DNA methylation tutorial: Data download and Alignment

by
Roseric Azondekon, PhD
University of Wisconsin Milwaukee

June 10, 2019

Background

In this tutorial, we show you how to download raw Bisulfite-seq DNA methylation sequence data from the European instance of the SRA, which can be accessed via https://www.ebi.ac.uk/ena. At ENA, the sequencing reads are directly available in FASTQ or SRA formats, which will be explained below.

For this tutorial, we need `FastQC`, `multiQC`, the `SRA toolkit`, a powerful suite of tools designed to interact with SAM and BAM files called `samtools`, and the `Bismark` aligner to align the Bisulfite-seq reads to the reference genome. All the above mentioned tools need to be installed and referenced in the environment variable `PATH`. Let's first check if this requirement is met:

```
In [ ]: fastqc --version
```

```
In [ ]: multiqc --version
```

```
In [ ]: fastq-dump --version
```

```
In [ ]: samtools --version
```

```
In [ ]: bismark --version
```

If at least one of the above commands produces an error, please, check your installation of the tool and try again.

Now let's create a working directory for our DNA methylation bisulfite-seq project.

```
In [ ]: mkdir -p tuto && cd tuto
```

1 Data Download

To download a set of SRA files: 1. Go to `https://www.ebi.ac.uk/ena`. 2. Search for the accession number of the project, e.g., SRP041828 (should be indicated in the published paper). 3. There are several ways to start the download, here we show you how to do it through the command line interface on GNU/Linux. - copy the link's address of the "SRA files" column (right mouse click), go to the command line, move to the target directory, type: `wget < link copied from the ENA` `website >` - If there are many samples as it is the case for the project referenced here (accession number: SRP041828), you can download the summary of the sample information from ENA by right-clicking on “TEXT” and copying the link location.

```
In [ ]: wget -O all_samples.txt "https://www.ebi.ac.uk/ena/data/warehouse\
        /filereport?accession=PRJNA246552&result=read_run&fields=study_accession,\
        sample_accession,secondary_sample_accession,experiment_accession,\
        run_accession,tax_id,scientific_name,instrument_model,library_layout,\
        fastq_ftp,fastq_galaxy,submitted_ftp,submitted_galaxy,sra_ftp,sra_galaxy,\
        cram_index_ftp,cram_index_galaxy&download=txt"
```

You may try to open the `all_samples.txt` file with LibreOffice or Excel to view it. For this project, we are only interested in the paired-end first 4 Bisulfite-seq samples (2 normal cells samples vs 2 breast cancer cells samples). Since the first line in `all_samples.txt` contains the header, we will generate another file containing only the first 4 lines of `all_samples.txt` with the following command:

```
In [ ]: sed '1d' all_samples.txt > all_samples2.txt
        head -4 all_samples2.txt > samples.txt
        rm all_samples2.txt
```

Now, let’s create a new folder for our SRA files.

```
In [ ]: mkdir -p sra_files
```

According to `https://www.ncbi.nlm.nih.gov/books/NBK158899/`, the FTP root to download files from NCBI is `ftp://ftp-trace.ncbi.nih.gov/` and the remainder path follow the specific pattern `/sra/sra-instant/reads/ByRun/sra/{SRR|ERR|DRR}/<first 6 characters of accession>/<accession>/<accession>.sra`.

Notice that the accession number for the SRA files are located in the 5th column “Run accession” in `all_samples.txt`. We proceed to the download of the SRA files of the samples listed in `samples.txt` with the following code:
(**Attention: The download may take a long time!**)

```
In [ ]: cut -f5 samples.txt | xargs -i bash -c \
                'v={}; FTPROOT=ftp://ftp-trace.ncbi.nih.gov/; \
                       REM=sra/sra-instant/reads/ByRun/sra/; \
                       url=${FTPROOT}${REM}${v:0:3}/${v:0:6}/${v}/${v}.sra; \
                       wget $url -P sra_files'
```

2 Converting SRA files to FASTQ files

Now that the download is complete, let’s convert the SRA files into FASTQ files with the following command:
(**Attention: This may take a long time!**)

```
In [ ]: cut -f5 samples.txt | xargs -i bash -c \
                'v={}; fastq-dump --outdir fastq/${v} --gzip \
                                  --skip-technical --split-3 sra_files/${v}.sra'
```

3 Quality Control of the FASTQ files

Up to this point, we have all our RNA-seq FASTQ files ready for Quality Control (QC) check. This is done with the `fastqc` tools developed by the Babraham Institute. Run the following command to perform QC check for all the samples: (**This may take some time!**)

```
In [ ]: cut -f5 samples.txt | xargs -i bash -c \
                'v={}; \
                 mkdir -p fastqc_reports/${v}; \
                 fastqc fastq/${v}/*fastq.gz -o fastqc_reports/${v}'
```

Next, let's summarize the QC reports (for all the samples) into one unique report using `multiqc`:

```
In [ ]: multiqc fastqc_reports --dirs -o multiQC_report/
```

Let's examine the summary `multiqc` report either by double-clicking on `multiQC_report/multiqc_report.html` or by executing the following code:

```
In [ ]: xdg-open multiQC_report/multiqc_report.html
```

4 Read Alignment

The assignment of sequencing reads to the most likely locus of origin is called read alignment or mapping and it is a crucial step in most types of high-throughput sequencing experiments.

The general challenge of short read alignment is to map millions of reads accurately and in a reasonable time, despite the presence of sequencing errors, genomic variation and repetitive elements. The different alignment programs employ various strategies that are meant to speed up the process (e.g., by indexing the reference genome) and find a balance between mapping fidelity and error tolerance.

4.1 Reference genome

Genome sequences and annotation are often generated by consortia such as (mod)ENCODE, The Mouse Genome Project, The Berkeley Drosophila Genome Project, and many more. The results of these efforts can either be downloaded from individual websites set up by the respective consortia or from more comprehensive data bases such as the one hosted by the University of California, Santa Cruz (UCSC) or the European genome resource (Ensembl).

Reference sequences are usually stored in plain text FASTA files that can either be compressed with the generic gzip command.

The reference sequences file can be obtained either from NCBI, ENSEMBL or UCSC Genome Browser.

For this DNA methylation (Bisulfite-seq) tutorial, we align the reads against the genome (DNA) reference sequences. We the genome refernce sequences and our gene annotation files from UCSC. This is very important as we intend to perform all downstream DNA methylation analysis using the `methylKit` package in `R` which works nominally with UCSC genome references.

```
In [ ]: # Download the latest human genome
        wget -P reference http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
```

4.2 Aligning reads using Bismark aligner

4.2.1 Generate genome index

This step has to be done only once per genome type (and alignment program). It may take a long time!.

```
In [ ]: bismark_genome_preparation --verbose ./reference
```

4.2.2 Alignment

This step has to be done for each individual FASTQ file.
This step may take a long time! (may take several days to complete)

```
In [ ]: # execute Bismark aligner
        cut -f5 samples.txt | xargs -i bash -c \
        'v={}; mkdir -p alignment_Bismark/${v}; \
        bismark --parallel 8 --gzip --fastq --output_dir alignment_Bismark/${v} \
        --genome ./reference -1 fastq/${v}/${v}_1.fastq.gz -2 fastq/${v}/${v}_2.fastq.gz'
```

4.2.3 Sorting BAM files and converting to SAM files

We sort sort the BAM files using the samtools sort command:

```
In [ ]: # Sorting the bam files and converting to
        for i in alignment_Bismark/*/*; do
            if [ "${i}" != "${i%pe.bam}" ];then
                samtools sort -l 0 \
                              -T $(dirname ${i})/$(basename ${i} \
                                          _1_bismark_bt2_pe.bam)_temp \
                              -O sam -@ 8 \
                              -o $(dirname ${i})/$(basename ${i} .bam).sort.sam ${i}
            fi
        done
```

Either SeqMonk or the the Integrative Genomics Viewer (IGV) can be used to visualize the resulting sorted SAM files.

We will later use the methylKit package to import the methylation data into R from the sorted SAM files.

5 Methylation extraction using Bismark methylation extractor

With the bismark_methylation_extractor command, we extract the methylation call for every single Cytosine analyzed. This process takes as input the resulting BAM file from Bismark aligner. The bismark_methylation_extractor command writes the position of every single Cytosine to a new output file, depending on its context (CpG, CHG or CHH), whereby methylated Cytosines are labelled as forward reads (+), non-methylated Cytosines as reverse reads (-).

SeqMonk, a genome viewer, can be used to visualize the output files.

We store the output of the Bismark methylation extractor in the methylation_data folder.

```
In [ ]: # Extract methylation data
        cut -f5 samples.txt | xargs -i bash -c \
                'v={}; mkdir -p bismark_methCalls/${v}; \
                    bismark_methylation_extractor --parallel 8 \
                            --gzip \
                            --bedGraph \
                            --buffer_size 40G \
                            --merge_non_CpG \
                            --comprehensive \
                            --output bismark_methCalls/${v} alignment_Bismark/${v}/*_pe.bam'
```

In another tutorial, we will analyze DNA methylation data from the generated sorted `SAM` files from this tutorial using the `MethylKit` package in `R`.