FASTQ files are the basic unit to start any bioinformatic process. These derive from the traditional *fasta* files that store genetic information. The difference resides in that FASTQ also contain quality information of the data obtained in ASCII code. The standard and commonest format is that produced by Illumina but bear in mind that different platforms (Sanger, Solexa) output slightly different files.

Below is an example of one of our FASTQ files. It contains information on all reads from a single run, and for each read there are 4 lines of information:

1. Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line). In case of Illumina sequences it also contains information about the position on the flow-cell in the following manner:
    1. @<instrument_name>:<run-id>:<flowcell-id>:<flowcell_lane>:<x-coordinate of the detected cluster>:<y-coordinate of the detected cluster>:<member of a pair>:<Y if the read is filtered,N otherwise>:<0 when control bits are on, even number otherwise>:<multiplex index>
    2. exome data is often generated using a pair-end approach which means sequences are retrieved from both ends of the same molecule. Then, sequence data from paired end approached come in two files:
        1. **_1_sequence.fq* stores all the forward sequences and will have a 1 in the field <number of a pair>
        2. **_2_sequnece.fq* stores all the reverse sequences and will have a 2 in the field <number of a pair>
2. Line 2 contains the raw sequence letters. The number of bases at line 2 corresponds to the number of bases read per run. Initially, illumina produced reads of about 30 bp. Right now it reaches to reads of 100 bp.
    1. A,C,T,G and N are the characters allowed in this line, with N used if the real base could not be determined.
3. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
4. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.
    1. Illumina reads follow ASCII quality code where ! represents the lowest quality while ~ is the highest.

```
sed -n 21,24p SN4570283_14827_P3F08_5220_1_sequence.fq
```

@D6L3XBQ1:283:C3UTFACXX:5:1101:1468:1865 1:N:0:GGCTAC
NTCTCACCTGAATGCCCCAACAGCTCTCTCTTAAACCTTCACCTACACGCCCTGCAGCCAGAAGACTCAGCCCT
GTATCTCTGCGCCAGCAGCCAAGACAC +
#1=DDFFFHHHHHJJJJJJJJJGIJIJJJJJJJGJIJJJJIJJJGJJJJIJJHIJJHHIIIIHHHHHFFFFFDCECCDDFDEDDDDDDDDD<BDBD
DDDDD

```
$ sed -n 21,24p SN4570283_14827_P3F08_5220_2_sequence.fq
```

@D6L3XBQ1:283:C3UTFACXX:5:1101:1468:1865 2:N:0:GGCTAC

GGGGCTCTTGGAGGAAATGTTCACCCGAGCCCTCCGTGGCCCCCACGGCTTCCTGGCAGGCCCCGAAGGTTTC
TGCACAGGAAAGCGGTGACTCTGCAAGG +
CCCFFFFFGHHGHJHIJJIIIIJJJGIIIJJJJJJIJJJJJJJIJJGIHFFFFEEDEEDDDDDDB?@BD9>CDDCDCDDDD?DC<CBD<B
@CDCCCC@CDD

ASCII codes translates into values from 33 to 126 which derives into Phred Scores (the standard Sanger variant to assess reliability of a base cal) from 0 to 93. However, not all platforms use all ASCII symbols:

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.................................
..................
........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX....
..................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII....
..................
.............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ....
..................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.................................
..................
''!"#$%&'()*+,-
./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
z{|}~
  |                         |    |        |                              |
|
  33                        59   64       73                            104
126''
  0.......................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
  0.2.....................26...31........41
```

- S - Sanger - Phred+33, raw reads typically (0, 40)
- X - Solexa - Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ - Phred+64, raw reads typically (3, 40) with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
- L - Illumina 1.8+ - Phred+33, raw reads typically (0, 41)

Wiki has a detailed explanation of FASTQ build, Phred quality scores, and softwares to deal with them.

For raw reads, the range of scores will depend on the technology and the base caller used, but will typically be up to 41 for recent Illumina chemistry. On average one expects to have above 30 to consider reads to have good quality, which can be assessed with FastQC

Bear in mind that other platforms like Roche, do not directly produce FASTQ files, bus SFF files, which in addition from sequence and quality information, also store signal strengths. There are softwares designed to deal with Roche's SFF files. But one can also convert it to FASTQ files using scripts provided by Roche (sff.extract) or other softwares like seq_crumbs created by users. There are useful

discussions about this topic at SeqAnswers I, SeqAnswers II, and Biostars