
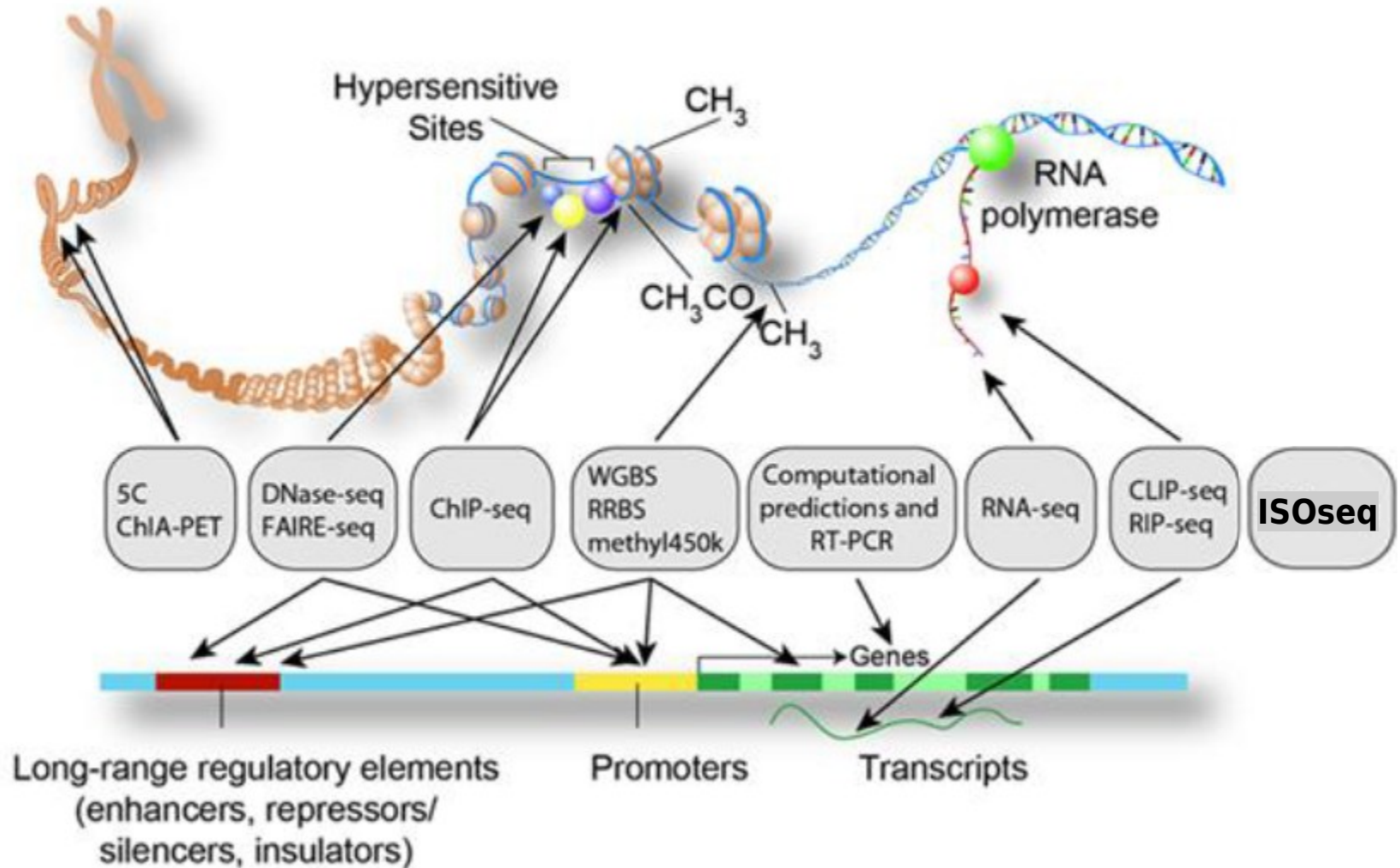


Syllabus and grading

#	Date	Short title	Lecturer	Subject
1	7/10/25	introduction	MR	Overview of Bioinformatics, sequence alignment
2	14/10/25	Linux/shell/ssh	AD	Introduction to Linux and the command line, bash scripting and ssh
3	21/10/25	QC+RNASeq	MR	Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
4	4/11/25	R (1)	AD	Introduction to the R programming language and Rstudio usage
5	11/11/25	R (2)	AD	Advances R subjects, introduction to Bioconductor
6	18/11/25	bedtools/vcftools/samtools fl	AD	Command line tool usage: bedtools, vcftools, samtools etc.
7	25/11/25	Denovo	MR	NGS for denovo genome and transcriptome assembly
8	2/12/25	Exome/SNP calling	AD	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
9	9/12/25	ChipSeq/chirp 	MR	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture (3C))
10	16/12/25	metabolomics	MR	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	13/1/26	presentations	MR+AD	Paper presentations by students
12	20/1/26	presentations	MR+AD	Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
13	27/1/26	final projects support	MR+AD	Support for the final project

Grade	100%
Presentation	30%
Exercises	20%
Final Project	50%

Functional Elements in the Genome



www.encodeproject.org

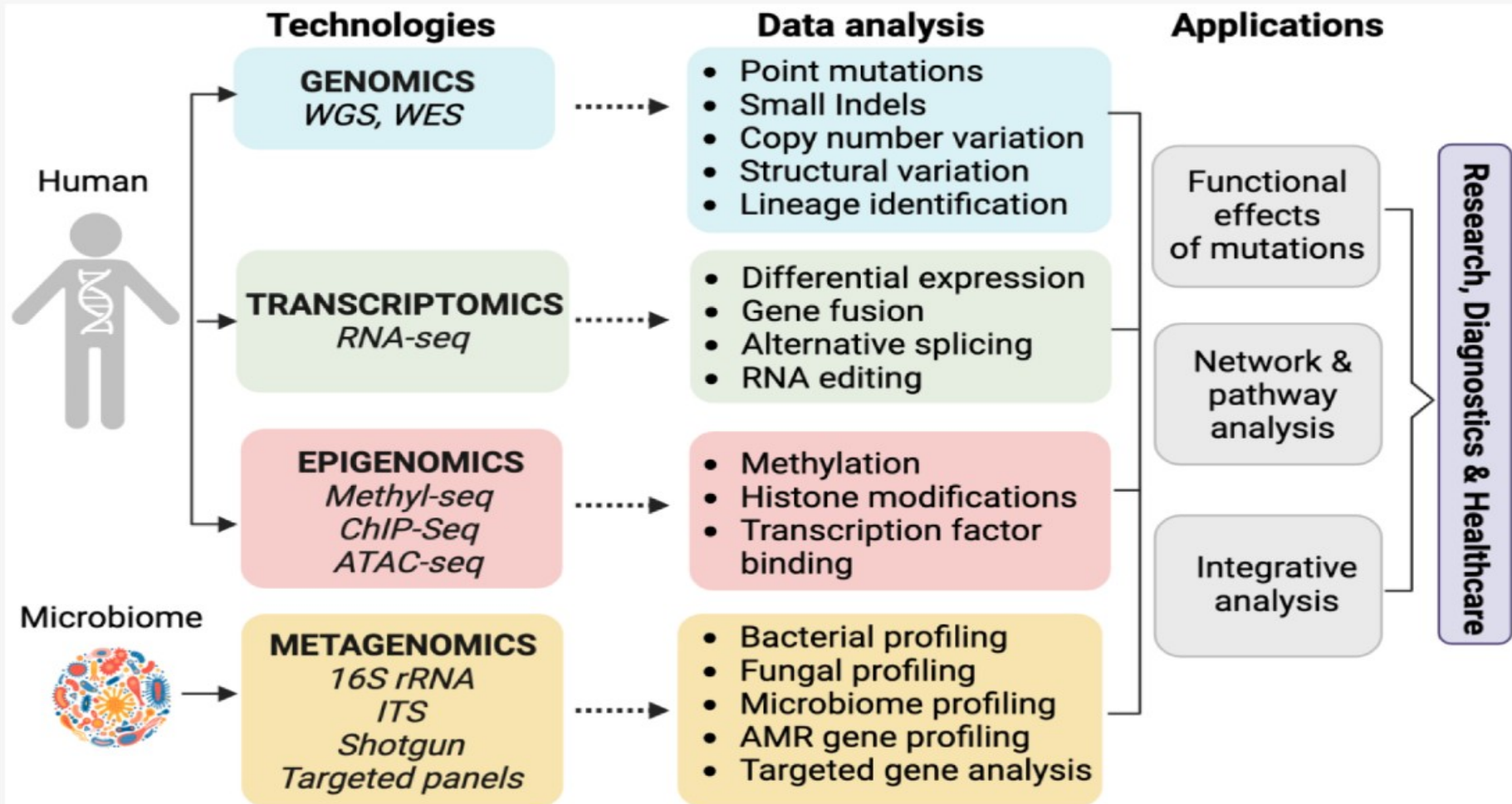
Check also 2020 NGS review at <https://www.nature.com/immersive/d42859-020-00099-0/pdf/d42859-020-00099-0.pdf>

<http://bioinformatics.ucdavis.edu>

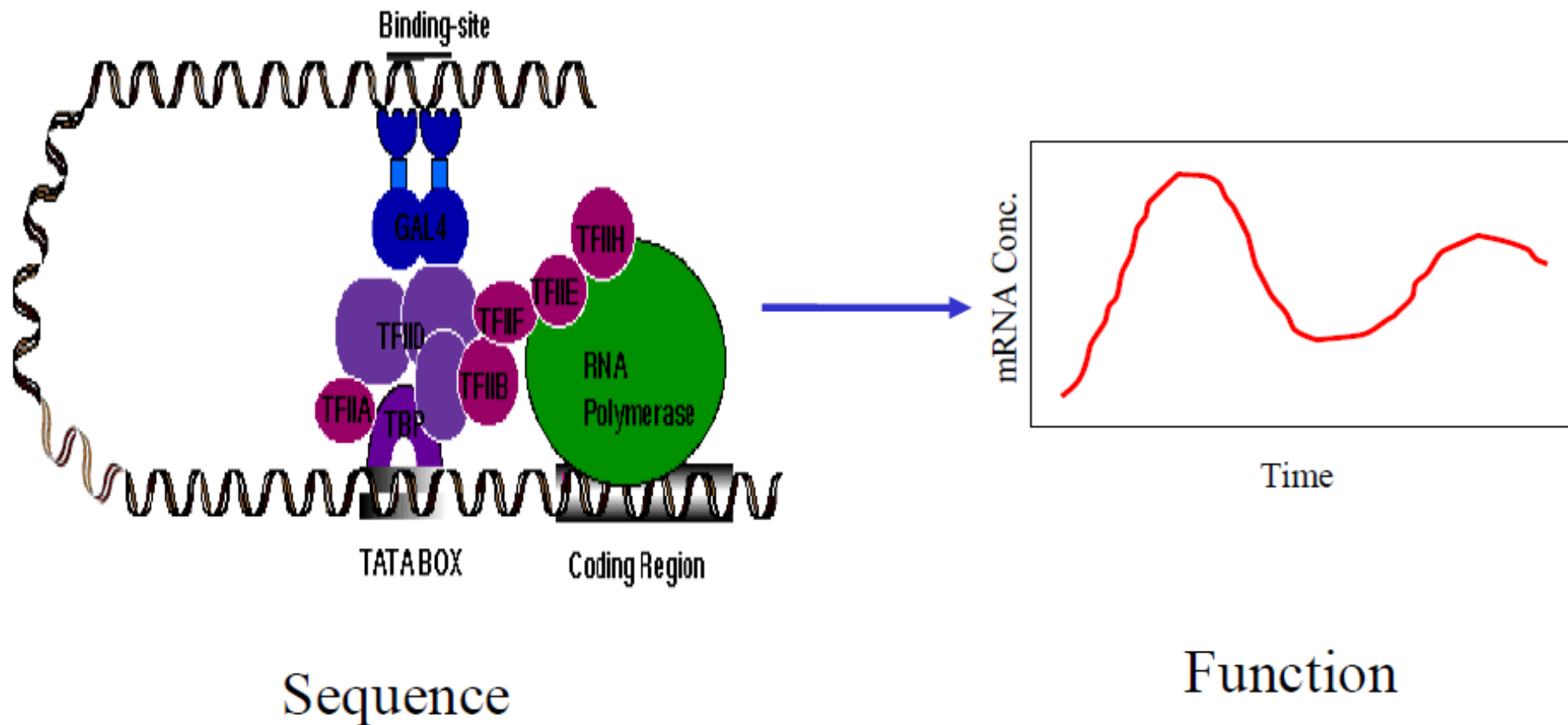
From: Next-Generation Sequencing Technology: Current Trends and Advancements

<https://www.mdpi.com/2079-7737/12/7/997>

Figure 3. Various approaches used for genome analysis and applications of NGS, including technological platforms, data analysis, and applications. WGS, whole-genome sequencing; WES, whole-exome sequencing; Seq, sequencing; ITS, internal transcribed spacer; ChIP, chromatin immunoprecipitation; ATAC, assay for transposase-accessible chromatin; AMR, anti-microbial resistance.

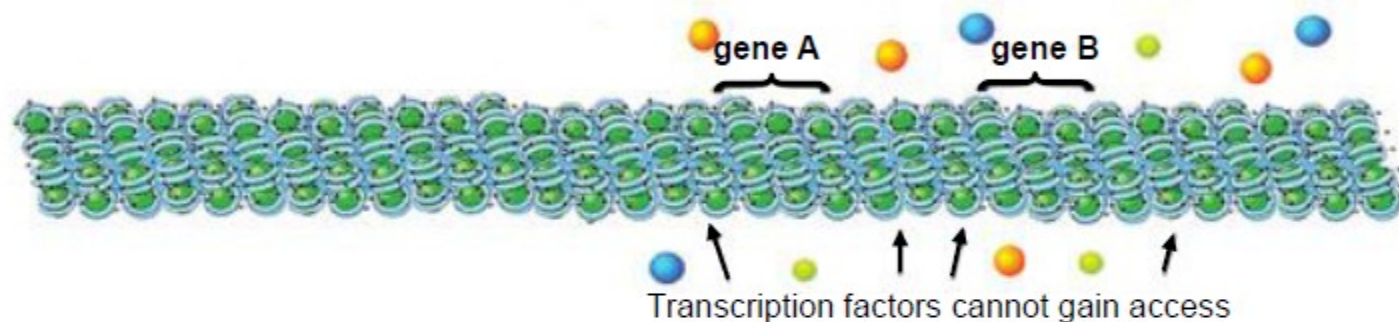


Gene Regulation

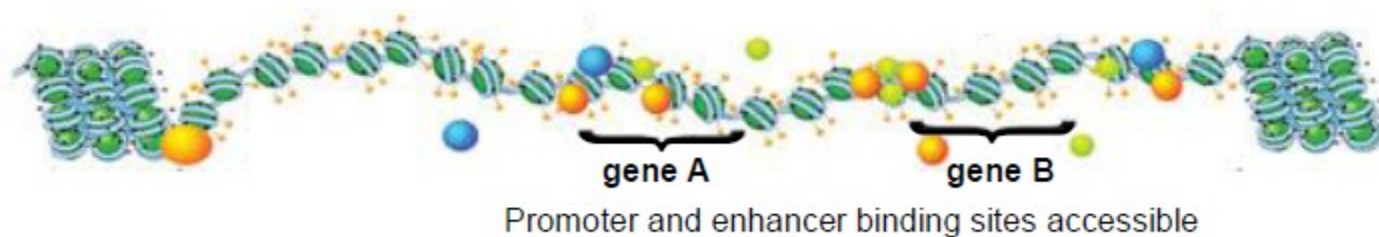


Chromatin Structure Determines Gene Status

Closed Chromatin



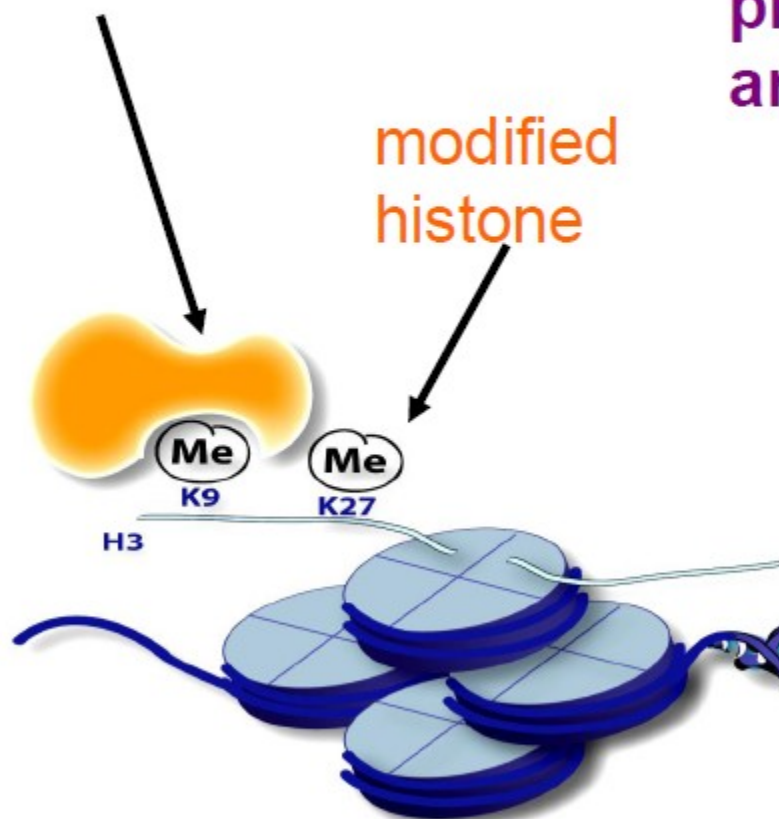
Open chromatin



DNA and Histone Modifications Create an Epigenome

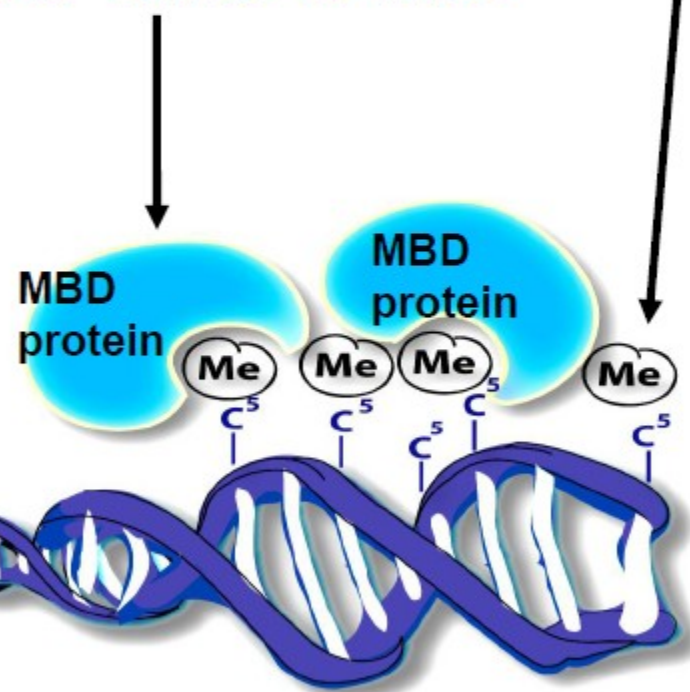
protein that modifies and/or binds to a modified histone

modified histone



protein that creates and/or binds to meC

meC



Methyl-CpG (5'-C-phosphate-G-3')-binding domain (MBD)

Coordinated Efforts to Decipher Epigenomes

- There is a wealth of publicly available data.
Don't be afraid to dig!
- NIH Roadmap Epigenomics Mapping Consortium
<http://www.roadmapepigenomics.org/>



- Encyclopedia of DNA Elements Consortium
(ENCODE)
- ENCODE data limited to cell-types at
<http://genome.ucsc.edu/index.html>

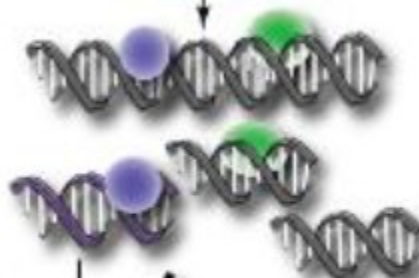


Key Steps in a ChIP Assay

Cross-linking



Fragmentation



Immunoprecipitation



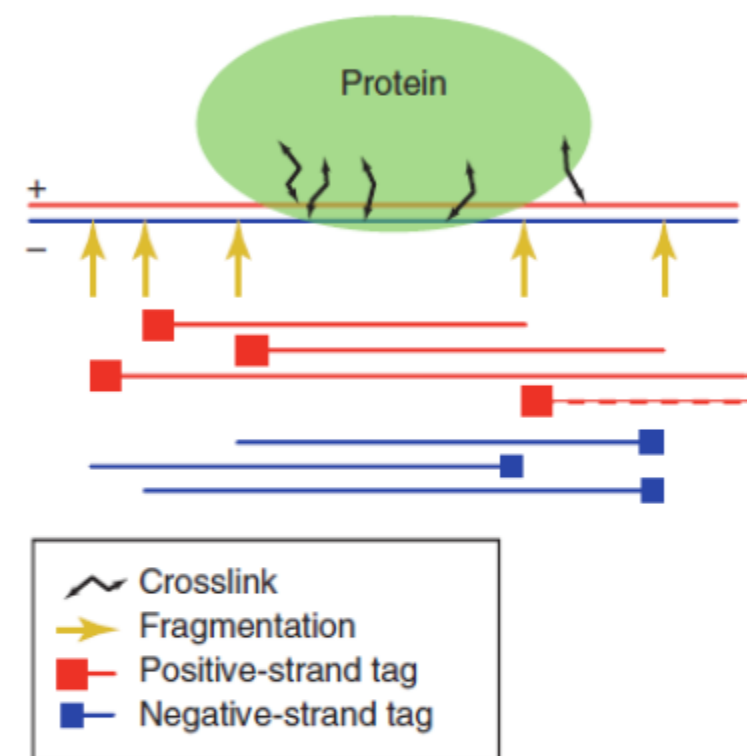
DNA purification



ChIP DNA

Input DNA

From Binding Site to Sequence to Peak

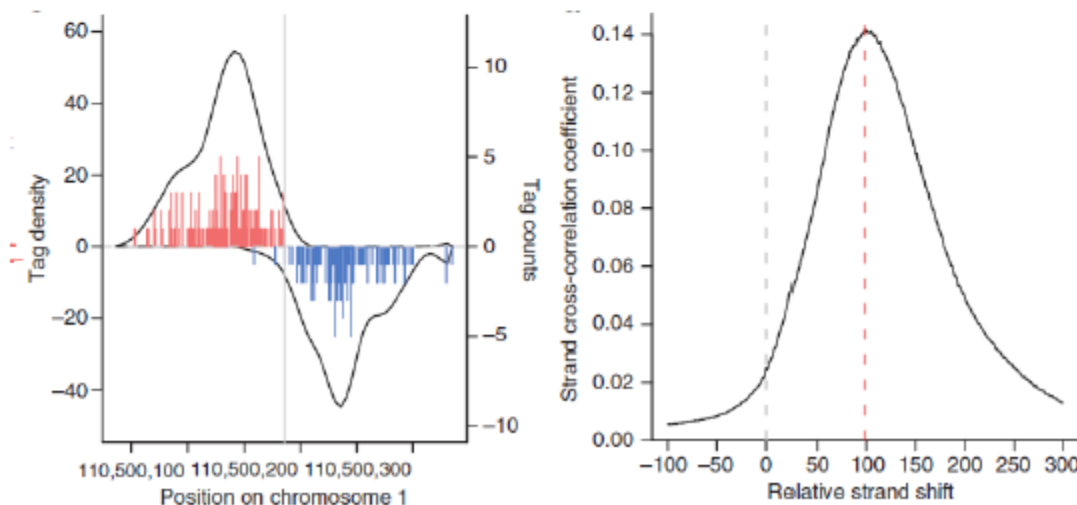


Kharchenko et al., 2008
Nat. Biotech. 26:1351

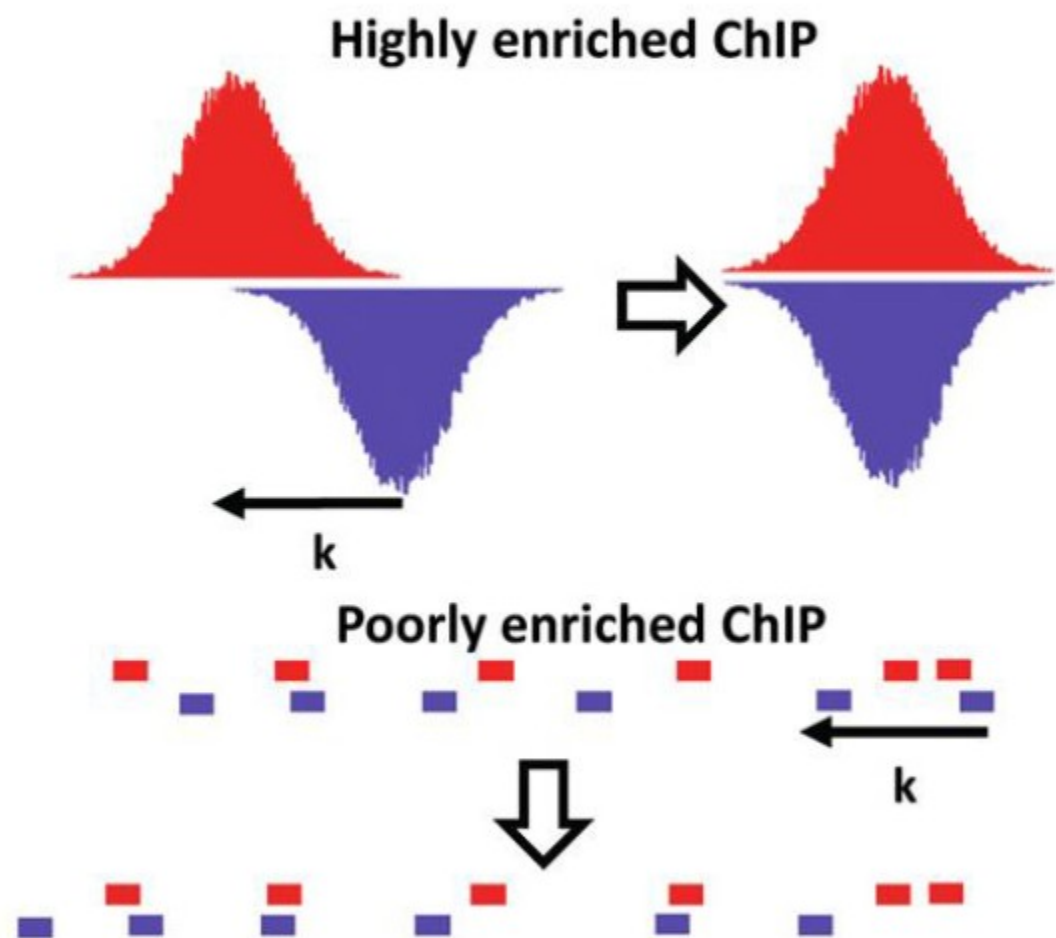
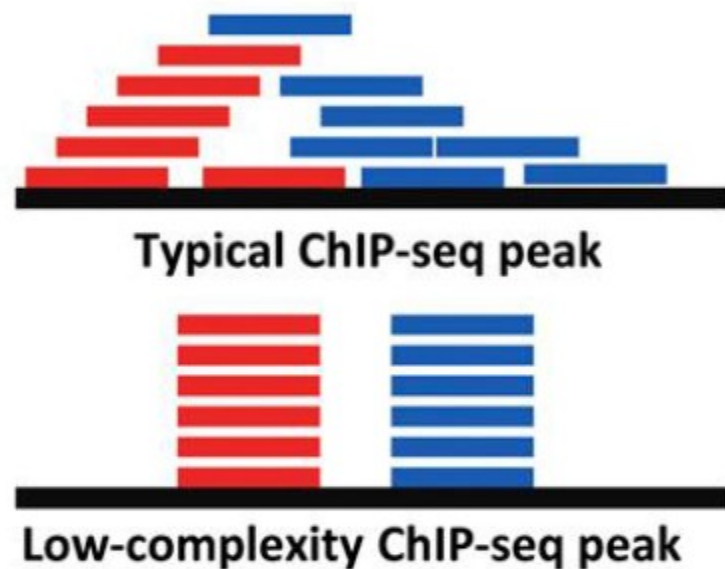
Short sequences are generated from each DNA molecule.

When mapped, a tag distribution is seen around a stable binding site.

Cross-correlation is calculated for the distance between positive- and negative-strand peaks

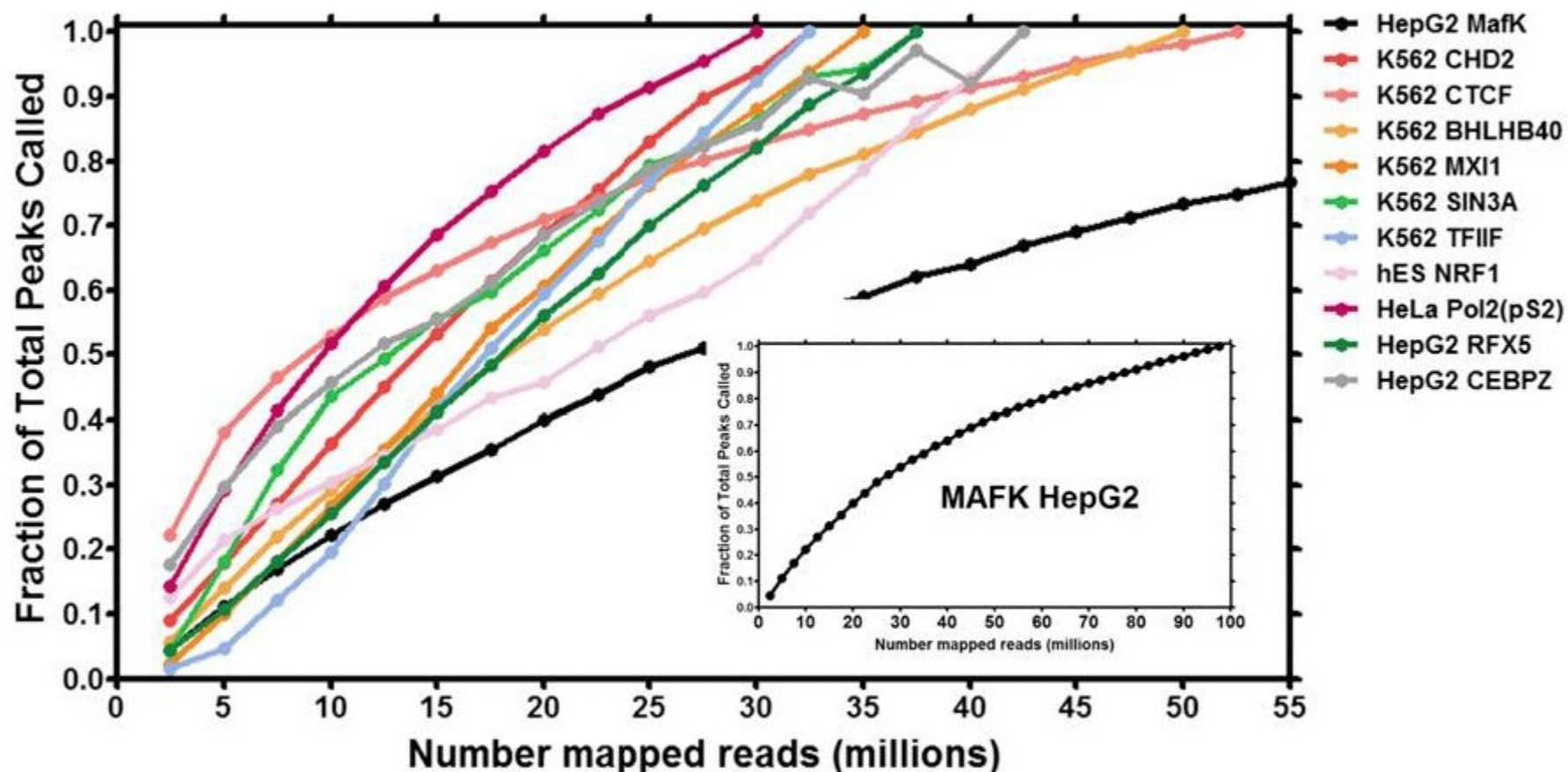


Library Complexity and Cross-Correlation



Landt, et al., 2012 Genome Res. 22:1813

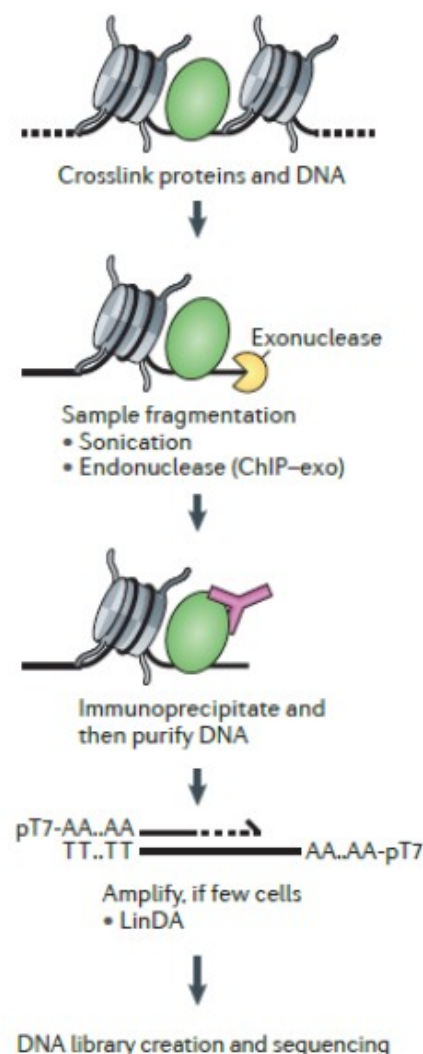
Called Peaks Increase With Sequencing Depth



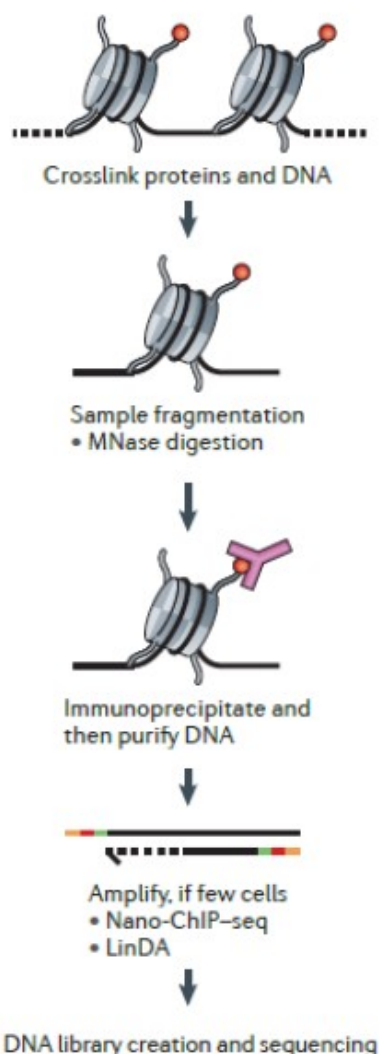
Landt, et al., 2012 Genome Res. 22:1813

Comparison of Experimental Protocols

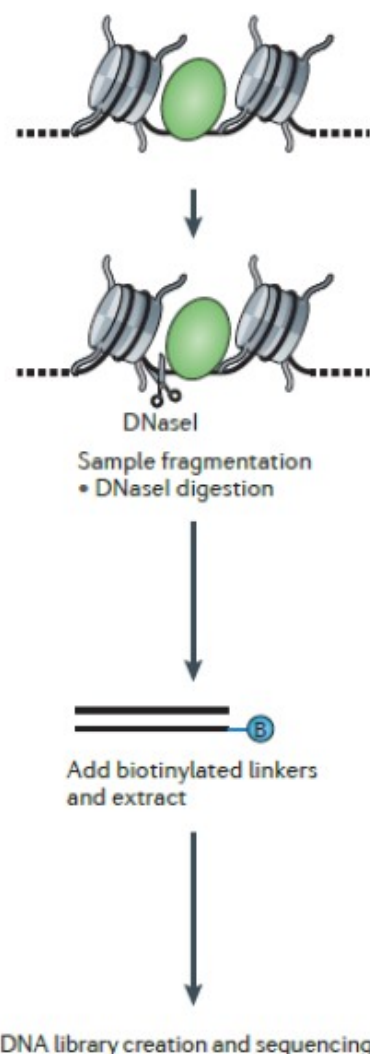
a DNA-binding protein ChIP-seq



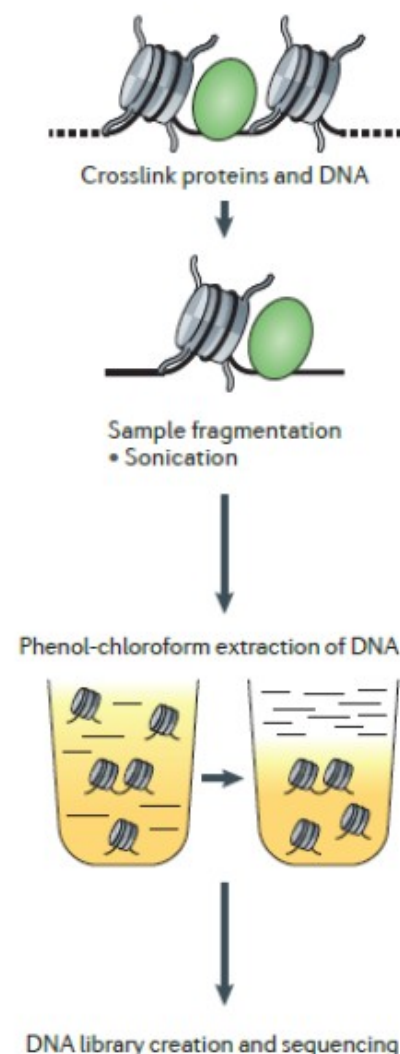
b Histone modification ChIP-seq



c DNase-seq



d FAIRE-seq

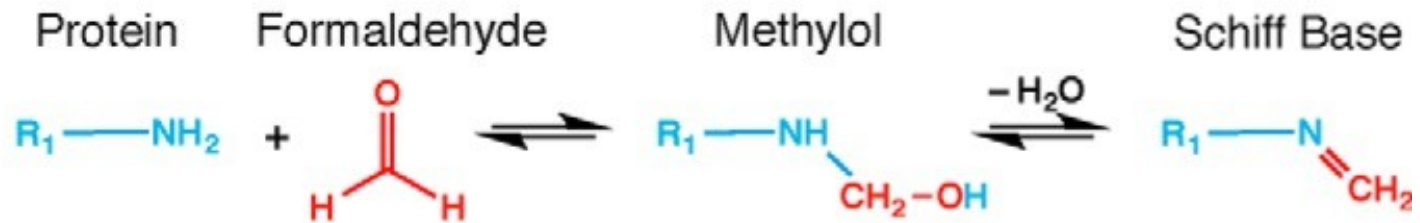


Furey, 2012 Nat. Rev. Gen. 13:840

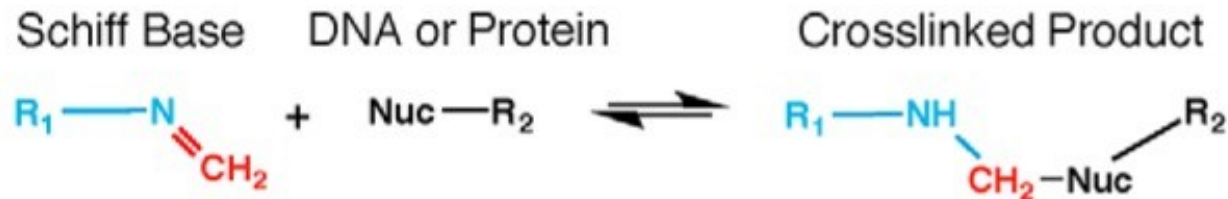
<http://bioinformatics.ucdavis.edu>

Chemical reactions during formaldehyde crosslinking of biomolecules

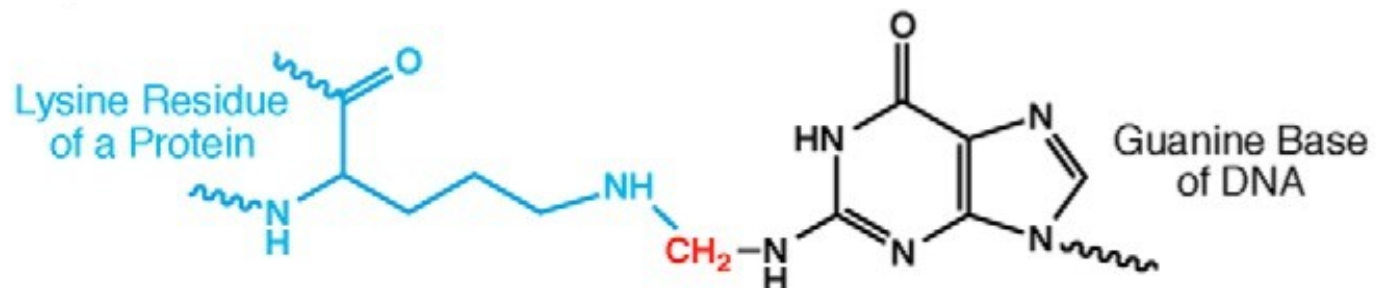
Step 1



Step 2

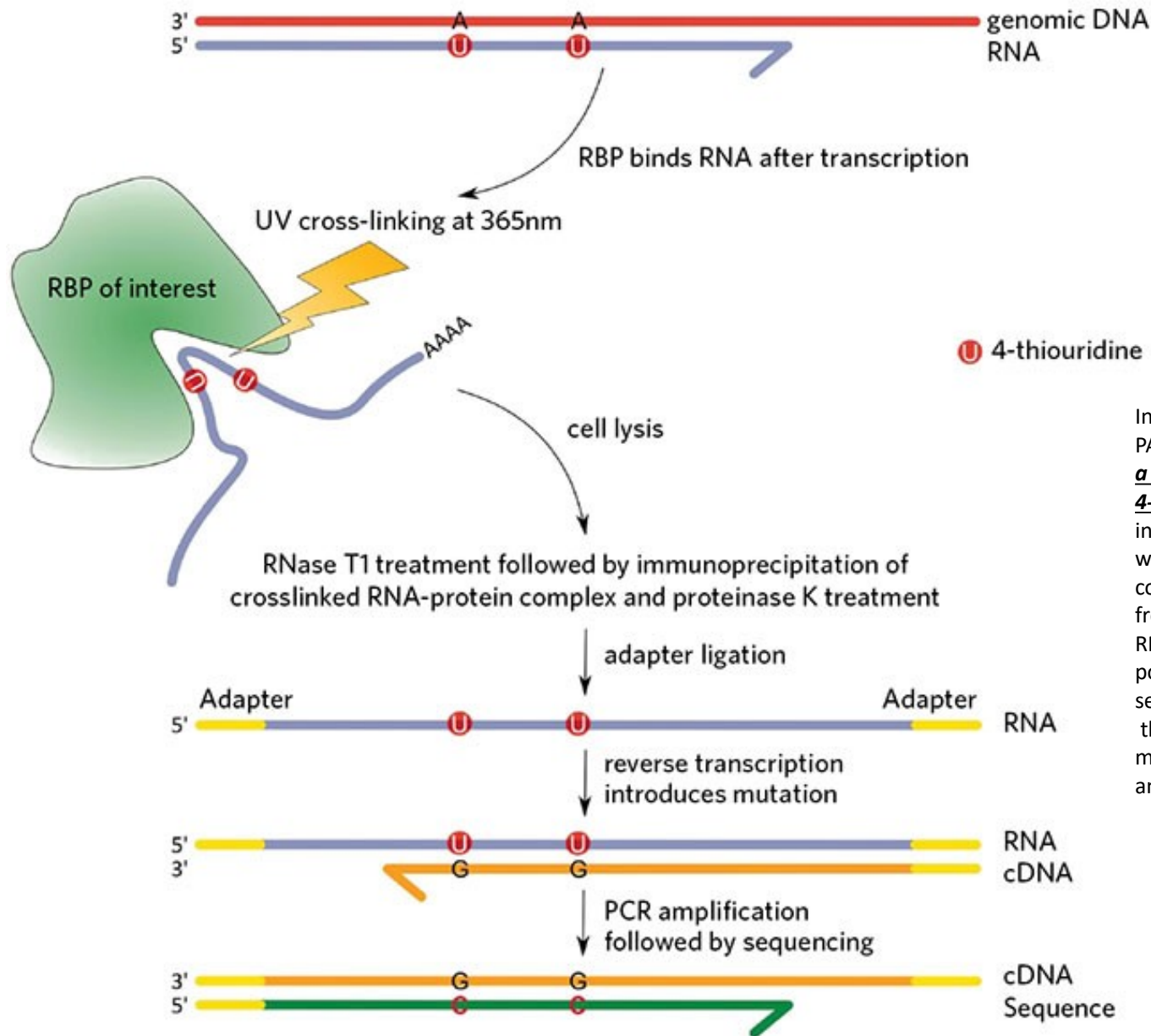


Example Protein-DNA Crosslink



ParClip

PhotoActivatable Ribonucleoside-enhanced CrossLinking ImmunoPrecipitation



In one variation on this technique, known as PAR-CLIP, cells are incubated with a light-reactive nucleoside analog, 4-thiouridine (U), that becomes incorporated into RNA. Irradiation with UV light crosslinks RNA-protein complexes, which are then isolated from cell lysates using antibodies. RNA located outside the protein binding pocket is degraded, and the remaining sequence is transcribed to DNA, a process that leads to a characteristic T to C mutation wherever the nucleoside analog incorporates.

ENCODE Guidelines For Controls and Replicates

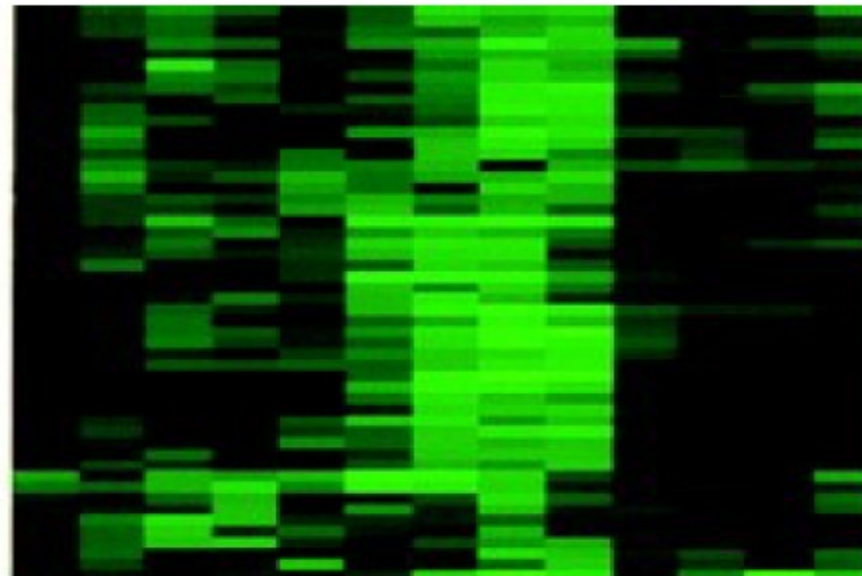
- Controls are Important!
 - Necessary to avoid non-uniform background (sonication, etc.)
 - Many cell lines have aneuploidy (genome size, copy number)
 - “Input” controls – similar prep, but no ChIP.
 - “IgG” – IP without the specific antibody
 - If amplification is done, must be done on all samples, including controls (and complexity needs to be evaluated after sequencing to ensure peaks aren't due to PCR artifacts)
- Replicates
 - Minimum of two biological replicates.
 - The number of mapped reads and identified targets should be within 2 fold between replicates
 - 80% of the top 40% of targets from one replicate should overlap the list of targets from the other replicate. OR
 - More than 75% of targets should be in common between each replicate

ENCODE Guidelines for Sequencing Depth

- The number of targets that can be identified varies substantially between cell types and experiments
- Depends on the TF, antibody, and peak-calling algorithm.
- Mammalian cells:
 - 10M uniquely mapped reads per replicate for point-source peaks (increased from previous requirement of 3M reads)
 - 20M uniquely mapped reads per replicate for broad-source peaks
- Other organisms need fewer reads (insects, yeast, etc.)
- Each replicate should be sequenced to similar depth. Controls to similar or greater depth.
- Complexity is important – low complexity libraries indicate PCR over-amplification, resulting in high false-positive rate (and failed experiment).
- FRiP (Fraction of Reads in Peaks) should be >1% of reads

Motif Finding Motivation

Clustering genes based on their expressions groups co-expressed genes



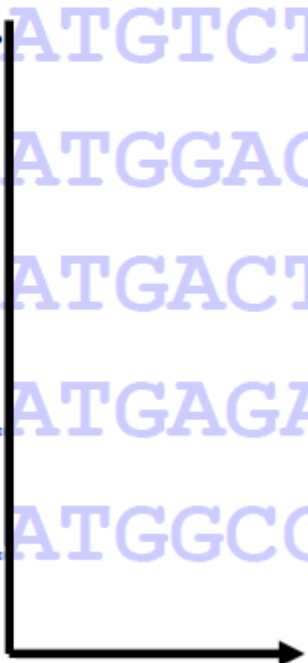
Assuming co-expressed genes are co-regulated, we look in their promoter regions to find conserved motifs, confirming that the same TF binds to them

Motifs vs Transcription Factor Binding Sites

- Motifs:
 - statistical or computational entities
 - predicted
- Transcription Factor Binding Sites (or more generally cis-regulatory elements)
 - biological entities
 - Real
- The hope is that TFBS are conserved, or otherwise significant computationally, so motifs can be used to find them

Finding Motifs in a Set of Sequences


GTGGCTGCACCCACGTGTATGC . . . ACGATGTCTC
ACATCGCATCACGTGACCAGT . . . GACATGGACG
CCTCGCACGTGGTGGTACAGT . . . AACATGACTA
CTCGTTAGGACCATCACGTGA . . . ACAATGAGAG
GCTAGCCCACGTGGATCTTGT . . . AGAATGGCCT



Can you see the motif?

Finding Motifs in a Set of Sequences

GGCTGCAC**CACGT**GTATGC . . . ACG**ATGTCTCGC**
ATCGCAT**CACGT**GACCAGT . . . GAC**ATGGACGGC**
TCG**CACGT**GGTGGTACAGT . . . AAC**ATGACTAAA**
CGTTAGGACCAT**CACGT**GA . . . ACA**ATGAGAGCG**
TAGCC**CACGT**GGATCTTGT . . . AGA**ATGGCCTAT**



Finding Motifs in a Set of Sequences



Motif Finding Problem


Given n sequences, find a motif (or subsequence) present in many

This is essentially multiple alignment. The difference is that multiple alignment is global

- longer overlaps
- constant site sizes and gaps
- NP-complete!

```
Escherichia coli      -----VKLTATTE--E--AKAGHINNADKYA---KGEFVGFPGDQVFP 125
Burkholderia cepacia -----VNGISTTH--R--AKAGHINADKIE---SGEELAEFGDQITP 225
Acetobacter xylinus  -----AKYLAEDN--A--AKAGHINAYAKKH---TQDHLLELDQDTE 241
Aquifex aeolicus     -----KKHYUTREKN--V--AKAGHINSAIKK---KQDYLLELDQDTE 252
Agrobacterium tumefaciens -----FRUTTERN--V--AKAGHINSLAH---TTEVTEVFDADHAP 371
Rhizobium radiobacter -----FRUTTERN--V--AKAGHINSLAH---TTEVTEVFDADHAP 371
Rhodospirillum rubrum -----VVRTTKNN--K--AKAGHINSAATKEL---KQELVVVFDADAVD 281
Nostoc punctiforme c583 -----ELKVMKRAE--AKGKKGALNQVTFLE---KQGTAVVFDADAVD 203
Arabidopsis thaliana 7120 c294 -----ELKVMKRAE--ATGKKGALNQVTFLE---KQGTAVVFDADAVD 203
Synechocystis 6803 p111377 -----ELKVMKRAE--ATGKKGALNQVTFLE---KQGTAVVFDADAVD 203
Arabidopsis thaliana 11357223 -----ELKVMKRAE--ATGKKGALNQVTFLE---KQGTAVVFDADAVD 203
Gossypium hirsutum 6446577 -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 529
Nostoc punctiforme c499 -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Nostoc punctiforme c526 -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Nostoc punctiforme c540 -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Synechococcus WH8102 -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Bacillus subtilis     -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Ferropasma acidimanus -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Thermoplasma acidophilum -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
Dictyostelium discoideum -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
ruler -----PELVVSEKKEKPGQKQKAKAGHNAVVRVAVLENAFPMNGLDQDQV 551
```

Definition and Representation

- Motifs: Short sequences
- IUPAC notation 
- Regular Expressions

- consensus motif

ACGGGTA

- degenerate motif

RCGGGTM

{G|A}CGGGT{A|C}

Single-Letter Codes for Nucleotides

Symbol	Meaning
G	G
A	A
T	T or U
C	C
U	U or T
R	G or A
Y	T, U or C
M	A or C
K	G, T or U
S	G or C
W	A, T or U
H	A, C, T or U
B	G, T, U or C
V	G, C or A
D	G, A, T or U
N	G, A, T, U or C

Position Specific Information

Seqs.

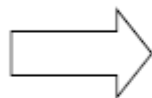
ACGGG

ATCGT

AAACC

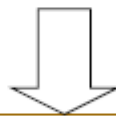
TTAGC

ATGCC



Alignment Matrix (Profile)

Pos	A	C	G	T
1	4	0	0	1
2	1	1	0	4
3	2	1	2	0
4	0	2	3	0
5	0	3	1	1



Position (Frequency) Weight Matrix

Pos	A	C	G	T	Conse
1	0.8	0	0	0.2	A
2	0.2	0.2	0	0.6	T
3	0.4	0.2	0.4	0	A G
4	0	0.4	0.6	0	G
5	0	0.6	0.2	0.2	C

Gibbs sampling

Find location AND description of commonly occurring substrings

“co-regulated genes”:

APPLEPEACHBANANAPPEARLEMONORANGEMELONKIWIGRAPELEMON
GAUDAEDAMLEERDAMPANAMATILSITBRIECHAMANBERTROQEFORT
OPELBWMTTOYOTAHYUNDAIMAZDAFIATRENAULTBAMANAFERRARI

Start with random positions for substrings

Gibbs sampling

Find location AND description of commonly occurring substrings

Step 1a:

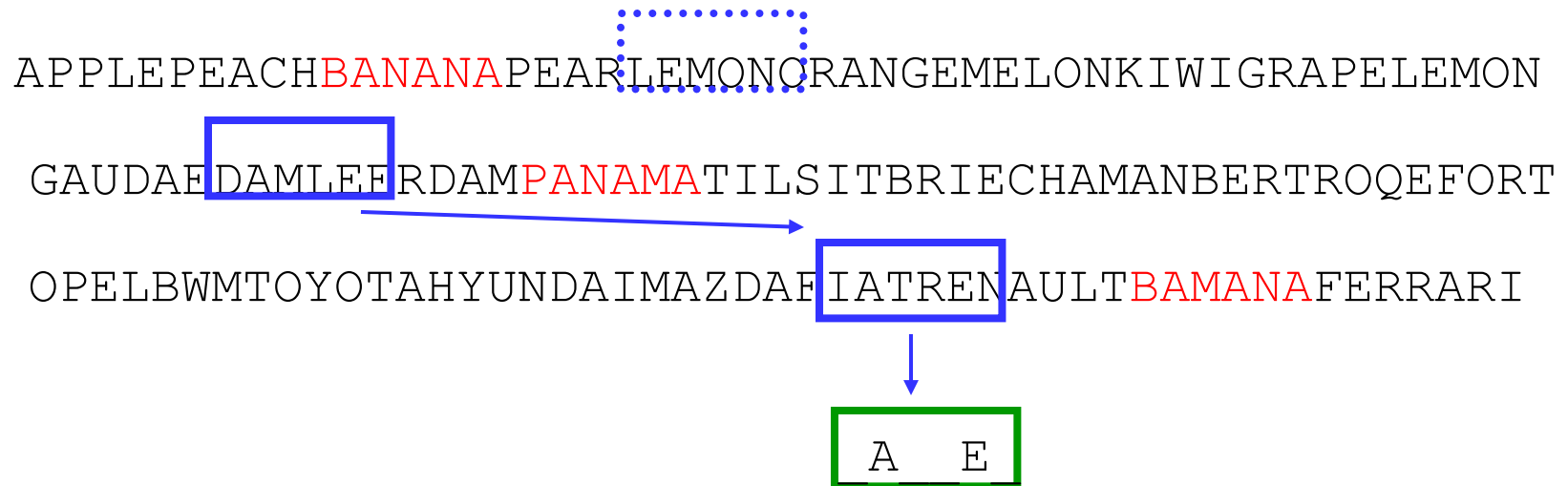
APPLEPEACHBANANAPPEARLEMONORANGEMELONKIWIGRAPELEMON
GAUDAEDAMLEFRDAMPANAMATILSITBRIECHAMANBERTROQEFORT
OPELBWMTYOYOTAHYUNDAIMAZDAFIATRENAULTBAMANAFERRARI

Pick one sequence

Gibbs sampling

Find location AND description of commonly occurring substrings

Step 1b:



Get statistics of all other substrings

Gibbs sampling

Find location AND description of commonly occurring substrings

Step 2a:

APPLEPEACHBANANAPPEARLEMONORANGEMELONKIWIGRAPELEMON
GAUDAE DAMLE ERDAMPANAMATILSITBRIECHAMANBERTROQEFORT
OPELBWMTYOYOTAHYUNDAIMAZDAFIATRENAULTBAMANAFERRARI

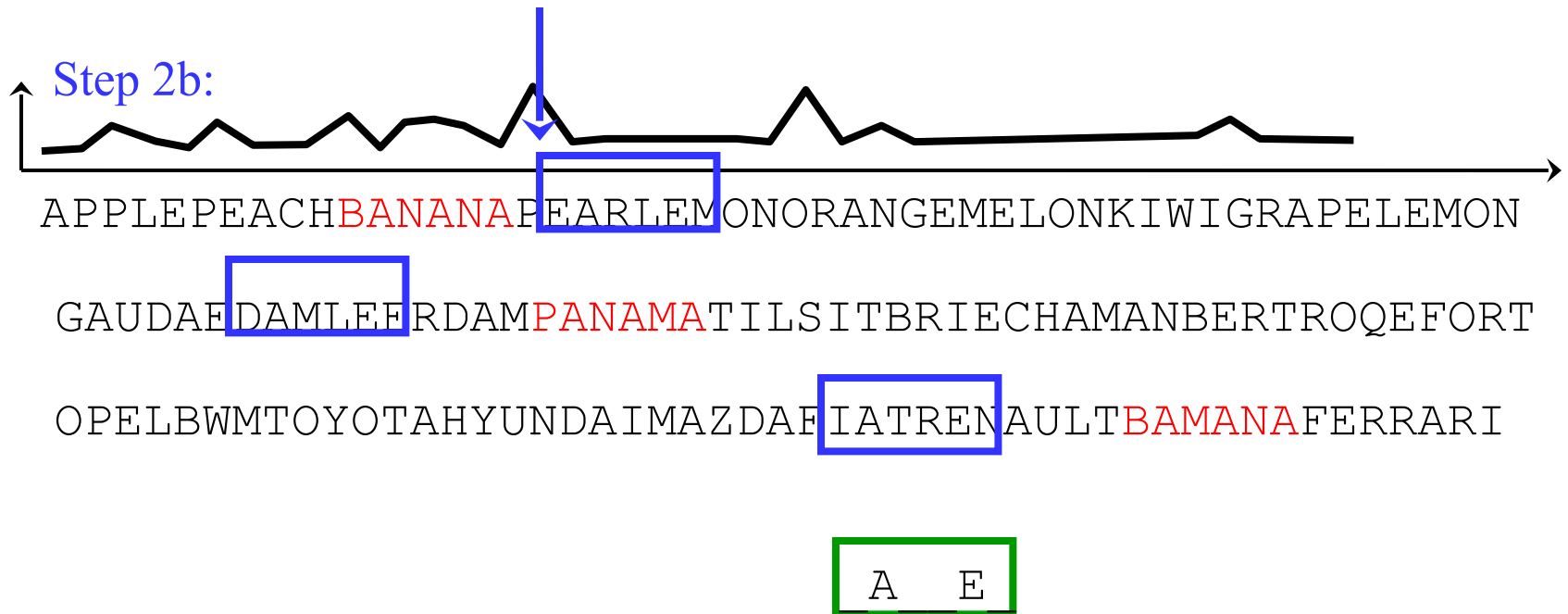
A E

A E

match description to all locations in sequence

Gibbs sampling

Find location AND description of commonly occurring substrings



Pick new location in sequence (probabilistic)

Gibbs sampling

Find location AND description of commonly occurring substrings

Repeat steps 1 and 2 until convergence:

APPLEPEACH **BANANA** PEARLEMONORANGEMELONKIWIGRAPELEMON
GAUDAEDAMLEERDAM **PANAMA** TILSITBRIECHAMANBERTROQEFORT
OPELBWMTYOYOTAHYUNDAIMAZDAFIATRENAULT **BAMANA** FERRARI

$\text{b } A_n A_n A$

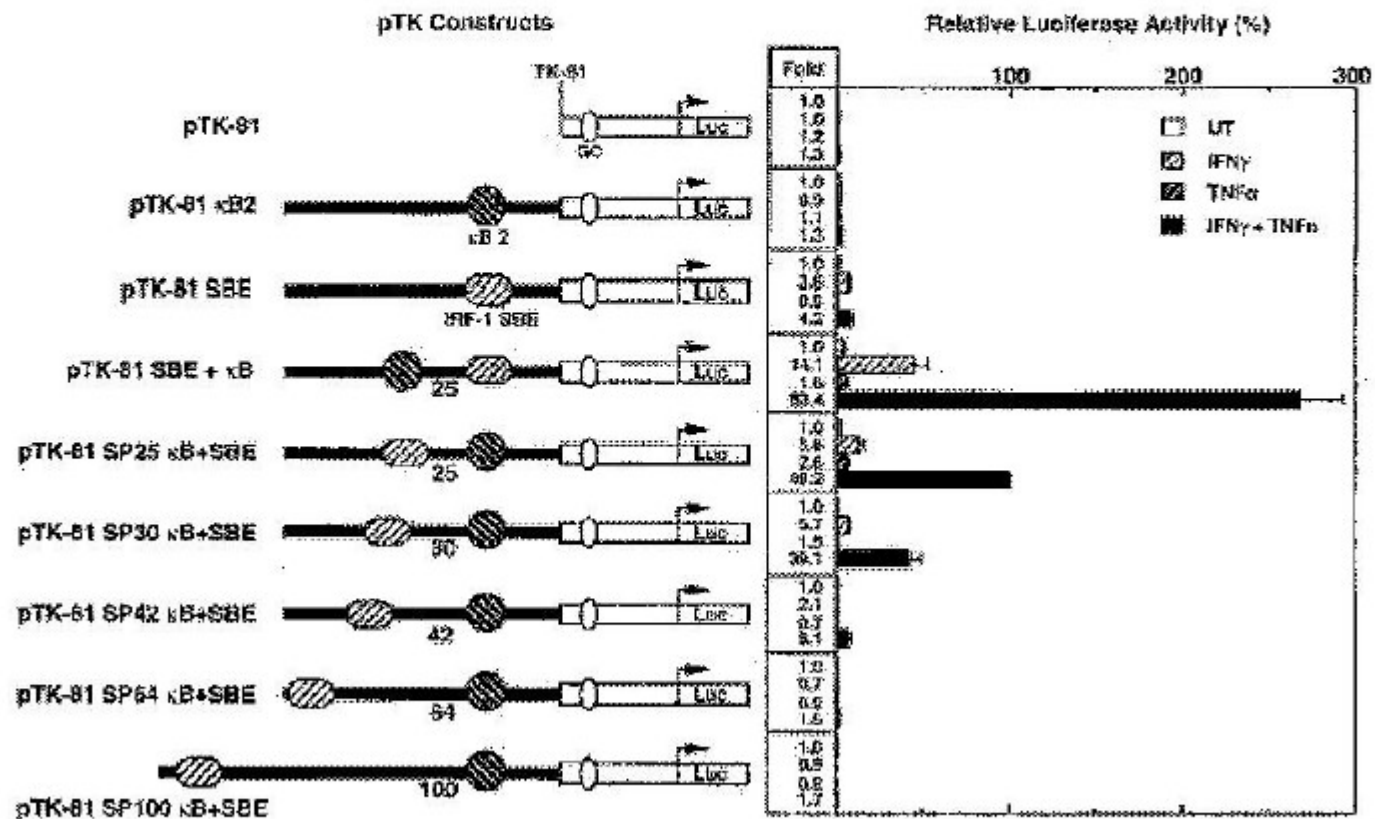
Multi-site Motif

- Two-site: Dimer, dyad
- Gapped Motif
- In general, a motif is an ordered set of binding sites

Table 3 • Dimer alignment
for MCM1 binding site

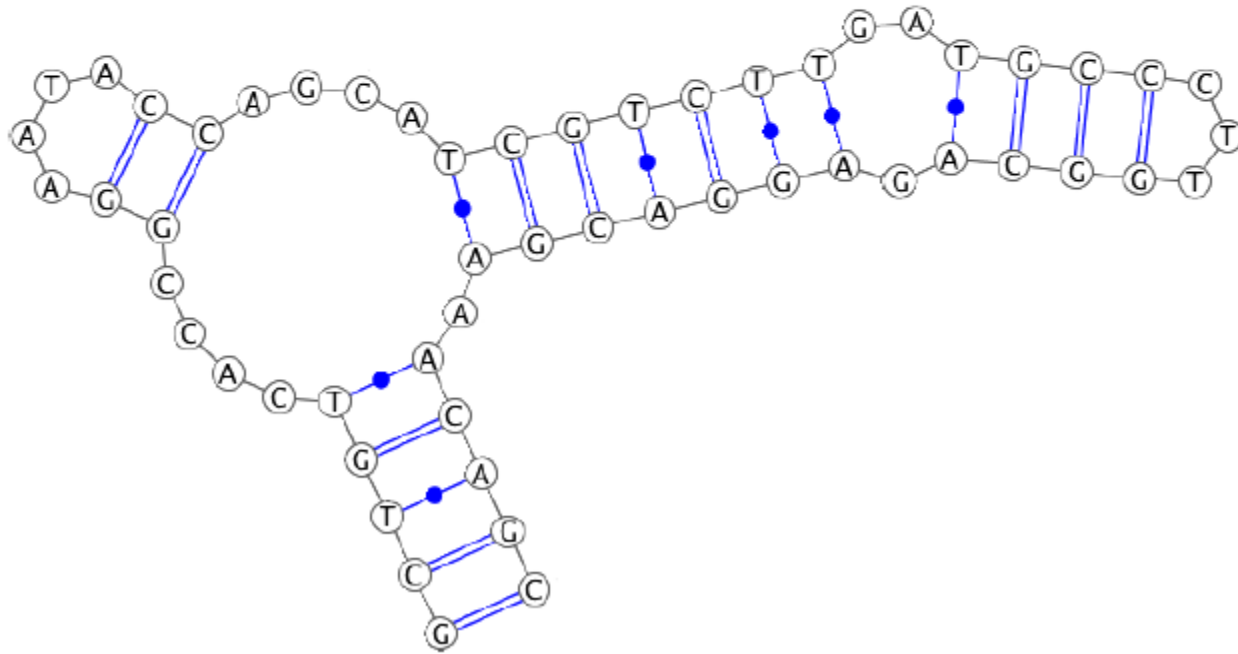
.ACC.....AGGA.
.ACC.....GGAA
..CCTA...AGGA.
.ACCT...AAGG..
..CCT.....GGAA
..CCTA...GGAA
TACC....AAGG..
.ACCT.....GGA.
.ACCT....AGGA.
TACC.....GGA.
TACC....AGGA.
.ACCT.....GGAA
TACC.....GGAA

Dependence of Simple Motif Pairs on Distance and Order Between Them



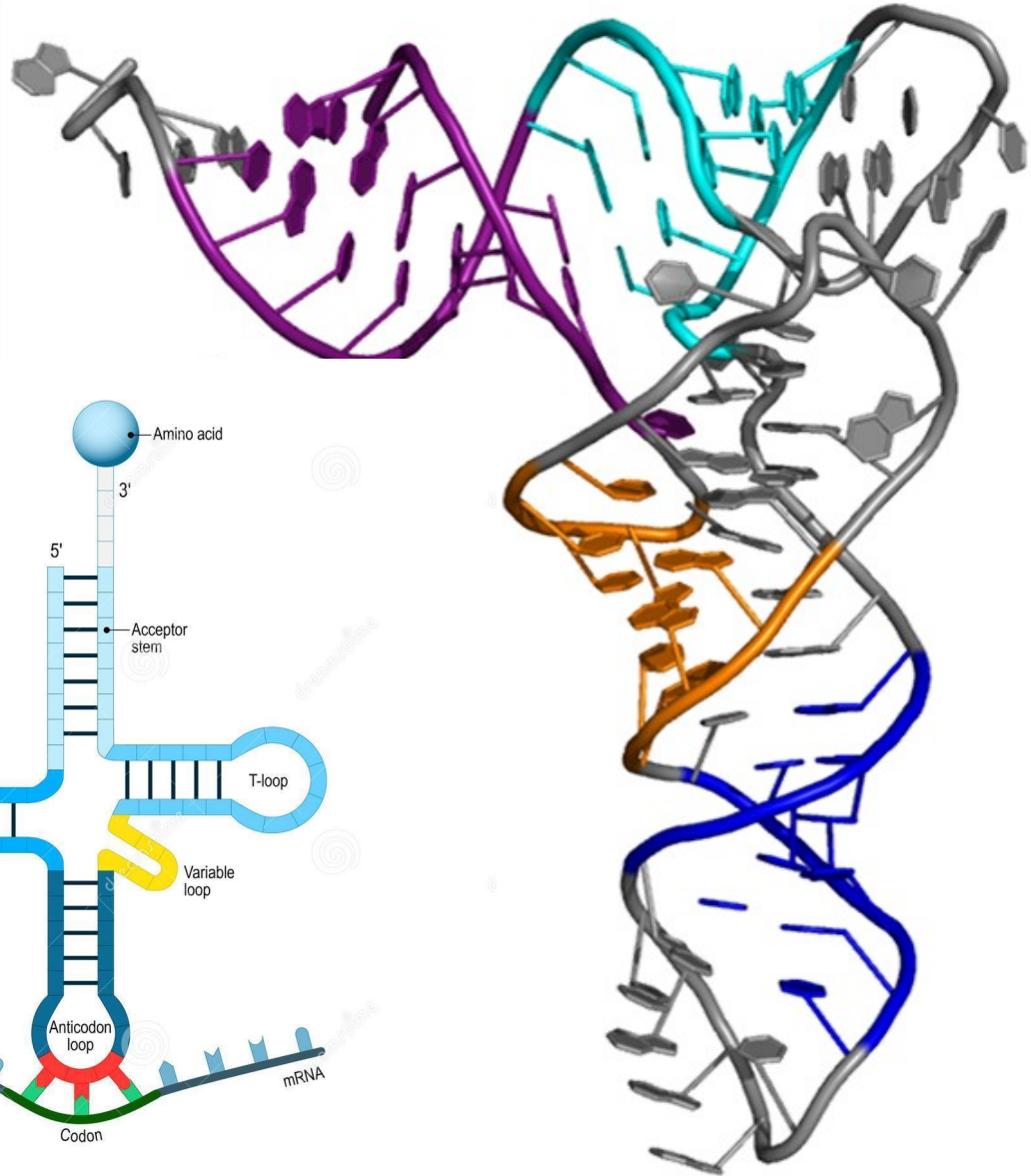
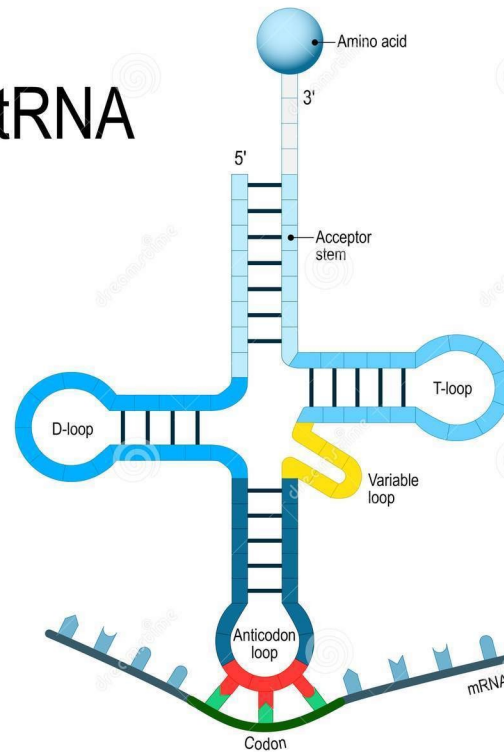
RNA SECONDARY STRUCTURE

Sequence → **Secondary Structure** → Tertiary Structure



Transfer RNA (tRNA)

tRNA



Download from
Dreamstime.com

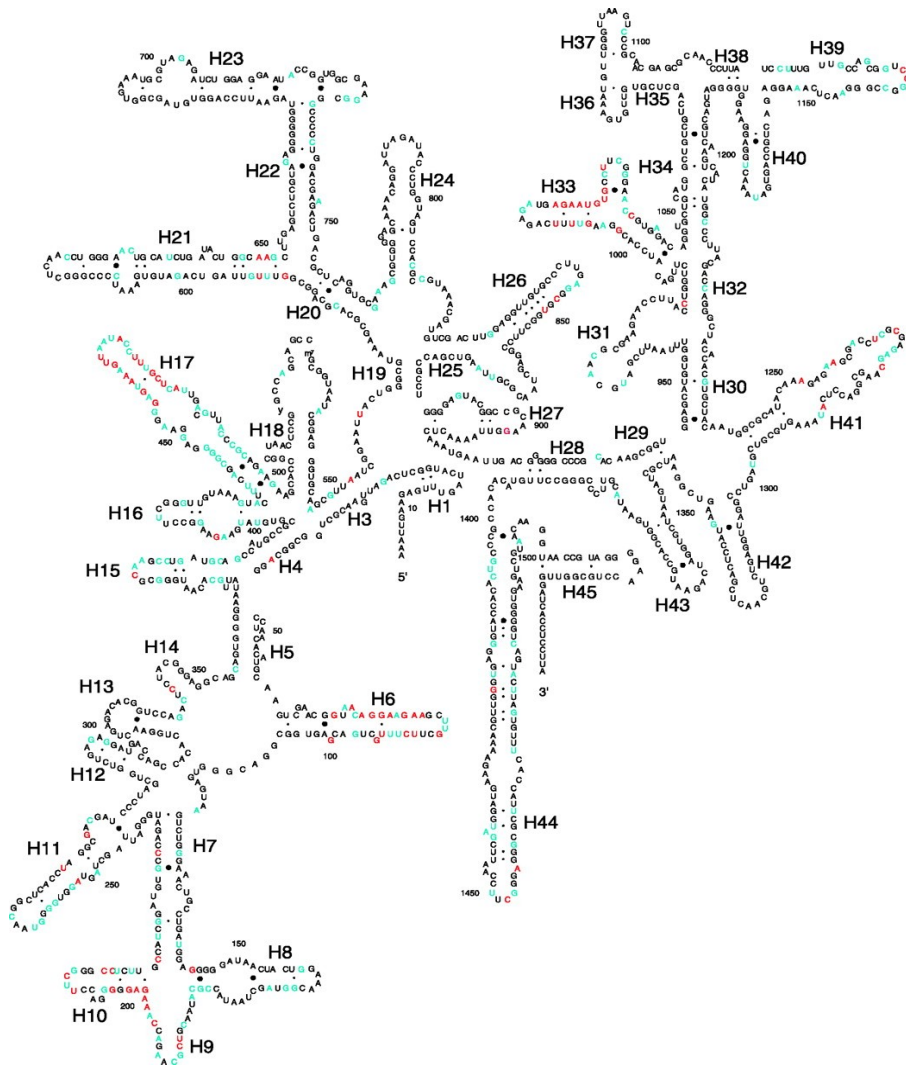
This watermarked image is for previewing purposes only.



113257763

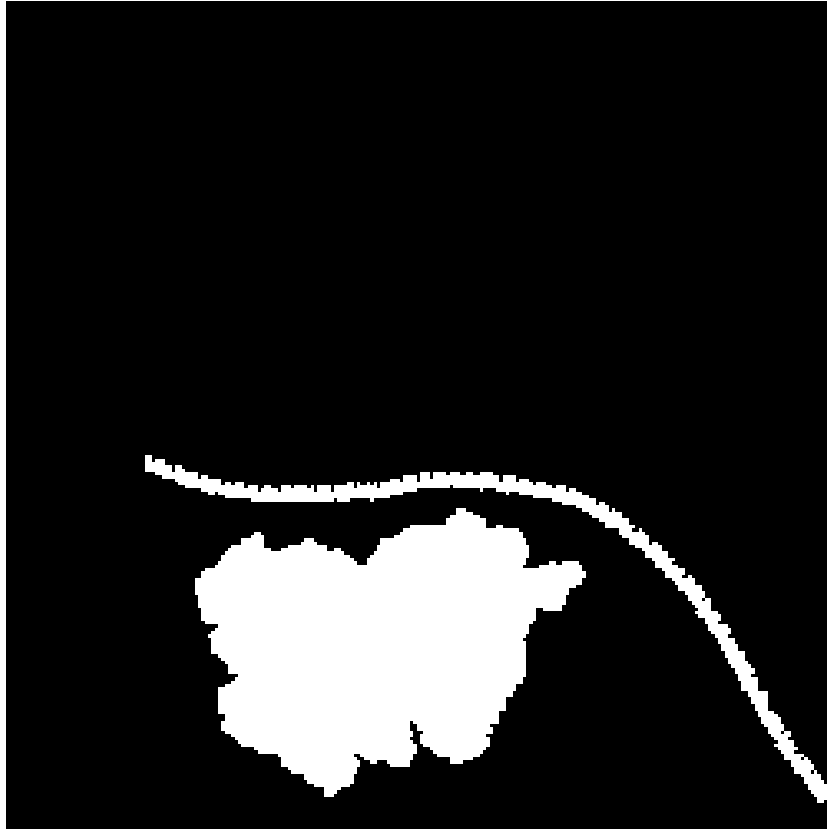
Designua | Dreamstime.com

Ribosomal RNA (rRNA)



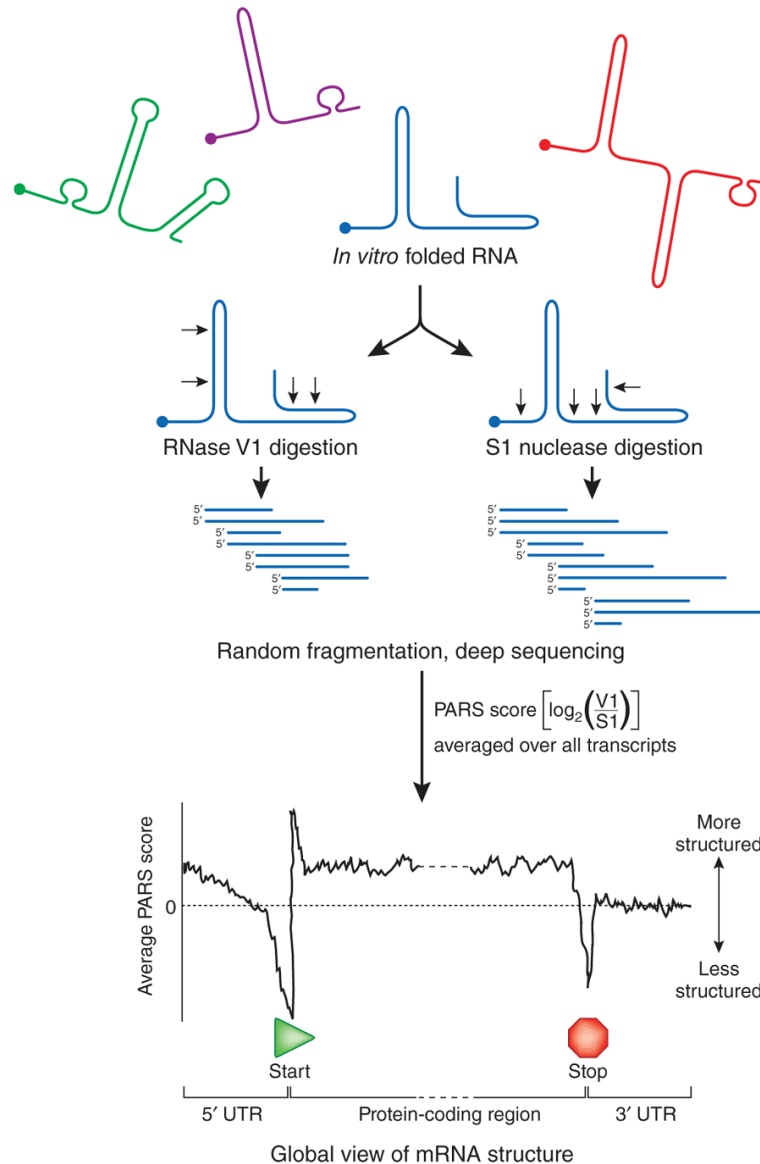
16S-rRNA (orange), proteins (blue)

rRNA+tRNA in Ribosome



By Bensaccount at en.wikipedia, CC BY 3.0, <https://commons.wikimedia.org/w/index.php?curid=8287100>

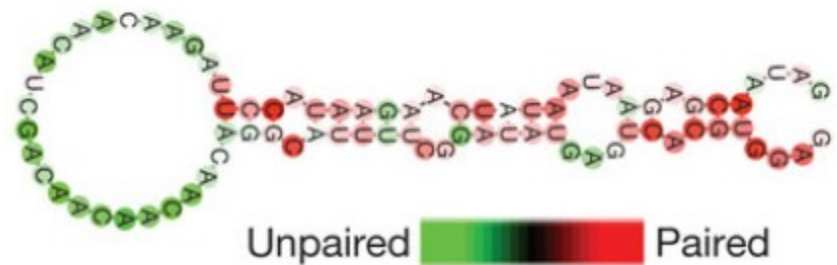
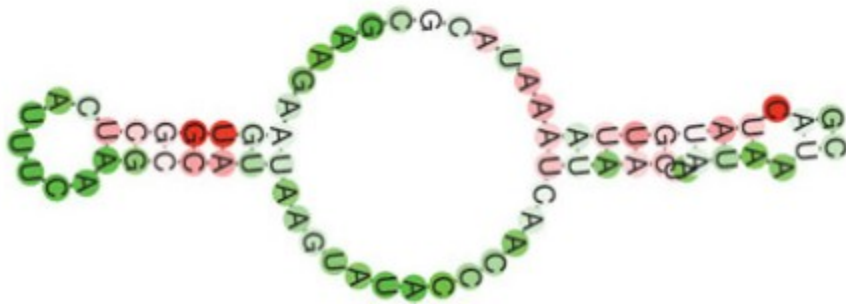
Parallel Analysis of RNA Structure (PARS)



PARS SCORE

Less Structure = more unpaired = score < 0

More structure = more paired = score > 0



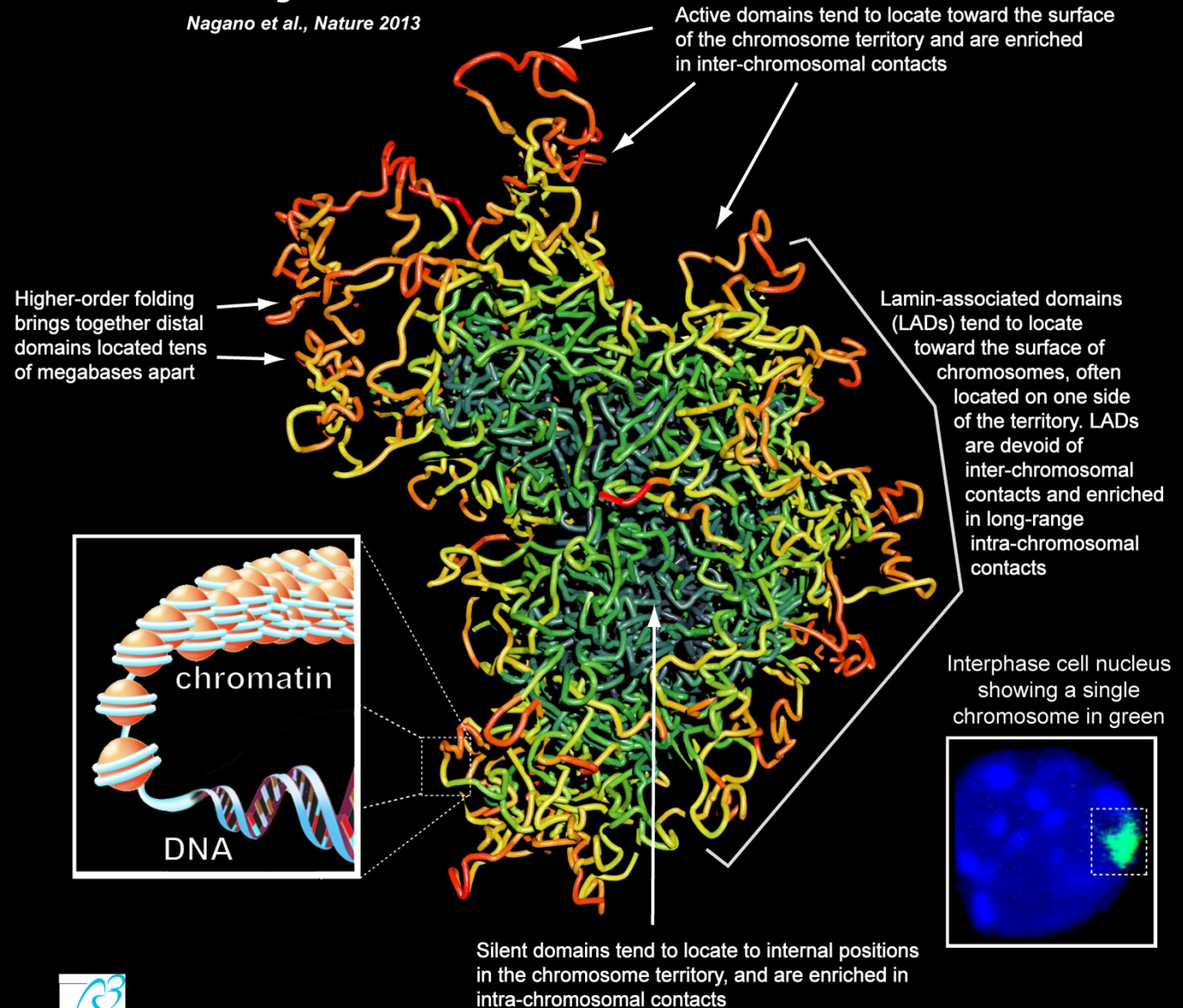
NGS Methods for RNA Secondary Structure

Method	Probing Type	Year	Key Feature
PARS	Enzymatic (RNases V1/S1)	2010	First method for genome-wide structural mapping (Structurome).
SHAPE-seq	Chemical (Flexibility)	2011	Measures nucleotide flexibility using an electrophile that acylates the $2' - OH$ of unstructured bases.
DMS-seq / Structure-seq	Chemical (Accessibility)	2014	Uses Dimethyl Sulfate (DMS) to probe accessible A and C bases <i>in vivo</i> .
icSHAPE	Chemical (In vivo)	2015	An improvement on SHAPE that uses a different chemical probe to provide higher signal-to-noise ratio <i>in vivo</i> .
DMS-MaP-seq	Chemical (Mapping)	2017	Combines DMS probing with techniques that map reverse transcriptase stops/misincorporations for high accuracy.
RING-seq	Enzymatic (Double-strand)	2019	Specialized method using a recombinant RNase to selectively cleave dsRNA regions, focusing on identifying regulatory structures.
Hi-LIP	Chemical (In vivo)	2021	Uses a cleavable chemical probe and specialized ligation to map RNA structures inside cells with high fidelity.

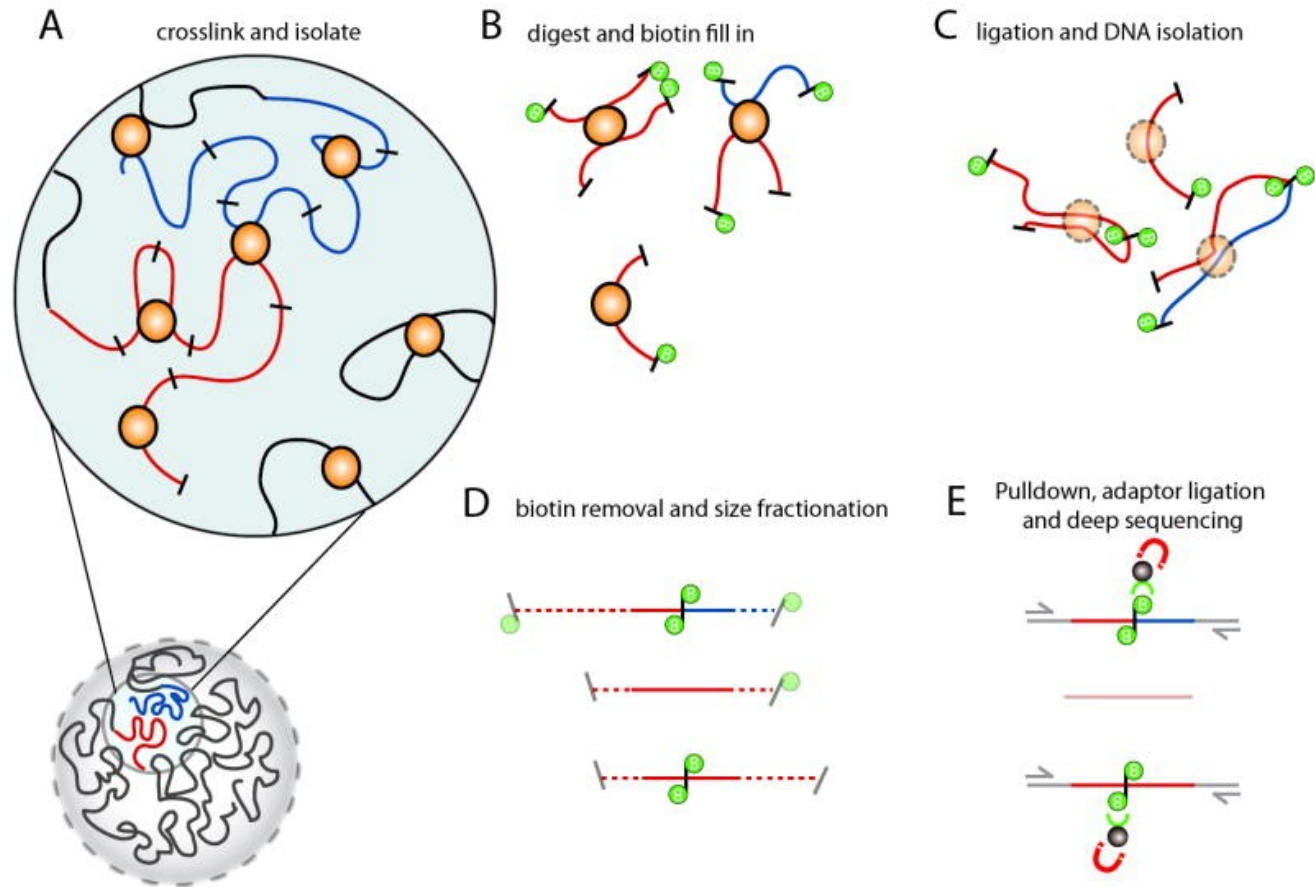
3D arrangement of Chromosomes

Chromosome Structure from single-cell Hi-C

Nagano et al., Nature 2013

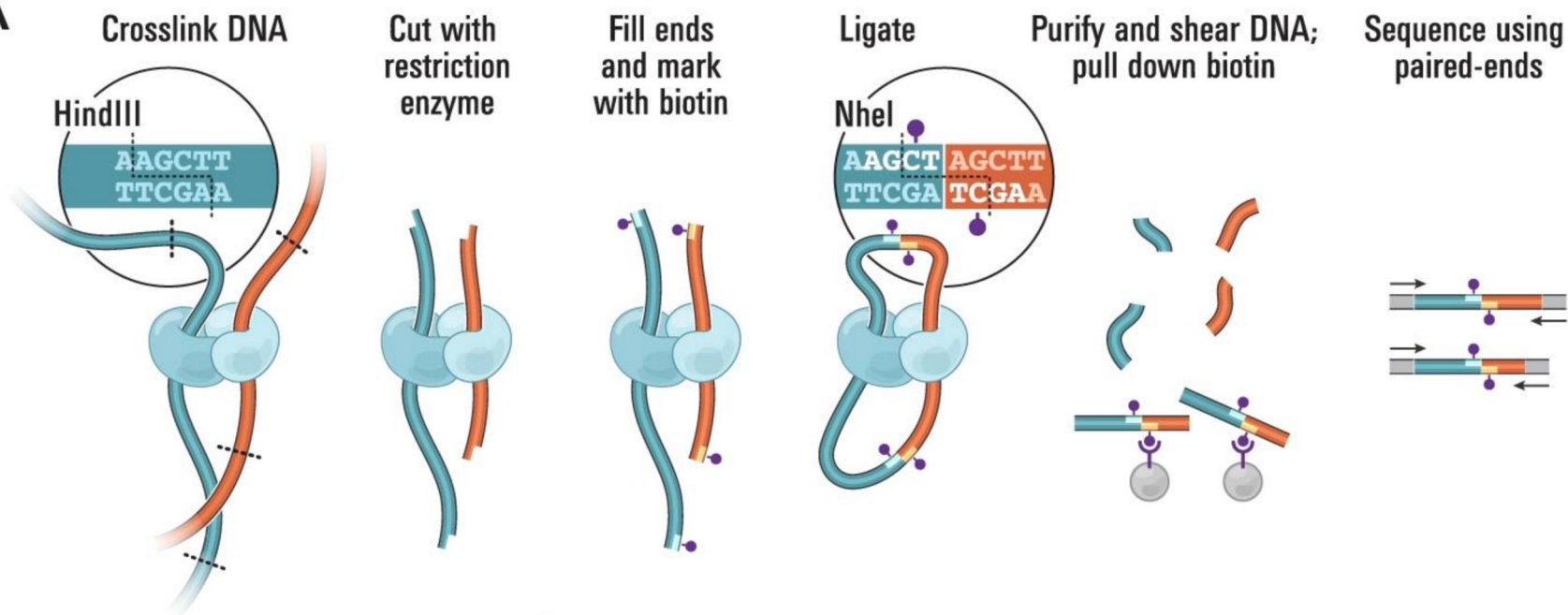


Overview of Hi-C technology

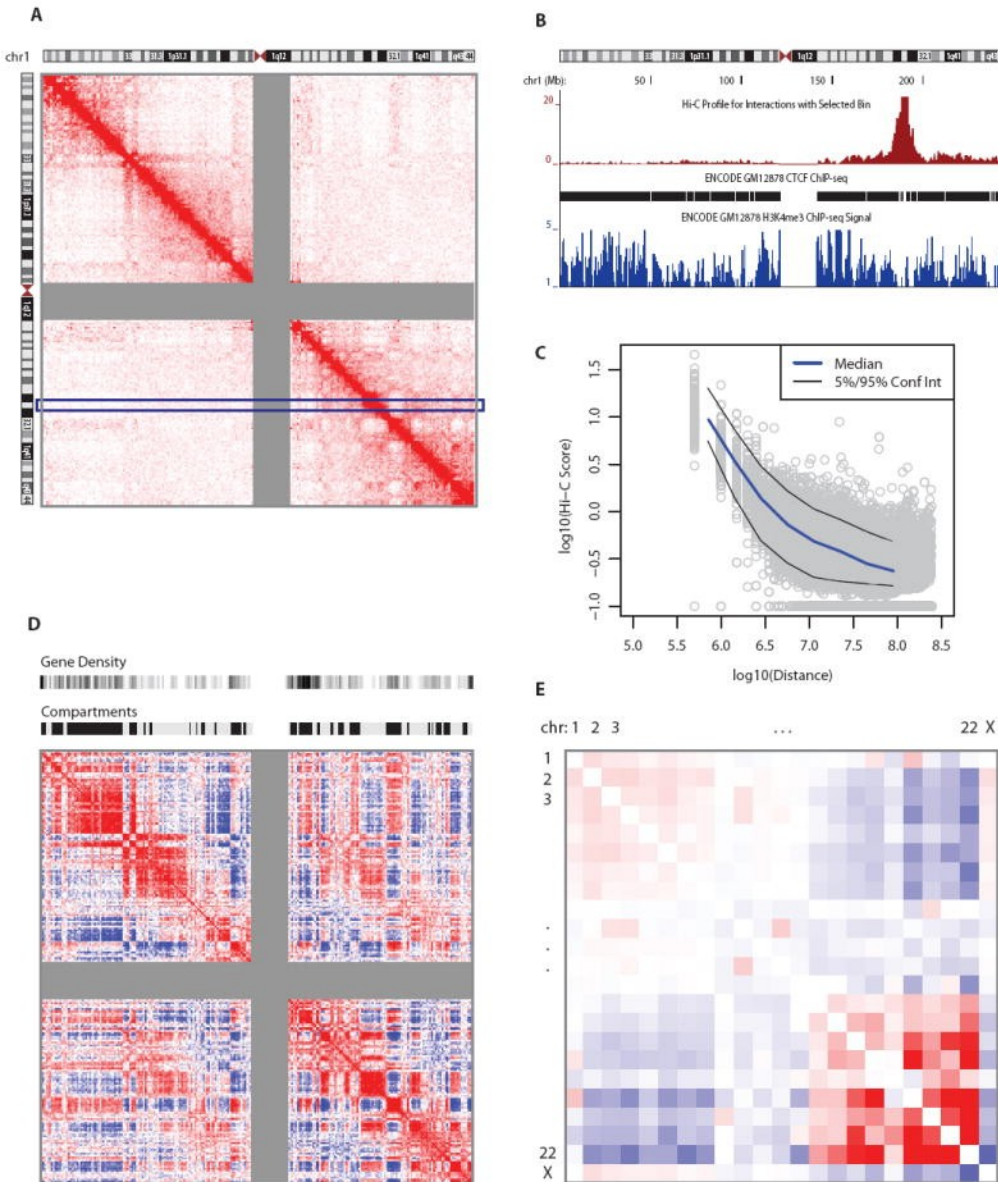


A) Hi-C detects chromatin interaction both within and between chromosomes by covalently crosslinking protein/DNA complexes with formaldehyde. **B)** The chromatin is digested with a restriction enzyme and the ends are marked with a biotinylated nucleotide. **C)** The DNA in the crosslinked complexes are ligated to form chimeric DNA molecules. **D)** Biotin is removed from the ends of linear fragments and the molecules are fragmented to reduce their overall size. **E)** Molecules with internal biotin incorporation are pulled down with streptavidin coated magnetic beads and modified for deep sequencing. Quantitation of chromatin interactions is achieved through massively parallel deep sequencing.

A



Hi-C data visualization and analysis



A) A heatmap of interactions between all 1 Mb bins along chr1 for GM06990 cells. The intensity of red color corresponds to the number of Hi-C interactions. **B)** A “4C profile” derived from one row of the Hi-C heatmap (blue box in A) showing all interactions between a fixed 1 Mb location at 190 Mb on chr1 and the rest of chr1. CTCF and H3K4me3 tracks from a similar cell line are displayed below as examples of other genomic datasets that can be compared with such an interaction profile. **C)** The \log_{10} of the Hi-C interaction counts of each pair of bins along chr1 is plotted versus the log of the genomic distance between each pair of bins. The median value of datapoints in the graph is indicated by a blue line while the 5% and 95% confidence intervals are shown as thin black lines. The slope of the median line from 500 kb to 10 Mb is -1, following the relationship expected for a fractal globule polymer structure of the chromatin. **D)** Red and blue “plaid” patterns show the compartmentalization of chr1 in two types of chromosomal domains. The data from A were transformed by first finding the observed interactions over the expected average pattern of decay away from the diagonal and then calculating a Pearson correlation coefficient between each pair of rows and columns. Regions highly correlated with one another in interaction are colored red and are likely to be classified by principle components analysis into the same compartment as shown above (black bands = open chromatin compartment; light grey bands = closed chromatin compartment). The compartment assignments correlate with the gene density profile, shown above the compartment profile (high gene density = black; low gene density = white). **E)** Whole chromosome interaction patterns show that longer chromosomes (chr1-10, chrX) are more likely to interact with one another and not with shorter chromosomes (chr14-22).

Methods for DNA Structure Assessment

Method	Full Name	Primary Function	Appearance Year
Hi-C	High-throughput Chromosome Conformation Capture	Maps long-range, genome-wide physical interactions to determine 3D folding and organization of chromatin.	2009
DNase-seq	DNase I Hypersensitive Sites sequencing	Uses DNase I enzyme to identify regions of open chromatin (active regulatory elements).	2008
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing	Uses a transposase enzyme to rapidly map accessible chromatin regions . (Modern standard, faster than DNase-seq).	2013
MNase-seq	Micrococcal Nuclease sequencing	Maps the precise location and occupancy of nucleosomes (the basic units of chromatin structure).	~2009
ChIP-seq	Chromatin Immunoprecipitation sequencing	While broad, it indirectly assesses structure by mapping binding sites of proteins (like histone modifications) that define active vs. repressed chromatin states.	2007

A (Non-Exhaustive) List of Useful References

ENCODE and modENCODE Guidelines For Experiments Generating ChIP, DNase, FAIRE, and DNA Methylation Genome Wide Location Data Version 2.0, July 20, 2011 (www.encodeproject.org)

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Landt et al., Genome Research, 2012, 22:1813.

ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Furey, Nat. Rev. Genetics 2012, 13:840

Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. O'Geen et al., 2011, Methods in Molecular Biology , 791:265

Design and analysis of ChIP-seq experiments for DNA-binding proteins. Kharchenko et al., 2008 Nature Biotechnology 26:1351

ChipSeq Exercise: tool installation

```
# install MEME
cd ~/tools
wget https://meme-suite.org/meme/meme-software/5.5.9/meme-5.5.9.tar.gz
tar xzf meme-5.5.9.tar.gz
cd meme-5.5.9
./configure --prefix=$HOME/meme --with-url=http://meme-suite.org --enable-build-libxml2 --enable-build-libxslt
make -j 4 install
```

```
# add line below to ~/.bashrc
export PATH=$HOME/meme/bin:$PATH
```

```
# add bedGraphToBigWig tool
cd ~/meme/bin
wget http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/bedGraphToBigWig
chmod u+x bedG*
```

```
# get and start Exercise
cd ~/tools
wget https://genomics-lab.fleming.gr/fleming/uoa/vm/ChIP-seq.zip
unzip ChIP-seq.zip
cd ChIP-seq/
google-chrome 20121016_ChIP-seq_Practical.pdf &
```

```
#build bowtie index (~15min)
bowtie-build bowtie_index/mm10.fa bowtie_index/mm10
```


ChIPSeq Exercise

alignment, direct output to sorted bam

```
bowtie -p 4 -m 1 -S bowtie_index/mm10 gfp.fastq | samtools view -bS - | samtools sort - > gfp.bam  
samtools index gfp.bam
```

```
bowtie -p 4 -m 1 -S bowtie_index/mm10 Oct4.fastq | samtools view -bS - | samtools sort - > Oct4.bam  
samtools index Oct4.bam
```

start igv , switch the current genome to “Mouse (GRCm38/mm10)”, load Oct4.bam , jump to gene “Lemd1”, zoom in

peak finding

```
macs -t Oct4.bam -c gfp.bam --format=BAM --name=Oct4 --gsize=138000000 --tsize=26 --diag --wig
```

motif finding, New instructions replacing page 12 to 15:

```
slopBed -i Oct4_summits.bed -g bowtie_index/mouse.mm10.genome -b 20 > Oct4_summits-b20.bed  
fastaFromBed -fi bowtie_index/mm10.fa -bed Oct4_summits-b20.bed > Oct4_summits-b20.fa  
~/meme/bin/meme Oct4_summits-b20.fa -o meme -dna -nmotifs 3  
goolge-chrome meme/meme.html
```