Changes to FASTQ Files Introduced by Illumina in CASAVA v1.8

File Naming Conventions

Illumina FASTQ files use the following naming scheme:

`<sample_name>_<barcode_sequence>_L<lane>_R<read_number>_<set_number>.fastq.gz`

Element	Requirements	Description
<sample_name>	Characters allowed: a-z, A-Z, 0-9, dash and underscore	Name of sample as provided by submitter.
<barcode_sequence>	ATGC	Barcode sequence associated with the sample
<lane>	Numerical	Lane number, preceded by an 'L' and 0-padded to three digits
<read>	1 or 2	Read number, preceded by an 'R'. (2 only found in paired runs.)
<set_number>	Numerical	If the read set is divided into multiple files, unique set numbers are used, 0-padded to three digits.

For example, the following is a valid FASTQ file name:

`NA10831_ATCACG_L002_R1_001.fastq.gz`

For a multiplexed run there may be one additional file for each lane containing reads with undetermined indexes. This file will have "lanex" (x=lane number) in place of <sample_name> and "Undetermined" in place of <barcode_sequence>. In the case of non-multiplexed samples, <barcode_sequence> will be replaced with "NoIndex".

Quality Score Encoding

The quality score encoding now conforms to the Sanger standard of ASCII+33. As with versions from 1.5 and up the characters associated with values of 0 (!) and 1 (") are never reported and the character associated with a value of 2 (#) is used as the read segment quality control indicator.

Sequence Identifier Format

Each entry in a FASTQ file consists of four lines:

- Sequence identifier
- Sequence
- Quality score identifier line (consisting of a +)[1]
- Quality score

Each sequence identifier, the line that precedes the sequence and describes it, needs to be in the following format:

```
@<instrument>:<run_number>:<flowcell_ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is_filtered>:<control_number>:<index_sequence>
```

Note the space between the y-pos and read elements. This space formally separates the two portions of the header line into the sequence identifier and description.

Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a-z, A-Z, 0-9, dash and underscore	Instrument ID
<run_number>	Numerical	Run number on instrument
<flowcell_ID>	Characters allowed: a-z, A-Z, 0-9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
space	space	Separates identifier from description portions of header.
<read>	Numerical	Read number. 1 can be single read or read 2 of paired-end
<is_filtered>[2]	Y or N	Y if the read is filtered, N otherwise
<control_number>[3]	Numerical	0 when none of the control bits are on, otherwise it is an even number
<index_sequence>[4]	ACTG	Index sequence

An example of a valid entry is as follows:

```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@
```

Notes:

1. The Quality score identifier line is intentionally left blank, other than the required “+” at the start of the line. This is allowed by the FASTQ format definition. As the format definition required this line to exactly match the sequence identifier/description it was redundant information.
2. The meaning of the filtering status indicator has been changed from “has this read passed filtering” to “has this read failed filtering”. Therefore the good reads will have an “N” here. Normally this is a non-issue for researchers as you will only be supplied with passed filter reads (that is, they will all have “N” here).
3. Control number applies to control DNA species, which may be spiked into the library before sequencing. These are not normally used and even if used would be filtered from the output before distributing data to the researcher. You should only see “0” here.
4. This is the sequence observed for the index read for this cluster. If mismatches are allowed during demultiplexing it may differ from the canonical barcode for the sample. A maximum of 1 mismatch may be allowed. This field will be blank for reads with an undetermined index sequence or for non-multiplexed samples.