# Reducing Whole-Genome Data Storage Footprint

Whole-genome data quality score resolution can be reduced without sacrificing score accuracy, or standard analysis and variant calling performance.

## Introduction

The quality of individual bases in sequencing data is usually assessed through the use of logarithmic quality scores. These scores constitute a significant fraction of the total sequencing data storage. As the total volume of sequencing data generated rapidly increases, it becomes important to assess whether the resolution of quality scores can be reduced to alleviate storage requirements.

This white paper examines a method to reduce the resolution of quality scores, enabling a more compact storage of raw sequence reads. Employing a quality scoring scheme with only eight levels of quality or less, the method was tested and found to be virtually lossless. The analysis results showed no significant differences in variant calling from those obtained with a full quality scale.

### Impact of Whole-Genome Base Quality Score Data on Data Storage Requirements

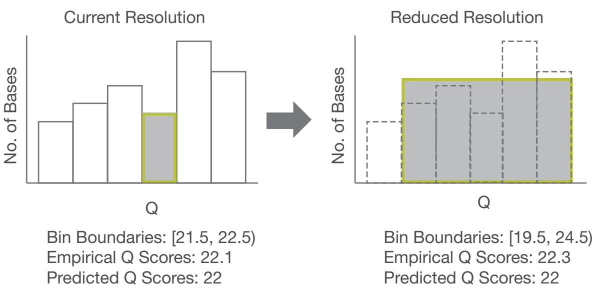
Base quality scores are an integral tool in the analysis of sequencing data, characterizing the level of confidence that can be assigned to the identity of an individual base call. They are routinely used by analysis applications to measure and improve the accuracy of results and determine the biological inferences that can be drawn from the raw sequencing data. For example, quality scores are used in many alignment and variant calling programs.

Quality scores have traditionally been expressed on a logarithmic scale known as Phred scale[1], where the quality score (Q) is derived from the probability of a basecalling error as:

$$Q= -10 \log_{10} p_{error}$$

Q scores are rounded to the nearest integer. High-quality bases can reach Q scores up to Q40 or above, depending on the treatment of the sample prior to sequencing and the sequencing technology itself.

Figure 1: Reducing Q Score Resolution



By replacing the quality scores between 19 and 25 with a new score of 22, data storage space can be freed without sacrificing Q score accuracy.

Q scores take up a large amount of the data storage footprint of a sequencing run. A base is usually expressed as one of four options (A, C, G, T), which corresponds to 2 bits of information. In contrast, 40 quality scores require 5.3 bits of storage, almost three times as much as the base call, before any additional compression is applied.

As the output of sequencing instruments increases, the storage and transfer costs become a much larger part of the total cost of sequencing. The question arises whether the information contained in the quality scores justifies the cost associated with their storage. In actuality, the expression of Q scores as integer values on the Phred scale is an arbitrary convention. The underlying accuracy of these scores is actually lower than the standard resolution of the Phred scale. As a result, a reduction in the scoring scheme to contain fewer levels of quality should yield results that show no significant difference from those obtained with a full quality scale. The reduced Q score output would be transparent across file formats and allow compression algorithms to operate more efficiently due to the reduced complexity of the file.

Several recent publications have explored methods to reduce the data footprint, for example the CRAM[2], cSRA[3], or SlimGene[4] formats. Some of these methods explore the concept of lossy compression of quality data. In many cases, the loss of information is based on alignments or other information that is only available after an initial analysis of the reads. However, quality score resolution can be reduced before alignments are available.

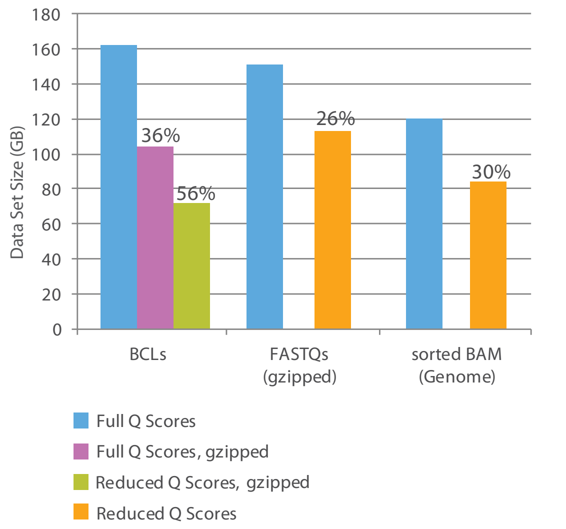
### Quality Score Reduced Resolution Method

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. First, a set of quality bins is selected. For example, the original scores 20-24 may form one bin, with the quality scores in that bin mapped to a new value of 22 (Table 1). This can be thought of as simply replacing all

Table 1: Q Scores Based upon an Optimized 8-level Mapping

| Old Quality Score | New Quality score |
|---|---|
| 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 |

Figure 2: Reduced Resolution Q Scores



Significant data size reductions in a 43× 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.

the occurrences of scores 20, 21, 23, 24 with a new score of 22 in the output sequence file. The choice of bins is empirically optimized to minimize the loss of resolution of the Q scores across most of the data, while simultaneously minimizing the storage footprint (Figure 1).

### Benefits of Reduced Quality Scores

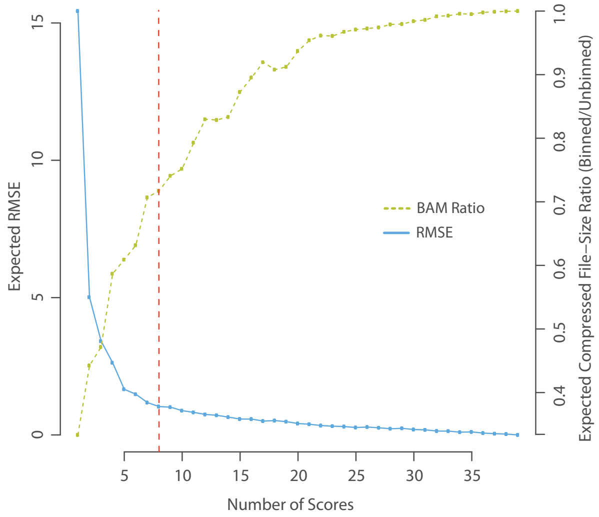
Reduced quality scores lead to a significant reduction in data storage footprints for all compressed sequence formats. We investigated the magnitude of the reduction by measuring the file sizes of a 43× human genome data set. The reduction in data size for compressed bcl files (Illumina raw sequence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

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



With an 8-level bining (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 qscores.

### 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.

### Simulation of SNP Calling with Reduced Scores

To determine that the reduced resolution would have no adverse effect on variant calling, a Monte Carlo simulation was performed to assess the effect of changing quality representation, creating simulated data sets at full and reduced resolution, and quantifing SNP calling performance[5].

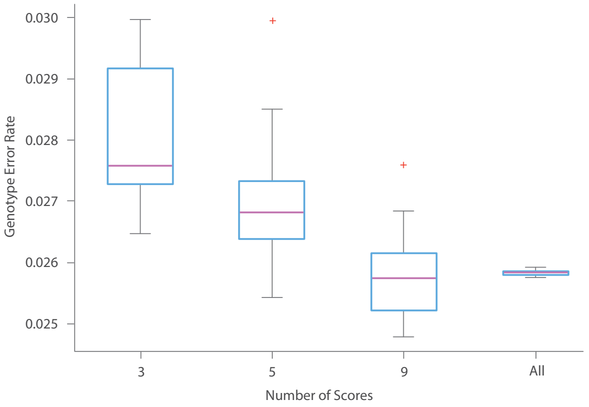
Realistic distributions of coverage and Q scores were used to generate a sample stack of aligned reads, including base calling errors, and input into a Bayesian allele caller based on the method in CASAVA 1.8[6]. For each "real" genotype, up to $10^8$ samples were generated to estimate the probability of genotype error. The simulation excluded variants other than SNPs (such as indels*) and assumed that basecall 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 there is a slight increase in median error at heterzygous error: 2.58% for 39 bins and 2.75% for 3 bins† (Figure 4). These error rates include sites that would normally

* 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.

### 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.

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[4] where the impact of reduced Q scores on a 50 Mbp portion of a human 30× 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 heterzygous 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 is currently unreleased). The samples were sequenced using TruSeq® chemistry on four lanes of a HiSeq® 2000 system, delivering just over 40× 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).

### 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.

## 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

1. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8: 186–194.
2. Fritz M H-Y, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput sequencing data using reference-based compression. Genome Research 21: 734–740.
3. http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=std, http://ftp-private.ncbi.nlm.nih.gov/sra/sdk/2.1.9/sra_sdk-2.1.9.txt
4. Kozanitis C, Saunders C, Kruglyak S, Bafna V, and Varghese G. (2011) Compressing Genomic Sequence Fragments Using SlimGene. Journal of Computational Biology 18: 401–413.
5. Bancarz I (2011) Simulating Quality and Depth of Sequence for Genotype Accuracy. Poster session, 19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB).
6. Illumina technical note, “Improved Accuracy for ELAND and Variant Calling.” Accessed from http://www.illumina.com/documents/products/technotes/technote_eland_variantcalling_improvements.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

Pub. No. 970-2012-013 Current as of 06 June 2012