diagenode

A Hologic Company

Premium RRBS kit V2

Spike-in controls analysis to estimate bisulfite conversion efficiency

USER GUIDE

Version 1 06_2022

Get the electronic version of this user manual on the product page

Please read this manual carefully before starting your analysis

Contents

Introduction

This document describes how to process demultiplexed RRBS V2 sequencing data to extract the information from the spike-in controls included in the Premium RRBS kit V2 in order to estimate the bisulfite conversion efficiency.

Sequences of the methylated and unmethylated spike-in controls, as well as the positions of the unmethylated cytosines present in the methylated spike-in control, can be downloaded from the 'Documents' section of the Premium RRBS Kit v2 page https://www.diagenode.com/en/p/premium-rrbs-kit-V2-x24:

- RRBS_methylated_control.fa: the sequence of the methylated spike-in control in FASTA format
- RRBS_unmethylated_control.fa: the sequence of the unmethylated spike-in control in FASTA format
- RRBS_control_unmC.bed: the positions of the unmethylated cytosines in the sequence of the methylated control in BED format

In the following chapters we will guide you through all the steps of data processing for which you will need these three files.
The commands and explanations are described for paired-end sequencing data, as recommended for RRBS V2 libraires. However, command lines to calculate the conversion rate from single-end reads can be found in the 'Appendix' section.

Note that we are working in Linux, as most software tools designed for RRBS data analysis are available for this platform. This also means that you will need to have some software tools installed on your computer, but nothing more than what you are already using for RRBS data analysis: a trimming tool, a bisulfite aligner/methylation detection tool, and basic Linux tools for simple file manipulation (such as awk). Other tools (such as bedtools and samtools in our example) are optional - the same function can be performed with basic Linux tools as well, albeit it might need a little bit more work and programming skills. In the examples below we use specific software tools but you can use your preferred programs, because the principle of the process is the same with every suitable tool. A list of the tools, their version numbers used at the time of analysis and a link to the software's website are provided below.

- TrimGalore! (v0.6.6): trimming tool based on cutadapt that has an RRBS specific mode (www.bioinformatics.babraham.ac.uk/projects/trim_galore/)

• Bismark (v0.23.0): a tool for bisulfite specific alignment and calculation of the per base methylation ratios: (www.bioinformatics.babraham.ac.uk/projects/bismark/)

• bedtools (v2.27.1): a software suit that includes a wide range of tools focused on BED file manipulations (https://github.com/arq5x/bedtools2)

• samtools (v1.9): a software for reading, writing, editing, indexing and viewing SAM and BAM files (http://www.htslib.org/doc/samtools.html)

Prerequisite

The commands described below supposed that the sequencing data has been demultiplexed and the UMI extracted using the demultiplex function from fumi_tools, as explained in the 'Samples demultiplexing' section (page 38) of the Premium RRBS kit V2 Manual:

https://www.diagenode.com/files/products/kits/RRBS-KIT-V2_manual.pdf.

Indexing

The provided FASTA files can be used as any other genome for alignment. First, an index must be created using the appropriate command for the specific aligner used. For analysis with Bismark the following commands are needed:

```
bismark_genome_preparation ./genomes/RRBS_methylated_control
bismark_genome_preparation ./genomes/RRBS_unmethylated_control
```

Where ./genomes/RRBS_methylated_control and ./genomes/RRBS_ unmethylated_control are the two folders where you have put the sequence files RRBS_methylated_control.fa and RRBS_unmethylated_control.fa, respectively.

Trimming

Before the alignment, the reads need to be trimmed to remove potential adapter contamination, low-quality bases and MspI-generated library bias. This can be done with TrimGalore!, which has an RRBS specific trimming mode:

```
trim_galore --non_directional --rrbs --paired MySample_R1.fastq
MySample_R2.fastq
```

Where MySample_R1.fastq and MySample_R2.fastq are the raw read pairs from your sample of interest. Trim_galore can also work on gzip-compressed fastq files (fastq.gz or fq.gz).

The output of this command will be files in the same format as the input, with 'val_1' and 'val_2' appended to R1 and R2 filenames, respectively.

```
MySample_R1_val_1.fastq

MySample_R2_val_2.fastq
```

These files contain quality and adapter-trimmed read pairs corrected for MspI-induced specific bias, i.e. the filled-in cytosines originating from the end repair were removed to avoid their inclusion in methylation calls.

Alignment

After trimming, the read pairs need to be aligned to the spike-in control sequences. The controls do not have separate indices, they are spiked in the sample of interest in a small amount. Consequently, a few read pairs from each sample of interest will map to the spike-in control sequences. Nevertheless, these few reads give a coverage of each control spike-in sequences enough to calculate an accurate conversion rate.

With Bismark the following command can be used for the alignment:

```
bismark -q --pbat --prefix Meth_ctrl  ./genomes/RRBS_methylated_control -
1 MySample_R1_val_1.fastq -2 MySample_R2_val_2.fastq

bismark -q --pbat --prefix unmeth_ctrl ./genomes/RRBS_unmethylated_control
-1 MySample_R1_val_1.fastq -2 MySample_R2_val_2.fastq
```

Optionally, multiple cores can be specified using the --multicore option.

Each of these commands will generate an alignment bam file, which will have the following name format:

```
Meth_ctrl.Mysample_R1_val_1_bismark_bt2_pe.bam

Unmeth_ctrl.Mysample_R1_val_1_bismark_bt2_pe.bam
```

Deduplication

The alignment files produced by Bismark contains the UMI sequences extracted at the demultiplexing step using fumi_tools. Read pairs aligned at the same genomic coordinates of the reference genome and carrying the same UMI sequence will be considered as PCR duplicates originating from the amplification of the libraries. These PCR duplicates can effectively and accurately be removed using the following commands:

1. Sort the alignment bam files by genomic coordinates using samtools:

```
samtools sort Meth_ctrl.Mysample_R1_val_1_bismark_bt2_pe.bam -
o Meth_ctrl.Mysample_sorted.bam

samtools sort Unmeth_ctrl.Mysample_R1_val_1_bismark_bt2_pe.bam
-o Unmeth_ctrl.Mysample_sorted.bam
```

2. Remove the duplicated read pairs using fumi_tools:

```
fumi_tools dedup -i Meth_ctrl.Mysample_sorted.bam -o
Meth_ctrl.Mysample_dedup.bam --threads 6 --memory 3G --paired

fumi_tools dedup -i Unmeth_ctrl.Mysample_sorted.bam -o
Unmeth_ctrl.Mysample_dedup.bam --threads 6 --memory 3G --paired
```

The -o parameter in both commands specifies the name of the ouput files. Consequently, the '_sorted.bam' alignment files contain the read pairs sorted by genomic coordinates while the '_dedup.bam' alignment files contain the read pairs deduplicated based on mapping coordinates and UMI sequences. The threads and memory parameters can be adjusted based on your computing capacities.

Methylation extraction

Subsequent to the alignment and deduplication, the methylation status of the cytosines included in the spike-in control sequences can be extracted by applying the bismark_methylation_extractor function of Bismark on the deduplicated alignment bam files:

```
bismark_methylation_extractor -p --gzip --bedGraph --CX
Meth_ctrl.Mysample_dedup.bam

bismark_methylation_extractor -p --gzip --bedGraph --CX
Unmeth_ctrl.Mysample_dedup.bam
```

The methylation extraction generates several files, including a bedGraph file, a coverage file and a cytosine report; several of these contain the methylation information and can be used for calculating the conversion efficiency of the bisulfite treatment.

Conversion efficiency calculation

The conversion rate of the methylated and unmethylated spike-in controls can be calculated per sample. However, given that the bisulfite conversion is performed on pools of samples (see STEP 5 and STEP 6 of the Premium RRBS kit V2 Manual: https://www.diagenode.com/files/products/kits/RRBS-KIT-V2_manual.pdf), the conversion rate per pool (i.e. from all the samples pooled in the same tube and whose bisulfite conversion has been performed simultaneously) is more representative of actual conversion efficiencies. Both options are presented below.

1. **Conversion rate per sample**

By default, bismark_methylation_extractor generates gzip compressed bedGraph, that can be decompressed using the following command line:

```
gunzip   Meth_ctrl.Mysample_dedup.bedGraph.gz

gunzip   Unmeth_ctrl.Mysample_dedup.bedGraph.gz
```

Basically, we need to average the methylation percentages of all the Cs to get the overall methylation level of the control sequence. The conversion rate is the inverse of the methylation ratio and can be calculated as:

Conversion rate = 100%-methylation%

For the **unmethylated spike-in control** we can use an *awk* command on the bedGraph file to average the methylation percentages (in column 4) and get the conversion rates:

```
awk   '{methperc+=$4;   allC++}   END   {print   100-methperc/allC}'
Unmeth_ctrl.Mysample_dedup.bedGraph
```

The same approach can be used for the **methylated spike-in control**. However, the positions of the unmethylated cytosines must be removed before calculation – as they are supposedly converted – which can be achieved by using the *intersectBed* function of bedtools:

```
intersectBed   -v   -a   Meth_ctrl.Mysample_dedup.bedGraph   -b
RRBS_control_unmC.bed  |  awk  '{methperc+=$4;  allC++}  END  {print  100-
methperc/allC}'
```

The numbers obtained by the last two commands correspond to the conversion rate (in percent) of the unmethylated spike-in control (should be higher than 98%) and methylated spike-in controls (should be lower than 2%), respectively.

2. Conversion rate per pool of samples

By default, bismark_methylation_extractor generates gzip compressed coverage files, that can be decompressed using the following command line:

```
gunzip   Meth_ctrl.Mysample_dedup.bismark.cov.gz

gunzip   Unmeth_ctrl.Mysample_dedup.bismark.cov.gz
```

The reads mapped to the spike-in controls from all the samples included in the pool will be merged into a single file to calculate an overall coverage of the methylated spike-in control and unmethylated spike-in control per pool. Subsequently, the conversion rate can be evaluated. These steps can be achieved in a single command line:

For the **unmethylated spike-in**:

```
awk '{meth+=$5;coverage+=($5+$6)} END {print
meth/coverage*100}' Unmeth*.bismark.cov
```

The .bismark.cov files of all the samples included in a pool can be specified individually, separated by space. Alternatively, if all the samples were pooled into one single pool, the wild cards ´*´, such as in the example, can be used to concatenate all bismark.cov files generated from the unmethylated spike-in control.

The same approach can be used for the **methylated spike-in control**. However, the positions of the unmethylated cytosines must be removed before calculation – as they are supposedly converted – which can be achieved by using the *intersectBed* function of bedtools:

```
sort -k2,2n Meth*.bismark.cov | intersectBed -v -a - -b
RRBS_control_unmC.bed |  awk '{meth+=$5;coverage+=($5+$6)} END
{print meth/coverage*100}'
```

Similarly to the unmethylated spike-in control, the .bismark.cov files of all the samples included in a pool can be listed individually, separated by space or can be specified using the wild cards ´*´, such as in the example, to concatenate all bismark.cov files generated from the methylated spike-in control.

Note that there are other ways to get the conversion efficiency, and you are free to use other tools and files. For example, a slightly more precise approach would be to count all the cytosines and compare the number of methylated ones with the number of unmethylated ones (this information is in the cytosine report), and calculate the conversion ratio from these cytosine numbers directly (rather than calculating from the methylation percentages). However, we did not aim to provide an exhaustive description in this guide. Instead, our goal was to offer a simple, robust, and user-friendly universal solution.

Appendix

Command line for single end reads

Although paired-end sequencing is recommended for RRBS V2, calculating the bisulfite conversion efficiency from the methylated and unmethylated spike-in controls can also be achieved from libraries sequenced in single-end. Some steps are identical to the processing of paired-end reads and are therefore not reiterated. The commands specific to single-end reads analysis are:

- Trimming

```
trim_galore --non_directional --rrbs MySample_R1.fastq
```

This command generates the output:

```
MySample_R1_trimmed.fastq
```

- Alignment

```
bismark --pbat --prefix Meth_ctrl -q ./genomes/RRBS_methylated_control MySample_R1_trimmed.fastq
bismark --pbat --prefix Unmeth_ctrl -q ./genomes/RRBS_unmethylated_control MySample_R1_trimmed.fastq
```

This command generates the outputs:

```
Meth_ctrl.Mysample_R1_trimmed_bismark.bam
Unmeth_ctrl.Mysample_R1_trimmed_bismark.bam
```

- Deduplication
    - Sorting the alignment bam file by coordinates:

```
samtools sort Meth_ctrl.Mysample_R1_trimmed_bismark.bam -o Meth_ctrl.Mysample_sorted.bam
samtools sort Unmeth_ctrl.Mysample_R1_trimmed_bismark.bam -o Unmeth_ctrl.Mysample_sorted.bam
```

    - Deduplication based on UMI:

```
fumi_tools dedup -i Meth_ctrl.Mysample_sorted_SE.bam -o Meth_ctrl.Mysample_dedup.bam --threads 6 --memory 3G
```

```
fumi_tools dedup -i Unmeth_ctrl.Mysample_sorted_SE.bam -o
Unmeth_ctrl.Mysample_dedup.bam --threads 6 --memory 3G
```

- Methylation extraction

```
bismark_methylation_extractor -s --gzip --bedGraph --CX
Meth_ctrl.Mysample_dedup.bam

bismark_methylation_extractor -s --gzip --bedGraph --CX
Unmeth_ctrl.Mysample_dedup.bam
```

Troubleshooting

Issue: I tried to run the commands as described here in this guide, but I cannot execute them.

Investigate the error message (if there is one), as it often provides useful information about the nature of the error. Check if you copied the commands correctly without typos. Check if your file names are correct and if your files are accessible (pointing to the correct folder). Check if you have permissions to execute programs; you might need to contact your system administrator to provide the proper rights to you. Check if your software tools are properly installed. Check the versions of the software tools; it is possible that for a different version you have to specify different/ additional settings - please consult the appropriate software manual.

Issue: I have checked all the above, and I am able to run the commands, yet I cannot generate the right files/results; e.g. the reads do not map, the trimming doesn't remove the artifacts, etc.

As a general rule, the data should be processed in the same way as you process them when you align to the genome (except for the final step of filtering/calculation - but the trimming, indexing, alignment, methylation extraction is the same). Note that the commands in the guide are just examples; your dataset might need special treatments. For example, if your files are compressed, you might need to decompress them. If you have used custom adapters in your library, you might need a custom trimming procedure. Please never hesitate to consult the appropriate software manuals and adapt the commands to the needs of your data. Note that the control sequences are rather short. If you use long reads designed for very long fragments (e.g. 2x150 bp reads), then the pairs can overlap, essentially there will be zero distance between read1 and read2. Some aligners interpret this as an error and discard such pairs, meaning no pairs will be aligned to the controls. Again, we advise you to consult the software manual in such cases, as it is likely that you will find a way to collect the discarded pairs, or you will find the right settings to alter this behavior and prevent the disposition of reads altogether.

Issue: I cannot use the files I have downloaded from the Diagenode website.

The files are tested and they work on Linux and other Unix-like platforms. Please make sure that the files are not modified in any way prior to usage. For example, the newline characters are different between some Linux, Mac and Windows systems (and probably other OSs). In some cases when you download/open the file, the newline characters are automatically replaced to match the system standards. Thus the newline characters could be replaced for example when you download the files on a Windows PC, and when you move the files to a Linux server, they will not be recognized by the Linux tools. There is a simple solution for that - the dos2unix package comes preinstalled with most Linux distributions, and it can convert the line breaks between different systems. Please check the dos2unix documentation on your computer.

Issue: Everything works, but after the filtering either the unmethylated Cs are not removed, or other (methylated) Cs are also removed.

Please make sure that you handle your data and software properly (refer to the appropriate software manual if needed), especially if you are using different software tools and/or files than what are mentioned in this guide. For example, the filtering can be done with join as well (a common line-by-line file comparison Linux tool), but it needs the join fields to be sorted lexicographically. If they are not sorted properly, the lines will not be removed. Also, check the coordinate systems in the files. For example, our BED file where we store the unmethylated cytosines complies to the standard BED format, i.e. it uses zero-based half-open coordinates (start is 0-based, end is 1-based). However, the coverage file of Bismark uses 1-based coordinates (both start and end are 1-based), while in the cytosine report each C is marked by only a single 1-based position. If you compare these files to our BED file, you should harmonize the coordinate systems, otherwise you will get improper line removal (unwanted lines will be removed and/or the target lines will not be removed).

Issue: My conversion ratios have unexpected values, they are too low/ too high.

First of all, like every biochemical process, the bisulfite conversion (plus the amplification, sequencing etc.) does not have a 100% efficiency/ accuracy, and therefore it is nigh impossible to reach 0%/100% values for the methylated/unmethylated values. Usually a circa 2% error rate is expected, so a 2%/98% conversion is absolutely normal. If you obtain very different values from these 2%/98%, first double check if you have done the analysis correctly, e.g. if the filtering of unmethylated cytosines was done properly or if you did not accidentally calculate the methylation ratios instead of the conversion ratios. If everything is correct and you are convinced that your samples have a case of under-/overconversion, then you might need to redo the experiment as your methylation ratios in your samples of interest are not reliable. If you have trouble doing the bisulfite conversion, please feel free to contact Diagenode's Customer Support Customer.Support@diagenode.com or through our web interface: Technical Support https://www.diagenode.com/en/pages/support .

Note that the out-of-place conversion ratio values can not only result from faulty conversion, but also from other problems in the workflow, like an imprecise amplification during library preparation or sequencing problems. These can cause misincorporations of Cs/Ts or read errors where Cs/Ts are not detected properly - all of these can lead to apparently incorrect conversion ratios. Unfortunately these also mean that your data is unreliable and the experiment must be redone. Before repeating the experiment make sure you find the origin of the problem, e.g. check the QC metrics of your sequencing run to find out if there is an unusually high error rate in the reads.

Issue: I have another problem that is not listed here.

Please contact our customer support, describing the problem in as much detail as possible (what files you are working with, what commands you used, what are the error messages, what is your operating environment, what is your experiment setup, etc.). Please free to contact Diagenode's Customer Support Customer.Support@diagenode.com or through our web interface: Technical Support https://www.diagenode.com/en/pages/support .

Revision history

Version	Date of modification	Description of modification
Version 1	06 2022	Creation of the manual

FOR RESEARCH USE ONLY.

Not intended for any animal or human therapeutic or diagnostic use.

NOTICE TO PURCHASERLIMITED LICENSE

The information provided herein is owned by Diagenode and/or its affiliates. Subject to the terms and conditions that govern your use of such products and information, Diagenode and/or its affiliates grant you a nonexclusive, nontransferable, non-sublicensable license to use such products and information only in accordance with the manuals and written instructions provided by Diagenode and/or its affiliates. You understand and agree that except as expressly set forth in the terms and conditions governing your use of such products, that no right or license to any patent or other intellectual property owned or licensable by Diagenode and/or its affiliates is conveyed or implied by providing these products. In particular, no right or license is conveyed or implied to use these products in combination with any product not provided or licensed to you by Diagenode and/or its affiliates for such use. Limited Use Label License: Research Use Only The purchase of this product conveys to the purchaser the limited, non-transferable right to use the product only to perform internal research for the sole benefit of the purchaser. No right to resell this product or any of its components is conveyed expressly, by implication, or by estoppel. This product is for internal research purposes only and is not for use in commercial applications of any kind, including, without limitation, quality control and commercial services such as reporting the results of purchaser's activities for a fee or other form of consideration. For information on obtaining additional rights, please contact info@diagenode.com.

TRADEMARKS

The trademarks mentioned herein are the property of Diagenode or their respective owners. Bioanalyzer is a trademark of Agilent Technologies, Inc. Agencourt and AMPure® are registered trademarks of Beckman Coulter, Inc. Illumina® is a registered trademark of Illumina®Inc; Qubit is a registered trademark of Life Technologies Corporation.

www.diagenode.com