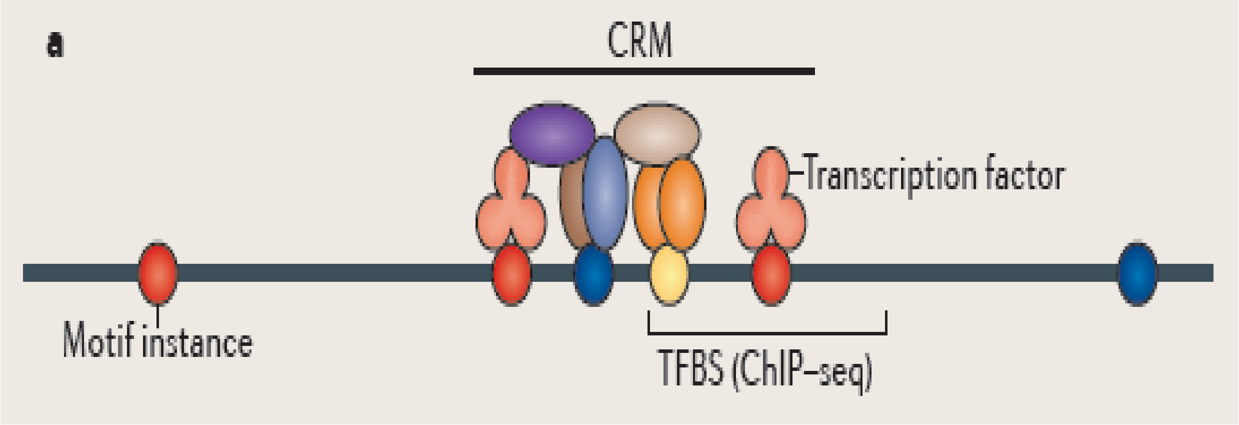
Genomic approaches towards finding *cis*-regulatory modules (CRM) in animals

Matthew I. Omoruyi

21.01.2013

Introduction

CRM is a stretch of DNA, usually 100 – 1000 DNA base pairs in length, where a number of transcription factors can bind and regulate expression of nearby genes



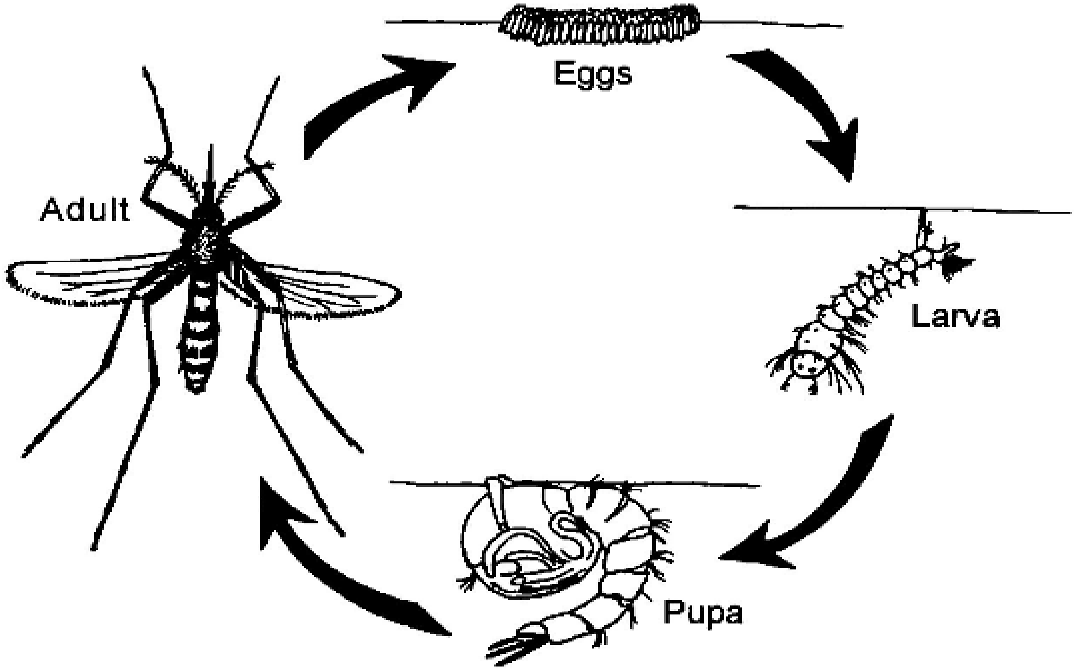
Introduction

They are typically located on the same DNA as the gene they control (cis)

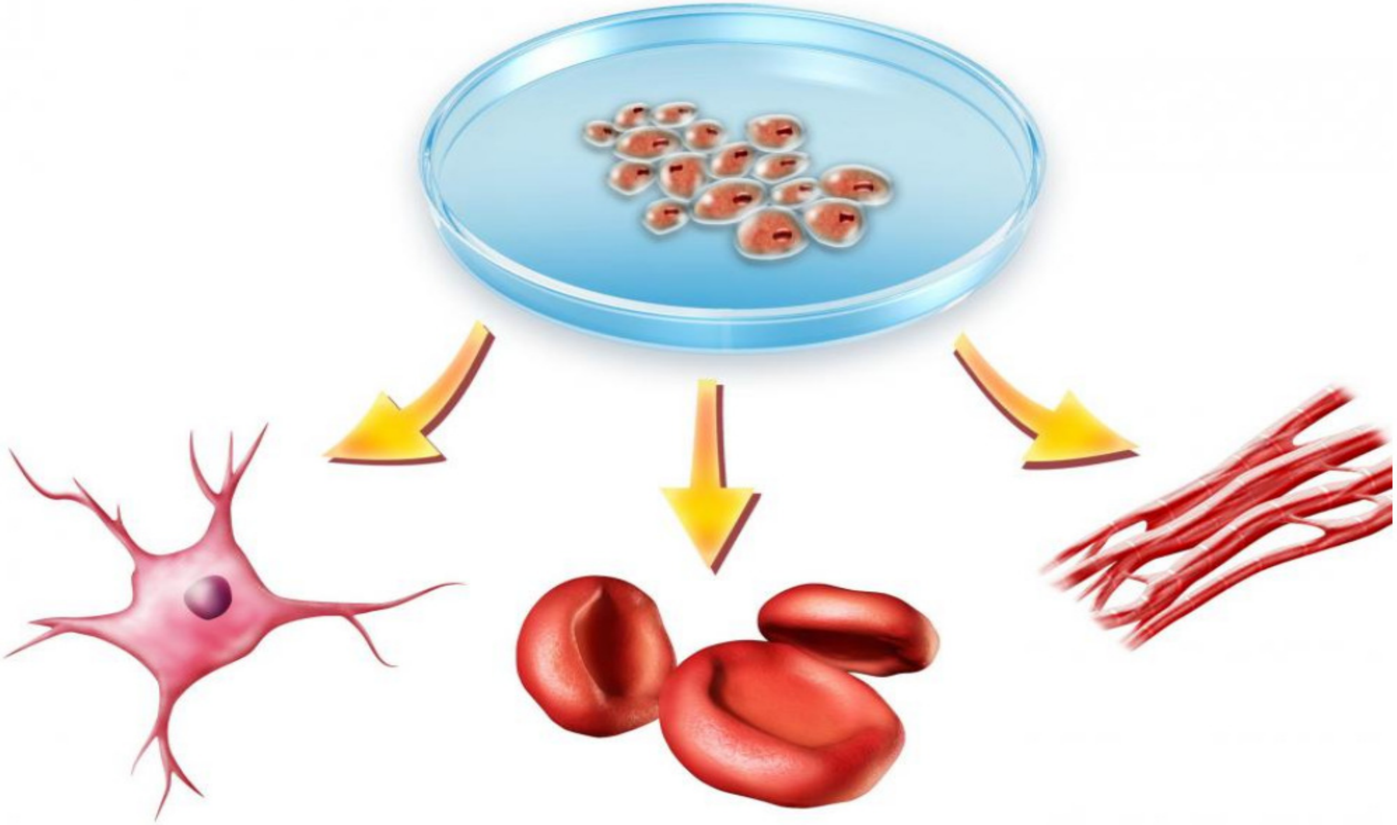
CRM includes, but not limited to the following

- ❖ Locus control regions
- ❖ Promoters
- ❖ Enhancers
- ❖ Silencers
- ❖ Boundary
- ❖ Control elements and
- ❖ Other modulators

Typical examples in understanding gene expression



The development of animals from zygote to adults requires the expression of a specific set of genes at each developmental stage



The differentiation of cells into distinct tissues and organs also requires the expression of a specific set of genes in each cell types

Gene expression

Importance of CRM

1. Expand our understanding of biology
2. Application in medicine (susceptibility to diseases)
3. Proper understanding of evolution

Methods used in predicting CRM in animals

1. Searching genomic DNA for clusters of motifs that are needed for the specific binding of transcription factors
2. Comparing homologous, non-coding DNA sequences between related species
3. Direct assays for DNA sequences with epigenetic features that are characteristic of regulatory regions

Information to help understand the methods

http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html

- The whole-genome sequences for over 85 microorganisms
- Humans, and a handful of other eukaryotic organisms

Combination of results from database similarity searches and gene-predicting algorithms to identify coding sequences with good but not complete accuracy

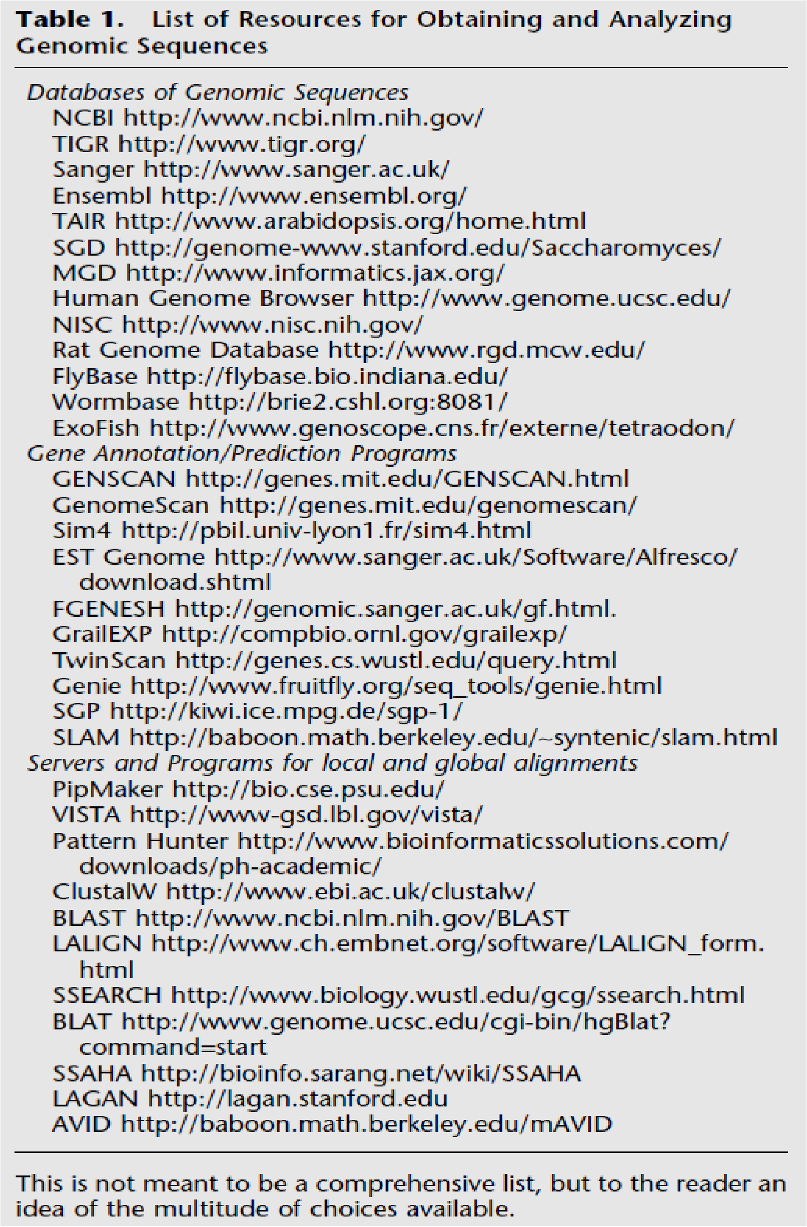
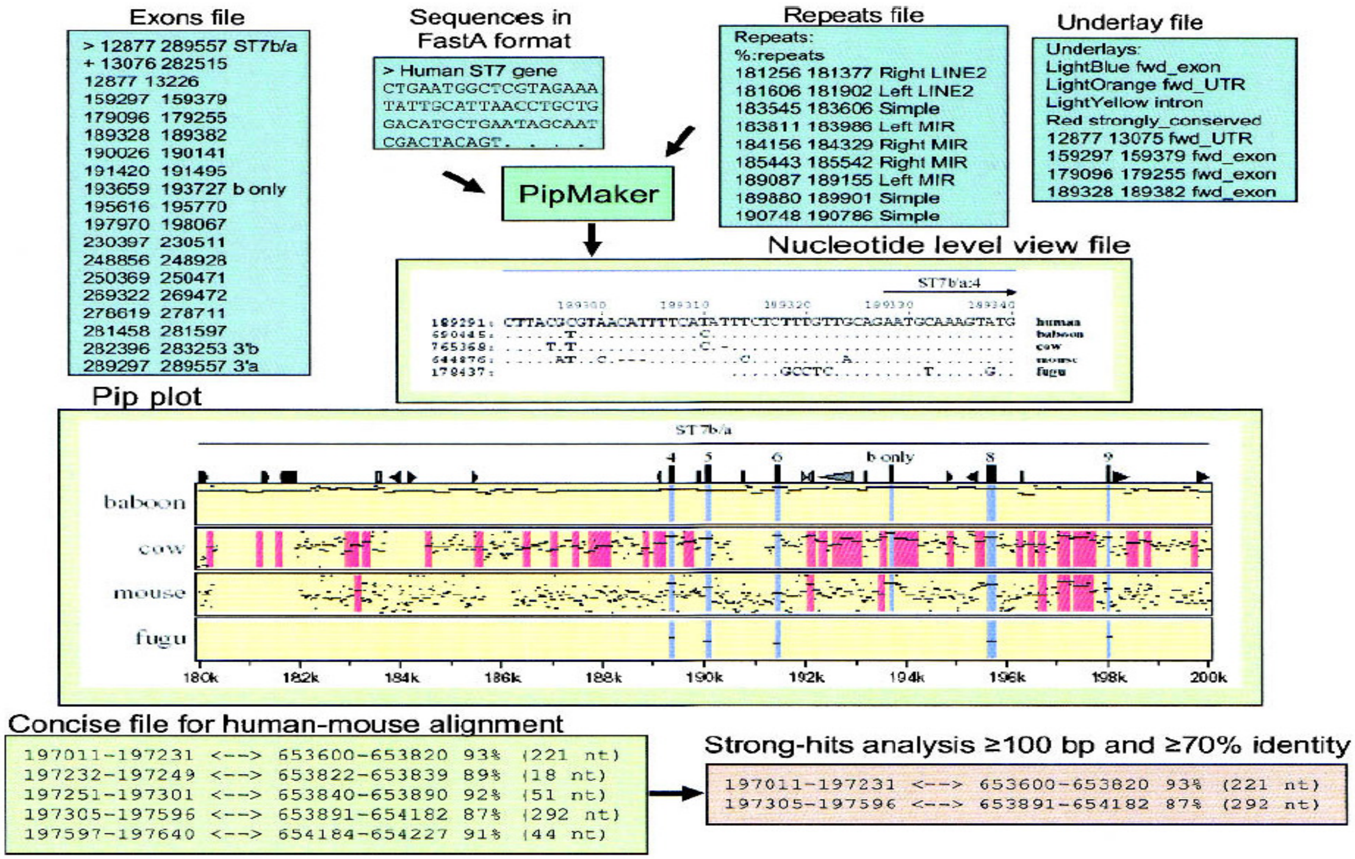


Table 1. List of Resources for Obtaining and Analyzing Genomic Sequences

Databases of Genomic Sequences	
NCBI	http://www.ncbi.nlm.nih.gov/
TIGR	http://www.tigr.org/
Sanger	http://www.sanger.ac.uk/
Ensembl	http://www.ensembl.org/
TAIR	http://www.arabidopsis.org/home.html
SGD	http://genome-www.stanford.edu/Saccharomyces/
MGD	http://www.informatics.jax.org/
Human Genome Browser	http://www.genome.ucsc.edu/
NISC	http://www.nisc.nih.gov/
Rat Genome Database	http://www.rgd.mcw.edu/
FlyBase	http://flybase.bio.indiana.edu/
Wormbase	http://brie2.cshl.org:8081/
ExoFish	http://www.genoscope.cns.fr/externe/tetraodon/
Gene Annotation/Prediction Programs	
GENSCAN	http://genes.mit.edu/GENSCAN.html
GenomeScan	http://genes.mit.edu/genomescan/
Sim4	http://pbil.univ-lyon1.fr/sim4.html
EST Genome	http://www.sanger.ac.uk/Software/Alfresco/download.shtml
FGENESH	http://genomic.sanger.ac.uk/gf.html.
GrailEXP	http://compbio.ornl.gov/grailexp/
TwinScan	http://genes.cs.wustl.edu/query.html
Genie	http://www.fruitfly.org/seq_tools/genie.html
SGP	http://kiwi.ice.mpg.de/sgp-1/
SLAM	http://baboon.math.berkeley.edu/~syntenic/slam.html
Servers and Programs for local and global alignments	
PipMaker	http://bio.cse.psu.edu/
VISTA	http://www-gsd.lbl.gov/vista/
Pattern Hunter	http://www.bioinformaticssolutions.com/downloads/ph-academic/
ClustalW	http://www.ebi.ac.uk/clustalw/
BLAST	http://www.ncbi.nlm.nih.gov/BLAST
LALIGN	http://www.ch.embnet.org/software/LALIGN_form.html
SSEARCH	http://www.biology.wustl.edu/gcg/ssearch.html
BLAT	http://www.genome.ucsc.edu/cgi-bin/hgBlat?command=start
SSAHA	http://bioinfo.sarang.net/wiki/SSAHA
LAGAN	http://lagan.stanford.edu
AVID	http://baboon.math.berkeley.edu/mAVID

This is not meant to be a comprehensive list, but to the reader an idea of the multitude of choices available.

Softwares



Exons file

```
> 12877 289557 ST7b/a
+ 13076 282515
12877 13226
159297  159379
179096  179255
189328  189382
190026  190141
191420  191495
193659  193727 b only
195616  195770
197970  198067
230397  230511
248856  248928
250369  250471
269322  269472
278619  278711
281458  281597
282396  283253 3'b
289297  289557 3'a
```

Sequences in FastA format

```
> Human ST7 gene
CTGAATGGCTCGTAGAAA
TATTGCATTAACCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. . . .
```

Repeats file

```
Repeats:
%:repeats
181256 181377 Right LINE2
181606 181902 Left LINE2
183545 183606 Simple
183811 183986 Left MIR
184156 184329 Right MIR
185443 185542 Right MIR
189087 189155 Left MIR
189880 189901 Simple
190748 190786 Simple
```

Underlay file

```
Underlays:
LightBlue fwd_exon
LightOrange fwd_UTR
LightYellow intron
Red strongly_conserved
12877 13075 fwd_UTR
159297 159379 fwd_exon
179096 179255 fwd_exon
189328 189382 fwd_exon
```

PipMaker

Nucleotide level view file

```
                                                              ST7b/a:4
             189300      189310      189320      189330      189340
189291: CTTACGCGTAACATTTTCATATTTCTCTTTGTTGCAGAATGCAAAGTATG  human
690445: ......T...........C...............................  baboon
765368: ....T.T...........C.-.............................  cow
644876: .....AT..C.---.........C.........A................  mouse
178437:                     .....GCCTC.........T.....G..  fugu
```

Pip plot

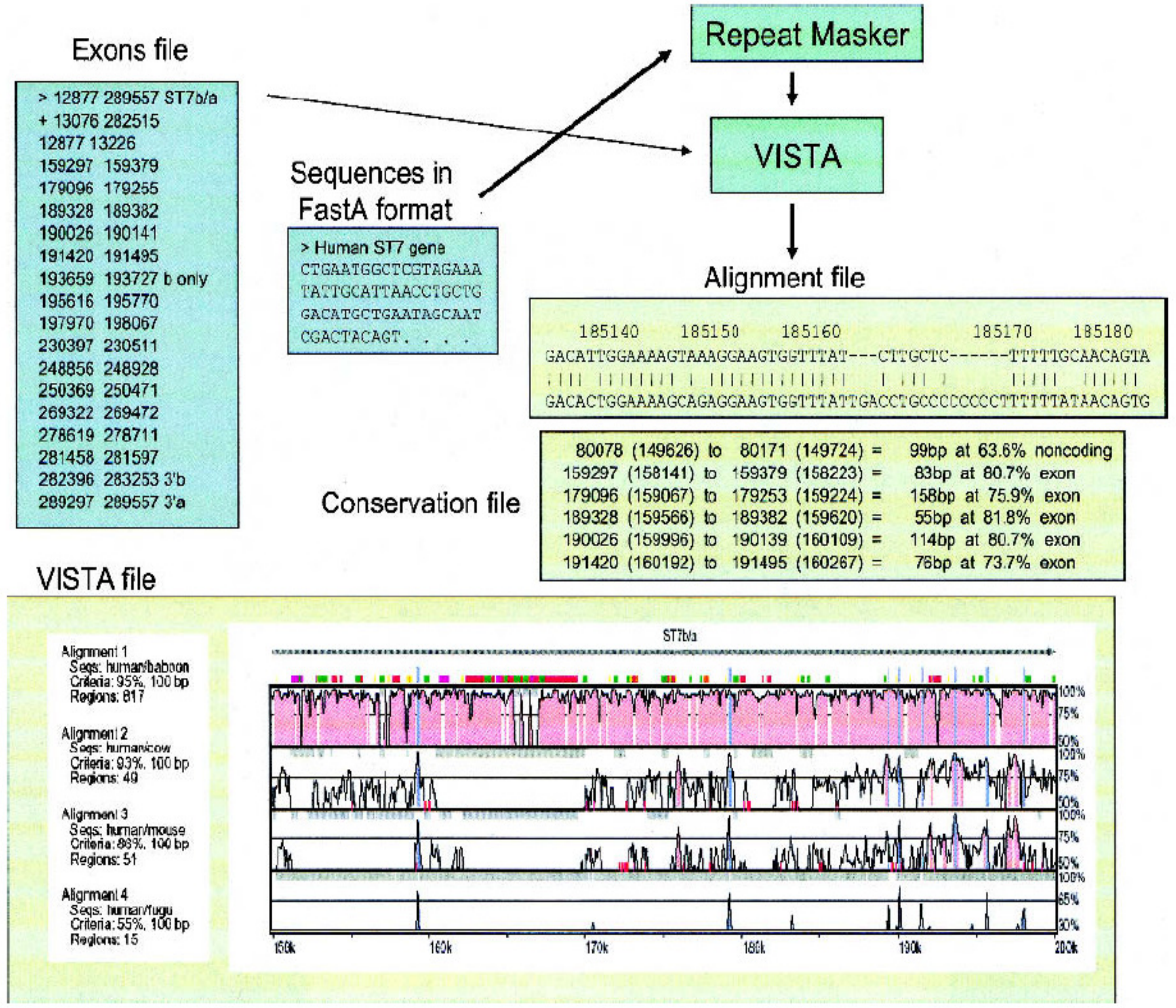
Concise file for human-mouse alignment

```
197011-197231 <--> 653600-653820 93% (221 nt)
197232-197249 <--> 653822-653839 89% (18 nt)
197251-197301 <--> 653840-653890 92% (51 nt)
197305-197596 <--> 653891-654182 87% (292 nt)
197597-197640 <--> 654184-654227 91% (44 nt)
```

Strong-hits analysis ≥100 bp and ≥70% identity

```
197011-197231 <--> 653600-653820 93% (221 nt)
197305-197596 <--> 653891-654182 87% (292 nt)
```

Figure 5. PipMaker: input and output files. Files for submission to PipMaker include Sequences (required), Repeats (recommended), Underlay (optional), and Exon annotations (optional). The Repeats file is made by simplifying RepeatMasker output using the program rmask2repeats (from the PipTools program package). The simplified version is shown. The coordinates in the Repeats file and Underlay file correspond to the coordinates in the Pip plot. PipMaker generates three multiple output files. The Pip plot shown is a subregion of the human *ST7* interval compared with the orthologous baboon, cow, mouse, or fugu sequences. Each panel represents a pairwise comparison between human sequence and that of the indicated species. Each alignment consists of a series of horizontal lines that represent the gap-free aligning segments that are graphed on a vertical



Exons file

```
> 12877 289557 ST7b/a
+ 13076 282515
12877 13226
159297  159379
179096  179255
189328  189382
190026  190141
191420  191495
193659  193727 b only
195616  195770
197970  198067
230397  230511
248856  248928
250369  250471
269322  269472
278619  278711
281458  281597
282396  283253 3'b
289297  289557 3'a
```

Sequences in FastA format

```
> Human ST7 gene
CTGAATGGCTCGTAGAAA
TATTGCATTAACCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. . . .
```

Repeat Masker

VISTA

Alignment file

```
   185140     185150     185160                185170     185180
GACATTGGAAAAGTAAAGGAAGTGGTTTAT---CTTGCTC------TTTTTGCAACAGTA
 |||  ||||||| | ||||||||||||||    | |||        |||||  ||||||
GACACTGGAAAAGCAGAGGAAGTGGTTTATTGACCTGCCCCCCCCTTTTTTATAACAGTG
```

Conservation file

```
 80078 (149626) to  80171 (149724) =   99bp at 63.6% noncoding
159297 (158141) to 159379 (158223) =   83bp at 80.7% exon
179096 (159067) to 179253 (159224) =  158bp at 75.9% exon
189328 (159566) to 189382 (159620) =   55bp at 81.8% exon
190026 (159996) to 190139 (160109) =  114bp at 80.7% exon
191420 (160192) to 191495 (160267) =   76bp at 73.7% exon
```

VISTA file

Alignment 1
Seqs: human/baboon
Criteria: 95%, 100 bp
Regions: 817

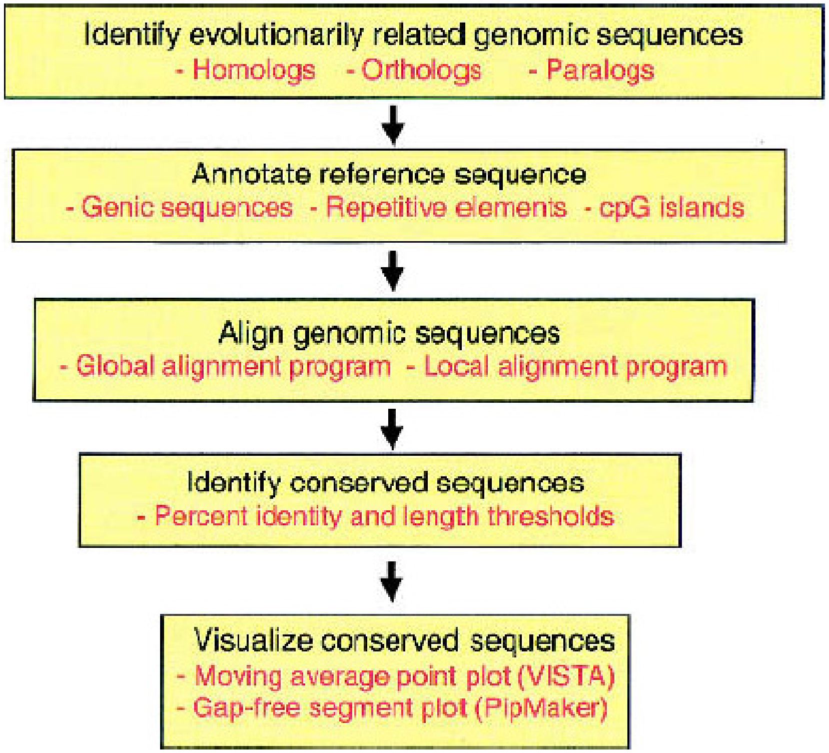
Alignment 2
Seqs: human/cow
Criteria: 93%, 100 bp
Regions: 49

Alignment 3
Seqs: human/mouse
Criteria: 86%, 100 bp
Regions: 51

Alignment 4
Seqs: human/fugu
Criteria: 55%, 100 bp
Regions: 15

Figure 7. VISTA: input and output files. Files for submission to VISTA include Sequences (required) and Exons (optional). Repeats are masked in the reference sequence using RepeatMasker upon its submission to VISTA. VISTA generates three output files. The VISTA plot shown here is a subregion of the human *ST7* interval compared with the orthologous baboon, cow, mouse, or fugu sequences. Conserved sequences represented as peaks [noncoding (red) and coding (blue)] are shown relative to their positions in the human genome (horizontal axes), and their percent identities (50%–100%) are indicated on the vertical axes. The locations of *ST7* exons are indicated by tall blue rectangles, and the direction of transcription is indicated by a horizontal arrow. The locations of repetitive elements are indicated by color rectangles (see Suppl. Fig. 2). The

How is this done?



Identify evolutionarily related genomic sequences
- Homologs - Orthologs - Paralogs

Annotate reference sequence
- Genic sequences - Repetitive elements - cpG islands

Align genomic sequences
- Global alignment program - Local alignment program

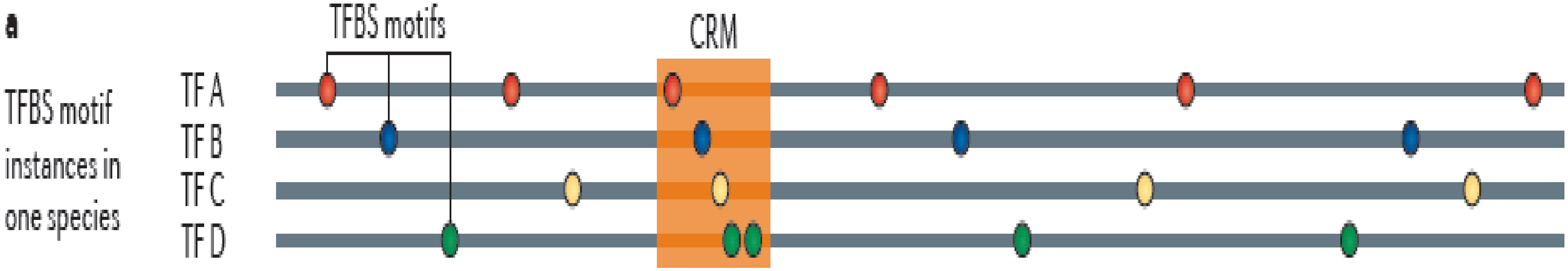
Identify conserved sequences
- Percent identity and length thresholds

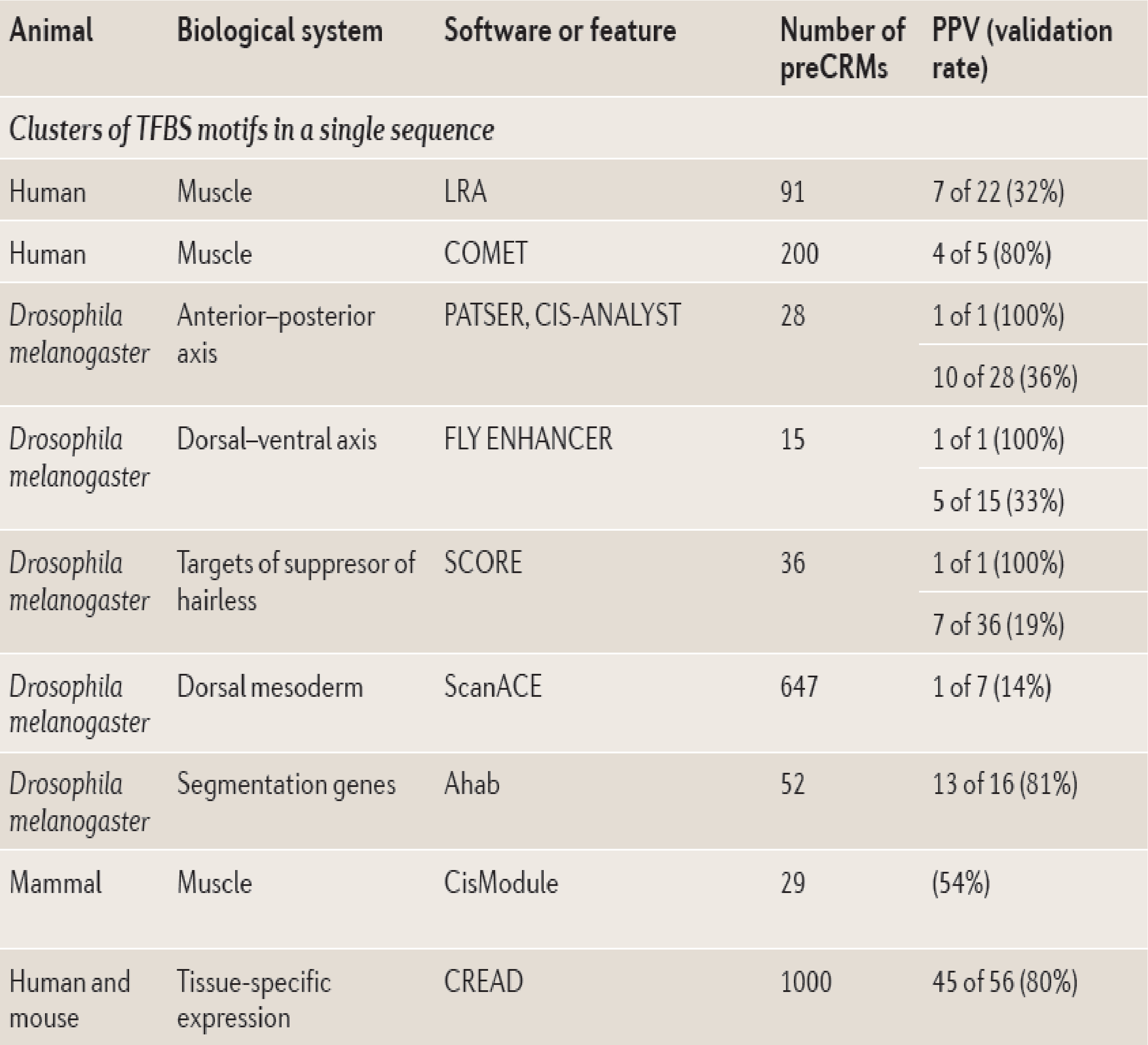
Visualize conserved sequences
- Moving average point plot (VISTA)
- Gap-free segment plot (PipMaker)

Figure 1. Multistep process of comparative sequence analysis. Evo-

Methods used in predicting CRM in animals

1. **Searching genomic DNA for clusters of motifs that are needed for the specific binding of transcription factors**





Animal	Biological system	Software or feature	Number of preCRMs	PPV (validation rate)
Clusters of TFBS motifs in a single sequence				
Human	Muscle	LRA	91	7 of 22 (32%)
Human	Muscle	COMET	200	4 of 5 (80%)
Drosophila melanogaster	Anterior–posterior axis	PATSER, CIS-ANALYST	28	1 of 1 (100%)
				10 of 28 (36%)
Drosophila melanogaster	Dorsal–ventral axis	FLY ENHANCER	15	1 of 1 (100%)
				5 of 15 (33%)
Drosophila melanogaster	Targets of suppresor of hairless	SCORE	36	1 of 1 (100%)
				7 of 36 (19%)
Drosophila melanogaster	Dorsal mesoderm	ScanACE	647	1 of 7 (14%)
Drosophila melanogaster	Segmentation genes	Ahab	52	13 of 16 (81%)
Mammal	Muscle	CisModule	29	(54%)
Human and mouse	Tissue-specific expression	CREAD	1000	45 of 56 (80%)

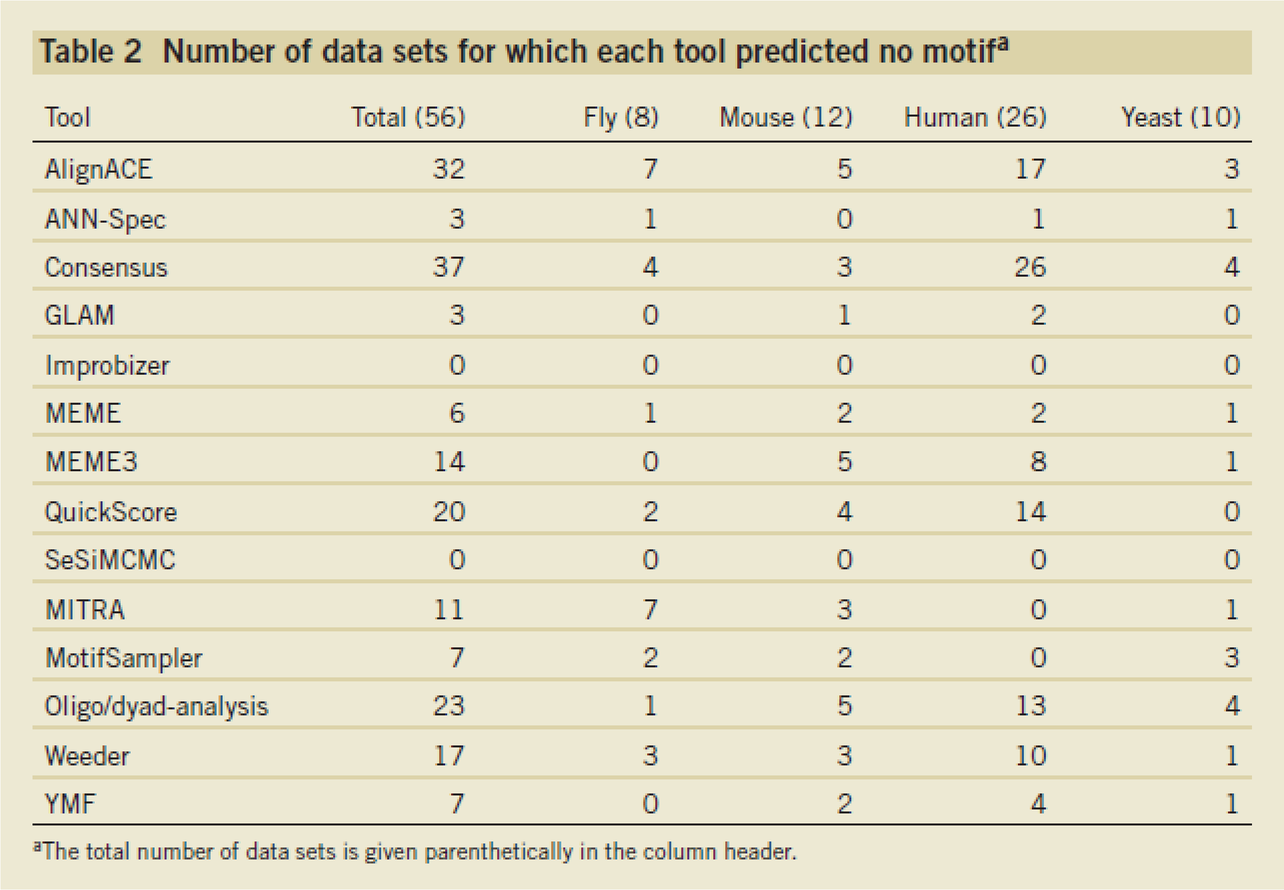
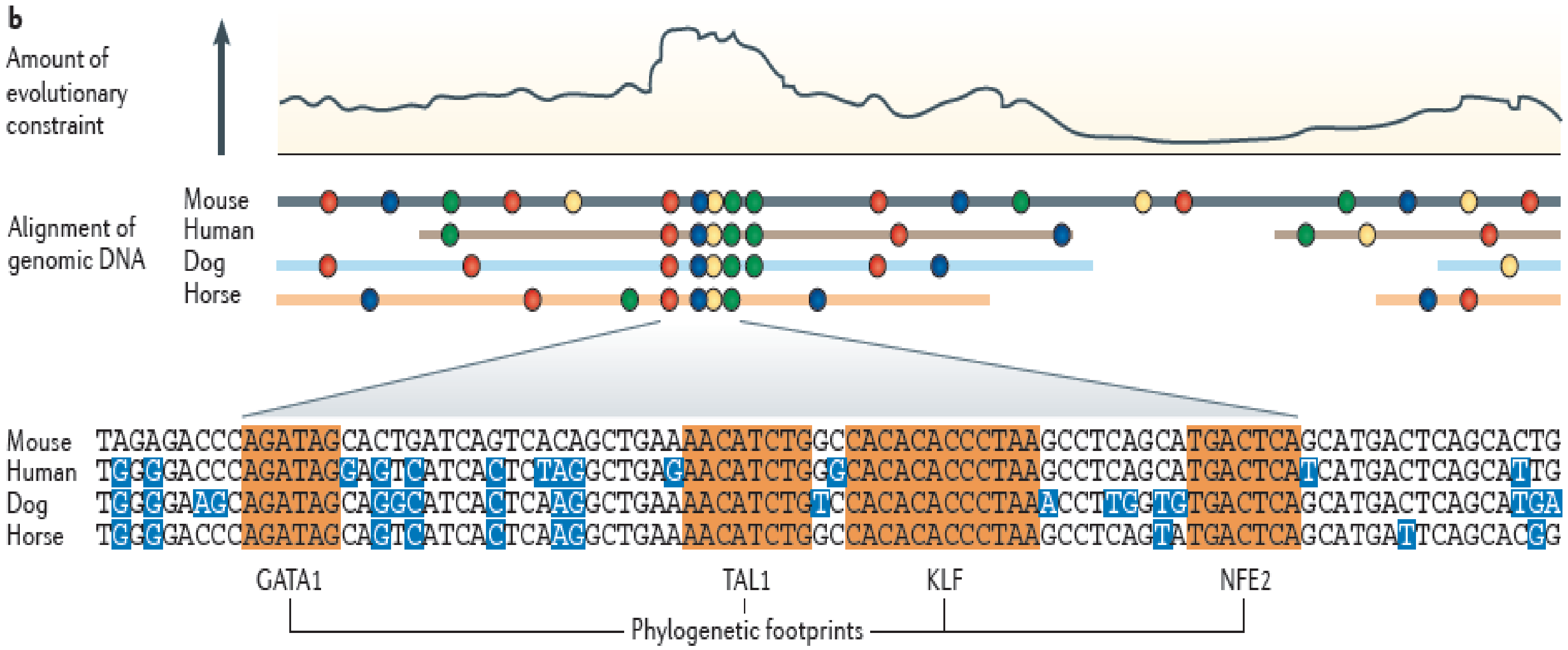


Table 2 Number of data sets for which each tool predicted no motif[a]

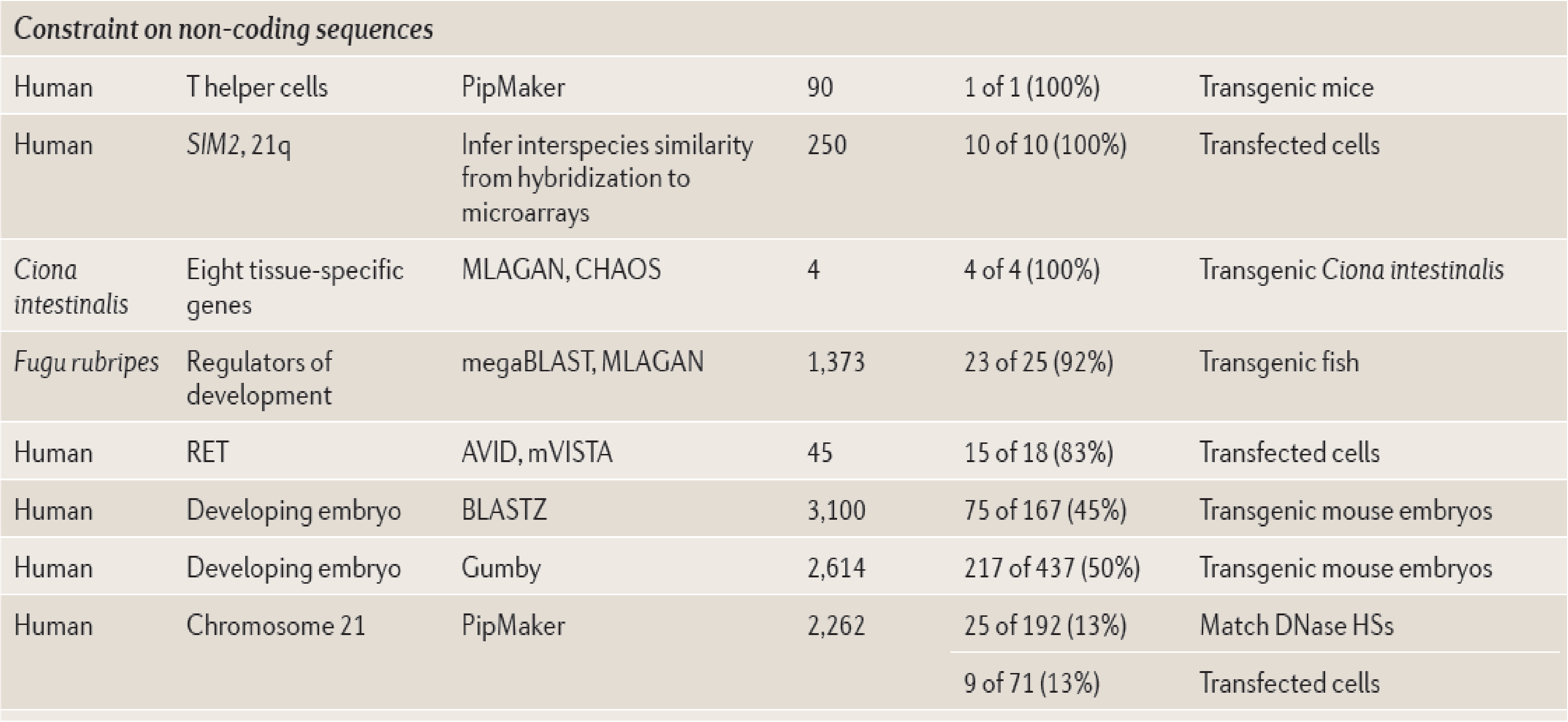
Tool	Total (56)	Fly (8)	Mouse (12)	Human (26)	Yeast (10)
AlignACE	32	7	5	17	3
ANN-Spec	3	1	0	1	1
Consensus	37	4	3	26	4
GLAM	3	0	1	2	0
Improbizer	0	0	0	0	0
MEME	6	1	2	2	1
MEME3	14	0	5	8	1
QuickScore	20	2	4	14	0
SeSiMCMC	0	0	0	0	0
MITRA	11	7	3	0	1
MotifSampler	7	2	2	0	3
Oligo/dyad-analysis	23	1	5	13	4
Weeder	17	3	3	10	1
YMF	7	0	2	4	1

[a] The total number of data sets is given parenthetically in the column header.

2. Comparing homologous, non-coding DNA sequences between related species


b
Amount of evolutionary constraint
Alignment of genomic DNA
Mouse
Human
Dog
Horse
Mouse TAGAGACCCAGATAGCACTGATCAGTCACAGCTGAAAACATCTGGCCACACACCCTAAGCCTCAGCATGACTCAGCATGACTCAGCACTG
Human TGGGACCCAGATAGGAGTCATCACTCTAGGCTGAGAACATCTGGGCACACACCCTAAGCCTCAGCATGACTCATCATGACTCAGCATTG
Dog TGGGAAGCAGATAGCAGGCATCACTCAAGGCTGAAAACATCTGTCCACACACCCTAAACCTTGGTGTGACTCAGCATGACTCAGCATGA
Horse TGGGACCCAGATAGCAGTCATCACTCAAGGCTGAAAACATCTGGCCACACACCCTAAGCCTCAGTATGACTCAGCATGATTCAGCACGG
GATA1
TAL1
KLF
NFE2
Phylogenetic footprints

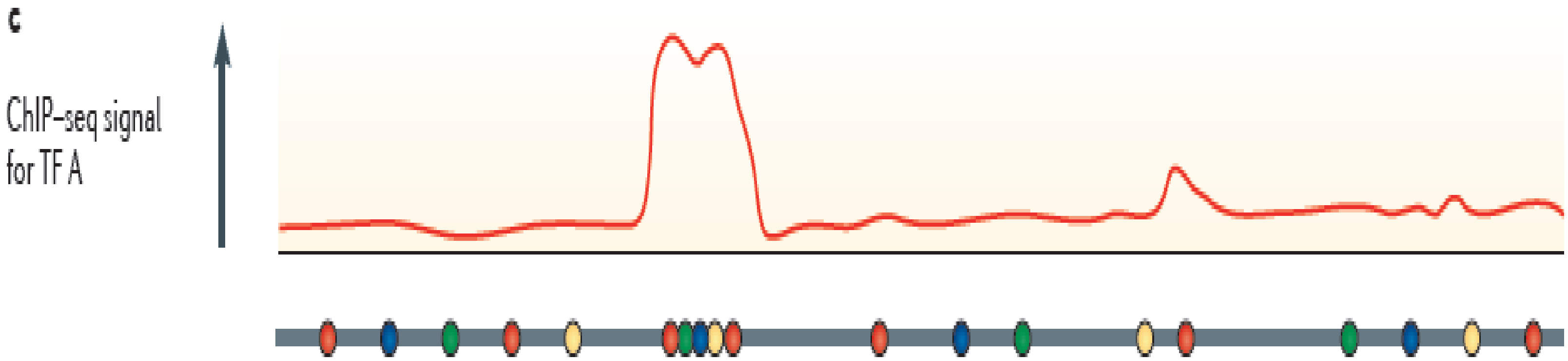
By comparing the genomic sequences of species at different evolutionary distances, one can identify coding sequences and conserved non-coding sequences with regulatory functions and determine which sequence are unique for a given specie



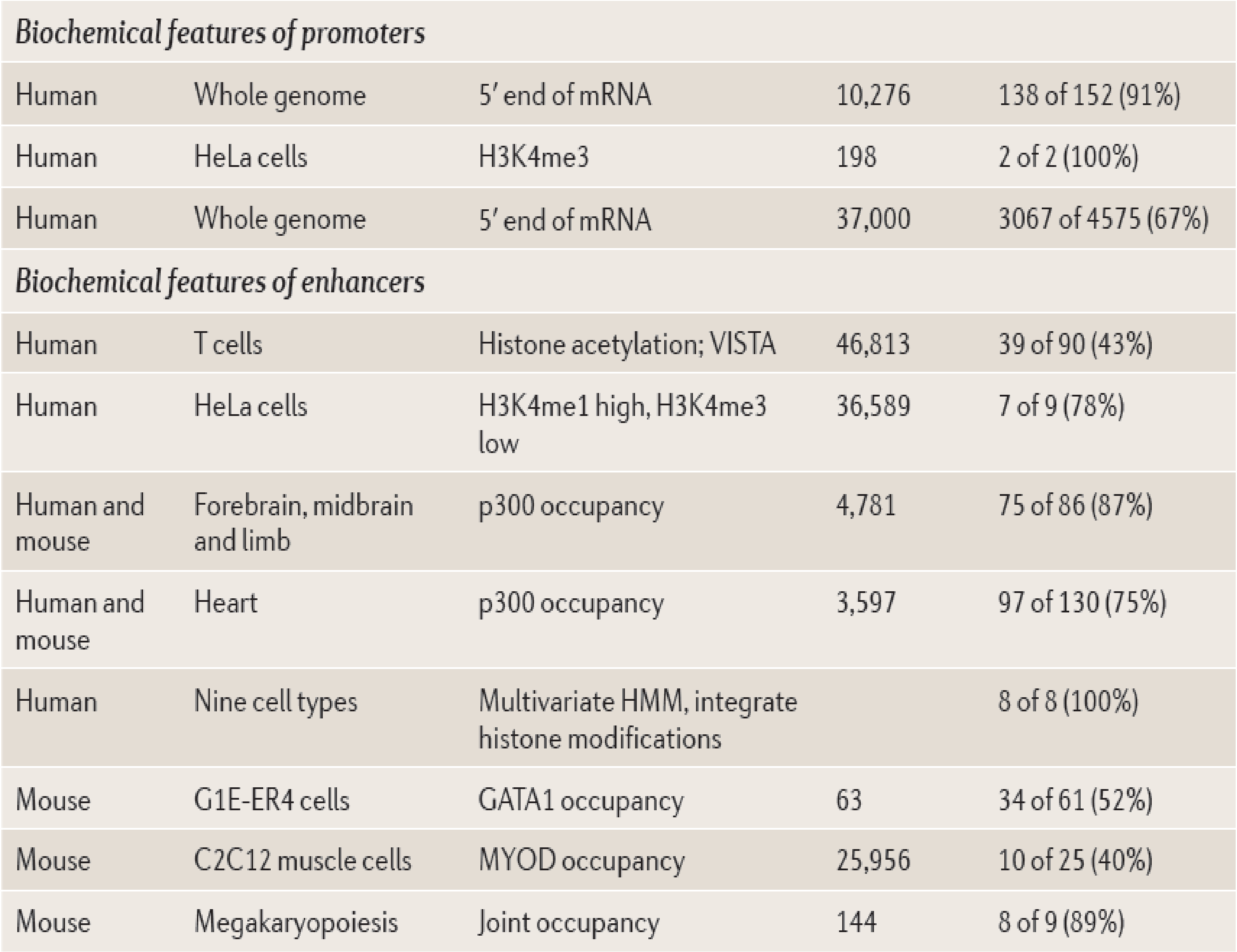
Constraint on non-coding sequences

Human	T helper cells	PipMaker	90	1 of 1 (100%)	Transgenic mice
Human	*SIM2*, 21q	Infer interspecies similarity from hybridization to microarrays	250	10 of 10 (100%)	Transfected cells
Ciona intestinalis	Eight tissue-specific genes	MLAGAN, CHAOS	4	4 of 4 (100%)	Transgenic *Ciona intestinalis*
Fugu rubripes	Regulators of development	megaBLAST, MLAGAN	1,373	23 of 25 (92%)	Transgenic fish
Human	RET	AVID, mVISTA	45	15 of 18 (83%)	Transfected cells
Human	Developing embryo	BLASTZ	3,100	75 of 167 (45%)	Transgenic mouse embryos
Human	Developing embryo	Gumby	2,614	217 of 437 (50%)	Transgenic mouse embryos
Human	Chromosome 21	PipMaker	2,262	25 of 192 (13%)	Match DNase HSs
				9 of 71 (13%)	Transfected cells

3.
Direct assays for DNA sequences with epigenetic features that are characteristic of regulatory regions



Epigenetic features are reversible features on a cell's DNA that affect gene expression without altering DNA
It is based on high-throughput sequencing and mapping to reference genomes



Biochemical features of promoters				
Human	Whole genome	5′ end of mRNA	10,276	138 of 152 (91%)
Human	HeLa cells	H3K4me3	198	2 of 2 (100%)
Human	Whole genome	5′ end of mRNA	37,000	3067 of 4575 (67%)
Biochemical features of enhancers				
Human	T cells	Histone acetylation; VISTA	46,813	39 of 90 (43%)
Human	HeLa cells	H3K4me1 high, H3K4me3 low	36,589	7 of 9 (78%)
Human and mouse	Forebrain, midbrain and limb	p300 occupancy	4,781	75 of 86 (87%)
Human and mouse	Heart	p300 occupancy	3,597	97 of 130 (75%)
Human	Nine cell types	Multivariate HMM, integrate histone modifications		8 of 8 (100%)
Mouse	G1E-ER4 cells	GATA1 occupancy	63	34 of 61 (52%)
Mouse	C2C12 muscle cells	MYOD occupancy	25,956	10 of 25 (40%)
Mouse	Megakaryopoiesis	Joint occupancy	144	8 of 9 (89%)

Pros and cons

Searching genomic DNA for clusters of motifs that are needed for the specific binding of transcription factors

1. Partial success under favourable conditions
2. Only small subset of CRMs is likely to be discovered by extreme evolutionary constraint
3. Does not work equally in all tissues
4. No sufficient specificity
5. Not designed to find CRMs that are active in only 1 specie or that are changing in a lineage specific manner

Comparing homologous, non-coding DNA sequences between related species

1. Partial success under favourable conditions
2. Only small subset of CRMs is likely to be discovered by extreme evolutionary constraint
3. Does not work equally in all tissues
4. No sufficient specificity
5. Not designed to find CRMs that are active in only 1 specie or that are changing in a lineage specific manner

Direct assays for DNA sequences with epigenetic features that are characteristic of regulatory regions

1. Epigenetic marks must be mapped in tissues and at times of development that are informative to the question at hand
2. May be bias.

Summary/conclusion

- *Cis*-regulatory modules are DNA sequence required to regulate gene expression
- The genome of both prokaryotes and eukaryotes are available in a vast number of databases
- These databases are used to predict the DNA sequence required for gene expression by different methods

Summary/conclusion

Given the limitations of methods based on sequence motifs and comparative genomics, direct measurement of diagnostic epigenetic features should lead to improved methods for CRM prediction. Particular epigenetic features are highly correlated with CRMs, and progress is being made in finding combinations of these features that may distinguish different types of CRM.

THANK YOU!