BIO306: Bioinformatics

Lecture 12

Epienetics and epigenome

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

Review last lecture

- What's the advantage of RNA-seq compared with microarray?
- What factors should we consider for RNA-seq data normalization?
- What's the advantage of single cell sequencing over bulk cells?

Origin of the term ‘epigenetics’

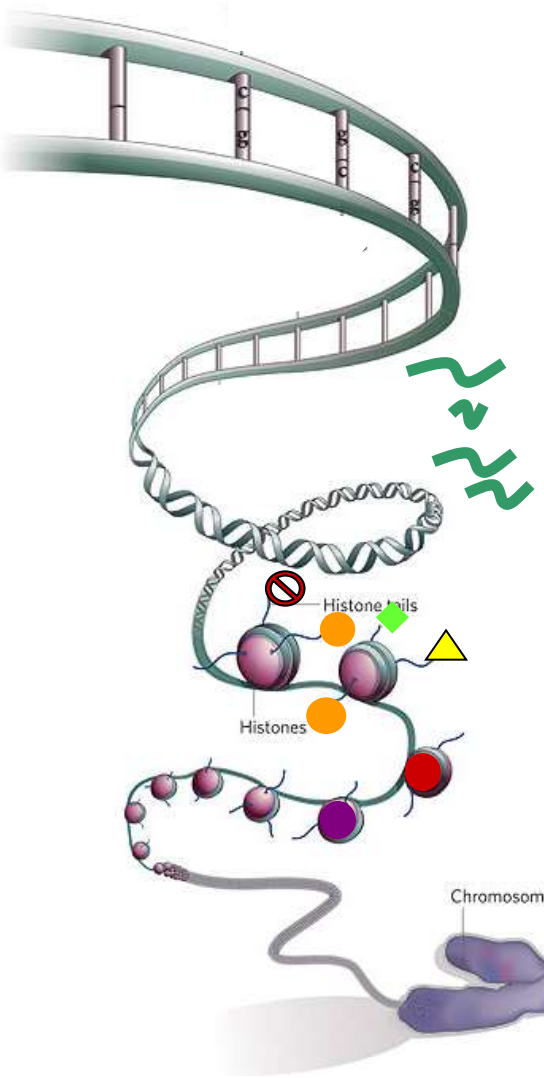
- C. H. Waddington coined the term epigenetics in 1942 as pertaining to epigenesis. Epigenesis in the context of the biology of that period referred to the differentiation of cells from their initial totipotent state in embryonic development.

- When Waddington coined the term, the physical nature of genes and their role in heredity was not known; he used it as a conceptual model of how genes might interact with their surroundings to produce a phenotype; he used the phrase "epigenetic landscape" as a metaphor for biological development.

Definition of epigenetics

- Definition: Epigenetics is the study of heritable changes in gene function that do not involve changes in the DNA sequence.
- Epigenetics most often denotes changes in a chromosome that affect gene activity and expression.
- The standard definition of epigenetics requires these alterations to be heritable, either in the progeny of cells or of organisms.

How epigenomes mark the genome?

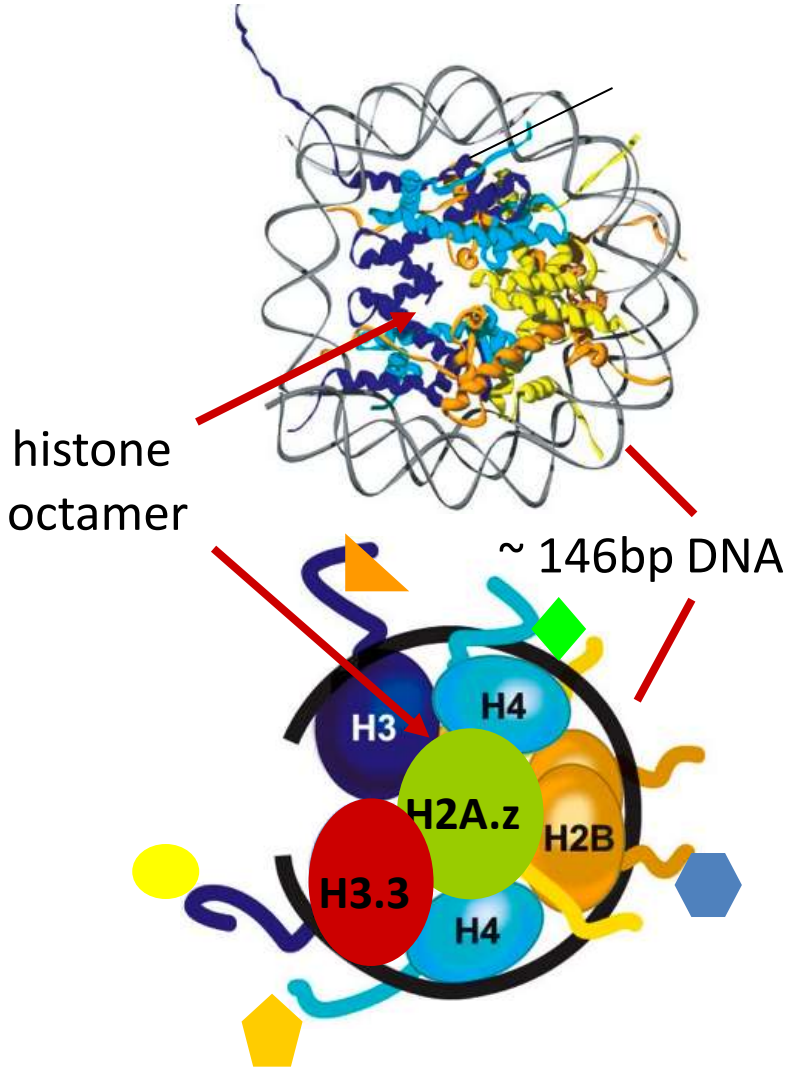


DNA methylation

smRNAs

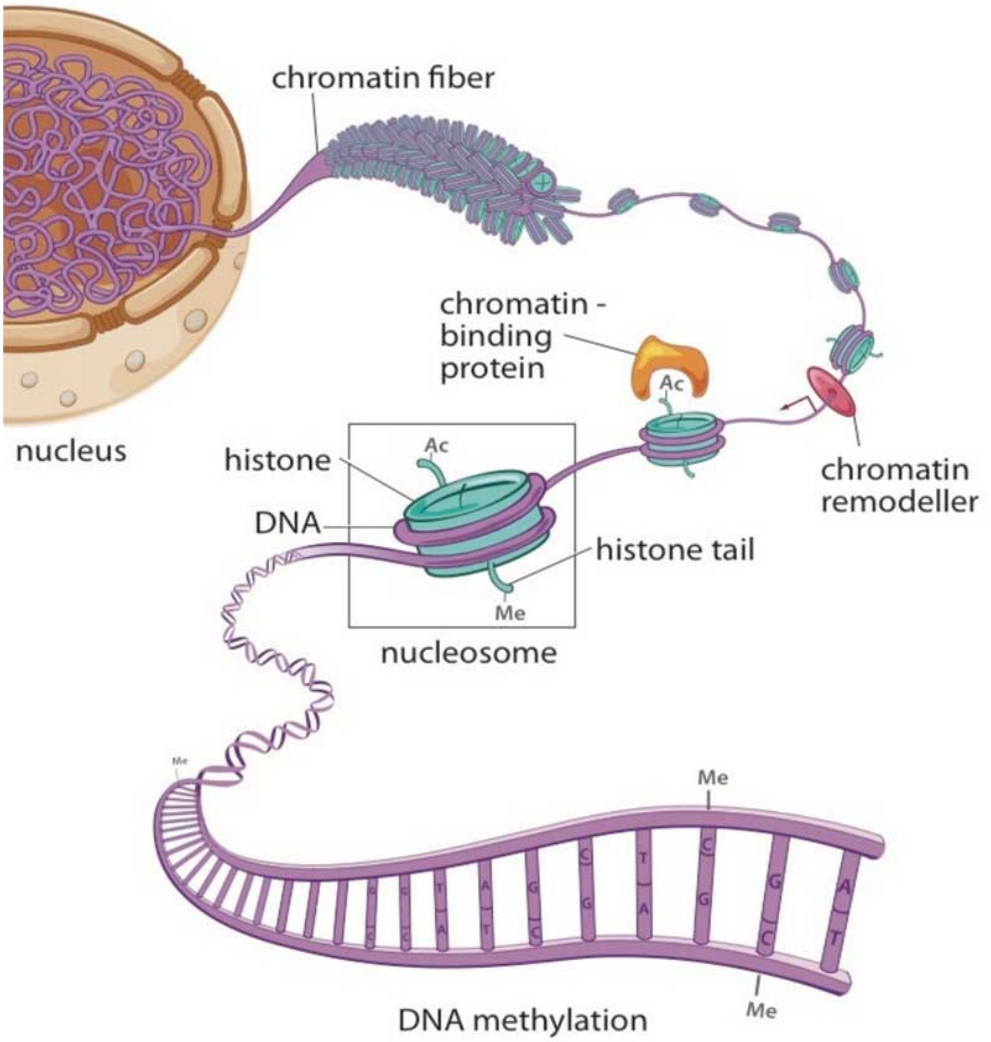
Histone modifications

Histone variants



Qiu et al. Nature (2006)

Epigenetics categories



DNA modification
DNA cytosine methylation

Chromatin related change
Histone modifications
Histone variants
Nucleosome positioning
Chromatin accessibility
Structure of chromatin
Higher order structure

RNA
mRNA modification
Non-coding RNA (smRNAs)

Protein
Prion
Transcription factor binding

From The Cell Biology of Stem Cells (2010)

Epigenetics assays using NGS

- Bisulfite-Sequencing (BS-seq)
  - DNA methylation
- MNase-seq
  - Nucleosome positioning
- Chromatin Immunoprecipitation sequencing (ChIP-seq)
  - Signatures of protein association
  - Histone variants and histone modification
- DNase-seq, ATAC-seq
  - Chromatin accessibility
- Hi-C, CHIA-PET
  - Chromatin architecture

Types of DNA methylation

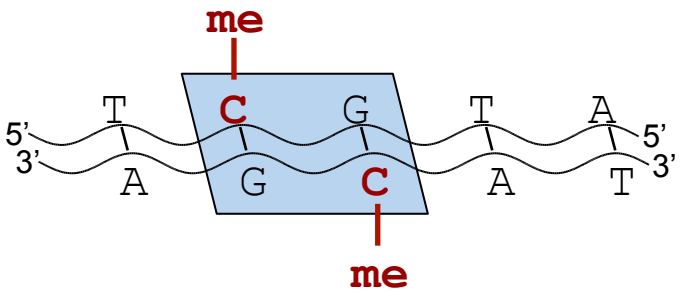
canonical

non-canonical (mammals)

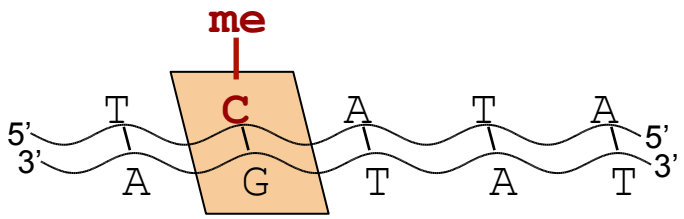
	Plants	Mammals
CG	symmetric	symmetric
CHG	symmetric	asymmetric
CHH	asymmetric	asymmetric

H = T/A/C

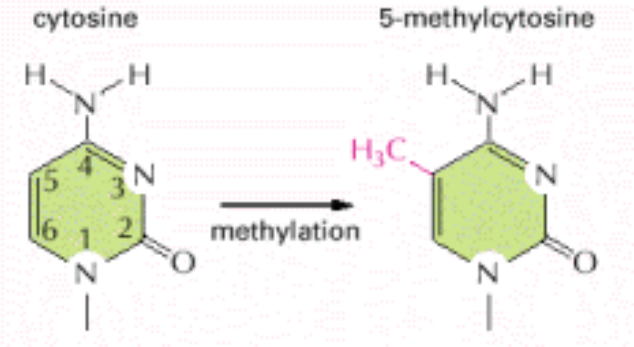
CG context



non-CG context

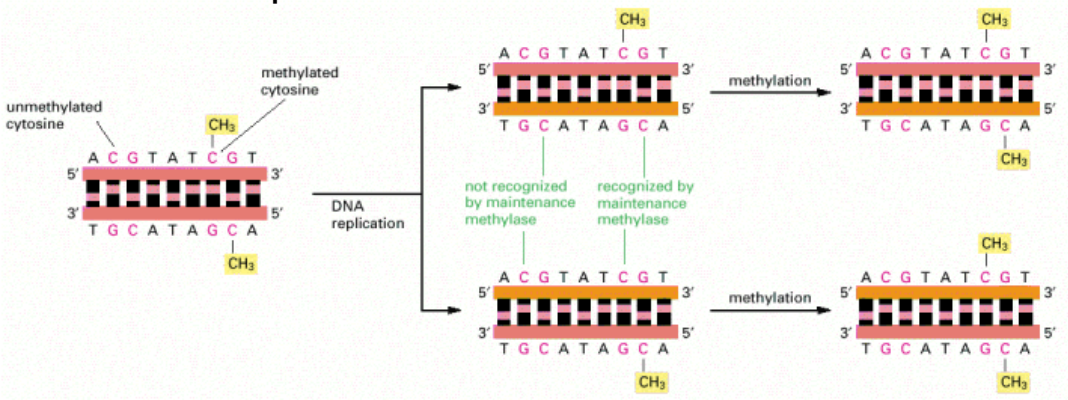


Canonical DNA methylation is inheritable



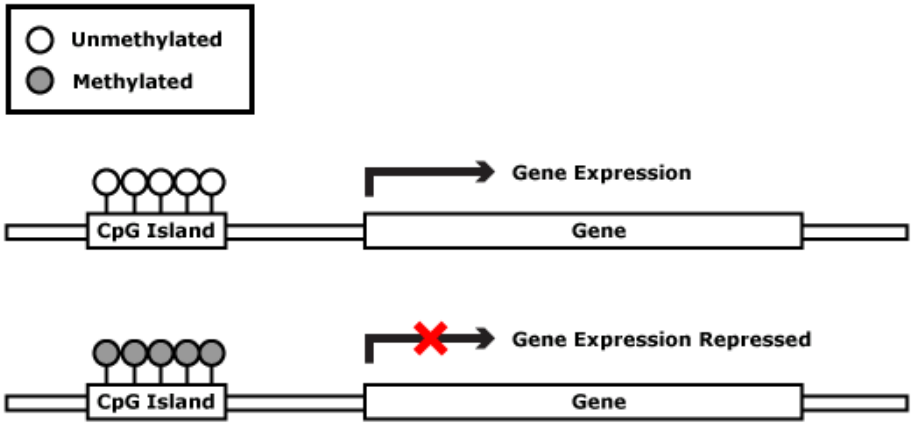
Mammals:

- Cytosine methylation
- Gene regulation
- CpG islands

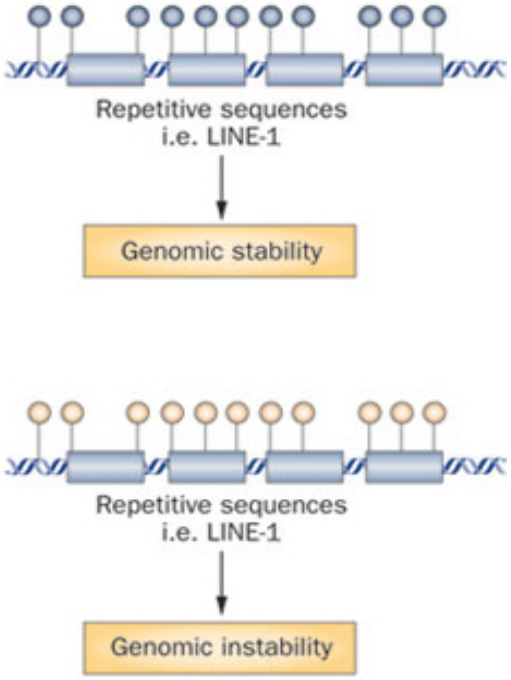


Cytosine methylation is passed on to daughter cells

Regulation by DNA methylation



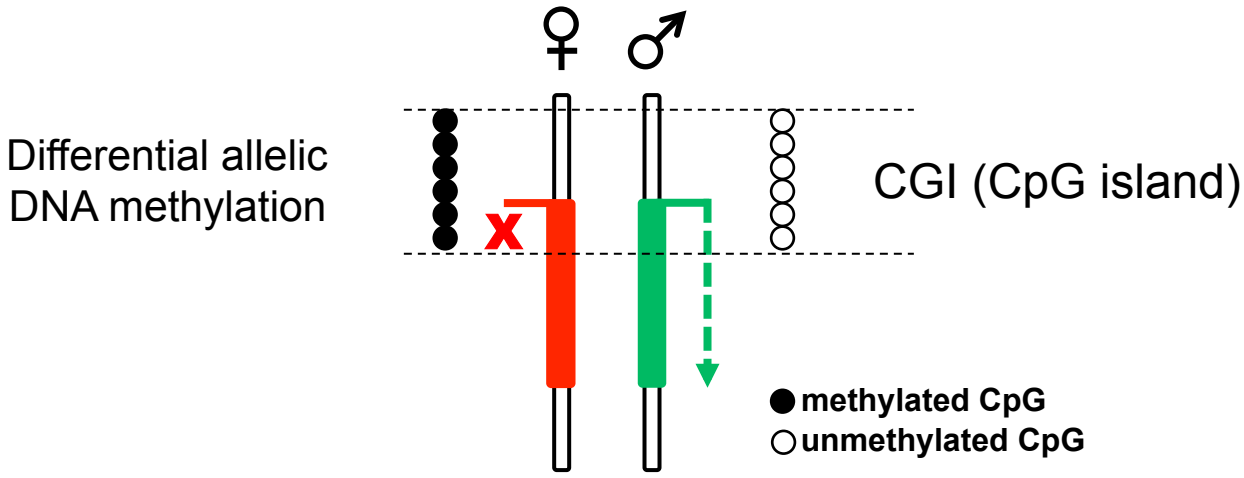
Silencing of gene expression
Tissue differentiation and embryonic development



Repeat activity
Genomic stability

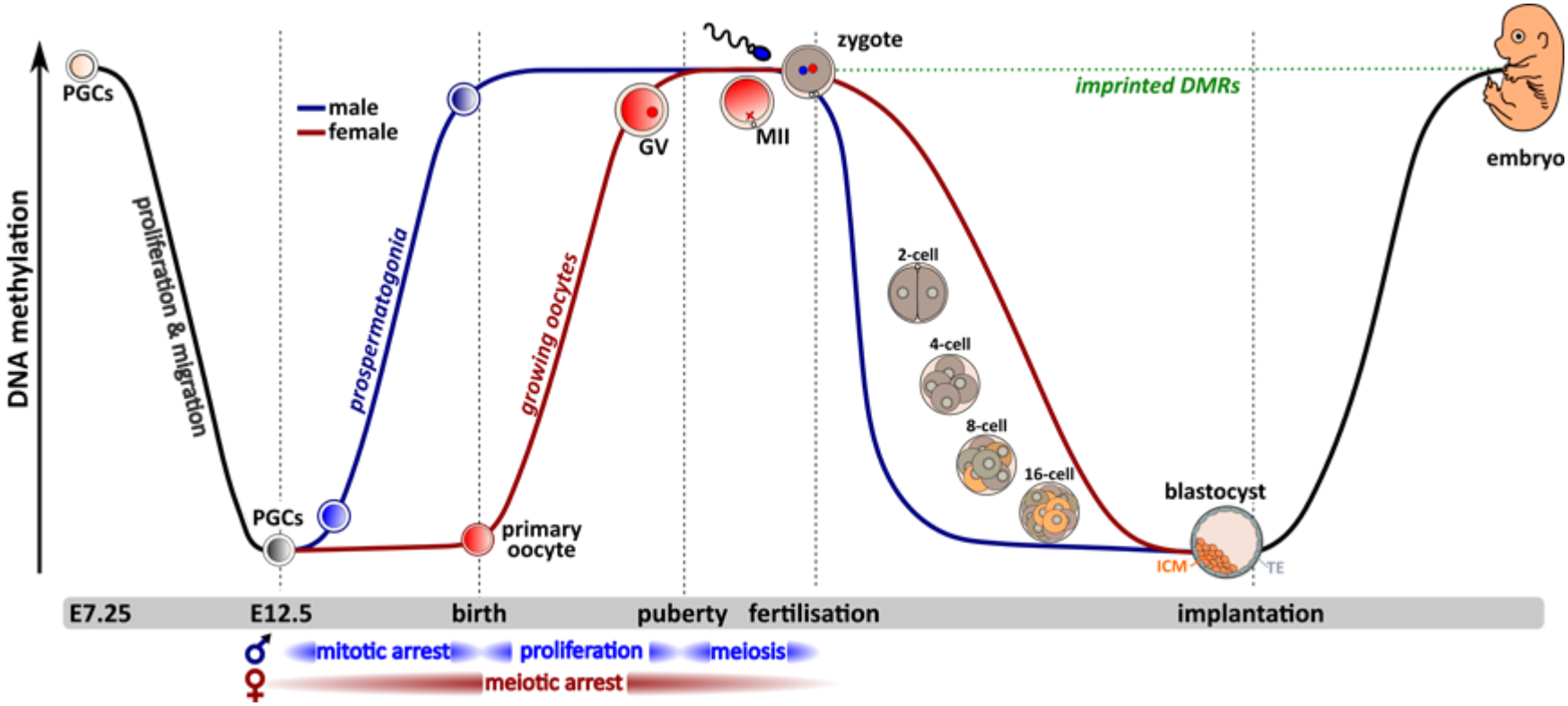
Faults in correct DNA methylation may result in
- early development failure
- epigenetic syndromes
- cancer

Imprinted Genes: mono-allelic expression

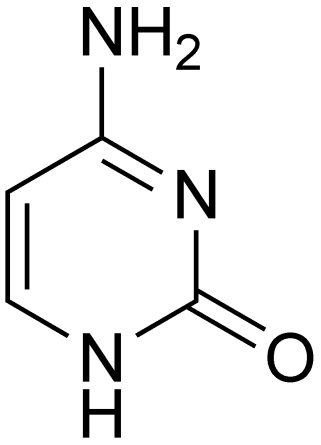


Imprinted Genes: Mono-allelic expression with parent-of-origin specificity.
Have key roles in energy metabolism, placenta functions.

DNA methylation is reset during reprogramming



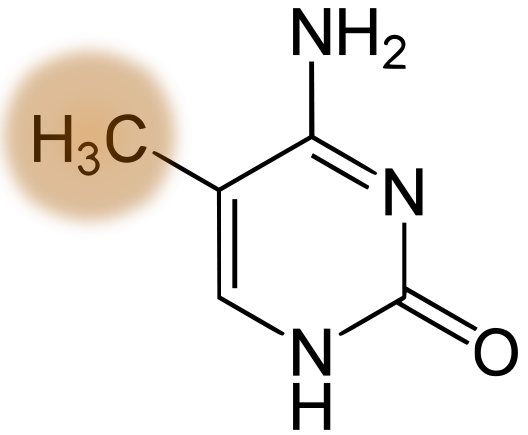
DNA methylation and demethylation



Cytosine

DNA methyl-transferases

DNA-demethylase(s)?
TETs?
Passive demethylation?



5-methyl Cytosine

Measuring DNA methylation by Bisulfite-sequencing

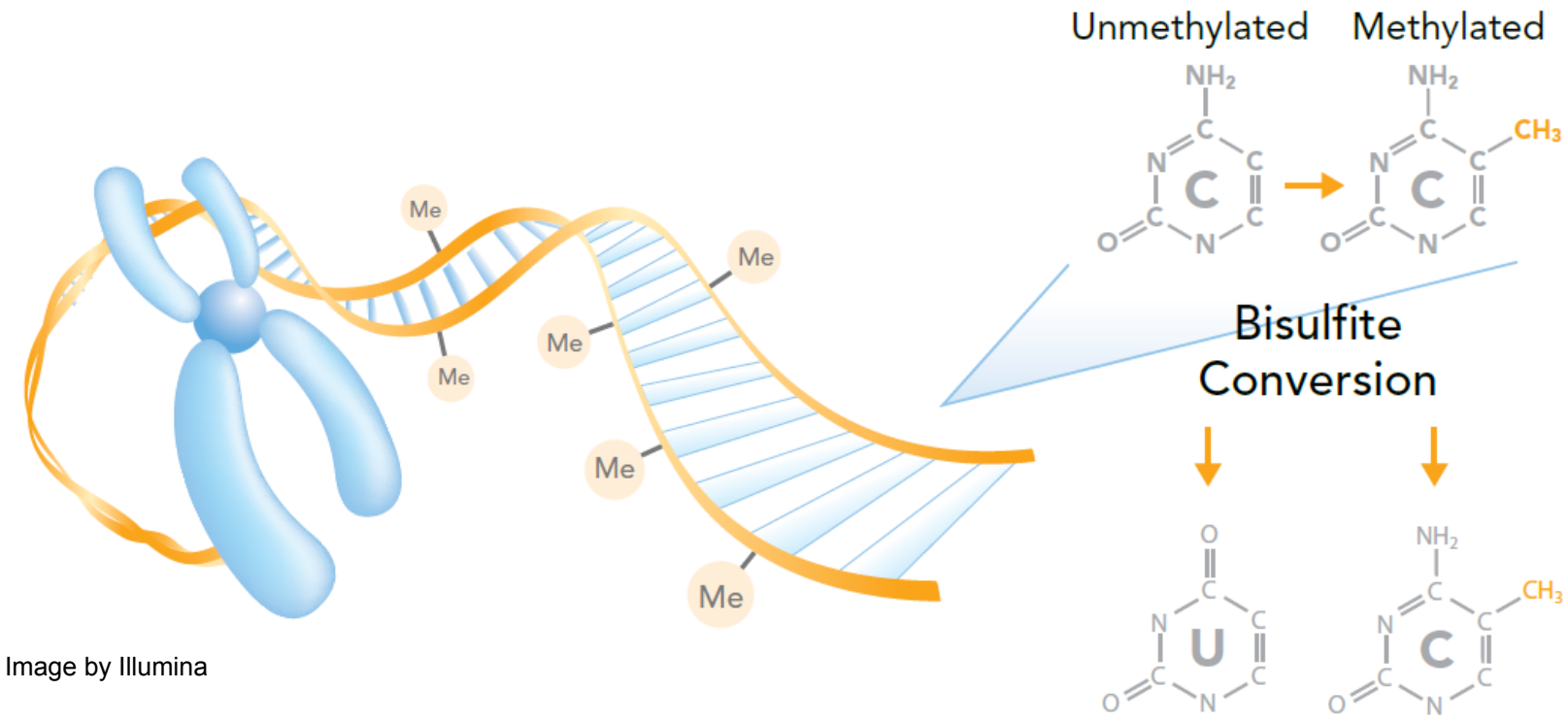
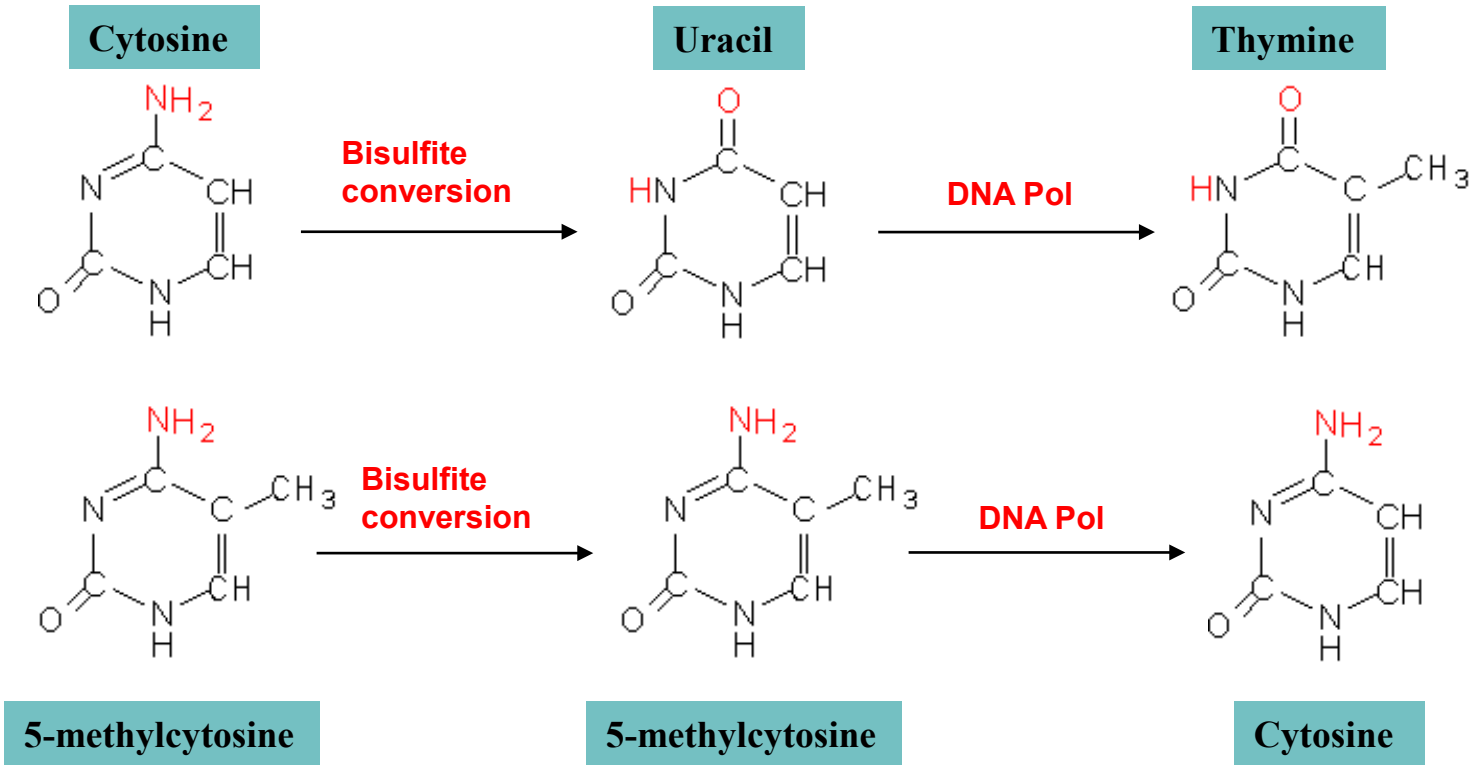


Image by Illumina

DNA Methylation with NGS : Bisulfite Conversion

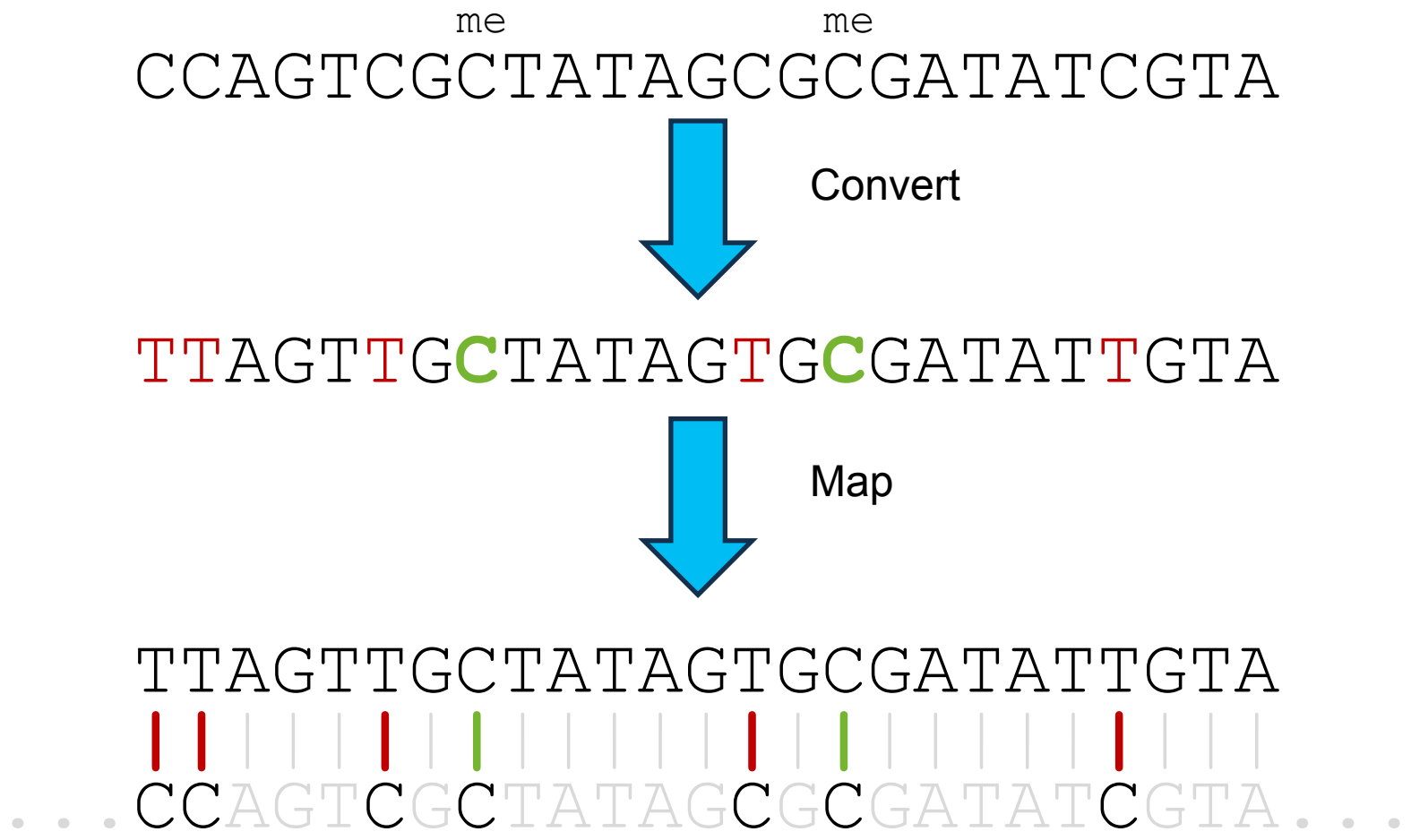


>>A C^m G T T C T C C A G T C>>

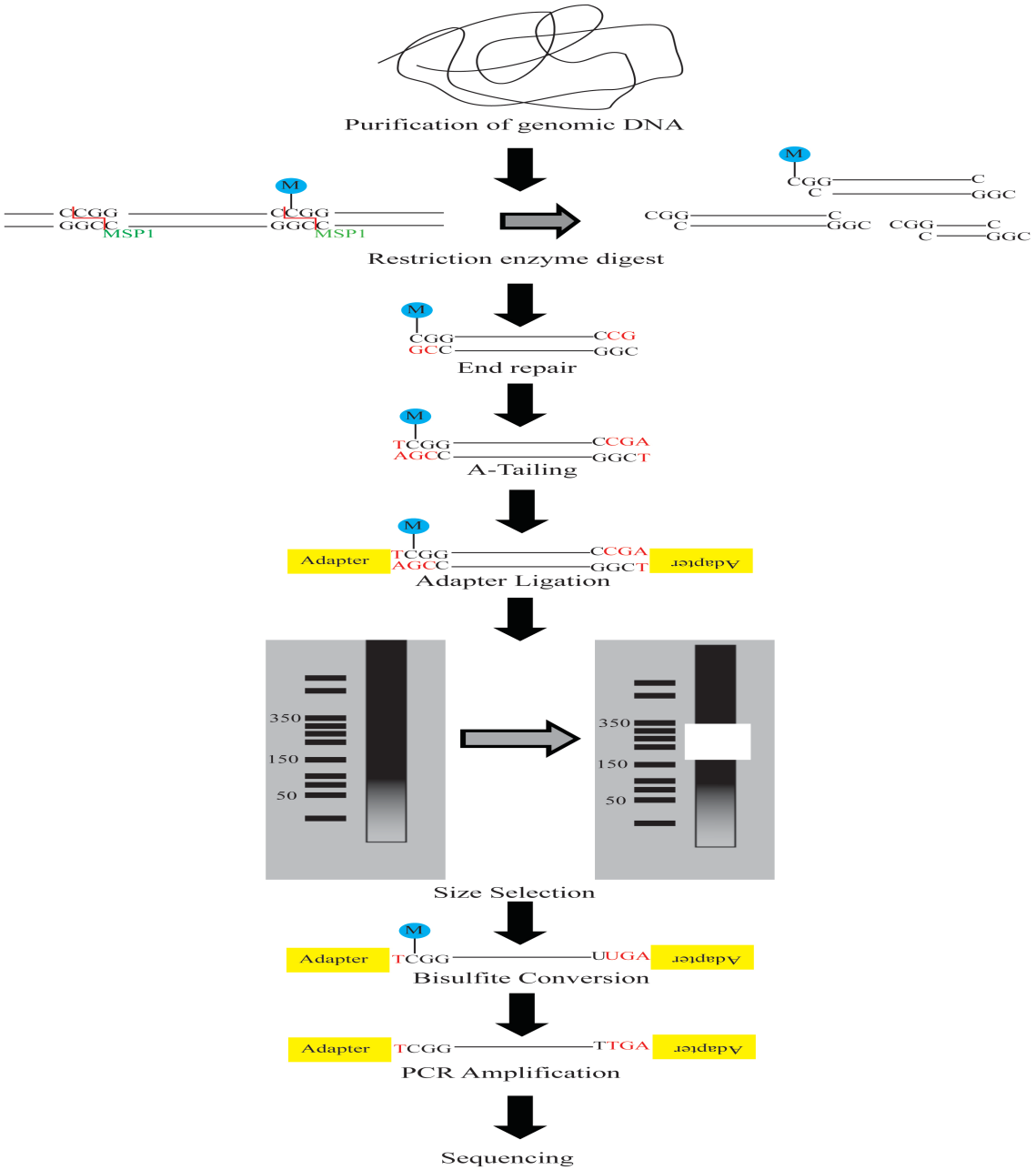
Bisulfite conversion

>>A C^m G T T T T T T A G T T>>

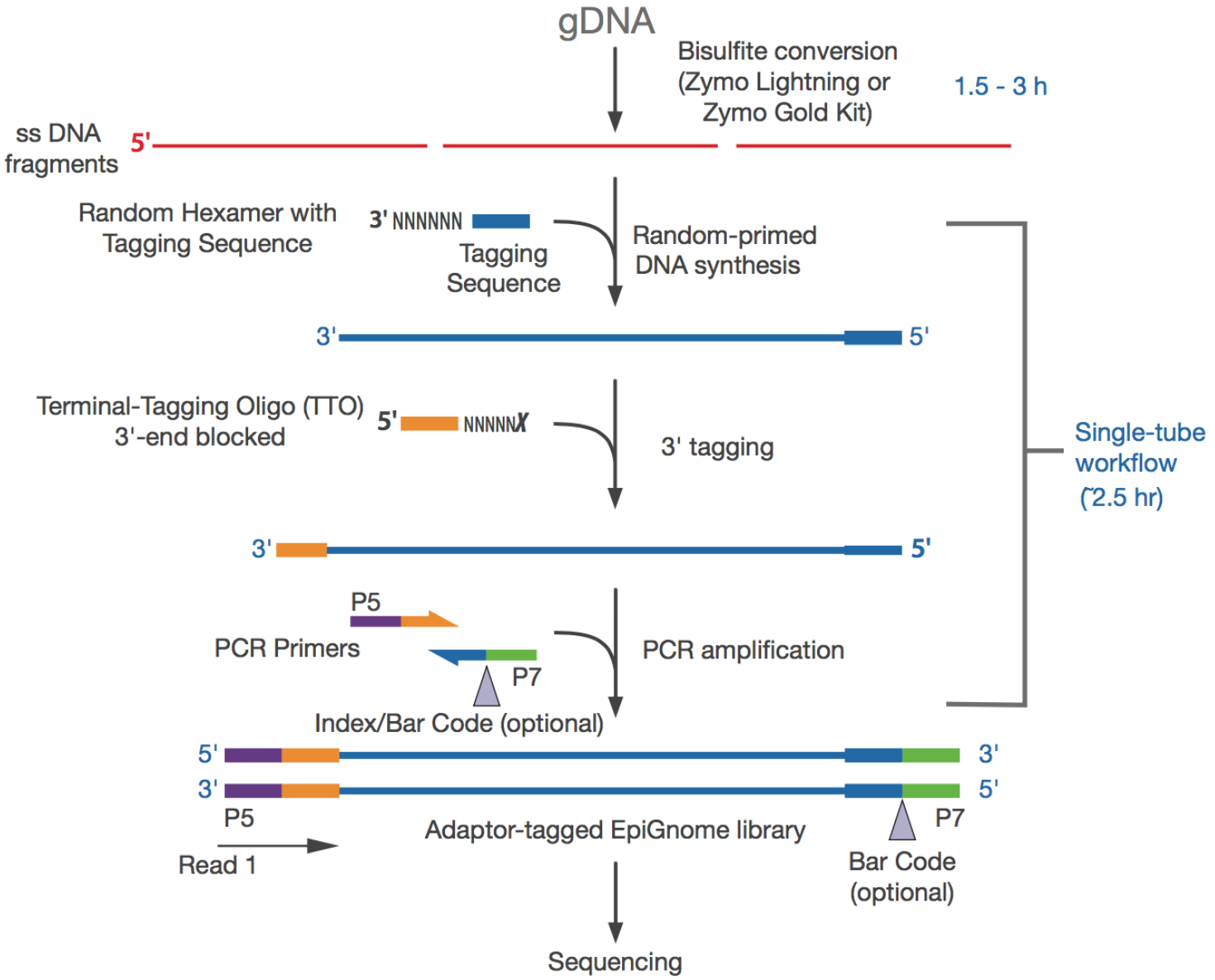
Bisulfite Informatics



Reduced representation BS-Seq (RRBS)

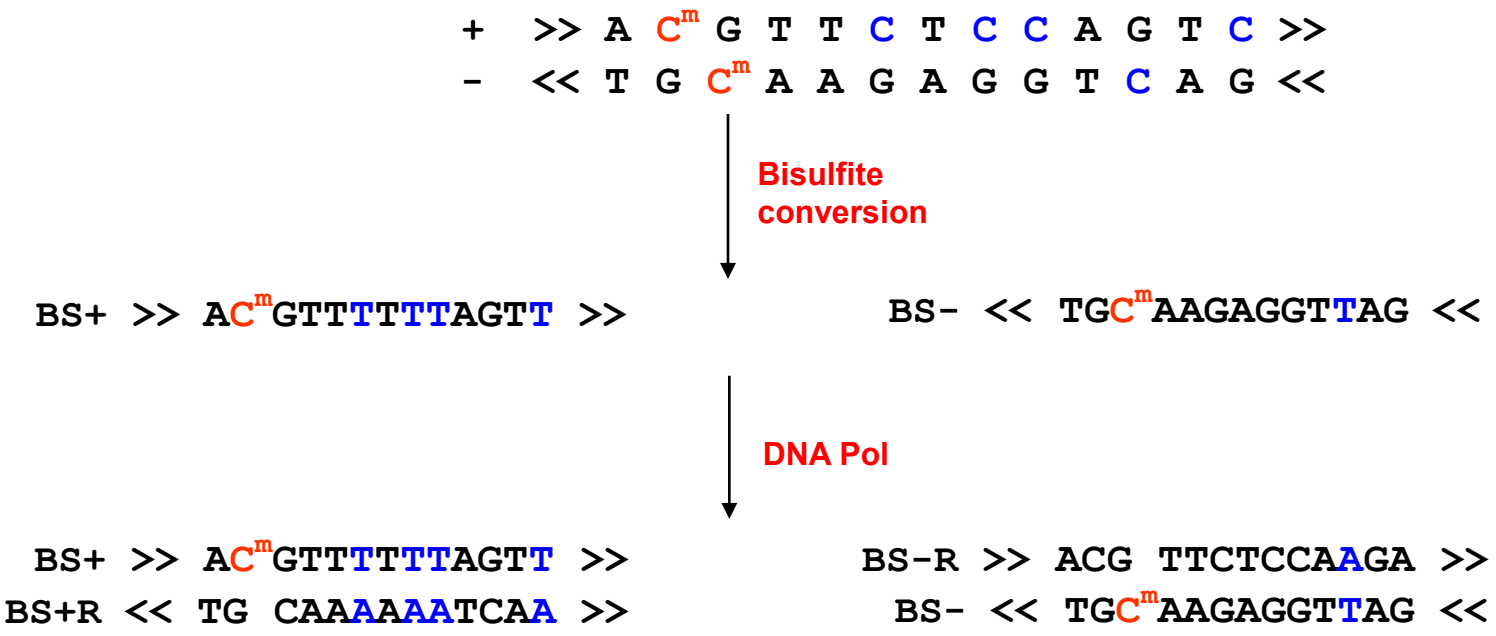


Whole Genome Bisulfite Sequencing (WGBS)



Mapping Bisulfite-converted reads

- Mutated sequence, increased search space



- Specialized aligners
  - BISMA, BSMAP, BS Seeker, MAQ bisulfite mode

Bismark workflow

Pre Alignment

FastQC — Initial quality control

Trim Galore — Adapter/quality trimming using Cutadapt; handles RRBS and paired-end reads; Trim Galore and RRBS User guide

Alignment

Bismark — Output BAM

Post Alignment

Deduplication — optional

Methylation extractor — Output individual cytosine methylation calls; optionally bedGraph or genome-wide cytosine report

M-bias analysis

bismark2report — Graphical HTML report generation

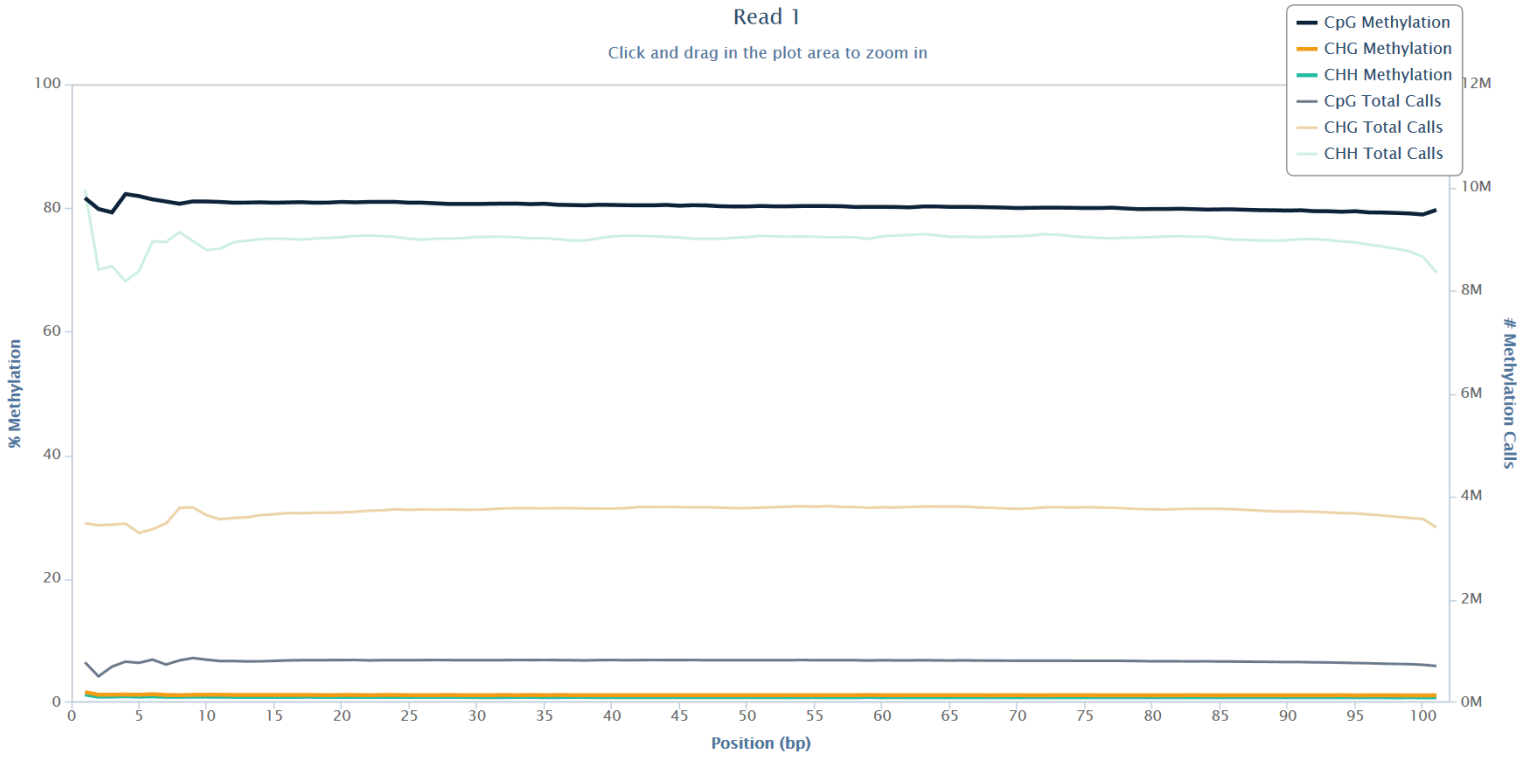
Example: http://www.bioinformatics.babraham.ac.uk/projects/bismark/PE_report.html

Epigenesys

protocol: *Quality Control, trimming and alignment of Bisulfite-Seq data*

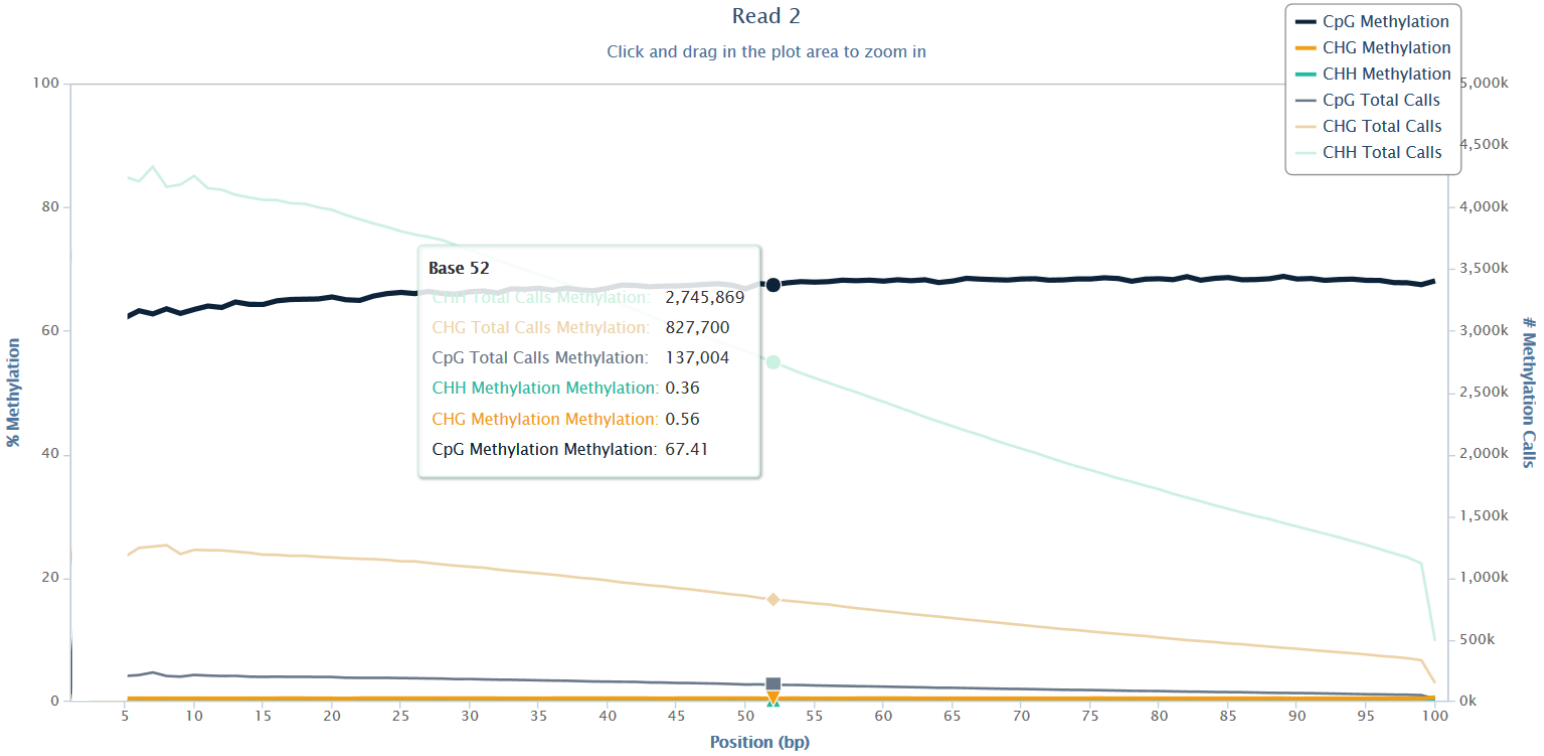
Mapped QC - Methylation bias

M-Bias Plot



good opportunity to look at conversion efficiency

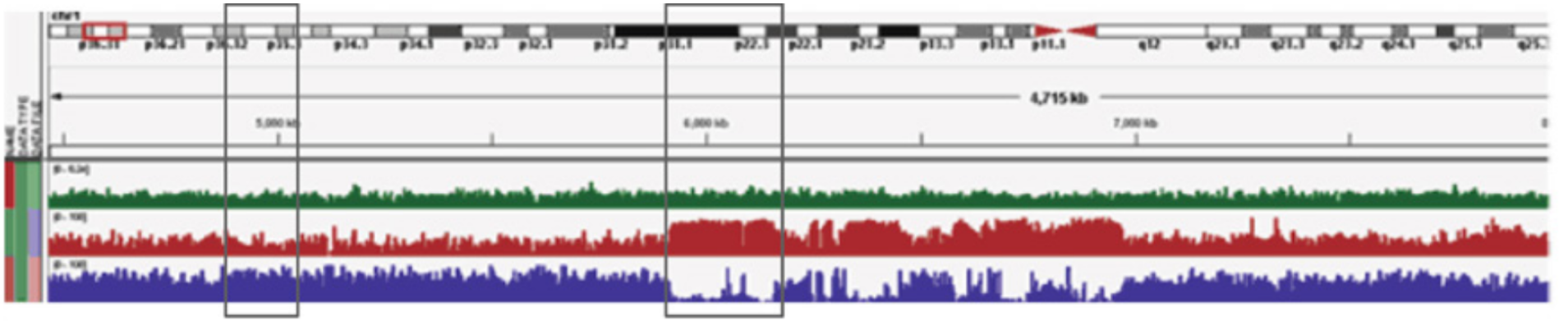
Artificial methylation calls in paired-end libraries



end repair + A-tailing

```
5′-        GGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCA        -3′
3′-        ACCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGG        -5′
```

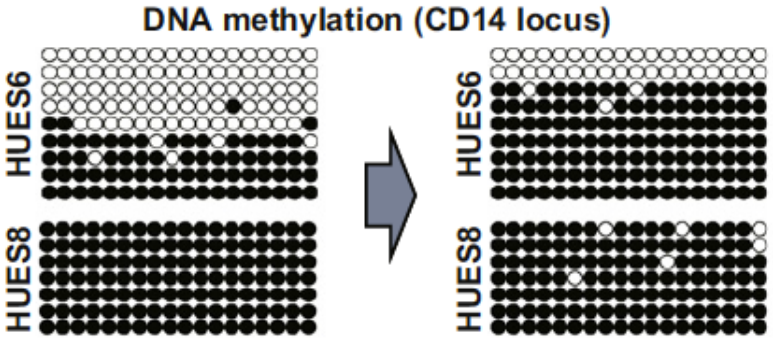
CpG Methylation Coverage

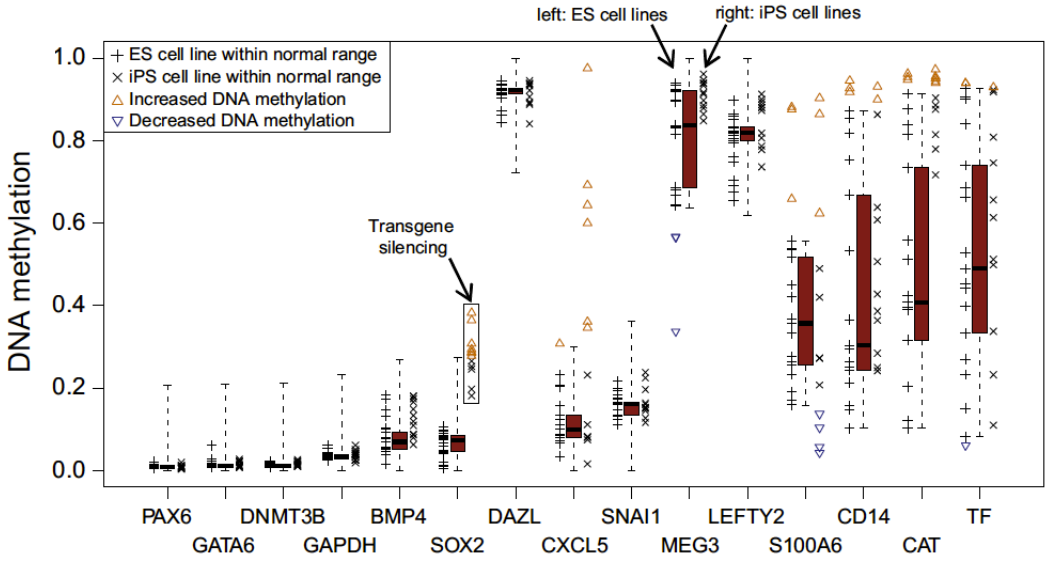


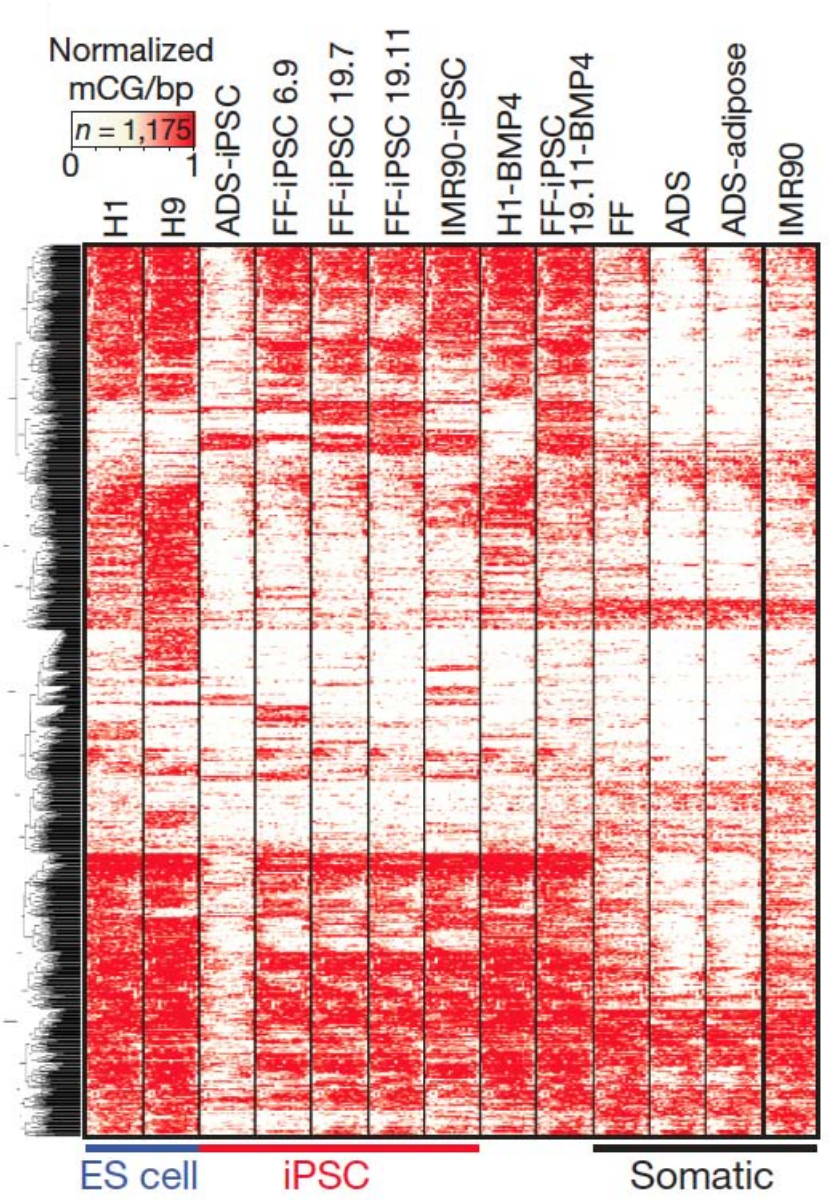
DNA Methylation : ES – iPS comparison

Bock *et al.*, Cell 2011 Lister *et al.*, Nature 2010

- CpG status across individual allele copies, variability





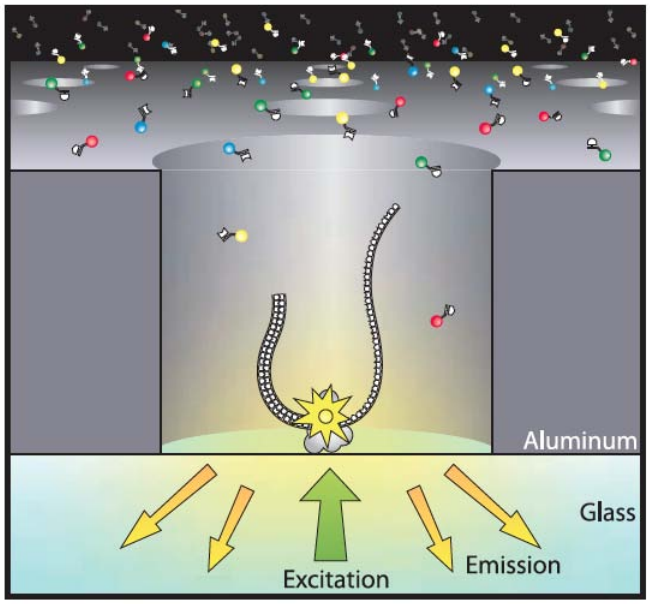


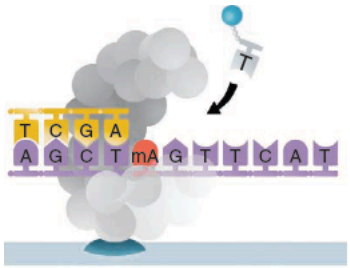
DNA Mehtylation : direct assessment

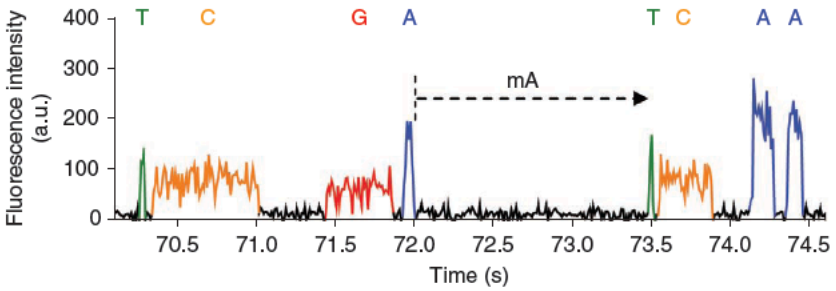
Direct detection of DNA methylation during single-molecule, real-time sequencing

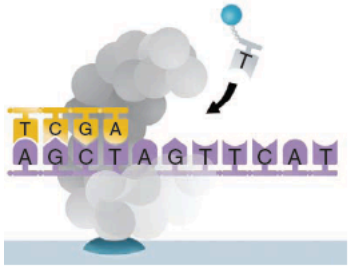
Nature Methods 2010

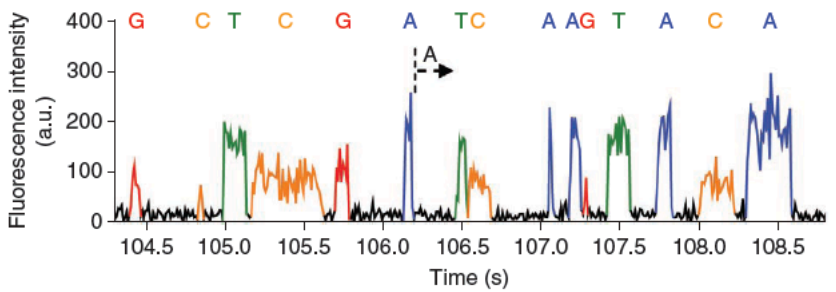
Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach & Stephen W Turner









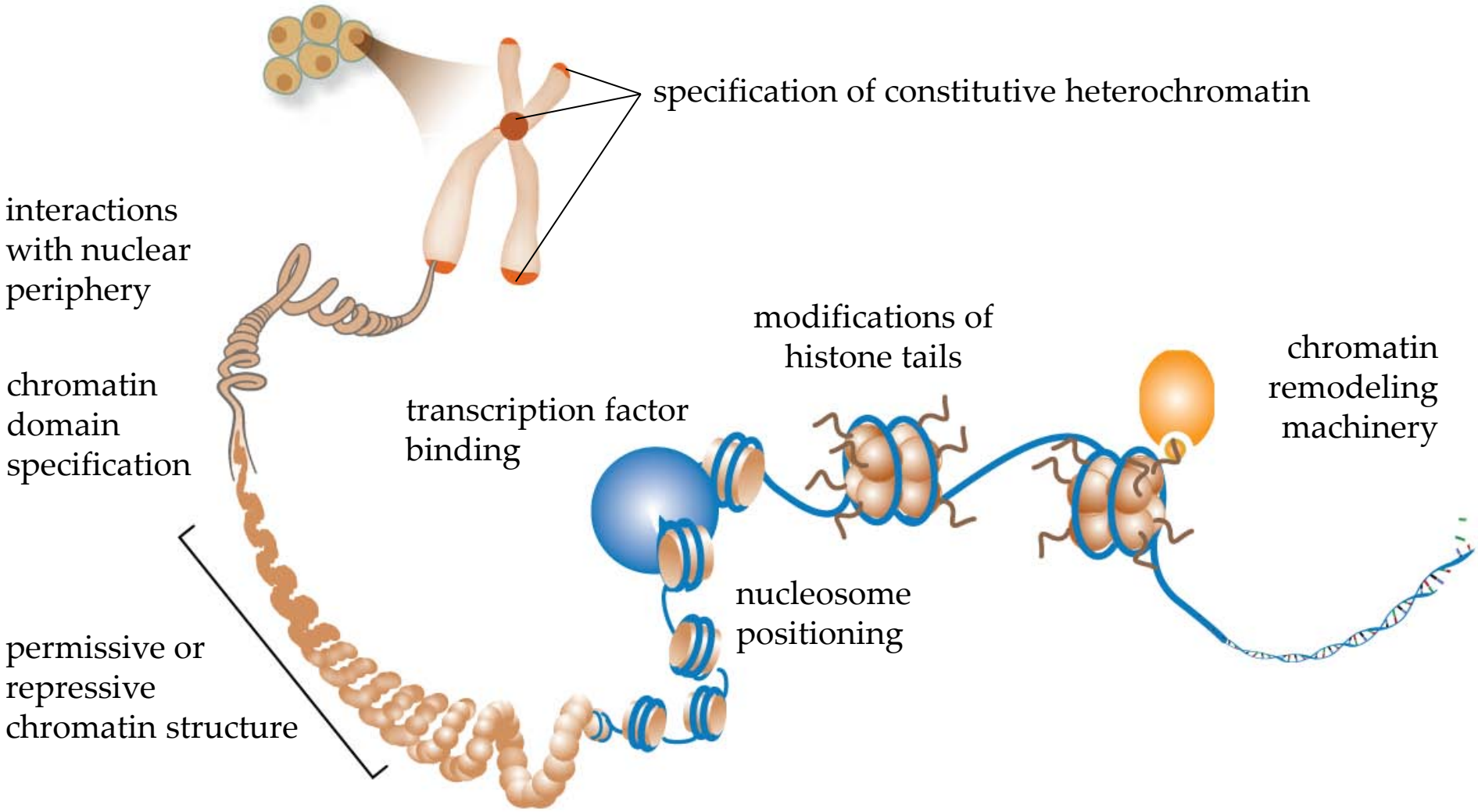


- PacBio Real-time sequencing technology, ZWG, circular templates
- Interpulse duration differences, complex dependency
- Unmethylated references, *de novo*

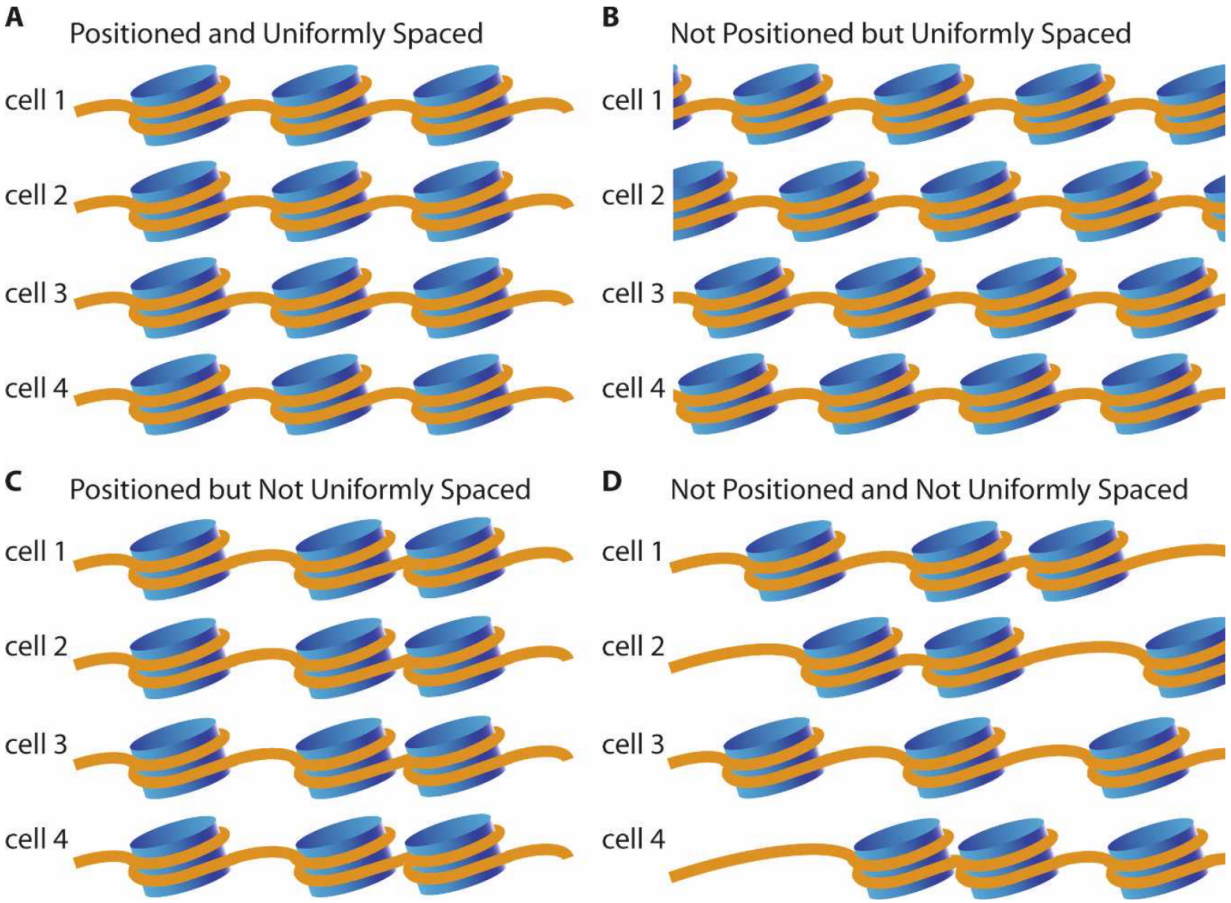
Epigenetics assays using NGS

- Bisulfite-Sequencing (BS-seq)
  - DNA methylation
- MNase-seq
  - Nucleosome positioning
- Chromatin Immunoprecipitation sequencing (ChIP-seq)
  - Signatures of protein association
  - Histone variants and histone modification
- DNase-seq, ATAC-seq
  - Chromatin accessibility
- Hi-C, CHIA-PET
  - Chromatin architecture

Chromatin structure and regulation



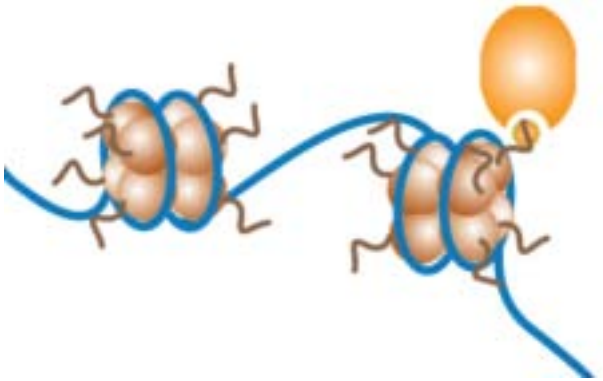
Possible models of nucleosome positioning

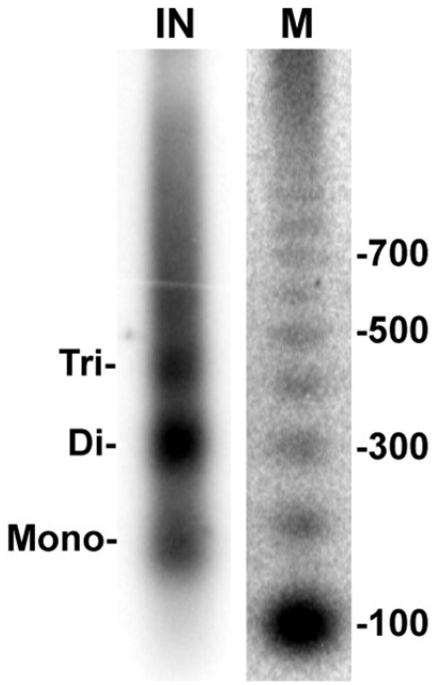


2008 Valouev et al. Genome Res

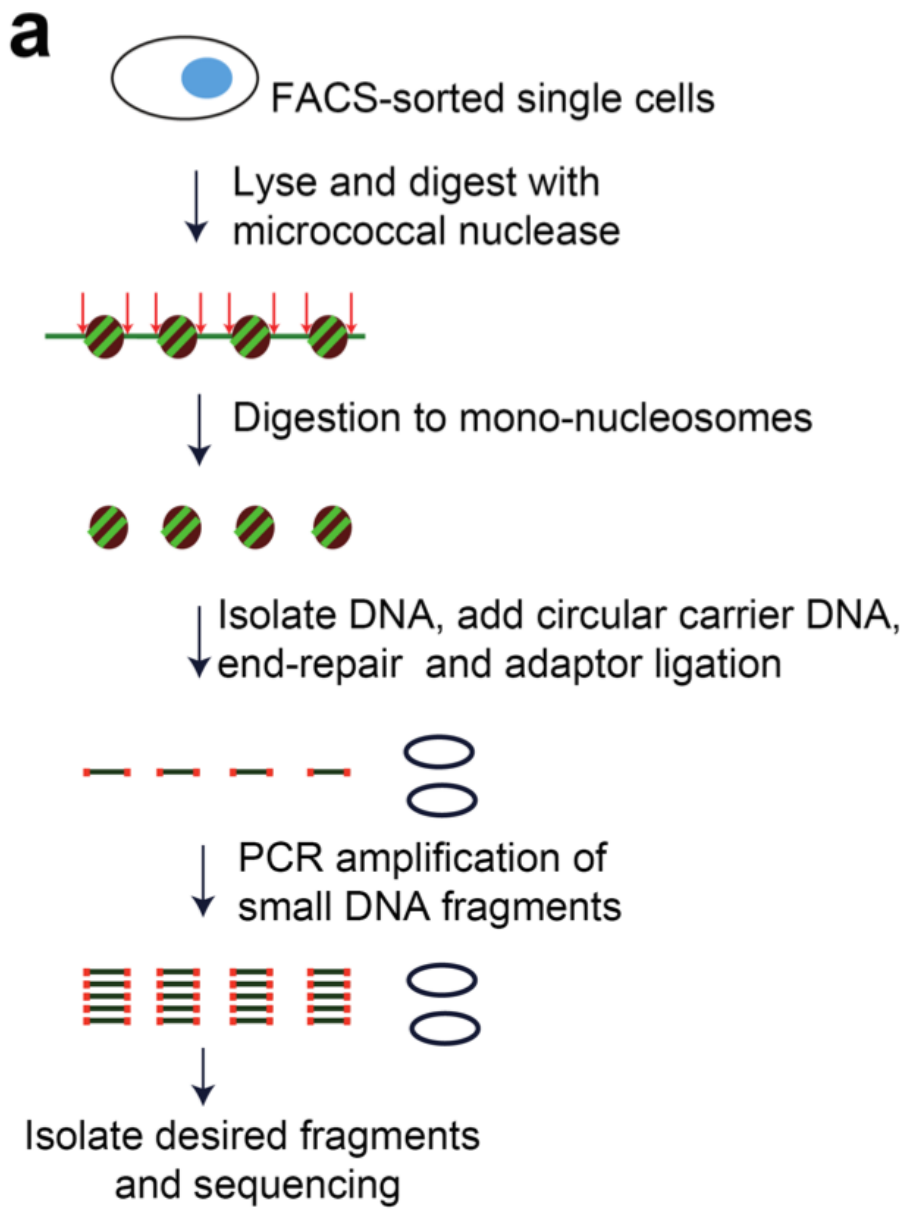
Nucleosome Positioning

- Which DNA fragments wrapped around nucleosomes?
- Digest linker DNA with Micrococcal Nuclease (MNase)
- Isolate mono-nucleosomal fragments
- Construct library and sequence



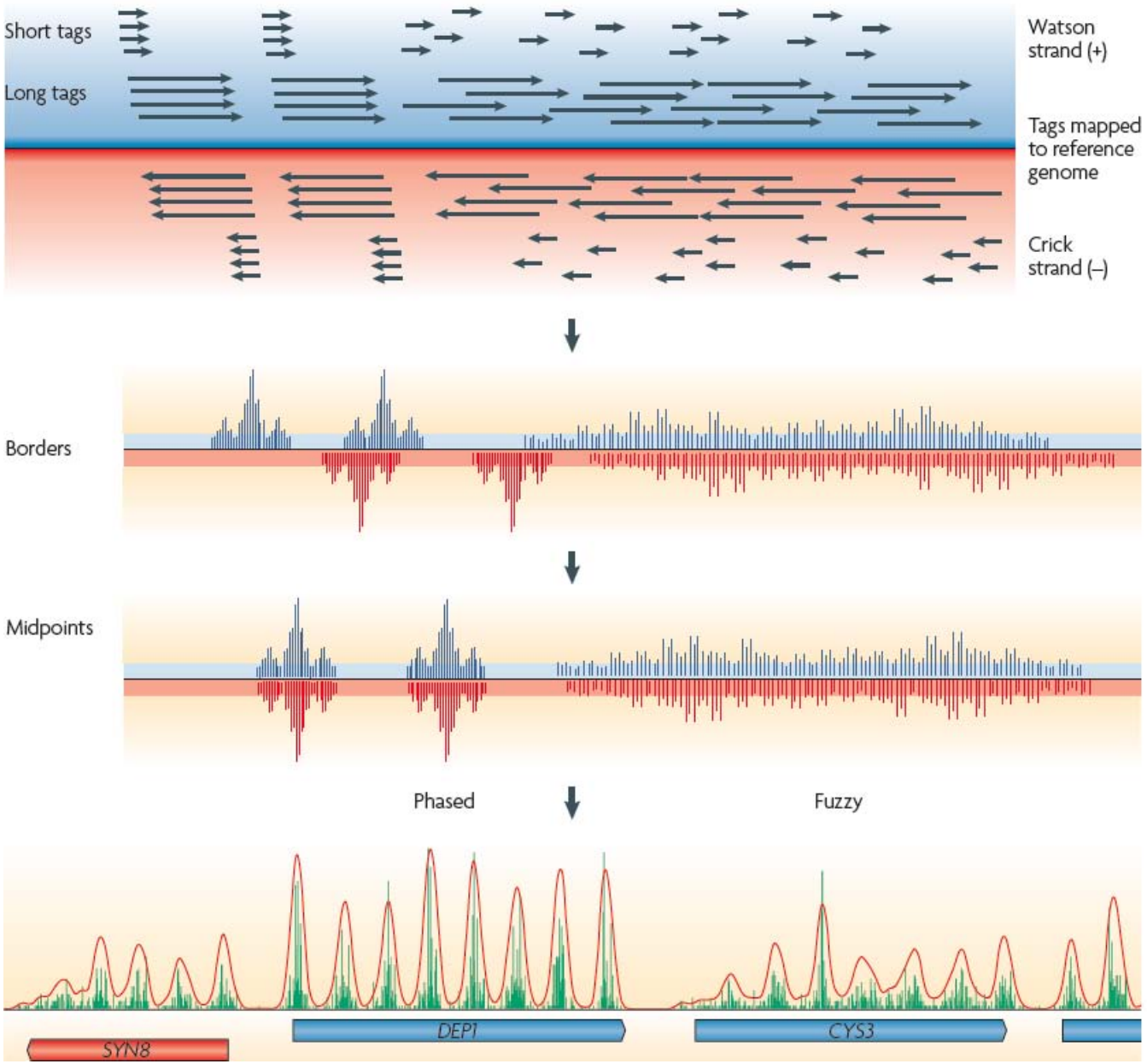


MNase-Seq

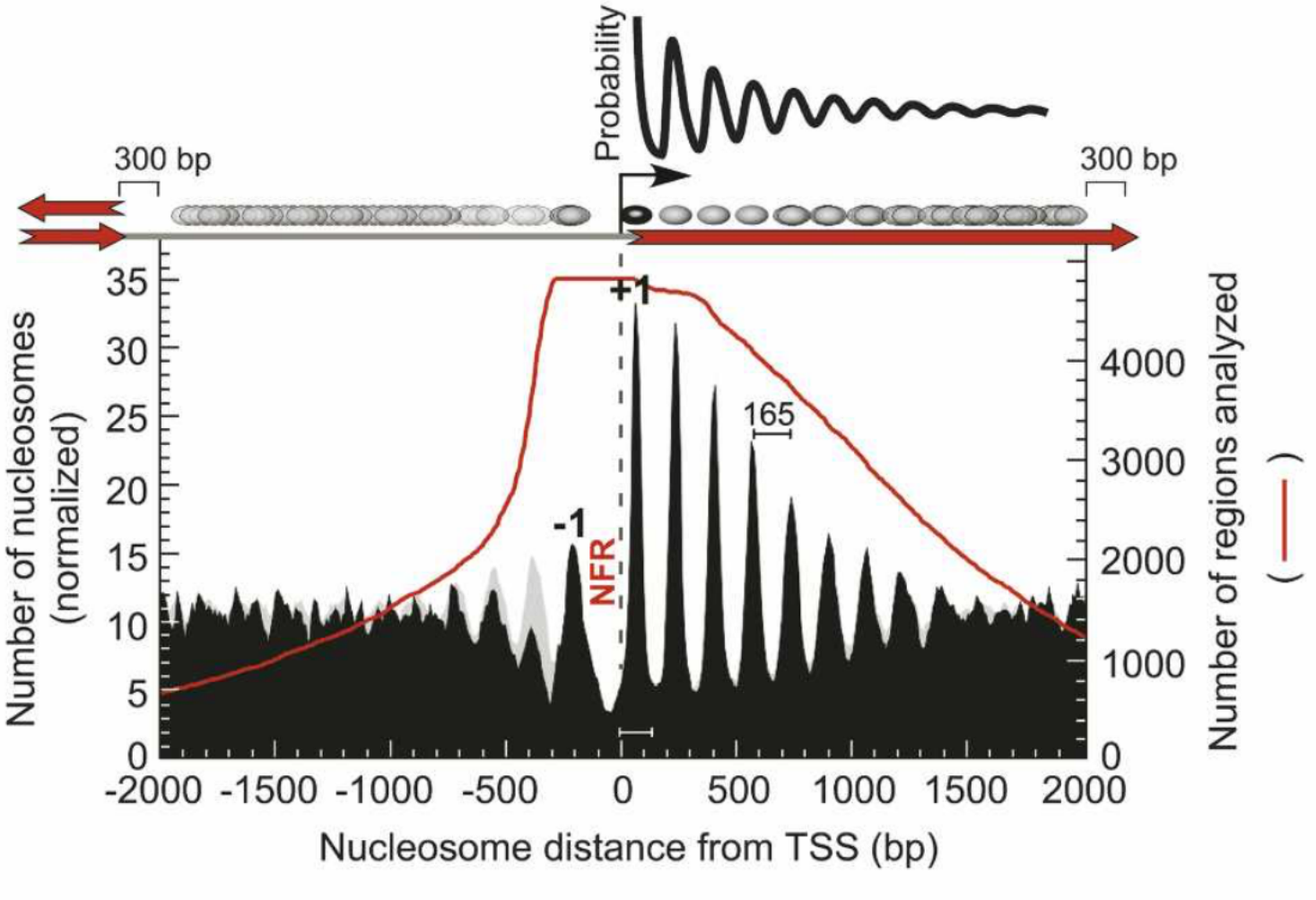

a
FACS-sorted single cells
Lyse and digest with micrococcal nuclease
Digestion to mono-nucleosomes
Isolate DNA, add circular carrier DNA, end-repair and adaptor ligation
PCR amplification of small DNA fragments
Isolate desired fragments and sequencing

Mononucleosome signals

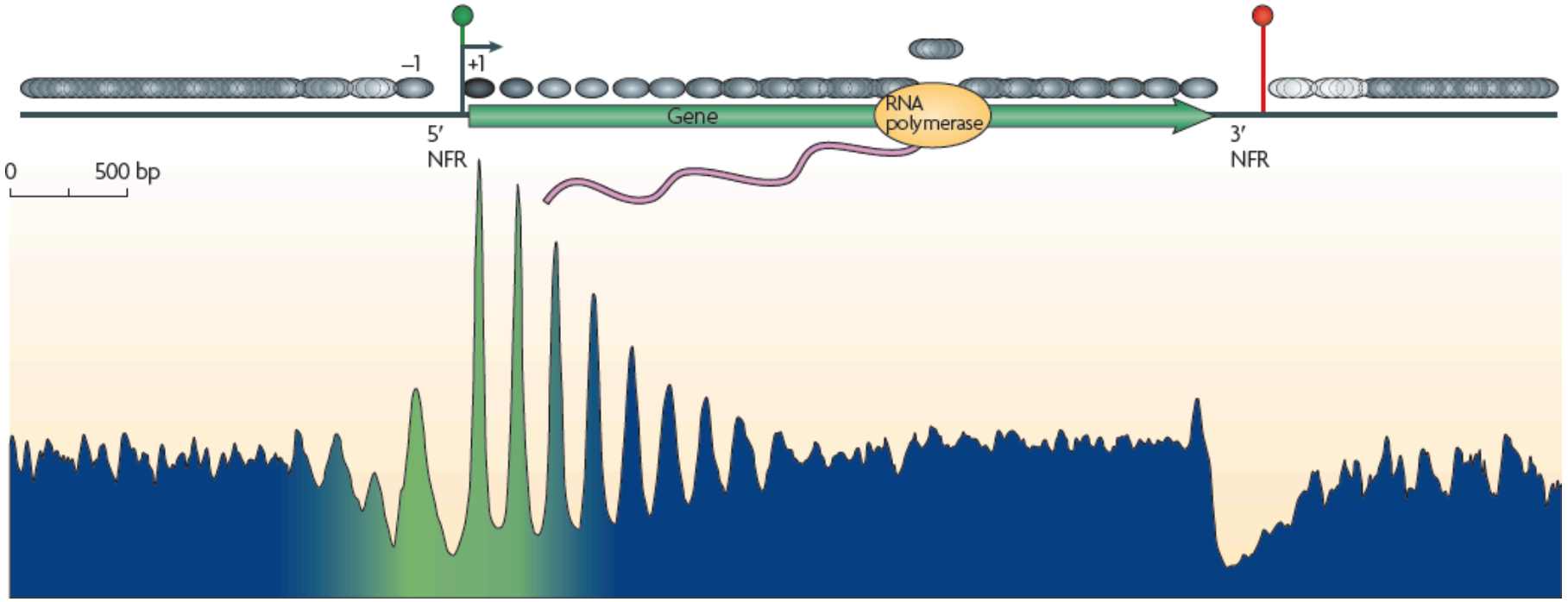
- Strand shift, followed by density estimation
- Direct detection of shifted strand patterns



Nucleosome positioning around TSS



Nucleosome positioning



- Well-positioned nucleosomes; downstream of active TSS
- Nucleosome Free Regions (NFR) upstream of TSS
- Turnover, Variants and modifications

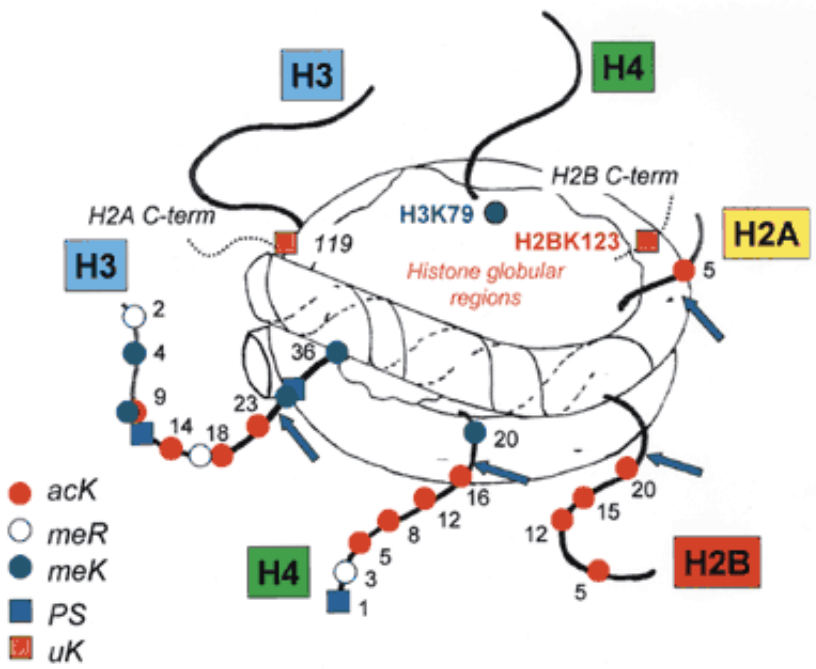
Jiang C., **Pugh BF.**, Nat. Rev. Genetics 2009

Epigenetics assays using NGS

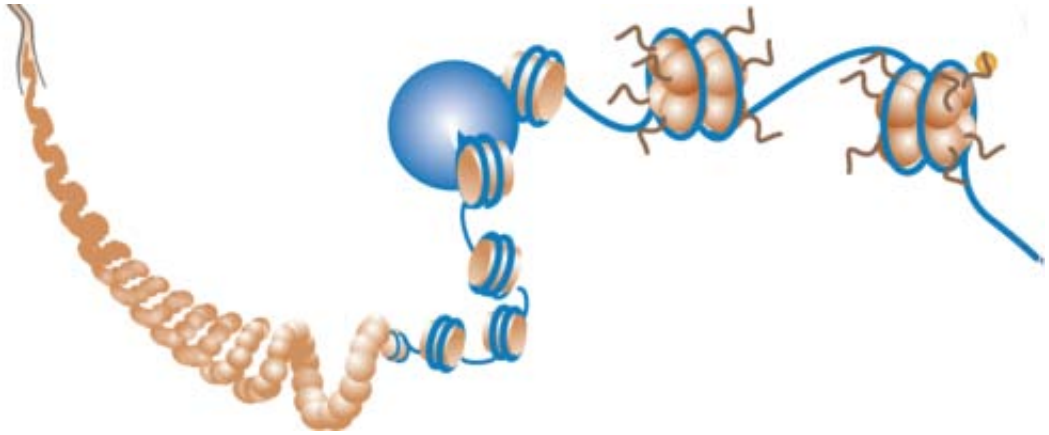
- Bisulfite-Sequencing (BS-seq)
  - DNA methylation
- MNase-seq
  - Nucleosome positioning
- Chromatin Immunoprecipitation sequencing (ChIP-seq)
  - Signatures of protein association
  - Histone variants and histone modification
- DNase-seq, ATAC-seq
  - Chromatin accessibility
- Hi-C, CHIA-PET
  - Chromatin architecture

Histone modifications and variations

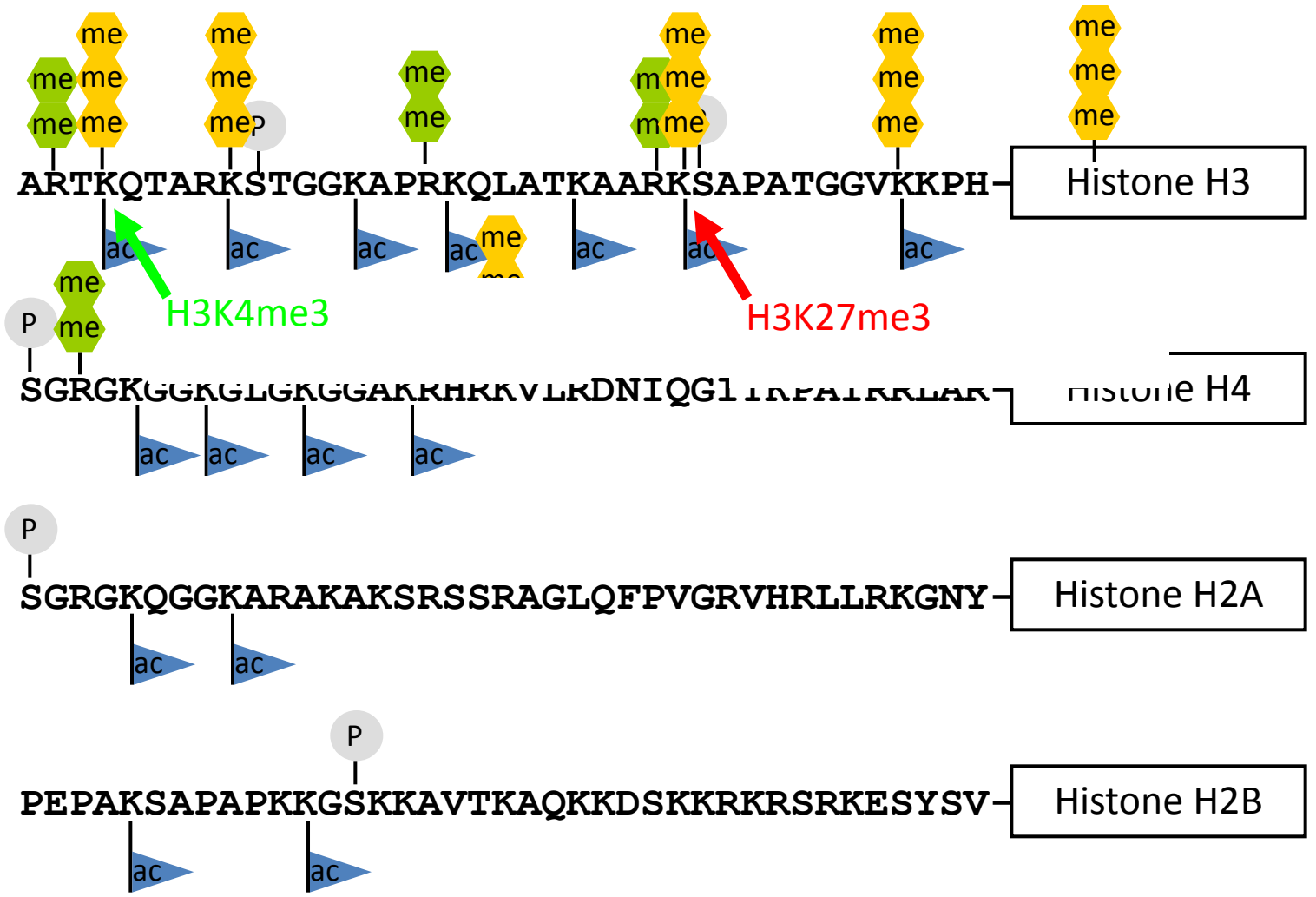
- Covalent modifications
  - acetylation, methylation, ubiquitination, phosphorylation,
- Histone variants
  - H2A.Z, H3.3, H2A.X,
- Functional differences



- Enrich for particular type of nucleosomes
- Use specific antibodies
- **Ch**romatin **I**mmuno**p**recipitation (ChIP)

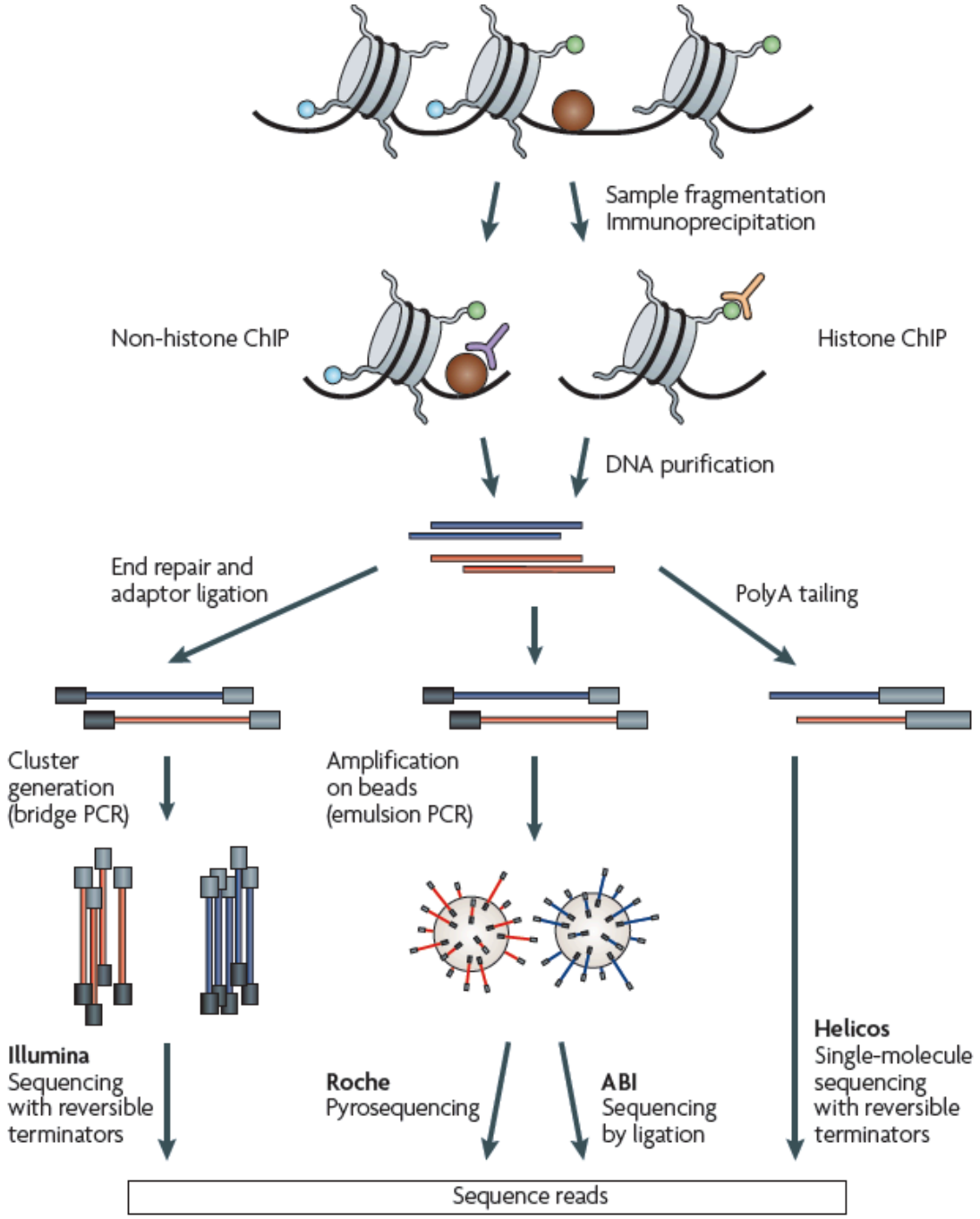


There are multiple histone modifications

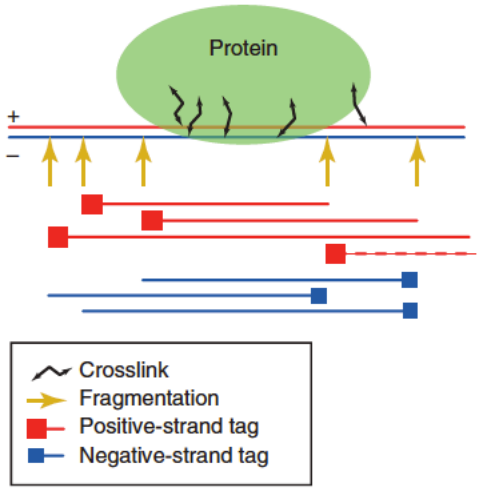


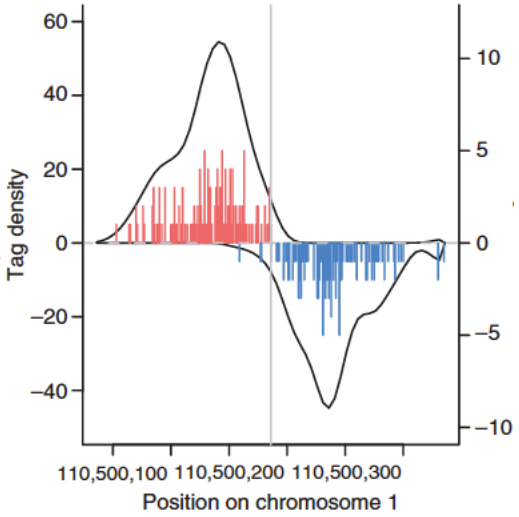
ChIP-seq

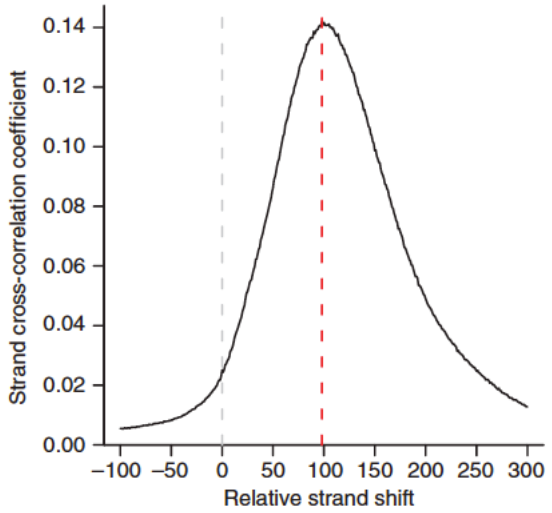
- DNA fragmentation
  - sonication
  - MNase
  - combinations
- Immunoprecipitation
  - Histone or DNA protein antibodies
  - Tags
- Crosslinking
- Amplification, library preparation



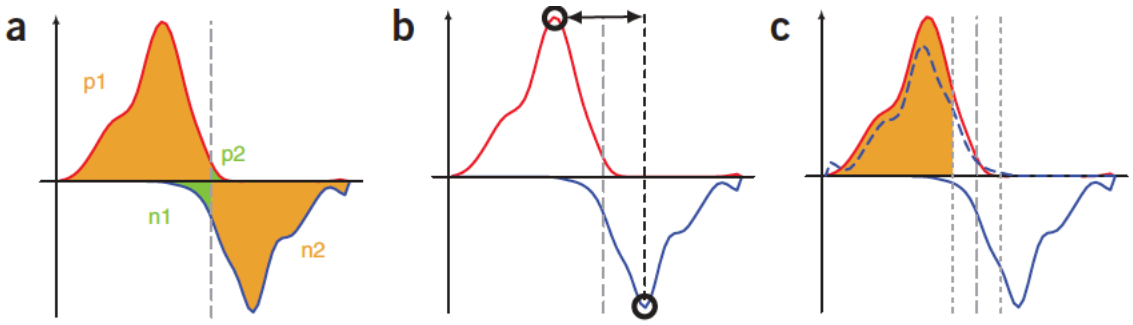
ChIP-seq : binding positions





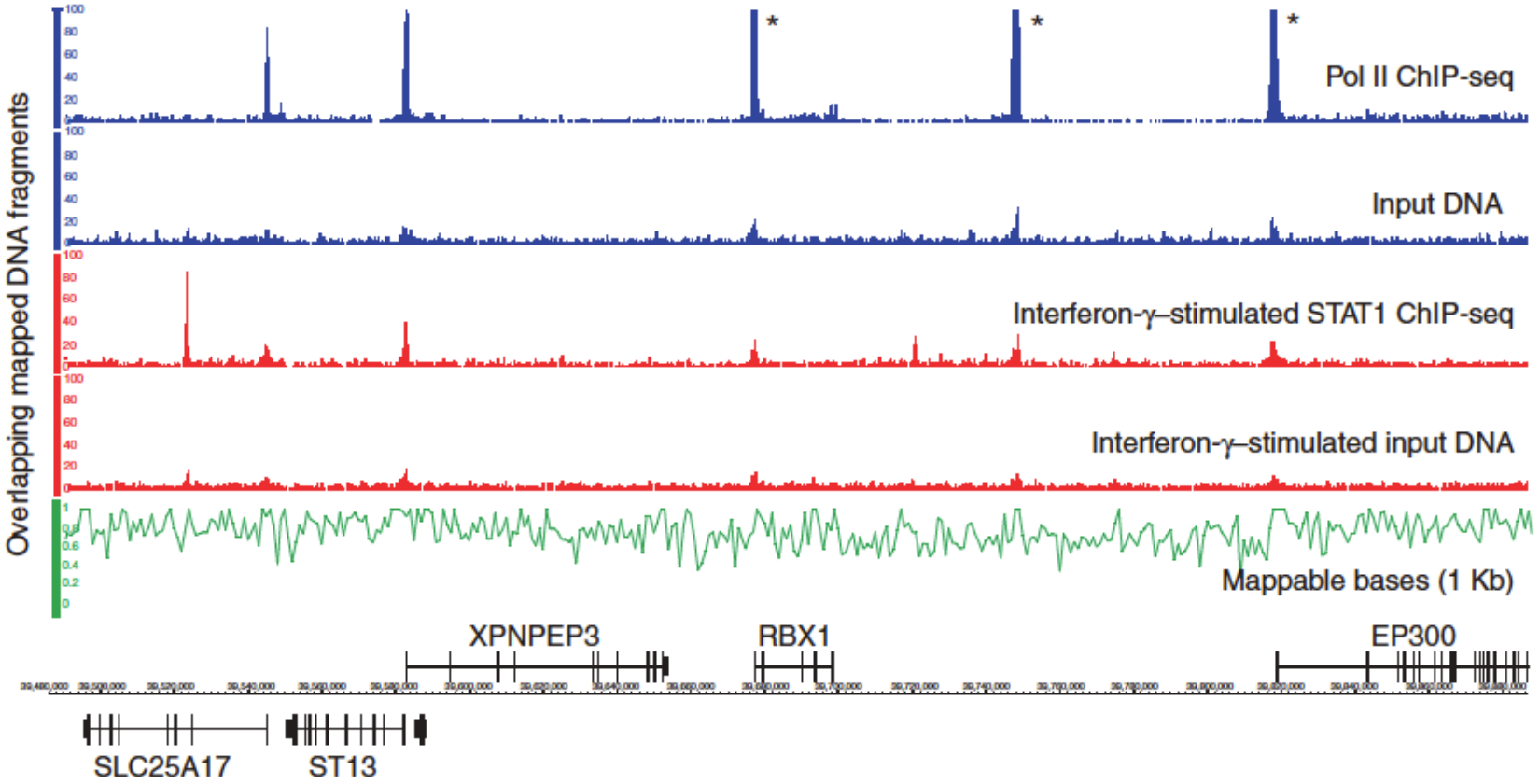


- Stranded pattern
- Peak detection methods
  - PeakSeq, MACS, SPP, QuEST



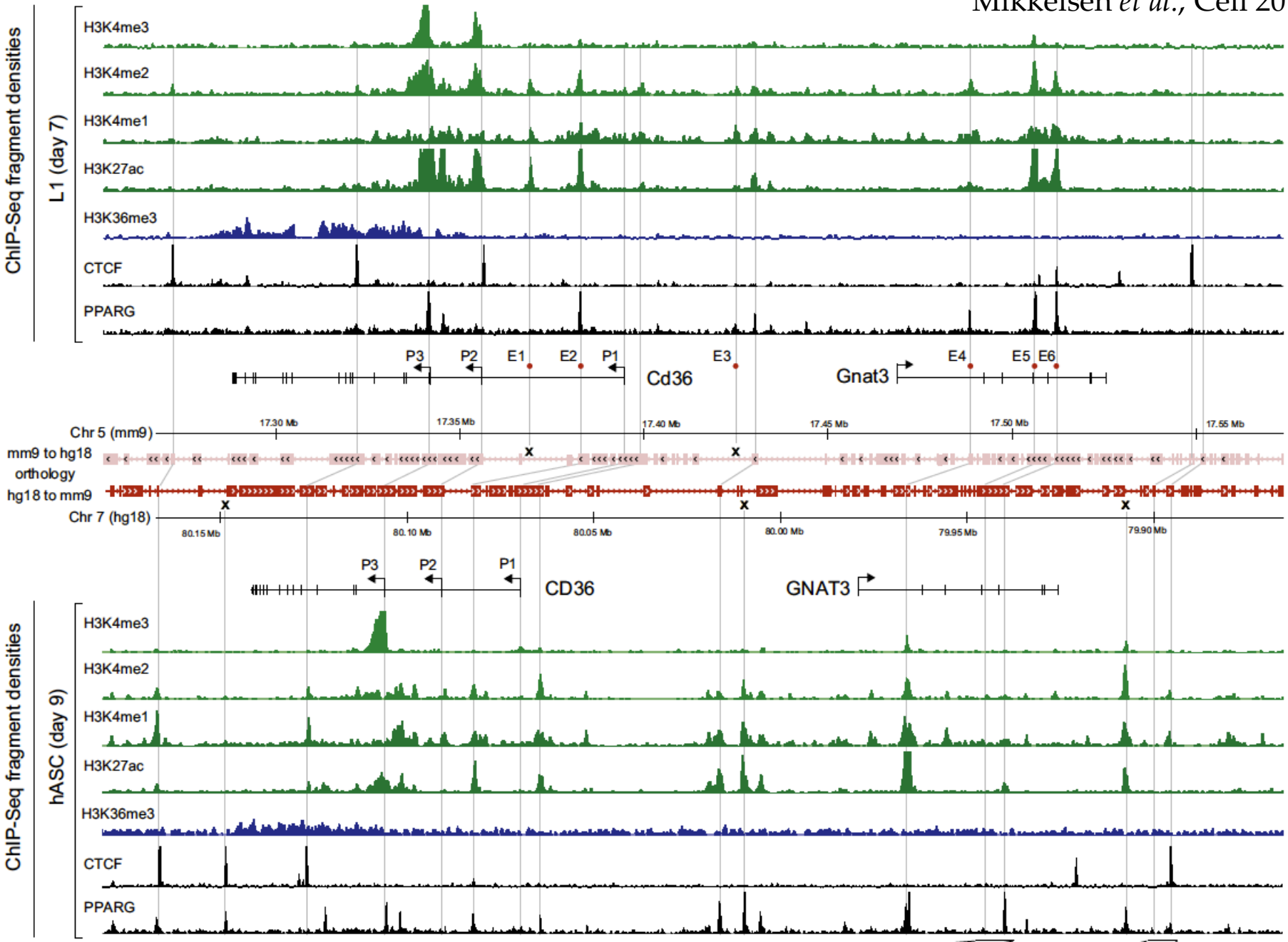
Assessing epitope enrichment

- Uneven background
  - seq. biases (fragmentation, GC, start)
  - mappability
- Controls
  - input chromatin
  - mock/non-specific Ab

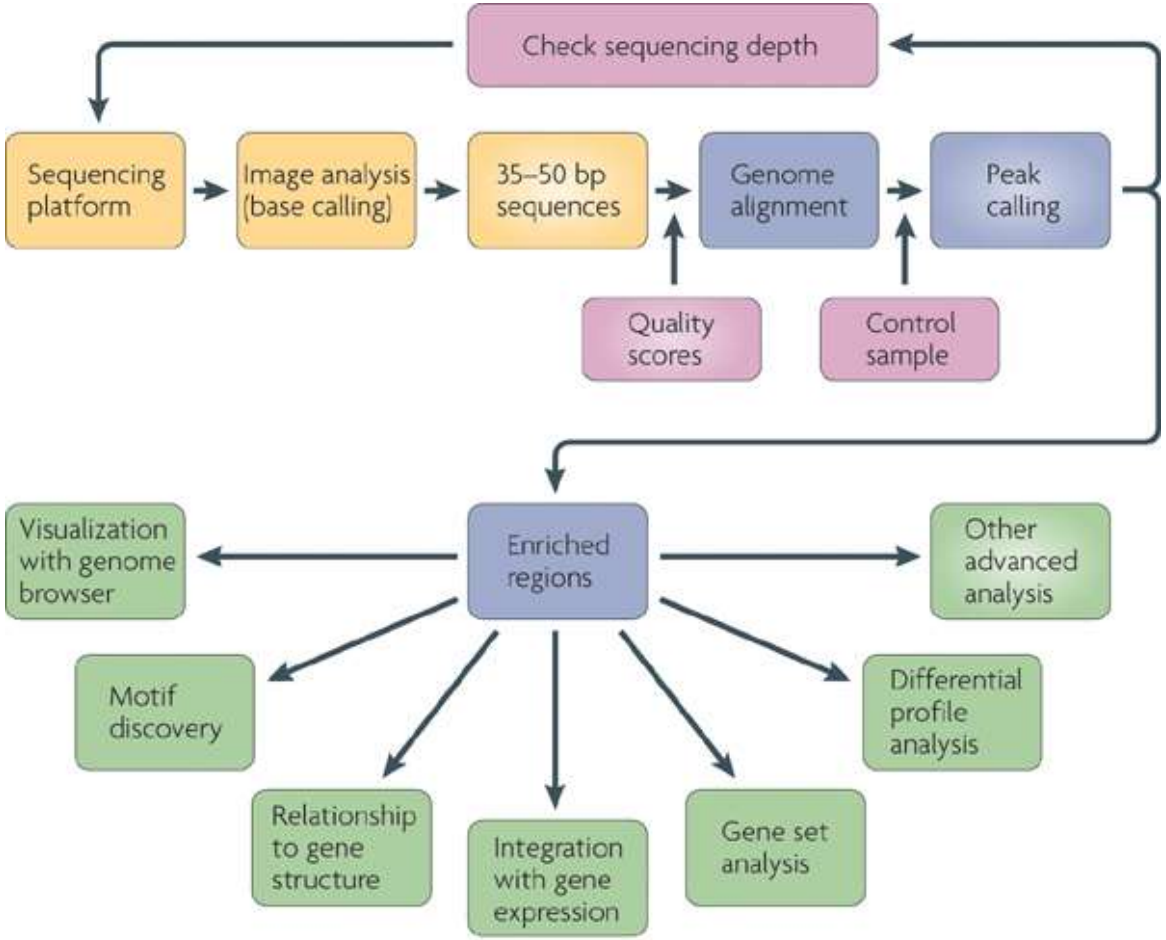


ChIP-seq example : adipogenesis

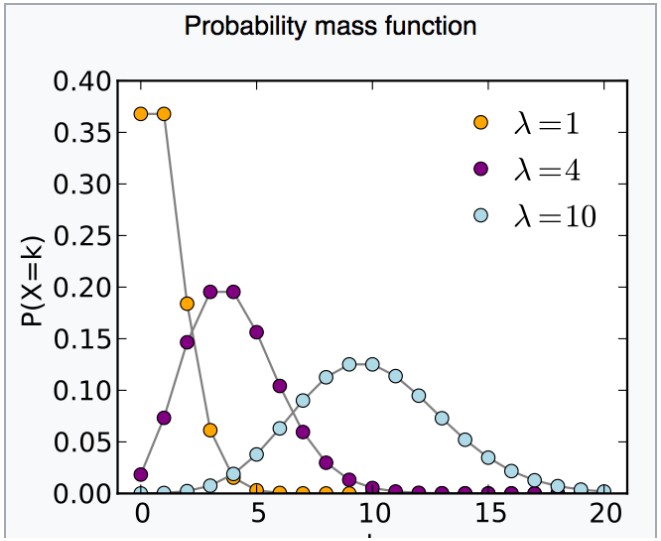
Mikkelsen *et al.*, Cell 2010



Pipeline for ChIP-seq data analysis



Number of reads from RNA-seq/ChIP-seq fit Poisson distribution



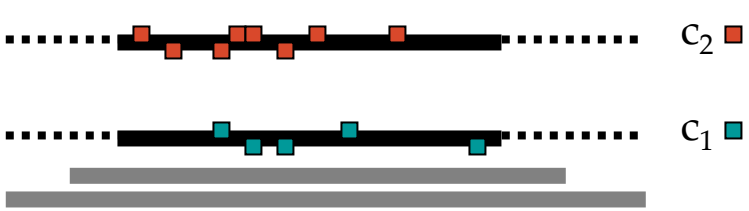
$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

where

- λ is the average number of events per interval
- e is the number 2.71828... (Euler's number) the base of the natural logarithms
- k takes values 0, 1, 2, …

ChIP-seq : estimating enrichment magnitude

- Log intensity ratios (M) : Log2[ChIP / input]
- Poisson process
  - Sequencing depth instead of time
  - Constant rate at a local scale
  - Estimate ratio of two Poisson rates



  - MLE $\hat{r} = \frac{\hat{\lambda}_1}{\hat{\lambda}_2} = R\frac{(c_1 + 0.5)}{(c_2 + 0.5)}$

$$R = \frac{S_2}{S_1}$$

  - Bounds $r^{(1-\alpha)} = R\frac{(c_1 + 0.5)}{(c_2 + 0.5)}F_{1-\alpha,2(c_1+0.5),2(c_2+0.5)}$

  - Sliding windows, conservative estimates, broad regions of enrichment
  - Spatial scale

- Testing for enrichment
  - binomial model
  - *background (R)* corrections

SICER for calling Histone peaks

1. Remove duplicate reads

2.1 The island approach

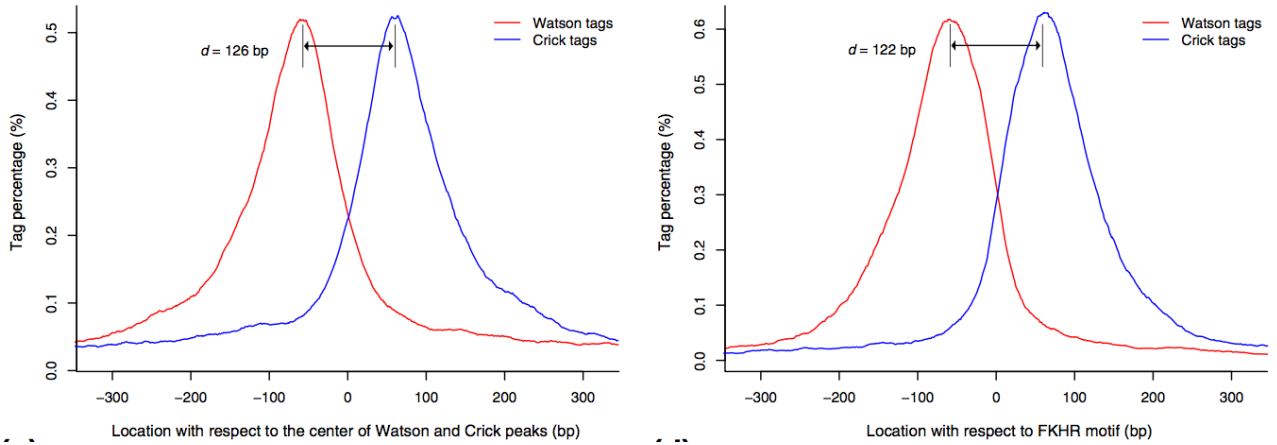
2.1.1 Scoring scheme We partition the genome of effective length L into non-overlapping windows of size w. We define the score s for a window with l reads to be $s(l) = -\log P(l, \lambda)$. $P(l, \lambda)$ is a Poisson distribution parameterized by the average number of reads in a window $\lambda = wN/L$, where N is the total number of reads in the ChIP-Seq library. Given this definition, the scores

Zang et al. 2008 Bioinformatics

MACS for calling TF peaks

1. Remove duplicate reads

2. Modeling the shift size of ChIP-Seq tags



3. Peak detection based on local reads distribution

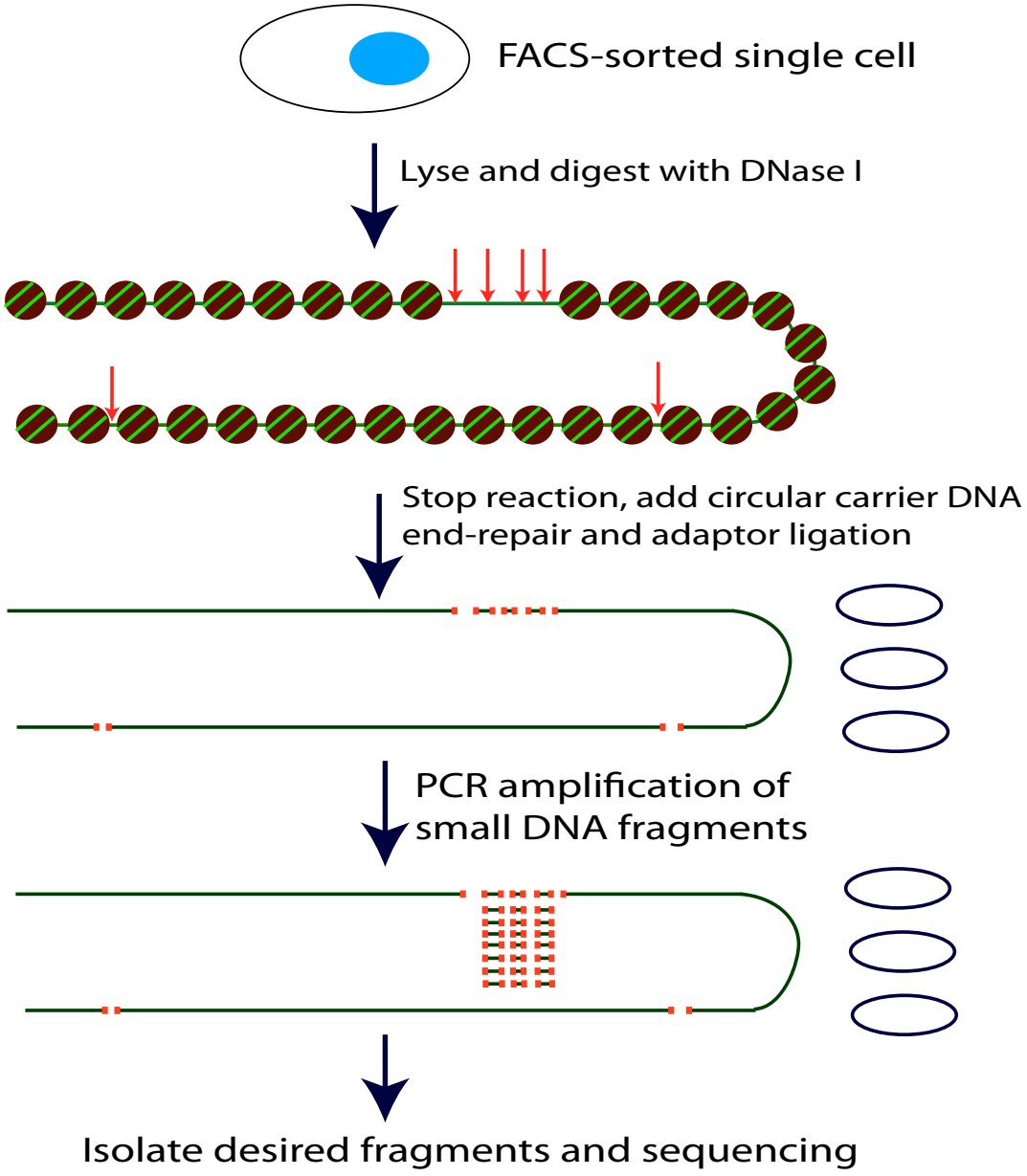
$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$$

Zhang et al. 2009 Genome Biol

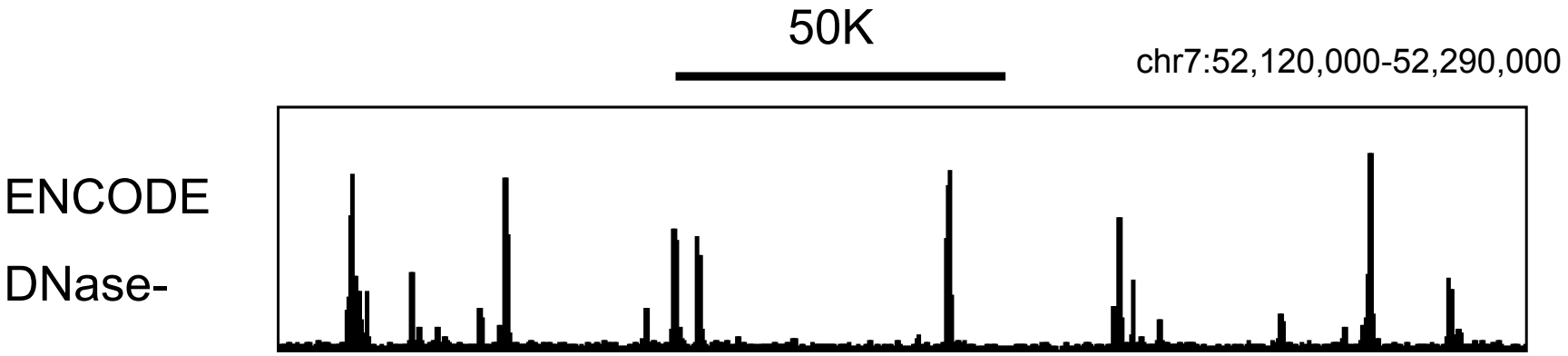
Epigenetics assays using NGS

- Bisulfite-Sequencing (BS-seq)
  - DNA methylation
- MNase-seq
  - Nucleosome positioning
- Chromatin Immunoprecipitation sequencing (ChIP-seq)
  - Signatures of protein association
  - Histone variants and histone modification
- DNase-seq, ATAC-seq
  - Chromatin accessibility
- Hi-C, CHIA-PET
  - Chromatin architecture

Schema of DNase-Seq



DNase-seq is a powerful tool for genome-wide mapping of regulatory elements

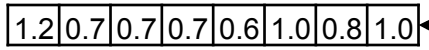


Boyle, *et al*. *Cell*, 2008

Recall: TFBS Position Weight Matrix (PWM)

Experimentally determined sites							
A	T	G	G	C	A	T	G
A	G	G	G	T	G	C	G
A	T	C	G	C	A	T	G
T	T	G	C	C	A	C	G
A	T	G	G	T	A	T	T
A	T	T	C	G	A	C	G
A	G	G	G	C	G	T	T
A	T	G	A	C	A	T	G
A	T	G	G	C	A	T	G
A	C	T	G	G	A	T	G

Alignment (count) Matrix								
A	9	0	0	1	0	8	0	0
C	0	1	1	1	7	0	3	0
G	0	2	7	8	1	2	0	8
T	1	7	2	0	2	0	7	2
Frequency Weight Matrix								
A	0.9	0.0	0.0	0.1	0.0	0.8	0.0	0.0
C	0.0	0.1	0.1	0.1	0.7	0.0	0.3	0.0
G	0.0	0.2	0.7	0.8	0.1	0.2	0	0.8
T	0.1	0.7	0.2	0.0	0.2	0.0	0.7	0.2
Cons	A	T	G	G	C	A	T	G



Information content of each column

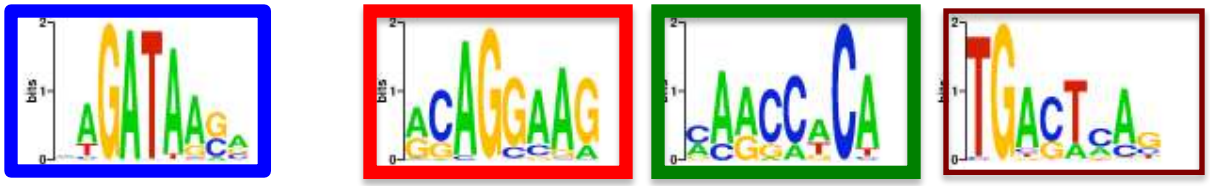
Information content of a motif = sum of all columns

= 1.2 + 0.7 + 0.7 +0.6 + 1.0 + 0.8 + 1.0 = 6.0

http://cs173.stanford.edu [BejeranoWinter12/13]

Motif discovery and enrichment analysis

Th2



GATA motif

Ets motif

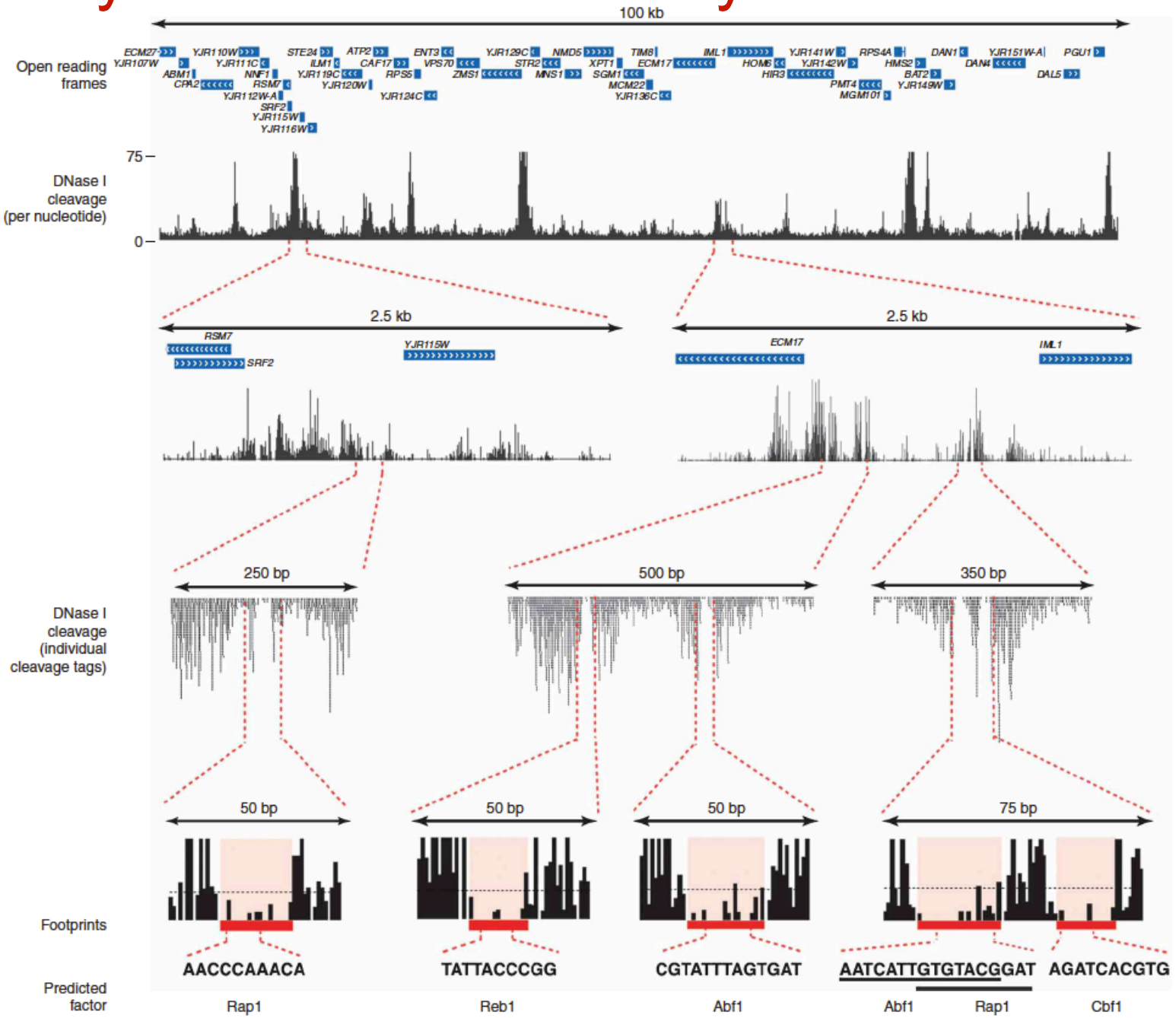
AP-1 motif

Runx motif

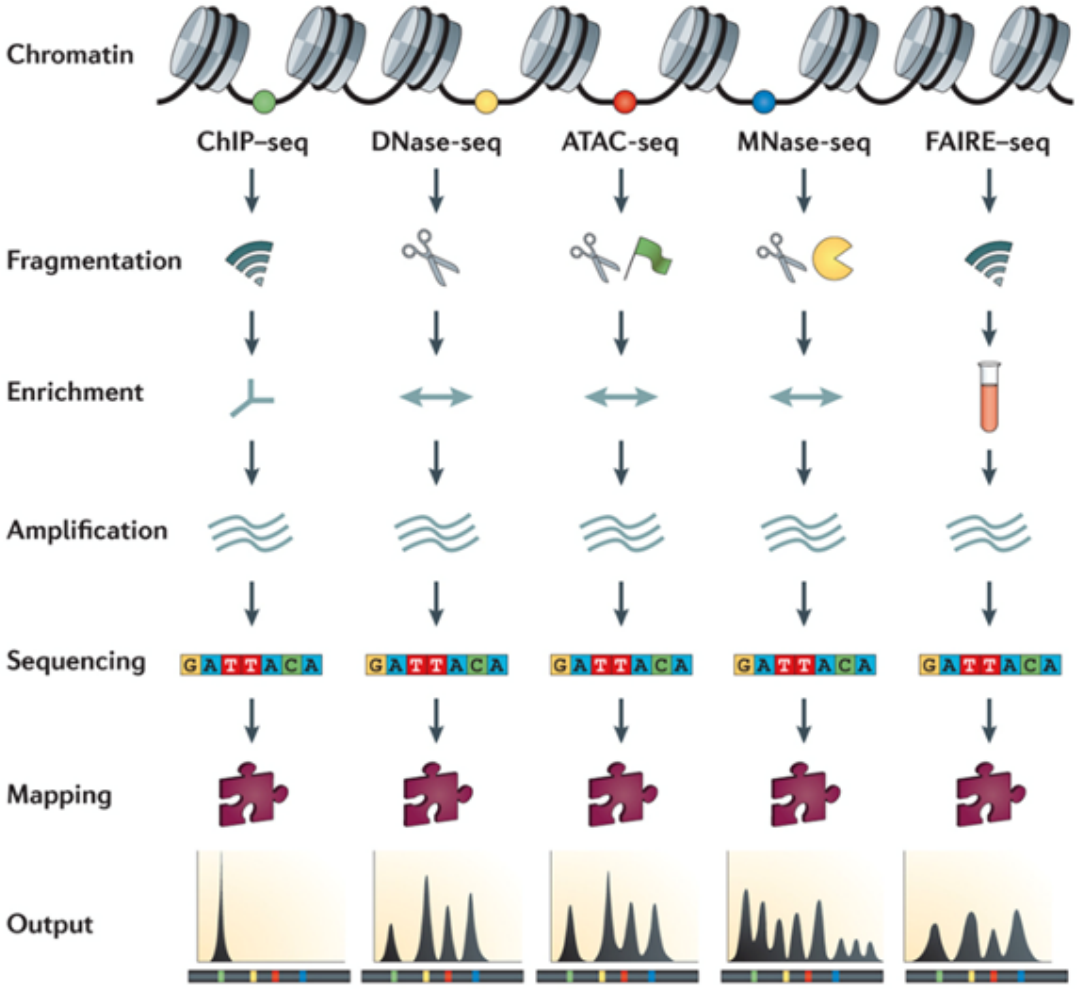
Wei *et al*. Immunity, 2011

Assays : DNase I sensitivity

Hesselberth *et al.*, Nat. Methods 200



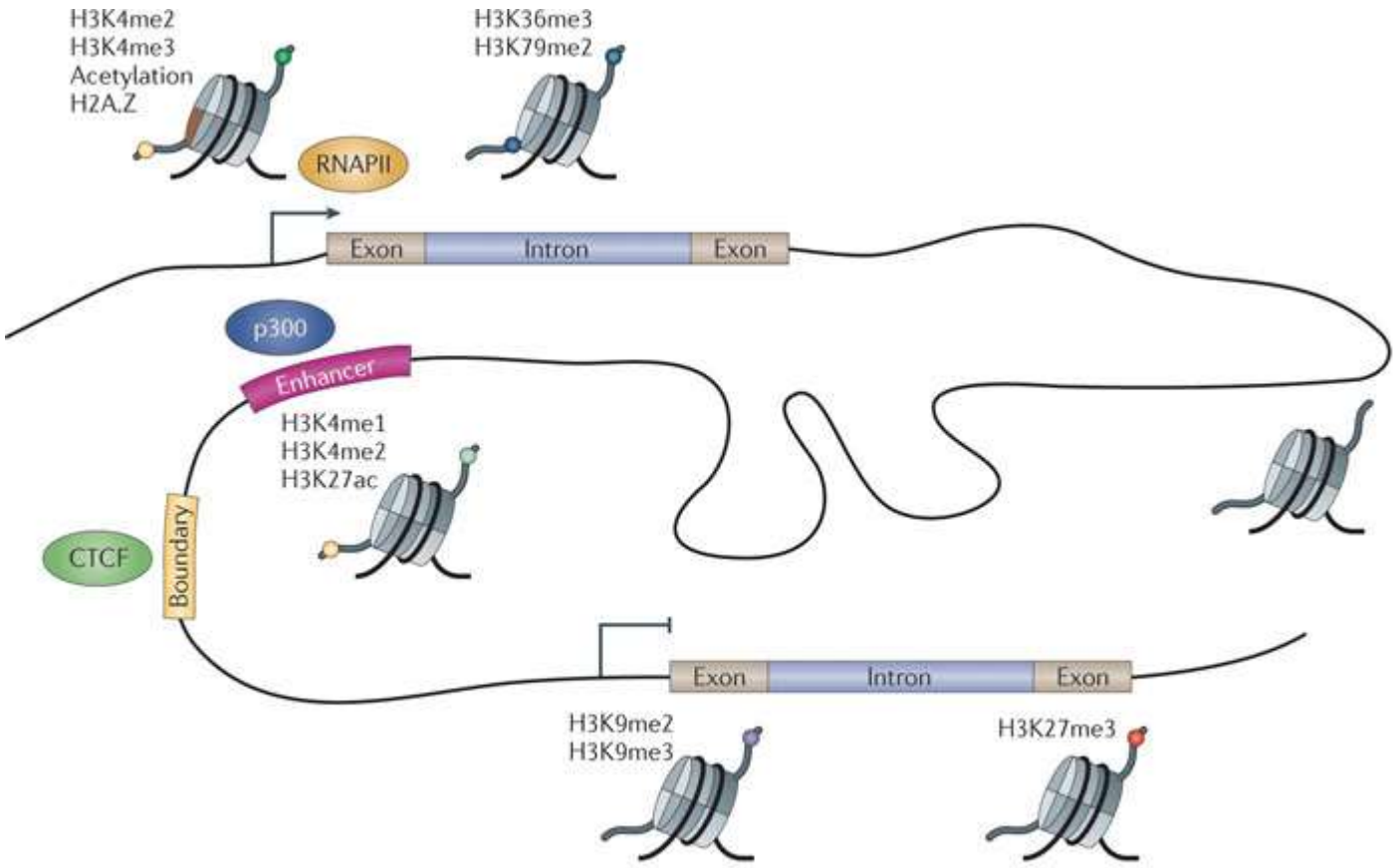
Summary of epigenetic locus study



Epigenetics assays using NGS

- Bisulfite-Sequencing (BS-seq)
  - DNA methylation
- MNase-seq
  - Nucleosome positioning
- Chromatin Immunoprecipitation sequencing (ChIP-seq)
  - Signatures of protein association
  - Histone variants and histone modification
- DNase-seq, ATAC-seq
  - Chromatin accessibility
- Hi-C, CHIA-PET
  - Chromatin architecture

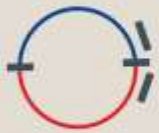
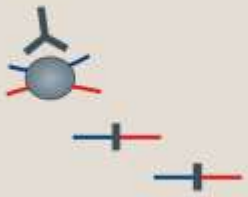
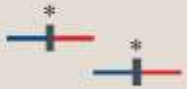
Regulatory function of chromatin interactions



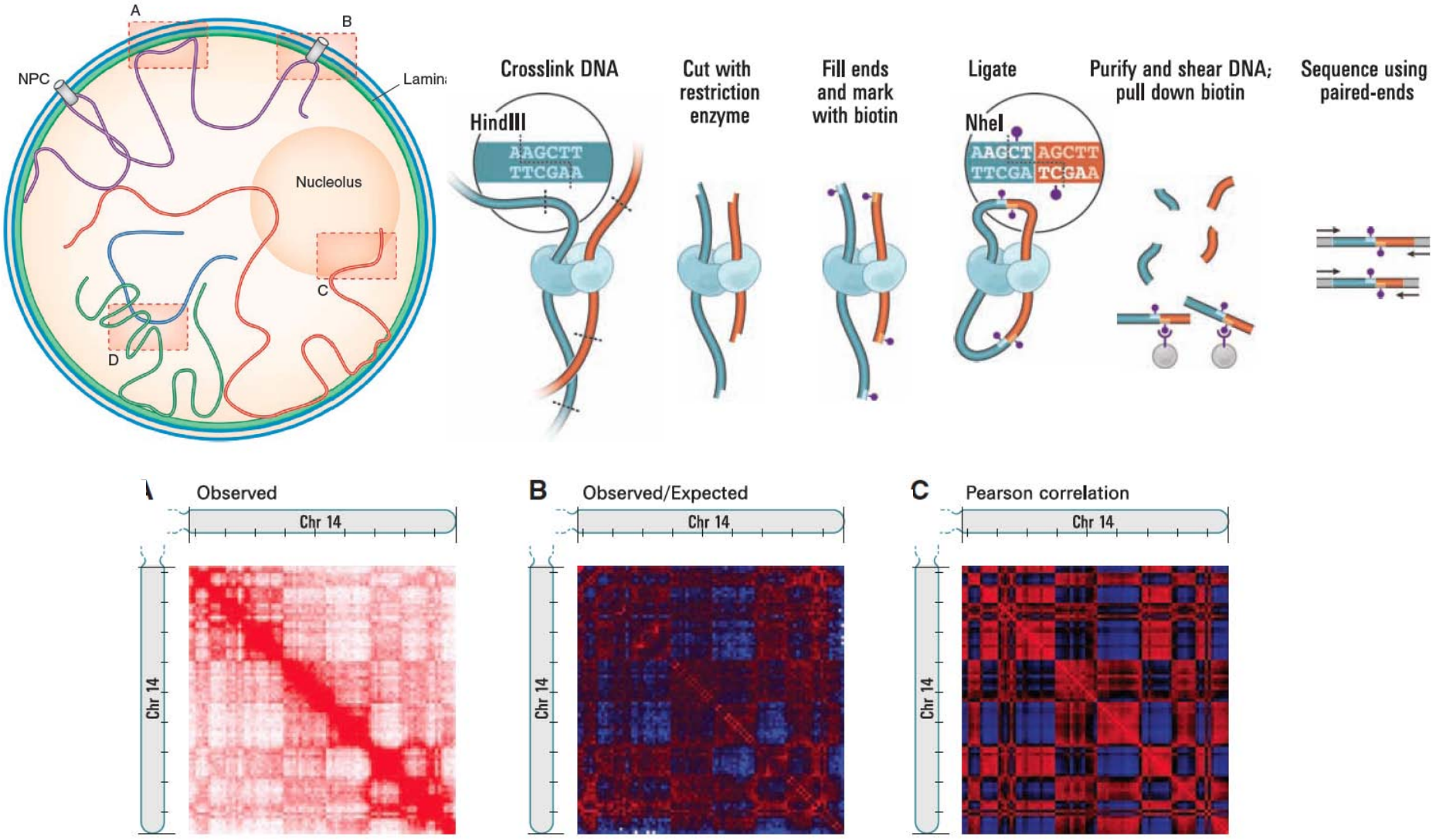
Chromosome Conformation Capture (3C)

a **3C: converting chromatin interactions into ligation products**

b **Ligation product detection methods**

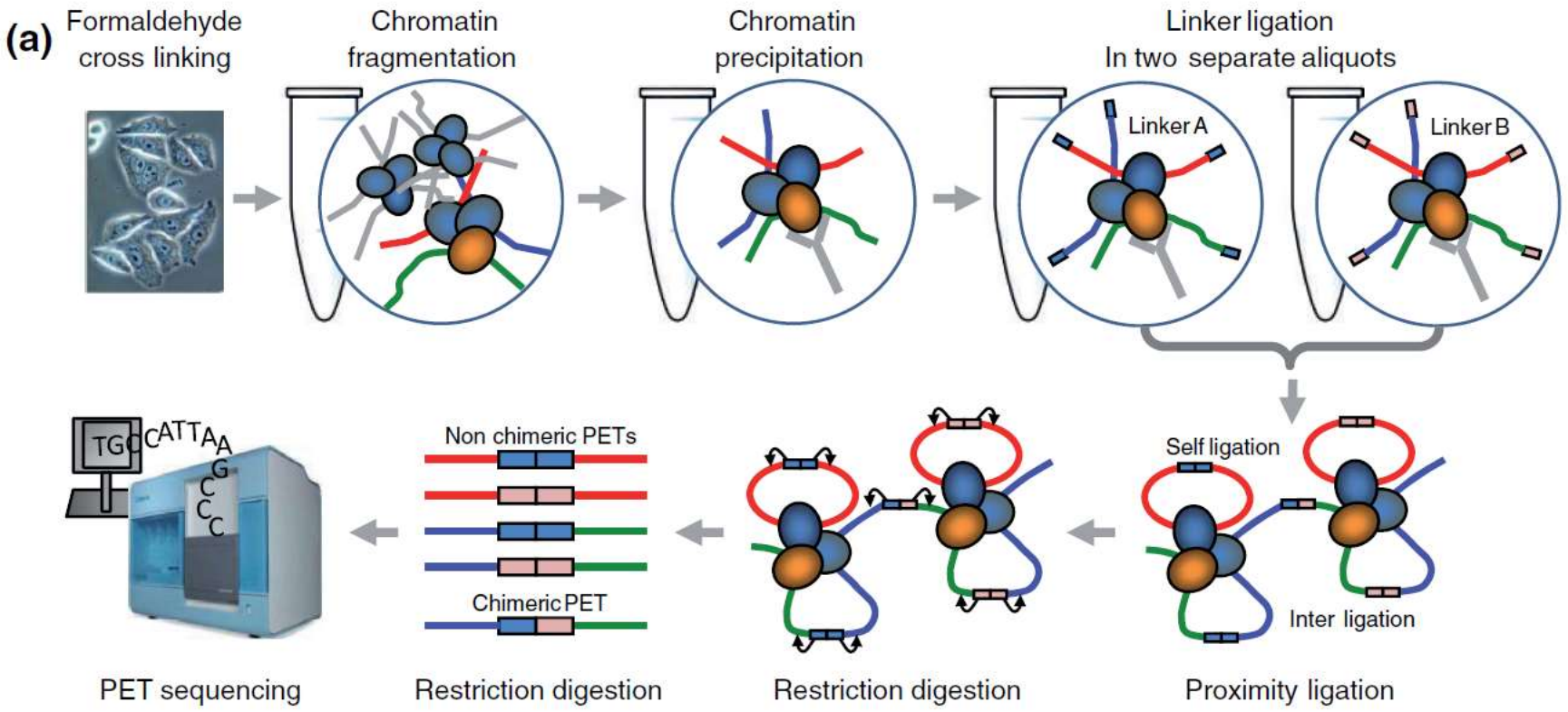
3C	4C	5C	ChIA–PET	Hi-C
One-by-one All-by-all	One-by-all	Many-by-many	Many-by-many	All-by-all
			• DNA shearing • Immunoprecipitation 	• Biotin labelling of ends • DNA shearing 
PCR or sequencing	Inverse PCR sequencing	Multiplexed LMA sequencing	Sequencing	Sequencing

Assays: 3D genome architecture

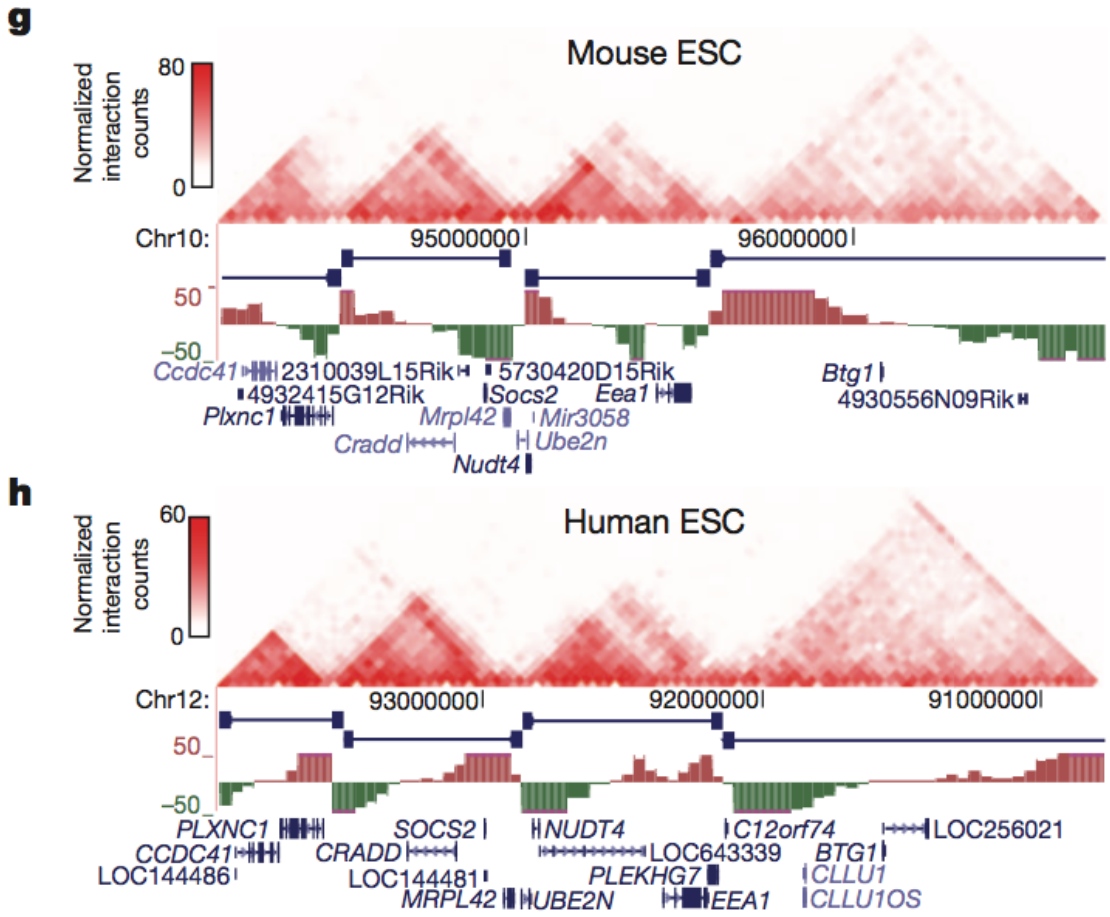


Lieberman-Aiden *et al.* **Dekker J**, Science 2009

ChIA-PET protocol



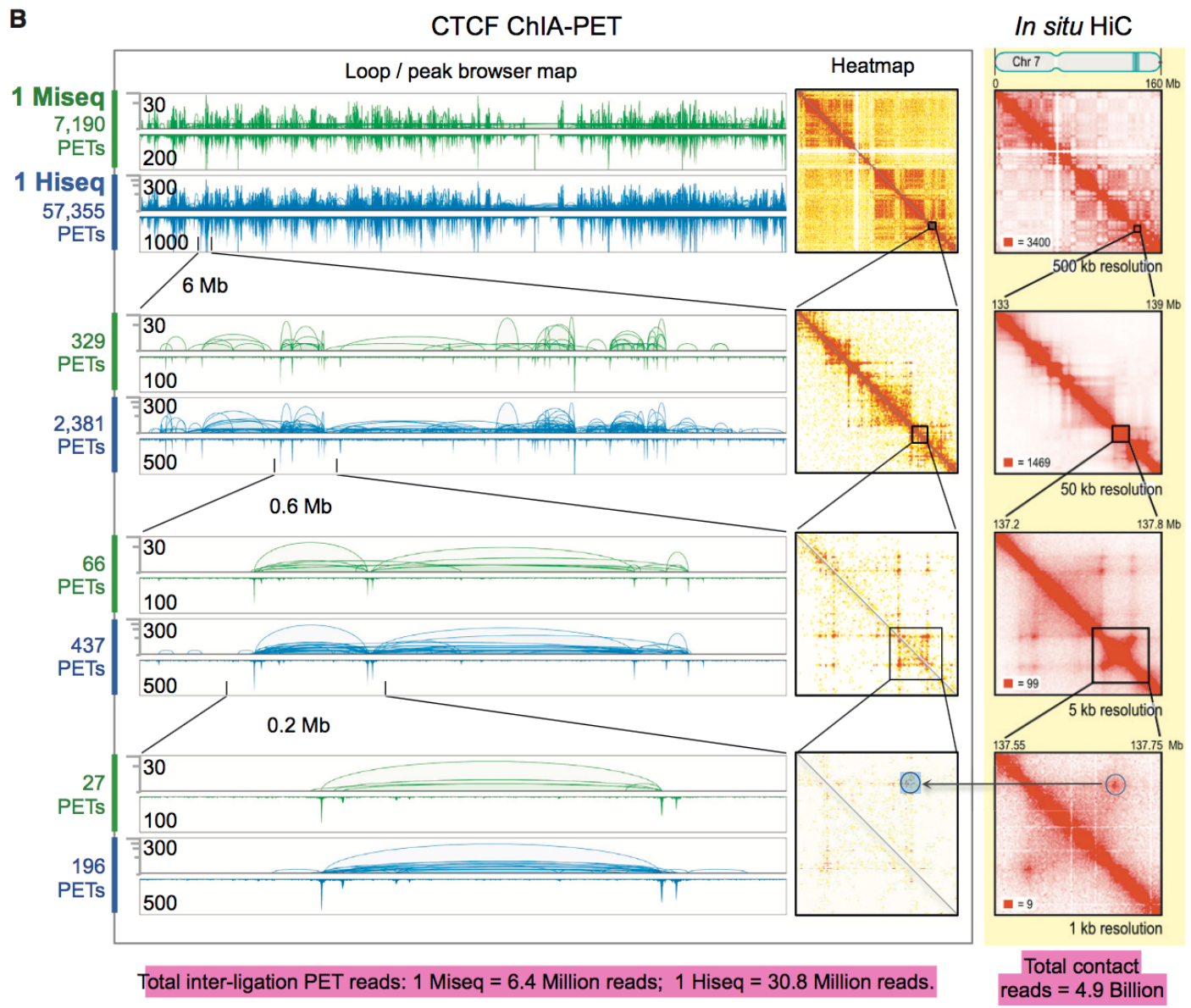
Topological associated domain (TAD)



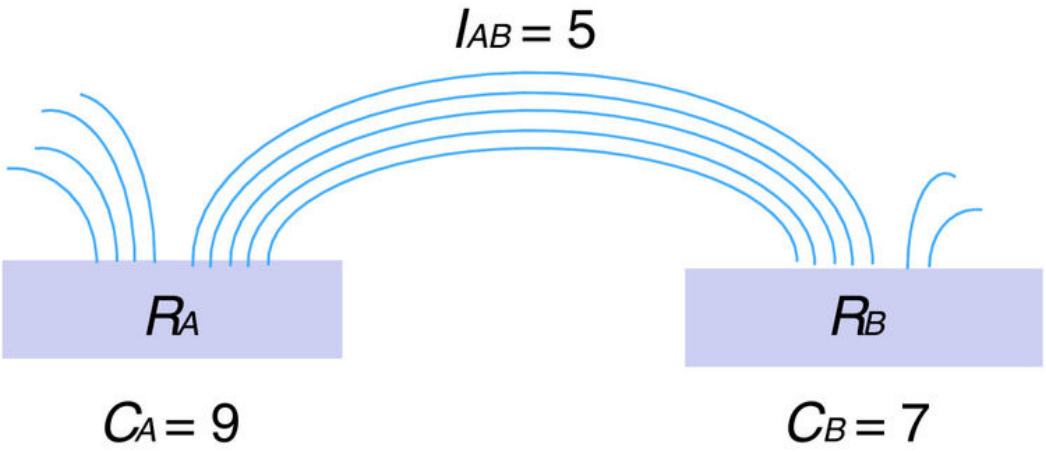
100Kb resolution

Dixon et al. 2012 Nature

Identification of chromatin loops



Identiying significant interaction



$$\Pr(I_{A,B} \mid N, c_A, c_B) = \frac{\binom{c_A}{I_{A,B}}\binom{2N - c_A}{c_B - I_{A,B}}}{\binom{2N}{c_B}}$$

Thank you for your attention