

DNA methylation segmentation of EpiAtlas data

Abdulrahman Salhab, Jörn Walter

Document version: 1.0

Last update 20th, April 2024

505 WGBS samples from data freeze Dec 2022 that are at least 10x coverage were used for segmentation.

Background

DNA methylation segmentation was performed for WGBS samples of EpiAtlas dataset to segment the genome into:

1. Short regions (500bp - 2kb) called low methylated regions (**LMRs**) and Unmethylated regions (**UMRs**) that represent distal and proximal regulatory regions, respectively [[Stadler et al. 2013](#)]. Together, they compose about 3-4% of the genome.
2. Large domains (10s kb - 10s Mb) called partially methylated domains (**PMDs**) and highly methylated regions/domains (**HMRs/HMDs**). PMDs are heterochromatic, gene-poor regions and characterized by highly disordered methylation, while HMRs are euchromatic and gene-rich regions. Together, they compose 96-97% of the genome.

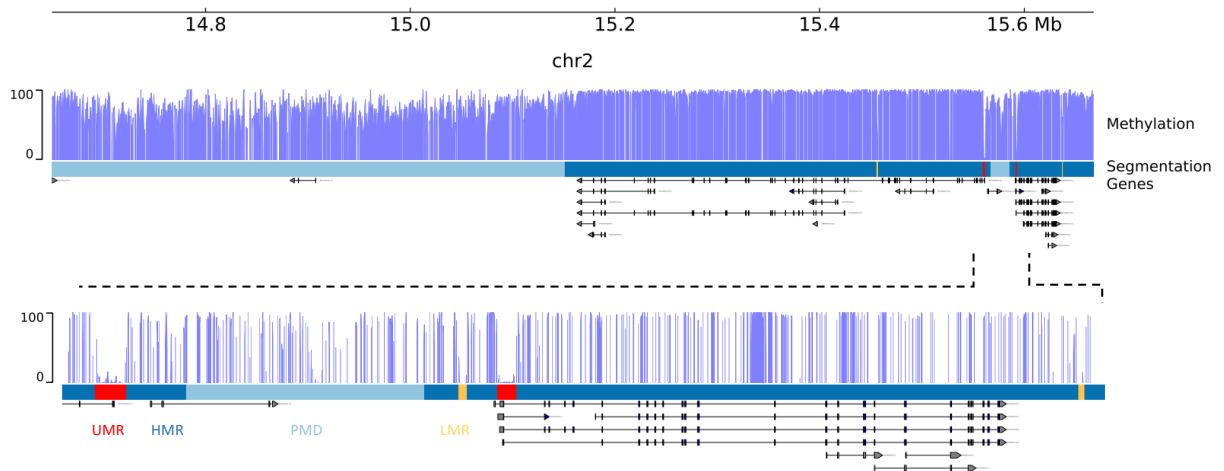


Figure 1. DNA methylation segmentation. Zoomed-in and zoomed-out snapshots of genome-wide segmentation of WGBS data. The segmentation track shows different genomic regions/domains. While PMDs (light brown) and HMDs (red) are observed as domains UMRs (dark blue) and LMRs (light blue) stretch over short regions which coincide with CpG-rich and CpG poor regions, respectively.

Segmentation Method

The segmentation was done individually for each sample using [MethylSeekR](#) tool. The method is based on a 2-state HMM to predict PMDs. The short segments LMRs and UMRs are defined by two parameters: a) methylation level less than 50%, and b) number of CpGs (UMRs should have ≥ 30 CpGs; LMRs should have < 30 CpGs). HMRs are defined as the rest of the genome after excluding gaps (centromeres and telomeres). However, if a UMR/LMR overlap 100% with a PMD, then it will not be reported. If a UMR/LMR partially overlaps a PMD then it will be reported. i.e. a very tiny part of the genome will then have two annotations simultaneously.

As an input for the segmentation pipeline the *cpg.bed.gz files from gemBS output were used. The pipeline compiles MethylSeekR outputs and generate a segmentation track (bed) for visualization in a genome browser.

Output for dissemination

- PREFIX.segment.bed.gz : A segmentation file BED format (can be visualized in a genome browser)

```
track name=IHECRE00003730.1.MethylSeekR description=IHECRE00003730.1.MethylSeekR visibility=1 itemRgb=On
chr1 10000 10583 HMR 55 . 10000 10583 5,113,176
chr1 10583 11864 UMR 42.8571 . 10583 11864 255,0,0
chr1 11864 78692 HMR 86.1817 . 11864 78692 5,113,176
chr1 78692 79796 LMR 15.2174 . 78692 79796 255,195,77
chr1 79796 89622 HMR 81.343 . 79796 89622 5,113,176
chr1 89622 93658 UMR 34.7148 . 89622 93658 255,0,0
chr1 93658 102328 HMR 85.1132 . 93658 102328 5,113,176
chr1 102328 103430 LMR 7.26325 . 102328 103430 255,195,77
chr1 103430 128975 HMR 61.9589 . 103430 128975 5,113,176
```

PREFIX=<epirr_id>.MSRvIHEC.<date>.GRCh38

Column 1: chromosome name

Column 2: start of the segment

Column 3: end of the segment

Column 4: segment type

Column 5: average CpG methylation in the corresponding segment.

Where and how can WGBS-segmentation be used for downstream analyses?

- LMRs are associated with distal regulatory regions (enhancers)
- UMRs are associated with proximal regulatory regions (promoters)
- In combination with ChromHMM LMRs/UMRs can be used to define enhancers /promoters (different classes?)
- PMDs and Het/polycomb/Quiescent ChromHMM states
- Cell type specific LMRs/UMRs/PMDs
- PMDs/HMRs can be used as a proxy for A/B compartments of Hi-C data
- Enrichment analysis – annotations
- Epilogos for visualization across many samples/cell types (Figure 2)

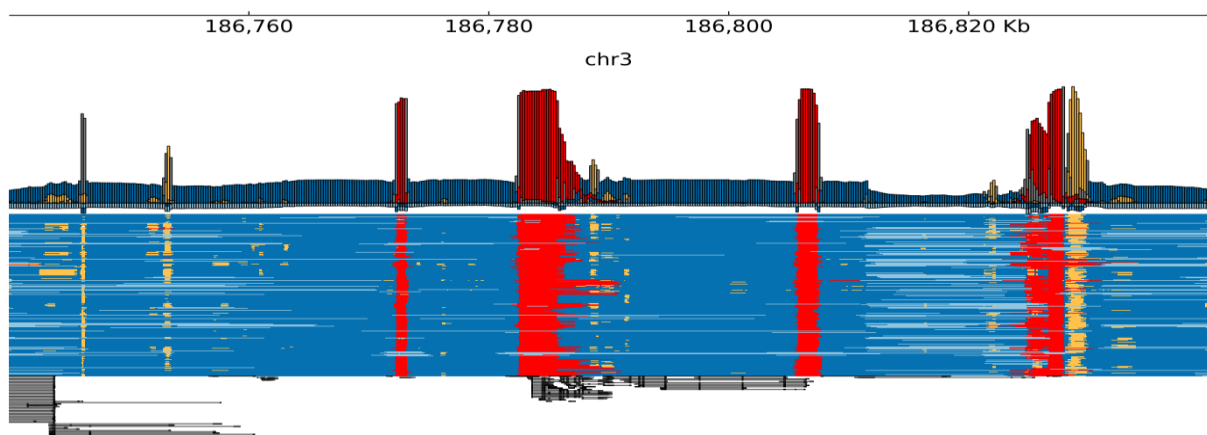


Figure 2. epilogos example generated from WGBS segmentation results of 505 samples

Notes when using UMRs/LMRs in downstream analyses:

- In some cell lines and cancer cells, PMDs lose methylation and become unmethylated (Figure 3). It is known that global demethylation, predominantly in PMDs, is a hallmark of cancer. So, the definition of PMD/UMR can be confusing in this case → use the segmentation with caution in the downstream analyses.

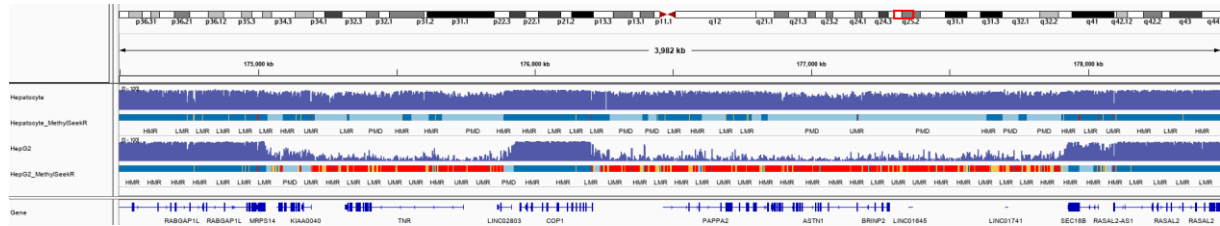


Figure 3. Loss of DNA methylation in PMDs of HepG2 cell line.

Contact

Abdulrahman Salhab abdosalhab@gmail.com

Jörn Walter j.walter@mx.uni-saarland.de