Annex

Ilumina fastq format

4 lines for each sequence:

1- Unique identifier, with the following format:
@<machine_id>:<lane>:<tile>:<x_coord>:<y_coord>#<index>/<read_#>
2- Sequence (A, T, C ,G or N (undetermined) only)
3- Orientation (always forward without mapping)
4- Quality value for each base, corresponding to a Phred-like score encoded in ASCII format, with an offset of 64 (e.g. “f” gives a value of 38).

example:

```
@HWI-ST132_403:1:1:17205:2161#AGGACC/1
CAATTAACACTCTAAGAGCTTCTTCTCCCTCGCTCTGGGTCTTTTTCACA
+
fffffefffedffffecceddaffceeeececeeeceecbeeededde`c
```

Control lane

A channel is loaded with the PhiX library. It is used as a control lane (to set the base-calling parameters) to guide the base-calling of other channels (see below).

PhiX error rate

The PhiX reads are mapped on the PhiX reference genome, the error rate is then estimated by the number mismatches, over the total number of bases of mapped PhiX reads.
The current Illumina specifications are the following:

- error rate below 1%	50bp reads
- error rate below 2%	100bp reads

On HiSeq2000:
The control lane for base-calling is set either on the PhiX channel or on a channel loaded with a homogenous genomic library.
Furthermore, at Fasteris, we use an indexed PhiX library for spiking in every channel of the flow-cell which allows the computation of the error rate for each channel by mapping the de-multiplexed PhiX reads. As sample libraries are also indexed, after de-multimplexing by index selection, no PhiX is present in the provided dataset.

N.B.: the calculated error rate of the spiked PhiX might not be directly extrapolated to other libraries in the channel. A difference in the size of the DNA colonies (insert length) and/or a difference in GC-content and/or the presence of secondary structures can influence the base-calling and possibly lead to a difference between PhiX error rate and other libraries error rates.

During the run, a real time error rate is also calculated by the HiSeq2000: the 24 first bases are mapped on the PhiX reference, allowing 1 mismatch. The error rate of the following bases is calculated by extending the match.

We observe improvement on the error rates with each new version of the base-calling pipeline and sequencing kits.

N.B.: the error rate increases with the cycle number. Therefore, a lower error rate allows longer sequencing.

Illumina quality score

In fastq file format, a Phred-like quality score is associated to each base of the sequence. It is calculated to reflect the probability that the base is called erroneously.
In Illumina's quality scoring, the quality score is typically quoted as QXX, where the “XX” is the score, and it means that a particular call has a probability of error (Pε) of Pε=10^(-XX/10). In other terms, the quality is QXX=-10xlog$_{10}$Pε .

e.g. Q30 is associated to an error rate of 1 in 1000 (0.1%) .

The highest base quality value is determined according to the algorithms that generate the matrices for the base-calling, and has been modified with the different releases and updates of the base-calling pipelines. The following table summarizes the main changes:

Quality scoring type	Solexa	Illumina	
CASAVA pipeline version	previous to 1.3	1.3 – 1.5	1.6 – 1.7
Highest quality value	40	35	38
Character associated	h	b	f

Currently Illumina promotes the use of the quality of the reads according to the percentage of bases having a base quality greater or equal than Q30. Our specifications concerning the Q30 are the following:

- ≥70% of the reads having a Q30 for reads longer than 75 bases.
- ≥75% of the reads having a Q30 for reads up to 50 bases.

N.B.: - *Some biases have been observed on Illumina quality score values for samples with a high/low GC content, or presenting strong secondary structures.*
- *The Solexa quality scoring are now obsolete and no longer supported.*