

Regulatory Genomics I: Regulatory Elements

Shamith Samarajiwa

- Group Leader (Computational Biology & Data Science)
MRC Cancer Unit,
University of Cambridge.

Computational Biology MPhil: Functional Genomics Lectures
February 19th 2021

wehi.edu.au



Overview

- Components involved in gene and genome regulation
- What are regulatory elements and where are they located?
- Regulatory element classes
- Non-coding RNA
- Transcription Factors: DNA binding regulatory proteins
- Computational identification of regulatory elements
- TF binding DNA motif discovery: What binds where and how to locate TF binding sites

Gene & Genome Regulation

- **Regulatory Elements & Transcription Factors (lecture 1)**
- **Polycomb Group (PcG) proteins**
- **Readers, Writers and Erasers**
 - SWI/SNF, BETs, 14-3-3, HATs, HMTs, Kinases, HDMs, HDACs etc.
- **DNA secondary structure**
- **Non coding RNA:**
 - miRNA, piRNA, snoRNA, Y-RNA etc.
 - lncRNA
- **Histone modifications (lecture 2)**
- **Nucleosome positioning & chromatin accessibility (lecture 2)**
- **Chromatin domains (Lecture 2)**
 - TADs
 - Hubs / Loops / Enhancer-Promoter Interactions
 - A/B compartments
- **DNA Methylation (lecture 4)**

Junk DNA

- It was initially believed that most of the genome consisted of non-functional DNA sequences or “Junk DNA”
- The term ‘junk DNA’ was first used in the 1960s (e.g., Ehret and de Haller 1963). It was formalized by Susumu Ohno in 1972.
- Showed to be false by projects such as ENCODE, Roadmap and Fantom

> Brookhaven Symp Biol. 1972;23:366-70.

So much "junk" DNA in our genome

S Ohno

PMID: 5065367

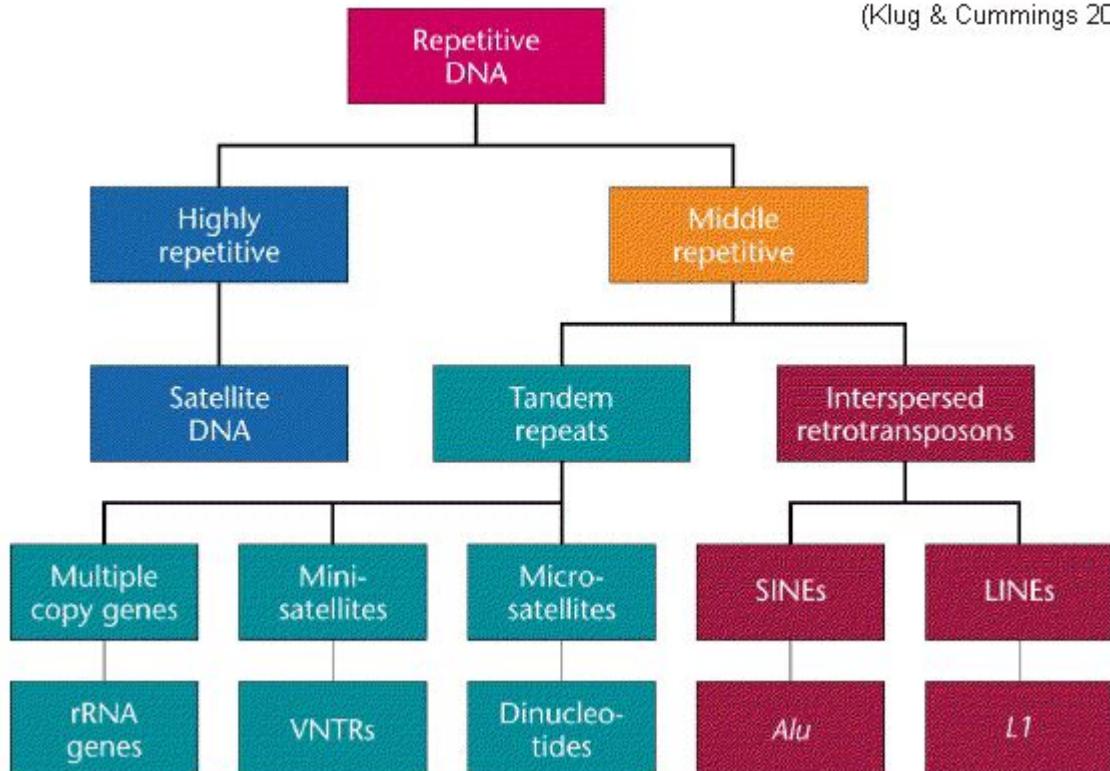
Repetitive DNA Elements



Barbara McClintock - discovered mobile genetic elements

- More than 50% of our genome consists of repetitive or transposable DNA elements.
- Involved in gene regulation, epigenetic catalysis, genome evolution, disease processes etc

(Klug & Cummings 2000)

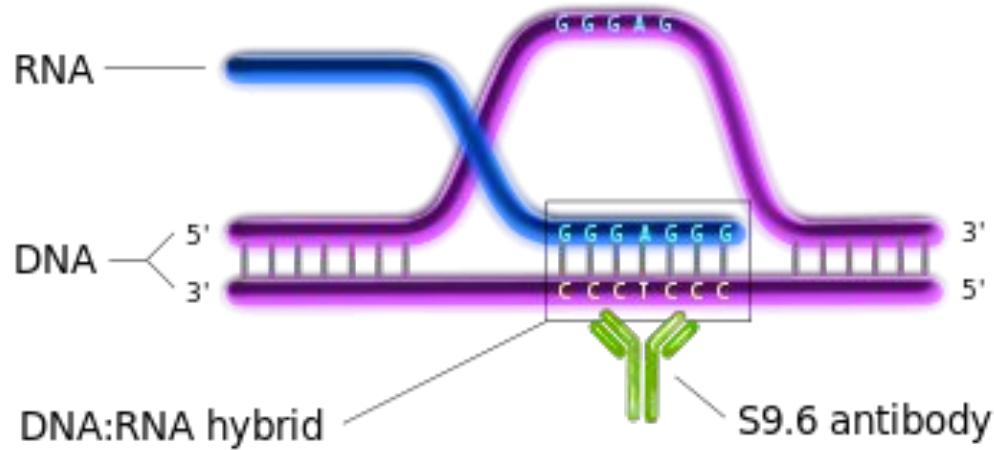


Non-Coding RNAs

Type	Abbreviation	Full name	Size
Housekeeping ncRNAs	rRNA	ribosomal RNA	120–4,500 nt
	tRNA	transfer RNA	76–90 nt
	snRNA	small nuclear RNA	100–300 nt
	snoRNA	small nucleolar RNA	60–400
	TERC	telomerase RNA	/
	tRF	tRNA-Derived Fragments	16–28 nt
	tiRNA	tRNA halves	29–50 nt
Regulatory ncRNAs	miRNA	microRNA	21–23 nt
	siRNA	small interfering RNA	20–25 nt
	piRNA	piwi-interacting RNA	26–32 nt
	eRNA	enhancer RNA	50–2,000 nt
	lncRNA	long non-coding RNAs	>200 nt
	circRNA	circular RNA	100–10,000 nt
	Y RNA	Y RNA	/

R-loops

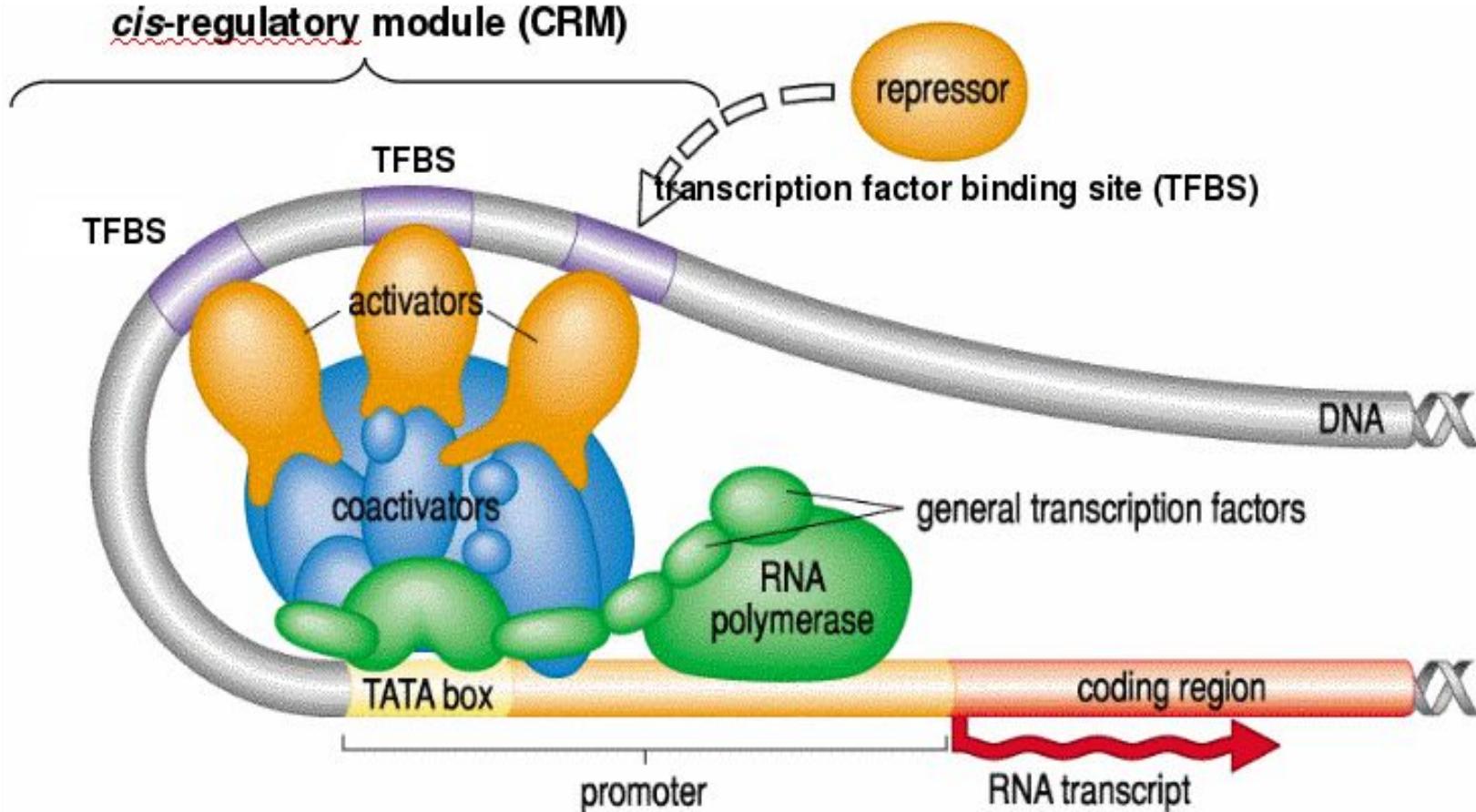
R-loops are three-stranded nucleic acid structures, composed of a DNA:RNA hybrid and the associated non-template single-stranded DNA.



(DNA-RNA immunoprecipitation) DRIP-seq

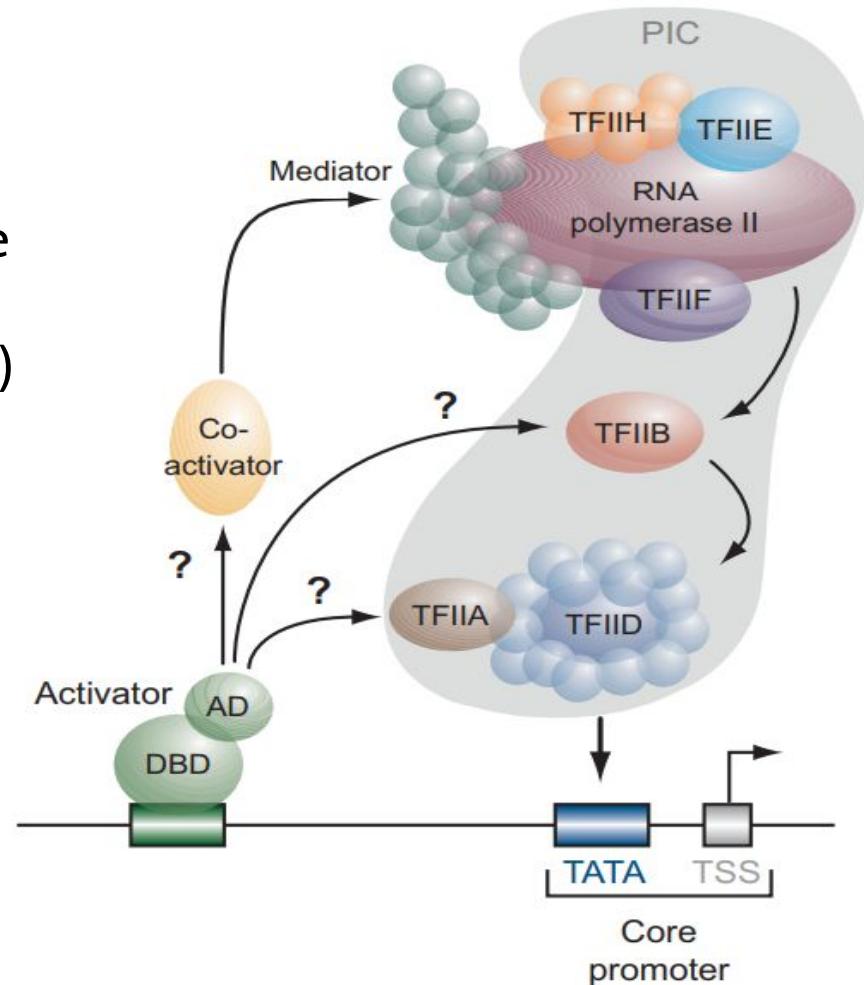
- **G-Quadruplexes** and **iMotifs** are other DNA based structural elements.
- Many roles in normal cellular function and in pathological states.

Regulatory DNA Elements



Pre Initiation Complex

- Transcriptional activators and coactivators, **Polymerase holoenzyme complex** & **Mediator complex** and GTFs form Pre Initiation Complex (PIC)



RNA Polymerases

RNA Polymerases (RNAP) are enzyme complexes that uses a DNA template to produce primary RNA transcripts.

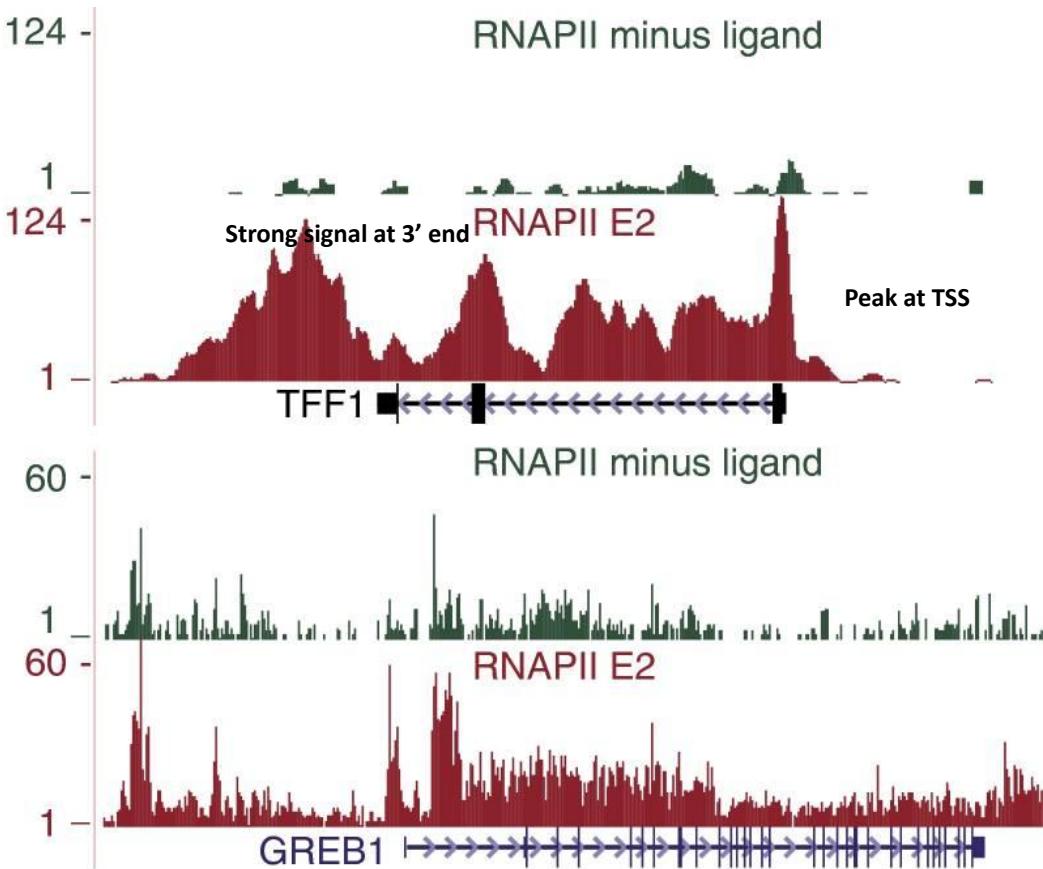


- **RNA polymerase I:** synthesizes a **pre-rRNA** 45S (35S in yeast), which matures into 28S, 18S and 5.8S rRNAs which will form the major RNA components of the ribosome.
- **RNA polymerase II:** synthesizes precursors of **mRNAs** and most **snRNAs** and **miRNAs**.
- **RNA polymerase III:** synthesizes **tRNAs**, **5S rRNA** and other small RNAs found in the nucleus and cytosol
- **RNA polymerase IV:** synthesizes siRNA in plants.
- **RNA polymerase V:** synthesizes RNAs involved in siRNA directed heterochromatin formation in plants

Measuring RNAP-II occupancy

- Stages of RNAP-II activity can be experimentally determined by ChIP-seq:

- Recruitment of the RNA Polymerase II complex
- Transcription initiation
- Clearance of RNAP-II
- Pausing
- Escape from pause
- Elongation
- Termination



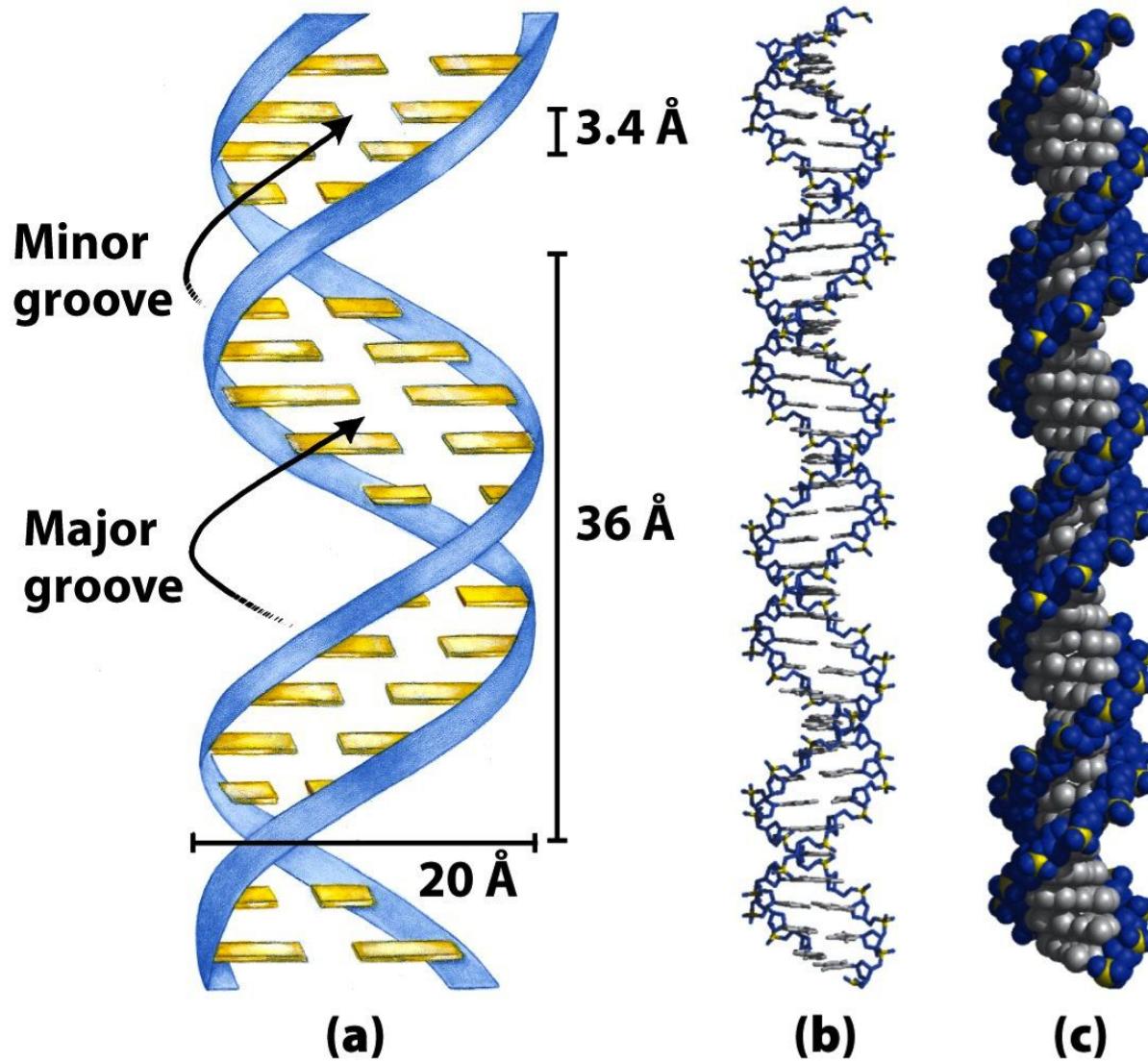


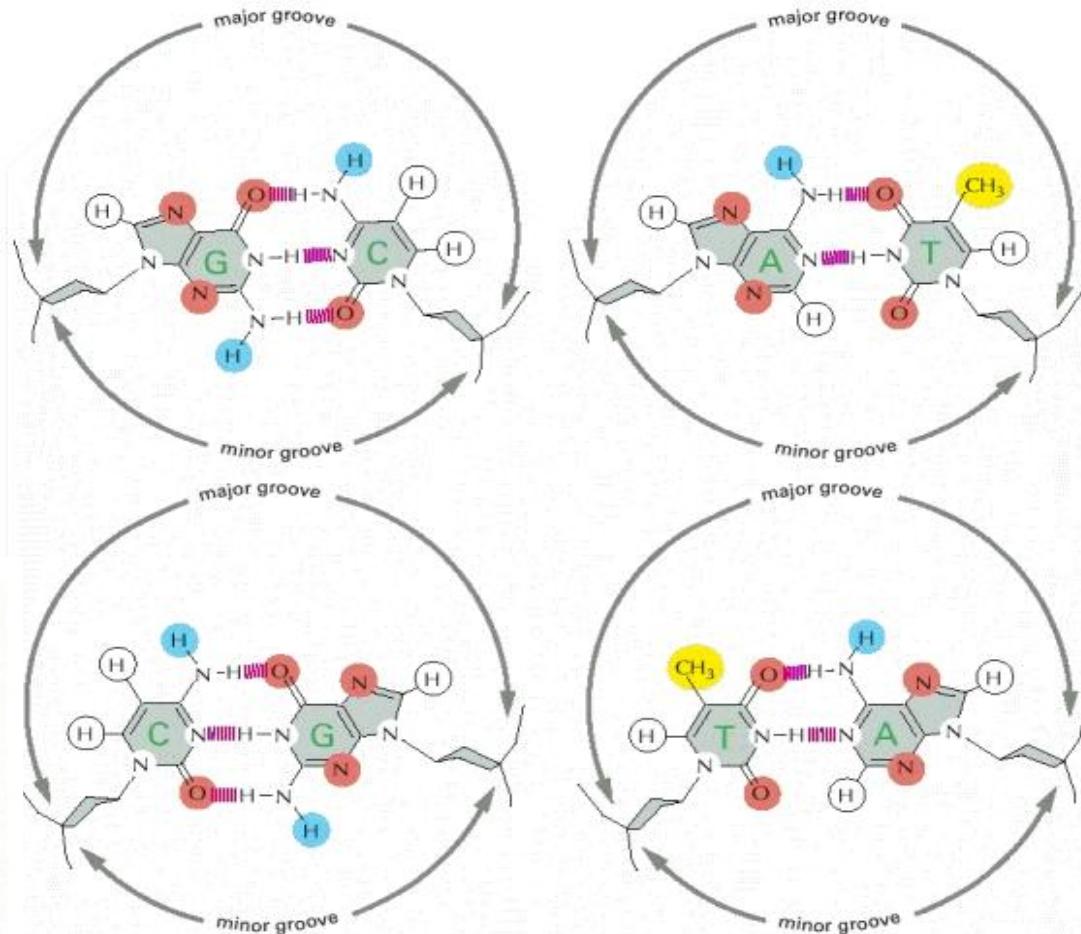
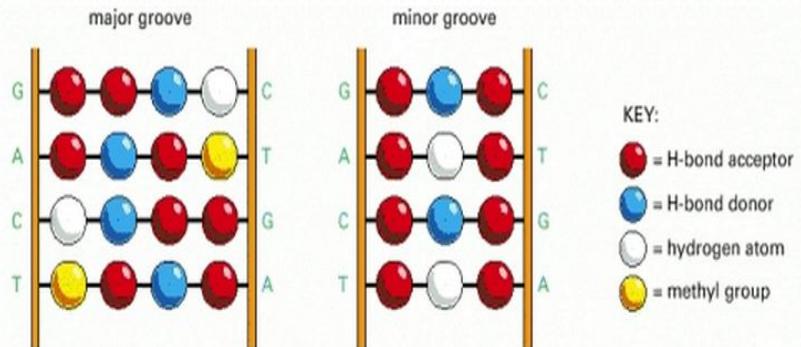
Figure 8-13

Lehninger Principles of Biochemistry, Fifth Edition

© 2008 W.H. Freeman and Company

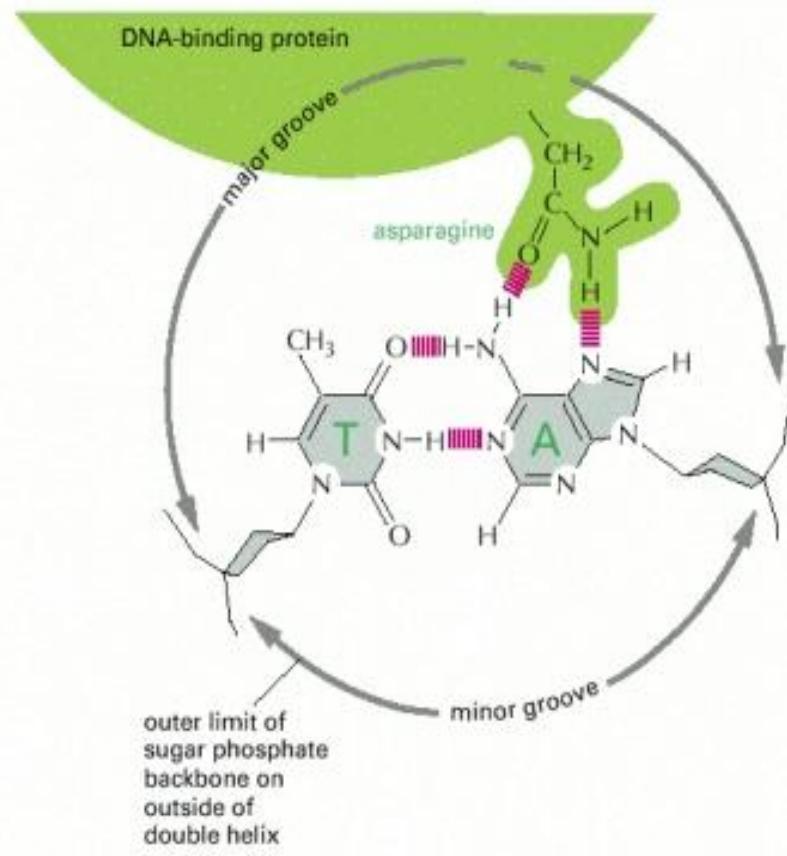
DNA recognition code

- The four possible configurations of base pairs are shown,
- potential **hydrogen bond donors** indicated in **blue**
- potential **hydrogen bond acceptors** in **red**
- hydrogen bonds of the base pairs are indicated as a series of short parallel **pink** lines.
- Methyl groups, which form **hydrophobic protuberances**, are shown in **yellow**
- Hydrogen atoms that are attached to carbons, and are therefore unavailable for hydrogen bonding, are **white**.



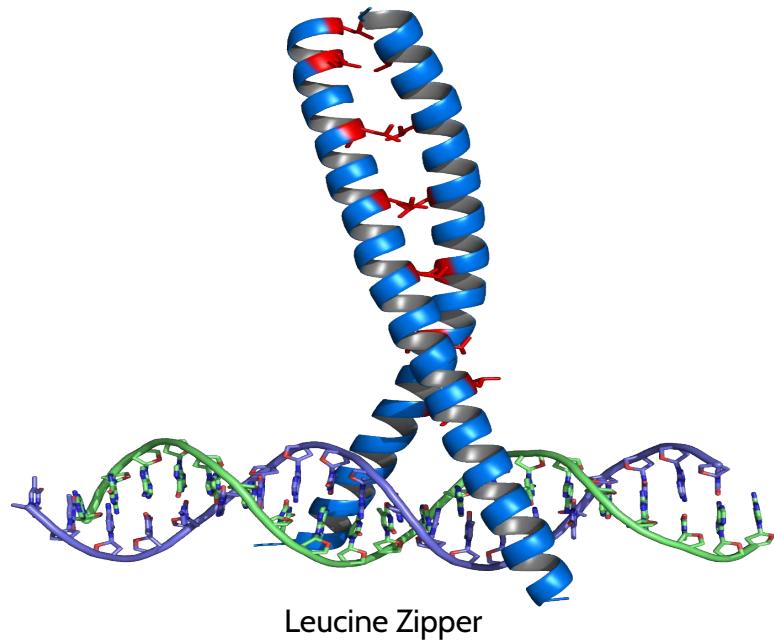
Transcription Factor-DNA interactions

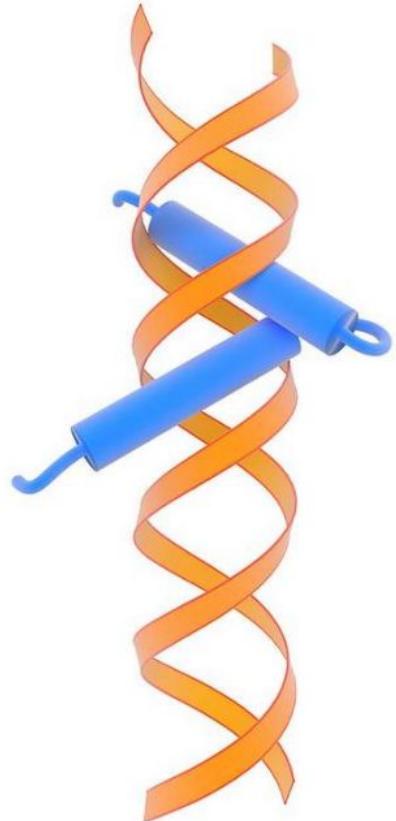
- Gene regulatory proteins recognise and bind to the **major groove** of DNA.
- Typically the protein-DNA interface would consist of 10-20 such amino acid contacts, each contributing to the **binding energy of the protein-DNA interface**.



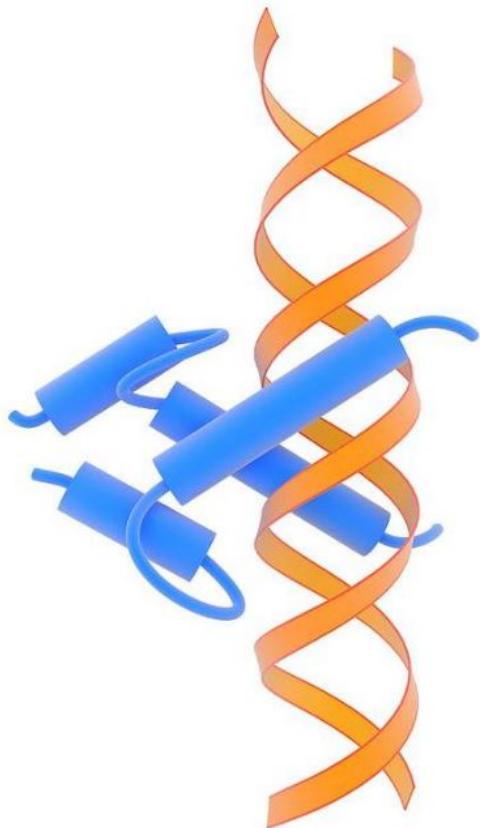
Transcription Factors (TFs)

- More than ~2000 in the human genome
- Contain one or more **DNA binding** and **Transactivation** domains
- Acts as single molecules, homo or heteromeric complexes.
- Classed by structural DNA recognition domains of TFs:
 - Helix-Turn-Helix
 - Zinc Finger
 - Leucine Zipper
 - Helix-Loop-Helix
 - Others





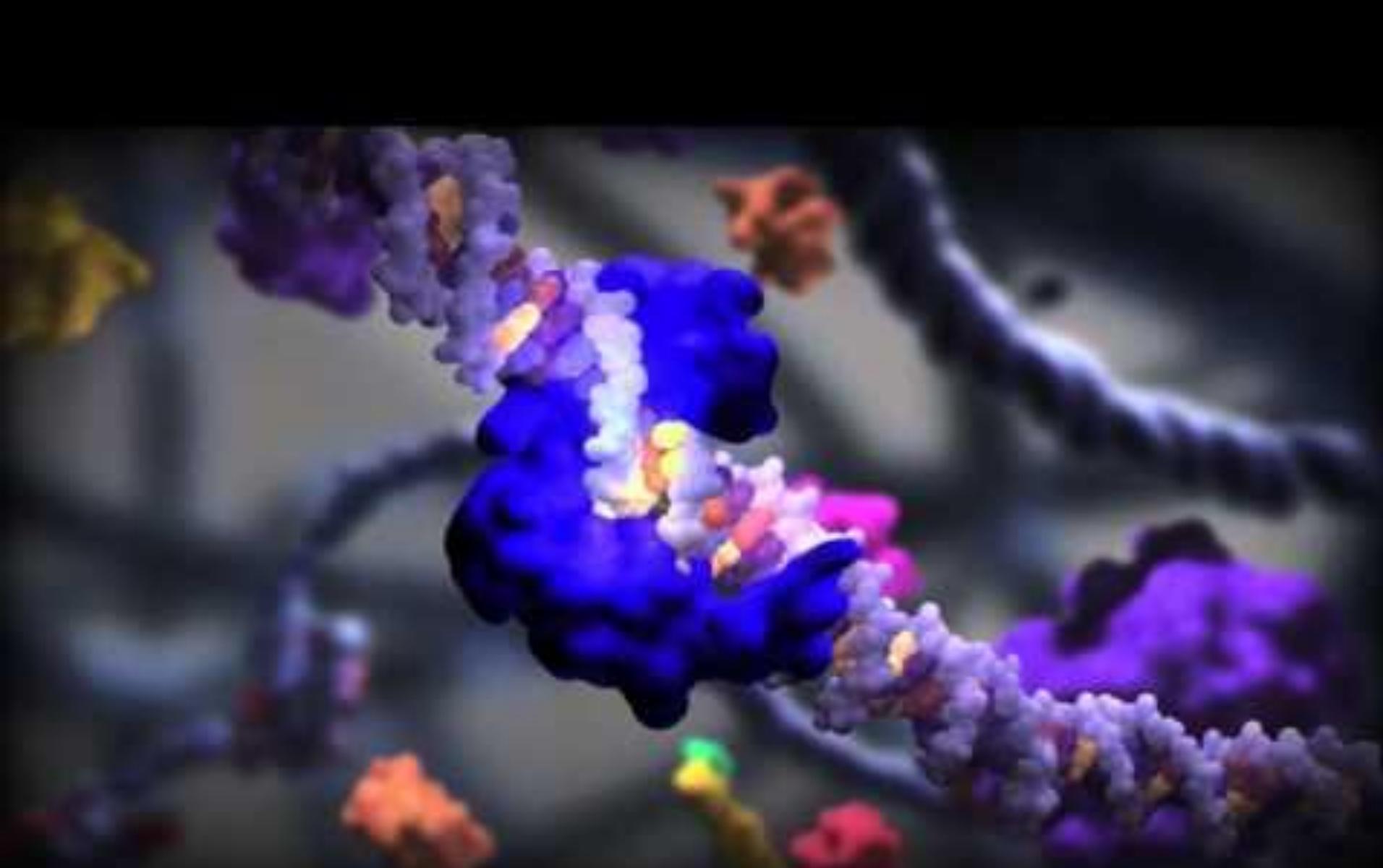
Helix-Turn-Helix



Helix-Loop-Helix



Zinc Finger



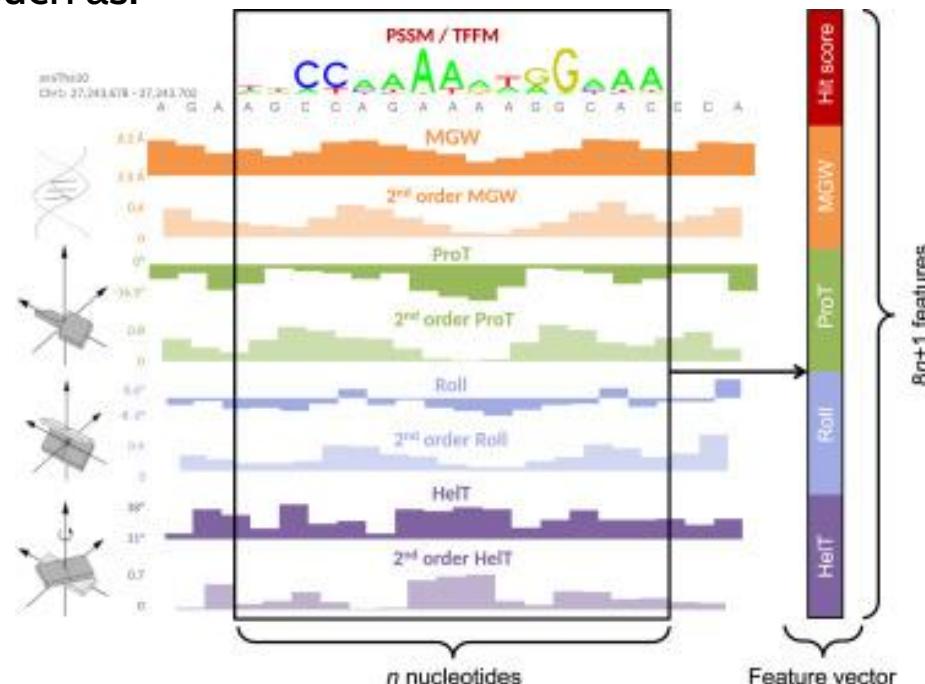
Specialised classes of TFs

- **Master Regulators:** Key transcription factors at the top of the regulatory hierarchy. They regulate multiple processes and phenotypes. (ex: TP53, NFkB, MYC etc.)
- **General Transcription Factors:** GTFs, RNA polymerase, and the mediator protein complex constitute the basic transcriptional apparatus that bind to the promoter, and starts transcription. **GTF** consists of TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH.
- **Pioneer factors:** TFs that can directly bind condensed chromatin and can have positive and negative effects on transcription. These are important in recruiting other transcription factors and histone modification enzymes as well as controlling DNA methylation. (ex: PU.1, SP1, AP2, GATA, FOXA1, KLF4)
- **Tissue specific factors:** expressed in specific tissue or cell type (ex: HNF factors in liver)

Transcription Factors can also recognize DNA shape independently of nucleotide recognition

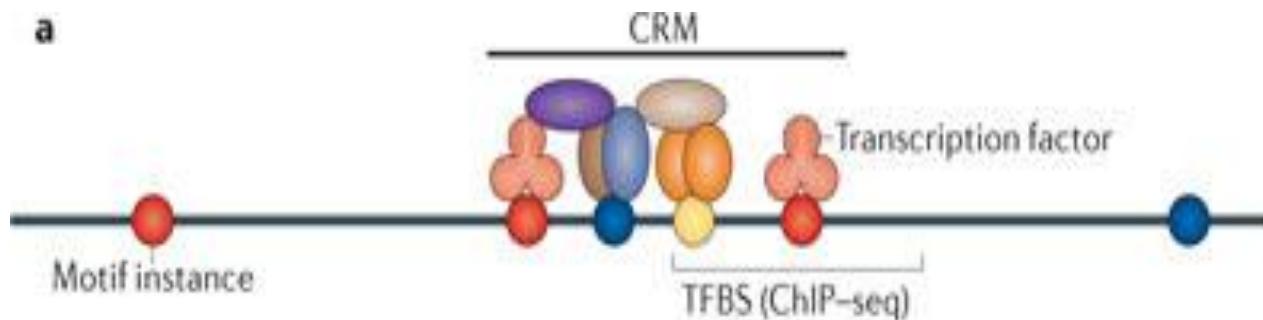
- Incorporating DNA sequence and shape information benefits the modeling of TF binding under complex *in vivo* conditions. This includes structural features associated with the minor groove such as:

- Helical twist
- Minor Groove width
- Propeller twist
- Roll



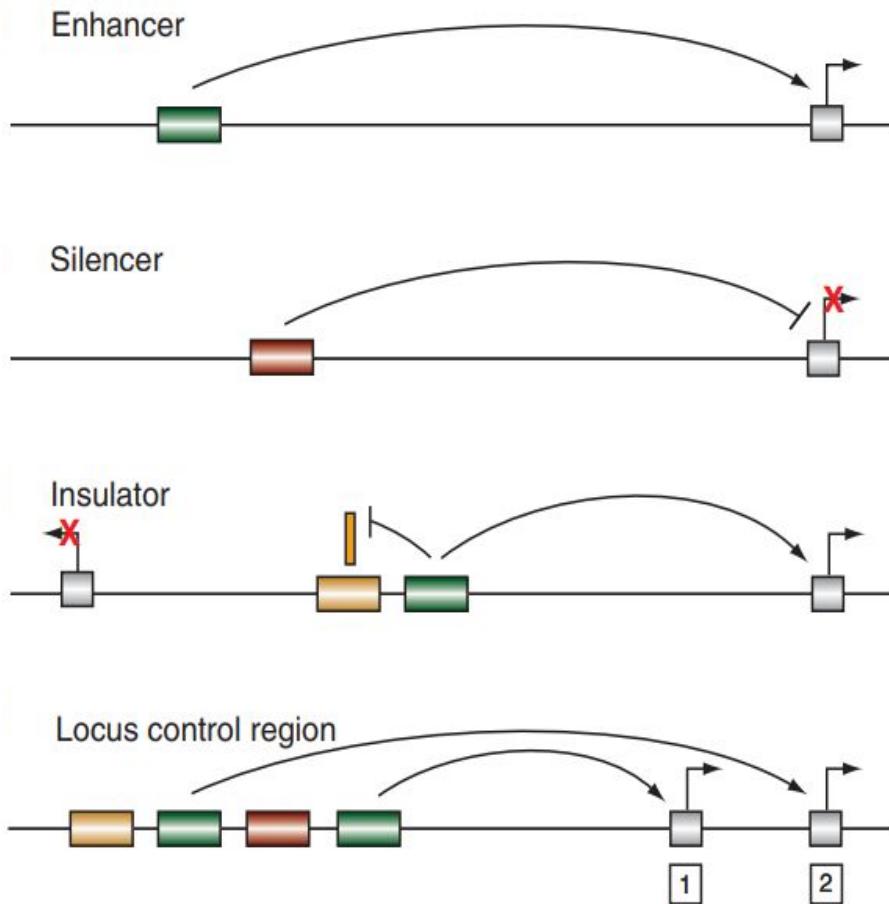
Cis Regulatory Modules (CRMs)

- Non-random clusters of TF binding sites that usually span a few 100bp.
- These dictate when, where and how genes are expressed.
- Can be **repressive** or **activating**, and output is the **combinatorial product** of multiple operations.
- Need to experimentally verify the function of CRMs to understand what the regulatory sequence does.
- Need robust computational methods to identify CRMs!



Classes of regulatory elements

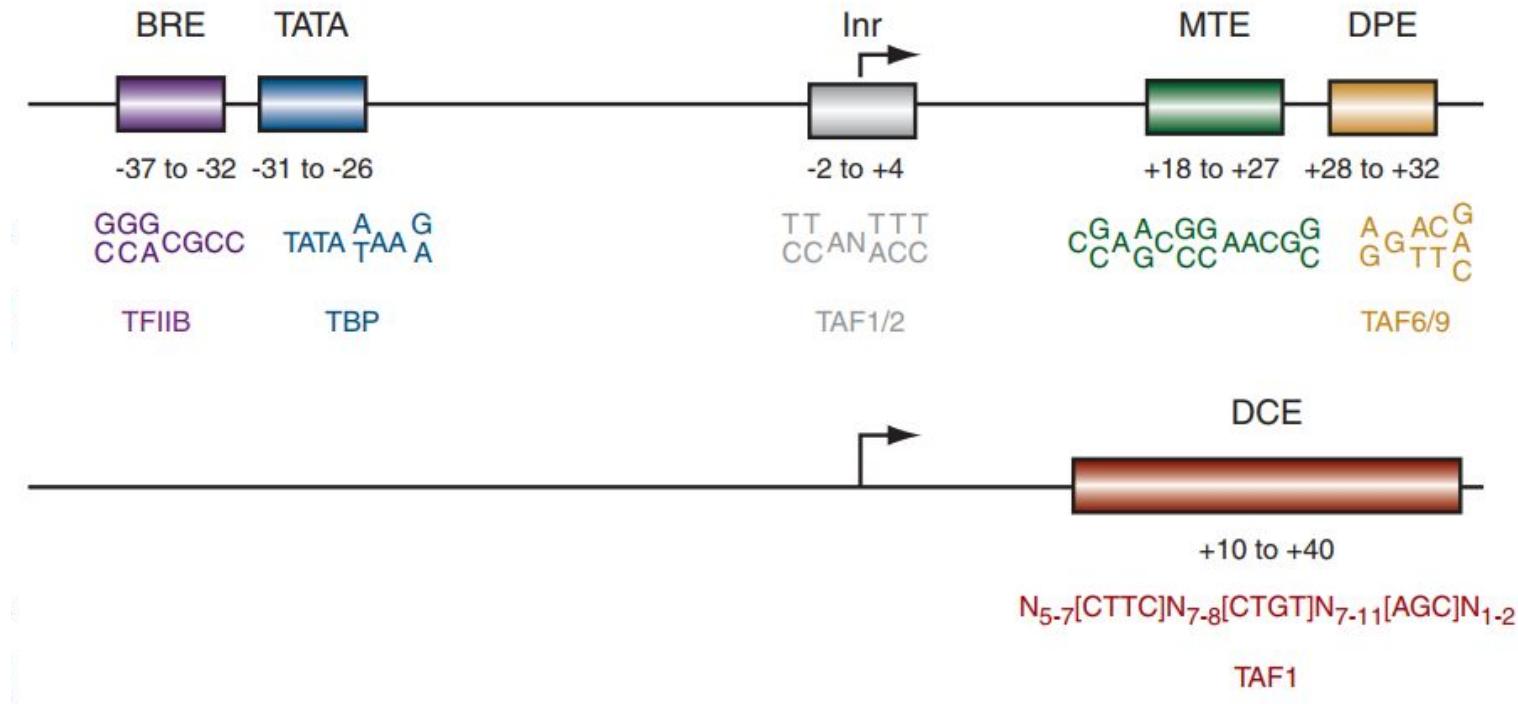
- Core Promoters
- Proximal Promoter
- Enhancers
 - Super enhancers
 - Stretch enhancers
 - Shadow or Distributed enhancers
 - Genic enhancers
 - Locus Control Regions
- Silencers
- Insulators



1. The Core Promoter

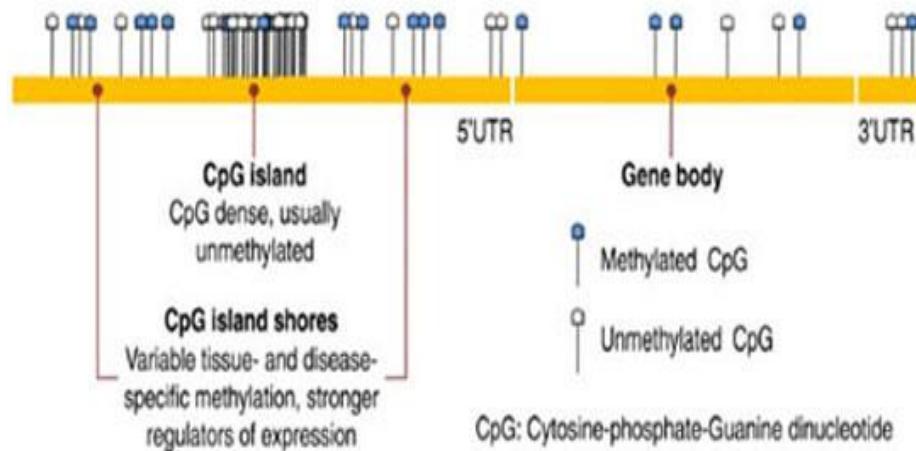
- **Core or minimal promoters** are regions present in **RNA PolIII** transcribed genes and are sufficient to mediate gene expression in a [reporter assay](#).
- **35-40 nt** up or downstream of TSS. May extend to 350 nt
- Contains core promoter motifs that interact with basal transcription machinery:
 - **TATA** box - binds TBP - bound by TFIID complex
 - **BRE** (B recognition element) - bound by TFIIB complex
 - **Inr** (Initiator element) - TFIID complex
 - **DPE** (Downstream promoter element)- TFIID complex
 - **MTE** (Motif ten element) - TFIID complex
 - **DCE** (Downstream core element)- TFIID complex
 - **CAAT** box - binds NF-Y & CEBP family proteins
 - **GC** box - binds SP1, EGR, MIG1, CREA (110 nt from TATA box)

Core promoter elements



2. Proximal promoters

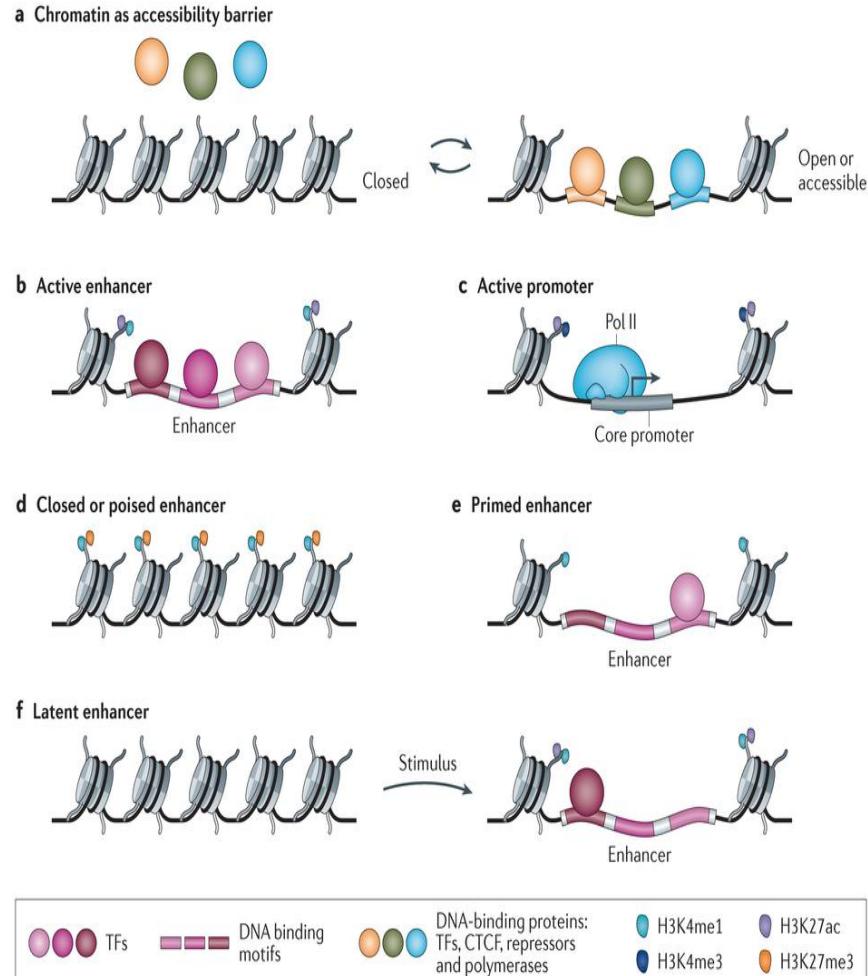
- Proximal promoters are immediately upstream (~1-5 kb) of the core promoter and contain multiple TFBS.
- **CpG islands:** are a class of proximal promoter with stretches of high GC content with over represented in CpG dinucleotides ranging between 0.5 -2 kb.
- Lacks TATA or DPE elements, contain multiple sites for GC box and SP1.
- Transcription at CGI promoters initiates from multiple weak TSS.
- Promoter-CGI **hyper methylation** can lead to transcriptional repression.



CpG Islands, shores, and methylation sites in relation to a hypothetical gene

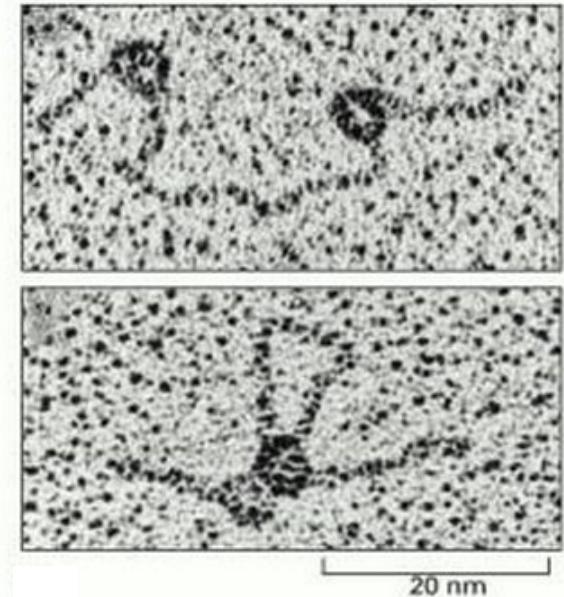
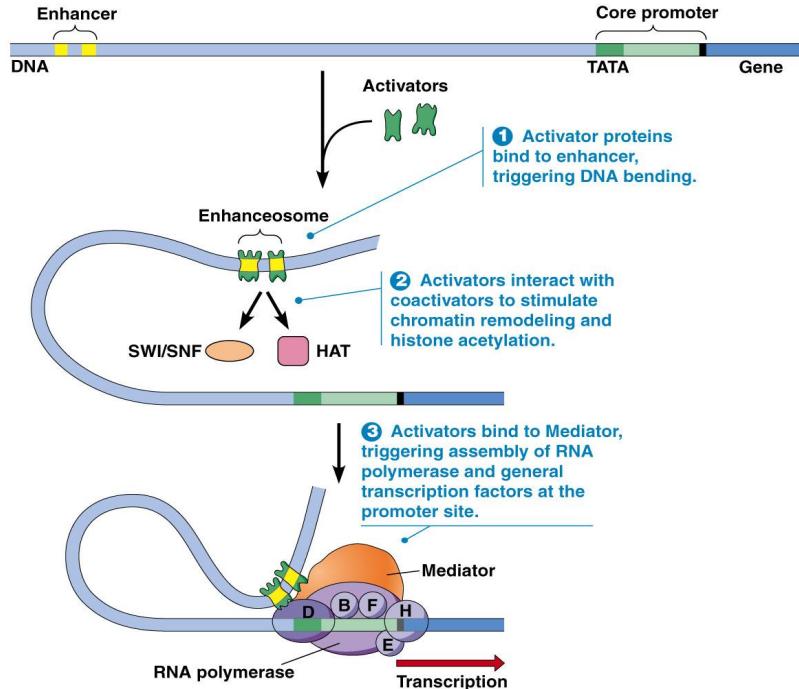
3. Enhancers

- The **core promoter** alone allows a basal level of transcription to occur. When the core promoter is removed from the gene, no transcription occurs.
- An **enhancer** alone cannot substitute for the promoter region, but combining an enhancer with a core promoter results in a **significantly higher level** of transcription than occurs with the promoter alone.
- This increase in transcription is observed even when the enhancer is
 - moved farther upstream
 - inverted in orientation
 - moved to the 3 prime side of the structural gene.
- A number of histone marks (high H3K4me1 and H3K27Ac, low H3K4me3) are indicative of enhancers.



Enhancers types

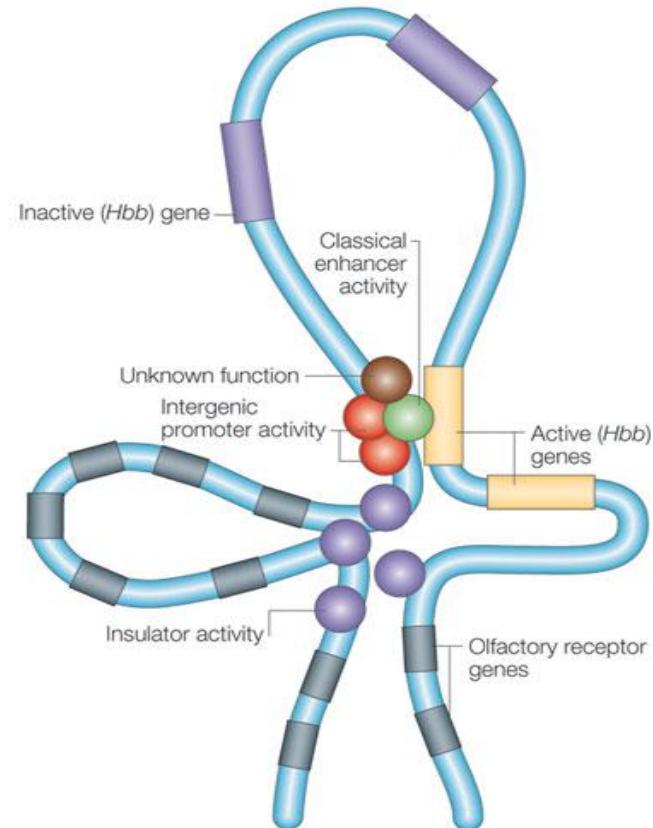
- **Enhancers:** Distal cis-regulatory elements that control spatiotemporal patterns and quantitative levels of gene transcription. Interact with promoters through looping.
- **Stretch Enhancers:** Cell type specific and enhance tissue specific expression
- **Super Enhancers:** Clusters of active enhancers bound by master transcription factors and the Mediator complex, have been found to drive high-level expression of genes that control cell identity.



Alberts, "Molecular Biology of the Cell", 2002

Locus Control Regions

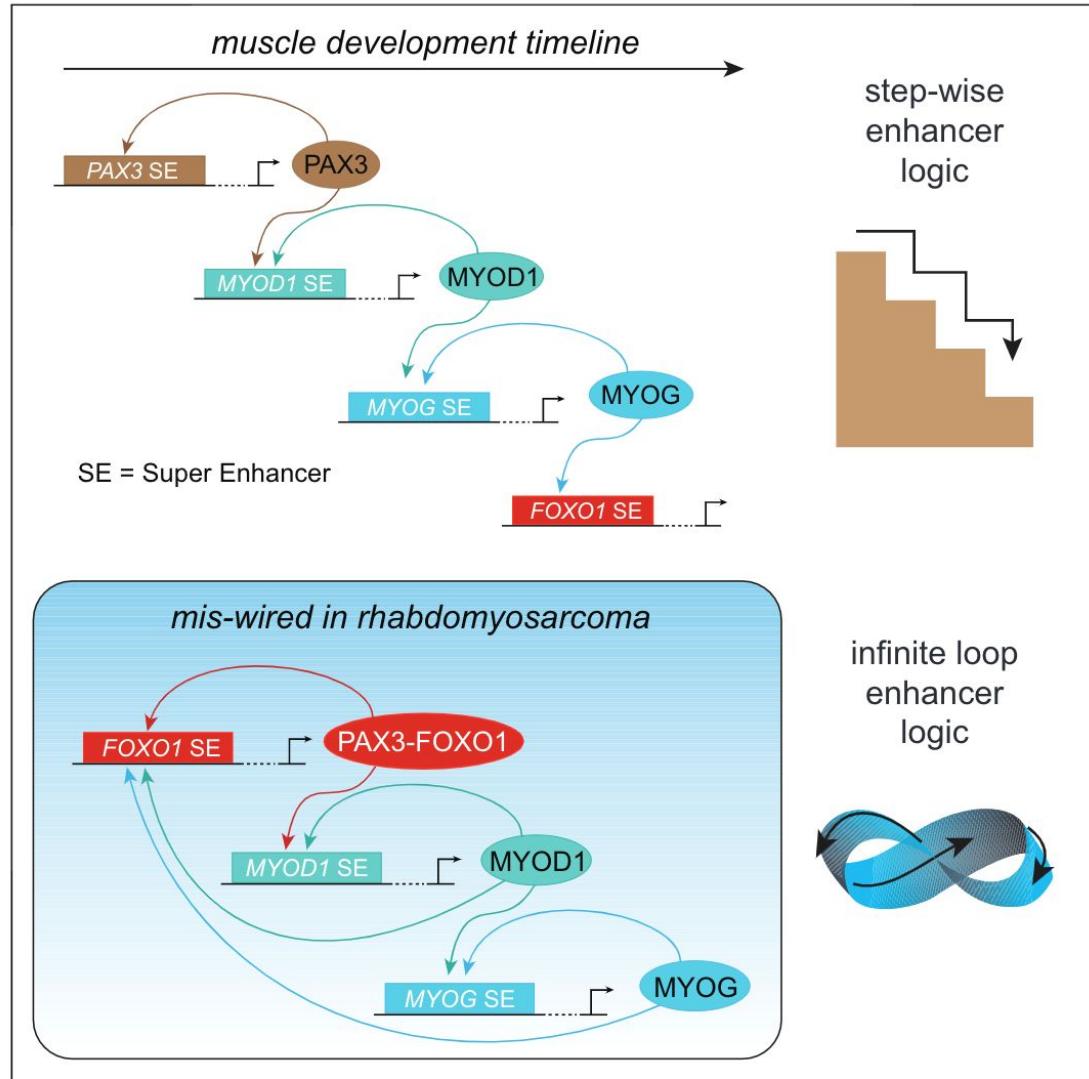
- Locus control regions (**LCRs**) are operationally defined by their ability to enhance the expression of **linked genes** to physiological levels in a tissue-specific and copy number-dependent manner at ectopic chromatin sites.
- Mouse globin gene cluster (*Hbb*) regulated by distal LCR (50kb away) by forming specific higher order looping structures



Copyright © 2005 Nature Publishing Group
Nature Reviews | Genetics

Enhancer “Rewiring”

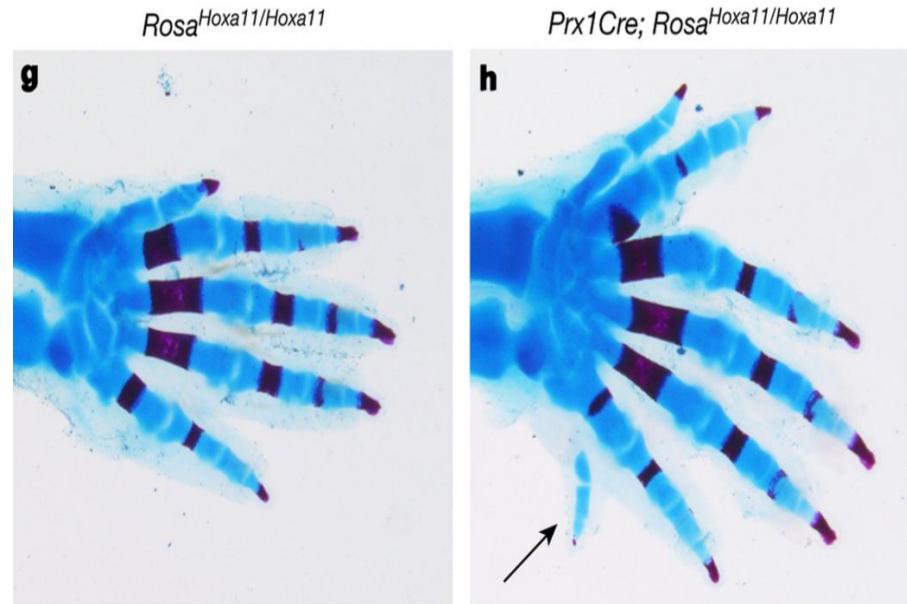
- TF activity establish enhancers with logical ordering during embryogenesis and development.
- In fusion-positive rhabdomyosarcoma, a cancer of the muscle lineage, the chief oncogene *PAX3-FOXO1* is driven by a translocated *FOXO1* super enhancer (SE) restricted to a late stage of myogenesis



Pentadactyl limb formation in tetrapods

- The earliest tetrapods were polydactyle with 6,7 or 8 digits per limb.
- Most living tetrapods — the four-legged land vertebrates — have five digits per limb or are **pentadactyle**. Mutations usually result in less than 5 digits.
- Mutually exclusive expression of *Hoxa11* and *Hoxa13* genes is required for maintaining pentadactyly.
- An enhancer has evolved (not present in fish) in the intron of *Hoxa11* gene (which is regulated by *Hoxa13*, and generates antisense transcription of *Hoxa11*).
- Mutation of this enhancer leads to *Hoxa11* expression in distal limb buds and triggers extra digit formation.

The authors propose that the evolution of *Hoxa11* regulation has contributed to the transition from polydactyly in extinct tetrapods to pentadactyly in extant tetrapods.



The hand skeleton of the *Rosa26^{Hoxa11/Hoxa11}; prx1Cre* mutant mouse.

Kherjamil et al., Nature 2016

4. Silencers and Repressors

- Silencers are regulatory regions that repress gene expression when bound by repressor TFs.

Candidate silencer elements for the human and mouse genomes

Naresh Doni Jayavelu, Ajay Jajodia, Arpit Mishra & R. David Hawkins [✉](#)

Nature Communications 11, Article number: 1061 (2020) | [Cite this article](#)



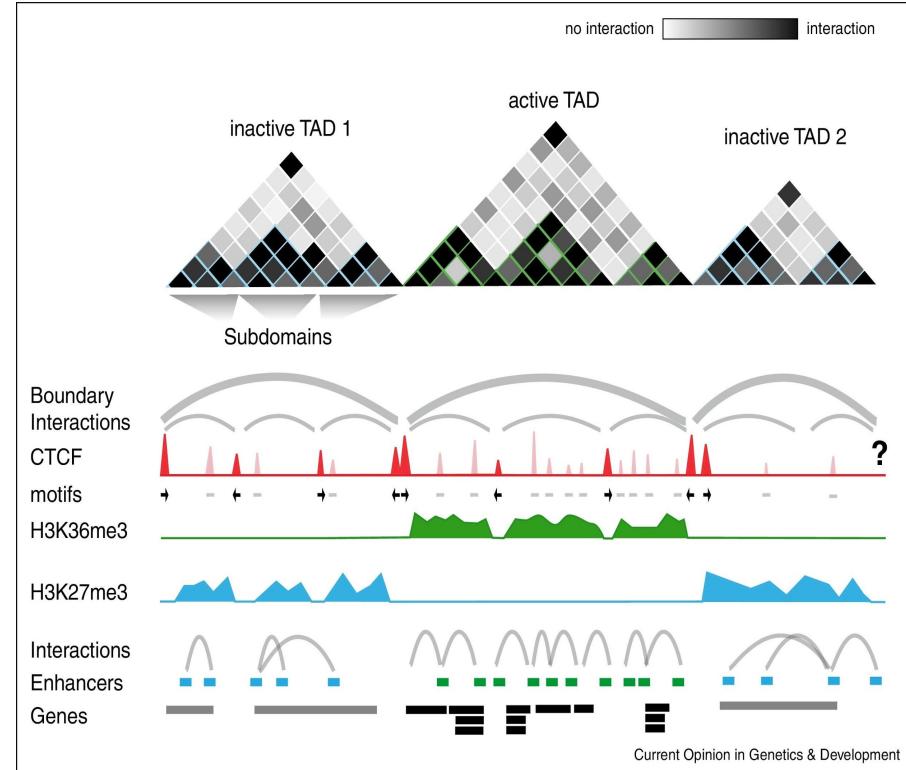
ARTICLE | VOLUME 77, ISSUE 2, P324-337.E8, JANUARY 16, 2020

Transcriptional Silencers in *Drosophila* Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts

Stephen S. Gisselbrecht ⁷ • Alexandre Palagi ⁷ • Jesse V. Kurland [●] ... Ye Zhan [●] Job Dekker [●] Martha L. Bulyk ⁸ [✉](#) [✉](#)

5. Insulators

- Insulator elements prevent the activation of a promoter by an enhancer placed between them.
- In vertebrates, insulator function has primarily been attributed to 3D chromatin structures mediated by CTCF-binding factor (CTCF).



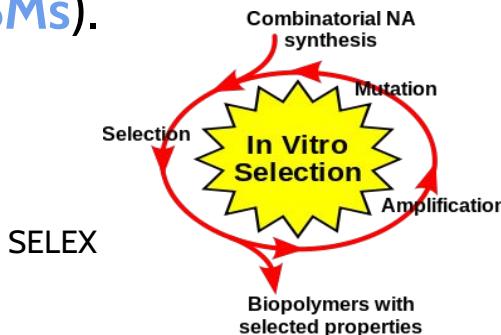
Tamer et al., Curr. Opin. Genet. Dev. 2016

Human Promoter Resources

Gene Promoter Resources	URL
Eukaryotic Promoter Database	http://epd.vital-it.ch/
DBTSS	http://dbtss.hgc.jp/
FANTOM5	https://fantom.gsc.riken.jp/data/
Ensembl regulatory build	http://www.ensembl.org/info/genome/funcgen/regulatory_build.html
UCSC (ENCODE) regulation tracks	http://genome.ucsc.edu/

In silico identification of DNA regulatory elements

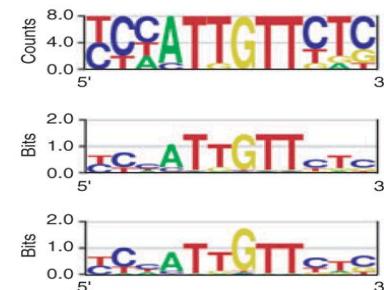
- Sequence **motifs**: short, recurring patterns in DNA or RNA associated with biological function.
- Sequence specific DNA binding sites for proteins (**nucleases, TFs**) or at the RNA level (**ribosome binding, mRNA processing and transcription termination**).
- Experimentally detected by ChIP, DNase footprinting, EMSA or reporter constructs.
- Binding affinities are explored using **SELEX** (Systematic Evolution of Ligands by *in vitro* sElectioN) assays or Protein Binding microarrays (**PBMs**).



HEM13	CCCA _A TTGTTCTC
HEM13	TTT _C TGGTTCTC
HEM13	TCA _A TTGTTAG
ANB1	CTCATTGTTGTC
ANB1	TCCATTGTTCTC
ANB1	CCTATTGTTCTC
ANB1	TCCATTGTTCGT
ROX1	CCAATTGTTTG

YCHATTGTTCTC

A	00270000010
C	464100000505
G	000001800112
T	422087088261



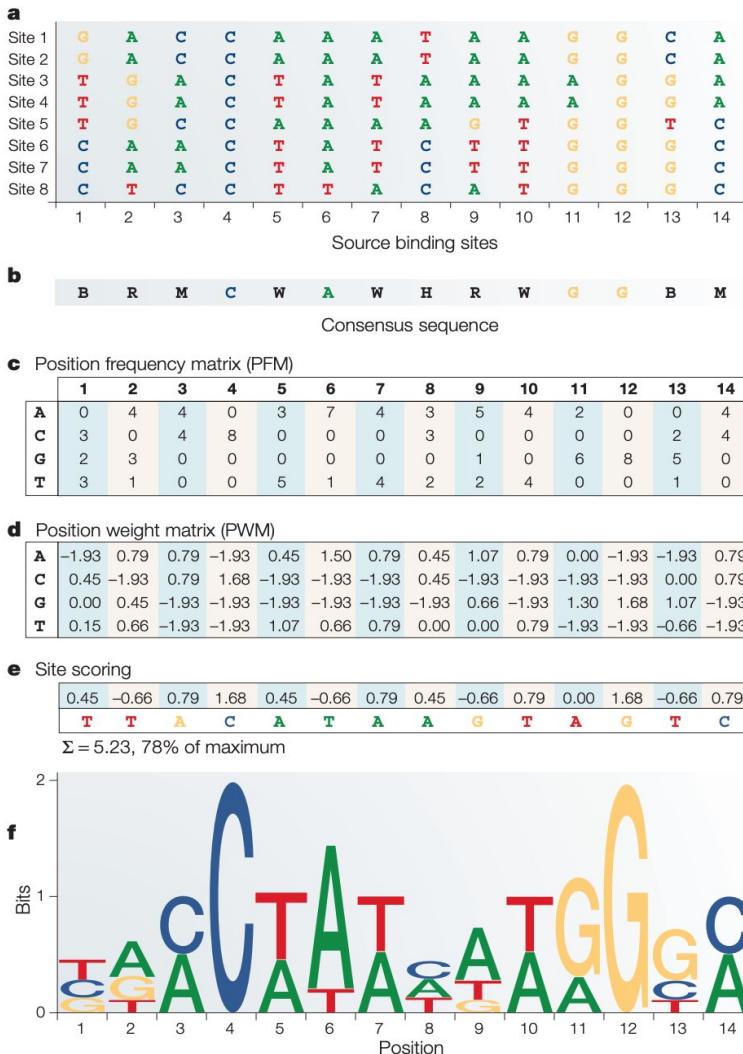
Scanning vs *de novo* motif identification

Don't scan a sequence with a **motif model** and expect all sites identified to be biologically active. Random matches will swamp the biologically relevant matches! This is a well known problem in motif searching, known as the "**Futility Theorem**" of motif finding. (Wasserman and Sandelin, *Nat. Rev. Genet.* 2004;5:276-87)

- **PWM** profile based **sequence scanning** methods. These methods uses prior information about TF binding sites and therefore can only be used to detect known Transcription Factor Binding Sites (**TFBS**).
- ***De novo*** motif identification – Pattern discovery methods:
 - **Word based** – occurrence of each ‘word’ of nucleotides of a certain length is counted and compared to a background distribution.
 - **Probabilistic** – seek the most **overrepresented** pattern using algorithmic approaches like **Gibbs sampling** and **Expectation Maximization**. These iteratively evolve an initial random pattern until a more specific one is found.
- Use *de novo* motif calling and alignment to build your own PWMs! **Biostrings** and **Motiv** Bioconductor packages have PFM to PWM conversion methods.

Representing binding sites: PFM and PWM

- Experimental binding data is collected and sequences are aligned.
- **Consensus sequence:** IUPAC codes that captures the sequence degeneracy of binding sites at each position.
- **Position Frequency Matrix (PFM):** Matrix encoding the nucleotide frequency at each position.
- **Position Specific Scoring Matrix (PSSM) or Position Weight Matrix (PWM):** Normalised frequency values of a PFM are converted into a log scale to get a PWM. For large and representative sequence collections the scores are proportional to binding energies.
- **Sequence Logo:** column specificity measured as information content for each nucleotide is scaled by the total bits of information multiplied by the relative occurrence of that nucleotide at that position.



Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

1

Corrected probability calculation:
$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \quad (1)$$

$f_{b,i}$ = counts of base b in position i ; N = number of sites; $p(b,i)$ = corrected probability of base b in position i ; $s(b)$ = pseudocount function

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (1) by expected background probabilities and converting the values to a log-scale (see equation 2).

2

PWM conversion:
$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

$p(b)$ = background probability of base b ; $p(b,i)$ = corrected probability of base b in position i ; $W_{b,i}$ = PWM value of base b in position i

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site (see equation 3).

3

Evaluation of sequences:
$$S = \sum_{i=1}^w W_{l_i, i} \quad (3)$$

l_i = the nucleotide in position i in an input sequence; S = PWM score of a sequence; w = width of the PWM

Probability values (1) can be used to determine the total information content (in bits) in each position (see equation 4).

4

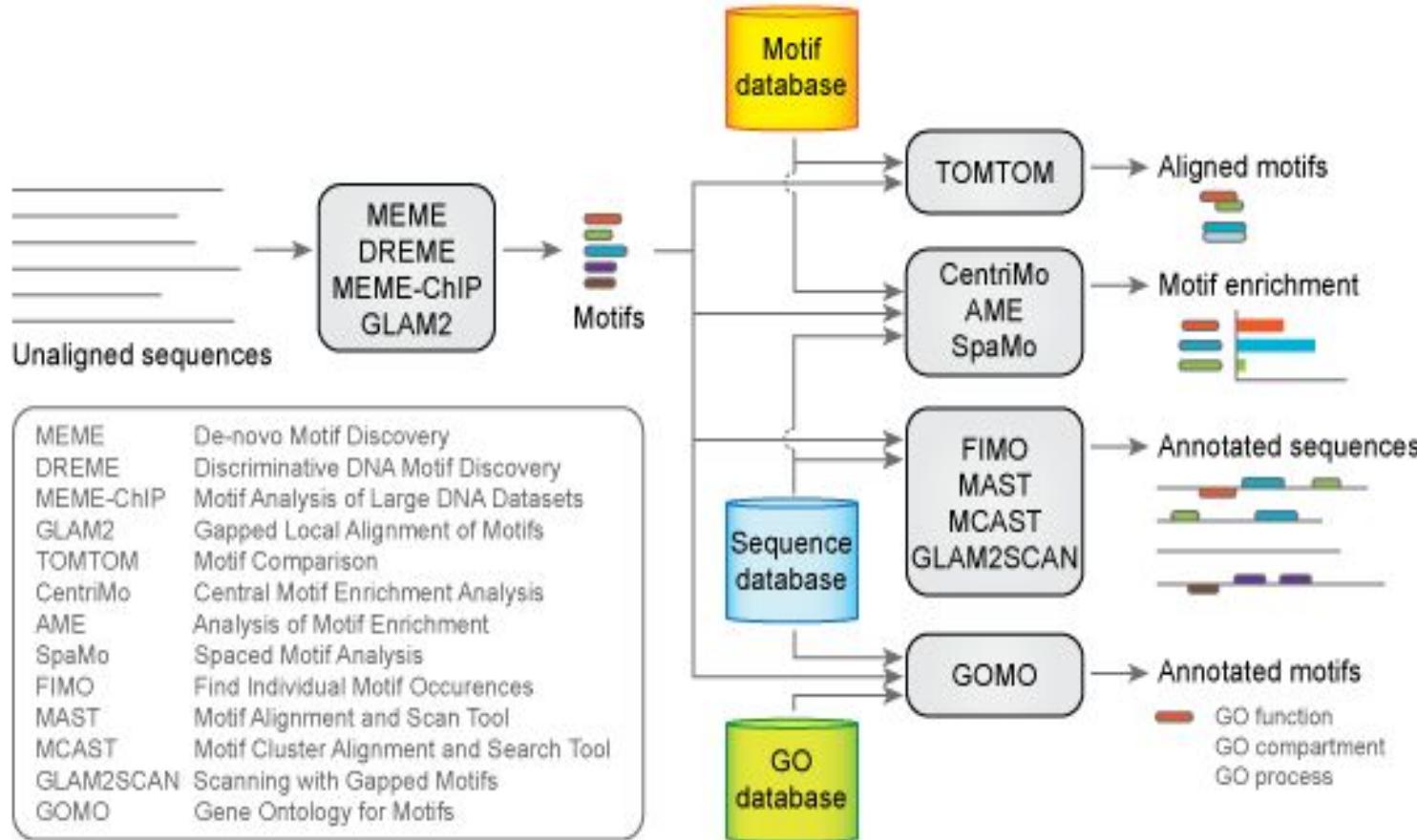
Information content calculation:
$$D_i = -\sum_b p(b,i) \log_2 p(b,i) \quad (4)$$

D_i = information content in position i ; $p(b,i)$ = corrected probability of base b in position i

Transcription Factor Resources

Transfac Professional	http://genexplain.com/transfac/
Jaspar	http://jaspar.genereg.net/
UniProbe	http://thebrain.bwh.harvard.edu/uniprobe/
Hocomoco	https://hocomoco11.autosome.ru/
Cistrome	http://cistrome.org/db/
ChIP-Atlas	https://chip-atlas.org/
OregAnno	http://www.oreganno.org/oregano/
FactorBook	https://www.factorbook.org/
hmChIP	http://jilab.biostat.jhsph.edu/database/cgi-bin/hmChIP.pl
ChipBase	http://deepbase.sysu.edu.cn/chipbase/index.php
SwissRegulon	https://swissregulon.unibas.ch/sr/
Human Protein-DNA Interactome	http://bioinfo.wilmer.jhu.edu/PDI/FAQ.html

Meme Suite



HOMER (Hypergeometric Optimization of Motif Enrichment)

<http://homer.salk.edu/homer/index.html>



- Large number of (Perl and C++) tools for ChIP-seq motif and peak analysis.
- Provides both *de novo* and PWM scanning based motif identification and enrichment analysis.
- User can specify custom background. (Randomly selected, GC or CGI matched backgrounds.)
- Uses a collection of ChIP-seq derived PWMs or user can specify PWM.
- Peak annotation, GO enrichment analysis, Extract peak sequences, Visualization.

Meta-Motif Analyzers

GimmeMotifs: a *de novo* motif prediction pipeline, especially suited for ChIP-seq datasets. It incorporates several existing motif prediction algorithms in an ensemble method to predict motifs and clusters these motifs using the weighted information content (WIC) similarity scoring metric.

<https://gimmemotifs.readthedocs.io/en/master/index.html>

BioProspector <http://motif.stanford.edu/distributions/bioprospector/>

GADEM <http://www.niehs.nih.gov/research/resources/software/gadem/index.cfm>

Improbizer <http://users.soe.ucsc.edu/~kent/>

MDmodule (included in the MotifRegressor Package) <http://www.math.umass.edu/~conlon/mr.html>

MEME <http://meme.sdsc.edu/>

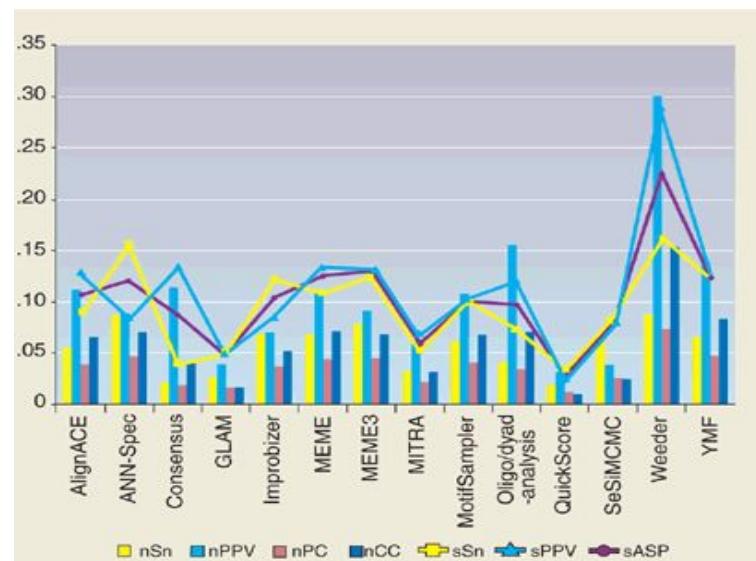
MoAn <http://moan.binf.ku.dk/>

MotifSampler <http://homes.esat.kuleuven.be/~sistawww/bioi/thijs/download.html>

Trawler <http://ani.embl.de/trawler/>

Weeder <http://159.149.160.51/modtools/>

Assessment of tools for TFBS discovery



Tompa et al., "Assessing computational tools for the discovery of transcription factor binding sites", Nature biotechnology 23 (1), 137-144

Table 1 Details about the operation principles, basic technical data and URLs of 13 analyzed tools

Program	Operating principle	Technical data	URL	Reference
AlignACE	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are overrepresented in the input set	Judges alignments sampled during the course of the algorithm using a maximum <i>a priori</i> log likelihood score, which gauges the degree of overrepresentation. Provides an adjuct measure (group specificity score) that takes into account the sequence of the motifs and the motifs for each motif found preferentially in association with the genes under consideration.	http://atlas.med.harvard.edu/	7
ANN-Spec	Models the DNA-binding specificity of a transcription factor using a weight matrix	Objective function based on log likelihood that transcription factor binds at least once in each sequence of the positive training data compared with the number of times it is estimated to bind in the background training data. Parameter fitting is accomplished with a gradient descent method, which includes Gibbs sampling of the positive training examples.	http://www.cbs.dtu.dk/~workman/ann-spec/	8
Consensus	Models motifs using weight matrices, searching for the matrix with maximum information content	Uses a greedy method, first finding the pair of sequences that share the motif with greatest information content, then finding the third sequence that can be added to the motif resulting in greatest information content.	http://bifrost.wustl.edu/consensus/	9
GLAM	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output	Since the basic algorithm could not find multiple motif instances per sequence, long sequences were fragmented into shorter ones, and the alignment was transformed into a weight matrix and used to scan the sequences to obtain the final site predictions.	http://zlab.bu.edu/glam/	10
The Improbizer	Uses expectation maximization to determine weight matrices of DNA motifs that occur improbably often in the input sequences	As a background (null) model it uses up to a second-order Markov model of background sequence. Optionally, Improbizer constructs a Gaussian model of motif placement, so that motifs that occur in similar positions in the input sequences are more likely to be found.	http://www.soe.ucsc.edu/~kent/improbizer	11
MEME	Optimizes the E-value of a statistic related to the information content of the motif	Rather than sum of information content of each motif column, statistic used is the product of the P values of column information content. This is a search column of performing expectation maximization from starting points derived from each subsequent occurring in the input sequences. MEME differs from MEME3 mainly in using a correction factor to improve the accuracy of the objective function.	http://meme.sdsc.edu/	12
MITRA	Uses an efficient data structure to traverse the space of IUPAC patterns	For each pattern, MITRA computes the hypergeometric score for the occurrences in the target sequences relative to the background sequences and reports the highest scoring patterns.	http://www.calit2.net/compbio/mitra/	13
MotifSampler	Matrix-based, motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model	The probabilistic framework is further exploited to estimate the expected number of motif instances in the sequence.	http://www.esat.kuleuven.ac.be/~dna/BiolSoftware.html	14

Table 1 continued on following page

Table 1 Continued

Program	Operating principle	Technical data	URL	Reference
Oligo/dyad-analysis	Detects overrepresented oligonucleotides with oligo-analysis ¹⁵ and spaced motifs with dyad-analysis ¹⁶	These algorithms detect statistically significant motifs by counting the number of occurrences of each word or dyad and comparing those with expectation. Most crucial parameter is choice of window size. It is also important the estimation of occurrence significance. In this study, a negative binomial distribution on word distributions was obtained from 1,000 random promoter selections of the same size as the test sets.	http://rsat.scmbb.ulb.ac.be/rsat/	15, 16
QuickScore	Based on an exhaustive searching algorithm that estimates probabilities of rare or frequent words in genomic texts	Incorporates an extended consensus method allowing well-defined mismatches and uses mathematical expressions for efficiently computing z-scores and P-values, depending on the statistical models used in their range of applicability. Special attention is paid to the drawbacks of numerical instability. The background model is Markovian, with order up to 3.	http://algo.inria.fr/dolley/QuickScore/	17
SeSiMCMC	Modification of Gibbs sampler algorithm that models the motif as a weight matrix, optionally with the symmetry of a palindrome or of a direct repeat, and optionally with spacers	Includes two alternating stages. The first one optimizes the weight matrix for a given motif and spacer length. The algorithm changes the positions of the motif occurrences in the sequences and infers the motif model from the current occurrences. These changes are used to optimize the likelihood of sequences as best as possible. The second stage optimizes the positions of motif occurrences. The optimization is organized via a Gibbs-like Markov chain, which samples positions in sequences one by one, until the Markov chain converges. The second stage looks for best motif and spacer lengths for obtained motif positions. It optimizes the common information content of motif and of distributions of motif occurrence positions.	http://favorov.hole.ru/gibbsfm/	18
Weeder	Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurrences (with substitutions) from input sequences	Each motif evaluated according to number of sequences in which it appears and how well conserved it is in each sequence, with respect to expected values derived from the oligo frequency analysis of the yeast genome. The sequences are taken from the yeast genome. Different combinations of 'canonical' motif parameters derived from the analysis of known instances of yeast transcription factor binding sites (length ranging from 6 to 12, number of substitutions from 1 to 4) are automatically tried by the algorithm in different runs. It also analyzes and compares the top-scoring motifs of each run with a simple clustering algorithm to determine which one could be most likely to correspond to a transcription factor binding site. Best instances of each motif are selected from sequences using a weight matrix built with sites found by consensus-based algorithm.	http://159.149.109.16/Tool/ind.php	19
YMF	Uses an exhaustive search algorithm to find motifs with the greatest z-scores	A P value for the z-score is used to assess significance of motif. Motifs themselves are short sequences over the IUPAC alphabet, with spacers ('N's) constrained to occur in the middle of the sequence.	http://bio.cs.washington.edu/software.html#ymf	20

Motif enrichment analysis (MEA)

- MEA determines which transcription factors control the transcription of a set of co-regulated genes by detecting enrichment of known binding motifs of TFs in the genes' regulatory regions.
- Is the motif *over or underrepresented* in the sample sequence set compared to a background sequence set?
- Identifying co-regulated gene sets is difficult. Use of Ontologies, pathways, GSEA etc is recommended.
- Picking the right background model will also contribute to the success of the motif enrichment analysis.

Background models

The background model select has a significant impact on detecting motifs

- All core-promoters from protein coding and/or non-coding genes etc.
- Window around TSS
- Higher order Markov model based backgrounds
- Open chromatin regions
- Shuffled sequence regions
- Known regulatory regions - ENCODE regulatory modules
- Exonic or Intronic regions- ex: 3rd exon
- Selecting sequences
 - Similar number or genome wide
 - Defined or variable length

MEME-ChIP

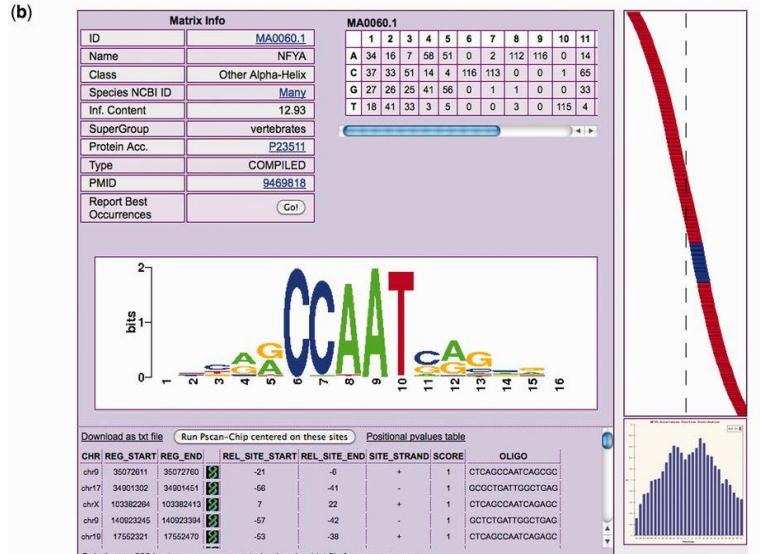
url: <http://meme.nbcr.net>

- Given a set of genomic regions, it performs
 - ab initio motif discovery -novel TF binding sites ([MEME](#), [DREME](#))
 - motif enrichment analysis -known TF enrichment ([Centrimo/AME](#))
 - motif visualization ([MAST](#) and [AMA](#))
 - binding affinity analysis
 - motif identification -compare to known motifs ([TOMTOM](#))
- Uses two algorithms for motif discovery:
 - MEME -expectation maximization (EM) to discover probabilistic models of DNA-binding by single TFs or TF complexes.
 - DREME -simpler, non-probabilistic model (regular expressions) to describe the short binding motifs.
- Motif identification:
 - [FIMO](#) -identify individual motifs

Machanick and Bailey, "MEME-ChIP: motif analysis of large DNA datasets." 2011 Bioinformatics

Pscan-ChIP

- Motif enrichment analysis using PWMs
- Profile databases and user defined background models.
- Optimized for ChIP-seq.
- Provides:
 - Ranked lists of enriched motifs.
 - Sequence logo's and motif enrichment distribution plots.

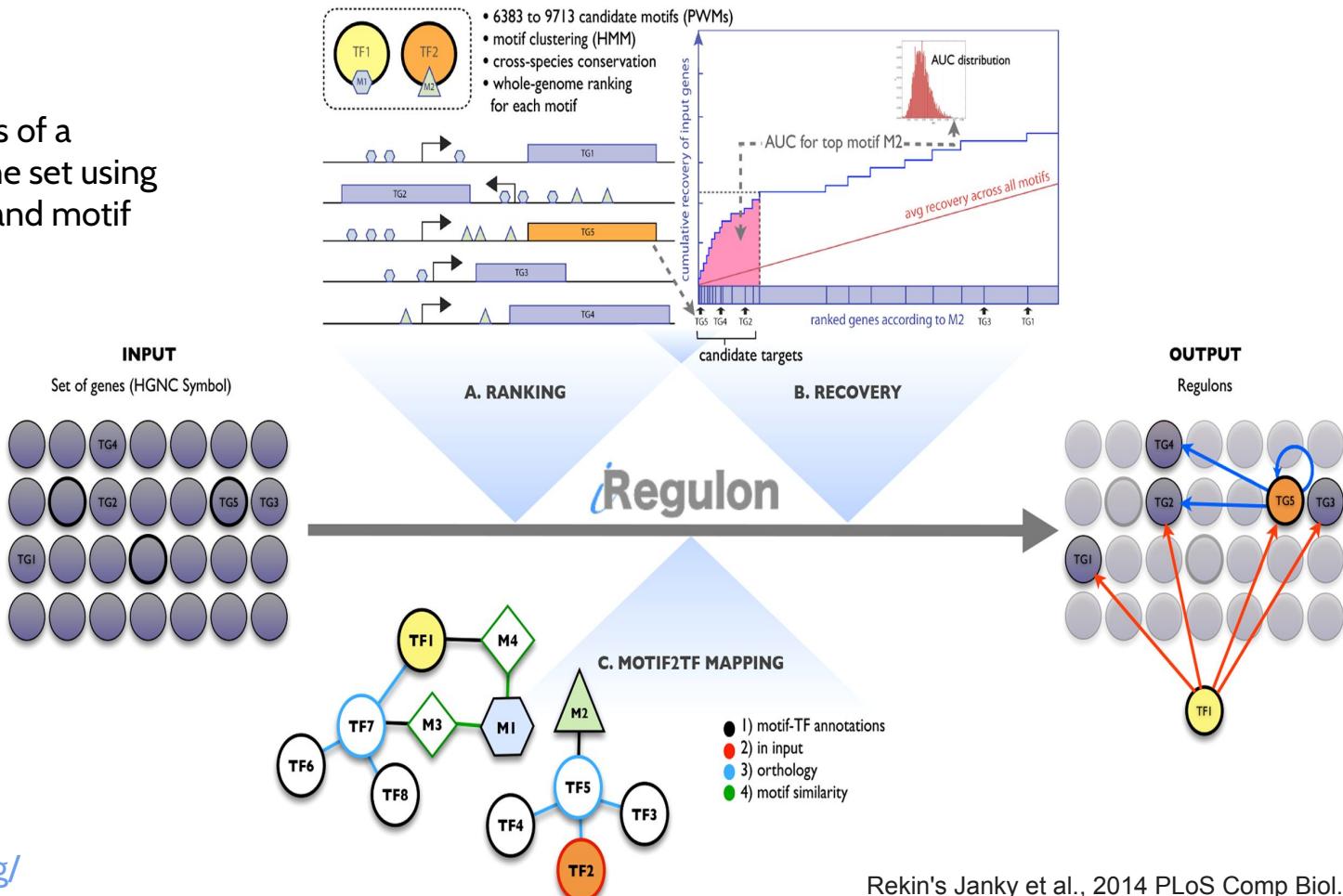


http://159.149.160.51/pscan_chip_dev/

"PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments." Zambelli et al., 2013 Nucleic Acids Res.

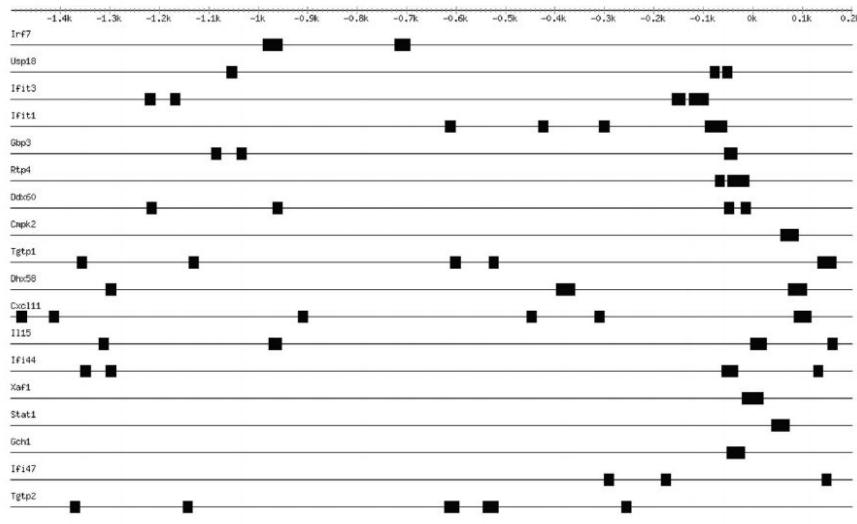
iRegulon

- Identify regulators of a co-expressed gene set using gene expression and motif enrichment.
- Cytoscape app.



Real world applications of Motif Enrichment Analysis

- Breast cancer metastasis is a key determinant of long-term patient survival.
- By comparing the transcriptomes of primary and metastatic tumor cells in a mouse model of spontaneous bone metastasis, we identified that a substantial number of genes suppressed in bone metastases are targets of the interferon regulatory factor IRF7.
- Restoration of *Irf7* activity in tumor cells or administration of interferon (which induces IRF7) led to reduced bone metastases and prolonged survival time.



nature
medicine

Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape

Bradley N Bidwell^{1,2,9}, Clare Y Slaney^{1,3,9}, Nimali P Withana^{1,4}, Sam Forster⁵, Yuan Cao^{1,2}, Sherene Loi⁶, Daniel Andrews^{1–3}, Thomas Mikeska^{1,4}, Niamh E Mangan⁵, Shamith A Samarajiwa^{5,7}, Nicole A de Weerd⁵, Jodee Gould⁵, Pedram Argani⁸, Andreas Möller^{1–4}, Mark J Smyth^{1,3}, Robin L Anderson^{1,3,4}, Paul J Hertzog⁵ & Belinda S Parker^{1–3}

References

Regulatory Elements

- Atkinson TJ, Halfon MS. "Regulation of gene expression in the genomic context." *Comput Struct Biotechnol J.* 2014 Jan 29;9:e201401001.
- Butler JE, Kadonaga JT. "The RNA polymerase II core promoter: a key component in the regulation of gene expression." *Genes Dev.* 2002 Oct 15;16(20):2583-92.
- Carninci P. et al., "Genome-wide analysis of mammalian promoter architecture and evolution." *Nat Genet.* 2006 Jun;38(6):626-35. Epub 2006 Apr 28. Erratum in: *Nat Genet.* 2007 Sep;39(9):1174.

Motifs

- Stormo GD. "DNA binding sites: representation and discovery." *Bioinformatics.* 2000 Jan;16(1):16-23.
- D'haeseleer P. "How does DNA sequence motif discovery work?" *Nat Biotechnol.* 2006 Aug;24(8):959-61.
- D'haeseleer P. "What are DNA sequence motifs?" *Nat Biotechnol.* 2006 Apr;24(4):423-5.
- Tompa M. et al., "Assessing computational tools for the discovery of transcription factor binding sites." *Nat Biotechnol.* 2005 Jan;23(1):137-44.
- Thijs G, "Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes." *J Comput Biol.* 2002;9(2):447-64.
- Wasserman WW, Sandelin A. "Applied bioinformatics for the identification of regulatory elements." *Nat Rev Genet.* 2004 Apr;5(4):276-87.
- Bidwell BN, et al., "Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape." *Nat Med.* 2012 Aug;18(8):1224-31.
- Bailey TL. et al., "MEME SUITE: tools for motif discovery and searching." *Nucleic Acids Res.* 2009 Jul;37(Web Server issue):W202-8. doi: 10.1093/nar/gkp335. Epub 2009 May 20.
- van Heeringen SJ, Veenstra GJ. "GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments." *Bioinformatics.* 2011 Jan 15;27(2):270-1. doi: 10.1093/bioinformatics/btq636. Epub 2010 Nov 15.

MEA

- Johnson DS, et al., "Genome-wide mapping of in vivo protein-DNA interactions." *Science.* 2007 Jun 8;316(5830):1497-502.
- Frith MC, "Detection of functional DNA motifs via statistical over-representation." *Nucleic Acids Res.* 2004 Feb 26;32(4):1372-81.
- McLeay RC, Bailey TL. "Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data." *BMC Bioinformatics.* 2010 Apr 1;11:165.
- Zambelli F, Pesole G, Pavesi G. "PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments." *Nucleic Acids Res.* 2013 Jul;41(Web Server issue):W535-43.
- Zambelli F, Pesole G, Pavesi G. "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era." *Brief Bioinform.* 2013 Mar;14(2):225-37. doi: 10.1093/bib/bbs016. Epub 2012 Apr 19.
- Machanick P, Bailey TL. "MEME-ChIP: motif analysis of large DNA datasets." *Bioinformatics.* 2011 Jun 15;27(12):1696-7.

Regulomes

- Samarajiwa SA & Kirschner K et al., "Phenotype specific analyses reveal distinct regulatory mechanism for chronically activated p53." *PLoS Genet.* 2015 Mar