Oct 19, 2012

Taking appropriate QC measures for RRBS-type or other -Seq applications with `Trim Galore!`

For all high throughput sequencing applications, we would recommend performing 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). Thorough quality control and taking appropriate steps to remove problems is vital for the analysis of almost all sequencing applications. This is even more critical for the 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 our brief guide to RRBS (RRBS Guide) we discuss the following points:

- poor qualities – affect mapping, may lead to incorrect methylation calls and/or mis-mapping
- adapter contamination – may lead to low mapping efficiencies, or, if mapped, may result in incorrect methylation calls and/or mis-mapping
- positions filled in during end-repair will infer the methylation state of the cytosine used for the fill-in reaction but not of the true genomic cytosine
- paired-end RRBS libraries (especially with long read length) yield redundant methylation information if the read pairs overlap
- RRBS libraries with long read lengths suffer more from all of the above due to the short size-selected fragment size

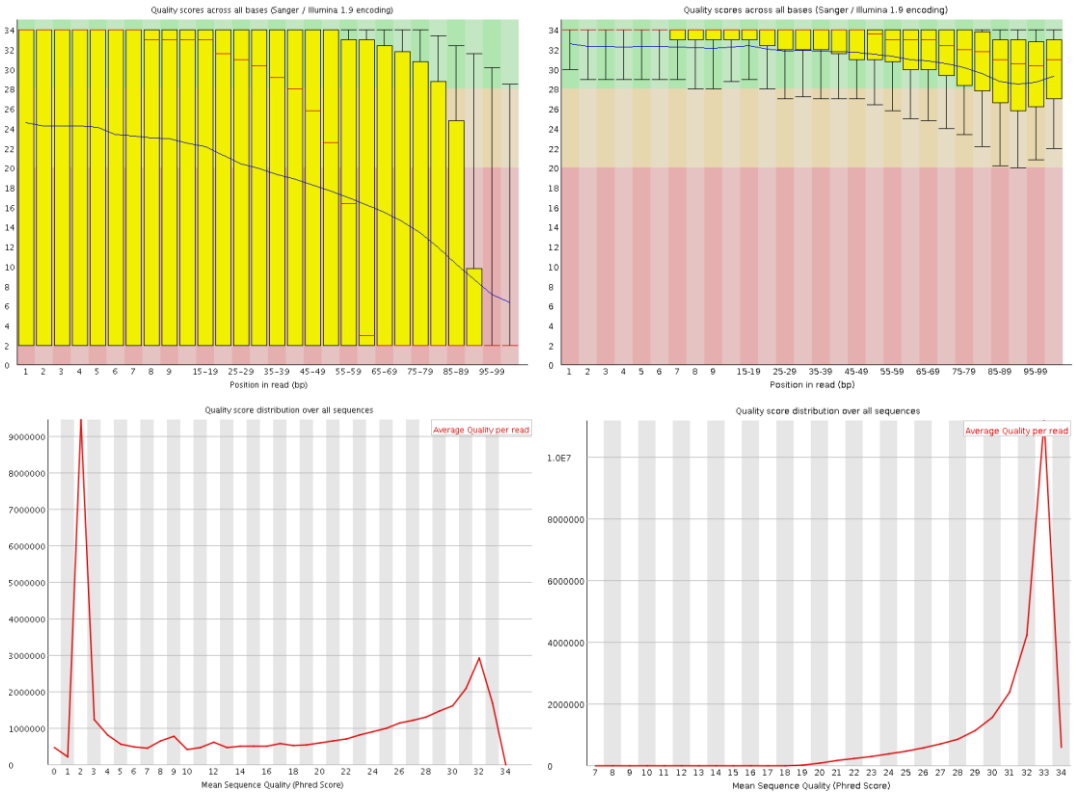
Poor base call qualities or adapter contamination are however just as relevant for 'normal', i.e. non-RRBS, libraries.

Adaptive quality and adapter trimming with `Trim Galore!`

We have tried to implement a method to rid RRBS libraries (or other 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.

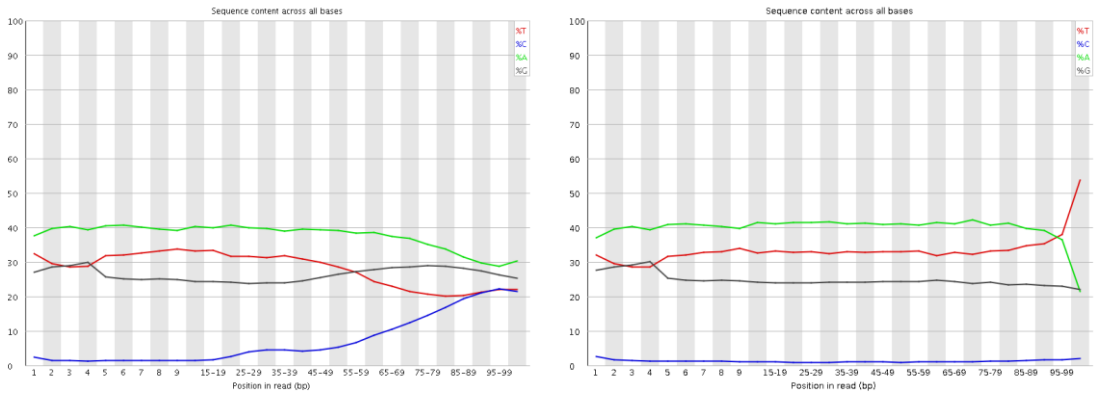
- In the first step, low-quality base calls are trimmed off from the 3' end of the reads before adapter removal. This efficiently removes poor quality portions of the reads. Here is an example of a dataset downloaded from the SRA which was trimmed with a Phred score threshold of 20 (data set DRR001650_1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



before trimming after trimming

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 for DNA material that was digested with MspI. In this mode, `Trim galore!` identifies sequences that were adapter-trimmed and removes another 2 bp from their 3' end. This is to avoid that the filled-in cytosine position 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.

- `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.

  Note that is not recommended to remove too short sequences if the analysed FastQ file is one of a pair of paired-end files since this confuses the sequence-by-sequence order of paired-end reads which is again required by many aligners. For paired-end files, Trim Galore! has an option '--paired' which runs a paired-end validation on both trimmed _1 and _2 FastQ files once the trimming has completed. This step removes entire read pairs if at least one of the two sequences became shorter than a certain threshold. If only one of the two reads is longer than the set threshold, e.g. when one read has very poor qualities throughout, this read can be written out to unpaired files (see option '--retain_unpaired') which may be aligned in a single-end manner.

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.

Full list of options for `Trim galore!`

USAGE:

```
trim_galore [options] <filename(s)>
```

General options:

`-h/--help`	Print this help message and exits.
`-v/--version`	Print the version information and exits.
`-q/--quality <INT>`	Trim low-quality ends from reads in addition to adapter removal. For RRBS samples, quality trimming will be performed first, and adapter trimming is carried in a second round. Other files are quality and adapter trimmed in a single pass. The algorithm is the same as the one used by BWA (Subtract INT from all qualities; compute partial sums from all indices to the end of the sequence; cut sequence at the index at which the sum is minimal). Default Phred score: 20.
`--phred33`	Instructs Cutadapt to use ASCII+33 quality scores as Phred scores (Sanger/Illumina 1.9+ encoding) for quality trimming. Default: ON.
`--phred64`	Instructs Cutadapt to use ASCII+64 quality scores as Phred scores (Illumina 1.5 encoding) for quality trimming.
`--fastqc`	Run FastQC in the default mode on the FastQ file once trimming is complete.
`--fastqc_args "<ARGS>"`	Passes extra arguments to FastQC. If more than one argument is to be passed to FastQC they must be in the form "arg1 arg2 etc.". An example would be: `--fastqc_args "--nogroup --outdir /home/"`. Passing extra arguments will automatically invoke FastQC, so `--fastqc` does not have to be specified separately.
`-a/--adapter <STRING>`	Adapter sequence to be trimmed. If not specified explicitly, the first 13 bp of the Illumina adapter '`AGATCGGAAGAGC`' are used by default.

-a2/--adapter2 <STRING> Optional adapter sequence to be trimmed off read 2 of paired-end files. This option requires '--paired' to be specified as well.

-s/--stringency <INT> Overlap with adapter sequence required to trim a sequence. Defaults to a very stringent setting of '1', i.e. even a single bp of overlapping sequence will be trimmed of the 3' end of any read.

-e <ERROR RATE> Maximum allowed error rate (no. of errors divided by the length of the matching region) (default: 0.1).

--gzip Compress the output file with gzip. If the input files are gzip-compressed the output files will be automatically gzip compressed as well.

--length <INT> Discard reads that became shorter than length INT because of either quality or adapter trimming. A value of '0' effectively disables this behaviour. Default: 20 bp.

For paired-end files, both reads of a read-pair need to be longer than <INT> bp to be printed out to validated paired-end files (see option --paired). If only one read became too short there is the possibility of keeping such unpaired single-end reads (see --retain_unpaired). Default pair-cutoff: 20 bp.

-o/--output_dir <DIR> If specified all output will be written to this directory instead of the current directory.

--no_report_file If specified no report file will be generated.

--suppress_warn If specified any output to STDOUT or STDERR will be suppressed.

RRBS-specific options (MspI digested material):

--rrbs Specifies that the input file was an MspI digested RRBS sample (recognition site: CCGG). Sequences which were adapter-trimmed will have a further 2 bp removed from their 3' end. This is to avoid that the filled-in C 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 further.

--non_directional Selecting this option for non-directional RRBS libraries will screen quality-trimmed sequences for 'CAA' or 'CGA' at the start of the read and, if found, removes the first two basepairs. Like with the

option '--rrbs' this avoids using cytosine positions that were filled-in during the end-repair step. '--non_directional' requires '--rrbs' to be specified as well.

--keep

Keep the quality trimmed intermediate file. Default: off, i.e. the temporary file is being deleted after adapter trimming. Only has an effect for RRBS samples since other FastQ files are not trimmed for poor qualities separately.

Note for RRBS using MseI:

If your DNA material was digested with MseI (recognition motif: TTAA) instead of MspI it is NOT necessary to specify --rrbs or --non_directional since virtually all reads should start with the sequence 'TAA', and this holds true for both directional and non-directional libraries. As the end-repair of 'TAA' restricted sites does not involve any cytosines it does not need to be treated especially. Instead, simply run Trim Galore! in the standard (i.e. non-RRBS) mode.

Paired-end specific options:

--paired

This option performs length trimming of quality/adapter/RRBS trimmed reads for paired-end files. To pass the validation test, both sequences of a sequence pair are required to have a certain minimum length which is governed by the option --length (see above). If only one read passes this length threshold the other read can be rescued (see option --retain_unpaired). Using this option lets you discard too short read pairs without disturbing the sequence-by-sequence order of FastQ files which is required by many aligners.

Trim Galore! expects paired-end files to be supplied in a pairwise fashion, e.g. file1_1.fq file1_2.fq SRR2_1.fq.gz SRR2_2.fq.gz ….

-t/--trim1

Trims 1 bp off every read from its 3' end. This may be needed for FastQ files that are to be aligned as paired-end data with Bowtie. This is because Bowtie (1) regards alignments like this:

```
R1 ---------------------------->
R2 <----------------------------
            or this:
R1 ----------------------->
R2       <-----------------
```

as invalid (whenever a start/end coordinate is contained within the other read).

`--retain_unpaired` If only one of the two paired-end reads became too short, the longer read will be written to either '`.unpaired_1.fq`' or '`.unpaired_2.fq`' output files. The length cutoff for unpaired single-end reads is governed by the parameters `-r1/--length_1` and `-r2/--length_2`. Default: OFF.

`-r1/--length_1 <INT>` Unpaired single-end read length cutoff needed for read 1 to be written to '`.unpaired_1.fq`' output file. These reads may be mapped in single-end mode. Default: 35 bp.

`-r2/--length_2 <INT>` Unpaired single-end read length cutoff needed for read 2 to be written to '`.unpaired_2.fq`' output file. These reads may be mapped in single-end mode. Default: 35 bp.