GIbPSs 1.0 Manual Part 2: Documentation of phredi_08.0.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 certain 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. This program analyzes one or several fastq files and determines distributions of Phred scores at each sequence position. The fastq infiles may be plain text files or gzip compressed files. The output enables you to identify such problematic sequence positions. The next program, **phred_pos_filter**, enables you to filter the affected sequences out of your data before you analyze them with **fdm, indloc** and the following programs of GIbPSs. **fdm** and **indloc** can handle sequences that contain N and erroneous base calls. Systematic error at specific sequence positions however can pose problems for the SNP calling algorithms in **indloc**. You should thus use **phredi** to analyze your sequence data and filter them with **phred_pos_filter** if necessary.

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 an additional file (see below) therein. Open a command prompt, cd to this directory, and call the program:

```
phredi_08.0.pl –i myfqinfiles.txt –o myoutfiles
```

or

```
perl phredi_08.0.pl –i myfqinfiles.txt –o myoutfiles
```

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: phredi_out
-p Phred score offset                            default: 33 (Sanger)
-P maximum number of processes in parallel       default: 1
```

```
-i fq_in.txt
```

Provide a filename when this file is stored in the same directory as the program. Provide a full path if not. This file contains the names of your fastq infiles, one per line. Provide pathes, when they are not stored in the same directory as the program. 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.

The fastq infiles must have this format: 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 sequences in one file should have the same length.

```
-o phredi_out
```

The program creates a subdirectory `phredi_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.

```
-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.

Outfiles

The program produces one outfile per sequence file. It contains a tab-delimited table with one line per sequence position and one column for each Phred score. The values are percentages of sequences with a given Phred score at a given position. You can use your favorite spreadsheet program to visualize the data (Figs. 1 and 2).

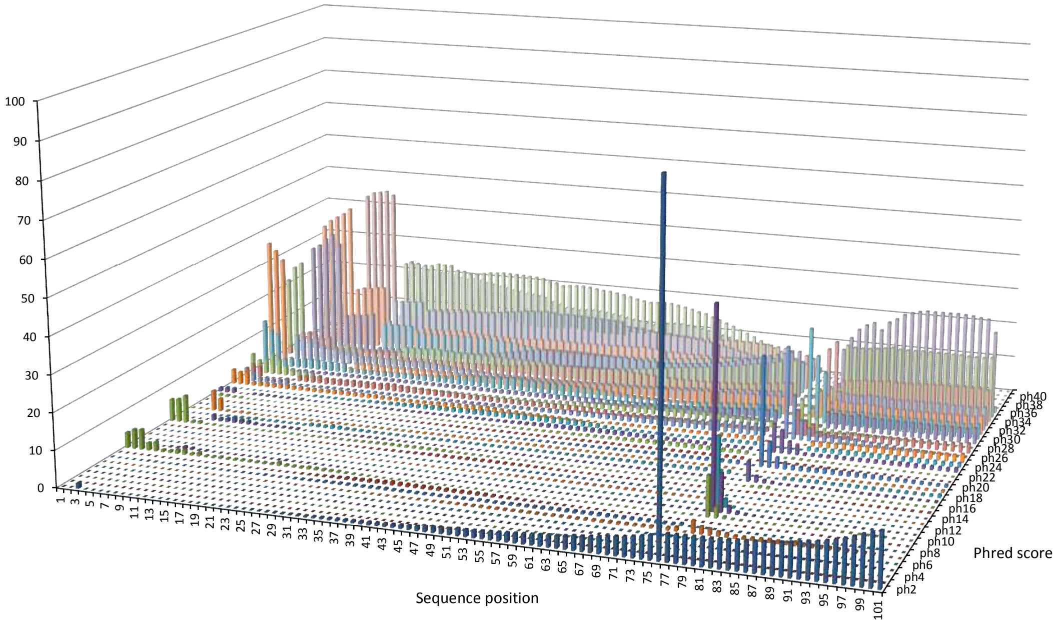

Fig. 1: Example output from phredi: A great proportion of sequences in this file has N or low quality base calls at position 76 and a few following positions. After having inspected these results, I have filtered the sequence data with **phred_pos_filter**. I used a filter that removed all sequences that have a Phred score below 10 at any position from 71 to 81.

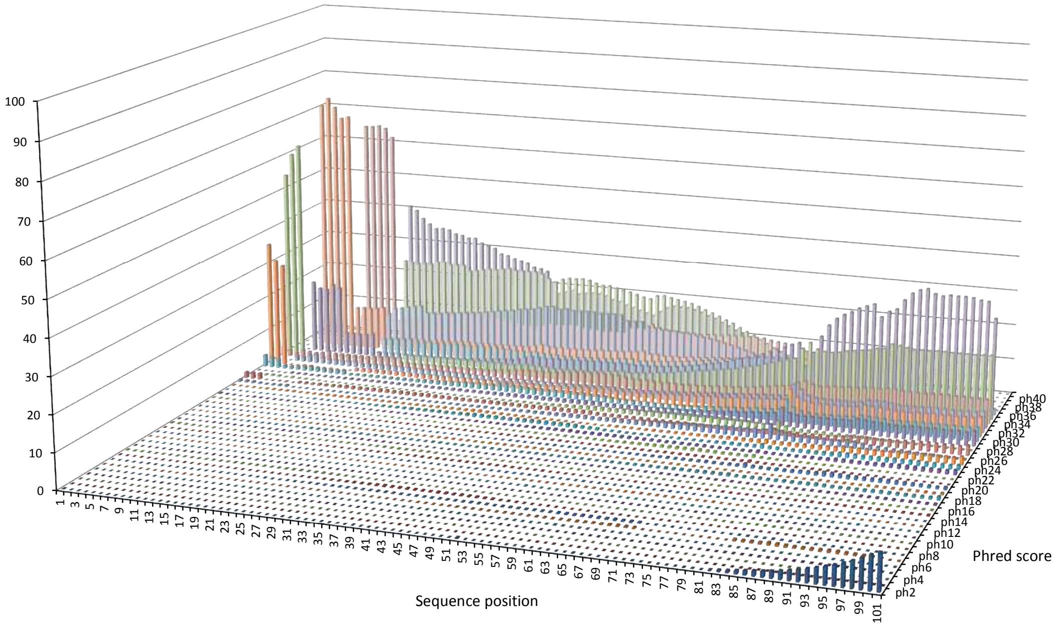

Fig. 2: Example output from phredi: The filtered sequence data reanalyzed with **phredi**.