

Base quality information

All DNA sequencing approaches are similar in a way that the sequencer has to interpret a signal that indicates the presence of a given nucleotide at a given position. The resulting raw data serves then as input into a software, the *basecaller*, that does the actual read out. In other words, it translates the signal into a nucleotide sequence.

The base calling is obviously not error free, as some signals suffer from a lot of noise (Trace A in figure 1). Others are pretty clear and easy to interpret (Trace B in figure 1). Note, Trace A and B cover the same stretch of the template sequence. As the four nucleotides cannot take up any information about the confidence in a certain base call, base qualities values have been introduced to propagate information about the sequence quality to downstream applications. The general procedure is outlined, exemplarily for Sanger sequencing in figure 1, but the same principle applies to other sequencing technologies.

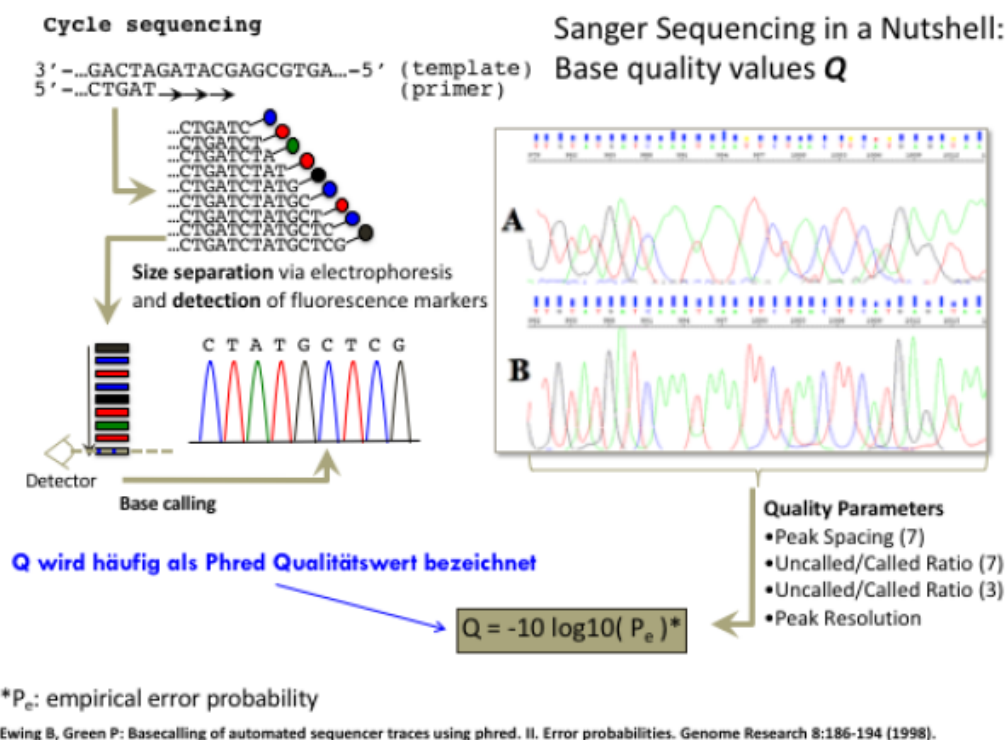


Figure 1: Base quality values and how they are inferred. Trace files (A, and B) represent the output from the sequencer which are subsequently interpreted by the basecaller. As it is nowadays no longer common to look at the individual traces by eye, software solutions exist to transfer the confidence in a base call into an empirical error probability. Base quality values have been initially proposed by [Green and Ewing 1998](#). They proposed a number of measures to assess the evenness of the chromatogram used for the base calling. The window size for which the measure is computed is given in parenthesis next to the evaluation criterion.

Base quality file formats

Base qualities are traditionally represented as the negative decadic logarithm of the error probability of a base call.

- Q=10 represents an error probability of one in 10
- Q=20 represents an error probability of one in 100
- Q=30 represents an error probability of one in 1000

Base qualities for individual sequencing reads typically range between 0 and maximally 40, although some technologies slightly differ in the range (Figure 2). Base qualities higher than 40 are typically only assigned to consensus sequences that are support from overlapping regions of more than one read.

Traditionally, base quality values have been stored as integer numbers. As DNA sequencing became cheaper, and, as a consequence, the amount of generated sequence data started to grow exponentially, it became common to store the information more memory efficient as ASCII characters (Fig. 2).



Figure 2: Projection of the base quality values to ASCII characters. Different sequencing platforms and different encoding versions use different encodings. Figure source: [FASTQ](#).

Relevance

Genome sequencing

Quality values used to be relevant for all applications of DNA sequencing. Nowadays, however, DNA sequencing has become cheap to an extent that read coverage, i.e. the number of reads covering a nucleotide in your template sequence, has taken away quite a bit of their importance, when it comes to the de-novo sequencing of genomes.


Transcriptomics / Genotyping

When it comes to applications where a uniformly high coverage is not guaranteed, e.g. when sequencing transcriptomes, or when one is looking for genetic variants, e.g. in the context of genotyping, then base qualities are still highly relevant.


Storing base quality information

FASTA

Base quality values are stored differently, depending on the sequence file format.

The  **FASTA** format cannot store sequence and base quality information within one file. In this case, each FASTA sequence file is accompanied by a corresponding base quality file. The headers of the corresponding sequence and base quality string must then be identical. Moreover, there must be a one-to-one relationship between nucleotides in the sequence and base quality values in the quality string.

FASTQ

The  **FASTQ** format stores sequence and quality information within one file. See figure 3 for an example.

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
! ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) **55CCF>>>>>>CCCCCCCC65
```

Header Information:

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



Figure 3: Example of a sequence read in FASTQ format. An explanation of the information given in the sequencing header is given below the file. Flowcell, lane, tile and x/y coordinates provide information about the physical location of the template cluster in the sequencer from which the sequence was derived

From:

<https://applbio.biologie.uni-frankfurt.de/teaching/wiki/> - Teaching

Permanent link:

<https://applbio.biologie.uni-frankfurt.de/teaching/wiki/doku.php?id=general:bioseqanalysis:sequencequalities>

Last update: 2021/10/20 12:54

