Understanding Illumina Quality Scores

Quality scores are an efficient way to communicate small error probabilities.

Highlights

- **A quality score is a prediction of the probability of an error in base calling**
- **Quality scores are generated by a quality table that uses a set of quality predictor values**
- **The quality table is updated when characteristics of the sequencing platform change**

What is a Quality Score?

A quality score (Q-score) is a prediction of the probability of an error in base calling. It serves as a compact way to communicate very small error probabilities.

A high quality score implies that a base call is more reliable and less likely to be incorrect. For example, for base calls with a quality score of Q40, one base call in 10,000 is predicted to be incorrect. For base calls with a quality score of Q30, one base call in 1,000 is predicted to be incorrect. Table 1 shows the relationship between the base call quality scores and their corresponding error probabilities.

Table 1: Q-Scores and Error Probabilities

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

How Are Quality Scores Generated?

During a sequencing run, a quality score is assigned to each base call for every cluster, on every tile, for every sequencing cycle. Illumina quality scores are calculated for each base call in a two-step process:

1. For each base call, a number of quality predictor values are computed. *Quality predictor values* are observable properties of clusters from which base calls are extracted. These include properties, such as intensity profiles and signal-to-noise ratios, measure various aspects of base call reliability. They have been empirically determined to correlate with the quality of the base call.
2. A *quality model*, also known as a *quality table* or *Q-table*, lists combinations of quality predictor values and relates them to corresponding quality scores; this relationship is determined by a calibration process using empirical data. To estimate a new quality score, the quality predictor values are computed for a new base call and compared to values in the pre-calibrated quality table.

Quality scores are recorded in base call files (*.bcl) that contain the base call and quality score per cycle. The quality scores are then converted to FASTQ files (*.fastq) in an encoded compact form.

How is a Quality Table Calibrated?

Calibration is a process in which a statistical quality table is derived from empirical data that includes various well-characterized human and non-human samples sequenced on a number of instruments. Using a modified version of the Phred algorithm[1], a quality table is developed and refined using characteristics of the raw signals and error rates determined by aligning reads to the appropriate references.

Why is Quality Score Binning Applied?

Q-scores comprise a large fraction of the total data storage space required for sequencing data. As the throughput of sequencing instruments continues to increase, data storage and transfer costs become a significant part of the total cost of sequencing. We have found that scores can be compressed into fewer quality bins, without affecting data quality or downstream analysis, such as alignment and variant calling.[2]

The resolution of Q-scores can be reduced in a number of ways, with the optimal approach depending on the quality distribution of the data generated by the sequencer. The most straightforward method begins with the creation of a high-resolution quality table. Q-scores are then mapped to a set of selected quality bins. For example, the original quality scores 20–24 may form one bin, and can all be mapped to a new value of 22. The choice of bins is empirically optimized to minimize the loss of quality score resolution across the data, while simultaneously minimizing the storage footprint. Table 2 shows existing bin boundaries in the first column and empirically mapped quality scores in the second column. Note that the bin boundaries listed in the first column of Table 2 may change with future quality calibrations.

Table 2: Quality Score Bins for Optimized 8-Level Mapping

Q-Score Bins	Example of Empirically Mapped Q-Scores*
N (no call)	N (no call)
2–9	6
10–19	15
20–24	22
25–29	27
30–34	33
35–39	37
≥ 40	40

By replacing the quality scores between 19 and 25 with a new score of 22, data storage space can be conserved.

*The mapped quality score for each bin (except "N") is subject to change depending on individual Q-tables.

Why Do Quality Tables Change?

The table that produces quality scores is typically updated when significant characteristics of the sequencing platform change, such as new hardware, software, or chemistry versions. For example, improvements in sequencing chemistry may require quality table recalibration to accurately score the new data.

Summary

Quality tables are created to provide quality scores for runs generated by specific instrument configurations and versions of chemistry. When significant characteristics of the sequencing platform change, the quality model may require recalibration. Therefore, to generate the most accurate quality scores, Illumina recommends using only the software version optimized specifically for the hardware and chemistry configuration currently in use.

References

1. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8: 186–194.
2. Illumina. (2014) Reducing whole-genome data storage footprint. (www.illumina.com/documents/products/whitepapers/whitepaper_datacompression.pdf)

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY OR FOR *IN VITRO* DIAGNOSTIC USE. NOT AVAILABLE IN ALL COUNTRIES OR REGIONS

Pub. No. 770-2012-058 Current as of 23 April 2014