Quality Predictors in RTA v1.12

RTA Software version 1.12 features updates to the quality predictors, and as a result RTA Software version 1.12 now has the most accurate representation of the quality of a read. This document provides background on the new model, what has changed from the previous model, and answers to questions customers may have.

Illumina Quality Scores

A quality score is a prediction of the **probability** of an error in base calling:

Quality (Q) = -10log$_{10}$ (probability that the base is wrong)

Examples of quality scores and error probabilities are provided below:

Q-Scores and Error Probabilities

Quality Score	Error Rate
Q40	1 error in 10,000
Q30	1 error in 1,000
Q20	1 error in 100
Q10	1 error in 10

Quality scores are produced by a model that uses quality predictors as inputs. Quality predictors are numbers correlated with the quality of a base call, and attempt to quantify concepts such as:

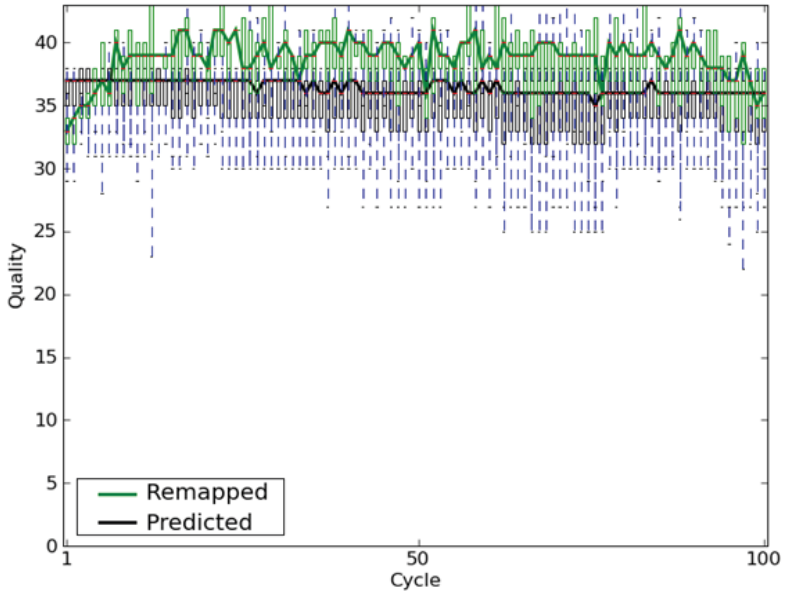
- **Is the signal for the called base much brighter than the others?**
- **Did the spot get suspiciously dim, compared to the beginning?**
- **Does the signal look clean in the next few cycles, and the previous few cycles?**

In RTA Software, we build a quality model and test how closely these predicted base quality values are to the actual base quality as assessed by remapping the read back to the reference.

Quality Prediction in Previous Version RTA v1.7

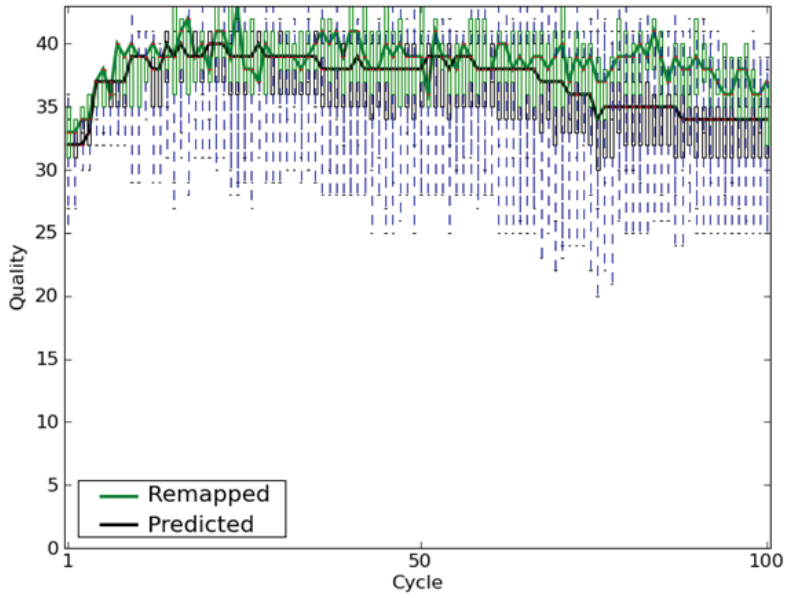
HCS1.1/RTA1.7 used a five-predictor model, which was really good at predicting quality, as can be seen when compared to quality scores obtained by re-mapping the data back to the Human reference (Figure 1). However, the model tended to under-estimate base quality in the main part of the read and over-estimate the quality at the beginning of the read.

Figure 1: Quality Scores by Cycle in RTA v1.7



Quality scores by cycle for a 100 cycle Human read. Five-predictor quality model used in HCS1.1/RTA1.7.
Green Line = Quality scores obtained by remapping data back to the Human reference.
Black Line = Predicted quality scores obtained from the five-parameter quality model used in RTA1.7.The model tends to under-estimate the base quality in the main part of the read and over-estimate quality at the beginning of the read.

Figure 2: Quality Scores by Cycle in RTA v1.12



RTA Software version 1.12 features the most accurate representation of the quality of a read.
Green Line = Quality scores by cycle for a 100 cycle Human read obtained by remapping data back to the Human reference.
Black Line = Predicted quality scores obtained from the new six-parameter quality model used in RTA1.12.

Quality Prediction in RTA Software version 1.12

HCS1.4/RTA1.12 uses a new six-predictor model, which accurately predicts the quality at the beginning of a read and more accurately reports the high quality of the main part of the read. As seen in Figure 1, the new six-parameter quality model featured in RTA Software version 1.12 closely matches the quality scores obtained by re-mapping the data for 3/4 of the run and then slightly underestimates the quality in the last cycles. Overall, the whole run shows data above Q30 and data in the Q40 range for most of the read.

The new model is also faster and provides Quality scores after around cycle 11 in read 2 of paired-end reads (compared to around cycle 25 with the previous model).

FAQs

What are predictors?

Numbers to attempt to quantify concepts such as: “Is the signal for the called base much brighter than the others?”, “Did the spot get suspiciously dim, compared to the beginning?”, “Does the signal look clean in the next few cycles, and the previous few cycles?”

Why is the quality observed at the beginning of the read lower in RTA1.12 than in RTA1.7?

The previous model did not properly account for the reduced, although still very high Q-values in the first few cycles

Why did we move to the 6-predictor model?

Although the 5-predictor model was very good at predicting quality, the 6-predictor model is more accurate and enables us to accurately predict the high percentage of Q40 data that was missed with the 5-predictor model. The new model is also faster and provides Quality scores after around cycle 11 in read 2 of paired-end reads (compared to around cycle 25 with the previous model).

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. 770-2011-010 Current as of 27 June 2011

illumina®