Example shell commands from our bisulfite sequencing data processing pipleline

these commands represent the minimum commands necessary to reproduce pipeline output. Our actual pipeline includes commands to move files into specific folders, execute the next script, and optimize processor and memory usage on our high-performance computing cluster. Each command assumes the required program modules have been loaded.

Genome conversion with Bismark

```
bismark_genome_preparation <path_to_folder_containing_reference_genome_*.fasta_file>
```

pre-trimming FastQC

this command runs FastQC and outputs (-o) to the indicated directory; assumes *.fastq files have been gzipped (*.gz)

```
fastqc \
-o <path_to_output_directory> \
<path_to_*.fastq_files>/*.fastq.gz
```

trimming

this command uses a for loop that instructs Trim Galore! to remove 10 bp from the 5' end (--clip_r1 10), using a minimum quality score of 30 (-q 30) based on ASCII+33 quality scores as Phred scores (Sanger/Illumina 1.9+ encoding) for quality trimming (--phred33). It also specifies the non-directional nature of our RRBS libraries (--non_directional --rrbs) and instructs Trim Galore! to gzip the output (--gzip) to a defined location (-o). --fastqc_args "-o" gives the command to run FastQC on the output and place the reports in a defined location.

```
for fq in *.fastq.gz ; \
do trim_galore \
$fq \
--clip_r1 10 \
-q 30 \
--phred33 \
--non_directional \
--rrbs \
--gzip \
-o <path_to_output_directory> \
--fastqc_args "-o <path_to_output_directory>" ; \
done
```

alignment

this command uses a for loop that instructs Bismark to use the Bowtie2 alignment algorithm (--bowtie2) on our non-directional libraries (--non_directional) trimmed as above, create the output as a *.bam file, aligning to a specified reference genome (--genome could be lambda-bacteriophage or human/mouse).

```
for sample in *fq.gz ; \
do bismark \
$sample \
--bowtie2 \
--non_directional \
--bam \
--temp_dir <path_to_temporary_output_directory> \
--genome <path_to_reference_genome> \
-o <path_to_output_directory> ; \
done
```

If needed for downstream applications, the resulting *.bam file can be sorted and indexed with samtools.

```
for file in *.bam ; \
do samtools sort $file ${file%.*}_sorted ; \
done

for sorted_file in *_sorted.bam ; \
do samtools index $sorted_file ; \
done
```

methylation extraction

this command instructs Bismark to perform methylation extraction on *_sorted.bam files from above, ignoring 1 base on the 3' end (--ignore_3prime 1) (to avoid an artifact of library preparation), using a specified reference genome, generating output as bedgraph and coverage files (--bedgraph) in addition to the other standard output, and gzipping (--gzip) the output in a defined location.

```
bismark_methylation_extractor \
*_sorted.bam \
--ignore_3prime 1 \
--genome_folder <path_to_reference_genome> \
--bedgraph \
--gzip \
--output <path_to_output_directory>
```