NovaSeq™ 6000 System Quality Scores and RTA3 Software

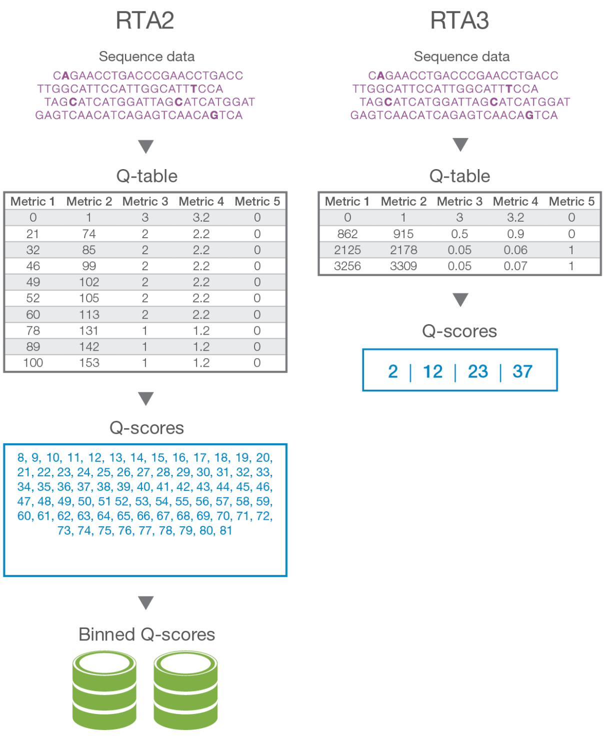
The NovaSeq 6000 System generates high-quality data comparable to the HiSeq X® Ten using more efficient storage of base calls and quality scores.

Introduction

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 Q-score of 30 (Q30) corresponds to a 0.1 percent error rate in base calling, and is widely considered a benchmark for high-quality data.[1] Q-scores not only provide a metric of base call data quality, they can also be used by secondary analysis tools. For example, a variant caller might weigh high-quality base calls more heavily, or simply discard base calls below a specific Q-score threshold.

NovaSeq 6000 System Q-scores are calculated through a process that is more streamlined than previous Illumina systems. This application note describes how Real Time Analysis 3 (RTA3) Software calculates Q-scores on the NovaSeq 6000 System and illustrates the benefits of the optimized RTA3 method. This application note also presents data from two experiments designed to evaluate NovaSeq data quality and accuracy.

For an in-depth description of Q-score calculations, read the Quality Scores for Next-Generation Sequencing or the Understanding Illumina Quality Scores technical notes.



Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
0	1	3	3.2	0
21	74	2	2.2	0
32	85	2	2.2	0
46	99	2	2.2	0
49	102	2	2.2	0
52	105	2	2.2	0
60	113	2	2.2	0
78	131	1	1.2	0
89	142	1	1.2	0
100	153	1	1.2	0

Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
0	1	3	3.2	0
862	915	0.5	0.9	0
2125	2178	0.05	0.06	1
3256	3309	0.05	0.07	1

Figure 1: Simplified Q-Scoring with RTA3—The simplified Q-Table of RTA3 enables faster data processing, reduced data file sizes, and simplified Q-score reporting chemistry figures.

NovaSeq Quality Score Calculations with RTA3

As with all Illumina systems, the NovaSeq 6000 System uses a platform-specific quality table (Q-table). The new Q-table was empirically developed by evaluating multiple features proven to be good predictors of quality on the NovaSeq System. Examples of these features include intensity, phasing, prephasing, and chastity values. To generate the Q-table for the NovaSeq System, three groups of base calls were determined, based on the clustering of these specific predictive features. Following grouping of the base calls, the mean error rate was empirically calculated for each of the three groups and the corresponding Q-scores were recorded in the Q-table alongside the predictive features correlating to that group. As such, only three Q-scores are possible with RTA3 and these Q-scores represent the average error rate of the group (Figure 1). Overall this results in simplified, yet highly accurate quality scoring. The three groups in the quality table correspond to marginal (< Q15), medium (~Q20), and high-quality (> Q30) base calls, and are assigned the specific scores of 12, 23, and 37 respectively.* Additionally, a null score of 2 is assigned to any no-calls.

Efficient Data Footprint and Faster Compute Architecture

The RTA3 method provides several significant advantages compared to previous versions, including a more efficient data footprint and faster computation. Due in large part to the simplified quality table, the NovaSeq 6000 System run folders require only ~0.4 bytes per base call compared to ~0.6 bytes per base call for the HiSeq X run folders.† More efficient data storage translates into lower storage costs and lower bandwidth requirements for sequencing data.[2] The total disk space footprint and the relative gain in efficiency are similar for

*These Q-scores may change with subsequent revisions of the quality table.

†These values are approximate, and can vary from run to run based on sample type and run quality.

For Research Use Only. Not for use in diagnostic procedures.

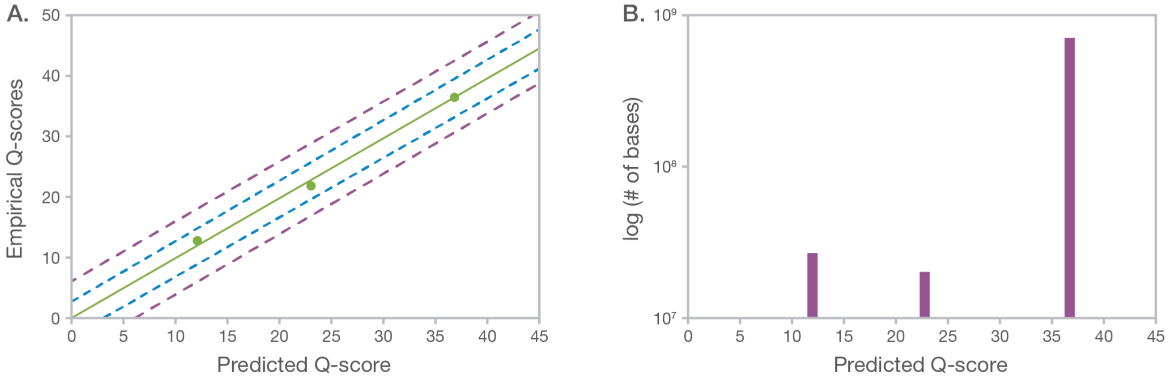


Figure 2: Comparison of NovaSeq Q-Scores and Empirical Q-Scores—(A)Empirical Q-scores are compared to NovaSeq Q-scores, showing high correlation. Note that RTA3 only reports three Q-scores: 12, 23, and 37. (B) A histogram of base call Q-scores falling into each quality group shows that the majority of base calls are above Q30 with a percent of total bases for each quality group of ~3.4%, ~2.6%, and ~94.0% respectively.

binary base call files, compressed FASTQ files, and BAM files. Additionally, technical enhancements enable faster RTA processing to align with the performance specifications of the NovaSeq System.

Testing NovaSeq Quality Score Accuracy

To validate the accuracy of the RTA3 method, we performed a whole-genome sequencing (WGS) run on the NovaSeq 6000 System using the well-characterized human sample NA12878. Empirical error rates were then calculated and compared to the Q-scores assigned by RTA3. A well-calibrated Q-table has empirical error rates correlating with the error rates predicted by Q-scores.[1]

Methods

WGS libraries were prepared from NA12878 genomic DNA (Coriell Institute for Medical Research) using the TruSeq™ DNA PCR-Free Library Prep Kit (Illumina, Catalog No. FC-121-3001) with an insert size of 450 bp. Sequencing was performed on the NovaSeq System with NovaSeq 6000 S2 Reagent Kit (Illumina, Catalog No. 20012860), using the 151 × 8 × 8 × 151 bp configuration.

To calculate the empirical error rates, each sequencing read was mapped to the human reference genome (HG38) using BaseSpace® Sequence Hub Whole-Genome Sequencing App v5.0.[3] Base calls were divided into quality groups, and the empirical scores were calculated from the observed error rate of those base calls. Known variants were excluded from the observed error rate calculation. The Q-score data were then plotted using Q-Q plot generation software.

Note: With this test the details of the aligner, including the details of soft-clipping and choosing which reads to align, play an import role in the measured error rate

Results

Empirical error rates were directly compared to assigned RTA3 Q-scores (Figure 2A).[‡] The diagonal green line represents the set of points where the empirical error rates exactly match the assigned Q-scores. Points above the green line indicate that RTA underestimated the true data quality, while points below the line indicate that RTA overestimated the true data quality. The blue and purple dashed lines indicate miscalibration by 3 and 6 units, respectively. This plot illustrates that the lowest quality value was underpredicted by one unit, the middle quality value was overpredicted by one unit, and the top quality value was accurate. These data show that the empirically determined error rates and the Q-scores assigned by RTA3 software are very well correlated and align nearly perfectly to one another.

We also assessed the number of base calls appearing in each quality group (Figure 2B). These data show that the vast majority of bases fall into the high-quality (Q30+) group. Overall, this test demonstrates that the RTA3 quality table is well-calibrated and shows high correlation with empirical error rates.

Comparison of NovaSeq and HiSeq X Data Quality and Variant Calling

To ensure that NovaSeq 6000 System sequencing data matches or exceeds HiSeq X data quality, the same human libraries were sequenced on both the NovaSeq and HiSeq X Systems. To quantify and compare variant calling accuracy, we used the well characterized NA12878 sample. In addition, these libraries were chosen because they show performance characteristics representative of the performance seen on the HiSeq X platform.

[‡] Data on file. Illumina Inc, 2017.

Methods

WGS libraries were prepared from NA12878 genomic DNA (Coriell Institute for Medical Research) using the TruSeq DNA PCR-Free Library Prep Kit (Illumina, Catalog No. FC-121-3001) with an insert size of 350 bp. Sequencing was performed on the HiSeq X System with the HiSeq X Ten Reagent Kit v2.5 (Illumina, Catalog No. FC-501-2501) and on the NovaSeq System with the NovaSeq 6000 S2 Reagent Kit (Illumina, Catalog No. 20012860), using the 2 × 150 bp run configuration. Secondary analysis was performed using BaseSpace Sequence Hub Whole-Genome Sequencing App v5.0,[3] also available as HiSeq Analysis Software v2.1.[4] Variant calling accuracy was assessed against PlatinumGenomes 2016 v1.0.[5] In both cases, the genome build was randomly downsampled to 30× coverage, using SAMBAMBA software.[6]

Results

Various primary and secondary analysis metrics for both platforms, including precision and recall for both single nucleotide variants (SNV), insertion-deletions (Indel), PhiX error rate, and more are summarized (Table 1). These data demonstrate that NovaSeq data quality and variant calling are equivalent to HiSeq X data quality and variant calling, with both systems showing both high-quality data and highly accurate variant calling.

Table 1: Comparison of NovaSeq and HiSeq X Data Quality and Variant Calling

	NovaSeq	HiSeq X
Autosome Mean Coverage	30.59	30.45
Autosome Callability	95.53%	95.46%
Autosome Exon Callability	98.47%	98.30%
SNV Precision	99.87%	99.88%
SNV Recall	97.07%	97.00%
Indel Precision	97.43%	97.65%
Indel Recall	95.49%	95.23%
PhiX Error Rate Read 1	0.35	0.41
PhiX Error Rate Read 2	0.61	1.41

Summary

The NovaSeq 6000 System uses a streamlined quality scoring method with RTA3 that enables many new performance improvements including faster data processing, reduced data storage footprint, and simplified Q-scoring. Our internal testing demonstrates that the NovaSeq Q-table generates highly accurate Q-scores with high correlation to empirically calculated error rates. Our internal testing also shows that the NovaSeq System produces high-quality data and variant calling comparable to those produced on the HiSeq X System.

References

1. Illumina (2011) Quality Scores for Next-Generation Sequencing. Accessed June 2017.
2. Illumina (2014) Reducing Whole-Genome Data Storage Footprint. Accessed June 2017.
3. Whole Genome Sequencing BaseSpace App. Accessed June 2017.
4. Illumina (2017) HiSeq Analysis Software v2.1 User Guide. Accessed June 2017.
5. Eberle MA, Fritzilas E, Krusche P, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 2017;27: 157-164.
6. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032-4.

For Research Use Only. Not for use in diagnostic procedures.

illumina®