ISTITUTO ITALIANO DI TECNOLOGIA
CENTER FOR GENOMIC SCIENCE

Campus IFOM—IEO

methylPipe: a library for the analysis of base-resolution DNA methylation data

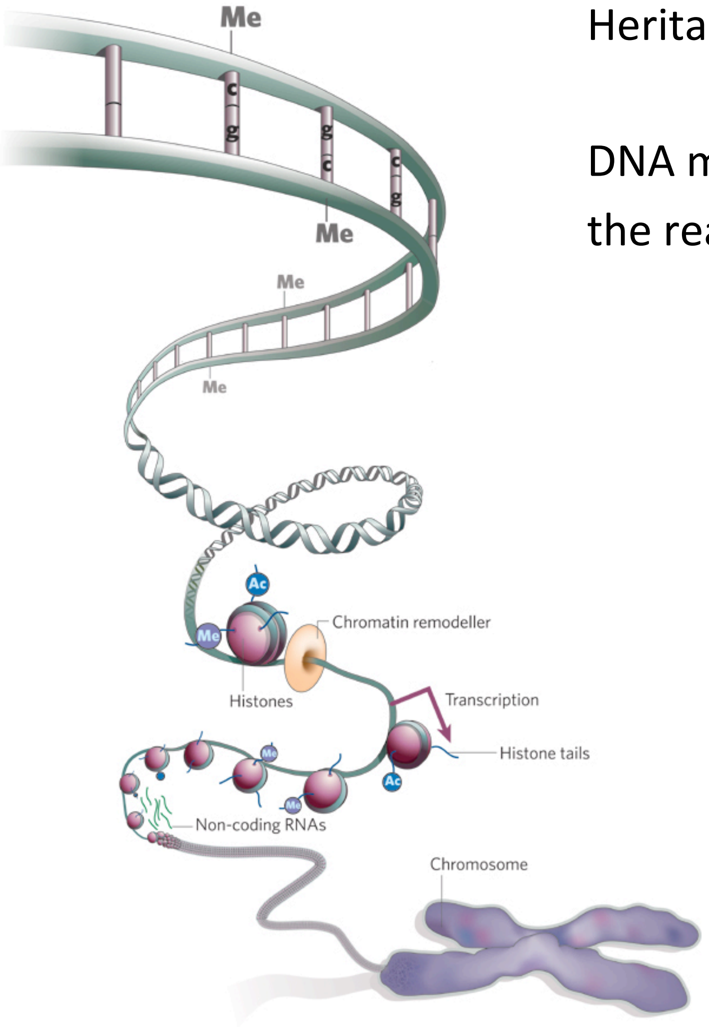
Bioconductor European Developers' Workshop 2012
University of Zurich

Mattia Pelizzola - Center for Genomic Science of IIT@SEMM

Outline of the presentation

- Background
- methylPipe overview
- Defined classes
- Profiling DNA methylation in a set of genomic regions
- Data visualization
- Identification of differentially methylated regions
- Work in progress

Eukaryotic epigenetics and DNA methylation



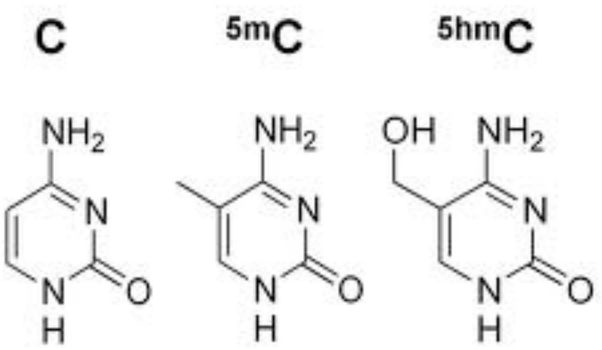
Nature 454, 711-715 (Aug 2008)

Heritable layer(s) of regulation superimposed on genome

DNA methylation and histone modifications can manipulate the readout of the underlying genetic information.

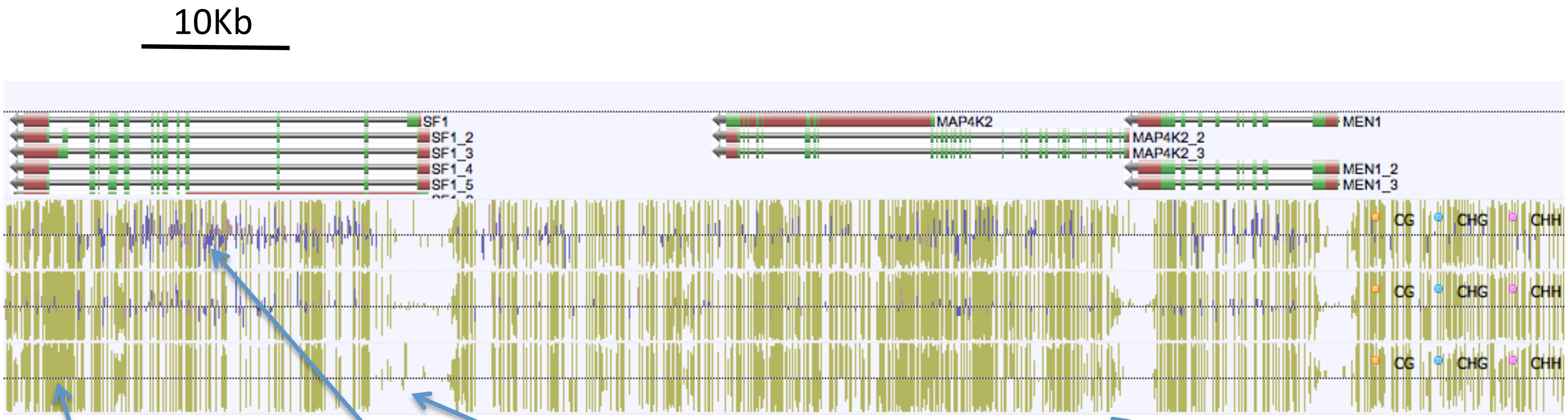
- Cell differentiation
- Tissue-specific gene regulation
- responsive to environment / diet
- varying with age
- Tumorigenesis
- Transposon silencing
- Modulation of binding of protein to DNA

DNA methylation



- Only C in specific sequence contexts (CG, CHG, CHH) can be methylated
- Strand specific
- Heterogeneous in cell populations
- Dynamic

DNA methylation: how the data look like



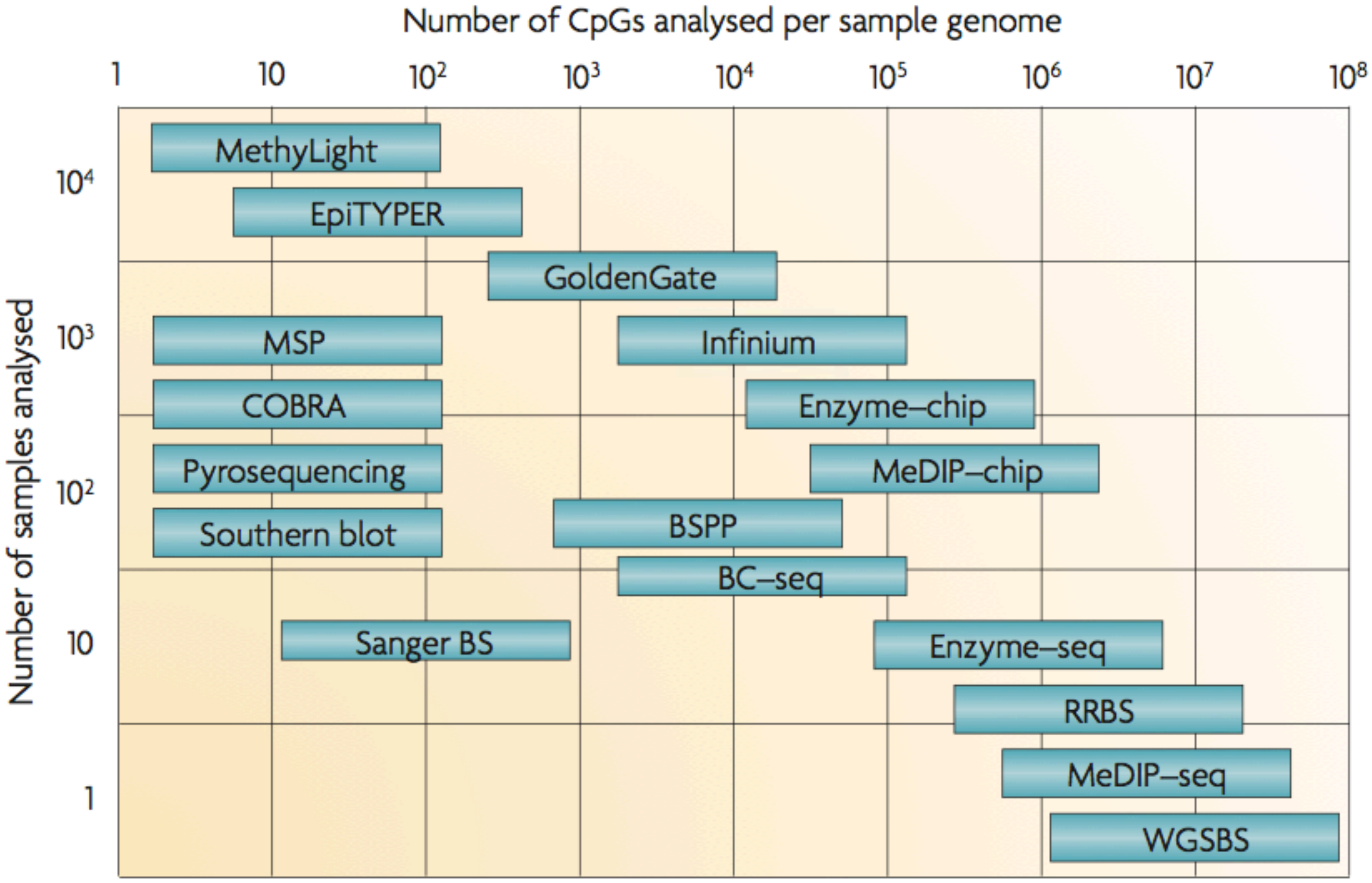
mC in the CG
Sequence context,
~4e7 mCpG in human

Hypomethylated promoter regions

mC in the nonCG
Sequence context (CHG, CHH)
~1e7 mC in pluripotent human cells
Almost absent in differentiated cells

Lister R et al, Nature 2009
http://neomorph.salk.edu/human_methylome/browser.html

Sample throughput versus genome coverage



Laird PW, Nature Review Genetics 2010

methylPipe overview

methylPipe is an R library that will soon be submitted to Bioconductor. The main functionalities cover:

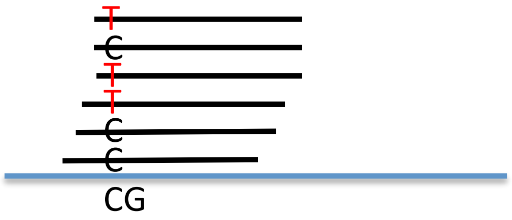
- **Storing and retrieving** low- and high-resolution genome-wide DNA methylation data for multiple samples
- Methods for **visualizing** DNA methylation profiles
- Identification of **differentially methylated regions** (pairwise or multi samples analysis, w/wo replicates)
- **Data integration** with other NGS and annotation data

methylPipe Classes

A number of classes are defined in methylPipe: BSdata, BSdataSet, GElist and GEset:

- **BSdata** is a reference class to collect DNA methylation data generated from a high-thoughput sequencing experiment for a given biological sample.
- The **BSdataSet** class allows collecting DNA methylation data for several samples for the same organism.
- The **GElist** class is used to store a collection of genomic regions and has additional components ready to be populated with data relevant to their DNA methylation status.
- Many GElist objects can be collected in an object of class **GEset**.

BSdata and BSdataSet classes



```
> library(methylPipe)
> library(BSgenome.Hsapiens.UCSC.hg18)

> BSprepare(files, fileout, tabixPath)
```

```
                       #C  #T
chr20  8179   +  CG    2   4
chr20  8180   -  CG    4   4
chr20  8426   +  CG    1   0
chr20  8427   -  CG    5   0
chr20  8432   +  CG    1   0
chr20  8433   -  CG    6   0
```

Bisulfite
Convertion rate

```
                       #C  #T   -10*log10P
chr20  8179   +  CG    2   4    20
chr20  8180   -  CG    4   4    48
chr20  8426   +  CG    1   0    14
chr20  8427   -  CG    5   0    84
chr20  8432   +  CG    1   0    14
chr20  8433   -  CG    6   0   102
```

1. Binomial p-value
2. Data compression (whole genome base-res human DNA methylome down to 500Mb)
3. TABIX indexing (fast and memory efficient access to the data, 2Mb index file)

```
> h1data= system.file('extdata', 'h1_chr20_CG_10k.gz', package='methylPipe')
> h1.db=BSdata(file=h1data, org=Hsapiens)
> imr90data= system.file('extdata', 'imr90_chr20_CG_10k.gz', package='methylPipe')
> imr90.db=BSdata(file=imr90data, org=Hsapiens)
> hsa.set= BSdataSet(list=list(h1=h1.db, imr90=imr90.db), org=Hsapiens)
```

GElist and GEset classes

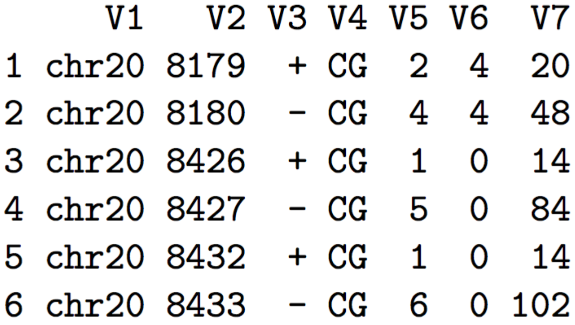
```
> example('GElist-class', 'methylPipe')
      GEist-> gel=GElist(start=c(1,10), end=c(5,12), chr=c('chr1','chr2'))

> Show(gel)
S4 Object of class GElist; 2 features

start :  1 10          ┐
end :  5 12            │ GRanges object
chr :  chr1 chr2       │
strand :  NA NA        ┘
transcript :  NA NA      Association with transcript ids
mClist : NA            ┐ List of mC- or C- positions for each GRange
Clist : NA             ┘
binmC : NA             ┐ Absolute and relative DNA methylation
binC : NA              │ for each bin in each GRange
binrC : NA             ┘
binscore : NA            Score for (each bin in) each GRange
nbins :  5               Number of bins each Grange has to be divided into

> geset=GEset(list=list(gel1=gel1, gel2=gel2))
```

Extracting data from genomes and BSdata objects

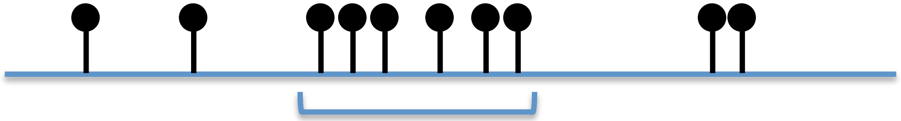
Extracting DNA methylation data for **one genomic region**:

> res= getmCdata(h1.db, chr='chr20', start=1, end=10000)

> head(res)

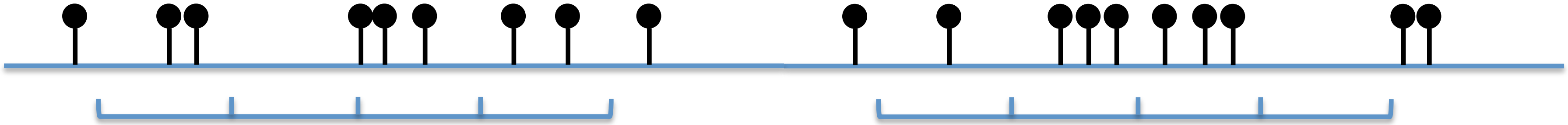


```
     V1   V2 V3 V4 V5 V6  V7
1 chr20 8179  + CG  2  4  20
2 chr20 8180  - CG  4  4  48
3 chr20 8426  + CG  1  0  14
4 chr20 8427  - CG  5  0  84
5 chr20 8432  + CG  1  0  14
6 chr20 8433  - CG  6  0 102
```



Extracting DNA methylation data for **many genomic regions, and every bin of**:

> resmC= MapBSdata2GElistBin(Object= gel, Sample= h1.db, context='CG')

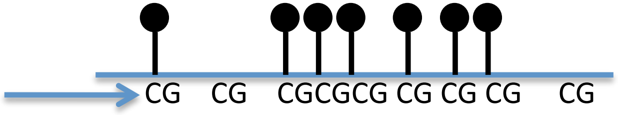


Extracting all **potential methylation sites** on the genome:

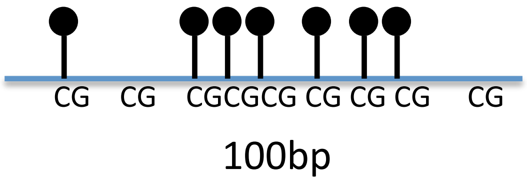
> resC=getCposChr(Object=gel, seqContext='CG', chrseq=unmasked(Hsapiens[['chr20']]))

> resC[[1]][[1]]

[1] 55 56 169 170 651 652 710 711 733 734 746 747

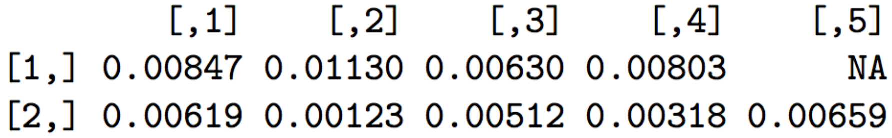


Determining absolute and relative DNA methylation



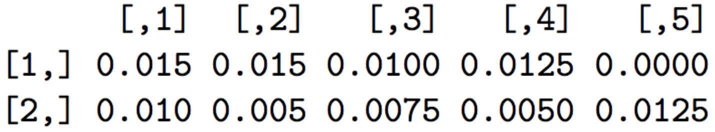
- ✓ Absolute DNA methylation= 0.07 mCG/bp
- ✓ Density of potential methylation sites: 0.09 CG/bp
- ✓ Relative DNA methylation= 100 * 0.07 / 0.09

> gel.h1= **profileDNAmetBin**(Object= gel, Sample=h1.db, mcCLASS='mCG')
> binmC(gel.h1, 'mCG')[1:2,]



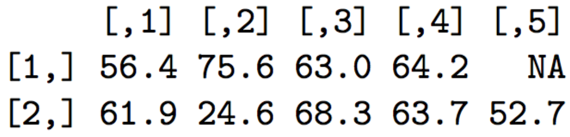
```
        [,1]    [,2]    [,3]    [,4]    [,5]
[1,] 0.00847 0.01130 0.00630 0.00803      NA
[2,] 0.00619 0.00123 0.00512 0.00318 0.00659
```

> binC(gel.h1, 'mCG')[1:2,]



```
      [,1]  [,2]   [,3]   [,4]   [,5]
[1,] 0.015 0.015 0.0100 0.0125 0.0000
[2,] 0.010 0.005 0.0075 0.0050 0.0125
```

> binrC(gel.h1, 'mCG')[1:2,]

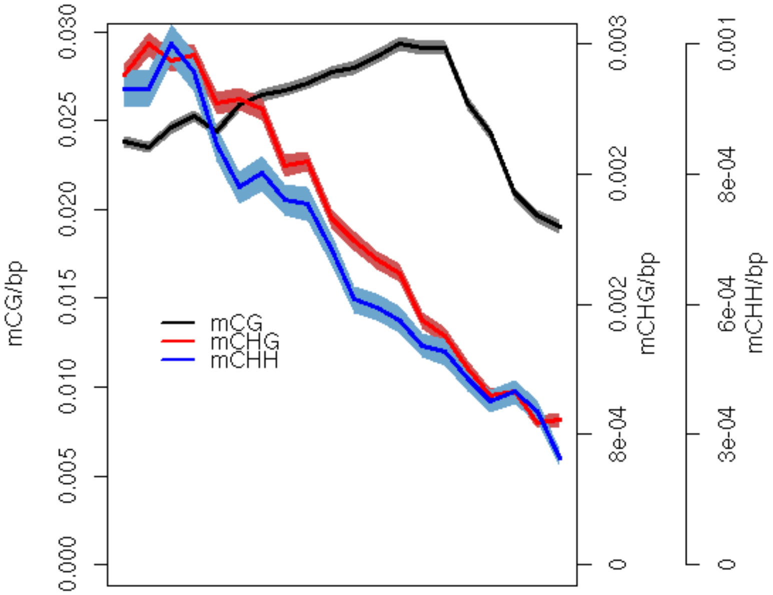


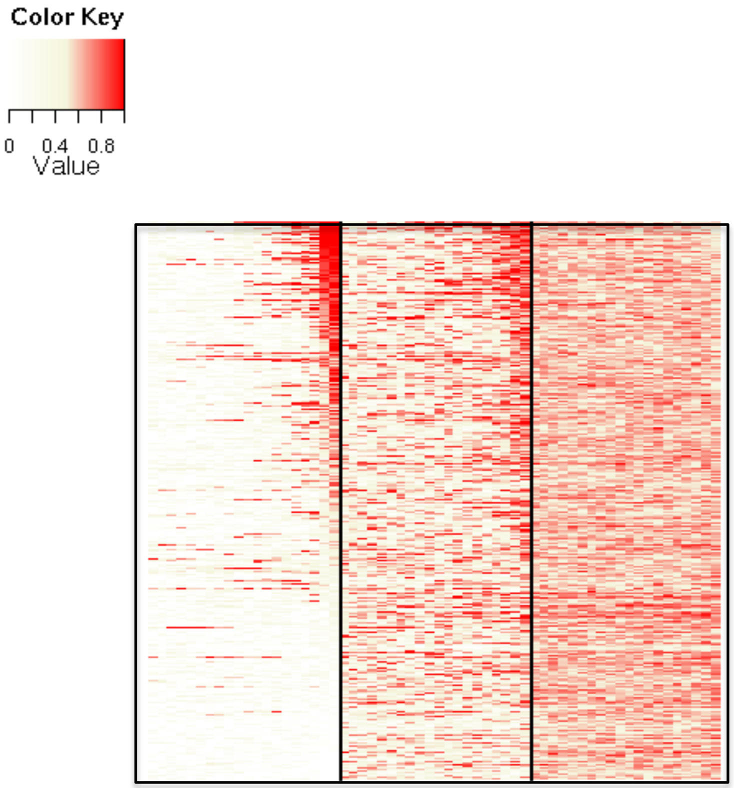
```
     [,1] [,2] [,3] [,4] [,5]
[1,] 56.4 75.6 63.0 64.2   NA
[2,] 61.9 24.6 68.3 63.7 52.7
```

Plotting DNA methylation profiles

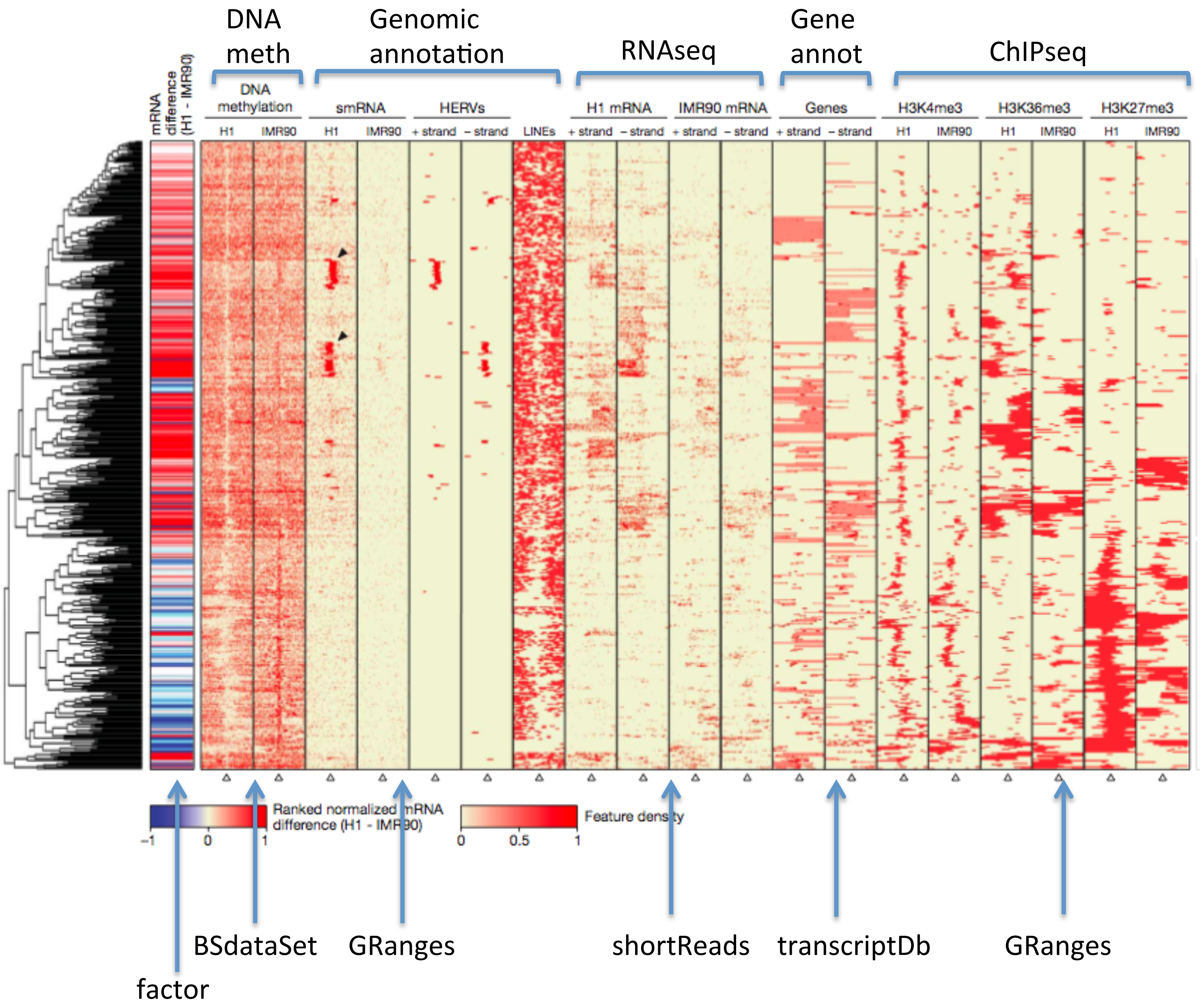
```
> plotME(object=gel.h1, mcClass='mCG', type='rC',  Xlab='', Ylabs='mCG/CG’,
+ leg=FALSE, legX=NULL, legY=NULL, confInt=TRUE, returnData=FALSE)
```

```
> heatmapME(object=gel.h1, mcClass='mCG', SFs=0.90, type='rC', clustRow=TRUE)
```





Data integration (GEsetHeatmap method)



DNA meth

Genomic annotation

RNAseq

Gene annot

ChIPseq

factor

BSdataSet

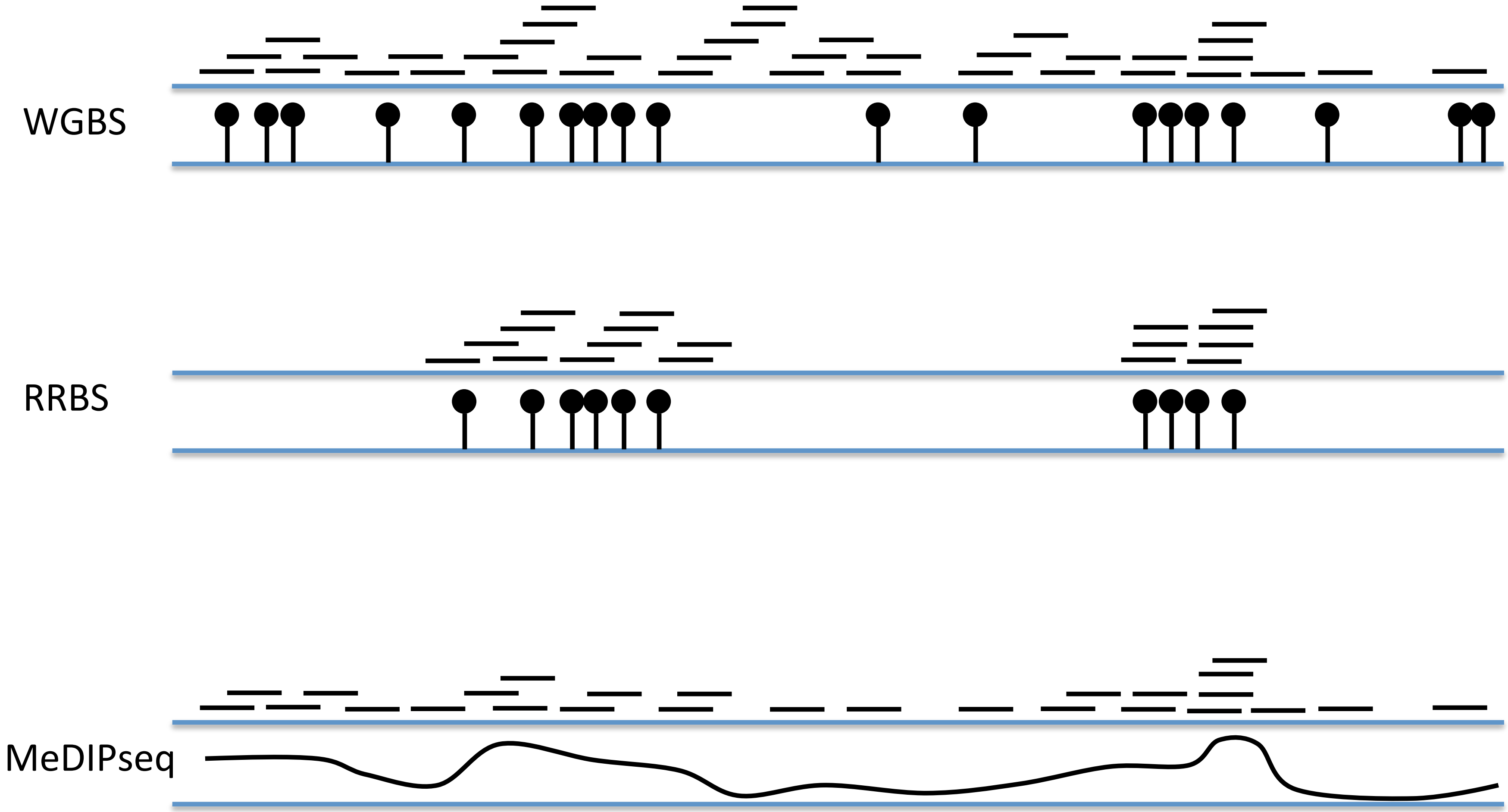
GRanges

shortReads

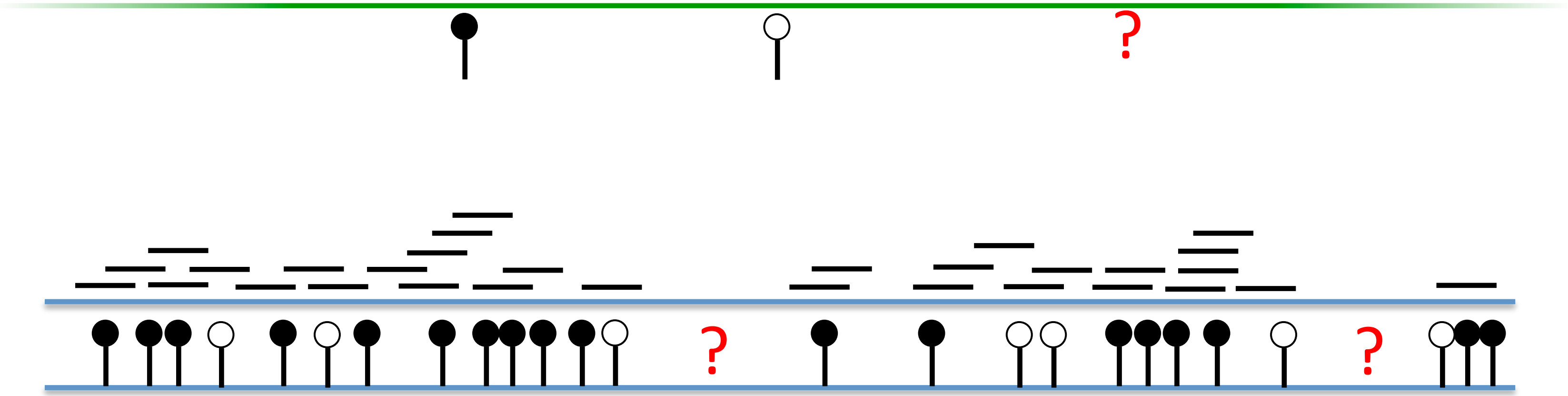
transcriptDb

GRanges

Accomodating data heterogeneous in terms of genome coverage and resolution



Dealing with methylated, unmethylated and uncovered Citoysines

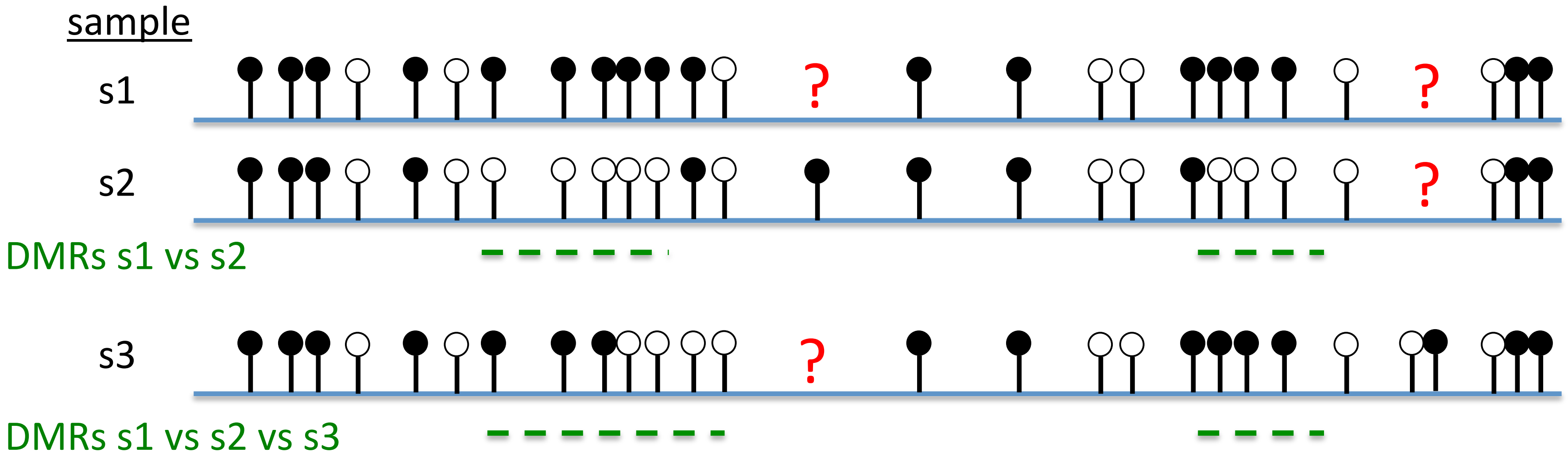


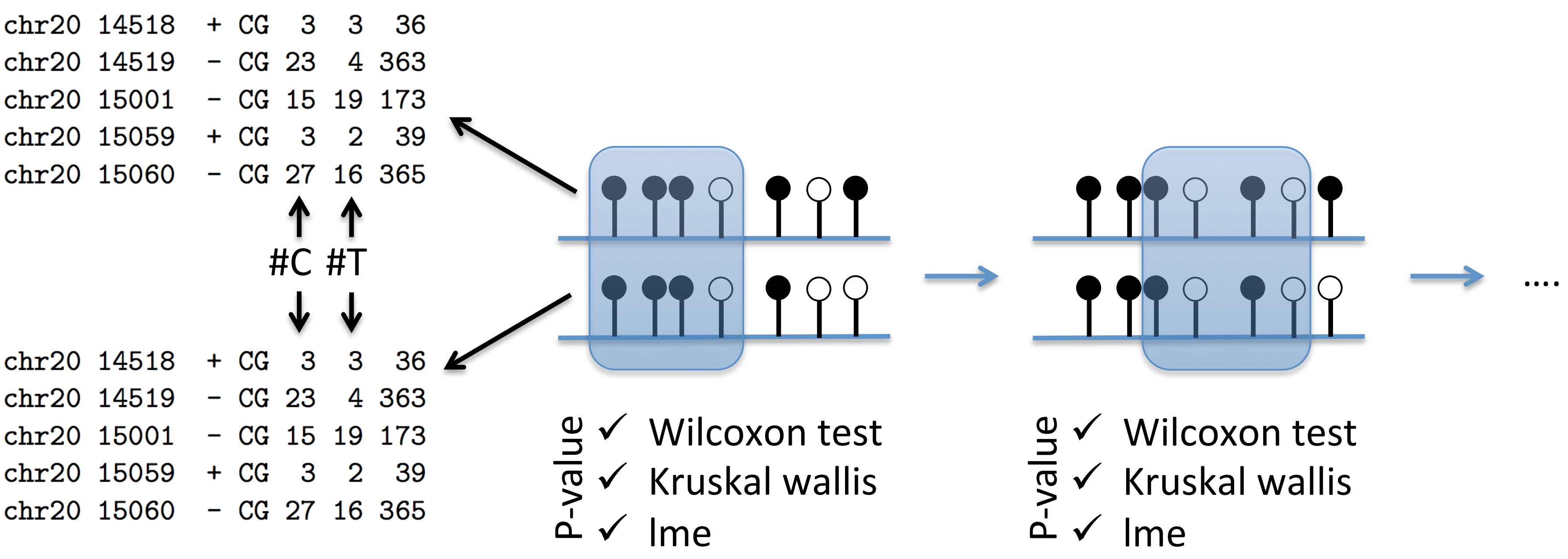
	Stem cells	Differentiated cells
mCG sites	~ 4e7 over 5e7	~ 4e7
mCHG	~ 5e6 over 1e8	~ 0
mCHH	~ 5e6 over 8e8	~ 0

In order to avoid storing too much data while maintaining the ability to identify methylated, unmethylated and uncovered Cytosines, methylPipe does the following:

1. only C positions with at least 1 **mC** read are stored
2. **Uncovered** regions are provided as a GRanges object
3. **Unmethylated** C are determined when profiling region(s) based on 1), 2) and the genome seq

Identification of differentially methylated regions (DMRs)



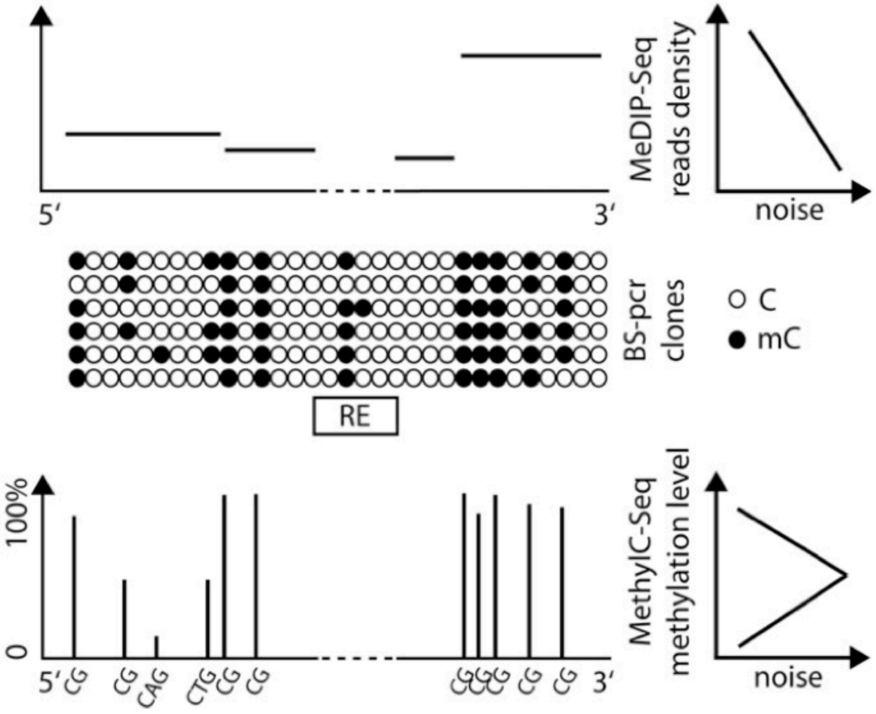


```
chr20 14518  + CG  3  3  36
chr20 14519  - CG 23  4 363
chr20 15001  - CG 15 19 173
chr20 15059  + CG  3  2  39
chr20 15060  - CG 27 16 365
```

```
chr20 14518  + CG  3  3  36
chr20 14519  - CG 23  4 363
chr20 15001  - CG 15 19 173
chr20 15059  + CG  3  2  39
chr20 15060  - CG 27 16 365
```

Work in progress

- Completing compliance to *transcriptDb* and *GRanges* objects
- Dealing with low resolution data (like MeDIP-seq)
- Accomodating 5hmC
- Improving graphic capabilities (*Gviz* and *ggplot2*)
- Implementing lme as method for the identification of DMRs
- Modeling spread vs signal and incorporating for identification of DMRs



Pelizzola M and Ecker JR, FEBS Letters 2010

Acknowledgements

Kamal Kishore (IIT)

Bruno Amati (IEO/IIT)

Ryan Lister (University of Western Australia)

Joseph Ecker (Salk Institute)

SEVENTH FRAMEWORK PROGRAMME

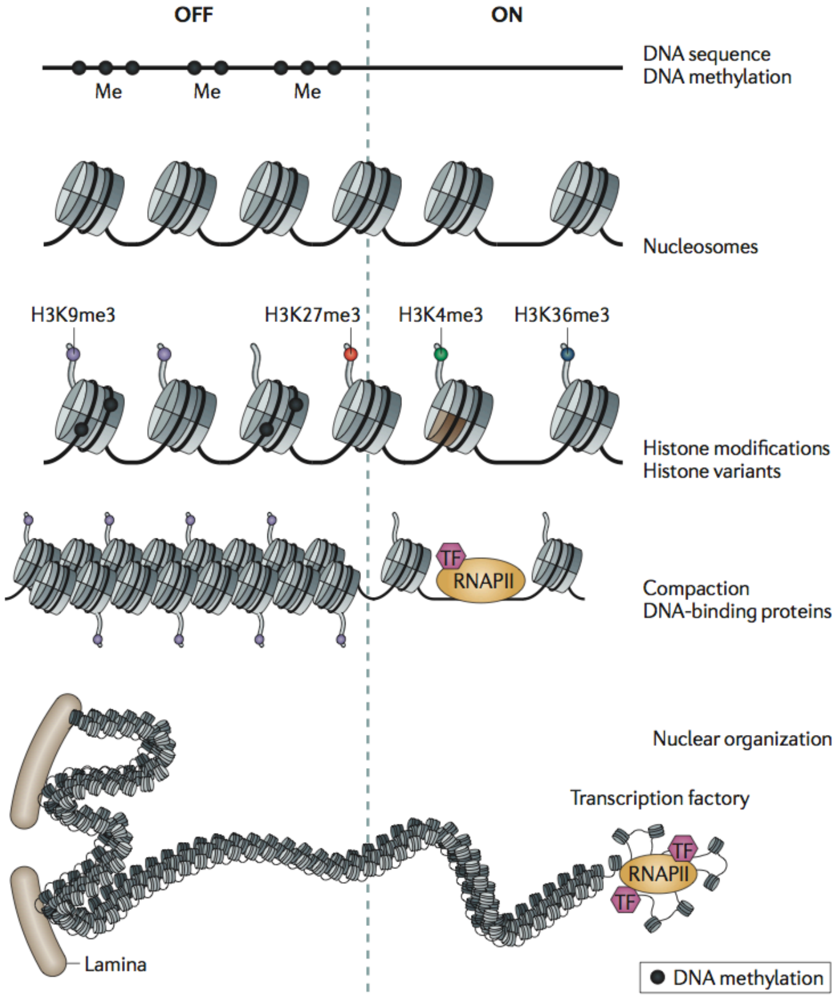
iit @SEMM ISTITUTO ITALIANO DI TECNOLOGIA
CENTER FOR GENOMIC SCIENCE

Campus IFOM—IEO

mattia.pelizzola@iit.it

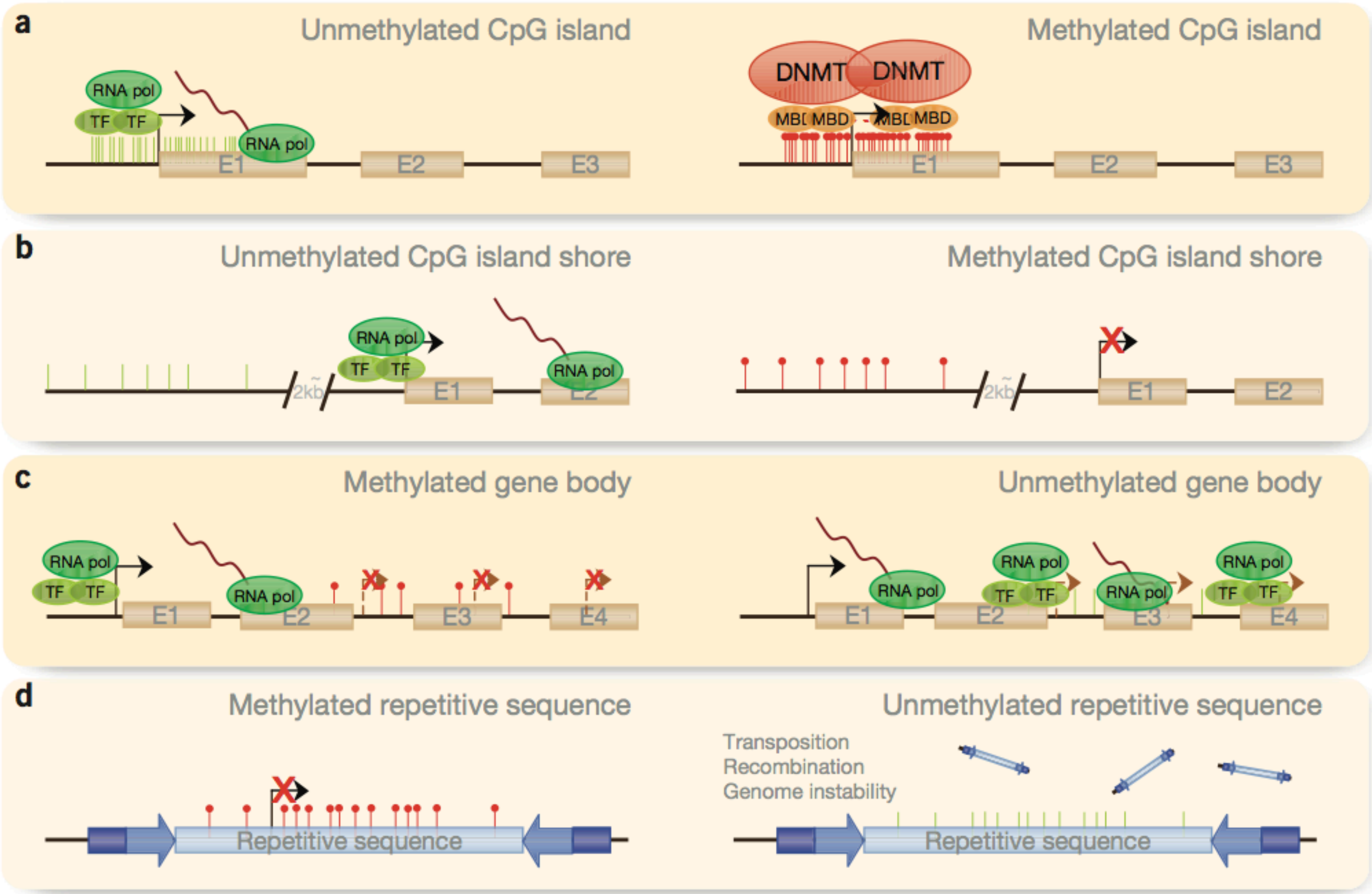
http://tiny.cc/comEpi

Layers of chromatin organization



Zhou VW *et al*, Nature Review Genetics 2010

Relevance of DNA methylation



Portela A and Esteller M, Nature Biotechnology 2010