Babraham Bioinformatics

September 07, 2016

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

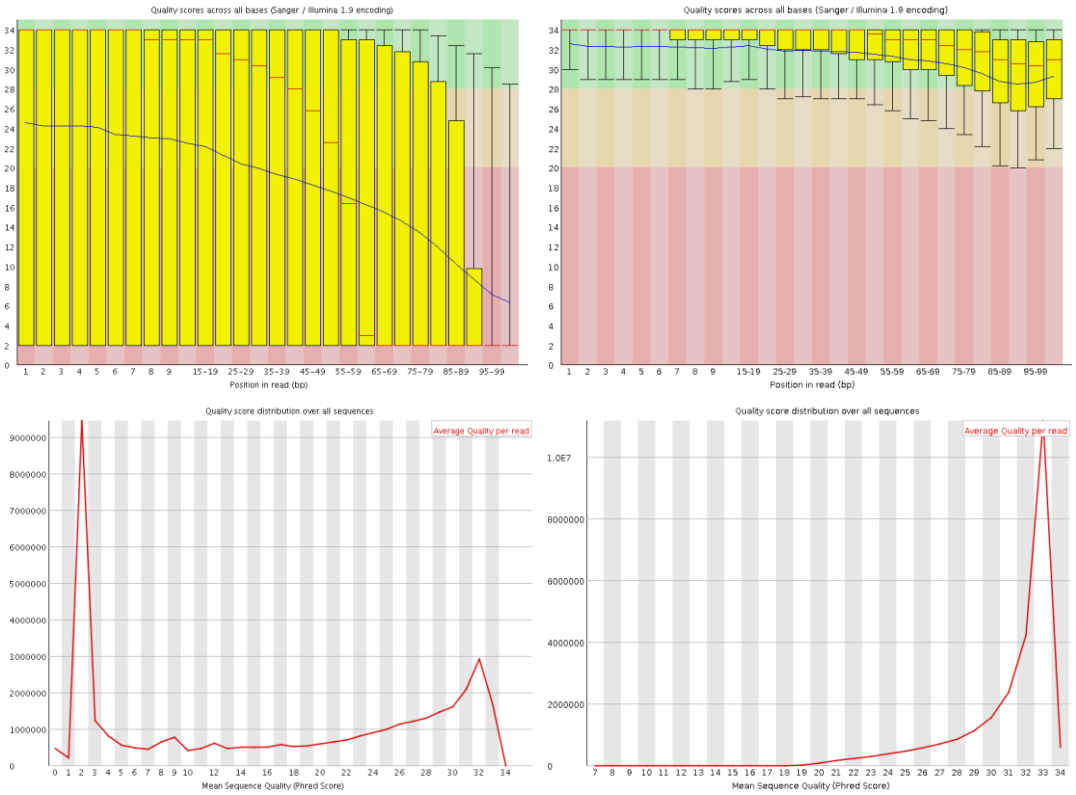
Poor base call qualities or adapter contamination are however just as relevant for 'normal', i.e. non-RRBS, libraries.

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).



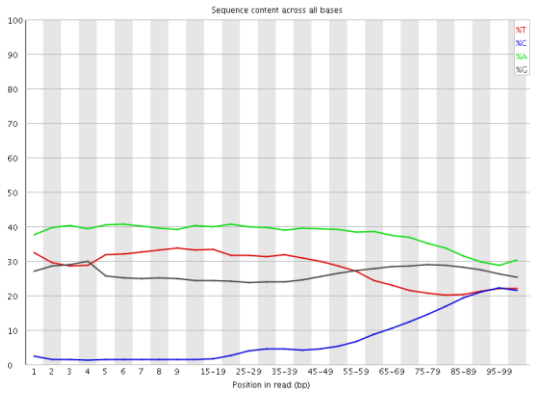
before trimming

after trimming

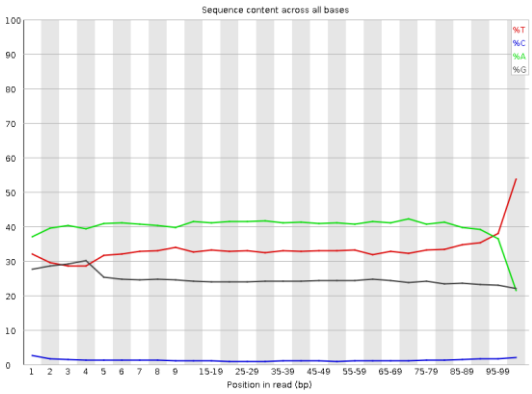
In the next step, Cutadapt finds and removes adapter sequences from the 3' end of reads. If no sequence was supplied it will attempt to auto-detect the adapter which has been used. For this it will analyse the first 1 million sequences of the first specified file and attempt to find the first 12 or 13bp of the following standard adapters:

```
Illumina:    AGATCGGAAGAGC
Small RNA:   TGGAATTCTCGG
Nextera:     CTGTCTCTTATA
```

If no adapter can be detected within the first 1 million sequences Trim Galore defaults to --illumina. The auto-detection behaviour can be overruled by specifying an adapter sequence or using --illumina, --nextera or --small_rna. (Please note the first 13 bp of the standard Illumina paired-end adapters ('AGATCGGAAGAGC') recognise and removes adapter from most standard libraries, including the TruSeq and Sanger iTag adapters). 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 singleton 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, Trim Galore will try to auto-detect whether the Illumina universal, Nextera transposase or Illumina small RNA adapter sequence was used. Also see `'--illumina'`, `'--nextera'` and `'--small_rna'`. If no adapter can be detected within the first 1

million sequences of the first file specified Trim Galore defaults to '--illumina'.

-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.

--illumina Adapter sequence to be trimmed is the first 13bp of the Illumina universal adapter 'AGATCGGAAGAGC' instead of the default auto-detection of adapter sequence.

--nextera Adapter sequence to be trimmed is the first 12bp of the Nextera adapter 'CTGTCTCTTATA' instead of the default auto-detection of adapter sequence.

--small_rna Adapter sequence to be trimmed is the first 12bp of the Illumina Small RNA 3' Adapter 'TGGAATTCTCGG' instead of the default auto-detection of adapter sequence. Selecting to trim smallRNA adapters will also lower the --length value to 18bp. If the smallRNA libraries are paired-end then -a2 will be set to the Illumina small RNA 5' adapter automatically ('GATCGTCGGACT') unless -a 2 had been defined explicitly.

--max_length <INT> Discard reads that are longer than <INT> bp after trimming. This is only advised for smallRNA sequencing to remove non-small RNA sequences.

-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.

--dont_gzip Output files won't be compressed with gzip. This overrides --gzip.

--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.
`--max_n COUNT`	The total number of Ns (as integer) a read may contain before it will be removed altogether. In a paired-end setting, either read exceeding this limit will result in the entire pair being removed from the trimmed output files.
`--trim-n`	Removes Ns from either side of the read. This option does currently not work in RRBS mode.
`-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.
`--clip_R1 <int>`	Instructs Trim Galore to remove `<int>` bp from the 5' end of read 1 (or single-end reads). This may be useful if the qualities were very poor, or if there is some sort of unwanted bias at the 5' end. Default: OFF.
`--clip_R2 <int>`	Instructs Trim Galore to remove `<int>` bp from the 5' end of read 2 (paired-end reads only). This may be useful if the qualities were very poor, or if there is some sort of unwanted bias at the 5' end. For paired-end BS-Seq, it is recommended to remove the first few bp because the end-repair reaction may introduce a bias towards low methylation. Please refer to the M-bias plot section in the Bismark User Guide for some examples. Default: OFF.
`--three_prime_clip_R1 <int>`	Instructs Trim Galore to remove <int> bp from the 3' end of read 1 (or single-end reads) AFTER adapter/quality trimming has been performed. This may remove some unwanted bias from the 3' end that is not directly related to adapter sequence or basecall quality. Default: OFF.
`--three_prime_clip_R2 <int>`	Instructs Trim Galore to remove <int> bp from the 3' end of read 2 AFTER adapter/quality trimming has been performed. This may remove some unwanted bias from the 3' end that is not directly related to adapter sequence or basecall quality. Default: OFF.

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.

Changelog:

07-09-16: Version 0.4.2 released

- Replaced 'zcat' with 'gunzip -c' so that older versions of Mac OSX do not append a .Z to the end of the file and subsequently fail because the file is not present. Dah...
- Added option '--max_n COUNT' to remove all reads (or read pairs) exceeding this limit of tolerated Ns. In a paired-end setting it is sufficient if one read exceeds this limit. Reads (or read pairs) are removed altogether and are not further trimmed or written to the unpaired output
- Enabled option '--trim-n' to remove Ns from both end of the reads. Does currently not work for RRBS-mode
- Added new option '--max_length ' which reads that are longer than bp after trimming. This is only advised for smallRNA sequencing to remove non-small RNA sequences

12-11-15: Version 0.4.1 released: Essential update for smallRNA libraries!

- Changed the Illumina small RNA sequence used for auto-detection to 'TGGAATTCTCGG' (from formerly 'ATGGAATTCTCG'). The reason for this is that smallRNA libraries have ssRNA adapters ligated to their -OH end, a signature of dicer cleavage, so there is no A-tailing involved. Thanks to Q. Gouil for bringing this to our attention
- Changed the length cutoff for sequences to 16bp (down from 20bp) for smallRNA libraries before sequences get removed entirely. This is because some 20-23bp long smallRNAs species that had removed T, TG, or TGG etc. might just about pass the 20bp cutoff
- Added a small description to the --help message for users of the NuGEN Ovation RRBS kit as to *NOT* use the --rrbs option (see --help)

06-05-15: Version 0.4.0 released

- Unless instructed otherwise Trim Galore will now attempt to auto-detect the adapter which had been used for library construction (choosing from the Illumina universal, Nextera transposase and Illumina small RNA adapters). For this the first 1 million sequences of the first file specified are analysed. If no adapter can be detected within the first 1 million sequences Trim Galore defaults to --illumina. The auto-detection behaviour can be overruled by specifying an adapter sequence or using --illumina, --nextera or --small_rna
- Added the new options '--illumina', '--nextera' and '--small_rna' to use different default sequences for trimming (instead of -a): Illumina: AGATCGGAAGAGC; Small RNA: TGGAATTCTCGG; Nextera: CTGTCTCTTATA
- Added a sanity check to the start of a Trim Galore run to see if the (first) FastQ file in question does contain information at all or appears to be in SOLiD colorspace format, and bails if either is true. Trim Galore does not support colorspace trimming, but users wishing to do this are kindly referred to using Cutadapt as a standalone program
- Added a new option '--path_to_cutadapt /path/to/cudapt'. Unless this option is specified it is assumed that Cutadapt is in the PATH (equivalent to '--path_to_cutadapt cutadapt'). Also added a test to

see if Cutadapt seems to be working before the actual trimming is launched
- Fixed an open command for a certain type of RRBS processing (was open() instead of open3())

16-07-14: Version 0.3.7 released

- Applied small change that makes paired-end mode work again (it was accidentally broken by changing @ARGV for @filenames when looping through the filenames...)

11-07-14: Version 0.3.6 released

- Added the new options '--three_prime_clip_r1' and '--three_prime_clip_r2' to clip any number of bases from the 3' end after adapter/quality trimming has completed
- Added a check to see if Cutadapt exits fine. Else, Trim Galore will bail a well
- The option '--stringency' needs to be spelled out now since using -s was ambiguous because of '--suppress_warn'

late 2013: Version 0.3.5 released

- Added the Trim Galore version number to the summary report

19-09-13: Version 0.3.4 released

- Added single-end or paired-end mode to the summary report
- In paired-end mode, the Read 1 summary report will no longer state that no sequence have been discarded due to trimming. This will be stated in the trimming report of Read 2 once the validation step has been completed

10-09-13: Version 0.3.3 released

- Fixed a bug what was accidentally introduced which would add an additional empty line in single-end trimming mode

03-09-13: Version 0.3.2 released

- Specifying --clip_R1 or --clip_R2 will no longer attempt to clip sequences that have been adapter- or quality-trimmed below the clipping threshold
- Specifying an output directory with --rrbs mode should now correctly create temporary files

15-07-13: Version 0.3.1 released

- The default length cutoff is now set at an earlier timepoint to avoid a clash in paired-end mode when '--retain_unpaired' and individual read lengths for read 1 and read 2 had been defined

15-07-13: Version 0.3.0 released

- Added the options '--clip_R1' and '--clip_R2' to trim off a fixed amount of bases at from the 5' end of reads. This can be useful if the quality is unusually low at the start, or whenever there is an undesired bias

at the start of reads. An example for this could be PBAT-Seq in general, or the start of read 2 for every bisulfite-Seq paired-end library where end repair procedure introduces unmethylated cytosines. For more information on this see the M-bias section of the Bismark User Guide

10-04-13: Version 0.2.8 released

- Trim Galore will now compress output files with GZIP on the fly instead of compressing the trimmed file once trimming has completed. In the interest of time temporary files are not being compressed
- Added a small sanity check to exit if no files were supplied for trimming. Thanks to P. for 'bringing this to my attention'

01-03-13: Version 0.2.7 released

- Added a new option '--dont_gzip' that will force the output files not to be gzip compressed. This overrides both the '--gzip' option or a .gz line ending of the input file(s)

07-02-13: Version 0.2.6 released

- Fixes some bugs which would not gzip or run FastQC correctly when the option '-o' had been specified
- When '--fastqc' is specified in paired-end mode the intermediate files '_trimmed.fq' are no longer analysed (only files '_val_1' and '_val_2')

19-10-12: Version 0.2.5 released

- Added option '-o/--output_directory' to redirect all output (including temporary files) to another folder (required for implementation into Galaxy)
- Added option '--no_report_file' to suppress a report file
- Added option '--suppress_warn' to suppress any output to STDOUT or STDERR

02-10-12: Version 0.2.4 released

- Removed the shorthand '-l' from the description as it might conflict with the paired-end options '-r1/--length1' or '-r2/--length2'. Please use '--length' instead
- Changed the reporting to show the true Phred score quality cutoff
- Corrected a typo in stringency...

31-07-12: Version 0.2.3 released

- Added an option '-e ERROR RATE' that allows one to specify the maximum error rate for trimming manually (the default is 0.1)

09-05-12: Version 0.2.2 released

- Added an option '-a2/--adapter2' so that one can specify individual adapter sequences for the two reads of paired-end files; hereby the sequence provided as '-a/--adapter' is used to trim read 1, and the sequence provided as '-a2/--adapter2' is used to trim read 2. This option requires '--paired' to be specified as well

20-04-12: Version 0.2.1 released

- Trim Galore! now has an option '--paired' which has the same functionality as the validate_paired_ends script we offered previously. This option discards read pairs if one (or both) reads of a read pair became shorter than a given length cutoff
- Reads of a read-pair that are longer than a given threshold but for which the partner read has become too short can optionally be written out to single-end files. This ensures that the information of a read pair is not lost entirely if only one read is of good quality
- Paired-end reads may be truncated by a further 1 bp from their 3' end to avoid problems with invalid alignments with Bowtie 1 (which regards alignments that contain each other as invalid...)
- The output may be gzip compressed (this happens automatically if the input files were gzipped (i.e. end in .gz))
- The documentation was extended substantially. We also added some recommendations for RRBS libraries for MseI digested material (recognition motif TTAA)

21-03-12: Version 0.1.4 released

- Phred33 (Sanger) encoding is now the default quality scheme
- Fixed a bug for Phred64 encoding that would occur if several files were specified at once

14-03-12: Version 0.1.3 released

- Initial stand-alone release; all basic functions working
- Added the option 'fastqc_args' to pass extra options to FastQC