Quality Control, trimming and alignment of Bisulfite-Seq data (Prot 57)

Felix Krueger[1], **Simon R Andrews**[1]

1. Bioinformatics Group, The Babraham Institute, Cambridge, CB22 3AT, United Kingdom

Email: felix.krueger@babraham.ac.uk, simon.andrews@babraham.ac.uk

Publication Date: 19 July 2012

Last reviewed: 09 July 2012, Christian Rohde, Department of Medicine, Hematology and Oncology, Domagkstr. 3, 48149 Münster, Germany

Introduction

Dramatic improvements and falling costs of high throughput sequencing have made bisulfite sequencing (BS-Seq) a viable option for the global analysis of DNA methylation (Bock et al, 2011; Li et al, 2010; Lister et al, 2009; Lister et al, 2011; Meissner et al, 2008; Stadler et al, 2011; Xie et al, 2012). The analysis of methylation obtained from BS-Seq is relatively straight forward, but care should be taken for initial quality control, trimming and suitable alignment of BS-Seq libraries since these are susceptible to a variety of errors or biases that one could probably get away with in other sequencing applications (discussed in (Krueger et al, 2012)).

This protocol will take you through each of the individual steps we routinely take for BS-Seq or Reduced Representation Bisulfite-Seq (RRBS) data [in our brief guide to RRBS we discuss some additional points that are specifically relevant for RRBS-type experiments (RRBS_Guide)]. Sequencing files that come fresh from the sequencer first undergo (1) initial quality control, are then subjected to (2) quality- and adapter-trimming before (3) the bisulfite reads are aligned to a genome. Optionally, results may be (4) filtered after the alignments have been performed. This procedure yields a final set of methylation data that can be analysed to answer your biological questions of interest.

(1) Quality control of high-throughput sequencing files

For all high throughput sequencing applications we recommend performing some quality control on the data to get an idea whether your experiment has worked as expected. It often enables you to tell straight away whether your dataset is of good quality or whether there were any fundamental problems with either the library or the sequencing itself. Quality control can point you towards taking appropriate steps to remove problems which is vital for the analysis of almost all sequencing applications.

For quality control purposes we use the tool `FastQC` (www.bioinformatics.babraham.ac.uk/projects/fastqc/), but there are several other programs around that serve a similar purpose (e.g. FastX toolkit, Prinseq, and others). To envoke `FastQC` on a sequence file using the command line you can simply type:

```
fastqc sequence_file.fastq
```

Alternatively you can also open sequence_file.fastq in `FastQC's` graphical interface. In the following I will give a brief outline which modules of a `FastQC` report deserve special attention for BS-Seq experiments.

Assessing sequence quality

Both the per-sequence and per-base quality plots yield valuable clues about the general (technical) quality of a sequencing experiment. Often, the quality of - especially very long - reads deteriorates towards later cycles. BS-Seq does not only rely on the correct mapping of reads like most other sequencing applications; it is furthermore also important that the reads are error-free since wrong base calls may lead to incorrect methylation calls. To avoid or reduce mis-mapping events or incorrect methylation calls we recommend using only the good quality portion of the data (see Fig. 1 and the quality trimming section below).

Base composition plot

The base composition plot can be very useful in detecting whether the bisulfite conversion has worked efficiently or whether the library suffers from potential contamination. Typical BS-Seq experiments in mammals tend to have an average cytosine content of ~1-2% throughout the entire sequence length (see Fig. 2). This may certainly be different for cell types or organisms with different methylation rates, especially in non-CG context (e.g. plants). If the occurrence of C seems to 'magically' increase to more than 20% towards later cycles you are almost certainly looking at adapter contamination of variable length (Fig. 2, left panel). In contrast to other -Seq applications, the low overall C content of BS-Seq libraries makes adapter contamination easy to spot and remove (see below). For RRBS, the base composition plot can also straight away tell you at a glance whether the experiment has worked well since all sequences should start with either CGG or TGG, depending on the methylation state of the first C (this only applies for directional libraries; for simplicity, I shall not discuss non-directional libraries here (more details can be found in the RRBS guide)).

GC-content

Since the typical GC content of (mammalian) BS-Seq libraries peaks between 20 and 30%, the per-base GC-content plot can be another way of spotting contaminating sequences. Adapter contamination can shift the GC profile to 40-60%, but this can usually be fixed by adapter trimming the sequence file (see below).

Duplication levels and over-represented sequences

A look at sequence duplication levels can quickly tell whether you have to expect a lot of duplicate alignments or whether your library looks sufficiently diverse. Whereas an estimated sequence duplication level of 10% for a mammalian shotgun BS-Seq experiment looks well diverse, a level of 80% is a strong indication that the sample is suffering heavily from PCR-duplication which should probably be removed before commencing with downstream analysis. If the over-represented sequence plot contains any sequences it may provide further clues of the potential source of contamination, typically Illumina adapters or primer sequences as a result of primer-dimers.

It is important to also consider the experimental setup when interpreting the duplication level plot: whereas an estimated sequence duplication level of 95% is pretty horrible for almost any application, it is quite normal for an RRBS library to feature duplication levels this high since all fragments are expected to line up perfectly at exactly the same genomic location numerous times (there are only so many MspI recognition sites in a genome). Similarly, sequencing a small genome with a ridiculously high number of reads (e.g. a full lane of HiSeq) may be expected to sequence the same sequences several times. In such cases, it can be (close to) impossible to say with certainty whether sequences are genuine different sequences from the same location in the genome or whether they are PCR duplicates. Therefore, at least for RRBS libraries a de-duplication step of the alignments is not desirable since it could very well remove a very large fraction of the aligned data altogether.

K-mer plot

The k-mer module often produces weird looking curves and many people are wondering what it actually means. Whilst the k-mer plot can be very powerful in detecting over-represented (chunks) of sequences that none of the other modules is able to detect reliably, for BS-Seq it is often flooded with pretty much any k-mer containing one or more Cs. This is because `FastQC` calculates an observed over expected ratio of individual k-mers by taking the overall frequency of all bases in the library into account. As C is normally very under-represented in the sequence file (~1-2%, see Fig. 2), the probability of encountering C-containing k-mers is so low that one can easily get stupidly high observed/expected ratios from just a few occurrences of C-containing k-mers. For BS-Seq, we therefore usually tend to simply ignore this plot altogether.

(2) Quality and adapter trimming

To get rid of potential problems described above, we subject all BS-Seq sequencing files to quality and adapter trimming before carrying out the actual read alignments. For trimming purposes we use `Trim Galore` (www.bioinformatics.babraham.ac.uk/projects/trim_galore/), a wrapper script that makes use of the publically available adapter trimming tool `Cutadapt`. `Trim Galore` which is preconfigured to use a set of stringent default parameters that we deem appropriate for BS-Seq files. It also has a set of additional parameters for RRBS files which is covered extensively in the `Trim Galore` documentation. Please note that all of the parameters can also be changed manually if desired.

`Trim Galore` is a command line utility that can be called by typing:

```
trim_galore sequence_file.fastq
```

This command automatically
- removes base calls with a Phred score of 20 or lower (assuming Sanger encoding)
- removes any signs of the Illumina adapter sequence from the 3' end (`AGATCGGAAGAGC`)
- removes sequences that got shorter than 20 bp

For paired-end files, `Trim Galore` can be called like so:

```
trim_galore --paired --trim1 file_1_1.fastq file_1_2.fastq
```

This command also trims low qualities and adapter contamination, but in addition it is aware of the paired-end nature of the files. Thus, it only removes sequences (from both FastQ files) if at least one of the paired sequences became shorter than the selected threshold (20 bp by default). The option '`--trim1`' trims one additional base pair from the 3' end of both reads, a step that is needed for subsequent alignments of completely overlapping long reads with Bowtie (1) (Fig. 3).

Quality trimming

Here is an example of the effect of quality trimming (Phred 20) for a dataset downloaded from the SRA (DRR001650; from (Kobayashi et al, 2012)).

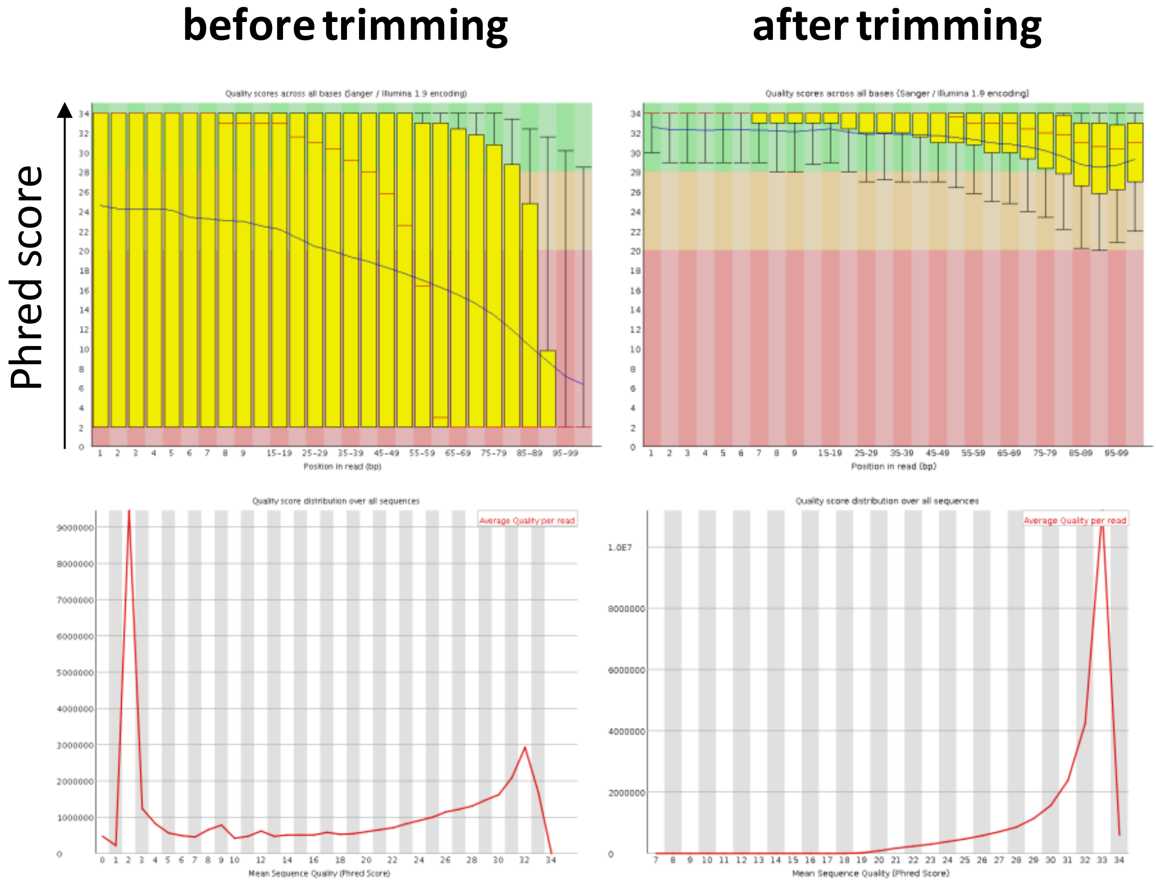

Fig. 1: Effect of quality trimming on a publically available dataset

Adapter trimming

In the next step, adapter sequences are removed from the 3' end of reads. If no specific sequence was supplied `Trim Galore` uses the first 13 bp of the standard Illumina paired-end adapters (`'AGATCGGAAGAGC'`), which recognises and removes adapters from most standard libraries [**See Comment 1**]. The stringency of the adapter removal process, i.e. the minimum number of required overlap of a sequence read with the adapter sequence, defaults to 1. This default setting is extremely stringent, i.e. an overlap with the adapter sequence of even a single bp is spotted and removed. This may appear unnecessarily harsh; however, as a reminder, adapter contamination in a BS-Seq setting may lead to mis-alignments and hence incorrect methylation calls, or result in the complete removal of the sequence because of too many mismatches in the alignment process. It is unlikely that the removed bits of sequence would have been involved in methylation calling anyway (since only the 4th and 5th adapter base would possibly be involved in methylation calls (for directional libraries that is)), however, it is quite likely that true adapter contamination – irrespective of its length – would be detrimental for the alignment, the methylation calling process, or both.

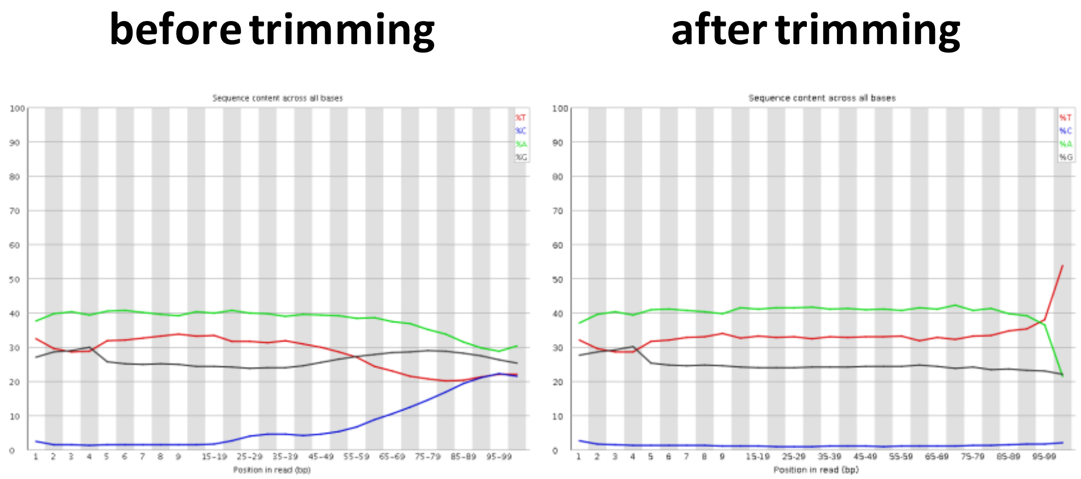

Fig. 2 Effect of adapter trimming on a publically available dataset. This example (same dataset as above) shows the dramatic effect of adapter contamination on the base composition of the analysed library. The C content (blue curve) rises from ~1% at the start of reads to around 22% **(!)** towards the end of reads. Adapter trimming effectively gets rid of most signs of adapter contamination. Note that the sharp decrease of A at the last position is a result of removing the adapter sequence very stringently, i.e. even a single trailing A at the end is removed.

(3) Alignment and methylation calling of bisulfite treated reads

Applying the aforementioned steps to both self-generated and downloaded data ensures that only the high quality portion of the data is used for alignments and further downstream analysis. We use `Bismark` (Krueger & Andrews, 2011) for BS-Seq or RRBS alignments (www.bioinformatics.babraham.ac.uk/projects/bismark/), however there is a considerable number of different aligners available as well (for a selection see (Krueger et al, 2012)). `Bismark` carries out read alignments and methylation calls at the same time, so the data is ready for analysis soon after the alignments have finished. The `Bismark` documentation provides an in-depth description of different aspects of BS-Seq and how the different Bismark modules work; so in the following section I am going to limit myself to showing typical example commands for single- or paired-end read alignments, and I'll discuss some points that are worth thinking about if the alignments don't turn out as expected.

Genome indexing

Before alignments can be carried out, the genome of interest needs to be bisulfite converted *in-silico* and indexed. If Bowtie (or Bowtie 2) are in the PATH environment, the following command will start the indexing process:

using Bowtie

```
bismark_genome_preparation /path/to/genome/
```

using Bowtie 2

```
bismark_genome_preparation --bowtie2 /path/to/genome/
```

The bisulfite genome index needs to be generated only once and can be used for all subsequent bisulfite alignments with Bismark. More information on the genome indexing step and its options may be found in the Bismark documentation.

Read alignments

The read alignment section here assumes that the data has been quality and adapter trimmed as outlined above. Untrimmed data may behave very differently to trimmed data in many respects, among them a much slower alignment speed, low mapping efficiency or a higher incidence of incorrect alignments.

In the following, I will first outline a typical choice of parameters for 100 bp single- or paired-end alignments, and then discuss a few options that are only relevant to paired-end alignments.

Single-end alignments

```
bismark -n 1 /path/to/bisulfite-genome/ file.fq
```

The seed mismatch parameter -n can sometimes be useful, especially when it is critical to obtain only reads with (up to) a certain number of mismatches. However, many users seem to mistake a high -n value with getting the highest number of alignments, which is not necessarily the case. A high -n value will however certainly increase the alignment time dramatically! If I recall it correctly the alignment times for the same 15M human sample reads (50 bp) used for the Bismark publication were ~30 mins, 50 mins and > 4 hrs for using -n values of 0, 1 or 2, respectively. Notably, the output was only marginally affected. So if possible, I would run initial alignments with a low -n value, such as 0 or 1 (the default setting is going to be changed to 1 for the next release), and only come back and change it if you are unhappy with the results.

The above command would still only allow reads to have a maximum number of 2 to 3 high-quality mismatches, owing to the default mismatch ceiling parameter `-e` (70 by default). Since we quality-trimmed the data to start with, all mismatches would have a rounded mismatch score of at least 20 (the lower threshold) or 30 (the upper cap). Thus, a read could have either up to 2 mismatches with a score of 30 or more, or up to 3 mismatches with a score of 20 before the total mismatch score would exceed the `-e` limit.

The above command is also suitable to align even fairly long reads within a reasonable time, whereby the stringency in this case is mainly conferred by trimming the data pretty conservatively. We have in fact been using this exact procedure for several lanes of 100 bp reads from the Illumina HiSeq recently and typically obtained mapping efficiencies of up to 75% (mouse genome).

Use of Bowtie 2 for alignments

Using `Bowtie 2` for bisulfite alignments offers the unique feature of aligning reads over indels. Even though `Bowtie 2` is supposed to be quicker for very long reads compared to `Bowtie`, we often see that running `Bismark` is considerably slower when run with `Bowtie 2`. So unless you would benefit from indel mapping we would suggest running `Bismark` in the default, i.e. `Bowtie`, mode.

Paired-end alignments

This would be pretty much the same command for paired-end alignments:

```
bismark -n 1 /path/to/bisulfite-genome/ -1 file_1.fq -2 file_2.fq
```

Low mapping efficiency

A frequent problem of paired-end alignments seems to be a low mapping efficiency. This can be a result of specifying the lower and upper fragment lengths too narrowly, e.g. with `-I 150` and `-X 300` (the defaults are 0 and 500). We would advise against such stringent settings since often the size-selected fragments turn out to be much smaller or larger than intended. Leaving `-I` at the default of 0 and setting `-X` to 1000 or so may offer the solution to a low mapping efficiency.

Another problem commonly arises when the read length is so long that both reads do completely contain each other, because such a scenario (Fig. 3A) is not considered a valid paired-end alignment by `Bowtie` (`Bowtie 2` considers it valid though). With ever increasing read lengths this has turned out to become a very frequently encountered problem. As hinted in the previous section, this problem can be overcome by shortening both reads by only a single base pair on the 3' ends of both reads (Fig. 3B). In recent 2x100 bp HiSeq lanes we saw an increase in mapping efficiencies of up to 30% (!) by running `Trim Galore` with `'--trim1'`; a very worthwhile step, indeed.

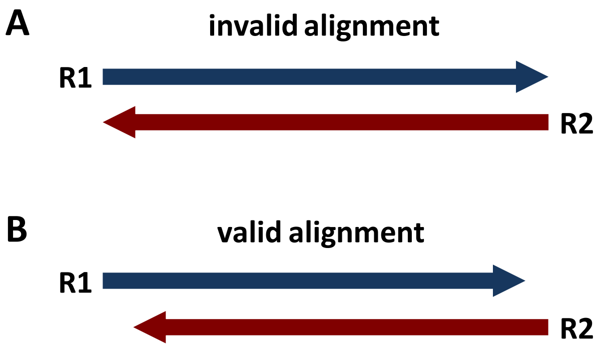

Fig. 3: Self-containing sequences are considered invalid paired-end alignments for `Bowtie`.

Using the methylation extractor

The methylation extractor produces tab delimited files with content-specific methylation information. These files serve as the endpoint of the `Bismark` package and can be easily processed further into various other formats (e.g. into `bedgraph` files using a bismark2bedgraph conversion script that is available from the Bismark project page) or read into visualisation and analysis tools such as `SeqMonk`). More information about the methylation extractor or its options can be found in the `Bismark` documentation.

Extracting methylation information from a single-end file

```
methylation_extractor -s file.fq_bismark.sam
```

Extracting methylation information from a paired-end file

```
methylation_extractor -p --no_overlap file_1.fq_bismark_pe.sam
```

We would like to recommend running the methylation extraction for paired-end files using the option '`--no_overlap` '. This is to avoid that overlapping parts of read pairs are not carried over to the methylation extraction file two times, as this would confer a coverage bias for the overlapping portion of the reads.

(4) Filtering reads after alignments

Filtering out reads with non-CG methylation

A number of studies chose in the past to filter out or truncate reads that contain too many (typically 3) methylated cytosines in non-CG context. This was done under the assumption that non-CG methylation is virtually non-existent; however there are a number of publications out by now suggesting that this is not necessarily the case [(Lister et al, 2009; Lister et al, 2011; Stadler et al, 2011; Xie et al, 2012)]. By removing a subset of reads or parts of reads that show intermediate or high methylation in non-CG context one risks biasing the results and removing potentially interesting data. We choose to not systematically bias our results by arbitrarily removing parts of the data, with the rationale being that non-bisulfite conversion should occur randomly and at a low frequency, and therefore a few methylated Cs here and there should not make a big difference on a global scale (whereas removing reads with non-CG methylation and then analysing non-CG methylation seems ... an interesting concept).

De-duplication

Mammalian genomes are so huge that it is rather unlikely to encounter several genuinely independent fragments which align to the very same genomic position. It is much more likely that such reads are a result of PCR amplification. For large genomes, removing duplicate reads is therefore a valid route to take. A de-duplication tool is available from the `Bismark` project page for this purpose. This de-duplication removes reads that have the same orientation, and start and end at the same position. Even after the removal of duplicates, any given position in the genome may theoretically still be covered by N reads, where N is the read length (this number may be even higher for paired-end reads).

Filtering for high read coverage

Before starting with the actual analysis of methylation we tend to ignore regions of the genome that are heavily over-represented in comparison to the rest of the genome. Such regions are commonly found close to telomeres or centromeres, or at regions that closely resemble repeat regions. A prime example for this is a specific region on mouse chromosome 2 that apparently looks very similar to satellite repeats (Fig. 4; around 8kb in total, covering 3 CpG islands). We found that up to 8% (!) of all reads of shotgun BS-Seq experiments may align to this specific location, e.g. more than 110 million reads from the Kono data aligned to this very region (Kobayashi et al, 2012). Interestingly, this region also shows 'spikes' of aligned reads in all sorts of other sequencing applications, such as ChIP-Seq, RNA-Seq, MeDIP-Seq or even Input controls. This is most likely caused by the fact that satellite regions are not part of the genome assembly, however a substantial number of satellite reads is present in any library. Thus, reads are apparently mis-placed to specific regions that most closely resemble satellite regions. We know that we can't get that many correct sequences from a region - therefore we must be mis-measuring our data in those regions and we can therefore ignore them.

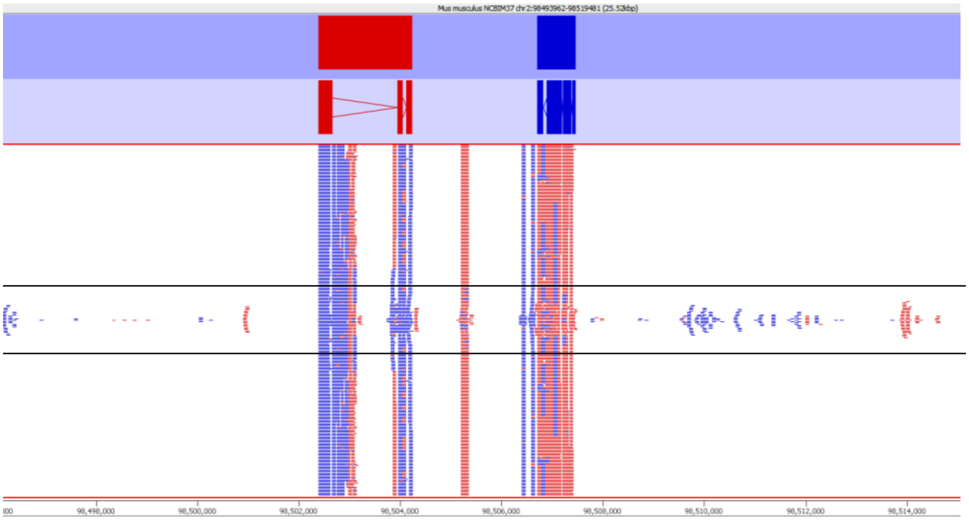

Fig. 4 SeqMonk screenshot from an example region on mouse chromosome 2 which features an enormous number of aligned reads in many different sequencing experiments. The black lines indicate the typical read coverage in the example whole genome BS-Seq dataset shown. The top 2 tracks show gene and mRNA annotations. Red and blue lines in the lower track indicate read alignments in forward or reverse orientation, respectively.

Reviewer comments:

Reviewed by Christian Rohde

In addition, removal of 3'-MspI-sites in RRBS data analysis is crucial, since its methylation state is determined by the cytosine nucleotide used for library preparation. This problem is addressed in Trim Galore with the --rrbs option. This option is very helpful for RRBS data analysis and explained in detail in the manual. Briefly, Trim Galore removes 2 additional nucleotides once an adapter sequence is detected. Thereby, it removes most of the 3'-MspI-sites. However, in case that the read length exactly matches the MspI-fragment length this problem cannot be addressed by Trim Galore. In such case the adapter is not yet sequenced and will not be removed. Unfortunately, the erroneous methylation information will remain in the data. In order to resolve these rare cases another round of post mapping 3'MspI-site detection and removal will be beneficial.

Keywords:

bisulfite sequencing
bisulfite alignments
quality control
adapter trimming
quality trimming
deduplication
FastQC
Bismark

References:

Bock C, Kiskinis E, Verstappen G, Gu H, Boulting G, Smith ZD, Ziller M, Croft GF, Amoroso MW, Oakley DH, Gnirke A, Eggan K, Meissner A (2011) Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144:** 439-452

Kobayashi H, Sakurai T, Imai M, Takahashi N, Fukuda A, Yayoi O, Sato S, Nakabayashi K, Hata K, Sotomaru Y, Suzuki Y, Kono T (2012) Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet* **8:** e1002440

Krueger F, Andrews SR (2011) Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*

Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* **9:** 145-151

Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, Zhang H, Chen Q, Luo R, Chen M, He Y, Jin X, Zhang Q, Yu C, Zhou G, Huang Y, Cao H, Zhou X, Guo S, Hu X, Li X, Kristiansen K, Bolund L, Xu J, Wang W, Yang H, Wang J, Li R, Beck S, Zhang X (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol* **8:** e1000533

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315-322

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM, Ecker JR (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471:** 68-73

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766-770

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490-495

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B (2012) Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148:** 816-831