L2.2

Epigenomes

Post-genomics

Genetics

Comparative (phylogenetic conservation indicates conserved function)

Human Genetic Variation (1000 Human Genomes - HapMap)

GWAS – Genome variations – phenotype correlation

Gene expression and phenotype

Functional Genomics

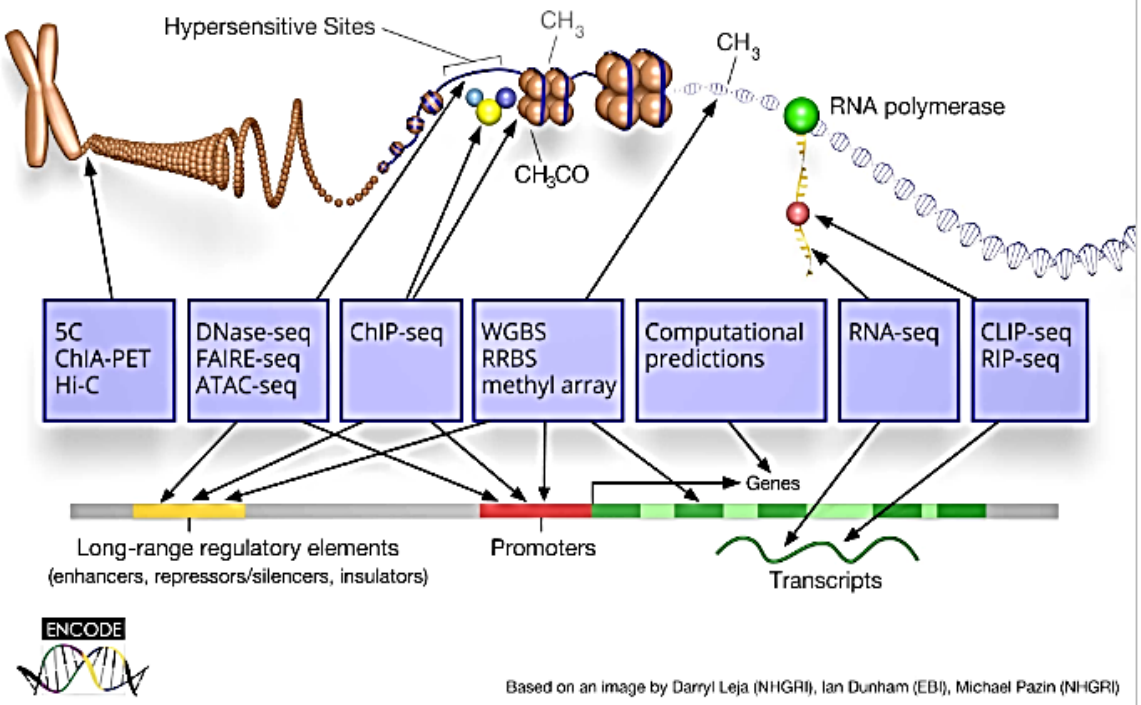
Epigenomics:	CpG methylation
	Histone modifications (PTMs)
	Chromatin status
	Protein-DNA mapping (e.g. Transcription Factors)
Transcriptomics:	Coding and noncoding RNAs

Sicuro | https://www.encodeproject.org

Home - PubMed - N | Università di Torino | Home Page - MedInT | Course: Advanced M | Innovative Training N | Dictionary | Definitio | Didattica | bioinformatics

ENCODE Data Encyclopedia Materials & Methods Help Search...

ENCODE: Encyclopedia of DNA Elements



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

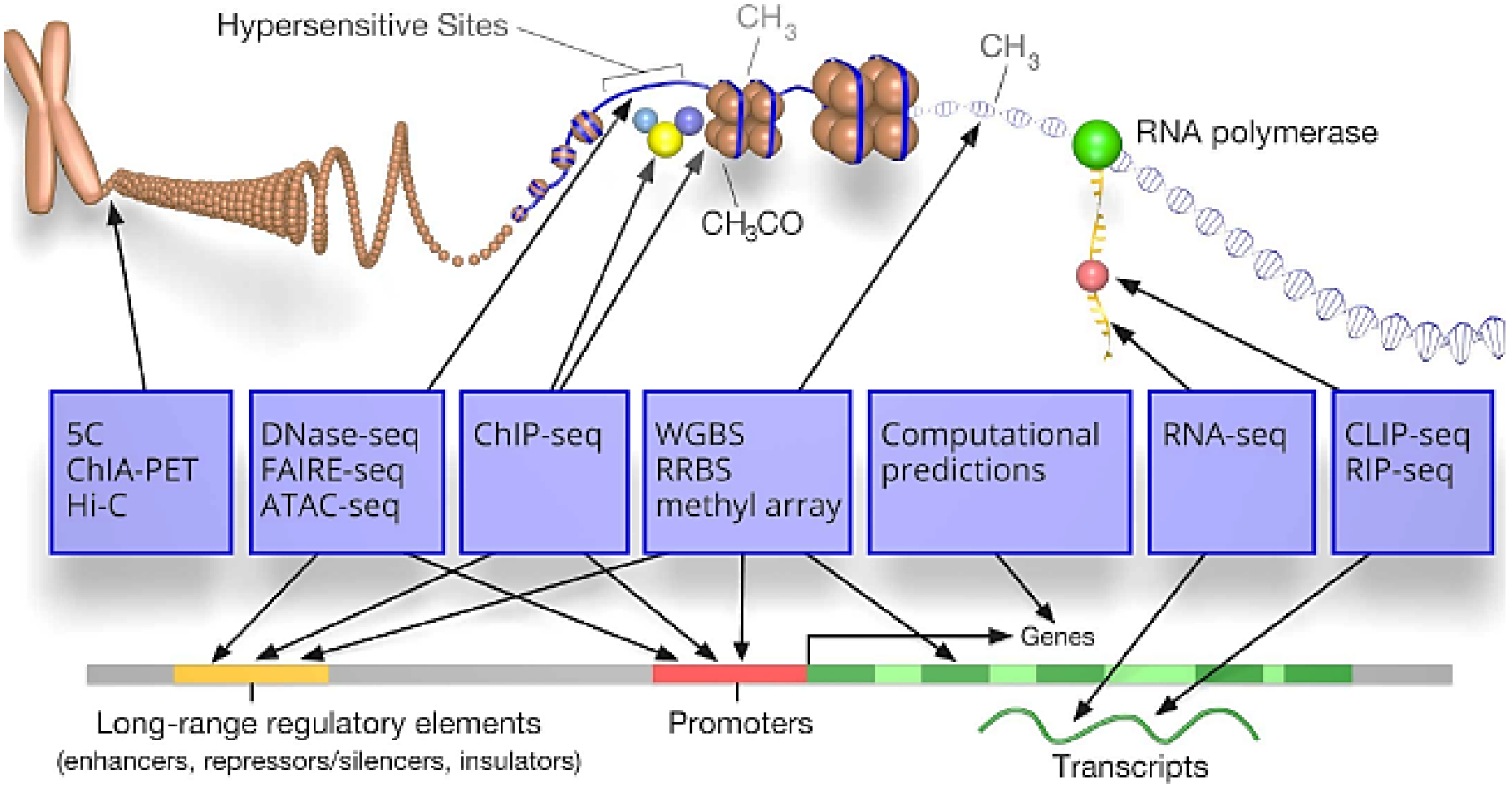
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Get Started

HUMAN MOUSE WORM FLY

[The ENCODE Project Website]

ENCODE: Encyclopedia of DNA Elements



Can we access information on chromatin **accessibility** **genome-wide** ?

Dnase-Seq
FAIRE-Seq (Formaldehyde Assisted Isolation of Regulatory Elements)
ATAC-Seq (Assay for Transposase-Accessible Chromatin)

All these methods are coupled to NGS, but they have been (and still are) also analyzed using microarrays

DNA microarrays

→ ordered arrays of single-stranded DNA probes designed to be complementary to a set of «genomic objects» (exons, promoters, enhancers, all nonrepetitive intergenic spaces, etc.)

Studies on histone modifications, W/E/R proteins, TF, CpG methylation have been made on different kind of « **tiling arrays** ».

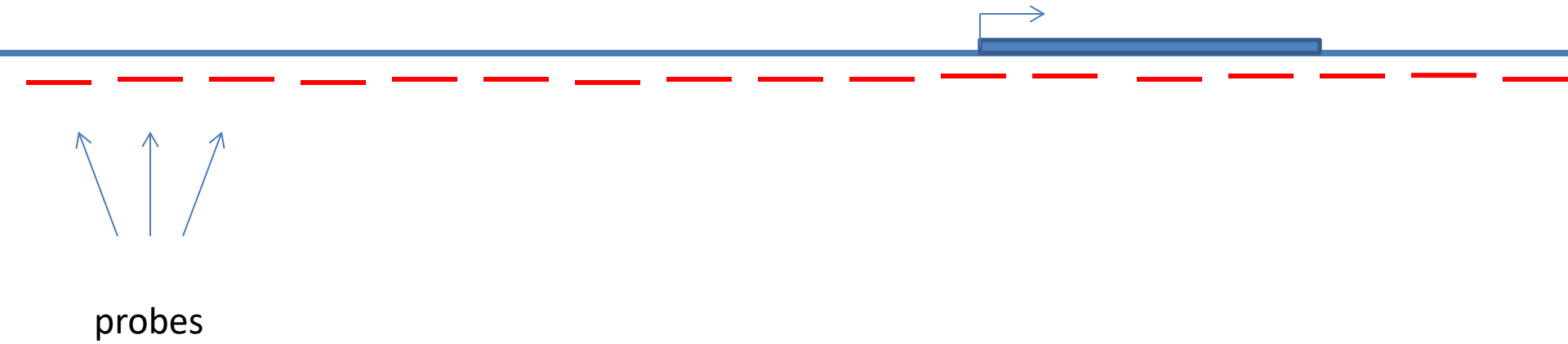
Technologically developed ca. 10 years before NGS, somehow cheaper, at least when NGS was in early phase.

Background help available

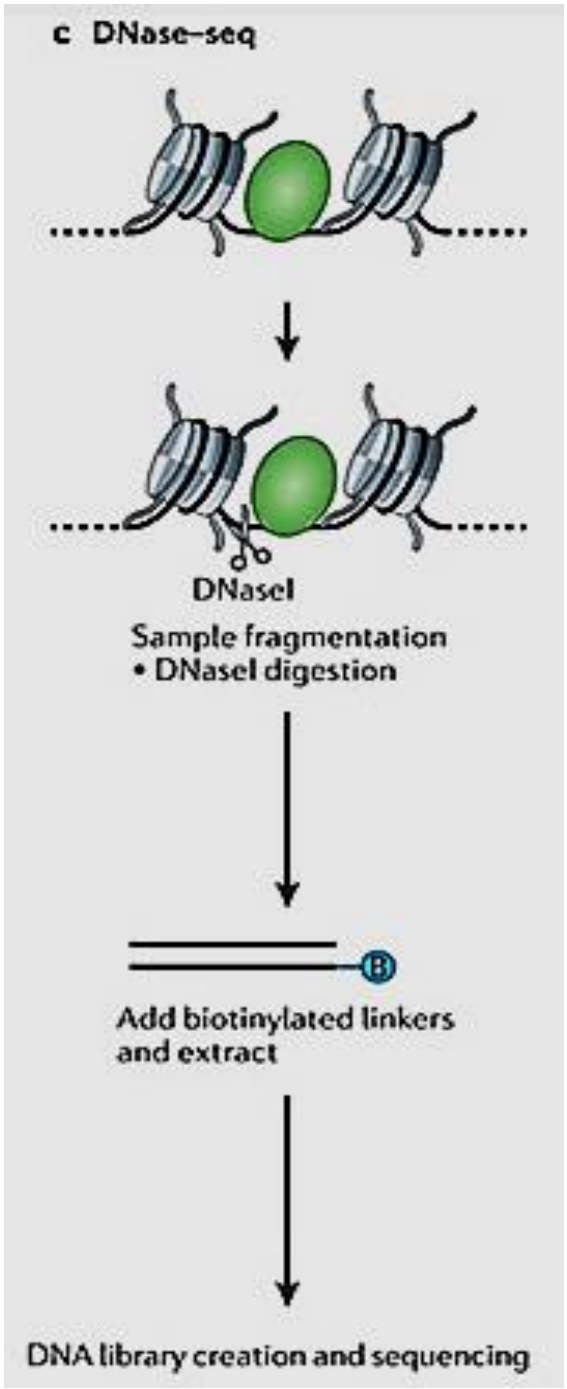
DNA microarrays

Tiling microarray

Probes «cover» all the genome, regardless of the annotated function



probes



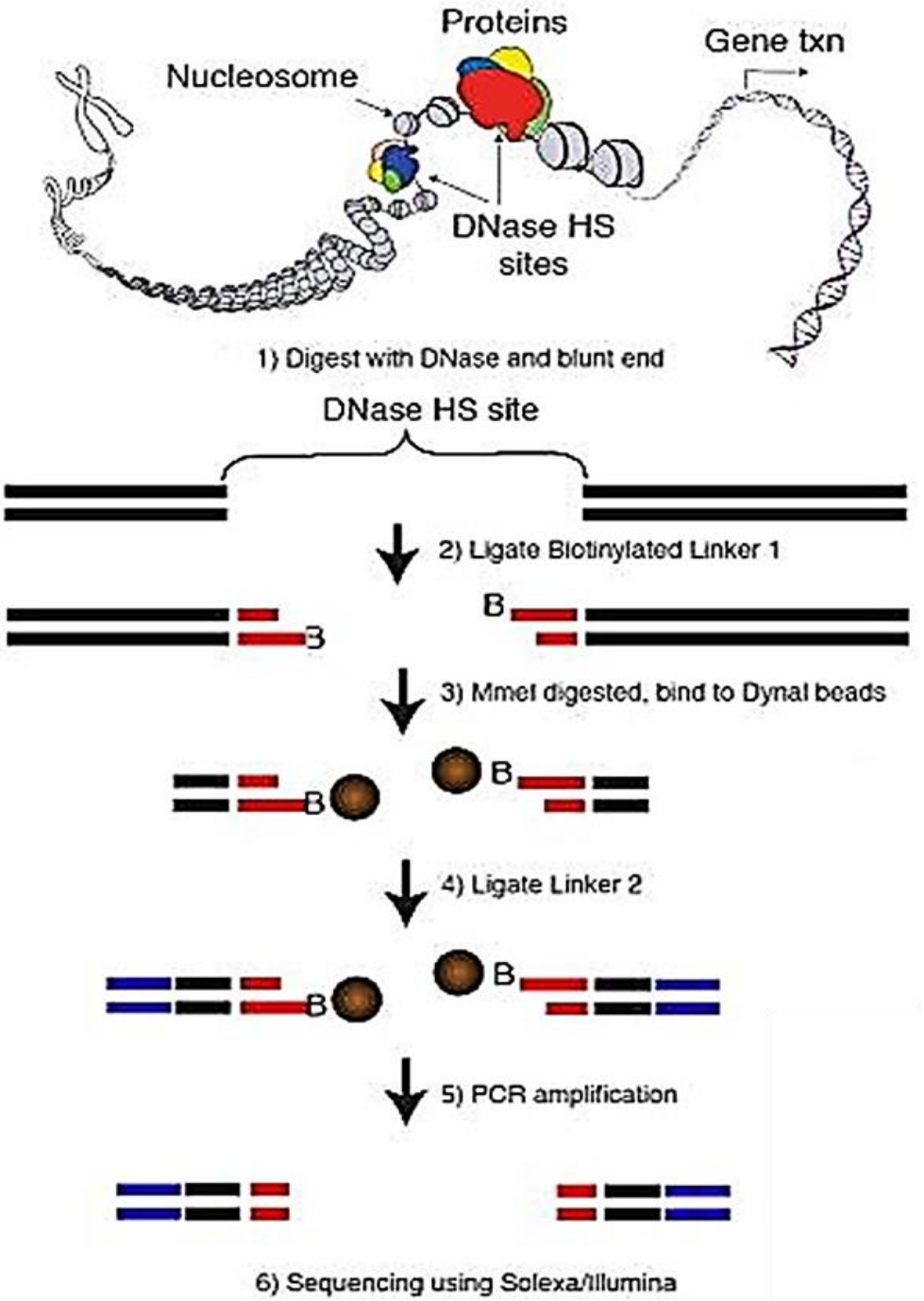
DNase I sensitivity assay

→ **DNase-Seq**

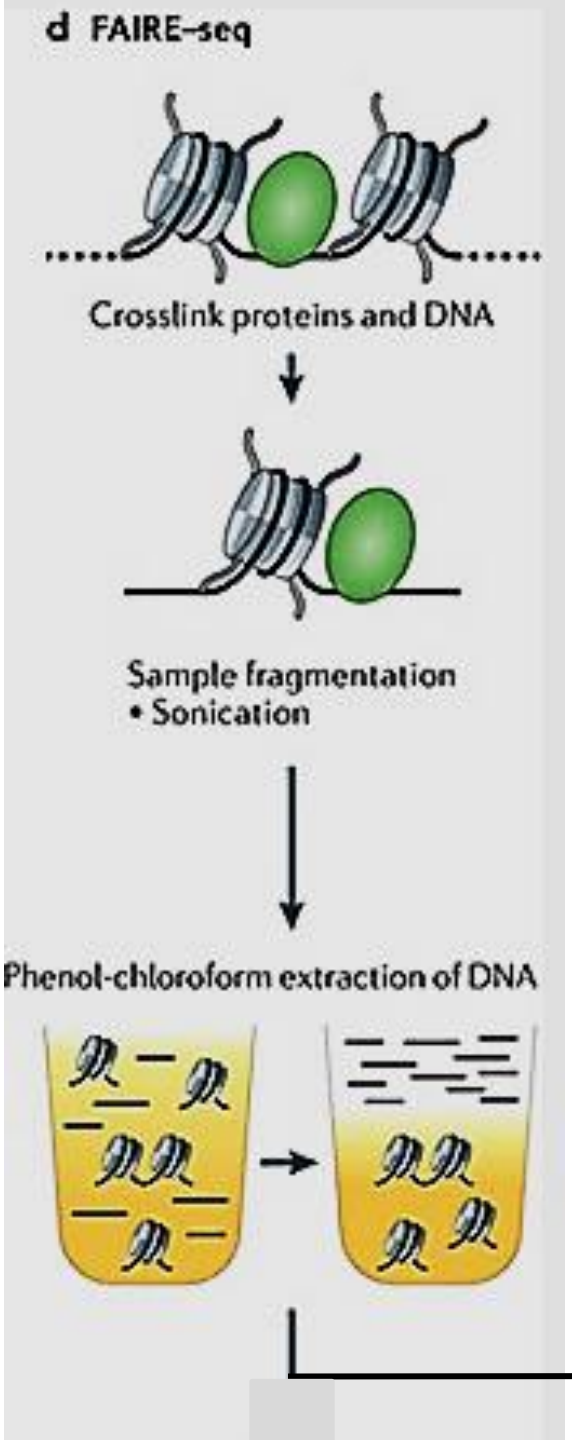
DNase–seq experiments combine traditional DHS assays with high-throughput sequencing to simultaneously identify all types of regulatory regions genome-wide.
The 5′ end of a sequence tag generated by DNase–seq indicates the site of a DNaseI digestion event, and regions of enrichment in digestion events are identified as DHS sites, each of which can contain binding sites of multiple factors.

Furey 2012, Nat Rev Genet. 13: 840-852

DNase-Seq



Song & Crawford 2010
Cold Spring Harbor Protocols

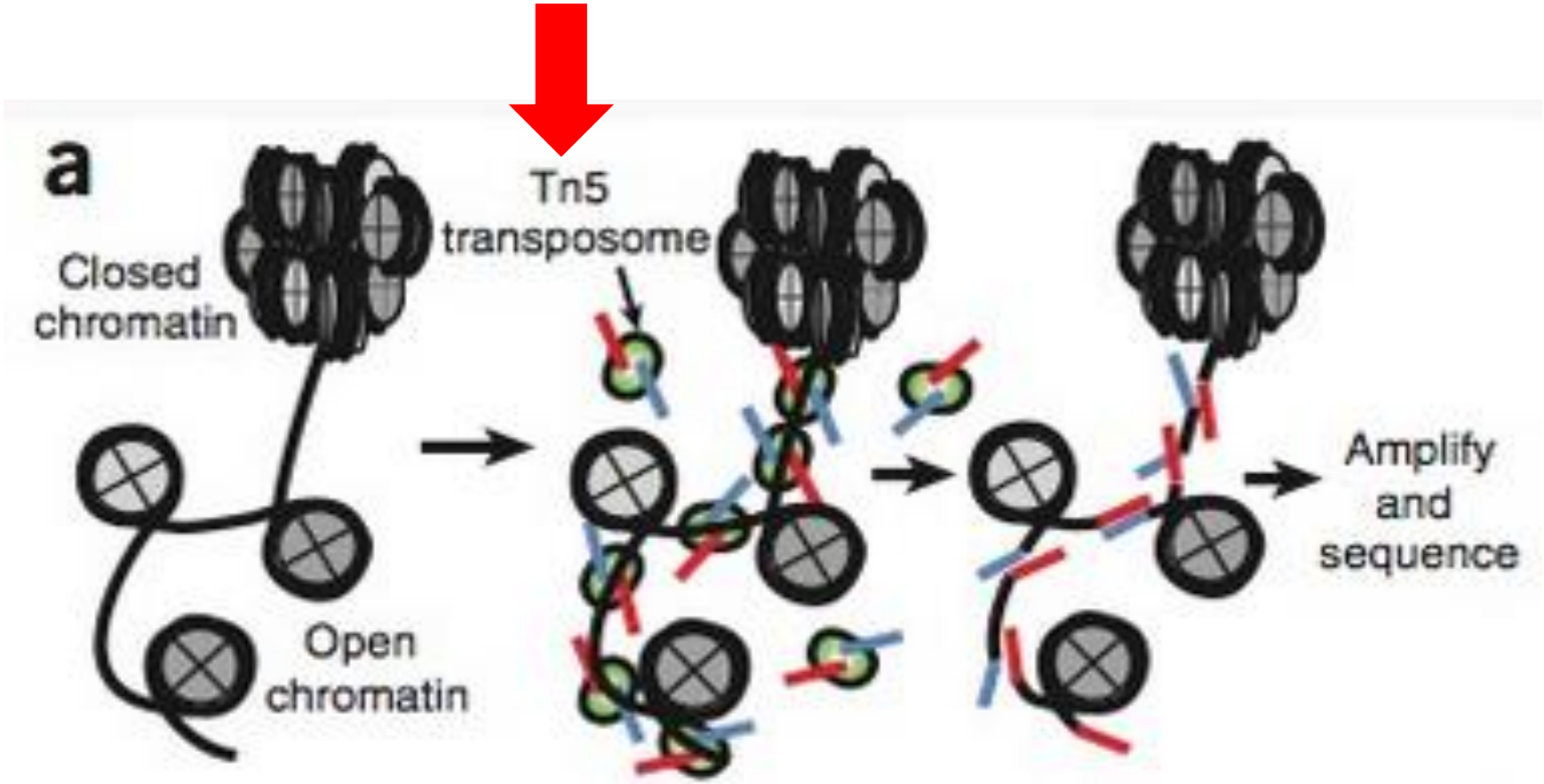


FAIRE–seq

The FAIRE–seq assay starts with formaldehyde crosslinking, similarly to ChIP, but then instead of using an antibody to target specific factors, DNA is sonicated and the extract is subjected to phenol-chloroform extraction. The nucleosome-depleted fraction of DNA is preferentially segregated to the aqueous phase. FAIRE-enriched DNA has been shown to correspond to regulatory regions.

Furey 2012, Nat Rev Genet. 13: 840-852

DNA library creation and sequencing

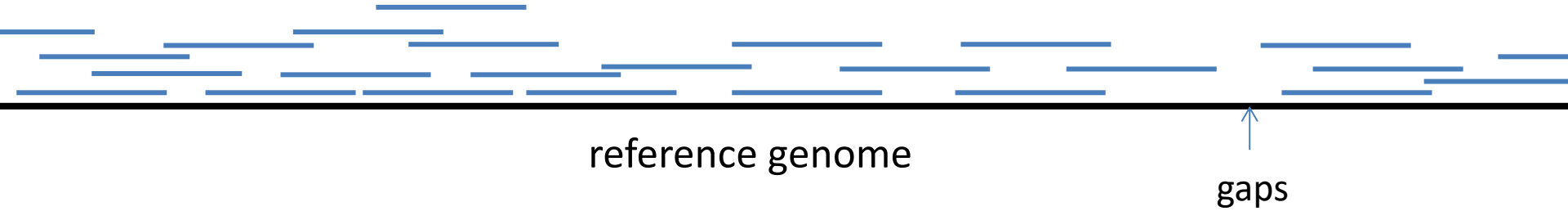


(a) ATAC-seq reaction schematic.
Transposase (green), loaded with sequencing adaptors (red and blue), inserts only in regions of open chromatin (between nucleosomes in gray) and generates sequencing-library fragments that can be PCR-amplified.

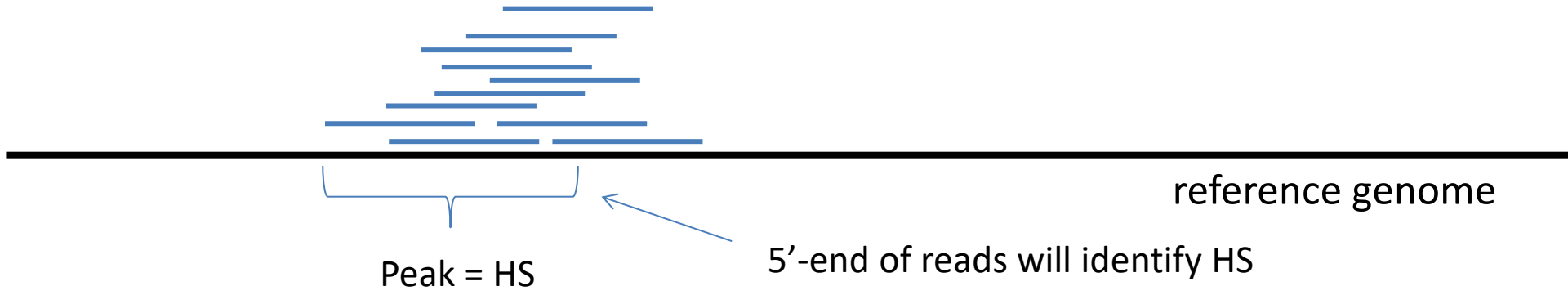
Buenrostro et al., Nat Methods. 2013, 10(12):1213-8.

Mapping data.

Results are very different from those obtained from NGS on genomic DNA, which are a «probabilistic» distribution of fragments on the reference genome.



ATAC-Seq will give only fragments from open regions



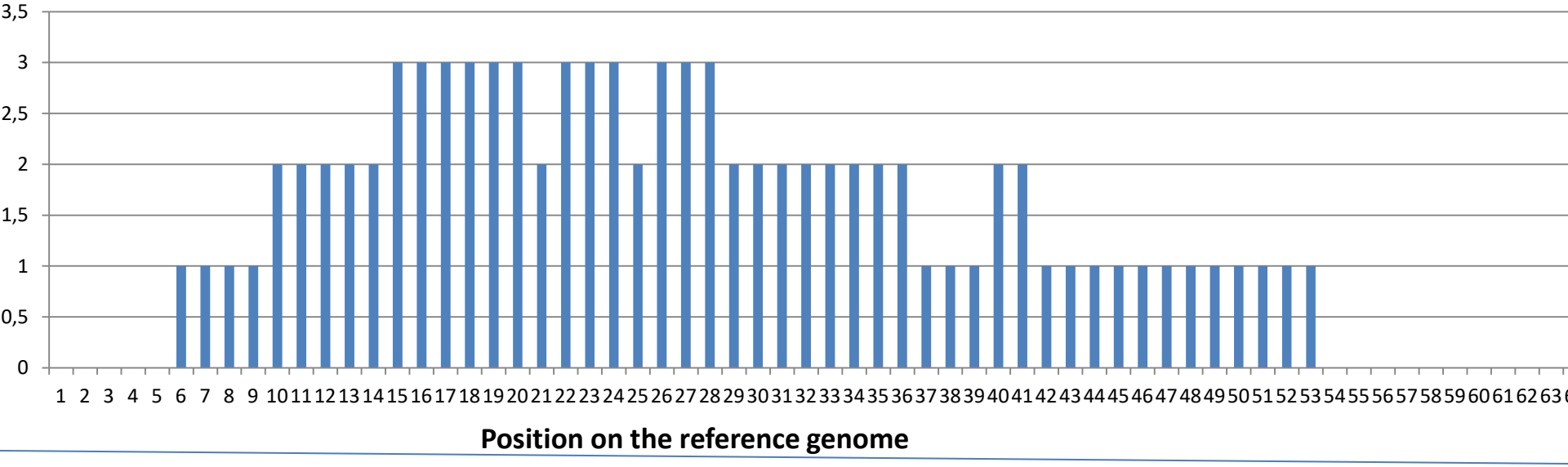
Reads alignment to reference genome, count hits at each position

```
                        tgggactgactggctg
              caaatcaatcatgg                                  reads
                     atcatgggactgact
          gcatgcaaatcaatc
cattgcatgcaaatc                              tgatcgctgactgc
```

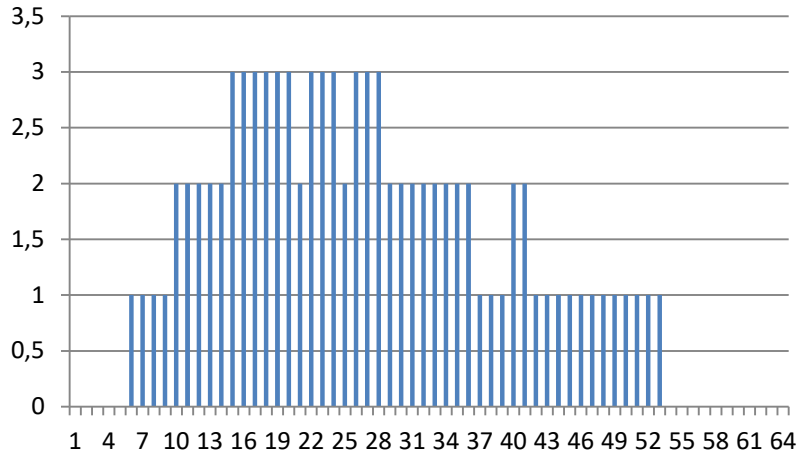
atccgtcattgcatgcaaatcaatcatgggactgactggctgatcgctgactgcctgcagtcgca

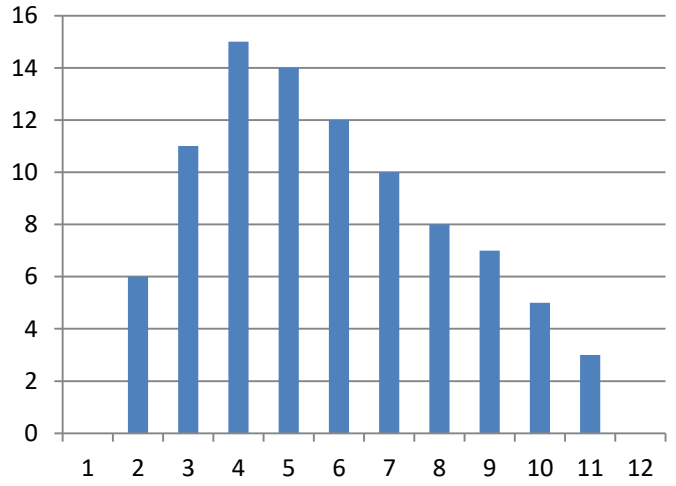
Counts/nucleotide:

000000111122222333333233323333222222221112211111111110000000000

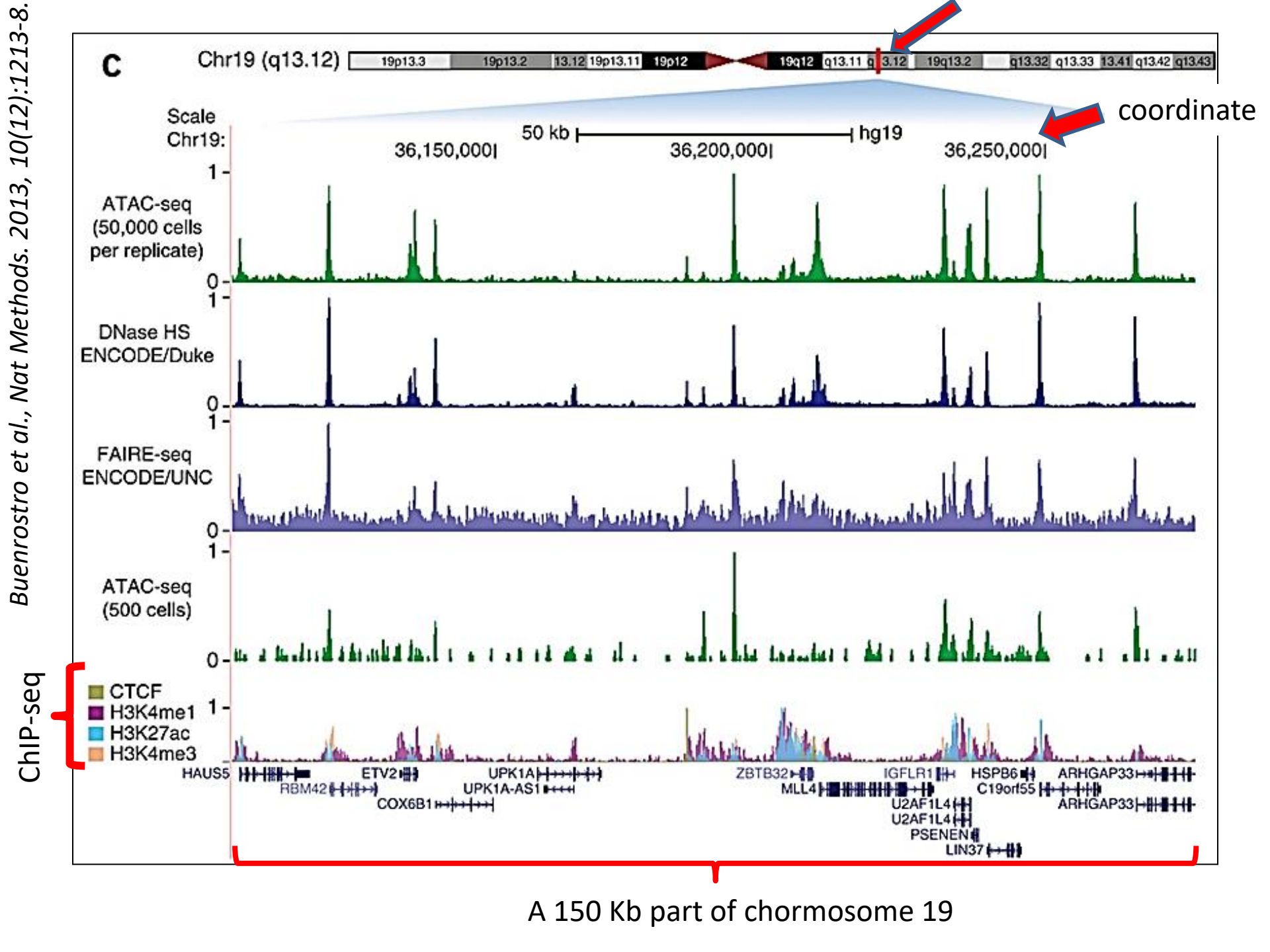


In order to simplify calculation, often nucleotides are grouped in **bins** of 5, 10 or more





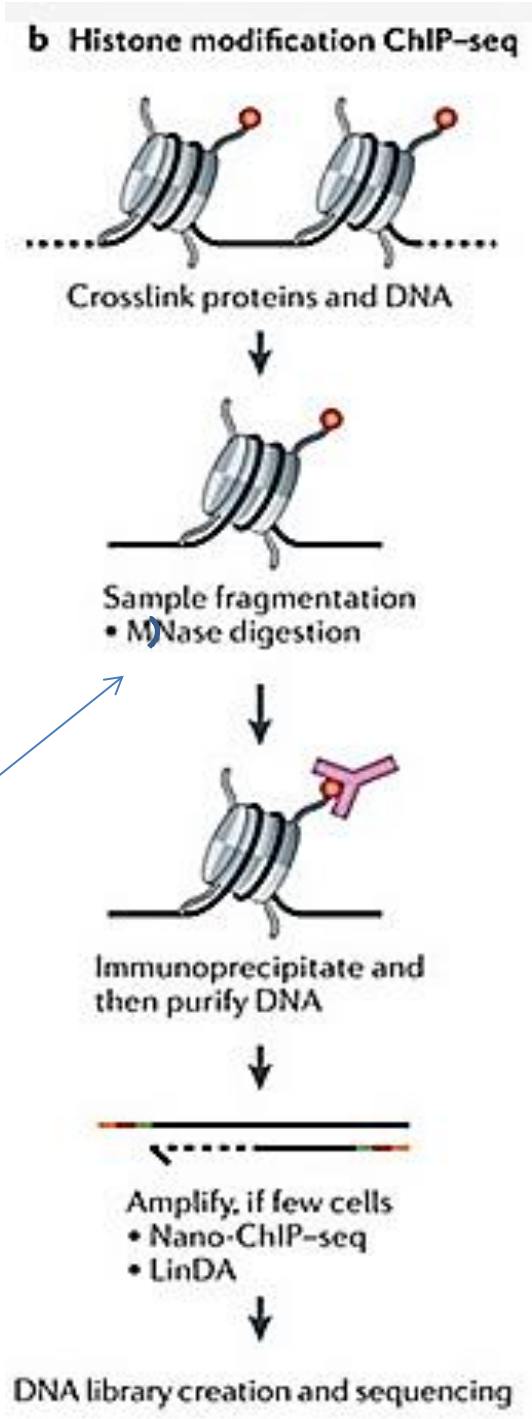
Bins=5bp


Buenrostro et al., Nat Methods. 2013, 10(12):1213-8.
c
Chr19 (q13.12)
coordinate
Scale
Chr19:
50 kb
hg19
36,150,000
36,200,000
36,250,000
ATAC-seq
(50,000 cells
per replicate)
DNase HS
ENCODE/Duke
FAIRE-seq
ENCODE/UNC
ATAC-seq
(500 cells)
ChIP-seq
CTCF
H3K4me1
H3K27ac
H3K4me3
HAUS5
ETV2
RBM42
COX6B1
UPK1A
UPK1A-AS1
ZBTB32
MLL4
IGFLR1
HSPB6
C19orf55
ARHGAP33
U2AF1L4
U2AF1L4
PSENEN
LIN37
A 150 Kb part of chormosome 19

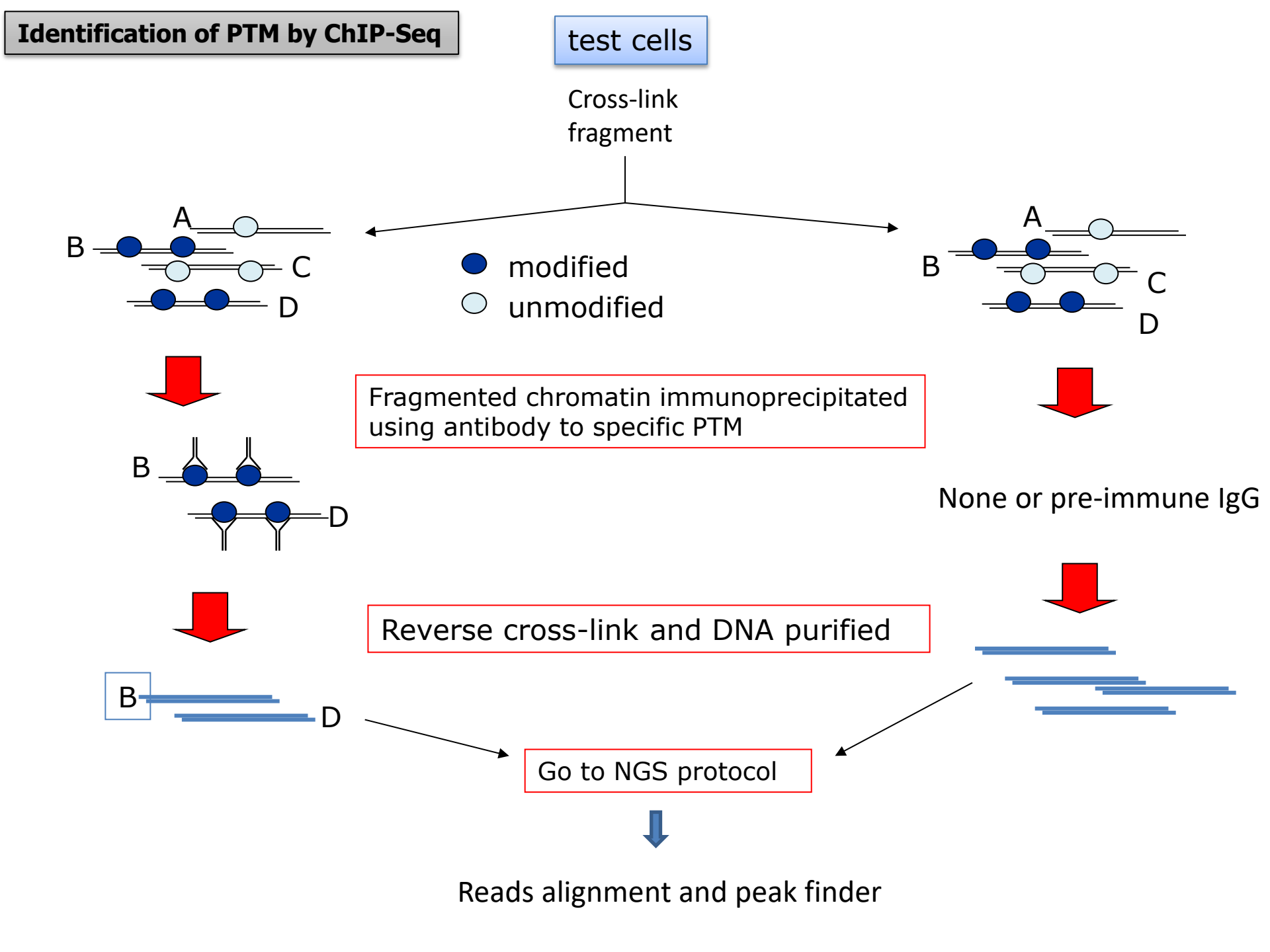
Histone PTMs are studied genome-wide using ChIP-Seq or **ChIP-on-chip** *What is this?*

Very **specific antibodies** are available, which not only recognize subtle differences (e.g. symmetric *versus* nonsymmetric arginine dimethylation) but also the peptide context in with PTM occurs (i.e. the exact aminoacid that is modified).

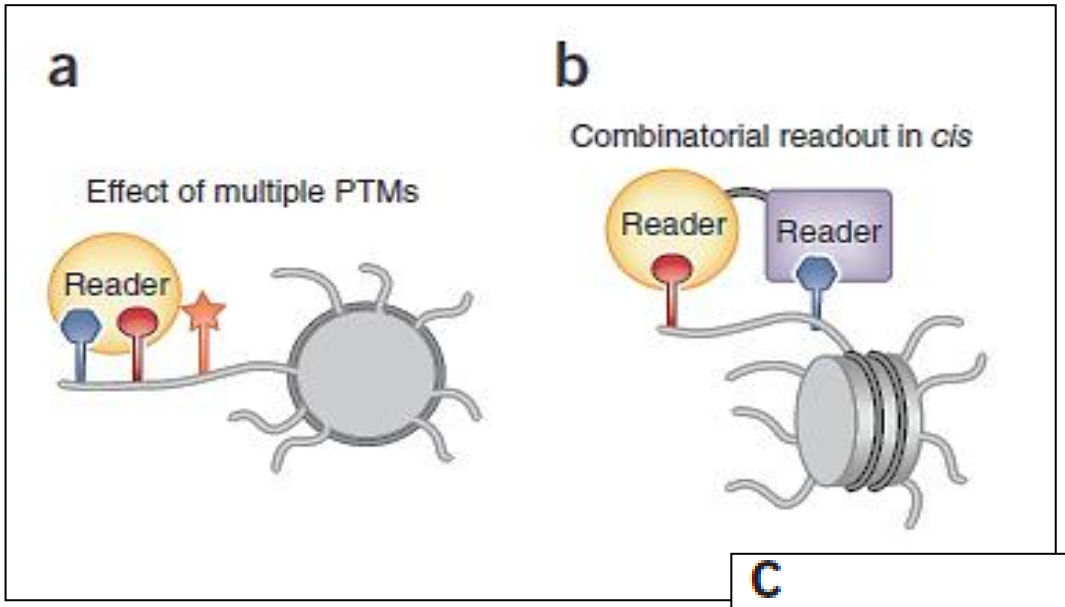
facultative

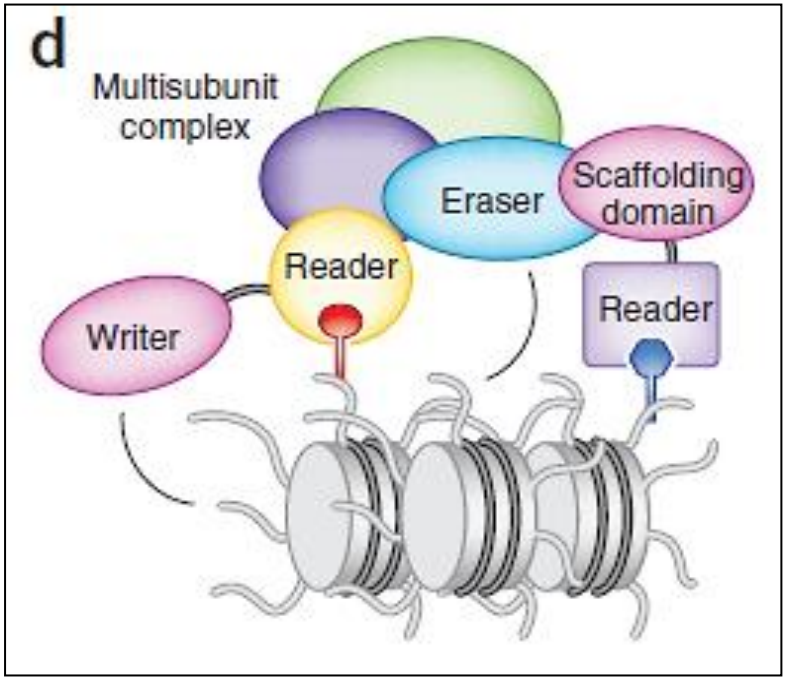


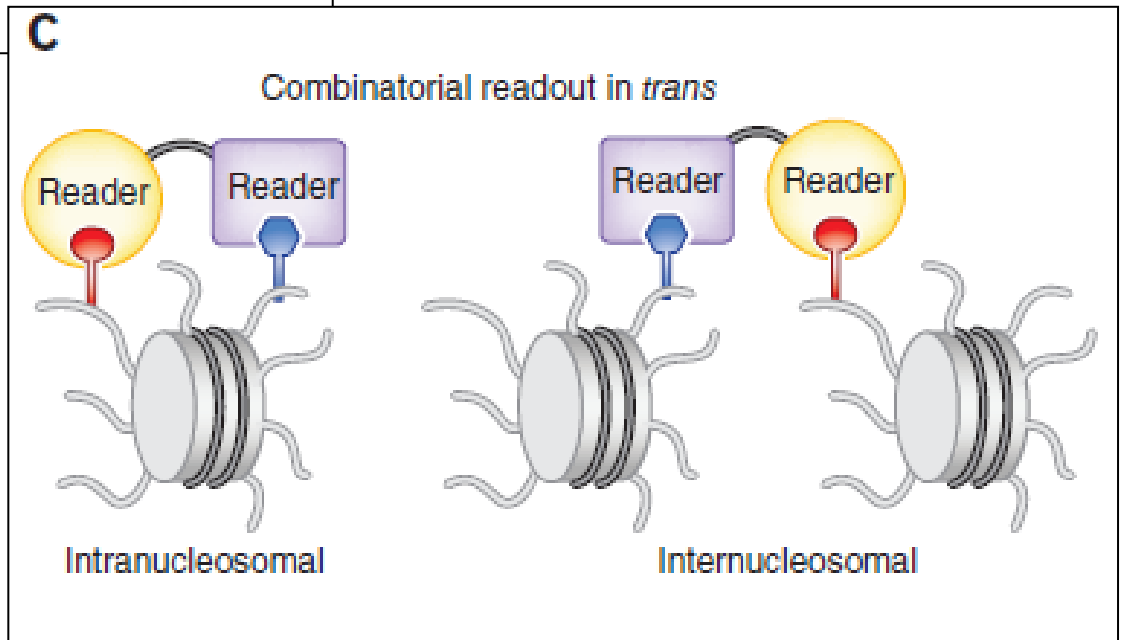
Identification of PTM by ChIP-Seq



Histone code means **multiple** PTMs

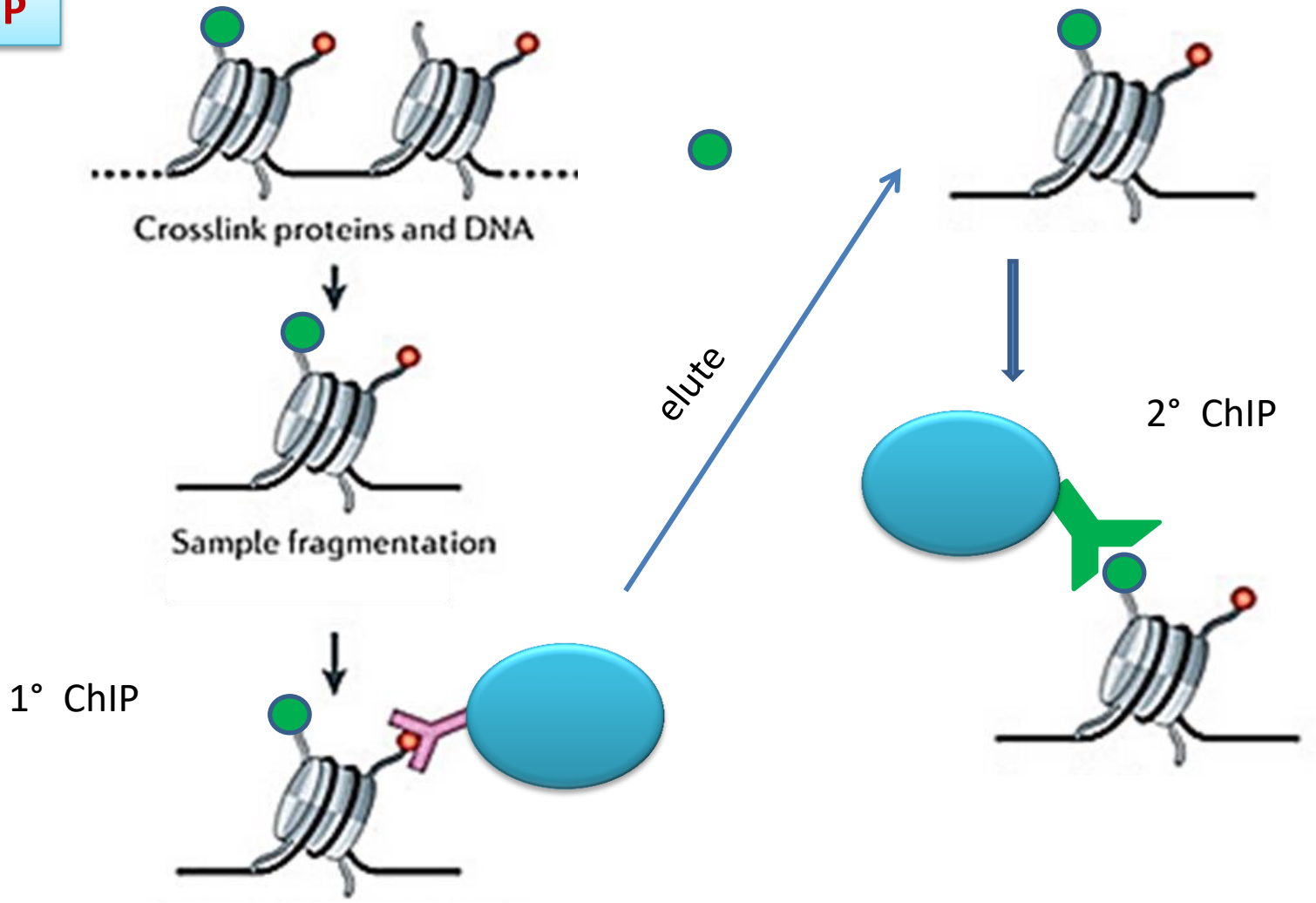






How can we be sure that several proteins are co-present on the same allele ?

Re-ChIP



Let us take a look of one of the first papers published, reporting PTMs analysis genome-wide using NGS (Illumina-Solexa first version)

Resource

Cell 129, 823–837 (2007)

Cell

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]

[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

[3] These authors contributed equally to this work and are listed alphabetically.

*Correspondence: zhaok@nhlbi.nih.gov

DOI 10.1016/j.cell.2007.05.009

SUMMARY

Histone modifications are implicated in influencing gene expression. We have generated high-resolution maps for the genome-wide distribution of 20 histone lysine and arginine methylations as well as histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF across the human genome using the Solexa 1G sequencing technology. Typical patterns of histone methylations exhibited at promoters, insulators, enhancers, and transcribed regions are identified. The monomethylations of H3K27, H3K9, H4K20, H3K79, and H2BK5 are all linked to gene activation, whereas trimethylations of H3K27, H3K9, and H3K79 are linked to repression. H2A.Z associates with functional regulatory elements, and CTCF marks boundaries of histone methylation domains. Chromosome banding patterns are correlated with unique patterns of histone modifications. Chromosome breakpoints detected in T cell cancers frequently reside in chromatin regions associated with H3K4 methylations. Our data provide new insights into the function of histone methylation and chromatin organization in genome function.

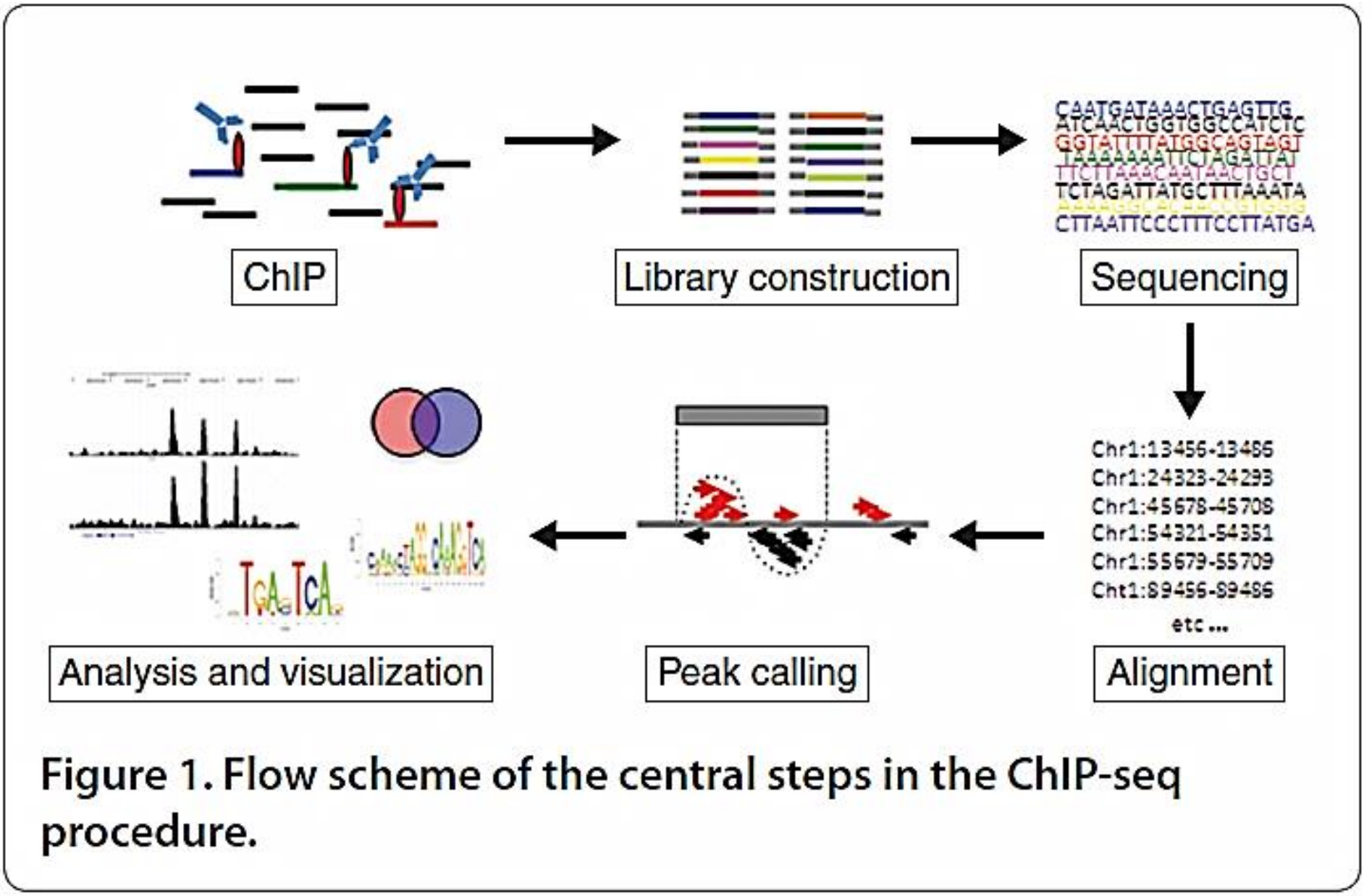
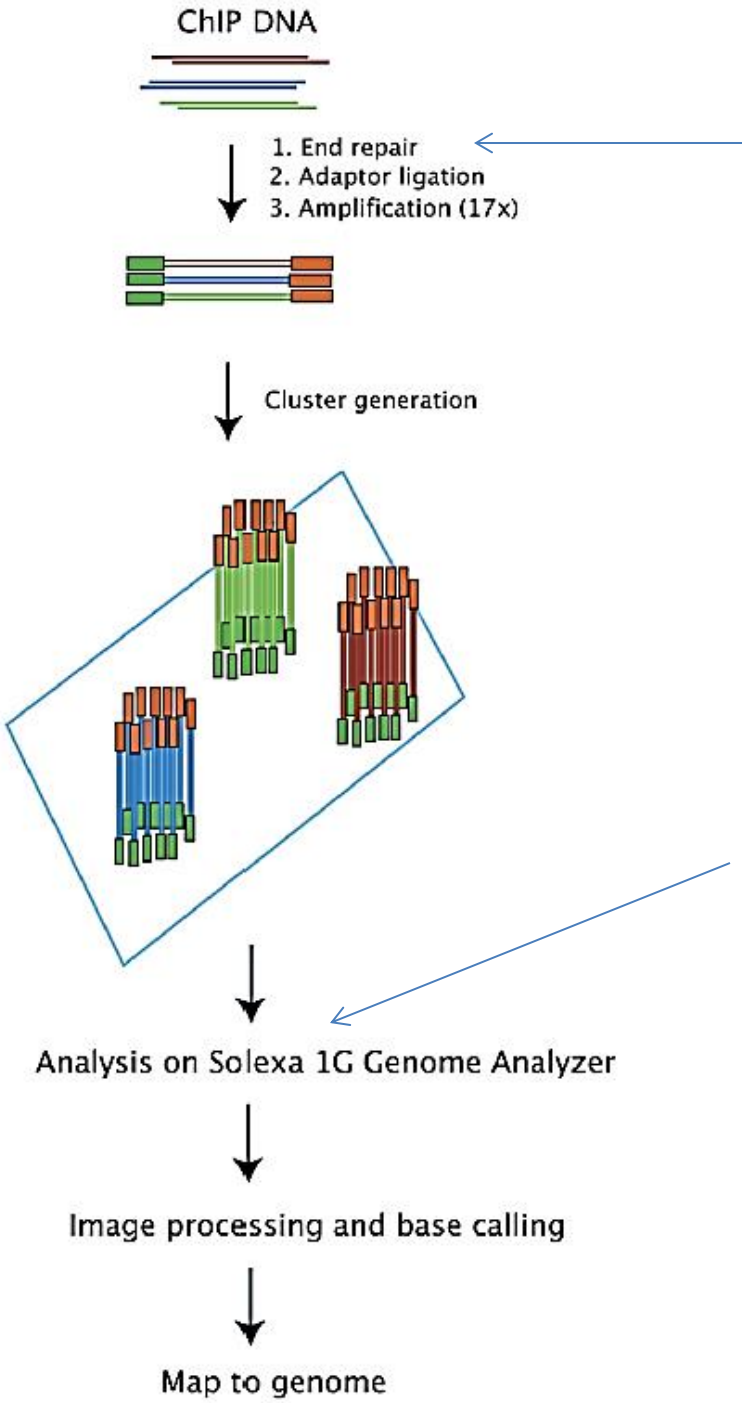


Figure 1. Flow scheme of the central steps in the ChIP-seq procedure.

Liu *et al*. *BMC Biology* 2010, **8**:56

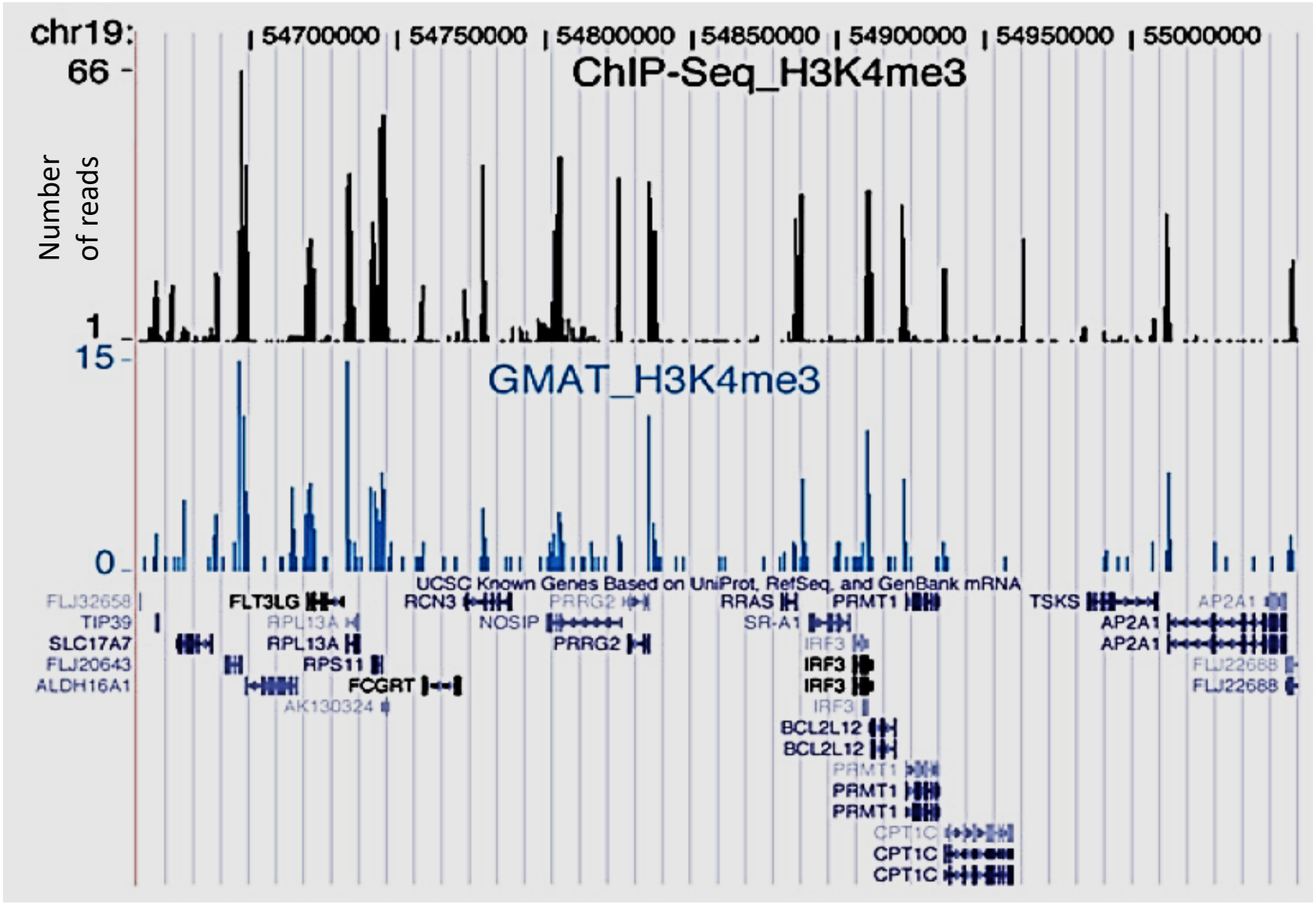


What is «end repair» ?

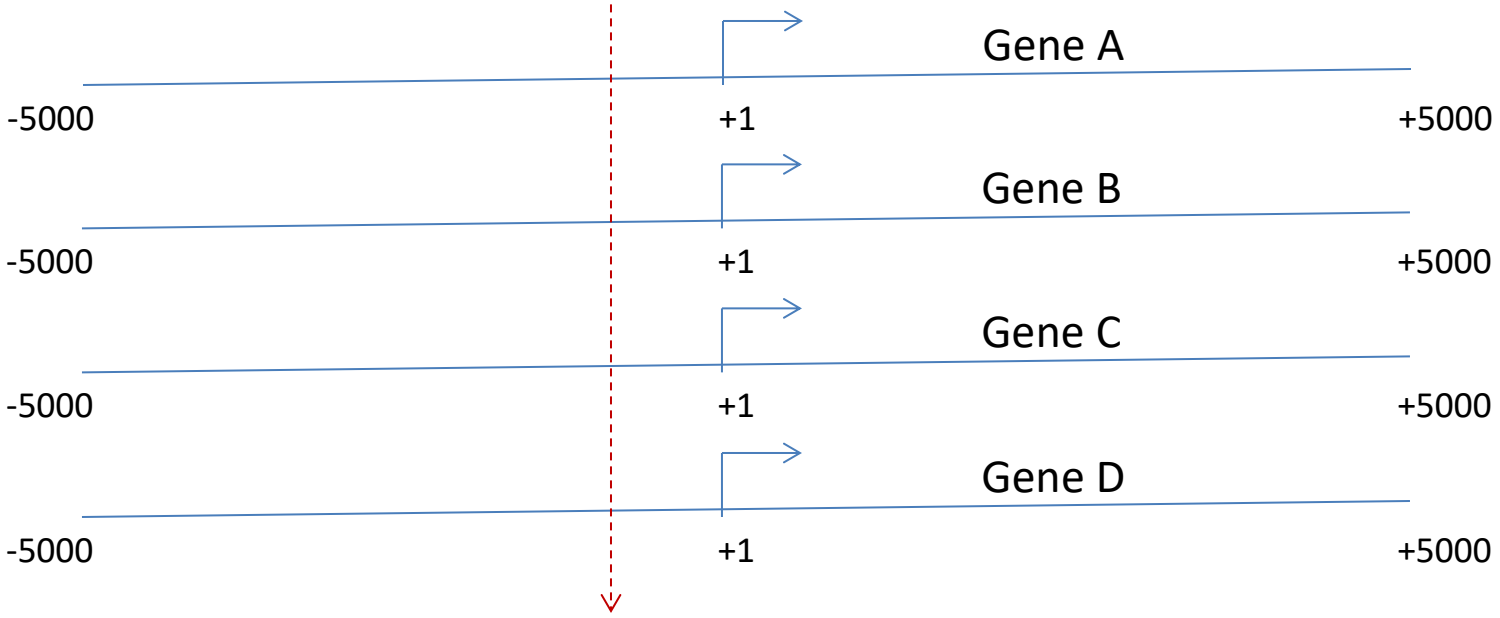
What is Solexa ?
Technology allowing sequencing of DNA on solid surfaces .
It was acquired by Illumina and applied to their NGS technology.

And map global:

Genome browser



Data are then collected from many similar position. In the figure, all RefSeq genes are taken into account and frequencies aligned using the TSS as central point, in a window of 5Kbp.
Then, the frequency of counts fro each position is calculated over the entire set of genes and plotted.



F(n) (the frequence of each nucleotide call)

Frequency representation

ChIP-Seq using anti-RNA Polymerase II (Pol II) antibody

ChIP-Seq using anti-H3 dimetyl-lysine 4 antibody

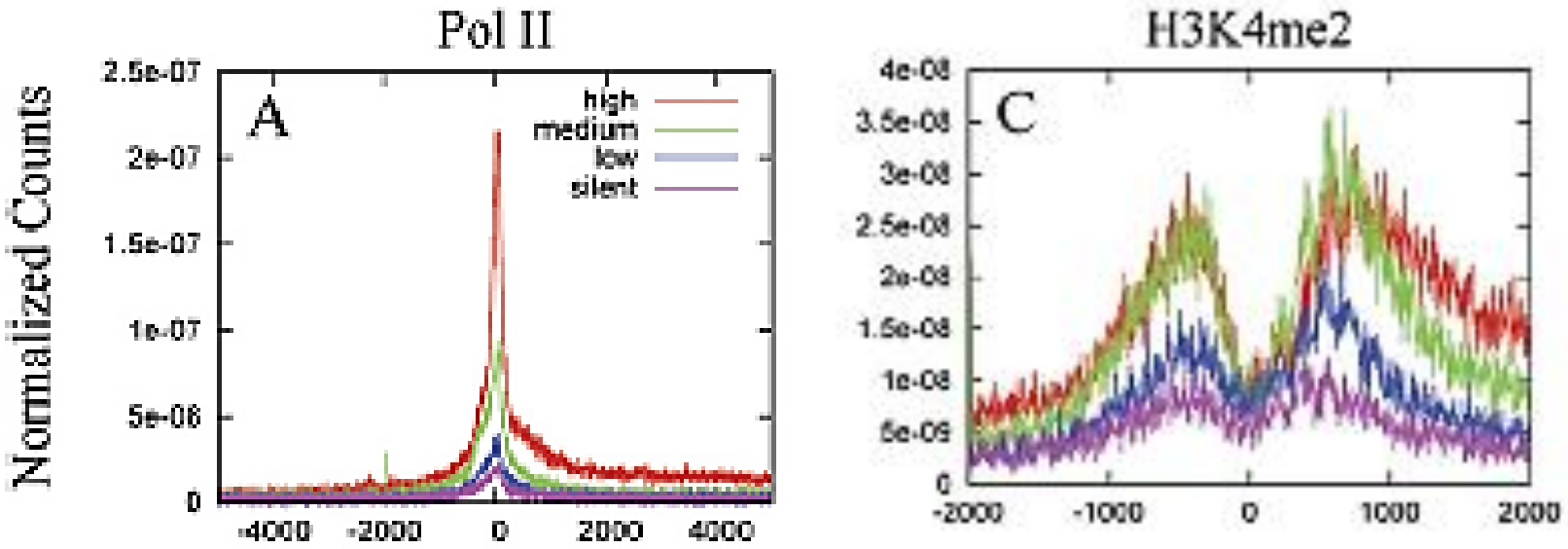
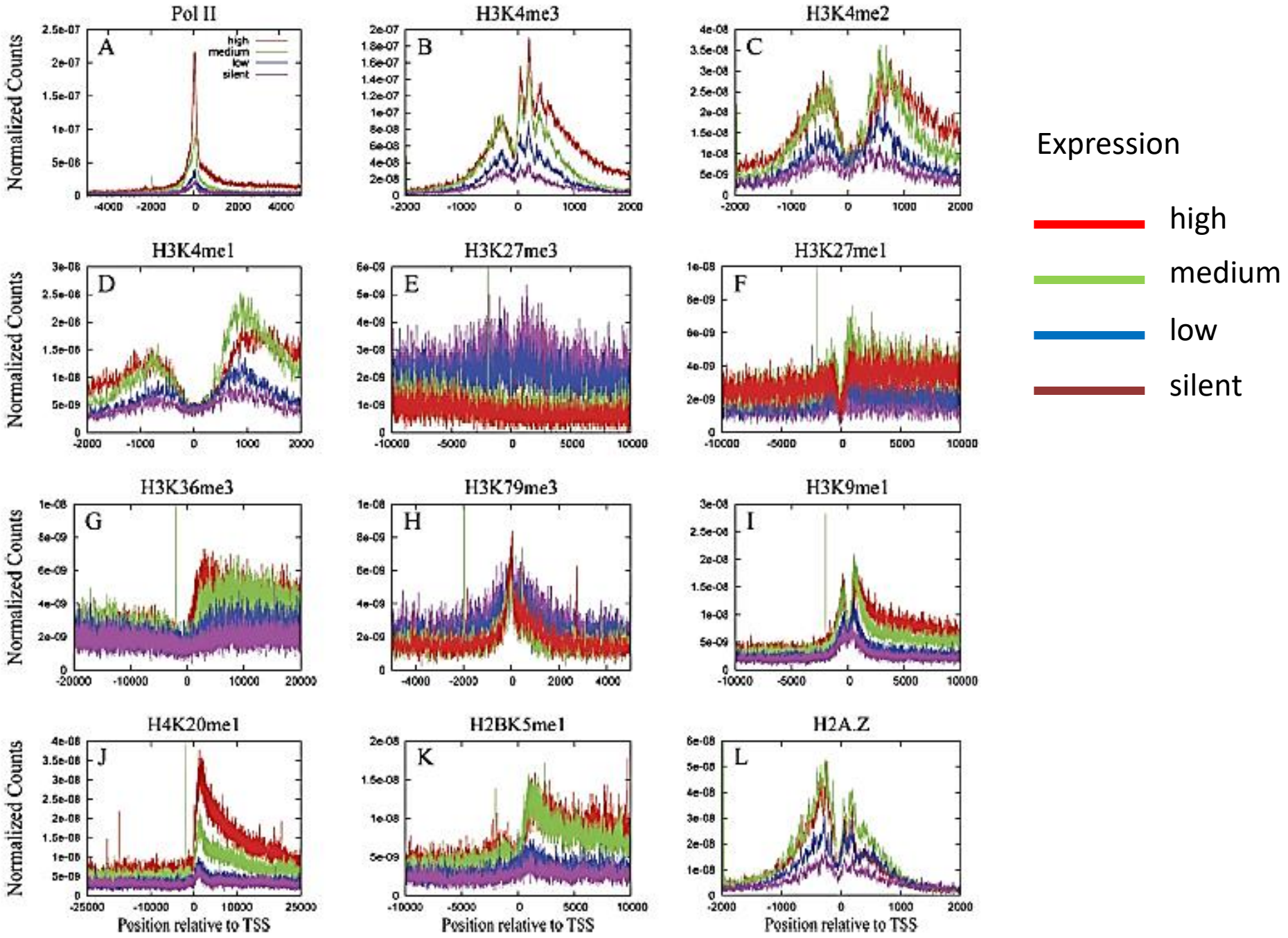
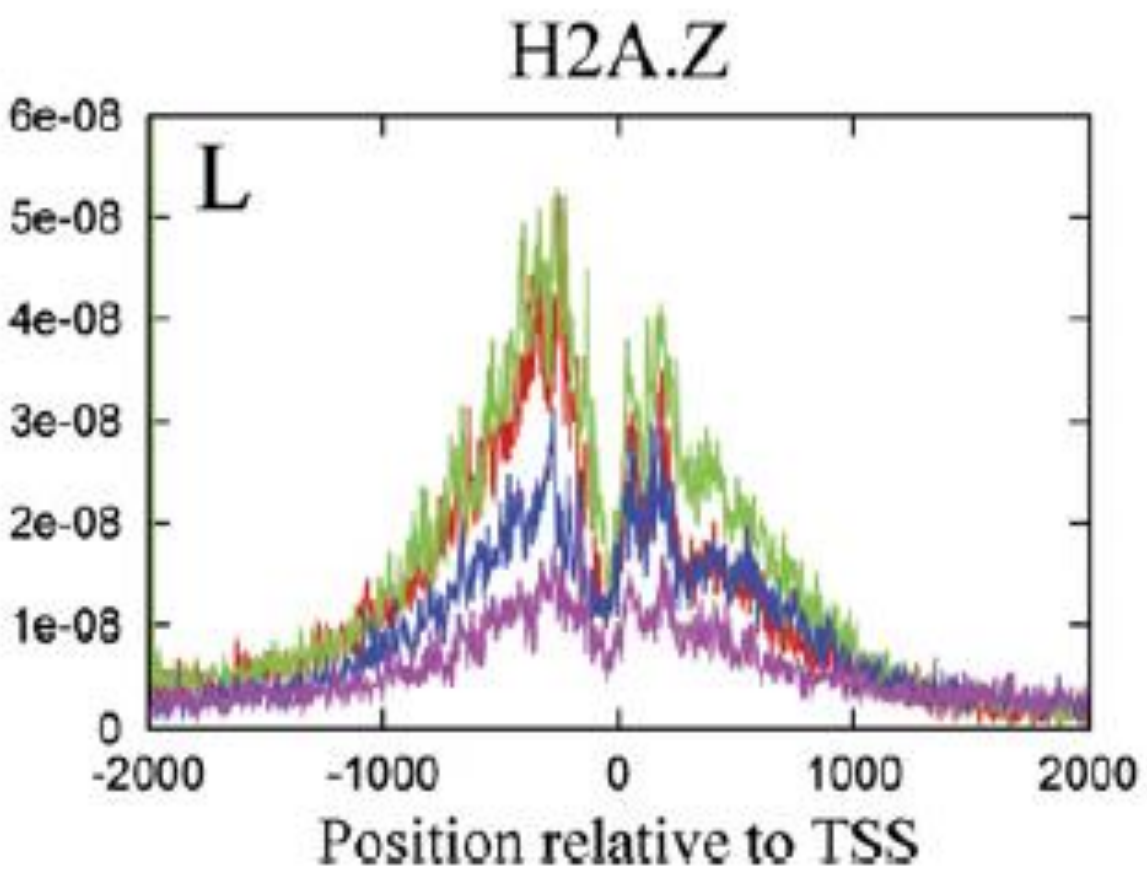


Figure 2. Histone Methylation near Transcription Start Sites
(A)–(L) Profiles of the histone methylation indicated above each panel across the TSS for highly active, two stages of intermediately active and silent genes are shown. Twelve thousand human genes were separated into twelve groups of one thousand genes according to their **expression level** (as measured on Affymetrix microarrays): **High – Medium – Low - Silent**


Pol II
H3K4me3
H3K4me2
H3K4me1
H3K27me3
H3K27me1
H3K36me3
H3K79me3
H3K9me1
H4K20me1
H2BK5me1
H2A.Z
A
B
C
D
E
F
G
H
I
J
K
L
high
medium
low
silent
Normalized Counts
Position relative to TSS
Expression
high
medium
low
silent



Expression

high

medium

low

silent

CpG methylation

a. **Cytonine-5-methylation** (+hydroxymethylation) is the only epigenetic modification concerning the DNA in higher organisms

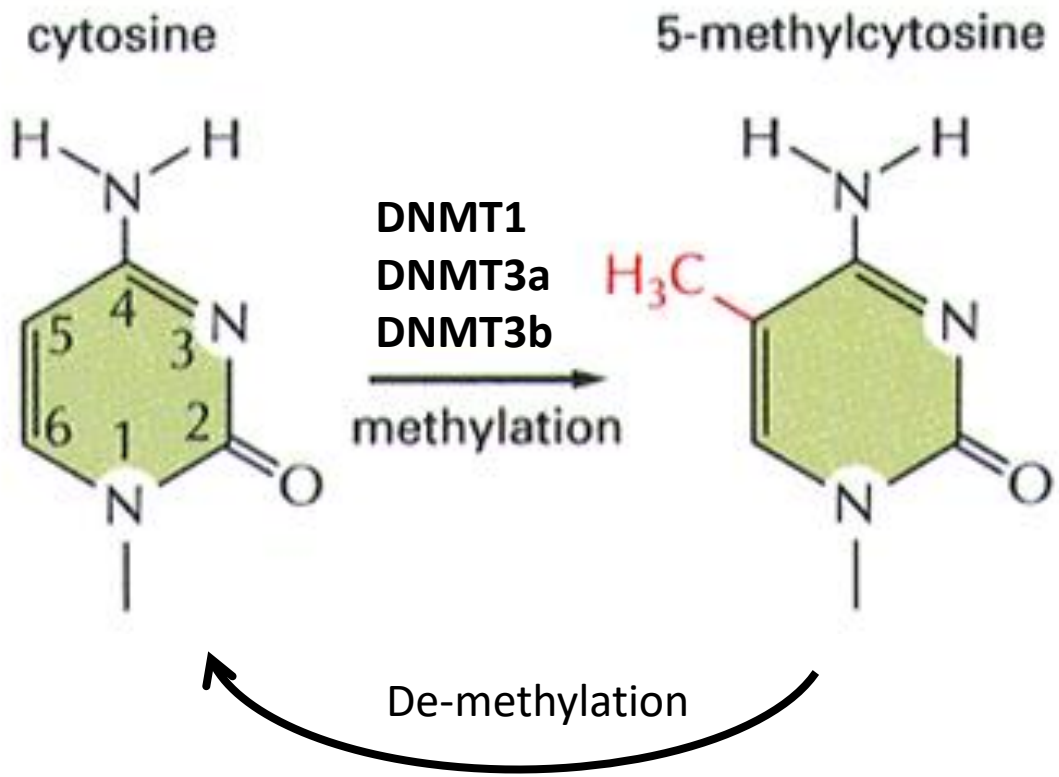
b. It occurs mostly at **CpG** dinucleotide (also at CpNpG in some cases, e.g. in plants and *see later in next chapter*)

c. Methylation of CpG is observed at regulatory regions of **silenced** genes

d. Hypermethylation is observed through the inactive X chromosome

e. Housekeeping genes (constitutively expressed) show unmethylated CpG islands at promoters

Important: Cytosine methylation is common in Mammals and Plants, but is not used in the same way in *S. cerevisiae* and *C. elegans*



1. **Oxidative pathway (Tet enzymes)**
2. **Dilution (DNMT1 off during repeated cell division)**

CpG methylation (hypermethylation, meaning that a region containing several CpG dinucleotides show several methylations) is usually associate with **silencing.** Possible mechanisms:

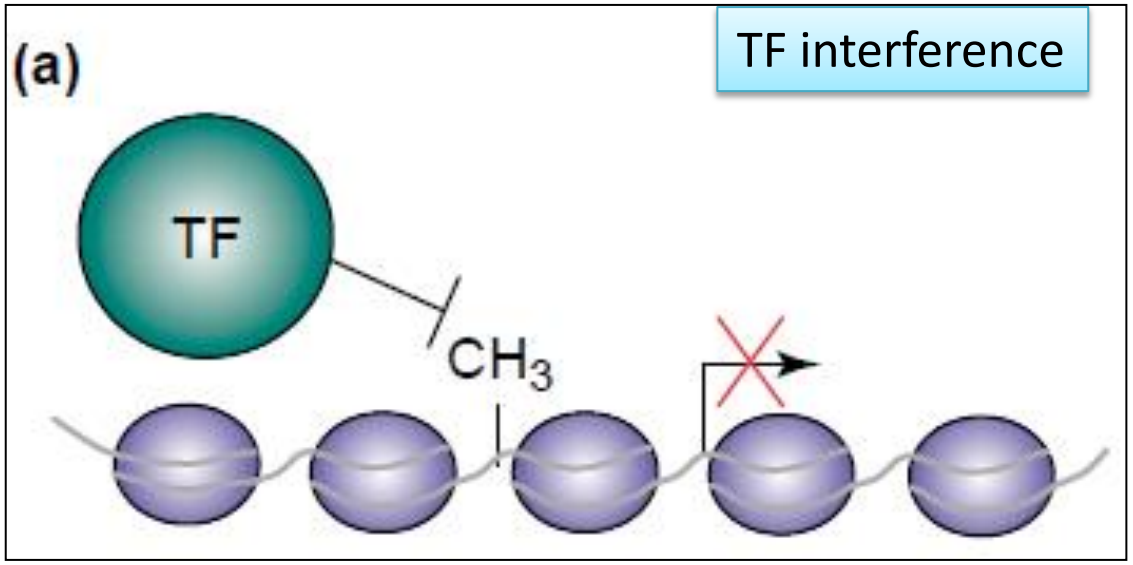


Figure 2. Mechanisms of DNA-methylation-mediated repression.
(**a**) DNA methylation in the cognate DNA-binding sequences of some (*not many, ndr*) transcription factors (TF) can result in inhibition of DNA binding. By blocking activators from binding targets sites, DNA methylation directly inhibits transcriptional activation

From Klose & Bird, 2006

Methyl-CpG binding proteins

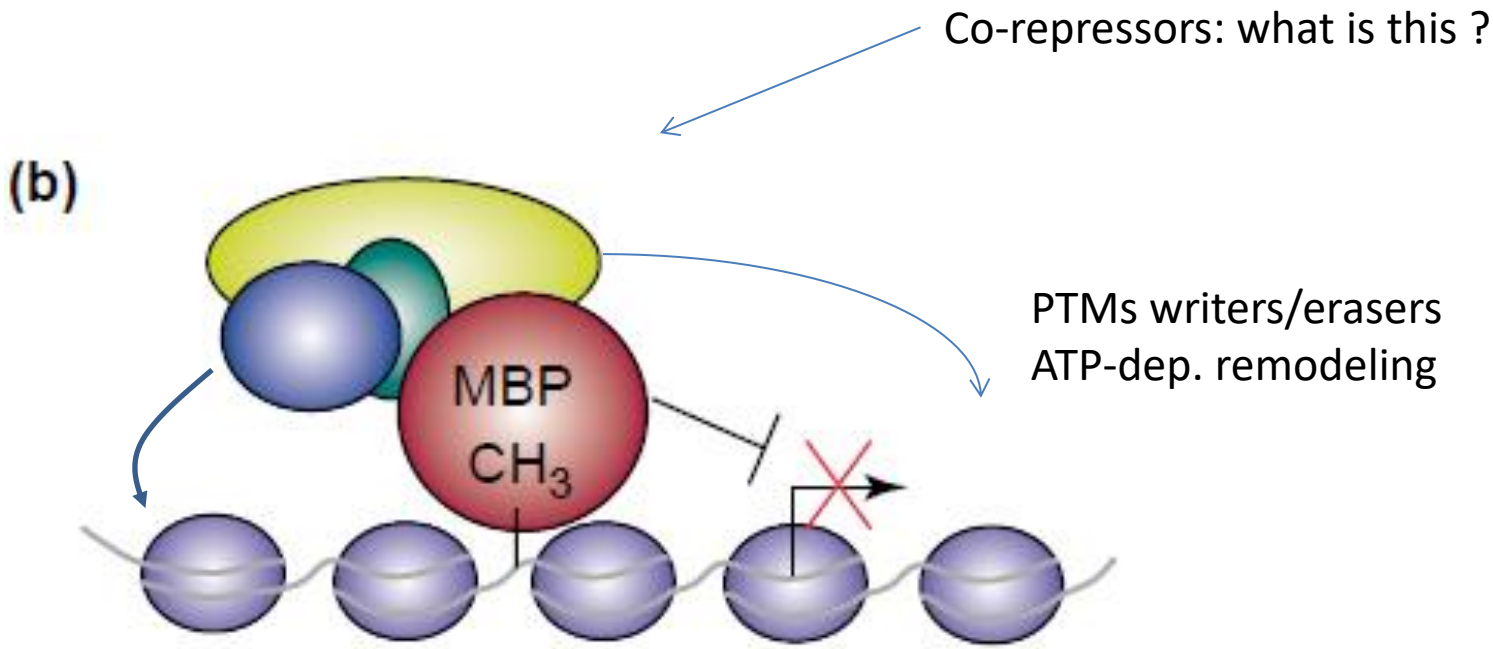


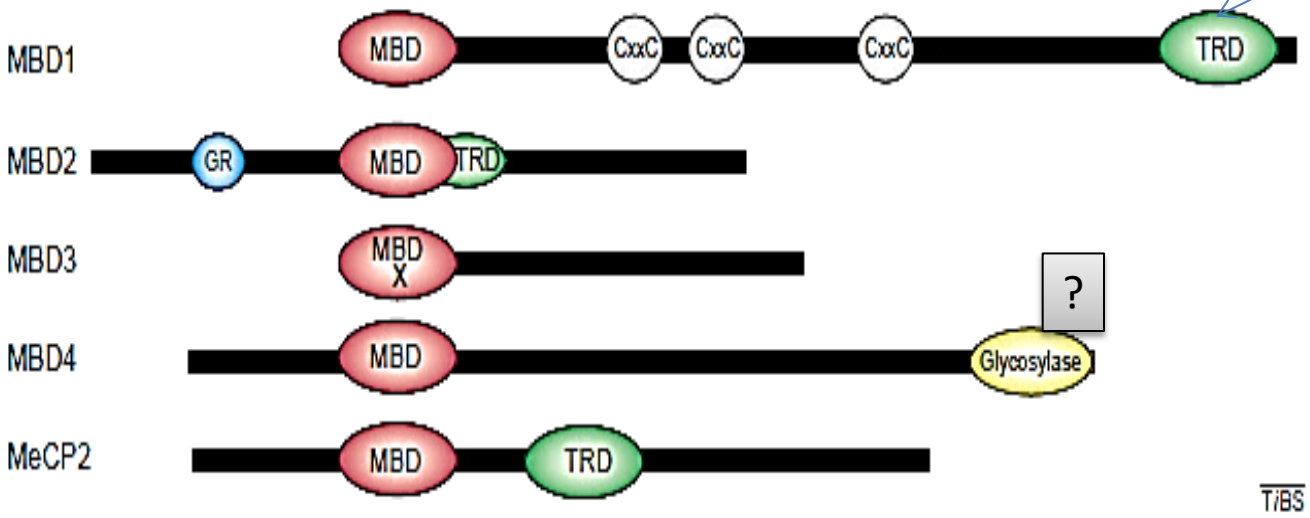
Figure 2. Mechanisms of DNA-methylation-mediated repression.
(**b**) Methyl-CpG-binding proteins (MBPs) directly recognize methylated DNA and recruit **co-repressor molecules** to silence transcription and to modify surrounding chromatin

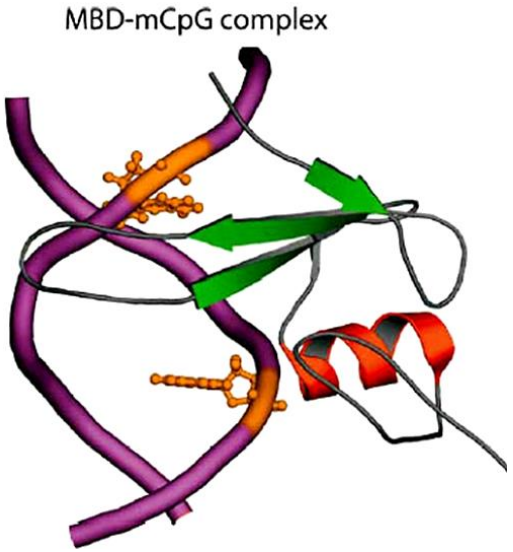
From Klose & Bird, 2006

Readers

Methyl DNA binding proteins have common and specific domains

How do we define a «transcriptional repressor domain ?»





A family of **methyl-CpG-binding proteins** (MBPs). Six mammalian MBPs have been characterized so far:

1. MBD1 uses its methyl-binding domain (MBD) to bind methylated DNA sequences. In addition, MBD1 contains three zinc-binding domains (CxxC), one of which binds specifically to non-methylated CpG dinucleotides, and a C-terminal transcriptional repression domain (TRD).
2. MBD2 possesses an MBD that overlaps with its TRD domain, and a GR repeat at its N terminus.
3. MBD3 contains a well-conserved MBD domain that does not recognize methylated DNA owing to crucial amino acid changes.
4. MBD4 binds methylated DNA through an MBD domain and has a C-terminal glycosylase domain that is important for is function in DNA repair.
5. MeCP2 is the founding member of the MBD protein family and contains a conserved MBD domain and an adjacent TRD domain.
6. The sixth protein, Kaiso, is unrelated and is not shown here.

From Klose & Bird, 2006, TiBS 31: 89-97.

CpG - Cytosine methylation

is a dynamic mark

Enzymes that catalyze methylation are SAM-dependent enzymes:

DNMT1 – maintenance enzyme
DNMT3a -
DNMT3b - } *de novo* methyl-transferases

DNMT3L is a similar gene that encodes a nonenzymatic protein, that has cofactor and regulator roles. It binds to unmethylated H3K4 .
DNMT3L also interacts with HDAC1.

A fourth homologue DNMT2 has little activity toward DNA, higher with RNA.

CpG methylation

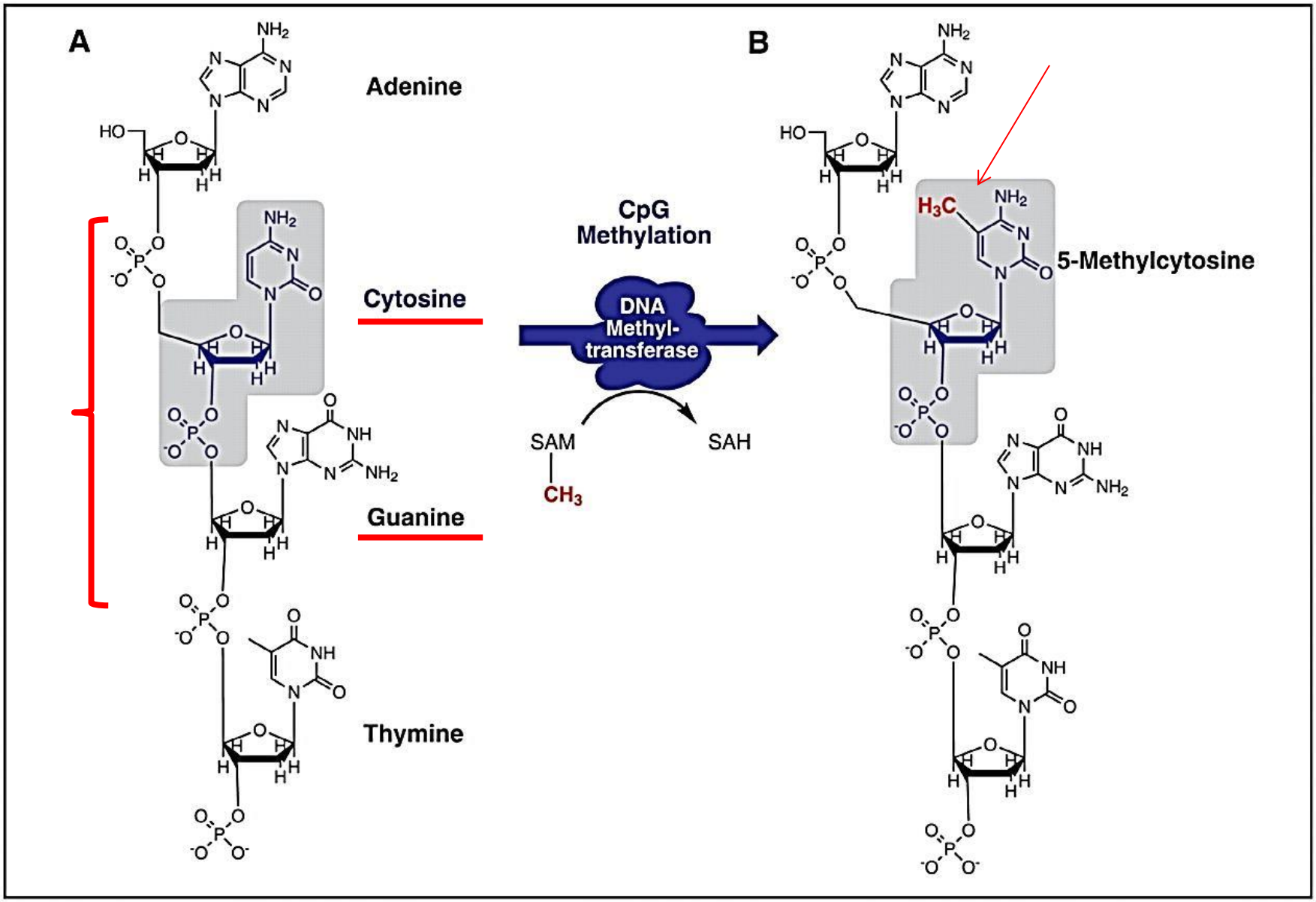
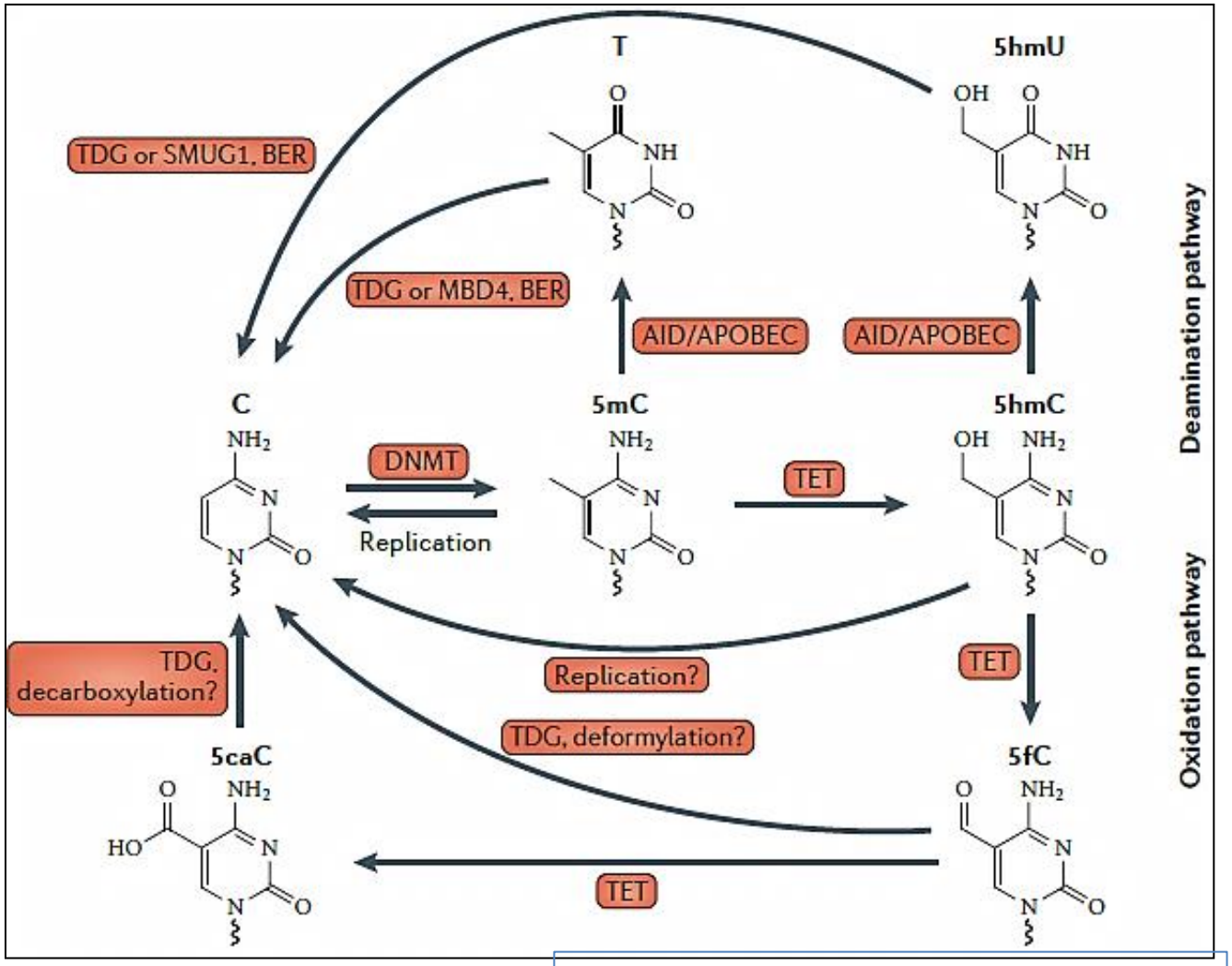


Figure 1 | **Potential pathways for DNA demethylation.**

1. Passive or «dilution»

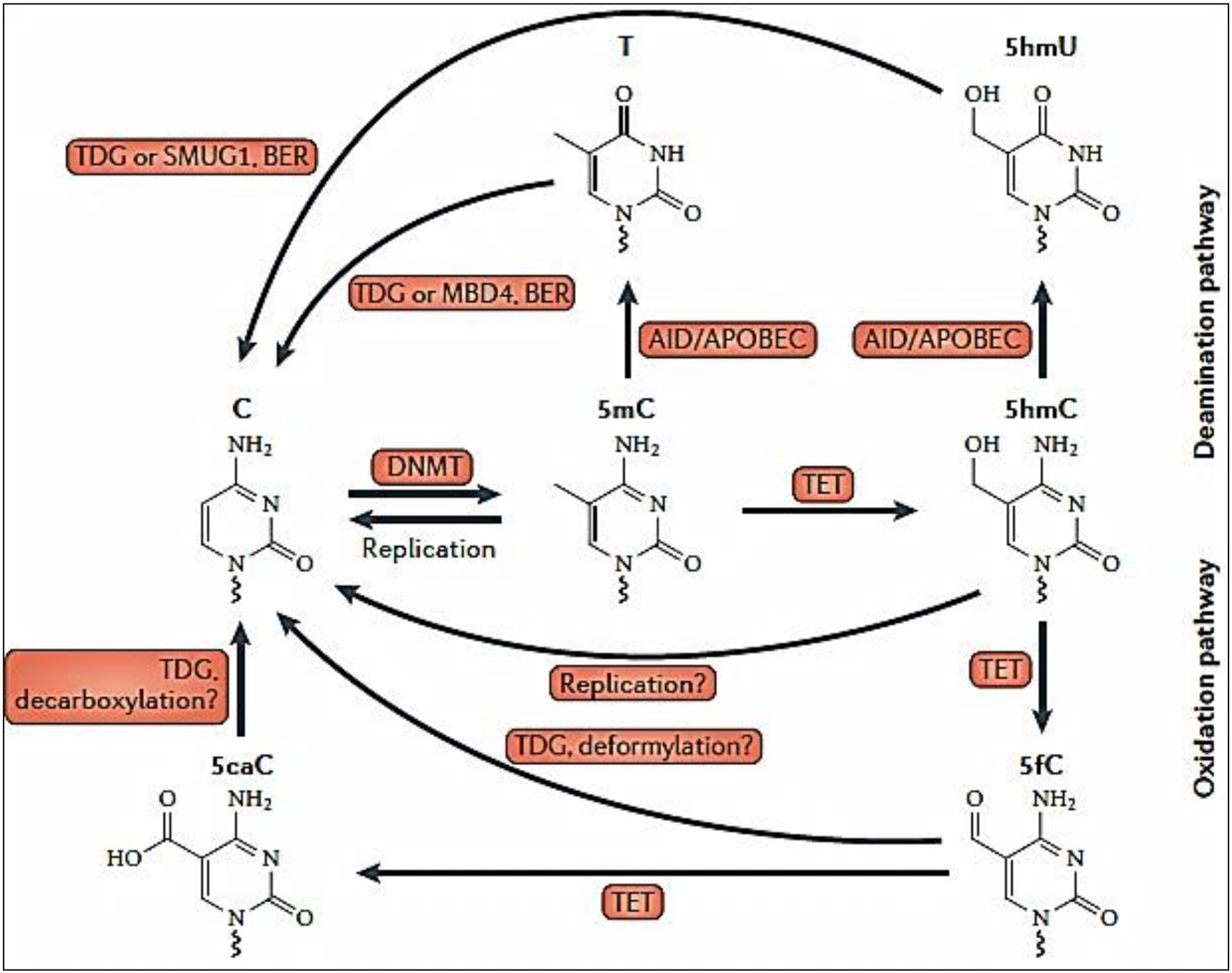
2. **Active** with 5-hydroxymethyl-cytosine as an intermediate:
2.1 Deamination of either 5mC or 5hmC by cytidine deaminase APOBEC produces mismatches that are recognized by DNA glycosylases, producing an abasic site that is then repaired by the base excision repair (BER) machinery.
2.2 5hmC is further oxidized to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by TET enzymes. Decarboxylation of 5caC may convert these intermediates directly back to cytosine.



Branco et al., Nat Rev Genet. (2012) 13: 7-13.

2.3 Thymine DNA glycosylase (TDG) has been shown to cleave 5fC and 5caC, again implicating the BER pathway in DNA demethylation.

5hmU, 5-hydroxymethyluracil; DNMT, DNA methyltransferase; MBD4, methyl-CpG-binding domain protein 4; SMUG1, strand-selective monofunctional uracil-DNA glycosylase 1.


T
5hmU
TDG or SMUG1, BER
TDG or MBD4, BER
AID/APOBEC
AID/APOBEC
Deamination pathway
C
5mC
5hmC
DNMT
Replication
TET
TDG, decarboxylation?
Replication?
TET
TDG, deformylation?
5caC
5fC
TET
Oxidation pathway

How to measure CpG methylation

1. Methylation-sensitive restriction enzymes
    1. HpaII*/MspI (CCGG)
    2. SmaI*/XmaI (CCCGGG)
    3. McrBC – recognize 2 methylated (G/A)pC (50-1000bp apart)

2. Bisulfite conversion (Methyl-C is not modified , C is converted to U)
    1. Enzymatic analysis (e.g. Cobra-next slide)
    2. Cloning alleles and Sanger sequencing

Most common methods to detect CpC methylation are based on **bisulfite**, which converts C (but not methyl-C) to Uridine. Uridine in DNA is then replaced by T in the following PCR. Conservation of C (in the case of methyl-C) or change in T are subsequently detected (simplest) by restriction site analysis, as in the **example** that follows:

Single CpG analysis

COBRA - *Co*mbined *B*isulfite *R*estriction *A*nalysis

Bisulfite-PCR ⇒ **Quantitation**

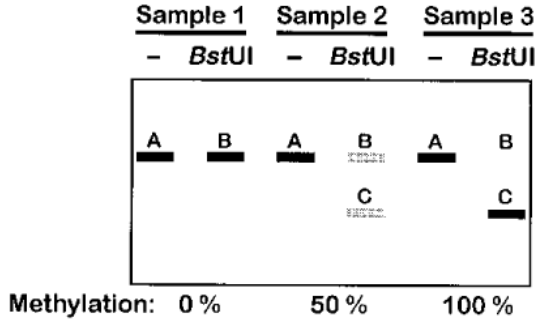
$C \rightarrow U \rightarrow T$

$^{5m}C \rightarrow {}^{5m}C \rightarrow C$

Restriction Digestion
PAGE Gel
Oligo Hybridization
Phosphorimager Quantitation

Example:

Methylated *Bst* UI Site	Unmethylated *Bst* UI Site
$^{5m}CG\,{}^{5m}CG$	CGCG
↓	↓
CGCG	TGTG
Bst UI Site Retained	*Bst* UI Site Lost



$$\% \text{ Methylation} = 100 \times \left(\frac{C}{B + C}\right)$$

Is it possible to study DNA CpG methylation genome-wide ?

Approach I

1) DNA immunoprecipitation using an Ab against 5-methyl-cytosine

2) DNA immunoprecipitation using tagged-MBD

followed by hybridization to microarrays or NGS

Approach II

Bisulfite conversion of the whole genomic DNA,

followed by NGS (WGBS)

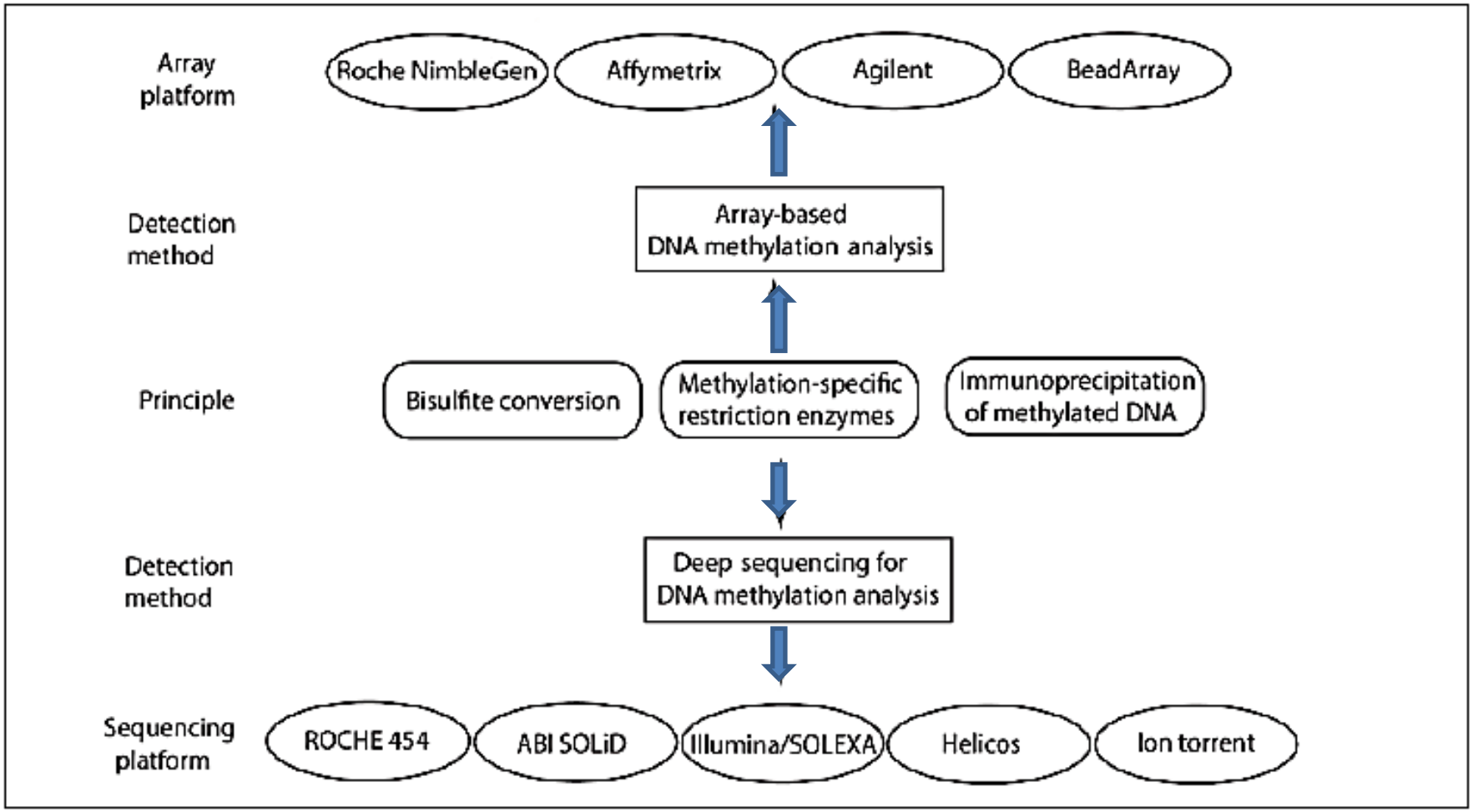


Figure 1. Methods for genome-wide DNA methylation analysis. Bisulfite-converted DNA, DNA methylation–specific immunoprecipitation methods, and DNA partitions generated by methylation-specific restriction enzymes can be used for both array-based and high-throughput deep sequencing–based genome-wide DNA methylation analysis. There are several different choices for both array platforms and sequencing technologies that can be used for DNA methylation analysis.

Gupta et al., 2010, Biotechniques 49: iii-xi

On microarrays...

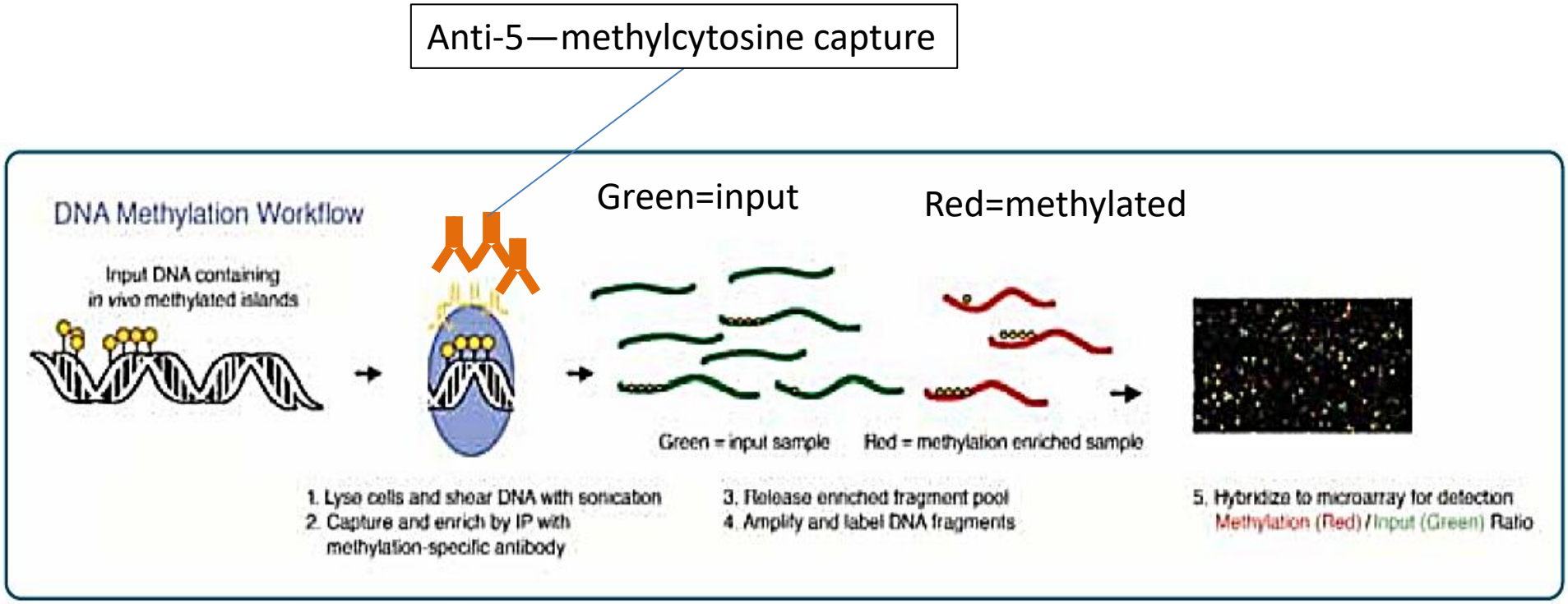


Figure 1. Affinity-based isolation of methylated DNA. A representative example of one method for isolation and enrichment of methylated DNA. Methylated regions of the genome (mDIP) from a genomic DNA sample are isolated with a monoclonal antibody to 5-methylcytosine. Isolated DNA is purified, Cyanine 5-labeled, and competitively hybridized against similarly Cyanine 3-labeled "input" genomic DNA onto a single microarray. Arrays are washed, scanned, and analyzed with Feature Extraction software. Relative DNA methylation levels for each probe/CpG Island are reflected in changes in Cyanine 5/Cyanine 3 ratios.

From the AGILENT website

nature genetics

2007 – the first methylome paper was made by 5-me-C IMPT followed by hybridization on microarray containing several probes for each gene promoter known at that time.

Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome

Michael Weber[1], Ines Hellmann[2,3], Michael B Stadler[1], Liliana Ramos[4], Svante Pääbo[2], Michael Rebhan[1] & Dirk Schübeler[1]

To gain insight into the function of DNA methylation at *cis*-regulatory regions and its impact on gene expression, we measured methylation, RNA polymerase occupancy and histone modifications at 16,000 promoters in primary human somatic and germline cells. We find CpG-poor promoters hypermethylated in somatic cells, which does not preclude their activity. This methylation is present in male gametes and results in evolutionary loss of CpG dinucleotides, as measured by divergence between humans and primates. In contrast, strong CpG island promoters are mostly unmethylated, even when inactive. Weak CpG island promoters are distinct, as they are preferential targets for *de novo* methylation in somatic cells. Notably, most germline-specific genes are methylated in somatic cells, suggesting additional functional selection. These results show that promoter sequence and gene function are major predictors of promoter methylation states. Moreover, we observe that inactive unmethylated CpG island promoters show elevated levels of dimethylation of Lys4 of histone H3, suggesting that this chromatin mark may protect DNA from methylation.

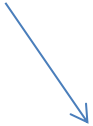
Bisulfite

Whole genome

DNA purification from samples

fragmentation

Bisulfite conversion



Adapter ligation library

NGS sequencing

Whole genome methylation by bisulfite modification is **very difficult** !

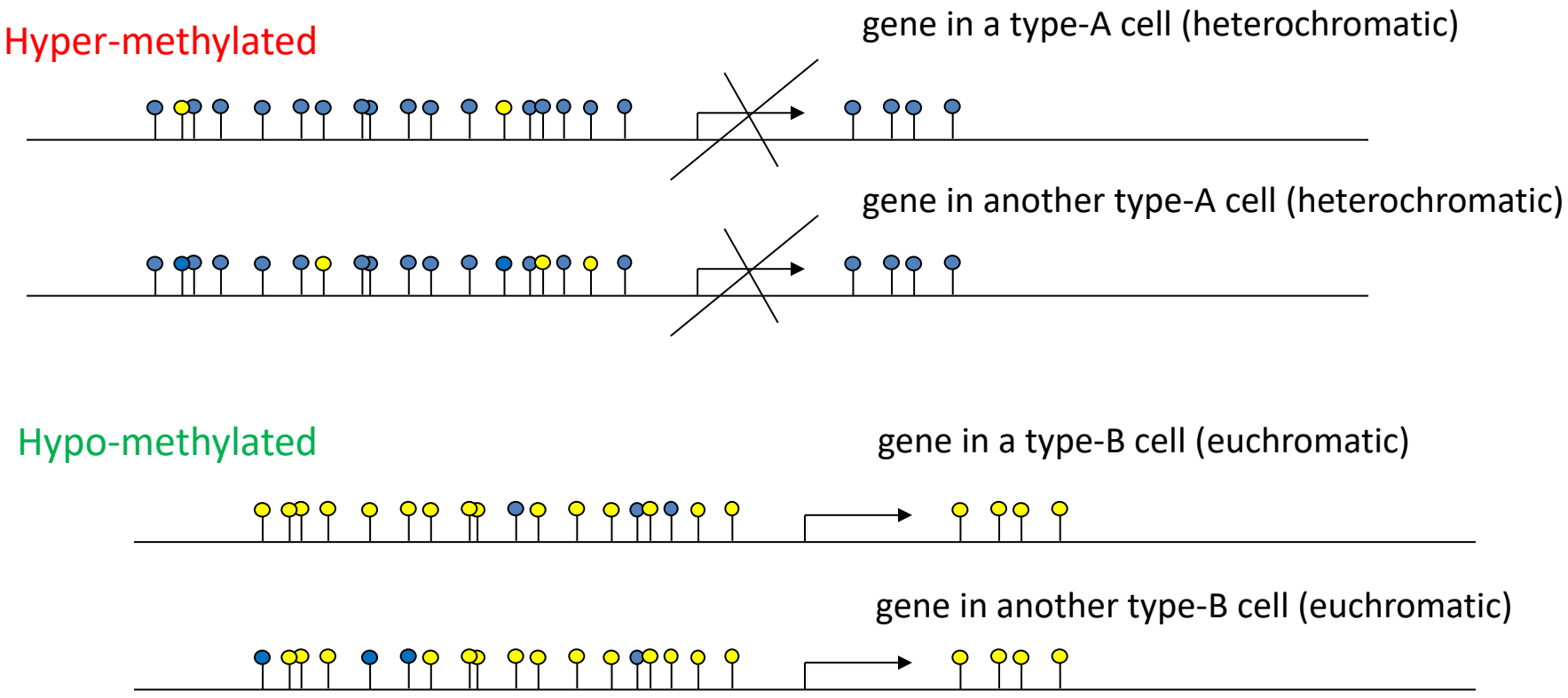
....remember that CpG methylation is stochastic, every cell has its own profile

Very **redundant** sequencing is needed

Aligning database should contain <u>all possible variations</u>

Reduced versions are possible:

Warning! – CpG methylation is not site-specific



Since CpG methylation of a given DNA fragment is different from cell to cell, sequencing needs either cloning + Sanger or NGS.

E.g. (Sanger sequencing) :

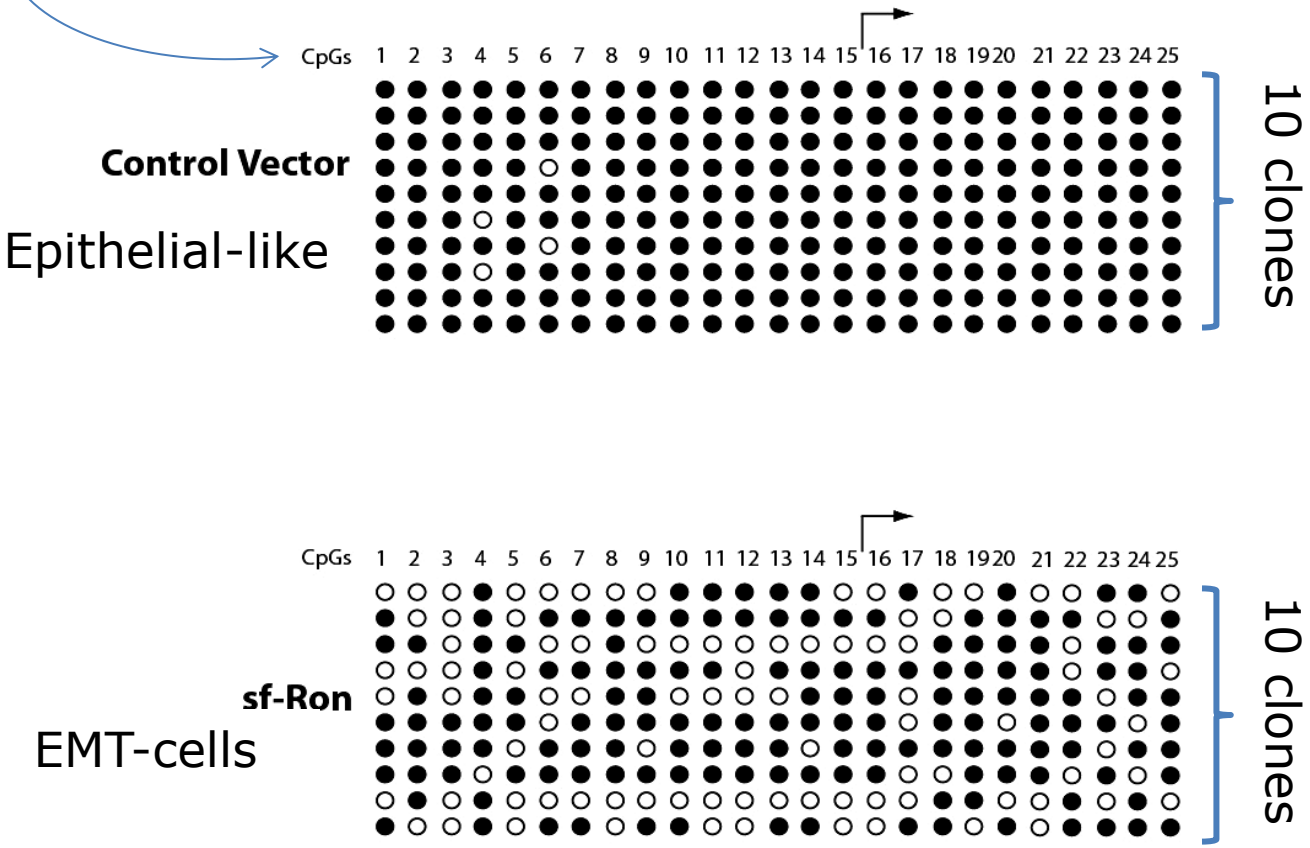
1) extract DNA from cells
2) bisulfite treatment
3) PCR the fragment using “side” primers
4) **clone** individual fragment in a plasmid vector
5) sequence a representative number of clones (if Sanger)

E-Cadherin gene promoter

Example

TSS

AGACCCTAGCAACTCCAGGCTAGAGGGTCACCGCGTCTATGCGAGGCCGGGTGGGCGGGCCGTCAGCTCCGCCCTG
GGGAGGGGTCCGCGCTGCTGATTGGCTGTGGCCGGCAGGTGAACCCTCAGCCAATCAGCGGTACGGGGGGCGGTG
CTCCGGGGCTCACCTGGCTGCAGCCACGCACCCCCTCTCAGTGGCGTCGGAACTGCAAAGCACCTGTGAGCTTGCG
GAAGTCAGTTCAGACTCCAGCCCGCTCCAGCCCGGCCCGACCCGACCGCACCCGGCGCCTGCCCTCGCTCGGCGTC
CCCGGCCAGCCATGGGCCCTTGGA



○ methylated

● unmethylated

From: Cardamone et al., 2009, PNAS 106(18):7420-5

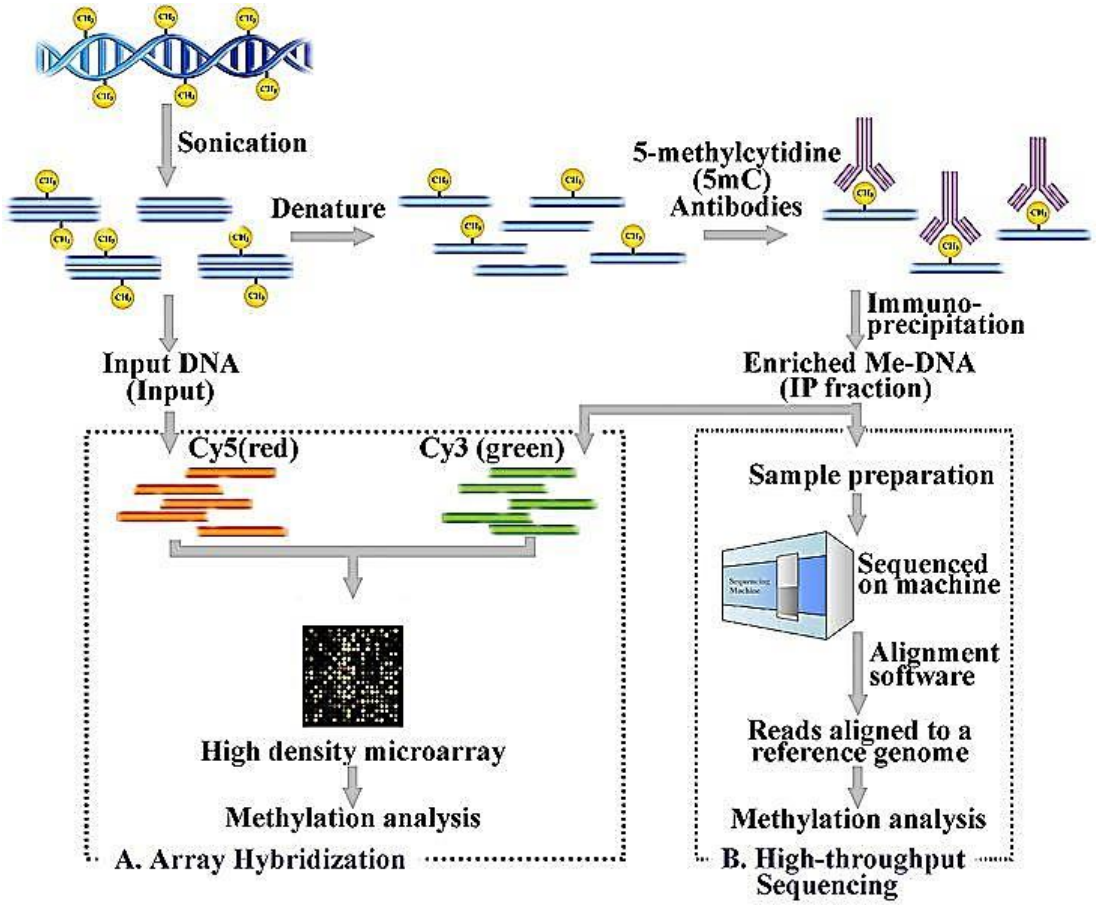
RRBS = reduced representation bisulfite sequencing

1

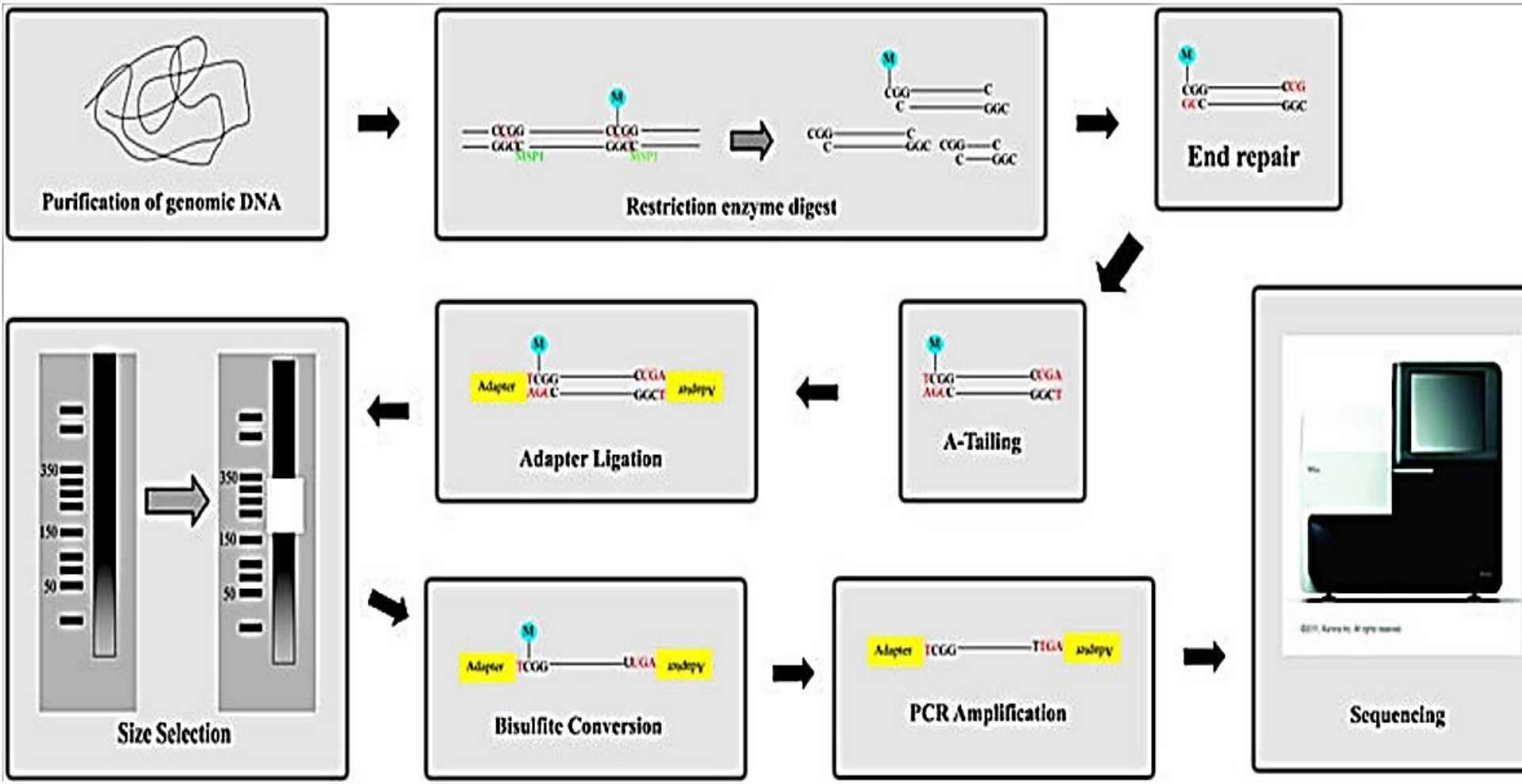
DNA cut with methylation-specific enzymes will enrich a 200-300 bp fraction of hyper-methylated DNA. This fraction is used to produce library → NGS

2

Or affinity-purified (MeDIP) methylated DNA fragments analyzed:



RRBS scheme


Purification of genomic DNA
Restriction enzyme digest
End repair
A-Tailing
Adapter Ligation
Size Selection
Bisulfite Conversion
PCR Amplification
Sequencing

Whole genome analysis

Vol 462 | 19 November 2009 | doi:10.1038/nature08514 nature

Human DNA methylomes at base resolution show widespread epigenomic differences

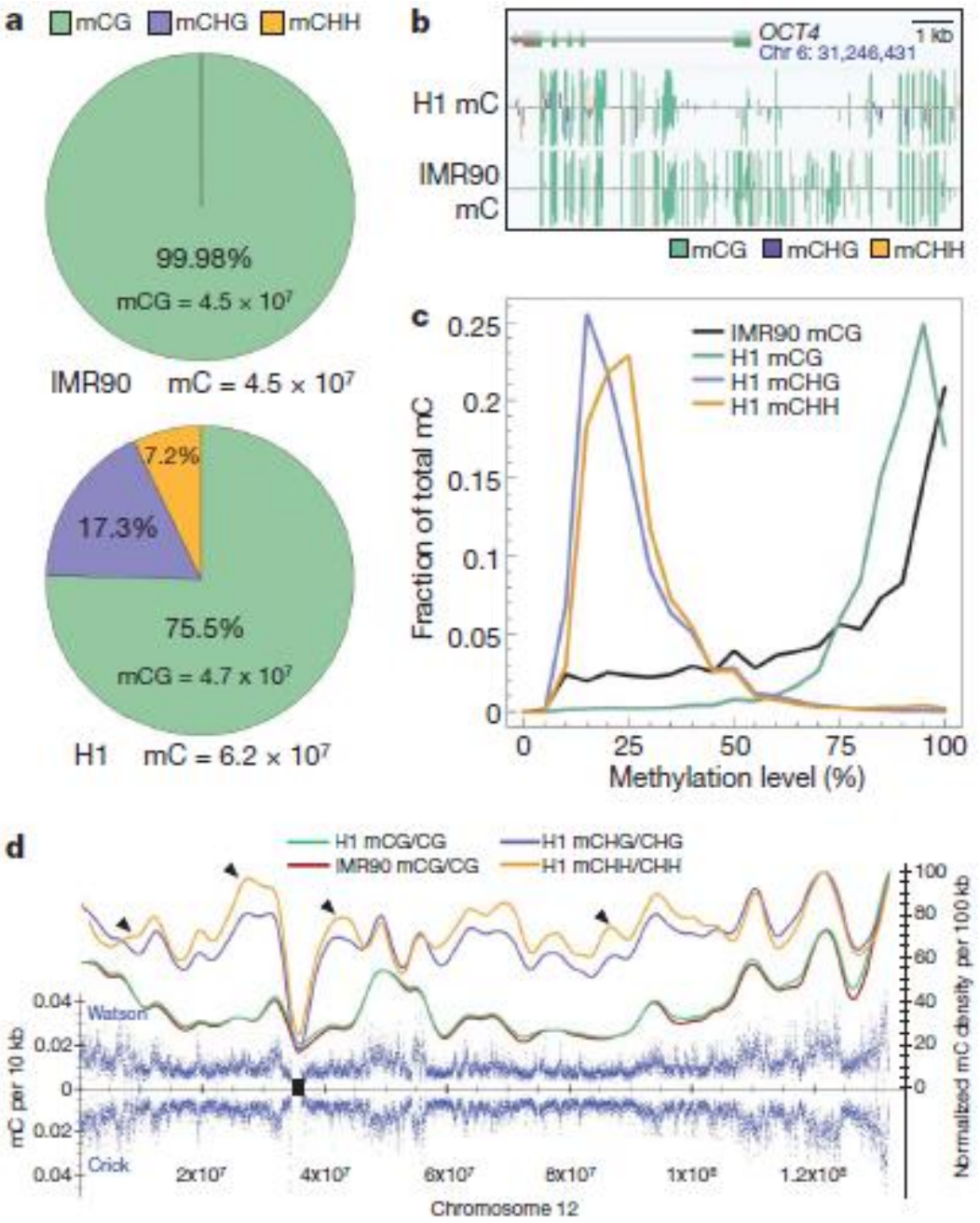
Ryan Lister[1]*, Mattia Pelizzola[1]*, Robert H. Dowen[1], R. David Hawkins[2], Gary Hon[2], Julian Tonti-Filippini[4], Joseph R. Nery[1], Leonard Lee[2], Zhen Ye[2], Que-Minh Ngo[2], Lee Edsall[2], Jessica Antosiewicz-Bourget[5,6], Ron Stewart[5,6], Victor Ruotti[5,6], A. Harvey Millar[4], James A. Thomson[5,6,7,8], Bing Ren[2,3] & Joseph R. Ecker[1]

DNA cytosine methylation is a central epigenetic modification that has essential roles in cellular processes including genome regulation, development and disease. Here we present the first genome-wide, single-base-resolution maps of methylated cytosines in a mammalian genome, from both human embryonic stem cells and fetal fibroblasts, along with comparative analysis of messenger RNA and small RNA components of the transcriptome, several histone modifications, and sites of DNA–protein interaction for several key regulatory factors. Widespread differences were identified in the composition and patterning of cytosine methylation between the two genomes. Nearly one-quarter of all methylation identified in embryonic stem cells was in a non-CG context, suggesting that embryonic stem cells may use different methylation mechanisms to affect gene regulation. Methylation in non-CG contexts showed enrichment in gene bodies and depletion in protein binding sites and enhancers. Non-CG methylation disappeared upon induced differentiation of the embryonic stem cells, and was restored in induced pluripotent stem cells. We identified hundreds of differentially methylated regions proximal to genes involved in pluripotency and differentiation, and widespread reduced methylation levels in fibroblasts associated with lower transcriptional activity. These reference epigenomes provide a foundation for future studies exploring this key epigenetic modification in human disease and development.

pdf

H1 are human embryonic **stem** cells

IMR90 are fetal lung fibroblasts



IMR90 are fetal lung fibroblasts

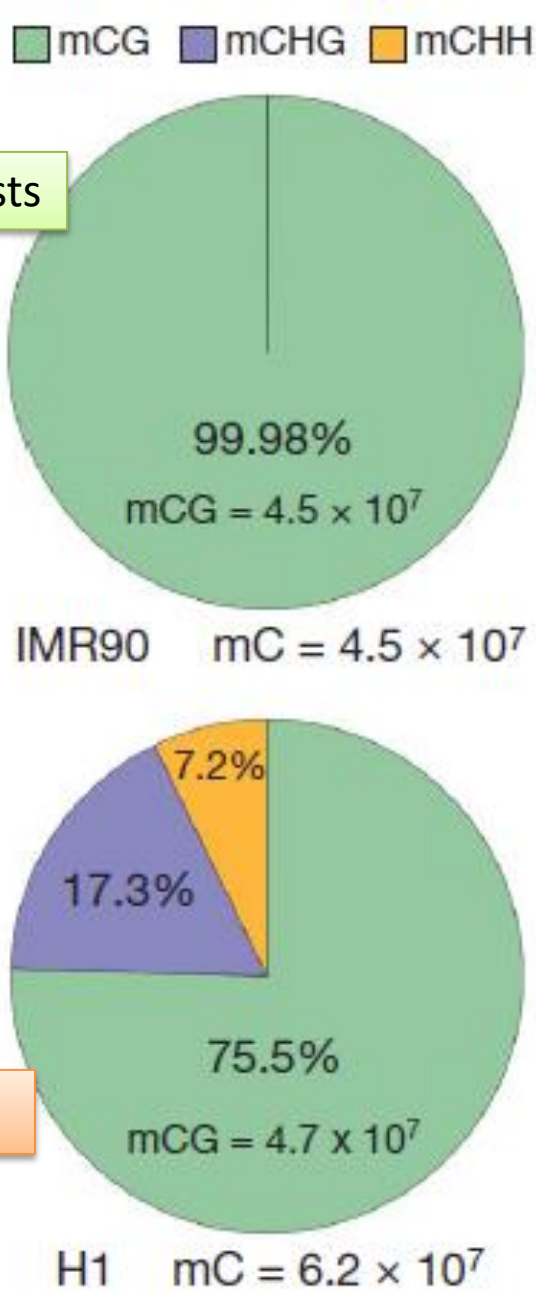
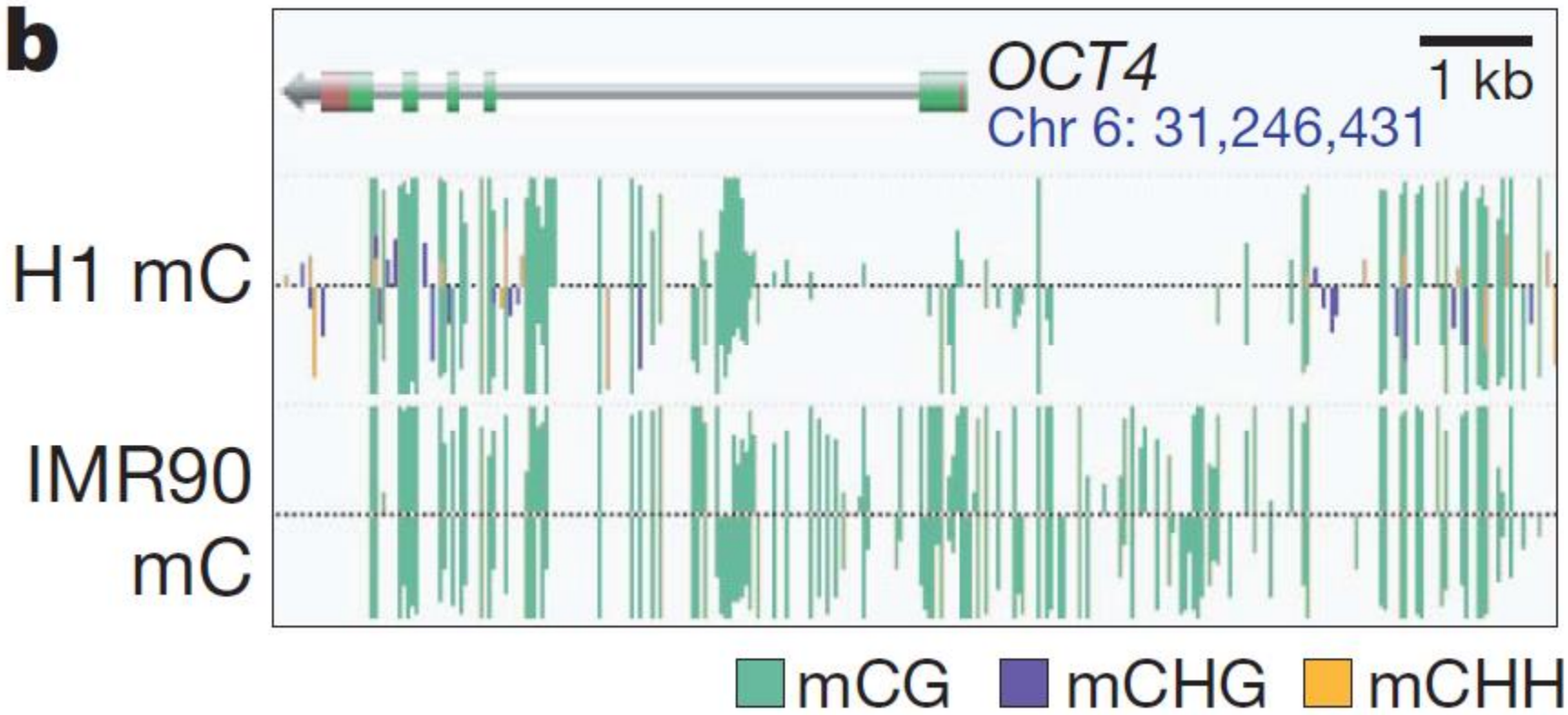
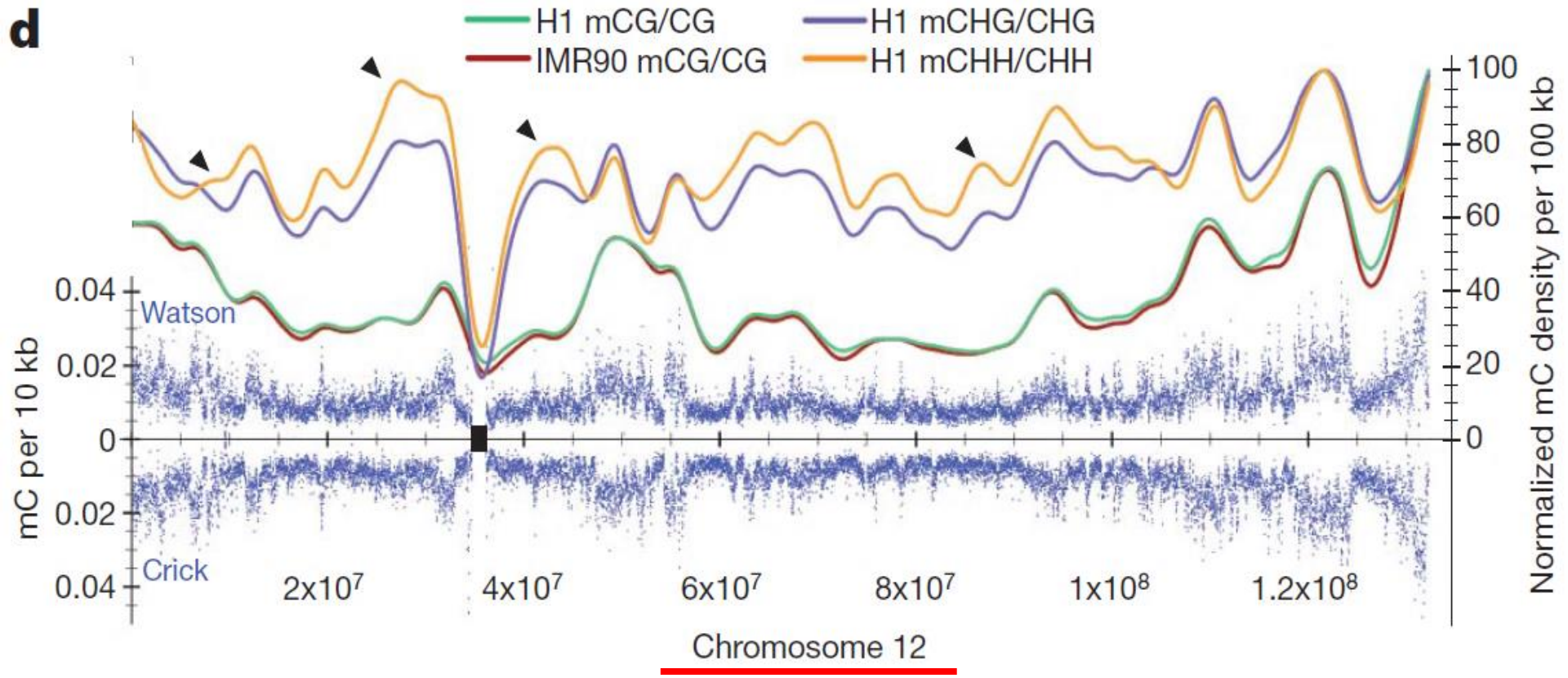


Figure 1 | Global trends of human DNA methylomes.
a, The percentage of methyl-cytosines identified for H1 and IMR90 cells in each sequence context.

H1 are human embryonic stem cells



b, AnnoJ browser representation of OCT4.



d, Blue dots indicate methyl-cytosine density in H1 cells in 10-kb windows throughout chromosome 12 (black rectangle, centromere). Smoothed lines represent the methyl-cytosine density in each context in H1 and IMR90 cells. Black triangles indicate various regions of contrasting trends in CG and non-CG methylation. mC, methyl-cytosine.

NATURE|Vol 462|19 November 2009

EPIGENOMICS

Methylation matters

Dirk Schübeler

Genome-wide maps of methylated cytosine bases at single-base-pair resolution in human cells reveal distinct differences between cell types. These maps provide a starting point to decode the function of this enigmatic mark.

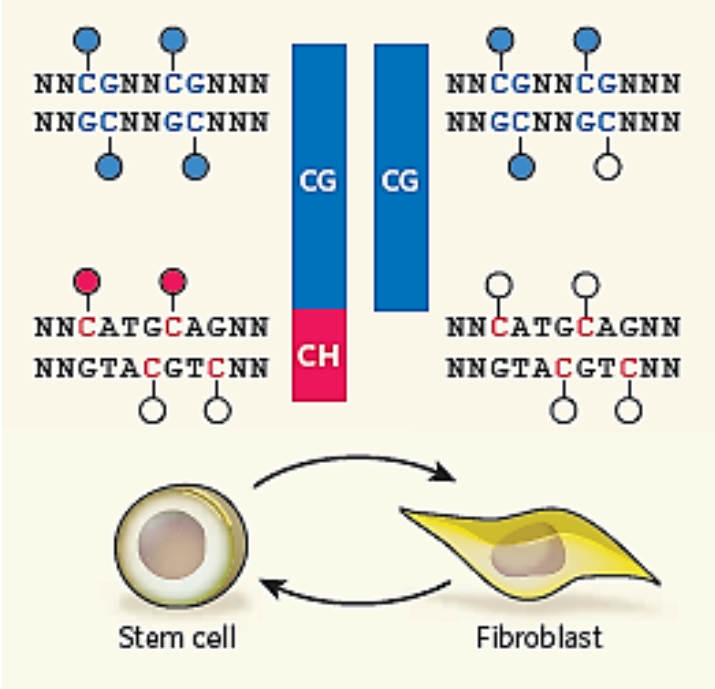


Figure 1 | DNA methylation patterns differ between stem cells and differentiated cells[1]. In stem cells, regions of DNA with CpG methylation (blue) are mostly uniformly methylated, whereas this modification is more heterogeneous in fibroblasts. Non-CpG methylation (red), which occurs primarily at CA nucleotides, is detected only in stem cells, yet is asymmetric and more scarce and patchy than CpG methylation. If fibroblasts are converted to induced pluripotent stem cells they regain non-CpG methylation. Filled circles, methylated cytosines; unfilled circles, unmethylated cytosines. H stands for A, C or T; N stands for any nucleotide.