Trimming Adapter Sequences From MiRna Reads

Peter von Rohr

15 January 2016

Disclaimer

This document summarizes information about how trimming is done for small RNA sequences.

Current Status

When analysing small RNA sequences with "EzAppNcpro", trimming is currently done using a function called "trimMirna()" which is stored in file "ngsMirna.R".

In what follows a short summary on trimming is given.

Function trimMirna

In function "trimMirna()" trimming is done using two programs.

1. prinseq-lite
2. flexbar

The program `prinseq-lite` runs first by pasting together the following system-command.

```
gunzip -c /home/petervr/myRepo/ezRun/inst/extdata/smRNA_250k/test1_R1.fastq.gz | \
  /usr/local/ngseq/stow/prinseq-lite-0.20.3/bin/prinseq-lite.pl                  \
    -no_qual_header               \  # empty header line for quality data in fastq files
    -trim_qual_right 20           \  # trim sequences by quality score from 3'-end with given threshold
    -trim_qual_type mean          \  # type of how quality score should be computed
    -trim_qual_window 4           \  # sliding window size used to compute quality score
    -fastq stdin                  \
    -out_bad test1_R1_qualtrimBad \
    -out_good test1_R1_qualtrim   \
    > /srv/local/scratch/PVR_test/ncPRO_Result/test1_R1.prinseq.out
```

Program prinseq-lite

Reference for prinseq-lite is available under http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3051327/. As seen from the above command, prinseq-lite does not do any trimming of the adapter sequence, because the adapter is nowhere specified. What prinseq-lite does, is to filter reads based on a given function or statistic (such as mean, min or max) of the quality score. Which type of statistic should be used is specified by parameter "-trim_qual_type". The statistic is computed on a sliding window the width of which is specified by the parameter "-trim_qual_window". Reads which are trimmed are written to the file indicated by the parameter "-out_bad", whereas the retained reads are written to a file given by the parameter "-out_good". The output caught from STDOUT was empty in all tests done so far.

Programm flexbar

The following command that is run after prinseq is the following which is using a program called "flexbar". The reference manual for flexbar is available at https://github.com/seqan/flexbar/wiki/Manual. The input of "flexbar" which is read from a file given by the parameter "–reads" corresponds to the reads that were retained after trimming with "prinseq-lite". The output after removing the adapter sequences are written to the file indicated by the option "–target"

```
/usr/local/ngseq/bin/flexbar      \
  --adapter-seq TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC \
  --adapter-trim-end RIGHT        \   # type of removal, left part of sequence remains
  --adapter-min-overlap 10        \   # minimum overlap between adapter and read
  --adapter-threshold 1.5         \   # allowed gaps and mismatch for removal
  --min-read-length 18            \   # minimum read length after removal for read to stay
  --max-uncalled 4                \   # allowed number of uncalled bases or N's per read
  --format i1.8                   \   # quality format (Illumina 1.8)
  --reads test1_R1_qualtrim.fastq \   # fasta/q files with reads that may contain barcodes/adapters
  --target test1_R1_allTrimmed    \   # prefix for output file names or paths
  > /srv/local/scratch/PVR_test/ncPRO_Result/test1_R1.flexbar.out
```

The output caught from STDOUT prints a summary and some count statistics from the processed read data.

New Trimming

Trimming as used in function "ncpro()" should be changed to use function "ezMethodTrim()" in "appTrim.R". Parameters of "ezMethodTrim()" are

```
  paired            a logical specifying whether the samples have paired ends.
  subsampleReads    an integer specifying how many subsamples there are. This will call
                       \code{ezMethodSubsampleReads()} if > 1.
  trimAdapter       a logical specifying whether to use a trim adapter.
  minTailQuality    an integer specifying the minimal tail quality to accept. Only used if > 0.
  minAvgQuality     an integer specifying the minimal average quality to accept. Only used if > 0.
  minReadLength     an integer specifying the minimal read length to accept.
  dataRoot          a character specifying the path of the data root to get the full column paths from.
```

The functions "trimMirna" and "ezMethodTrim()" are not called with the same parameters. The former has parameters "input", "output" and "param" where the first is the input filename, the second is the name of the output file and the last is a list of parameters. The arguments of "ezMethodTrim" have the same name, but are expected to be instances of of the reference class "EzDataset".

Function ezMethodTrim()

Function parameters of function `ezMethodTrim()` are

- input
- output
- param

The first two parameters (input and output) are expected to be instances of the reference class "EzDataset". The function parameter `param` is expected to be a list with components that are listed above as parameters of "ezMethodTrim()".

Comparison

The two trimming routines “trimMirna” und “ezMethodTrim” will be compared by running them on the same dataset and comparing the results after both trimming runs.

These tests are done on the same test data sets as were already used for updating “ncpro”.

Reference Class EzTrimTester

To make the comparison as easy as possible, a reference class called “EzTrimTester” was created. This does all the preparation and has as runMethod a function that is the same as the one for the ncpro app until the trimming is done. As a result the following files are generated.

```
rw-rw-r-- 1 petervr SG_Users 26292548 Jan 15 16:18 test1_R1.fastq
-rw-rw-r-- 1 petervr SG_Users     1688 Jan 15 16:18 test1_R1.flexbar.out
-rw-rw-r-- 1 petervr SG_Users        0 Jan 15 16:18 test1_R1.prinseq.out
-rw-rw-r-- 1 petervr SG_Users 25323095 Jan 15 16:19 test2_R1.fastq
-rw-rw-r-- 1 petervr SG_Users     1688 Jan 15 16:19 test2_R1.flexbar.out
-rw-rw-r-- 1 petervr SG_Users        0 Jan 15 16:18 test2_R1.prinseq.out
```

where the files with extension .flexbar.out are the output of the flexbar program and the files ending in .fastq are the trimmed readfiles.