FastQ Format

File extensions : file.fastq, file.sanfastq, file.fq

The official documentation for FastQ format can be found here.

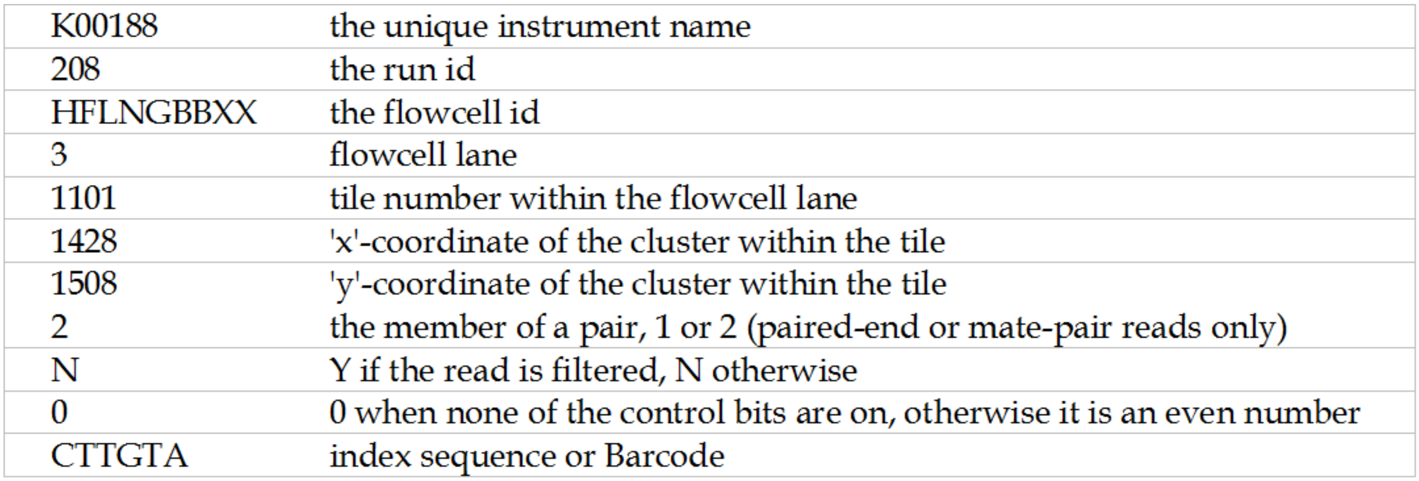
This is the most widely used format in sequence analysis as well as what is generally delivered from many sequencers. Many analysis tools require this format because it contains much more information than FastA.
The format is similar to FastA though there are differences in syntax as well as integration of quality scores. Each sequence requires at least 4 lines:

1. The first line is the sequence header which starts with an ‘@’ (not a ‘>’!).
    - Everything from the leading ‘@’ to the first whitespace character is considered the sequence identifier.
    - Everything after the first space is considered the sequence description
2. The second line is the sequence.
3. The third line starts with ‘+’ and can have the same sequence identifier appended (but usually is empty).
4. The fourth line are the quality scores
5.

The FastQ sequence identifier generally adheres to a particular format, all of which is information related to the sequencer and its position on the flowcell. The sequence description also follows a particular format and holds information regarding sample information. For example, if the first line of sequence is the following:

```
Line 1: @K00188:208:HFLNGBBXX:3:1101:1428:1508 2:N:0:CTTGTA
```

This is how to interpret the information:



K00188	the unique instrument name
208	the run id
HFLNGBBXX	the flowcell id
3	flowcell lane
1101	tile number within the flowcell lane
1428	'x'-coordinate of the cluster within the tile
1508	'y'-coordinate of the cluster within the tile
2	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read is filtered, N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
CTTGTA	index sequence or Barcode

Below is an example of a FastQ file (ignore the line numbers):

```
1  @HWUSI-EAS300R:7:1:7:674#0/1
2  CTTTTGTAGTTTACAAATCATGAATAATTTATAGAGTTAGTAACTTATAATTAATATACCTAGGAAATAAGTTA
3  +HWUSI-EAS300R:7:1:7:674#0/1
4  abbb_[V‘a_abaababa‘abbaaba_bbbabbaab‘aaaabaaba‘baabbbabaaa‘\bbaaa‘aaa_b‘‘a
5  @HWUSI-EAS300R:7:1:9:1897#0/1
6  ACTTAATTATTTATGCTTTTCTCTACTCTGCACGGCATGCAAATGCAATATAGATGCAAGGCGAGCCGAAACAA
7  +HWUSI-EAS300R:7:1:9:1897#0/1
8  aaab\‘bb^bababbaababbbbbababbbbabaaaaabbbb[_bba_‘^aabaabaaaa__aaaaaa[Saa_‘
9  @HWUSI-EAS300R:7:1:13:849#0/1
10  GAAATATTGCGTAGCCGGAAACAAAAGAGTGCAAATACATTTCGACGATGATGAGAGAGCTATTCAGGGCTGTA
11  +HWUSI-EAS300R:7:1:13:849#0/1
12  a^‘aaaabbb^‘bab‘a_Xaa_aa^_a__‘_aaaa‘ba‘aaa_‘_aaa‘aZ_a^a]_^aR_a^_^‘_‘]^_‘\Y
13  ......
```

For the above sequence then, the first four lines is the first sequence read:

```
@HWUSI-EAS300R:7:1:7:674#0/1
CTTTTGTAGTTTACAAATCATGAATAATTTATAGAGTTAGTAACTTATAATTAATATACCTAGGAAATAAGTTA
+
abbb_[V‘a_abaababa‘abbaaba_bbbabbaab‘aaaabaaba‘baabbbabaaa‘\bbaaa‘aaa_b‘‘a
```

And again, every read occupies four lines in a fastq file, so in the case of the above sequence:

```
{line 1} @:name of the read (always starts with ’@’)
{line 2} CTT..:the sequence of the read (IUPAC)
{line 3} +:the name again (only useful with multiline fastq entries)
{line 4} abbb_[V’a_….: the base quality of the DNA sequence; one character for every
nucleotide
```

How to interpret the quality scores?

What exactly does the fourth line of each sequence tell us? So what does a quality of ’a’ or ’b’ mean?

You would think there would be a simple answer to this questions, but thanks to Illumina the answer is: ”it depends “. Accordingly we now have different types of fastq files depending on the version of the Illumina instrument used. Nonetheless, basically every base quality entry gives you the probability that the corresponding DNA nucleotide is wrong (thus a sequencing error). So a quality of 0.001 means that on average 1 in 1000 bases having this quality is wrong.

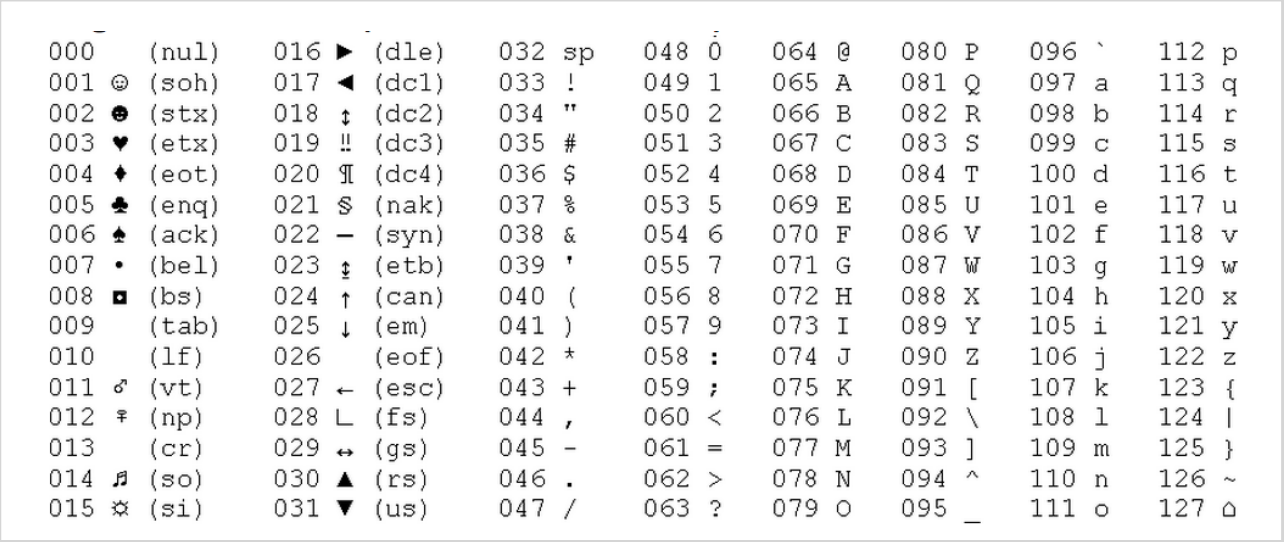
The encoding of the quality involves a few steps

- get the decimal value of the quality character from an ASCII table (see below)
- subtract the offset (sanger=33 illumina=64)
- voila: the quality of the base; usually a number between 0-40 [=dec(ascii)]
- if you want, you can go on and calculate the probability of being a sequencing error:

$$p = 10^{-\frac{dec(ascii)}{10}}$$

The ASCII table

In informatics every character has a number. For the computer they are actually interchangeable. The binary ’01000001’ refers at the same time to the number ’65’ and to the character ’A’. Only the context decides which definition will be used. It is therefore fairly straight forward to translate numbers between 0-127 to characters. Below is an ASCII table showing the codes;



000		(nul)	016	►	(dle)	032	sp	048	0	064	@	080	P	096	`	112	p
001	☺	(soh)	017	◄	(dc1)	033	!	049	1	065	A	081	Q	097	a	113	q
002	☻	(stx)	018	↕	(dc2)	034	"	050	2	066	B	082	R	098	b	114	r
003	♥	(etx)	019	‼	(dc3)	035	#	051	3	067	C	083	S	099	c	115	s
004	♦	(eot)	020	¶	(dc4)	036	$	052	4	068	D	084	T	100	d	116	t
005	♣	(enq)	021	§	(nak)	037	%	053	5	069	E	085	U	101	e	117	u
006	♠	(ack)	022	▬	(syn)	038	&	054	6	070	F	086	V	102	f	118	v
007	•	(bel)	023	↨	(etb)	039	'	055	7	071	G	087	W	103	g	119	w
008	◘	(bs)	024	↑	(can)	040	(	056	8	072	H	088	X	104	h	120	x
009		(tab)	025	↓	(em)	041	)	057	9	073	I	089	Y	105	i	121	y
010		(lf)	026		(eof)	042	*	058	:	074	J	090	Z	106	j	122	z
011	♂	(vt)	027	←	(esc)	043	+	059	;	075	K	091	[	107	k	123	{
012	♀	(np)	028	∟	(fs)	044	,	060	<	076	L	092	\	108	l	124	|
013		(cr)	029	↔	(gs)	045	-	061	=	077	M	093	]	109	m	125	}
014	♫	(so)	030	▲	(rs)	046	.	062	>	078	N	094	^	110	n	126	~
015	☼	(si)	031	▼	(us)	047	/	063	?	079	O	095	_	111	o	127	⌂

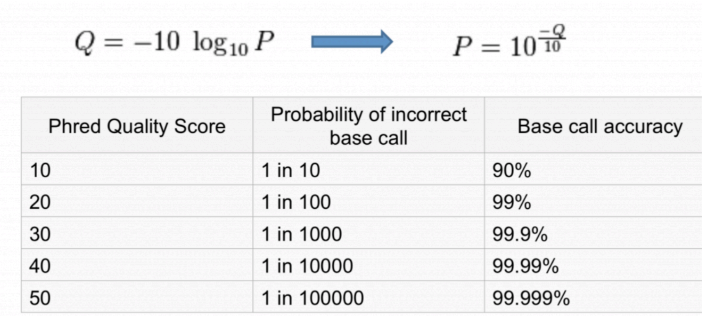
An example of decoding a FastQ

```
@read1
TTACGTTTTTT
+read1
87ba7777777
```

Base ’A’ in Illumina encoding (offset = 64)

- the ‘A’ has the quality ‘b’
- b⇒98 (from ascii table)
- 98−64=34; my base quality for A is therefore 34
- error probability p = $10^{-34/10}$ = 0.000398
- This error probability can be interpreted as 1 in 2511 (= 1/0.000398) base pairs being wrongly sequenced (= sequencing error)

You can get a general sense of the quality score for a base using the table below:



$$Q = -10\ \log_{10} P \longrightarrow P = 10^{\frac{-Q}{10}}$$

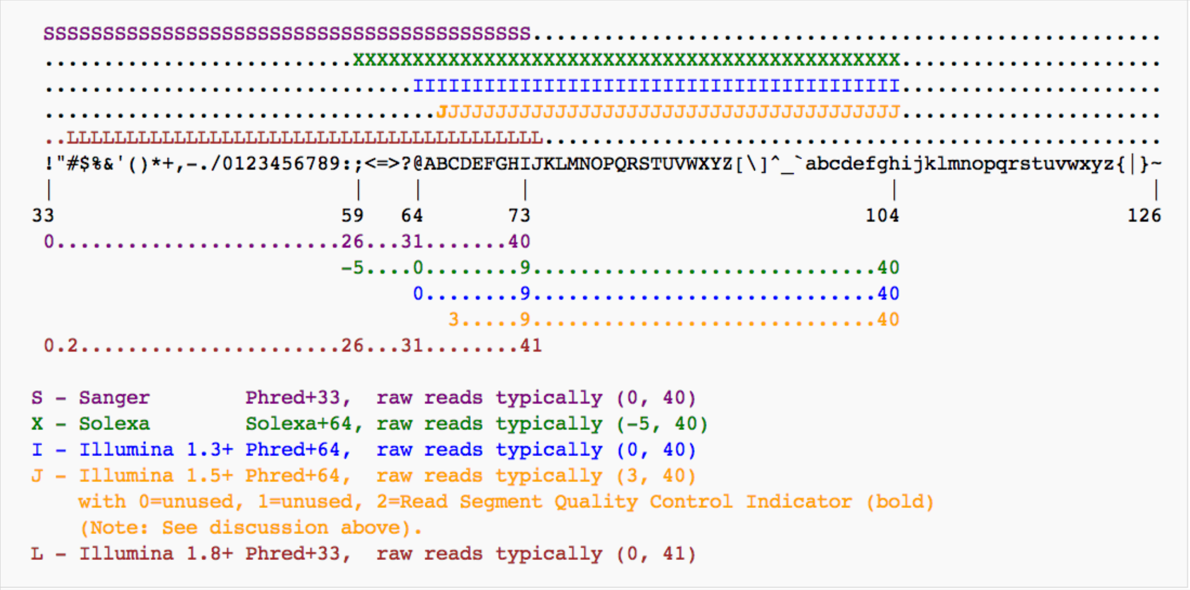
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

What does Phred mean in this figure?

It basically is the quality score as described above. For an explanation of Phred scores, check out this link: Phred scores.

Can you tell from the FastQ files which platform the sequence was collected?

Kind of. If any of the bases has a negative quality with Illumina (64) than the encoding is Sanger (33). Therefore if you find any ascii character < 64 (e.g.: 1,2,3,4,5,6,7,8,9) in your fastq file than your encoding is Sanger; if not then it is Illumina. Or it could be from a different instrument. Below is a figure that sort of maps instrument type onto ascii code. Do not worry if you do not completely understand the figure. It all can be a little confusing sometimes.



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
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)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

What software use FastQ?

Nearly everything works with this format. Some common examples are:

- Aligners
  - Bowtie, Tophat2
- Assemblers
  - Velvet, Spades
- QC tools
  - Trimmomatic, FastQC

I think it's a shorter list to tell you what does not work with FastQ files. Please note that there are tools available to convert FastQ to FastA in the event that FastQ is incompatible with the tool you're using:

- Blast
- Multiple Sequence Aligners
- Any reference sequence

How are these files generated?

- Most sequencers generate this format either by default (Illumina) or are easily converted.
- This can also be generated from a few different file formats (BAM, SFF, HDF5), though they all were some form of FastQ at some point.