Principles and Methods in Regulatory Genomics

Carl Herrmann

Bioinfo 1 — MoBi Bachelor 5. FS

WS 2023/2024

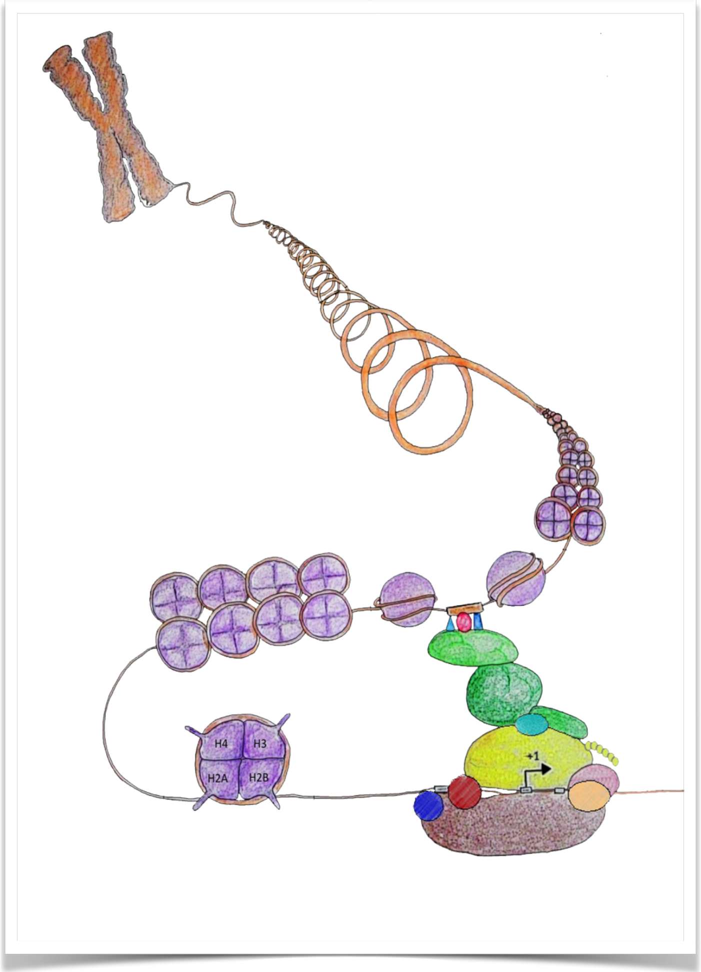
IPMB

Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

What this is all about ...

- **How are the right genes expressed in the right cells at the right time? And why is it sometimes not the case?**
- What are the **molecular actors** of transcriptional regulation ?
  - Role of transcription factors ?
  - Role of epigenetic modifications ?
  - Role of 3D chromatin structure ?
- What kind of **experimental data** can we use to understand transcriptional regulation ?
- How can we use **bioinformatics tools** to make predictions ?



Content of the lecture

- **Part 1**
  - Introduction to regulatory genomics
  - Main actors and available data types

- **Part 2**
  - Models for binding of transcription factors
  - Experimental approaches for the determination of TFBS / databases

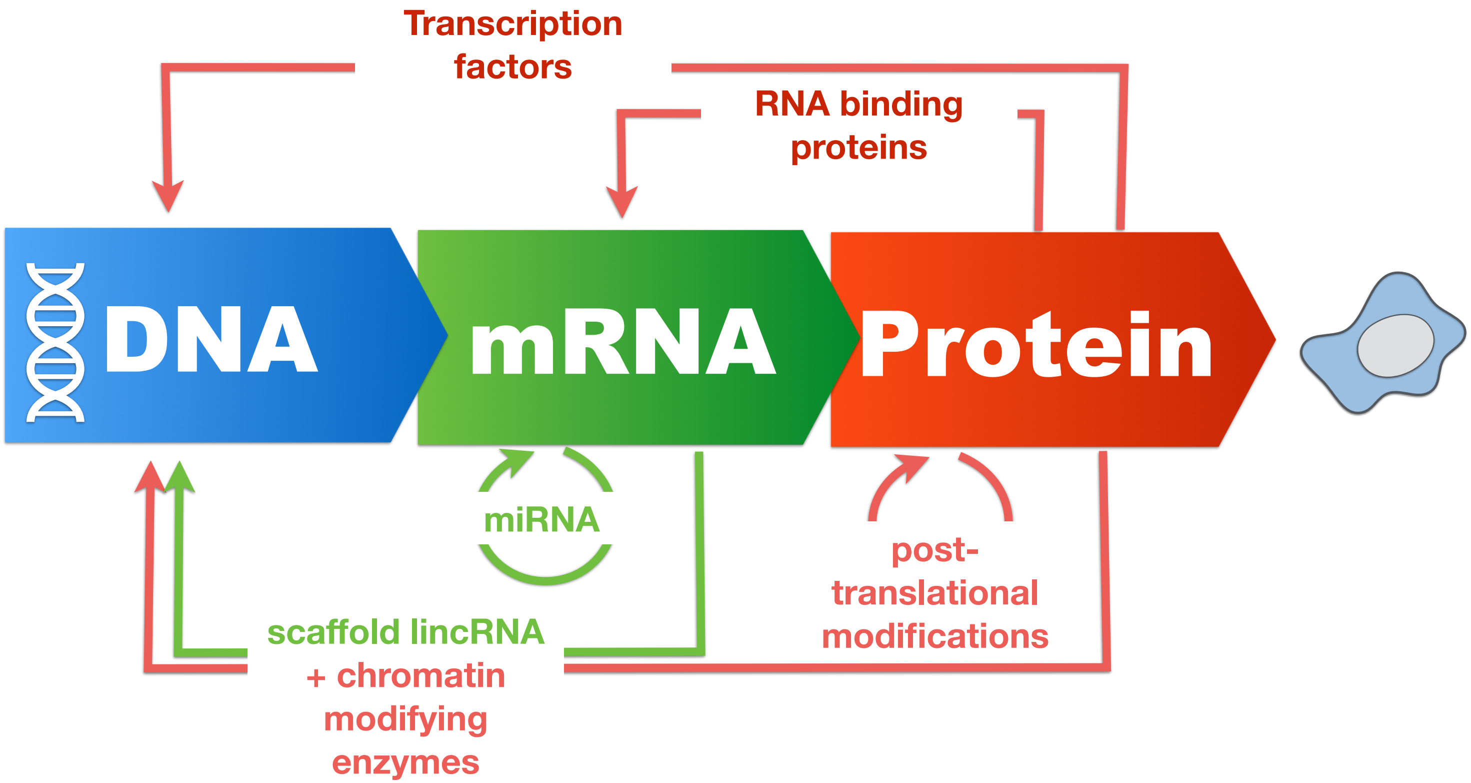
- **Part 3**
  - Bioinformatic prediction of TFBS
  - Role of background models

- **Part 4**
  - Improving predictions: Phylogenetic footprinting

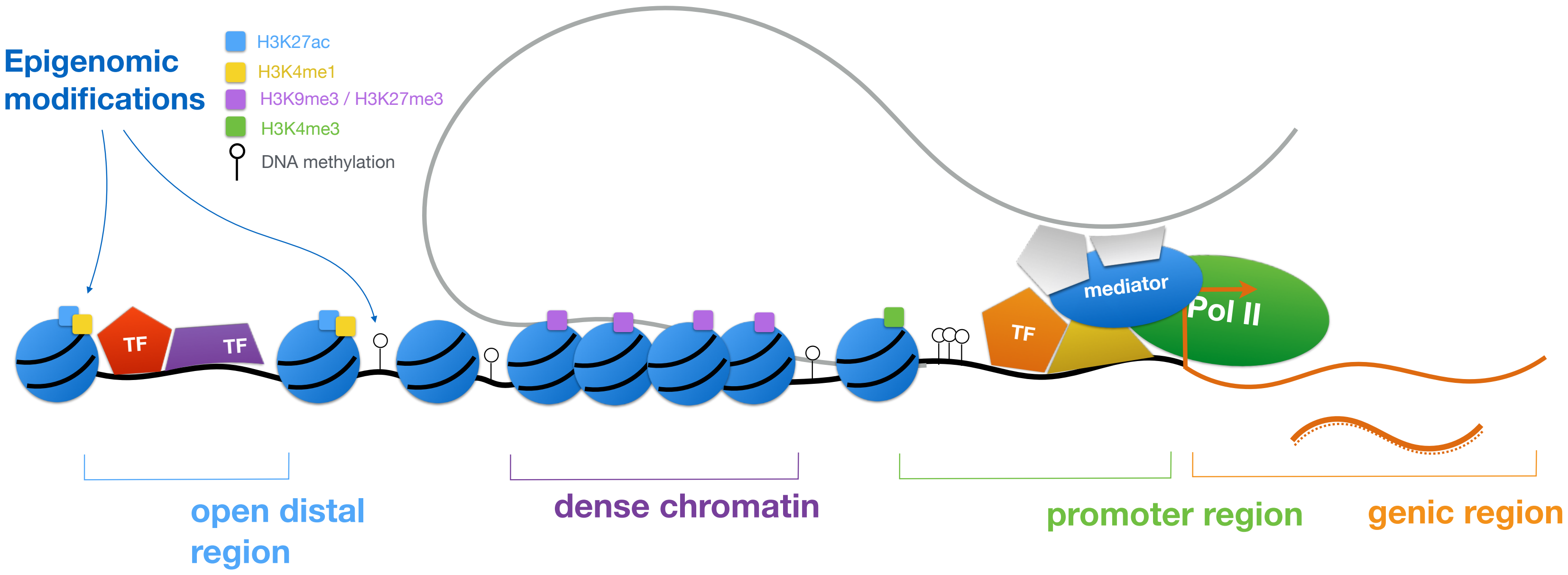
- **Part 5**
  - Epigenetic regulation
  - Hidden Markov Models / ChromHMM

- **Part 6**
  - Motif discovery
  - Expectation Maximization

The Dogma in the genomics area



Transcriptional regulation



Epigenomic modifications

- 🟦 H3K27ac
- 🟨 H3K4me1
- 🟪 H3K9me3 / H3K27me3
- 🟩 H3K4me3
- ⚲ DNA methylation

open distal region

dense chromatin

promoter region

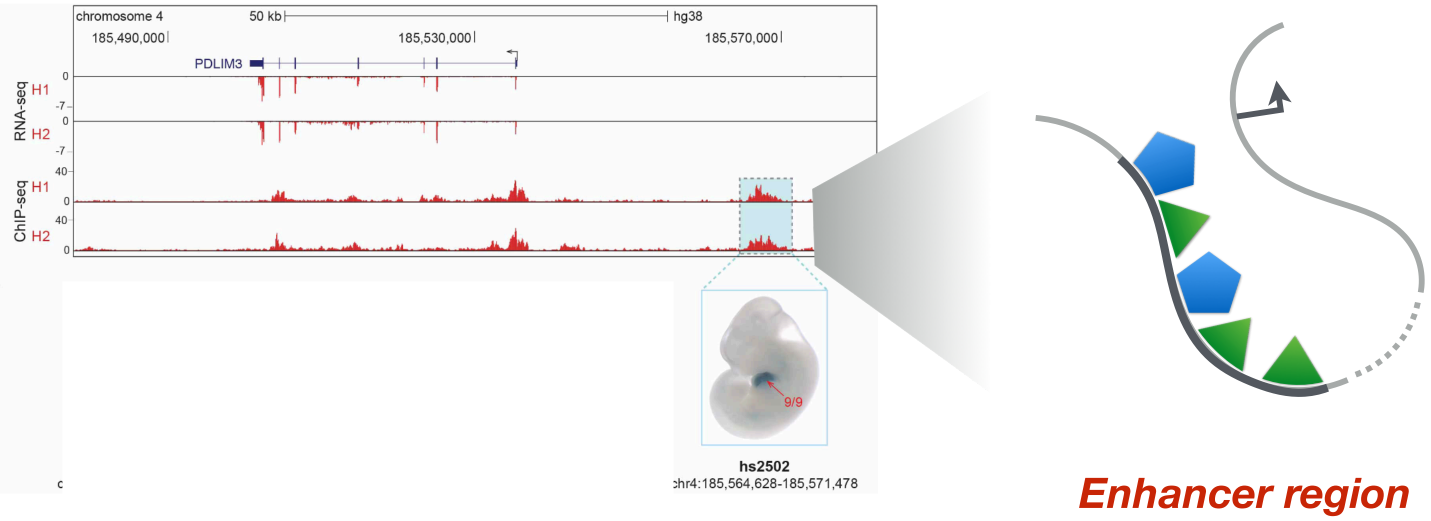
genic region

combinatorial interplay of multiple components

Transcriptional regulation

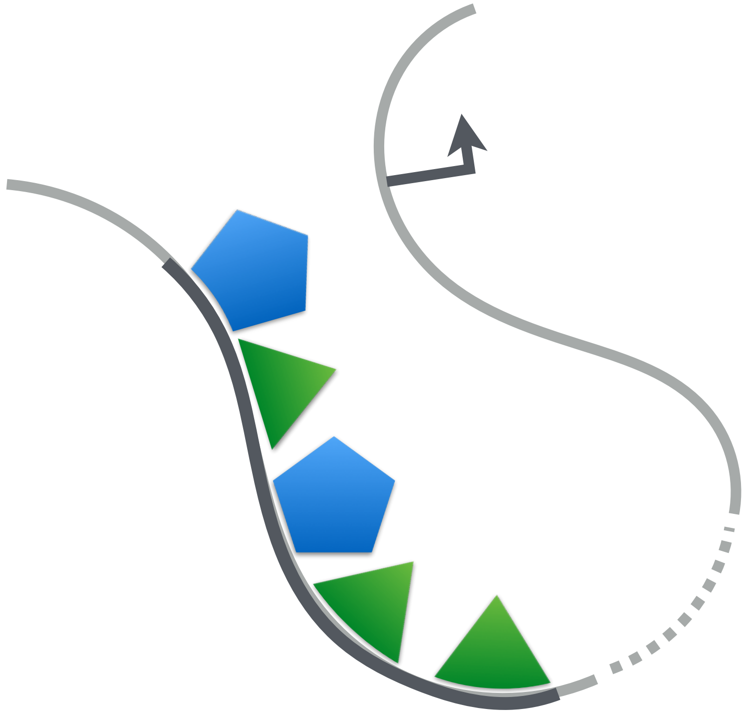
- **Enhancers** are regulatory elements which can be located far from the target genes
- **Multiple binding sites** for different transcription factors
- typical length: few kb → several hundred kb ("superenhancers")
- Organisational principles ("grammar") remains unclear (see exercises)

Identification of heart enhancers

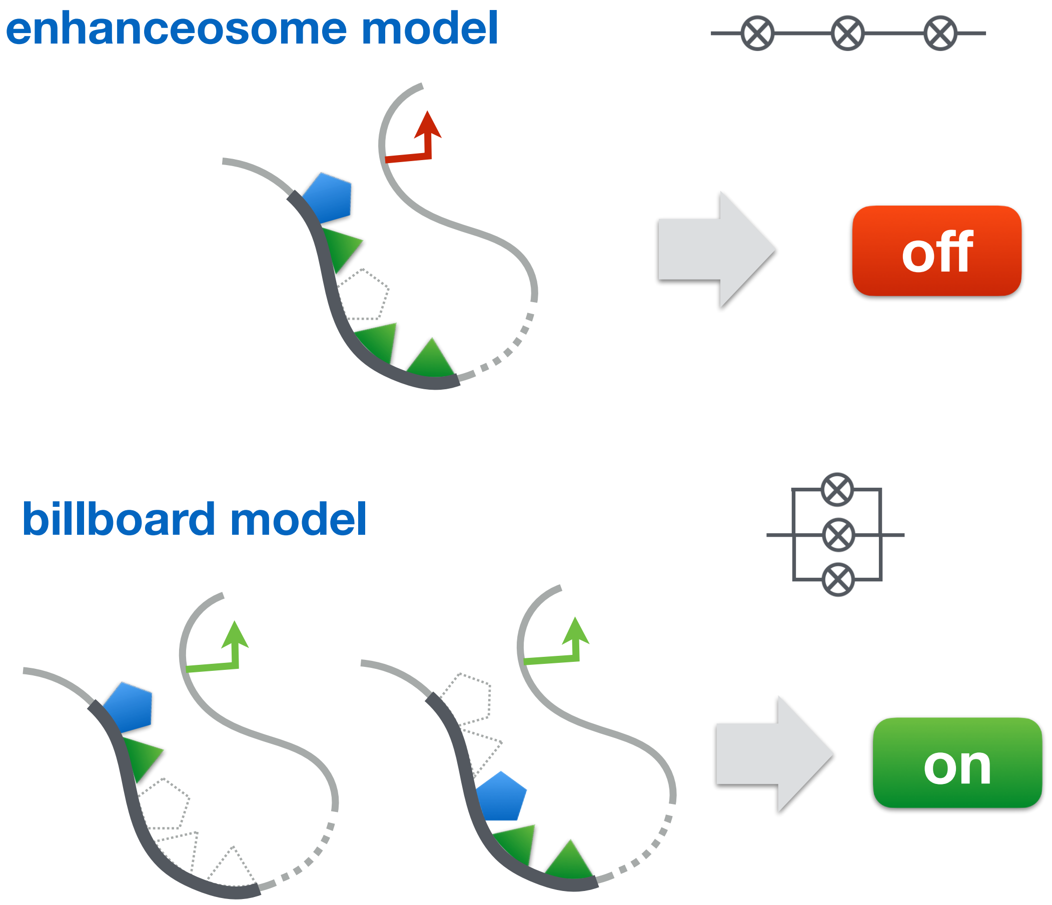


[Spurrell et al., 2019]

Transcriptional regulation



Enhancer region



enhanceosome model

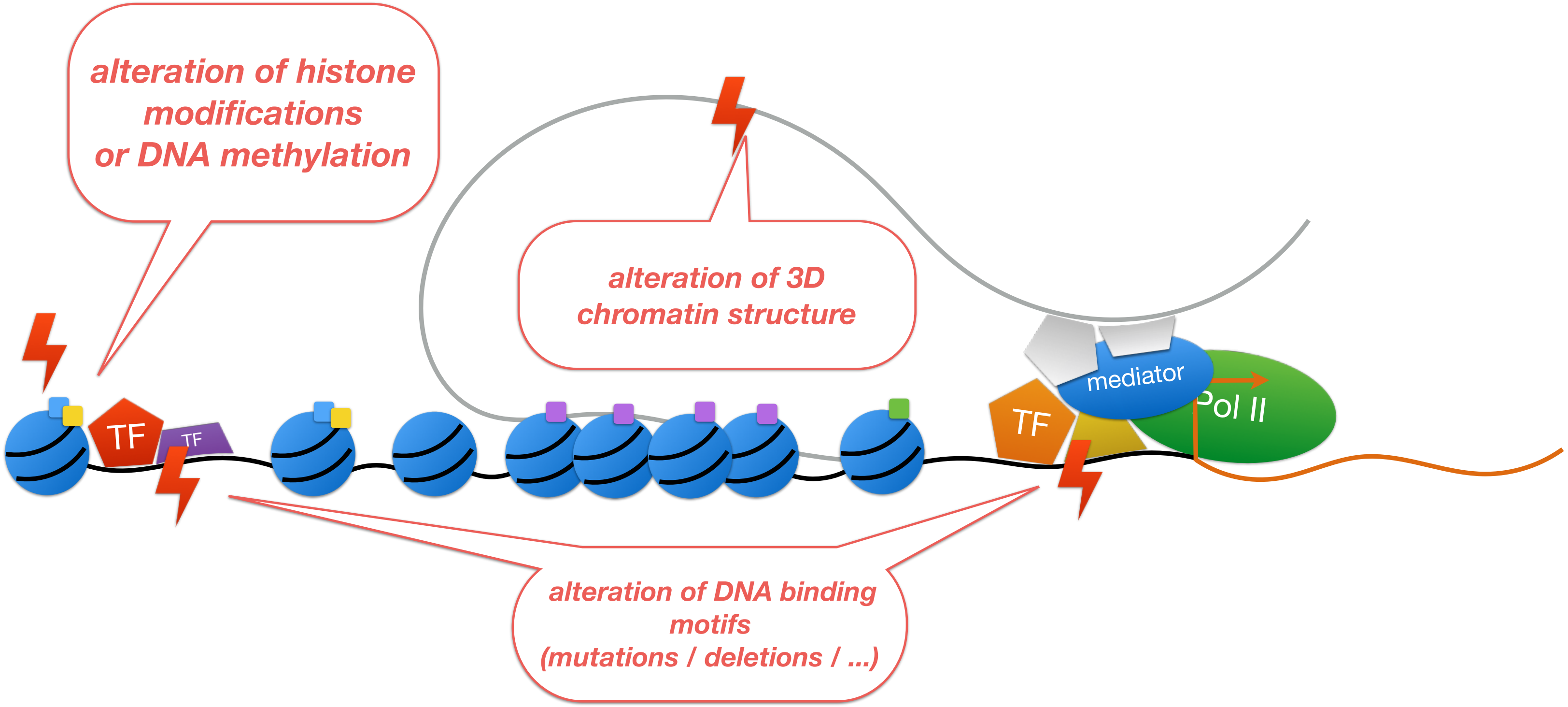
off

billboard model

on

[Arnosti, Kulkarni, 2005]

Transcriptional deregulation



alteration of histone modifications or DNA methylation

alteration of 3D chromatin structure

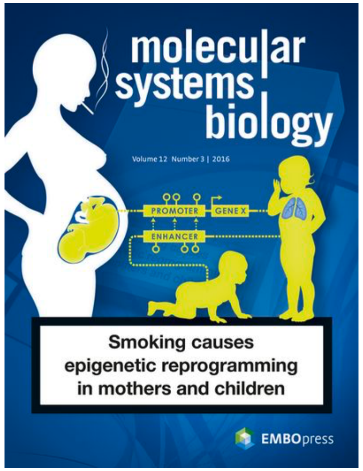
alteration of DNA binding motifs (mutations / deletions / ...)

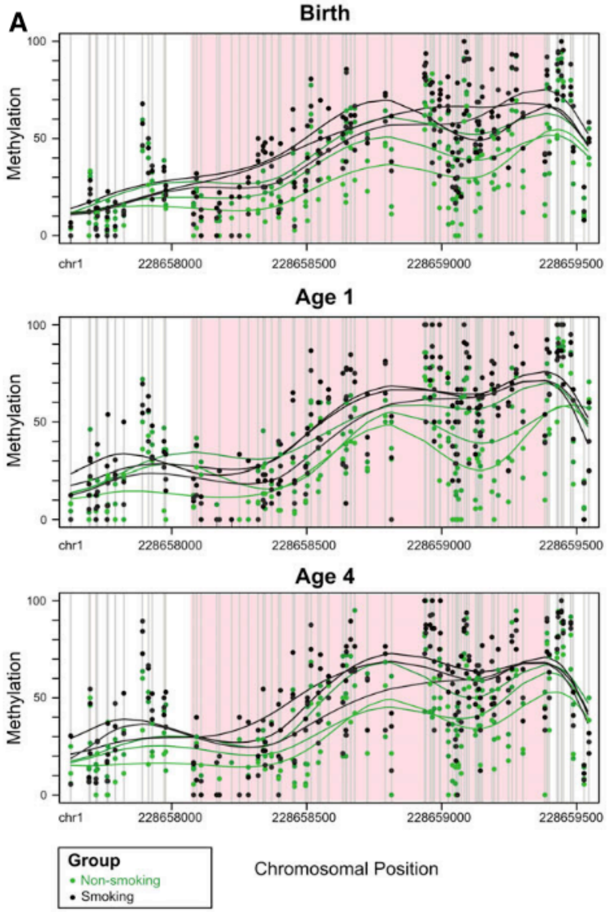
complex interplay of multiple components

multiple sources of potential deregulation

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

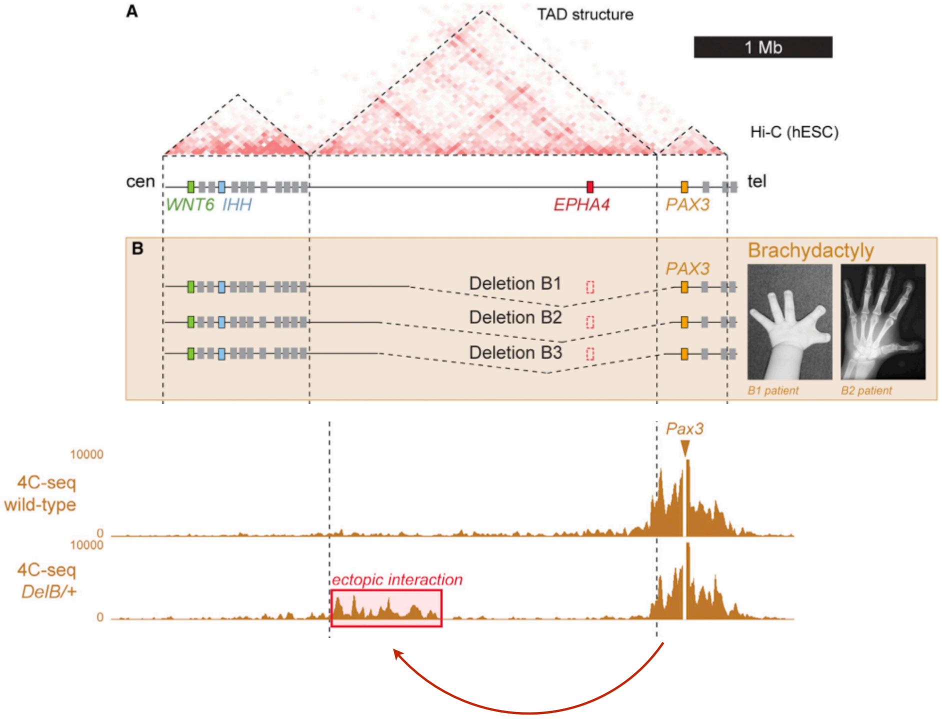

molecular systems biology
Volume 12 Number 3 | 2016
PROMOTER
GENE X
ENHANCER
Smoking causes epigenetic reprogramming in mothers and children
EMBOpress



[Bauer et al., MSB (2016)]

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

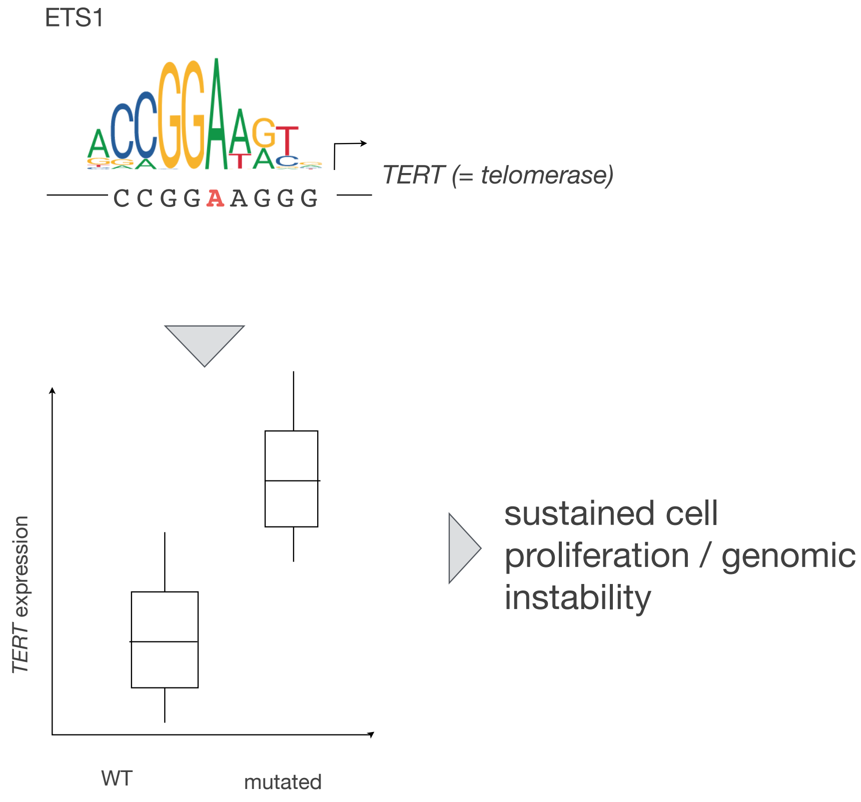


new regulatory interaction

[Lupiañez,..., Mundlos, Cell (2015)]

Sequence deregulation

- Point mutations affect transcription factor binding sites in cancer genomes
- Creation/disruption of binding sites
  → deregulation of target gene expression
- Example: somatic point mutation in TERT promoter creates new binding site for ETS1 transcription factor



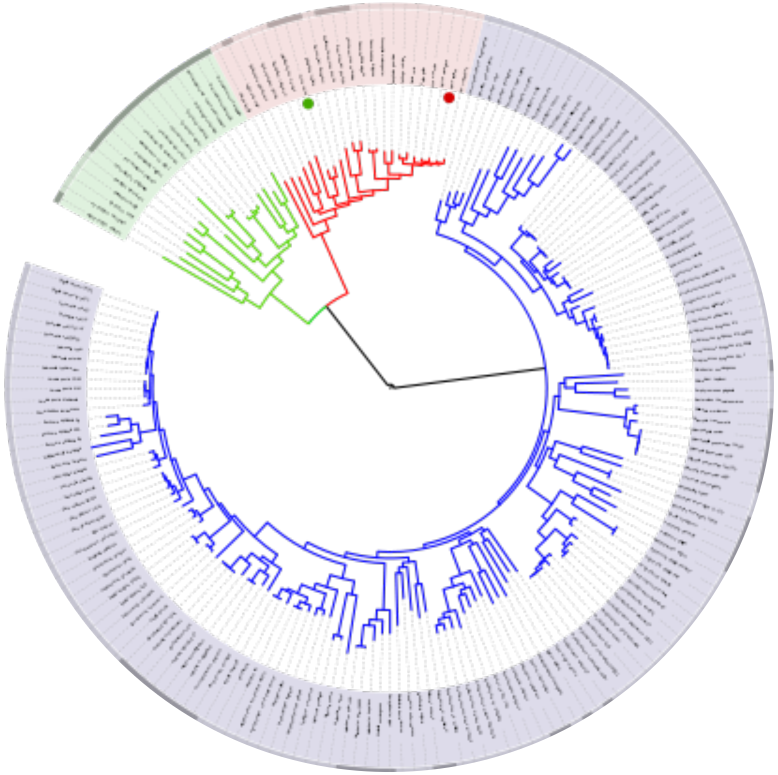
sustained cell proliferation / genomic instability

[Killela et al., PNAS (2013)]

[Huang et al., Science (2013)]

Bigger genome = more evolved ?

Genome size

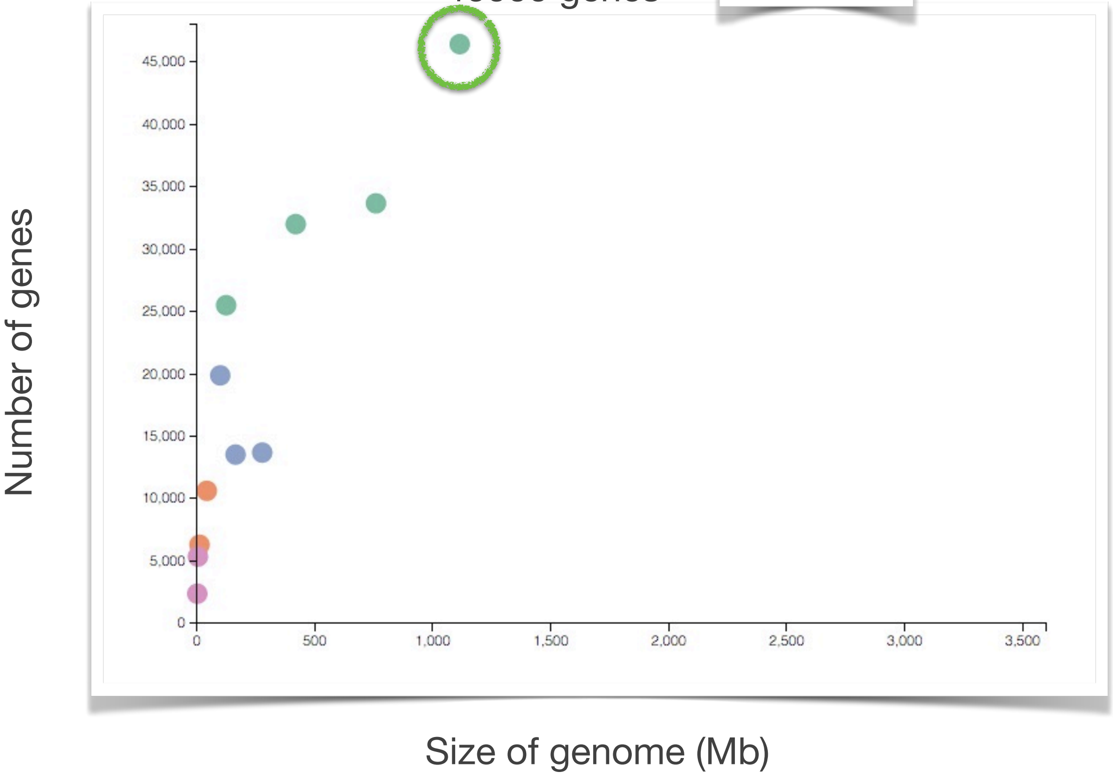


Eukaryotes
Archae
Bacteria

[interactive Tree of Life]

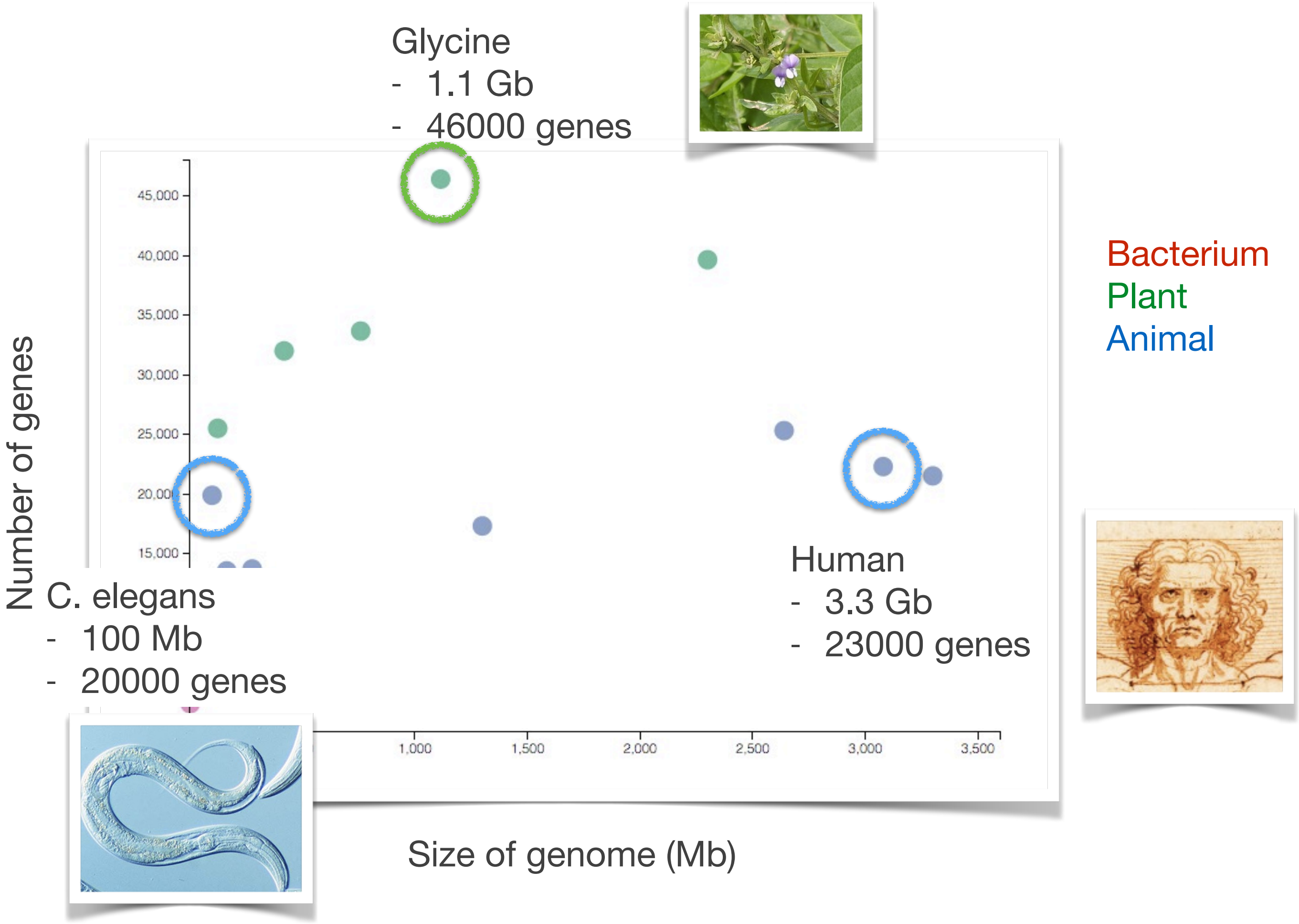
More genes = more complex ?

Glycine
- 1.1 Gb
- 46000 genes



Bacterium
Plant
Animal

More genes = more complex ?



Glycine
- 1.1 Gb
- 46000 genes

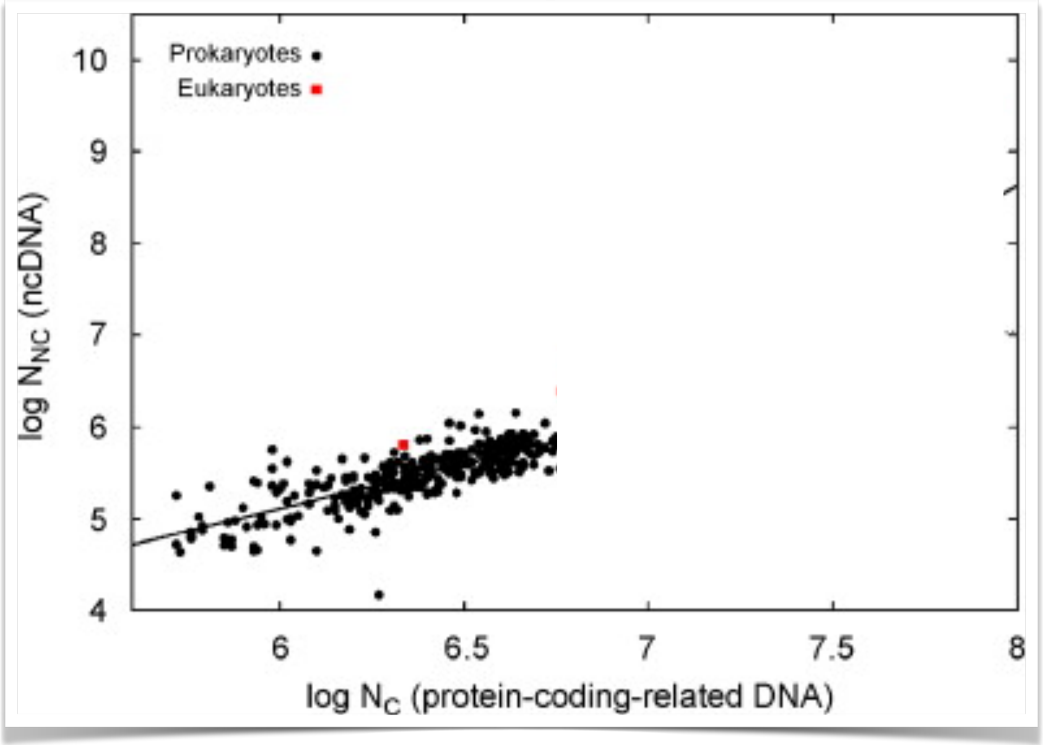
C. elegans
- 100 Mb
- 20000 genes

Human
- 3.3 Gb
- 23000 genes

Bacterium
Plant
Animal

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

Bigger non-coding genome = higher complexity

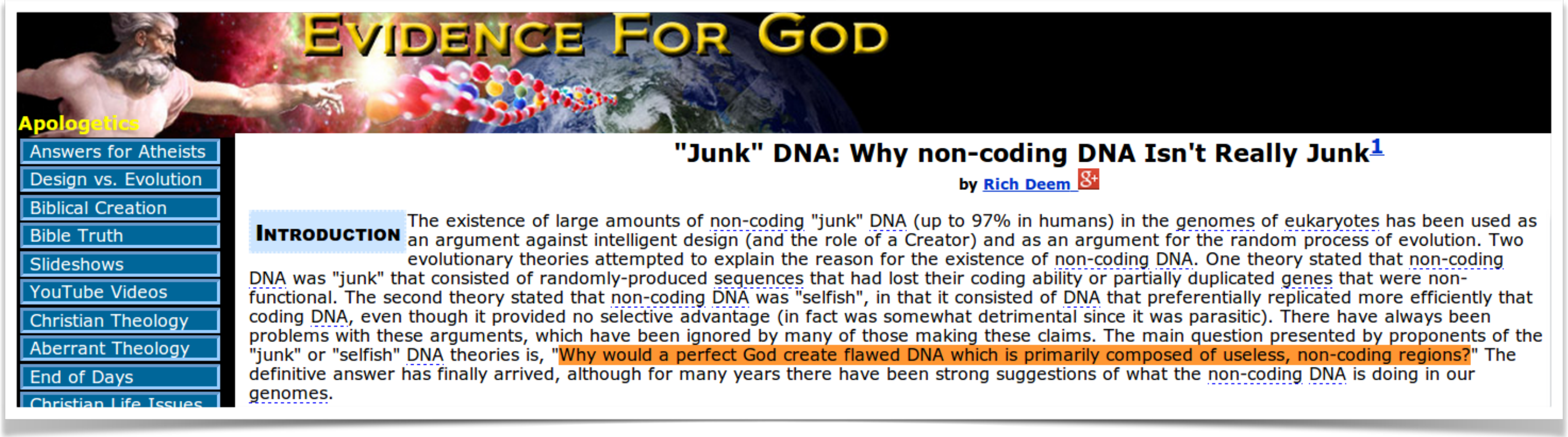


Proportion of **non-coding DNA** correlates with organismal complexity

[Ahnert, Fink, Zinovyev, 2008]

Junk DNA ?

ICR INSTITUTE FOR CREATION RESEARCH



"Junk" DNA: Why non-coding DNA Isn't Really Junk[1]

by Rich Deem

Introduction The existence of large amounts of non-coding "junk" DNA (up to 97% in humans) in the genomes of eukaryotes has been used as an argument against intelligent design (and the role of a Creator) and as an argument for the random process of evolution. Two evolutionary theories attempted to explain the reason for the existence of non-coding DNA. One theory stated that non-coding DNA was "junk" that consisted of randomly-produced sequences that had lost their coding ability or partially duplicated genes that were non-functional. The second theory stated that non-coding DNA was "selfish", in that it consisted of DNA that preferentially replicated more efficiently that coding DNA, even though it provided no selective advantage (in fact was somewhat detrimental since it was parasitic). There have always been problems with these arguments, which have been ignored by many of those making these claims. The main question presented by proponents of the "junk" or "selfish" DNA theories is, "Why would a perfect God create flawed DNA which is primarily composed of useless, non-coding regions?" The definitive answer has finally arrived, although for many years there have been strong suggestions of what the non-coding DNA is doing in our genomes.

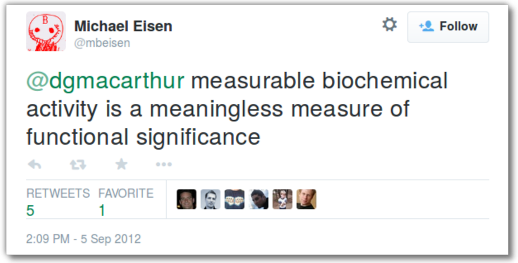
Creationists: An Intelligent Designer would not have filled the human genome with useless, junk DNA

Evolutionarists: Junk DNA is no junk at all…

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.*"



Michael Eisen
@mbeisen

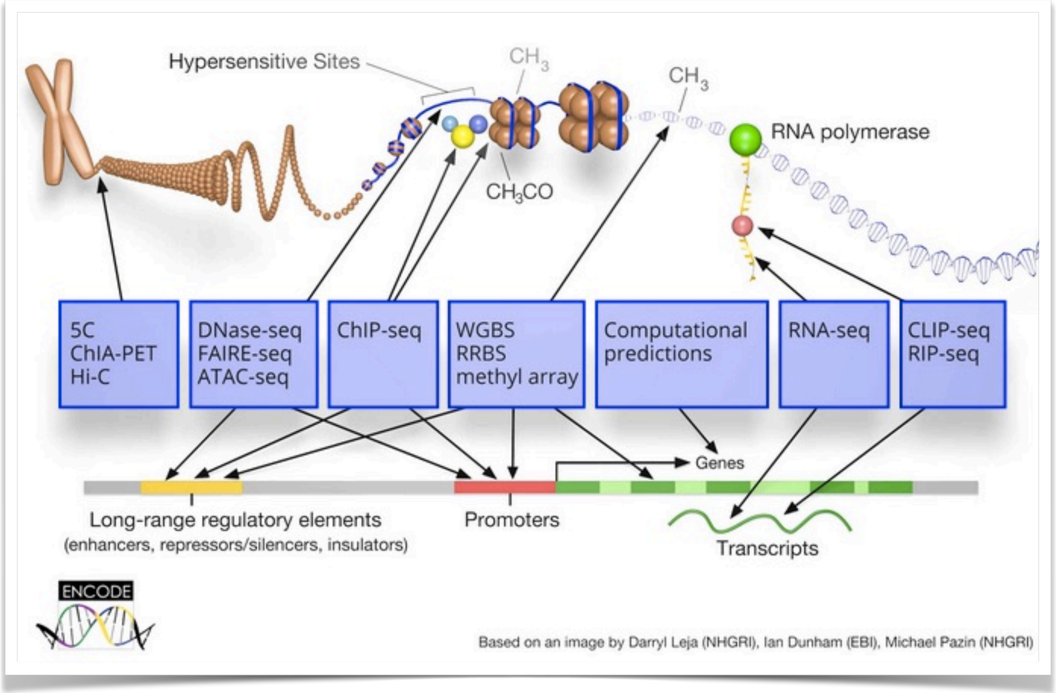
Follow

@dgmacarthur measurable biochemical activity is a meaningless measure of functional significance

RETWEETS 5 FAVORITE 1

2:09 PM - 5 Sep 2012

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



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

https://www.encodeproject.org/

1. data types

Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Experimental methods

- Sequence contribution:

- Chromatin structure and epigenetic

- Three-dimensional DNA looping

- Readout: gene expression

Experimental methods

- **Sequence contribution:**
  → ChIP-seq: transcription factor binding sites

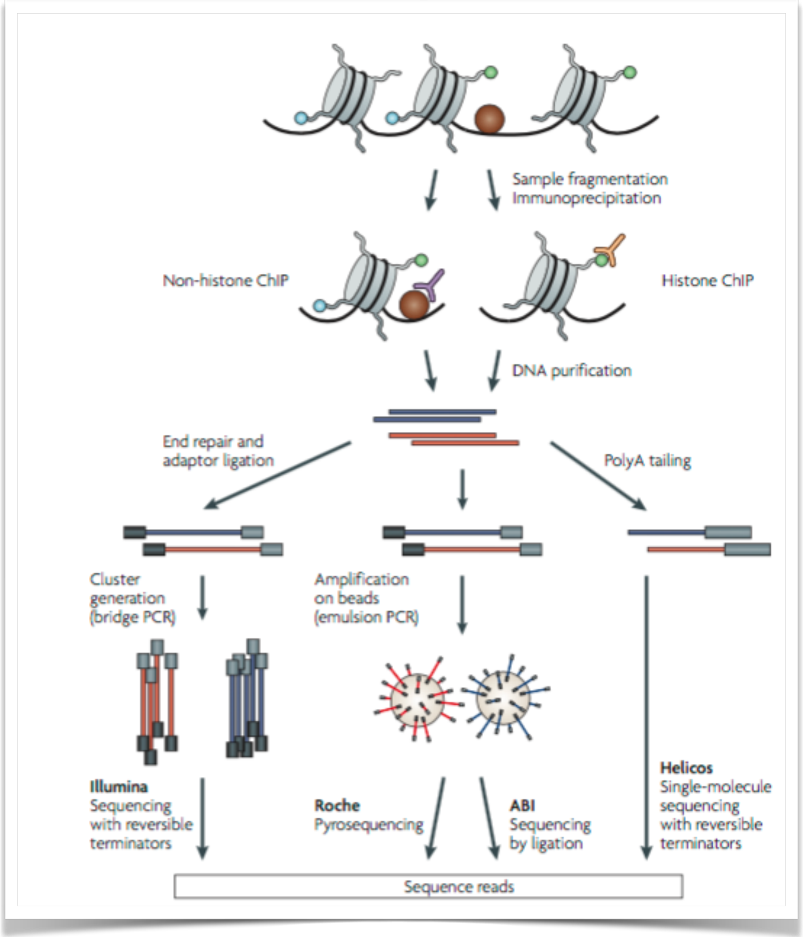
- **Chromatin structure and epigenetic**
  → ChIP-seq : post-translational histone modifications
  → whole genome bisulfite sequencing, arrays : DNA methylation
  → ATAC-seq, DNAse-seq, FAIRE-seq : open chromatin region

- **Three-dimensional DNA looping**
  → 3C/4C/Hi-C : interacting chromatin regions

- **Readout: gene expression**
  → RNA-seq : expression of transcribed elements

Chromatin Immunoprecipitations

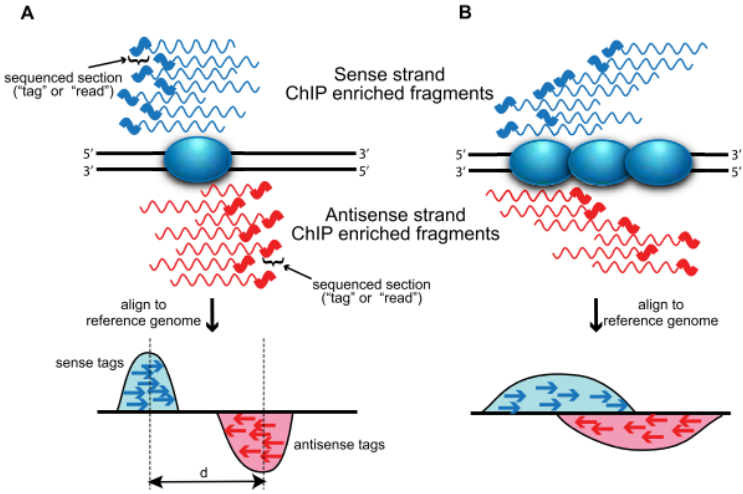
- Chromatin immunoprecipitation (ChIP) yields **DNA fragments**, that are
  - bound by the protein of interest
  - marked by a specific chemical modification (acetylation, methylation,.)
- Identification of the fragments :
  - sequencing (ChIP-seq)
    → genome-wide
  - PCR/qPCR
    → targeted experiment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


Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads

[Park, Nat.Rev. 2009]

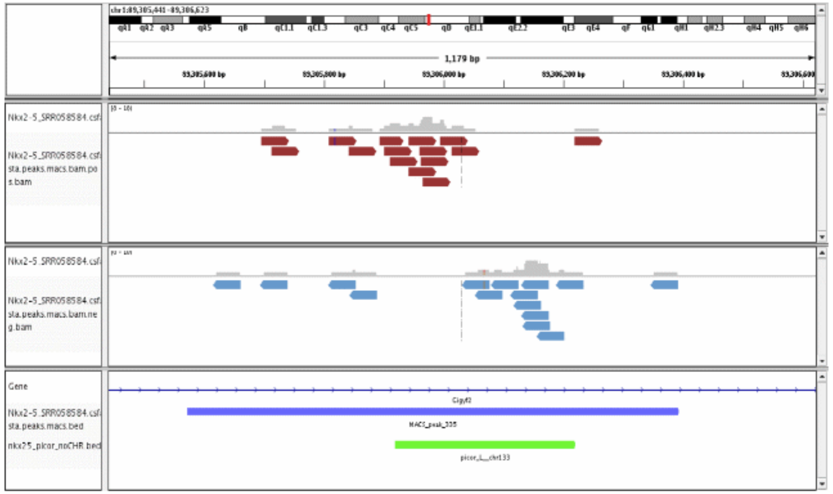
ChIP-sequencing

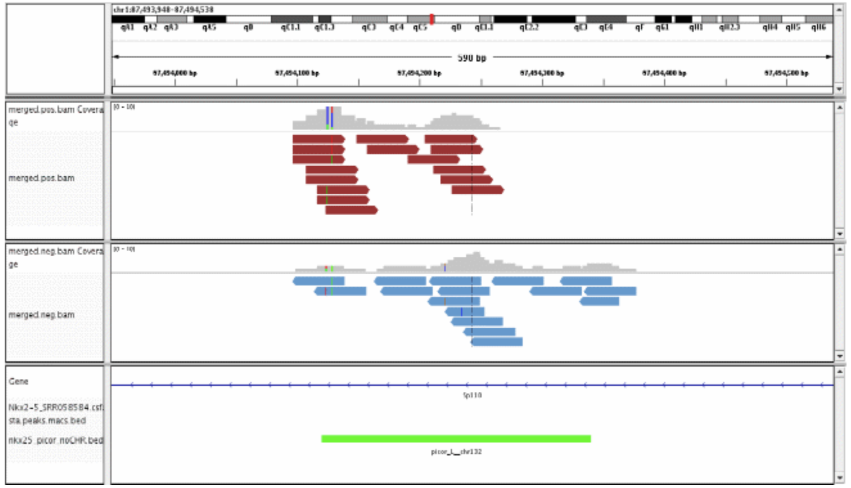
Sharp signal



Broad signal

[Wilbanks & Facciotti
PLoS One (2010)]]

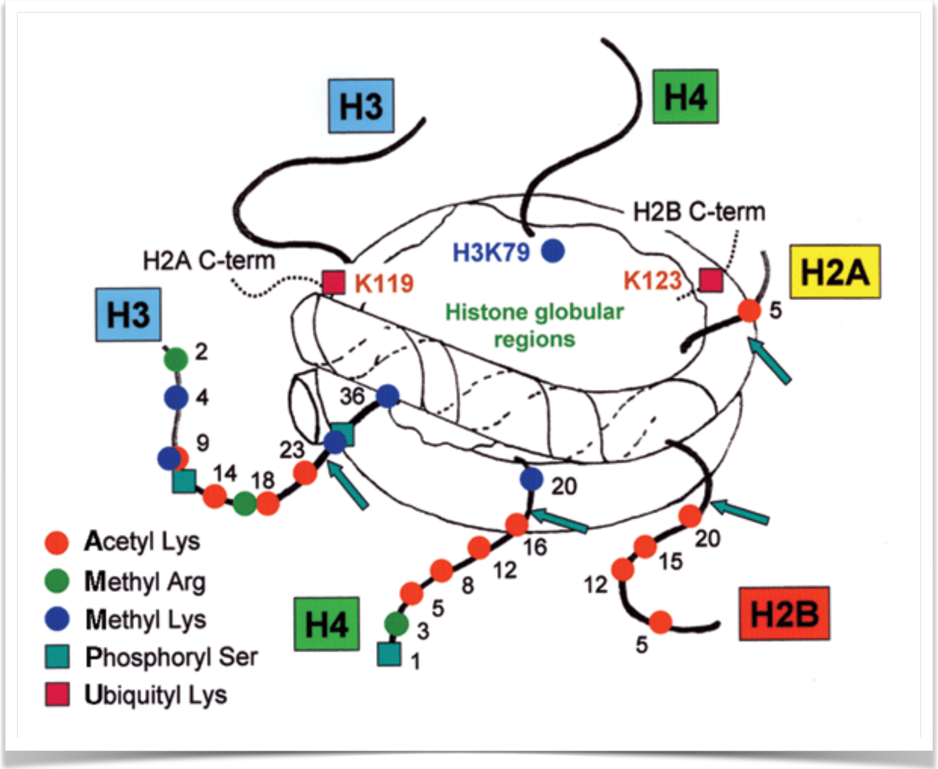

1,179 bp
Nkx2-5_SRR058584.csf
Nkx2-5_SRR058584.csfasta.peaks.macs.bam.pos.bam
Nkx2-5_SRR058584.csf
Nkx2-5_SRR058584.csfasta.peaks.macs.bam.neg.bam
Gene
Nkx2-5_SRR058584.csfasta.peaks.macs.bed
nkx25_picor_noCHR.bed



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

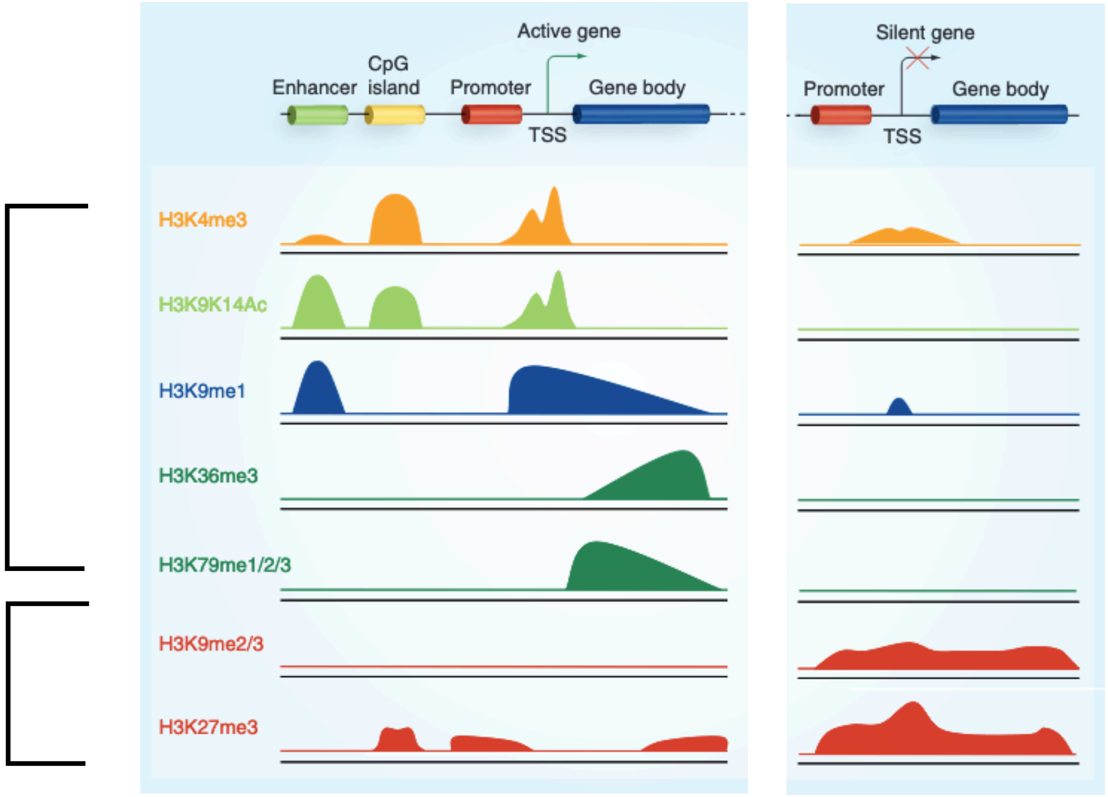
- they **modify the physical properties of the DNA-nucleosome interactions**



nomenclature: H3K27ac = acetylation of lysine 27 on histone 3

Histone modifications

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

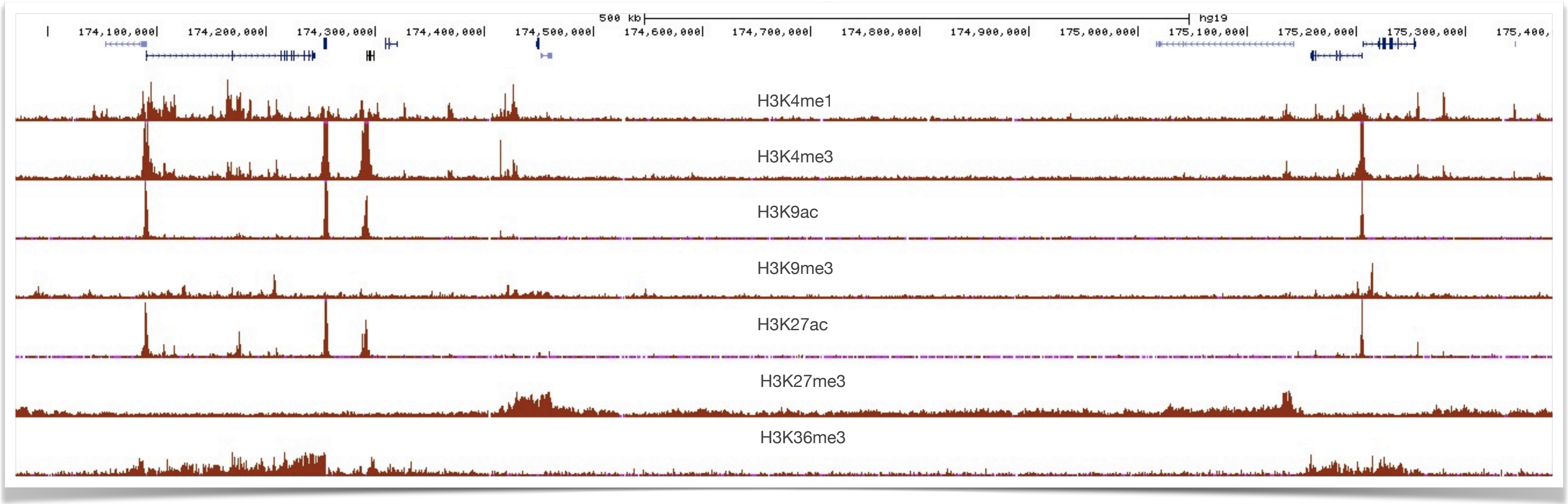


active marks

repressive marks

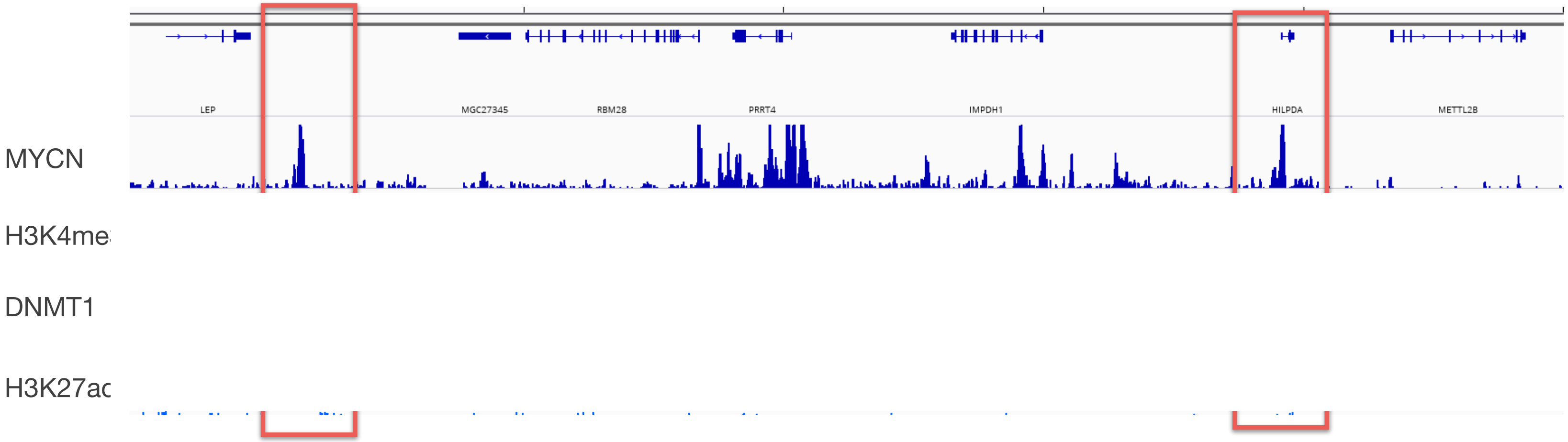
[Lin, Shannon, Hardy, 2010]

Histone modifications



- Histone marks have **distinct signal profiles**
  - sharp signal : narrow peaks of enrichment at specific loci (H3K4me3 = promoters, H3K27ac = enhancers,…)
  - broad signal : wide regions of enrichment (H3K36me3 = transcribed genes; H3K27me3 = repressed regions)

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

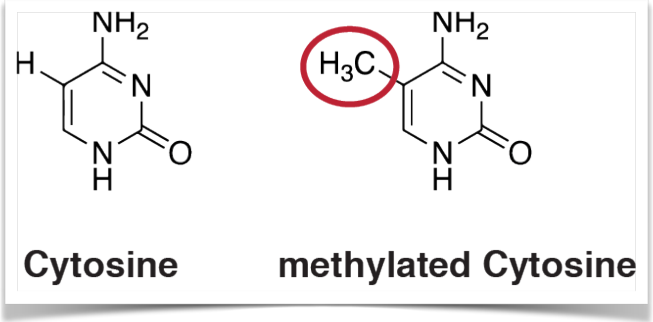


distal
MYCN peak
(away from gene promoters)

MYCN peak
and H3K4me3
enrichment
at promoter

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

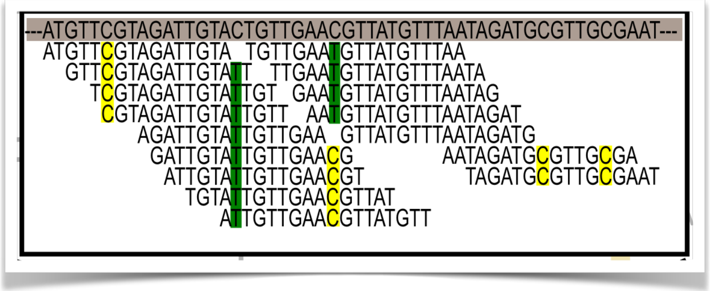


- unmethylated CpG are identified by the presence of a **mismatch TpG**
- 2 approaches:
  - array based: hybridization to CpG probes on array
  - sequencing: whole genome bisulfite-sequencing

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

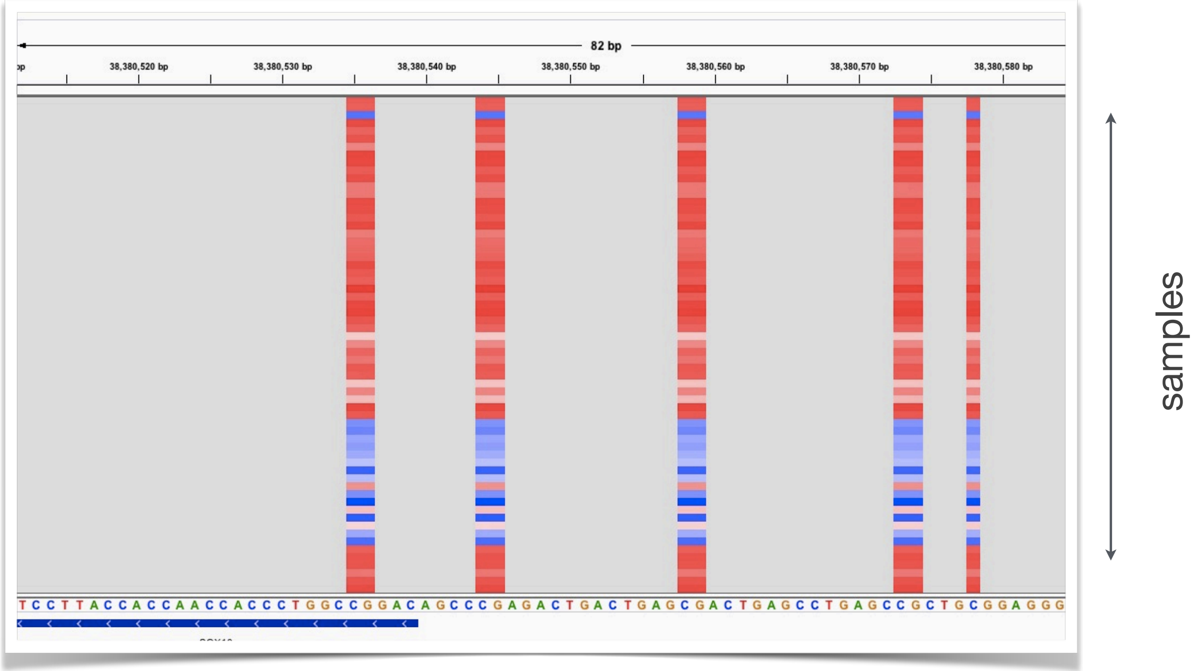
- **Sequencing based methods** (whole-genome bisulfite sequencing WGBS)
  - unmethylated C → T
  - methylated C → C
- Shearing, conversion and sequencing (Illumina X-10)
- Information about the 28 million CpGs



```
---ATGTTCGTAGATTGTACTGTTGAACGTTATGTTTAATAGATGCGTTGCGAAT---
   ATGTTCGTAGATTGTA TGTTGAATGTTATGTTTAA
     GTTCGTAGATTGTATT TTGAATGTTATGTTTAATA
       TCGTAGATTGTATTGT GAATGTTATGTTTAATAG
        CGTAGATTGTATTGTT AATGTTATGTTTAATAGAT
           AGATTGTATTGTTGAA GTTATGTTTAATAGATG
            GATTGTATTGTTGAACG        AATAGATGCGTTGCGA
             ATTGTATTGTTGAACGT         TAGATGCGTTGCGAAT
               TGTATTGTTGAACGTTAT
                  ATTGTTGAACGTTATGTT
```

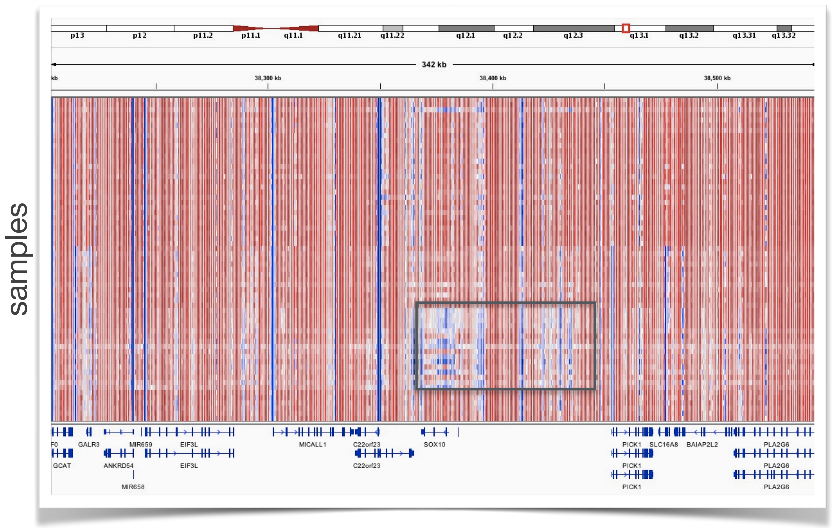
Example DNA methylation

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


82 bp
38,380,520 bp
38,380,530 bp
38,380,540 bp
38,380,550 bp
38,380,560 bp
38,380,570 bp
38,380,580 bp
T C C T T A C C A C C A A C C A C C C T G G C C G G A C A G C C C G A G A C T G A C T G A G C G A C T G A G C C T G A G C C G C T G C G G A G G G
samples

Example DNA methylation

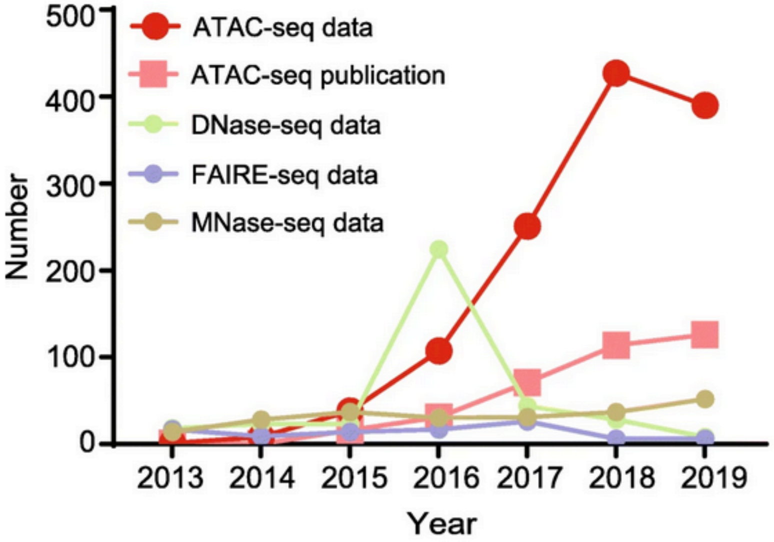
- Genome appears **highly methylated** at a global scale except at gene promoters or CpG islands (blue stripes)
- Some samples have differential methylation (grey box)
- Impact on gene regulation:
  → DNAme is a **repressive mark** which inhibits binding of transcription factors and expression of genes
- in cancer:
  - *hypomethylation* : aberrant expression of oncogenes
  - *hypermethylation* : aberrant repression of tumor suppressors



Chromatin accessibility

Goal: map open and accessible regions in the genome
- **open promoters → active (?) genes**
- **open distal regulatory regions → active enhancers**
- **association with disease genetic variants**

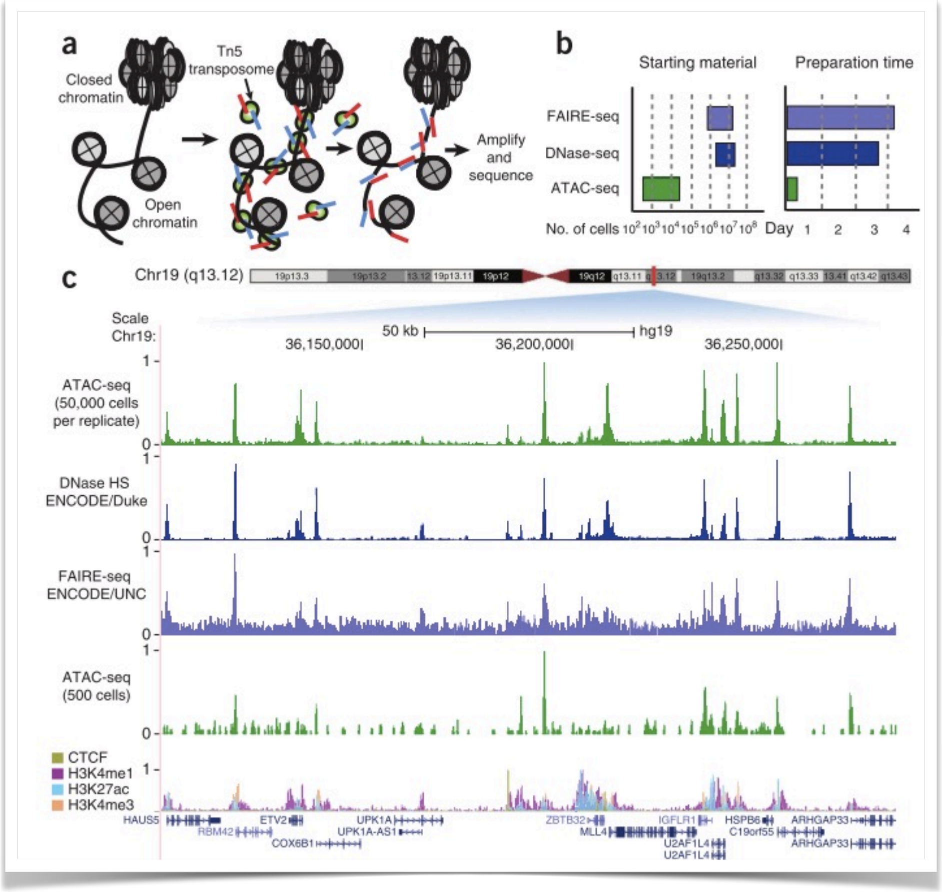
- **DNAse-seq**: fragment the genome using DNAse1 and sequence fragments
- **FAIRE-seq**: cross-linking of nucleosome associated DNA + phase-separation and isolation of nucleosome-free regions + sequencing
- **MNase-seq**: uses a endo-exonuclease to cleave unbound DNA regions + sequencing of regions bound by nucleosome or transcriptions factors
- **ATAC-seq**: uses a Tn5 transposases with sequencing adapters



[Yan et al., Genome Biology (2020)]

Chromatin accessibility

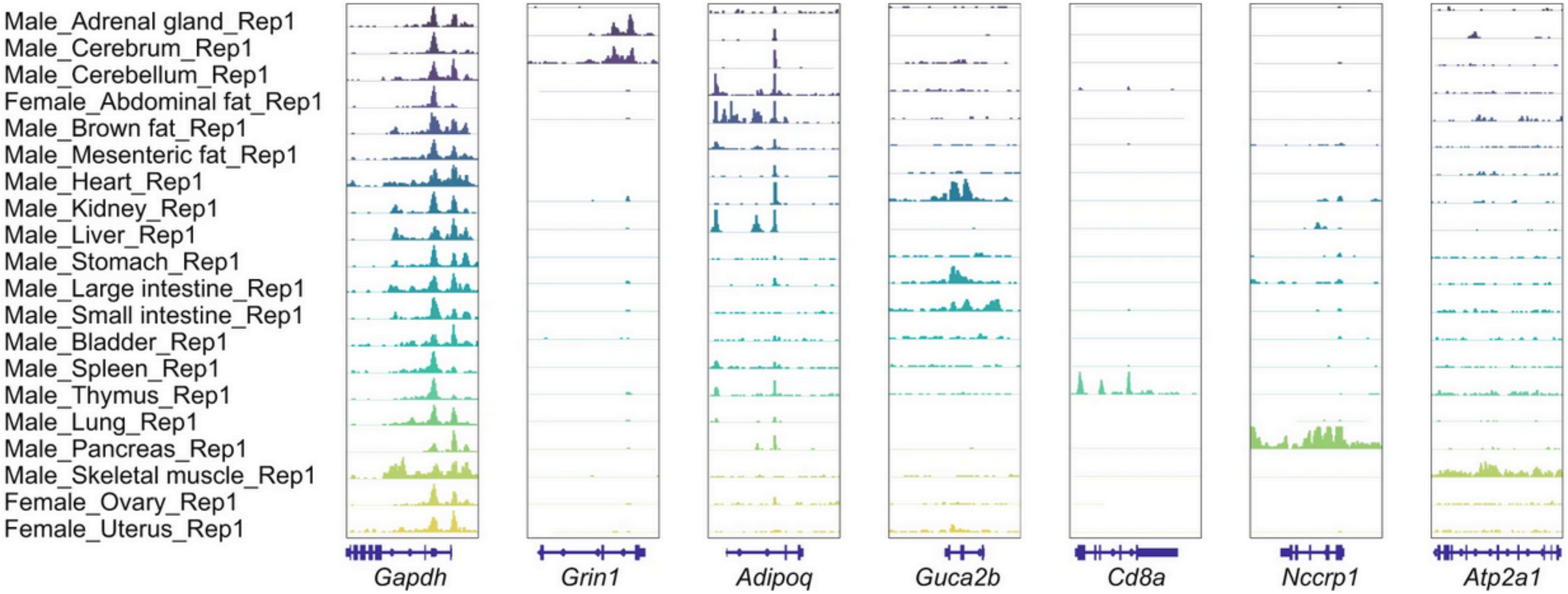
- **ATAC-seq**: using Tn5 transposase prepared with sequencing primers
- requires a small number of input material (~10,000 cells)
- easily adapter to single-cell sequencing
- identification of open chromatin regions (peaks)



[Greenleaf (2013)]

Accessibility atlas

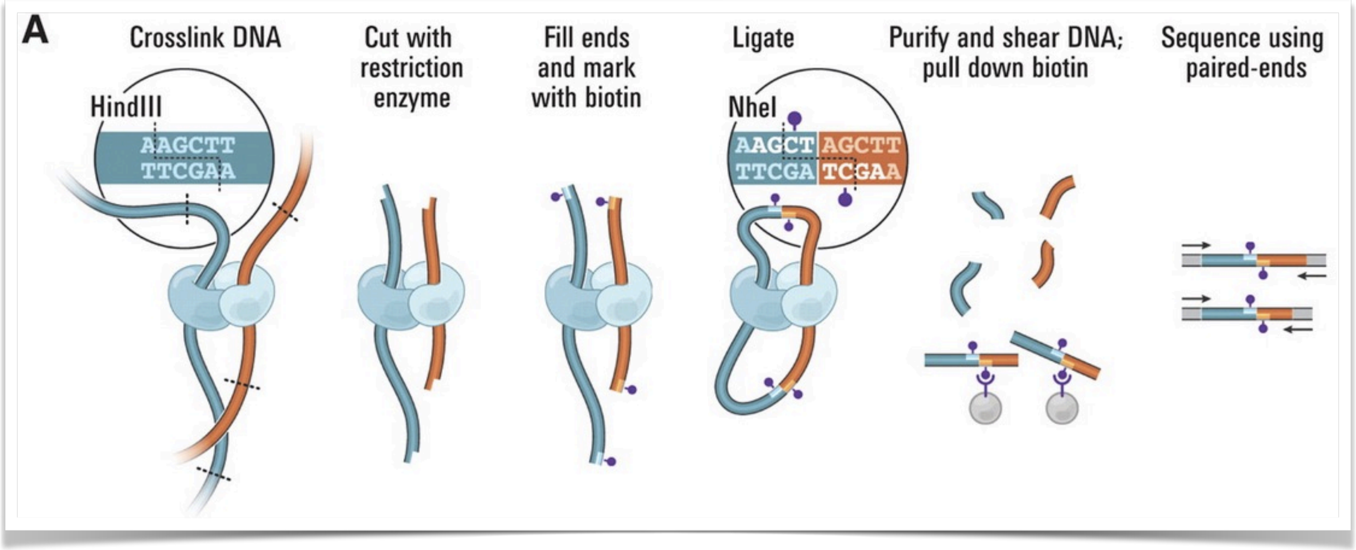
- Patterns of chromatin accessibility are **cell-type specific**



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

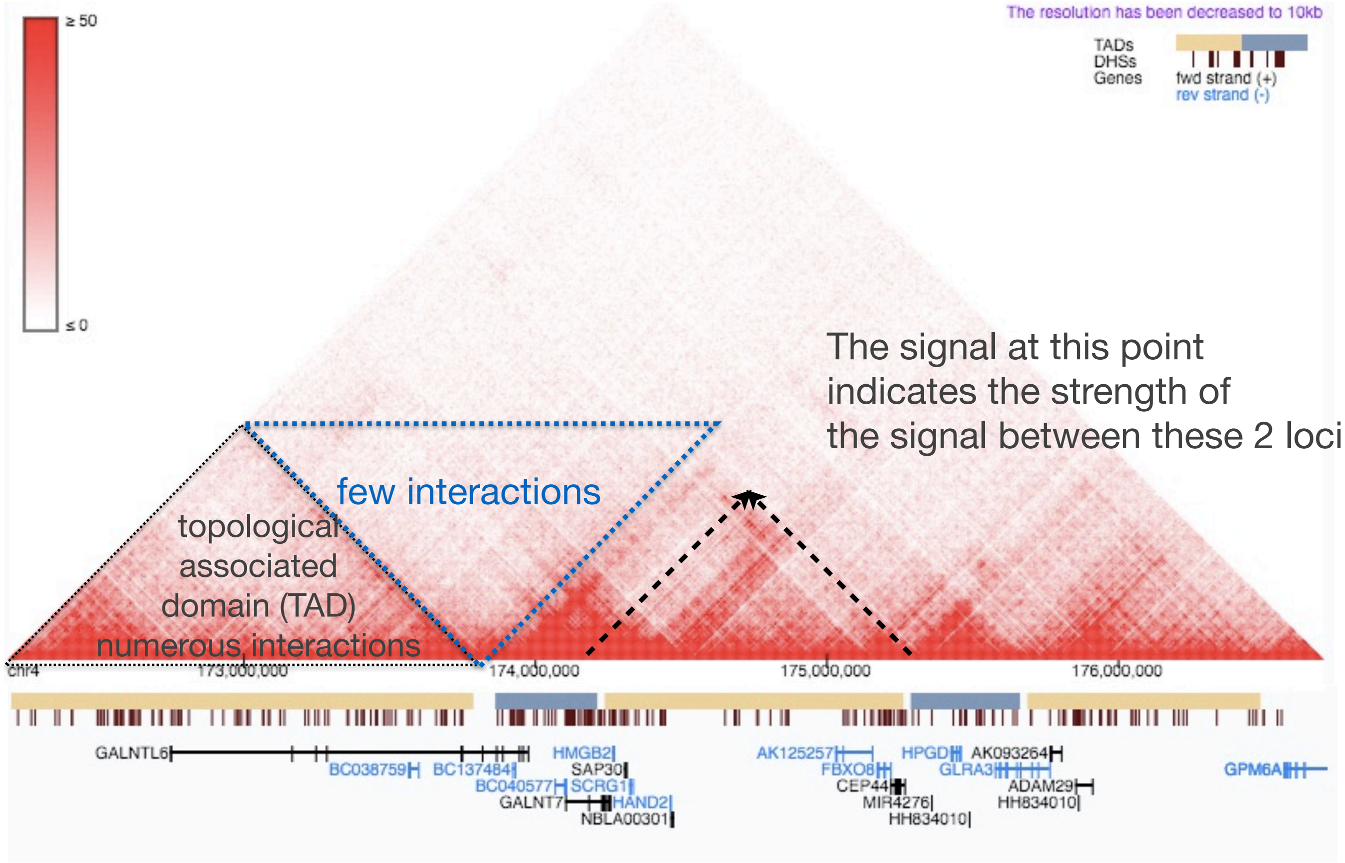
Mapping chromatin interactions

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


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

[Liebermann-Aiden, 2009]

Hi-C and topological domains



few interactions

topological associated domain (TAD) numerous interactions

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

Topological associated domains

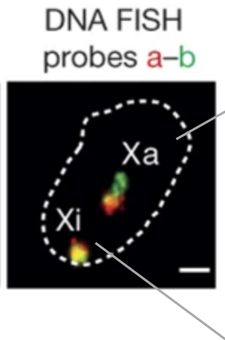
[http://promoter.bx.psu.edu/hi-c/view.php]

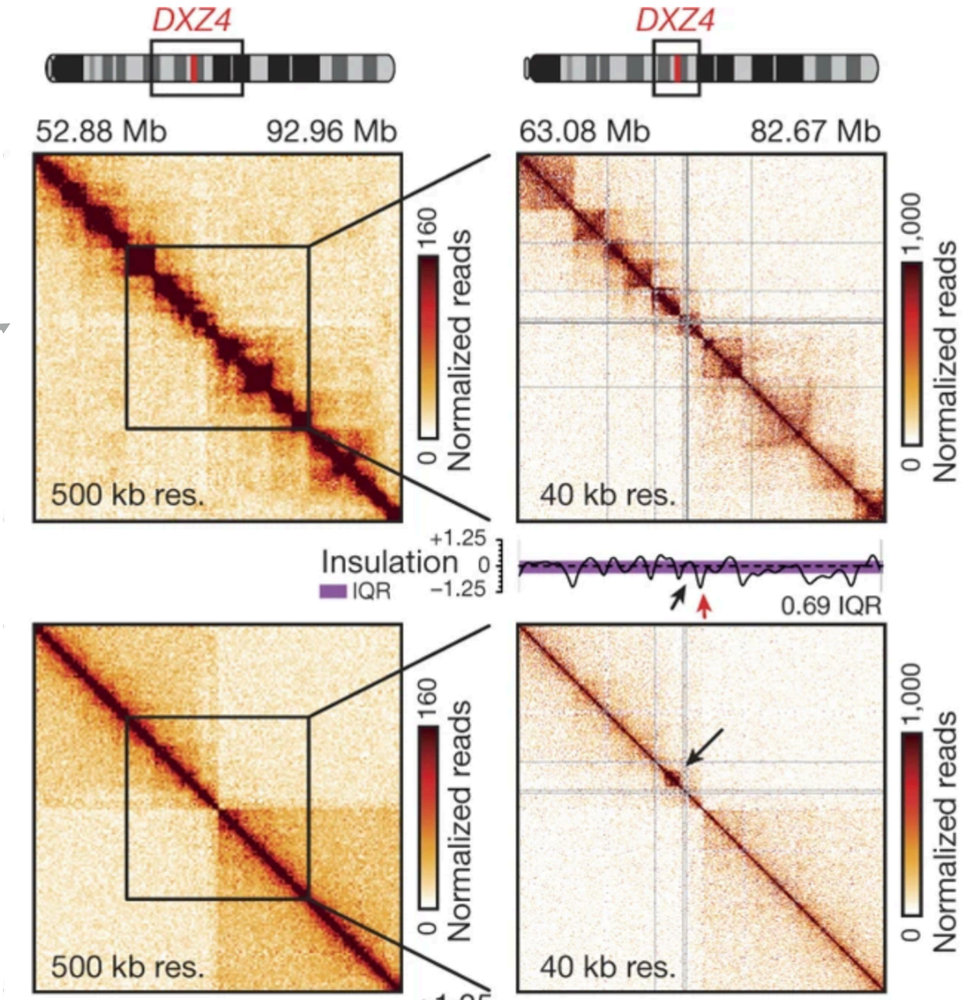
[Dixon (2012,2015)]

Chromatin organization and cell state

Allele specific Hi-C
in neural progenitor cells

Active/inactive X allele

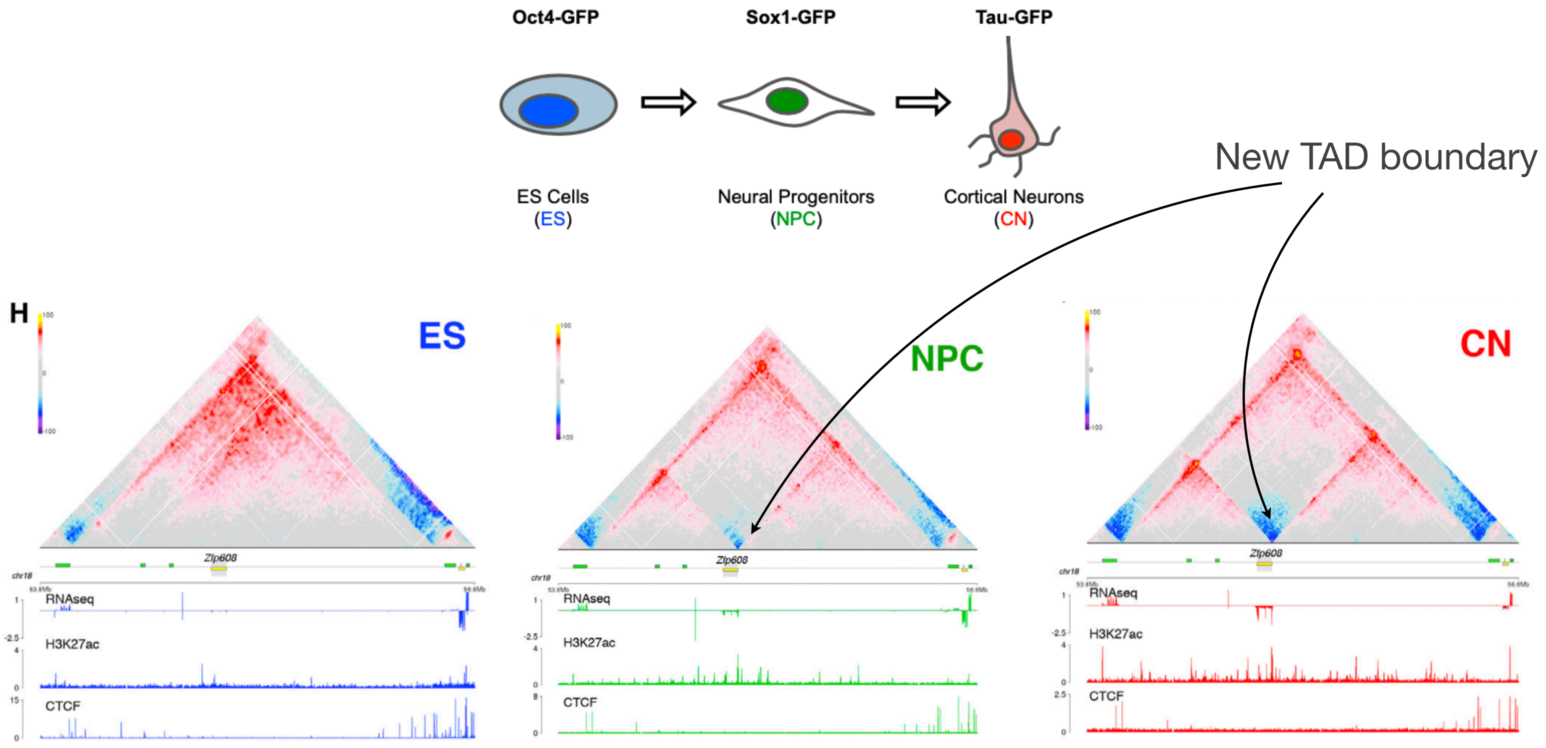




[Georgetti et al., Nature 2016]

Chromatin organization and differentiation

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



New TAD boundary

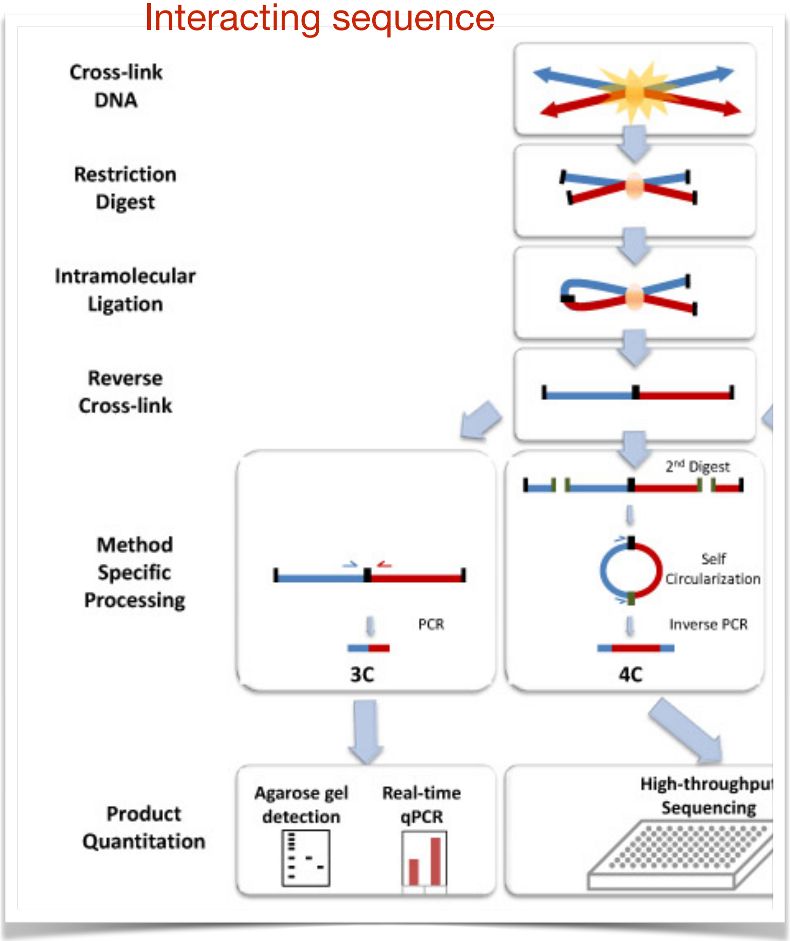
[Bonev et al., Cell (2017)]

Closer look : 4C

- 4C allows to interrogate a **specific locus** ("viewpoint") for all its interaction partners
- used to identify all promoter-enhancer interactions for a given gene

Viewpoint sequence

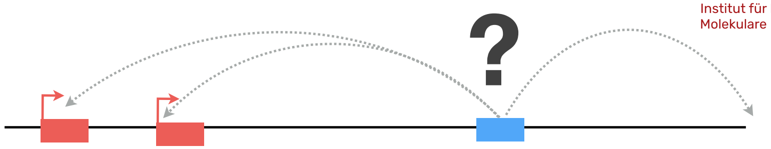
Interacting sequence

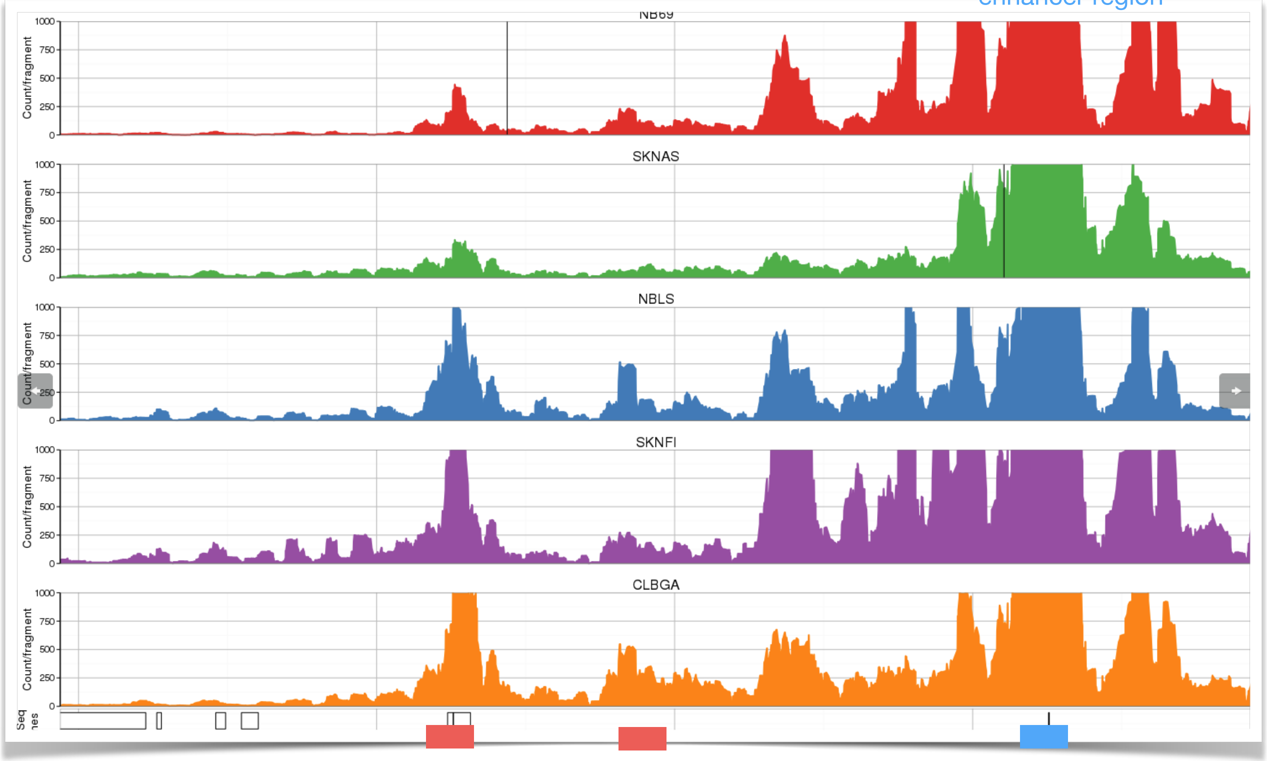


3C = one-vs-one

4C = one-vs-all

Closer look: 4C



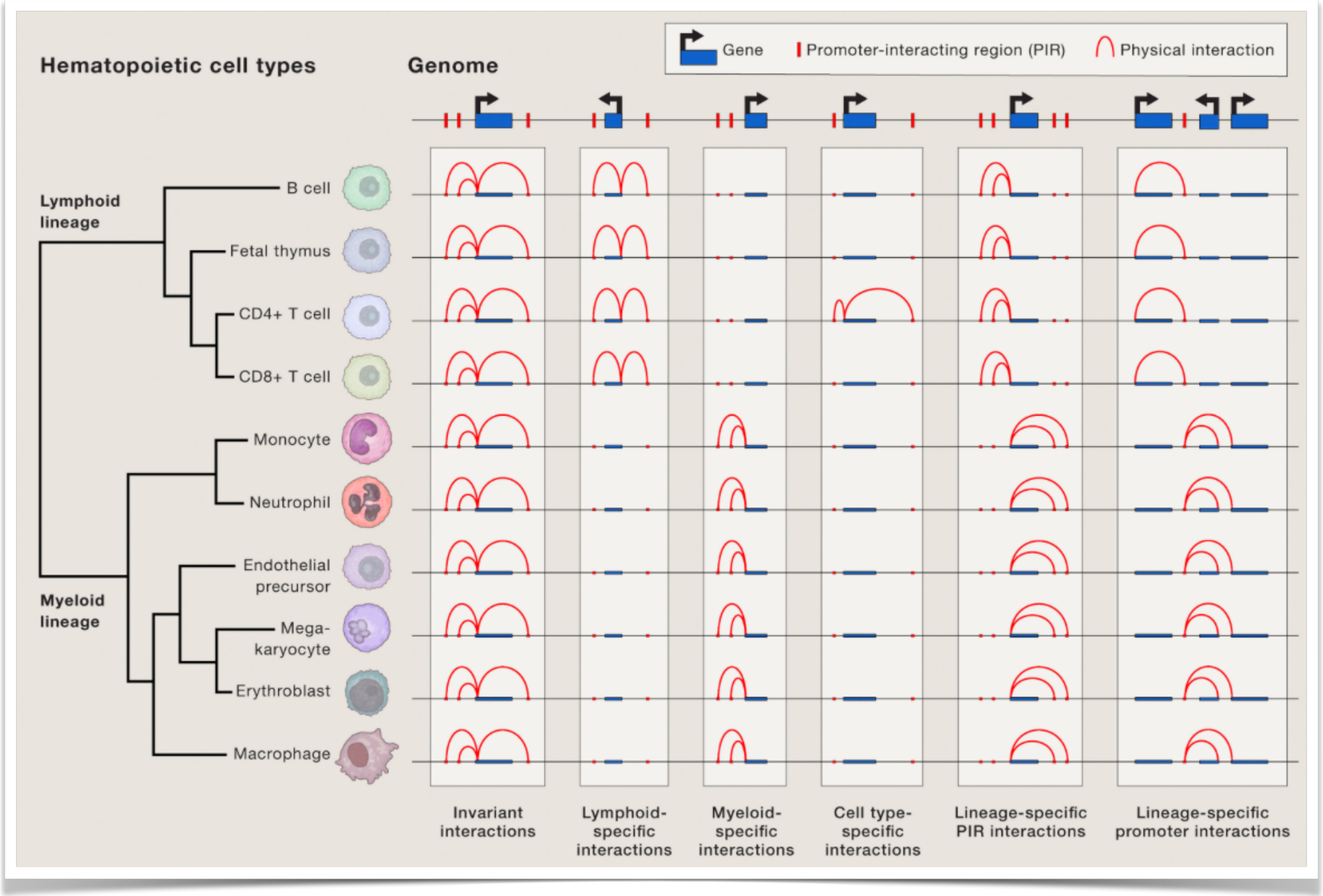


[Dreidax, Gartlgruber (DKFZ)]

Interacting regions
(here: promoters)

View point
(here : enhancer region)

Chromatin interactions shape the cell identity

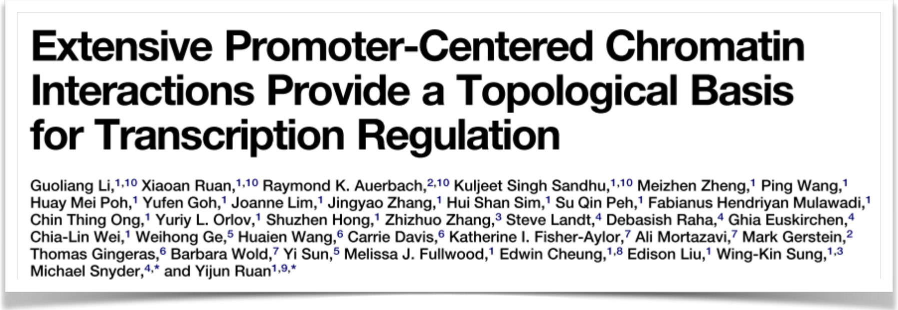


[Spurell et al., Cell 2016]

A complex picture

- Long range interactions show numerous **enhancer-promoter** interactions but also many **promoter-promoter** interactions
- *Hypothesis* : Promoters can act in trans and be used as **enhancers** for distal genes

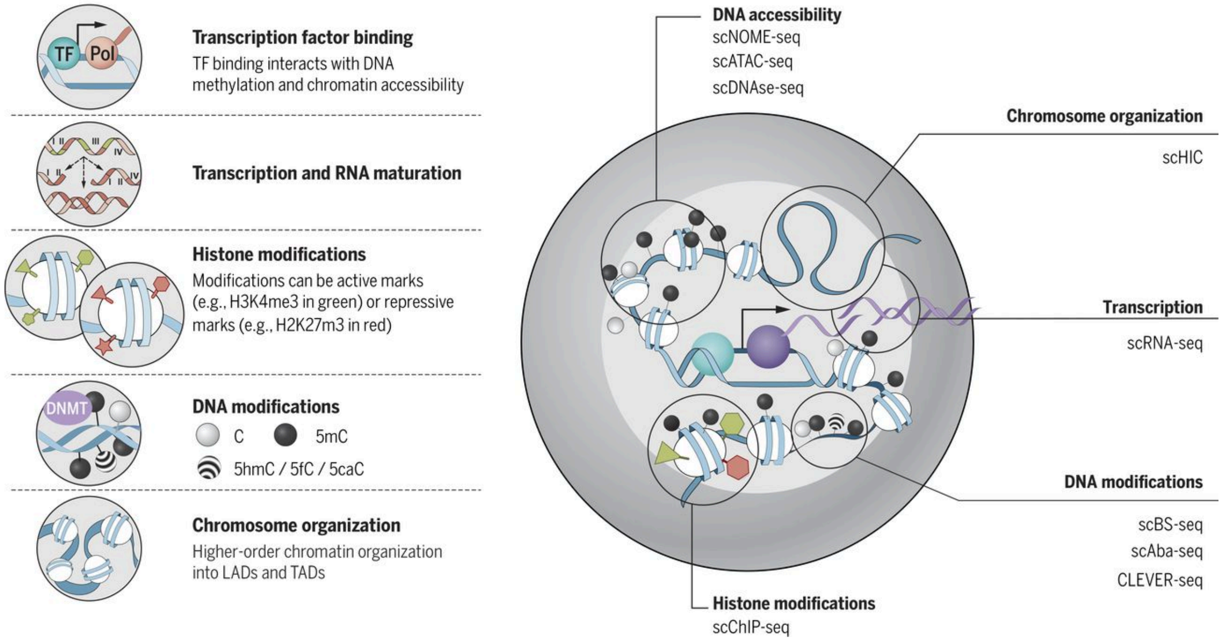
> *"Intriguingly, the multigene complexes illustrated in this study are, in principle, akin to the* ***bacterial operon*** *as a mechanism for* ***coordinated transcriptional regulation of related genes****, suggesting the possibility of a* ***chromatin-based operon mechanism*** *(chro-operon or chroperon) for spatiotemporal regulation of gene transcription in eukaryotic nuclei."*



Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation

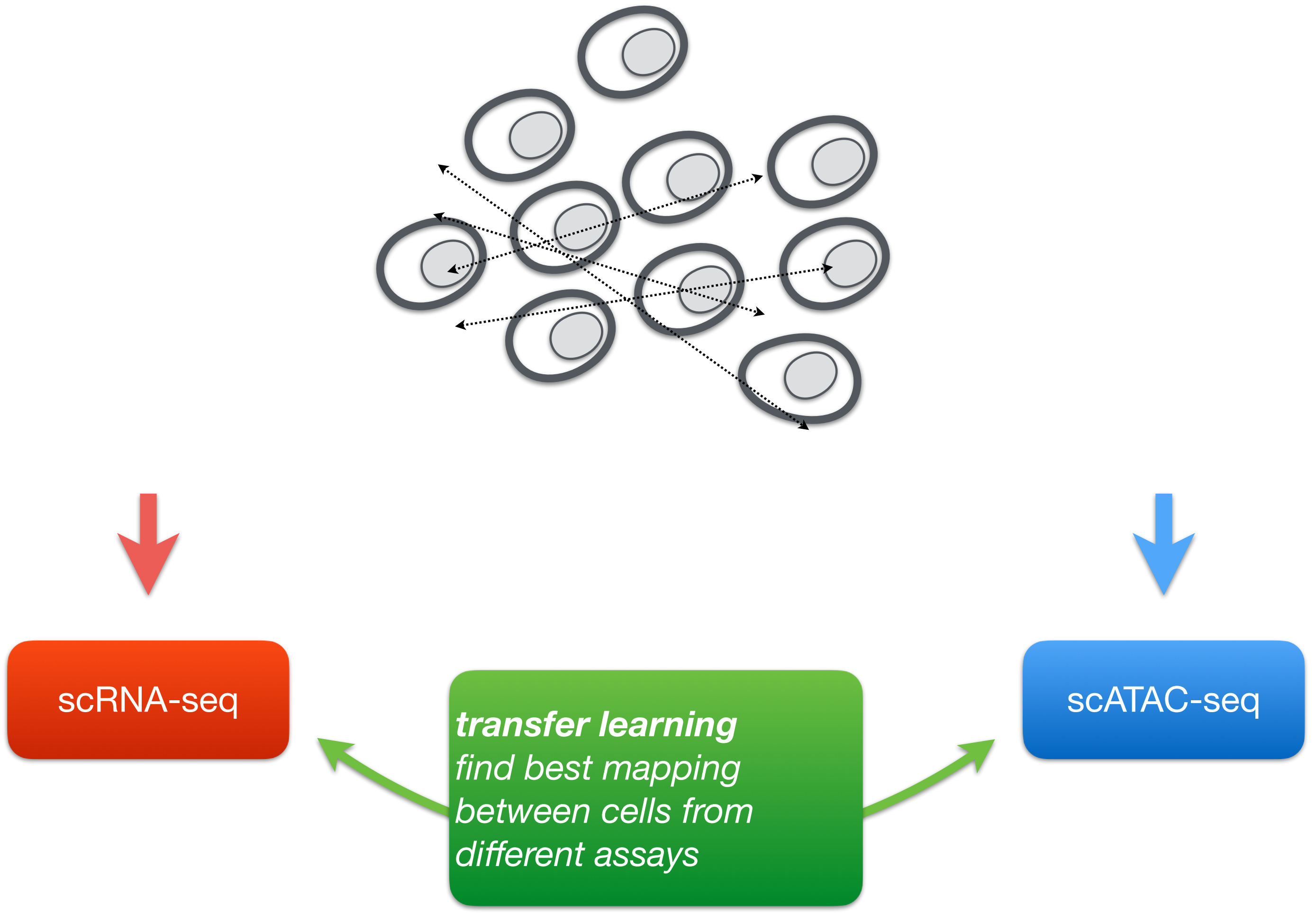
Guoliang Li,[1,10] Xiaoan Ruan,[1,10] Raymond K. Auerbach,[2,10] Kuljeet Singh Sandhu,[1,10] Meizhen Zheng,[1] Ping Wang,[1] Huay Mei Poh,[1] Yufen Goh,[1] Joanne Lim,[1] Jingyao Zhang,[1] Hui Shan Sim,[1] Su Qin Peh,[1] Fabianus Hendriyan Mulawadi,[1] Chin Thing Ong,[1] Yuriy L. Orlov,[1] Shuzhen Hong,[1] Zhizhuo Zhang,[3] Steve Landt,[4] Debasish Raha,[4] Ghia Euskirchen,[4] Chia-Lin Wei,[1] Weihong Ge,[5] Huaien Wang,[6] Carrie Davis,[6] Katherine I. Fisher-Aylor,[7] Ali Mortazavi,[7] Mark Gerstein,[2] Thomas Gingeras,[6] Barbara Wold,[7] Yi Sun,[5] Melissa J. Fullwood,[1] Edwin Cheung,[1,8] Edison Liu,[1] Wing-Kin Sung,[1,3] Michael Snyder,[4,*] and Yijun Ruan[1,9,*]

Single-cell regulatory genomics


TF
Pol
Transcription factor binding
TF binding interacts with DNA methylation and chromatin accessibility
Transcription and RNA maturation
Histone modifications
Modifications can be active marks (e.g., H3K4me3 in green) or repressive marks (e.g., H2K27m3 in red)
DNMT
DNA modifications
C
5mC
5hmC / 5fC / 5caC
Chromosome organization
Higher-order chromatin organization into LADs and TADs
DNA accessibility
scNOME-seq
scATAC-seq
scDNAse-seq
Chromosome organization
scHIC
Transcription
scRNA-seq
DNA modifications
scBS-seq
scAba-seq
CLEVER-seq
Histone modifications
scChIP-seq

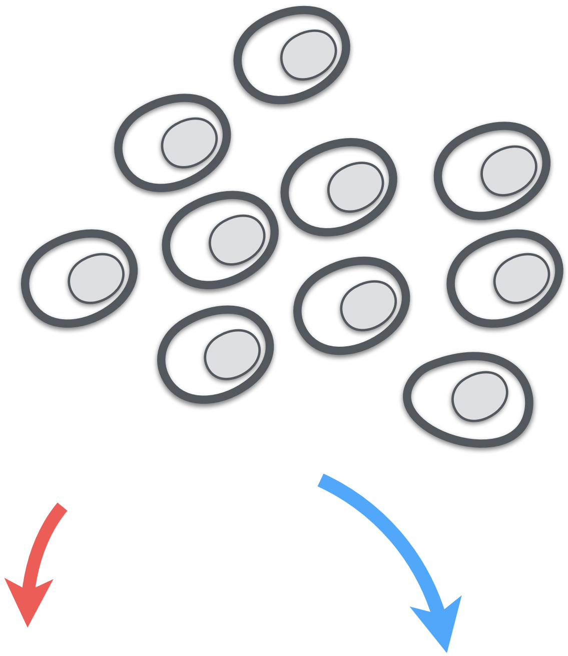
[Kelsey et al., Science (2017)]

Single-cell multi-omics



transfer learning
find best mapping between cells from different assays

Single-cell multi-omics



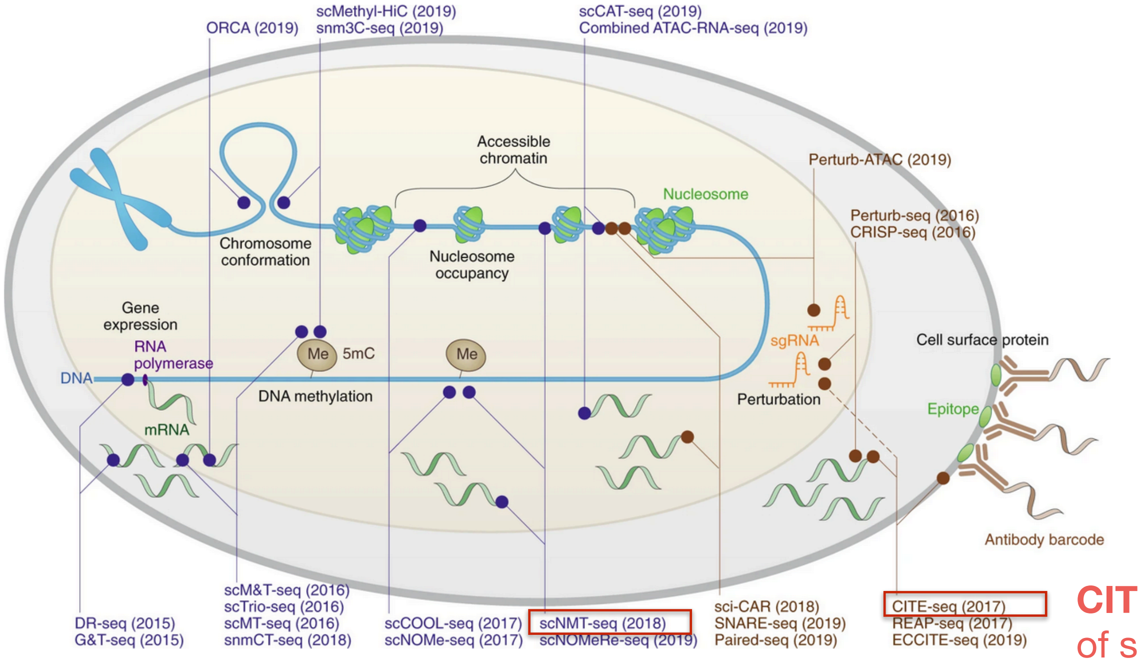
scRNA-seq

scATAC-seq

Accessibility & expression

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

Single-cell multi-omics

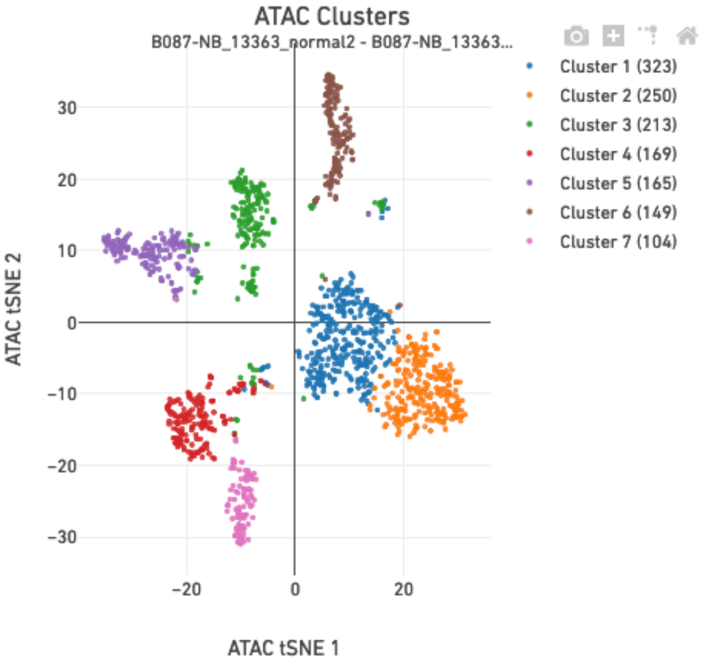


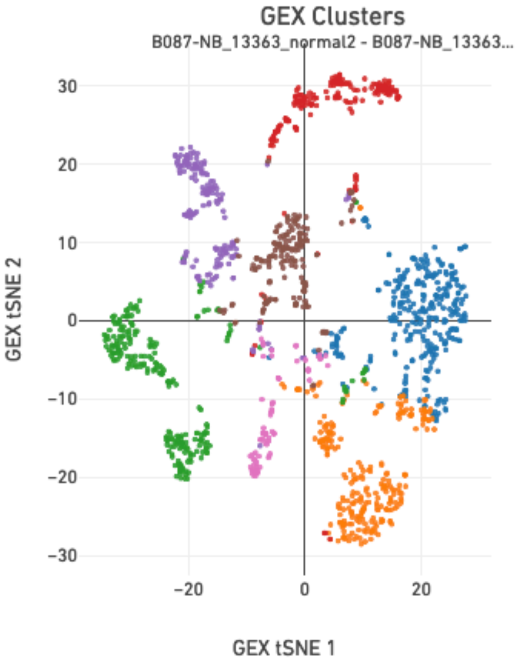
CITE-seq: identification of surface proteins + scRNA-seq

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

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

single-cell Multiome: ATAC / Expression

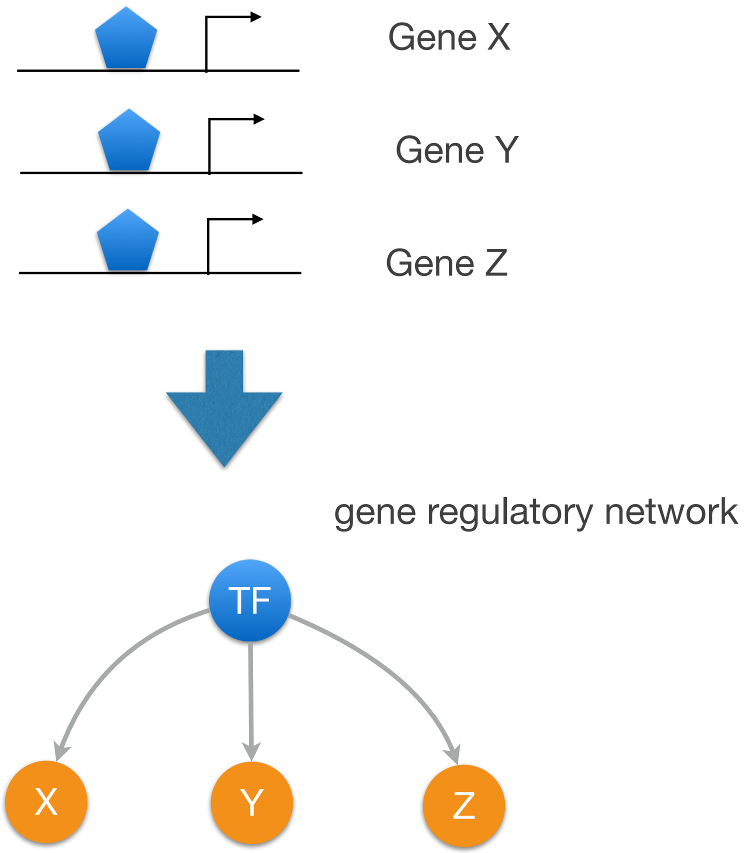




cluster structure is slightly different between scATAC and scRNA!

Gene regulatory networks

- Datasets can be integrated to **predict the list of target genes** for a transcription factor
- Summarized as **gene regulatory networks (GRN)**
- These can be
  - cell type specific
  - dependent on the developmental stage
- Many **computational methods** to infer GRNs
  - based on expression data only
  - based on multi-modal data (expression + chromatin accessibility)
  - based on single-cell data
- **Activity of the transcription factor** can be computed based on the expression of its target genes



Take-home - Part 1

- Transcriptional regulation ensures proper tissue and time-specific expression of genes
- Regulatory elements are located in the non-coding part of the genome (promoters, enhancers,...)
- Larger genomes allow a more complex combinatorial gene regulation (multiple enhancers, ...)
- TR is a complex interplay of multiple components: sequence driven, epigenetic, 3D conformation, ...
- Alterations in any of these components can lead to deregulation and are related to diseases
- Experimental assays allow collection of data on each of these aspects → GRN
- Single-cell genomics allow a cell-type specific view on transcriptionl regulation.