FASTQ format explained

2017-05-14

This post is part 2 of a series on file formats, written for the 2017 UK-KBRIN Essentials of Next Generation Sequencing Workshop at the University of Kentucky. The conference website is hosted [here](http://www.endophyte.uky.edu/ngs/).

# FastQ sequence format

FASTQ was originally developed by the Wellcome Trust Sanger Institute to bind together FASTA sequences with their respective quality data. It is now the standard for high-throughput sequencing output.

## The format

FASTQ is a four-line per sequence format. If it looks like the raw sequence of your read takes up more than four lines, you probably have word-wrapping enabled.

@SEQ\_ID  
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT  
+  
!''\*((((\*\*\*+))%%%++)(%%%%).1\*\*\*-+\*''))\*\*55CCF>>>>>>CCCCCCC65

### Line 1

@SEQ\_ID is analogous to the >SEQ\_ID line of a FASTA file. Like a FASTA title line, everything before the space is the sequence *identifier* and everything after the first space is the optional sequence *description*.

Note that different sequence sources have different standards for the description line here. Sequences off of an illumina machine will follow this format:

@Instrument\_name:flow\_cell\_lane:tile\_number\_in\_flowcell:x-coordinate:y-coordinate#index/pair1

So for example,

@myInstrument:1:3:100:200#2/2 describes a read off of myInstrument. The read was on tile number 3 of flowcell lane 1, located at X=100 Y=200, index number 2, pair #2 of a paired-end read.

Note that these standards do change: for example, in Casava 1.8 extra information is added and the description is as follows:

@Instrument\_name:runID:flow\_cell\_ID:flow\_cell\_lane:tile\_number\_in\_flowcell:x-coordinate:y-coordinate pair:filtered/not\_filtered:index\_sequence.

If you need to examine the description line of your reads, make sure you know what format is being used.

### Line 2

Line 2 is the actual sequence.

### Line 3

Line 3 begins with the + character followed by the sequence identifier (and description). This identifeir and description is optional: the + is not.

### Line 4

Line 4 contains the quality values for the sequence detailed in line 2. This line must therefore contain the exact same number of characters as line 2: each base has a corresponding sequence score.

Ordered from lowest quality (!) to highest quality (~), these are the quality score characters

!"#$%&'()\*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^\_`abcdefghijklmnopqrstuvwxyz{|}~