WGBS data

WGBS (Whole genome bisulfite sequencing) permits to determine the DNA methylation status of single cytosines in the whole genome (in the sense that estimates the ratio of methylated C in a certain position on the genome grouping reads from different cells), in our case we focus only on cytosines belonging to CpG sites.

This is an example of WGBS data of H1 cells downloaded from encode (after dropping some columns):

```
##             chr     Cpos strand reads prop
##        1: chr1    10468      +     0    0
##        2: chr1    10469      -     2  100
##        3: chr1    10470      +     0    0
##        4: chr1    10471      -     2  100
##        5: chr1    10483      +     0    0
##       ---
## 58304908: chrY 56887580      -     3  100
## 58304909: chrY 56887581      +     5    0
## 58304910: chrY 56887582      -     3    0
## 58304911: chrY 56887700      +     2  100
## 58304912: chrY 56887701      -     6   50
```

chr and **Cpos** identifies the position of the C of a CpG site.

strands: indicates which of the 2 strands, "+"" is forward and "-" is reverse.

reads: number of reads for a site.

prop: percentage of methylated reads for a site.

Often I aggregate the reads of the strands for each site, in order to obtain more reads for each site. The assumption is that most of times in each cell the methylation of a single site is the same for both strands.

```
##             chr     Cpos reads  prop
##        1: chr1    10468     2 100.0
##        2: chr1    10470     2 100.0
##        3: chr1    10483     2 100.0
##        4: chr1    10488     2 100.0
##        5: chr1    10492     2 100.0
##       ---
## 29152452: chrY 56887220    10 100.0
## 29152453: chrY 56887399    13 100.0
## 29152454: chrY 56887579     8 100.0
## 29152455: chrY 56887581     8   0.0
## 29152456: chrY 56887700     8  62.5
```

MSR computation details

First of all I aggregate the reads from both strands.

Since we need a binary value for each site in order to calculate the MSR, we have to transform the proportion vector into a binary vector, and this can be done in different ways:

- Use a 50% threshold: assign 0 if prop < 0.5 and 1 otherwise
- Assign according to a threshold such that at the end the proportion of ones is equal to the original methylation rate (calculated as the mean of the prop vector).
- Sample the value from a Bernoulli distribution with p = prop

I usually choose the last method.

One problem is that there are sites with no reads, so these are missing values. We can also choose a minimum number of reads a site must have in order to be not considered a missing value. If the number of missing values is sufficiently small the MSR can be still calculated with small error.

CpG sites distribution in human genome

The frequency of CpG dinucleotides in human genomes is 0.98% (in the sense that about 1 nucleotide over 100 belongs to a CpG site), less than one-quarter of the expected frequency.

CpG islands are regions (~ > 300bp) with a high frequency of CpG sites (~ >3%) (there is not a precise definition). The total number is ~ 28.000 (~ 50.000 if you include repeat sequences).

In general they show significant lower methylation levels with respect to low CpG density regions. Although they have a relatively high CpG density, they contains only about 1-2% of all CpG sites, so their influence on the overall methylation rate is small.

It seems that CpG islands have a functional importance, for example the methylation of CpG islands seems to results in stable silencing of gene expression.

Anyway the methylation of CpG islands is similar between cells from different tissues, it does not show evident tissue-specific patterns.

Cell types

H1: a line of Human Embryonic Stem Cells, they can be propagated indefinitely in vitro and they have the potential to differentiate into a variety of cell lineages.

HeLa: an immortal cell line first derived from cervical cancer cells, used extensively in scientific study since they are remarkably durable and prolific.

K562: myelogenous leukemia cell line (cancer of the white blood cells).

Stomach: tissue cells from stomach

Enhancers

Enhancers are short (50-1500 bp) regions of DNA that can be bound by proteins to increase the likelihood that transcription of a particular gene will occur. They can be located up to 1,000,000 bp away from the gene.

They are located in regions with low density of CpG sites (as usual ~1%), so an enhancer often contains few of them (or none).

A reference on methylation: DNA Methylation and Its Basic Function: https://www.nature.com/articles/npp2012112.pdf