GlbPSs 1.0 Manual Part 3: Documentation of phred_pos_filter_07.0.pl

5/22/2015 Andreas Hapke,
Institute of Anthropology, Johannes Gutenberg University Mainz, Germany, ahapke2@gmail.com

Fastq files from a next-generation sequencing run may contain a fraction of sequences with N or low quality base calls at specific sequence positions where other sequences have high quality base calls. Such systematic error arises from technical problems (bubbles) in specific parts of the flow cell during specific cycles. You can use **phredi** to analyze your fastq files and to identify systematic error at certain sequence positions. The program **phred_pos_filter** then enables you to remove the affected sequences from the dataset with custom filters. It can handle single or paired fastq infiles (and even a mixture of both). The fastq infiles may be plain text files or gzip compressed files. The program produces filtered fastq outfiles as plain text files or gzip compressed files.

Usage

Store your fastq-infiles in one or several directories anywhere on your computer. Create a directory and store a copy of the program and additional files (see below) therein. Open a command prompt, cd to this directory, and call the program:

```
phred_pos_filter_07.0.pl –i myfqinfiles.txt –o myoutfiles –f myfilters.txt
```

or

```
perl phred_pos_filter_07.0.pl –i myfqinfiles.txt –o myoutfiles –f
myfilters.txt
```

Several optional command flags enable you to make settings:

```
-i name of text file with fastq infilenames      default: fq_in.txt
-o name of outfile directory                     default: pf_out
-f name of text file with filters                default: filter.txt
-p Phred score offset                            default: 33 (Sanger)
-P maximum number of processes in parallel       default: 1
-z 1: produce gzip compressed outfiles, 0: don't default: 0
```

Additional files: `-i –f`

Provide filenames when these files are stored in the same directory as the program. Provide pathes when not.

```
-i fq_in.txt
```

This file contains the names of your fastq infiles. Provide pathes, when they are not stored in the same directory as the program. When you have single fastq infiles, provide one name or path per line. When you have paired fastq infiles, provide both names/pathes of a pair in one line separated by any number of whitespace characters. See the example files `fq_in_pair.txt` and `fq_in_single.txt`. The program assumes that the first file contains forward reads and the second (if any) reverse reads. Names of gzip compressed fastq infiles must have the extension `".gz"`. Plain text files may have any other or no extension. Each gzip compressed file must contain one single fastq file. The program is not able to unpack a tarball.

Format of fastq infiles: four lines per sequence entry: header, sequence, third line, Phred symbols, no additional blank lines. Paired reads must be in corresponding order in paired files. All forward reads and all reverse reads should have the same length. Forward and reverse reads may have different lengths.

`-o pf_out`

The program creates a subdirectory `pf_out` in the directory where it resides and stores outfiles therein. You may provide another name with this flag. The program issues an error message and stops execution when this directory already exists.

`-f filter.txt`

This text file contains your filter-settings. It must contain three values per line, separated by any number of whitespace characters:

```
f   75   10
f   76   10
f   77   10
r   20   20
r   21   20
```

The first value specifies a filter for forward (f) or reverse (r) reads. (Use `f` when you have single fastq infiles.) The next two values specify a sequence position and a minimum acceptable Phred score. Make sure that all positions for forward/reverse reads truly exist in every forward/reverse read. **The program will behave unexpectedly when it cannot verify a position because the sequence is too short.** The program treats a sequence as bad and discards it when the Phred score at any specified position is smaller than the respective minimum acceptable Phred score. When you have paired reads, you may apply filters to forward reads or reverse reads or both. The program discards both reads of a pair when any of them is bad.

`-p Phred score offset`

Per default, the program uses 33 for Sanger. Provide another appropriate value if necessary.

`-P maximum number of processes in parallel`

The program can process several fastq infiles or pairs of fastq infiles in parallel. The limiting factor for speed may be the hard drive though – test this variable carefully.

`-z produce gzip compressed outfiles`

Use `-z 1` to produce gzip compressed fastq outfiles and `-z 0` to produce plain text files. Writing of plain text files is faster. Writing of gzip compressed files saves a lot of disk space.

Outfiles

The program stores all outfiles in the specified outfile directory. During the run, it produces several temporary outfiles, which it later deletes again. Do not manipulate files in the outfile directory before the run is finished. The program produces a report file and one fastq outfile per fastq infile. The report file `pf_rep.txt` contains a tab-delimited table with one line per fastq infile or pair of fastq infiles:

`f_infile:`	Name of the forward-fastq infile (or single fastq infile).
`n_seq:`	number of sequences or sequence pairs
`n_badseq:`	number of bad sequences or pairs
`prop_badseq:`	proportion of bad sequences or pairs
`n_goodseq:`	number of good sequences or pairs

The program stores good sequences or pairs of good sequences in fastq outfiles in the same order as in the infiles. The names of fastq outfiles correspond to the fastq infiles with a slight modification: The program removes the file extension and then appends `_pf.fq` to the filenames of plain text files and `_pf.fq.gz` to those of gzip compressed files. It verifies if paired outfiles have different names after this operation. If not, it adds some characters to the beginning of the filenames to make them different:

```
R1_0_*      R2_0_*
R1_1_*      R2_1_*
...
```