Regulatory Sequence Analysis

dr. ir. Stein Aerts
Laboratory of Neurogenetics
Dept. Human Genetics

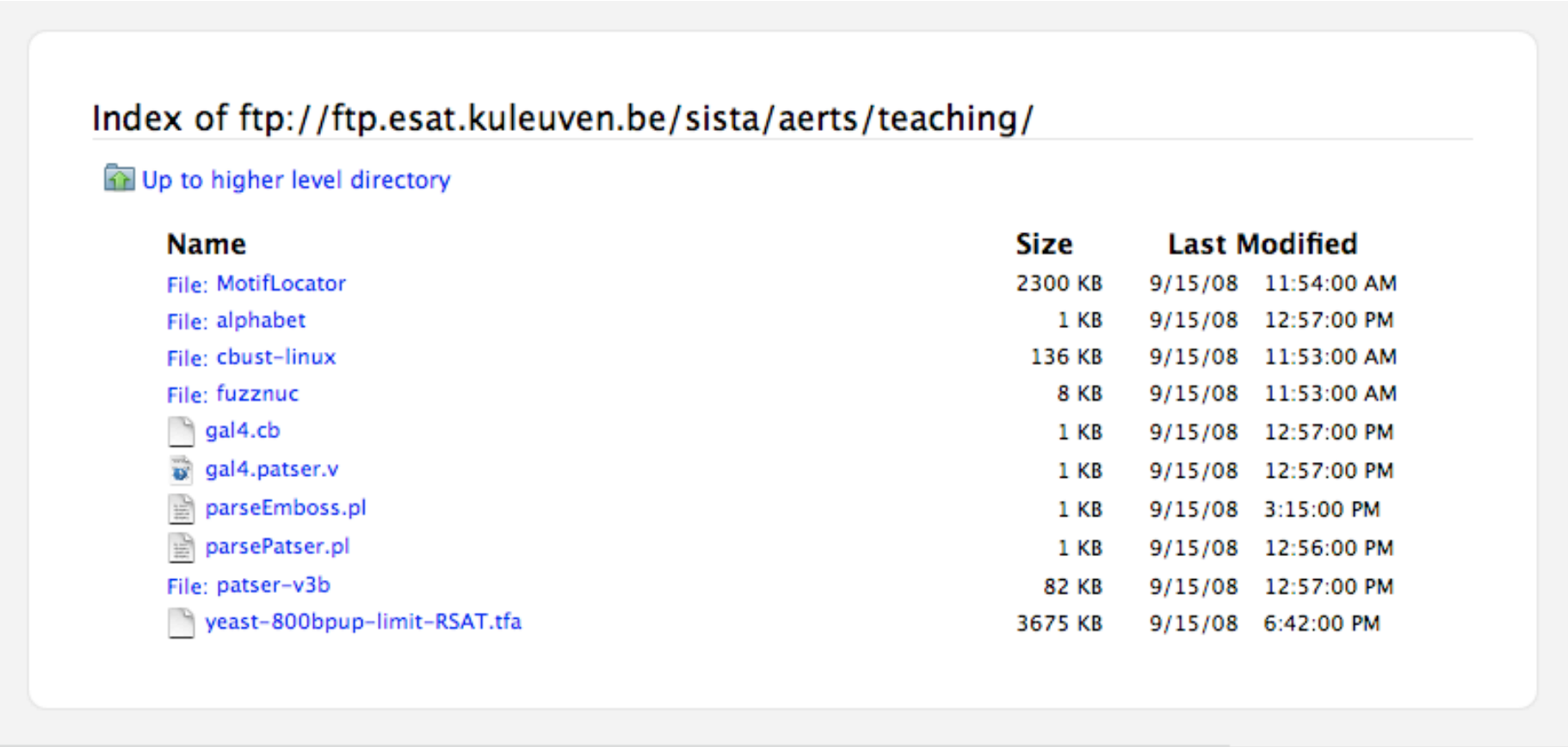


Readings

- Genes & Signals
  - Ptashne & Gann, CSHL press 2002
- Reviews
  - W.W. Wasserman et al Nature Reviews Genetics 2004
  - M. Bulyk Genome Biology 2003
  - J. Nardone et al. Nature Immunology 2004
  - GuhaThakurta et al NAR 2006
- Tools
  - RSAT: Thomas-Chollier et al NAR 2008
  - TOUCAN: Aerts et al NAR 2003, 2005

Material

- TOLEDO
  - Presentation slides
- FTP server: data files



Index of ftp://ftp.esat.kuleuven.be/sista/aerts/teaching/

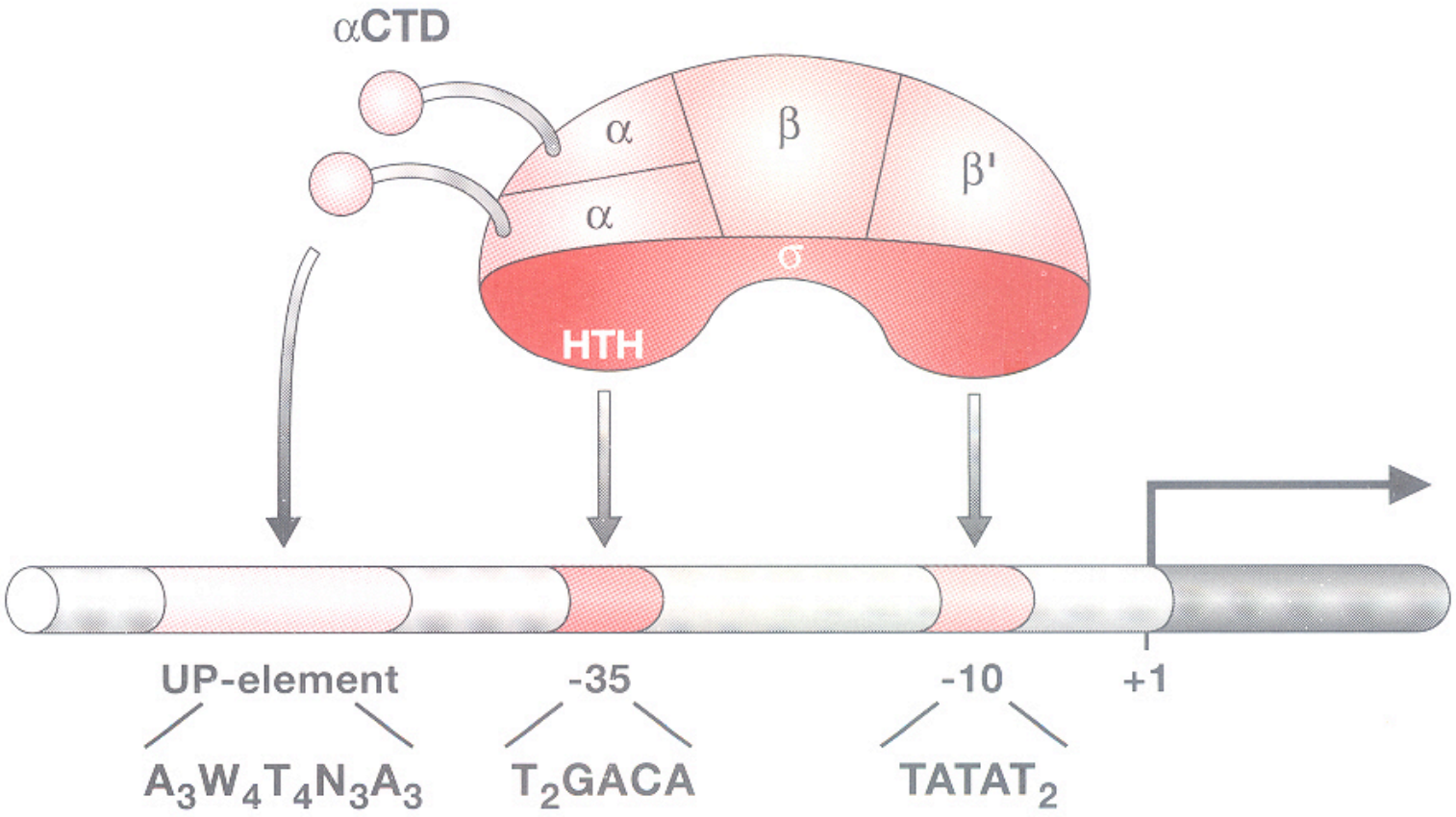
Up to higher level directory

Name	Size	Last Modified
File: MotifLocator	2300 KB	9/15/08 11:54:00 AM
File: alphabet	1 KB	9/15/08 12:57:00 PM
File: cbust-linux	136 KB	9/15/08 11:53:00 AM
File: fuzznuc	8 KB	9/15/08 11:53:00 AM
gal4.cb	1 KB	9/15/08 12:57:00 PM
gal4.patser.v	1 KB	9/15/08 12:57:00 PM
parseEmboss.pl	1 KB	9/15/08 3:15:00 PM
parsePatser.pl	1 KB	9/15/08 12:56:00 PM
File: patser-v3b	82 KB	9/15/08 12:57:00 PM
yeast-800bpup-limit-RSAT.tfa	3675 KB	9/15/08 6:42:00 PM

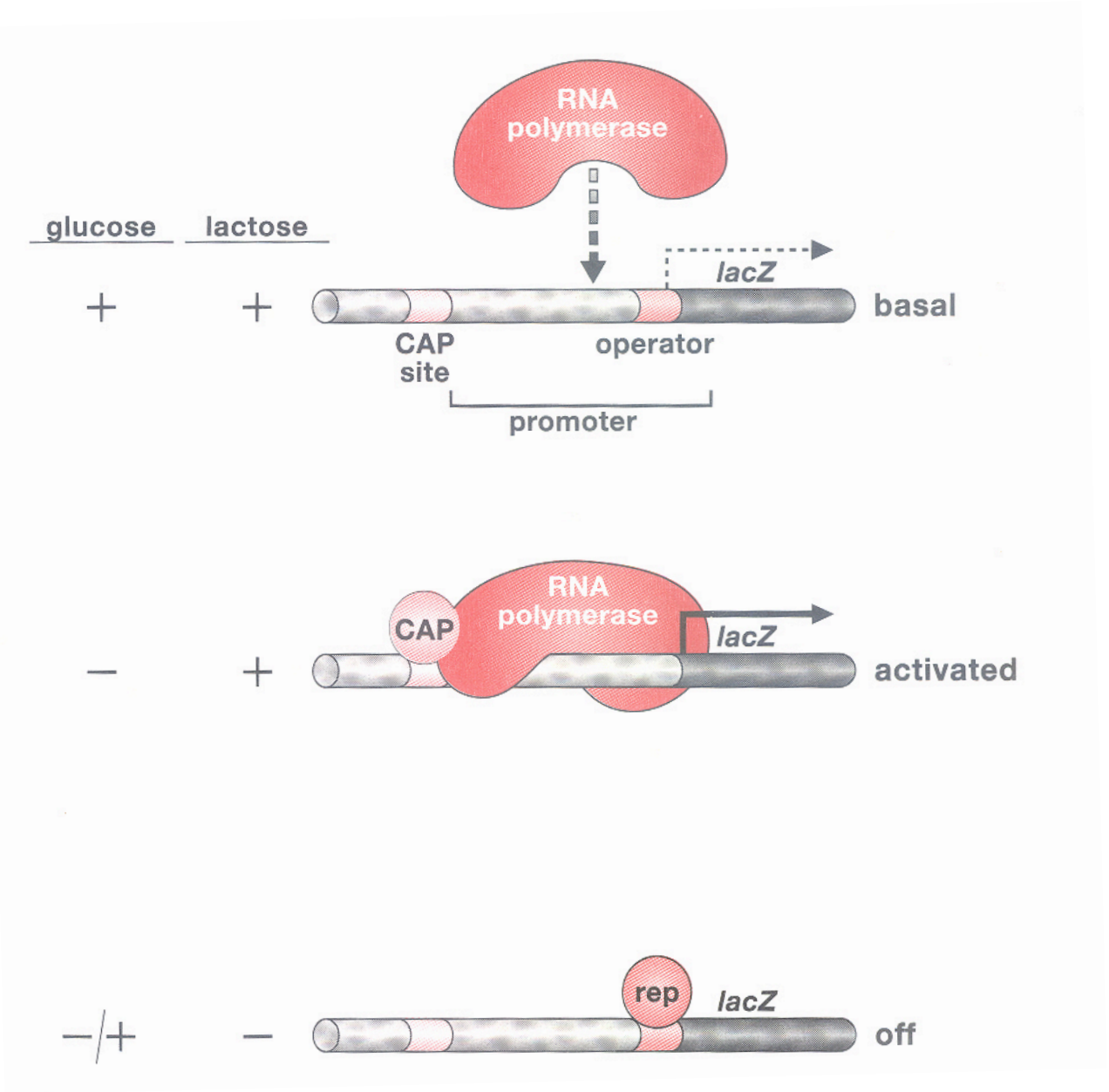
Agenda

- Transcriptional regulation
  - Bacteria
  - Yeast
  - Higher eukaryotes
- Regulatory sequence analysis
  - Motif representation, motif detection, phylogenetic footprinting, module detection
  - *De novo* motif & module discovery by phylogenetic footprinting and by gene co-regulation

Prokaryotic promoter



The *lac* genes



Prokaryotic promoter

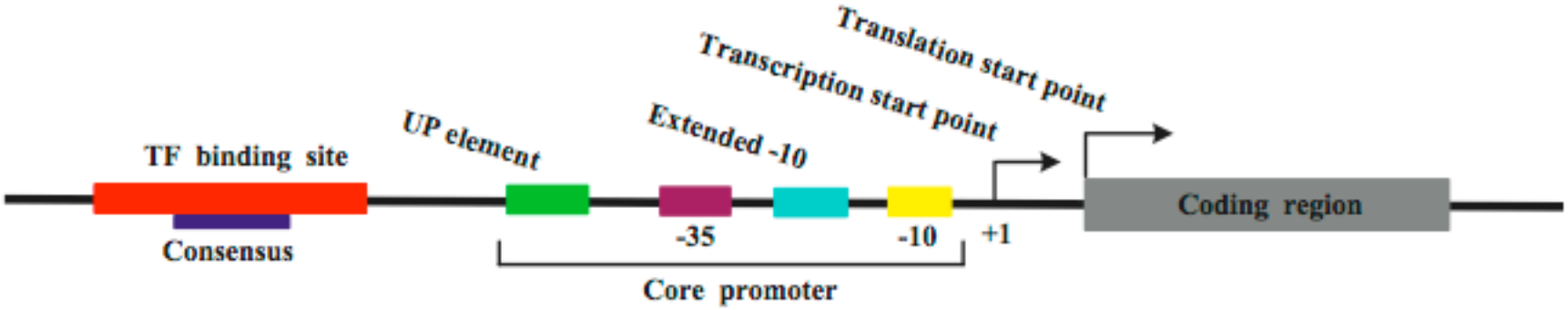


Figure 1. Structure of a prokaryotic promoter. A promoter is a region of DNA on the genome where RNA polymerase and TF bind to initiate transcription. The +1 indicates the base pair where transcription initiates, and it is commonly called transcription start point. Base pairs upstream of the transcription start point are assigned positive numbers, while those downstream are shown with negative numbers. The core promoter consists of –10, –35 and extended –10 and UP elements. The –10, –35 and extended –10 elements are recognized by domains 2, 4 and 3 of the RNA polymerase σ subunit, respectively [1, 246]. The UP element, located upstream of the –35 element, is recognized by the C-terminal domains of the RNA polymerase α subunits [247]. Sometimes, the TF binding site may overlap the core promoter sequence. A consensus sequence is often located in the TF-binding site.

What is known in Ecoli - RegulonDB

RegulonDB
Escherichia coli K12 Transcriptional Network

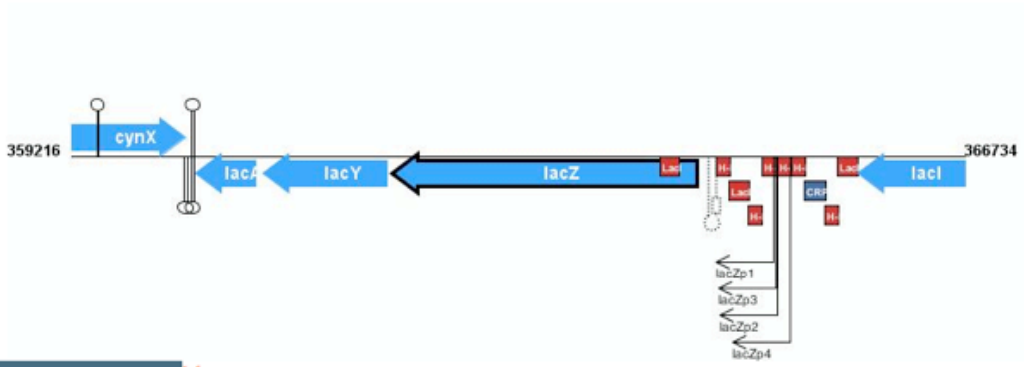
Main Page | Using RegulonDB | Tools | Downloads | About RegulonDB

Search: Gene Go

Genome Browser | Nebulon NetView | NetWork Tool | Export to: XML

Gene *lacZ* in *Escherichia coli K12*

Object color & Code



GENE : Info

Name:	*lacZ*
Synonym(s):	*ECK0341,b0344*
Map position(nucleotides):	362455 <— 365529 Genome Browser
Strand:	reverse
Sequence:	Get nucleotide sequence
GC content %:	56.26
Note(s):	The *lacZ* gene is one of the structural genes of the *lac* operon in *E. coli*. Transcription of this operon is both negatively and positively regulated. Negative regulation occurs when the *lac* repressor (*lacI*) binds to the operator, preventing transcription by RNA polymerase. Positive regulation is mediated by the cyclic AMP receptor protein, CRP (also referred to as the catabolite activator protein, CAP). When cyclic AMP binds to CRP, CRP becomes activated, binding to DNA near the *lac* promoter and facilitating transcription. Synthesis of *lac* operon proteins is induced when *E. coli* is grown on lactose in the absence of glucose. Induction occurs when the physiological inducer allolactose binds to the *lac* repressor, preventing it from binding to the operator. Non-physiological analogs, such as thiogalactosides, can also function as inducers. If *E. coli* is grown in the presence of glucose, *lac* gene expression is not needed and a control mechanism known as catabolite repression occurs. This refers to the ability of glucose to inhibit expression of the operon. As cellular cyclic AMP levels drop in response to uptake of glucose, CRP is not activated and operon expression is inhibited. By this mechanism, *lac* genes are only expressed if lactose is present and glucose is absent. The *lac* operon has been extensively studied both genetically and biochemically (reviewed by several authors in Miller JH,1978.
External database links:	
ECOCYC:	EG10527
GENBANK:	1786539
GENBANK_RefSeq:	16128329
OU-MICROARRAY:	b0344
GENE CONTEXT TOOL	

Yeast

Yeast

- DNA wrapped around histones: nucleosomes, inside nucleus
- expression of eukaryotic gene is more complicated than expression of bacterial gene (splicing, modifications, transported out of nucleus)
  - but we focus on transcription initiation
- also complicated by histones and nuclei
  - histone modifications (can affect protein binding to DNA)
- regulators move from cytoplasm to nucleus

Model case: the GAL genes

review

EMBO reports **7**, 5, 496–499 (2006)
doi:10.1038/sj.embor.7400679

Yeast Gal4: a transcriptional paradigm revisited

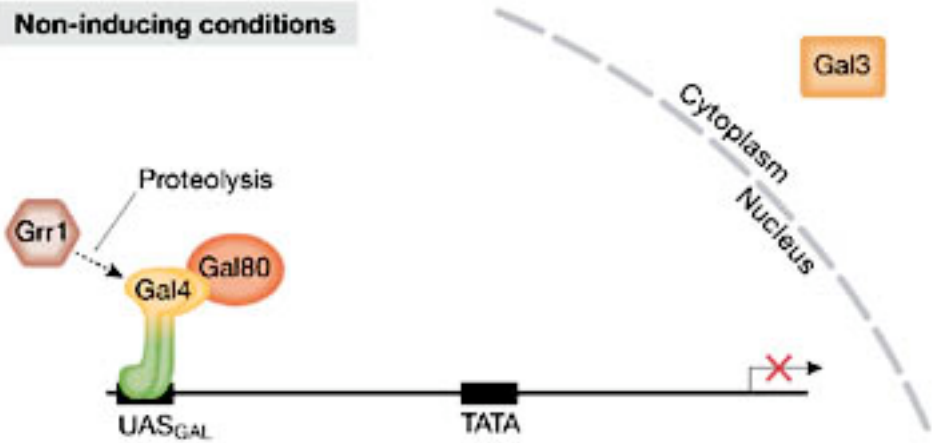
Ana Traven[1], Branka Jelicic[2] & Mary Sopta[2]

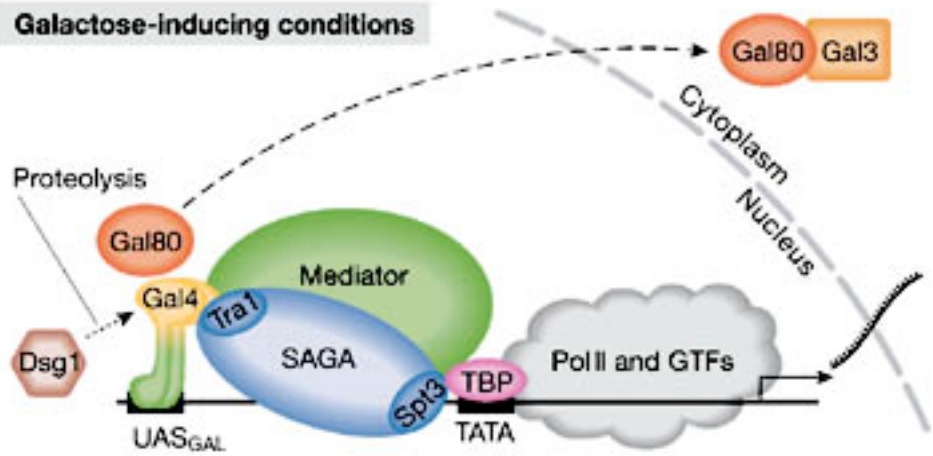
[1] St Vincent's Institute, 9 Princes Street, Fitzroy, Vic 3065, Austral
[2] Department of Molecular Biology, Rudjer Boskovic Institute, Bije

To whom correspondence should be addressed
Mary Sopta Tel:+38 5 1456 0948; Fax: +38 5 1456 1177;
msopta@irb.hr

Received 21 November 2005; Accepted 6 February 2006.







SCPD http://rulai.cshl.edu/SCPD/

Get mapped sites | Get putative sites | Get intergenic region

Retrieve sequence Start<-ATG -500 ATG->End 50

```
       ORF        YML051W
      GENE        GAL80
        XX
    FACTOR        GAL4
COORDINATE        (-172, -156)
  SEQUENCE        CGGCGCACTCTCGCCCG
 REFERENCE        EMBO J. 5:603-608 (1986)

        XX
                >YML051W         GAL80
                                          GAL4
                AATAATATCCTATATTTTCTTCATTTACCGGCGCACTCTCGCCCGAACGA      -151

                CCTCAAAATGTCTGCTACATTCATAATAACCAAAAGCTCATAACTTTTTT      -101

                TTTTGAACCTGAATATATATACATCACATATCACTGCTGGTCCTTGCCGA      -51

                CCAGCGTATACAATCTCGATAGTTGGTTTCCCGTTCTTTCCACTCCCGTC      -1

                ATGGACTACAACAAGAGATCTTCGGTCTCAACCGTGCCTAATGCAGCTCC      +50

//
```

GAL4

Get regulated genes | Get sites | Get consensus | Get matrix | Get affinity data

Get genomewise distribution | Sort by copy No | Sort by function category

```
>YBR018C        AGCGCTCGGACAACTGTTGACC
>YBR018C        ATACTTCGGAGCACTGTTGAGCG
>YBR019C        CGGCGGCTTCTAATCCG
>YBR019C        TCGGAGGGCTGTCGCCCG
>YBR019C        CGGAGGAGAGTCTTCCG
>YBR019C        ATTGTTCGGAGCAGTGCGGCGCG
>YBR020W        CGCGCCGCACTGCTCCGAACAAT
>YBR020W        CGGAAGACTCTCCTCCG
>YBR020W        CGGGCGACAGCCCTCCGA
>YBR020W        CGGATTAGAAGCCGCCG
>YLR081W        TATCGGGGCGGATCACTCCGAAC
>YLR081W        CACCGGCGGTCTTTCGTCCGTGC
>YML051W        CGGCGCACTCTCGCCCG
>YOR120W        TCGGGGCAGACTATTCCGG
```

```
>YBR018C
CGGACAACTGTTGACCG
>YBR018C
CGGAGCACTGTTGAGCG
>YBR019C
CGGCGGCTTCTAATCCG
>YBR019C
CGGAGGGCTGTCGCCCG
>YBR019C
CGGAGGAGAGTCTTCCG
>YBR019C
CGGAGCAGTGCGGCGCG
>YLR081W
CGGGGCGGATCACTCCG
>YLR081W
CGGCGGTCTTTCGTCCG
>YML051W
CGGCGCACTCTCGCCCG
>YOR120W
CGGGGCAGACTATTCCG
```

- Consensus: CGGNNNNNNNNNNNCCG

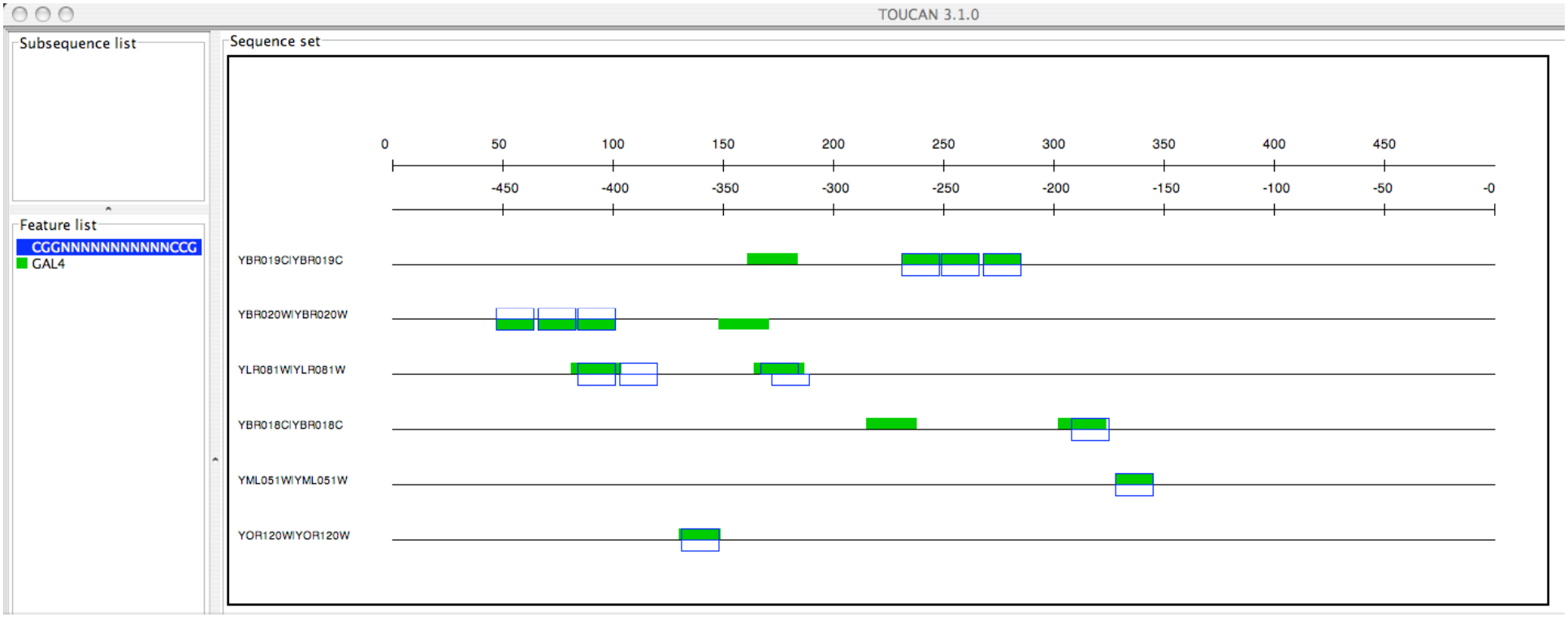
Gal4 targets in TOUCAN

- SGD gene names from S.cer promoter database http://rulai.cshl.edu/SCPD/
- SGD Gene ID (YBR018C, etc.) from BioMart http://www.biomart.org/
- 500bp upstream sequence from BioMart or UCSC Genome Browser (Tables) http://genome.ucsc.edu
- FASTA format

```
> seqId1
AGGTTGCAATTTCTTTTTCTATTAGTAGCTAAAAATGGGTCACGTGATCTATATTCGAAA
GGGGCGGTTGCCTCAGGAAGGCACCGGCGGTCTTTCGTCCGTGCGGAGATATCTGCGCC
> seqId2
CCGAACCGAGATTAGTTAAGCCCTTCCCATCTCAAGATGGGGAGCAAATGGCATTATACTC
CTGCTAGAAAGTTAACTGTGCACATATTCTTAAATTATACAACATTCTGGAGAGCTAT
```

- Load into TOUCAN software
  - Add known sites in GFF format
- Find Gal4 motif by consensus matching

TOUCAN


TOUCAN 3.1.0
Subsequence list
Feature list
CGGNNNNNNNNNNNCCG
GAL4
Sequence set
0
50
100
150
200
250
300
350
400
450
-450
-400
-350
-300
-250
-200
-150
-100
-50
-0
YBR019C|YBR019C
YBR020W|YBR020W
YLR081W|YLR081W
YBR018C|YBR018C
YML051W|YML051W
YOR120W|YOR120W

GFF

GFF format Index ▷

GFF (General Feature Format) lines are based on the GFF standard file format. GFF lines have nine required fields that *must* be tab-separated. If the fields are separated by spaces instead of tabs, the track will not display correctly. For more information on GFF format, refer to http://www.sanger.ac.uk/Software/formats/GFF.

Here is a brief description of the GFF fields:

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
7. **strand** - Valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - If the feature is a coding exon, *frame* should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. **group** - All lines with the same group are linked together into a single item.

Example:
Here's an example of a GFF-based track. Click here for a copy of this example that can be pasted into the browser without editing. NOTE: Paste operations on some operating systems will replace tabs with spaces, which will result in an error when the GFF track is uploaded. You can circumvent this problem by pasting the URL of the above example (http://genome.ucsc.edu/goldenPath/help/regulatory.txt) instead of the text itselfinto the custom annotation track text box.

```
track name=regulatory description="TeleGene(tm) Regulatory Regions"
chr22  TeleGene enhancer  1000000  1001000  500 +  .  touch1
chr22  TeleGene promoter  1010000  1010100  900 +  .  touch1
chr22  TeleGene promoter  1020000  1020000  800 -  .  touch2
```

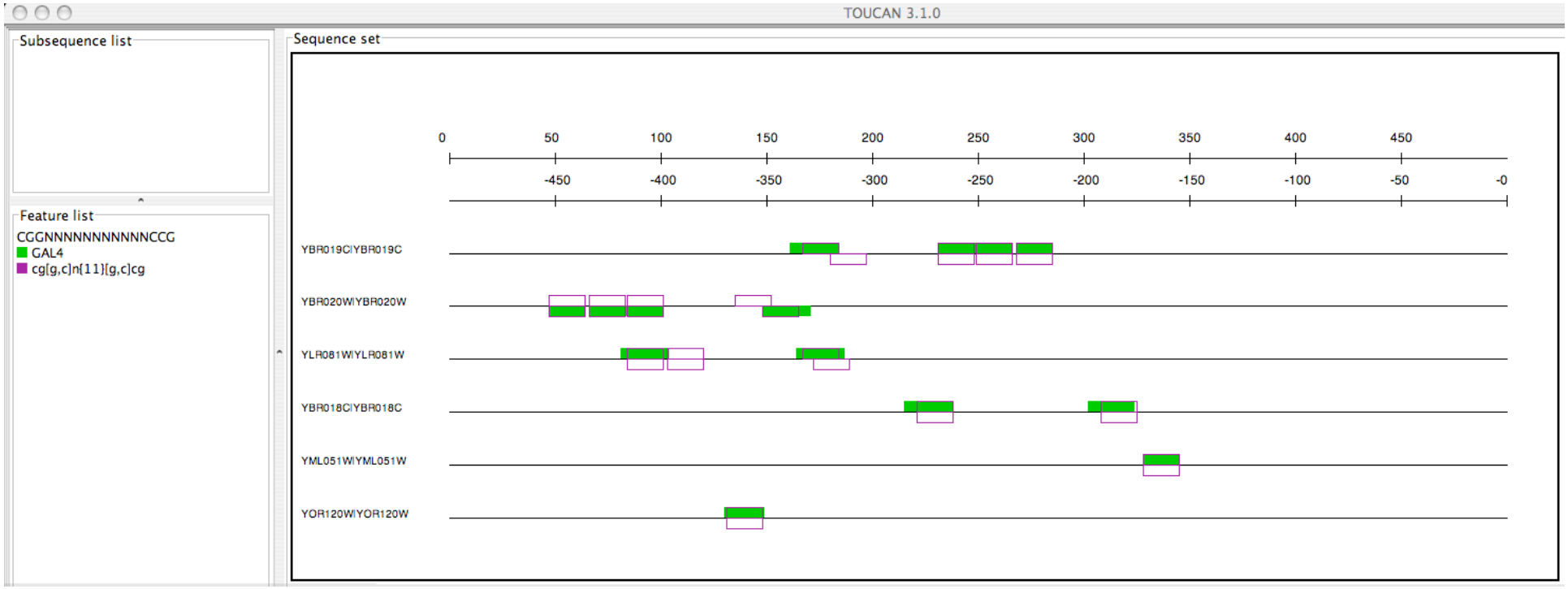
http://genome.ucsc.edu/FAQ/FAQformat#format3

General Feature Format GFF

Tab delimited (\t)

```
YBR019C|YBR019C Manual  misc_feature    269     285     0.0     +       .       type "GAL4";score "0.0";occ "cggcggcttctaatccg";note "Single Motif"
YBR019C|YBR019C Manual  misc_feature    249     266     0.0     +       .       type "GAL4";score "0.0";occ "tcggagggctgtcgcccg";note "Single Motif"
YBR019C|YBR019C Manual  misc_feature    232     248     0.0     +       .       type "GAL4";score "0.0";occ "cggaggagagtcttccg";note "Single Motif"
YBR019C|YBR019C Manual  misc_feature    162     184     0.0     +       .       type "GAL4";score "0.0";occ "attgttcggagcagtgcggcgcg";note "Single Motif"
YBR020W|YBR020W Manual  misc_feature    48      64      0.0     -       .       type "GAL4";score "0.0";occ "cggcggcttctaatccg";note "Single Motif"
YBR020W|YBR020W Manual  misc_feature    67      84      0.0     -       .       type "GAL4";score "0.0";occ "tcggagggctgtcgcccg";note "Single Motif"
YBR020W|YBR020W Manual  misc_feature    85      101     0.0     -       .       type "GAL4";score "0.0";occ "cggaggagagtcttccg";note "Single Motif"
YBR020W|YBR020W Manual  misc_feature    149     171     0.0     -       .       type "GAL4";score "0.0";occ "attgttcggagcagtgcggcgcg";note "Single Motif"
YLR081W|YLR081W Manual  misc_feature    82      104     0.0     +       .       type "GAL4";score "0.0";occ "caccggcggtctttcgtccgtgc";note "Single Motif"
YLR081W|YLR081W Manual  misc_feature    82      104     0.0     +       .       type "GAL4";score "0.0";occ "caccggcggtctttcgtccgtgc";note "Single Motif"
YLR081W|YLR081W Manual  misc_feature    165     187     0.0     +       .       type "GAL4";score "0.0";occ "tatcggggcggatcactccgaac";note "Single Motif"
YBR018C|YBR018C Manual  misc_feature    216     238     0.0     +       .       type "GAL4";score "0.0";occ "atacttcggagcactgttgagcg";note "Single Motif"
YBR018C|YBR018C Manual  misc_feature    303     324     0.0     +       .       type "GAL4";score "0.0";occ "agcgctcggacaactgttgacc";note "Single Motif"
YML051W|YML051W Manual  misc_feature    329     345     0.0     +       .       type "GAL4";score "0.0";occ "cggcgcactctcgcccg";note "Single Motif"
YOR120W|YOR120W Manual  misc_feature    131     149     0.0     +       .       type "GAL4";score "0.0";occ "tcggggcagactattccgg";note "Single Motif"
```

- Use as consensus:
  - “regular expression”: CG[G,C]N{11}[G,C]CG
  - Or in IUPAC code: CGSNNNNNNNNNNSCG



IUPAC code table

Nucleic acid codes

code	description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

- Exercise 1

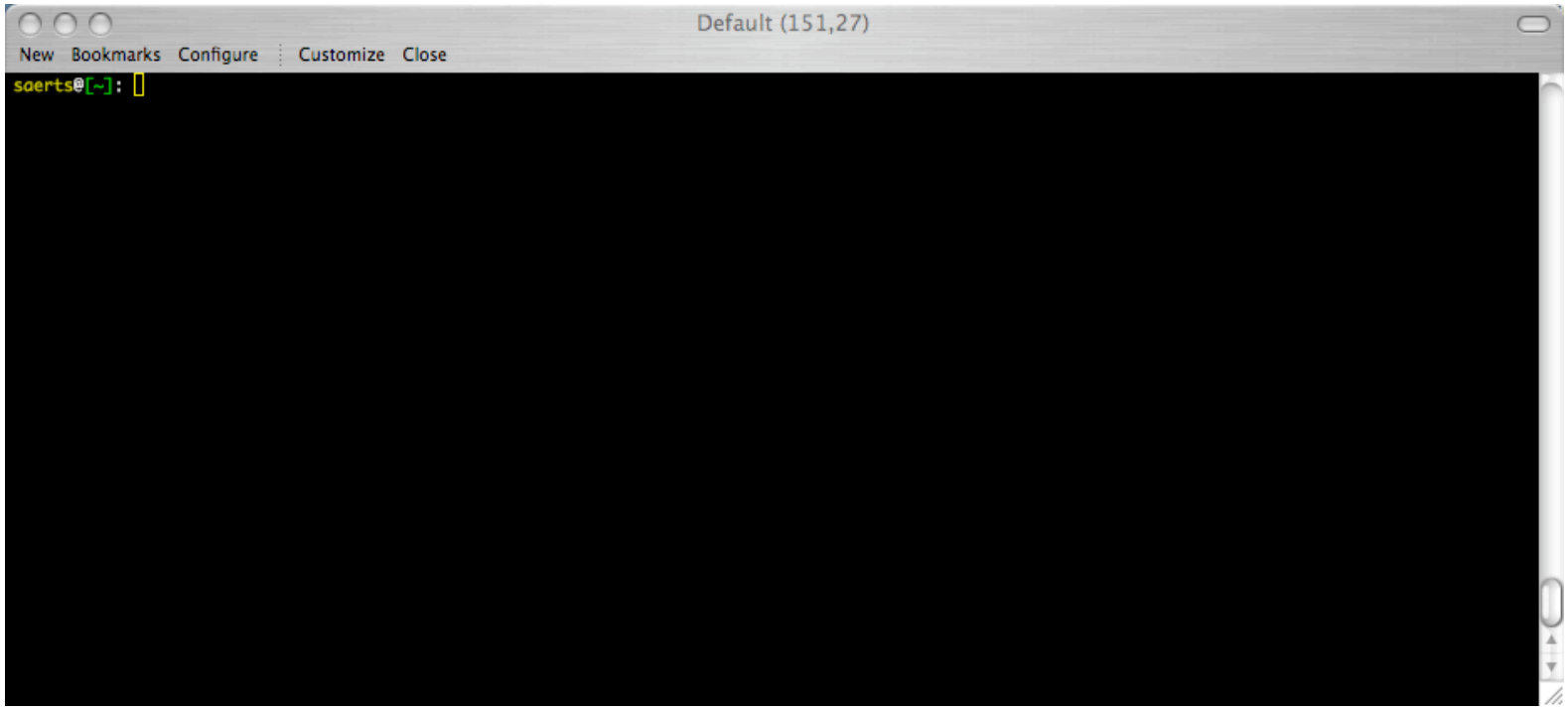
Genome-wide predictions

- RSAT online
- For more flexibility: use programs locally: LINUX command line
  - Scan a whole genome, multiple genomes
  - Perform “batch” jobs, using many motifs or many genomes
  - Integrate different search strategies
  - Etc.

Linux

- Reboot PC, choose **Ubuntu Linux** as Operating System (OS)
- Operating system (OS)
- Unix-like
  - Note: Mac OS X (Darwin) is also unix-like
- Free software, open source development
  - typically all underlying source code can be freely modified, used, and redistributed by anyone
- Linux kernel originally written in 1991 by Linus Torvalds
- Used in servers, playstation 2/3, …
- Supports dozens of programming languages

The linux terminal



Genome-wide predictions

- EMBOSS
  - http://emboss.sourceforge.net/
  - A high-quality package of free, Open Source software for molecular biology
  - Runs on LINUX (but not installed in PC rooms..)
  - Fuzznuc
    - Do “fuzznuc -help”
    - fuzznuc -sequence chr2.fa -pattern CGGNNNNNNNNNNCCG -mismatch 0 -complement 1 -outfile chr2.fuzznuc
- Alternative: make/use a perl script

FastaRegExpGff.pl

```
#!/usr/bin/perl

my $seqFile = $ARGV[0];
my $site=$ARGV[1];
my $sitelen=$ARGV[2];
my $name=$ARGV[3];
my %seqList;
open(FILE,$seqFile);
my $seq=0;
while(<FILE>){
    chomp;
    if($_=~/^>(.*)/){
        if($seq){
            $seqList{$seqId}=uc $seq;
        }
        $seq="";
        $seqId=$1;
    }
    elsif($_=~/[a,c,g,t,A,C,G,T]*/){
        $seq=$seq.$_;
    }
}
$seqList{$seqId}=$seq;
$seqlen=length($seq);
foreach $seqId (keys %seqList){
    $seq=$seqList{$seqId};
#    print ">".$seqId."\n'".$seq."'\n";
    while ($seq =~ /($site)/g) {
        $end=pos $seq;
        $start=$end-$sitelen+1;
        $occ=$1;
        print $seqId."\t"."ExactMatch\tMisc_feature\t".$start."\t".$end."\t0.0\t-\t.\ttype \"".$name."\"; occ=\"".$occ."\"\n";
    }

    my $rc = $seq;
    $rc =~ tr/ACGTacgt/TGCAtgca/;
    $rc = reverse $rc;
#    print ">".$seqId."\n'".$rc."'\n";
    while ($rc =~ /($site)/g) {
        $temp=pos $rc;
        $start=$seqlen-$temp+1;
        $end=$start+$sitelen-1;
        print $seqId."\t"."ExactMatch\tMisc_feature\t".$start."\t".$end."\t0.0\t-\t.\ttype \"".$name."\"; occ=\"".$occ."\"\n";
    }
}
```

perl FastaRegExpGff.pl gal4-500bp-5seq.tfa "CGG\w{11}CCG" 17 gal4

perl FastaRegExpGff.pl gal4-500bp-5seq.tfa "CGG[A,C,G,T]{11}CCG" 17 gal4

On Gal4 target sequences

```
saerts@[dme-srv-03][~/teach/yeast/gal4]: perl ~/perl/FastaRegExpGff.pl  gal4-500bp-5seq.tfa "CGG\w{11}CCG" 17 gal4
YBR018C|YBR018C ExactMatch      Misc_feature    309     325     0.0     +       .       type "gal4"; occ="CGGACAACTGTTGACCG"
YBR018C|YBR018C ExactMatch      Misc_feature    309     325     0.0     -       .       type "gal4"; occ="CGGACAACTGTTGACCG"
YML051W|YML051W ExactMatch      Misc_feature    329     345     0.0     +       .       type "gal4"; occ="CGGCGCACTCTCGCCCG"
YML051W|YML051W ExactMatch      Misc_feature    329     345     0.0     -       .       type "gal4"; occ="CGGCGCACTCTCGCCCG"
YBR019C|YBR019C ExactMatch      Misc_feature    232     248     0.0     +       .       type "gal4"; occ="CGGAGGAGAGTCTTCCG"
YBR019C|YBR019C ExactMatch      Misc_feature    250     266     0.0     +       .       type "gal4"; occ="CGGAGGGCTGTCGCCCG"
YBR019C|YBR019C ExactMatch      Misc_feature    269     285     0.0     +       .       type "gal4"; occ="CGGCGGCTTCTAATCCG"
YBR019C|YBR019C ExactMatch      Misc_feature    269     285     0.0     -       .       type "gal4"; occ="CGGCGGCTTCTAATCCG"
YBR019C|YBR019C ExactMatch      Misc_feature    250     266     0.0     -       .       type "gal4"; occ="CGGCGGCTTCTAATCCG"
YBR019C|YBR019C ExactMatch      Misc_feature    232     248     0.0     -       .       type "gal4"; occ="CGGCGGCTTCTAATCCG"
YOR120W|YOR120W ExactMatch      Misc_feature    132     148     0.0     +       .       type "gal4"; occ="CGGGGCAGACTATTCCG"
YOR120W|YOR120W ExactMatch      Misc_feature    132     148     0.0     -       .       type "gal4"; occ="CGGGGCAGACTATTCCG"
YLR081W|YLR081W ExactMatch      Misc_feature    85      101     0.0     +       .       type "gal4"; occ="CGGCGGTCTTTCGTCCG"
YLR081W|YLR081W ExactMatch      Misc_feature    104     120     0.0     +       .       type "gal4"; occ="CGGAGATATCTGCGCCG"
YLR081W|YLR081W ExactMatch      Misc_feature    168     184     0.0     +       .       type "gal4"; occ="CGGGGCGGATCACTCCG"
YLR081W|YLR081W ExactMatch      Misc_feature    173     189     0.0     -       .       type "gal4"; occ="CGGGGCGGATCACTCCG"
YLR081W|YLR081W ExactMatch      Misc_feature    104     120     0.0     -       .       type "gal4"; occ="CGGGGCGGATCACTCCG"
YLR081W|YLR081W ExactMatch      Misc_feature    85      101     0.0     -       .       type "gal4"; occ="CGGGGCGGATCACTCCG"
```

Save to file: use “> file_name”

perl FastaRegExpGff.pl gal4-500bp-5seq.tfa "CGG\w{11}CCG" 17 gal4 > gal4-500bp-5seq-cons.gff

On chromosome

```
saerts@[dme-srv-03][~/teach/yeast/chrom]:  perl ~/perl/FastaRegExpGff.pl chr2.fa "CGG\w{11}CCG" 17 gal4 > chr2.regexp.gff
saerts@[dme-srv-03][~/teach/yeast/chrom]: cat chr2.regexp.gff | head
chr2    ExactMatch      Misc_feature    4125    4141    0.0     +       .       type "gal4"; occ="CGGTAAGCGTATCACCG"
chr2    ExactMatch      Misc_feature    11140   11156   0.0     +       .       type "gal4"; occ="CGGGCCTTTCTATACCG"
chr2    ExactMatch      Misc_feature    30297   30313   0.0     +       .       type "gal4"; occ="CGGACCCACACTATCCG"
chr2    ExactMatch      Misc_feature    45259   45275   0.0     +       .       type "gal4"; occ="CGGGTCGAAATTTGCCG"
chr2    ExactMatch      Misc_feature    59859   59875   0.0     +       .       type "gal4"; occ="CGGCACATTCCTTACCG"
chr2    ExactMatch      Misc_feature    76469   76485   0.0     +       .       type "gal4"; occ="CGGCAAGGGCAGCACCG"
chr2    ExactMatch      Misc_feature    89651   89667   0.0     +       .       type "gal4"; occ="CGGGTTTGGCGAGGCCG"
chr2    ExactMatch      Misc_feature    144565  144581  0.0     +       .       type "gal4"; occ="CGGGGATTAAGATACCG"
chr2    ExactMatch      Misc_feature    160205  160221  0.0     +       .       type "gal4"; occ="CGGCTCCAGTGATGCCG"
chr2    ExactMatch      Misc_feature    170373  170389  0.0     +       .       type "gal4"; occ="CGGGAAACTTTTTTCCG"
```

> write standard output (stdout) to file

cat: concatenate files and print on the standard output

| (pipe) : A pipe is a means by which the output from one process becomes the input to a second. In technical terms, the standard output (stout) of one command is sent to the standard input (stdin) of a second command.

head: output the first part of files

BED format

BED format Index ▷

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray).
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

Example:
Here's an example of an annotation track that uses a complete BED definition:

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

http://genome.ucsc.edu/FAQ/FAQformat#format1

On chromosome

```
saerts@[dme-srv-03][~/teach/yeast/chrom]:  perl ~/perl/FastaRegExpGff.pl chr2.fa "CGG\w{11}CCG" 17 gal4 > chr2.regexp.gff
saerts@[dme-srv-03][~/teach/yeast/chrom]: cat chr2.regexp.gff | head
chr2    ExactMatch      Misc_feature    4125    4141    0.0     +       .       type "gal4"; occ="CGGTAAGCGTATCACCG"
chr2    ExactMatch      Misc_feature    11140   11156   0.0     +       .       type "gal4"; occ="CGGGCCTTTCTATACCG"
chr2    ExactMatch      Misc_feature    30297   30313   0.0     +       .       type "gal4"; occ="CGGACCCACACTATCCG"
chr2    ExactMatch      Misc_feature    45259   45275   0.0     +       .       type "gal4"; occ="CGGGTCGAAATTTGCCG"
chr2    ExactMatch      Misc_feature    59859   59875   0.0     +       .       type "gal4"; occ="CGGCACATTCCTTACCG"
chr2    ExactMatch      Misc_feature    76469   76485   0.0     +       .       type "gal4"; occ="CGGCAAGGGCAGCACCG"
chr2    ExactMatch      Misc_feature    89651   89667   0.0     +       .       type "gal4"; occ="CGGGTTTGGCGAGGCCG"
chr2    ExactMatch      Misc_feature    144565  144581  0.0     +       .       type "gal4"; occ="CGGGGATTAAGATACCG"
chr2    ExactMatch      Misc_feature    160205  160221  0.0     +       .       type "gal4"; occ="CGGCTCCAGTGATGCCG"
chr2    ExactMatch      Misc_feature    170373  170389  0.0     +       .       type "gal4"; occ="CGGGAAACTTTTTTCCG"
```

CONVERT TO BED:

cat chr2.regexp.gff | cut -f1,4,5 > chr2.regexp.gff.bed

Or with name:

cat chr2.regexp.gff | cut -f9 | cut -f2 -d “” > names

paste chr2.regexp.gff.bed names > chr2.regexp.gff.names.bed

```
saerts@[dme-srv-03][~/teach/yeast/chrom]: cat chr2.regexp.gff | cut -f1,4,5,9 | cut -f1 -d';' | sed 's/type \"//g' | cut -f1 -d '"'   | head
chr2    4125    4141    gal4
chr2    11140   11156   gal4
chr2    30297   30313   gal4
chr2    45259   45275   gal4
chr2    59859   59875   gal4
chr2    76469   76485   gal4
chr2    89651   89667   gal4
chr2    144565  144581  gal4
chr2    160205  160221  gal4
chr2    170373  170389  gal4
```

Basic commands

- **`ls`**, Give a listing of the current directory. Try also `ls -l`
- **`cp`**, Copy file from source to destination
- **`mv`**, Move file from source to destination. If both are the same directory, the file is renamed
- **`vi`**, Edit a file. `vi` is one of the most powerful text editors
- **`chmod`**, Change file permissions
- **`mkdir`**, **`rmdir`** Make/Remove a directory
- **`cd`**, Change directory
- **`rm`**, Remove a file. Can also remove directory tree
- **`man ls`**, Get help for `ls`. All commands have help

Manipulating files

- **cat**
  - Concatenate program. Can be used to concatenate multiple files together into a single file, or, much more frequently, to send the contents of a file to the terminal for viewing.
- **more**
  - Scroll through a file page by page. Very useful when viewing large files. Works even with files that are too big to be opened by a text editor.
- **less**
  - A version of **more** with more features.
- **head**
  - View the head (top) of a file. You can control how many lines to view.
- **tail**
  - View the tail (bottom) of a file. You can control how many lines to view. You can also use **tail** to view a growing file.
- **wc**
  - Count words, lines and/or characters in one or more files.
- **tr**
  - Substitute one character for another. Also useful for deleting characters.
- **sort**
  - Sort the lines in a file alphabetically or numerically.
- **uniq**
  - Remove duplicated lines in a file.
- **cut**
  - Remove sections from each line of a file or files.
- **fold**
  - Wrap each input line to fit in a specified width.
- **grep**
  - Filter a file for lines matching a specified pattern. Can also be reversed to print out lines that don't match the specified pattern.
- **gzip (gunzip)**
  - Compress (uncompress) a file.
- **tar**
  - Archive or unarchive an entire directory into a single file.
- **pico**
  - Run the pico text editor (good for beginners).
- **emacs**
  - Run the Emacs text editor (good for experts).

Create custom track in UCSC

- UCSC Genomes
  - Add regexp BED as custom track
- UCSC Tables
  - Create custom track with only exons
  - Select all regexp records that don’t overlap with exons

Which genes?

- Make a track with 500bp upstream sequence for all genes
- Intersect this track with Gal4_noExon track
  - save BED file
  - cat *file* | cut -f4 | cut -f1 -d'_'
  - Or paste in Excel

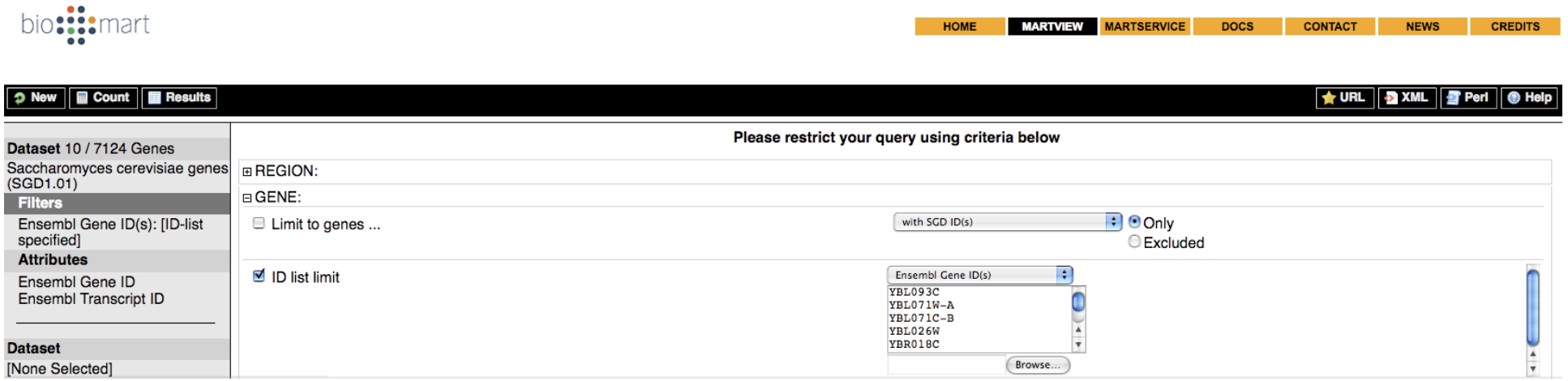
```
chr2    45412   45912   YBL093C_up_500_chr2_44913_r_up_500_chr2_45413_r 0       -
chr2    88972   89472   YBL071W-A_up_500_chr2_89473_f_up_500_chr2_88973_f       0       +
chr2    90051   90551   YBL071C-B_up_500_chr2_89552_r_up_500_chr2_90052_r       0       -
chr2    169584  170084  YBL026W_up_500_chr2_170085_f_up_500_chr2_169585_f       0       +
chr2    275986  276486  YBR018C_up_500_chr2_275487_r_up_500_chr2_275987_r       0       -
chr2    278811  279311  YBR019C_up_500_chr2_278312_r_up_500_chr2_278812_r       0       -
chr2    277979  278479  YBR020W_up_500_chr2_278480_f_up_500_chr2_277980_f       0       +
chr2    644504  645004  YBR210W_up_500_chr2_645005_f_up_500_chr2_644505_f       0       +
chr2    702297  702797  YBR240C_up_500_chr2_701798_r_up_500_chr2_702298_r       0       -
chr2    785958  786458  YBR293W_up_500_chr2_786459_f_up_500_chr2_785959_f       0       +
```

	A	B	C	D	E	F	G
1	chr2	45412	45912	YBL093C_up_	0	-	YBL093C
2	chr2	88972	89472	YBL071W-A_	0	+	YBL071W-A
3	chr2	90051	90551	YBL071C-B_u	0	-	YBL071C-B
4	chr2	169584	170084	YBL026W_up	0	+	YBL026W
5	chr2	275986	276486	YBR018C_up	0	-	YBR018C
6	chr2	278811	279311	YBR019C_up	0	-	YBR019C
7	chr2	277979	278479	YBR020W_up	0	+	YBR020W
8	chr2	644504	645004	YBR210W_up	0	+	YBR210W
9	chr2	702297	702797	YBR240C_up	0	-	YBR240C
10	chr2	785958	786458	YBR293W_up	0	+	YBR293W

=LEFT(D1,FIND("_",D1)-1)

YBL093C
YBL071W-A
YBL071C-B
YBL026W
YBR018C
YBR019C
YBR020W
YBR210W
YBR240C
YBR293W

ID -> Gene (BioMart)



Genome-wide prediction: validation of predicted targets

http://go.princeton.edu/cgi-bin/GOTermFinder/GOTermFinder

Result Table

Terms from the Process Ontology of gene_association.sgd with p-value <= 0.01

Gene Ontology term	Cluster frequency	Genome frequency of use	Corrected P-value	FDR	False Positives	Genes annotated to the term
galactose metabolic process	7 out of 301 genes, 2.3%	11 out of 6370 genes, 0.2%	5.40e-05	0.00%	0.00	GAL1, GAL4, GAL2, GAL10, GAL80, GAL7, GAL3
cellular response to nutrient	6 out of 301 genes, 2.0%	13 out of 6370 genes, 0.2%	0.00573	0.00%	0.00	GAL1, USV1, GAL4, MIG2, GAL80, GAL3
regulation of transcription by carbon catabolites	6 out of 301 genes, 2.0%	13 out of 6370 genes, 0.2%	0.00573	0.00%	0.00	GAL1, USV1, GAL4, MIG2, GAL80, GAL3
monosaccharide catabolic process	12 out of 301 genes, 4.0%	62 out of 6370 genes, 1.0%	0.00847	0.00%	0.00	ENO2, MDH1, GAL7, TKL1, GCY1, HXK1, GAL1, YJR096W, GAL10, ZWF1, SOL3, GAL3

Overlap with experimentally found binding regions?

- Compare with ChIP-chip data
  - Make custom track for ChIP-chip regions where GAL4 binds
    - TransRegCodeProbe filter: “AND tfList like ‘%GAL4%’
    - Intersect this track with gal4_noExon track
    - Produce hyperlinks or BED

Galaxy

http://main.g2.bx.psu.edu/

- Retrieve data from UCSC or Ensembl; output data from UCSC Tables to Galaxy
- Use “Get data” to paste a chromosome or region (BED format)
  - chr2 1 813136 chr2
  - Then use “Fetch Sequences - Extract genomic sequence” to get sequence of this region (whole chrom)
- Better: upload chrN.fa
- Galaxy has EMBOSS programs available
  - run fuzznuc on chromosome, choose GFF as output, use only + strand (otherwise start > end), repeat for reverse complementary motif
  - Make sure “chrN” is name in first column
  - Convert formats: GFF-to-BED

Position weight matrix

Position weight matrix

GAL4

Get regulated genes | Get sites | Get consensus | Get matrix | Get affinity data
Get genomewise distribution | Sort by copy No | Sort by function category

```
>YBR018C        AGCGCTCGGACAACTGTTGACC
>YBR018C        ATACTTCGGAGCACTGTTGAGCG
>YBR019C        CGGCGGCTTCTAATCCG
>YBR019C        TCGGAGGGCTGTCGCCCG
>YBR019C        CGGAGGAGAGTCTTCCG
>YBR019C        ATTGTTCGGAGCAGTGCGGCGCG
>YBR020W        CGCGCCGCACTGCTCCGAACAAT
>YBR020W        CGGAAGACTCTCCTCCG
>YBR020W        CGGGCGACAGCCCTCCGA
>YBR020W        CGGATTAGAAGCCGCCG
>YLR081W        TATCGGGGCGGATCACTCCGAAC
>YLR081W        CACCGGCGGTCTTTCGTCCGTGC
>YML051W        CGGCGCACTCTCGCCCG
>YOR120W        TCGGGGCAGACTATTCCGG
```

```
>YBR018C
CGGACAACTGTTGACCG
>YBR018C
CGGAGCACTGTTGAGCG
>YBR019C
CGGCGGCTTCTAATCCG
>YBR019C
CGGAGGGCTGTCGCCCG
>YBR019C
CGGAGGAGAGTCTTCCG
>YBR019C
CGGAGCAGTGCGGCGCG
>YLR081W
CGGGGCGGATCACTCCG
>YLR081W
CGGCGGTCTTTCGTCCG
>YML051W
CGGCGCACTCTCGCCCG
>YOR120W
CGGGGCAGACTATTCCG
```

Get regulated genes | Get sites | Get consensus | Get matrix | Get affinity data
Get genomewise distribution | Sort by copy No | Sort by function category

```
>GAL4    5        118      SCPD

A    0    0    0    4    1    1    7    0    5    1    0    2    0    2    0    0    0
T    0    0    0    0    1    1    1    0    5    2    7    2    1    6    0    0    0
G    0   10    9    4    5    3    2    3    0    3    1    1    4    1    1    0   10
C   10    0    1    2    3    5    0    7    0    4    2    5    5    1    9   10    0
```

Frequencies, weights

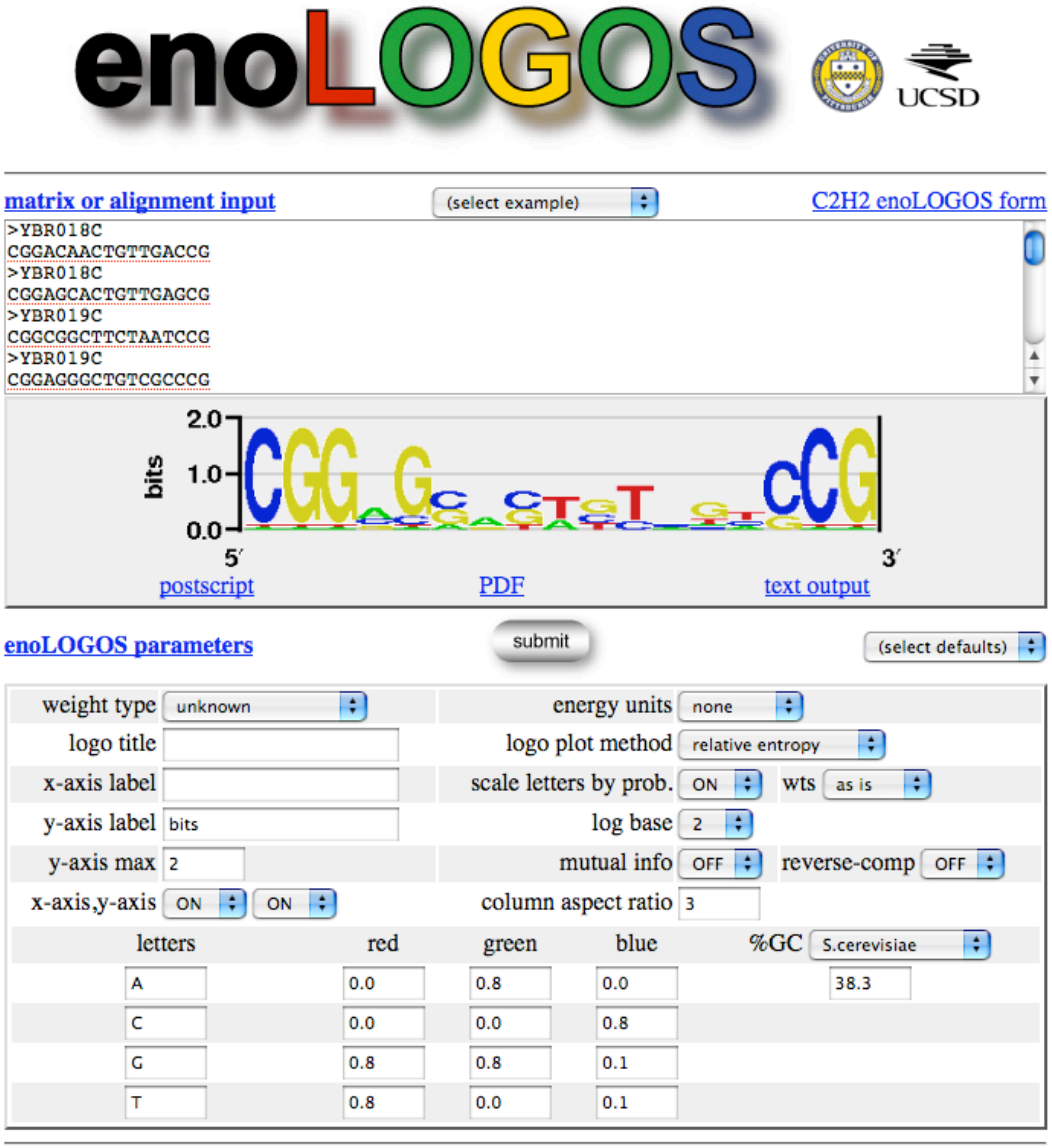
```
>GAL4    5        118      SCPD

A    0    0    0    4    1    1    7    0    5    1    0    2    0    2    0    0    0
T    0    0    0    0    1    1    1    0    5    2    7    2    1    6    0    0    0
G    0   10    9    4    5    3    2    3    0    3    1    1    4    1    1    0   10
C   10    0    1    2    3    5    0    7    0    4    2    5    5    1    9   10    0
```

0.023	0.932	0.023	0.023
0.023	0.023	0.932	0.023
0.023	0.114	0.841	0.023
0.386	0.205	0.386	0.023
0.114	0.295	0.477	0.114
0.114	0.477	0.295	0.114
0.659	0.023	0.205	0.114
0.023	0.659	0.295	0.023
0.477	0.023	0.023	0.477
0.114	0.386	0.295	0.205
0.023	0.205	0.114	0.659
0.205	0.477	0.114	0.205
0.023	0.477	0.386	0.114
0.205	0.114	0.114	0.568
0.023	0.841	0.114	0.023
0.023	0.932	0.023	0.023
0.023	0.023	0.932	0.023

-2.398	1.316	-2.398	-2.398
-2.398	-2.398	1.316	-2.398
-2.398	-0.788	1.213	-2.398
0.435	-0.201	0.435	-2.398
-0.788	0.167	0.647	-0.788
-0.788	0.647	0.167	-0.788
0.969	-2.398	-0.201	-0.788
-2.398	0.969	0.167	-2.398
0.647	-2.398	-2.398	0.647
-0.788	0.435	0.167	-0.201
-2.398	-0.201	-0.788	0.969
-0.201	0.647	-0.788	-0.201
-2.398	0.647	0.435	-0.788
-0.201	-0.788	-0.788	0.821
-2.398	1.213	-0.788	-2.398
-2.398	1.316	-2.398	-2.398
-2.398	-2.398	1.316	-2.398

$$\ln \frac{(n_{i,j} + p_i) \,/\, (N+1)}{p_i} \;;$$

- http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi



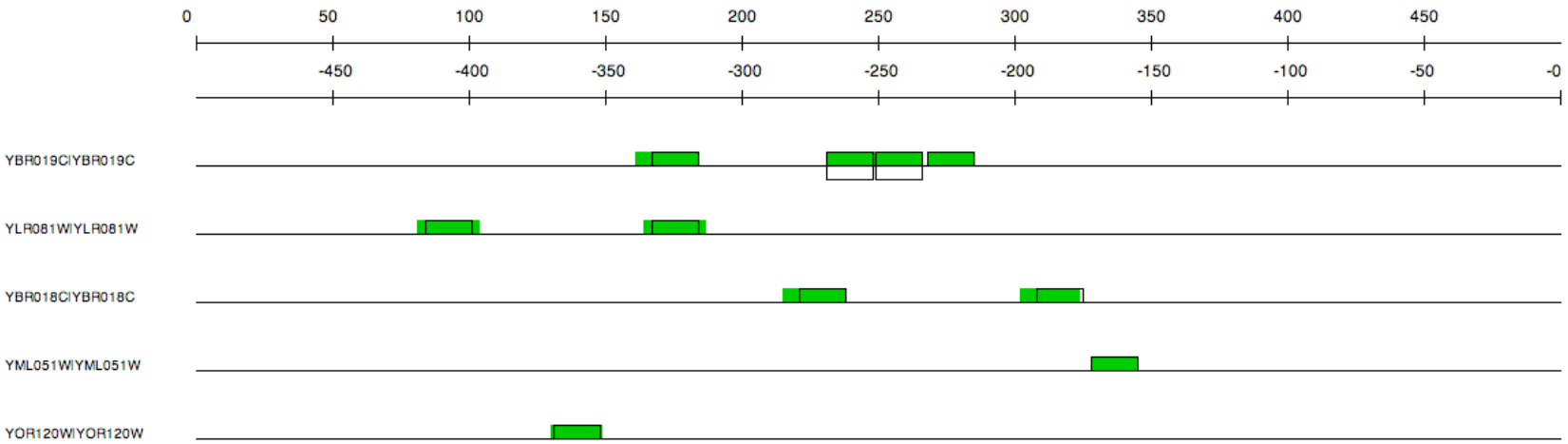
Score an instance with a PWM

			A	C	G	T	MAX	MIN
	1.316	C	-2.398	1.316	-2.398	-2.398	1.31567679	-2.398
	1.316	G	-2.398	-2.398	1.316	-2.398	1.31567679	-2.398
	1.213	G	-2.398	-0.788	1.213	-2.398	1.21302264	-2.398
	0.435	A	0.435	-0.201	0.435	-2.398	0.43531807	-2.398
	0.167	C	-0.788	0.167	0.647	-0.788	0.64662716	-0.788
	-0.788	A	-0.788	0.647	0.167	-0.788	0.64662716	-0.788
	0.969	A	0.969	-2.398	-0.201	-0.788	0.96940056	-2.398
	0.969	C	-2.398	0.969	0.167	-2.398	0.96940056	-2.398
	0.647	T	0.647	-2.398	-2.398	0.647	0.64662716	-2.398
	0.167	G	-0.788	0.435	0.167	-0.201	0.43531807	-0.788
	0.969	T	-2.398	-0.201	-0.788	0.969	0.96940056	-2.398
	-0.201	T	-0.201	0.647	-0.788	-0.201	0.64662716	-0.788
	0.435	G	-2.398	0.647	0.435	-0.788	0.64662716	-2.398
	-0.201	A	-0.201	-0.788	-0.788	0.821	0.82098055	-0.788
	1.213	C	-2.398	1.213	-0.788	-2.398	1.21302264	-2.398
	1.316	C	-2.398	1.316	-2.398	-2.398	1.31567679	-2.398
	1.316	G	-2.398	-2.398	1.316	-2.398	1.31567679	-2.398
SUM	11.259						15.5217066	-32.717
W-Wmin/Wmax-Wmin	0.91162331							

AGCGCT**CGGACAACTGTTGACCG**

Toucan - find instances back?

- MotifLocator, threshold=0.9, bg0



Genome-wide prediction of Gal4 targets using a PWM

- RSAT
- Command line
  - MotifLocator or Patser standalone version
    - On all 500bp upstream sequences from Ensembl or UCSC or RSAT
    - On chromosome N
- Difference with regular expression search?
  - How many overlap/ differ
    - Excel, grep, …

```
#INCLUSive Motif Model
#
#ID = GAL4-SCPD
#W = 17
0       10      0       0
0       0       10      0
0       1       9       0
4       2       4       0
1       3       5       1
1       5       3       1
7       0       2       1
0       7       3       0
5       0       0       5
1       4       3       2
0       2       1       7
2       5       1       2
0       5       4       1
2       1       1       6
0       9       1       0
0       10      0       0
0       0       10      0
```

Patser command-line

- echo chr2.seq | ~/software/patser/patser-v3b -m gal4.patser.v -v -c -l 8.631 | ~/perl/parsePatser.pl 17

Exercise

- Exercise 3

PWM determination

- SELEX
- Protein binding microarrays
- Bacterial one-hybrid

PWMs for all human & mouse homeodomains

Cell
Volume 133, Issue 7, 27 June 2008, Pages 1266-1276

Article | Figures/Tables | References | PDF (2124 K)

doi:10.1016/j.cell.2008.05.024 Cite or Link Using DOI

Resource

Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences

Michael F. Berger[1, 3, 8], Gwenael Badis[5, 8], Andrew R. Gehrke[1, 8], Shaheynoor Talukder[5, 8], Anthony A. Philippakis[1, 3, 6], Lourdes Peña-Castillo[4], Trevis M. Alleyne[5], Sanie Mnaimneh[4], Olga B. Botvinnik[1, 7], Esther T. Chan[5], Faiqua Khalid[4], Wen Zhang[5], Daniel Newburger[1], Savina A. Jaeger[1], Quaid D. Morris[4, 5], Martha L. Bulyk[1, 2, 3, 6], and Timothy R. Hughes[4, 5],

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA [2]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA [3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA [4]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, Canada [5]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada [6]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA [7]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 21 November 2007; revised 10 March 2008; accepted 12 May 2008. Published: June 26, 2008. Available online 26 June 2008.

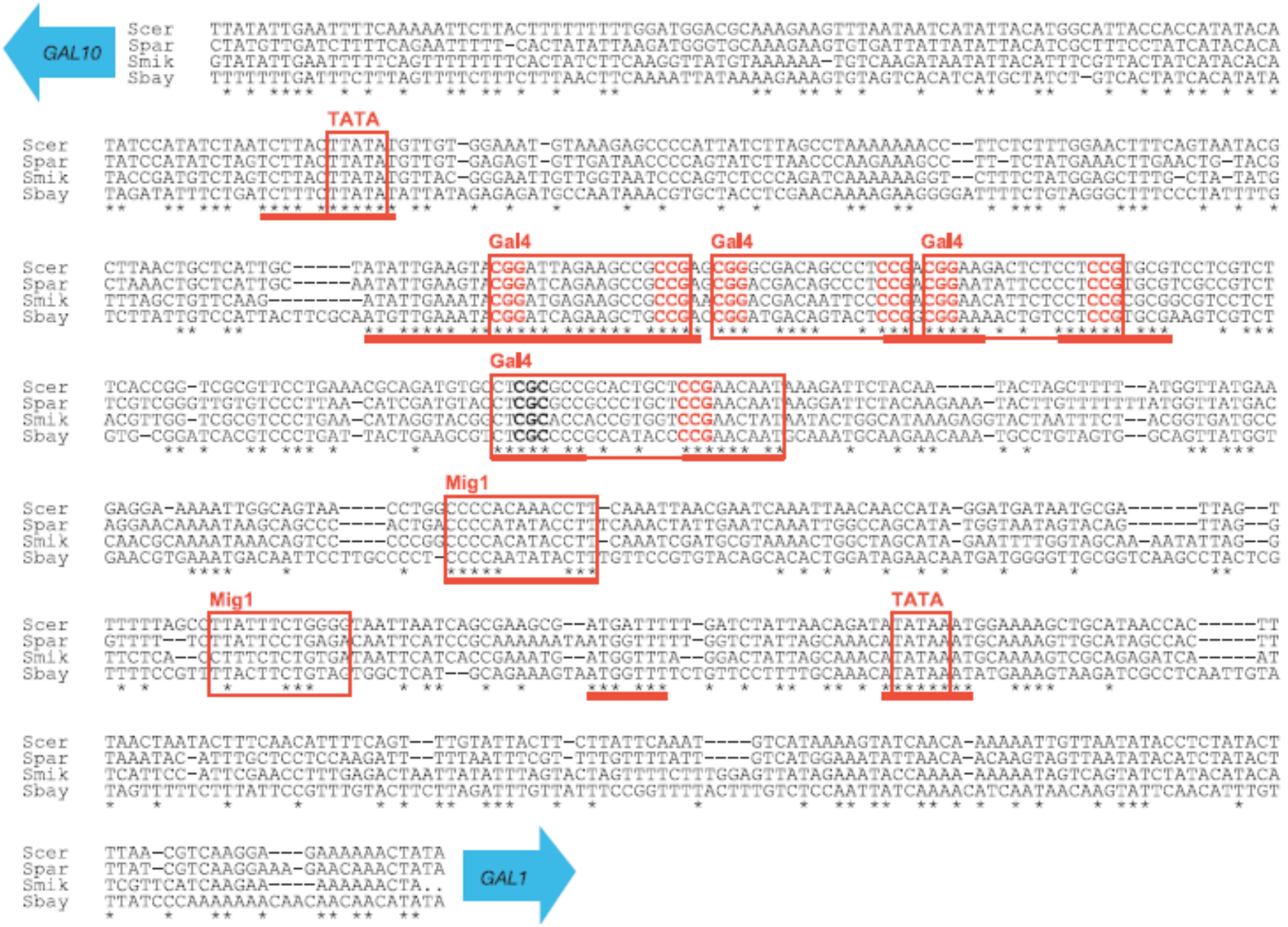
Refers to: **Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites**
Cell, Volume 133, Issue 7, 27 June 2008, Pages 1277-1289
Marcus B. Noyes, Ryan G. Christensen, Atsuya Wakabayashi, Gary D. Stormo, Michael H. Brodsky, Scot A. Wolfe
PDF (2581 K)

Phylogenetic footprinting

Use sequence conservation in scanning

- Basic assumption: functional sequences evolve more slowly than non-functional sequences, as they are subject to selection pressure.
- Basic approach:
  - Identify conserved regions by sequence alignment algorithms
  - Restrict motif finding in conserved regions.

GAL4



Kellis et al Nature 2003

Sequencing and comparison of yeast species to identify genes and regulatory elements

Manolis Kellis*†, Nick Patterson*, Matthew Endrizzi*, Bruce Birren* & Eric S. Lander*‡

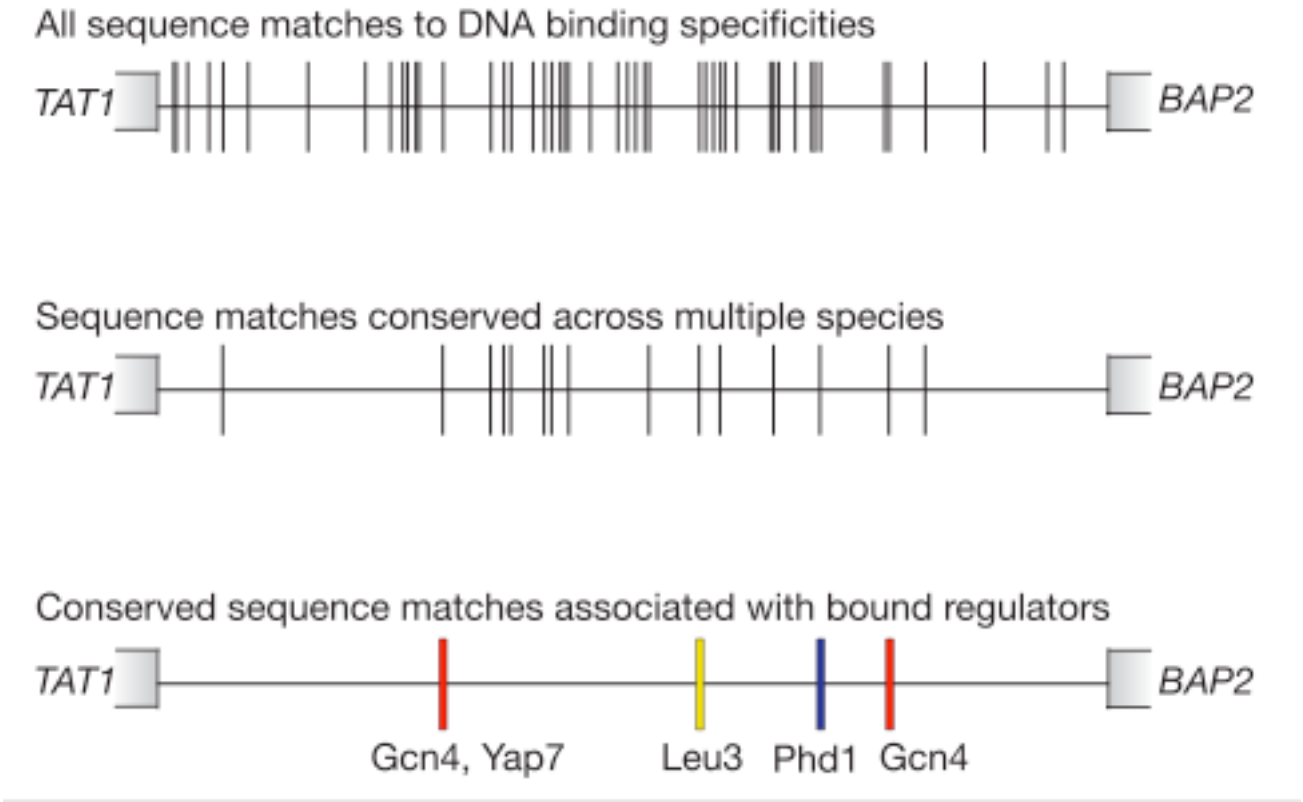
** Whitehead/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
† Department of Computer Science and ‡ Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Identifying the functional elements encoded in a genome is one of the principal challenges in modern biology. Comparative genomics should offer a powerful, general approach. Here, we present a comparative analysis of the yeast *Saccharomyces cerevisiae* based on high-quality draft sequences of three related species (*S. paradoxus*, *S. mikatae* and *S. bayanus*). We first aligned the genomes and characterized their evolution, defining the regions and mechanisms of change. We then developed methods for direct identification of genes and regulatory motifs. The gene analysis yielded a major revision to the yeast gene catalogue, affecting approximately 15% of all genes and reducing the total count by about 500 genes. The motif analysis automatically identified 72 genome-wide elements, including most known regulatory motifs and numerous new motifs. We inferred a putative function for most of these motifs, and provided insights into their combinatorial interactions. The results have implications for genome analysis of diverse organisms, including the human.

- GAL4
  - 96 times in intergenic regions, 415 times in genic regions
  - conservation rate (12,5%) is 5-fold higher than equivalent random motifs (2.4%)
  - Intergenic GAL4 are 12,5% conserved while genic motifs only 3.1%
    - By contrast: random motifs are less frequently conserved intergenic (3.1%) than genic (7.0%)

Use sequence conservation in scanning

- Harbison et al
- Criterium for conserved motif: the aligned sequence of at least 2 other *sensu stricto* yeast species have a motif match (>60% of motif maximum)



Use sequence conservation in scanning

- + add phylogenetic tree
  - Evolutionary model
  - Monkey

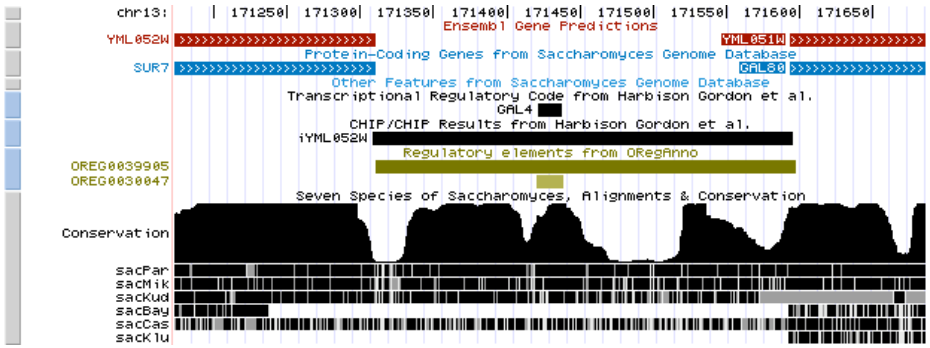
    (Moses et al. Genome Biology 2004)

Table 2

Performance of different scores in recognizing functional and nonfunctional sites

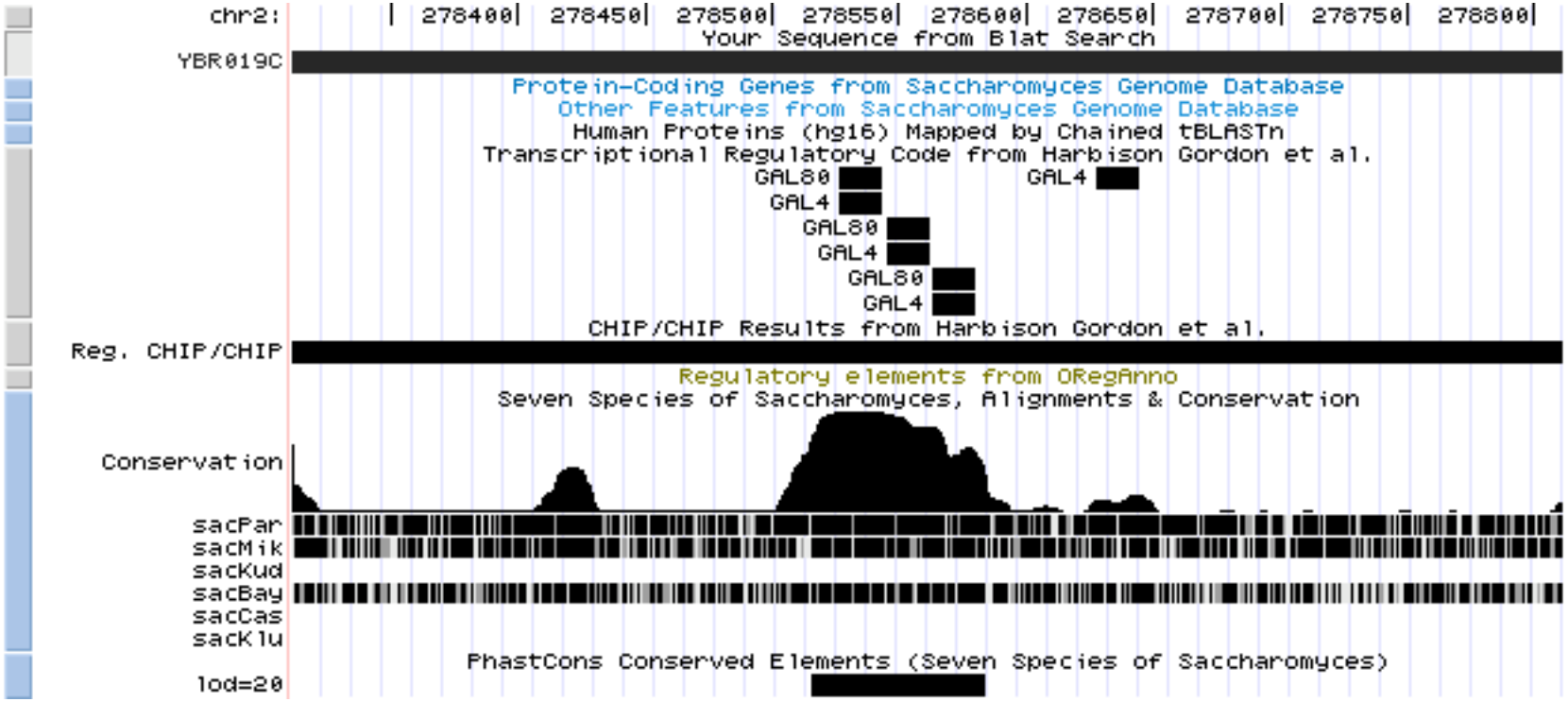
Percent of binding sites assigned a lower *p*-value	Gal4p	
	Positives (n = 10)	Negatives (n = 10)
Halpern-Bruno vs simple	90%	30%
Halpern-Bruno vs *S. cerevisiae* alone	100%	40%
Simple vs *S. cerevisiae* alone	100%	30%

- PhastCons scores



Exercise

- Intersect Gal4 predicted binding sites with PhastCons conserved elements



Seven Species of Saccharomyces, Alignments & Conservation

Conservation score statistics

Capitalize coding exons based on SGD Genes show all bases

Place cursor over species for alignment detail. Click on 'B' to link to browser for aligned species, click on 'D' to get DNA for aligned species.

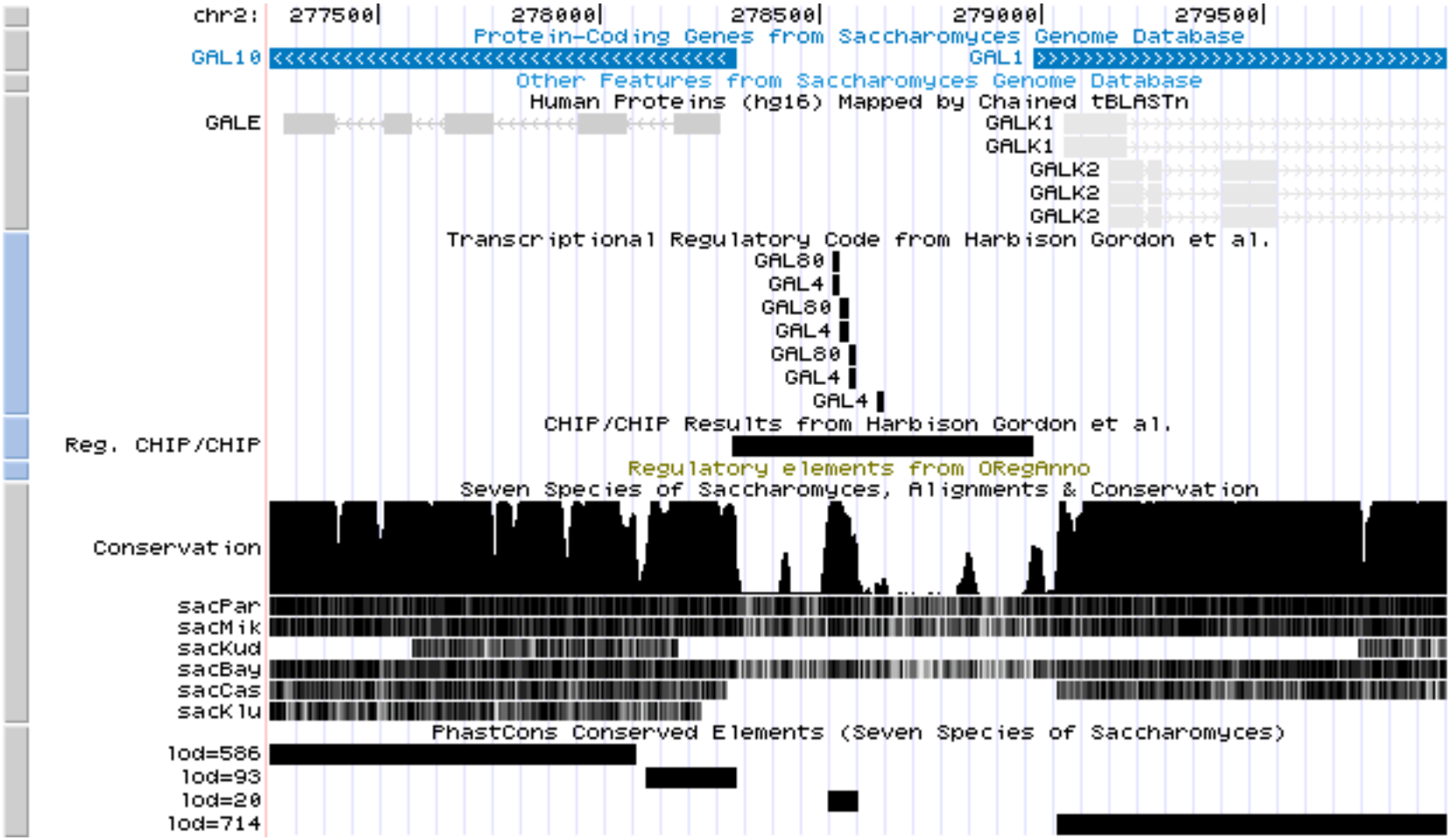
Alignment block 1 of 1 in window, 278312 - 278811, 500 bps

```
B D  S. cerevisiae  ttata---------ttgaattttcaaaaattcttacttttttttttggatggacgcaaagaagtttaat--
           sacPar   ctatg---------ttgatcttttcagaattttt-cactatattaagatgggtgcaaagaagtgtgat--
           sacMik   gtata---------ttgaattttcagtttttttttcactatcttcaaggttatgtaaaaaatgtcaa---
           sacBay   atacgtttatttttttgatttctttagttttctttctttaacttcaaaattataaaagaaagtgtagtca

     S. cerevisiae  aatcatattacatggcattaccaccatatacatatccatatctaatcttacttatatgttgtggaaa--t
           sacPar   tattatattacatcgctttcctatcatacacatatccatatctagtcttacttatatgttgtgagag--t
           sacMik   gataatattacatttcgttactatcatacacatacgatgtctagtcttacttatatgttacgggaat-t
           sacBay   catcatgctatct---gtcactatcacatatatagatatttctgatctttcttatatattatagagagat

     S. cerevisiae  gtaaagagccccattatcttagcctaaaaaaa--ccttctctttggaactttcagtaatacgcttaactg
           sacPar   gttgataaccccagtatcttaacccaagaaag---cctttctatgaaacttg-aactgtacgctaaactg
           sacMik   gttggtaatcccagtctcccagatcaaaaaag--gtctttctatggagcttt--gctatatgtttagctg
           sacBay   gccaataaacgtgctacctcgaacaaaagaaggggattttctgtagggctttccctattttgtcttattg

     S. cerevisiae  ctcatt-----gctatattgaagtacggattagaagccgccgagcgggcgacagccctccgacggaagac
           sacPar   ctcatt-----gcaatattgaagtacggatcagaagccgccgagcggacgacagccctccgacggaatat
           sacMik   ttca--------agatattgaaatacggatgagaagccgccgaacggacgacaattccccgacggaacat
           sacBay   tccattacttcgcaatgttgaaatacggatcagaagctgccgaccggatgacagtactccggcggaaaac

     S. cerevisiae  tctcctccgtgcgtcctcgtcttcaccggtcgcgttcctgaaacgcagatgtgcctcgcgccgcactgct
           sacPar   tcccctccgtgcgtcgccgtcttcgtcggttgtgtccctt-aacatcgatgtacctcgcgccgccctgct
           sacMik   tctcctccgcgcggcgtcctctacgttggtcgcgtccctg-aacataggtacggctcgcaccaccgtggt
           sacBay   tgtcctccgtgcgaagtcgtctgtgcggatcacgtccctg-attactgaagcgtctcgccccgccatacc

     S. cerevisiae  ccgaacaataaagattctacaa-----tactag--cttttatggttatg-aagaggaaaaattggcagta
           sacPar   ccgaacaataaggattctacaa-gaaatacttgtttttttatggttatgacaggaacaaaataagcagcc
           sacMik   ccgaactataatactggcataaagaggtactaa--tttctacggtgatgcccaacgcaaaataaacagtc
           sacBay   ccgaacaatgcaaatgcaagaa-caaatgcctg--tagtggcagttatggtgaacgtgaaat-gacaatt

     S. cerevisiae  acctggccccacaaacc-ttcaaattaacgaatcaaattaacaaccataggatgataat------gcg--
           sacPar   cactgaccccatataccttttcaaactattgaatcaaattggccagcatatggtaatagt------aca--
           sacMik   ccccggccccacatacc-ttcaaatcgatgcgtaaaactggctagcatagaattttggta-----gcaaa
           sacBay   ccttgcccctccccaatatactttgttccgtgtacagcacactgg-atagaacaatgatggggttgcg--

     S. cerevisiae  --attag--------ttttttagccttatttctggggtaattaatcagc--gaagcgat
           sacPar   --gttag--------ggttttc--cttattcctgagacaattcatccgcaaaaaataat
           sacMik   atattag--------gttctca--cctttctctgtgataattcatcacc--gaaatgat
           sacBay   --gtcaagcctactcgttttccgttttacttctgtagtggctcatgcag--aaagtaat
```

Words of caution

- Non-coding DNA alignment is more difficult
  - Incorrect alignment
  - Difficult to find the orthologous motif
  - Allow motif movement within N bp window
  - ‘preserved motifs’ instead of ‘conserved’
  - Motif is present in orthologous promoters, not necessarily aligned
- Motif conservation vs motif diversity
  - evolution of animal morphology by cis-evolution

Motif clusters



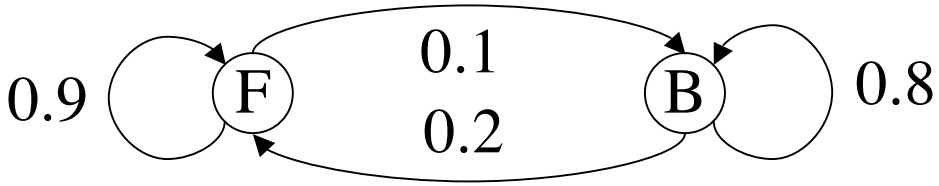
Motif Clusters

- Sliding window
  - E.g., 500 bp
  - Stepsize 250 bp
  - minimally 3 sites
- Hidden Markov Model
  - E.g., cister, comet, cluster-buster, ahab, stubb, …

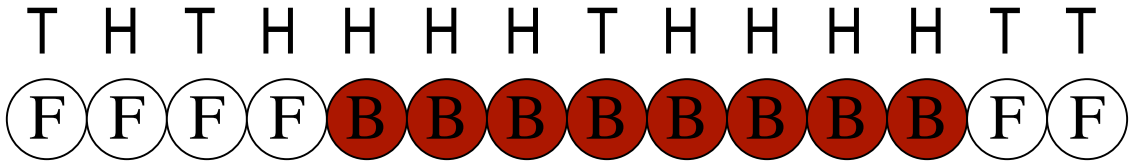
Hidden Markov Models (HMM)

Coin flip

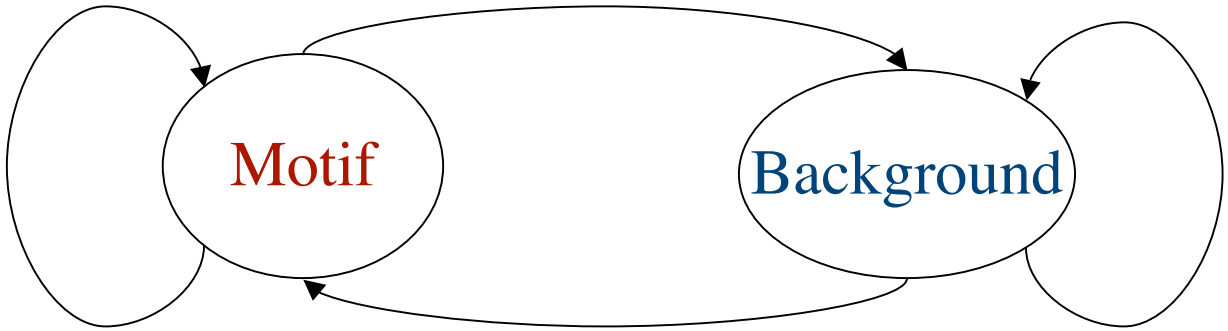
- Emission: P(H| F)=0.5 , P(H| B)=0.9
- Transition

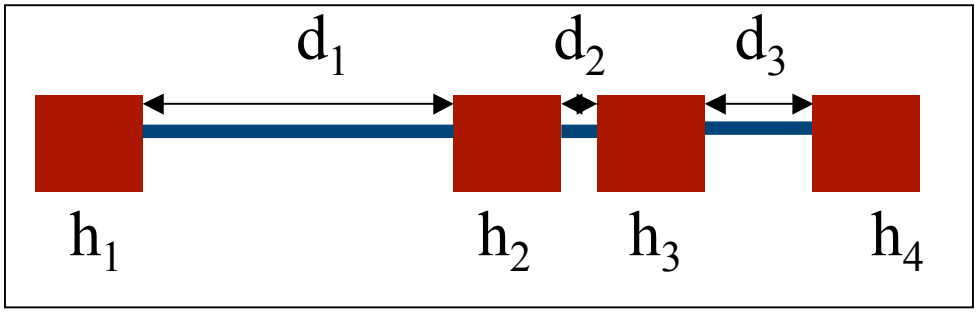


- Observe
- States



HMM for motif cluster





$$M = h_1 + h_2 + h_3 + h_4 + g(d_2 + d_3 + d_4)$$

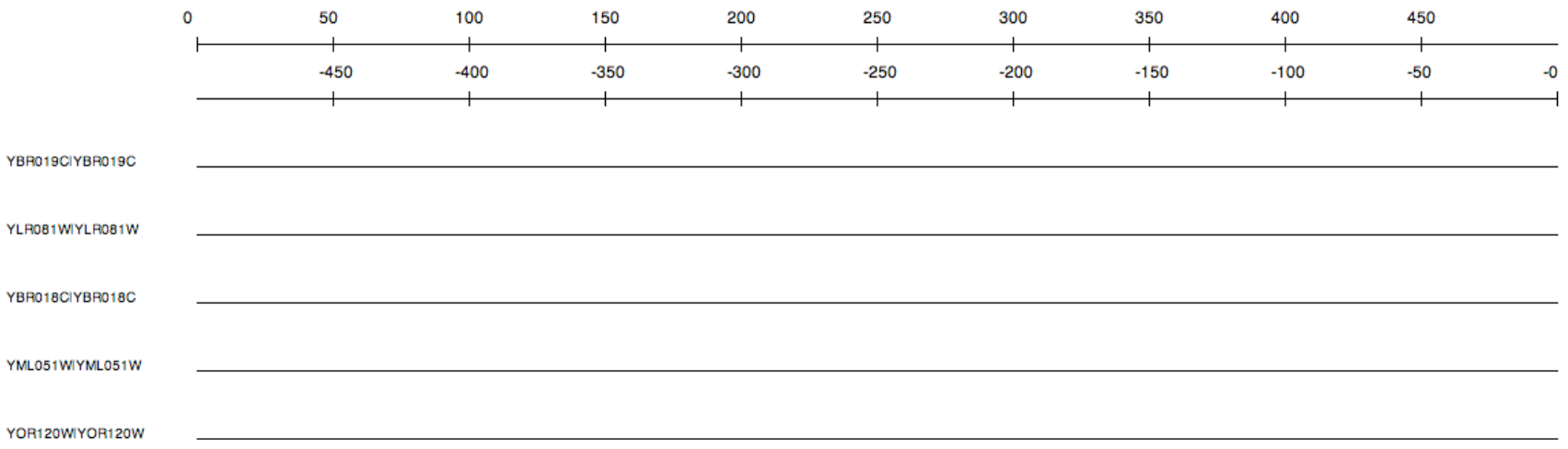
The "match score" (M) for the cluster is the sum of the log-odds scores.

- Exercise 4
  - Run Cluster-Buster on yeast chromosome

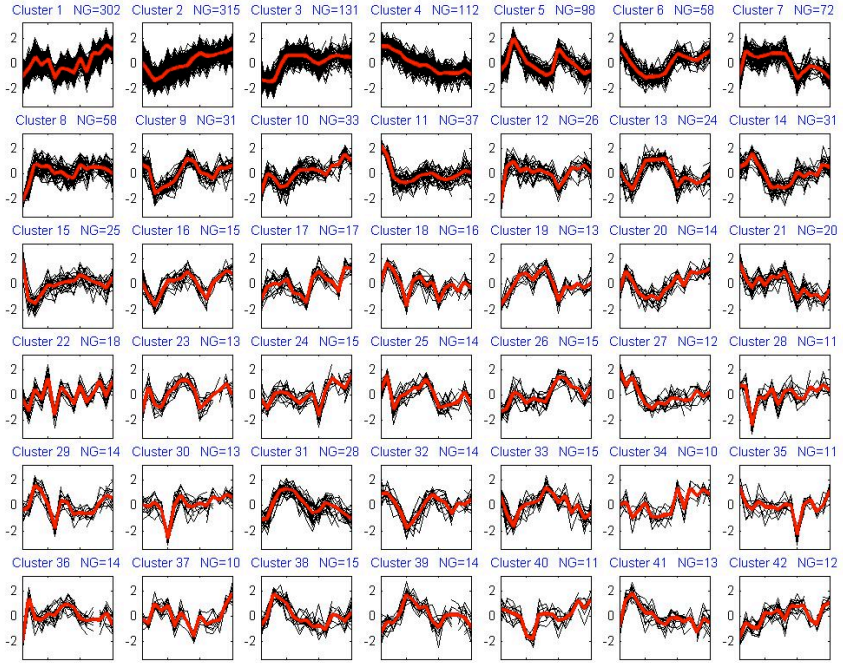
Motif discovery

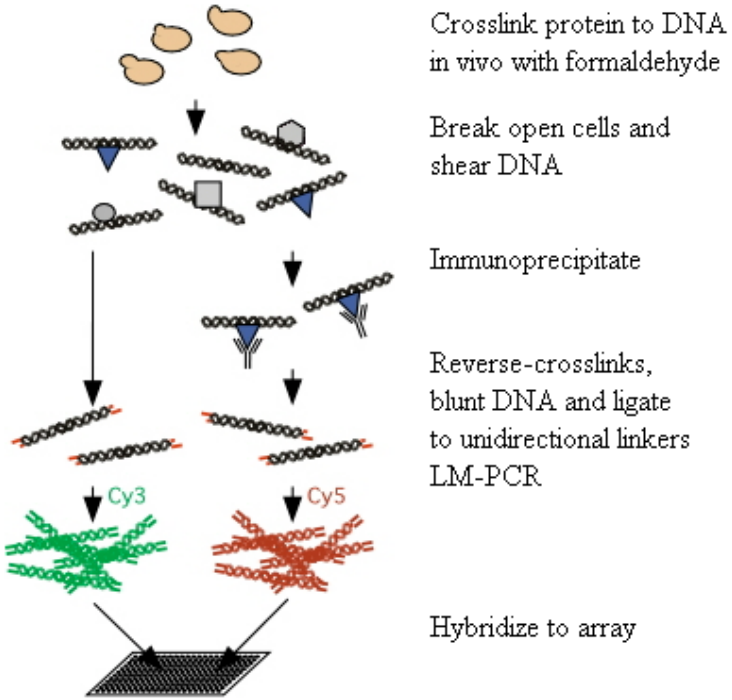
Motif discovery

- Given: set of co-regulated genes
- Question: what are the regulatory motifs?



Co-regulated genes: microarrays & ChIP-chip/seq





GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC

CACATCGCATCACGTGACCAGT...GACATGGACGGC

GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA

TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG

CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT

GTA...GATGGCTGCACCACGTGTATGC

CGA...CACATCGCATCACGTGACCAGT

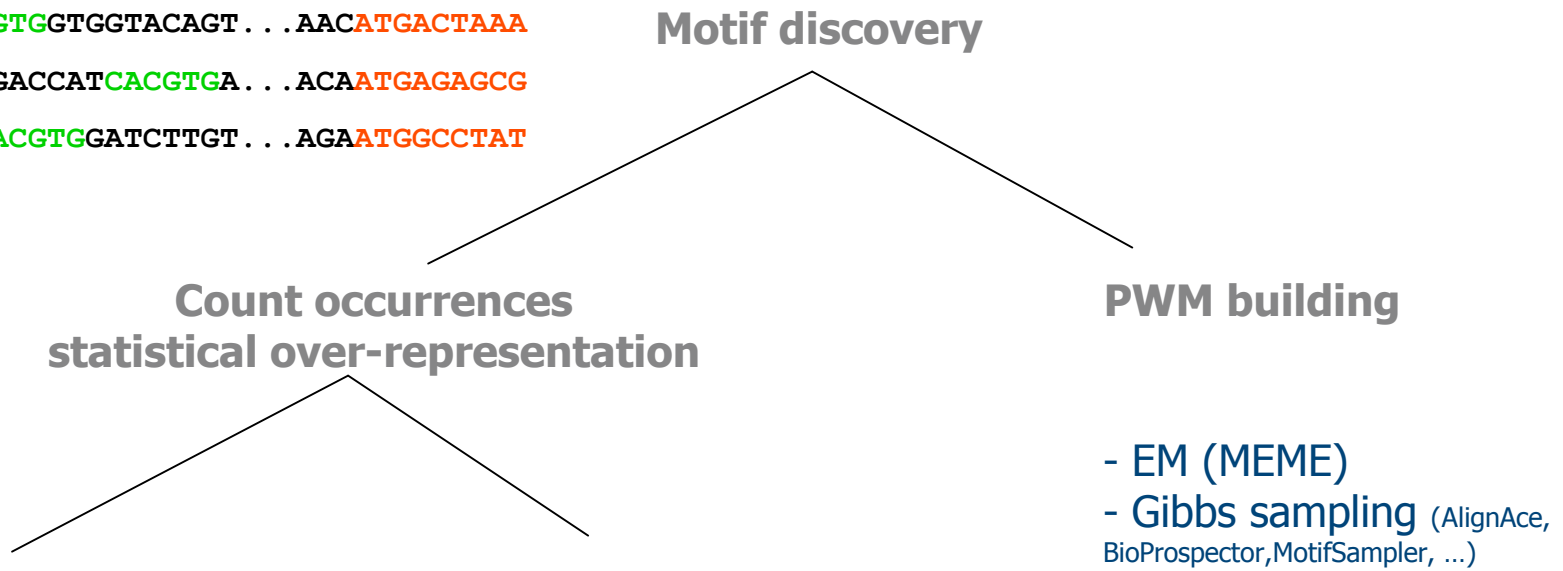
TTA...GCCTCGCACGTGGTGGTACAGT

GGC...TCTCGTTAGGACCATCACGTGA

CTG...CGCTAGCCCACGTGGATCTTGT

Motif discovery

```
GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC
CACATCGCATCACGTGACCAGT...GACATGGACGGC
GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA
TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG
CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT
```



- EM (MEME)
- Gibbs sampling (AlignAce, BioProspector,MotifSampler, …)

- all possible k-mers
- 4^k possibilities
- binomial analysis w/ background frequencies

- library of PWM's (TRANSFAC, JASPAR)
- binomial analysis w/ background frequencies

van Helden J, et al.. J Mol Biol. 1998 4;281(5):827-42.

Aerts S et al. Nucleic Acids Res. 2003 15;31(6):1753-64.

- *Lawrence CE, et al. Science. 1993 8;262(5131):208-14.*
- *Bailey T et al. Proceedings of ISMB, pp. 28-36, 1994.*

Statistical over-representation

- Binomial probability

Example

On a ten question multiple choice test, with 4 options per question, the probability of getting 5 and only 5 answers correct if the answers are guessed can be calculated like so:

$p = 0.25$
$q = 1 - p = 0.75$
$k = 5$
$n = 10$

$$\begin{aligned} P(pass) &= \binom{n}{k} p^k q^{n-k} \\ &= \binom{10}{5} (0.25)^5 (0.75)^5 \\ &\approx 5.8\% \end{aligned}$$

Therefore if somebody guesses 10 answers on a multiple choice test with 4 options, they have about a 5.8% chance of getting 5 and only 5 correct answers. If 5 or more correct answers are needed to pass, then the probability of passing can be calculated by adding the probability of getting 5 (and only 5) answers correct, 6 (and only 6) answers correct, and so on up to 10 answers correct. This total probability is approximately 7.8%.

$$E(occ\{b\}) = F_e\{b\} \times 2 \times \sum_{i=1}^{S} (L_i - w + 1) = F_e\{b\} \times T$$

$$P(occ\{b\} = n) = \frac{T!}{(T-n)! \times n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)}$$

$$P(occ\{b\} \geq n) = \sum_{j=n}^{T} P(occ\{b\} = j)$$

Oligo-analysis example

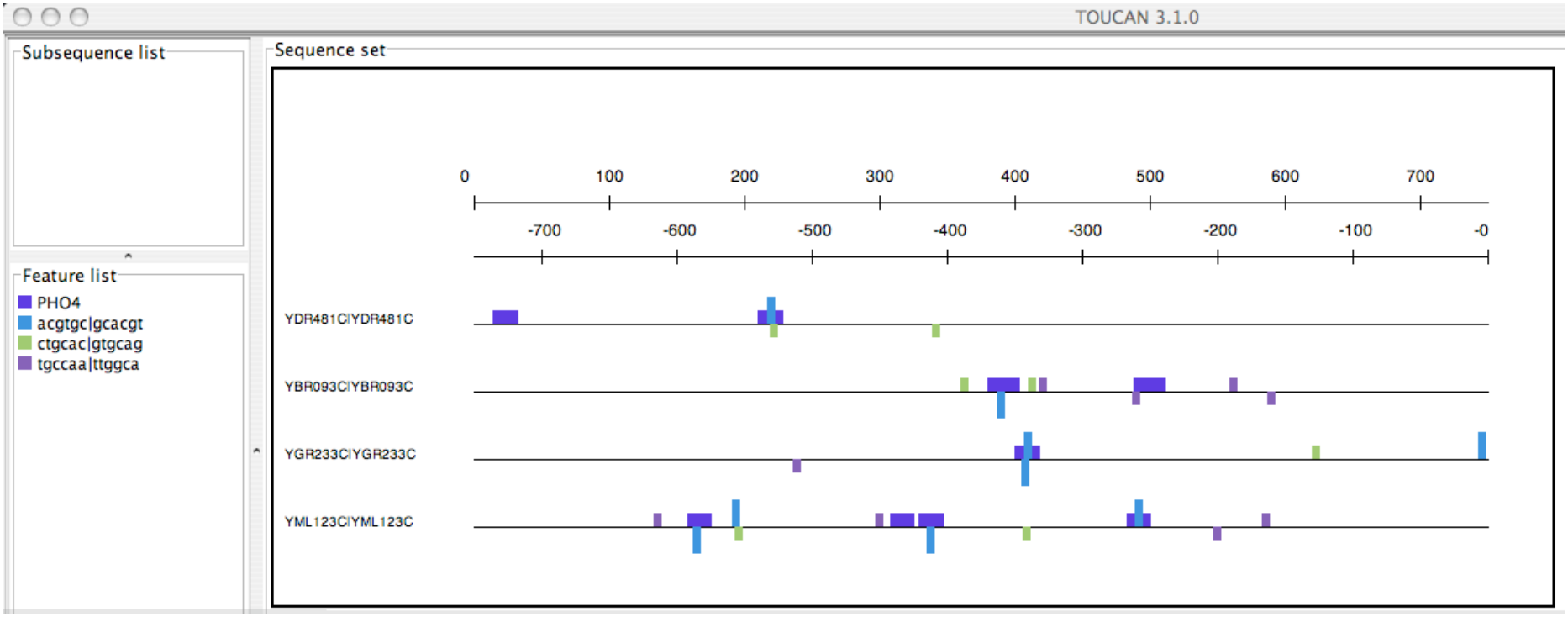
Family	Hexanucleotide analysis result: Sequence	Ms	Occ	Exp	*sig*	Sites previously characterized: Consensus	Bound factors
NIT	ATAAGA	6	20	6.0	**2.0**		Gln3p, Nillp,
	GATAAG	6	26	3.0	9.1	GATAAG	Gzf3p, Uga43p
	AGATAA	7	17	6.1	0.4		(Zn finger)
	CTGATA	6	11	3.1	0.1		
	CCGCGC	2	6	0.7	0.8	–	–
	CGGCAC	4	6	0.8	0.5	–	–
	ACATCT	4	11	2.9	0.4	–	–
MET	CACGTG	9	26	2.0	**7.0**		Cbflp-Met4p-
	TCACGT	9	19	2.9	**6.1**	TCACGTG	Met28p complex
	GTCACG	6	8	1.4	0.7		(bHLH-bLZ-bLZ)
	TGTGGC	7	10	2.4	0.5		
	CTGTGG	8	11	2.1	**1.6**	AAAACTGTGG	Met31p, Met32p
	ACTGTG	9	12	3.2	0.6		(Zn finger)
	AACTGT	10	17	5.5	0.9		
	ATATAT	19	82	42.3	0.8	–	–
	TATATA	11	80	43.9	0.2		
	GCTTCC	7	12	3.5	0.2	–	–
PHO	CGCACG	5	6	0.5	**1.5**		
	GCACGT	5	10	0.8	**4.4**		Pho4p
	CACGTG	5	12	0.9	**1.8**	GCACGTGGG	(bHLH)
	ACGTGG	5	8	0.7	**2.8**	(high affinity)	
	CGTGGG	3	5	0.5	0.5		
	CACGTT	5	7	1.2	0.3	GCACGTTTT	Pho4p
	ACGTTT	5	11	2.6	0.8	(medium affinity)	(bHLH)
	CTGCAC	4	8	1.0	**1.7**	–	–

van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 1998 Sep 4;281(5):827-42.

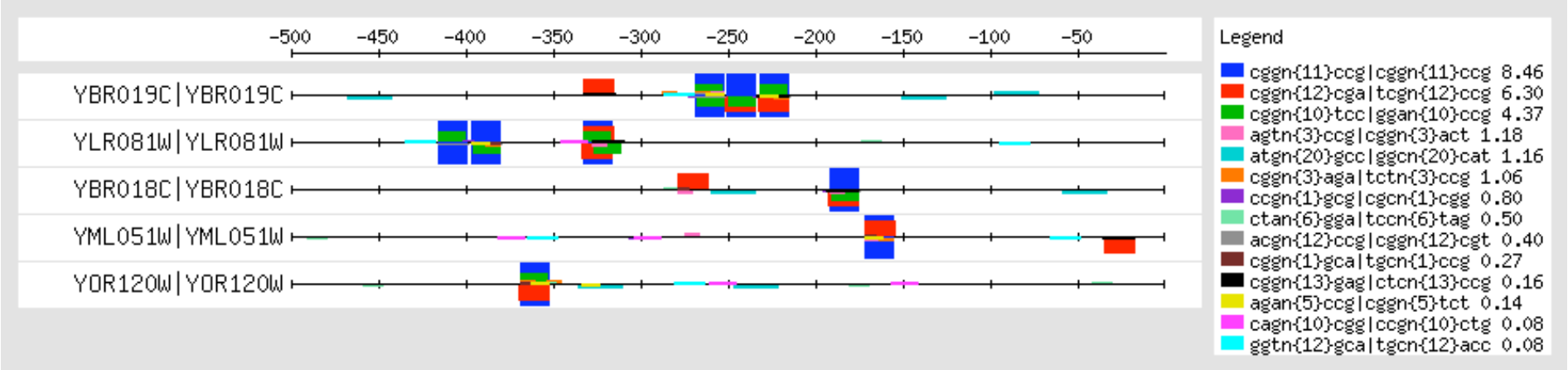
Gal4

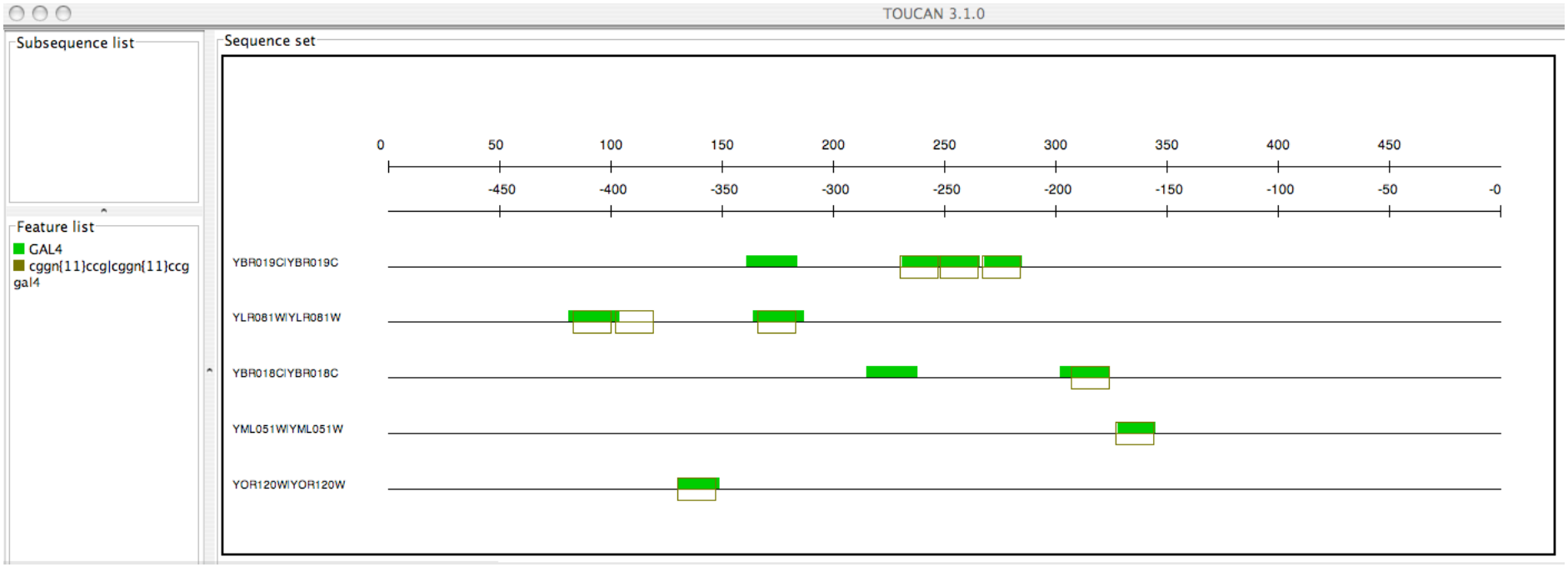
No over-represented hexamers found.

Pho4


TOUCAN 3.1.0
Subsequence list
Feature list
PHO4
acgtgc|gcacgt
ctgcac|gtgcag
tgccaa|ttggca
Sequence set
0
100
200
300
400
500
600
700
-700
-600
-500
-400
-300
-200
-100
-0
YDR481C|YDR481C
YBR093C|YBR093C
YGR233C|YGR233C
YML123C|YML123C

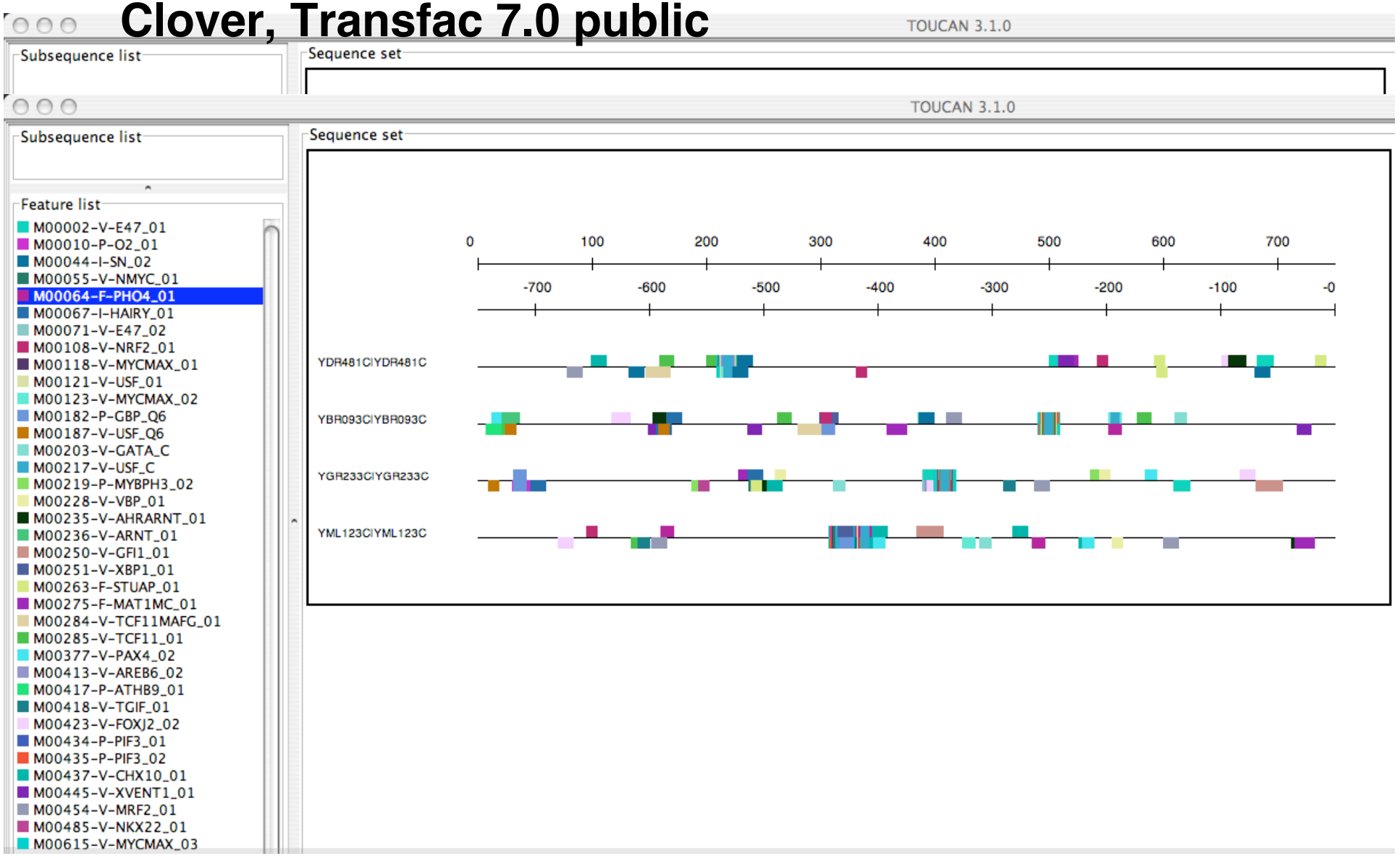
Dyad-analysis, Gal4





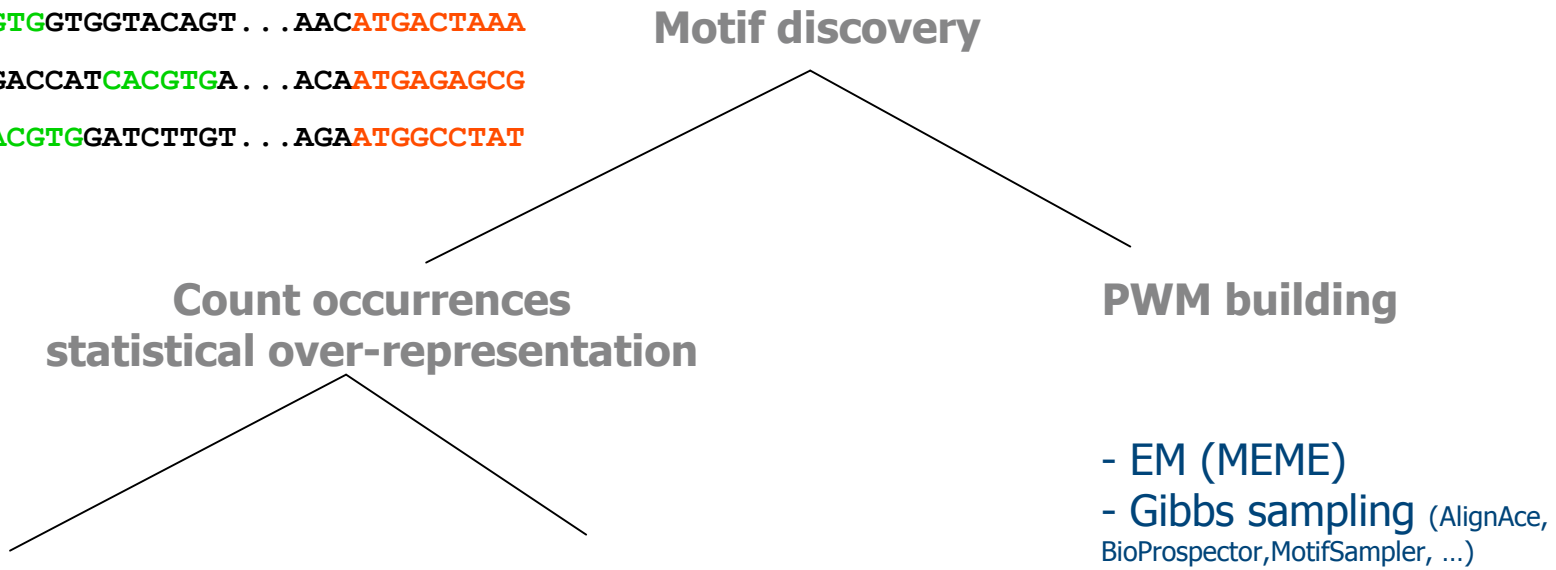
PWM over-representation

Clover, Transfac 7.0 public



Motif discovery

```
GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC
CACATCGCATCACGTGACCAGT...GACATGGACGGC
GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA
TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG
CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT
```



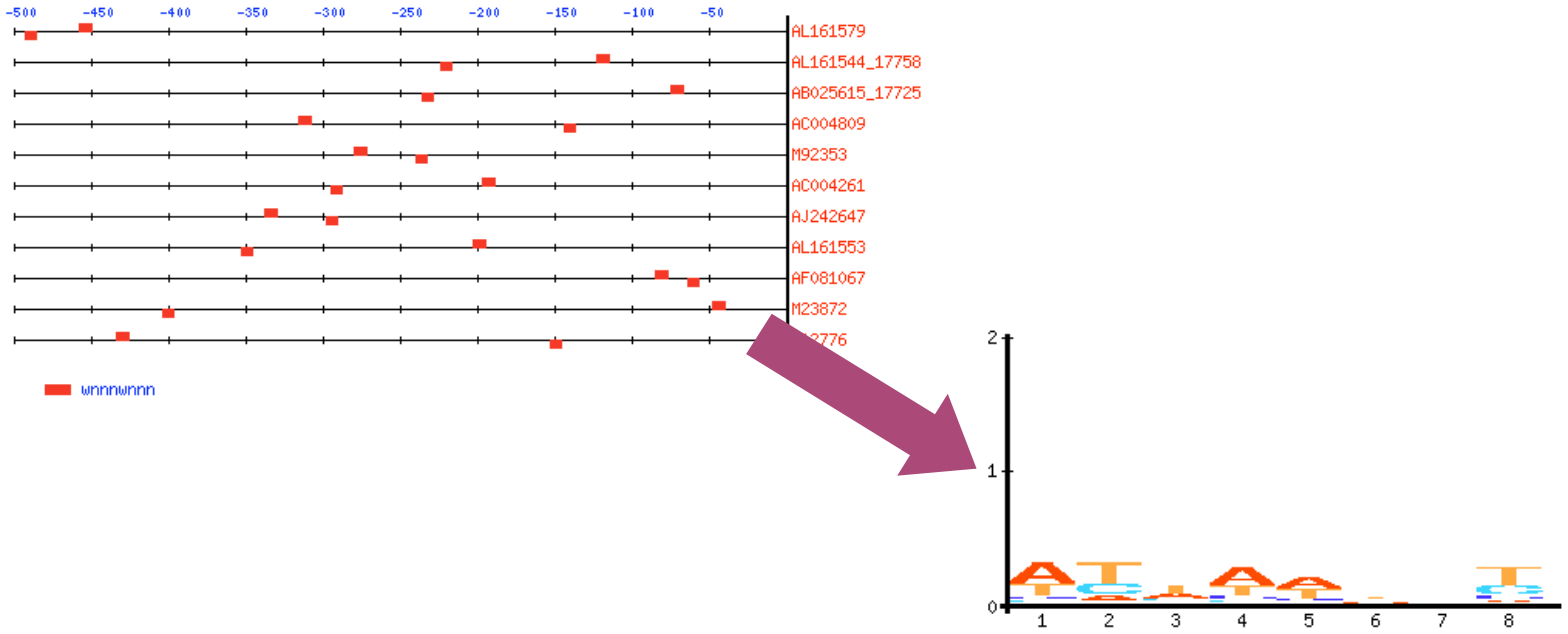
van Helden J, et al.. J Mol Biol. 1998 4;281(5):827-42.

Aerts S et al. Nucleic Acids Res. 2003 15;31(6):1753-64.

- *Lawrence CE, et al. Science. 1993 8;262(5131):208-14.*
- *Bailey T et al. Proceedings of ISMB, pp. 28-36, 1994.*

Gibbs sampling

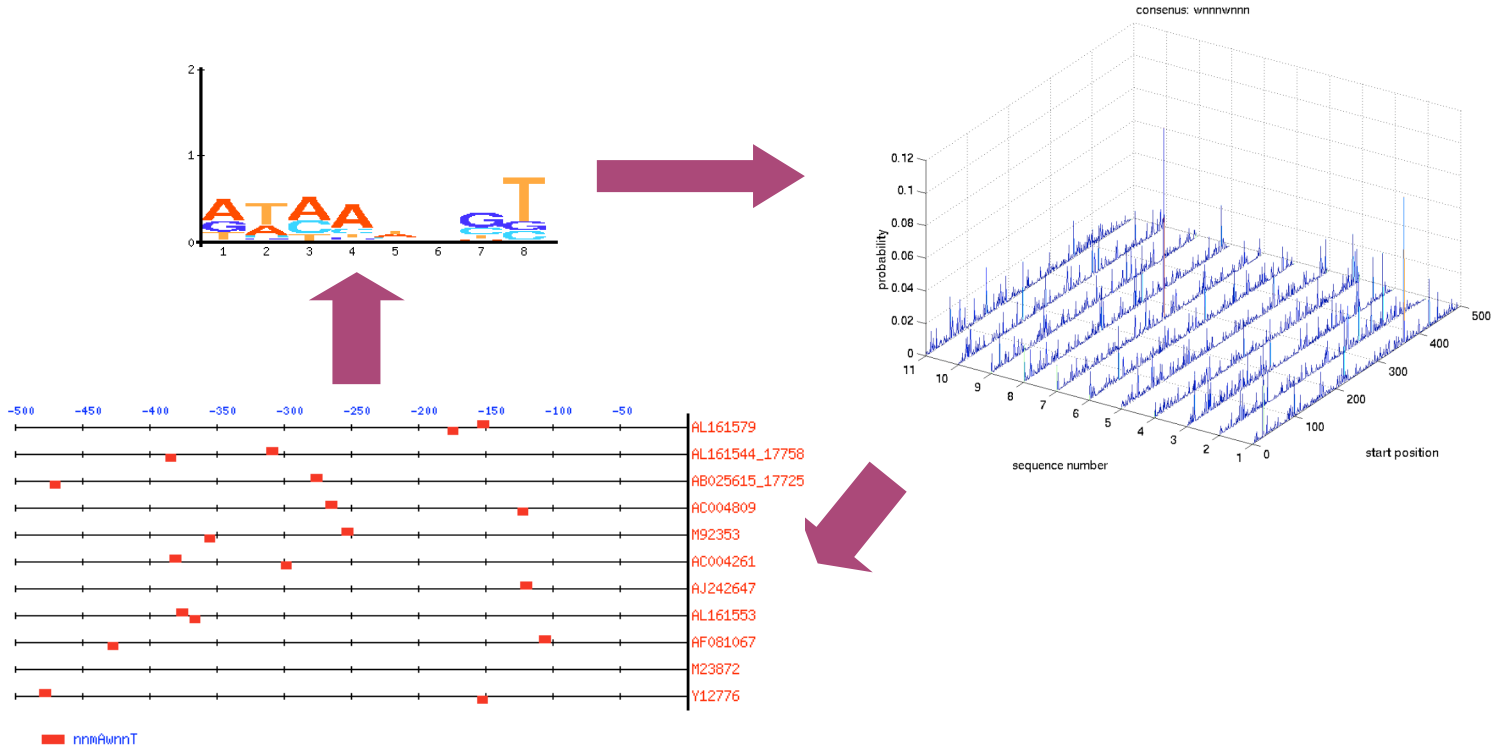
- Start from random set of motif positions
- Create initial motif model



Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol. 2002;9(2):447-64.

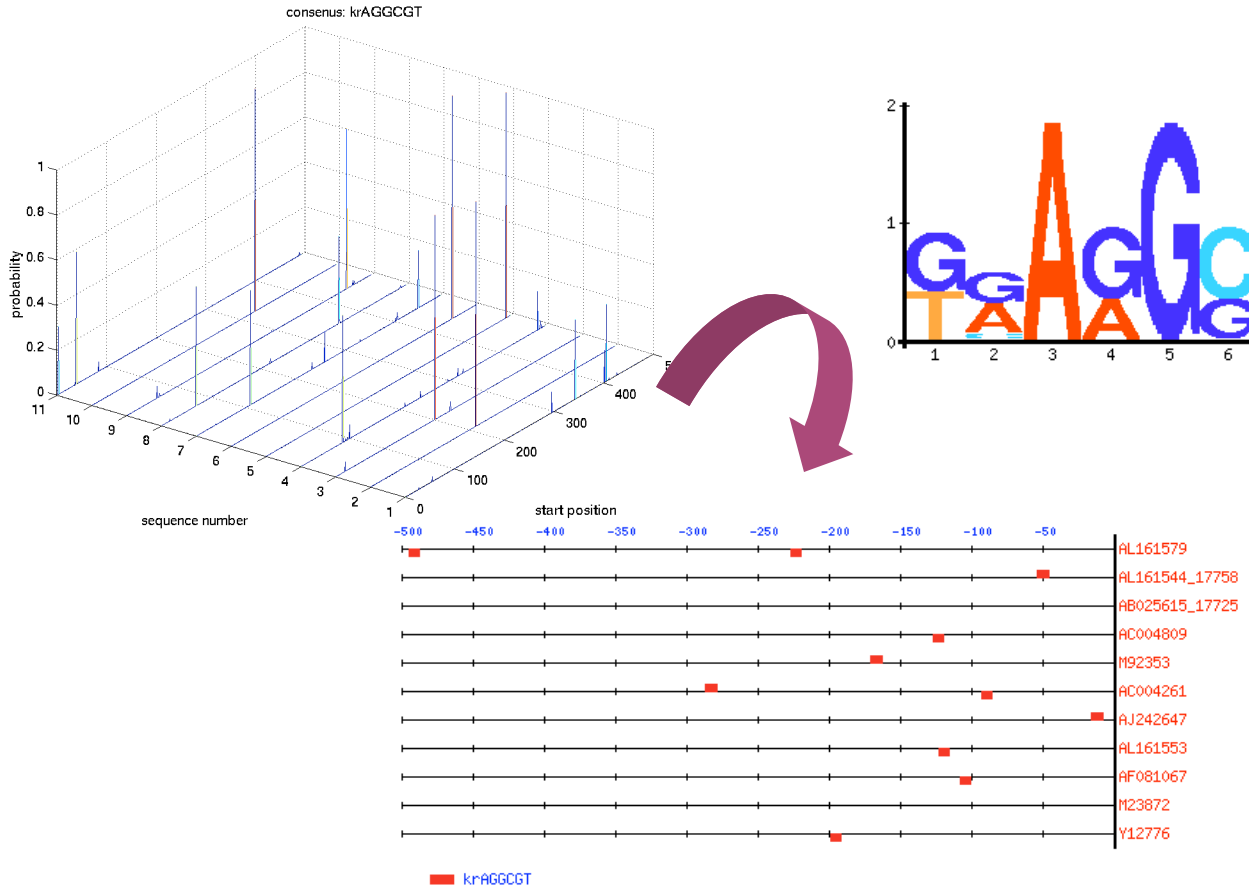
Gibbs sampling (ctd.)

1. Score sequences with current motif model and divide by score of the background model
2. Calculate distribution = probability of position being motif start
3. Sample new alignment position
4. Iterate for fixed number of steps



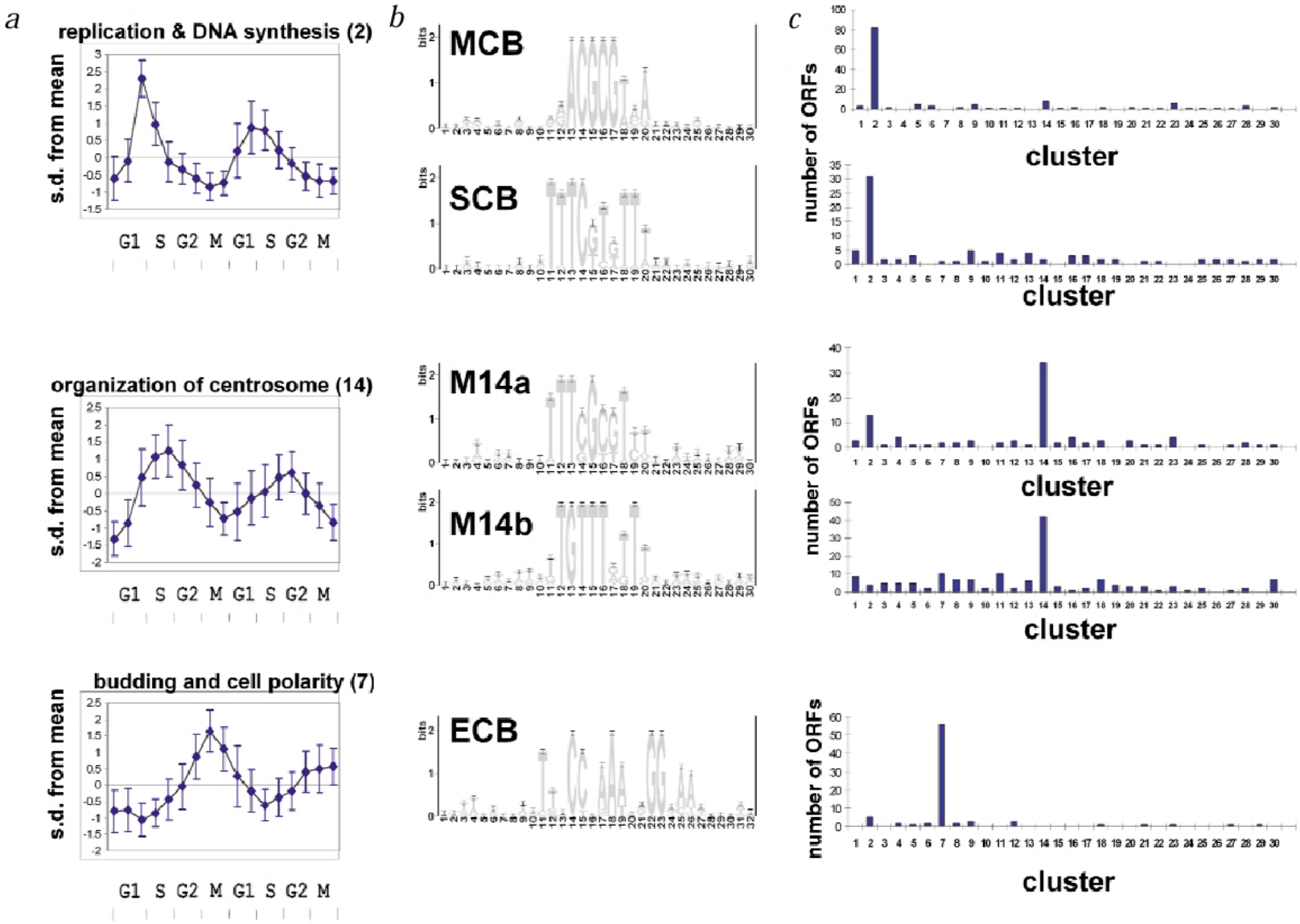
Gibbs sampling (ctd.)

Select best scoring positions from probability distribution to create motif and alingment



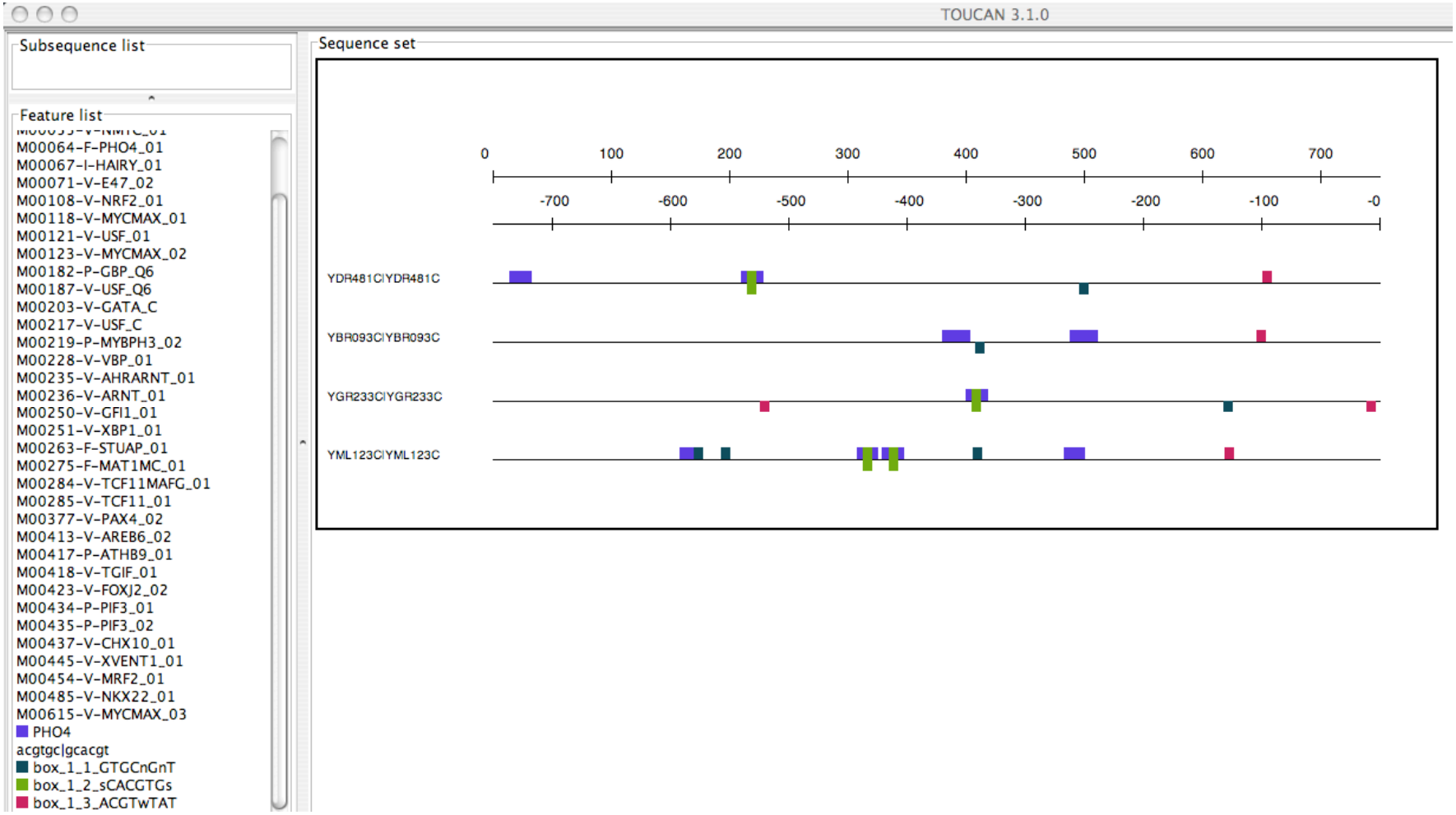
Variations: background models, gapped motifs, ChIP-chip data (MDScan)

Gibbs sampling example



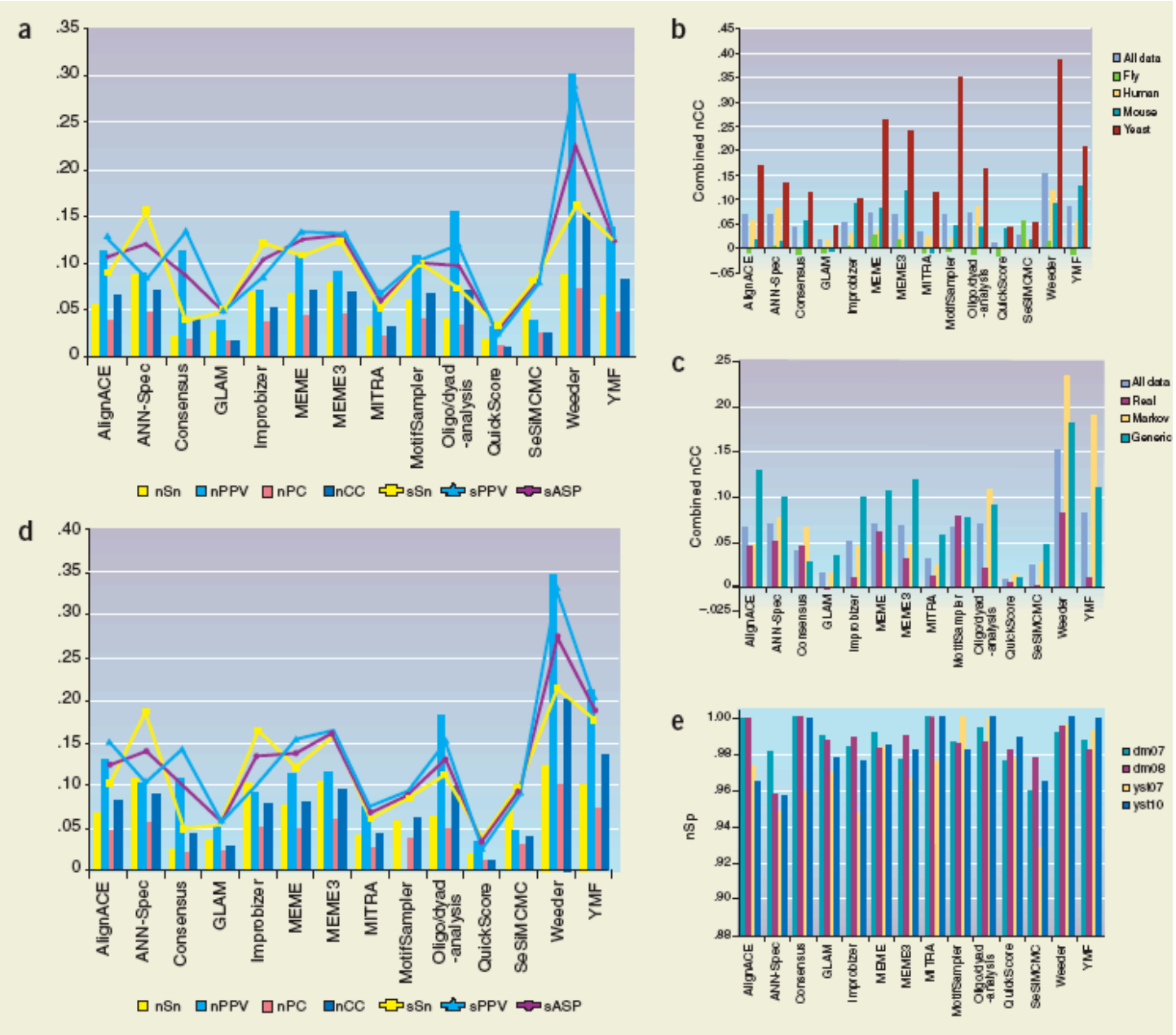
Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999 Jul;22(3):281-5.

PHO4


TOUCAN 3.1.0
Subsequence list
Feature list
M00064-F-PHO4_01
M00067-I-HAIRY_01
M00071-V-E47_02
M00108-V-NRF2_01
M00118-V-MYCMAX_01
M00121-V-USF_01
M00123-V-MYCMAX_02
M00182-P-GBP_Q6
M00187-V-USF_Q6
M00203-V-GATA_C
M00217-V-USF_C
M00219-P-MYBPH3_02
M00228-V-VBP_01
M00235-V-AHRARNT_01
M00236-V-ARNT_01
M00250-V-GFI1_01
M00251-V-XBP1_01
M00263-F-STUAP_01
M00275-F-MAT1MC_01
M00284-V-TCF11MAFG_01
M00285-V-TCF11_01
M00377-V-PAX4_02
M00413-V-AREB6_02
M00417-P-ATHB9_01
M00418-V-TGIF_01
M00423-V-FOXJ2_02
M00434-P-PIF3_01
M00435-P-PIF3_02
M00437-V-CHX10_01
M00445-V-XVENT1_01
M00454-V-MRF2_01
M00485-V-NKX22_01
M00615-V-MYCMAX_03
PHO4
acgtgc|gcacgt
box_1_1_GTGCnGnT
box_1_2_sCACGTGs
box_1_3_ACGTwTAT
Sequence set
0
100
200
300
400
500
600
700
-700
-600
-500
-400
-300
-200
-100
-0
YDR481C|YDR481C
YBR093C|YBR093C
YGR233C|YGR233C
YML123C|YML123C

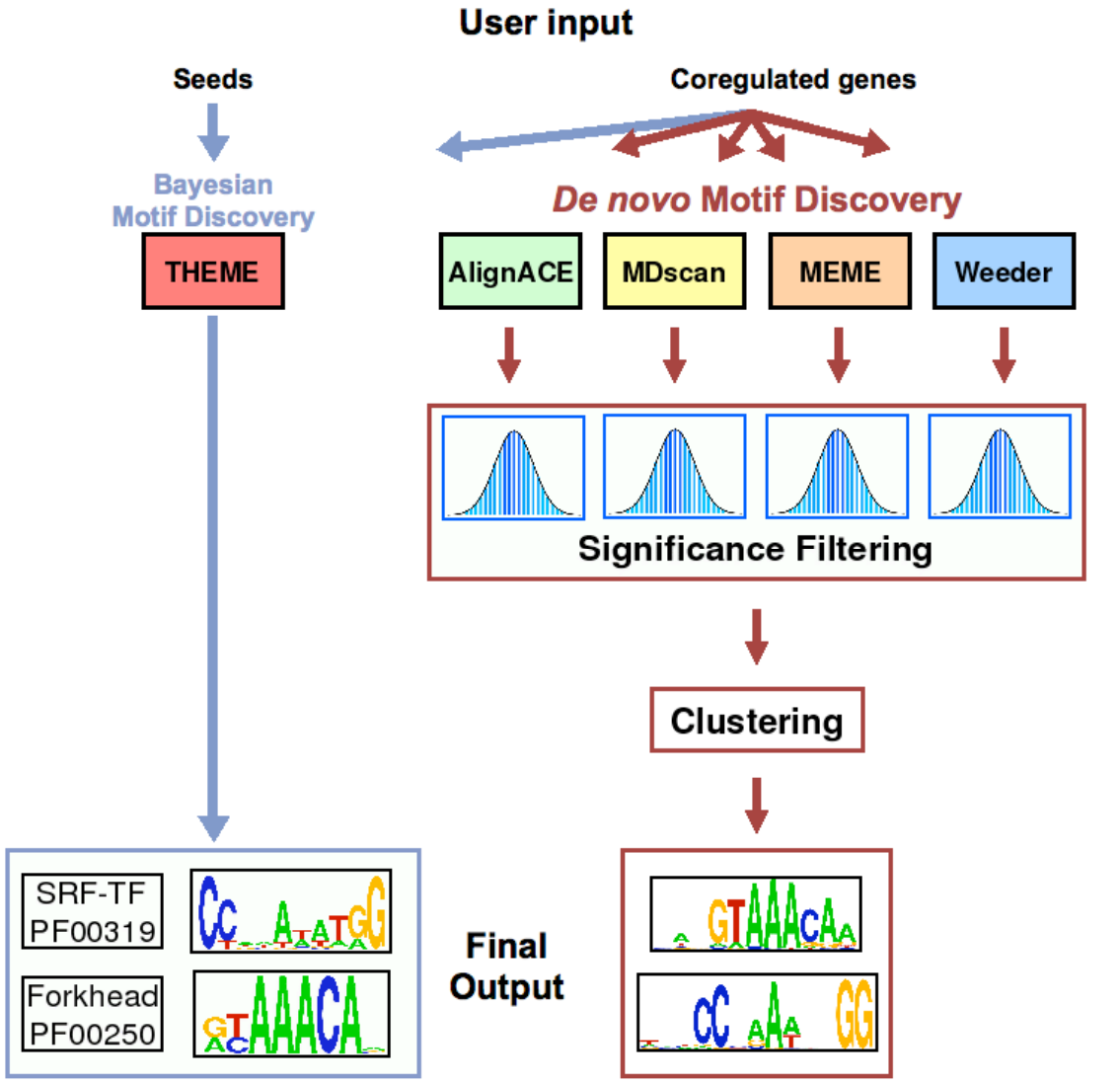
Assessing tools for motif discovery

Tompa et al. *Nature Biotechnology* (2005) 23:1



Combining multiple programs

http://fraenkel.mit.edu/webmotifs/



Use motif discovery to build PWM



Harbison et al

Letters to Nature

Nature **431**, 99-104 (2 September 2004) | doi:10.1038/nature02800; Receive
2004

Transcriptional regulatory code of a eukaryotic genome

Christopher T. Harbison[1,2,5], D. Benjamin Gordon[1,5], Tong Ihn Lee[1], Nicola J. Rinaldi[1,2], Kenzie D. Macisaac[3], Timothy W. Danford[3], Nancy M. Hannett[1], Jean-Bosco Tagne[1], David B. Reynolds[1], Jane Yoo[1], Ezra G. Jennings[1], Julia Zeitlinger[1], Dmitry K. Pokholok[1], Manolis Kellis[1,3,4], P. Alex Rolfe[3], Ken T. Takusagawa[3], Eric S. Lander[1,2,4], David K. Gifford[3,4], Ernest Fraenkel[1,3] & Richard A. Young[1,2,4]

1. Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA
2. Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
3. MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, Massachusetts 02139, USA
4. Broad Institute, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA
5. These authors contributed equally to this work

Harbison et al

- ChIP for 203 proteins
  - Rich media
  - For 84 regulators: at least 1 of 12 other environmental conditions
- 147 regulators that bind more than 10 probes
  - PWM from 6 motif discovery methods, clustering, statistical tests
  - Significant motifs for 116 regulators
  - Single motif for 65 regulators
    - 17 previously had motif
    - For 21 there was no prior specificity information
  - Supplement with other motifs from the literature that also pass the conservation test: 102 regulators
- Regulatory map: motifs bound by regulators at high confidence AND conserved among sensu stricto Sacch. Species.

Using sequence conservation in motif discovery

- Converge
  - Static alignment
  - Conservation in joint probability model (EM)
- PhyloCon
  - Produces many local alignments
  - These are assembled using greedy algorithm
  - patterns that are both conserved in orthologous genes and present in many of the co-regulated promoters
- PhyME
- orthoMEME
- PhyloGibbs
- EMnEM
- FootPrinter (not alignment based)

B

Alignment of conserved regions

```
YGL125W                                                  LEU3
S. cerevisiae    GAAAAAATAACAGCGACTTTTCTCCCGGTAGCGGGCCGTCGTTTAGTCATTCTATCCCTC
S. mikatae       AAAACATAACAGCGAATTTTCCTCCCGGTAGCGGGCCTTCGTTTAGTCATTCTCTCTCTT
S. bayanus       AAAAAATAACAGCGACTTTTCCCCCCGGTAGCGGGCCGTCGTTTAGTCATTCTCTCTCCC
S. kudriavzevii  GAAAAAAAACAACGGCGGCCTCCCCCGGTAGCGGGCCGTCGTTTAGTCATTCTCTCTCTC
                  ***** **** **   *** * *********************************

YOR108W                                  LEU3
S. cerevisiae    GCCATCATGGTCCGGTAACGGTCGTAGTGAATGACTCATATTTTTCCATCTCTTT
S. mikatae       GCCATCAAGGTCCGGTAACGGTCGTAGTGAATGACTCACATTTTCTTCGTTATTC
S. bayanus       ACCATTACGGTCCGGTAACGGACTTAGTGAATGATTCATCTTTTCTTCTTTTTTC
S. kudriavzevii  GTCGTTAAGGTCCGGTAACGGCCCTCAGCGAATGATTCATAATTTCATTTTTTTC
                 ***** * ***********  * **********  ***  ****    *** ***

YMR108W                                          LEU3
S. cerevisiae    AACGCCTAGCCGCCGGAGCCTGCCGGTACCGGCTTGGCTTCAGTTGCTGATCTCGG
S. mikatae       CACAATGACACATACCTAACAGCCGGTACCGGCTTGAATGCCGCCGTTGGCTTCGG
S. bayanus       ATCTTCTAGTCACCGCAGTCTGCCGGTACCGGCTTGAATTCCGCCGTTGATCCTGG
S. kudriavzevii  CACATCTCTAGTCCGCGCTCTGCCGGTACCGGCTTAGACTAGCCACGAATCTCGGC
                  **  ***  * ****   ***********  ****  *** ** * **     **
```

Alignment of profiles

| | | |

Use sequence conservation in discovery (phylocon)

Research article

Highly accessed | Open Access

An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*

Kenzie D MacIsaac* [1,2], **Ting Wang*** [3], **D Benjamin Gordon**[1,4], **David K Gifford**[1,2], **Gary D Stormo**[3] and **Ernest Fraenkel**[1,2,5]

1 Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA
2 MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar St., Cambridge, MA 02139, USA
3 Department of Genetics, Washington University Medical School, St. Louis, MO 63110, USA
4 Agilent Technologies, 245 1st Street, Cambridge, MA 02142, USA
5 Biological Engineering Division, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

author email corresponding author email * Contributed equally

BMC Bioinformatics 2006, **7**:113 doi:10.1186/1471-2105-7-113

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/7/113

Received: 9 September 2005
Accepted: 7 March 2006
Published: 7 March 2006

© 2006 MacIsaac et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

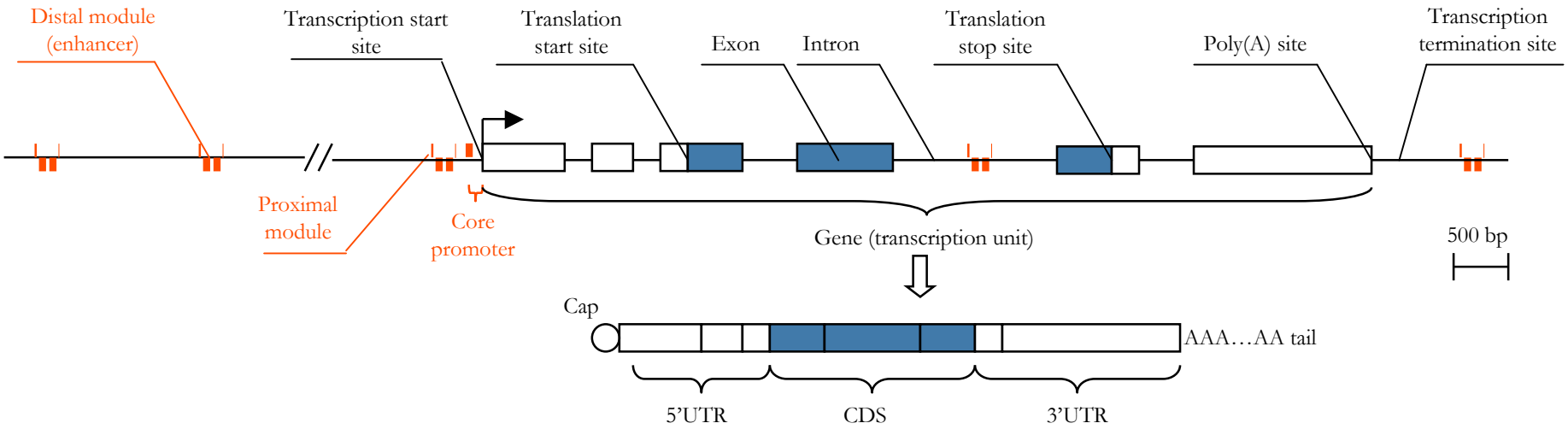
Improve motif prediction

- Use gene expression data
  - E.g., Segal et al regulatory modules
  - REDUCE (regression)
- Nucleosome positioning data

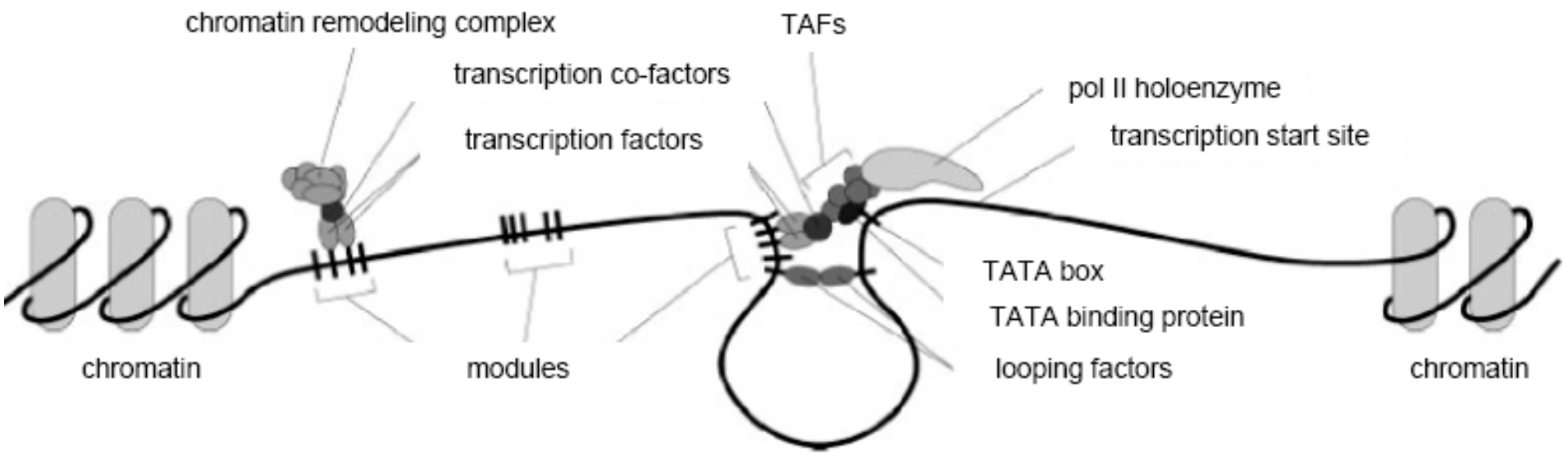
Higher eukaryotes

- Genes expressed at many different times and places during development
  - More signal integration
    - Simultaneous presence of multiple activators, each of which may be controlled by a separate signal
  - More combinatorial control
    - Many activators work in various combinations with other activators (and repressors) to regulate different sets of genes
  - Control by alternative sets of regulators
    - Same gene expressed in response to different sets of signals

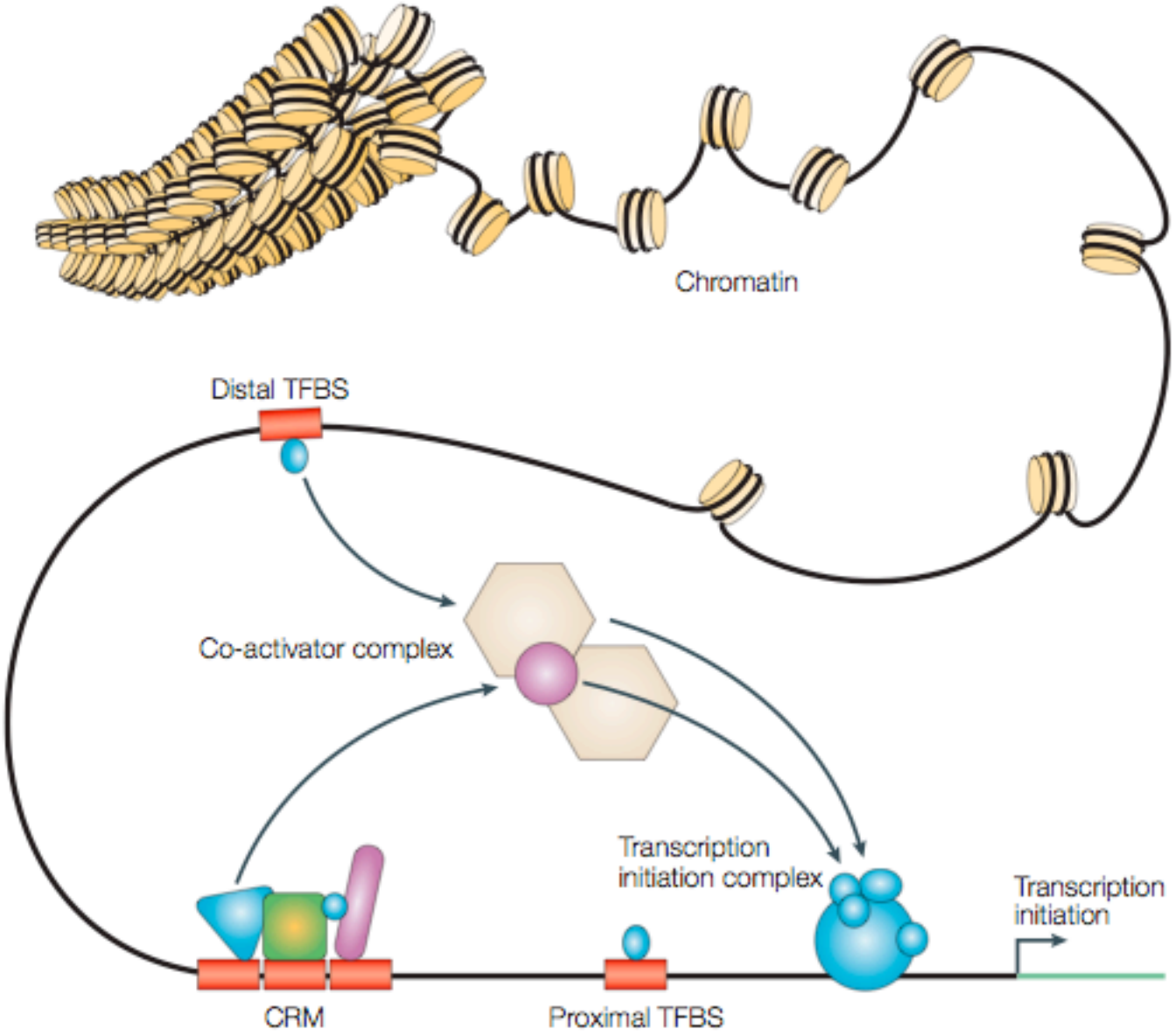
Cis-regulatory system



Cis-regulatory system



Wray et al. 2003


Chromatin
Distal TFBS
Co-activator complex
Transcription
initiation complex
Transcription
initiation
CRM
Proximal TFBS

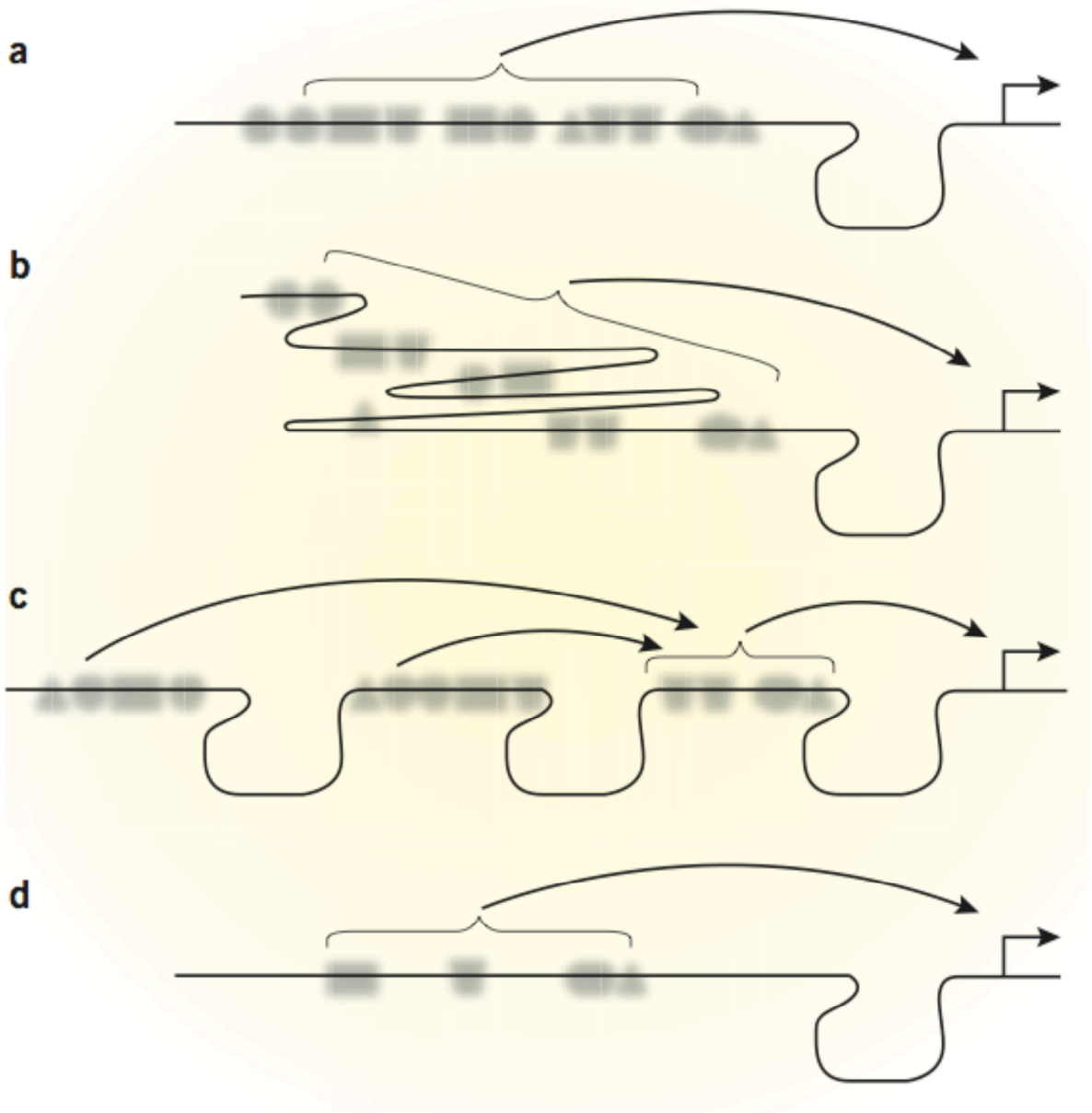
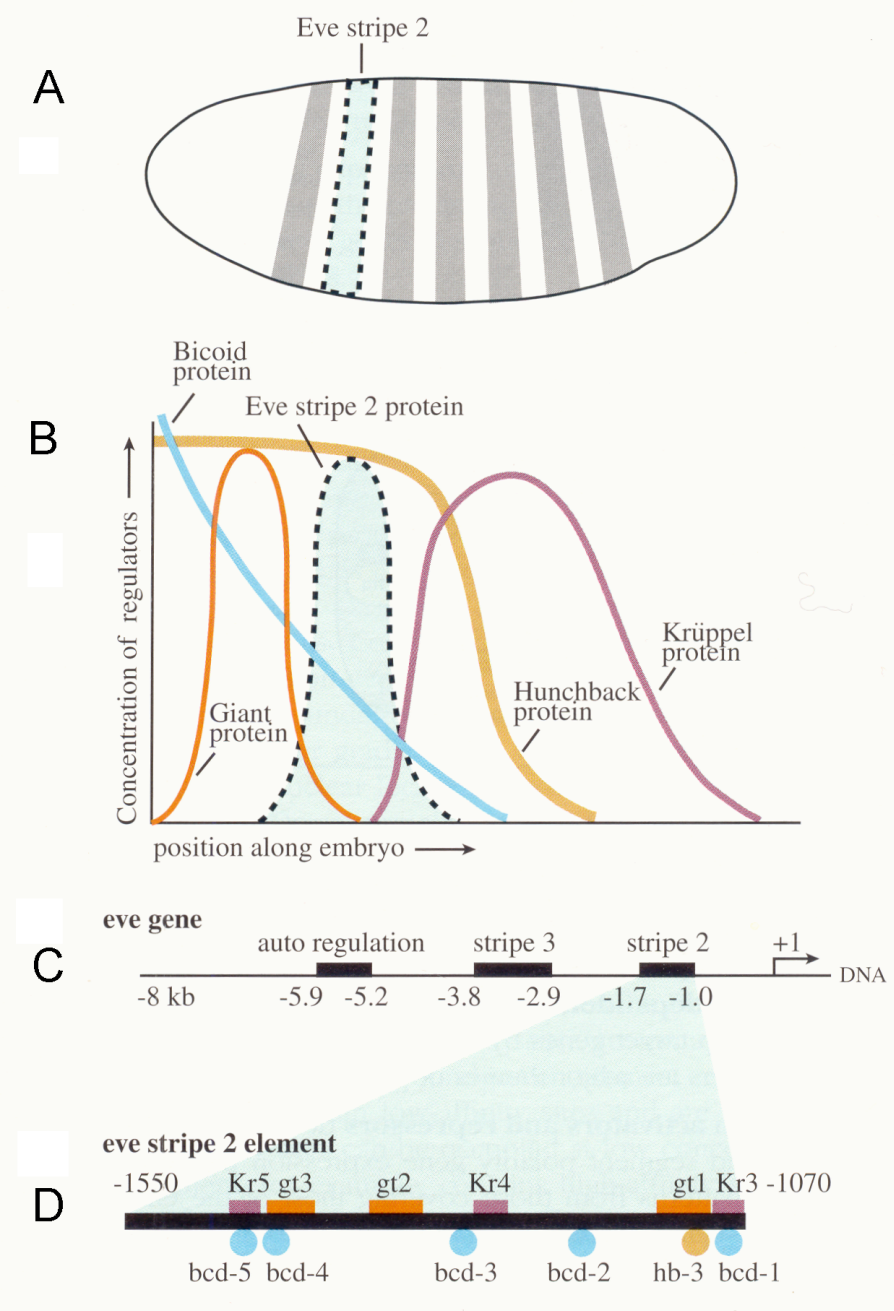


Figure 1 Different arrangements of *cis*-regulatory modules (CRMs). (**a**) 'Typical' CRM containing a compact arrangement of clustered binding sites. (**b**) Diffuse CRM in which binding sites are distributed over a large DNA segment. (**c**) Composite CRM. (**d**) Simple compact CRM without a dense cluster of binding sites.

M. Halfon, *Nat Genet* 2006

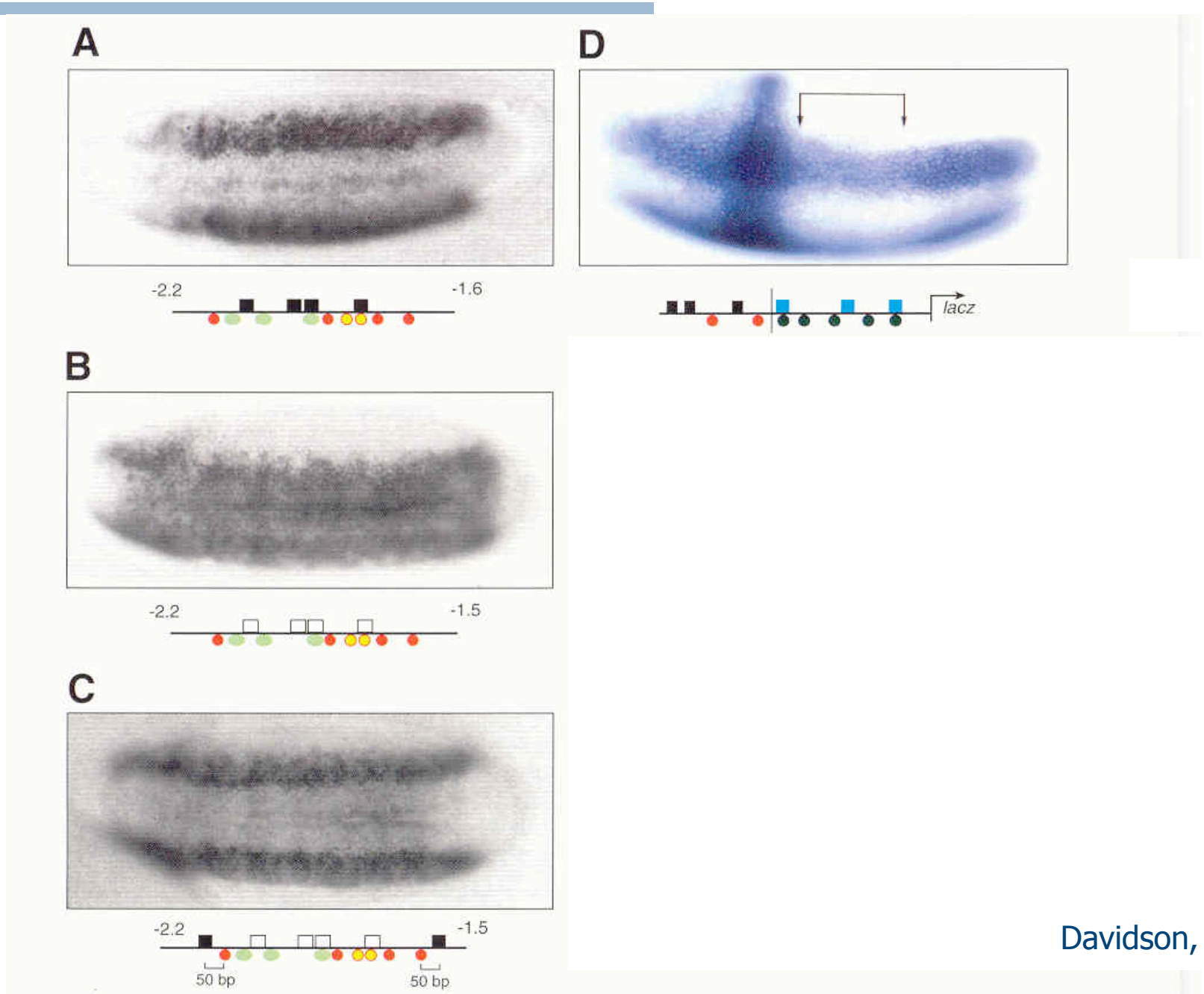
Module example: *eve*



- Features of CRMs:
  - 4-5 different factors per module
  - Length several 100's bp
  - TFs in competition-> TFBS may overlap
  - Multiple binding sites for same factor
  - Often conserved
  - Autonomous
  - Independent of location and orientation

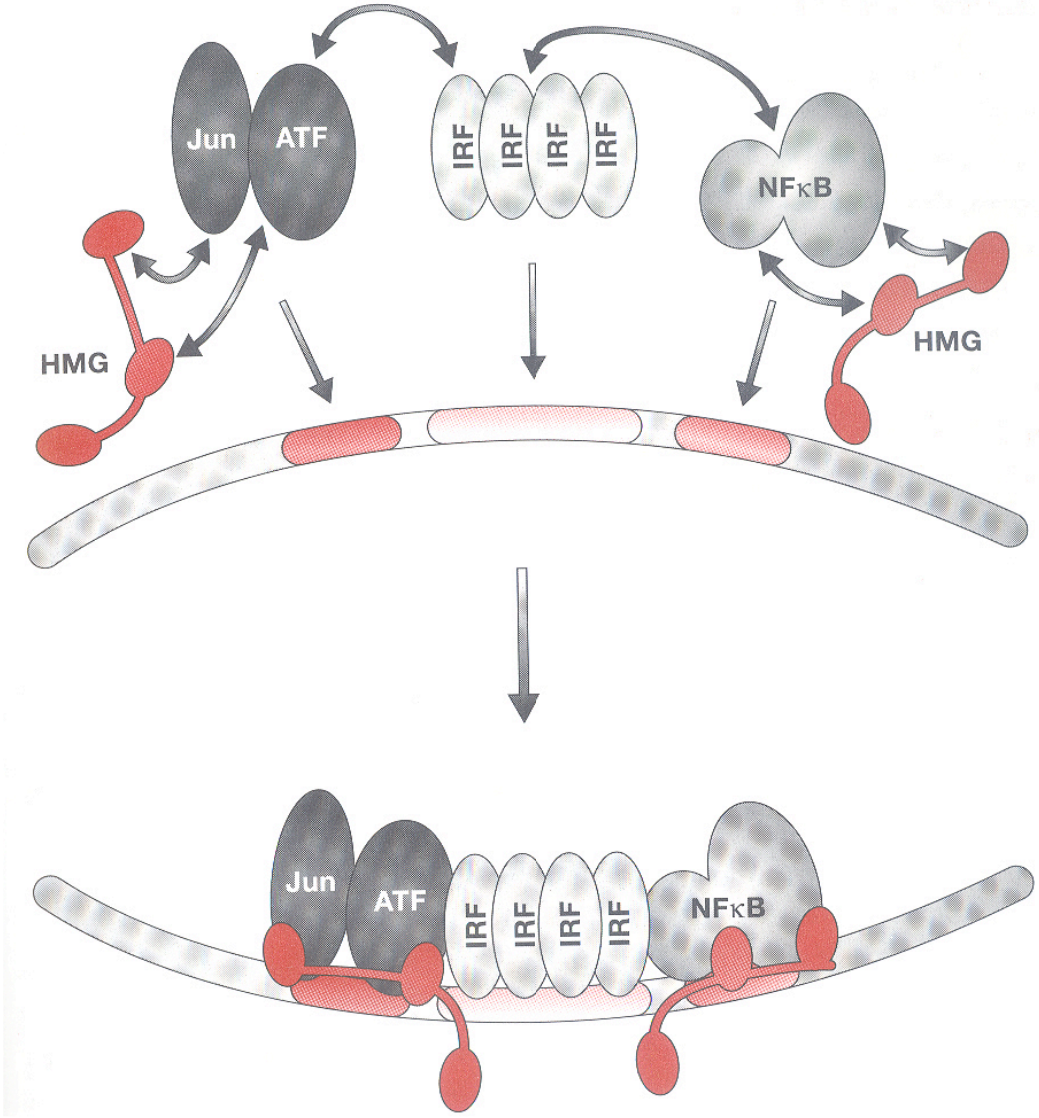
Caroll, 2001

Module example: *rho* and *eve*

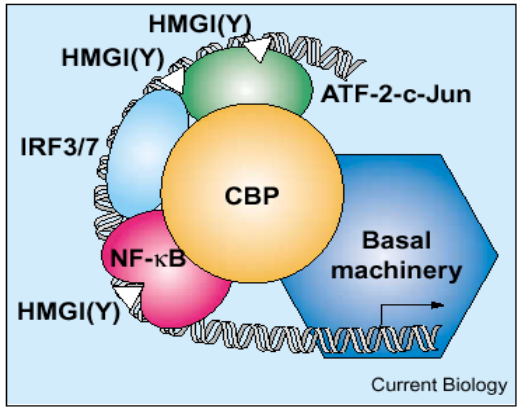


Davidson, 2001

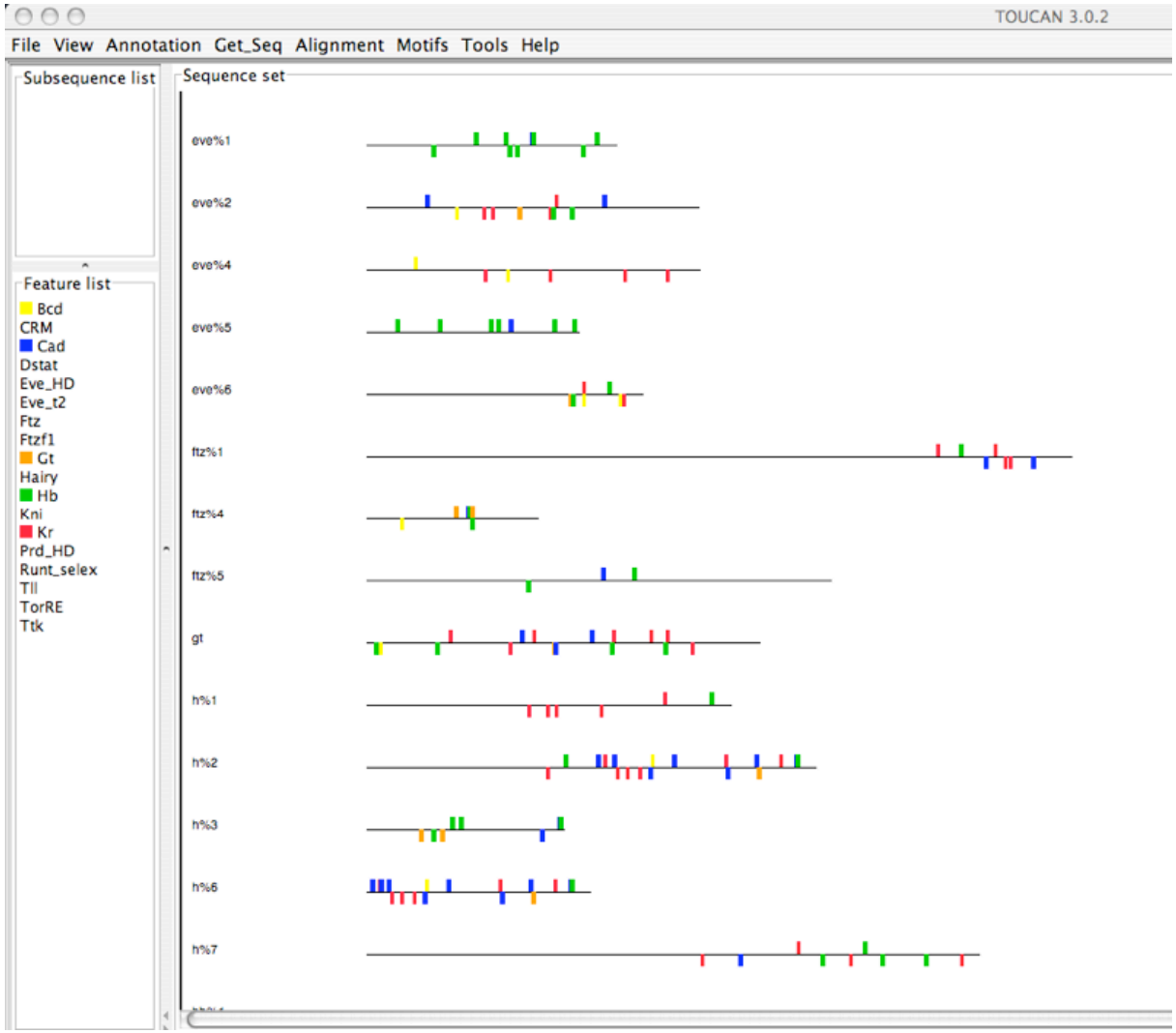
Module example: human interferon beta enhancer



“Enhanceosome” versus “billboard” model



Courey AJ (2001) Curr Biol. 11(7):R250-2



Example of billboard model: Functional conservation despite lack of seq conserv.

OPEN ACCESS Freely available online

PLoS GENETICS

Sepsid *even-skipped* Enhancers Are Functionally Conserved in *Drosophila* Despite Lack of Sequence Conservation

Emily E. Hare[1☯], Brant K. Peterson[1,2☯], Venky N. Iyer[1], Rudolf Meier[3], Michael B. Eisen[1,2,4,5*]

1 Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, California, United States of America, **2** Center for Integrative Genomics, University of California Berkeley, Berkeley, California, United States of America, **3** Department of Biological Sciences, National University of Singapore, Singapore, **4** Genomics Division, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **5** California Institute for Quantitative Biosciences, Berkeley, California, United States of America

Abstract

The gene expression pattern specified by an animal regulatory sequence is generally viewed as arising from the particular arrangement of transcription factor binding sites it contains. However, we demonstrate here that regulatory sequences whose binding sites have been almost completely rearranged can still produce identical outputs. We sequenced the *even-skipped* locus from six species of scavenger flies (Sepsidae) that are highly diverged from the model species *Drosophila melanogaster*, but share its basic patterns of developmental gene expression. Although there is little sequence similarity between the sepsid *eve* enhancers and their well-characterized *D. melanogaster* counterparts, the sepsid and *Drosophila* enhancers drive nearly identical expression patterns in transgenic *D. melanogaster* embryos. We conclude that the molecular machinery that connects regulatory sequences to the transcription apparatus is more flexible than previously appreciated. In exploring this diverse collection of sequences to identify the shared features that account for their similar functions, we found a small number of short (20–30 bp) sequences nearly perfectly conserved among the species. These highly conserved sequences are strongly enriched for pairs of overlapping or adjacent binding sites. Together, these observations suggest that the local arrangement of binding sites relative to each other is more important than their overall arrangement into larger units of *cis*-regulatory function.

Citation: Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid even-skipped Enhancers Are Functionally Conserved in *Drosophila* Despite Lack of Sequence Conservation. PLoS Genet 4(6): e1000106. doi:10.1371/journal.pgen.1000106

Editor: Norbert Perrimon, Harvard Medical School, Howard Hughes Medical Institute, United States of America

Received February 28, 2008; **Accepted** May 22, 2008; **Published** June 27, 2008

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose.

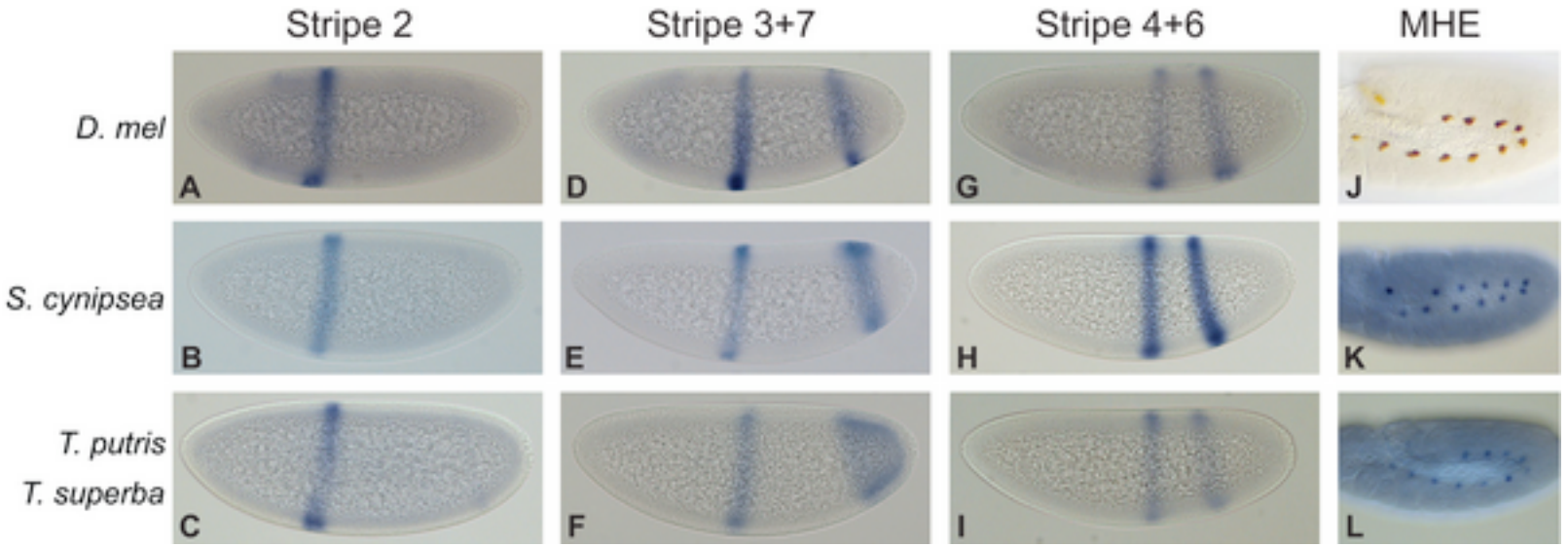
Funding: This work was supported by NIH grant HG002779 to MBE by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by NSF Grant No. 0334948 to RM. EEH was supported by an NSF Graduate Research Fellowship. Sequencing was conducted under a Community Sequencing Grant from the Department of Energy's Joint Genome Institute.

Competing Interests: The authors have declared that no competing interests exist.

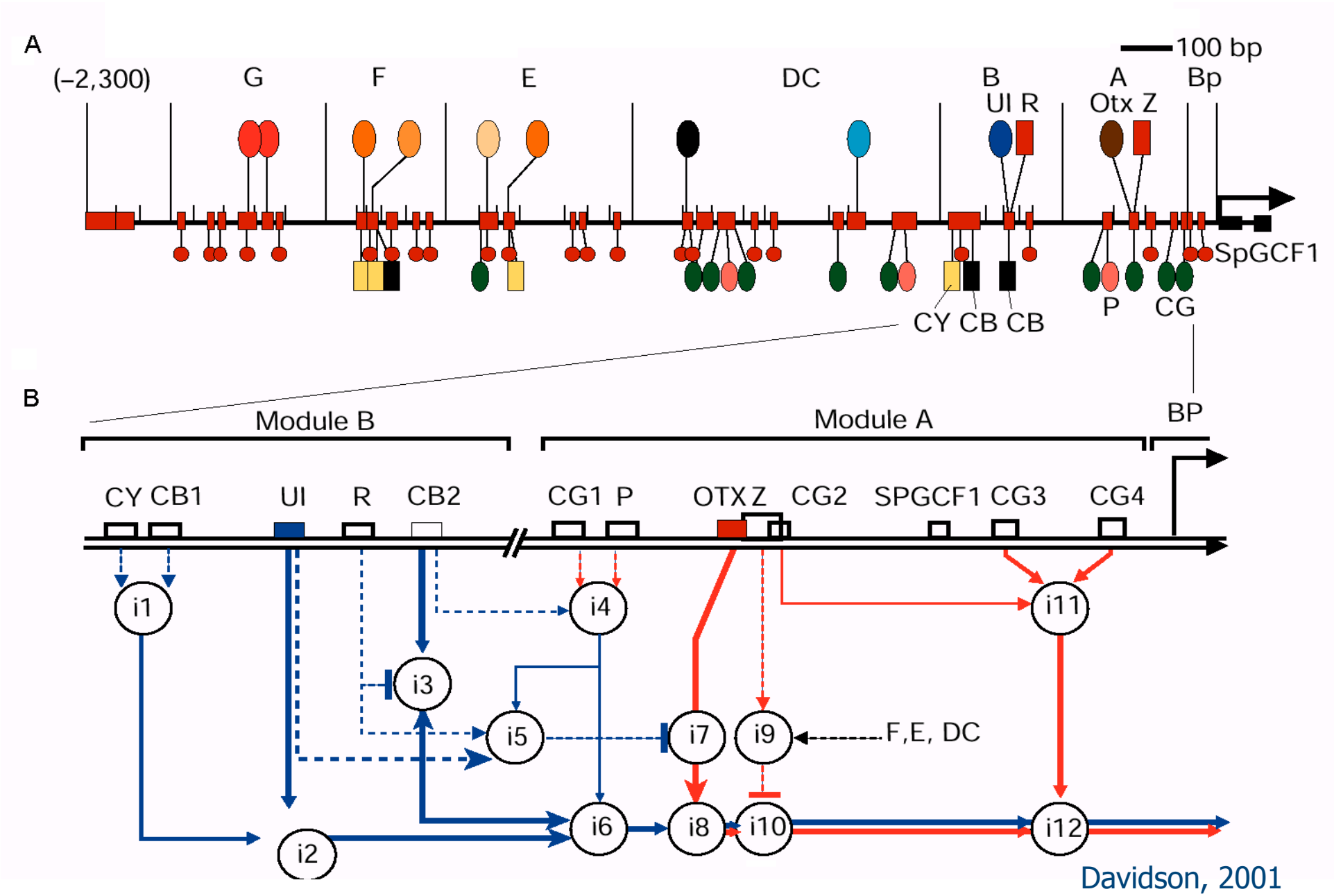
* E-mail: mbeisen@lbl.gov

☯ These authors contributed equally to this work


Stripe 2
0
400
800
1200
1600
2000
2400
2800
3200
3600
FOOTPRINTS
D. mel
D. ere
D. pse
D. vir
D. moj
D. gri
S. punctum
S. cynipsea
D. sp.
T. superba
T. minor
T. putris
*
BCD
ln(p) < -6
HB
ln(p) < -6
GT
ln(p) < -5.5
KR
ln(p) < -6
SLP1
ln(p) < -6

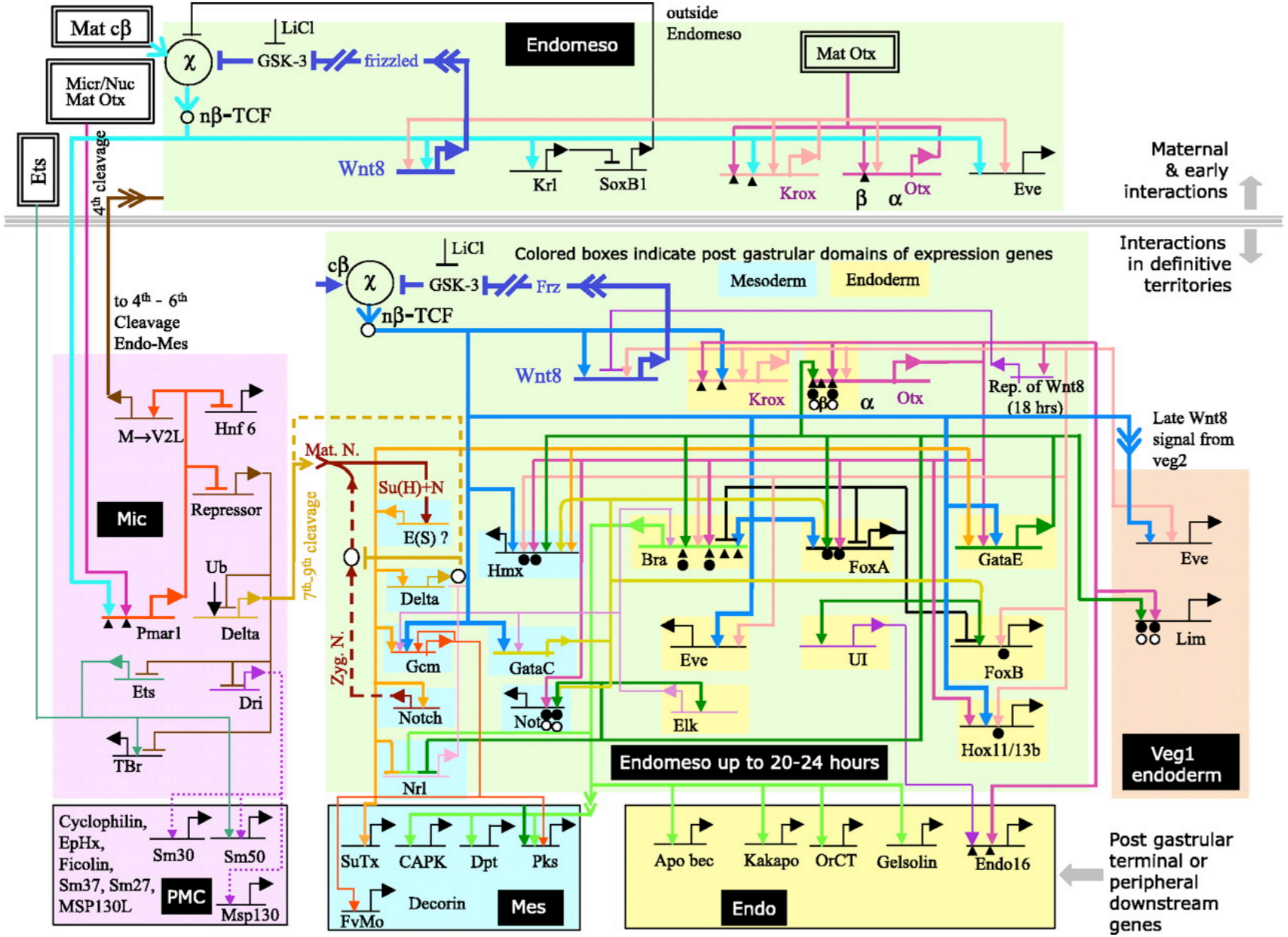

Stripe 2
Stripe 3+7
Stripe 4+6
MHE
D. mel
S. cynipsea
T. putris
T. superba
A
B
C
D
E
F
G
H
I
J
K
L

Example: Endo16 (*S.purpuratus*)

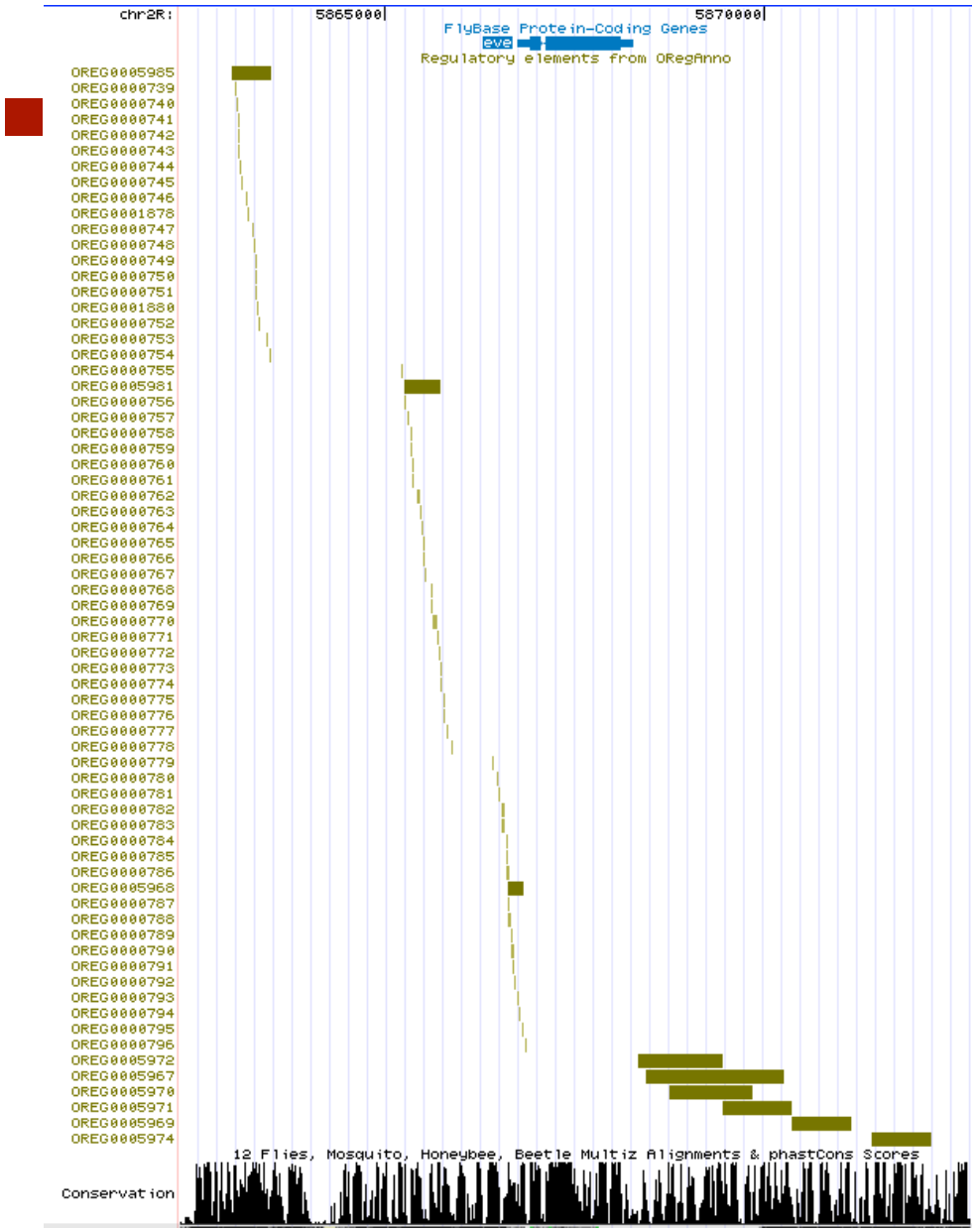


Davidson, 2001

Gene-regulatory network



What is known: cis-regulatory sequences



CRMs

RedFly: 655 CRMs

TFBS

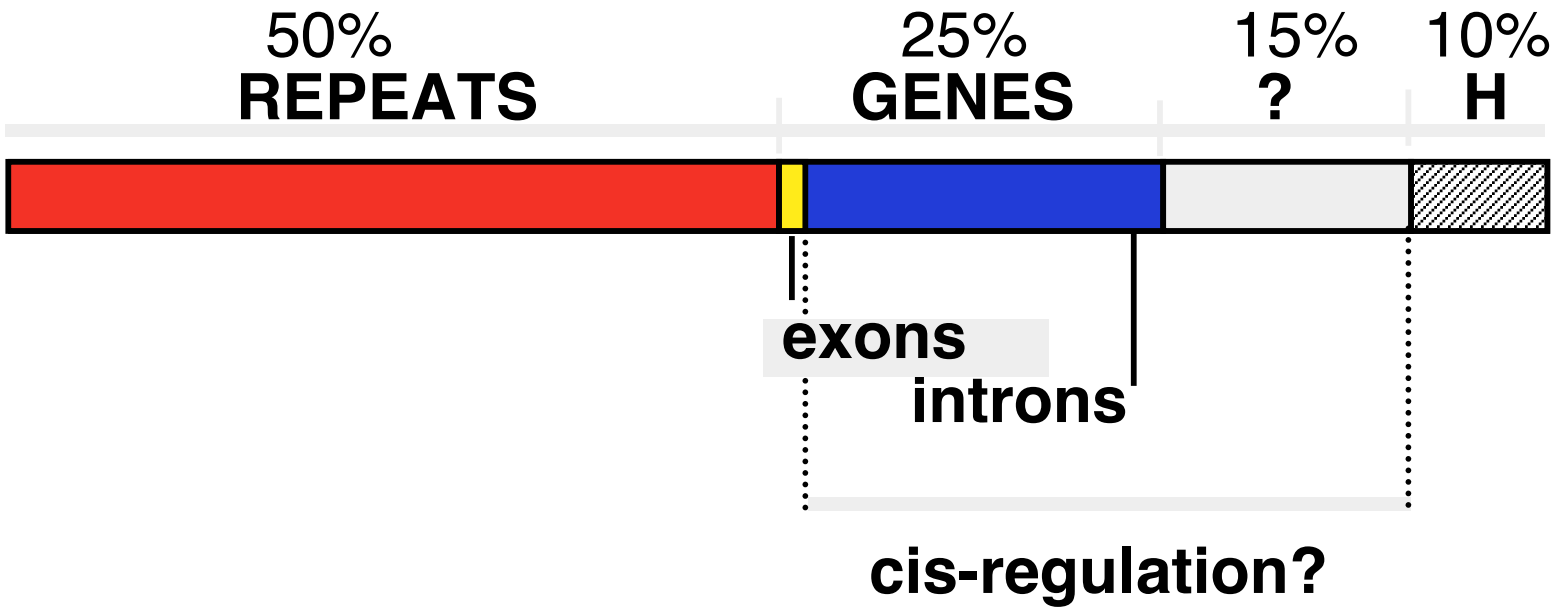
FlyReg: 1344 TFBS

DNase I hypersensitivity

All curated elements

ORegAnno: 2089 records

1-2 % of the human genome encodes proteins



H = largely unsequenced heterochromatin

M. Markstein http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Markstein.ppt

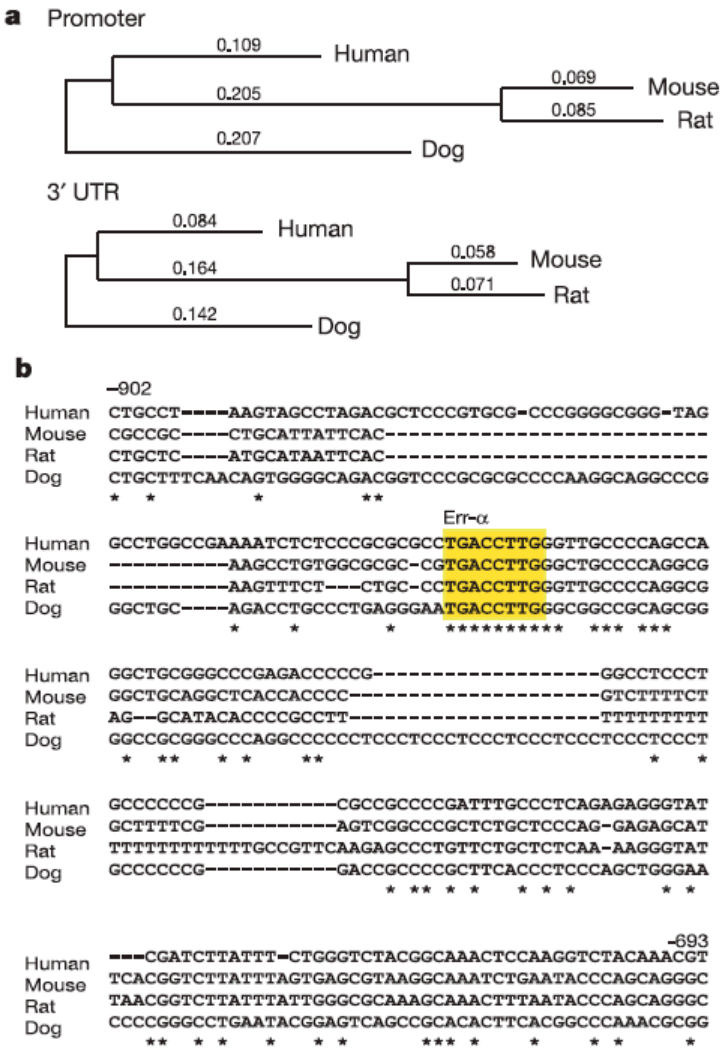
Problems

- Problem: (very) large sequence space and “futility theorem”
  - 1000 fold excess of false positive predictions
- Solutions
  - Analyse only proximal promoters
    - Kellis et al
    - E2F case
  - Use phylogenetic footprinting
  - Use motif combinations
    - Scanning (cfr yeast)
    - discovery
  - Strong statistics with gene expression data (e.g., Lever)
    - Over-representation alone is not sufficient (signal/noise to low)
  - Use ChIP data

Genome-wide analysis of conserved motifs

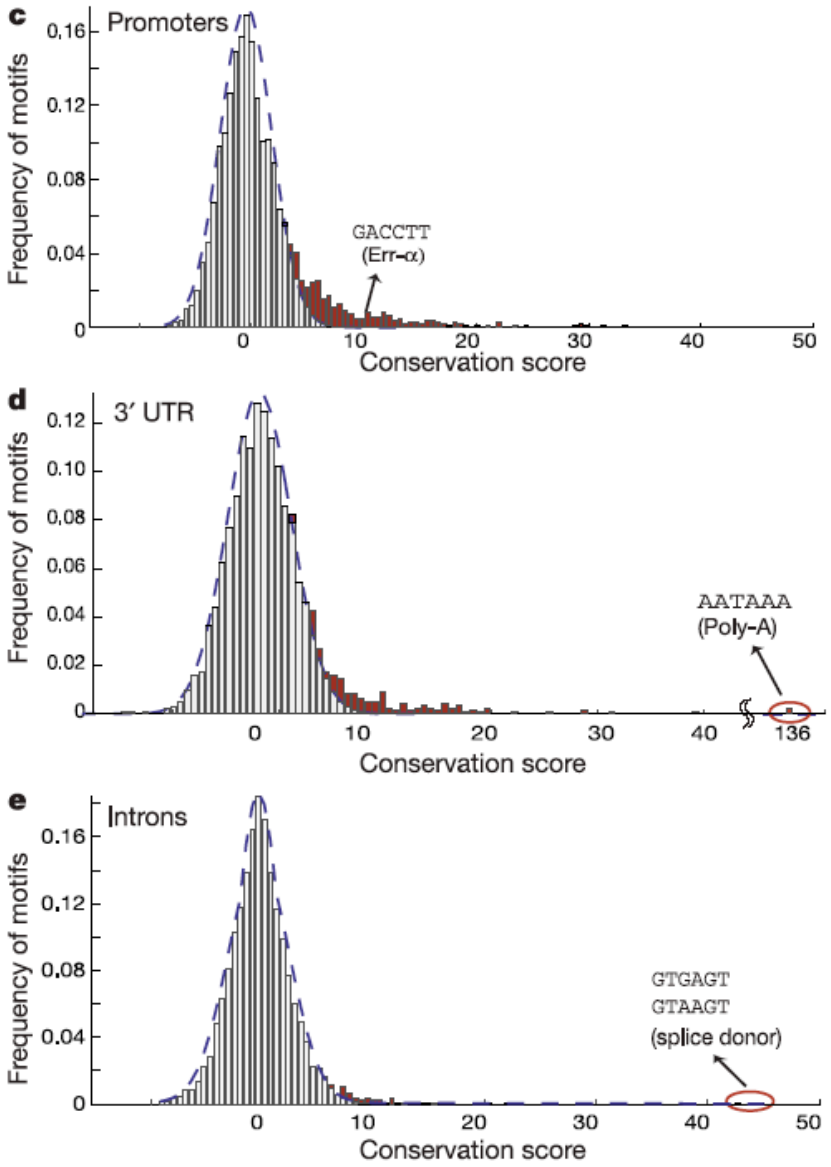
- Combination of
  - Proximal promoters
  - TFBS conservation
- Promoter alignments (4kb centered around annotated TSS)
- Example: Err-α
  - 434 occurrences in human promoters
    - 162 conserved across all 4 species
    - 37% conservation rate
  - Random 8-mer: 6.8% cons. rate
- Motif Conservation Score (MCS)
  - **binomial probability** of observing *k* conserved instances out of *n* given the probability p_0 of conservation for any one instance
  - Err-α: MCS = 25.2
- TRANSFAC motifs
  - 63% MCS>3
  - 50% MCS>5
- Random motifs:
  - 1.6% MCS>3
  - 0% MCS>5

Xie et al. (2005) *Nature* 434, 338-345

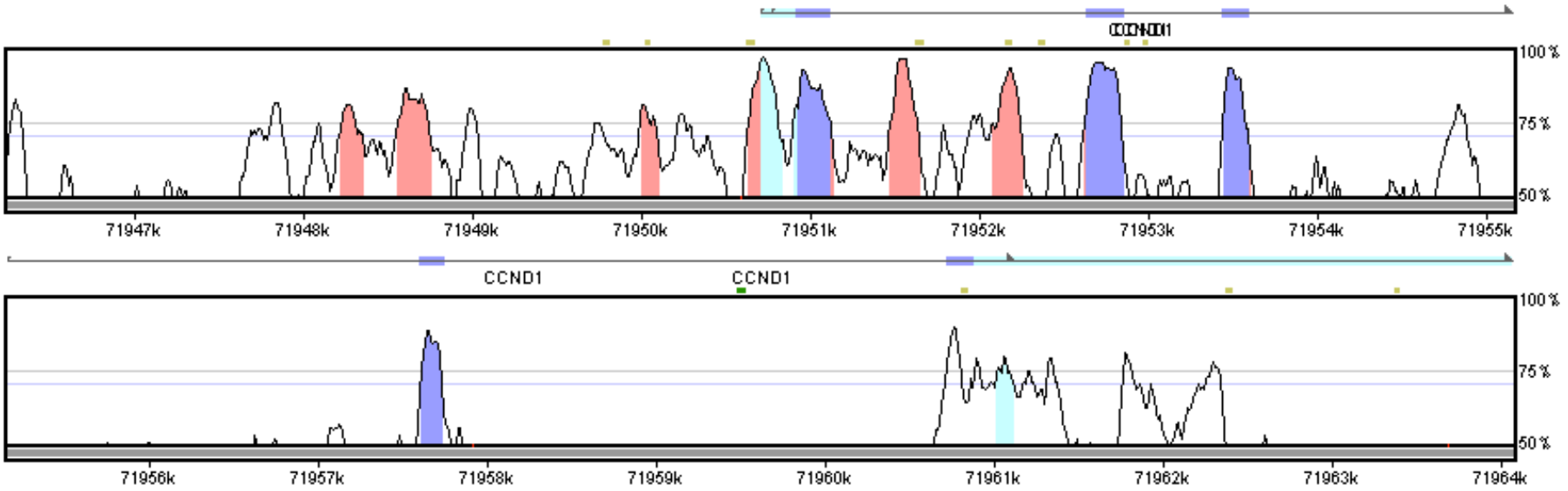


Genome-wide analysis of conserved motifs (ctd.)

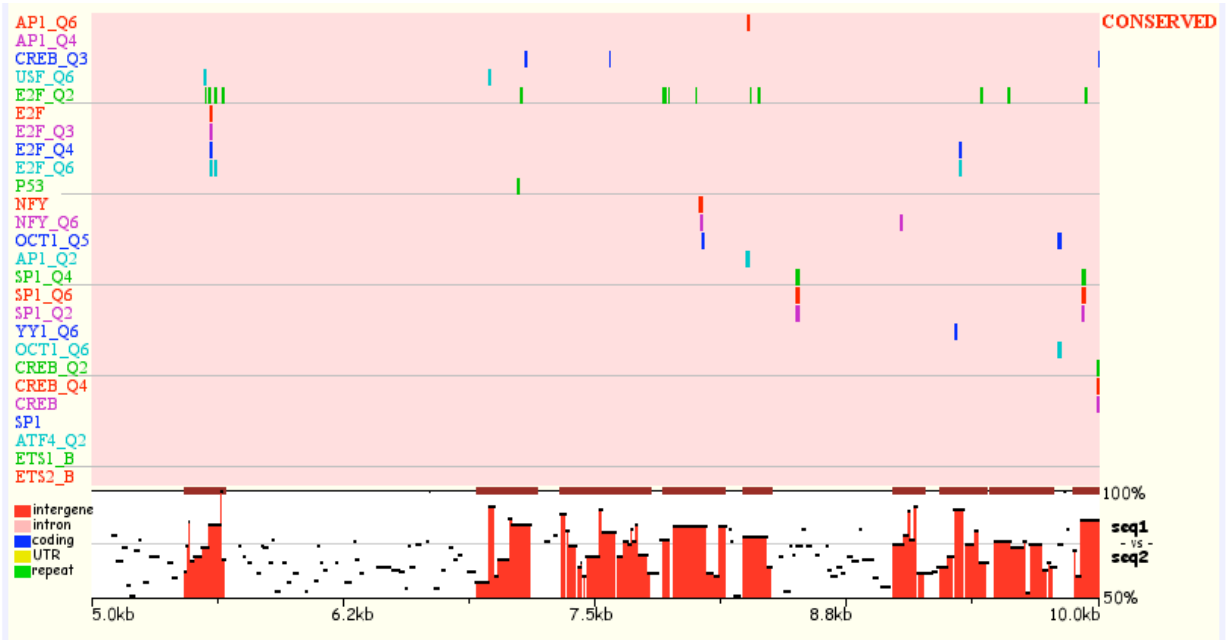
- New motifs
  - 6-18 bp w/ allowed degeneracy
  - 174 motifs MCS>6
    - 59 w/ matches to known motifs
  - Validation:
    - correlation with tissue-specific gene expression
    - Positional bias of the motifs relative to the TSS



Conserved non-coding sequences (CNS)



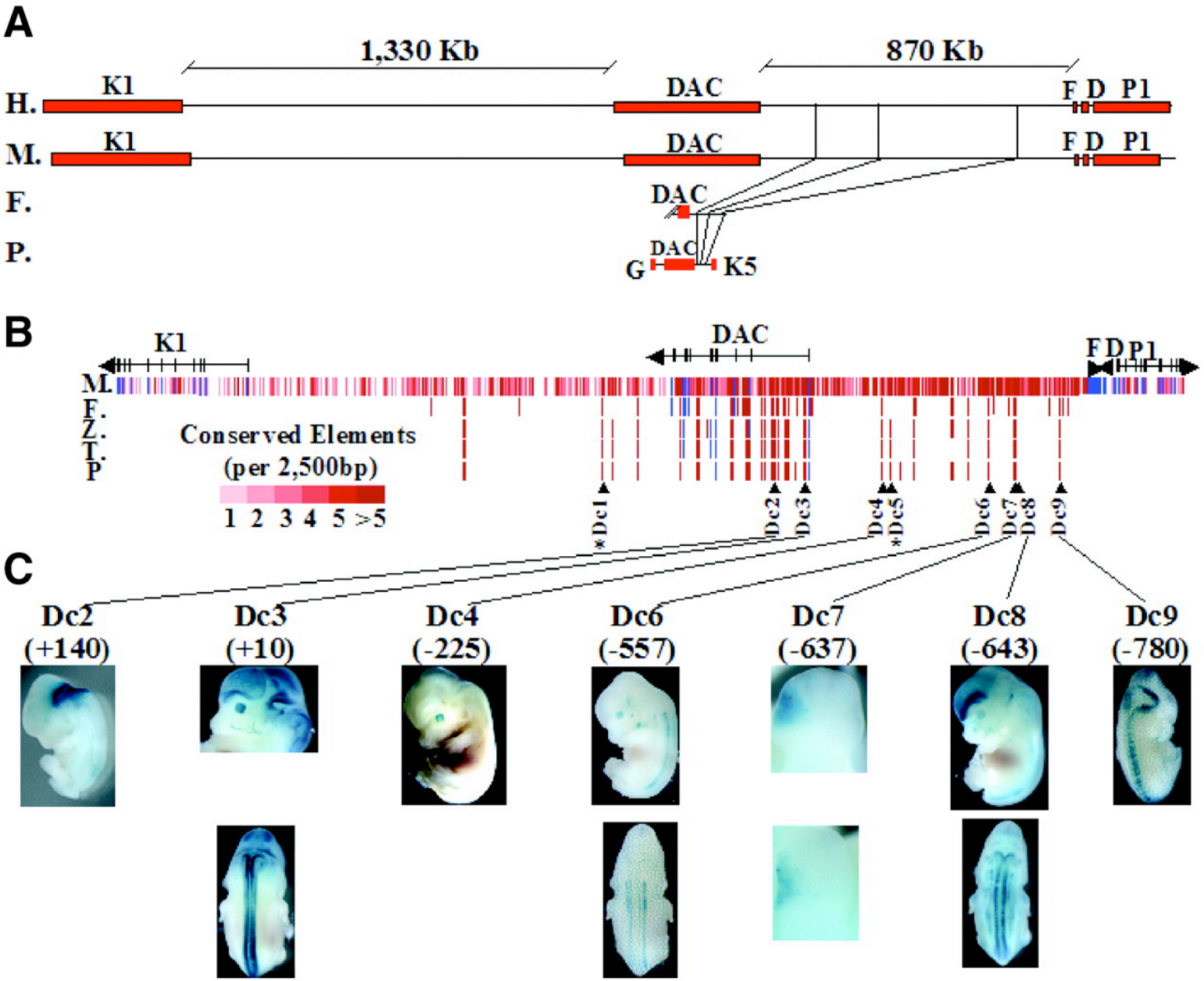
AVID, LAGAN, BLASTZ



TraFac, ConSite, rVISTA

Conserved non-coding sequences

Nobrega et al. (2003) *Science* 302 (5644)



Ultraconserved sequences

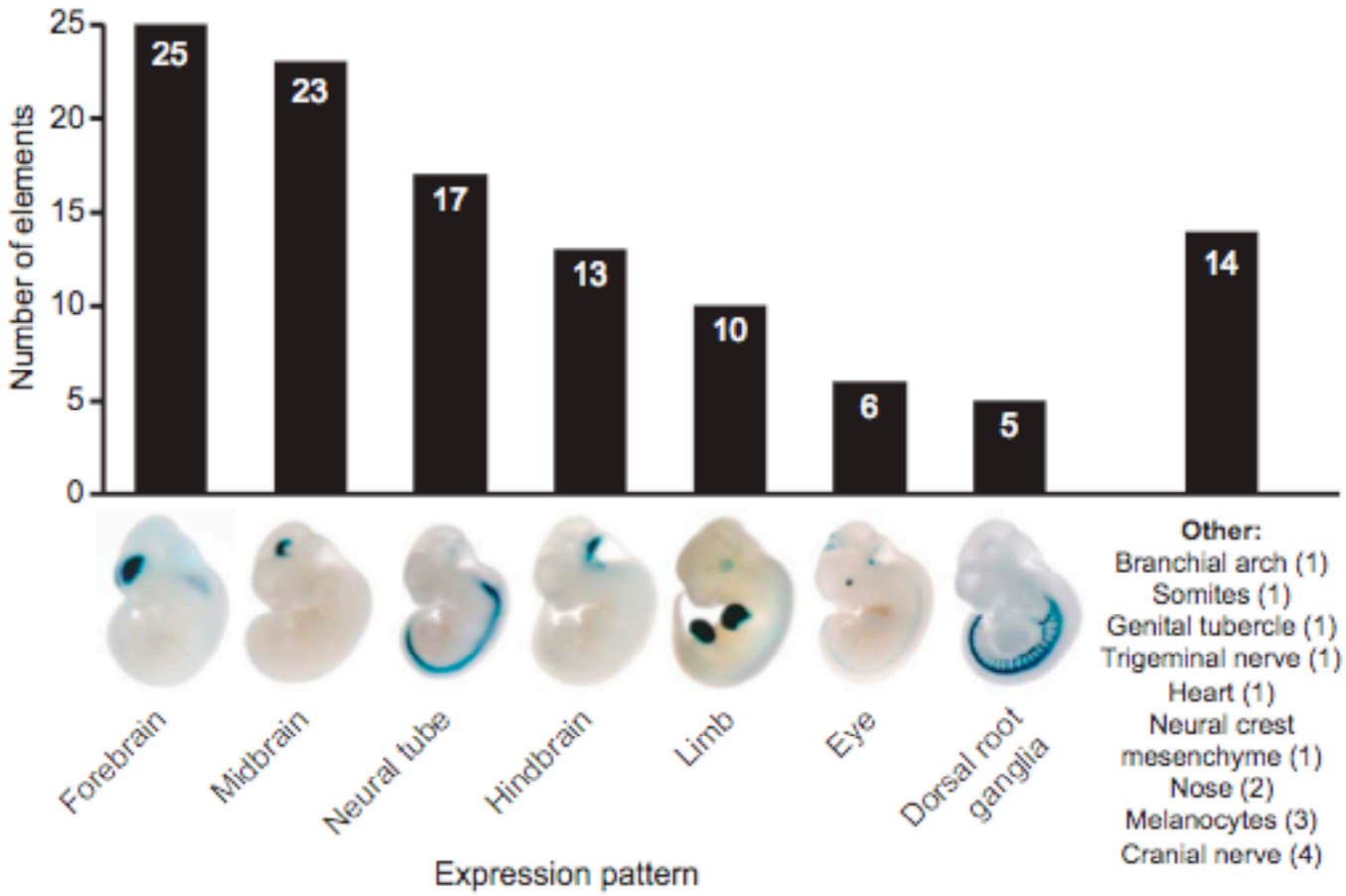
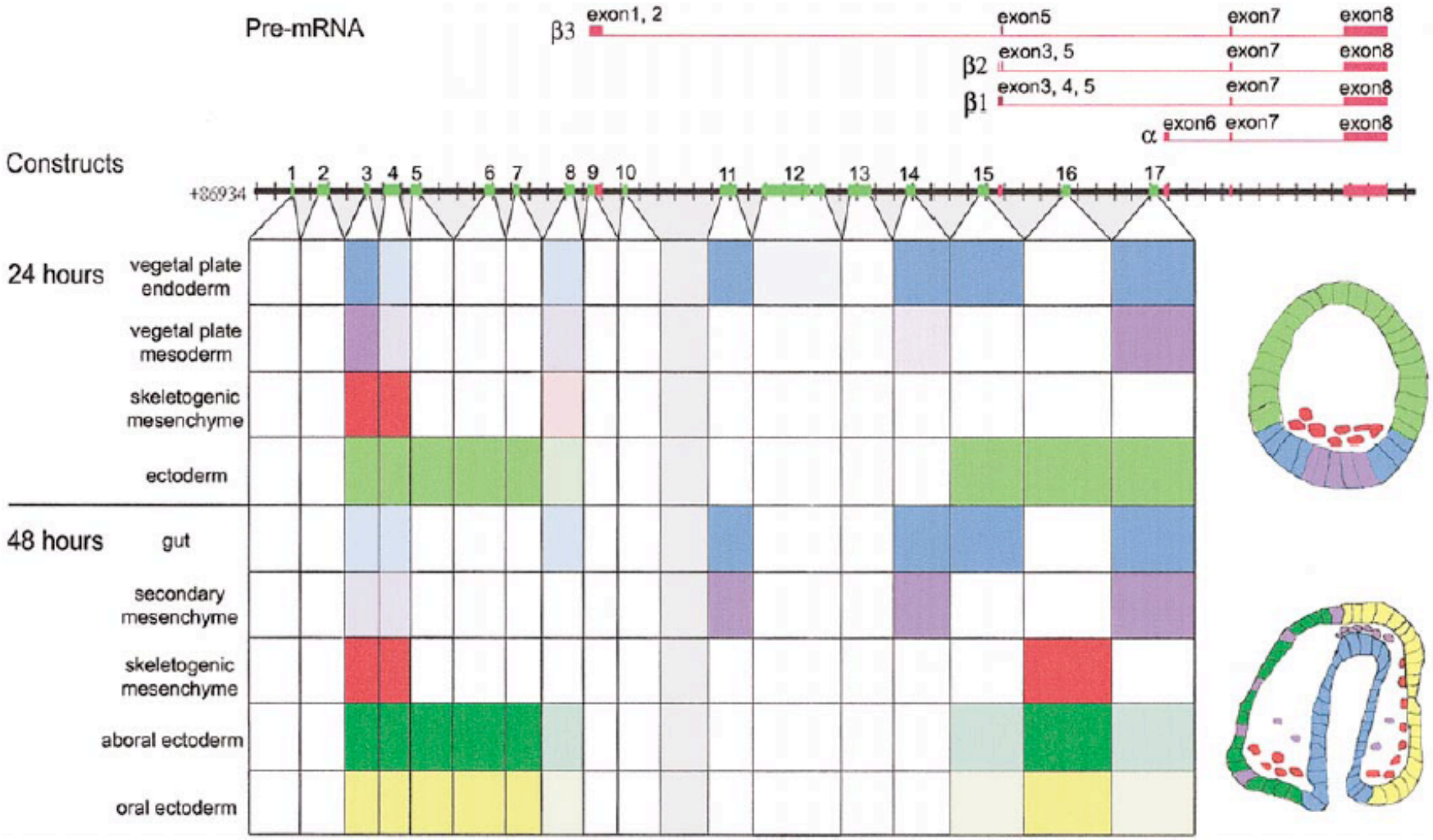


Figure 3 | Grouping of positive expression patterns captured in the transgenic mouse enhancer assay. The total number of elements displaying a given anatomical pattern is depicted by the height of the bars in the chart. A representative transgenic embryo is provided for each expression pattern. Elements with reproducible staining in more than one structure are included in each respective category.

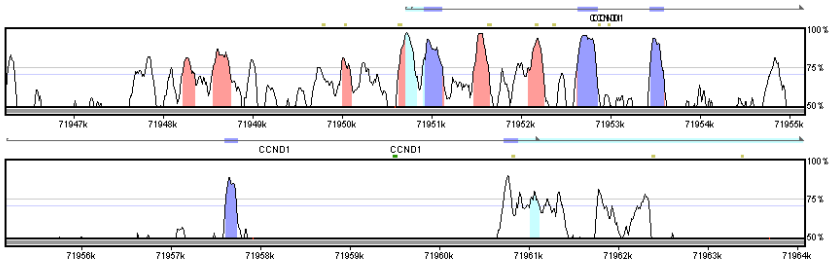
Conserved non-coding sequences

Strongylocentrotus purpuratus - Lytechinus variegatus: otx gene Yuh et al. (2002) *Dev Biol* 246:1



Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. Dev Biol. 2002 Jun 1;246(1):148-61.

Conserved non-coding sequences (CNS)



AVID, LAGAN, BLASTZ



Nobrega et al. (2003) *Science* 302 (5644)

Woolfe et al. (2005) *PLoS Biology* 3(1)

Yuh et al. (2002) *Dev Biol* 246:1

$$E(occ\{b\}) = F_e\{b\} \times 2 \times \sum_{i=1}^{S}(L_i - w + 1) = F_e\{b\} \times T$$

$$P(occ\{b\} = n) = \frac{T!}{(T-n)!\times n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)}$$

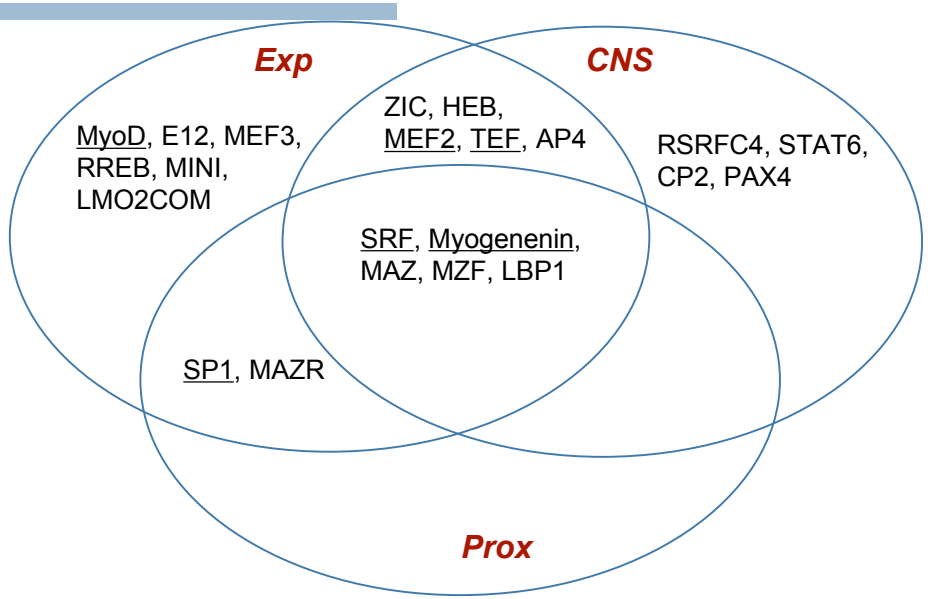
$$P(occ\{b\} \geq n) = \sum_{j=n}^{T} P(occ\{b\} = j)$$

Statistical over-representation

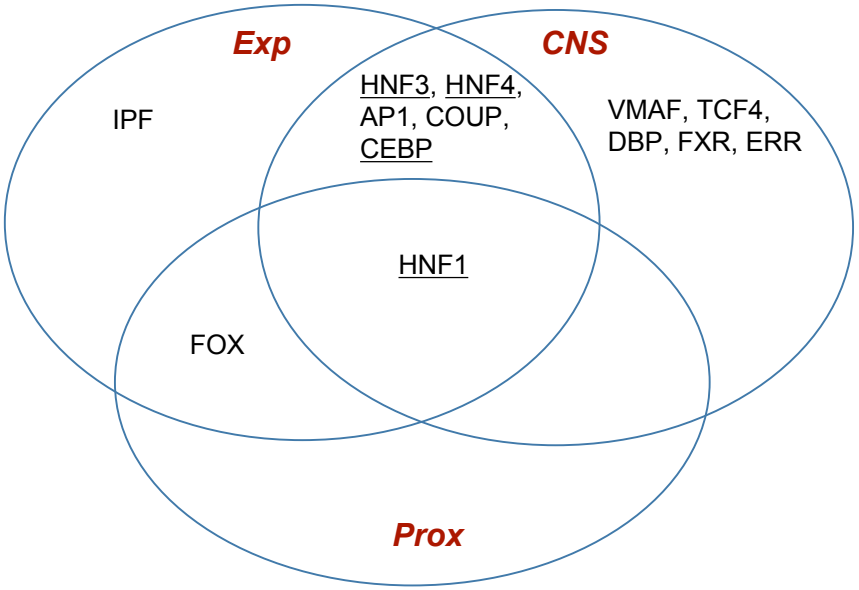
Aerts et al. (2003) *Nucleic Acids. Res.* 15;31(6):1753-64

Statistical over-representation

Wasserman & Fickett (1998)



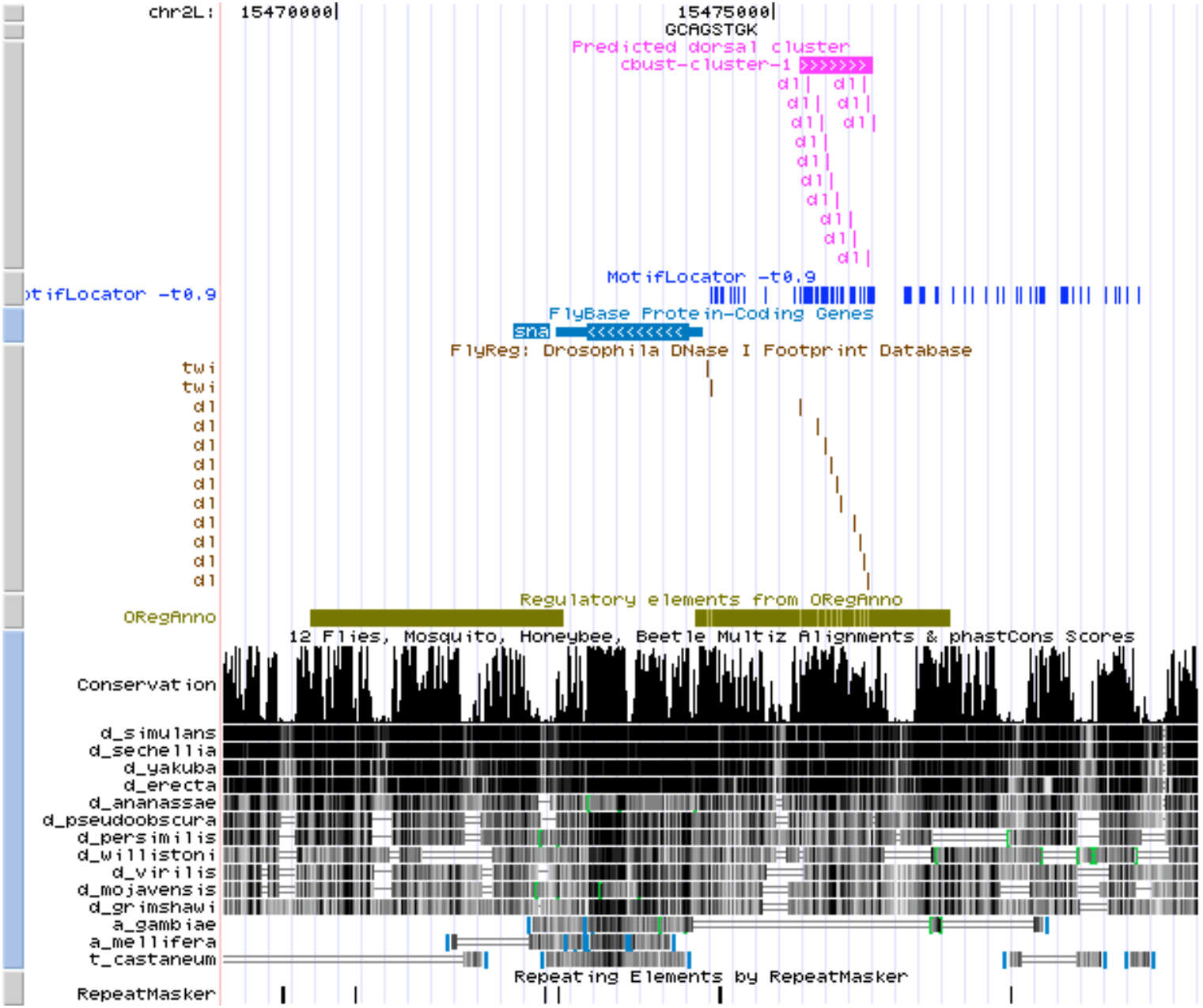
Krivan & Wasserman (2001)



Aerts et al. (2003) *Nucleic Acids. Res.* 15;31(6):1753-64

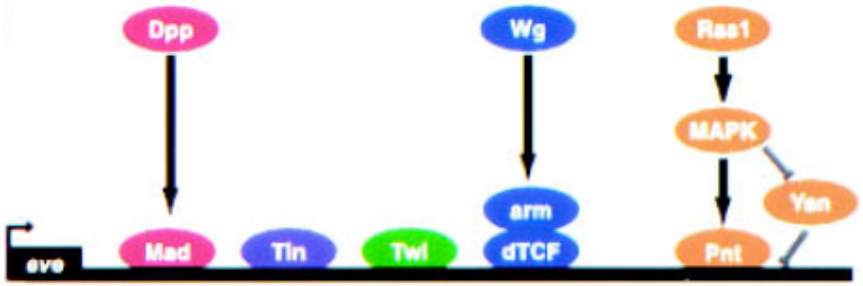
cis-regulatory modules

Homotypic clusters

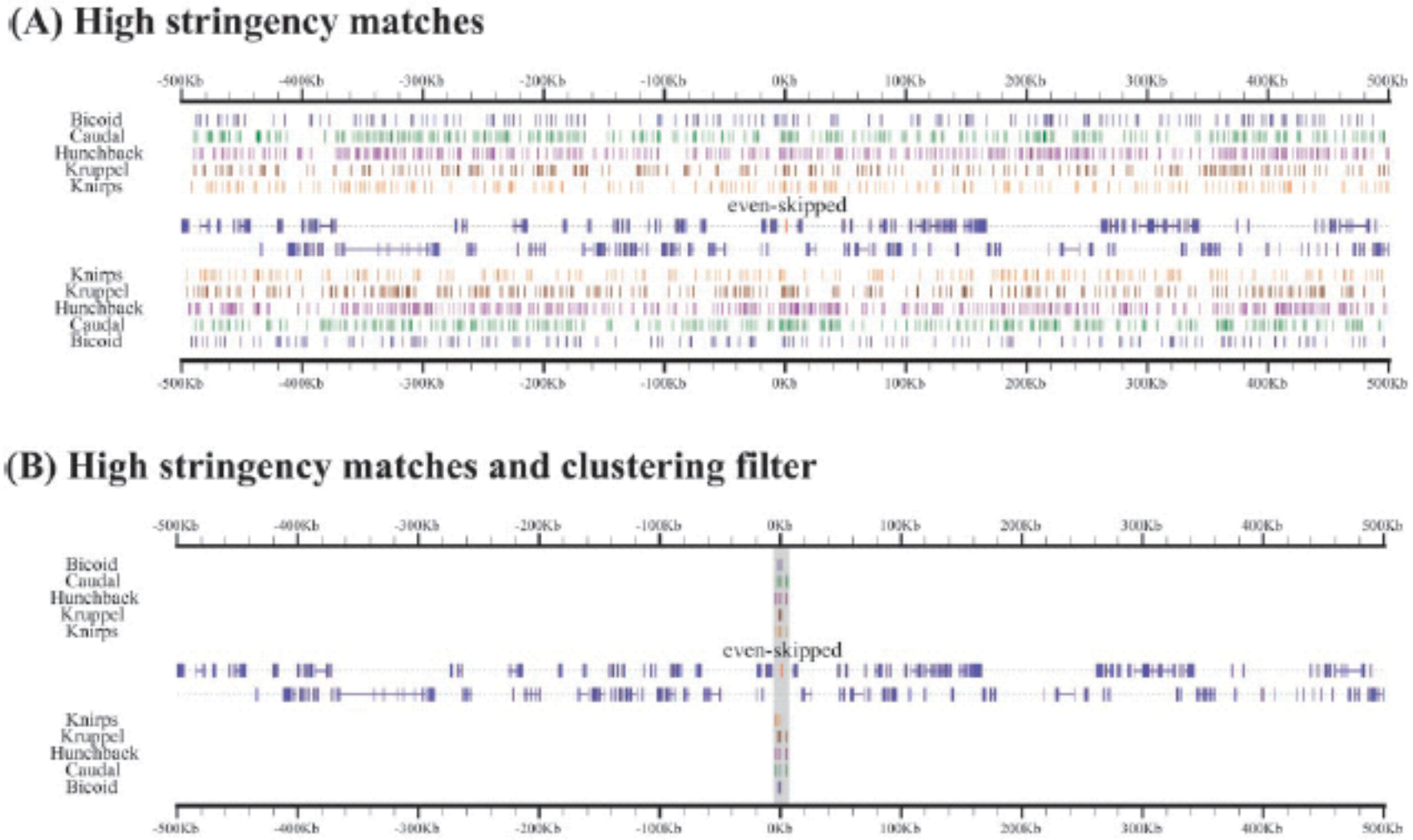


Heterotypic clusters

- Given:
  - A known, functional combination of transcription factors
  - The PWM for each TF
- Question:
  - Genome-wide occurrences of cis-regulatory modules
- Example: Eve MHE

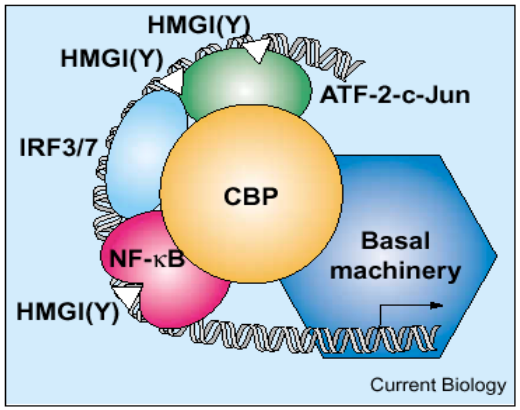


  - Score each PWM independently
  - Search for **co-occurence** within 500 bp windows
    - 647 elements
    - Filter: 2 binding sites for each PWM, location, expression data, etc.
    - Halfon et al. 2002, *Genome Research*

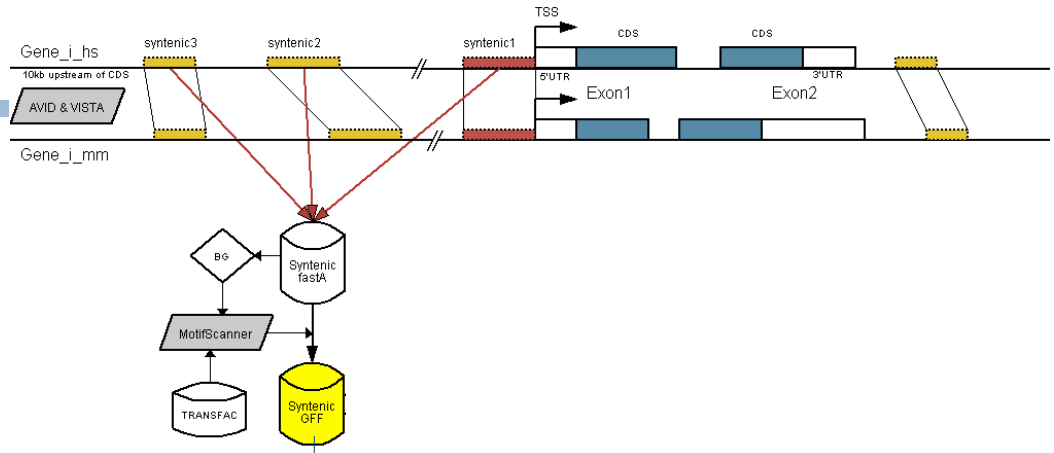


Berman BP et al. 2002 (PNAS)
Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.

ModuleScanner



Courey AJ (2001) Curr Biol. 11(7):R250-2



M00750-V$HMGIY_Q6
M00054-V$NFKAPPAB_0
M00747-V$IRF1_Q6

TOP 10

High scoring genes:

- IFNβ (nr 4)
- EH-domain containing protein 1 (MHC class I recycling)
- IL-1β precursor (inflammatory response)
- NF-κβ inhibitor α
- semaphorin 3B precursor

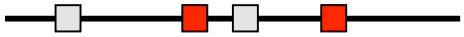
Significant GO terms	Corrected p value
apoptosis	0.002046759
negative regulation of cell proliferation	0.003660635
protein amino acid dephosphorylation	0.004239594
response to pest/pathogen/parasite	0.005201112
protein phosphatase activity	0.006733237
innate immune response	0.010171468
cytokine activity	0.012253669
response to stress	0.014523294
phosphoric monoester hydrolase activity	0.015017083
cell communication	0.031928904

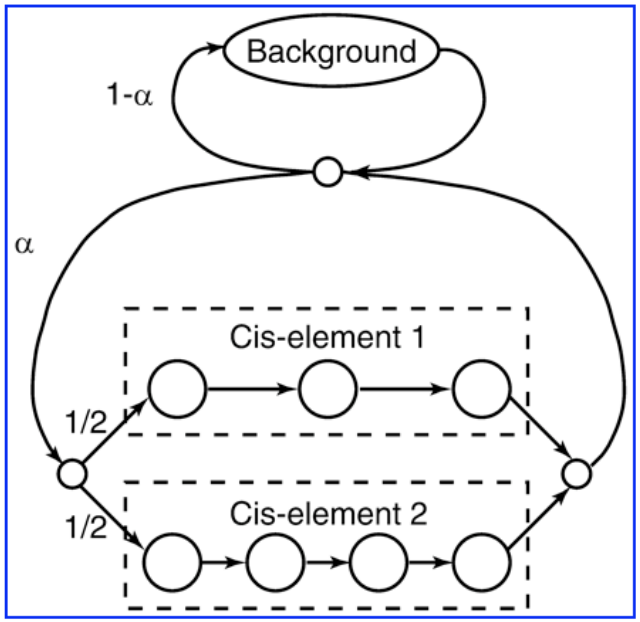
HMM for module scanning

DNA sequence segment

$$\text{score}\left(\text{segment}\right) = \ln\left[\frac{\text{Prob}\left(\text{segment} \mid \text{cluster model}\right)}{\text{Prob}\left(\text{segment} \mid \text{random model}\right)}\right]$$

Cluster model:



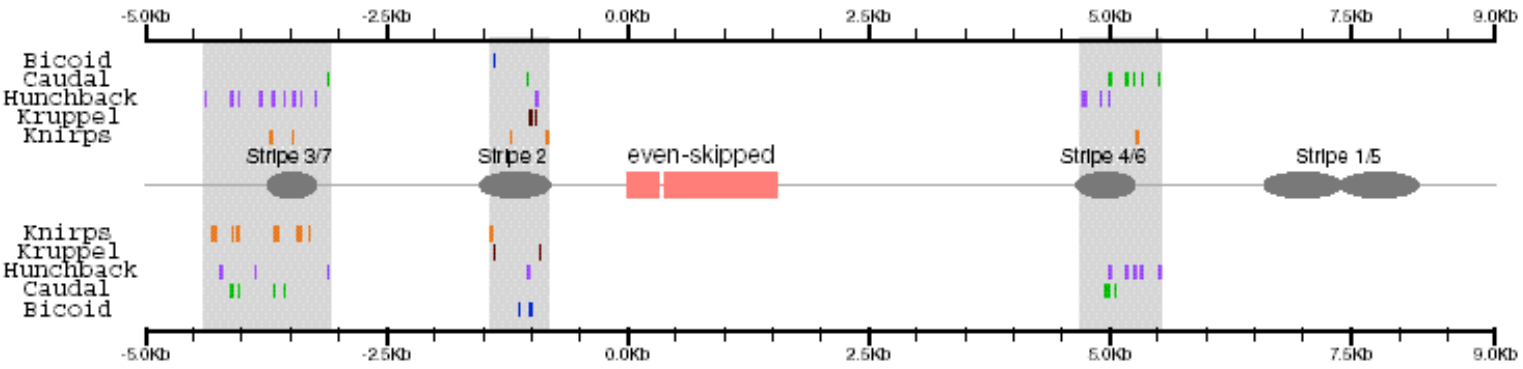


HMM: Cister, COMET, Ahab, Stubb, ClusterBuster, MCAST

Combined statistical significance: MSCAN

Example: Cis-regulatory module in *drosophila*

Cis-regulatory modules often involve clusters of binding sites for one or more transcription factors (e.g., drosophila EVE gene).



Clusters of 13 or more pattern matches in a window of 700 bp

Use sequence conservation for CRM scanning

- Scan only conserved sequences (CNS)
  - ModuleScanner
- Aligned sites
  - As a filter
    - e-cisAnalyst
  - Incorporated in the HMM
    - Stubb
- ‘Preserved’ sites
  - As a filter
  - Combine HMM scores for orthologous regions

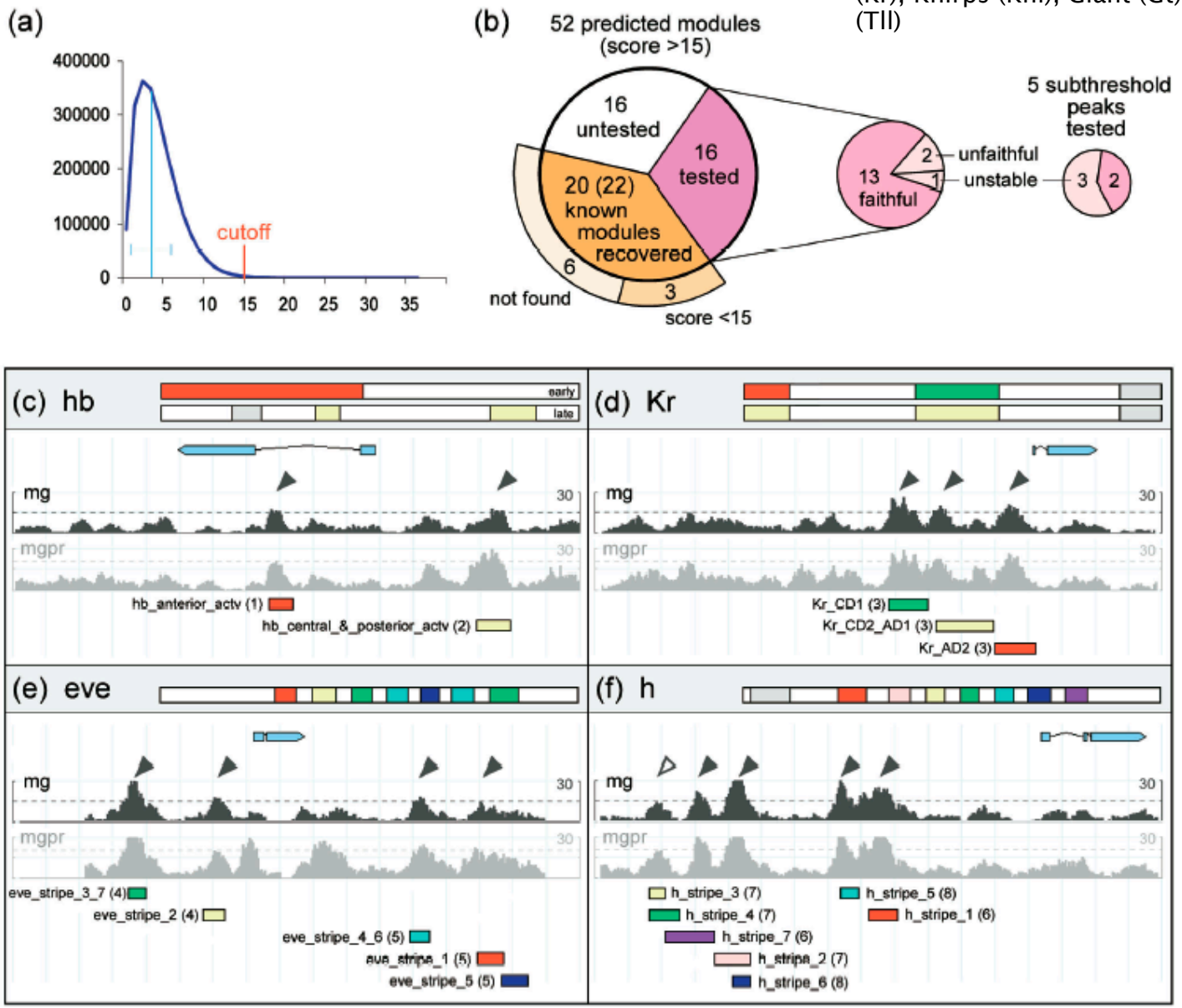
Genome-wide target identification by CRM scanning

Examples CRM scanning

- Homotypic clusters
  - Su(H) YGTGDGAA: multiple instances in 100bp-5000bp window
    - ~ 40 candidate targets
    - Validated Him
    - Rebeiz et al *PNAS* 2002
  - Dorsal GGGWD$_{\{0,1\}}$W$_{\{3\}}$CCM, ≥ 3 sites, windows of 400bp
    - 15 clusters genome-wide
    - Validated sog (Phm, Ady)
    - Markstein et al. PNAS 2001
  - Bicoid PWM, ≥ 6 low affinity sites 550 bp, ≥ 2 medium aff 200bp, ≥ 1 high aff
    - Scored 10 candidate genes
    - 11/14 tested clusters recapitulate endogeneous expression
- Heterotypic clusters
  - [Cad,Bcd,Kr,Kni,Hb] PWMs, $p_{site} < 0.0003$, ≥15 sites, 700 bp window
    - 37 clusters genome-wide, in 49 unique genes (11 with correct expression)
    - Validated gt (Berman et al. *PNAS* 2002)
    - 15/37 with early-embryonic enhancer activity (Berman et al. Genome Biol. 2004)
  - [dTcf,Mad,Tin,Twi,Pnt], each ≥ 2 (except dTcf), in 500bp window
    - 33 clusters genome-wide
    - Validated hbr
    - Halfon et al. *Genome Research*, 2002
  - [Bcd, Hb, Cad, TorRE, Stat92E, Kr, Kni, Gt, Tll], Hidden Markov Model (HMM), window 500bp
    - Schroeder et al. PLoS Biology 2004
    - 52 modules in 29 selected genes
    - 13/16 tested modules produce faithful expression patterns

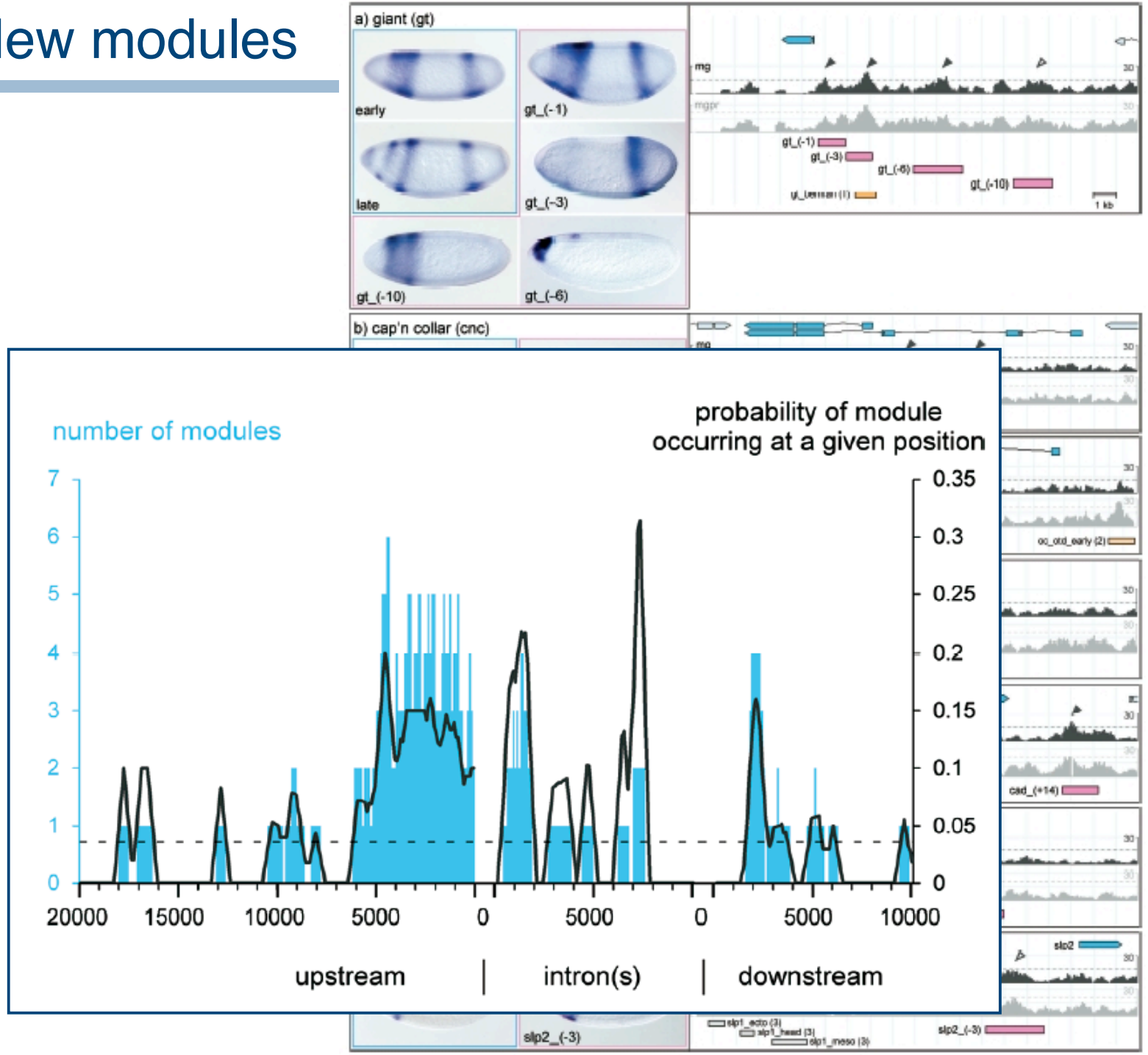
Recovery of known modules

Bicoid (Bcd), Hunchback (Hb), Caudal (Cad), the Torso-response element (TorRE), and Stat92E (D-Stat); Kruppel (Kr), Knirps (Kni), Giant (Gt), and Tailless (Tll)



Schroeder et al. 2004, PLOS Biology

New modules


a) giant (gt)
early
late
gt_(-10)
gt_(-1)
gt_(-3)
gt_(-6)
b) cap'n collar (cnc)
slp2_(-3)
number of modules
probability of module
occurring at a given position
0
1
2
3
4
5
6
7
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
20000
15000
10000
5000
0
5000
0
5000
10000
upstream
intron(s)
downstream

Classification, contingency table, metrics

		Condition (as determined by "Gold standard")		
		Positive	***Negative***	
Test outcome	***Positive***	True Positive	False Positive (Type I error, P-value)	→ Positive predictive value
	Negative	False Negative (Type II error)	True Negative	→ Negative predictive value
		↓ Sensitivity	↓ Specificity	

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error
true positive rate (TPR)
eqv. with hit rate, recall, sensitivity
$TPR = TP / P = TP / (TP + FN)$
false positive rate (FPR)
eqv. with false alarm rate, fall-out
$FPR = FP / N = FP / (FP + TN)$
accuracy (ACC)
$ACC = (TP + TN) / (P + N)$
specificity (SPC)
$SPC = TN / (FP + TN) = 1 - FPR$
positive predictive value (PPV)
eqv. with precision
$PPV = TP / (TP + FP)$
negative predictive value (NPV)
$NPV = TN / (TN + FN)$
false discovery rate (FDR)
$FDR = FP / (FP + TP)$
Matthews Correlation Coefficient (MCC)
\mbox{MCC} = (\mbox{TP}\; \mbox{TN} - \mbox{FP}\; \mbox{FN})/ \sqrt{P N P' N'}

Source: Fawcett (2004).

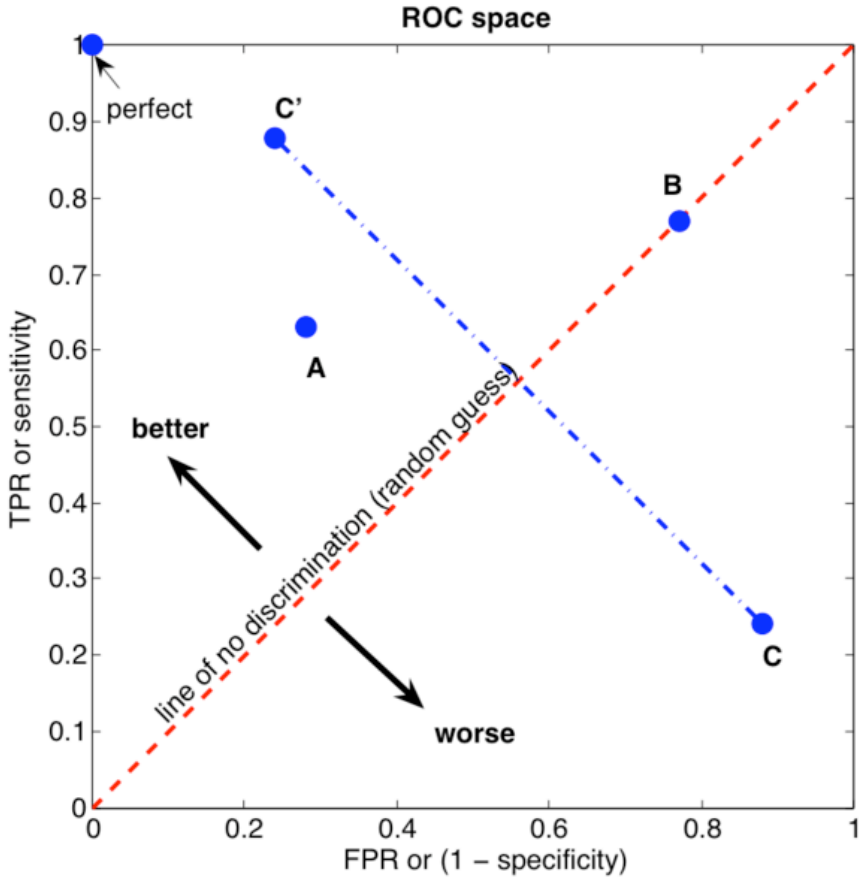
		Patients with bowel cancer (as confirmed on endoscopy)		
		Positive	***Negative***	?
FOB test	***Positive***	TP = 2	FP = 18	= TP / (TP + FP) = 2 / (2 + 18) = 2 / 20 ≡ **10%**
	Negative	FN = 1	TN = 182	= TN / (TN + FN) 182 / (1 + 182) = 182 / 183 ≡ **99.5%**
		↓ = TP / (TP + FN) = 2 / (2 + 1) = 2 / 3 ≡ **66.67%**	↓ = TN / (FP + TN) = 182 / (18 + 182) = 182 / 200 ≡ **91%**	

the Fecal occult blood (FOB) screen test is used in 203 people to look for bowel cancer:

Related calculations

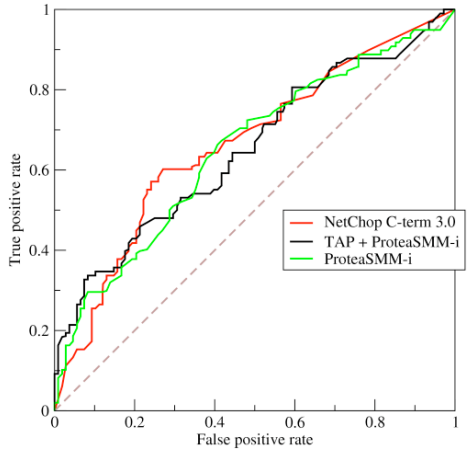
- False positive rate (α) = FP / (FP + TN) = 18 / (18 + 182) = 9% = 1 - specificity
- False negative rate (β) = FN / (TP + FN) = 1 / (2 + 1) = 33% = 1 - sensitivity
- Power = sensitivity = 1 − β

Receiver Operating Characteristic (ROC)



Discrete classifiers, such as decision tree or rule set, yield numerical values or **binary label**. When a set is given to such classifiers, the result is a single point in the ROC space.

For other classifiers, such as naive Bayesian and neural network, they produce **probability values** representing the degree to which class the instance belongs to. For these methods, setting **a threshold value** will determine a point in the ROC space. For instance, if probability values below or equal to a threshold value of 0.8 are sent to the positive class, and other values are assigned to the negative class, then a confusion matrix can be calculated. Plotting the ROC point for each possible threshold value results in a curve.



A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

TPR = 0.63
FPR = 0.28
ACC = 0.68

B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

TPR = 0.77
FPR = 0.77
ACC = 0.50

C

TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

TPR = 0.24
FPR = 0.88
ACC = 0.18

C'

TP=88	FP=24	112
FN=12	TN=76	88
100	100	200

TPR = 0.88
FPR = 0.24
ACC = 0.82

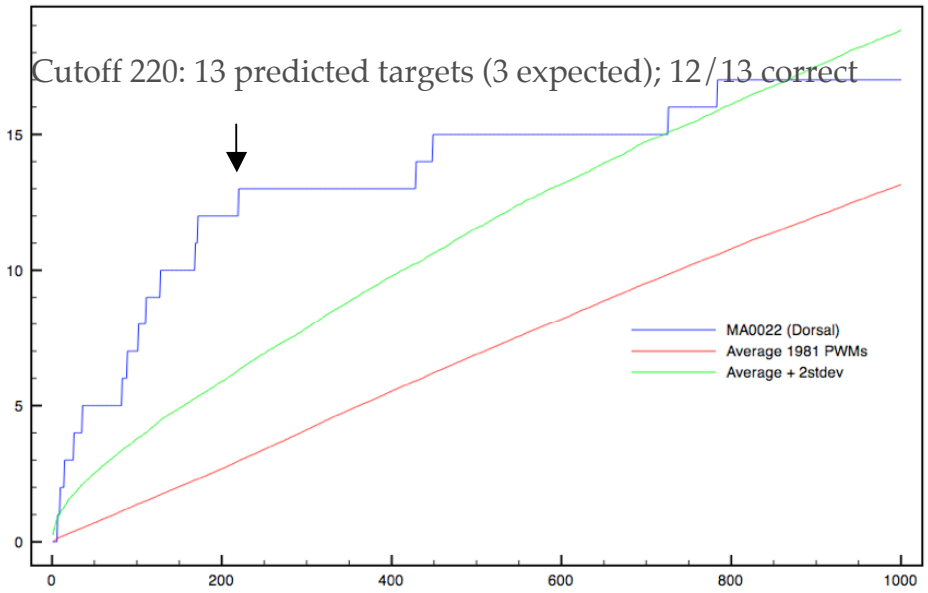
Drosophila example: dorsal

39 dl targets		dl downstream	
twi	7	twi	7
sna	10	sna	10
vnd	15	vnd	15
Mef2	26	Mef2	26
rho	36	rho	36
bun	61	sim	83
Trim9	64	ed	89
sim	83	fog	102
fog	102	ths	111
ths	111	sog	128
sog	128	dpp	169
dpp	169	tup	172
tup	172	Doc3	220
CadN	209	Mdr49	429
Doc3	220	zfh1	449
vn	228	Ect4	726
zfh1	449	Neu3	784
stumps	463	wts	1326
Dr	774	Socs36E	1360
Neu3	784	C15	1494

Use expression data

80 co-expressed genes, downstream of Dorsal

26/39 known direct targets present



Precision: 92% (lower bound)

Recall: 46% (upper bound)

Lever

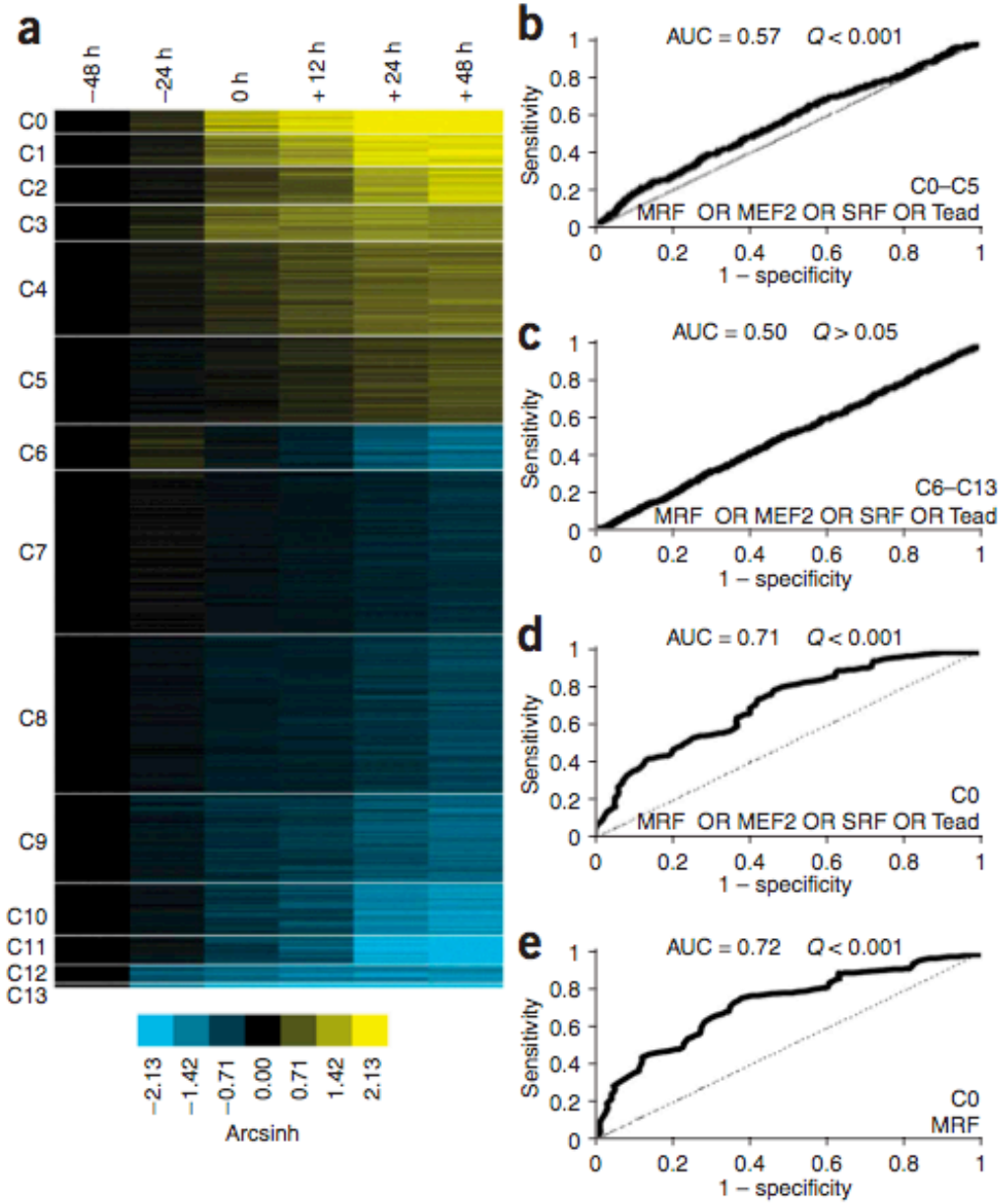
Systematic identification of mammalian regulatory motifs' target genes and functions

Jason B Warner[1,7], Anthony A Philippakis[1,3,4,7], Savina A Jaeger[1,7], Fangxue Sherry He[1,6], Jolinta Lin[1,5] & Martha L Bulyk[1–4]

We developed an algorithm, Lever, that systematically maps metazoan DNA regulatory motifs or motif combinations to sets of genes. Lever assesses whether the motifs are enriched in *cis*-regulatory modules (CRMs), predicted by our PhyloCRM algorithm, in the noncoding sequences surrounding the genes. Lever analysis allows unbiased inference of functional annotations to regulatory motifs and candidate CRMs. We used human myogenic differentiation as a model system to statistically assess greater than 25,000 pairings of gene sets and motifs or motif combinations. We assigned functional annotations to candidate regulatory motifs predicted previously and identified gene sets that are likely to be co-regulated via shared regulatory motifs. Lever allows moving beyond the identification of putative regulatory motifs in mammalian genomes, toward understanding their biological roles. This approach is general and can be applied readily to any cell type, gene expression pattern or organism of interest.

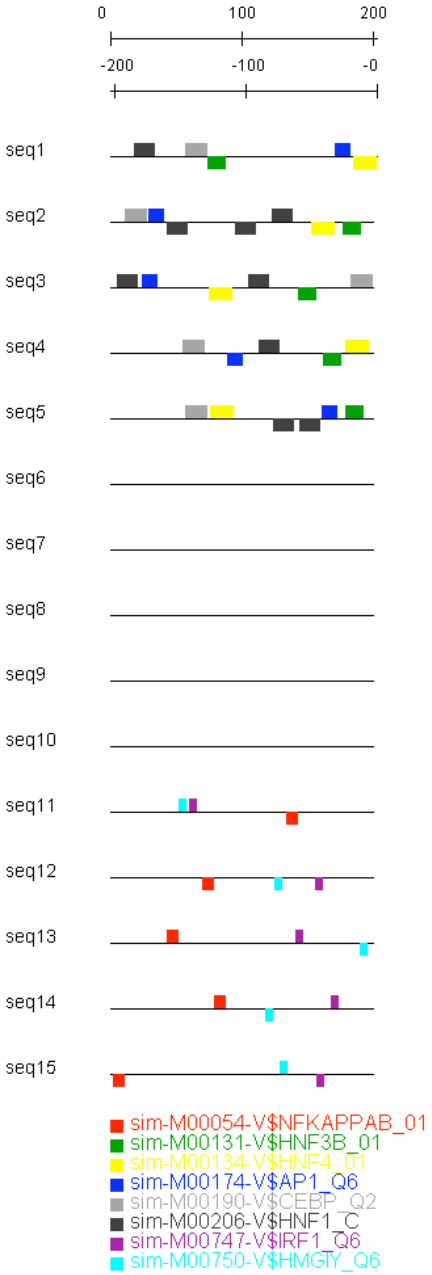
regulated by the CRMs. Although algorithms have been developed recently for evaluating the regulatory importance of CRM binding site composition[5,6], thus far they have been unable to evaluate the vast sequence regions beyond the proximal promoter that must be considered in mammalian genomes.

Because of these complications, analyses of transcriptional regulatory elements in mammals have focused either on the prediction of CRMs starting with a collection of known co-regulatory transcription factors whose DNA binding specificities are known and a set of genes that the transcription factors may regulate[1,3,7,8], or on the computational identification of 'motif dictionaries'[9–12]. However, with the advent of high-throughput methods for assembling motif dictionaries, from either chromatin immunoprecipitations[13] or protein binding microarrays[14–16], the major computational problem to solve will shift from motif prediction to identifying and associating CRMs to both specific genes and biological processes[17].



Module discovery

Module score function



$\mathcal{M}=(\Theta_1, \Theta_2,\ldots, \Theta_n)$ Module model

$$\text{PWM}=\Theta=\begin{bmatrix} 0,4 & 0,1 & \ldots & 0,2 \\ 0,2 & 0,1 & \cdots & 0,6 \\ 0,3 & 0,7 & \ldots & 0,1 \\ 0,1 & 0,1 & \ldots & 0,1 \end{bmatrix}$$

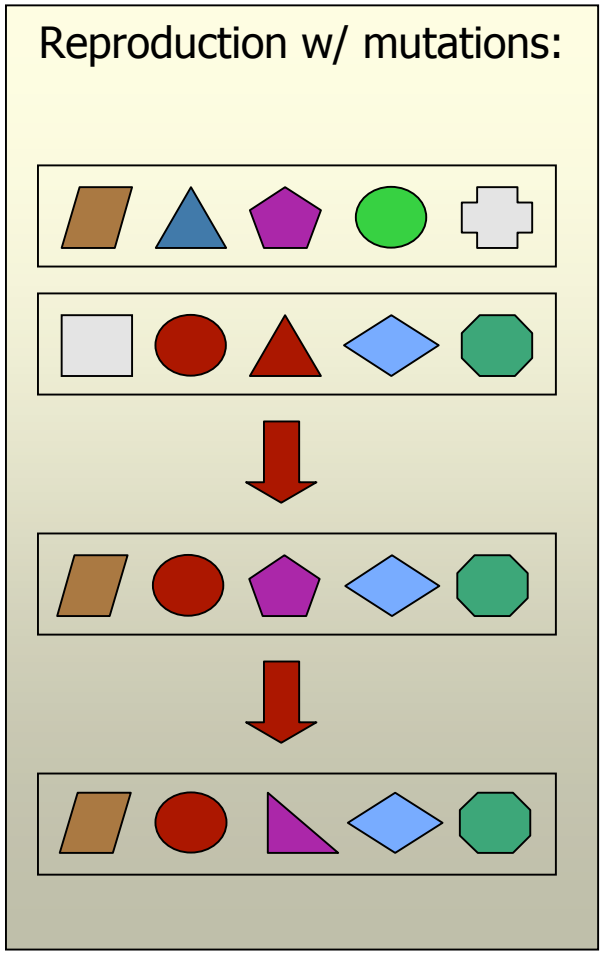
$$\mathcal{S}_{\mathcal{M}}(\mathbf{set}) = \sum_{i=1}^{n} \mathcal{S}_{\mathcal{M}}(seq_i)$$

$$\mathcal{S}_{\mathcal{M}}(seq) = \max_{T} p(\mathbf{m}) \times \sum_{\mathbf{x}\in\mathbf{m}} W(\mathbf{x})$$

T = all valid modules m's
- within window of length maxLength
- overlapAllowedYesNo

n=5 ⇒ $400^5=10^{13}$ combinations ⇒ Efficient search algorithm

Genetic algorithm



Population: generate *p* random modules

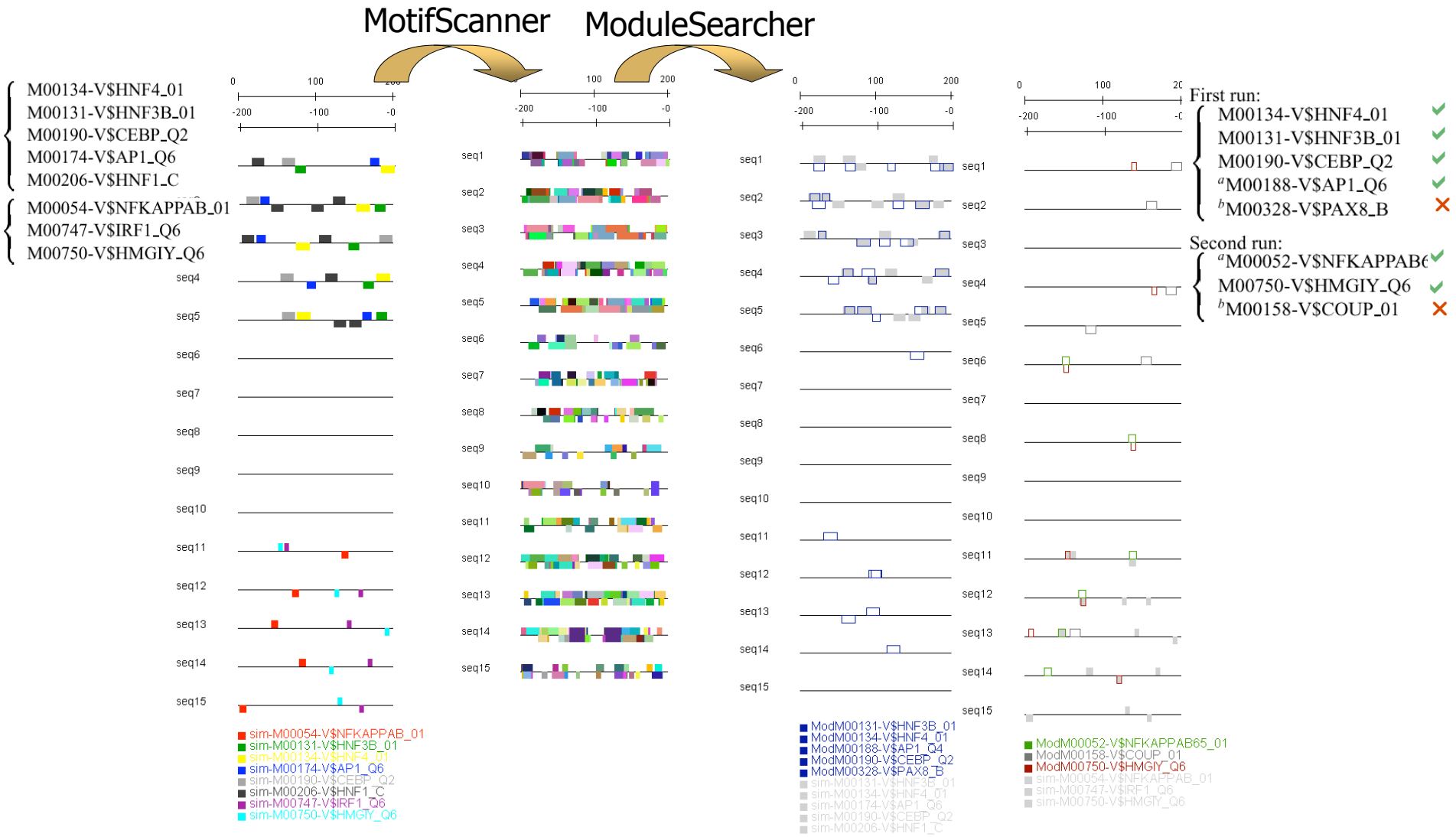
Evaluation: fitness value for each module
s best scoring modules are survivors

g generations

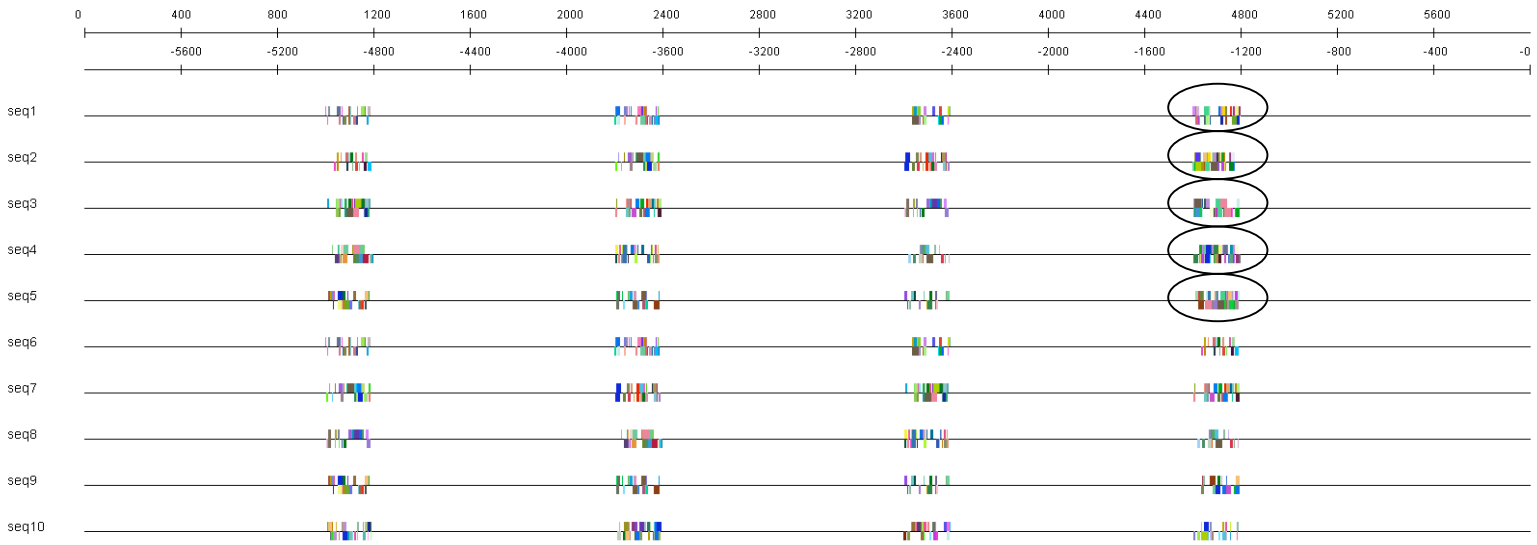
Reproduction:
survivors are paired to create *p* new modules by crossover

Mutation: replace a factor with another factor with mutation probability ρ

Test on simulated data

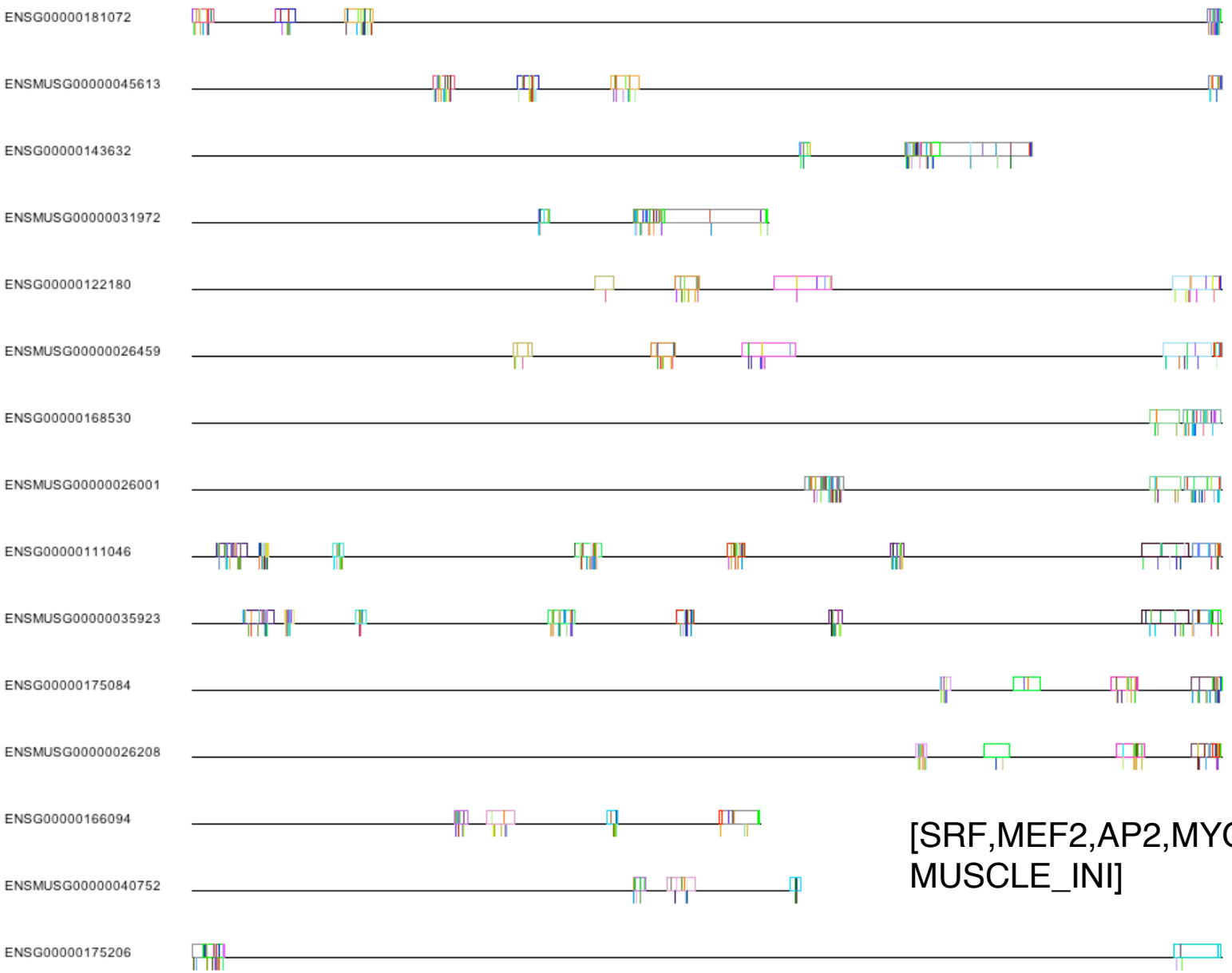


Test on simulated data



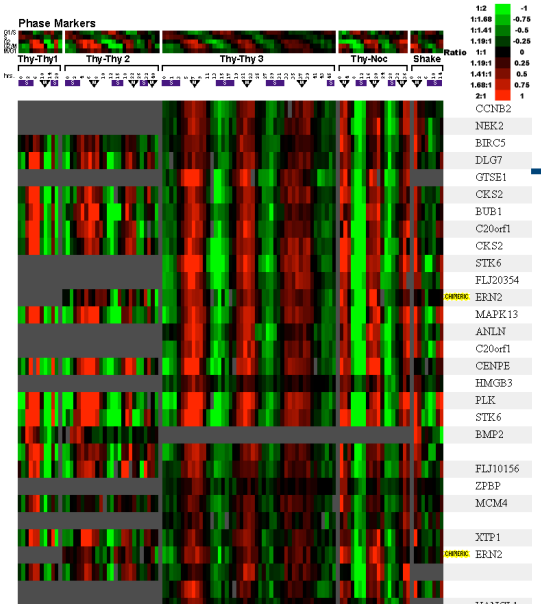
- M00134-V$HNF4_01
- M00131-V$HNF3B_01
- M00190-V$CEBP_Q2
- M00188-V$AP1_Q6
- M00206-V$HNF1_Cc

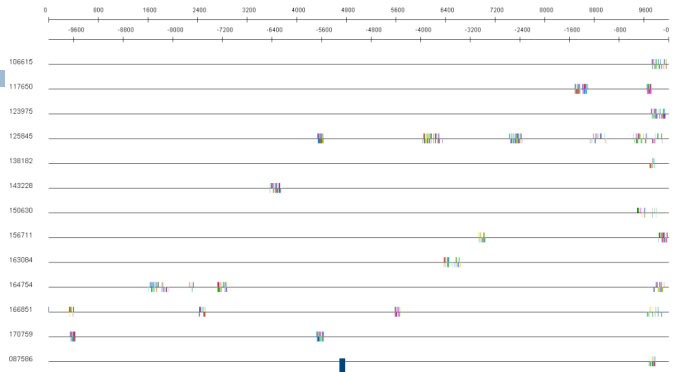
ModuleSearcher + phylogenetic footprinting



[SRF,MEF2,AP2,MYOGENIN, MUSCLE_INI]

Module validation





M00116-V$CEBPA_01
M00264-V$STAF_02
M00287-V$NFY_01
M00671-V$TCF4_Q5

Apply Genome-wide scanning

Significant GO terms	Corrected *p* value
mitosis	0.000435808
M phase of mitotic cell cycle	0.000452022
cytokinesis	0.000468573
nuclear division	0.001257531
regulation of cell cycle	0.001395887
protein serine/threonine kinase activity	0.010734361
obsolete	0.024402181
cell proliferation	0.026461498

High scoring genes:

- PLK
- NEK2
- CDC2
- CKS1B (CDC2 associated protein)

Exercise / competition motif and CRM discovery

- Simulated set
  - 5 PWMs
  - 10 seq with 1 instance of each PWM
  - 10 seq generated from 0th order background model
  - What and where are the motifs, and which are the 10 sequences?
    - Upload GFF to **http://med.kuleuven.be/cme-mg/lng/student/** !!
    - Put method, parameters, explanation in "comment" field
    - Scored by calculating the nucleotide-level correlation with the answer vectors

End

Parse Fuzznuc output to GFF

```
########################################
# Program: fuzznuc
# Rundate: Fri Aug 29 16:05:59 2008
# Report_format: seqtable
# Report_file: chr2.fuzznuc
########################################

#=======================================
#
# Sequence: chr2     from: 1   to: 813136
# HitCount: 52
#
# Pattern: CGGNNNNNNNNNNCCG
# Mismatch: 0
# Complement: No
#
#=======================================

  Start      End Mismatch Sequence
   4125     4141        . CGGTAAGCGTATCACCG
  11140    11156        . CGGGCCTTTCTATACCG
  30297    30313        . CGGACCCACACTATCCG
  45259    45275        . CGGGTCGAAATTTGCCG
  59859    59875        . CGGCACATTCCTTACCG
  76469    76485        . CGGCAAGGGCAGCACCG
  89651    89667        . CGGGTTTGGCGAGGCCG
 144565   144581        . CGGGGATTAAGATACCG
 160205   160221        . CGGCTCCAGTGATGCCG
```

```perl
#!/usr/bin/perl -w
my $format=$ARGV[0];
while(<STDIN>){
        $line=$_;
        chomp $line;
        if ($line =~/Sequence:\s(.*)\s+from/){
                $seq=$1;
        }
        if ($line =~/Pattern:\s(.*)/){
                $pattern=$1;
        }
        if ($line =~/\s+(\d+)\s+(\d+)\s+([\d\.])\s+([acgtACGT]+)/){
                $start=$1;
                $end=$2;
                $mismatch=$3;
                $site=$4;
                if($start>$end){
                        $temp=$end;
                        $end=$start;
                        $start=$temp;
                        $strand="-";
                }
                else{
                        $strand="+";
                }
                if($format eq "BED"){
                    print $seq."\t".$start."\t".$end."\t.\t.\t".$strand."\n";
                }
                elsif($format eq "GFF"){
                    print $seq."\tfuzznuc\tmisc_feature\t".$start."\t".$end."\t".$mismatch."\t".$strand."\t.\tid \"".$pattern."\"; site \"".$site."\";\n";
                }
        }
}
```

Parse fuzznuc output to BED format

- cat chr2.fuzznuc | grep CGG | grep -v Pattern | tr ' ' '-' | cut -f2,4 -d '-' | tr '-' ’\t’
- Cat chr2.fuzznuc l perl parseEmboss.pl BED
- Or after making parseEmboss.pl script executable
  - chmod u=rwx
  - Cat fuzznuc_output l parseEmboss.pl BED
  - Because perl script starts with #!/usr/bin/perl