FASTQ format

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity. It was originally developed at the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data, but has recently become the *de facto* standard for storing the output of high throughput sequencing instruments such as the Illumina Genome Analyzer [1] .

Format

A FASTQ file normally uses four lines per sequence. Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line). Line 2 is the raw sequence letters. Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again. 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.

A minimal FASTQ file might look like this:

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

The original Sanger FASTQ files also allowed the sequence and quality strings to be wrapped (split over multiple lines), but this is generally discouraged as it can make parsing complicated due to the unfortunate choice of "@" and "+" as markers (these characters can also occur in the quality string).

Illumina sequence identifiers

Sequences from the Illumina software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Versions of the Illumina pipeline since 1.4 appear to use **#NNNNNN** instead of **#0** for the multiplex ID, where **NNNNNN** is the sequence of the multiplex tag.

NCBI Short Read Archive

FASTQ files from the NCBI Short Read Archive often include a description, e.g.

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

In this example there is an NCBI-assigned identifier, and the description holds the original identifier from Solexa/Illumina (as described above) plus the read length.

Also note that the NCBI have converted this FASTQ data from the original Solexa/Illumina encoding to the Sanger standard (see encodings below).

Variations

Quality

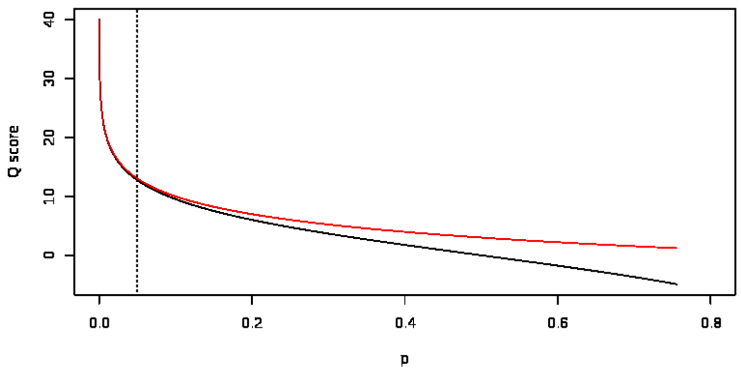
A quality value *Q* is an integer mapping of *p* (i.e., the probability that the corresponding base call is incorrect). Two different equations have been in use. The first is the standard Sanger variant to assess reliability of a base call, otherwise known as Phred quality score:

$$Q_\text{sanger} = -10\,\log_{10} p$$

The Solexa pipeline (i.e., the software delivered with the Illumina Genome Analyzer) earlier used a different mapping, encoding the odds *p*/(1-*p*) instead of the probability *p*:

$$Q_\text{solexa-prior to v.1.3} = -10\,\log_{10} \frac{p}{1-p}$$

Although both mappings are asymptotically identical at higher quality values, they differ at lower quality levels (i.e., approximately $p > 0.05$, or equivalently, $Q < 13$).



Relationship between *Q* and *p* using the Sanger (red) and Solexa (black) equations (described above). The vertical dotted line indicates $p = 0.05$, or equivalently, $Q \approx 13$.

At times there has been disagreement about which mapping Illumina actually uses. The user guide (Appendix B, page 122) for version 1.4 of the Illumina pipeline states that: "The scores are defined as Q=10*log10(p/(1-p)) [*sic*], where p is the probability of a base call corresponding to the base in question"[2] . In retrospect, this entry in the manual appears to have been an error. The user guide (What's New, page 5) for the latest version of the Illumina pipeline (v.1.5) lists this description instead: "Important Changes in Pipeline v1.3 [*sic*]. The quality scoring scheme has changed to the Phred [i.e., Sanger] scoring scheme, encoded as an ASCII character by adding 64 to the Phred value. A Phred score of a base is: Q_phred=-10 log_{10}(e), where *e* is the estimated probability of a base being wrong[3] .

Encoding

- Sanger format can encode a Phred quality score from 0 to 93 using ASCII 33 to 126 (although in raw read data the Phred quality score rarely exceeds 60, higher scores are possible in assemblies or read maps). Also used in SAM format[4] . Coming to the end of February 2011, Illumina's newest version (1.8) of their pipeline CASAVA will produce directly fastq in Sanger format, according to the announce on seqanswers.com forum[5] .
- Solexa/Illumina 1.0 format can encode a Solexa/Illumina quality score from -5 to 62 using ASCII 59 to 126 (although in raw read data Solexa scores from -5 to 40 only are expected)
- Illumina 1.3+ format can encode a Phred quality score from 0 to 62 using ASCII 64 to 126 (although in raw read data Phred scores from 0 to 40 only are expected).
- The Phred scores 0 to 2 in Illumina 1.5+ have a slightly different meaning. The values 0 and 1 are no longer used and the value 2, encoded by ASCII 66 "B", is used also at the end of reads as a *Read Segment Quality Control Indicator* [6] . The Illumina manual[7] (page 30) states the following: *If a read ends with a segment of mostly low quality (Q15 or below), then all of the quality values in the segment are replaced with a value of 2 (encoded as the letter B in Illumina's text-based encoding of quality scores)... This Q2 indicator does not predict a specific error rate, but rather indicates that a specific final portion of the read should not be used in further analyses.* Also, the quality score encoded as "B" letter may occur internally within reads at least as late as pipeline version 1.6, as shown in the following example:

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTTGGGGGAGACATTTTTGTGATTGCCTTGAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BBBBBBBBBBBBBBBBBBBBBBBB
```

An alternative interpretation of this ASCII encoding has been proposed[8] .

For raw reads, the range of scores will depend on the technology and the base caller used, but will typically be up to 40. For aligned sequences and consensuses higher scores are common.

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

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).
```

Color space

For SOLiD data, the sequence is in color space, except the first position. The quality values are those of the Sanger format. Alignment tools differ in their preferred version of the quality values: some include a quality score (set to 0, i.e. '!') for the leading nucleotide, others do not. The sequence read archive includes this quality score.

File extension

There is no standard file extension for a FASTQ file, but .fq, .fastq, and .txt are commonly used.

Format converters

- Biopython version 1.51 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
- EMBOSS version 6.1.0 patch 1 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
- BioPerl version 1.6.1 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
- BioRuby version 1.4.0 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
- MAQ [9] can convert from Solexa to Sanger (use this patch [10] to support Illumina 1.3+ files).

References

[1] Cock et al (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, doi:10.1093/nar/gkp1137 (http://dx.doi.org/10.1093/nar/gkp1137)

[2] Sequencing Analysis Software User Guide: For Pipeline Version 1.4 and CASAVA Version 1.0, dated April 2009 PDF (http://genomecenter.ucdavis.edu/dna_technologies/documents/pipeline_1_4.pdf)

[3] Sequencing Analysis Software User Guide: For Pipeline Version 1.5 and CASAVA Version 1.0, dated August 2009 PDF (http://illumina.ucr.edu/illumina_docs/Pipeline1.5/Pipeline1.5_CASAVA1.0_User_Guide_15006500_A.pdf)

[4] Sequence/Alignment Map format Version 1.0, dated August 2009 PDF (http://samtools.sourceforge.net/SAM1.pdf)

[5] Seqanswer's topic of skruglyak, dated January 2011 website (http://seqanswers.com/forums/showthread.php?s=ba8c7dfba863815f637c0bf45882f14b&t=8895)

[6] Illumina Quality Scores, Tobias Mann, Bioinformatics, San Diego, Illumina (http://seqanswers.com/forums/showthread.php?t=4721)

[7] [Using Genome Analyzer Sequencing Control Software, Version 2.6, Catalog # SY-960-2601, Part # 15009921 Rev. A, November 2009]http://watson.nci.nih.gov/solexa/Using_SCSv2.6_15009921_A.pdf

[8] SolexaQA project website (http://solexaqa.sourceforge.net/questions.htm#illumina)

[9] http://maq.sourceforge.net

[10] http://sourceforge.net/tracker/index.php?func=detail&aid=2824334&group_id=191815&atid=938893

External links

- MAQ (http://maq.sourceforge.net/fastq.shtml) webpage discussing FASTQ variants
- (http://imtech.res.in/raghava/crag/types.html) (CRAG) Instrument Output Formats
- Galaxy fastq tools (http://bitbucket.org/galaxy/galaxy-central/src/tip/tools/fastq/)
- Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing
- Fastqc (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) quality control tool for high throughput sequence data

Article Sources and Contributors

FASTQ format *Source*: http://en.wikipedia.org/w/index.php?oldid=408555083 *Contributors*: Cjfields1, Drilnoth, Jeberle, Komich, Magioladitis, Minimac, Narayanese, Peak, Sanders muc, Sealox, Torst, 41 anonymous edits

Image Sources, Licenses and Contributors

File:Probability_metrics.svg *Source*: http://en.wikipedia.org/w/index.php?title=File:Probability_metrics.svg *License*: Public Domain *Contributors*: User:Sealox

License