INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2017/M40222
April 2017, Hobart, AU

Source	**Requirements**
Status	**Input document**
Title	**Unified representation of sequencing quality values**
Author	Jan Voges (Leibniz Universität Hannover), Claudio Alberti (EPFL), Mikel Hernaez (UIUC), Tom Paridaens (UGent), James Bonfield (Wellcome Trust Sanger Institute), Paolo Ribeca (The Pirbright Institute), Jaime Delgado (UPC)

1 Introduction

This document summarizes the different sets of descriptors used by the technology submitted to Core Experiment 2 “Quality Values”. CE2 is described in the MPEG and ISO/TC 276/WG 5 documents N16724/N153 and N16727/N156.

2 Tools

2.1 QVZ

QVZ computes the optimal quantizers that minimize the rate for a given distortion target under a specific distortion metric [1].

2.1.1 Description

Order-1 Markov statistics are computed on a block basis. Given these statistics, a set of quantizers minimizing the rate for a given distortion target T under a specific distortion metric is computed. The relative position of quality values within a read is used. Furthermore, if quantizer indices and the respective codebooks are transmitted, the read lengths must be known at the decoder.

2.1.2 Compression

An order-1 range encoder is used to encode the modified quality values. A previous version of QVZ encoded quantizer indices and the respective codebooks instead of the modified quality values.

2.1.3 Random access

Random access is not implemented but in theory possible on a block basis (where a block contains N reads).

2.1.4 Parameters

- Distortion metric M
- Distortion target T
- Block size N

2.2 CALQ

CALQ quantizes and compresses quality values in aligned sequence reads based on a metric which is computed per genomic locus. Specifically, the locus-wise genotype certainty is used as metric.

2.2.1 Description

For each genomic locus l, CALQ computes a certainty level and uses this value to select a quantizer from a set of quantizers to quantize all quality values at locus l.

2.2.2 Compression

An order-1 range encoder is used to compress quantizer indices.
An order-1 range encoder is used to compress the quantizer identifiers.

2.2.3 Random access

Random access is possible on a block basis (where a block contains N reads).

2.2.4 Parameters

- Block size N
- Certainty level computation algorithm

2.3 CARGO

CARGO concatenates the quality value strings into a single stream. A second stream is used to store the quality vector sizes.

2.3.1 Description

After the parsing of the record description, quality values are represented as vectors of variable sizes. The current implementation for this data type is to have two separate streams. One stream contains the actual quality values. A second stream contains the quality value vector sizes.

2.3.2 Compression

CARGO uses PPMd level 4 for the encoding quality values. The sizes are compressed with gzip level 4.

2.3.3 Random access

Random access is possible via a range query by using the general mechanism implemented in CARGO. In a generalized description, this in random access on a block basis.

2.3.4 Parameters

- Big block size B
- Small block size S
- Compression buffer size C for PPMd

2.4 CRAM

CRAM concatenates the numerical quality values [2]. The encoder selects a codec from a set of codecs to encode the quality values (mixture of experts). The quality value vector lengths are stored separately.

2.4.1 Description

CRAM concatenates the numerical quality values with no separator. The quality value vector lengths are stored separately. Quality value vectors belonging to reverse complemented reads are added in reverse order. In the case of a "*" in the SAM QUAL field, a vector of quality values 255 is added.

2.4.2 Compression

The encoder tries a set of codecs and learns what is best (with a speed versus size tradeoff controlled by compression level 1 to 9). Standard CRAM typically ends up using the order-1 rANS codec.
The extended set of codecs consists of:

- rANS (order-0 or order-1)
- Dedup (escape mechanism to state "copy QVs from the last record") + rANS
- BSC codec (mix of BWT and entropy coding)
- Fqz codec (adaptive QV statistics + byte-wise arithmetic coder)

2.4.3 Random access

Specified to the encoding program, as number of records (sequences) and/or number of bases per slice. Defaults to 10,000 sequences, but results were given at 1k, 10k, and 100k sequences.

2.4.4 Parameters

The encoder can be adjusted to indicate which compression codecs may be used, as well as random access granularity.

2.5 Crumble

Mixes pileup/column-oriented reductions with row-oriented reductions. The simplified qualities are then fed to CRAM.

2.5.1 Description

Vertical compression
Crumble changes the quality values before encoding. Crumble uses a variant caller derived from Gap5's consensus algorithm. Furthermore, there are numerous heuristics to spot potentially poor alignments which may indicate that an intelligent caller such as GATK HaplotypeCaller could do realignment and change the bases within any specific column. To do this it has a fast short-tandem-repeat (STR) finder and if it needs to keep a column then it extends this column over the entire STR plus some slop either side so that any realignment will still be used confidences that we have retained.
There are additional and optional whole-read heuristics, enabled at various compression levels. These are designed to cope with data that may be in the wrong location. The output from this stage is to keep the whole column of qualities or to discard the whole column. It optionally also has code to amend the quality of bases that actively disagree with the call.

Horizontal compression
For horizontal compression, the P-block algorithm is used. Horizontal compression applies to all qualities in all reads.

BD/BI auxiliary tags

Crumble has options to entirely discard auxiliary tags or for BD and BI tags to quantize them down to a constant value or one of two distinct values. These fields are additional quality streams generated by the GATK quality recalibration process. They are used within the HaplotypeCaller, but have minimal impact on Illumina data. Calling PacBio, IonTorrent and ONT data sets requires these fields to be present, but high fidelity is not necessary.

2.5.2 Compression

Crumble is independent from the compressor, but when coupled to CRAM it will probably use rANS order 1. Results were submitted for small-slice rANS and large-slice Fqz codec, with the latter showing little benefit.

2.5.3 Random access

Determined by the CRAM encoding step.

2.5.4 Parameters

Many Crumble parameters to control compression levels and heuristics, plus all the CRAM parameters for the final encode.

2.6 IMEC (AQUA)

Block-based lossless compression using a coding toolset (experts).

2.6.1 Description

Uses a pool of experts that all predict the quality values for each position of a read. Errors are coded. The tool that provides the smallest block size will be selected.

2.6.2 Compression

CABAC is used for the encoding of all parameters and residue values.

2.6.3 Random access

Random access is possible at a CABAC window size granularity.

2.6.4 Parameters

- Block size N
- Prediction window size (in number of reads)
- Actual list of experts

2.7 FAPEC

FAPEC uses differential coding followed by entropy coding.

2.7.1 Description

For the first quality value line, differences between consecutive characters (i.e. quality values) are coded. For subsequent lines, the first 14 characters are coded versus the same character of the previous lines. For the next characters of the line the consecutive differences are coded.

2.7.2 Compression

FAPEC's entropy coding kernel is used for the coding of residues.

2.7.3 Random access

Random access is possible on a block basis (where a block contains N reads).

2.7.4 Parameters

- Block size S, in bytes, leading to non-uniform N reads per block

3 Main concepts

Context-based compression; random (i.e., non-sequential) access is supported on a block level.

3.1 Raw reads

We assume that we encode quality values as concatenated vectors of 8-bit integers.
The read length must be available at the decoder.

Descriptor	Semantics	Comments
QV	The quality value	8-bit integer

3.2 Aligned reads

3.2.1 Horizontal mode

The same descriptors used for raw reads are used.

3.2.2 Vertical mode

This mode applies only when using quantization of quality values. It requires that the absolute position of each quality value on the reference sequence is available at the decoder.

Descriptor	Semantics	Comments
QuantizationIdx	Quantization index to reconstruct a (quantized) QV	N bit
CodebookIdx	The index of the codebook used for all QVs at a given absolute position	N bit

4 References

[1] G. Malysa, M. Hernaez, I. Ochoa, M. Rao, K. Ganesan, and T. Weissman, "QVZ: lossy compression of quality values," *Bioinformatics*, vol. 31, no. 19, pp. 3122–3129, 2015.

[2] J. K. Bonfield, "The Scramble conversion tool," *Bioinformatics*, vol. 30, no. 19, pp. 2818–2819, 2014.