diagenode

Innovating Epigenetics Solutions

Premium RRBS spike-in controls - data processing

USER GUIDE

Version 1 08_2018

Introduction

This document describes how to process the sequencing data from the spike-in controls included in the RRBS kit in order to estimate the bisulfite conversion efficiency. You can find the sequences of the controls and the methylation status of each cytosine in the Premium RRBS Kit Manual www.diagenode.com/files/products/kits/Premium_RRBS_kit_manual.pdf. However, to ease the data processing, we provide three files that can be downloaded from the Premium RRBS Kit page www.diagenode.com/en/p/premium-rrbs-kit-x24-24-rxns:

- 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.

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 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. The list of tools we use (we also provide the version numbers we used for the example and a link to the software's website where you can find manuals, downloadables and further information):

- TrimGalore! (v0.4.5): trimming tool based on cutadapt that has an RRBS specific mode (www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
- Bismark (v0.19.0): a tool for bisulfite specific alignment and calculation of the per base methylation ratios (www.bioinformatics.babraham.ac.uk/projects/bismark/)
- bedtools (v2.26.0): a software suit that includes a wide range of tools focused on BED file manipulations

Preparation (trimming, indexing)

You can use the provided FASTA files as you would any other genome for alignment. First you need to build an index from them using the appropriate command for your preferred aligner. For example with Bismark the following commands are usually 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.

Before the alignment, the reads need to be trimmed in order to remove the RRBS library specific artifacts. This can be done with TrimGalore!, which has an RRBS specific trimming mode:

```
trim_galore --rrbs MySample.fastq
```

Where MySample.fastq is the raw reads from your sample of interest. The output of this command will be a file in the same format as the input, with a 'trimmed' tag in its name: MySample_trimmed.fastq.

Alignment

The next step after the preparations is to align the reads in your sample to the control sequences. The controls do not have separate indices, they are spiked in your sample of interest in a small amount, which means that if you align the reads in your samples, reads will map to the controls - these are usually very few, which is normal considering the length of the controls compared to the length of the genome of your target organism. Nevertheless, these few reads are still enough to give a high coverage on the control sequences. With Bismark the following command can be used for the alignment:

```
bismark --prefix meth_ctrl ./genomes/RRBS_methylated_control MySample_
trimmed.fastq
```

```
bismark --prefix unmeth_ctrl ./genomes/RRBS_unmethylated_control
MySample_trimmed.fastq
```

Methylation extraction

As it was seen before, the control sequences should be used in the same way as your genome of interest. The alignment step generated an alignment file (in BAM format by default), from which you need to extract the methylation information for each cytosine. Still using Bismark, the following commands can generate files that contain this information:

```
bismark_methylation_extractor --comprehensive --merge_non_CpG
--bedGraph --CX --cytosine_report --CX --genome_folder ./genomes/
RRBS_methylated_control meth_ctrl.MySample_trimmed_bismark_bt2.bam
```

```
bismark_methylation_extractor --comprehensive --merge_non_CpG
--bedGraph --CX --cytosine_report --CX --genome_folder ./genomes/
RRBS_unmethylated_control unmeth_ctrl.MySample_trimmed_bismark_bt2.
bam
```

Conversion efficiency calculation (filtering, averaging)

The previous step generates a number of files, like a bedGraph file and a cytosine report; several of these contain the methylation information and can be used for calculating the conversion efficiency. In our example we will use the bedGraph files (find its format description in the Bismark manual).

Basically we need to average the methylation percentages of all the Cs to get the overall methylation level of the control sequence. Because the conversion ratio is the inverse of the methylation ratio, we need to calculate the 100%-methylation% value. In the case of the methylated control, we also need to filter it to remove the unmethylated cytosines (which are supposed to be converted).

For the unmethylated control we can use a simple awk command on the bedGraph file to average the methylation percentages (in column 4) and get the conversion ratio in the end:

```
awk '{methperc+=$4; allC++} END {print 100-methperc/allC}' unmeth_
ctrl.MySample_trimmed_bismark_bt2.bedGraph
```

For the methylated control we use exactly the same approach. However, we need a preceding filtering step. We use the intersectBed tool from the bedtools suit to remove the unmethylated cytosines (provided in a BED file) from the bedGraph file:

```
intersectBed -v -a meth_ctrl.MySample_trimmed_bismark_bt2.bedGraph
-b RRBS_control_unmC.bed | awk '{methperc+=$4; allC++} END {print
100-methperc/allC}'
```

In both cases the number we obtain at the end as the final output is the conversion efficiency for the relevant 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.

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. If you have paired-end reads, you should align the reads in paired-end mode. Please never hesitate to consult the appropriate software manuals and adapt the commands to the needs of your data. A remark for paired-end reads: 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 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.

www.diagenode.com