trim_galore-0.4.1 using DE

The DE Quick Start tutorial provides an introduction to basic DE functionality and navigation.
Please work through the tutorial and add your comments on the bottom of this page, or send comments per email to upendra@cyverse.org. Thank you.

Rationale and background:

Trim Galore is an app to automate quality and adapter trimming as well as quality control, with some added functionality to remove biased methylation positions for RRBS sequence files (for directional, non-directional (or paired-end) sequencing). It's main features are:

- For adapter trimming, Trim Galore! uses the first 13 bp of Illumina standard adapters ('AGATCGGAAGAGC') by default (suitable for both ends of paired-end libraries), but accepts other adapter sequence too.
- For MspI-digested RRBS libraries, Trim Galore! performs quality and adapter trimming in two subsequent steps. This allows it to remove 2 additional bases that contain a cytosine which was artificially introduced in the end-repair step during the library preparation.
- For any kind of FastQ file other than MspI-digested RRBS, Trim Galore! can perform single-pass adapter- and quality trimming.
- The Phred quality of basecalls and the stringency for adapter removal can be specified individually.
- Trim Galore! can remove sequences if they become too short during the trimming process. For paired-end files Trim Galore! removes entire sequence pairs if one (or both) of the two reads became shorter than the set 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.
- Trim Galore! can trim paired-end files by 1 additional bp from the 3' end of all reads to avoid problems with invalid alignments with Bowtie 1.
- Trim Galore! accepts and produces standard or gzip compressed FastQ files.
- FastQC can be run on the resulting output files once trimming has completed (optional).

Trim_galore makes use of the publicly 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.

Prerequisites

1. A CyVerse account. (Register for an CyVerse account here - user.cyverse.org.)
2. Input
   a. Sequence file in fastq format (either single end or paired end reads).
   b. Paired-end specific options:
      i. 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).
      ii. Retain unpaired reads (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).
      iii. Unpaired single-end read length cut-off for read 1 (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).
      iv. Unpaired single-end read length cut-off for read 2 (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).
      v. Trim 1bp from 3'end (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).

3. Parameters
   a. quality (Default Phred score: 20).
   b. phred33 (Sanger/Illumina 1.9+ encoding).
   c. phred64 (Illumina 1.5 encoding).
   d. fastqc (Run FastQC in the default mode on the FastQ file once trimming is complete).
   e. Adapter sequence to be trimmed (If not specified explicitly, the first 13 bp of the Illumina adapter 'AGATCGGAAGAGC' are used by default).
   f. adapter2 (Optional adapter sequence to be trimmed off read 2 of paired end files. This option requires '--paired' to be specified as well).
   g. stringency (Overlap with adapter sequence required to trim a sequence (very stringent setting of '1', i.e. even a single bp of overlapping sequence will be trimmed of the 3' end of any read).
   h. Maximum allowed error rate (no. of errors divided by the length of the matching region) (default: 0.1).
   i. Length (Discard reads that became shorter than length INT because of either quality or adapter trimming. A value of '0' effectively disables this behavior. Default: 20 bp). For paired-end files, both reads of a read-pair need to be longer than 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.
   j. Clip R1 (Instructs Trim Galore to remove 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.
   k. Clip R2 Instructs Trim Galore to remove 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.

l. 3' Clip R1 (Instructs Trim Galore to remove 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).
m. 3' Clip R2 (Instructs Trim Galore to remove 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).
n. RRBS-specific options (MspI digested material):
    i. 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).
    ii. 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.
    iii. 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.

Test/sample data:

The test data are provided for testing trim_galore-0.4.1 is in here - /iplant/home/shared/iplantcollaborative/example_data/trim_galore:

Use the following inputs/outputs and parameters for testing trim_galore-0.4.1

1. Input/Outputs
    a. ATreads.fq
    b. ATreads2.fq
2. Optional arguments/parameters
    a. Paired (select Paired check box)
    b. Quality type - sanger
    c. fastqc (select fastqc box)

    Leave the rest of the options as default

Output Reports:

After successful completion of the run, expect the following files as output:

1. ATreads_unpaired_1.fq
2. ATreads_val_1_fastqc.html
3. ATreads_val_1_fastqc.zip
4. ATreads_val_1.fq
5. ATreads.fq_trimming_report.txt
6. ATreads2_unpaired_2.fq
7. ATreads2_val_2_fastqc.html
8. ATreads2_val_2_fastqc.zip
9. ATreads2_val_2.fq
10. ATreads2.fq_trimming_report.txt

More information about Scythe-0.991 can be found at trim-galore manual.