Differential Methylation Analysis for Bisulfite Sequencing using DSS

Supplementary Materials

Hao Feng[1], Hao Wu[1]

[1] Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, GA 30322, USA

Corresponding Author:
Hao Wu, Email: hao.wu@emory.edu

1. Quality control of BS-seq data using FastQC

Quality control step can help identify potential contamination or errors during library building or sequencing step. Therefore, a quality control step using software FastQC is recommended. FastQC is java software that is freely available through Babraham Institute's website (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Common sequencing file extensions like fastq or compressed fastq are supported. Below is a quick example for running FastQC in command line.

```
your/fastqc/folder/fastqc sample1.fastq sample2.fastq
```

After running FastQC, a report with multiple modules will be generated. The following command line is useful for users who would like to view the help manual:

```
your/fastqc/folder/fastqc --help
```

Alternatively, FastQC provides graphical user interface (GUI) for users who are unfamiliar with command line tools. The GUI version tutorial is available online (https://www.youtube.com/watch?v=bz93ReOv87Y).

The FastQC report contains useful information for determining how/which sequencing reads will be trimmed or removed. Users can refer to FastQC website for detailed information on interpreting the report.

2. Trimming the BS-seq data

Raw sequencing data (fastq or compressed fastq) may contain adapters or the low-quality bases on sequencing reads. In these situations, trimming can help remove those. Trim Galore find the remove adapter sequences and trim low-quality bases. Below is the example code for trimming paired-end library sequencing files, discarding low quality reads from the 3' with a phred64 scale threshold of 20.

```
your/trim_galore/folder/trim_galore --paired -q 20 -phred64
sample_R1.fq.gz sample_R2.fq.gz
```

The full manual can be found at https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md. There are alternative trimming programs like Trimmomatic or seqtk.

3. BS-seq data mapping and methylation calling

Mapping the BS-seq data is the most important step for detecting the methylation signal on each CpG site. The output can then be analyzed by DSS to determine DML/DMR. Bismark provides such functions of mapping the BS-seq data and calling the methylation signal. Below is the example for using Bismark for mapping. Here, the input is fastq (syntax '-q') and the output

folder is specified (syntax '-o'). The installed Bowtie path is also specified (syntax '—path_to_bowtie').

```
/your/bismark/folder/bismark -q --path_to_bowtie /your/bowtie/path/
/your/bsmark/hg19/Bismark_refGenome/folder/ -o /your/output/folder
/your/bs_seq/fastq/folder/sample1.fastq
```

After this mapping step above, user needs to run the 'bismark_methylation_extractor' function to extract the methylation signal, using on the output mapped SAM/BAM file. Here, the input SAM file is from a single-end read data (syntax '-s'). The output format follows the bedGraph convention (syntax '--bedGraph').

```
/your/bismark/folder/bismark_methylation_extractor -s --bedGraph --counts --
buffer_size 10G /your/output/folder/sample1.sam
```

Syntax and software version could change over time. More detailed user guidelines and up-to-date script/software can be found at Bismark's website (https://www.bioinformatics.babraham.ac.uk/projects/bismark/).