Quality Scores for Next-Generation Sequencing

Assessing sequencing accuracy using Phred quality scoring.

Introduction

A next-generation sequencing experiment consists of a series of discrete steps that uniquely contribute to the overall quality of a data set. Sequencing quality metrics can provide important information about the accuracy of each step in this process, including library preparation, base calling, read alignment, and variant calling. Base calling accuracy, measured by the Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform. It indicates the probability that a given base is called incorrectly by the sequencer.

Historically used to determine Sanger sequencing accuracy, Phred originated as an algorithmic approach that considered Sanger sequencing metrics, such as peak resolution and shape, and linked them to known sequence accuracy through large multivariate lookup tables. This method proved to be highly accurate[1] across a range of sequencing chemistries and instruments, making it the quality scoring standard for commercial sequencing technologies.

While next-generation sequencing metrics vary from those of Sanger sequencing (e.g., no electropherogram peak heights), the process of generating a Phred quality scoring scheme is largely the same. Parameters relevant to a particular sequencing chemistry are analyzed for a large empirical data set of known accuracy. The resulting quality score lookup tables are used to calculate a quality score for *de novo* next-generation sequencing data (in real time on Illumina platforms), possessing an equivalent meaning to the historical metrics familiar to most Sanger sequencing users.

Calculating Phred Quality Scores

Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)[2].

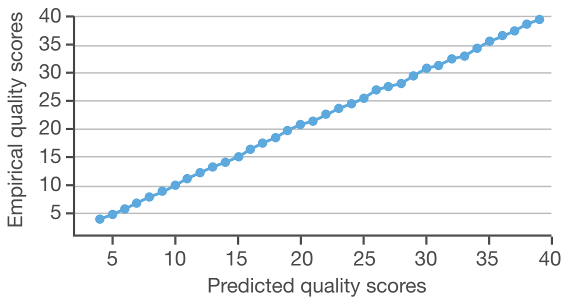
$$\mathbf{Q = -10 \log_{10} P}$$

For example, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times (Table 1). This means that the base call accuracy (i.e., the probability of a correct base call) is 99.9%. A lower base call accuracy of 99% (Q20) will have an incorrect base call probability of 1 in 100, meaning that every 100 bp sequencing read will likely contain an error. When sequencing quality reaches Q30, virtually all of the reads will be perfect, having zero errors and ambiguities. This is why Q30 is considered a benchmark for quality in next-generation sequencing. By comparison, Sanger sequencing systems generally produce base call accuracy of ~99.4%, or ~Q20[3]. Low Q scores can increase false-positive variant calls, which can result in inaccurate conclusions and higher costs for validation experiments.

Illumina Data Quality

Illumina Q score calculations have been shown to be very similar to the actual data quality observed in human genome sequencing[4]. Figure 1 shows that predicted and empirical quality scores from a HiSeq 2000 run are well correlated. Q scores can reveal how much of the data from a given run is usable in a resequencing or assembly experiment. Sequencing data with lower quality scores can result in a significant portion of the reads being unusable, resulting in wasted time and expense. PhiX quality scores for the MiSeq® and HiSeq® systems show that nearly all bases have scores > Q30 for single and paired-end reads (Figure 2). Comparison of *E. coli* whole-genome sequencing data shows that this high data quality is consistent across both platforms (Table 2).

Figure 1: High Correlation of Empirical and Predicted Q Scores



Illumina sequencing Q scores are highly accurate. This example shows that predicted Q scores for a HiSeq 2000 run correlate well to empirically derived Q scores.

Table 1: Quality Scores and Base Calling Accuracy

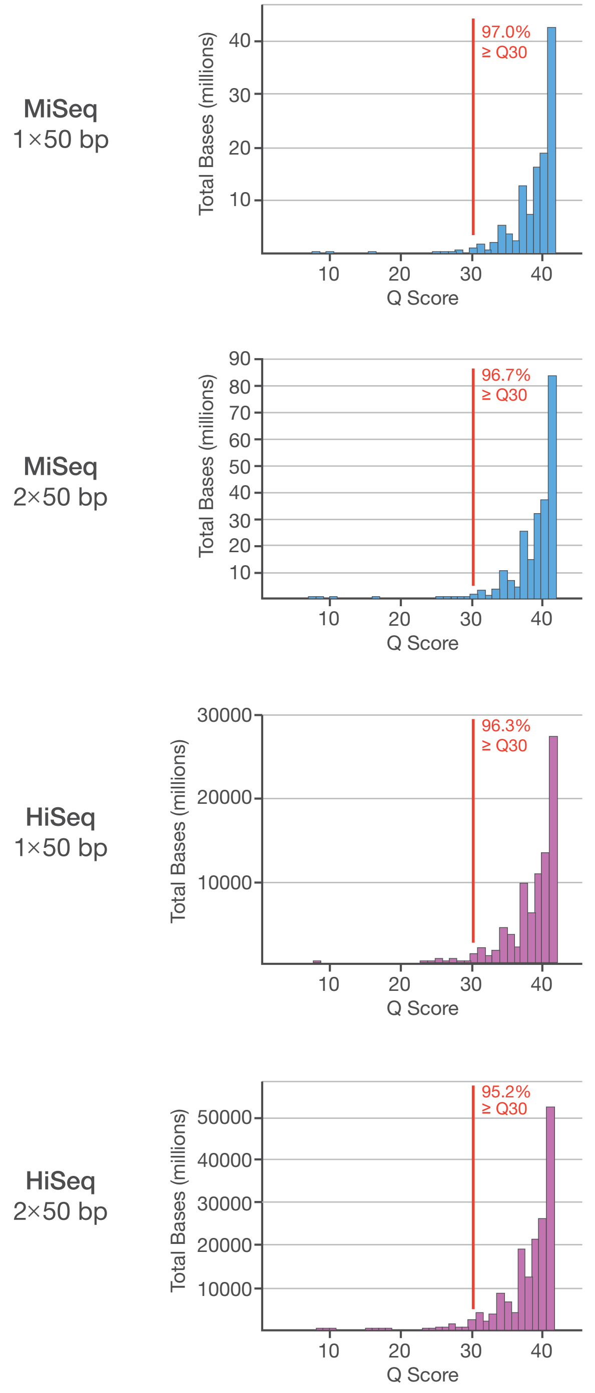
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Table 2: MiSeq vs HiSeq 2000 *E.coli* K12 MG1655 Data Comparison

Metric	MiSeq System		HiSeq System	
	Read 1	Read 2	Read 1	Read 2
% Bases Q ≥ 30	91.9	87.5	89.3	86.1
% Total Bases Q ≥ 30	89.7		87.7	

A whole-genome sequencing run (2 × 150 bp) of *E. coli* K12 MG1655 performed on the MiSeq system yielded 1.7 Gb of high-quality data. MiSeq data were trimmed to 2 × 100 bp to allow for a direct comparison with 2 × 100 bp reads from the HiSeq 2000 platform.

Figure 2: PhiX Quality Scores for HiSeq 2000 and MiSeq



At both 1 x 50 bp and 2 x 50 bp read lengths, virtually all bases are above Q30 across both the HiSeq and MiSeq systems.

Accurate Sequencing Chemistry

Illumina sequencing by synthesis (SBS) technology delivers the highest percentage of error-free reads, with a vast majority of bases having quality scores above Q30. In many cases, even higher quality scores of Q35–Q40 are available. The latest version of the chemistry, TruSeq™ SBS and Cluster Generation v3 reagents, have been optimized for accurate base calling even within difficult-to-sequence regions of the genome, such as repeats, homo polymers, and high GC regions. TruSeq v3 chemistry is available for the HiSeq and MiSeq systems. The unparalleled TruSeq accuracy is ideal for next-generation sequencing in clinical environments that demand the highest standard of quality[5]. Since the release of the original Illumina Genome Analyzer™ system, SBS technology has been used in the widest range of sequencing applications, resulting in more than 2,000 peer-reviewed publications in just five years—a feat unmatched for any other life science technology.

SBS chemistry uses four fluorescently labeled nucleotides to sequence up to billions of clusters on the flow cell surface in parallel. During each sequencing cycle, a single labeled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain. The dNTPs contain a reversible blocking group that serves as a terminator for polymerization, so after each dNTP incorporation, the fluorescent dye is imaged to identify the base and then enzymatically cleaved to allow incorporation of the next nucleotide. Since all four reversible terminator-bound dNTPs (A, C, T, G) are present as single, separate molecules, natural competition minimizes incorporation bias, which can be problematic with serial nucleotide incorporation chemistry used in Sanger sequencing. Base calls are made directly from signal intensity measurements during each cycle, greatly reducing raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that eliminates sequence-context specific errors, enabling robust base calling across the genome, including repetitive sequence regions and homo polymers.

Summary

Q scores are used to measure base calling accuracy, one of the most common metrics for assessing sequencing data quality. Low Q scores can lead to increased false-positive variant calls, resulting in inaccurate conclusions and higher costs for validation experiments. Illumina's sequencing chemistry delivers unparalleled accuracy, with a vast majority of bases scoring Q30 and above. This level of accuracy is ideal for a range of sequencing applications, including clinical research.

References

1. Richterich P. (1998): Estimation of errors in "raw" DNA sequences: a validation study. Genome Res. 8(3):251–259.
2. Ewing B, Green P. (1998): Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8(3):186–194.
3. http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_040402.pdf (as of 3/12/2012).
4. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, et al. (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nature Genetics. 42:931–936.
5. Morgan JE, Carr IM, Sheridan E, Chu CE, Hayward B, et al. (2010) Genetic diagnosis of familial breast cancer using clonal sequencing. Hum Mutat. 31(4):484–91.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2011-030 Current as of 31 October 2011