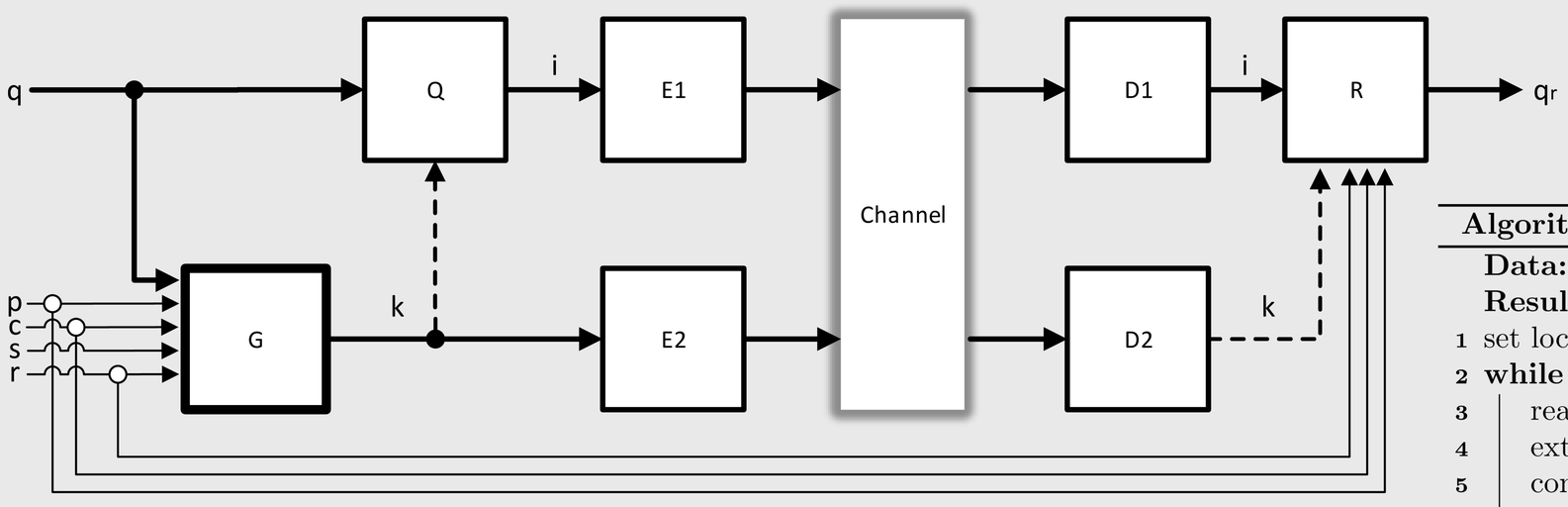
CALQ: compression of quality values of aligned sequencing data

Jan Voges[1] • Jörn Ostermann[1] • Mikel Hernaez[2]
{voges,office}@tnt.uni-hannover.de mhernaez@illinois.edu

Codec structure



Operating principle: exploit the alignment information to infer the acceptable quantization coarseness for quality values

Algorithm 1: CALQ

Data: SAM file and the reference genome with length L
Result: compressed quality values

```
set locus index: l ← 0 ;
while l < L do
    read all reads covering locus l;
    extract CIGAR, POS, QUAL and SEQ for each of these reads;
    compute the genotype uncertainty (see box below);
    compute the index k from the genotype uncertainty;
    quantize the quality values;
    l ← l + 1;
end
compress the quantized quality values using an entropy encoder;
compress the indexes k using an entropy encoder;
```

Legend

p:	mapping position	q:	quality values	G:	genotyping module
c:	CIGAR strings	k:	quantizer indexes	Q, R:	quantization & reconstruction modules
s:	read-out nucleotide sequences	i:	quality value (QV) indexes	E1, E2:	entropy encoding modules
r:	reference sequence names			D1, D2:	entropy decoding modules

Inference of the genotype uncertainty

Goal: Compute the posterior distribution of the genotype G, given the observable nucleotides parameterized by the observable quality values.

Observable data: $\mathbf{n} = \{n_i\}_{i=1}^{N}$ $\mathbf{q} = \{q_i\}_{i=1}^{N}$

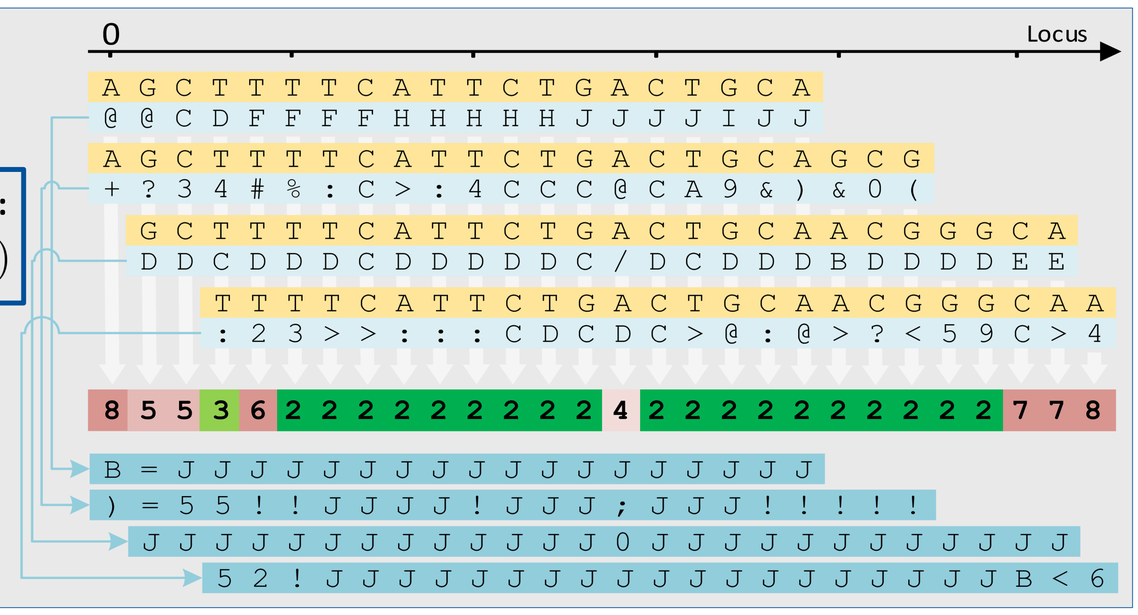
Posterior probability: $P(G|\mathbf{n};\mathbf{q}) \propto P(\mathbf{n}|G;\mathbf{q}) \cdot P(G)$

Quantizer index computation: $k \leftarrow f(M(P(\mathbf{n}|G;\mathbf{q})))$

Likelihood: $P(\mathbf{n}|G;\mathbf{q}) = \prod_{i=1}^{N} P(n_i|G;q_i)$

$$P(n_i|G=(a_1,\ldots,a_h);q_i) = \sum_{\alpha=1}^{h} \frac{P(n_i|A_\alpha=a_\alpha;q_i)}{h}$$

$$P(n_i|A_\alpha=a_\alpha;q_i) = \begin{cases} 1-10^{-\frac{q_i}{10}}, & n_i = a_\alpha \\ \frac{10^{-\frac{q_i}{10}}}{|\mathcal{A}|-1}, & n_i \neq a_\alpha \end{cases}$$



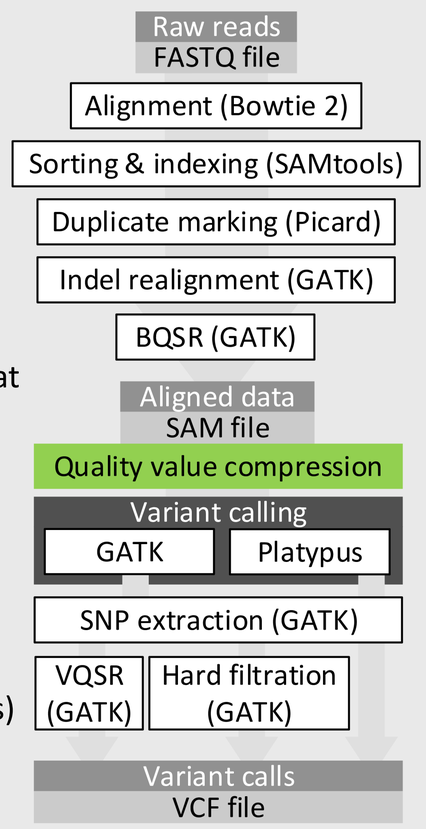
Evaluation

Measuring the impact of lossy quality value compression on variant calling:

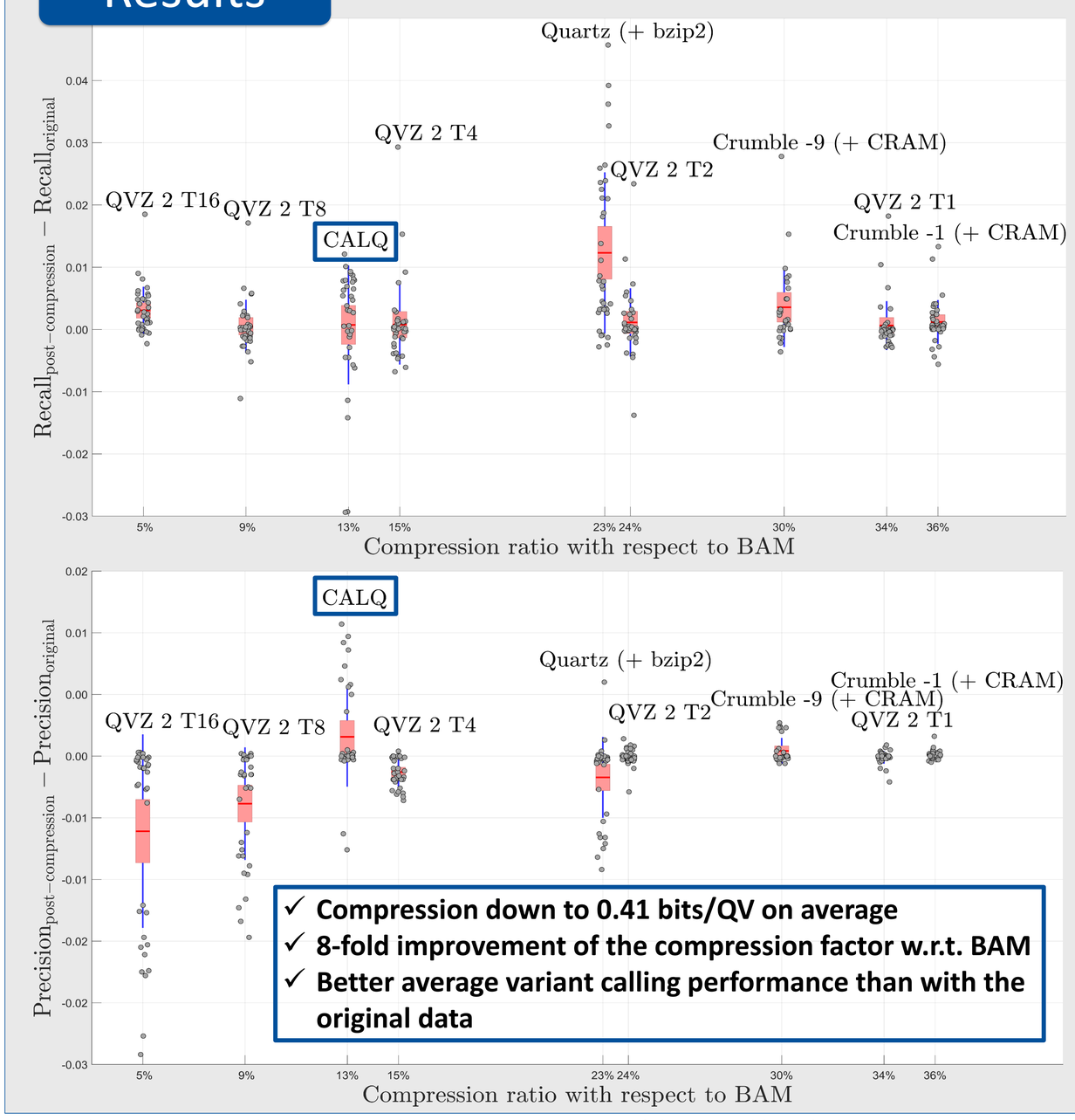
- 3 different variant calling pipelines (GATK + VQSR, GATK + hard filtration, Platypus)
- Comparison of the called variants to the NIST consensus set
  - Restriction to high-confidence regions
  - Recall: the proportion of called variants that are included in the consensus set
  - Precision: the proportion of consensus variants that are called

Data: chromosomes 11 and 20 from individual NA12878, re-sequenced with Illumina HiSeq 2000 @ 14x, Ion Torrent @ 10x, and Illumina HiSeq X @ 49x

➢ 6 (pipelines and filtering values) x 3 (datasets) x 2 (chromosomes) = 36 Recall/Precision vs. bits/QV points per evaluated tool



Results



- ✓ **Compression down to 0.41 bits/QV on average**
- ✓ **8-fold improvement of the compression factor w.r.t. BAM**
- ✓ **Better average variant calling performance than with the original data**

Boxplot semantics: red line: mean, red box: 95 % standard error of the mean, blue whiskers: standard deviation

References

[1] Alberti, C. et al. (2016). An Evaluation Framework for Lossy Compression of Genome Sequencing Quality Values. In Proceedings 2016 Data Compression Conference (DCC 2016), pages 221–230.
[2] McKenna, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303.
[3] Rimmer, A. et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), 912–918.
[4] Hernaez, M., Ochoa, I., and Weissman, T. (2016). A Cluster-Based Approach to Compression of Quality Scores. In 2016 Data Compression Conference (DCC), pages 261–270.
[5] Yu, Y. W., Yorukoglu, D., Peng, J., and Berger, B. (2015). Quality score compression improves genotyping accuracy. Nature Biotechnology, 33(3), 240–243.

Leibniz Universität Hannover
tnt

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

[1]Leibniz Universität Hannover
Institut für Informationsverarbeitung http://www.tnt.uni-hannover.de
Appelstr. 9A, 30167 Hannover, Germany

[2]University of Illinois at Urbana-Champaign
http://www.igb.illinois.edu Carl R. Woese Institute for Genomic Biology
1206 W Gregory Dr, Urbana, IL 61801, USA