**FASTQ format**

From Wikipedia, the free encyclopedia

**FASTQ format** is a text-based [format](http://en.wikipedia.org/wiki/File_format) for storing both a biological sequence (usually [nucleotide sequence](http://en.wikipedia.org/wiki/Nucleotide_sequence)) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single [ASCII](http://en.wikipedia.org/wiki/ASCII) character for brevity. It was originally developed at the [Wellcome Trust Sanger Institute](http://en.wikipedia.org/wiki/Wellcome_Trust_Sanger_Institute) to bundle a [FASTA](http://en.wikipedia.org/wiki/FASTA_format) 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](http://en.wikipedia.org/wiki/Illumina_%28company%29) Genome Analyzer.[[1]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-Cock_et_al_2009-1)

**Contents**

* [1 Format](http://en.wikipedia.org/wiki/FASTQ_format#Format)
  + [1.1 Illumina sequence identifiers](http://en.wikipedia.org/wiki/FASTQ_format#Illumina_sequence_identifiers)
  + [1.2 NCBI Sequence Read Archive](http://en.wikipedia.org/wiki/FASTQ_format#NCBI_Sequence_Read_Archive)
* [2 Variations](http://en.wikipedia.org/wiki/FASTQ_format#Variations)
  + [2.1 Quality](http://en.wikipedia.org/wiki/FASTQ_format#Quality)
  + [2.2 Encoding](http://en.wikipedia.org/wiki/FASTQ_format#Encoding)
  + [2.3 Color space](http://en.wikipedia.org/wiki/FASTQ_format#Color_space)
  + [2.4 Compression](http://en.wikipedia.org/wiki/FASTQ_format#Compression)
* [3 File extension](http://en.wikipedia.org/wiki/FASTQ_format#File_extension)
* [4 Format converters](http://en.wikipedia.org/wiki/FASTQ_format#Format_converters)
  + [4.1 Command line conversions](http://en.wikipedia.org/wiki/FASTQ_format#Command_line_conversions)
* [5 See also](http://en.wikipedia.org/wiki/FASTQ_format#See_also)
* [6 References](http://en.wikipedia.org/wiki/FASTQ_format#References)
* [7 External links](http://en.wikipedia.org/wiki/FASTQ_format#External_links)

**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](http://en.wikipedia.org/wiki/FASTA_format) 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 FASTQ file containing a single sequence might look like this:

@SEQ\_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!''\*((((\*\*\*+))%%%++)(%%%%).1\*\*\*-+\*''))\*\*55CCF>>>>>>CCCCCCC65

The character '!' represents the lowest quality while '~' is the highest. Here are the quality value characters in left-to-right increasing order of quality ([ASCII](http://en.wikipedia.org/wiki/ASCII)):

!"#$%&'()\*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^\_`abcdefghijklmnopqrstuvwxyz{|}~

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).[[2]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-2) An example of a tools that break the 4 line convention is vcfutils.pl from samtools.[[3]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-3)

**Illumina sequence identifiers**

Sequences from the [Illumina](http://en.wikipedia.org/wiki/Solexa) 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.

With Casava 1.8 the format of the '@' line has changed:

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

|  |  |
| --- | --- |
| **EAS139** | the unique instrument name |
| **136** | the run id |
| **FC706VJ** | the flowcell id |
| **2** | flowcell lane |
| **2104** | tile number within the flowcell lane |
| **15343** | 'x'-coordinate of the cluster within the tile |
| **197393** | 'y'-coordinate of the cluster within the tile |
| **1** | the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)* |
| **Y** | Y if the read is filtered, N otherwise |
| **18** | 0 when none of the control bits are on, otherwise it is an even number |
| **ATCACG** | index sequence |

**NCBI Sequence Read Archive**

FASTQ files from the [NCBI](http://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information)/[EBI](http://en.wikipedia.org/wiki/European_Bioinformatics_Institute) [Sequence Read Archive](http://en.wikipedia.org/wiki/Sequence_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](http://en.wikipedia.org/wiki/Solexa) (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**

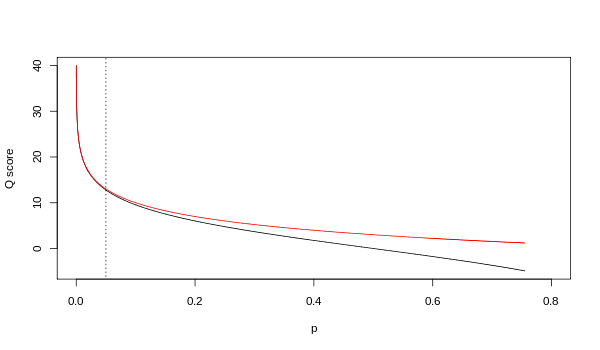
A quality value *Q* is an integer mapping of *p* (i.e., the probability that the corresponding base call is incorrect). Two different equations have been in use. The first is the standard Sanger variant to assess reliability of a base call, otherwise known as [Phred quality score](http://en.wikipedia.org/wiki/Phred_quality_score):

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](http://en.wikipedia.org/wiki/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).

[](http://en.wikipedia.org/wiki/File:Probability_metrics.svg)

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* ≈ 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*](http://en.wikipedia.org/wiki/Sic)], where p is the probability of a base call corresponding to the base in question".[[4]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-Illumina_User_Guide_1.4-4) In retrospect, this entry in the manual appears to have been an error. The user guide (What's New, page 5) for version 1.5 of the Illumina pipeline lists this description instead: "Important Changes in Pipeline v1.3 [[*sic*](http://en.wikipedia.org/wiki/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_\text{phred} = -10 \log_\text{10} e, where *e* is the estimated probability of a base being wrong.[[5]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-Illumina_User_Guide_1.5-5)

**Encoding**

* Sanger format can encode a [Phred quality score](http://en.wikipedia.org/wiki/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.[[6]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-Sequence.2FAlignment_Map_format-6) Coming to the end of February 2011, Illumina's newest version (1.8) of their pipeline CASAVA will directly produce fastq in Sanger format, according to the announcement on seqanswers.com forum.[[7]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-Upcoming_changes_in_CASAVA_topic-7)
* Solexa/Illumina 1.0 format can encode a Solexa/Illumina quality score from -5 to 62 using [ASCII](http://en.wikipedia.org/wiki/ASCII) 59 to 126 (although in raw read data Solexa scores from -5 to 40 only are expected)
* Starting with Illumina 1.3 and before Illumina 1.8, the format encoded a [Phred quality score](http://en.wikipedia.org/wiki/Phred_quality_score) from 0 to 62 using [ASCII](http://en.wikipedia.org/wiki/ASCII) 64 to 126 (although in raw read data Phred scores from 0 to 40 only are expected).
* Starting in Illumina 1.5 and before Illumina 1.8, the Phred scores 0 to 2 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*.[[8]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-8) The Illumina manual[[9]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-9) (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.[[10]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-10) Also, in Illumina runs using PhiX controls, the character 'B' was observed to represent an "unknown quality score". The error rate of 'B' reads was roughly 3 phred scores lower the mean observed score of a given run.

* Starting in Illumina 1.8, the quality scores have basically returned to the use of the Sanger format (Phred+33).

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. Since the maximum observed quality score was previously only 40, various scripts and tools break when they encounter data with quality values larger than 40. For processed reads, scores may be even higher. For example, quality values of 45 are observed in reads from Illumina's Long Read Sequencing Service (previously Moleculo).

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................

..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................

...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................

.................................**J**JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................

LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................

!"#$%&'()\*+,-./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)

**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.

**Compression**

Quality values account for about half of the required disk space in the FASTQ format (before compression), and therefore the compression of the quality values can significantly reduce storage requirements and speed up analysis and transmission of sequencing data. Both lossless and lossy compression are recently being considered in the literature. For example, the algorithm QualComp [[11]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-11) performs lossy compression with a rate (number of bits per quality value) specified by the user. Based on rate-distortion theory results, it allocates the number of bits so as to minimize the MSE (mean squared error) between the original (uncompressed) and the reconstructed (after compression) quality values. Other algorithms for compression of quality values include SCALCE [[12]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-12) and Fastqz.[[13]](http://en.wikipedia.org/wiki/FASTQ_format#cite_note-13) Both are lossless compression algorithms that provide an optional controlled lossy transformation approach. For example, SCALCE reduces the alphabet size based on the observation that “neighboring” quality values are similar in general.

**File extension**

There is no standard [file extension](http://en.wikipedia.org/wiki/File_extension) for a FASTQ file, but .fq and .fastq, are commonly used.

**Format converters**

* [Biopython](http://en.wikipedia.org/wiki/BioPython) version 1.51 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
* [EMBOSS](http://en.wikipedia.org/wiki/EMBOSS) version 6.1.0 patch 1 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
* [BioPerl](http://en.wikipedia.org/wiki/BioPerl) version 1.6.1 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
* [BioRuby](http://en.wikipedia.org/wiki/BioRuby) version 1.4.0 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
* [BioJava](http://en.wikipedia.org/wiki/BioJava) version 1.7.1 onwards (interconverts Sanger, Solexa and Illumina 1.3+)
* [MAQ](http://maq.sourceforge.net) can convert from Solexa to Sanger (use this [patch](http://sourceforge.net/tracker/index.php?func=detail&aid=2824334&group_id=191815&atid=938893) to support Illumina 1.3+ files).
* [fastx\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit/) The included fastq\_quality\_converter program can convert Illumina to Sanger

**Command line conversions**

**FASTQ to FASTA format:**

zcat input\_file.fastq.gz | awk 'NR%4==1{printf ">%s\n", substr($0,2)}NR%4==2{print}' > output\_file.fa

**Illumina FASTQ 1.8 to 1.3**

sed -e '4~4y/!"#$%&'\''()\*+,-.\/0123456789:;<=>?@ABCDEFGHIJ/@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\\]^\_`abcdefghi/' myfile.fastq # add -i to save the result to the same input file

**Illumina FASTQ 1.3 to 1.8**

sed -e '4~4y/@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\\]^\_`abcdefghi/!"#$%&'\''()\*+,-.\/0123456789:;<=>?@ABCDEFGHIJ/' myfile.fastq # add -i to save the result to the same input file

**See also**

* [FASTA format](http://en.wikipedia.org/wiki/FASTA_format)
* [Phred quality score](http://en.wikipedia.org/wiki/Phred_quality_score)
* [List of file formats for molecular biology](http://en.wikipedia.org/wiki/List_of_file_formats#Biology)

**References**

+ Cock et al (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, [doi](http://en.wikipedia.org/wiki/Digital_object_identifier):[10.1093/nar/gkp1137](http://dx.doi.org/10.1093%2Fnar%2Fgkp1137)

+ Cock, P. J. A.; Fields, C. J.; Goto, N.; Heuer, M. L.; Rice, P. M. (2009). ["The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants"](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217). *Nucleic Acids Research* **38** (6): 1767–1771. [doi](http://en.wikipedia.org/wiki/Digital_object_identifier):[10.1093/nar/gkp1137](http://dx.doi.org/10.1093%2Fnar%2Fgkp1137). [PMC](http://en.wikipedia.org/wiki/PubMed_Central) [2847217](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217). [PMID](http://en.wikipedia.org/wiki/PubMed_Identifier) [20015970](http://www.ncbi.nlm.nih.gov/pubmed/20015970). [edit](http://en.wikipedia.org/w/index.php?title=Template:Cite_doi/10.1093.2Fnar.2Fgkp1137&action=edit&editintro=Template:Cite_doi/editintro2)

+ <http://cnr.lwlss.net/SeqConsensus/> accessed 14 March 2014

+ 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)

+ 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)

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

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

+ Illumina Quality Scores, Tobias Mann, Bioinformatics, San Diego, Illumina [[1]](http://seqanswers.com/forums/showthread.php?t=4721)

+ [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>

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

+ Ochoa, Idoia, et al. "QualComp: a new lossy compressor for quality scores based on rate distortion theory." BMC bioinformatics 14.1 (2013): 187. <http://www.biomedcentral.com/1471-2105/14/187/>

+ Hach F, Numanagi ́c I, Alkan C, Sahinalp SC:SCALCE: boosting sequencecompression algorithms using locally consistent encoding.Bioinformatics2012,28(23):3051–3057.

+ fastqz.<http://mattmahoney.net/dc/fastqz/>

**External links**

* [MAQ](http://maq.sourceforge.net/fastq.shtml) webpage discussing FASTQ variants
* [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
* [PRINSEQ](http://prinseq.sourceforge.net/) can be used for QC and to filter, reformat, or trim sequence data (web-based and command line versions)