CALQ: Coverage-adaptive lossy compression of high-throughput sequencing quality values

Jan Voges[1], Mikel Hernaez[2], and Jörn Ostermann[1]

[1]Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover, Hannover.
[2]Department of Electrical Engineering, Stanford University, Stanford.

Introduction

Next-Generation Sequencing (NGS) machines produce a multitude of reads of fragments of DNA material. During the sequencing process, a quality value is assigned to each nucleotide in a read. These quality values express the confidence that the corresponding nucleotide has been read out correctly. After the raw data have been generated, one of the most common subsequent processing step is the reference-based alignment of the reads. As a result, the raw data is further extended to include all the information generated during the alignment process and then stored in the form of BAM files. Moreover, since all the information contained in the raw data is also contained in the BAM file, these files have become the baseline for performing further analysis on the sequencing data. Therefore, efficient storage and transmission of these prohibitively large files is becoming of uttermost importance for the advancement towards precision medicine.

It has been shown that quality values can take up to 80% of the lossless compressed size [2]. To further reduce the file sizes, Illumina proposed a binning method to reduce the number of different quality values from 42 to 8. With this proposal, Illumina opened the doors for allowing lossy compression of the quality values. Previously proposed lossy compressors for quality values are primarily focused on the raw data; and even though those compressors could be easily applied to BAM files, they do not exploit the extra alignment information stored in these files. To the knowledge of the authors, the proposed method CALQ is the first compressor of its type, which exploits alignment information to minimize (or even eliminate) the effect that lossy compression has in downstream applications. Thus, high compression is achieved with a negligible impact on downstream analyses.

Methods

Broadly described, the proposed method first infers the "genotype certainty" at each genomic locus l from the observable data using a statistical model, where the immediate observable data are the read-out nucleotides and the associated quality values of all reads overlapping locus l. The genotype certainty can be regarded as the algorithms' confidence that a given genotype is the correct one. Then, it uses the computed per-locus genotype certainty to determine the level of distortion that the quality values at each genomic locus can handle without affecting subsequent downstream analyses such as variant calling.

To the knowledge of the authors, CALQ is the first compressor of its type, which exploits alignment information to minimize (or even eliminate) the effect that lossy compression has in downstream applications. Thus, high compressions are achieved with a negligible impact on downstream analyses.

Results

For the set of simulations we used the paired-end run ERR174324 of the NA12878 individual. This run was sequenced by Illumina as part of their Platinum Genomes project. The coverage of this data set is 14×. We tested CALQ on chromosomes 11 and 20.

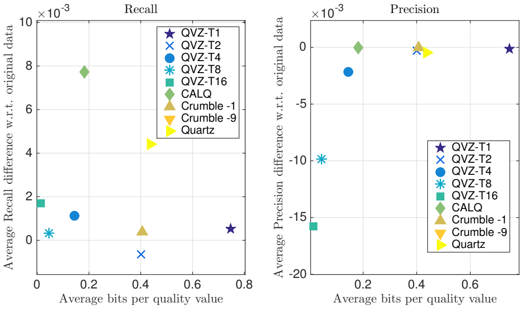

Figure 1: Recall and Precision results for the Illumina data set.

Figure 1 shows the bits per quality value versus the average Precision and Recall over the GATK best practices variant calling pipeline achieved by the proposed algorithm CALQ. It also shows the performance achieved by QVZ 2 [2] (for five different compression modes), Crumble [1] (for two different modes), and Quartz [3]. The values shown in the figure are the result of averaging over four VQSR filtering values (i.e., 90, 99, 99.9, 100) as well as over both chromosomes.

From the figure we can observe that CALQ achieves the best performance in terms of both Recall and Precision. Moreover, CALQ achieves a considerably higher average Recall than that of the lossless case. This means that the variant caller identifies more true positives with CALQ quality values than with the original ones. Note that this is also true for the rest of the lossy compressors, although in a minor scale. This suggests that by applying a lossy compressor more true positives are discovered. Similar findings were shown in [2, 3].

Regarding the Precision, CALQ also achieves the best results, yielding a performance marginally above the lossless case. This improvement with respect to the lossless case is also observed for QVZ 2 T1 and Crumble -1. However, both incur in more bits per quality score. In this regard, CALQ achieves a compressed size of less than 0.2 bits per quality value which is an order of magnitude less than the state-of-the-art lossless compressors.

References

[1] J. K. Bonfield, "Crumble" (https://github.com/jkbonfield/crumble).

[2] I. Ochoa et al., "Effect of lossy compression of quality scores on variant calling", *Briefings in Bioinformatics*, 2016.

[3] Y. W. Yu et al., "Quality score compression improves genotyping accuracy", *Nature Biotechnology*, 2015.