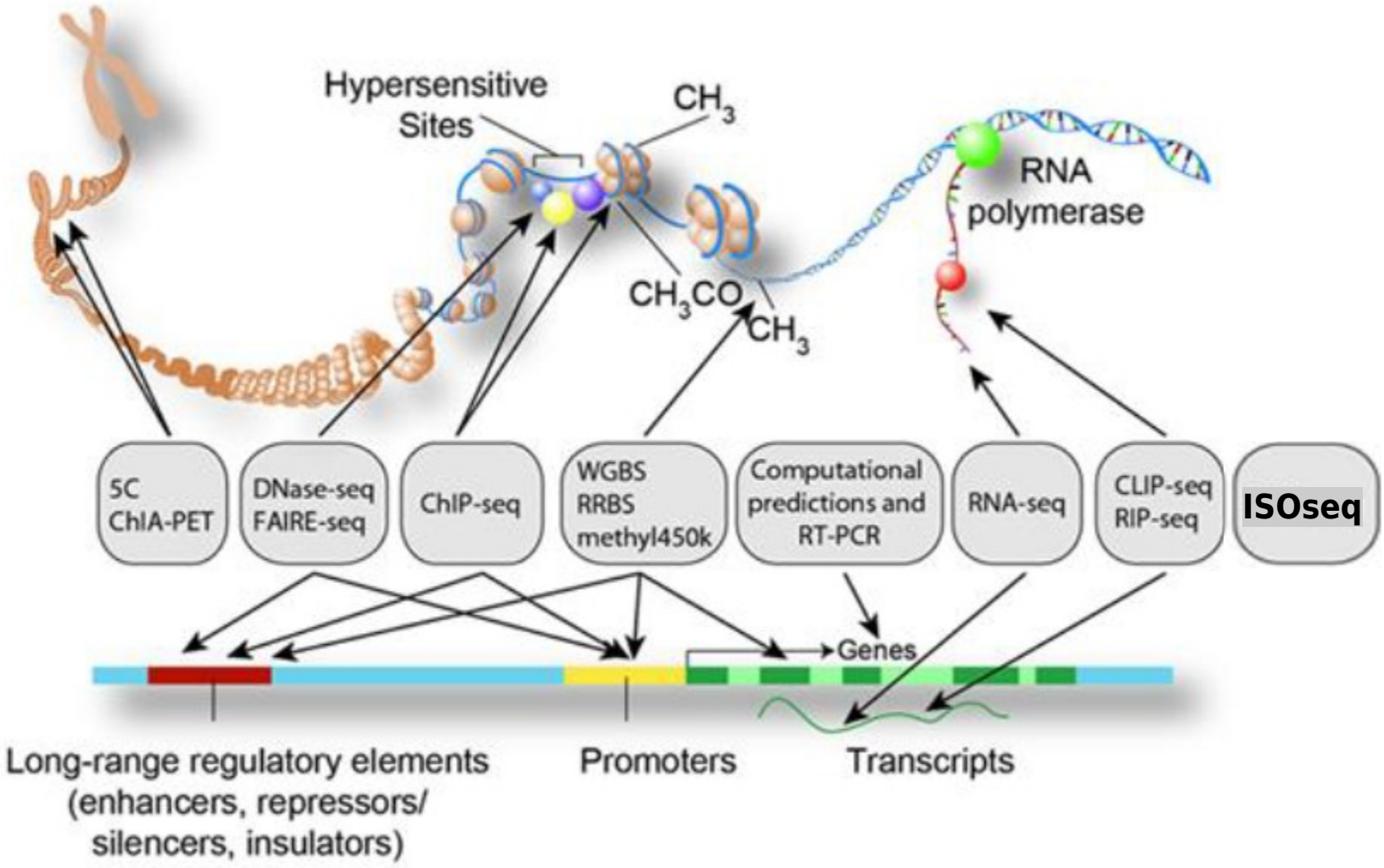
Syllabus and grading

#	Date	Short title	Lecturer	Other actions	Subject
1	Tuesday, October 10, 2023	introduction	MR		Overview of Bioinformatics, sequence alignment
2	Tuesday, October 17, 2023	Linux/shell/ssh	AD		Introduction to Linux and the command line, bash scripting and ssh
3	Tuesday, October 24, 2023	R (1)	AD		Introduction to the R programming language and Rstudio usage
4	Tuesday, October 31, 2023	QC+RNASeq	MR		Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
5	Tuesday, November 7, 2023	R (2)	AD		Advances R subjects, introduction to Bioconductor
6	Tuesday, November 14, 2023	bedtools/vcftools/samtools	AD		Command line tool usage: bedtools, vcftools, samtools etc.
7	Tuesday, November 21, 2023	Denovo	MR		NGS for denovo genome and transciptome assembly
8	Tuesday, November 28, 2023	ChipSeq/chirp	MR	assign presentations	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture (3C))
9	Tuesday, December 5, 2023	metabolomics	MR		Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
10	Tuesday, December 12, 2023	Exome/SNP calling	AD	assign final projects	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	Tuesday, December 19, 2023	presentations	MR+AD		Paper presentations by students
12	Tuesday, January 9, 2024	presentations	MR+AD		Paper presentations by students
13	Tuesday, January 16, 2024	final projects support	MR+AD		Support for the final project

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

Functional Elements in the Genome

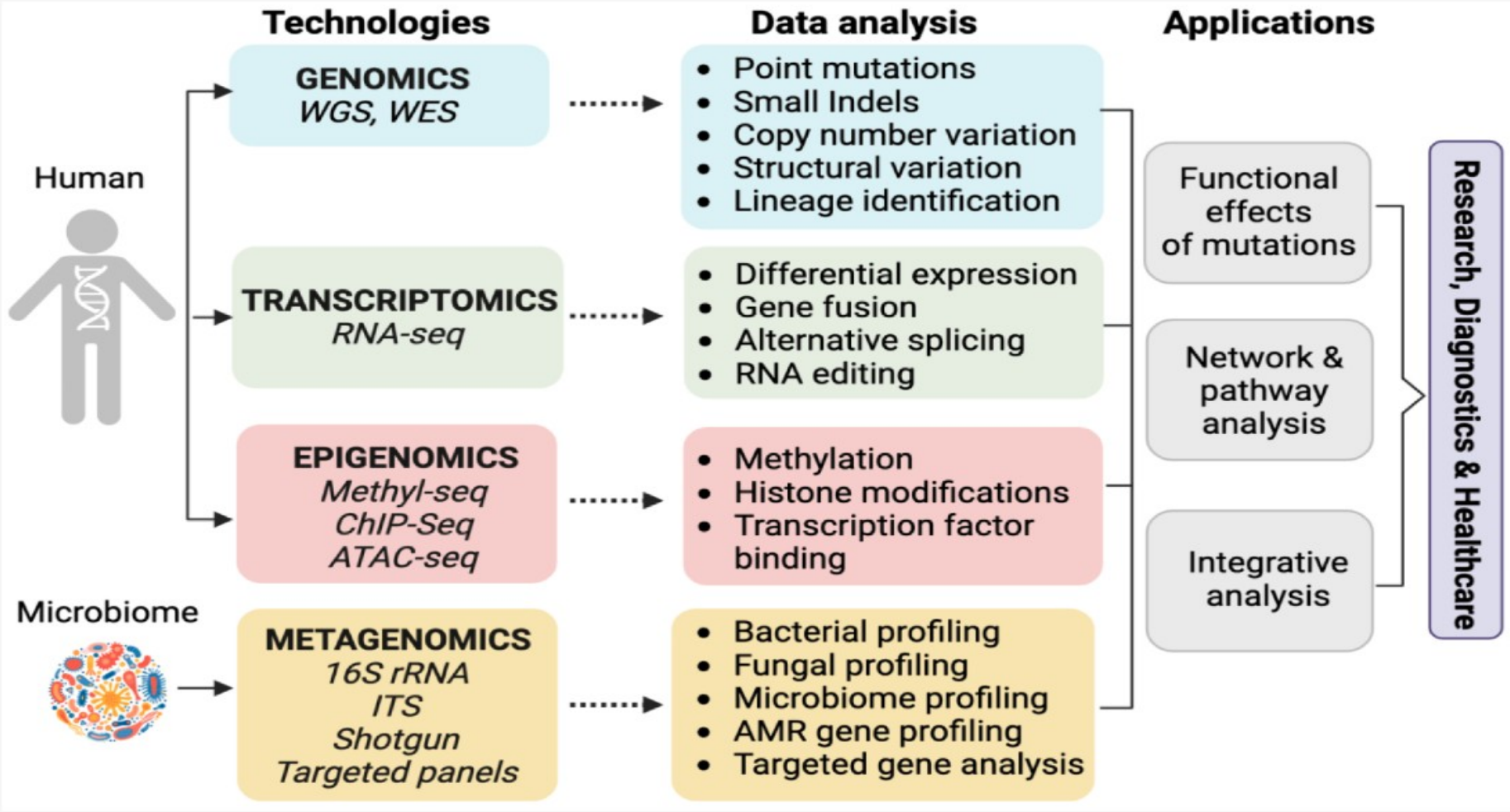


www.encodeproject.org

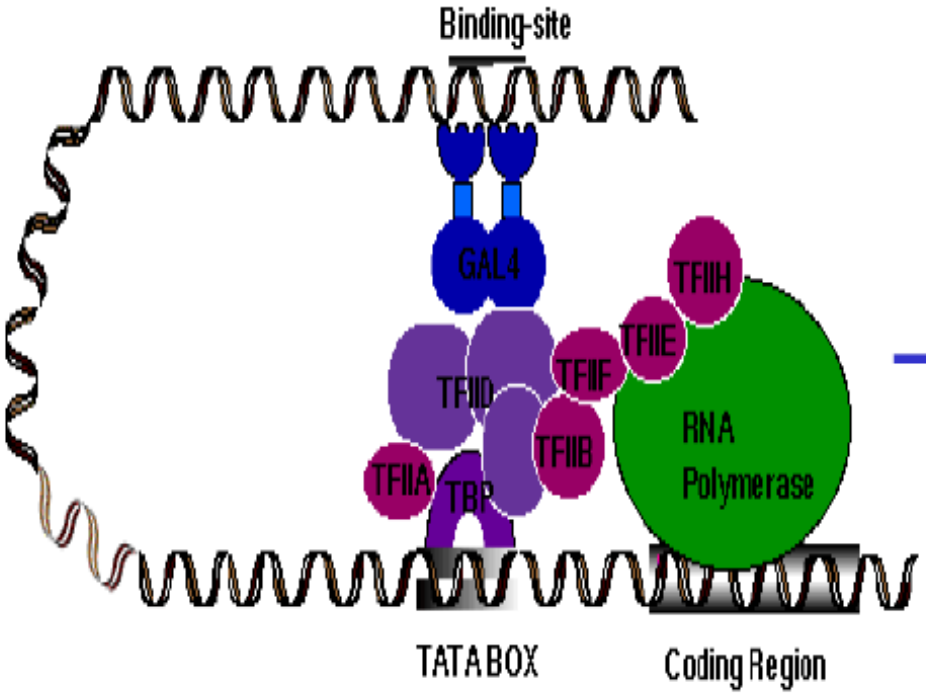
Check also 2020 NGS review at https://www.nature.com/immersive/d42859-020-00099-0/pdf/d42859-020-00099-0.pdf

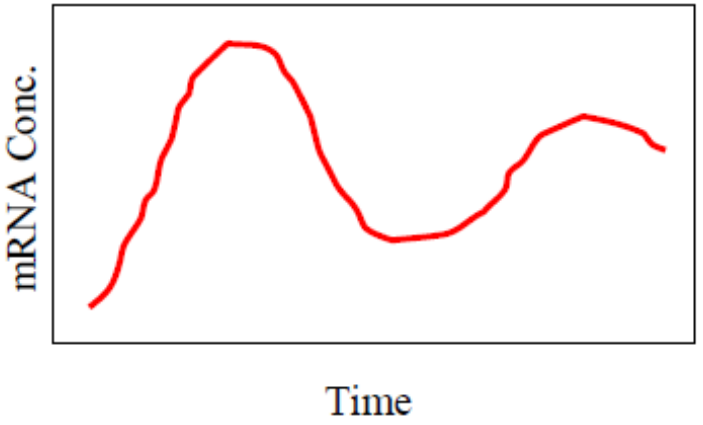
From: Next-Generation Sequencing Technology: Current Trends and Advancements
https://www.mdpi.com/2079-7737/12/7/997

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



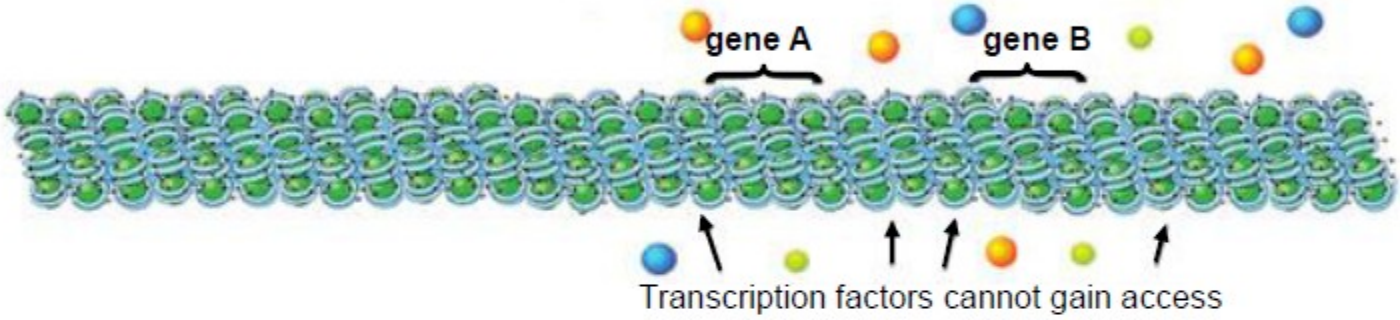


Sequence

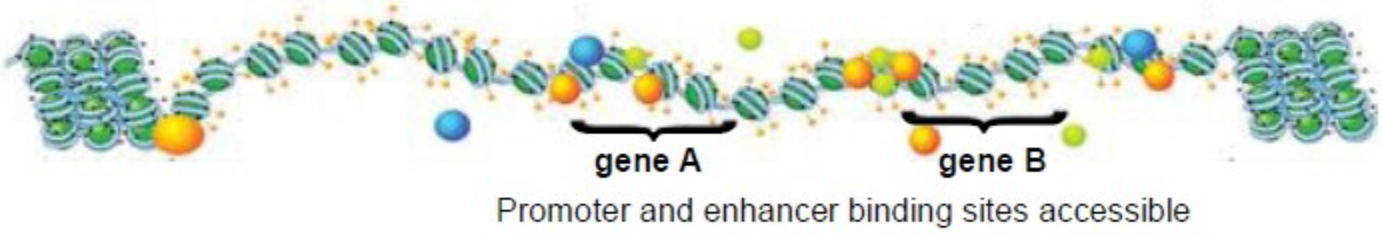
Function

Chromatin Structure Determines Gene Status

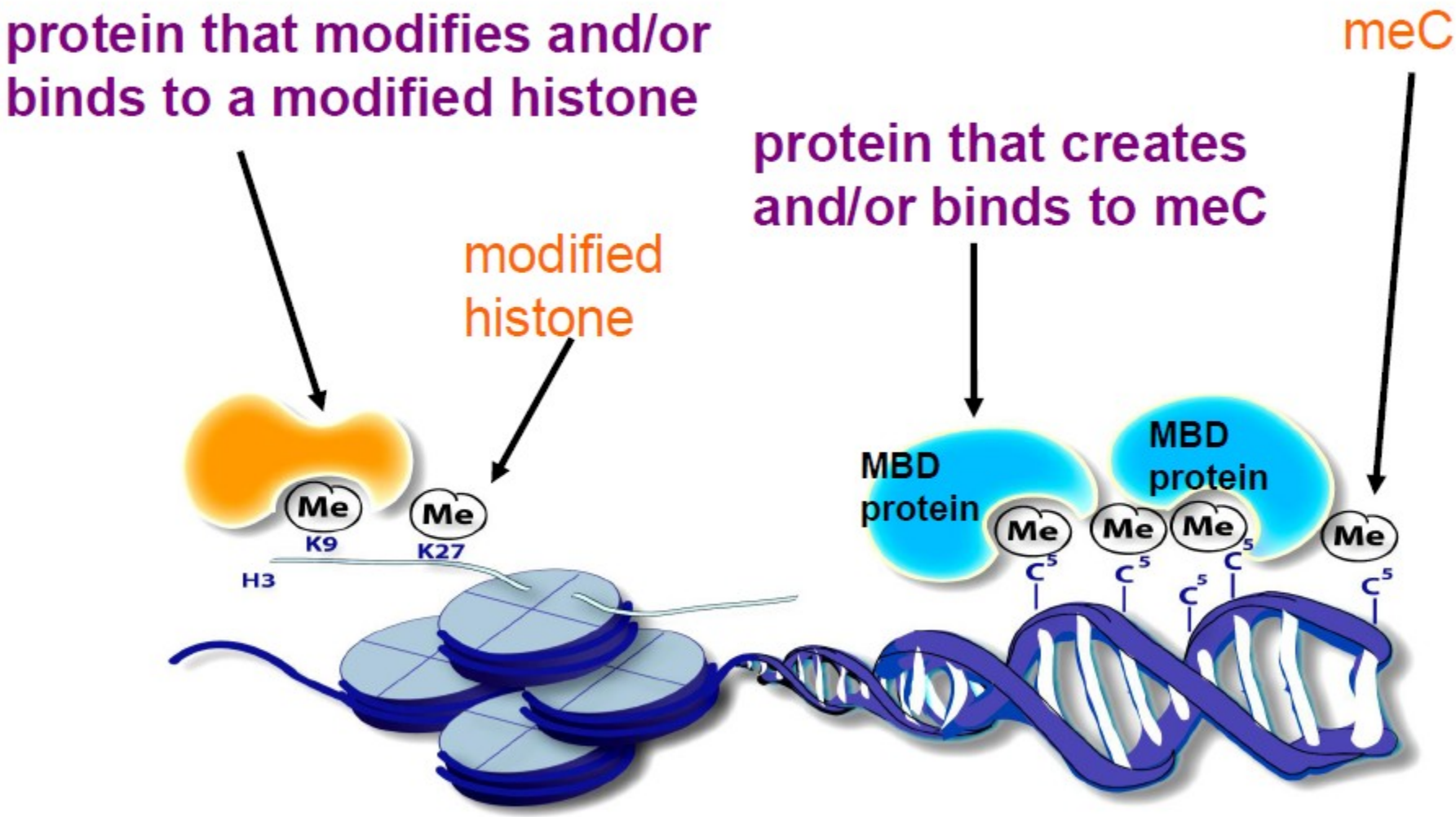
Closed Chromatin



Open chromatin



DNA and Histone Modifications Create an Epigenome

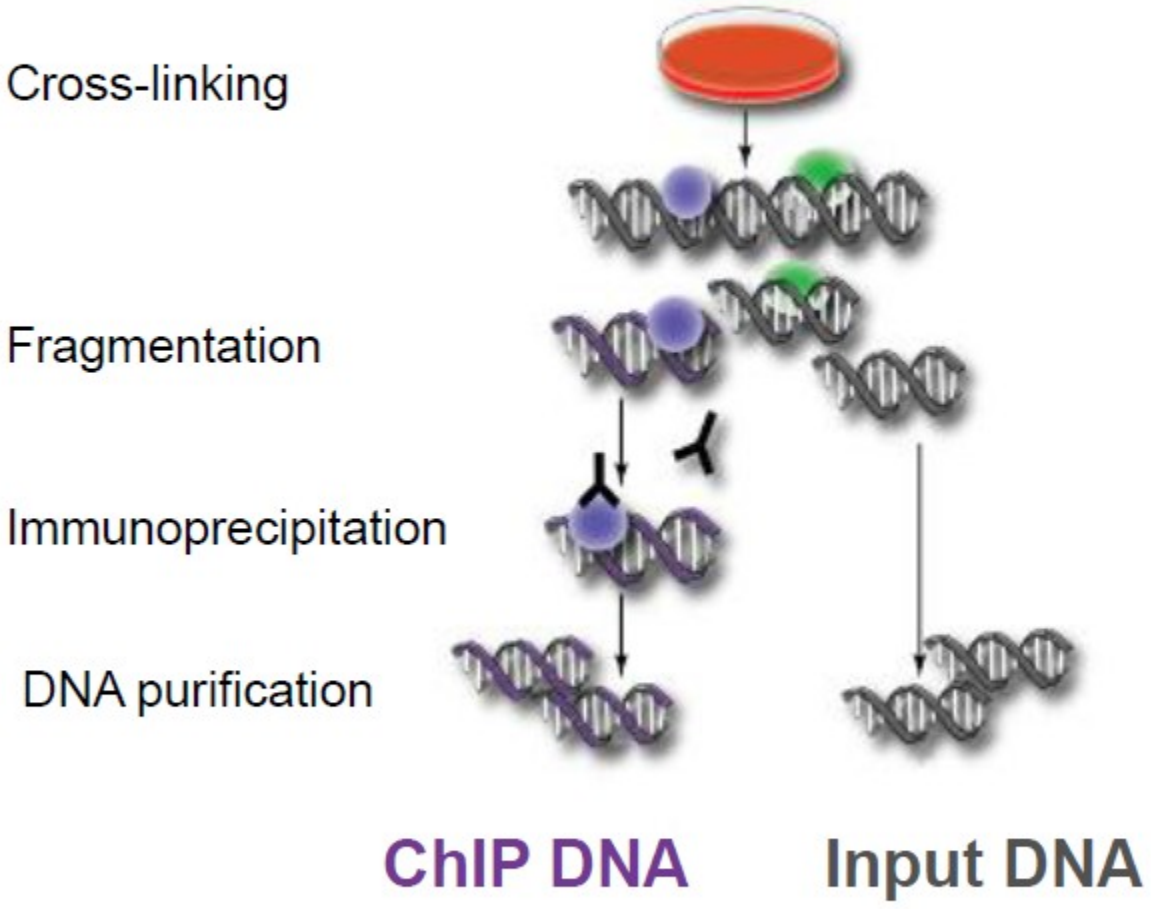


Methyl-CpG $_{(5'-C-phosphate-G-3')}$-binding domain (MBD)

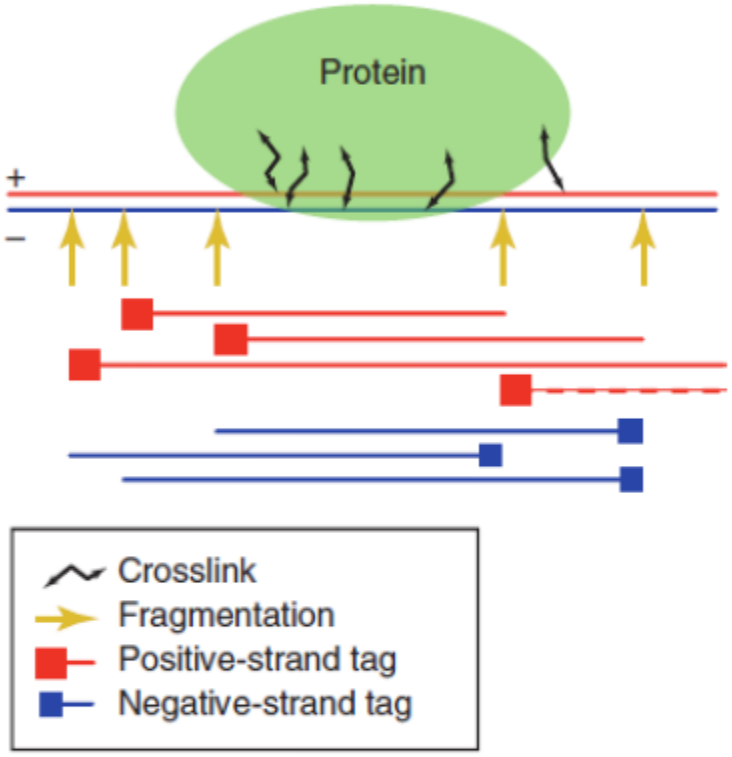
Coordinated Efforts to Decipher Epigenomes

- 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



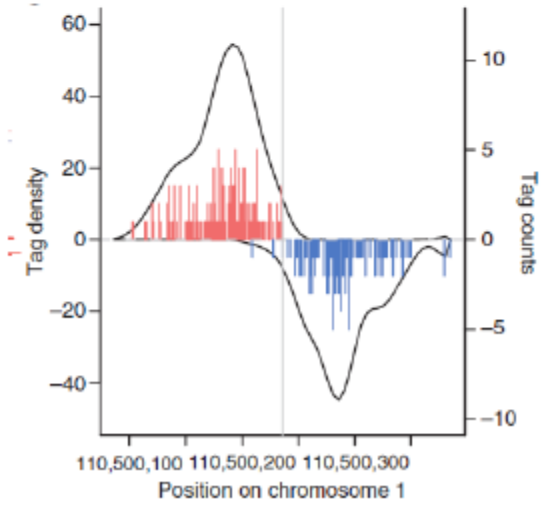
From Binding Site to Sequence to Peak

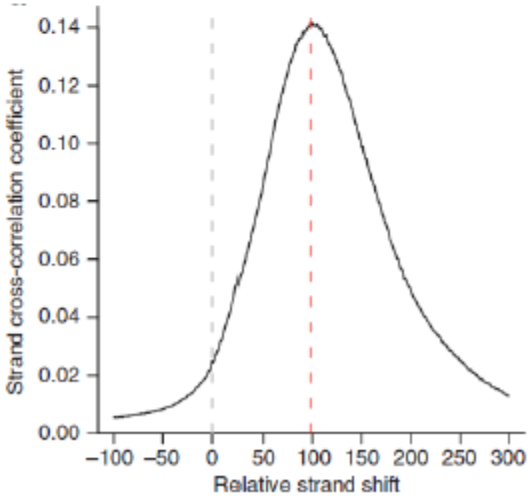


Short sequences are generated from each DNA molecule.

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



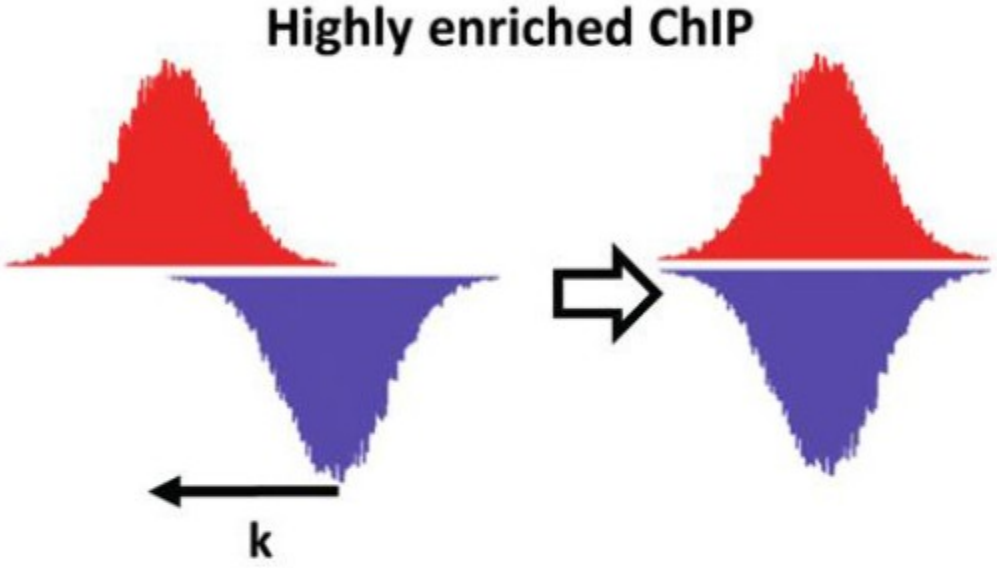


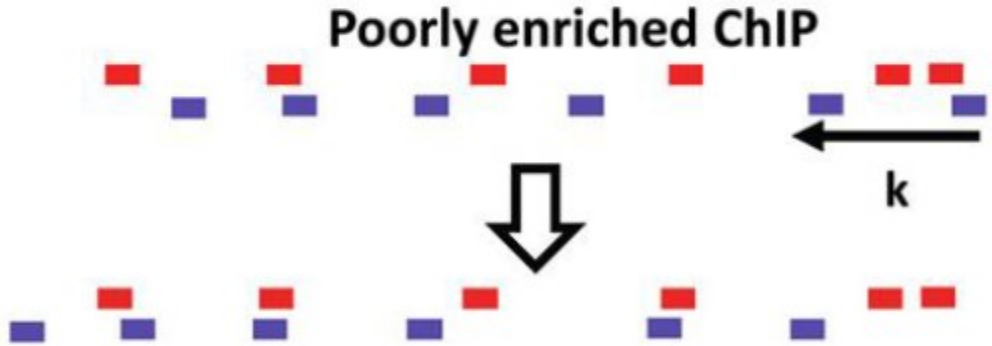
Kharchenko et al., 2008
Nat. Biotech. 26:1351

Library Complexity and Cross-Correlation

Typical ChIP-seq peak

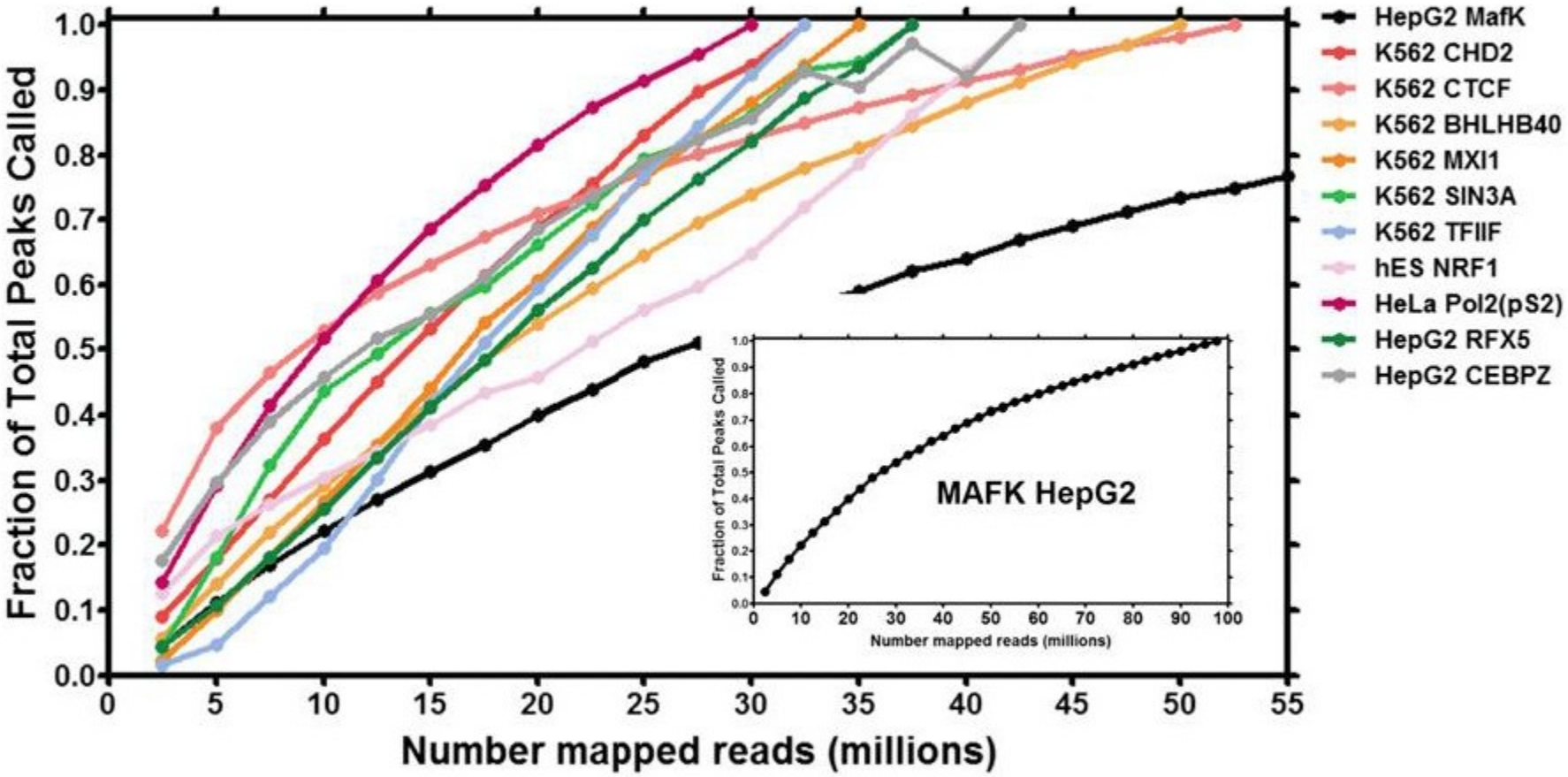
Low-complexity ChIP-seq peak





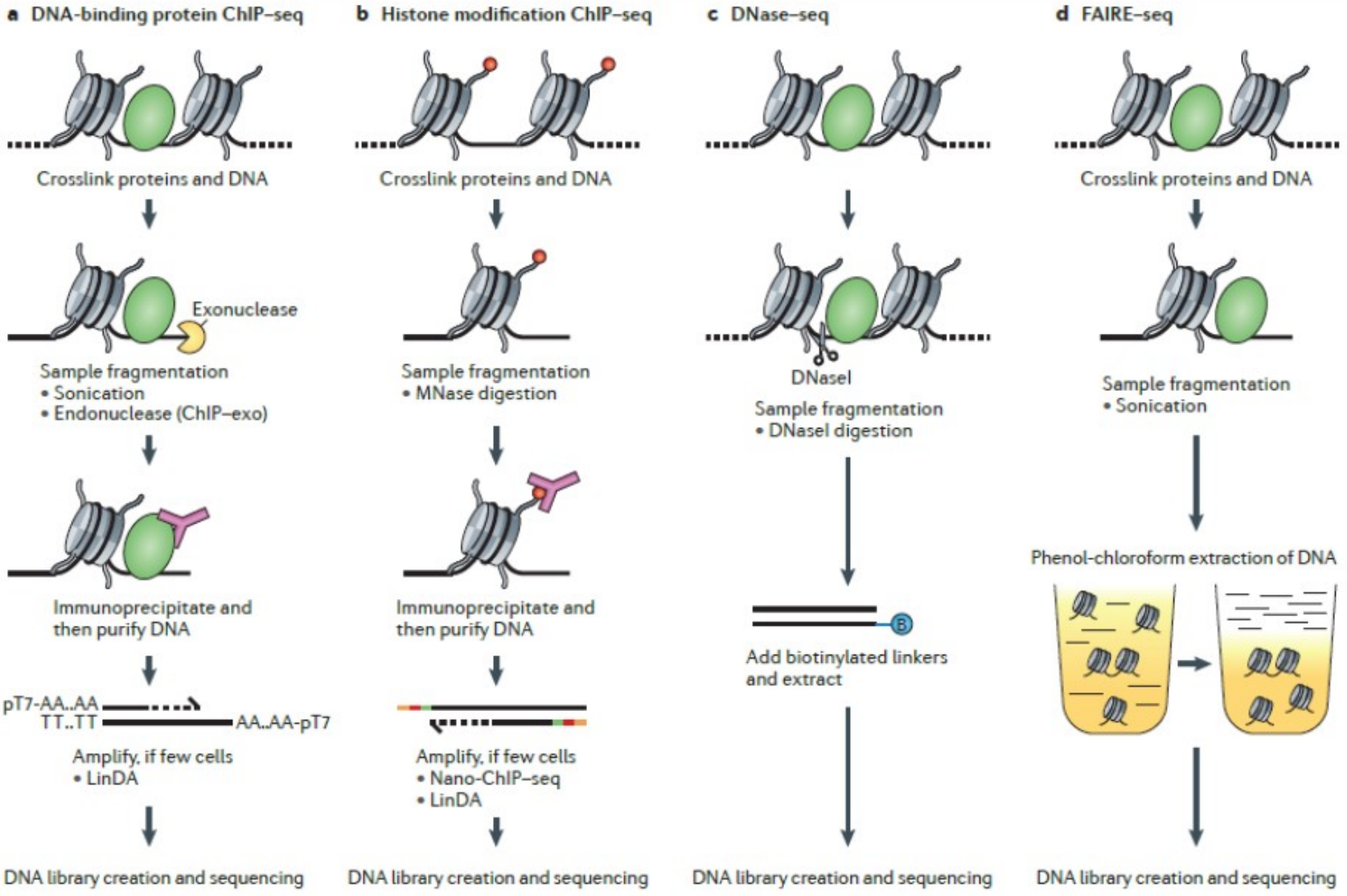
Landt, et al., 2012 Genome Res. 22:1813

Called Peaks Increase With Sequencing Depth



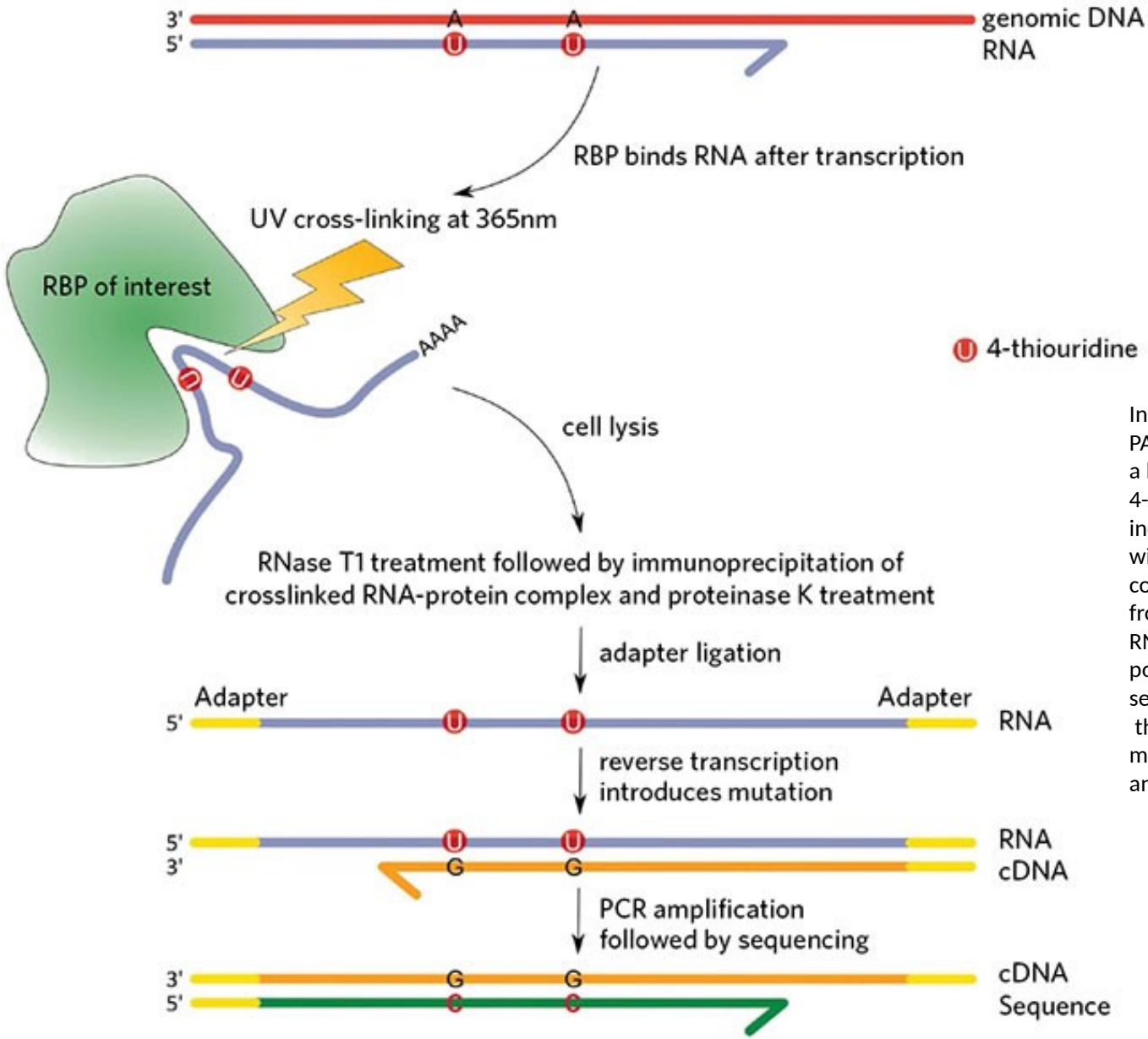
Landt, et al., 2012 Genome Res. 22:1813

Comparison of Experimental Protocols



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

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

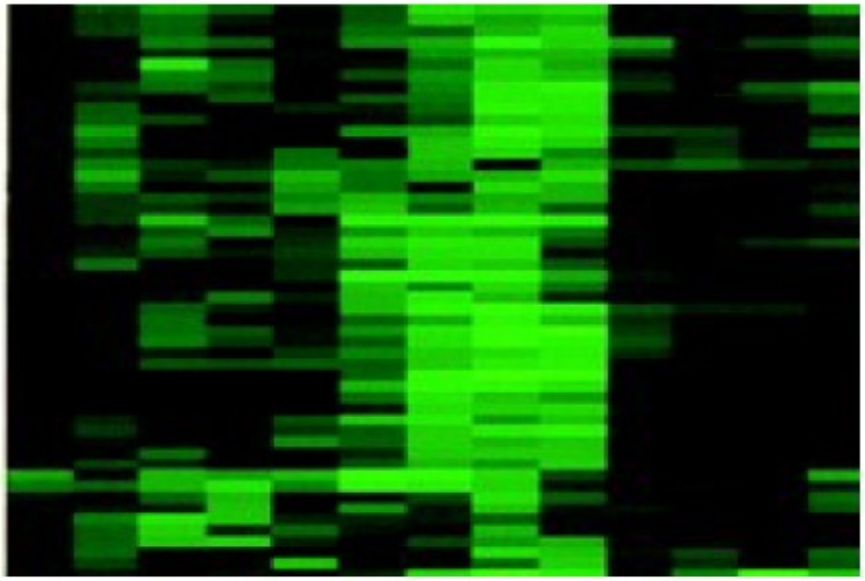
- Controls are Important!
  - Necessary to avoid non-uniform background (sonication, etc.)
  - Many cell lines have aneuploidy (genome size, copy number)
  - "Input" controls – similar prep, but no ChIP.
  - "IgG" – IP without the specific antibody
  - If amplification is done, must be done on all samples, 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

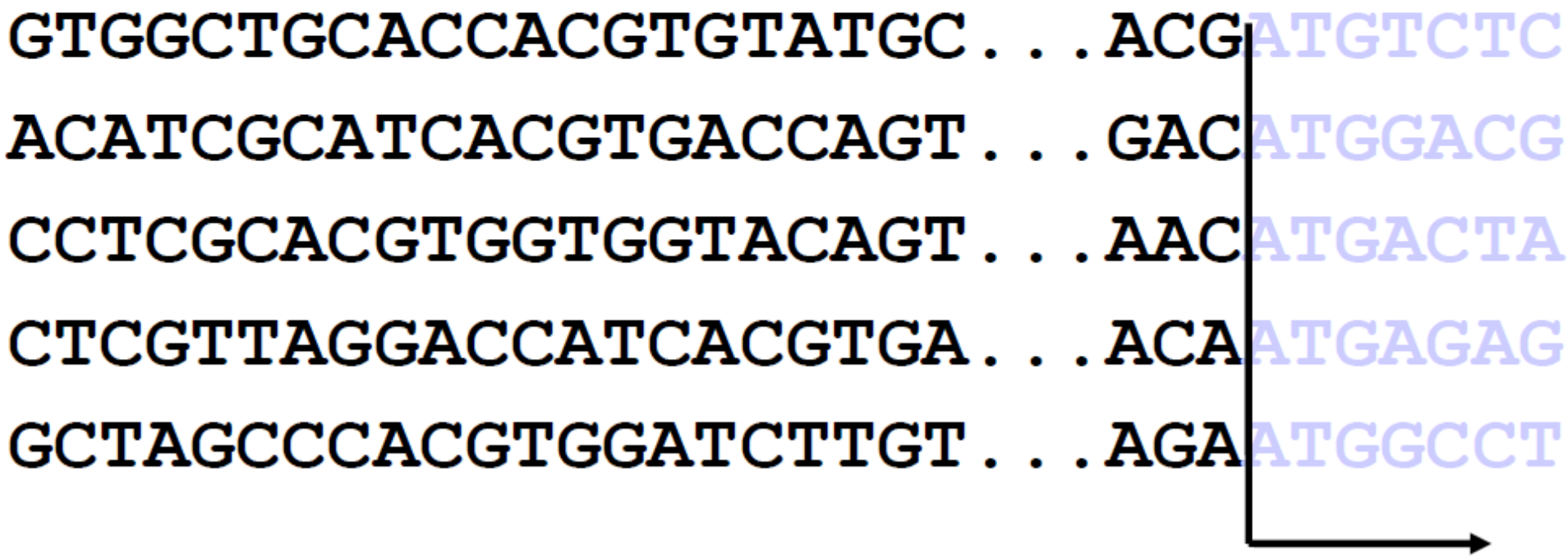


Assuming co-expressed genes are co-regulated, we look in their promoter regions to find conserved motifs, confirming that the same TF binds to them

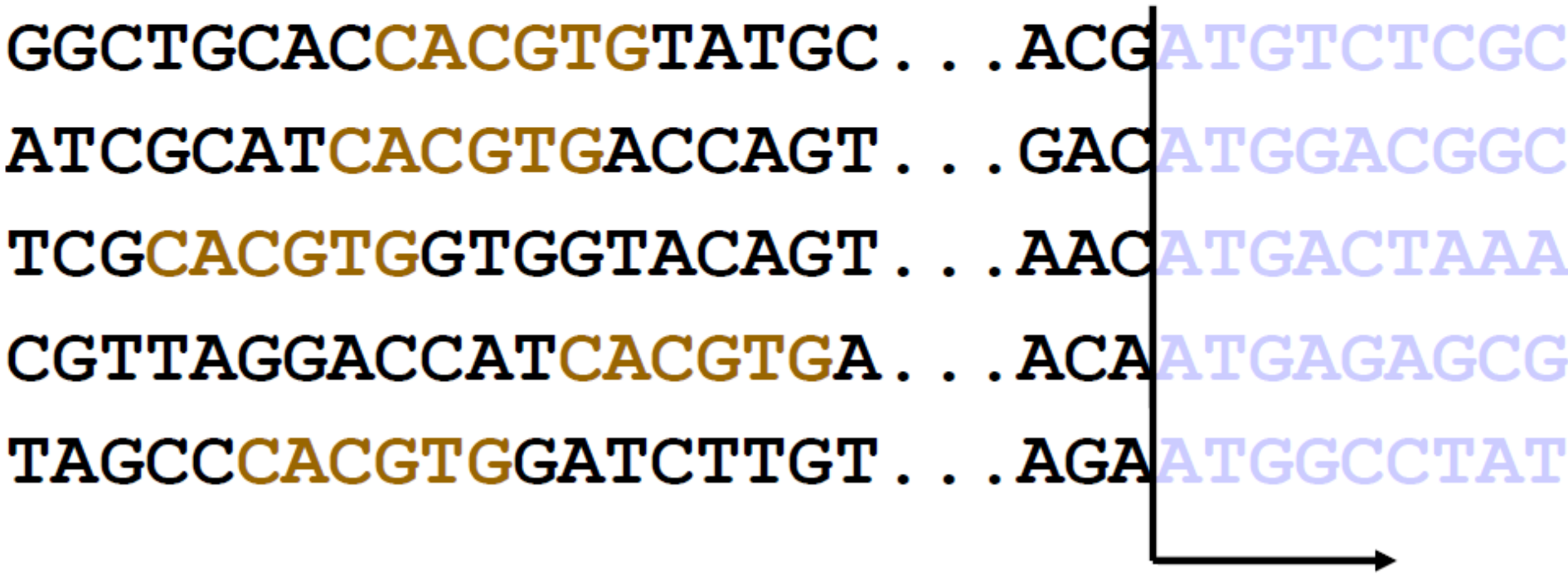
Motifs vs Transcription Factor Binding Sites

- 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



Finding Motifs in a Set of Sequences



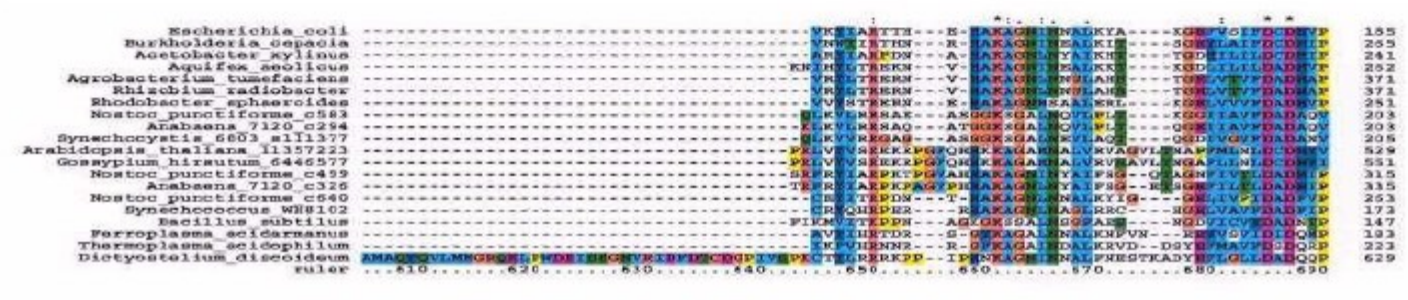
Finding Motifs in a Set of Sequences

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!



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.

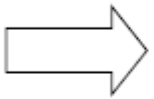
ACGGG

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

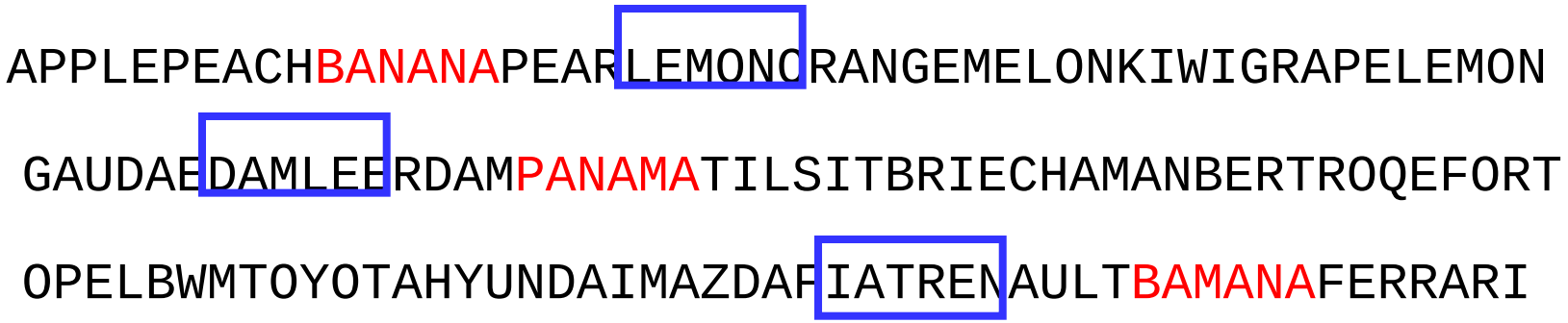
Position (Frequency) Weight Matrix

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 occuring substrings

"*co-regulated genes*":

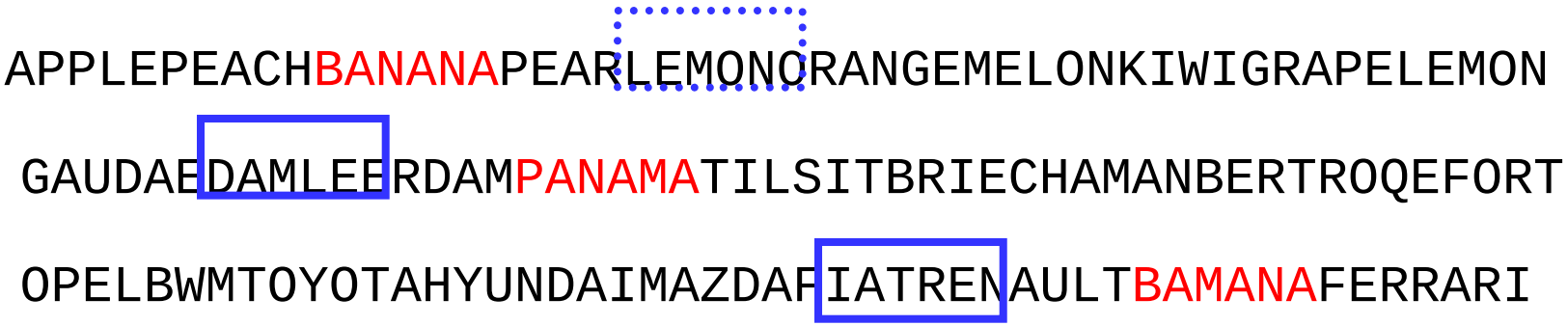


Start with random positions for substrings

Gibbs sampling

Find location AND description of commonly occuring substrings

Step 1a:

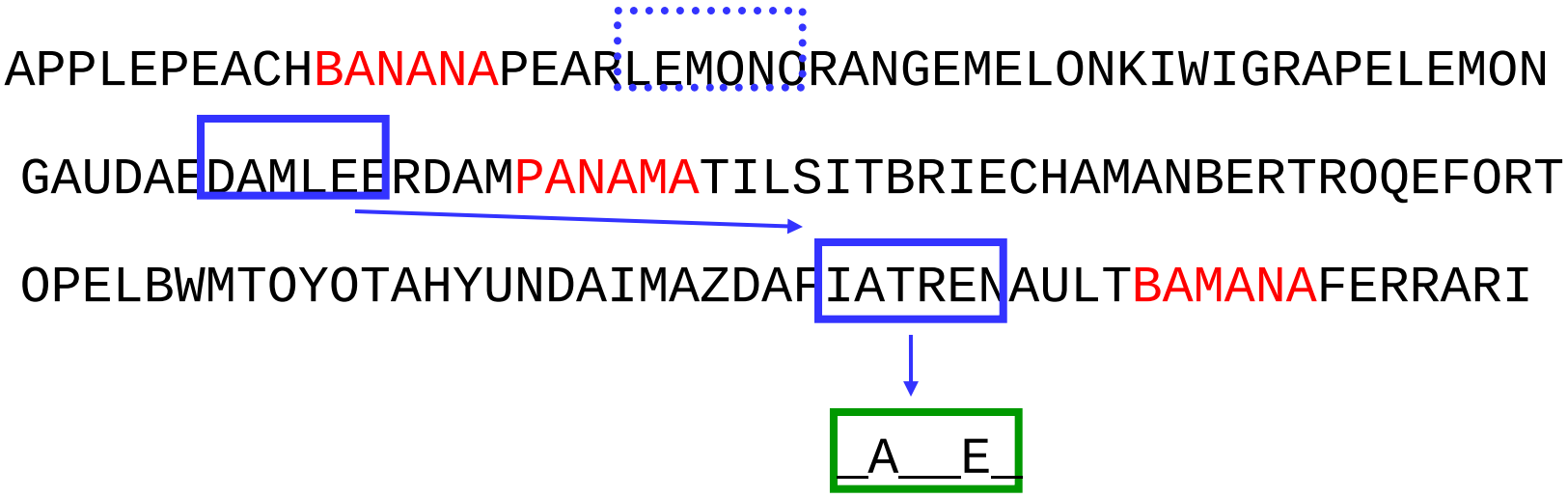


Pick one sequence

Gibbs sampling

Find location AND description of commonly occuring substrings

Step 1b:

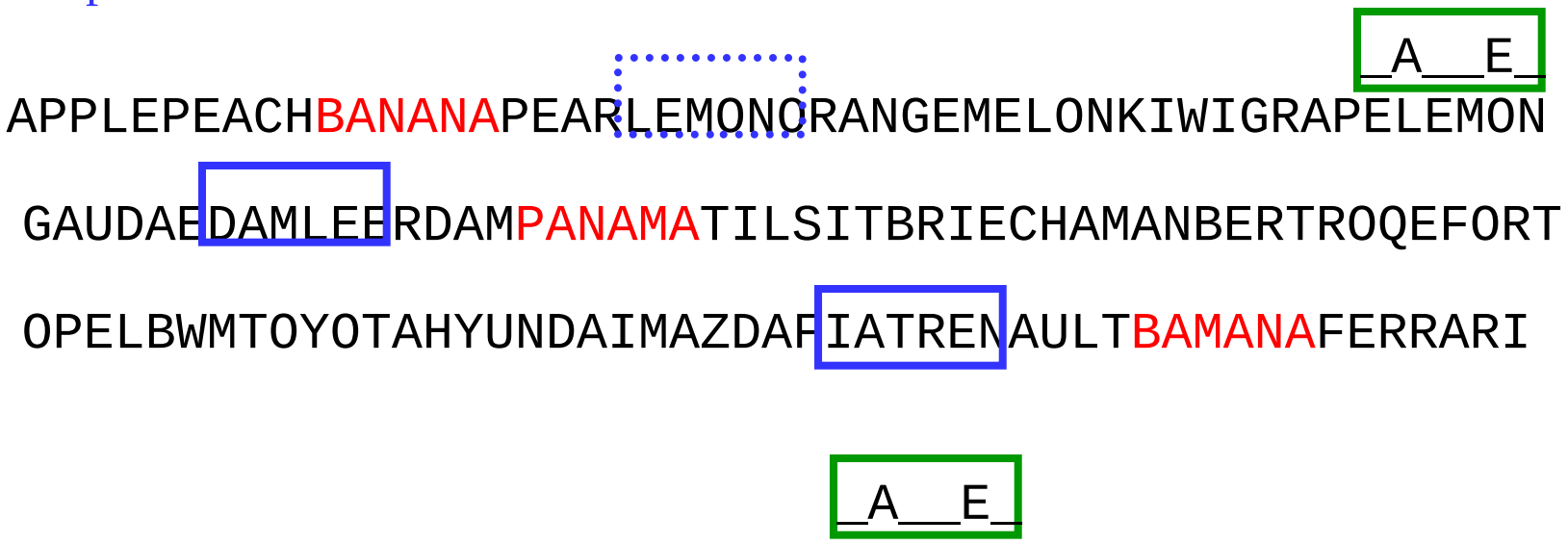


Get statistics of all other substrings

Gibbs sampling

Find location AND description of commonly occuring substrings

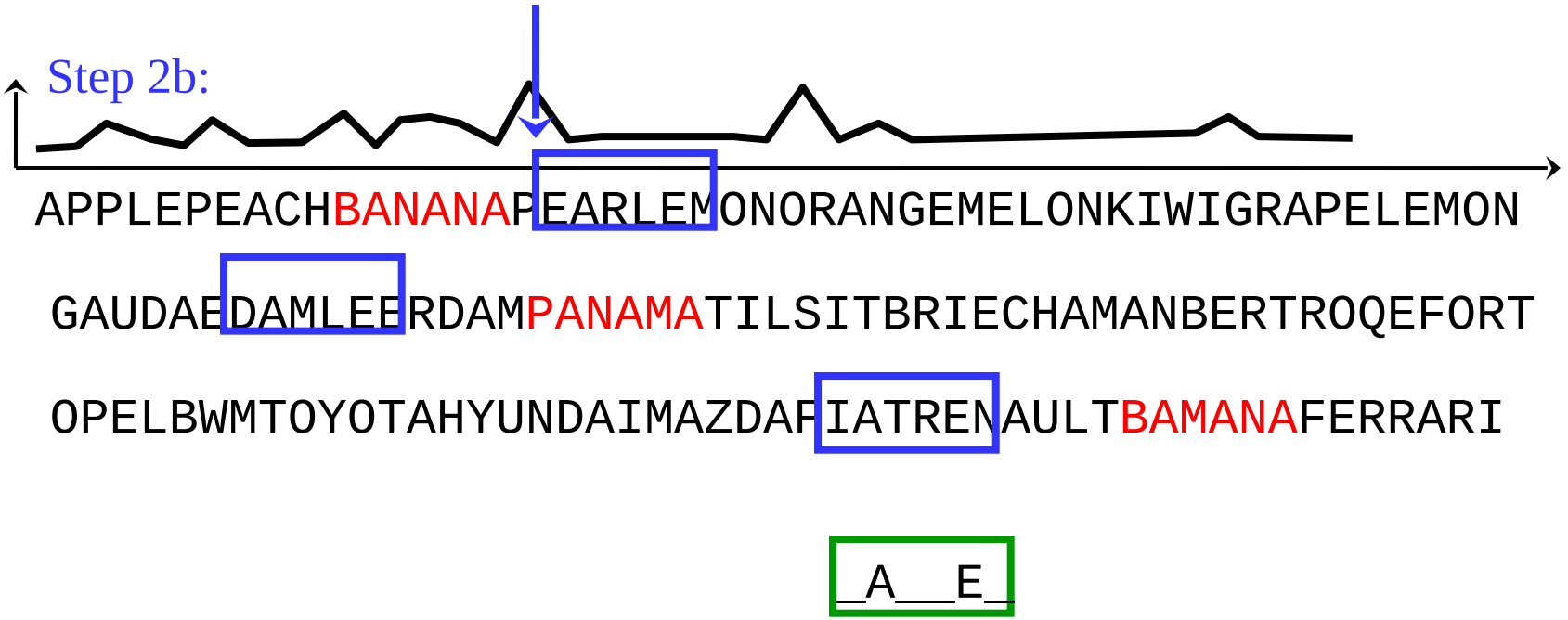
Step 2a:



match description to all locations in sequence

Gibbs sampling

Find location AND description of commonly occuring substrings

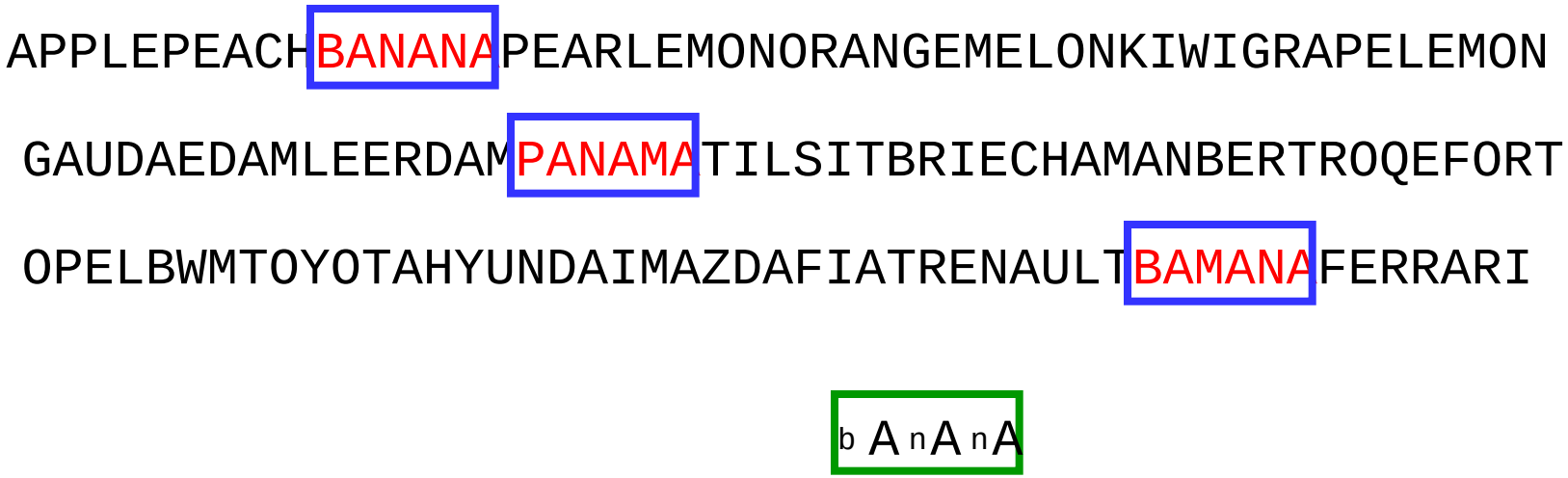


Pick new location in sequence (probabilistic)

Gibbs sampling

Find location AND description of commonly occuring substrings

Repeat steps 1 and 2 until convergence:



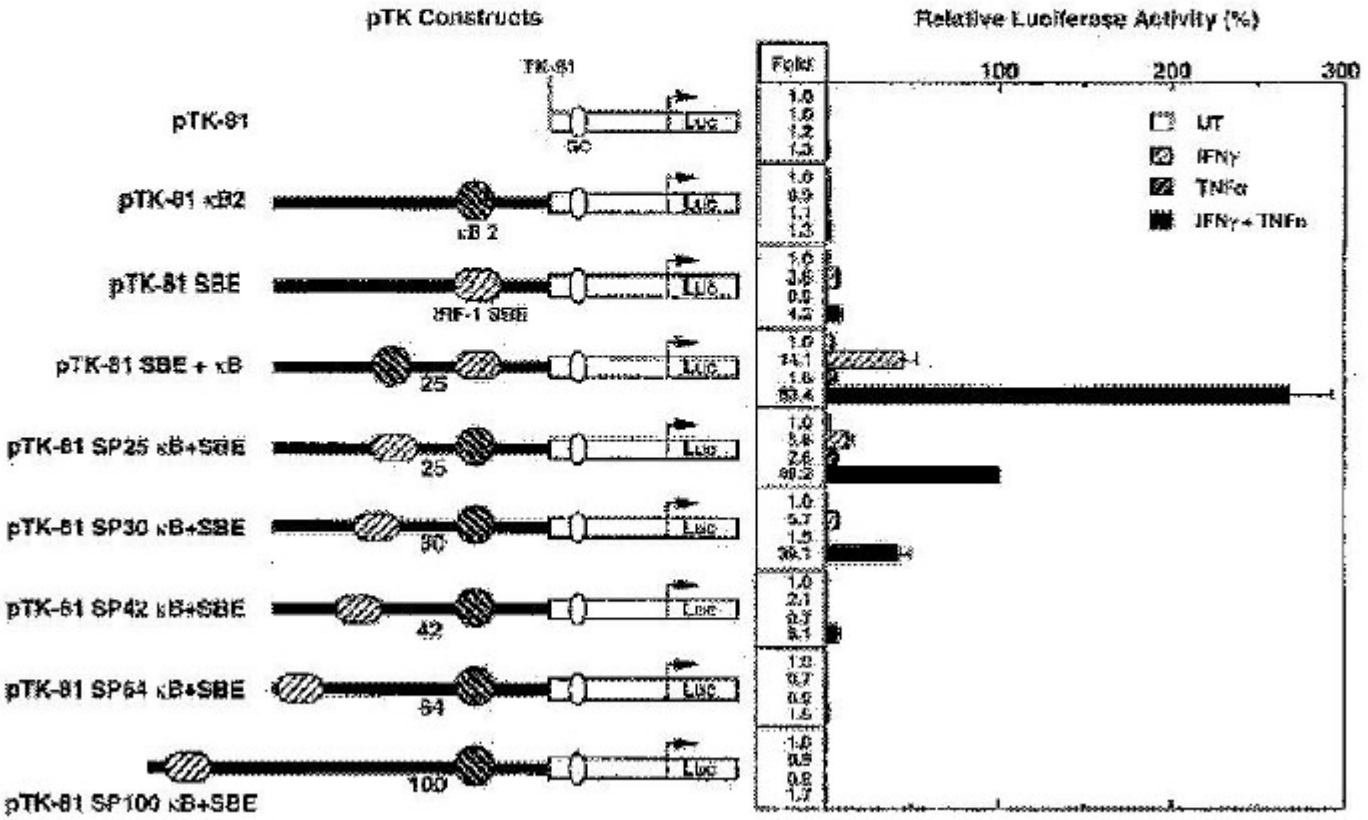
Multi-site Motif

- 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
```

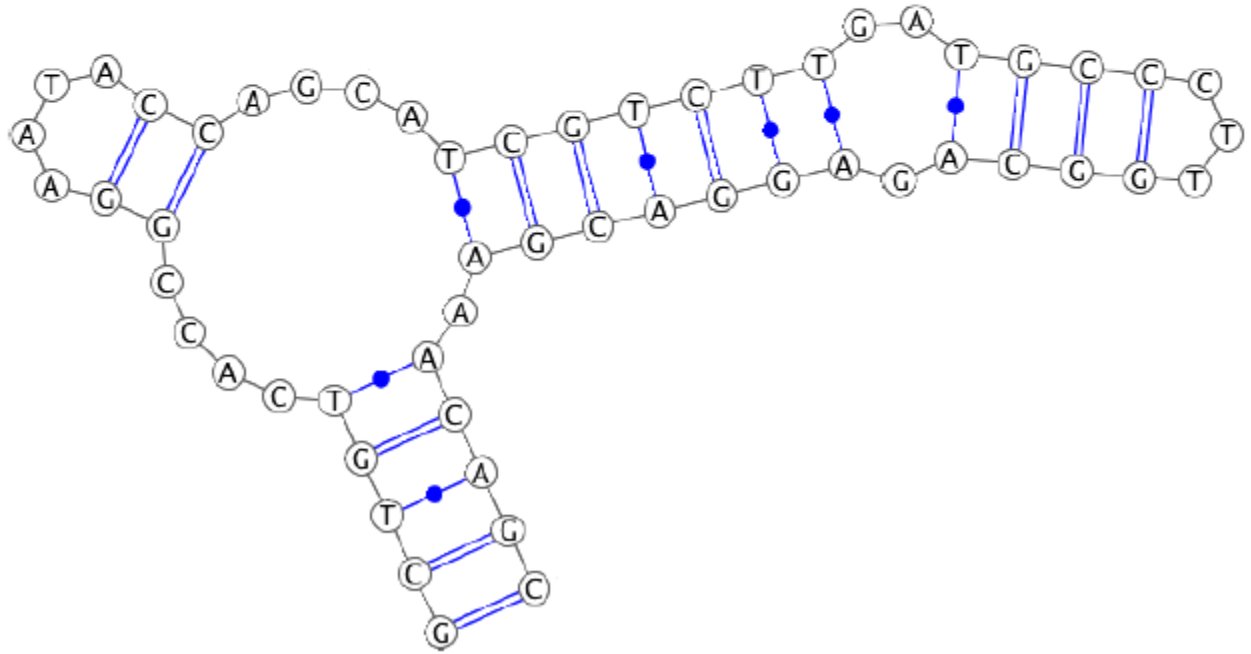
Dependence of Simple Motif Pairs on Distance and Order Between Them



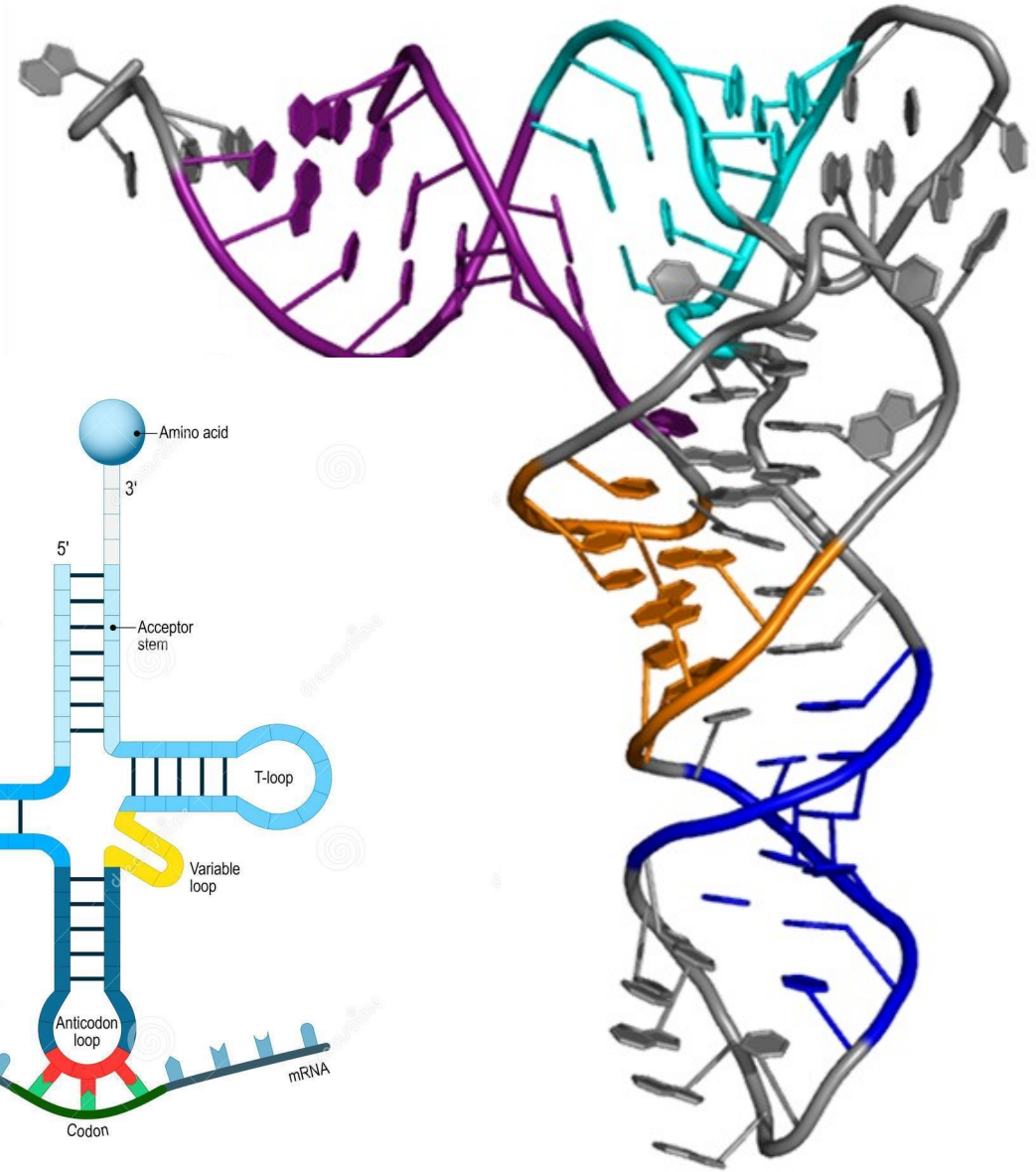
Ohmori et al., 1997

RNA SECONDARY STRUCTURE

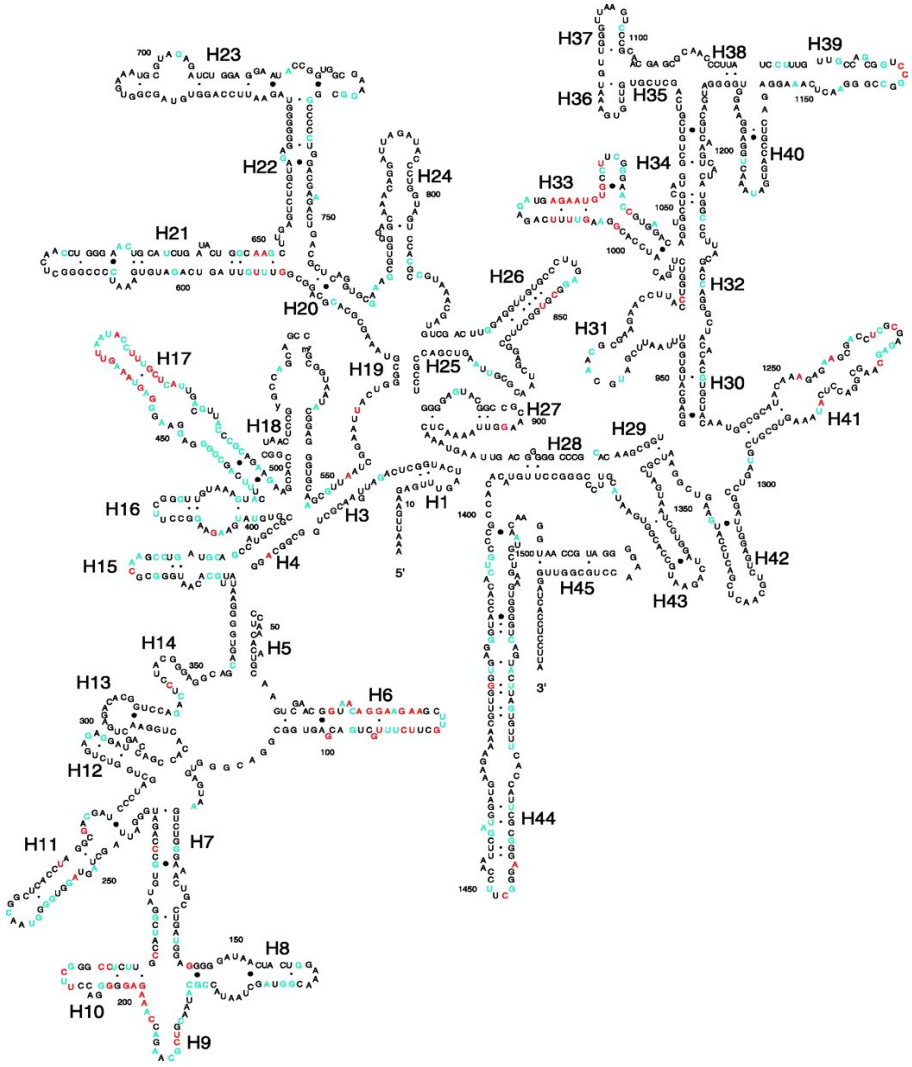
Sequence → **Secondary Structure** → Tertiary Structure

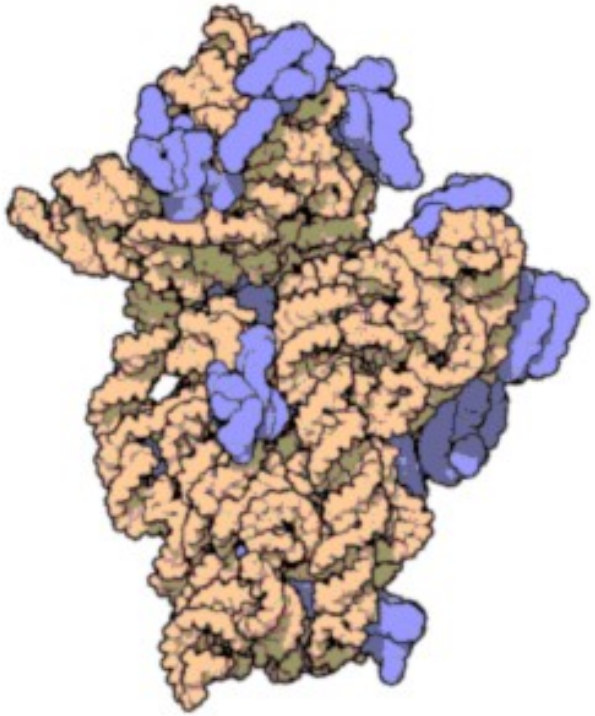


Transfer RNA (tRNA)



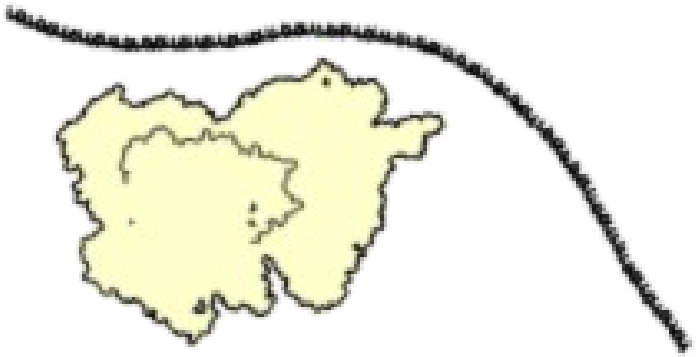
Ribosomal RNA (rRNA)



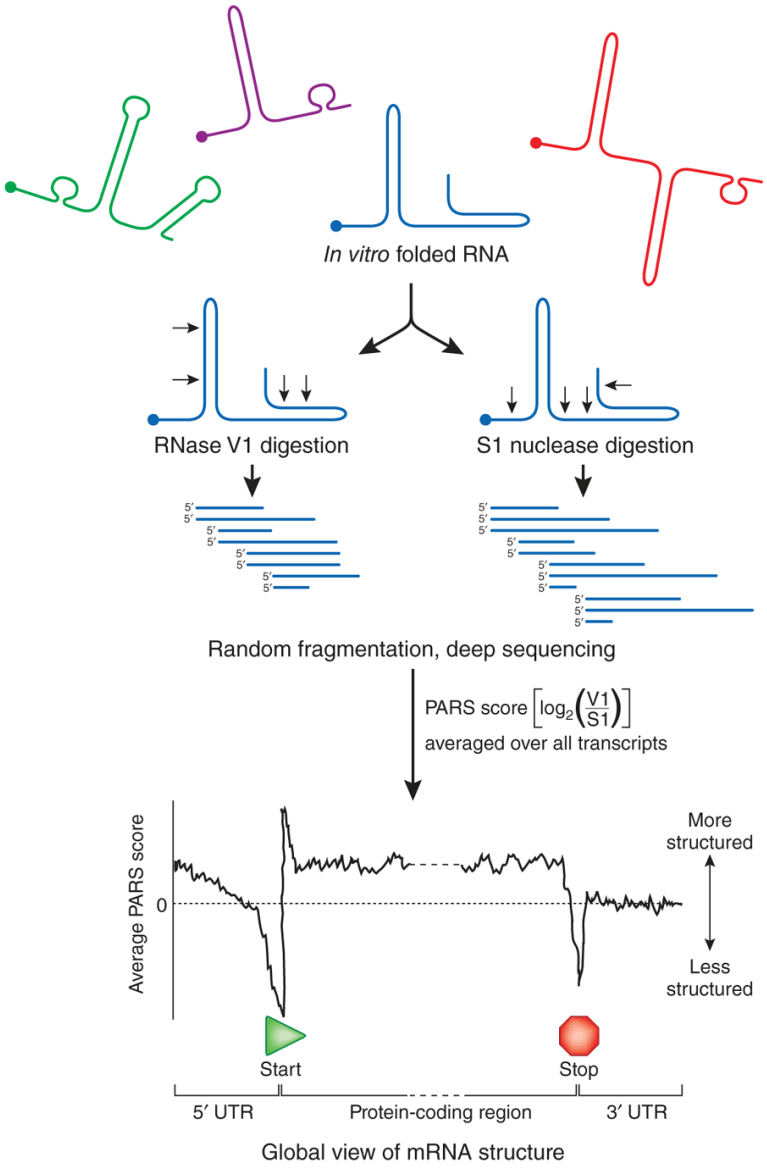


16S-rRNA (orange), proteins (blue)

rRNA+tRNA in Ribosome



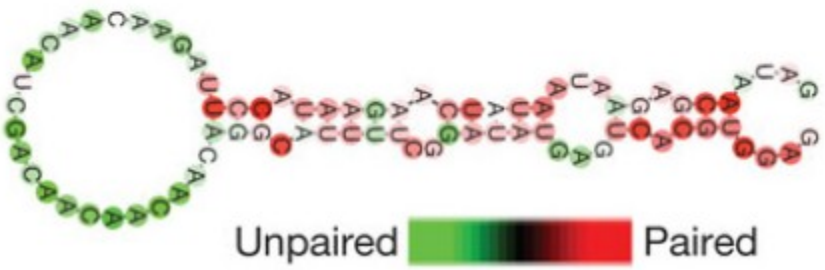
Parallel Analysis of RNA Structure (PARS)



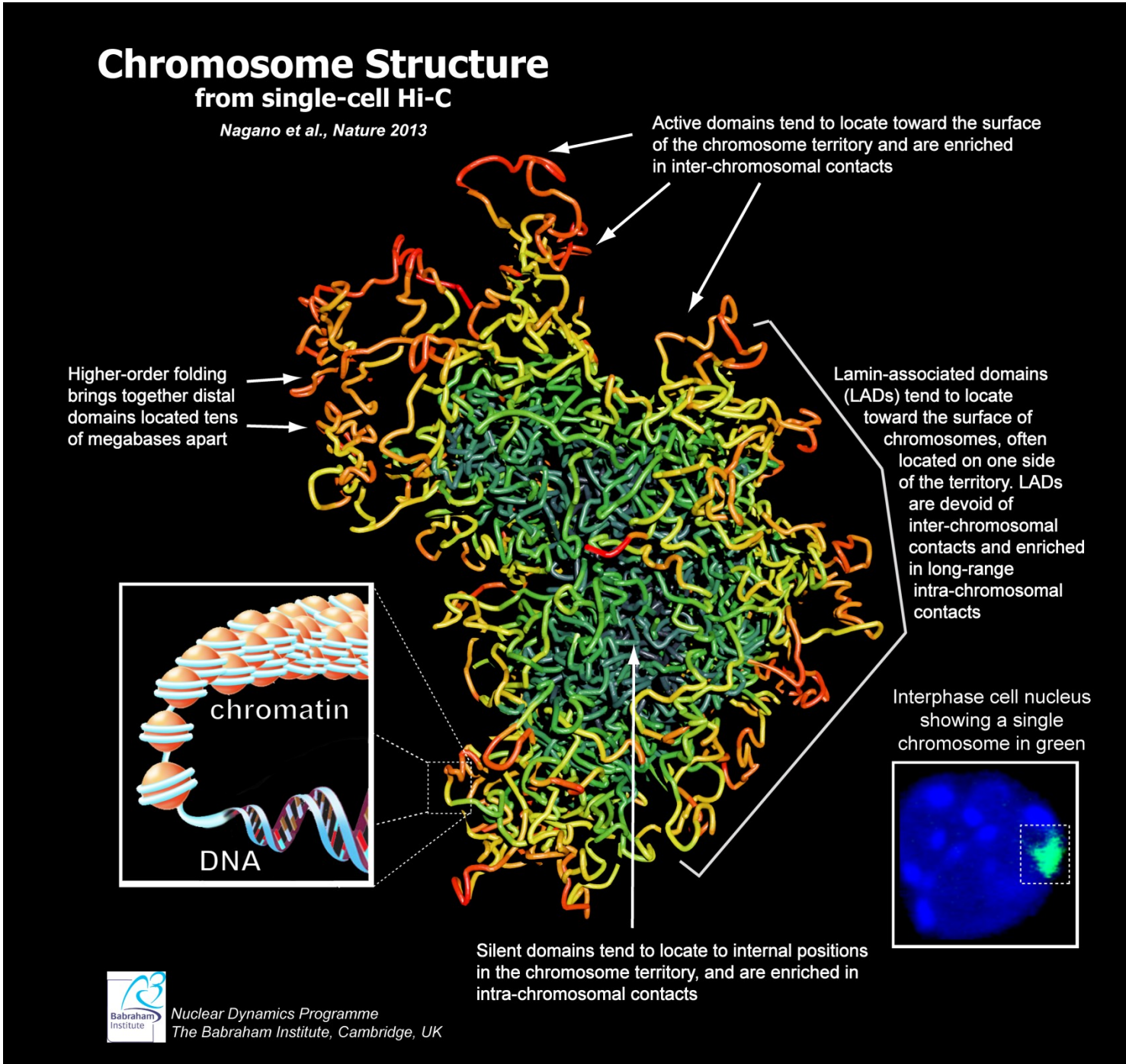
PARS SCORE

Less Structure = more unpaired = score < 0

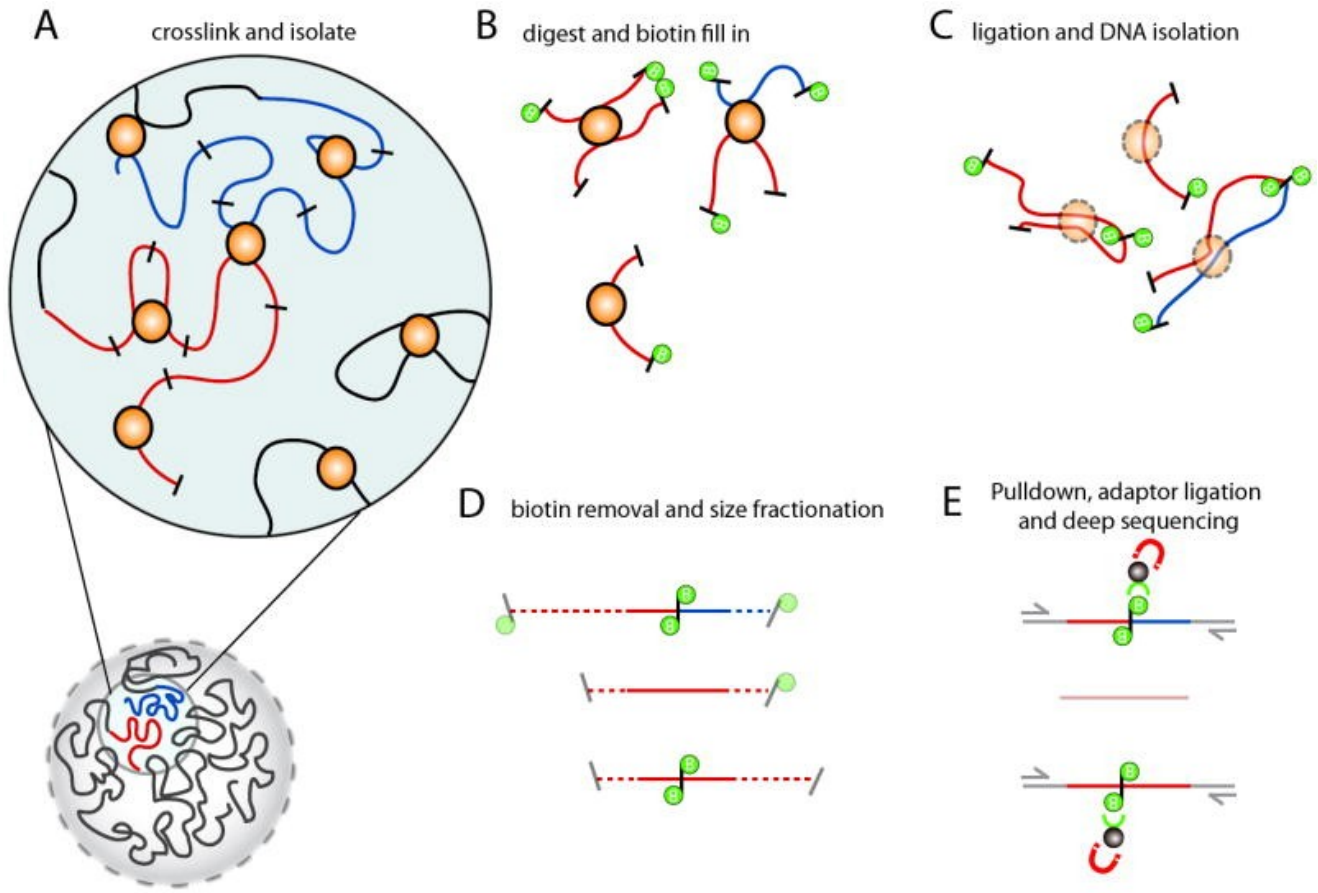
More structure = more paired = score > 0



3D arrange-ment of Chromo-somes

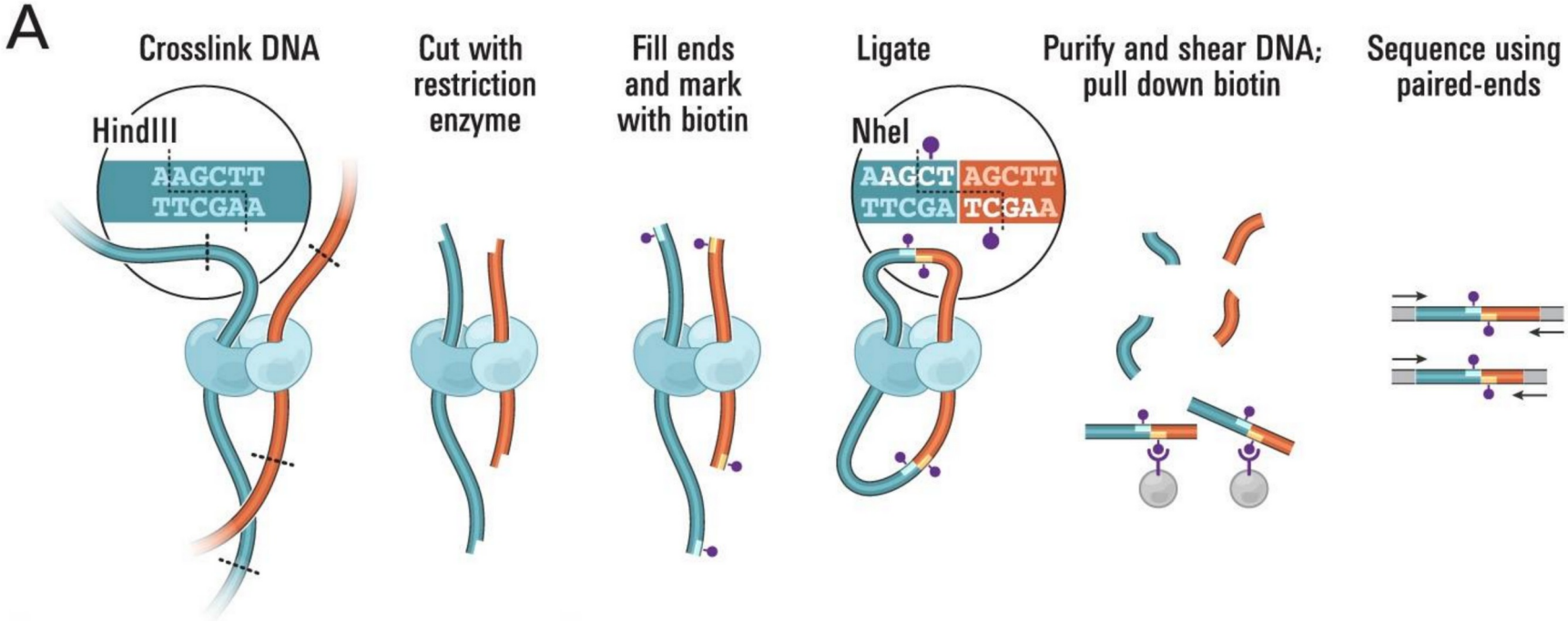


Overview of Hi-C technology

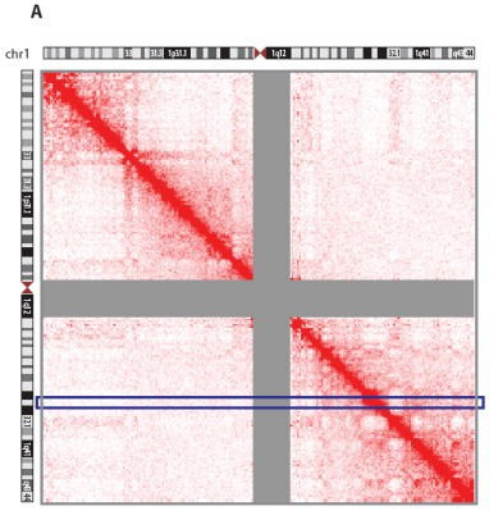


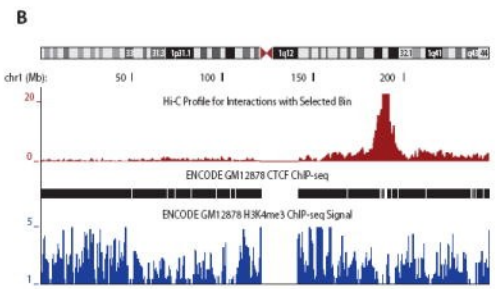
A) Hi-C detects chromatin interaction both within and between chromosomes by covalently crosslinking protein/DNA complexes with formaldehyde. **B)** The chromatin is digested with a restriction enzyme and the ends are marked with a biotinylated nucleotide. **C)** The DNA in the crosslinked complexes 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.

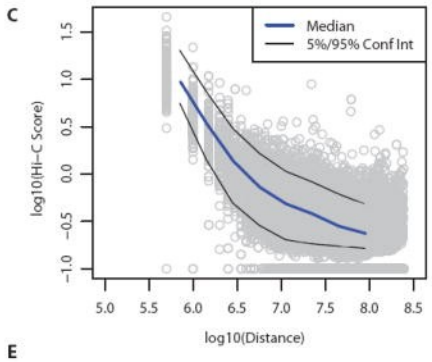
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3874846/

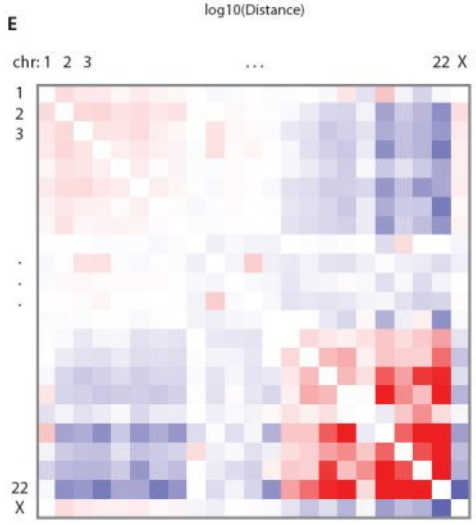

A
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

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 log10 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).

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 PeakAnalyzer
cd ~/tools
wget  http://www.bioinformatics.org/ftp/pub/peakanalyzer/PeakAnalyzer_1.4.tar.gz
tar xzf PeakAnalyzer_1.4.tar.gz

#install MEME
cd ~/tools
wget http://meme-suite.org/meme-software/5.0.2/meme-5.0.2.tar.gz
tar xzf meme-5.0.2.tar.gz
cd meme-5.0.2
./configure --prefix=$HOME/meme --with-url=http://meme-suite.org --enable-build-libxml2 --enable-build-libxslt
make install

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

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/
evince 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 -o - - > gfp.bam
samtools index gfp.bam

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

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

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
firefox meme/meme.html

Choose your paper for presentation at: https://tinyurl.com/52u2rv5a (contains URLs to papers)

	Paper	source/year
1	A method for multiple-sequence-alignment-free protein structure prediction using a protein language model	Nature Machine Intelligence 2023
2	A self-supervised deep learning method for data-efficient training in genomics	Communications Biology 2023
3	Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models	bioRxiv 2023
4	The landscape of biomedical research	bioRxiv 2023
5	trRosettaRNA: automated prediction of RNA 3D structure with transformer network	Nature Comm. 2023
6	Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction	Nature Communications 2023
7	Large language models encode clinical knowledge	Nature 2023
8	Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement	Nature Comm. 2021
9	Geometric deep learning of RNA structure	Science 2021
10	Data-driven discovery of innate immunomodulators via machine learning-guided high throughput screening	Chemical Science 2023
11	A draft for the human PanGenome	Nature 2023
12	Artificial Intelligence for Autonomous Molecular Design: A Perspective	Molecules 2021
13	Antibody-Antigen Docking and Design via Hierarchical Equivariant Refinement	ICML 2022
14	NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning	Frontiers in Immunology 2022
15	Discriminating physiological from non-physiological interfaces in structures of protein complexes: a community-wide study	Proteomics 2023
16	End-to-end accurate and high-throughput modeling of antibody-antigen complexes	MLSB 2022
17	Predicting structures of large protein assemblies using combinatorial assembly algorithm and AlphaFold2	bioRxiv 2023
18	When will RNA get its AlphaFold moment?	NAR 2023
19	scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data	Nature Machine Intelligence 2022
20	Physics-informed machine learning	Nature Reviews Physics 2021
21	Accelerating science with human-aware artificial intelligence	Nature Human Behaviour 2023
22	Neural networks and the chomsky hierarchy	ICLR 2023
23	Unifying Large Language Models and Knowledge Graphs: A Roadmap	arXiv 2023
24	A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics	Proc of IEEE 2023
25	A Survey on Transformers in Reinforcement Learning	Machine Learning Research 2023
26	A Survey on Model Compression for Large Language Models	arXiv 2023
27	Generative Agents: Interactive Simulacra of Human Behavior	arXiv 2023
28	Deep learning of causal structures in high dimensions under data limitations	Nature Machine Intelligence 2023