9.1 FASTQ Files (NGS Sequence Format)

Fastq files were create to record sequence and quality information within a text file.

FASTQ files consist of 4 lines per record:

1. Sequence Identifier with additional information
2. Base Sequences (ACGTN)
3. A Separator (+) - originally this was a copy of the first line
4. Base Quality Score encoding - Uses ASCII characters to encode PHRED quality.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Illumina Sequence Identifiers

The first line of the Illumina FASTQ sequence provide additional information about the which center, machine, flowcell, lane, read, barcode, and additional cluster information.

Entry	Data	Description
1	Instrument Name	Unique Name of Instrument
2	Run ID	Run number on the machine
3	Flowcell Barcode	Unique Flowcell Barcode
4	Flowcell Lane	Flowcell Lane Number
5	Tile Number	Tile number within the Flowcell
6	X-coord	X-Coordinate of the cluster tile
7	Y-coord	Y Coordinate of the cluster tile
8	Pair Member	Pair number for paired-end or mate-pair only
9	Filter Status	Y if read was filtered and didn't PASS, N otherwise
10	Control Bits Set	See Illumina Documentation
11	Index Sequence	Sample DNA Barcode used for Multiplexing

Template

```
@<Instrument Name>:<Run ID>:<Flowcell Barcode>:<Flowcell Lane>:<Tile Number>:<X-coord>:<Y-coord> <Pair Member>:<Filter Status>:<Control
```

Actual Header

```
@A00740:683:HY23VDSX5:1:1101:5113:1016 1:N:0:TAGGATGA+CGAGTAAT
```

Parsing the actual header:

1. **Instrument Name:** A00740
2. **Run ID:** 683
3. **Flowcell Barcode:** HY23VDSX5
4. **Flowcell Lane:** 1
5. **Tile Number:** 1101
6. **X-Coord:** 5113
7. **Y-Coord:** 1016
8. **Pair Member:** 1
9. **Filter Status:** N
10. **Control Bit Set:** 0

11. **Index Sequence:** TAGGATGA+CGAGTAAT

Why do we care about any of this?

This basic information is used within downstream pipelines to apply the proper operation of your pipeline, QC and processing of the sequences.

FASTQ -> BAM (Raw Sequence to Mapped Reads)

When mapping and processing raw sequences reads, we don't actually have any information about where those reads map within the reference genome. The Variant calling pipelines uses **Read Group** information, which is partially constructed from the FASTQ header, to determine what processes are necessary and in which order.

The **Read Group** information is embedded into the alignment files. There are multiple parts of a **Read Group** that are integral to the alignment/map file.

RG	Description	Required
ID	Unique Identier for the Reads `<Flowcell Barcode>.<Lane>`	Yes
PU	Platform Unit `<Flowcell Barcode>.<Lane>.<Sample Barcode>`	No
SM	Sample Name or Identifier	Yes
PL	Platform Technology (Illumina/PACBIO/...)	Yes
LB	Library - Identify which library sample came from	Yes
BC	Barcode Sequence identifying sample or library	No
CN	Name of Sequencing Center	No
DS	Description	No
DT	Date of sequencing	No
FO	Flow Order	No
KS	Array of nucleotide bases corresponding to the key sequence of each read	No
PG	Program used to process the read group	No
PI	Predicted Median Insert Size	No
PM	Platform Model	No

- FASTQ Format - Wikipedia
- Illumina FASTQ Files Explained
- Illumina FASTQ Format BaseSpace
- Sequence Alignement/Map File Format (SAM/BAM)
-