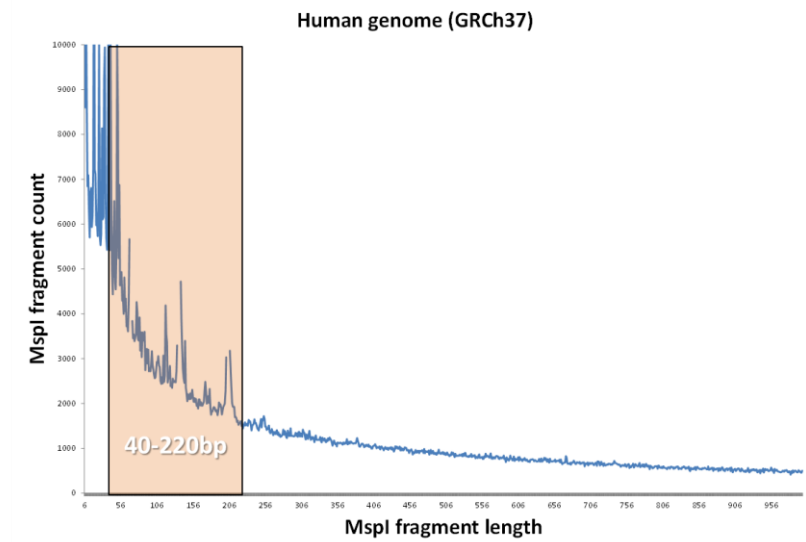


Reduced Representation Bisulfite-Seq – A Brief Guide to RRBS

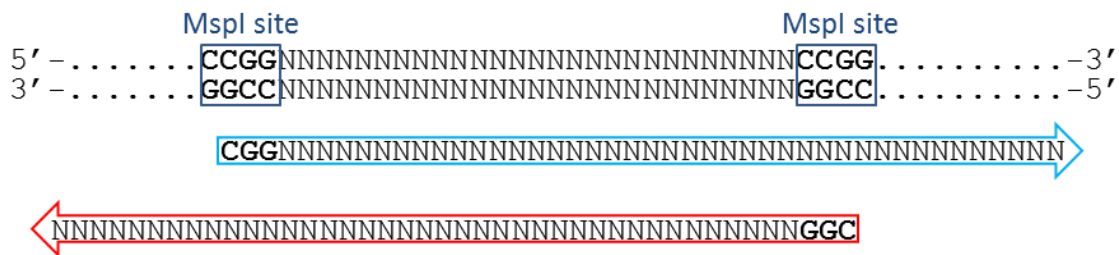
What is RRBS?

Typically, RRBS samples are generated by digesting genomic DNA with the restriction endonuclease *MspI*. This is followed by end-repair, A-tailing, adapter ligation and finally bisulfite conversion. Often, the library is also size-selected for fragments between 40 and 220bp in length. This fragment size has been shown to be plentiful in the sample and yield information on the vast majority of CpG islands (CGIs) in the human or mouse genome. Fig. 1 shows that quite a few *MspI*-*MspI* fragments (generated *in silico* for the mouse genome) are even shorter than 40bp. Since the size selection process is not as good as it is in theory, often a sizeable number of fragments below 40bp can end up in the RRBS library.



The fairly small fragment size of RRBS fragments can become a potential problem especially for sequencing reads with high read length (e.g. > 75bp or >100bp). If the read length is

longer than the MspI-MspI fragment itself, the sequencing read may continue to read into the adapter sequence on the 3' end:

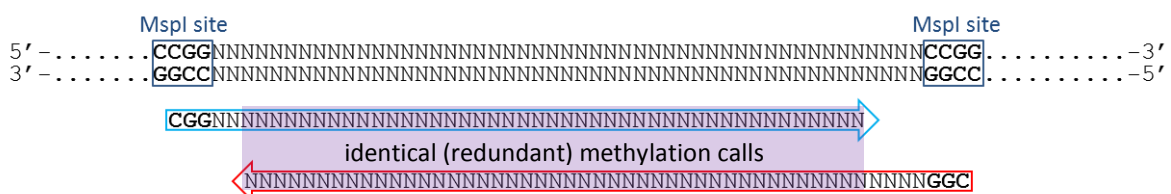


Such read-through adapter contamination may result in a lower mapping efficiency if the read does not align at all, or it may lead to false alignments which can result in incorrect methylation calls. As a simple rule, the longer the read length the higher the proportion of reads with adapter contamination. If such adapter contamination is not spotted and removed appropriately, a longer read length is most likely resulting in a lower mapping efficiency!

If the read length is longer than the MspI fragment one will also read (and perform a methylation call) for a cytosine that has been filled in with a predefined methylation state during the end-repair step. This is discussed further below.

Single-end or paired-end?

It seems to be a common misconception, that paired-end reads yield methylation results for both the forward and the reverse strand. In reality, a paired-end read results from PCR amplification of either the original top strand (OT), or the original bottom strand (OB). Thus, the other ends that are sequenced in the second round are sequences from the strand complementary to OT (CTOT), or complementary to OB (CTOB). These complementary strands are also informative for the same strand as their partner reads. As a consequence, paired-end reads that overlap in the middle yield redundant methylation information for the same strand:

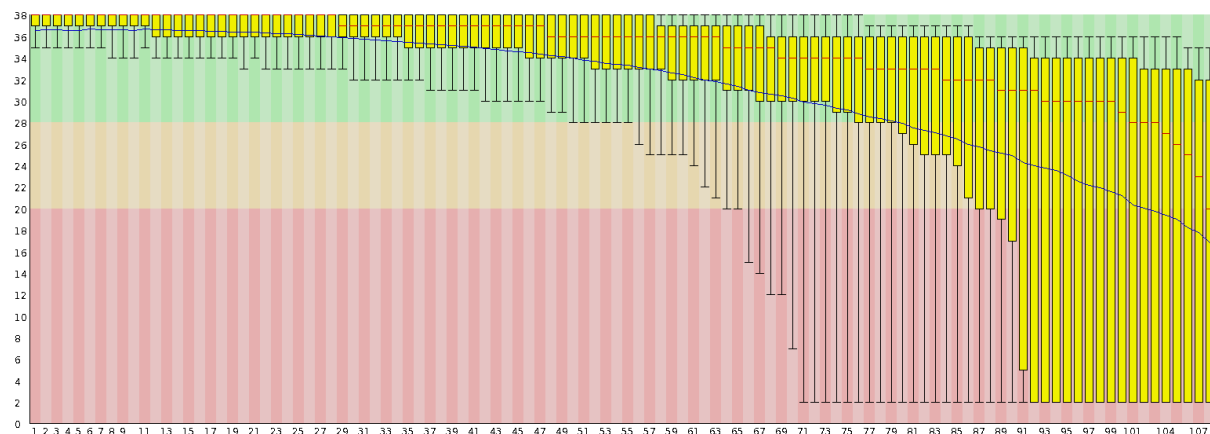


Granted, the paired-end nature might result in a somewhat increased mapping efficiency of paired-end reads over single-end reads. However, in addition to reading into the adapter on the other side, paired-end reads face the additional problem of generating potentially redundant methylation information. Redundant methylation calls need to be discarded if positions are filtered for a certain coverage by independent reads, since regions of overlap for paired-end reads would be over-represented. In short, because of the redundant overlapping parts paired-end RRBS reads are not

simply ‘twice as many reads therefore twice as many methylation calls’. Single-end experiments with the same number of reads as both paired-end reads added together are more likely to yield more genuine methylation information, as long as the read length is long enough to allow for a fair mapping efficiency for single-end reads (40-50bp reads are probably long enough to get mapping efficiencies in the range of 60 to 70%).

Other read length effects

Current sequencing on the Illumina platform often produces data whose quality deteriorates towards later cycles. We have received feedback from numerous sources, or downloaded data from public archives, which looks similar to this read quality profile:



Up to a read length of around 60-70bp, the basecall quality of these reads is excellent ($> \text{Phred } 30$). After that, however, Phred scores tend to drop dramatically in a fairly large number of sequences, which means the rates at which bases are called erroneously increases. Base call errors in reads can result in reads not being aligned at all (reduced mapping efficiency), incorrect methylation calls or, in the worst case, mis-alignments (which will most likely also generate incorrect methylation calls). Some of these aspects have been recently discussed in the following review: Krueger F., Kreck B. et al., DNA Methylome Analysis Using Short Bisulfite Read Data, 2012).

In a nutshell:

RRBS reads suffer disproportionately from problems associated with long read lengths because they are, compared to other -Seq applications, size-selected for rather short fragment sizes. In the following sections, I will discuss some further aspects that need to be considered when analysing RRBS samples, and I will introduce our way of dealing with all sorts of read length related problems or experimentally introduced biases for RRBS libraries.

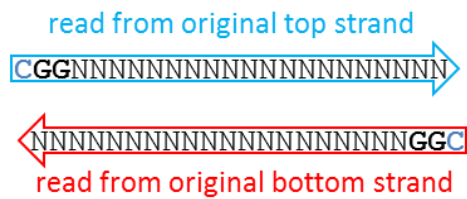
Directional sequencing, the most wide-spread way of (RR)BS, only ever sequences reads originating from the original top (OT) or original bottom (OB) strands. For simplicity, I have refrained from drawing out all four bisulfite DNA strands for the illustration below.

[illegible]

After the adapters are attached, the sequences are treated with sodium bisulfite, which converts unmethylated cytosines into thymines. Thus, the first three bases of (almost) all RRBS reads are either **CGG** or **TGG**, depending on their genomic methylation state. This applies to reads from both the OT and OB strand, and as nearly all reads in a directional RRBS experiment start with one of these two options, every read provides information on at least one CpG right in the start.

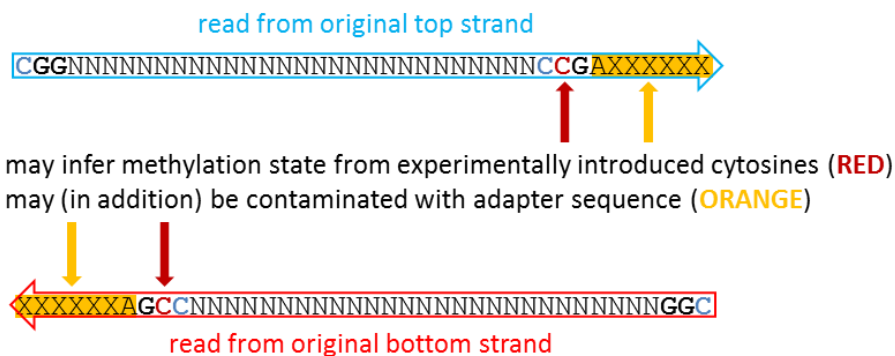
For directional libraries one can then discriminate the following two cases:

A) The read length is shorter than the MspI fragment



In this case, the entire read can be used for alignments and methylation calls. The first position resembles the true genomic methylation state (which can be C or T).

B) The read length is longer than the MspI fragment size

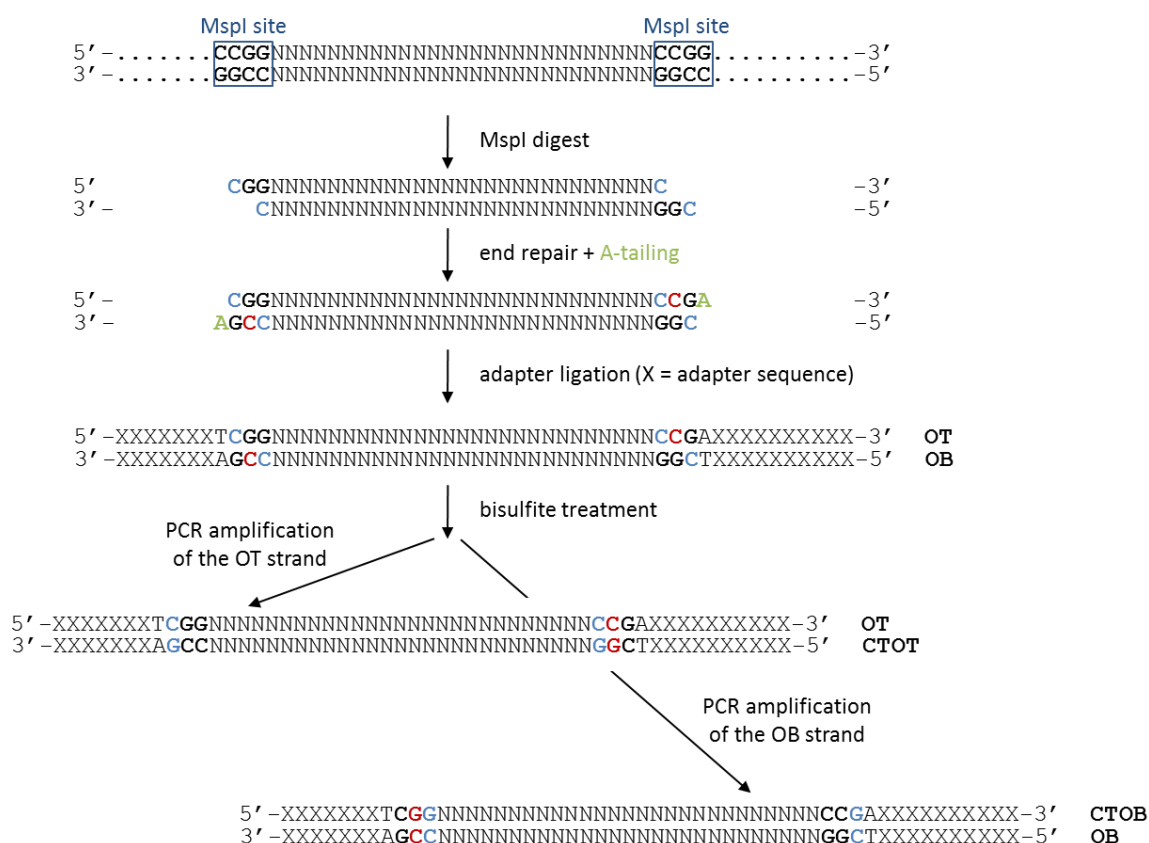


In this case, the sequencing read will contain the position which has been filled in during the end repair step (marked in **RED**), as well as read into the adapter sequence on the 3' end of the read (marked in **ORANGE**). Retaining either the biased position or adapter contamination in the sequence read is highly undesirable.

Non-directional or paired-end RRBS libraries

Non-directional bisulfite sequencing is less common, but has been performed in a number of studies (Cokus et al. (2008), Popp et al. (2010), Smallwood et al. (2011), Hansen et al. (2011), Kobayashi al. (2012)). In this type of library, sequence reads may originate from all four possible bisulfite DNA strands (original top (OT), complementary to OT (CTOT), original bottom (OB) or complementary to OB (CTOB)) with roughly the same likelihood.

Paired-end reads do by definition contain one read from one of the original strands as well as one complementary strand. Please note that the CTOT or CTOB strands in a read pair are reverse-complements of the OT or OB strands, respectively, and thus they carry methylation information for the exact same strand as their partner read but not for the other original DNA strand. Similar to directional single-end libraries, the first read of directional paired-end libraries always comes from either the OT or OB strand. The first read of non-directional paired-end libraries may originate from any of the four possible bisulfite strands.

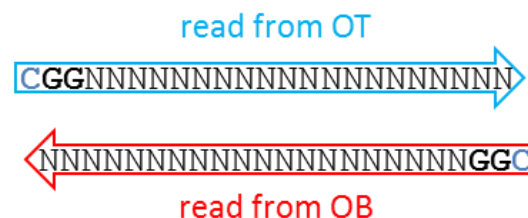


Again, **cytosines in blue** retain the original genomic methylation state, whereas **cytosines in red** are introduced experimentally during the fragment end-repair reaction. This can be accomplished with either unmethylated or methylated cytosines, the trend seems to be that unmethylated cytosines are being used primarily now.

After bisulfite conversion, the first three bases of non-directional RRBS reads that originated from the OT or OB strands will also be either **CGG** or **TGG**, depending on their genomic methylation state. In addition, however, non-directional libraries may contain reads which originated from the CTOT or CTOB strands. These reads will have **CAA** or **CGA** at the start, depending on whether unmethylated or methylated cytosines were used for the end-repair reaction, respectively. (Theoretically, the sequence could also be **CAG** or **CGG**, but this would assume that a C in CHH context was methylated on the other strand, and this is arguably very rarely the case for CpG-rich sequences. For simplicity it is therefore left out here). In either case, the second base would incur a methylation call of a base that does no longer reflect the genomic methylation state, which is illustrated below.

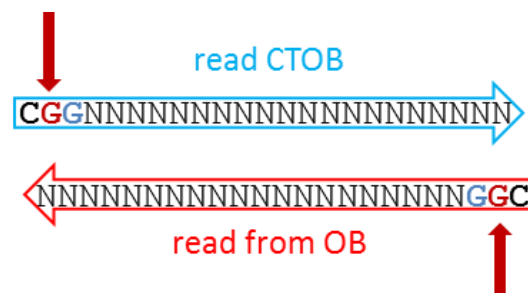
For non-directional libraries, one can discriminate the following four cases:

A) The read length is shorter than the MspI fragment, OT or OB alignments



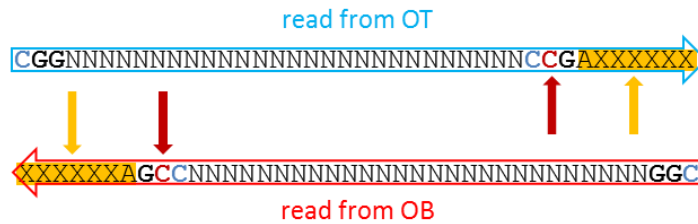
In this case, the entire read can be used for alignments and methylation calls. The first position resembles the true genomic methylation state (which can be C or T).

B) The read length is shorter than the MspI fragment, CTOT or CTOB alignments



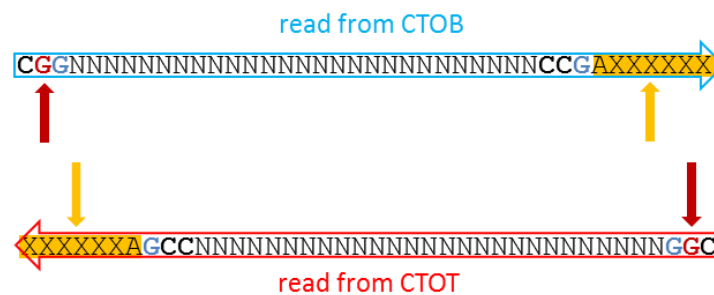
In this case, the read will start with **CAA** or **CGA** (the filled-in and the other cytosine in CHG context on the opposing strand are expected to be fully bisulfite converted), whereby the position marked in **RED** infers the methylation state from a cytosine that was experimentally introduced. The positions in **BLUE** carries genomic methylation information. As a consequence, methylation information of the second base would bias the results depending on the methylation state of the cytosine used for end-repair and needs to be excluded from methylation analysis.

C) The read length is longer than the fragment length, OT or OB alignments



Analogous to directional libraries, the sequencing read will contain the position that was filled in during the end-repair step (marked in **RED**), as well as read into the adapter sequence on the 3' end of the read (marked in **ORANGE**). Retaining either the biased position or adapter contamination in the sequence read is highly undesirable.

D) The read length is longer than the fragment length, CTOT or CTOB alignments



- reads may infer methylation state from experimentally introduced cytosines (RED)
- reads may (in addition) be contaminated with adapter sequence (ORANGE)

Similar to case **B)** the read does now contain the filled-in base at position 2 in the read (typically CAA or CGA, **RED**), as well as adapter contamination at the 3' end (**ORANGE**). Retaining either the biased position or adapter contamination in the sequence read is highly undesirable.

As for almost all high throughput sequencing applications we would recommend to perform some quality control on the data, as it can often straight away point you towards the next steps that need to be taken (e.g. with FastQC, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). As outlined above, we believe that thorough QC and taking appropriate steps to remove problems is absolutely critical for proper analysis of RRBS libraries since they are susceptible to a variety of errors or biases that one could probably get away with in other sequencing applications. In summary, the examples discussed here were:

- ### Exploiting the filled-in position to determine the bisulfite conversion efficiency

Diagram illustrating bisulfite conversion of DNA. The top strand, labeled "bisulfite conversion complete", shows a sequence of 'N's (non-CpG sites) followed by a 'TGA' sequence (CpG site) and then 'X's (non-CpG sites). A red arrow points to the 'T' in 'TGA'. The bottom strand, labeled "bisulfite conversion failed", shows the same sequence, but the 'T' in 'TGA' is highlighted in blue, indicating it was not converted to 'C'.

$$\% \text{ non-conversion} = C_X / (T_Y + C_X) * 100$$

(with X and Y being the number of times these residues were observed).

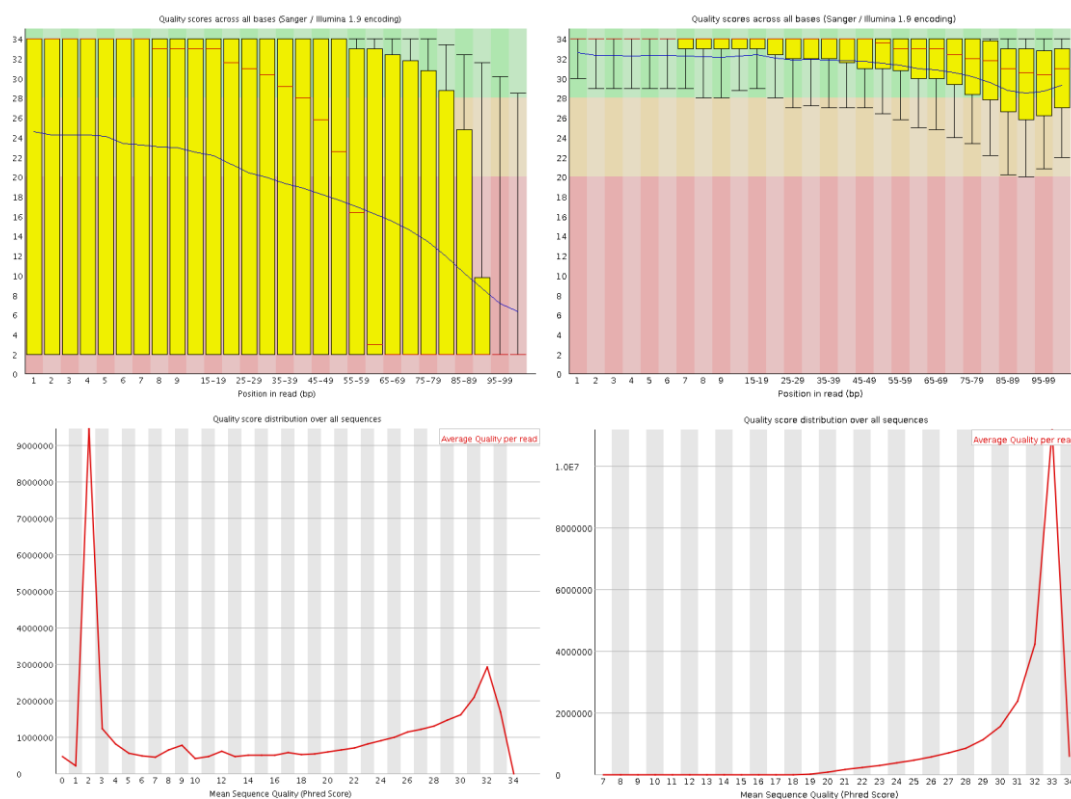
To make this calculation more reliable one can vary the required overlaps with the adapter sequence, e.g. requiring 3, 5, 7bp etc., before the non-conversion rate of the filled-in position is determined.

Adaptive quality and adapter trimming with Trim Galore!

We have tried to implement a method to rid RRBS libraries (or potentially all kinds of sequencing datasets) of potential problems in one convenient process. For this we have developed a wrapper script (`trim_galore`) that makes use of the publically available adapter trimming tool `Cutadapt` and `FastQC` for optional quality control once the trimming process has completed.

Even though `Trim Galore` works for any (base space) high throughput dataset (e.g. downloaded from the SRA) this section describes its use mainly with respect to RRBS libraries.

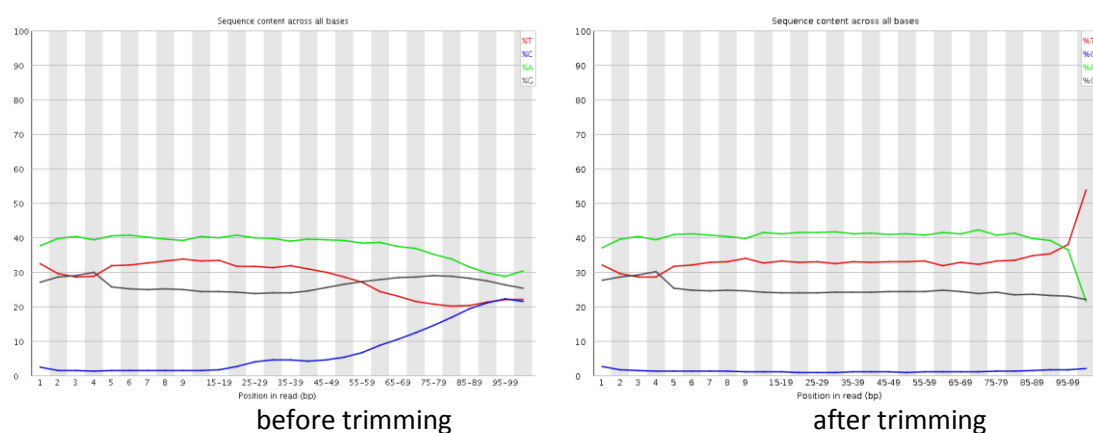
- In the first step, low-quality base calls are trimmed off from the 3' end of the reads before adapter removal. This efficiently removes poor quality portions of the reads. Here is an example of a dataset downloaded from the SRA which was trimmed with a Phred score threshold of 20 (data set DRR001650_1 from Kobayashi et al., 2012).



before trimming

after trimming

In the next step, *Cutadapt* finds and removes adapter sequences from the 3' end of reads. If no sequence was supplied it will use the first 13 bp of the standard Illumina paired-end adapters ('AGATCGGAAGAGC'), which recognises and removes adapters from most standard libraries. To control the stringency of the adapter removal process one gets to specify the minimum number of required overlap with the adapter sequence, else it will default 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 may in a bisulfite-Seq setting lead to mis-alignments and hence incorrect methylation calls, or result in the removal of the sequence as a whole because of too many mismatches in the alignment process. Tolerating adapter contamination is most likely detrimental to the results, but we realize that this process may in some cases also remove some genuine genomic sequence. 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 or methylation call process, or both.



This example (same dataset as above) shows the dramatic effect of adapter contamination on the base composition of the analysed library, e.g. the C content rises from ~1% at the start of reads to around 22% (!) towards the end of reads. Adapter trimming with *Cutadapt* gets rid of most signs of adapter contamination efficiently. 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.

– *Trim Galore* also has an '--rrbs' option, which identifies sequences that were adapter-trimmed and removes another 2 bp from the 3' end of Read 1 (or the only read in single-end experiments). This is to avoid that the filled-in cytosine position (marked in **RED** in the above examples) close to the second *MspI* site in a sequence is used for methylation calls. Sequences which were merely trimmed because of poor quality will not be shortened any further. Read 2 however does not technically need 3' trimming this since the end of Read 2 is not affected by the artificial methylation states introduced by the [end-repair] fill-in reaction. Instead, the first couple of positions of Read 2 suffer from the same fill-in problems as standard paired-end libraries

(<https://sequencing.qcfail.com/articles/library-end-repair-reaction-introduces-methylation-biases-in-paired-end-pe-bisulfite-seq-applications/>). Read 2 reads will thus be trimmed by 2bp from the 5' end instead (issue on Github (<https://github.com/FelixKrueger/TrimGalore/issues/3>)).

- `Trim Galore` also has a `'--non_directional'` option, which will screen adapter-trimmed sequences for the presence of either CAA or CGA at the start of sequences and clip off the first 2 bases if found. If CAA or CGA are found at the start, no bases will be trimmed off from the 3' end even if the sequence had some contaminating adapter sequence removed (in this case the sequence read likely originated from either the CTOT or CTOB strand).

- Lastly, since quality and/or adapter trimming may result in very short sequences (sometimes as short as 0 bp), `Trim Galore` can filter trimmed reads based on their sequence length (default: 20 bp). This is to reduce the size of the output file and to avoid crashes of alignment programs which require sequences with a certain minimum length.

Paired-end reads: Most alignment programs require both paired-end `FastQ` files to contain sequence pairs in exactly the same order. To not bring this sequence-by-sequence order out of sync, `Trim Galore` trims paired-end `FastQ` files sequentially but does not discard too short sequences straight away. Instead, once both files have been trimmed it performs a validation of both paired-end files at the same time where it can discard entire sequence pairs if one of the two sequences has become too short during the trimming. This process retains the sequence-by-sequence order of paired-end files, and results in the final output files of paired-end trimming, termed `_val_1.fq` and `_val_2.fq` (for validated). `Trim Galore` also allows you to keep singleton reads if they fulfil certain length criteria (see the `Trim Galore` help for more details).

Applying these steps to both self-generated and downloaded data can ensure that you really only use the high quality portion of the data for alignments and further downstream analyses and conclusions.