Babraham Bioinformatics

Exercises: QC and Mapping of BS-Seq data

Version 1.4

22 May 2017

Licence

Introduction

In this session we will use small subsets of publically available Bisulfite-Seq data in FastQ format, perform initial QC and trimming steps and align the data to the mouse genome. The basic steps we're going to perform are:

- Initial quality control with FastQC
- Adapter and quality trimming using TrimGalore (wrapper around Cutadapt)
- Alignment of reads to the mouse genome using Bismark
- Deduplicate aligned data and extract methylation calls (used for analysis later on)
- Generate graphical HTML analysis report

Software

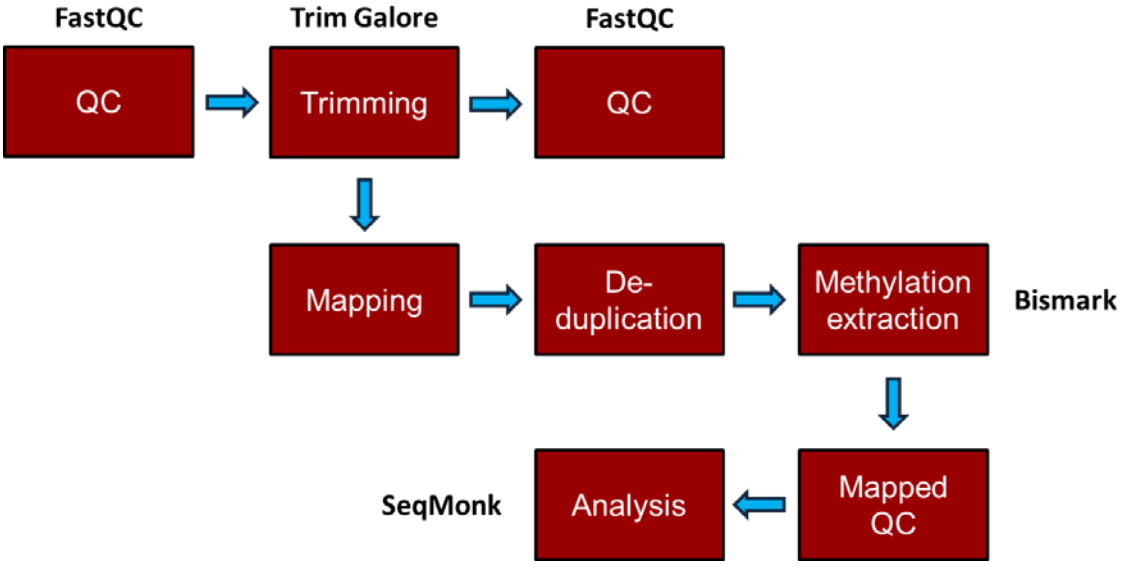
The data processing described in this practical is intended to be run as a set of command line tools and should work on Linux or Mac platforms.

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
- Cutadapt (https://code.google.com/p/cutadapt/)
- Bismark (http://www.bioinformatics.babraham.ac.uk/projects/bismark/)
- Bowtie (http://bowtie-bio.sourceforge.net/)
- Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/)

Data

The data in this practical comprises the first 200000 reads from a single-end data set (GSM1337541) and a paired-end dataset (GSM1027571). The data used in this course may be downloaded from the Babraham Bioinformatics web site (http://www.bioinformatics.babraham.ac.uk/training.html).

Exercise outline



Exercise 1 – Initial QC

The data for this section is in the QC_and_Mapping subfolder of your main data folder. You should be able to see three files in there:

```
$ cd QC_and_Mapping
$ ls
Sample1_PE_R1.fastq.gz
Sample1_PE_R2.fastq.gz
Sample2_SE.fastq.gz
```

In this exercise, we are going to run a full primary analysis of bisulfite treated reads, starting from raw FastQ files as they might come off an Illumina sequencer from any sequencing facility. Since modern sequencers generate huge data volumes (~30M sequences for a MiSeq, ~200M for a lane on the HiSeq and ~450M for a NextSeq) the analysis of a full run may take up to a couple of days to complete (depending on several parameters such as read length, genome size, repeat content, alignment parameters etc).

The two files for Sample1 are Read 1 (R1) and Read 2 (R2) of a paired-end (PE) experiment. Sample2 is a single-end (SE) experiment.

First we would like to get an initial idea about the sequencing quality of the files we are going to use. You can run FastQC on all samples in one go by using its non-interactive mode with the command:

```
$ fastqc *fastq.gz
```

Once the analysis has completed you can look at the html reports using Firefox, e.g.:

```
$ firefox Sample1_PE_R1_fastqc.html &
```

What can you tell about the files without knowing the origin of or having aligned the data? Try to find answers to the following questions:

- What does the sequence quality look like?
- Does the data appear to contain read-throughs into the Illumina adapter sequence?

- Are there overrepresented sequences/contaminants?
- Is there a noticeable difference between Read 1 and Read 2 for the paired-end experiment?
- Is there anything else you can spot from the QC reports that might affect subsequent steps?

Exercise 2: Adapter and quality trimming

In Exercise 1 we have seen that the sequencing quality in our test samples looks fairly good. Especially with longer reads or RRBS data one should still get into the habit of trimming the data to improve mapping efficiency and reduce the chance of misalignments.

Sample 1 is a paired-end experiment, so it is important to make this known to the trimming program with the option --paired:

```
$ trim_galore --paired --fastqc Sample1_PE_R1.fastq.gz
Sample1_PE_R2.fastq.gz
```

This command automatically removes:

- base calls with a Phred score of 20 or lower (assuming Sanger encoding)
- any traces of standard Illumina adapter sequence from the 3' end
- Sequence pairs where either read became too short as a result of trimming (<20 bp)
- Keeps Reads 1 and Read 2 of sequence pairs in sync (required by all aligners)

From Exercise 1 we know that Sample2 looks like a single-end RRBS experiment. Trim Galore has an additional option (`--rrbs`) that aims to remove filled-in artificial cytosines at the 3' end of reads. The command to be used is:

```
$ trim_galore --rrbs --fastqc Sample2_SE.fastq.gz
```

This command automatically removes:

- base calls with a Phred score of 20 or lower (assuming Sanger encoding)
- any traces of standard Illumina adapter sequence from the 3' end
- Sequences which became too short as a result of trimming (<20 bp)
- If adapter contamination is found, 2 additional base pairs are being removed from the 3' end to remove unwanted artificially introduced unmethylated cytosine residues

Exercise 3: Read alignments using Bismark

Alignments and methylation calling are carried out with Bismark. Both paired-end reads of Sample1 and the rather short single-end reads of Sample2 are aligned using Bowtie 2. The subfolder 'Genomes' already contains the reference sequence files for the mouse genome (NCBIM37 build) as well as the bisulfite genome indexes for Bowtie and Bowtie 2, so we can use them straight away (the indexing of a human or mouse sized genome takes ~2 hours, so we have already prepared this in advance).

To run Bismark for Sample 1 (paired-end), type:

```
$ bismark ../Genomes/ -1 Sample1_PE_R1_val_1.fq.gz -2
Sample1_PE_R2_val_2.fq.gz
```

To run Bismark for Sample 2 (single-end), enter:

```
$ bismark ../Genomes/ Sample2_SE_trimmed.fq.gz
```

The alignments should finish within a couple of minutes each. If you wanted to find out how well the alignment step went you can look at the Bismark mapping reports in the terminal window, like so:

```
$ cat Sample1_PE_R1_val_1_bismark_bt2_PE_report.txt
$ cat Sample2_SE_trimmed_bismark_bt2_SE_report.txt
```

Exercise 4: Deduplication

The deduplication is carried out using a script which is part of the Bismark package, called deduplicate_bismark. In its default mode it will use the first alignment for a given genomic region, which is basically equivalent to using a random alignment per position. The script works out whether the file to be deduplicated is a single-end or paired-end file. The parameter `--bam` ensures that the output file is a BAM file (and not an uncompressed SAM file).

```
$ deduplicate_bismark --bam Sample1_PE_R1_val_1_bismark_bt2_pe.bam
```

Note that Sample2 is an RRBS file and should not be deduplicated since this might remove a lot of valid alignments.

Exercise 5: Methylation extraction

Extracting methylation calls from the Bismark BAM files automatically generates a small report detailing the methylation calls that were found during the process (which might or might not look somewhat different to the stats in the Bismark mapping report, depending on whether or not the file was heavily duplicated). The methylation extractor also detects if the input file is a single-end or a paired-end file, and will set the option `--no_overlap` for paired-end files automatically. The option `--bedGraph` will produce the coverage files, which are very intuitive to read and will serve as input file for the R-package bsseq later in the differential methylation part of this course. The coverage files could also be submitted to GEO as processed data when it comes to publishing your data.

The methylation extractor command for Sample 1 is:

```
$ bismark_methylation_extractor --bedGraph --gzip
Sample1_PE_R1_val_1_bismark_bt2_pe.deduplicated.bam
```

This command

- generates strand and context specific cytosine output files
- counts overlapping parts of read pairs only once
- generates an M-bias report
- produces a bedGraph and coverage file
- generates an overall count report for (splitting report)

The methylation extractor command for Sample 2 is:

```
$ bismark_methylation_extractor --bedGraph --gzip
Sample2_SE_trimmed_bismark_bt2.bam
```

This command

- generates strand and context specific cytosine output files
- generates an M-bias report
- produces a bedGraph and coverage file
- generates an overall count report for (splitting report)

Exercise 6: Bismark HTML report

The module `bismark2report` generates a visual HTML report from the Bismark alignment, deduplication, methylation extraction and M-bias reports. It attempts to automatically detect all relevant files in the current working directory folder for you (which should work just fine here), but files may also be specified if desired. Just run:

```
$ bismark2report
```

The report is meant to get you an impression very quickly of how well the experiment has worked. To view the reports, type:

```
$ firefox Sample1_PE_R1_val_1_bismark_bt2_PE_report.html &
$ firefox Sample2_SE_trimmed_bismark_bt2_SE_report.html &
```

Take a look at the reports and try to answer the following questions:

- Was the mapping efficiency good or were there any major problems?
- Does the average methylation level look as expected?
- Can you comment and the level of bisulfite conversion efficiency in the samples? (You might want to look at the levels of non-CG methylation for this)
- Are there any notable technical methylation biases in the results? If so what might be the causes and what could you do to remove them?

The Bismark report should look something like this:

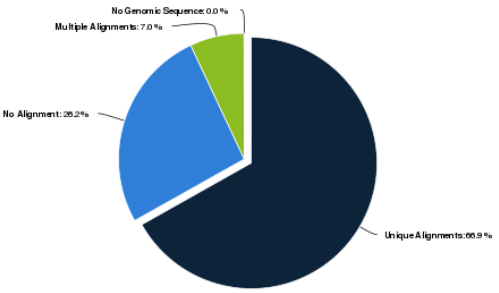
Bismark Processing Report

Sample1_PE_R1.fastq.gz and Sample1_PE_R2.fastq.gz

Data processed at 17:15 on 2014-12-10

Alignment

Sequence pairs analysed in total	200,000
Paired-end alignments with a unique best hit	133,718
Pairs without alignments under any condition	52,309
Pairs that did not map uniquely	13,973
Genomic sequence context not extractable (edges of chromosomes)	0



Cytosine Methylation

Total C's analysed	4,946,951
Methylated C's in CpG context	222,397
Methylated C's in CHG context	9,398
Methylated C's in CHH context	20,320
Methylated C's in Unknown context	0
Unmethylated C's in CpG context	92,677
Unmethylated C's in CHG context	1,282,894
Unmethylated C's in CHH context	3,319,265
Unmethylated C's in Unknown context	2
Percentage methylation (CpG context)	70.6%
Percentage methylation (CHG context)	0.7%
Percentage methylation (CHH context)	0.6%
Methylated C's in Unknown context	N/A%