DNA methylation Haplotypes

Jiantao Shi

5 December 2018

Abstract

DNA methylation haplotypes represent methylation status of cytosines along single DNA molecules. Few published tools can extract DNA methylation haplotypes conveniently. Here we present a Java tool that could extract DNA methylation haplotypes from BAM files generated by popular aligners (BSMAP, BISMARK and MAQ) for bisulfite sequences.

Contents

1 Input files

mHaplotype requires indexed BAM files and CpG location files to run.

1.1 BAM files

Sorted and indexed BAM files are standard output of most pipelines for bisulfite sequences. Currently, mHaplotype mainly support BAM files generated by BSMAP, BISMARK and MAQ.

1. BSMAP is one of the fastest aligner for bisulfite sequences. Please use -R option when running BSMAP, which generates tag ZS:Z in resulting BAM file, in which ++ or +- for reads from watson strand and -+ or -- for reads from crick strand. You may refer to BSMAP publication for details.
2. BISMARK is another aligner for bisulfite sequences with rich QC information. mHaplotype check SAM flag of each read. Reads with flag 99 or 147 will be parsed as watson strand, and 83 or 163 as crick strand.
3. MAQ is not designed for bisulfite sequences but has been used by some groups. If the aligner is specificed as MAQ, then strand information is infered from SAM flags. Specificly, if read reverse strand is detected, the read is parsed as crick strand, otherwise watson strand.

1.2 CpG location files

mHaplotype require a folder with CpG location files. They must be named as chr14.txt. Each file contains two columns, second column must be CpG location.

1.3 Interval

mHaplotype process one interval at a time. An interval, in the format of 14:57248000-57293348, is needed in command line.

2 Usage

mHaplotype is designed to capture standard input so that it could work with samtools, which could be used to filter out low quality reads using option -F 3840, such as not primary alignment, read fails platform/vendor quality checks, read is PCR or optical duplicate and supplementary alignment. When running mHaplotype without any option, help will be printed, as shown below.

```
java -Xmx4g -jar mHaplotype.jar
  -T bam2Hap
  -A [BSMAP, MAQ, BISMARK]
  -C CpG position folder
  -i Interval string.
```

```
    -O Output file name.
```

We have included example BAM file in folder `exampleData` and one CpG postion file in folder `CpG/hg19`. An typical command looks like this:

```
samtools view -F 3840 exampleData/GEO_OTX2_Cancer.bam |java -Xmx4g -jar mHaplotype.jar -A BSMAP -T bam2Hap -C
```

3 output files

3.1 Text file

`mHaplotype` output a text file with three columns: Genomic interval, Haplotype, Counts. Genomic interval is defined as the first and last CpG site position for each haplotype. The above command generate output file `GEO_OTX2_Cancer.txt`. The first few lines of this file is listed below.

```
14:57263949-57264186    11111111111 1
14:57279428-57279463    00  38
14:57279428-57279463    01  2
```

3.2 Tabix indexed file

The haplotype file could very large especially for WGBS data. Fortunately, it is a genomic position-based file and could be indexed by Tabix for fastq query. The bash script below convert unsorted haplotype file to indexed haplotype file.

```
cat GEO_OTX2_Cancer.txt | sed 's/[:-]/\t/g' |sort -k1,1 -k2,2n | bgzip > GEO_OTX2_Cancer.gz
tabix -b 2 -e 3 -p bed GEO_OTX2_Cancer.gz
```