



**Table 1: Q-Score Bins for an Optimized 8-Level Mapping**

Quality Score Bins	Example of Empirically Mapped Quality 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 is conserved.  
 \*The mapped quality score of each bin (except "N") is subject to change depending on individual Q-tables.

**Benefits of Reduced Quality Scores**

Reduced quality scores lead to a significant reduction in data storage footprints for all compressed sequence formats. We investigated the magnitude of the reduction by measuring the file sizes of a 43x human genome data set. The reduction in data size for compressed \*.bcl files (Illumina raw sequence format) is typically > 50% and the resulting sorted BAM files are reduced by ~30% (Figure 2).

**Test Methodology**

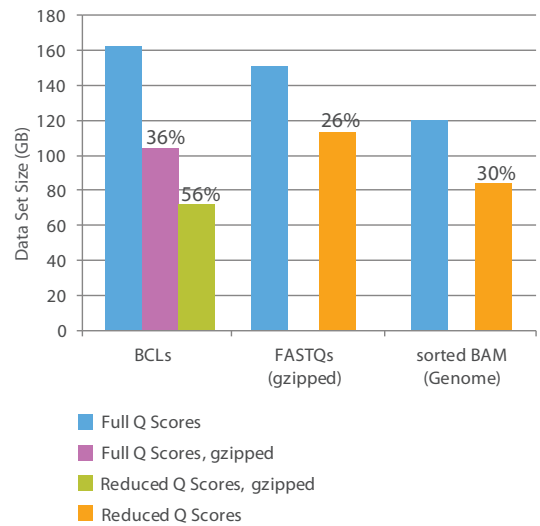
The following investigations were performed to demonstrate the negligible impact of reducing the quality scale to eight levels on analysis results:

- Directly compared the distribution of the old and new scores and quantified the root-mean-square error (RMSE) introduced by the loss of resolution. This approach is completely application-independent.
- Compared the results of simulated data sets of sequencing reads at different Q-score resolutions with called SNPs using a probabilistic SNP caller.
- Used actual sequencing data from human whole-genome sequencing and analyzed the data at full and reduced resolution using the Illumina CASAVA analysis pipeline and the widely used BWA and GATK tools.

**Comparison of Reduced and Full Resolution Quality Distribution**

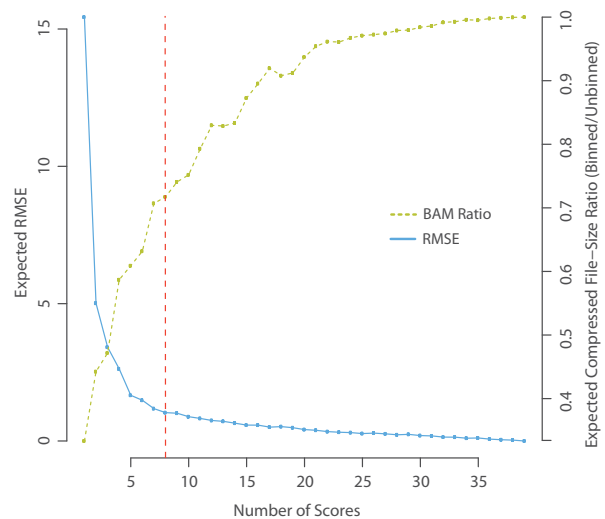
We find that with an 8-level binning (dotted line in Figure 3), the RMS difference between full and reduced distributions is 1.03, or only around 1 Phred score. This low deviation is thus no larger in magnitude than deviations from the underlying accuracy of predicted scores and on the same order of magnitude as the rounding errors (0.5) introduced with a full scale of scores.

**Figure 2: Reduced Resolution Q-Scores**



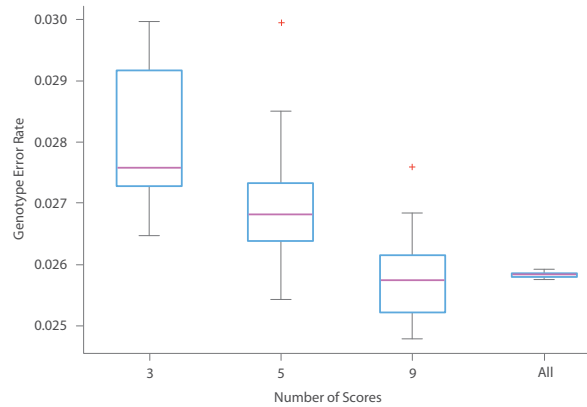
Significant data size reductions in a 43x human genome data set can be accomplished with compression (gzip) and reduced quality scores. The percentages shown are in comparison to the data file sizes of full resolution genomes.

**Figure 3: Information Loss and File-Size Ratio as a Function of Quality Score Bin Number**



With an 8-level binning (red dotted line), the RMS difference between full and reduced distributions was 1.03 or approximately 1 Phred score. Bin boundaries are from minimizing expected error on scaled Q-scores.

**Figure 4: Heterozygous Errors for Varying Q-Score Subsets**



Reducing the number of Q-score bins resulted in a slight increase in median net error (+ 0.17%), even when scores were reduced to 3 bins.

**Table 2: Reduced Resolution Q-Scores**

Resolution	Sensitivity (%)	Conflicts*	Specificity (%)
Full (Elandv2e+CASAVA Variant Calling)	95.29	5,419	99.999788
Reduced (Elandv2e+CASAVA Variant Calling)	95.56	5,940	99.999792
Full (BWA + GATK)	98.40	16,766*	99.999365
Reduced (BWA + GATK)	98.40	17,400*	99.999341

\* Absolute conflict numbers cannot be directly compared, due to the different filters and thresholds used by different tools. It is the relative performance with full and reduced Q-score resolution that is of interest.

### Simulation of SNP Calling with Reduced Q-Scores

To determine that reduced resolution would have no adverse effect on variant calling, a Monte Carlo simulation was performed to create simulated data sets at full and reduced resolution, and to quantify SNP calling performance<sup>5</sup>.

Realistic distributions of coverage and Q-scores were used to generate a sample stack of aligned reads, including base calling errors, as input into a Bayesian allele caller based on the method in CASAVA 1.8<sup>6</sup>. For each “real” genotype, up to 10<sup>8</sup> samples were generated to estimate the probability of genotype error. The simulation excluded variants other than SNPs (such as indels\*) and assumed that base call errors and Q-scores were independent of position in the genome. In particular, alignment effects were not modeled.

A simulation using a reduced number of Q-score bins (20 samplings each with 3, 5, and 9 score bins) showed that there is a slight increase in median error at heterozygous error: 2.58% for 39 bins and 2.75% for three bins<sup>1</sup> (Figure 4). These error rates include sites that would normally not be called in real data, because of very low coverage or a missed allele (i.e. the call rate was forced to be 100%).

### Evaluation of Real Sequencing Data with Different Analysis Pipelines

One way to assess the impact of reduced resolution scores on analysis is to take actual data sets, analyze them using common software packages, and compare the results between full and reduced resolution. An investigation along these lines has recently been published<sup>4</sup> where the impact of reduced Q-scores on a

50 Mbp portion of a human 30x chromosome 2 data set was investigated in detail. With 8 bins of Q-scores, the authors found only a small fraction of discordant SNPs (< 1%), concluding that discordant positions come from marginal decisions between heterozygous and homozygous calls at low coverage. Almost all discordant positions agree with dbSNP and it is not clear which call is correct.

To confirm these results, we took three sets of data from a human trio (mother NA19238, father NA19239, and child NA19240—this data set is currently unreleased). The samples were sequenced using TruSeq<sup>®</sup> chemistry on four lanes of a HiSeq<sup>®</sup> 2000 system, delivering just over 40x coverage per genome. The data were aligned with ELANDv2e and variants were called using CASAVA 1.8.2.

The data was also analyzed with a BWA/GATK workflow. We determined the rate of autosomal Mendelian SNP conflicts in the child as a measure of overall variant calling accuracy. Again we observed no significant difference in accuracy (Table 2).

### Summary

We find no significant differences in either the underlying quality distributions or variant calling performance on human whole-genome sequencing data when we reduce the resolution of high-quality Illumina Q-scores to 8 levels or bins. Variant calling performance of both the BWA+GATK packages and ELANDv2e +CASAVA remains unaffected by the loss of resolution. We propose to enable reduced resolution scores as one of the possible output formats of Illumina sequencers.

### References

- Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8: 186–194.
- Fritz M H-Y, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput sequencing data using reference-based compression.

\* The indel error rate of Illumina sequencing technology is very low. The reduced resolution framework might not be suitable for sequencing platforms with higher indel error rates.  
 † Even though the 9-score bin used in this simulation is different than the 8-score bin formally implemented, the simulation results are not expected to be qualitatively different. In the simulation, genotype calls were attempted irrespective of coverage.



