MOL3021 / MOL3022

Lecture 05

Gene regulation

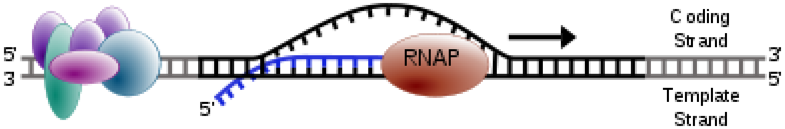
Based on slides from Finn Drabløs

Learning goals

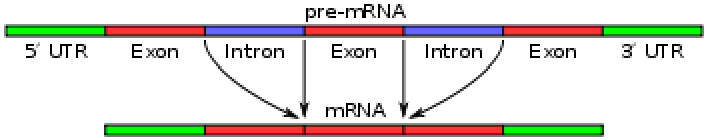
- Know the main concepts of gene regulation by transcription factors
- Understand how transcription factor binding sites can be identified with ChIP-seq
- Understand how transcription factor binding sites can be predicted computationally
- Know some main tools and resources for analyzing transcription factor data

Regulation from gene to protein

- Transcript produced from DNA by polymerase II (Pol II)

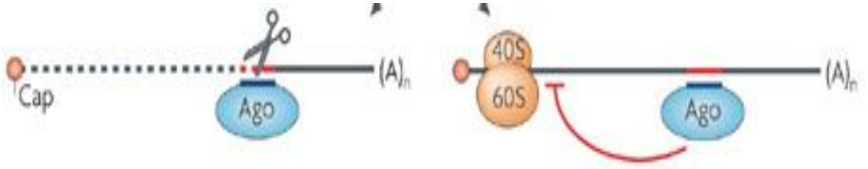


- Transcript spliced and processed into messenger RNA (mRNA)



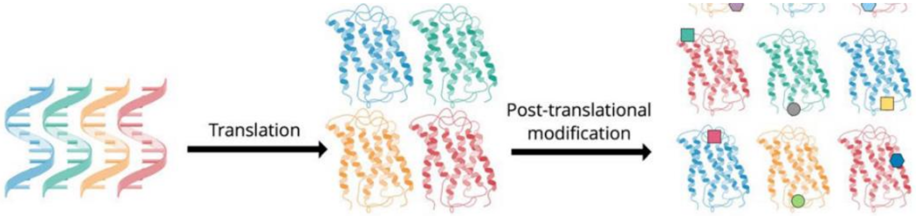
- mRNA transported out of the nucleus to the cytoplasm

- Adjustment of mRNA levels by non-coding RNA in the cytoplasm



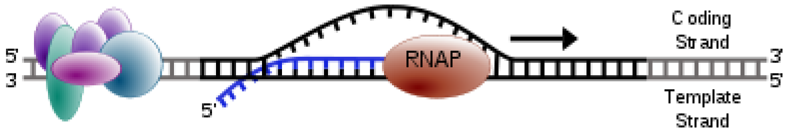
- Translation of mRNA to protein

- Regulation of protein function, by post translational modifications (PTMs)

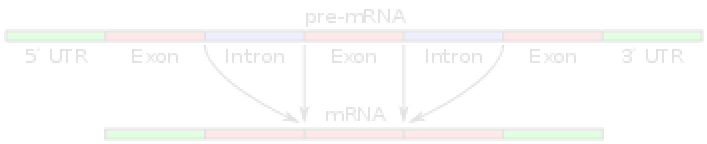


Regulation from gene to protein

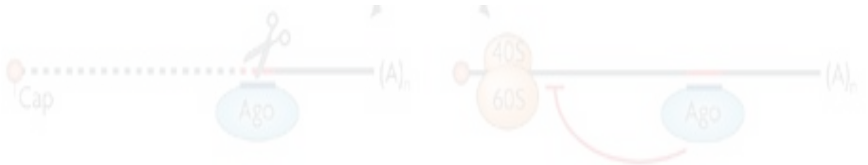
- **Transcript produced from DNA by polymerase II (Pol II)**



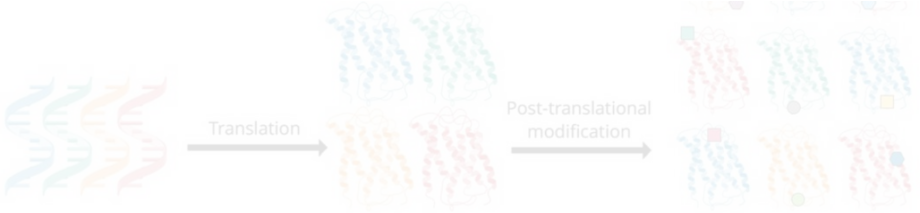
- Transcript spliced and processed into messenger RNA (mRNA)



- mRNA transported out of the nucleus to the cytoplasm
- Adjustment of mRNA levels by non-coding RNA in the cytoplasm



- Translation of mRNA to protein
- Regulation of protein function, by post translational modifications (PTMs)



The functional genome

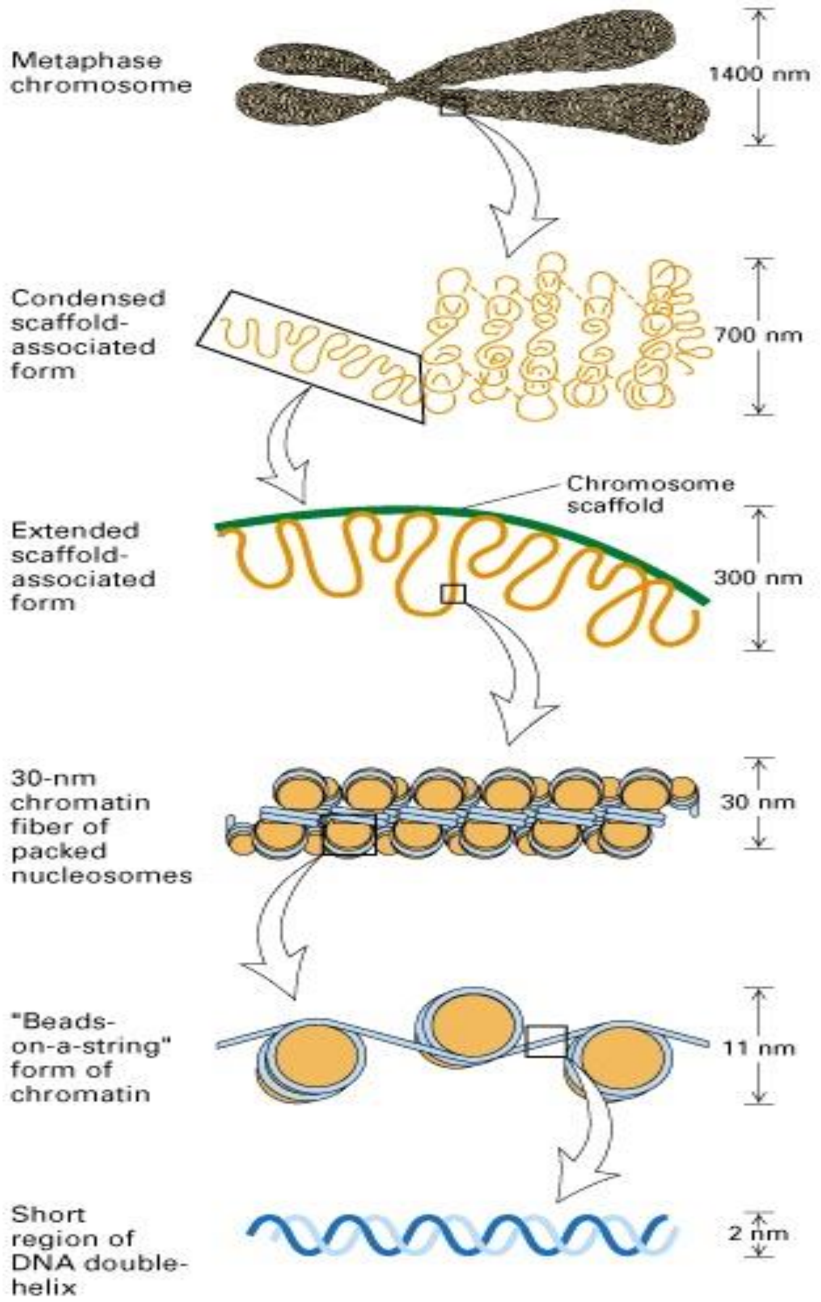
- Protein coding sequence.
  - constitutes 2% of the human genome (exons)
  - Function of the rest of the genome still uncharacterised

- Non-coding transcripts
  - Can arise from introns or be transcribed independently from other parts of the genome
  - Cytoplasm: Affect mRNA levels and translation rate
  - Nucleus: Regulate transcription from DNA
  - Many types and classes: miRNA, lncRNA, eRNA, pseudo-genes…

- **Regulatory elements**
  - **Genomic regions affecting the level of RNA transcribed from DNA**
  - **Affect both coding and non-coding transcripts.**
  - **Constitute a binding platform for transcription factor proteins (TF) (and other regulatory proteins),**
  - **Main purpose: regulate the recruitment and activity of the transcript producing Pol II complex.**

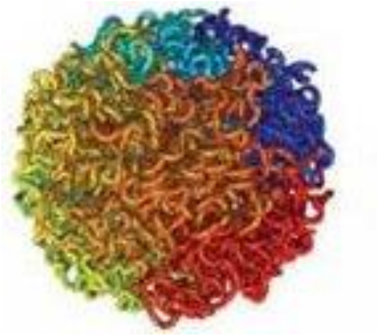
Chromatin structure

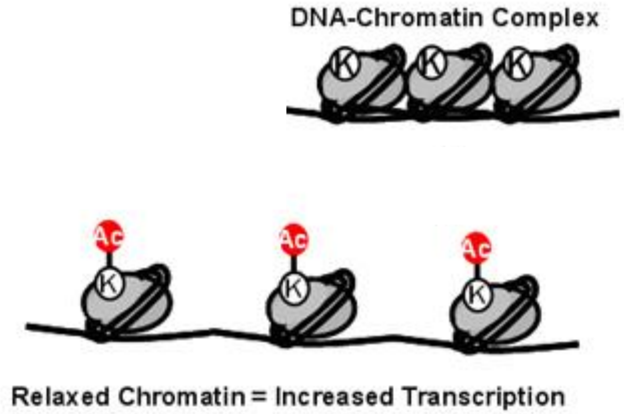
- Level of regulation in addition to DNA sequence
- Chromatin structure decides cell-type specificity by defining regulatory elements
- Interacts with transcription factors, regulatory proteins and RNA.
- Consist of DNA wrapped around nucleosomes
- Nucleosomes are again organized into higher order fibres and structures


Metaphase chromosome
1400 nm
Condensed scaffold-associated form
700 nm
Chromosome scaffold
Extended scaffold-associated form
300 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
"Beads-on-a-string" form of chromatin
11 nm
Short region of DNA double-helix
2 nm

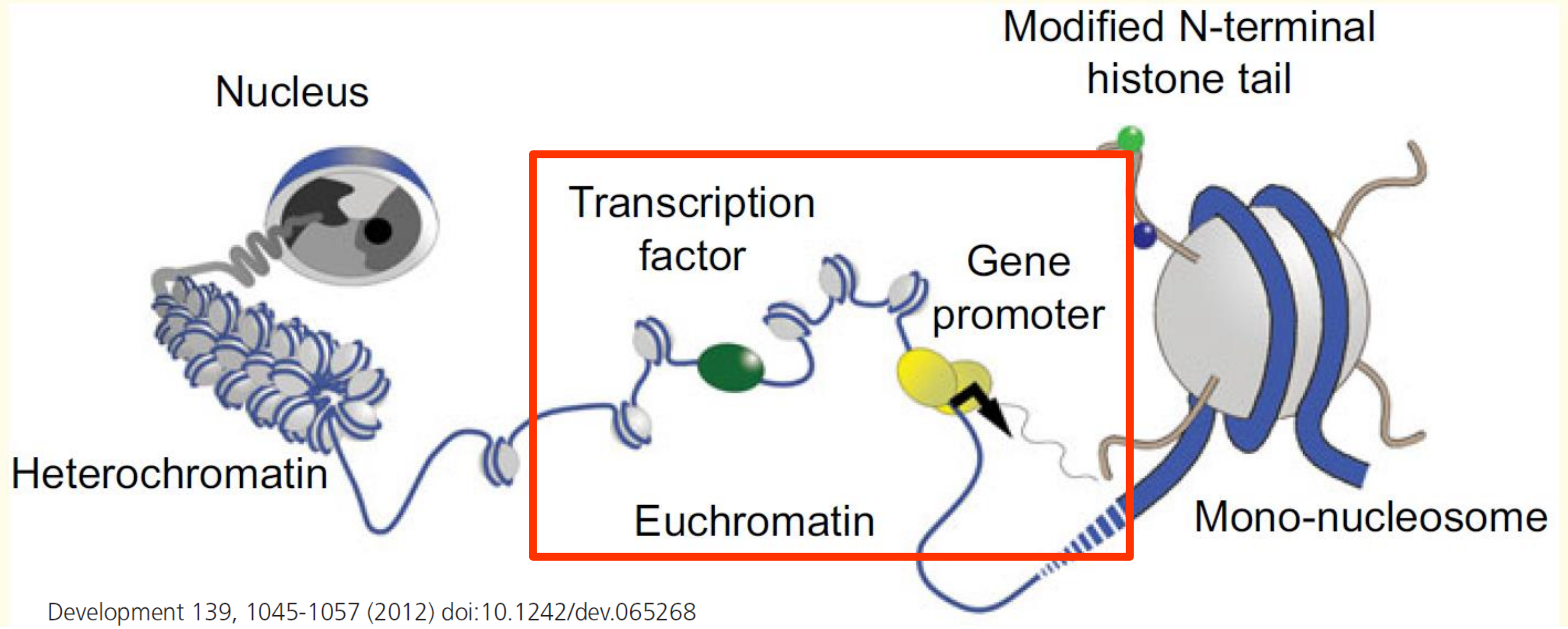
Chromatin structure

- Characteristic X-form of chromosomes only during mitosis (cell-division)
- Nucleosomes packaging define genomic chromatin compartments:
  - *Tightly packed (silent, inaccessible heterochromatin)*
  - *Loosely packed (active, accessible euchromatin)*
- Genomic active and silent regions vary between cell-types
  - *Foundation of cell-type specificity*
- Regulatory elements are generally associated with regions of active chromatin.

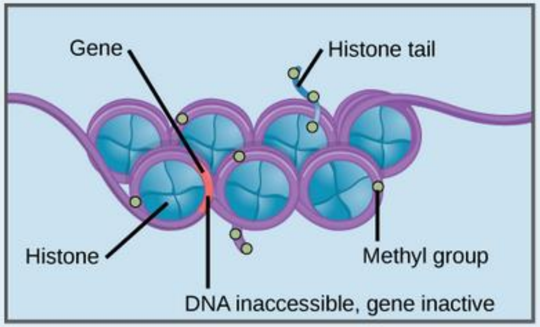




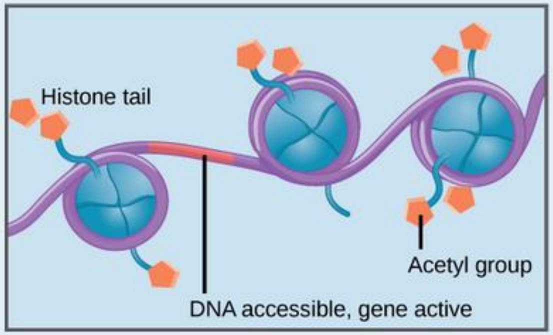
Organization – Regulatory Elements



Development 139, 1045-1057 (2012) doi:10.1242/dev.065268
© 2012. Published by The Company of Biologists Ltd



Methylation of DNA and histones causes nucleosomes to pack tightly together. Transcription factors cannot bind the DNA, and genes are not expressed.

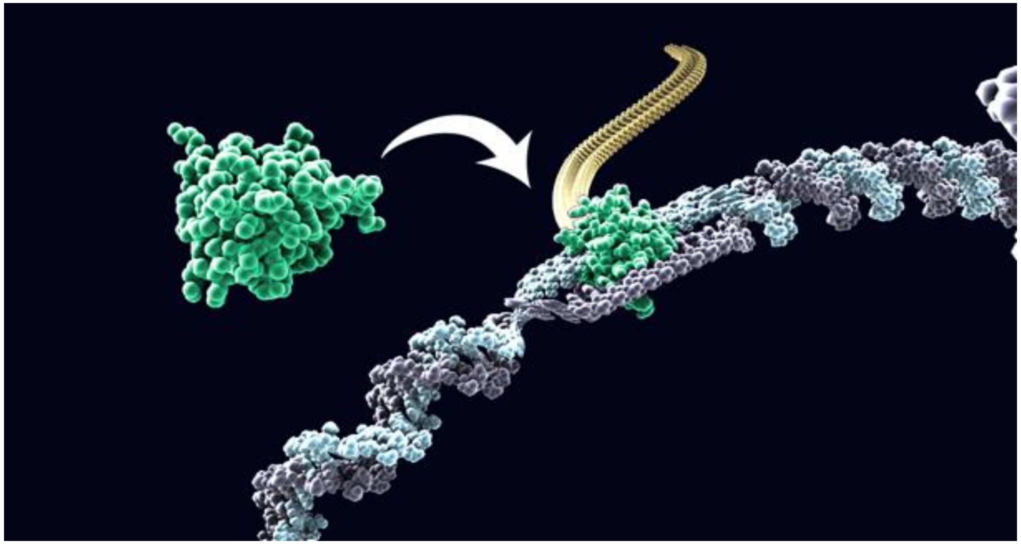


Histone acetylation results in loose packing of nucleosomes. Transcription factors can bind the DNA and genes are expressed.

http://cnx.org/contents/GFy_h8cu@10.53:rZudN6XP@2/Introduction

Transcription factors

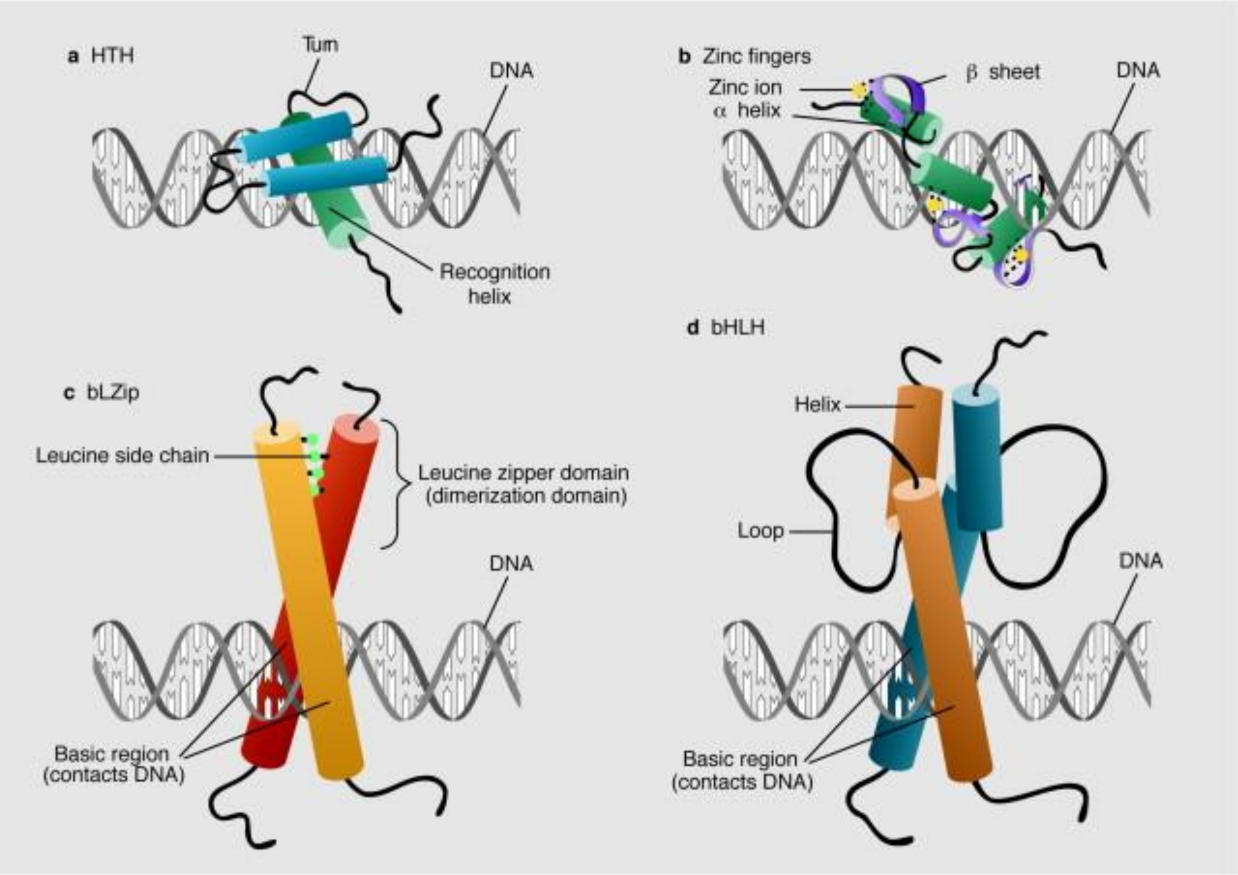
- Proteins translated in the cytoplasm (as other proteins),
- Transported back to the nuclueus to carry out their function on DNA.
- TFs bind directly to DNA in regulatory elements.
  - Bind to a 6-25 bp long **DNA motif**
- Current number estimate of TFs in human is 1500-2000.



Important TF classes

- **Most TFs belong to one of the main classes**

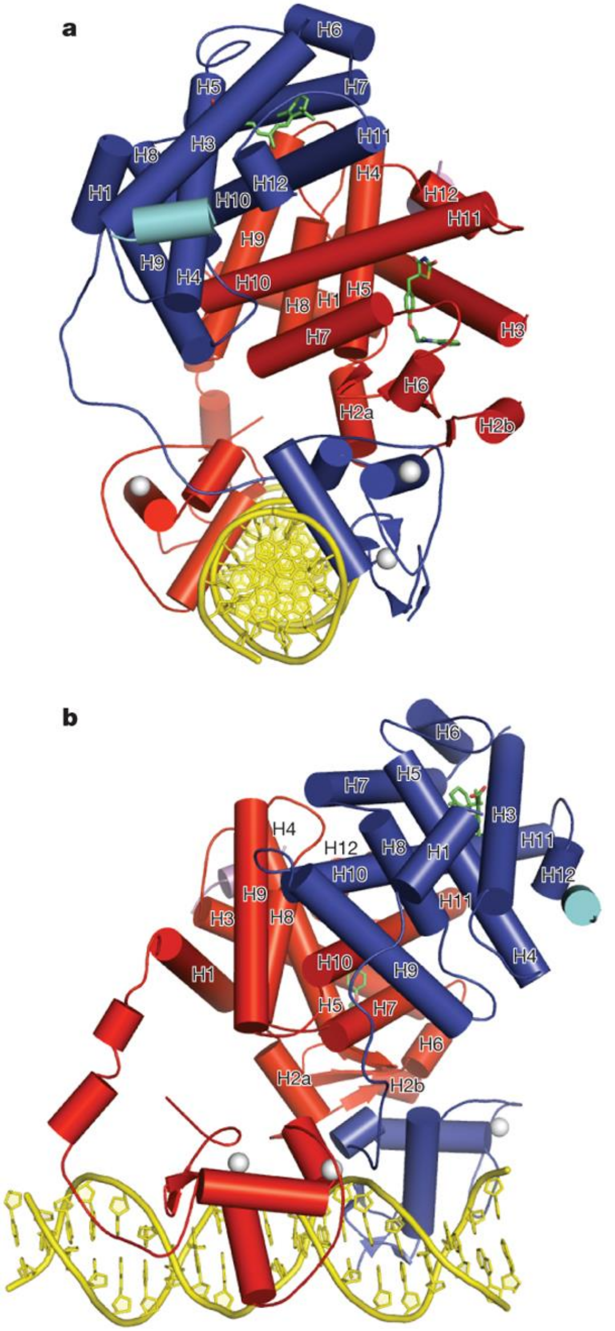
- Helix-turn-helix (HTH)
- Zinc fingers
  - With subtypes
- Basic leucine zipper (bZip)
- Basic helix-loop-helix (bHLH)



Athanasios G Papavassiliou

Example of a TF structure

- A heterodimer, each part with a DNA binding domain (DBD) and a ligand binding domain (LBD)
  - PPAR – Peroxisome proliferator-activated receptor
  - RXR – Retinoid X receptor
- Binds to DNA with zinc finger domains
- Here the part involved in ligand binding is large, compared to the DNA-binding region
  - This may vary between TFs



NATURE | Vol 456 | 20 November 2008

TF – DNA binding

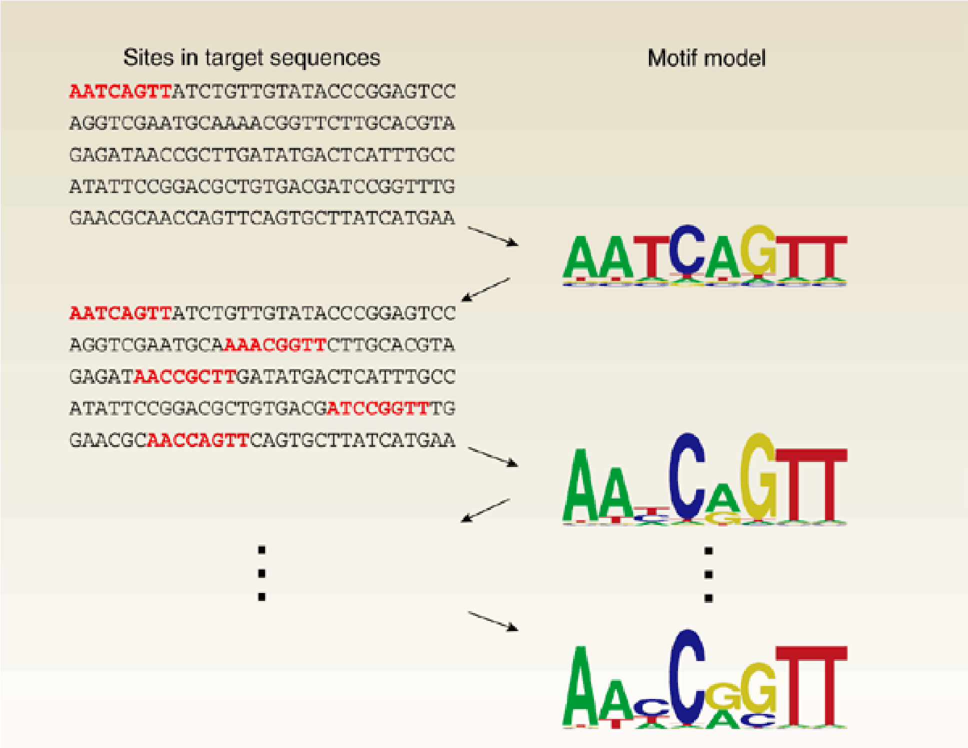
- **Motifs**

  – TFs which binds DNA show preference for certain DNA-sequences.

  – The sequences for a certian TF resemble each other, but are not identical

  – Statistical models to describe motifs and their sequence variation

  – Use models to find binding sites


Sites in target sequences
Motif model
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
AATCAGTT
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
AA CAGTT
AA CGGTT

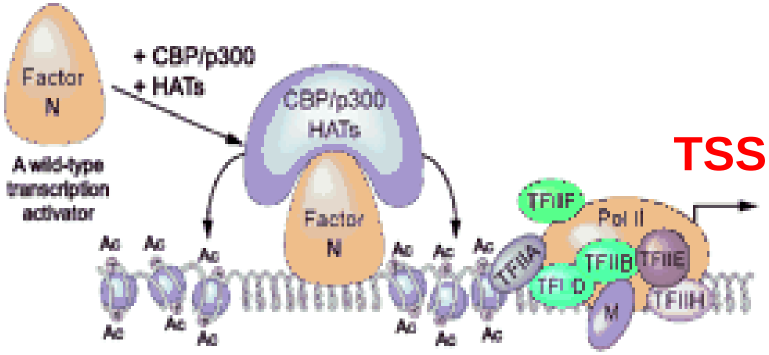
Promoters

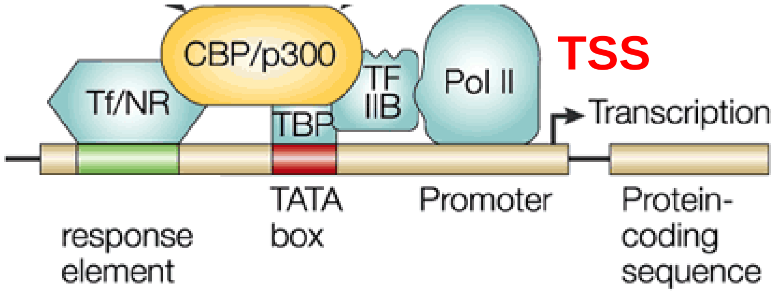
- Most transcripts (genes) are enriched for TF binding sites immediately up and downstream of TSS.

- This region of enrichment is called the **promoter.**

- The promoter region is typically defined 2000bp upstream and 200bp downstream of TSS

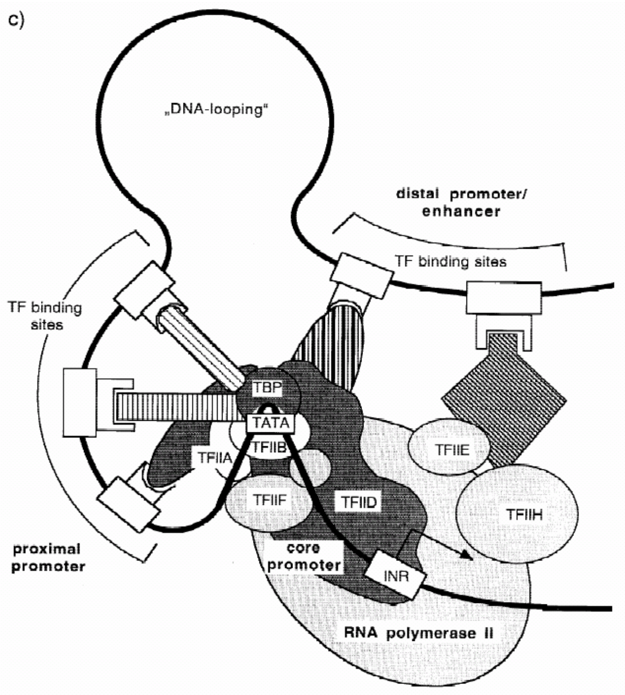
- But this is not an absolute measure, and enrichment of transcription factor binding often extend 5-10 kb upstream and in the first intron downstream of TSS.

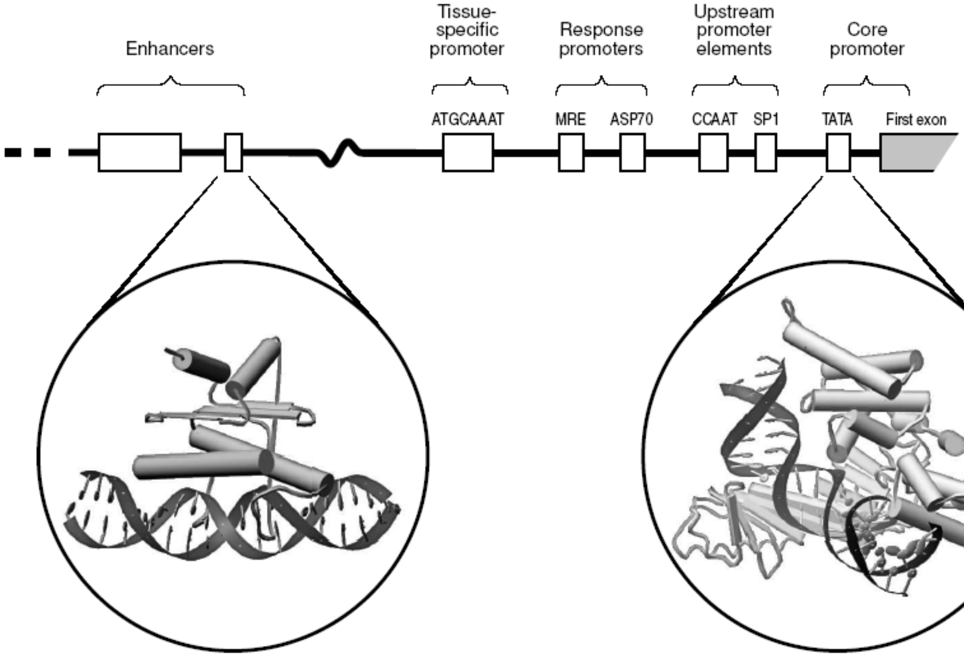




TFs and regulatory regions

- Proteins (transcription factors, TFs) recognise binding sites (sequence motifs) in **gene regulatory regions**
  - In particular **promoters** and **enhancers**
- The transcription factors stabilise the transcription complex

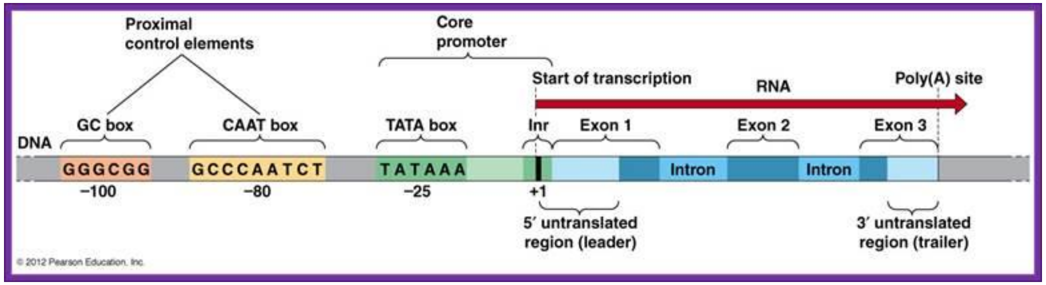




Michael Lones
Thomas Werner

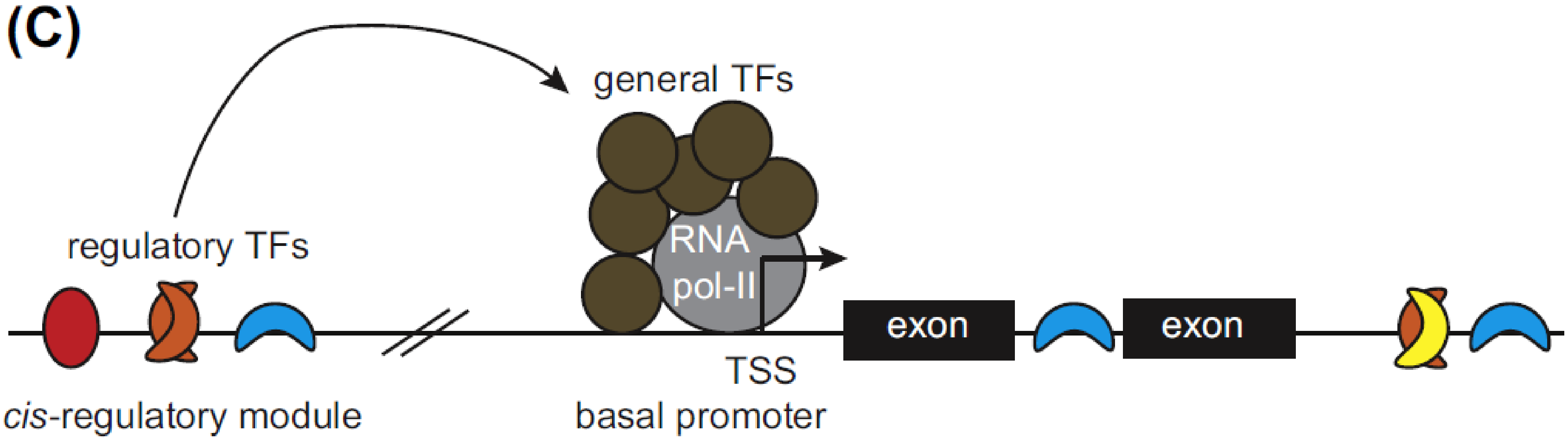
General and Regulatory TFs

- General TFs in the core / basal promoter
  - TATA Box, right next to TSS
- Regulatory TFs everywhere else …
  - Mainly in promoters and enchancers
  - Also in introns and UTRs



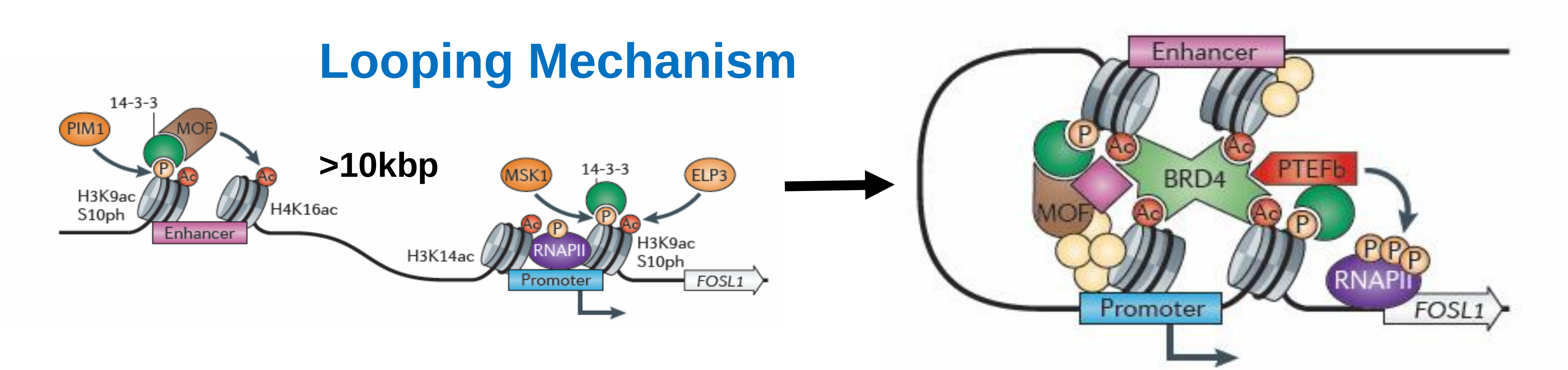
http://mol-biol4masters.masters.grkraj.org/html/Gene_Structure5B-Eukaryotic_Promoter_Structure_for_RNA_Polymerase_II.htm

  - Multiple regulatory TFs often form a *cis*-regulatory module



Enhancers

- Distal regulatory elements that affect gene transcription (primarily in an activating way) at longer distances (> 10kb).
- Many enhancers affect one or a few nearby genes.
- Some enhancers have been shown to work at much longer distances, even affecting genes on different chromosomes.
- Note: A long genomic distance does not imply that the genomic regions are located far apart in space in the nucleus.

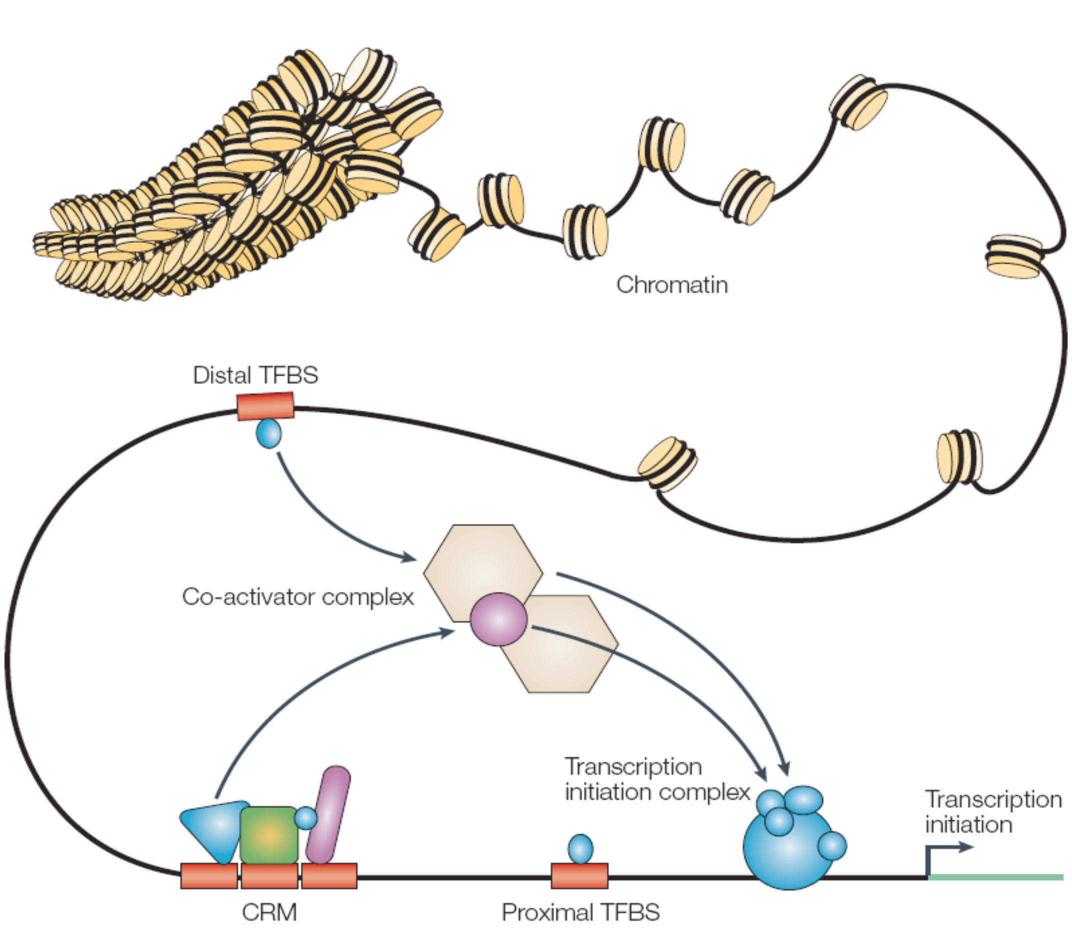


Looping Mechanism

>10kbp

Ong,Nature,2011

A complex regulatory system



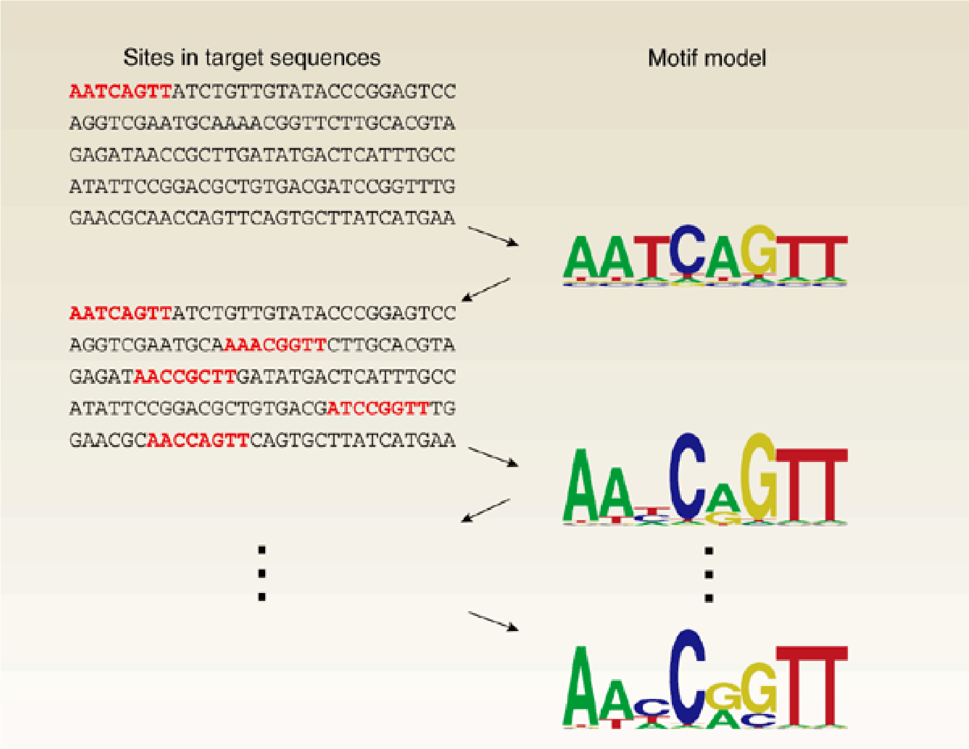
- Regulation over large distances
  - DNA looping / chromatin folding
- Several TFs act together – *cis*-regulatory modules (CRM)
- Combinatorial complexity of regulation
  - TFs may be both repressive and activating
  - Cooperativity
  - Competition

Finding TF binding sites

- **Computational approaches**
  - Easy to make predictions
  - Many predictions will be false positive (mainly) or false negative
  - Predictions may be very **general,** not for a specific set of conditions
- ***Experimental approaches***
  - May tell you where a TF actually binds under **specific** conditions
    - E.g. cell type, stimulation of cell, status of nutrients etc
  - May be experimentally challenging
  - ChIP-seq most common experimental approach
- ***Combined approaches***
  - E.g. filter computational binding sites by open chromatin

Computational Approaches

- **Motifs**
  - TFs which binds DNA show preference for certain DNA-sequences.
  - The sequences for a certian TF resemble each other, but are not identical
  - Statistical models to describe motifs and their sequence variation
  - Use models to find binding sites



AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA

AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCA**AAACGGTT**CTTGCACGTA
GAGAT**AACCGCTT**GATATGACTCATTTGCC
ATATTCCGGACGCTGTGACG**ATCCGGTT**TG
GAACGC**AACCAGTT**CAGTGCTTATCATGAA

Motif models

- **Mismatch model (MM, Hamming distance)**
  - Number of mismatches vs consensus
  - Very fast to compute
- **IUPAC representation**
  - Consensus using ambiguity codes
  - More expressive than Hamming distance
- **Position Weight Matrix (PWM)**
  - Probability of occurrence at each position
  - Flexible, no hard cutoff

- Positions are uncorrelated in all these models

Mismatch (MM) / IUPAC model

- Mismatch
  - Distance is number of mismatches between test sequence and consensus sequence
  - Positive match if mismatch distance is less than cutoff
- IUPAC
  - Make IUPAC consensus sequence from alignment of example sequences (related to a simple regular expression)

```
CONS  AACGGATAA
TEST  AACGCATTA
```

Mismatch distance = 2

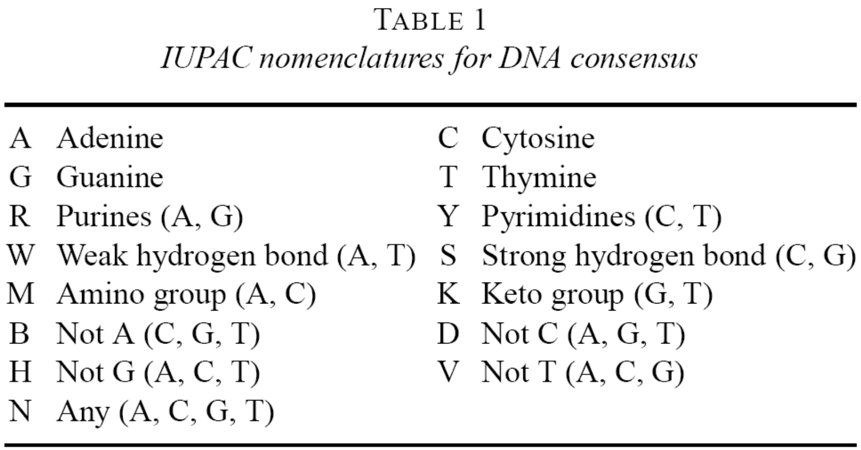


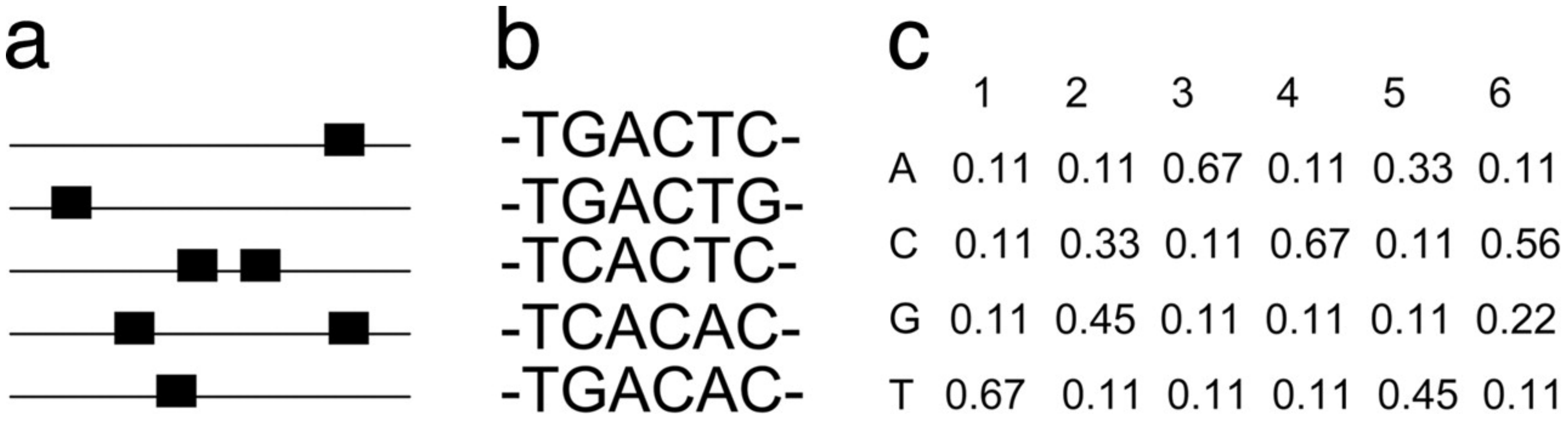
TABLE 1

IUPAC nomenclatures for DNA consensus

A	Adenine	C	Cytosine
G	Guanine	T	Thymine
R	Purines (A, G)	Y	Pyrimidines (C, T)
W	Weak hydrogen bond (A, T)	S	Strong hydrogen bond (C, G)
M	Amino group (A, C)	K	Keto group (G, T)
B	Not A (C, G, T)	D	Not C (A, G, T)
H	Not G (A, C, T)	V	Not T (A, C, G)
N	Any (A, C, G, T)		

```
SEQ   AATTGA
SEQ   AGGTCC
SEQ   AGGATG
SEQ   AGGCGT
CONS  ARKHBN
TEST  AGTAAA
```

Position Weight Matrix (PWM)



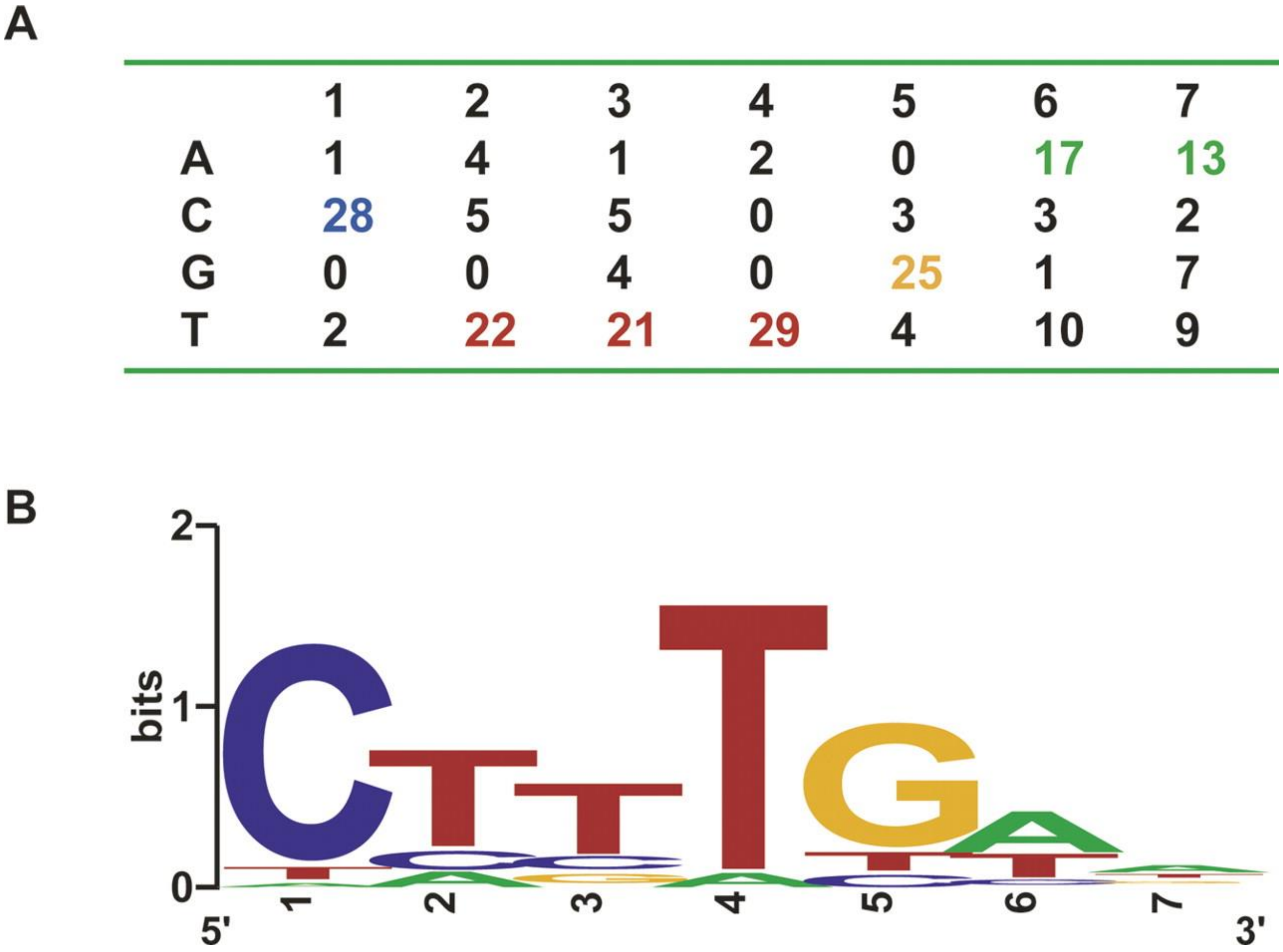
-TGACTC-
-TGACTG-
-TCACTC-
-TCACAC-
-TGACAC-

	1	2	3	4	5	6
A	0.11	0.11	0.67	0.11	0.33	0.11
C	0.11	0.33	0.11	0.67	0.11	0.56
G	0.11	0.45	0.11	0.11	0.11	0.22
T	0.67	0.11	0.11	0.11	0.45	0.11

Use a set of known TF binding sites (a),
align these binding sites (b),
and count the relative occurrence of each base at each position (c)

PNAS | **July 5, 2005** | vol. 102 | no. 27 | **9481–9486**

Making a sequence logo

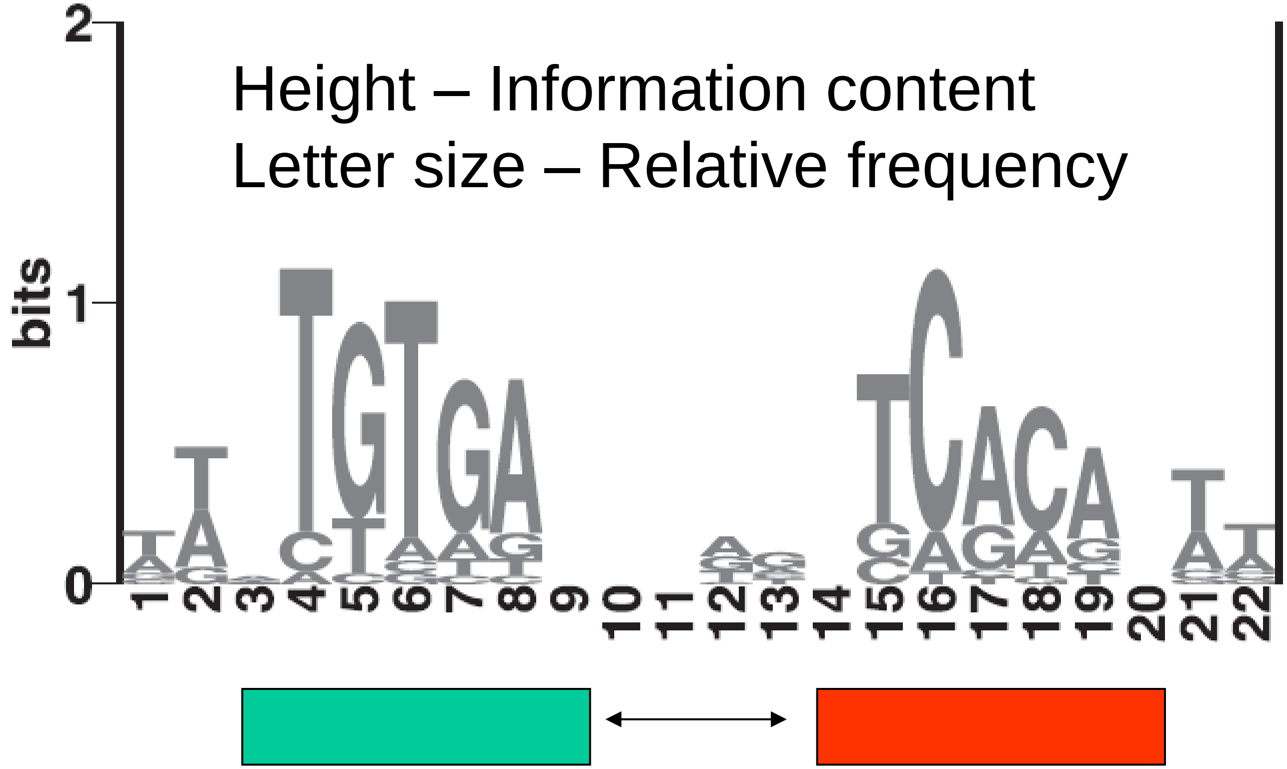


A

	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

Genome Res. 2007 17: 1438-1447

An interesting sequence logo



Height – Information content

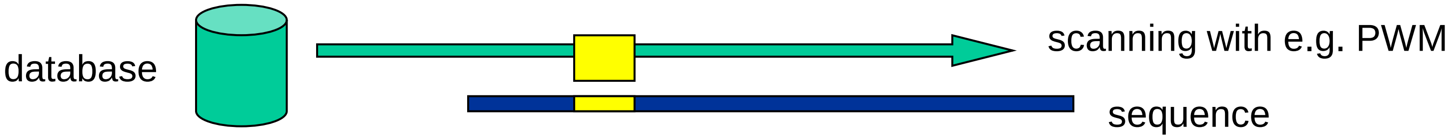
Letter size – Relative frequency

Palindromic motif

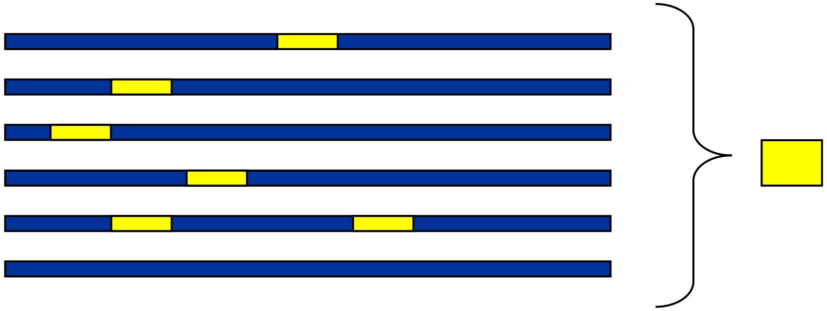
- Was it a cat I saw? (Same both ways …)
- TGTGA – TCACA (Reverse complement!)
- ACACT – AGTGT (Base pairing, symmetry)

Main strategies for finding motifs

- Scanning methods
  - Search sequences for known (e.g. experimental) motifs

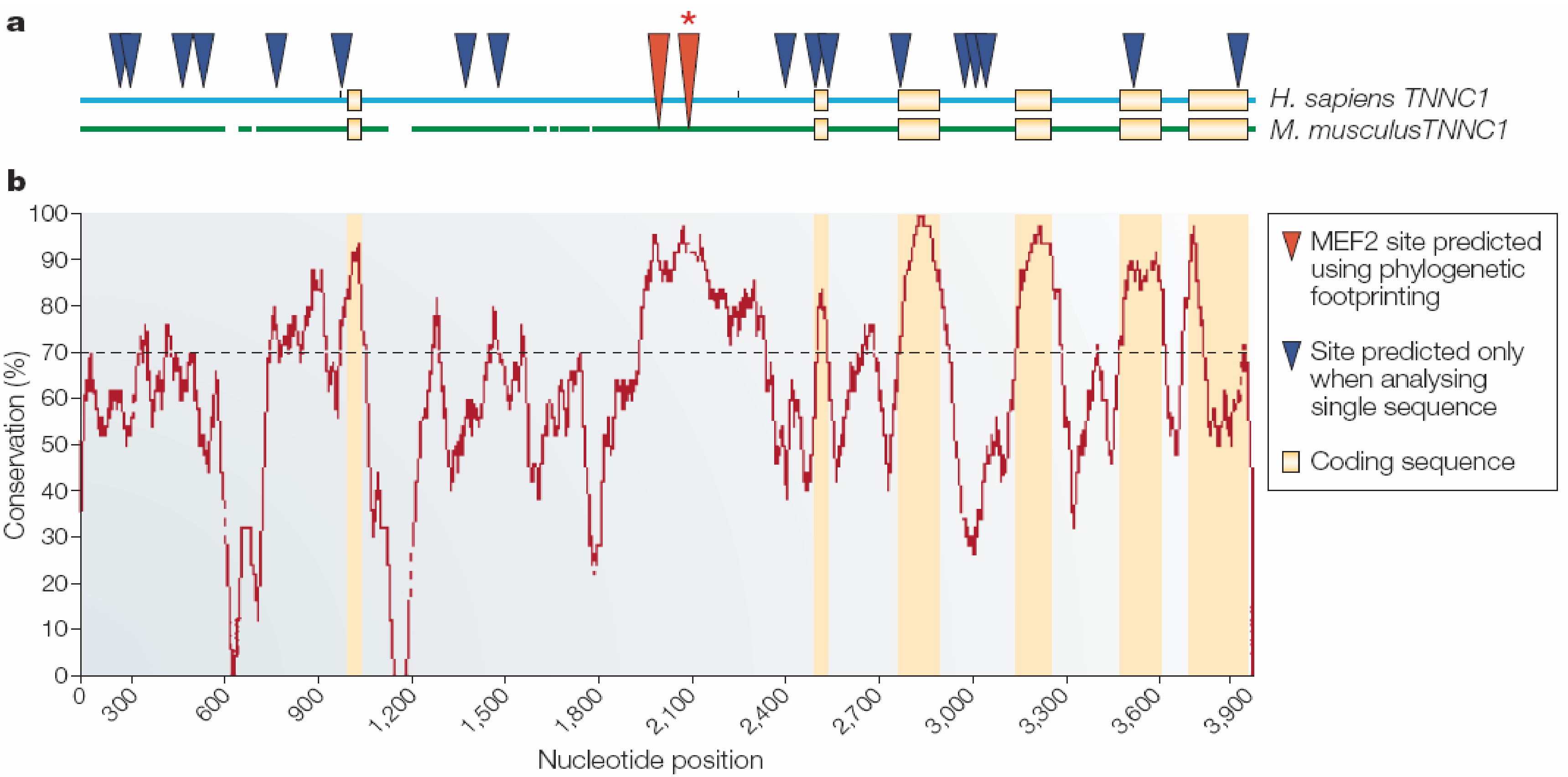


1. *De novo* motif discovery methods
   - Example: Promoters of co-regulated genes
     - Gene expression data (RNA-seq or microarray)
   - Experimental TF binding data (ChIP-chip or ChIP-seq)
2. Search for conserved / overrepresented motifs in the data set
     - Word counting / consensus sequence discovery
     - Position-specific weight matrix (PWM) based



Phylogenetic footprinting

- May be used to improve performance
- Looks for non-coding regions that are conserved
  - May miss species-specific binding sites



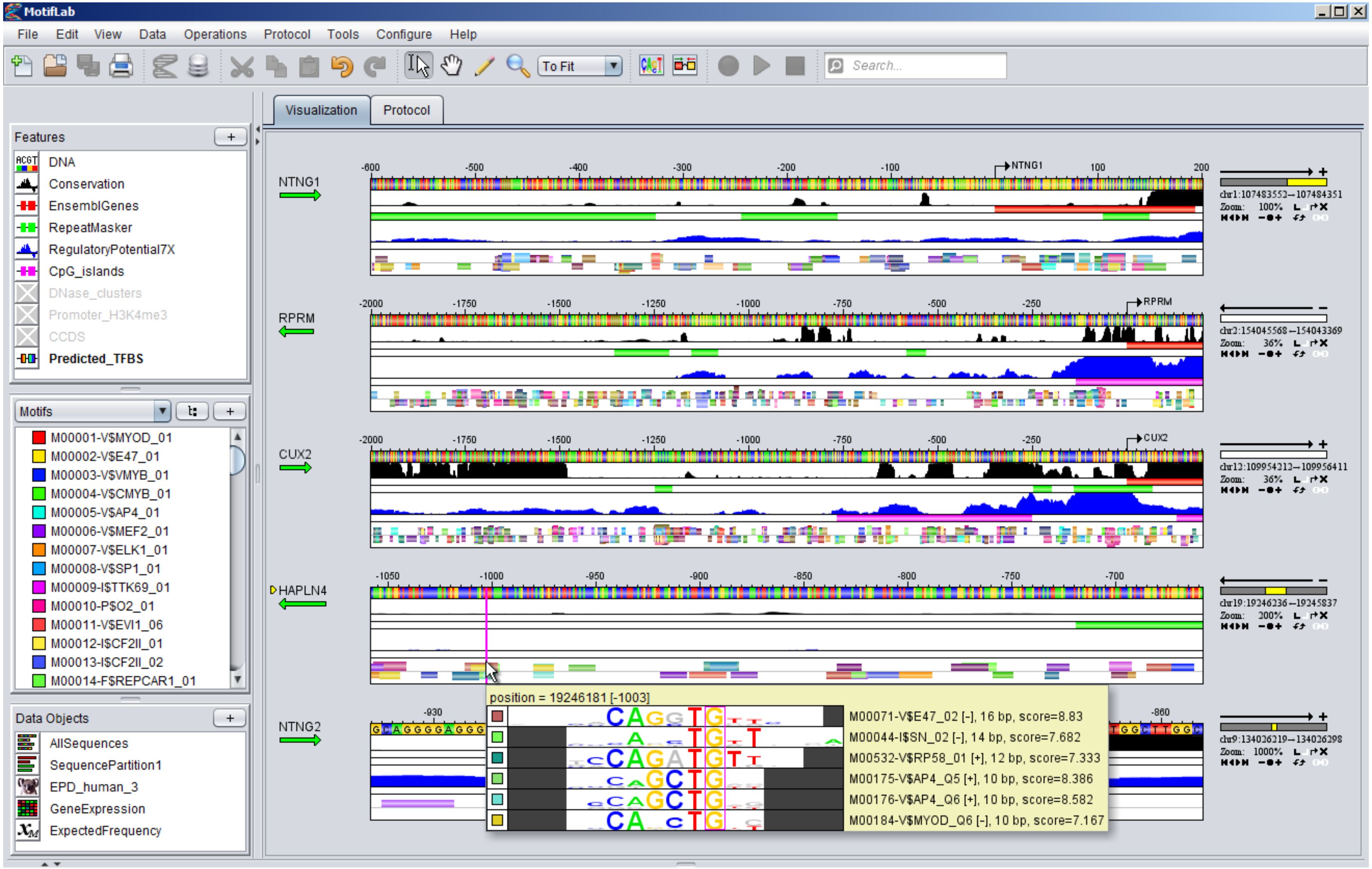
Some important resources

- Motif discovery tools
  - MEME Suite – Find e.g. shared motifs in unaligned sequences, using different models (one occurrence; zero or one; any number of occurrences)
    - http://meme-suite.org/
- Motif databases
  - Jaspar – Open source, high quality
    - http://jaspar.genereg.net/
  - TRANSFAC – Open, commercial with more motifs
    - http://www.gene-regulation.com/
- Visualisation tools, genome browsers, in particular with experimental data (ChIP-chip, ChIP-seq)
  - UCSC genome browser
    - http://genome.ucsc.edu/
  - Ensembl genome browser
    - http://www.ensembl.org/
- *Many* more resources, tools etc

De novo motif discovery

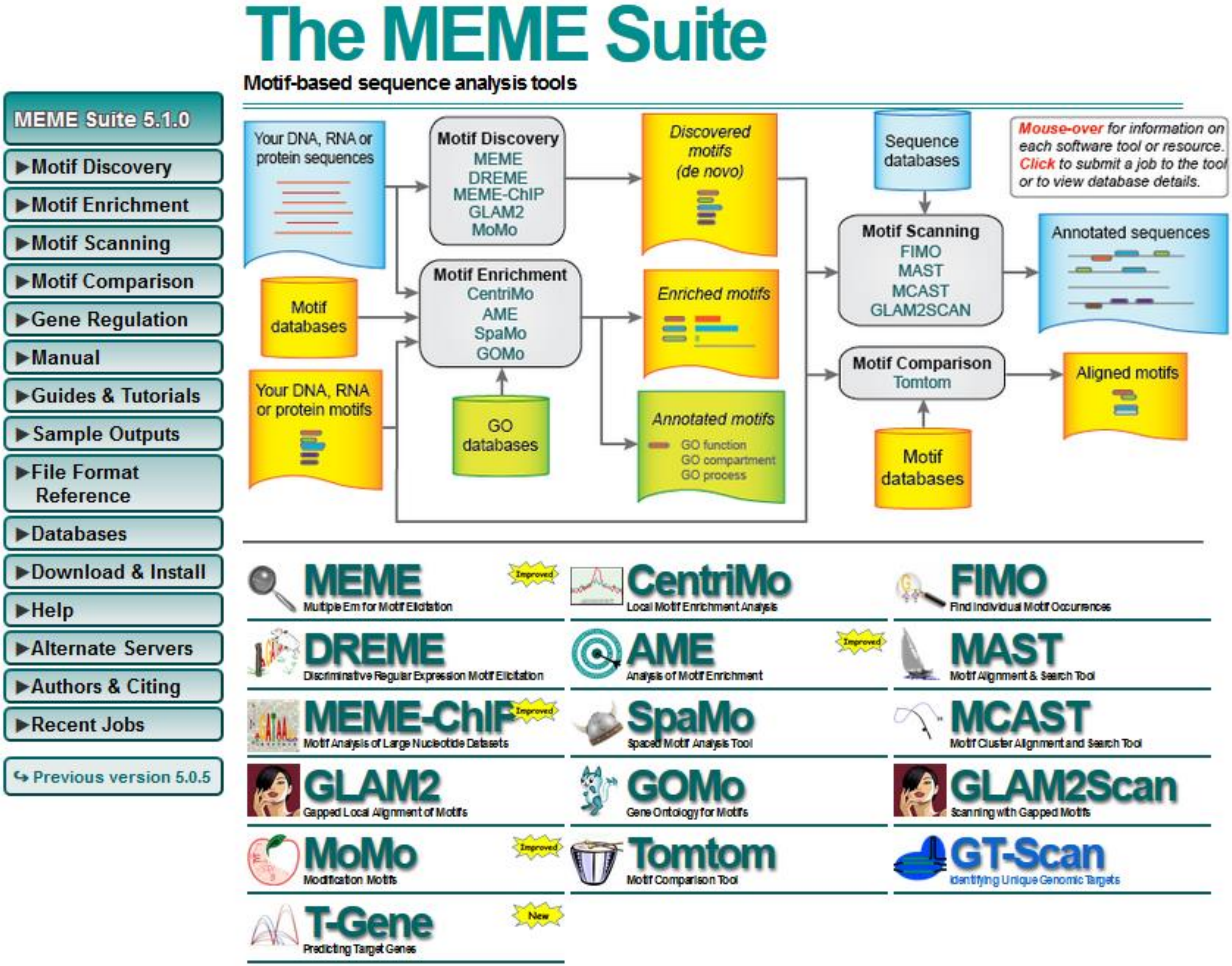
- Focus on the most likely regulatory region
  - Promoter region for gene sets can be downloaded using UCSC Table Browser or Ensembl BioMart
    - <https://genome.ucsc.edu/cgi-bin/hgTables>
    - <http://www.ensembl.org/biomart/martview>
  - Something like -1000 to +200 is often used
    - But this is not fixed ...
- Filter out repeats (why?)
  - This can be done with RepeatMasker
    - <http://www.repeatmasker.org/>
  - It may also be possible to download pre-masked regions
  - Make sure that the masking (N, n, lower-case, ...) is understood by your motif discovery tool

MotifLab



position = 19246181 [-1003]

M00071-V$E47_02 [-], 16 bp, score=8.83
M00044-I$SN_02 [-], 14 bp, score=7.682
M00532-V$RP58_01 [+], 12 bp, score=7.333
M00175-V$AP4_Q5 [+], 10 bp, score=8.386
M00176-V$AP4_Q6 [+], 10 bp, score=8.582
M00184-V$MYOD_Q6 [-], 10 bp, score=7.167



The MEME Suite

Motif-based sequence analysis tools

MEME Suite 5.1.0

- ▶Motif Discovery
- ▶Motif Enrichment
- ▶Motif Scanning
- ▶Motif Comparison
- ▶Gene Regulation
- ▶Manual
- ▶Guides & Tutorials
- ▶Sample Outputs
- ▶File Format Reference
- ▶Databases
- ▶Download & Install
- ▶Help
- ▶Alternate Servers
- ▶Authors & Citing
- ▶Recent Jobs

↪ Previous version 5.0.5

Mouse-over for information on each software tool or resource. *Click* to submit a job to the tool or to view database details.

MEME Improved
Multiple Em for Motif Elicitation

CentriMo
Local Motif Enrichment Analysis

FIMO
Find Individual Motif Occurrences

DREME
Discriminative Regular Expression Motif Elicitation

AME Improved
Analysis of Motif Enrichment

MAST
Motif Alignment & Search Tool

MEME-ChIP Improved
Motif Analysis of Large Nucleotide Datasets

SpaMo
Spaced Motif Analysis Tool

MCAST
Motif Cluster Alignment and Search Tool

GLAM2
Gapped Local Alignment of Motifs

GOMo
Gene Ontology for Motifs

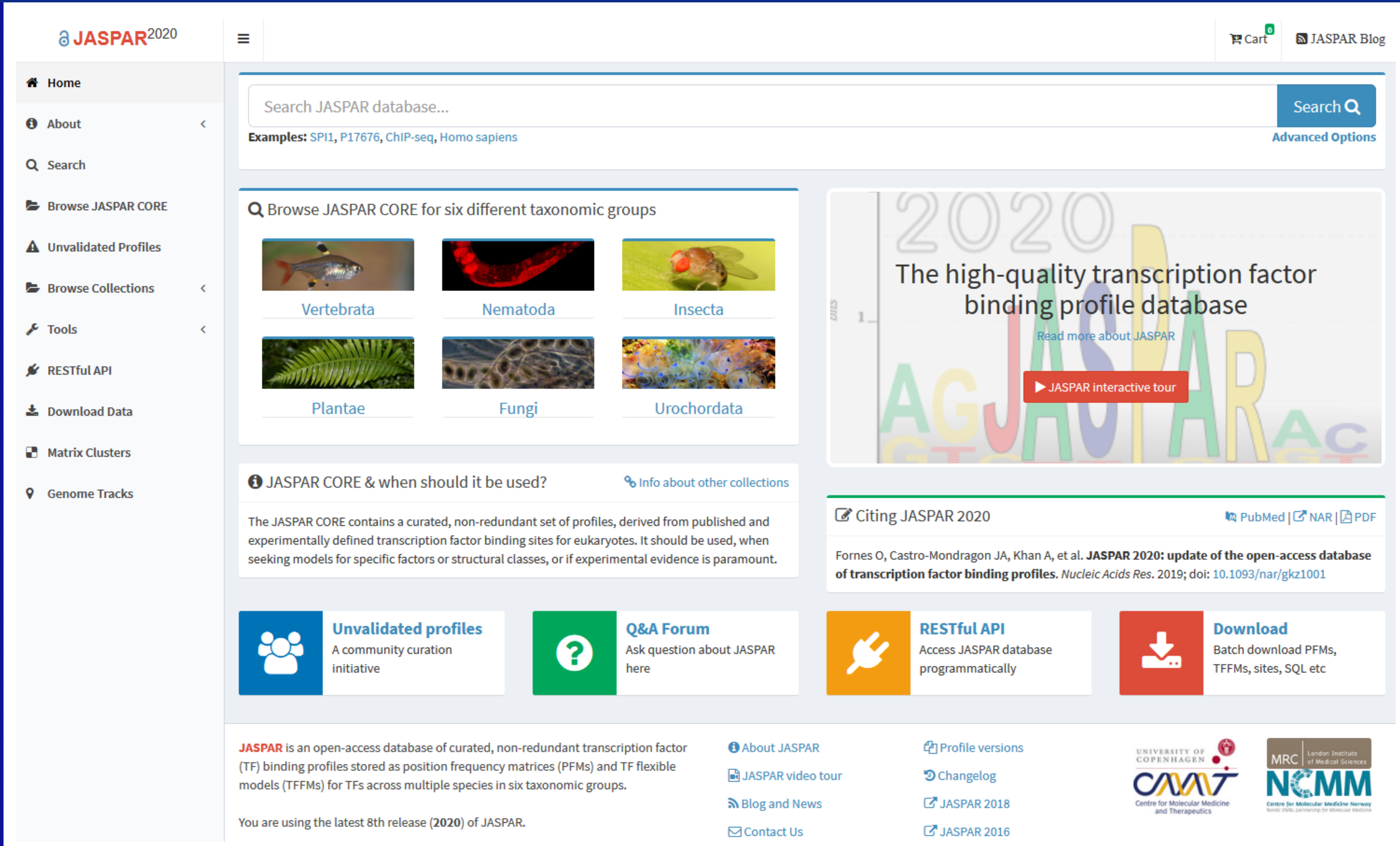
GLAM2Scan
Scanning with Gapped Motifs

MoMo Improved
Modification Motifs

Tomtom
Motif Comparison Tool

GT-Scan
Identifying Unique Genomic Targets

T-Gene New
Predicting Target Genes



JASPAR2020

☰

Cart 0 | JASPAR Blog

- Home
- About
- Search
- Browse JASPAR CORE
- Unvalidated Profiles
- Browse Collections
- Tools
- RESTful API
- Download Data
- Matrix Clusters
- Genome Tracks

Search JASPAR database... [Search]

Examples: SPI1, P17676, ChIP-seq, Homo sapiens

Advanced Options

Browse JASPAR CORE for six different taxonomic groups

Vertebrata | Nematoda | Insecta

Plantae | Fungi | Urochordata

2020

The high-quality transcription factor binding profile database

Read more about JASPAR

▶ JASPAR interactive tour

JASPAR CORE & when should it be used?

Info about other collections

The JASPAR CORE contains a curated, non-redundant set of profiles, derived from published and experimentally defined transcription factor binding sites for eukaryotes. It should be used, when seeking models for specific factors or structural classes, or if experimental evidence is paramount.

Citing JASPAR 2020

PubMed | NAR | PDF

Fornes O, Castro-Mondragon JA, Khan A, et al. **JASPAR 2020: update of the open-access database of transcription factor binding profiles.** *Nucleic Acids Res.* 2019; doi: 10.1093/nar/gkz1001

Unvalidated profiles
A community curation initiative

Q&A Forum
Ask question about JASPAR here

RESTful API
Access JASPAR database programmatically

Download
Batch download PFMs, TFFMs, sites, SQL etc

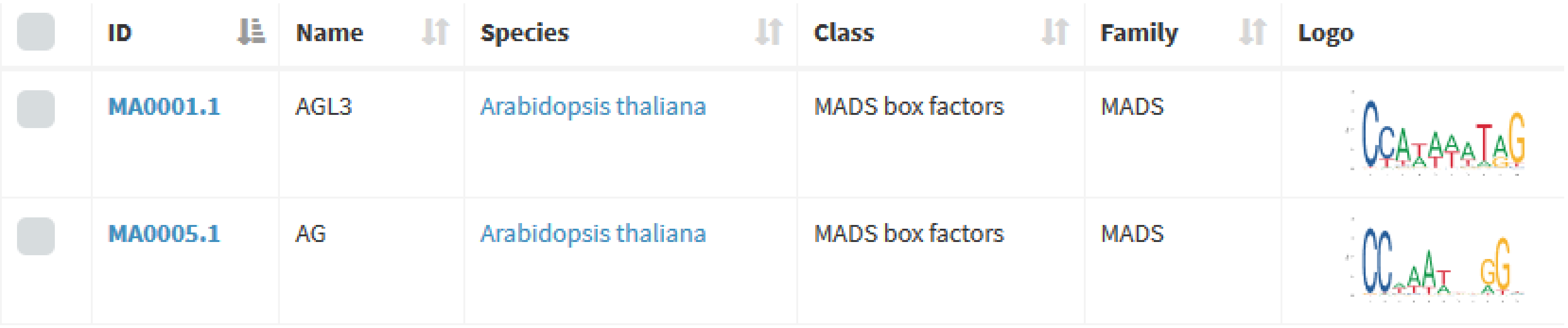
JASPAR is an open-access database of curated, non-redundant transcription factor (TF) binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups.

You are using the latest 8th release (**2020**) of JASPAR.

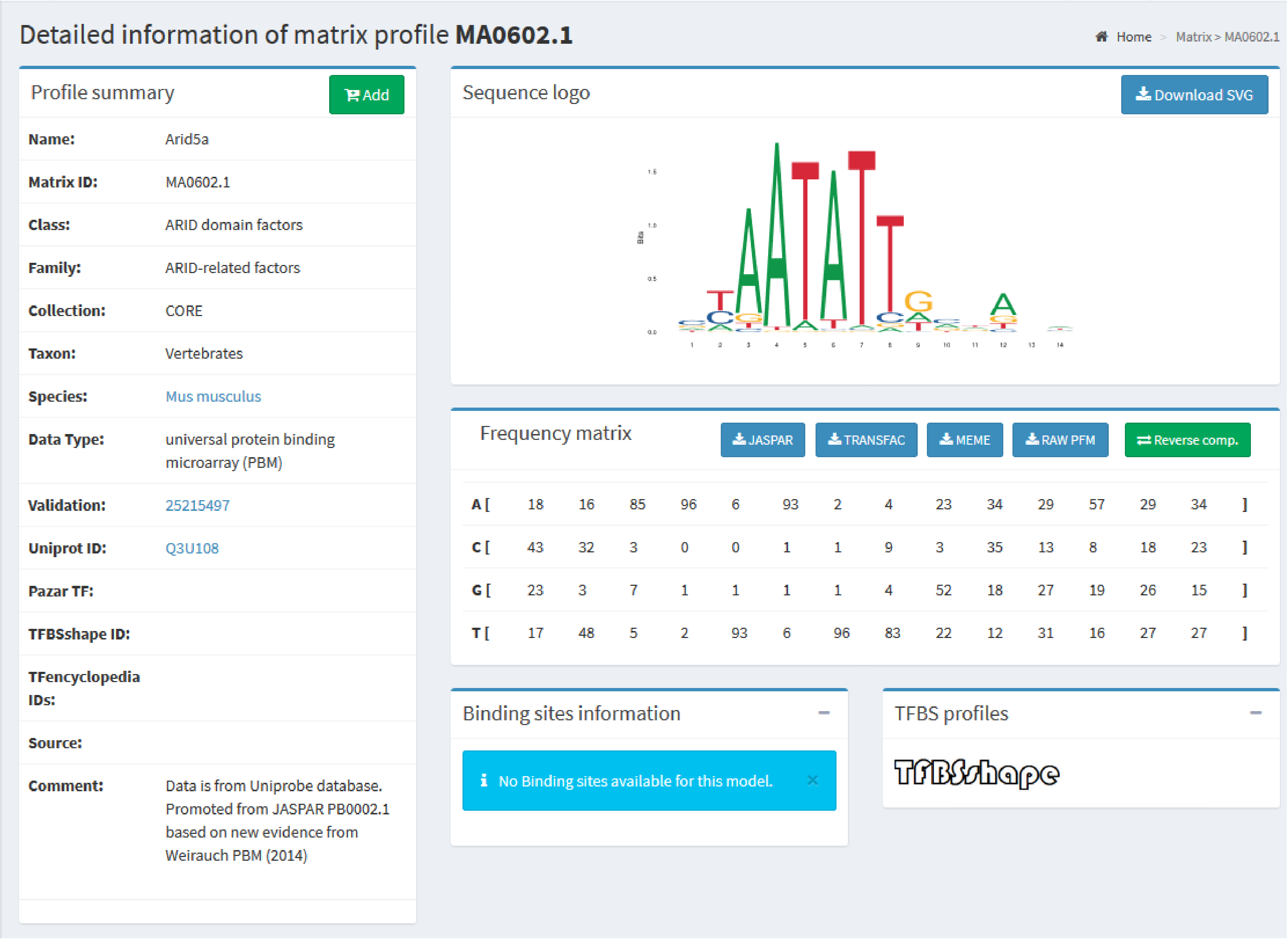
- About JASPAR
- JASPAR video tour
- Blog and News
- Contact Us

- Profile versions
- Changelog
- JASPAR 2018
- JASPAR 2016

UNIVERSITY OF COPENHAGEN
CMMT
Centre for Molecular Medicine and Therapeutics
MRC London Institute of Medical Sciences
NCMM
Centre for Molecular Medicine Norway



□	ID	Name	Species	Class	Family	Logo
□	MA0001.1	AGL3	Arabidopsis thaliana	MADS box factors	MADS	
□	MA0005.1	AG	Arabidopsis thaliana	MADS box factors	MADS	



Detailed information of matrix profile **MA0602.1**

Profile summary

Name:	Arid5a
Matrix ID:	MA0602.1
Class:	ARID domain factors
Family:	ARID-related factors
Collection:	CORE
Taxon:	Vertebrates
Species:	Mus musculus
Data Type:	universal protein binding microarray (PBM)
Validation:	25215497
Uniprot ID:	Q3U108
Pazar TF:	
TFBSshape ID:	
TFencyclopedia IDs:	
Source:	
Comment:	Data is from Uniprobe database. Promoted from JASPAR PB0002.1 based on new evidence from Weirauch PBM (2014)

Sequence logo

Frequency matrix

A [	18	16	85	96	6	93	2	4	23	34	29	57	29	34	**]**
C [	43	32	3	0	0	1	1	9	3	35	13	8	18	23	**]**
G [	23	3	7	1	1	1	1	4	52	18	27	19	26	15	**]**
T [	17	48	5	2	93	6	96	83	22	12	31	16	27	27	**]**

Binding sites information

ℹ No Binding sites available for this model. ×

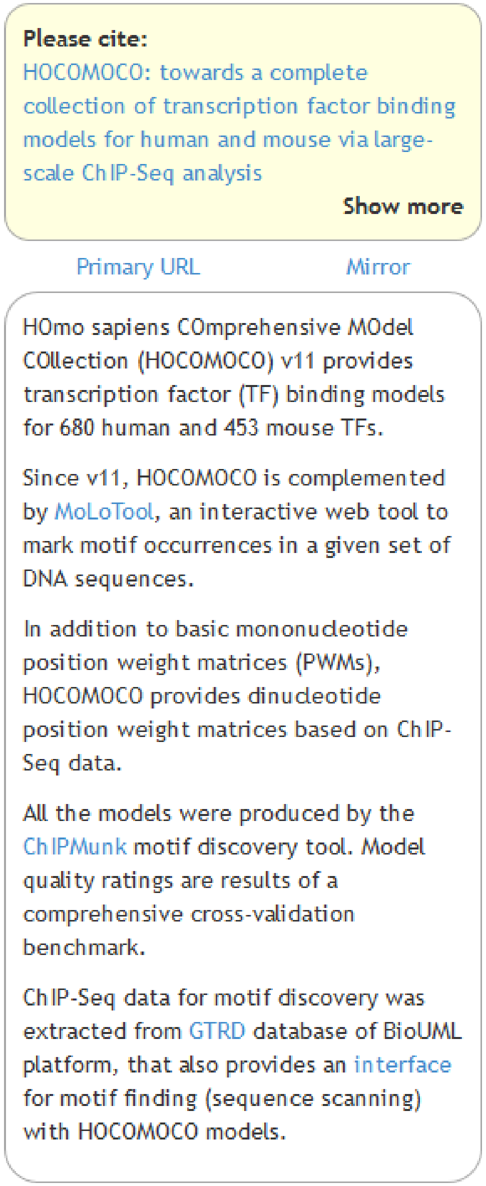
TFBS profiles

TFBSshape

TRANSFAC

	TRANSFAC Professional 2020.1	TRANSFAC Public
Factors	48,084	6,133
DNA sites	50,912	7,915
miRNAs	1,771	-
mRNA sites	67,823	-
Genes	102,900	2,397
ChIP fragments	103,548,181	-
Promoters	441,771	-
Matrices	9,962	398
References	40,648	(flat file)

HOCOMOCO



Please cite:
HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis

Show more

Primary URL Mirror

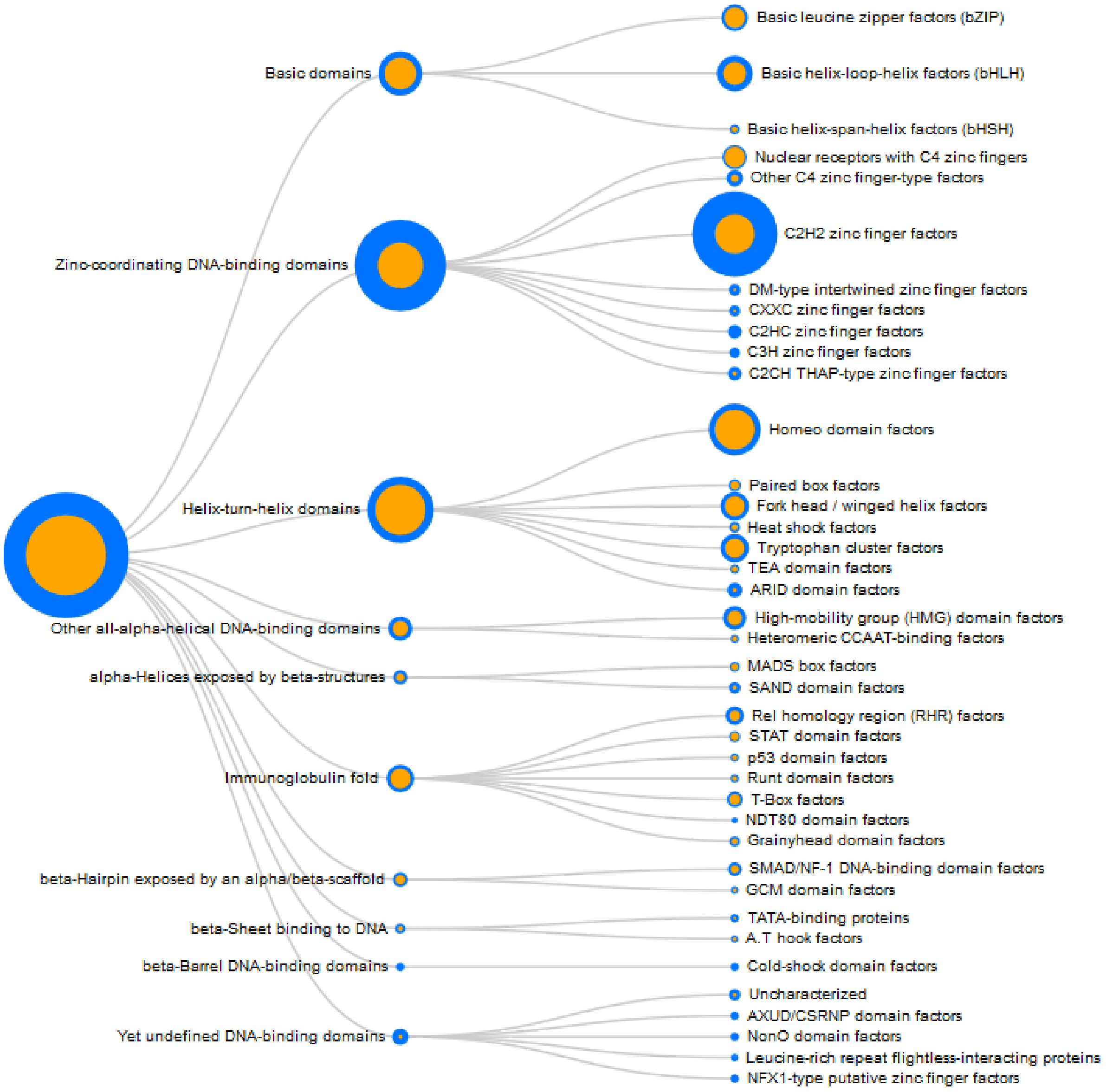
HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) v11 provides transcription factor (TF) binding models for 680 human and 453 mouse TFs.

Since v11, HOCOMOCO is complemented by MoLoTool, an interactive web tool to mark motif occurrences in a given set of DNA sequences.

In addition to basic mononucleotide position weight matrices (PWMs), HOCOMOCO provides dinucleotide position weight matrices based on ChIP-Seq data.

All the models were produced by the ChIPMunk motif discovery tool. Model quality ratings are results of a comprehensive cross-validation benchmark.

ChIP-Seq data for motif discovery was extracted from GTRD database of BioUML platform, that also provides an interface for motif finding (sequence scanning) with HOCOMOCO models.


Basic domains
Zinc-coordinating DNA-binding domains
Helix-turn-helix domains
Other all-alpha-helical DNA-binding domains
alpha-Helices exposed by beta-structures
Immunoglobulin fold
beta-Hairpin exposed by an alpha/beta-scaffold
beta-Sheet binding to DNA
beta-Barrel DNA-binding domains
Yet undefined DNA-binding domains
Basic leucine zipper factors (bZIP)
Basic helix-loop-helix factors (bHLH)
Basic helix-span-helix factors (bHSH)
Nuclear receptors with C4 zinc fingers
Other C4 zinc finger-type factors
C2H2 zinc finger factors
DM-type intertwined zinc finger factors
CXXC zinc finger factors
C2HC zinc finger factors
C3H zinc finger factors
C2CH THAP-type zinc finger factors
Homeo domain factors
Paired box factors
Fork head / winged helix factors
Heat shock factors
Tryptophan cluster factors
TEA domain factors
ARID domain factors
High-mobility group (HMG) domain factors
Heteromeric CCAAT-binding factors
MADS box factors
SAND domain factors
Rel homology region (RHR) factors
STAT domain factors
p53 domain factors
Runt domain factors
T-Box factors
NDT80 domain factors
Grainyhead domain factors
SMAD/NF-1 DNA-binding domain factors
GCM domain factors
TATA-binding proteins
A.T hook factors
Cold-shock domain factors
Uncharacterized
AXUD/CSRNP domain factors
NonO domain factors
Leucine-rich repeat flightless-interacting proteins
NFX1-type putative zinc finger factors