Crumble: reference free lossy compression of sequence quality values.

James K Bonfield [1,*], Shane A McCarthy [1,2] and Richard Durbin [1,2]

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK and
[2] Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK.

[*] To whom correspondence should be addressed.

Abstract

Motivation: The bulk of space taken up by NGS sequencing CRAM files consists of per-base quality values. Most of these are unnecessary for variant calling, offering an opportunity for space saving.
Results: On the CHM1+CHM13 test set, a 17 fold reduction in quality storage can be achieved while maintaining variant calling accuracy.
Availability: Crumble is OpenSource and can be obtained from https://github.com/jkbonfield/crumble.
Contact: jkb@sanger.ac.uk
Supplementary information: Supplementary data are available.

1 Introduction

The rapid reduction of costs for genome sequencing (Wetterstrand, 2016) has led to a corresponding growth in storage costs, far outstripping Moore's Law for CPU and Kryder's Law for storage. This has led to considerable research into data compression (Numanagić *et al.*, 2016).

The most significant component in data storage cost is the per-nucleotide confidence values, which carry information about the likelihood of each base call being in error. Hence this has been the focus of lossy compression research with two main orthogonal strategies: "horizontal" and "vertical".

"Horizontal" compression smooths qualities along each sequence in turn, as implemented in libCSAM (Cánovas *et al.*, 2014), QVZ (Malysa *et al.*, 2015) and FaStore (Roguski *et al.*, 2017) or via quantisation (Illumina, 2014). This type of compression can be applied before alignment and is entirely reference free.

"Vertical" compression takes a slice through an aligned dataset in the SAM format (Li *et al.*, 2009) to determine which qualities to keep and which to discard, as used in CALQ (Voges *et al.*, 2017), or via hashing techniques on unaligned data in Leon (Benoit *et al.*, 2015) and GeneCodeq (Greenfield *et al.*, 2016). Traditional loss measures, such as mean squared error, will appear very high, but these tools focus on minimising the changes in post-processed data (variant calling).

We present Crumble as a mixture of both horizontal and vertical compression. It operates on coordinate sorted aligned SAM, BAM or CRAM files. While this approach does not explicitly use a reference, the sequence aligner does which may result in some reference bias.

2 Methods

A variant caller evaluates the sequence base calls overlapping each genome locus along with their associated qualities to determine whether that site represents a variant. Irrespective of whether the call is a variant, if the same call is made with comparable confidence both with and without sequence quality values present then it can be concluded that the qualities are not necessary in that column.

This requires running the variant caller twice to assess the change, but if limited to sites with high confidence calls the need for a second test can be avoided. We implemented a fast, but naïve, variant caller derived from Gap5's consensus algorithm (Bonfield and Whitwham, 2010). This caller is independent of the major downstream variant callers.

Even when deemed unnecessary, qualities cannot be entirely discarded as tools expect them to exist. By replacing the qualities for bases that agree with a confident consensus call with constant high values, the entropy of the quality signal is reduced. Quality values for bases that disagree with a confident consensus call may optionally be set to a constant low value, heavily quantised, or left intact.

There are sites where any variant caller may incorrectly give the wrong call with a high confidence. We do not wish to replace qualities in such regions. We therefore have a set of heuristics to try to find potentially unreliable calls and retain verbatim the confidence values for these locations and surrounding bases depending on sequence context. Similarly there may be places where an entire read needs to have qualities retained as there is evidence for it being misplaced or being part of a large structural rearrangement.

The heuristics used in Crumble to identify where confidence values should be retained vary by compression level requested, but include:

- **Concordant soft clipping**: many reads having soft clipped bases at the same site often indicates a large insertion (absent in the reference) or contamination.
- **Excessive depth**: possible contamination or collapsed repeat. Variant calls often appear unusually good in such data, even when wrong.
- **Low mapping quality**: possibly caused through poor reference. We optionally can also store quality values for the reads with high mapping quality that colocate with many low mapping quality reads.
- **Unexpected number of variants**: we assume data from a single diploid sample with at most two alleles at each locus. More than two alleles implies misaligned data, duplication or contamination.

Table 1. Effect of lossy quality compression on 50x Syndip data

Category	Original	Crumble-1	Crumble-9p8	Crumble*
Qual size (MB)	5207	614	235	229
SNP False Positive	6495	-249	-240	-479
SNP False Negative	4767	+3	-75	+1
Indel False Positive	3680	-12	+22	-32
Indel False Negative	6912	+6	-61	-53

Comparison of calls filtered by QUAL$\geq$30 to the Syndip truth set. Crumble* refers to parameters optimised for this data set: “crumble -9p8 -u30 -Q60 -D100”. The false positive/negative values of the GATK calls on the crumbled data set are shown relative to the GATK called lossless dataset. The lossless call set has 264,692 SNP true positives and 35,612 indel true positives. The quality sizes are absolute for all files.

- **Low quality variant calls**: typically a single base loci where the consensus is unclear. If a heterozygous indel in a short tandem repeat, the extents of the repeat govern the region in which to retain qualities.

Finally for the quality values that we deem necessary to keep, we provide horizontal compression via the P-block algorithm from CSAM. This is most useful on older Illumina data sets with over 40 distinct levels of quality values.

The nature of the Crumble algorithm makes it amenable to streaming and it does not require large amounts of memory to operate.

3 Results

Analysis of how quality compression affects variant calling was performed on Syndip (Li *et al.*, 2017), an Illumina sequenced library artificially constructed from the haploid cell lines CHM1 and CHM13, with an associated high quality truth set based on two PacBio assemblies (Schneider *et al.*, 2017).

The input BAM file (ERR1341796) had previously been created with GATK best practices including IndelRealigner and Base Quality Score Recalibration steps. To test the impact on raw variant calling, we ran GATK HaplotypeCaller (Poplin *et al.*, 2017), filtering to calls of quality 30 or above, without use of Variant Quality Score Recalibration.

Table 1 shows the lossless results on the CHM pair along with the changes caused by lossy compression using a variety of Crumble options. We chose the minimal compression level, an expected maximum compression level and a set of manually tuned parameters optimised for this data set. The manual tuning traded false positives and false negatives in an attempt to get a call set comparable or better than the original in all regards. It is unknown if the tuned parameters are appropriate for all data sets. More complete comparisons including against other tools are available in the online Supplementary Materials.

On the original BAM file with $\sim$50x coverage we observed a 17 fold reduction in the size of CRAM compressed quality values, while achieving a 7% drop in SNP false positive rate (higher precision) and comparable false negative rates (recall). Indels were marginally improved in both recall and precision. At a sub-sampled 15x coverage the impact is more noticeable; we see a 3% drop in SNP false positive rates and a 12% reduction in SNP false negatives. Indel calls were more comparable, with 1% higher false positives and 3% lower false negatives (see Supplementary Materials).

It is likely these gains to both SNP precision and recall only apply to data coming from a single individual, but they demonstrate the efficacy of lossy quality compression.

4 Conclusion

We have demonstrated that Crumble, when combined with CRAM, can greatly reduce file storage costs while having a minimal, if not beneficial, impact on variant calling accuracy of individual samples. For maximum compression Crumble also permits discarding read identifiers and some auxiliary tags, typically yielding files in the size of 5-10Gb for a 30x whole genome processed with Crumble -9p8.

Acknowledgements

We would like to thank Yasin Memari for help testing and evaluating an earlier version of the program.

Funding

This work was funded by the Wellcome Trust (WT098051).

References

Benoit, G. *et al.* (2015). Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph. *BMC Bioinformatics*, **16**(1), 288.

Bonfield, J. K. and Whitwham, A. (2010). Gap5–editing the billion fragment sequence assembly. *Bioinformatics*, **26**(14), 1699–1703.

Cánovas, R. *et al.* (2014). Lossy compression of quality scores in genomic data. *Bioinformatics*, **30**(15), 2130–2136.

Greenfield, D. L. *et al.* (2016). GeneCodeq: quality score compression and improved genotyping using a Bayesian framework. *Bioinformatics*, **32**(20), 3124–3132.

Illumina (2014). Reducing whole-genome data storage footprint. Technical report.

Li, H. *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, H. *et al.* (2017). New synthetic-diploid benchmark for accurate variant calling evaluation. *bioRxiv. doi:10.1101/223297*.

Malysa, G. *et al.* (2015). QVZ: lossy compression of quality values. *Bioinformatics*, **31**(19), 3122–3129.

Numanagić, I. *et al.* (2016). Comparison of high-throughput sequencing data compression tools. *Nature Methods*, **13**(12), 1005–1008.

Poplin, R. *et al.* (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv. doi:10.1101/201178*.

Roguski, Ł. *et al.* (2017). FaStore–a space-saving solution for raw sequencing data. *bioRxiv*, page 168096.

Schneider, V. A. *et al.* (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, **27**(5), 849–864.

Voges, J. *et al.* (2017). CALQ: compression of quality values of aligned sequencing data [version 1; not peer reviewed]. Number 6(ISCB Comm J):1382 (poster).

Wetterstrand, K. A. (2016). DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcostsdata, accessed 6th Oct 2017.