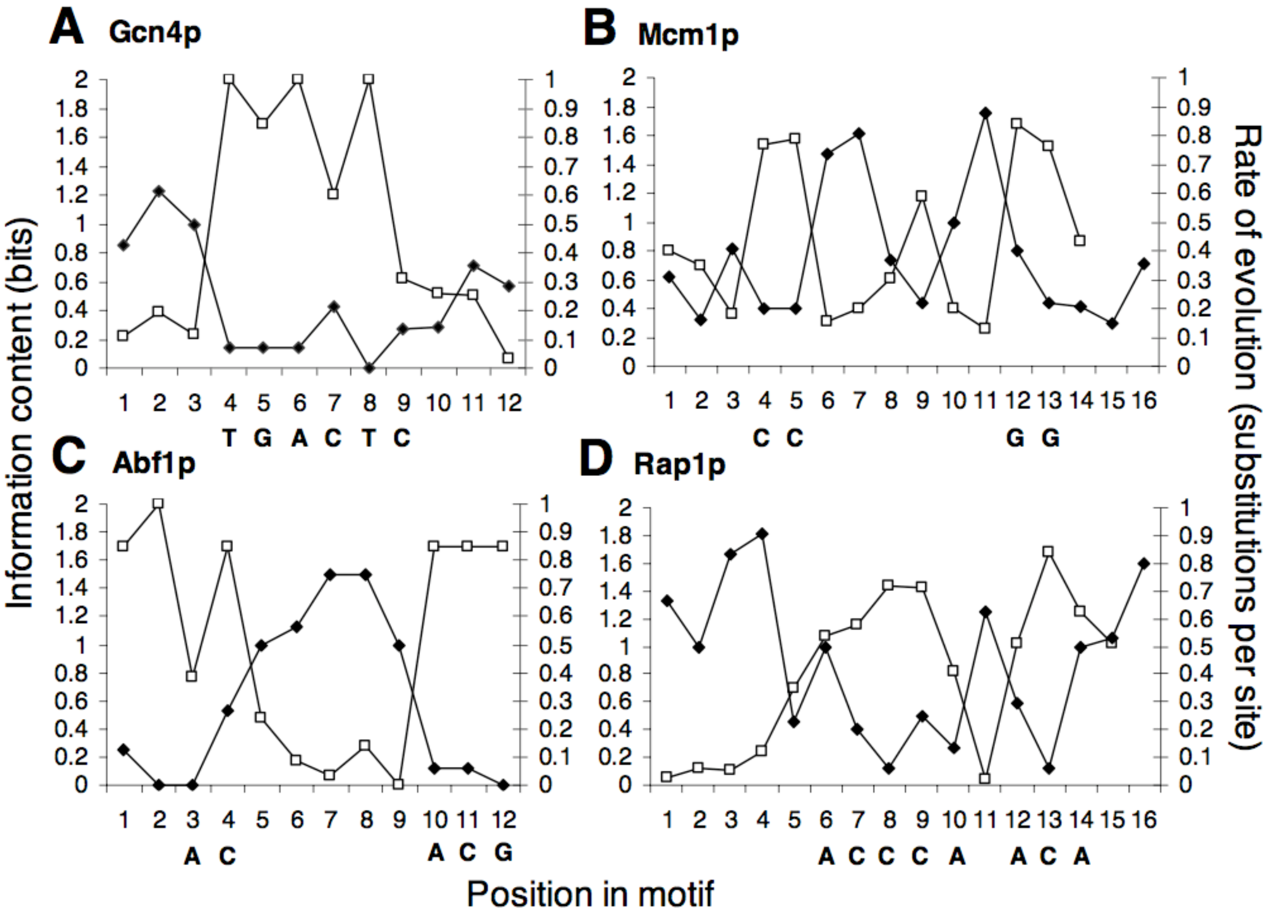
What kinds of constraints act on TF binding sites?

1. Productive contact between protein-DNA (constraint on *sequence* of binding site)

Sites of contact evolve slower (under more constraint)



Variation within a site across species parallels variation across sites within a genome



Open symbols: Information content Closed symbols: Substitutions per site

What kinds of constraints act on TF binding sites?

1. Productive contact between protein-DNA (constraint on *sequence* of binding site)

2. Distance from transcription start site (constraint on *position* of the binding site)
   also may be restricted by placement of nucleosome-depleted regions

What kinds of constraints act on TF binding sites?

1. Productive contact between protein-DNA (constraint on *sequence* of binding site)

2. Distance from transcription start site (constraint on *position* of the binding site)
   also may be restricted by placement of nucleosome-depleted regions

3. Spacing between elements if cooperative TF interactions (constraint of *position)*

What kinds of constraints act on TF binding sites?

1. Productive contact between protein-DNA (constraint on *sequence* of binding site)

2. Distance from transcription start site (constraint on *position* of the binding site)
   also may be restricted by placement of nucleosome-depleted regions

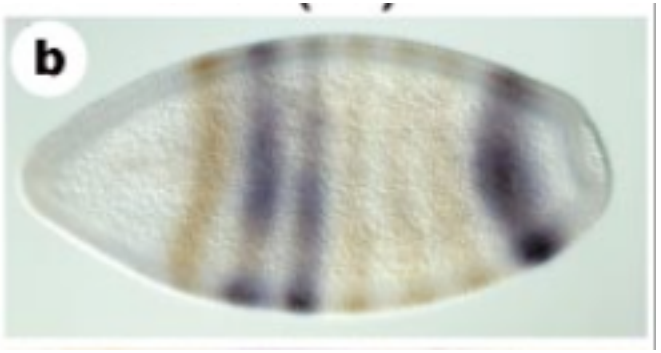
3. Spacing between elements if cooperative TF interactions (constraint of *position)*

How do Regulatory Regions evolve?

1. Conserved regulation but evolution of regulatory regions (stabilizing selection)

   - Binding-site turnover: non-conserved sites but conserved regulation
     Seems to be very prevalent across many organisms

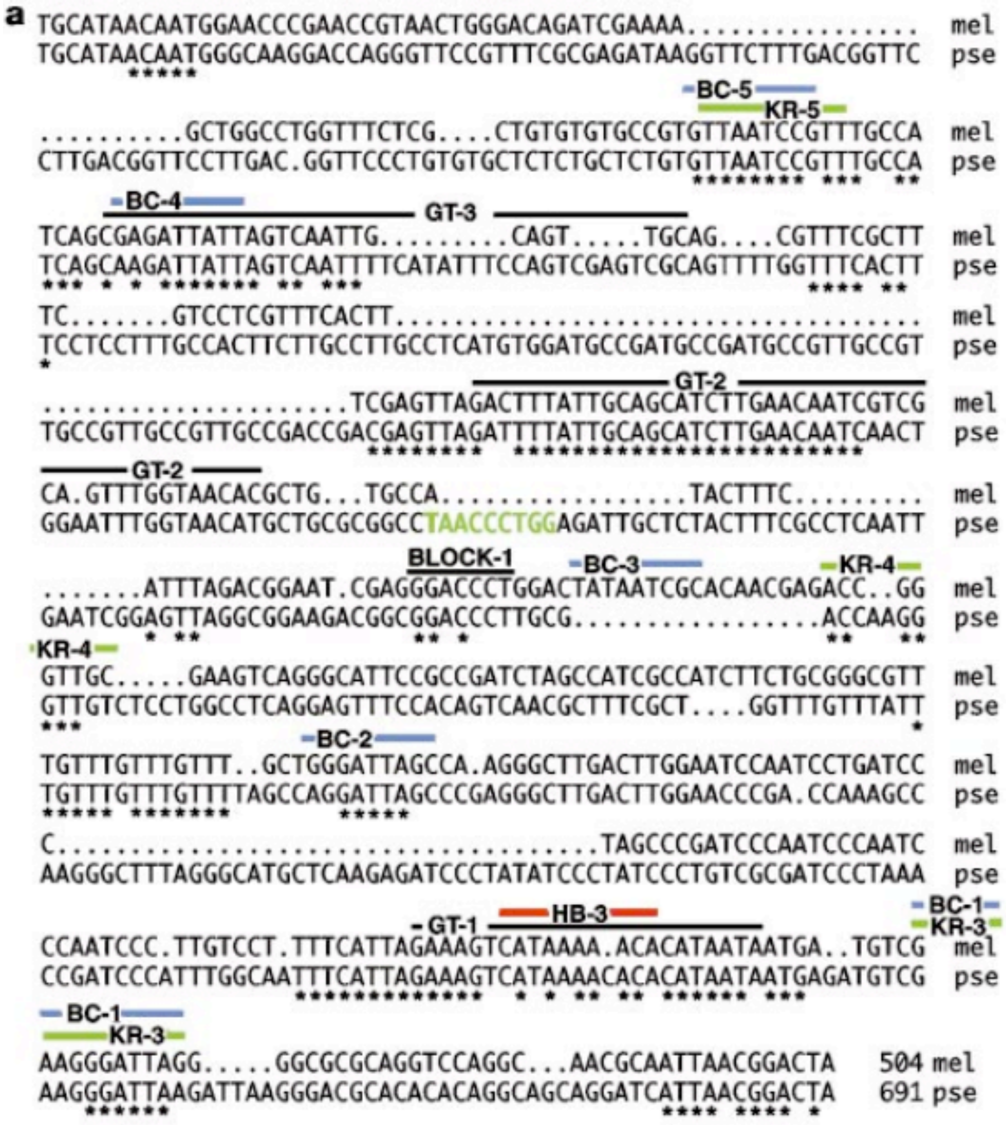
Ludwig *et al.* Nature. 2000



Four TFs act combinatorially
To determine Eve2 patterns

Eve stripe 2 expression
highly conserved across
species.

None of 16 binding sites in
stripe 2 enhancers
is perfectly conserved
across 13 species

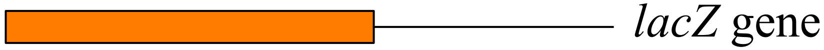


```
a TGCATAACAATGGAACCCGAACCGTAACTGGGACAGATCGAAAA................  mel
  TGCATAACAATGGGCAAGGACCAGGGTTCCGTTTCGCGAGATAAGGTTCTTTGACGGTTC  pse
       *****
                                                   -BC-5-
                                                    KR-5
  ..........GCTGGCCTGGTTTCTCG....CTGTGTGTGCCGTGTTAATCCGTTTGCCA  mel
  CTTGACGGTTCCTTGAC.GGTTCCCTGTGTGCTCTCTGCTCTGTGTTAATCCGTTTGCCA  pse
                                             ********* ***  **
     -BC-4-
   ------------------ GT-3 --------------
  TCAGCGAGATTATTAGTCAATTG.........CAGT.....TGCAG....CGTTTCGCTT  mel
  TCAGCAAGATTATTAGTCAATTTTCATATTTCCAGTCGAGTCGCAGTTTTGGTTTCACTT  pse
  *** * * ******* ** ***                               **** **
  TC.......GTCCTCGTTTCACTT....................................  mel
  TCCTCCTTTGCCACTTCTTGCCTTGCCTCATGTGGATGCCGATGCCGATGCCGTTGCCGT  pse
  *
                                  ------------ GT-2 -----------
  .........................TCGAGTTAGACTTTATTGCAGCATCTTGAACAATCGTCG  mel
  TGCCGTTGCCGTTGCCGACCGACGAGTTAGATTTTATTGCAGCATCTTGAACAATCAACT  pse
                          ********  ************************
  ----- GT-2 -----
  CA.GTTTGGTAACACGCTG...TGCCA.................TACTTTC.........  mel
  GGAATTTGGTAACATGCTGCGCGGCCTAACCCTGGAGATTGCTCTACTTTCGCCTCAATT  pse
                         BLOCK-1     -BC-3-             -KR-4-
  .......ATTTAGACGGAAT.CGAGGGACCCTGGACTATAATCGCACAACGAGACC..GG  mel
  GAATCGGAGTTAGGCGGAAGACGGCGGACCCTTGCG..................ACCAAGG  pse
          *  **              ** *                     **    **
  KR-4-
  GTTGC.....GAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTCTGCGGGCGTT  mel
  GTTGTCTCCTGGCCTCAGGAGTTTCCACAGTCAACGCTTTCGCT....GGTTTGTTTATT  pse
  ***                                                       *
                   -BC-2-
  TGTTTGTTTGTTT..GCTGGGATTAGCCA.AGGGCTTGACTTGGAATCCAATCCTGATCC  mel
  TGTTTGTTTGTTTTAGCCAGGATTAGCCCGAGGGCTTGACTTGGAACCCGA.CCAAAGCC  pse
  ***** ********       *****
  C.......................................TAGCCCGATCCCAATCCCAATC  mel
  AAGGGCTTTAGGGCATGCTCAAGAGATCCCTATATCCCTATCCCTGTCGCGATCCCTAAA  pse
                               -HB-3-                       -BC-1-
                      - GT-1 ----------------               -KR-3-
  CCAATCCC.TTGTCCT.TTTCATTAGAAAGTCATAAAA.ACACATAATAATGA..TGTCG  mel
  CCGATCCCATTTGGCAATTTCATTAGAAAGTCATAAAACACACATAATAATGAGATGTCG  pse
                 **************  * * ** ** ****** ***
  -BC-1-
  -KR-3-
  AAGGGATTAGG.....GGCGCGCAGGTCCAGGC...AACGCAATTAACGGACTA   504 mel
  AAGGGATTAAGATTAAGGGACGCACACACAGGCAGCAGGATCATTAACGGACTA   691 pse
     ******                                **** **** *
```

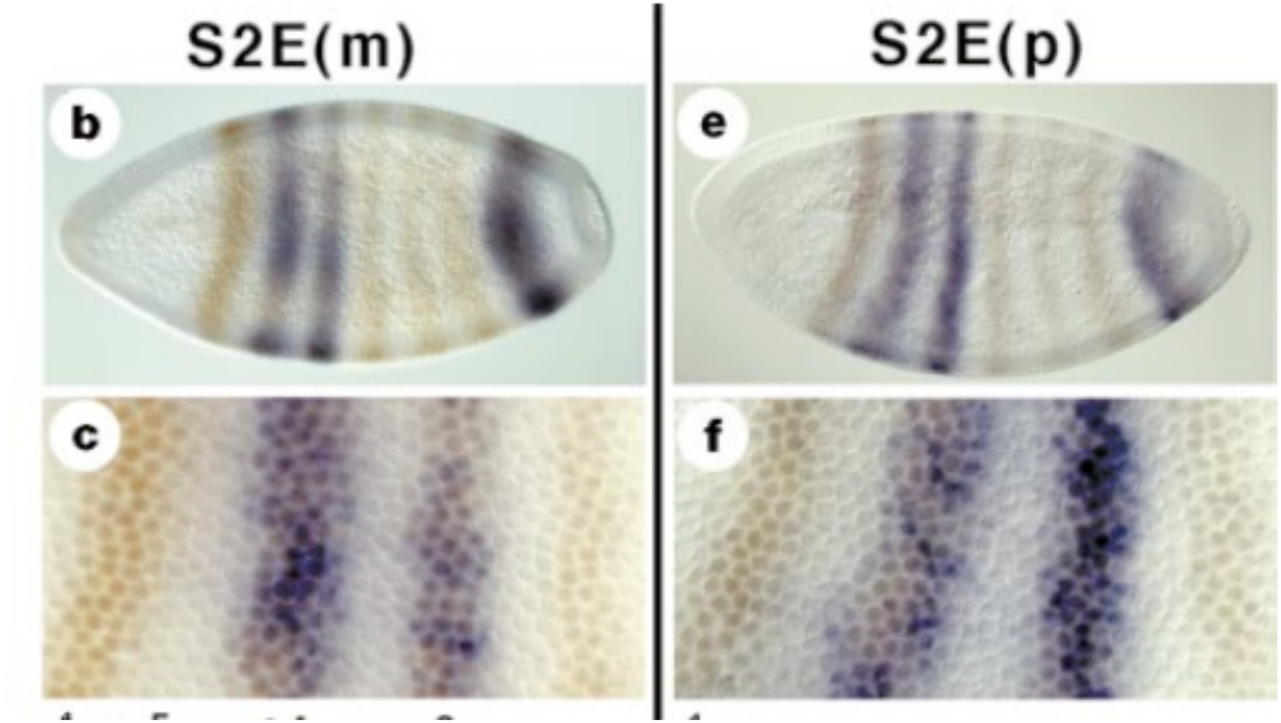
Ludwig *et al.* Nature. 2000

Evidence for stabilizing selection in a eukaryotic enhancer element.

Native *D. pseudoobscura* enhancer works well in *D. melanogaster*



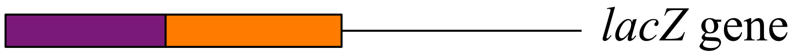
lacZ gene



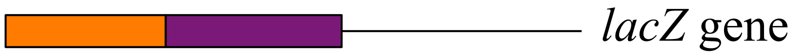
Ludwig *et al.* Nature. 2000

Evidence for stabilizing selection in a eukaryotic enhancer element.

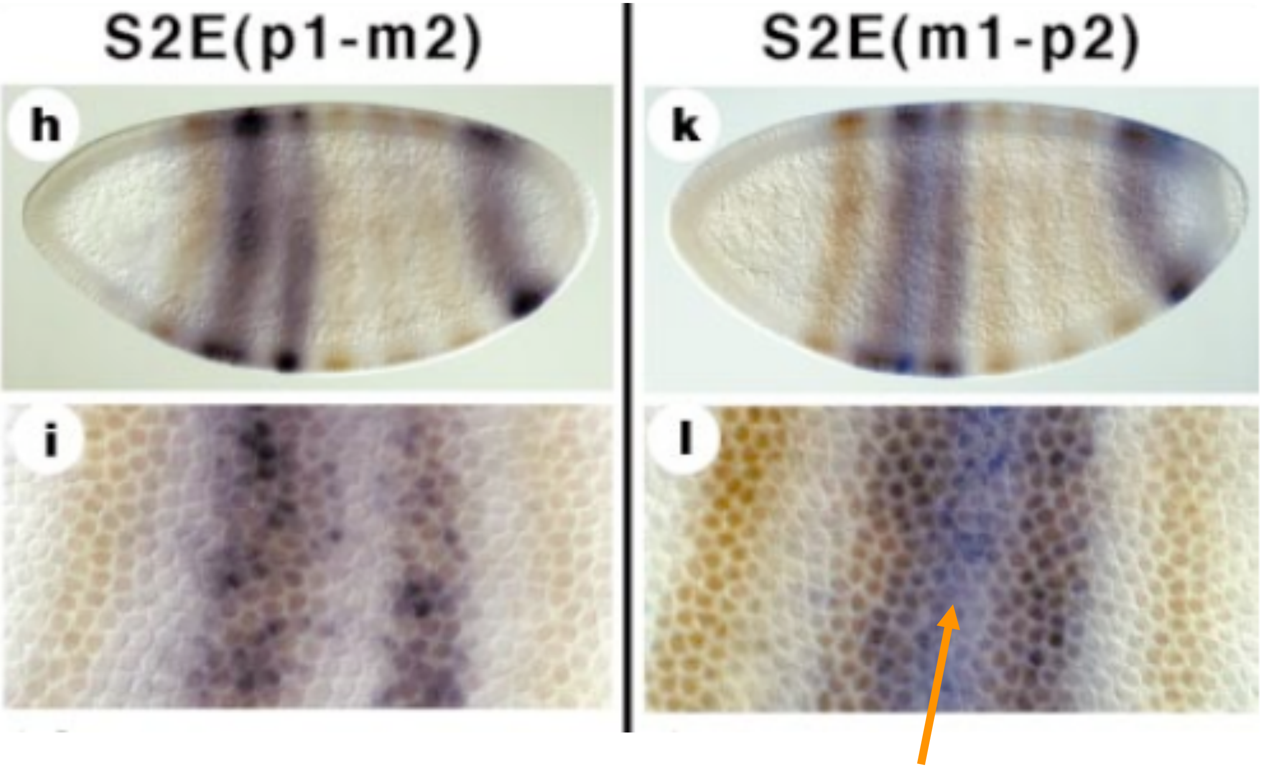
But hybrid enhancers (mel-pseudo or pseudo-mel from 5’ to 3’) are defective



lacZ gene

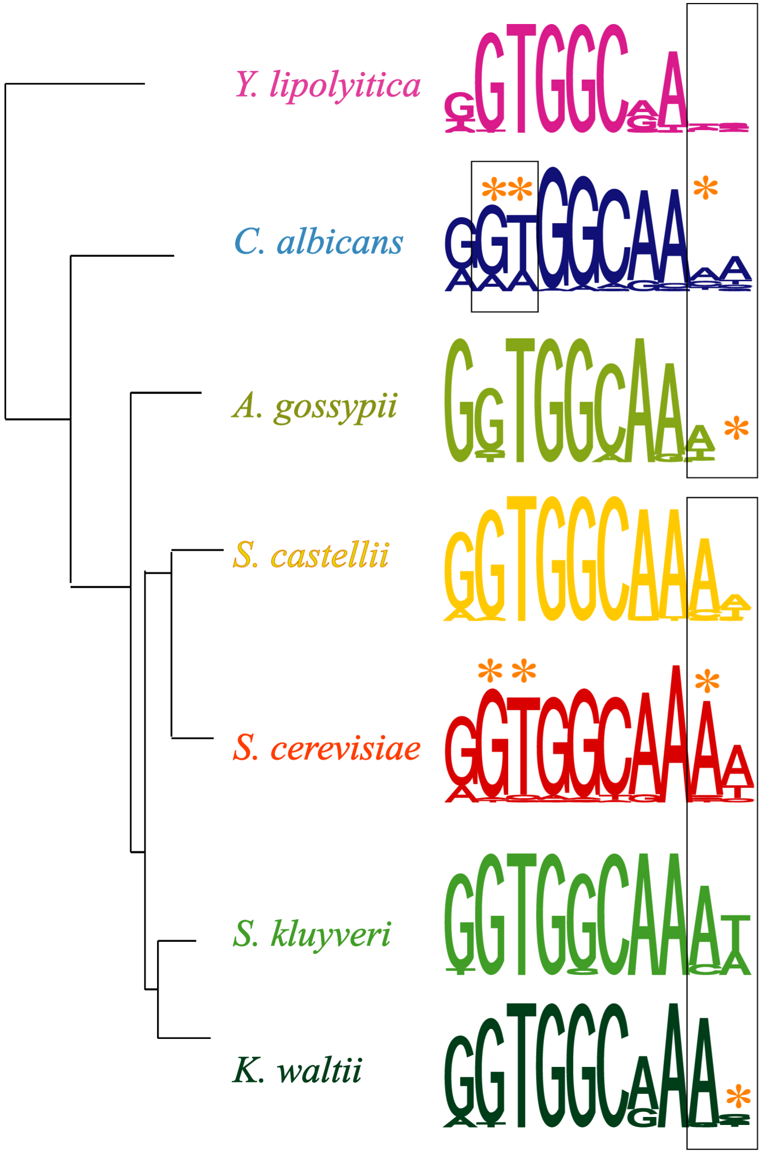


lacZ gene



They argue for stabilizing selection and binding-site turnover across the enhancer

Co-evolution of Rpn4 sites upstream proteosome genes & Rpn4 binding specificity



	# Genes	# Elements
Y. lipolyitica	25	46
C. albicans	26	39
A. gossypii	18	26
S. castellii	27	31
S. cerevisiae	30	32
S. kluyveri	15	18
K. waltii	24	27

What kinds of constraints act on TF binding sites?

1. Productive contact between protein-DNA (constraint on *sequence* of binding site)

2. Distance from transcription start site (constraint on *position* of the binding site)

   also may be restricted by placement of nucleosome-depleted regions

3. Spacing between elements if cooperative TF interactions (constraint of *position)*

How do Regulatory Regions evolve?

1. Conserved regulation but evolution of regulatory regions (stabilizing selection)

   - Binding-site turnover: non-conserved sites but conserved regulation

     Seems to be very prevalent across many organisms

   - Co-evolution between binding site and TF specificity