EpiGnome™ Methyl-Seq Bioinformatics User Guide Rev. 0.1

Introduction

This guide contains data analysis recommendations for libraries prepared using Epicentre's EpiGnome™ Methyl-Seq Kit, and sequenced on an Illumina® sequencer. EpiGnome prepares whole genome bisulfite sequencing libraries (WGBS).

The following test data set is available for download, and can be used as an example data set with this guide. The data set was generated using EpiGnome Methyl-Seq Kit with 50 ng of Coriell gDNA (GM12878) as input into bisulfite conversion. The libraries were sequenced with paired-end 75 bp reads. The data set contains 10,000 reads from Read 1 and Read 2.

http://www.epibio.com/wgbs_sample/Sample_R1.fastq.gz

http://www.epibio.com/wgbs_sample/Sample_R2.fastq.gz

IMPORTANT: This document provides information on data analysis for EpiGnome™ Methyl-Seq libraries that has been demonstrated internally and may be of interest to customers. The information is provided as-is and is not an Epicentre product and is not accompanied by any rights or warranties. Customers using or adapting this information should obtain any licenses required and materials from authorized vendors. Epicentre products mentioned herein are for research use only unless marked otherwise. While customer feedback is welcomed, this user guide is not supported by Epicentre Technical Support or Field Application Scientists.

Overview

Bisulfite treatment of DNA converts non-methylated cytosine nucleotides to uracil, which are read as thymine (T) when sequenced. Methylated cytosines are not converted and still read as cytosine (C). Therefore, this method provides information about the methylation state of each nucleotide. This user guide provides guidance for analyzing bisulfite treated samples using the open-source software packages Bismark, Bowtie, Trimmomatic and SAMtools.

Note: While this document provides instructions on the analysis of Whole Genome Bisulfite Sequencing (WGBS) samples, it is not intended to be a comprehensive guide. Please also note that Bismark, Trimmomatic and SAMtools are not Illumina/Epicentre supported products.

The workflow described here assumes the user has access to a UNIX server with command-line access, 16GB of RAM *64-bit architecture (64GB RAM preferred*), and basic knowledge of UNIX. In addition, the user may require additional assistance from the IT department for software installation and permissions.

This Workflow can be applied to FASTQ file(s) either from the HiSeq® or MiSeq® Sequencers. Instructions on how to generate FASTQ files are given in the guides below:

CASAVA 1.8.2 User Guide:
http://supportres.illumina.com/documents/myillumina/a557afc4-bf0e-4dad-9e59-9c740dd1e751/casava_userguide_15011196d.pdf

MiSeq User Guide:
http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseqsystem_userguide_15027617_h.pdf

The resource below can also be useful for help in generating FASTQ file(s):
http://support.illumina.com/sequencing/sequencing_software/casava/questions.ilmn

Software Installation

Information and guides to install the required tools are listed below:

- Bismark (http://www.bioinformatics.babraham.ac.uk/projects/bismark/)
- Bowtie (http://bowtie-bio.sourceforge.net)
- Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
- SAMtools (http://samtools.sourceforge.net)
- Trimmomatic (http://www.ncbi.nlm.nih.gov/pubmed/22684630)

The following versions of these packages were used in the development of this workflow:

Bismark (version 1.8.1)

Bowtie (version 0.12.7)

Bowtie (version 2.1.0)

Trimmomatic (version 0.30)

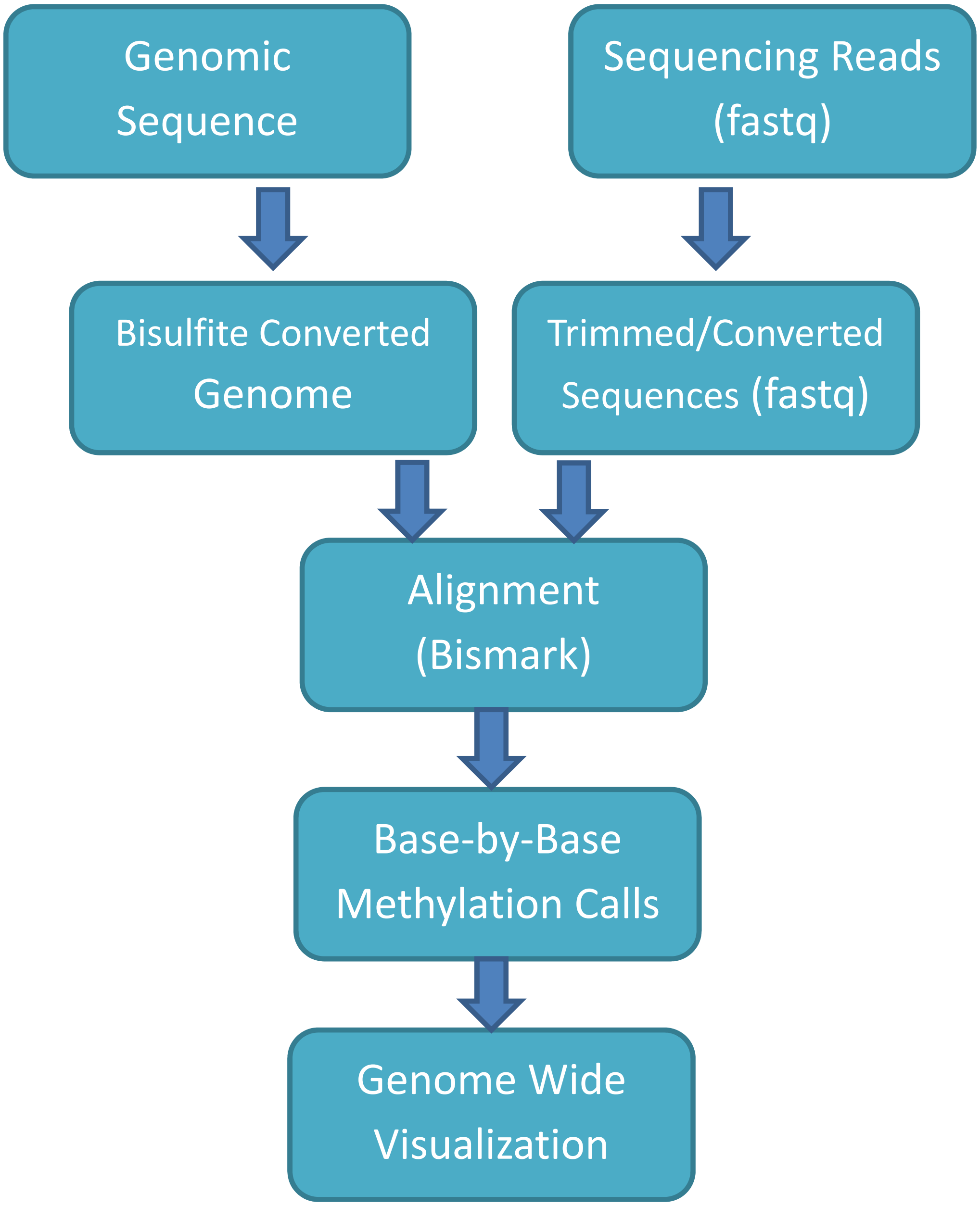
SamTools (version 0.1.18)

Though more recent versions of the programs will also be compatible with this pipeline, this workflow intended to work with the versions listed above.

Workflow

1. Filtering poor quality reads, and reads with adapter sequences (Trimmomatic)
2. Generation of bisulfite converted genome (Bismark)
3. Genome Alignment (Bismark - Bowtie)
4. Methylation calls (Bismark)
5. Generation of genome wide tracks for visualization (SAMtools)

Workflow Schematic


Genomic Sequence
Sequencing Reads (fastq)
Bisulfite Converted Genome
Trimmed/Converted Sequences (fastq)
Alignment (Bismark)
Base-by-Base Methylation Calls
Genome Wide Visualization

STEP 1 - Filtering poor quality reads and reads with adapter sequences (Trimmomatic)

Bisulfite treatment is harsh and can damage DNA. Subsequently, the shorter fragments are favored in the sample prep, generating more adaptor at the 3' end of the reads than typically observed in a standard DNA sample prep. The trimming step filters out the adaptor sequences from the fastq files, significantly improve alignment rates. In extreme cases, alignment rates can improve from 40% to 75% with this step.

As a quality control step, reads with poor quality should be filtered as well as any residual adapter sequences. Further instructions on trimmomatic can be found here:

http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.30.pdf

NOTE: It is recommended to trim the first 6 bases off the 5' end from each read (R1 and R2), as higher error rates are typically observed in the first 6 bases.

Find the location of the fastq file under the run directory (typically this would be called Unaligned):

```
> cd <run>/Unaligned/Project_A/Sample1
```

Note: Make sure you have only one fastq file. In case of a PE run then only one file for each read1 and read2. If you instead have multiple files you can combine them via UNIX
"cat" command.

```
>cat *R1.fastq.gz > Sample_R1.fastq.gz
>cat *R2.fastq.gz > Sample_R2.fastq.gz
```

```
> java -classpath /illumina/thirdparty/Trimmomatic-0.30/trimmomatic-0.30.jar
org.usadellab.trimmomatic.TrimmomaticPE -phred33 -threads 30 -trimlog r.log Sample_R1.fastq.gz
Sample_R2.fastq.gz trimmed_Sample_R1.fastq unpaired_Sample_R1.fastq trimmed_Sample_R2.fastq
unpaired_Sample_R2.fastq LEADING:30 HEADCROP:6 TRAILING:30 ILLUMINACLIP:adapter.fa:2:40:15
SLIDINGWINDOW:4:15 MINLEN:16
```

Briefly, the current trimming parameters are: (consult the trimmomatic guide for a more detailed description)

ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.

The fasta file adapter.fa used to clip the adapter sequences is listed below:

```
>adaptor1
AGATCGGAAGAGCACACGTCTGAAC
>adaptor2
AGATCGGAAGAGCGTCGTGTAGGGA
```

SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5' end and clips the read once the average quality within the window falls below a threshold.

MAXINFO: An adaptive quality trimmer which balances read length and error rate to maximize the value of each read

LEADING: Cut bases off the start of a read, if below a threshold quality

TRAILING: Cut bases off the end of a read, if below a threshold quality

CROP: Cut the read to a specified length by removing bases from the end

HEADCROP: Cut the specified number of bases from the start of the read

MINLEN: Drop the read if it is below a specified length

AVGQUAL: Drop the read if the average quality is below the specified level

TOPHRED33: Convert quality scores to Phred-33

Note: Refer to the trimmomatic user guide for guidance on the above parameters.

Step 2 - Generation of bisulfite converted genome (Bismark)

Genome Reference Files

In order to align WGBS libraries, the user will need to download the desired genome. Illumina provides these resources in the iGenomes repository for the following organisms:

- Arabidopsis_thaliana
- Bos_taurus
- Caenorhabditis_elegans
- Canis_familiaris
- Drosophila_melanogaster
- Equus_caballus
- Escherichia_coli_K_12_DH10B
- Escherichia_coli_K_12_MG1655
- Gallus_gallus
- Homo_sapiens
- Mus_musculus
- Mycobacterium_tuberculosis_H37RV
- Pan_troglodytes
- PhiX
- Rattus_norvegicus
- Saccharomyces_cerevisiae
- Sus_scrofa

The iGenomes repository can be accessed from Illumina's FTP site : ftp://ussd-ftp.illumina.com

For example, download the genome sequences for the human hg19 build from the iGenomes repository with the following commands:

```
> wget --ftp-user=igenome --ftp-password=G3nom3s4u ftp://ussd-ftp.illumina.com/Homo_sapiens/UCSC/hg19/Homo_sapiens_UCSC_hg19.tar.gz
```

Login using the following credentials:

- Username: igenome
- Password: G3nom3s4u

The downloaded file then needs to be unpacked using the following command:

```
> tar xvzf Homo_sapiens_UCSC_hg19.tar.gz
```

Unpacking will make its own folder
Homo_sapiens/UCSC/hg19

Within this folder you will find the genome sequences in FASTA format:

- Homo_sapiens/UCSC/hg19/Sequence

These files, one per chromosome, will be used to generate a converted genome.

Genome Conversion (Bismark)

The conversion steps are described below. Correct the paths as appropriate for your system. The example shown is for UCSC hg19 build within Illumina’s iGenomes. It can be easily modified for paths on your system.

1 Move to your personal genome folder and create the directory structure. The example here is for UCSC’s hg19 build.

```
cd <my_personal_genomes_folder>
mkdir Bisulfite
mkdir Bisulfite/hg19
```

2 If you have the chromosome files for the human genome in iGenomes already, create symbolic links to them. Adjust this path for your system:

```
chromosomesPath=/iGenomes/Homo_sapiens/UCSC/hg19/Sequence/Chromosomes
```

3 Ensure the path correct is created:

```
ls $chromosomesPath
```

4 Move into your FASTA folder and create all the symbolic links:

```
cd Bisulfite/hg19
ln -s $chromosomesPath/chr1.fa chr1.fa
ln -s $chromosomesPath/chr2.fa chr2.fa
ln -s $chromosomesPath/chr3.fa chr3.fa
ln -s $chromosomesPath/chr4.fa chr4.fa
ln -s $chromosomesPath/chr5.fa chr5.fa
ln -s $chromosomesPath/chr6.fa chr6.fa
ln -s $chromosomesPath/chr7.fa chr7.fa
ln -s $chromosomesPath/chr8.fa chr8.fa
ln -s $chromosomesPath/chr9.fa chr9.fa
ln -s $chromosomesPath/chr10.fa chr10.fa
ln -s $chromosomesPath/chr11.fa chr11.fa
ln -s $chromosomesPath/chr12.fa chr12.fa
ln -s $chromosomesPath/chr13.fa chr13.fa
ln -s $chromosomesPath/chr14.fa chr14.fa
ln -s $chromosomesPath/chr15.fa chr15.fa
ln -s $chromosomesPath/chr16.fa chr16.fa
ln -s $chromosomesPath/chr17.fa chr17.fa
ln -s $chromosomesPath/chr18.fa chr18.fa
ln -s $chromosomesPath/chr19.fa chr19.fa
ln -s $chromosomesPath/chr20.fa chr20.fa
ln -s $chromosomesPath/chr21.fa chr21.fa
ln -s $chromosomesPath/chr22.fa chr22.fa
ln -s $chromosomesPath/chrM.fa chrM.fa
ln -s $chromosomesPath/chrX.fa chrX.fa
ln -s $chromosomesPath/chrY.fa chrY.fa
```

5 Run the Bismark tool (this may take several hours to finish):

```
/path/to/bismark/bismark_genome_preparation \
    --verbose /path/to/fasta/files/hg19 \
    --path_to_bowtie /path/to/bowtie/linux64/
```

STEP 3 - Genome Alignment (Bismark)

Running Bismark

Further instructions are found in Bismark's User Guide:

http://www.bioinformatics.babraham.ac.uk/projects/bismark/Bismark_User_Guide_v0.8.3.pdf

USAGE: bismark [options] <genome_folder> {-1 <Read1> -2 <Read2> | <singles>}

Example:

```
/path_to_bismark/bismark_v0.8.1/bismark --bowtie2 --path_to_bowtie /mypath/bowtie2-2.1.0 -p 32
/path_to_genome/BISMARK/BS_genome_hg19
--1 trimmed_R1_fastq.gz --2 trimmed_R1_fastq.gz
```

Ouput files

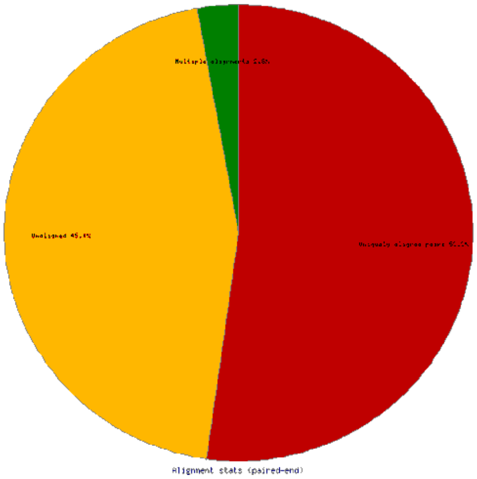
resultsTrimmed_fastq.gz_bismark_bt2_pe.sam = SAM FILE containing the alignment

R1.fastq.gz_bismark_bt2_PE_report.txt = Bismark's alignment report

Example:

```
Bismark report for: LANE_r1.fastq_trimmed.txt and LANE_r2.fastq_trimmed.txt (version: v0.7.2)
Bowtie was run against the bisulfite genome of /illumina/thirdparty/MeSeq/BS_genome_hg19/ with the specified options: -q -n
2 -l 32 -k 2 --best --maxins 500 --chunkmbs 512

Option '--directional' specified: alignments to complementary strands will be ignored (i.e. not performed)
Final Alignment report
======================
Sequence pairs analysed in total:	221418062
Number of paired-end alignments with a unique best hit: 158616794
Mapping efficiency:	71.6%
Sequence pairs with no alignments under any condition:	53367213
Sequence pairs did not map uniquely:	9434055
Sequence pairs which were discarded because genomic sequence could not be extracted:	6

Number of sequence pairs with unique best (first) alignment came from the bowtie output:
CT/GA/CT:	79135722	((converted) top strand)
GA/CT/CT:	0	(complementary to (converted) top strand)
GA/CT/GA:	0	(complementary to (converted) bottom strand)
CT/GA/GA:	79481072	((converted) bottom strand)

Number of alignments to (merely theoretical) complementary strands being rejected in total:	0

Final Cytosine Methylation Report
=================================
Total number of C's analysed:	5914886431

Total methylated C's in CpG context:	202188950
Total methylated C's in CHG context:	21952223
Total methylated C's in CHH context:	28234455

Total C to T conversions in CpG context:	171831796
Total C to T conversions in CHG context:	1453737795
Total C to T conversions in CHH context:	4036941212

C methylated in CpG context:	54.1%
C methylated in CHG context:	1.5%
C methylated in CHH context:	0.7%
```

trimmed_R1.fastq.gz_bismark_PE.alignment_overview.png   :Chart showing alignments.



Step 3.1 Removing PCR duplicates

As a byproduct of PCR, it is possible that fragments are replicated and sequenced. The following steps allow the user to remove duplicated reads from the fastq files.

1- Converts sam file to bam

```
>samtools view –S resultsTrimmed_fastq.gz_bismark_bt2_pe.sam –b –o
resultsTrimmed_fastq.gz_bismark_bt2_pe.bam
```

2- Removes PCR duplicates

```
>samtools -rmdup –S resultsTrimmed_fastq.gz_bismark_bt2_pe.bam
nodup_resultsTrimmed_fastq.gz_bismark_bt2_pe.bam
```

3- Converts bam file back to sam file (can be used in next step, methylation calls)

```
>samtools view  nodup_resultsTrimmed_fastq.gz_bismark_bt2_pe.bam –S –o

nodup_resultsTrimmed_fastq.gz_bismark_bt2_pe.sam
```

STEP 4 – Methylation calls (Bismark)

USAGE: bismark_methylation_extractor [options] <filenames>

A typical command to extract context-dependent (CpG/CHG/CHH) methylation could look like this:

```
bismark_methylation_extractor -s –comprehensive resultsTrimmed_fastq.gz_bismark_bt2_pe.sam
```

This will produce three output files:

(a) CpG_context_ resultsTrimmed_fastq_bismark.txt

(b) CHG_context_ resultsTrimmed_fastq _bismark.txt

(c) CHH_context_ resultsTrimmed_fastq _bismark.txt

STEP 5 – Generation of genome wide tracks for visualization

Generation of bedGraph files

For more information on generating bedgraph files, review the following section from Bismark's User Guide.

III Bismark methylation extractor

Briefly, below is a typical command including the optional bedGraph --counts:

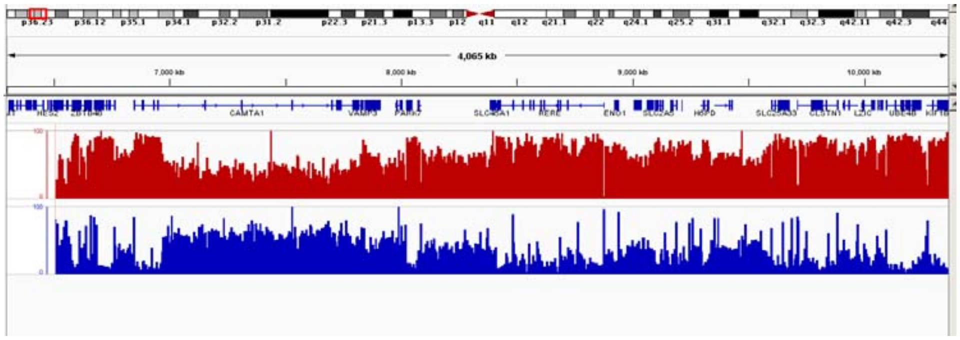
```
bismark_methylation_extractor -s --bedGraph --counts --buffer_size 10G
s_1_sequence.txt_bismark.sam
```

IGV Visualization

The generated bedgraph file can be loaded into a genome browser. For example, the bedgraph file(s) generated above can be visualized using the Integrative Genome Viewer (http://www.broadinstitute.org/igv/)

NOTE: For large genome regions we recommend uploading separate coverage track as BAM coverage is shown only for smaller regions.

Example: Region of chromosome 1 showing areas of high methylation in red, and areas that lack methylation in blue.


p36.23
p36.12
p35.1
p34.1
p32.2
p31.2
p22.3
p21.3
p13.3
p12
q11
q12
q21.1
q22
q24.1
q25.2
q31.1
q32.1
q32.3
q42.11
q42.3
q44
4,065 kb
7,000 kb
8,000 kb
9,000 kb
10,000 kb
HES2
CAMTA1
VAMP3
PARK7
SLC45A1
RERE
ENO1
SLC2A5
H6PD
SLC25A33
CLSTN1
LZIC
UBE4B
100
0
100
0