Principles and Methods in Regulatory Genomics

Carl Herrmann

Master Medizininformatik - Hochschule Heilbronn

WS 2022/2023

Health Data Science Unit

Medizinische Fakultät Heidelberg

Our primary interests

- Understanding the mechanisms of **regulatory genomics** in development and disease (especially cancer)
  - neuroblastoma / glioblastoma
    epigenomics
    role of transcription factors

- **Methods development** for integration of omics datasets (especially single-cell)
  - molecular signature extraction
    single-cell multi-omics

- Integration of clinical and omics data using ML approaches

- Ashwini Sharma (postdoc)
- Carlos Ramirez (postdoc)
- Andres Quintero (PhD)
- Ana Luisa Costa (PhD)
- Daria Doncevic (PhD)
- Youcheng Zhang (PhD)
- Carl Herrmann

www.hdsu.org

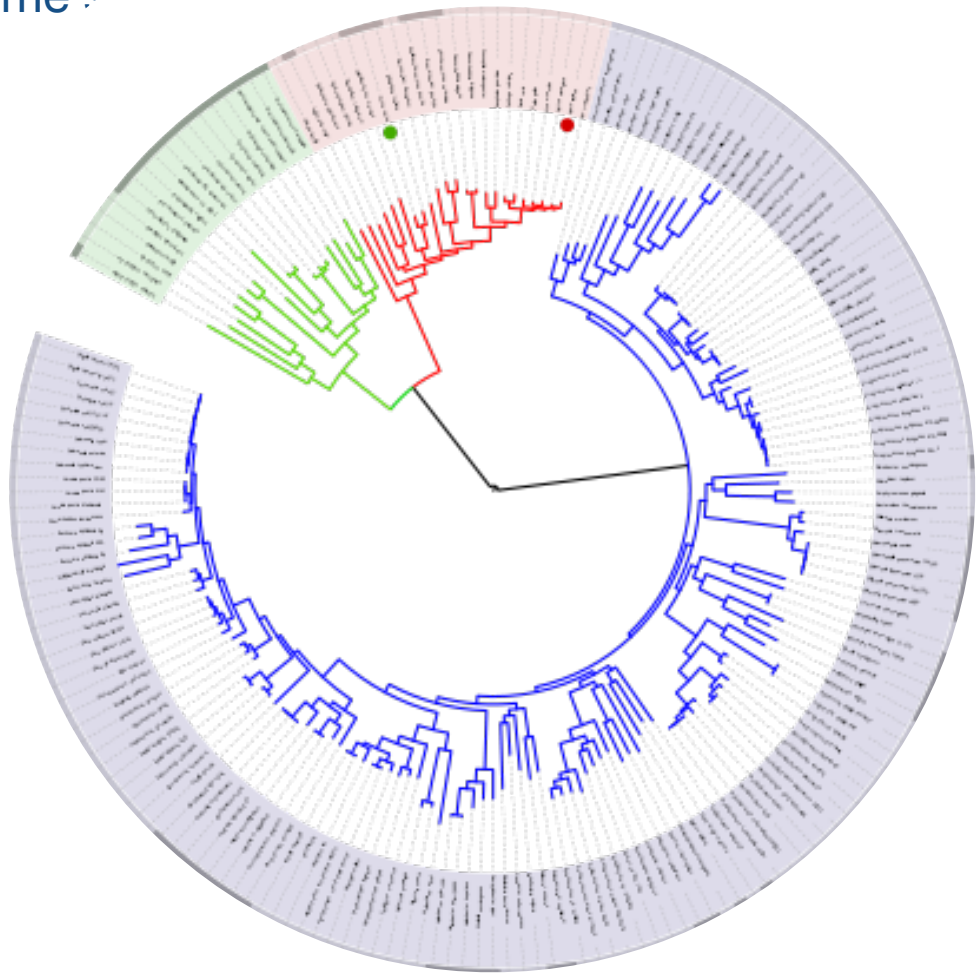
Content of the lecture

1. Introduction to regulatory genomics
2. Available data types
3. Transcription factors
4. Improving regulatory predictions
5. Integrative models
6. Chromatin networks (→ bayesian networks)
7. Conclusion

1. Introduction to regulatory principles

Bigger genome = more evolved ?

Genome s

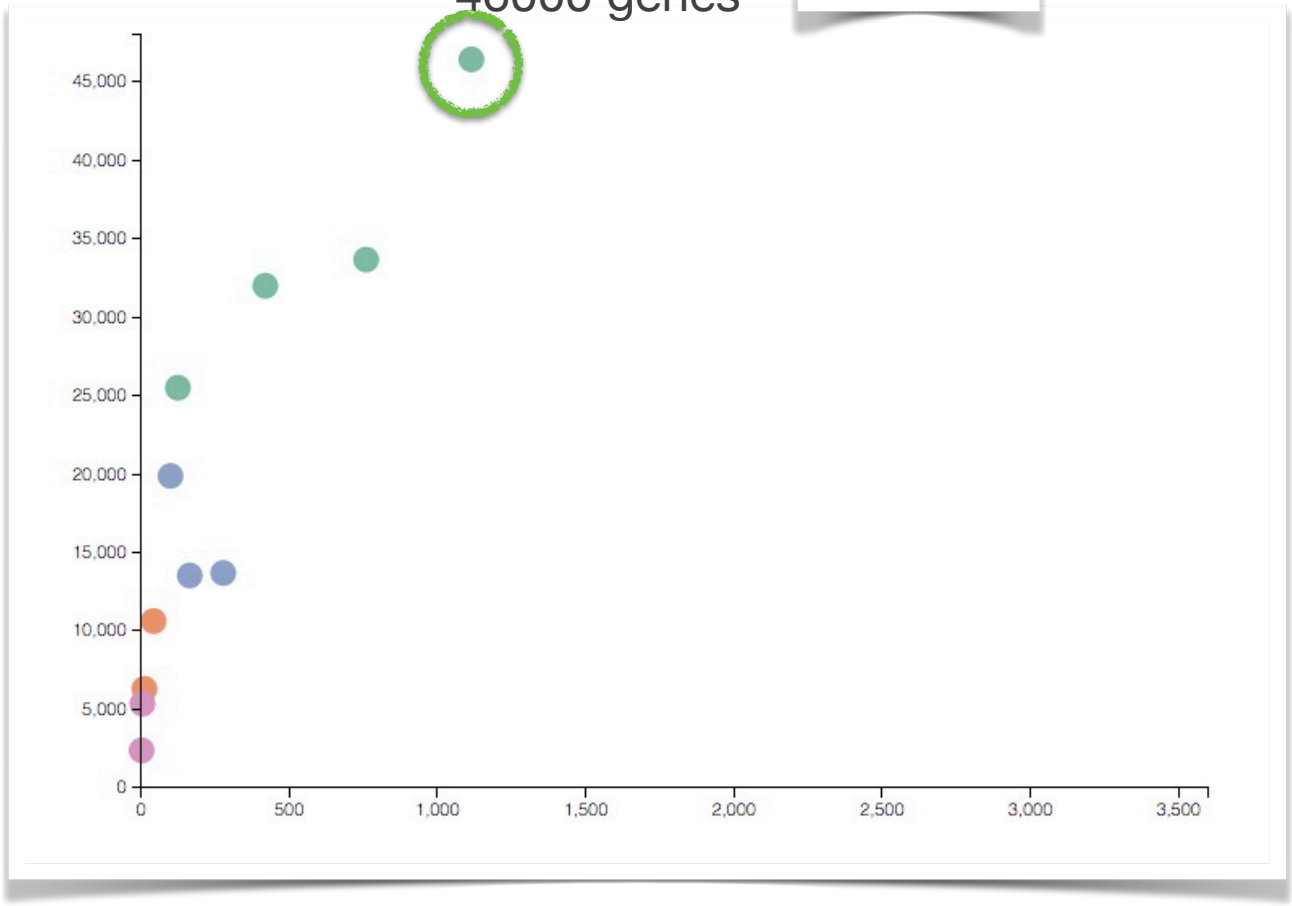


Eukaryotes
Archae
Bacteria

[interactive Tree of Life]

More genes = more complex ?

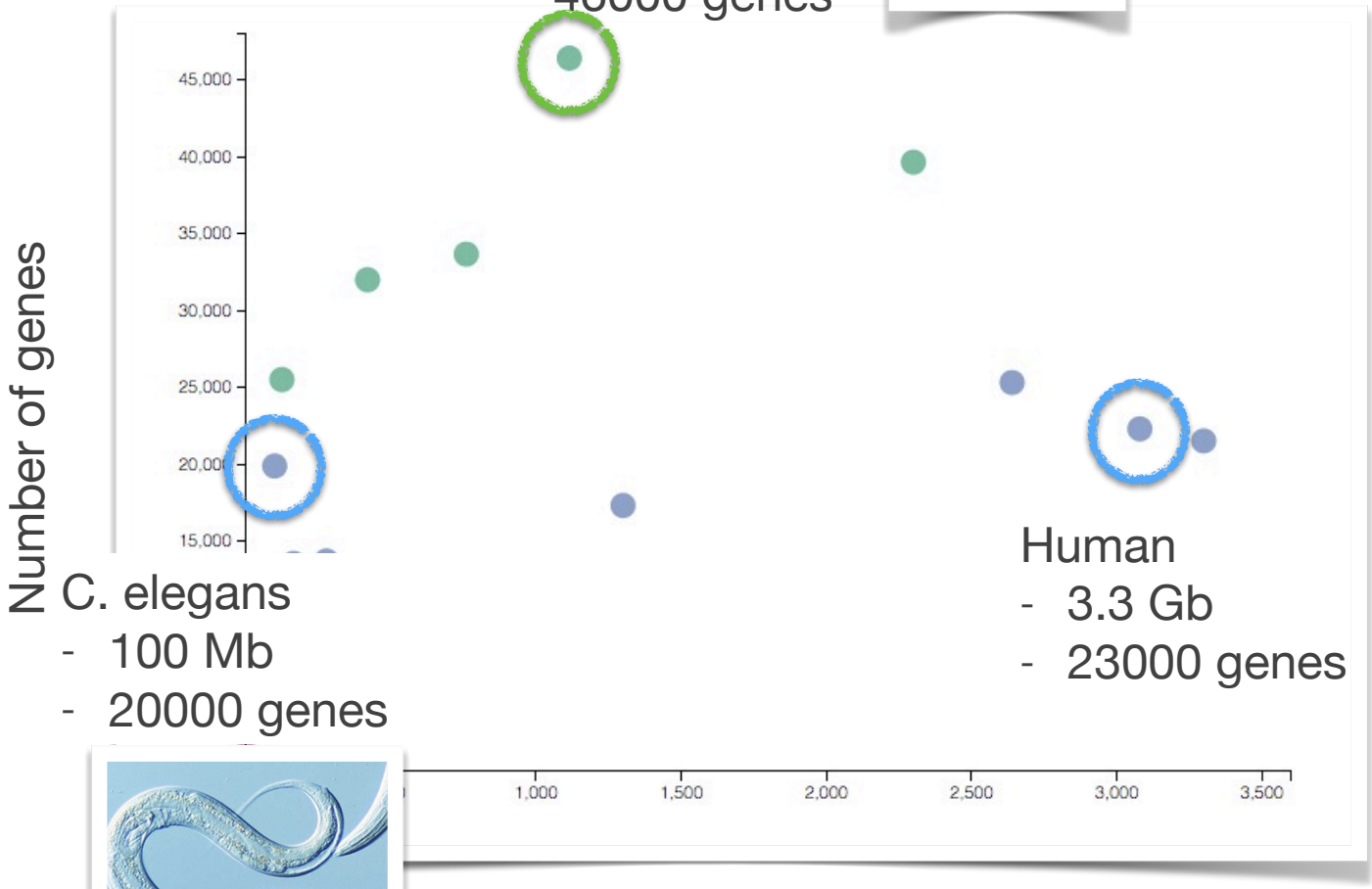
Glycine
- 1.1 Gb
- 46000 genes



Bacterium
Plant
Animal

More genes = more complex ?

Glycine
- 1.1 Gb
- 46000 genes



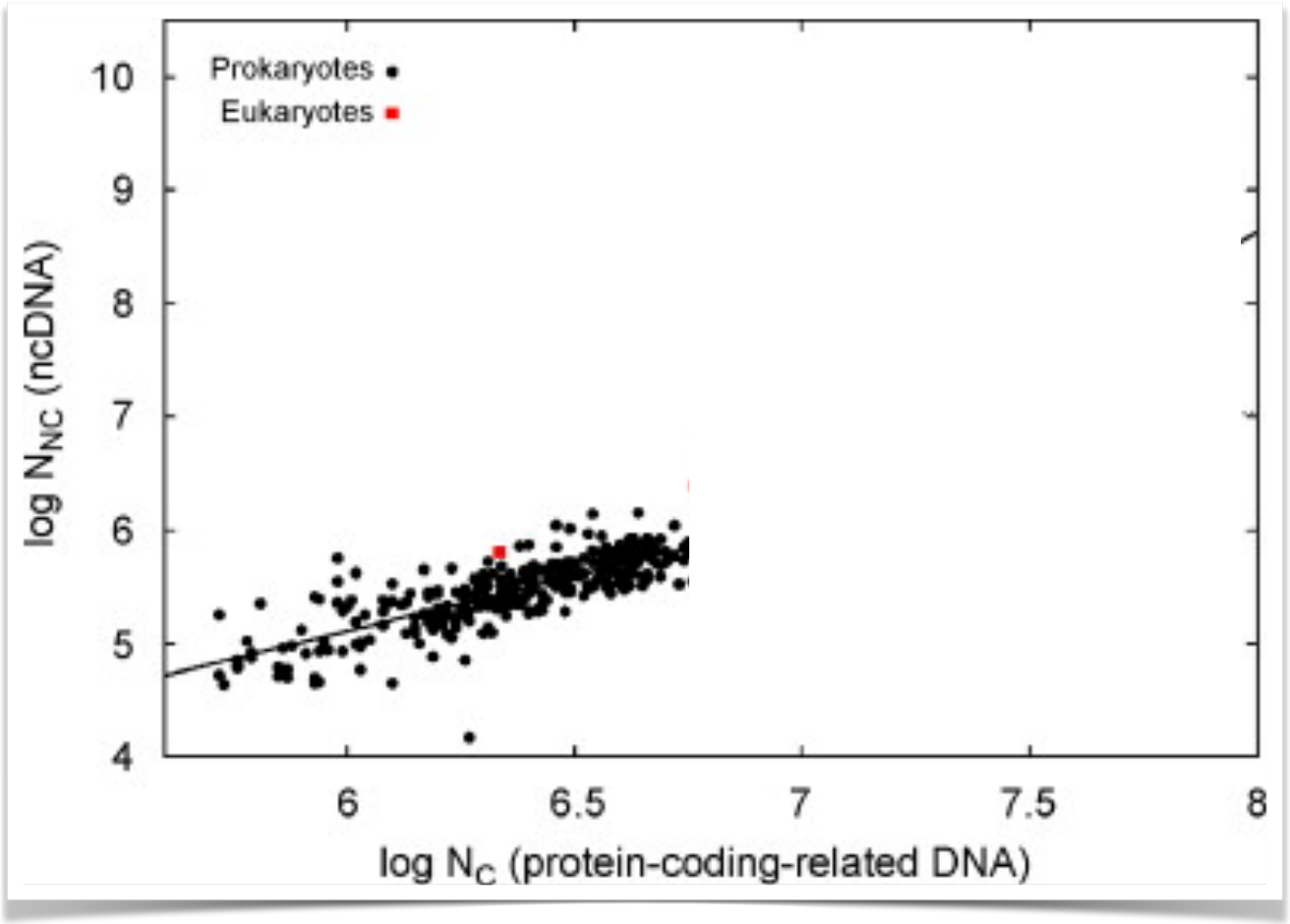
Bacterium
Plant
Animal

C. elegans
- 100 Mb
- 20000 genes

Human
- 3.3 Gb
- 23000 genes

[https://gf.neocities.org/gs/genes.html]

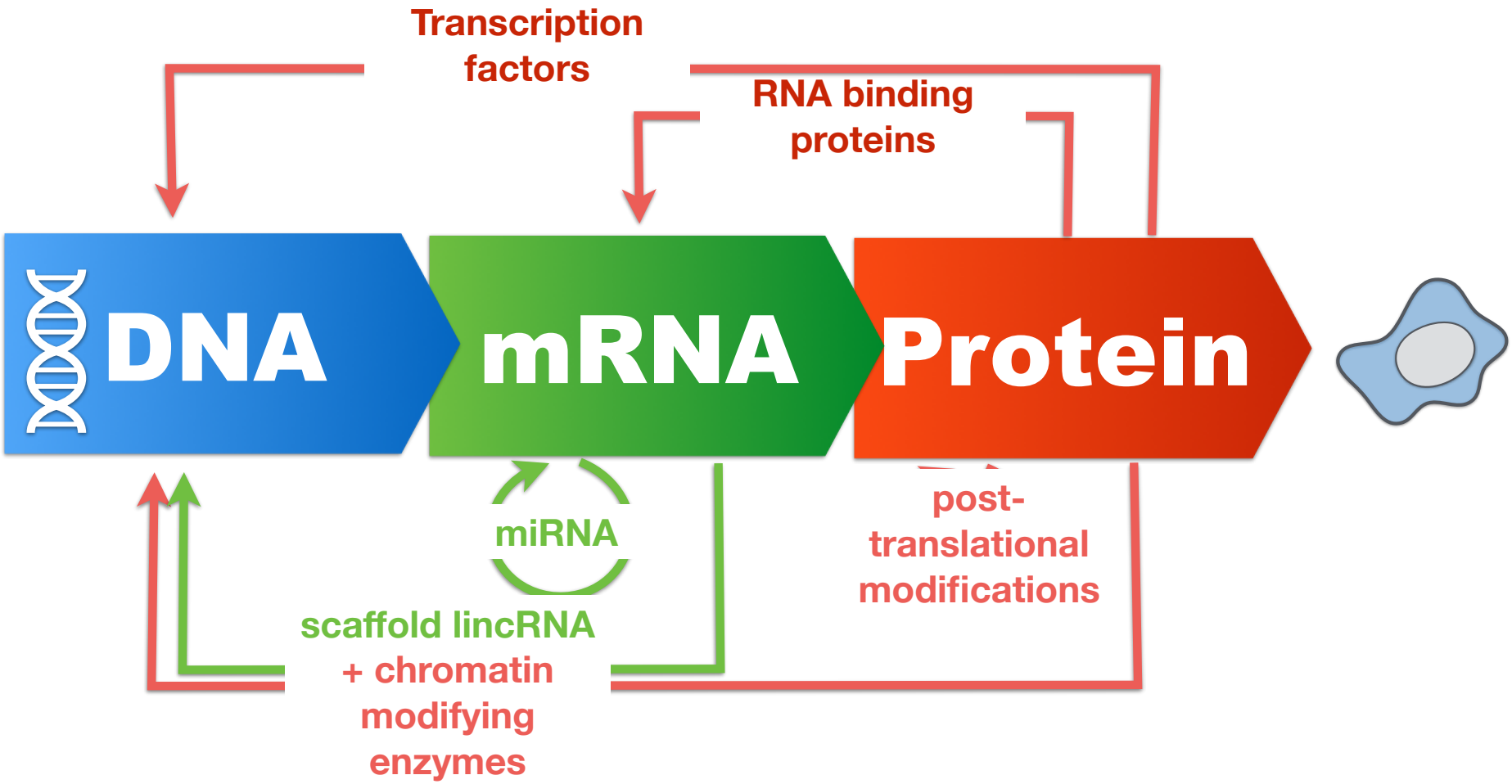
Bigger non-coding genome = higher complexity



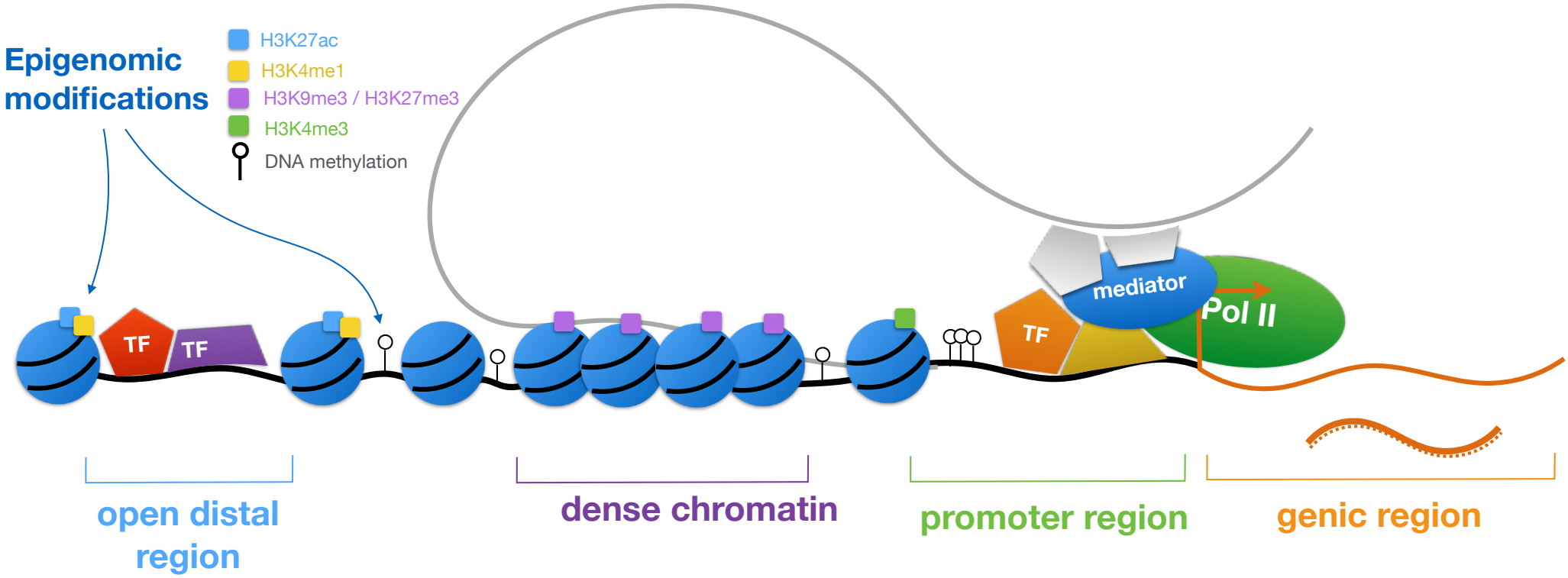
Proportion of **non-coding DNA** correlates with organismal complexity

[Ahnert, Fink, Zinovyev, 2008]

The Dogma in the genomics area



Transcriptional regulation

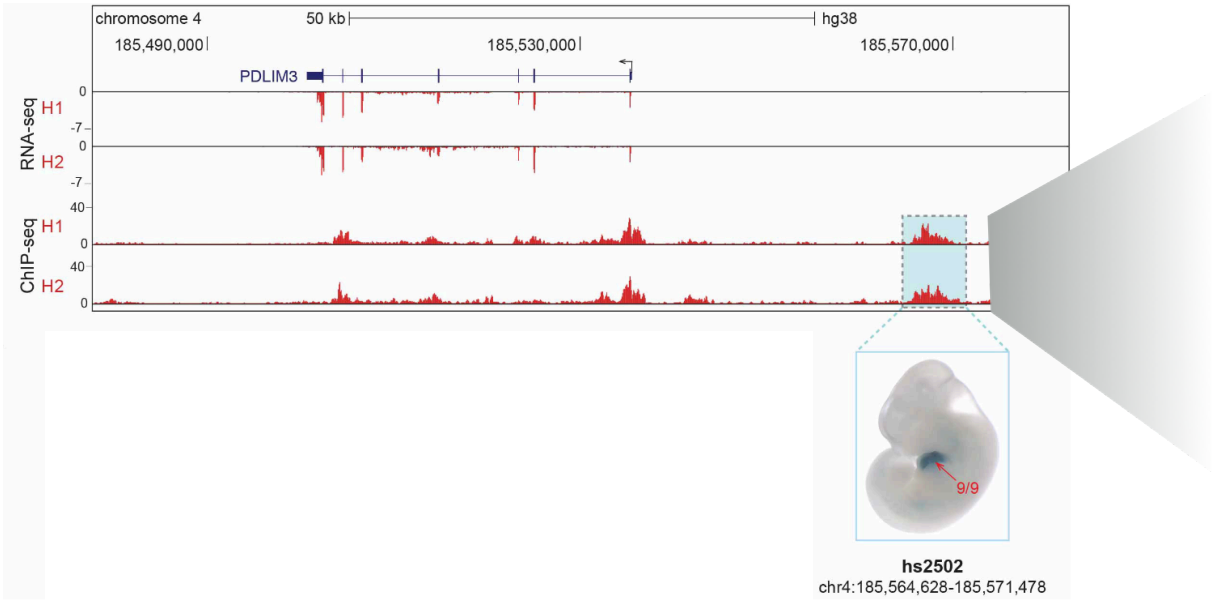


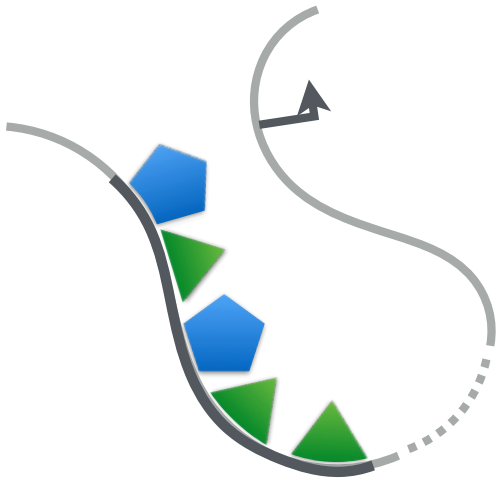
combinatorial interplay of multiple components

Transcriptional regulation

- **Enhancers** are regulatory elements which can be located far from the target genes
- **Multiple binding sites** for different transcription factors
- typical length: few kb → several hundred kb ("superenhancers")
- Organisational principles ("grammar") remains unclear (see exercises)

Identification of heart enhancers

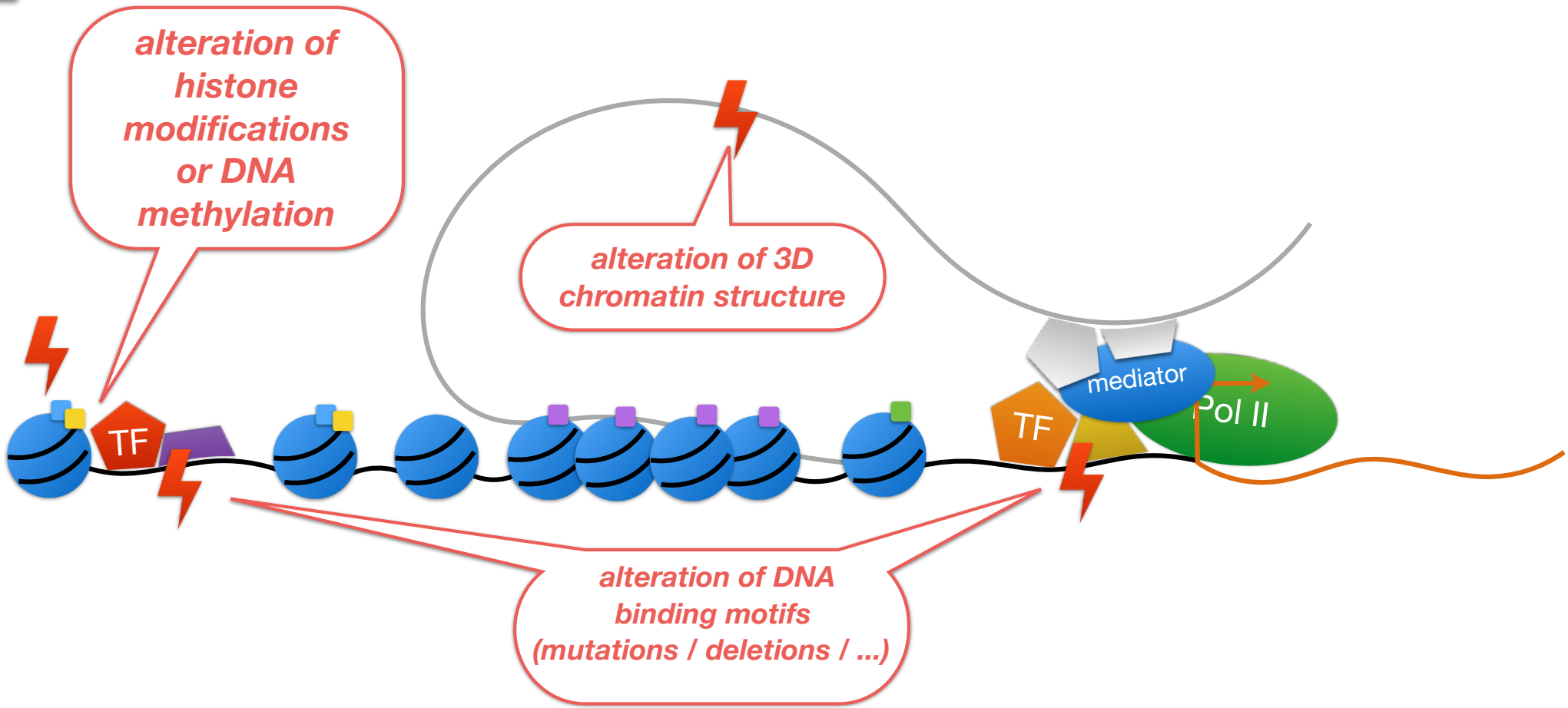




Enhancer region

[Spurrell et al., 2019]

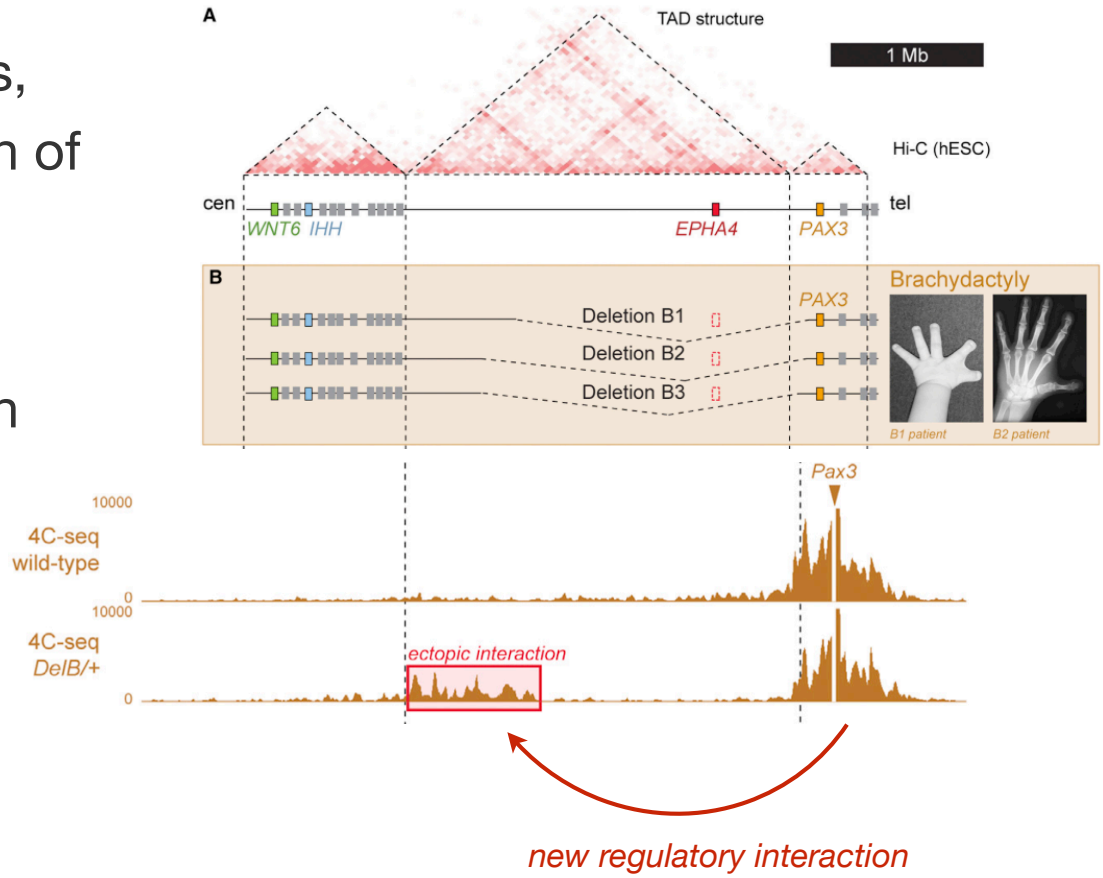
Transcriptional deregulation



complex interplay of multiple components
multiple sources of potential deregulation

Conformational deregulation

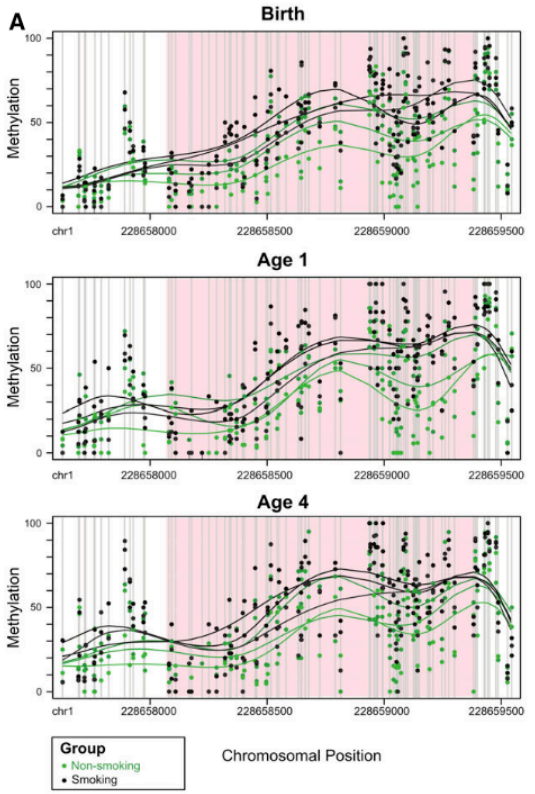
- Chromatin conformation defines **domains**, separated by **insulators**
- Genomic alterations (deletions, inversions...) lead to disruption of 3D conformation
  → ectopic gene activation
- "Enhancer hijacking" has been described in cancer

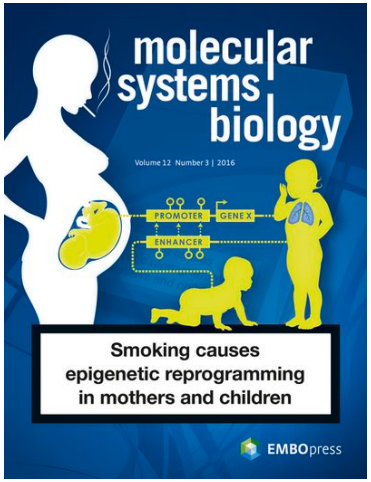


[Lupiañez,..., Mundlos, Cell (2015)]

Epigenetic deregulation

- **Epigenetic marks** (e.g. histone marks or DNA methylation) can encode external environmental cues
- Maternal smoking affects DNA methylation in children at regulatory sites (**differential methylated regions** DMR)
- These regions control **developmental genes** involved e.g. in lung development
  → higher susceptibility to lung diseases





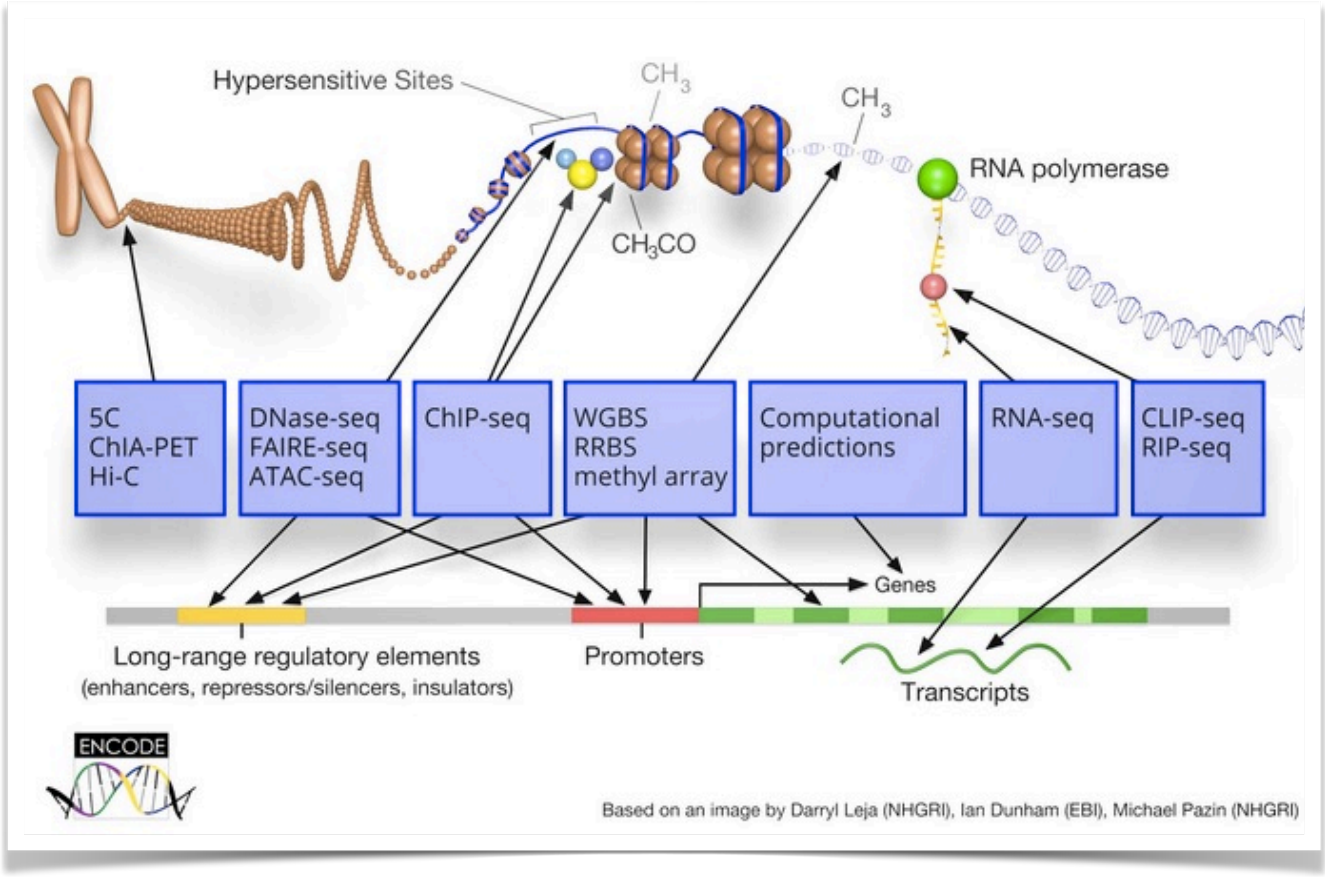
[Bauer et al., MSB (2016)]

2. Which data are available?

Exploring the genome's activity

- Large scale consortia (ENCODE, Roadmap, …) have systematically explored the **activity** of the genome using experimental assays

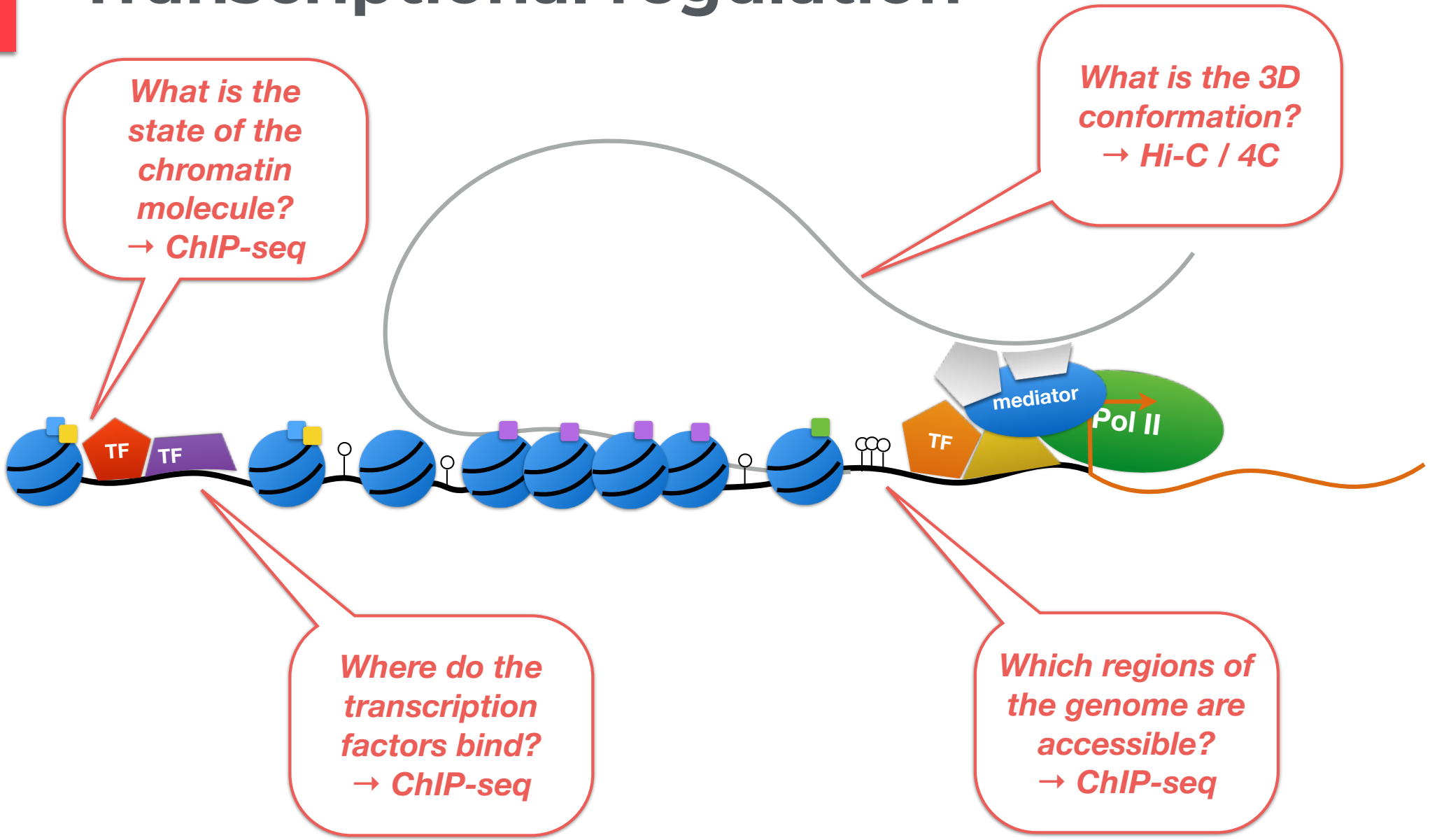
"***The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/ or chromatin-associated event in at least one cell type***.
99% is within 1.7kb of at least one of the biochemical events measured by ENCODE."



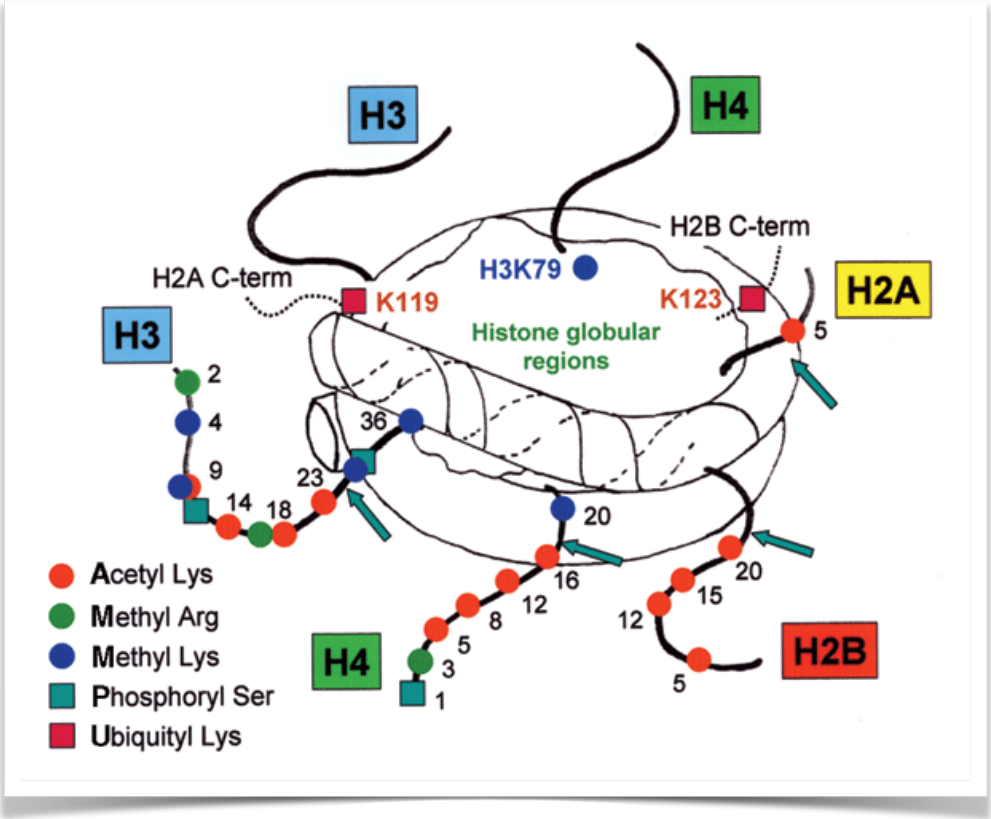
https://www.encodeproject.org/

https://www.encodeproject.org/matrix/?type=Experiment

Transcriptional regulation



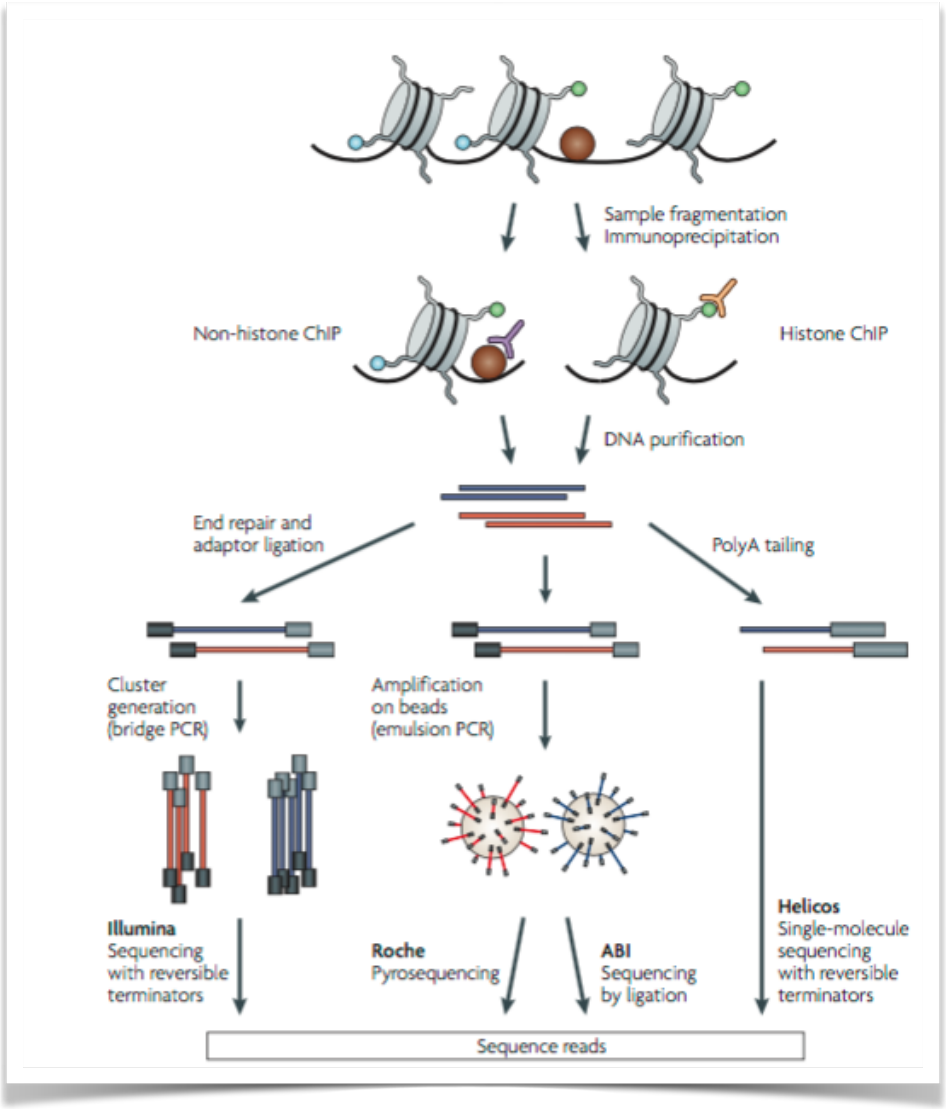
ChIP-seq for histone modifications



- histones are subject to **post-translational modifications** at their N-terminal tail
  - Lysine methylation
  - Lysine/arginine acetylation
  - Serine phosphorylation
  - ubiquitylation
- they **modify the physical properties of the DNA-nucleosome interactions**

nomenclature: H3K27ac = acetylation of lysine 27 on histone 3

Chromatin Immunoprecipitations

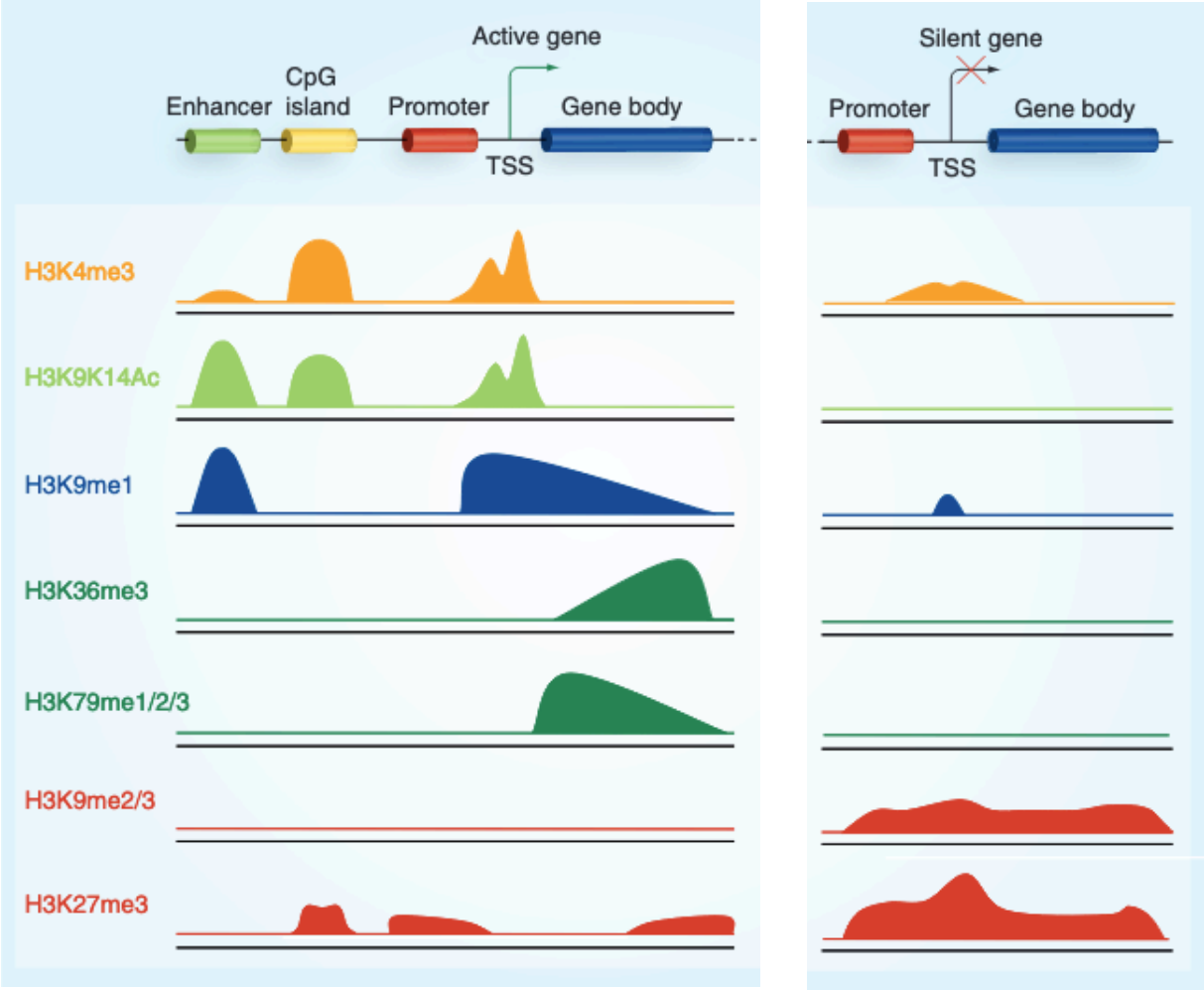


[Park, Nat.Rev. 2009]

- Chromatin immunoprecipitation (ChIP) yields **DNA fragments**, that are
  - bound by the protein of interest
  - marked by a specific chemical modification (acetylation, methylation,.)
- Identification of the fragments :
  - sequencing (ChIP-seq)
    → genome-wide
  - PCR/qPCR
    → targeted experiment
- Important aspect
  - Quality/Specificity of the antibody ?
  - DNA fragment (~200-300bp)
    → binding site (~10 bp) ?

Histone modifications

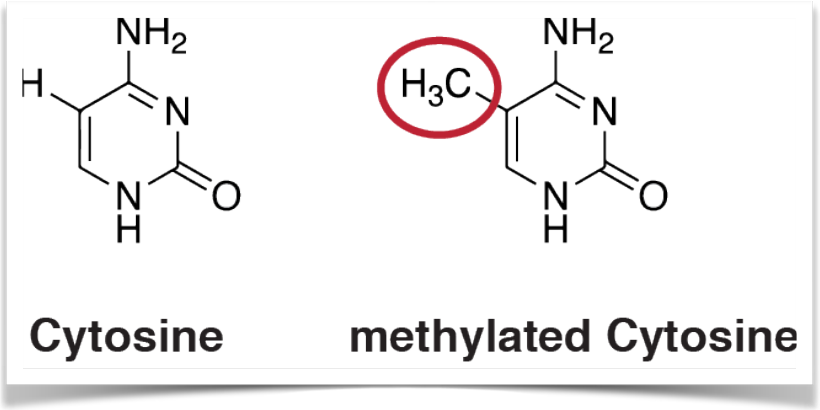
histone modifications are a good proxy of gene expression and presence of regulatory elements



[Lin, Shannon, Hardy, 2010]

Measuring DNA methylation

- DNA methylation occurs mainly on **cytosines in CpG** dinucleotides in the human genome (28 million in human genome!)
- DNA methylation is revealed by using **bisulfite conversion** (HSO_3^-):
  - unmethylated cytosines are converted
    C → U → T
  - methylated cytosines are protected
    mC → mC

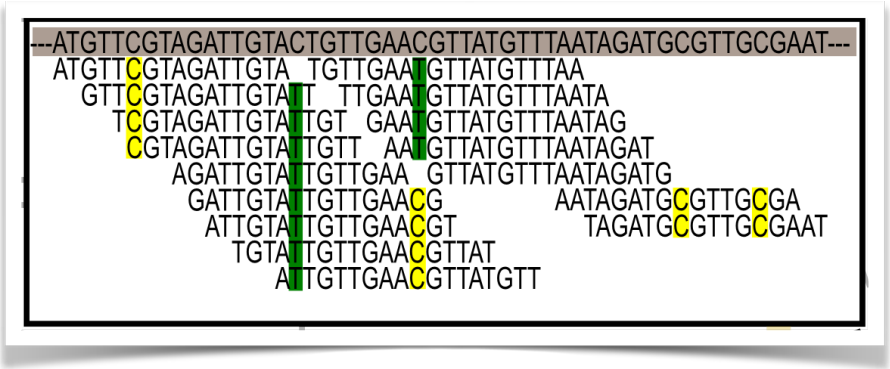


- unmethylated CpG are identified by the presence of a **mismatch TpG**
- 2 approaches:
  - array based: hybridization to CpG probes on array
  - sequencing: whole genome bisulfite-sequencing

Measuring DNA methylation

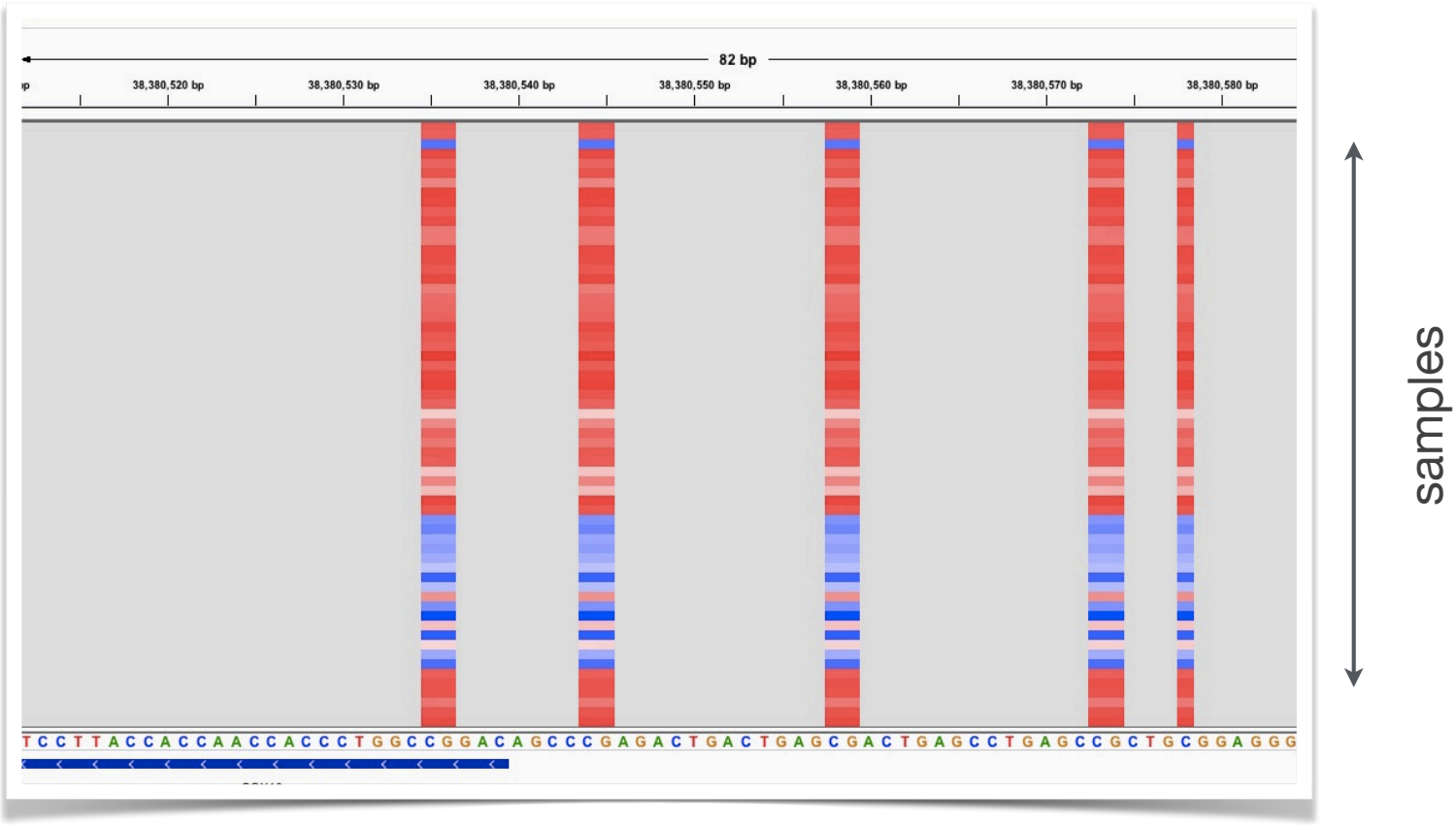
- **Array based methods**
- CpG containing probes on array
  - 27K probes
  - 450K probes
  - 800K (EPIC)
- all probes contain a methylated (C) and unmethylated (T) version
- Cheap but sparse

- **Sequencing base methods** (whole-genome bisulfite sequencing WGBS)
  - unmethylated C → T
  - methylated C → C
- Shearing, conversion and sequencing (Illumina X-10)
- Information about the 28 million CpGs



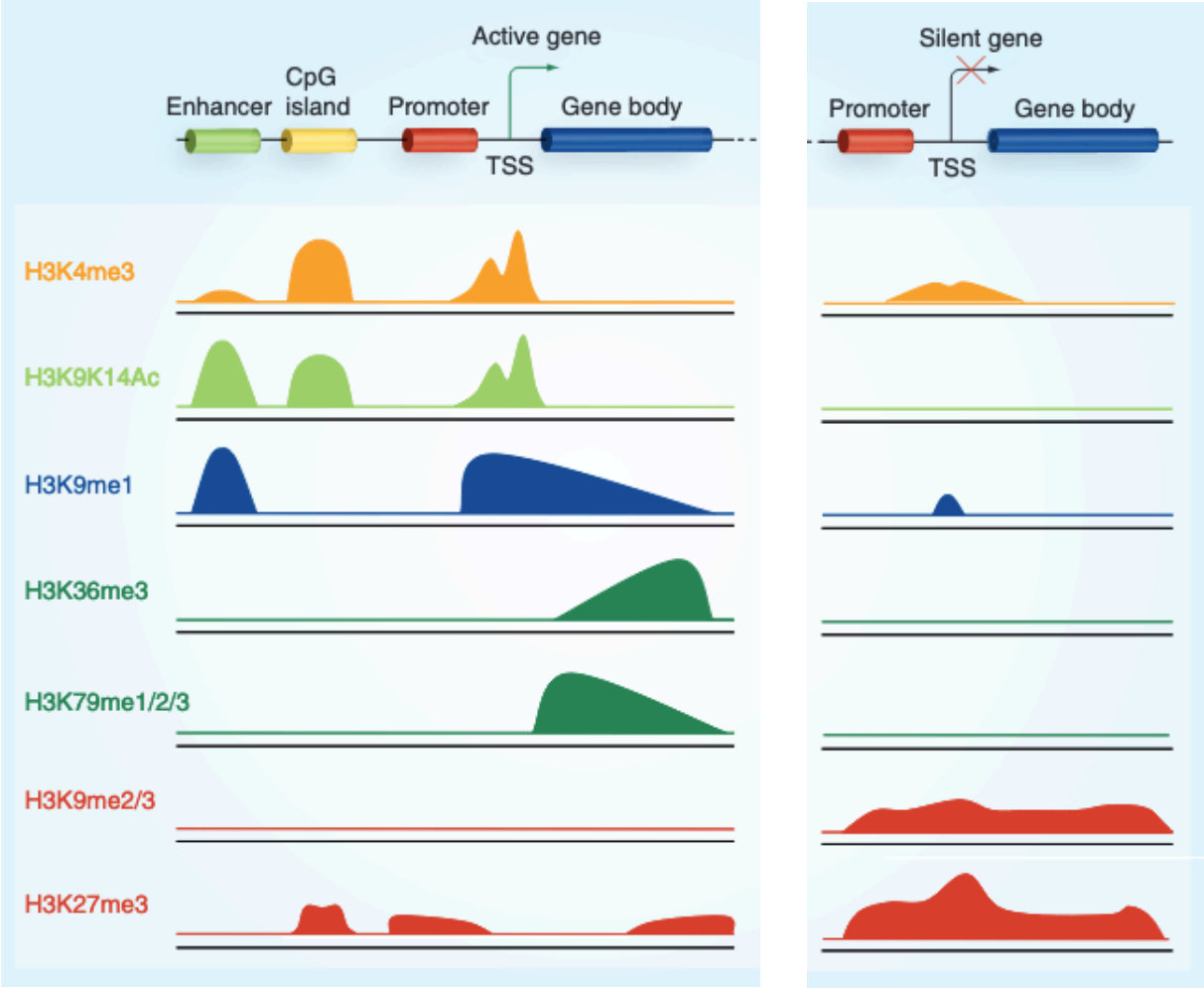
Example DNA methylation

- Whole genome bisulfite sequencing provide information about all CpGs in the genome
- Vertical bars = CpG positions; red = high methylation (100%); blue = no methylation (~10%)



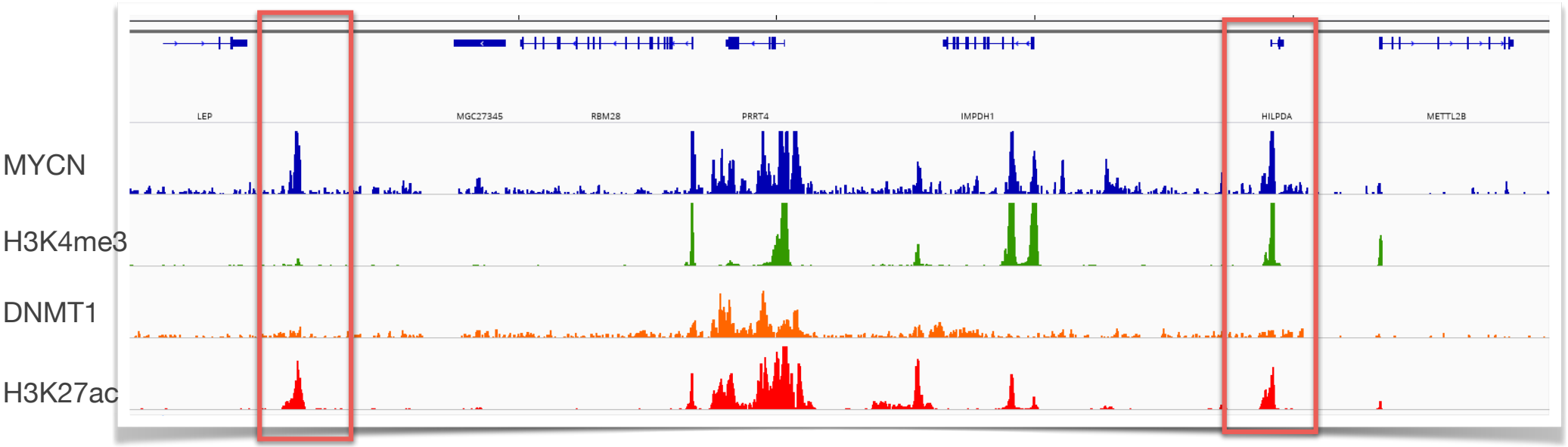
Histone modifications

histone modifications are a good proxy of gene expression and presence of regulatory elements



[Lin, Shannon, Hardy, 2010]

Example of ChIP-seq signal for transcription factors / DNA-binding proteins

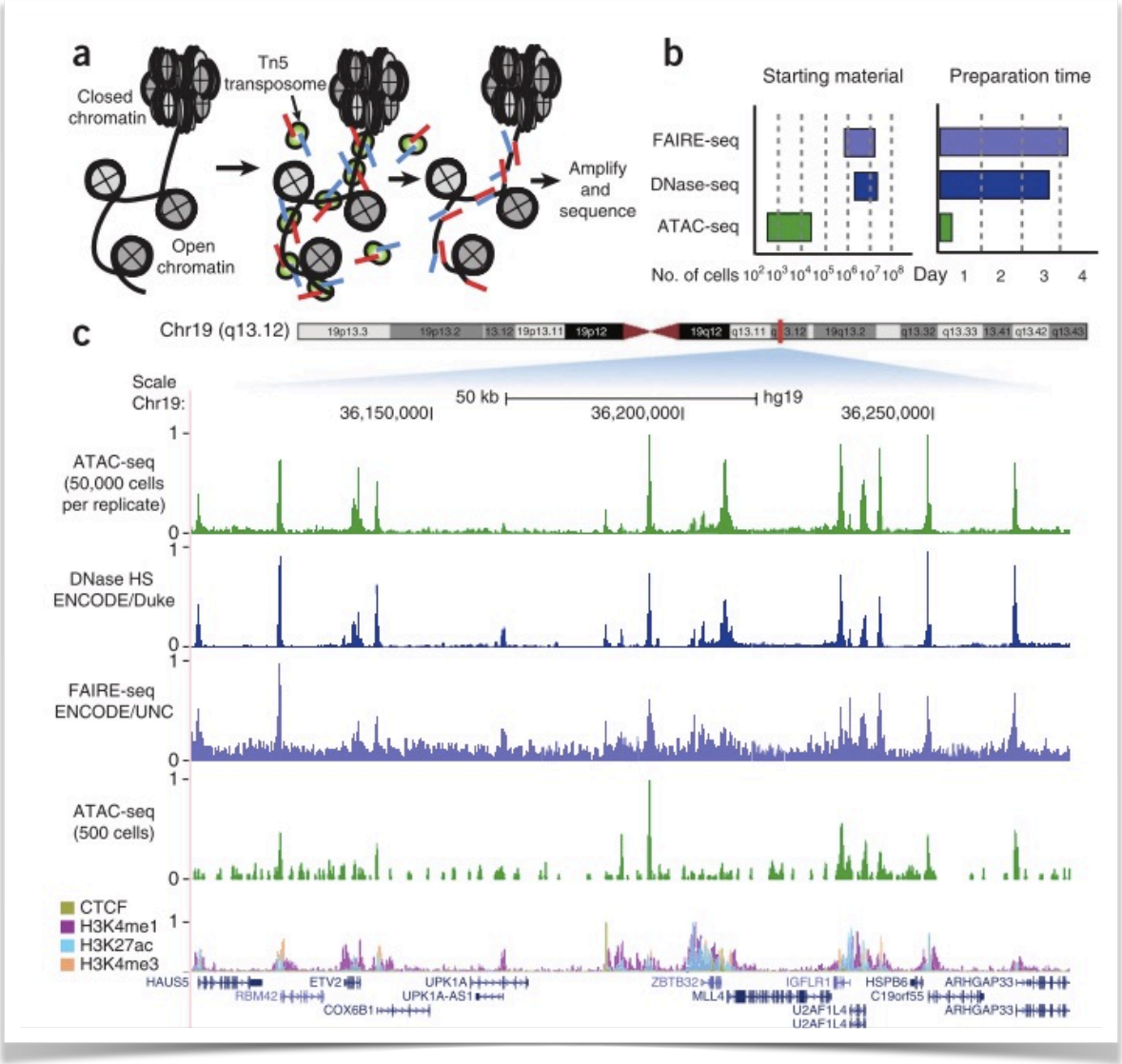


distal MYCN peak (away from gene promoters)

MYCN peak and H3K4me3 enrichment at promoter

Chromatin accessibility

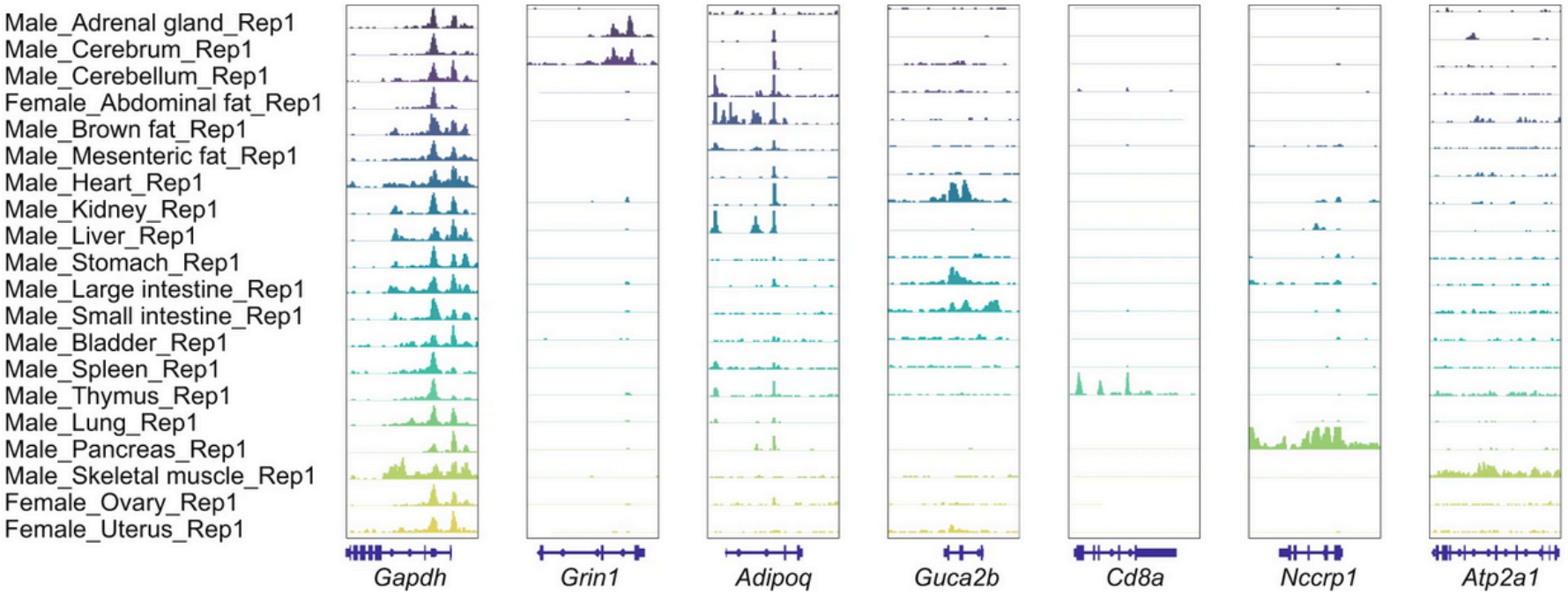
- **ATAC-seq**: using Tn5 transposase prepared with sequencing primers
- requires a small number of input material (~10,000 cells)
- easily adapter to single-cell sequencing
- identification of open chromatin regions (peaks)



[Greenleaf (2013)]

Accessibility atlas

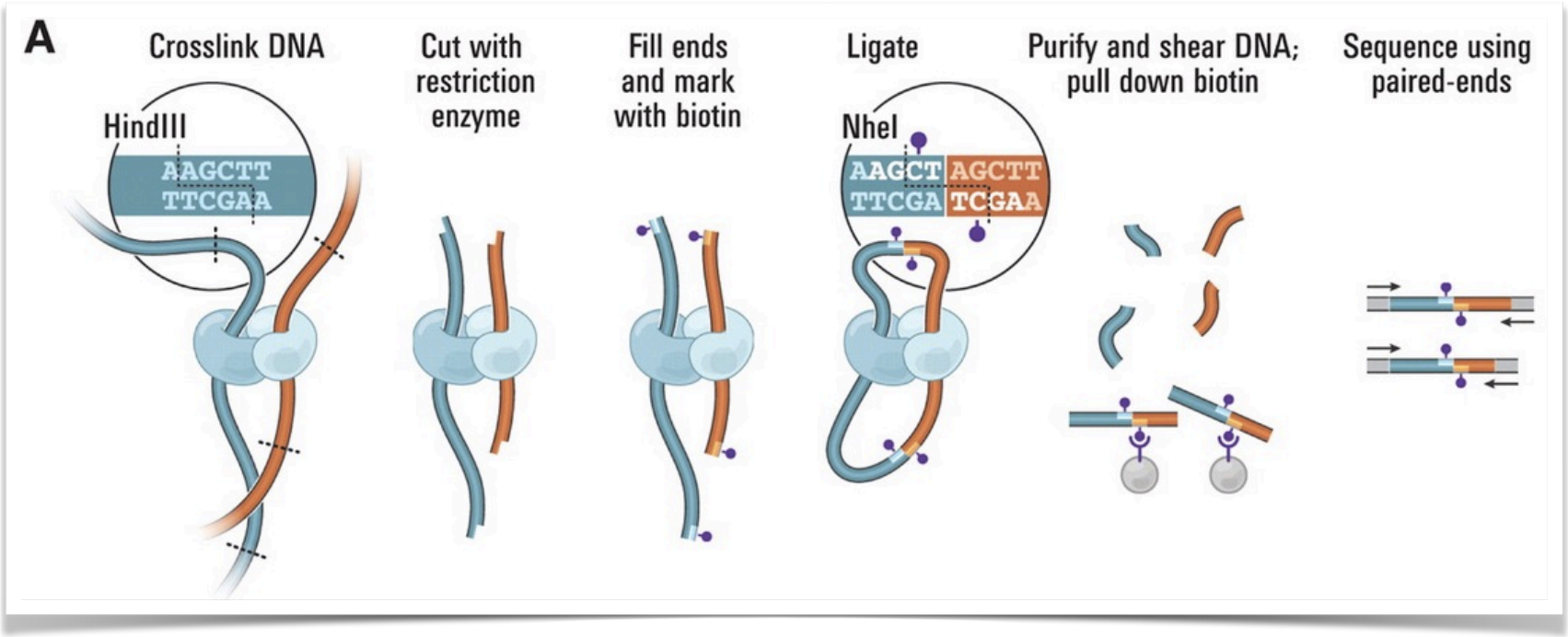
- Patterns of chromatin accessibility are **cell-type specific**



[Liu et al., Scientific Data (2019)]

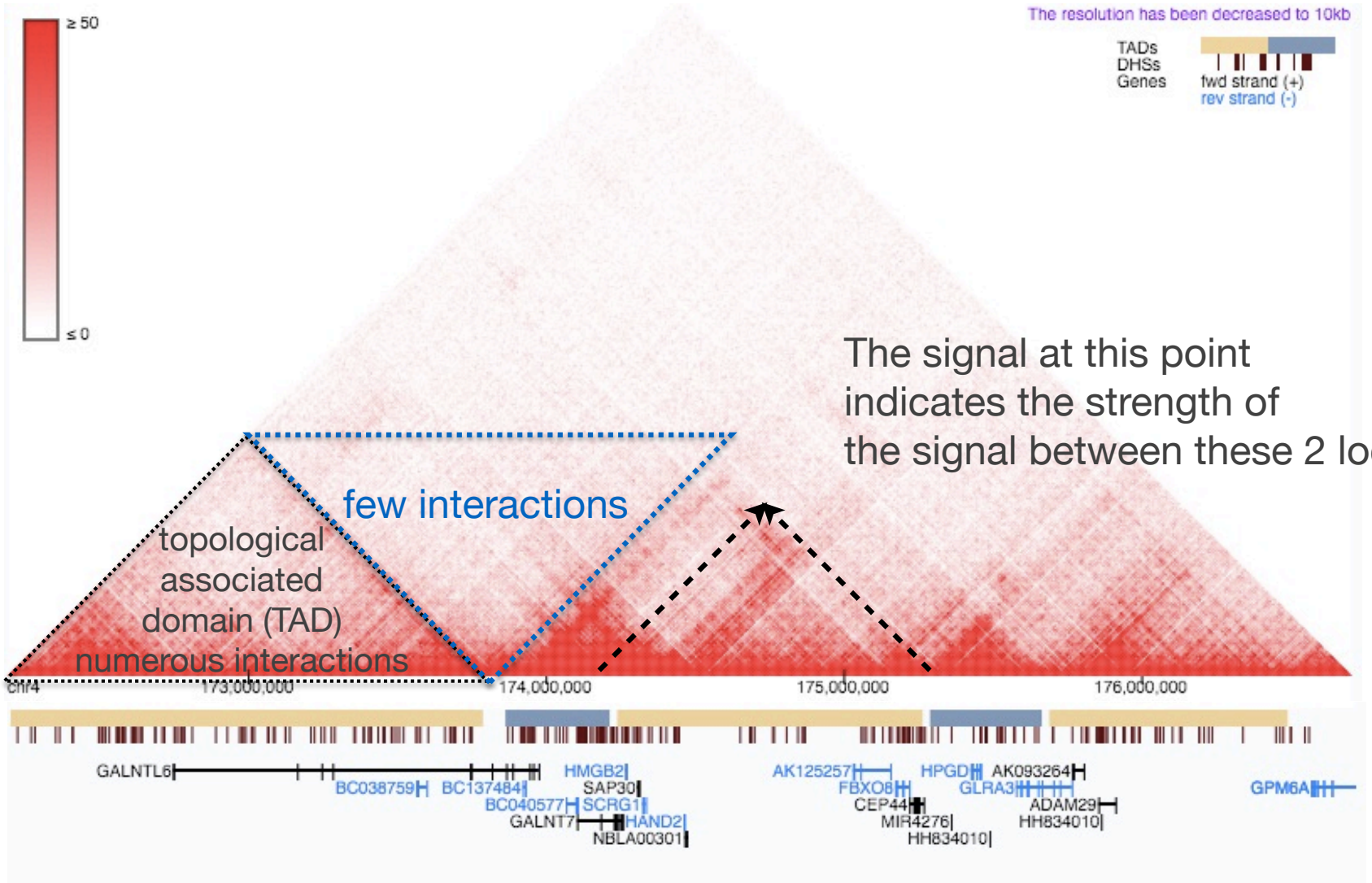
Mapping chromatin interactions

- DNA looping allows interactions between distal DNA loci
- Identification of interacting regions through **“chromatin conformation capture”** methods (3C / 4C / Hi-C)



[Liebermann-Aiden, 2009]

Hi-C and topological domains



The signal at this point indicates the strength of the signal between these 2 loci

Topological associated domains

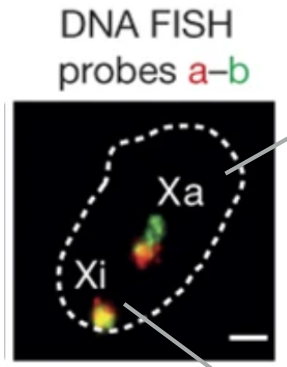
[http://promoter.bx.psu.edu/hi-c/view.php]

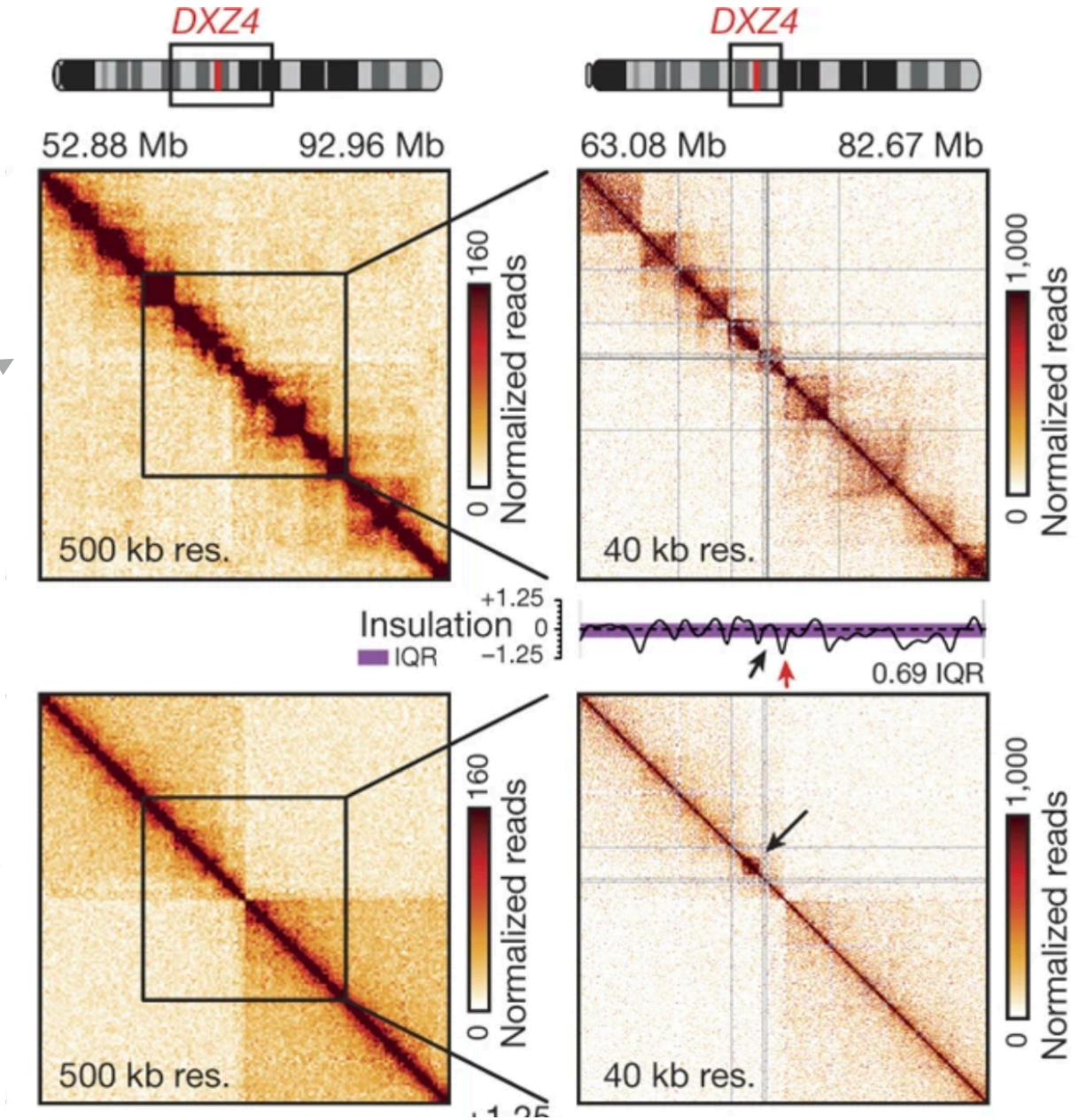
[Dixon (2012,2015)]

Chromatin organization and cell state

Allele specific Hi-C
in neural progenitor cells

Active/inactive X allele

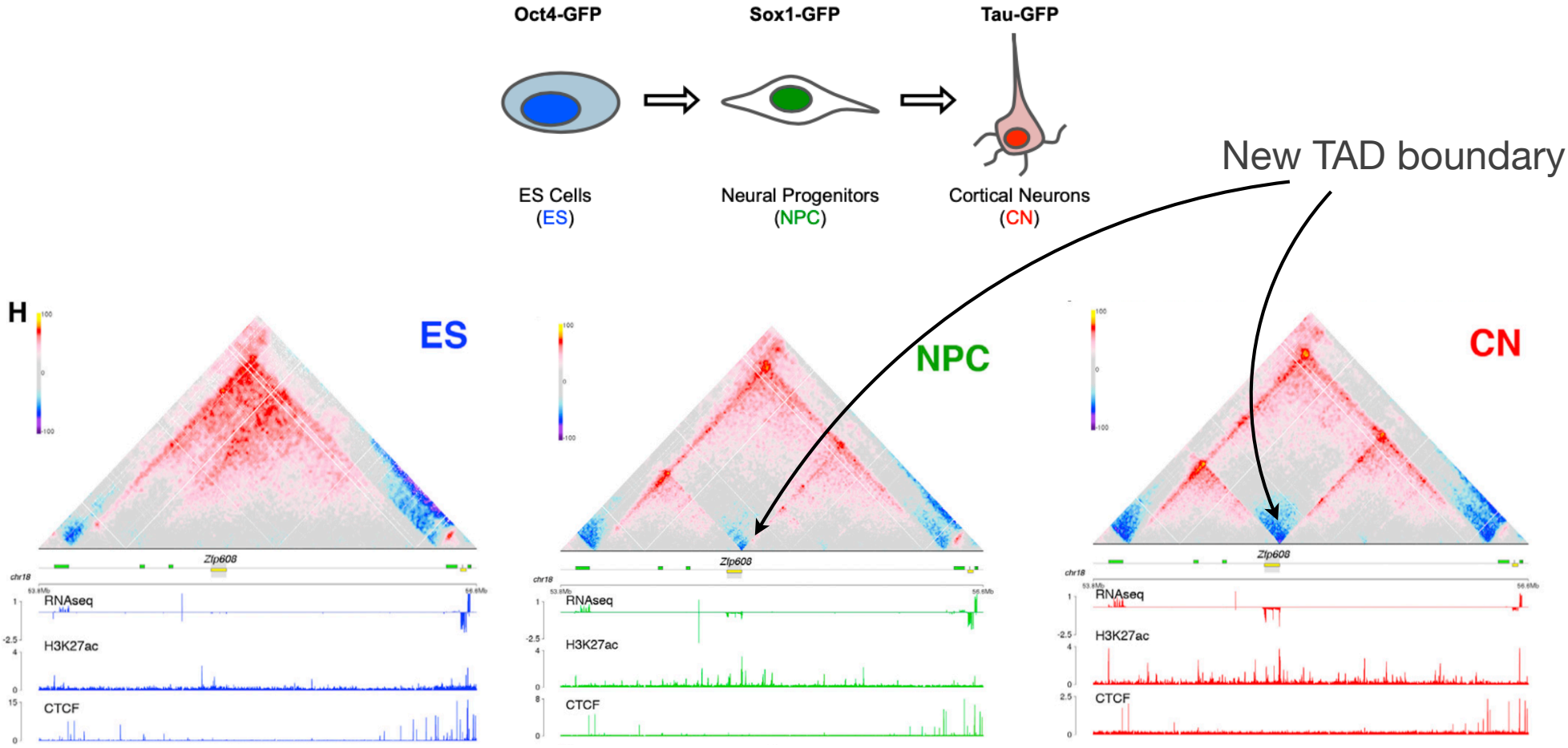




[Georgetti et al., Nature 2016]

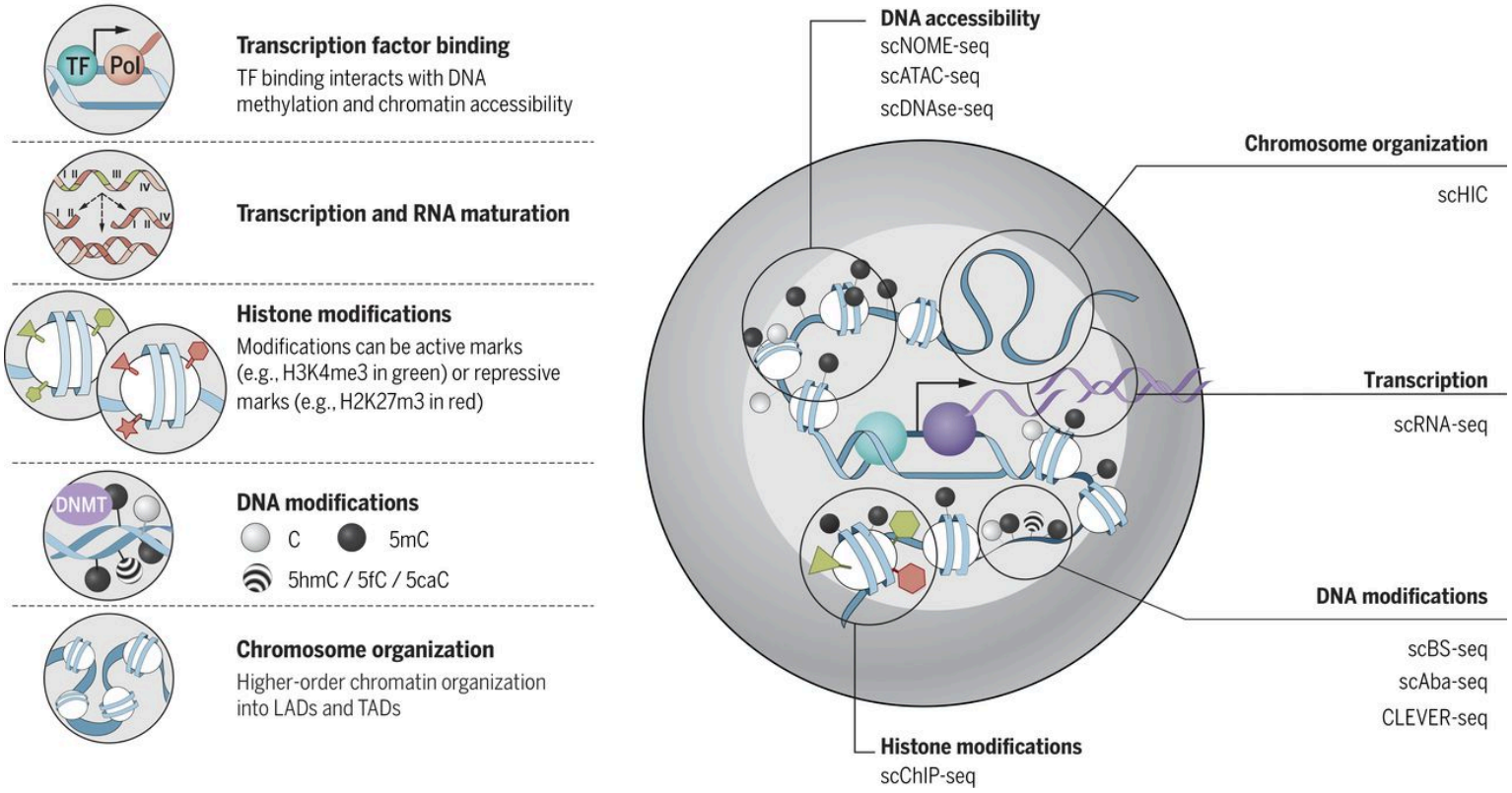
Chromatin organization and differentiation

- Changes in chromatin conformation occur during cell differentiation (e.g. neural development)



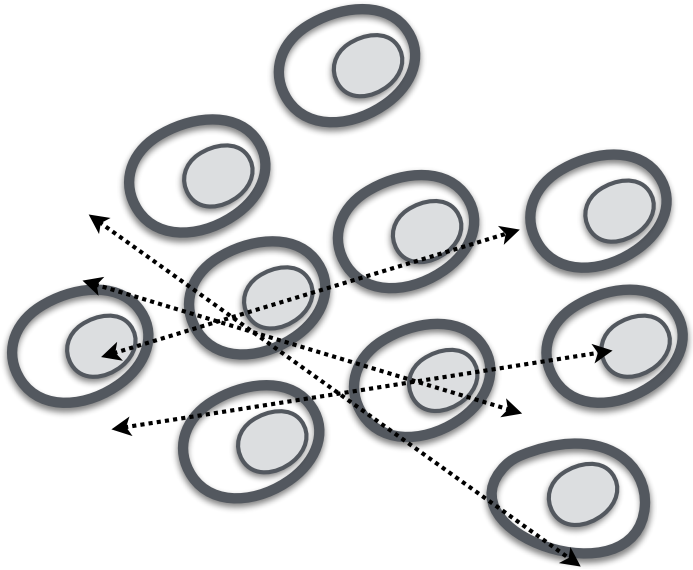
[Bonev et al., Cell (2017)]

Single-cell regulatory genomics



[Kelsey et al., Science (2017)]

Single-cell multi-omics

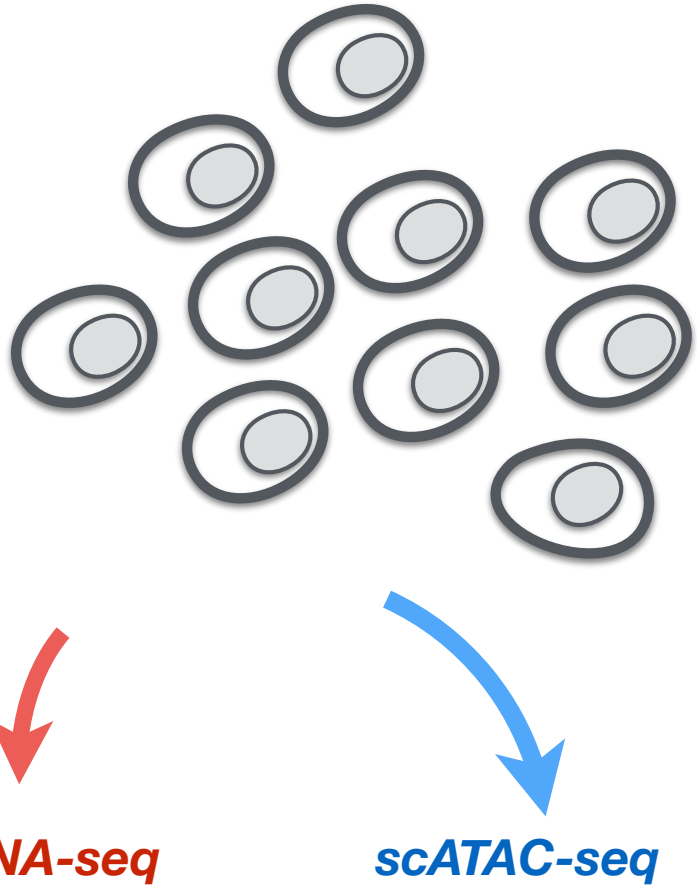


scRNA-seq

transfer learning
find best mapping between cells from different assays

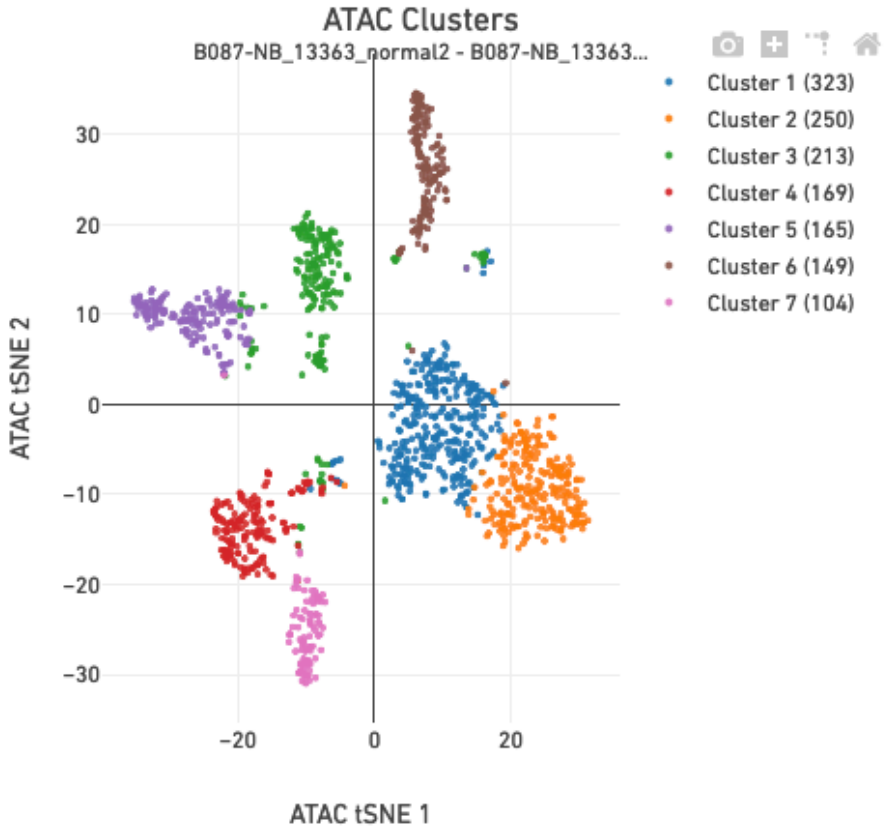
scATAC-seq

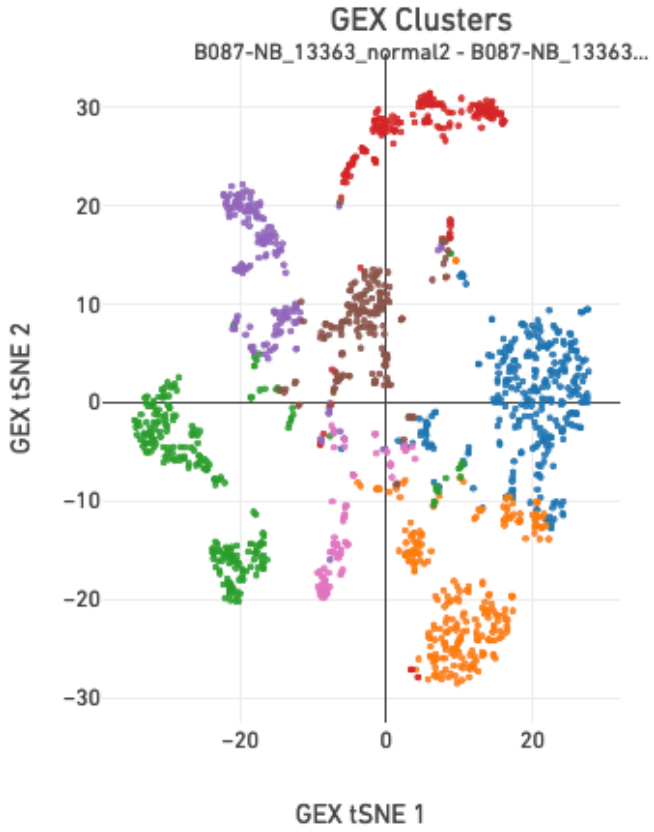
Single-cell multi-omics



Accessibility & expression
[Cao et al. 2018]
[Clark et al. 2017]
[scCAT (Liu et al. 2019)]

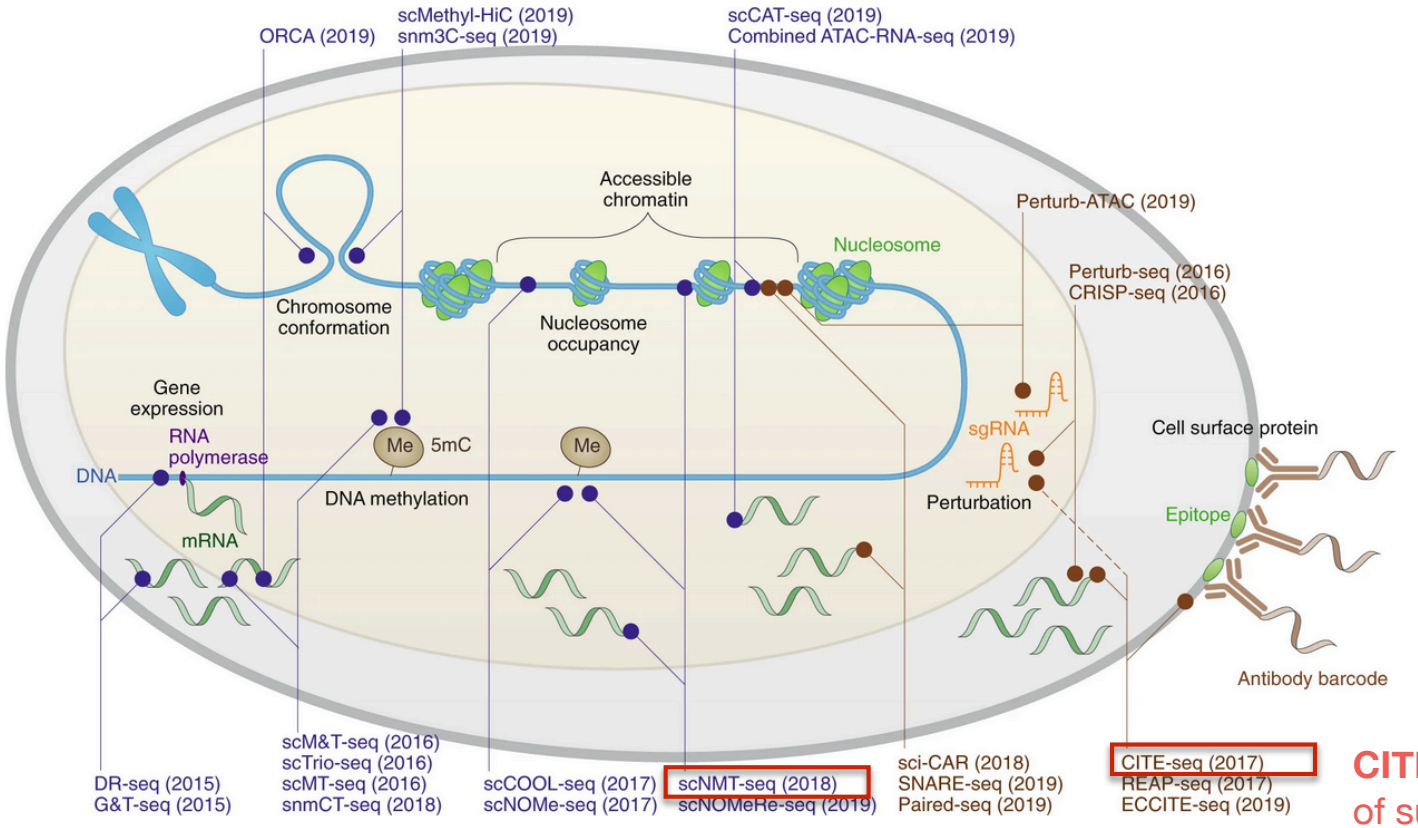
single-cell Multiome: ATAC / Expression





cluster structure is slightly different between scATAC and scRNA!

Single-cell multi-omics



CITE-seq: identification of surface proteins + scRNA-seq

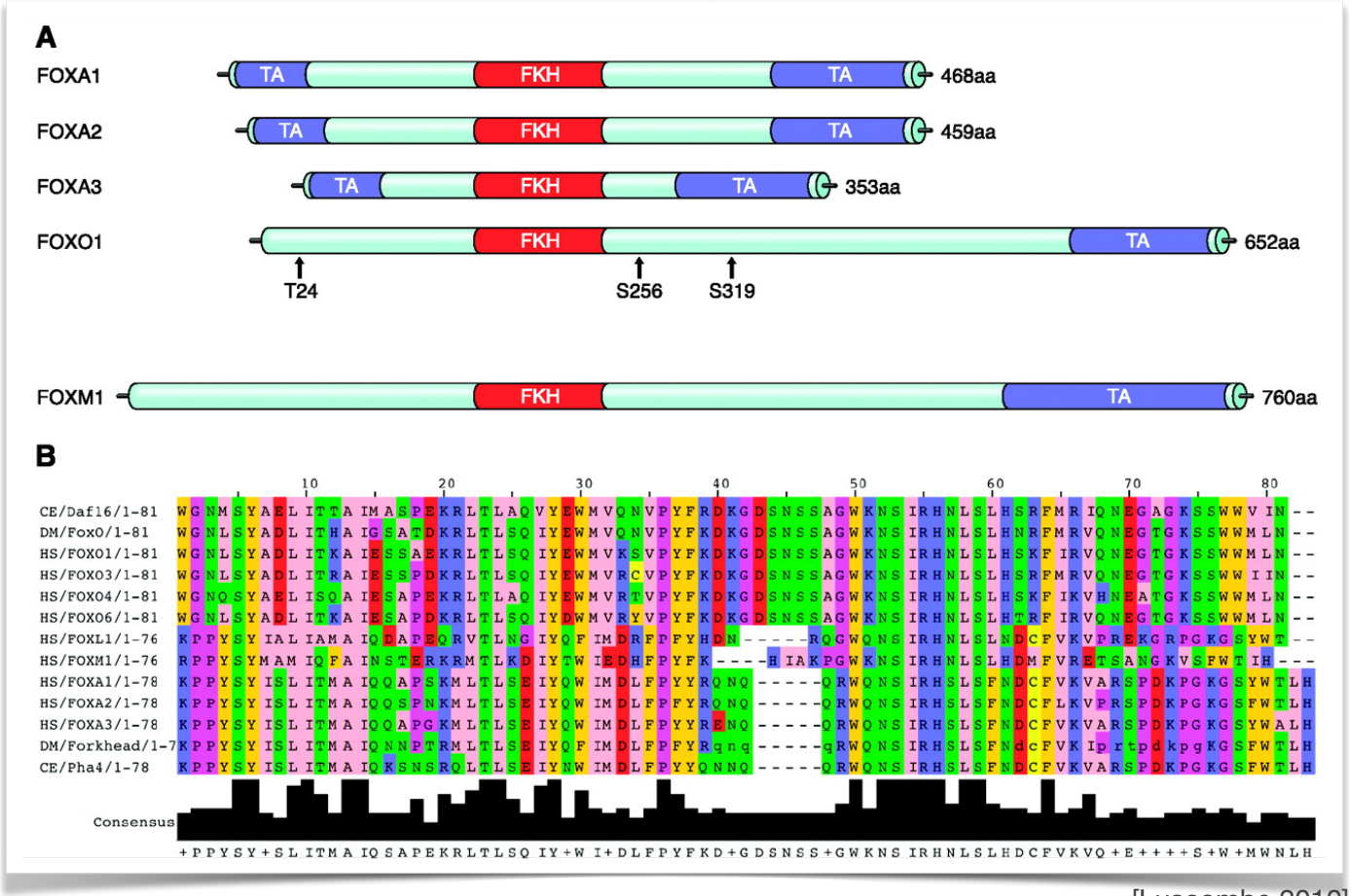
scNMT-seq: identification of DNA-methylation + accessible DNA

[Zhou et al., Nature Methods (2020)]

3. transcription factors

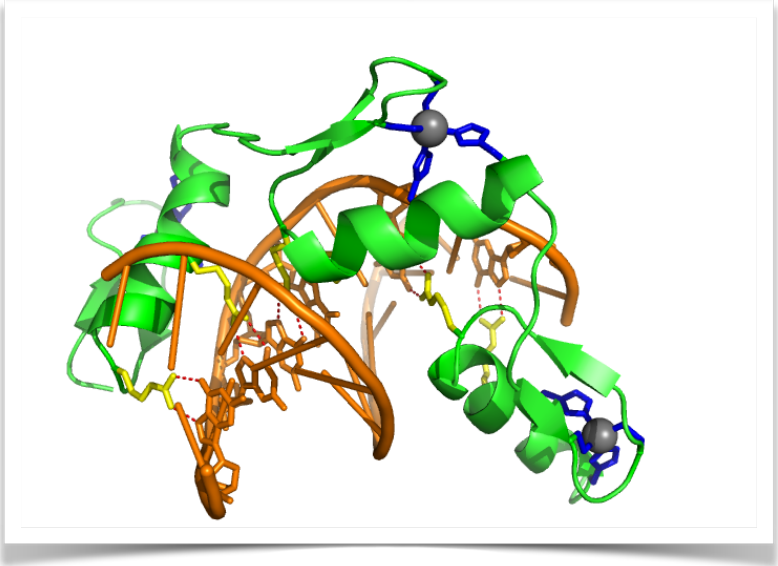
DNA binding domains

- Transcription factors contain a **DNA binding domain** (DBD) and a **transcriptional activator** (TA)
- Homologous TFs share similar DBDs (here: forkhead)



[Luscombe 2010]

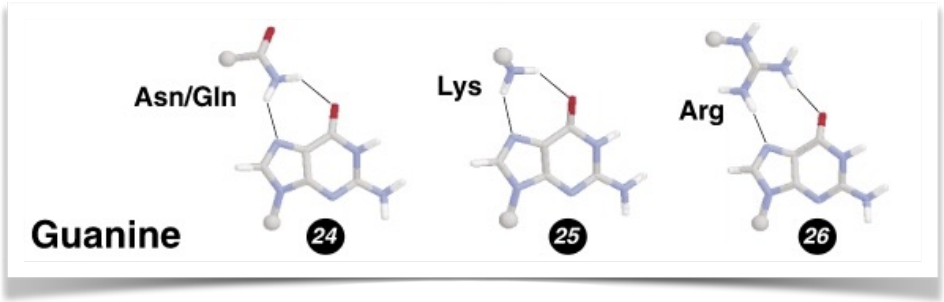
Protein DNA interactions



- majority of protein-DNA interactions for TF occur through a **alpha-helix** fitting into the major groove (=DNA binding domain)
- **hydrogen bonds** with specific bases
- stabilization of the protein-DNA complex is ensured by additional structures (helix, beta-sheet) via **van der Walls** interactions

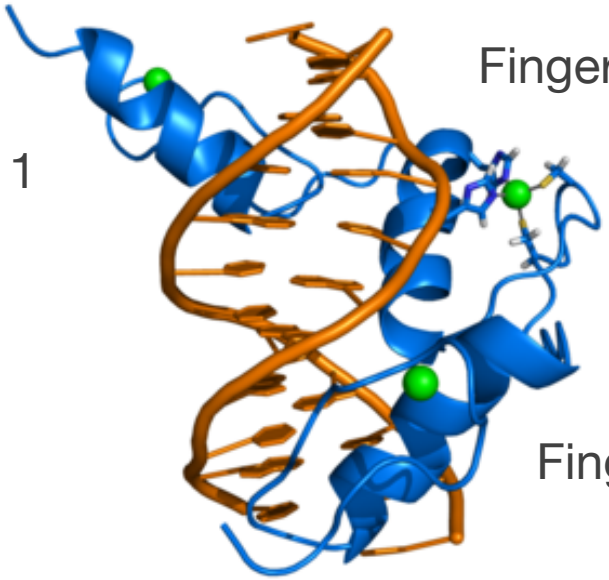
Amino acids	Mode of interaction	Recognised base
Hydrogen bond		
[ARG, LYS]	Multiple-donor	G/complex
[HIS]	Multiple-donor (bifurcate)	G
[SER]	Multiple-donor (bifurcate)	G
	Acceptor+donor	complex
[ASN, GLN]	Acceptor+donor	A/complex
[ASP, GLU]	Multiple-acceptor	complex
van der Waals contacts		
[PHE, PRO]	Ring-stacking	A, T
[THR]	Methyl contact	T
[GLY, ALA, VAL,	-	many (non-
LEU, ISO, TYR]	-	specific)
No base contact		
[CYS, MET, TRP]	-	-

[Luscombe et al., NAR (2001)]

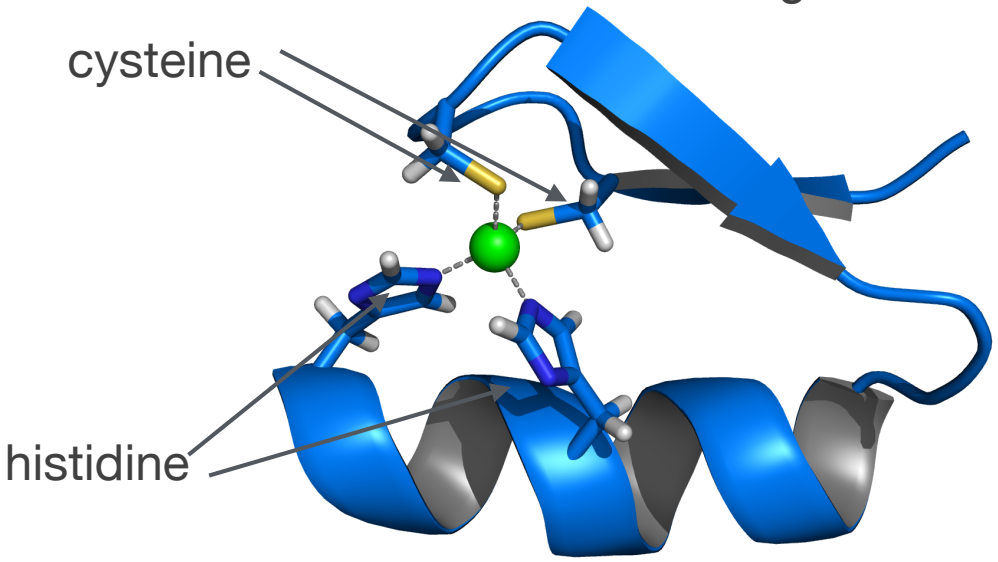


[Cheng et al., JMB (2003)]

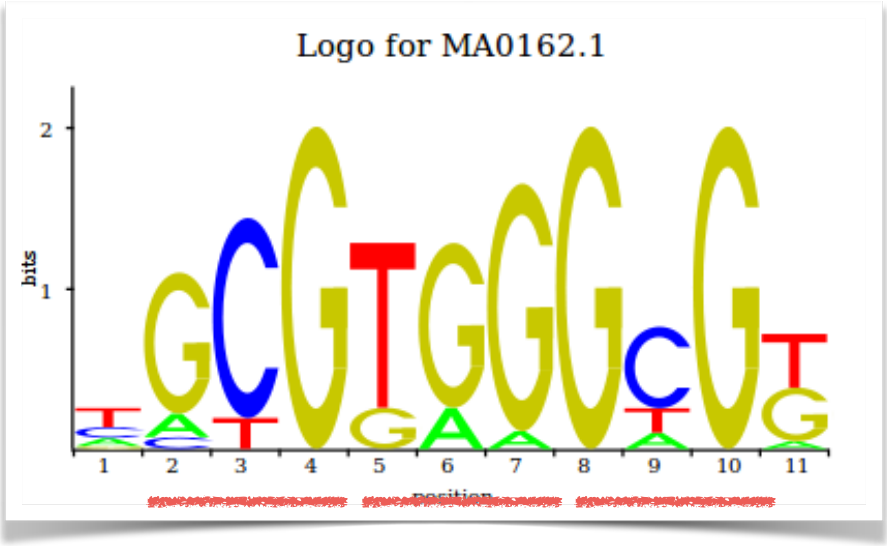
Structural family: Zinc coordinating



Egr1/Zif268



Cys2His2 Fold (“Zinc finger”)
→ one of the most common family
of transcription factors in mammalians



Finger 1 Finger 2 Finger 3

Characterizing binding affinities

How can we represent the binding sites ?

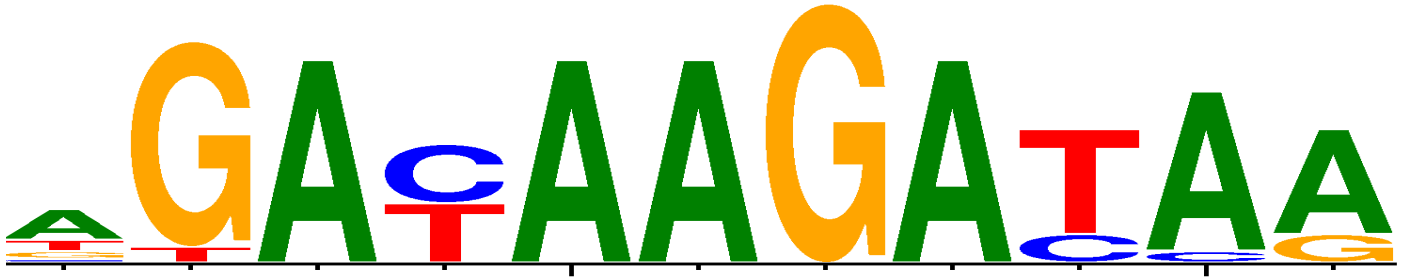
- count frequencies of nucleotides at each position
- normalize to obtain **position frequency matrix (PFM)**


```
G G A C A A G A T A A
A G A C A A G A T A G
A G A C A A G A T A G
G G A C A A G A T A G
T G A C A A G A T C A
C G A C A A G A C A A
A T A C A A G A C A A
T G A T A A G A T A A
A G A T A A G A T A A
T G A T A A G A T A A
A G A T A A G A T A A
A G A T A A G A T A A
A G A T A A G A T A A
A G A T A A G A C A A
```

$f_{i,j}$

a|	0.57	0.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.93	0.79
c|	0.07	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.21	0.07	0.00
g|	0.14	0.93	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.21
t|	0.21	0.07	0.00	0.50	0.00	0.00	0.00	0.00	0.79	0.00	0.00

Predicting binding sites in sequences

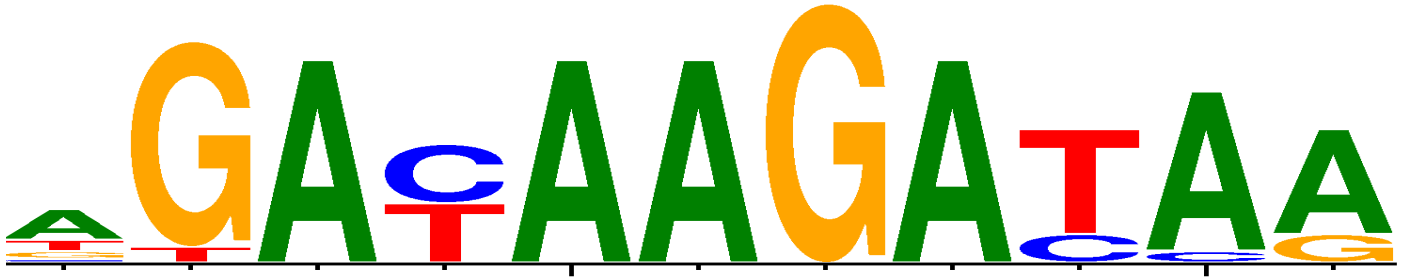


a|	0.55	0.02	**0.95**	0.02	**0.95**	0.95	0.02	**0.95**	0.02	0.88	0.75
c|	0.08	0.02	0.02	**0.48**	0.02	**0.02**	0.02	0.02	**0.22**	**0.08**	0.02
g|	0.15	**0.88**	0.02	0.02	0.02	0.02	**0.95**	0.02	0.02	0.02	**0.22**
t|	**0.22**	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02

T G A C A C G A C C G

p(S|M) = 0.22 * 0.88 * 0.95 * 0.48 * 0.95 * 0.02 * 0.95 * 0.95 * 0.22 * 0.08 * 0.22
= **4.5e-6**

Predicting binding sites in sequences



a|	0.55	0.02	**0.95**	0.02	**0.95**	0.95	0.02	**0.95**	0.02	0.88	0.75
c|	0.08	0.02	0.02	**0.48**	0.02	**0.02**	0.02	0.02	**0.22**	**0.08**	0.02
g|	0.15	**0.88**	0.02	0.02	0.02	0.02	**0.95**	0.02	0.02	0.02	**0.22**
t|	**0.22**	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02

T G A C A C G A C C G

$p(S|M) = \mathbf{4.5e\text{-}6}$

$p(S|B) = p_A^3\ p_C^4\ p_G^3\ p_T$

$= \mathbf{1.9e\text{-}7}$

$$LLR = \log \frac{P(S|M)}{P(S|B)}$$

LLR = 3.2

Matrix logos

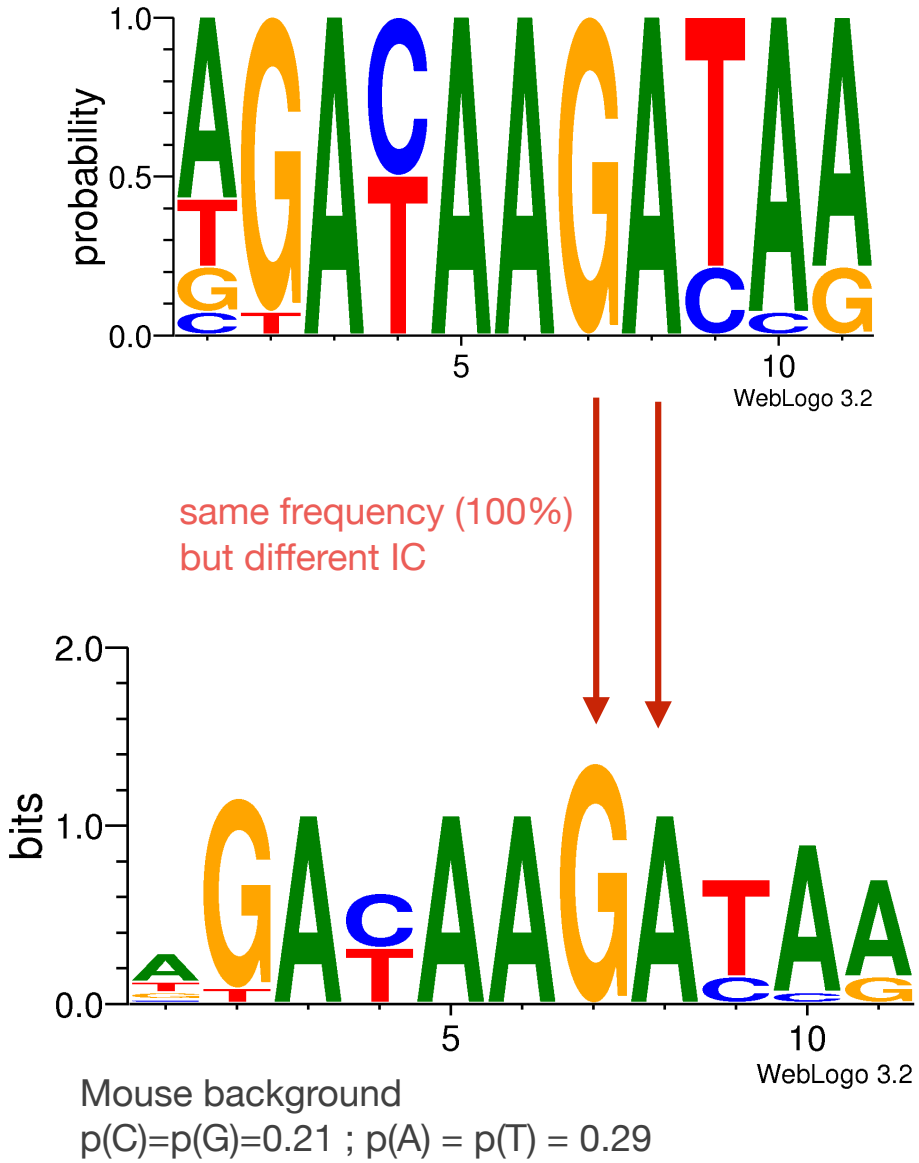
- Information content of the matrix:

$$IC = \sum_{j=1}^{L} \sum_{i \in A,C,G,T} f'_{i,j} \log_2 \frac{f'_{i,j}}{p_i}$$

- Information content of a column:

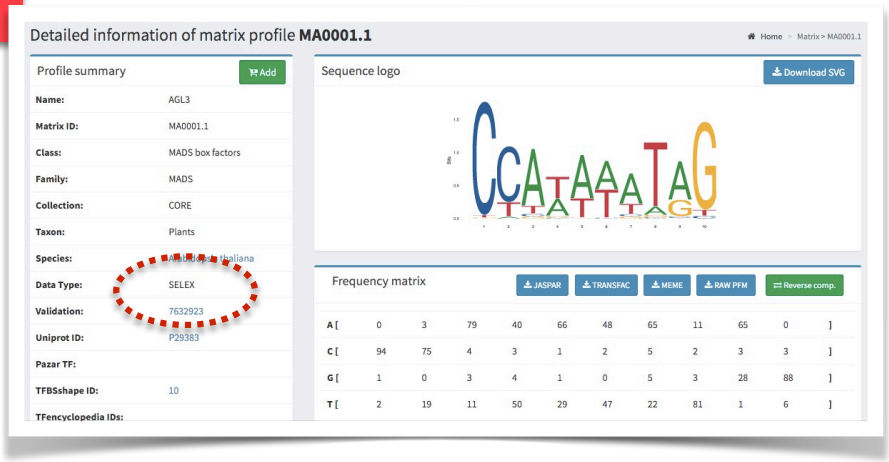
$$IC^j = \sum_{i \in A,C,G,T} f'_{i,j} \log_2 \frac{f'_{i,j}}{p_i}$$

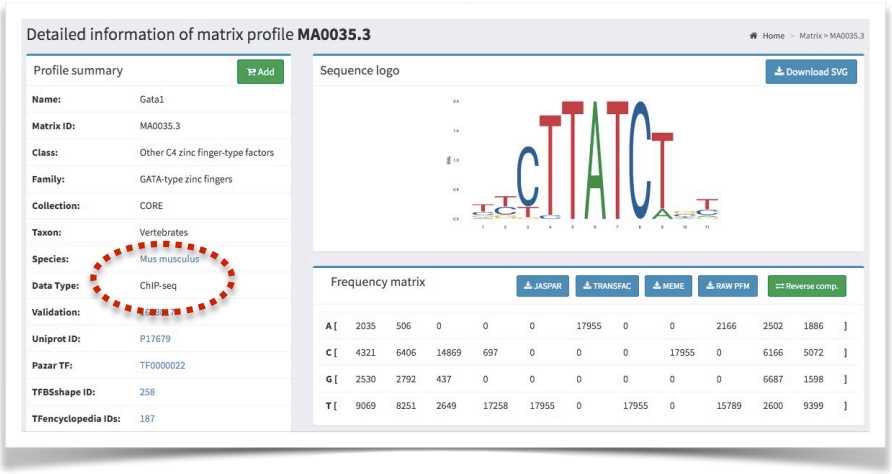
- Conventions:
  - height of column represents IC
  - relative sizes proportional to frequencies



Mouse background
p(C)=p(G)=0.21 ; p(A) = p(T) = 0.29

Different sources

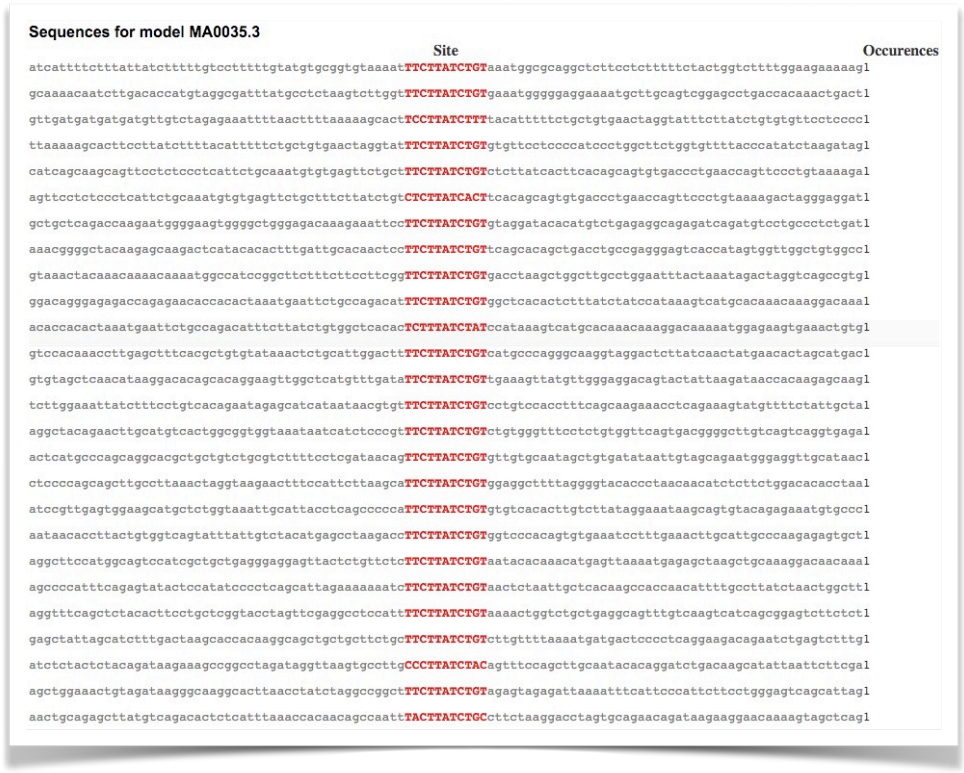




Sequences for model MA0001.1

Site	Occurences
acaaCCATATATAGtagccactgaat	1
ccaccCCATATATAGtagtgcgggtggtg	1
CCATAAATAGataggcagactgtcgctgt	1
gtaaacataCCATAAATAGga	1
ttcaagaaactgCCATAAATAGcgat	1
tagaggtttttgtgCCATAAATAGgt	1
ccCCATAAATAGgaatatcgggatga	1
ttgccattaatagattataCCATATATGG	1
tatcaacaacgataccaacCCATATATGG	1
tttCCAAATATAGaaggtgtggaaag	1
tCCAAATATAGtaaaatcgagtcgcggat	1
gactggggcCCAAATATAGcatgttc	1
atcattagcttttacttaCCATAAATGG	1
attcttttgCCATAAATGGtaactcg	1
CCATAAATGGcaagtctgtcgaataacgg	1
cCCATAAATGGcagggtattagcacg	1
CCAAAAATAGatatgtgtcgtaacagctt	1
CCAAAAATAGggggacaatggaagtgggg	1
CCAAAAATAGgccagacgtgtttacaacg	1
CCAAAAATAGttaaataatgtcatacatt	1
ctacaccttCCAAAAATAGtaatct	1
ttgCCAAATATGGggttagagtgttc	1
gtctttaCCAAAAATGGtgatcctgt	1
ttgCCAAAAATGGagcgtttaccaat	1
atccaCCATTTATAGaaagttcaggaggc	1

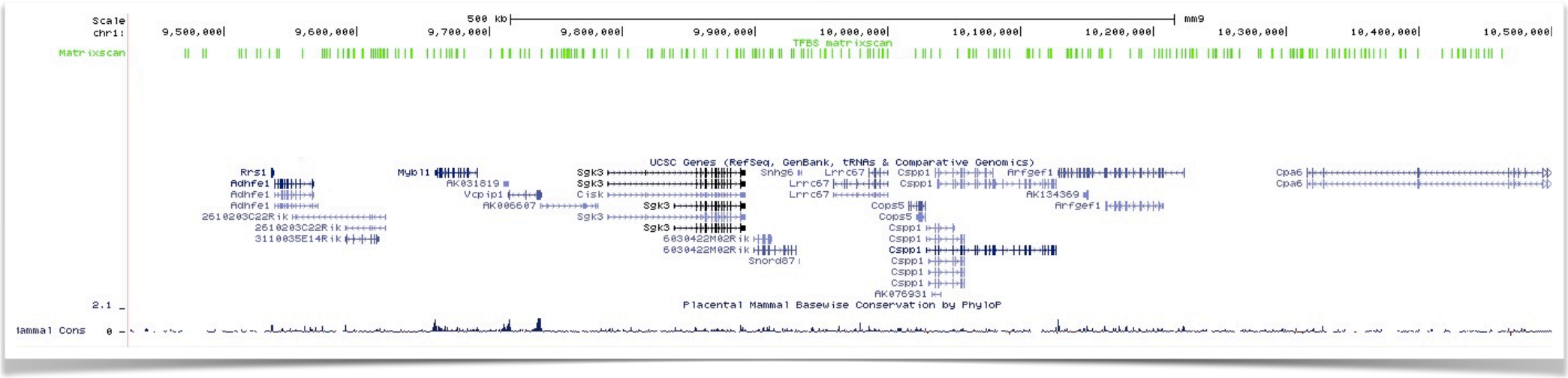
ChIP-seq: real binding site is hidden in much longer sequence
→ lower resolution



Sequences for model MA0035.3

Site	Occurences
atcattttcttattatcttttgtccttttgtatgtgcggtgtaaaatTTCTTATCTGTaaatggcgcaggctcttcctcttttctactggtcttttggaagaaaaag	1
gcaaaacaatcttgacaccatgtaggcgatttatgcctctaagtcttggtTTCTTATCTGTgaaatgggggaggaaaatgcttgcagtcggagcctgaccacaaactgact	1
gttgatgatgatgatgttgtctagagaaattttaacttttaaaaagcactTCCTTATCTTTtacatttttctgctgtgaactaggtatttcttatctgtgtgttcctcccc	1
ttaaaaagcacttccttatctttacatttttctgctgtgaactaggtatTTCTTATCTGTgtgttcctccccatccctggcttctggtgttttacccatatctaagatag	1
catcagcaagcagttcctctccctcattctgcaaatgtgtgagttctgctTTCTTATCTGTctcttatcacttcacagcagtgtgaccctgaaccagttccctgtaaaaga	1
agttcctctccctcattctgcaaatgtgtgagttctgctttcttatctgtCTCTTATCACTtcacagcagtgtgaccctgaaccagttccctgtaaaagactaggggaggat	1
gctgctcagaccaagaatgggagaagtggggctgggagacaaagaaattcctTTCTTATCTGTgtaggatacacatgtctgagaggcagagatcagatgtcctgccctctgat	1
aaacggggctacaagagcaagactcatacacacttttgattgcacaactccTTCTTATCTGTtcagcacagctgacctgccgagggagtcaccatagtggttggctgtggcc	1
gtaaactacaaacaaaacaaaatggccatccggcttctttcttccttcggTTCTTATCTGTgacctaagctggcttgcctggaatttactaaatagactaggtcagccgtg	1
ggacagggagagaccagagaacaccacactaaatgaattctgccagacatTTCTTATCTGTggctcacactctttatctatccataaagtcatgcacaaacaaaggacaaa	1
acaccacactaaatgaattctgccagacatttcttatctgtggctcacacTCTTTATCTATccataaagtcatgcacaaacaaaggacaaaaatggagaagtgaaactgtg	1
gtccacaaaccttgagctttcacgctgtgtataaactctgcattggacttTTCTTATCTGTcatgcccagggcaaggtaggactcttatcaactatgaacactagcatgac	1
gtgtagctcaacataaggacacagcacaggaagttggctcatgtttgataTTCTTATCTGTtgaaagttatgttgggaggacagtactattaagataaccacaagagcaag	1
tcttggaaattatctttcctgtcacagaatagagcatcataataacgtgtTTCTTATCTGTcctgtccaccttcagcaagaaacctcagaaagtatgttttctattgctat	1
aggctacagaacttgcatgtcactggcggtggtaaataatcatctcccgtTTCTTATCTGTctgtgggtttcctctgtggttcagtgacggggcttgtcagtcaggtgaga	1
actcatgcccagcaggcacgctgctgtctgcgtcttttcctcgataacagTTCTTATCTGTgttgtgcaatagctgtgatataatttgtagcagaatgggaggttgcataac	1
ctccccagcagcttgccttaaactaggtaagaactttccattcttaagcaTTCTTATCTGTggaggctttaggggtacaccctaacaacatctcttctggacacacctaa	1
atccgttgagtggaagcatgctctggtaaattgcattacctcagccccaTTCTTATCTGTgtgtcacacttgtcttataggaaataagcagtgtacagagaaatgtgccc	1
aataacaccttactgtggtcagtatttattgtctacatgagcctaagaccTTCTTATCTGTggtcccacagtgtgaaatcctttgaaacttgcattgcccaagagagtgct	1
aggcttccatggcagtccatcgctgctgagggaggagttactctgttctcTTCTTATCTGTaatacacaaacatgagttaaaatgagagctaagctgcaaaggacaacaaa	1
agccccatttcagagtatactccatatccccctcagcattagaaaaaaatcTTCTTATCTGTaactctaattgctcacaagccaccaacattttgccttatctaactggctt	1
aggtttcagctctacacttcctgctcggtacctagttcgaggcctccattTTCTTATCTGTaaaactggtctgctgaggcagtttgtcaagtcatcagcggagtcttctct	1
gagctattagcatctttgactaagcaccacaaggcagctgctgcttctgcTTCTTATCTGTcttgttttaaaatgatgactcccctcaggaagacagaatctgagtctttg	1
atctctactctacagataagaaagccggcctagataggttaagtgccttgCCCTTATCTACagtttccagcttgcaatacacaggatctgacaagcatattaattcttcga	1
agctggaaactgtagataagggcaaggcacttaacctatctaggccggctTTCTTATCTGTagagtagagattaaaatttcattcccattcttcctgggagtcagcattag	1
aactgcagagcttatgtcagacactctcatttaaaccacaacagccaattTACTTATCTGCcttctaaggacctagtgcagaacagatagaaggaacaaaagtagctcag	1

Predicting TFBS on real sequences



- Predicting TFBS on a 1 Mb portion of Mouse chromosome 1
- Software : Matrix-Scan ; Matrix : **HNF4a**
- Threshold to call TFBS : $p \leq 1e\text{-}4$
- Background : Markov model order=3 estimated on input sequence
- Output : **259 predicted TFBS**

RSAT Metazoa
http://rsat.eu/