jmpegg 1.0 – the pure java implementation of the MPEG-G ISO/IEC 23092 standard.

D. Repchevsky[1,2], D. Naro[1], R. Royo[1,2], S. Llorente[3], J. Delgado[3], J.L. Gelpi[1,2,4]

[1] Barcelona Supercomputing Center (BSC), Barcelona Spain.
[2] Spanish National Bioinformatics Institute - Instituto Nacional de Bioinformatica (INB-ISCIII), Barcelona Spain.
[3] Universitat Politècnica de Catalunya (UPC), Barcelona Spain.
[4] Universitat de Barcelona (UB), Barcelona Spain.

High-throughput sequencing technologies consistently produce overwhelming amount of genomic data. This data is extremely important for the biomedical research and has an enormous impact on the progress in healthcare system. Considering the amount of generated sequencing data, data storage capacity became the principal concern for organizations like European Genome-phenome Archive (EGA) or Database of Genotypes and Phenotypes (dbGaP). The efficient and standard storage format is an important part of forging interoperability between different organizations. Several global initiatives such as ELIXIR and GA4GH already established the strategic partnership in standardization of genomic data formats and APIs.

On the other hand, other initiatives to provide a secure and efficient compressed format have appeared, such as MPEG-G (ISO/IEC 23092), developed by the MPEG working group of ISO (ISO/IEC JTC1 SC29/WG11). MPEG-G, also supported by ISO/TC 276 on Biotechnology, is making efforts for a better alignment and integration with approaches taken by other initiatives such as GA4GH.

ISO/IEC 23092 MPEG-G Standard

MPEG-G is the result of technologies proposed and evaluated by MPEG core experiment N16526 which evaluated the performance of compression methods and formats proposed by different groups and organizations. These experiments led to the current MPEG-G specification which consists of five parts. Each part covers one aspect of the standard. We concentrated on two main parts of the specification:

Part 1: Transport and Storage of Genomic Information

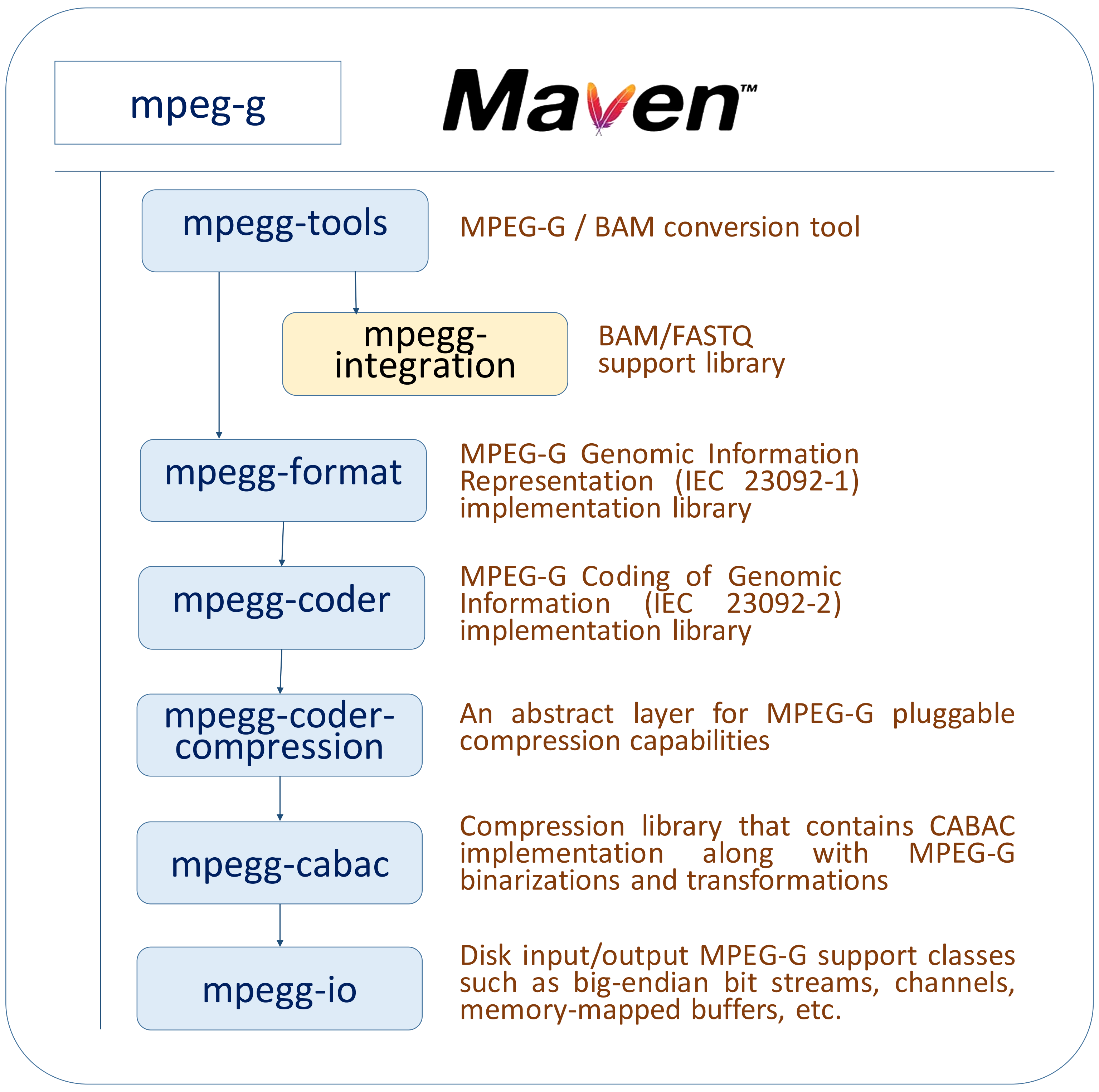
Specifies MPEG-G container file format. The defined structural blocks convey information such as genome references, units of encoded information, metadata and security information.

- Selective access to compressed data
- Data streaming
- Compressed file concatenation and aggregation

Part 2: Coding of Genomic Information

Specifies MPEG-G data decoding and decompression.

- Reference-less and reference based coding
- Reads classification (perfectly mapped, substitutions only, unmapped, etc.)
- Lossless and lossy quality values
- Tokenized read names
- Block level transformations (RLE, LEMPEL-ZIV)
- Sub-symbol level transformations
- Context-adaptive binary arithmetic coding (CABAC)


mpeg-g

Maven™

mpegg-tools — MPEG-G / BAM conversion tool

mpegg-integration — BAM/FASTQ support library

mpegg-format — MPEG-G Genomic Information Representation (IEC 23092-1) implementation library

mpegg-coder — MPEG-G Coding of Genomic Information (IEC 23092-2) implementation library

mpegg-coder-compression — An abstract layer for MPEG-G pluggable compression capabilities

mpegg-cabac — Compression library that contains CABAC implementation along with MPEG-G binarizations and transformations

mpegg-io — Disk input/output MPEG-G support classes such as big-endian bit streams, channels, memory-mapped buffers, etc.

jmpegg 1.0 library

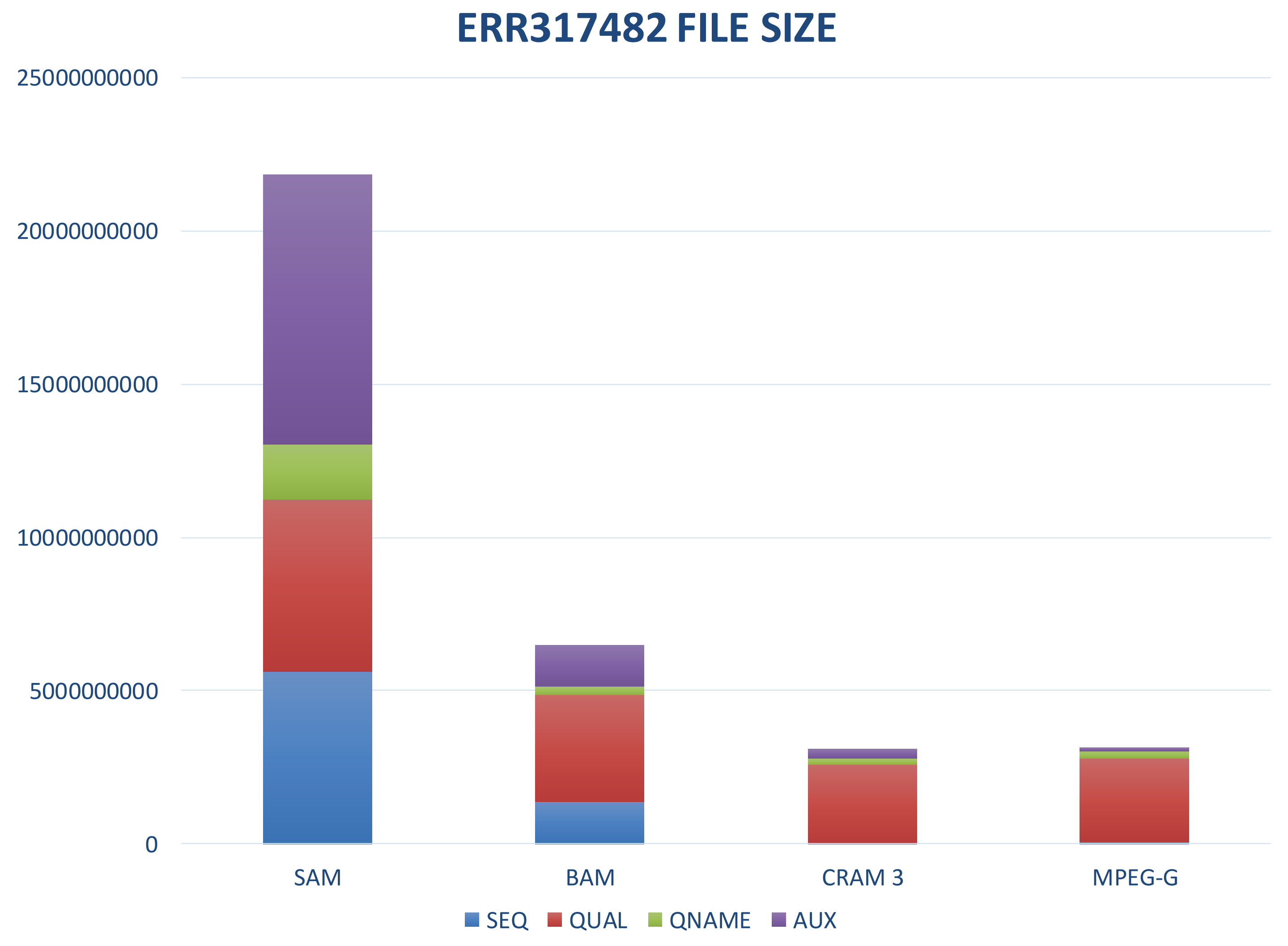
The jmpegg library uses Apache Maven build system and consists of several modules. The library implements ISO/IEC DIS 23092-1 and ISO/IEC DIS 23092-2 parts of the specification. MPEG-G Part 2 can be arbitrarily divided into two parts: the coding and the compression and is implemented in separated modules.

Compression of ERR317482 WGS* (bytes)

	SEQ	QUAL	QNAME	TOTAL
SAM	5.620.980.674	5.620.980.674	1.808.847.284	21.843.078.465
BAM	1.379.499.539	3.480.836.321	289.786.572	6.492.665.954
CRAM 3*	---	2.578.307.514	229.640.572	3.108.947.072
MPEG-G	59.904.176	2.715.591.522	243.032.431	3.123.280.143

* ISO/TC 276/WG 5 N155
Database for Evaluation of Genomic Information Compression and Storage
ID:05 – "http://www.ebi.ac.uk/ena/data/view/ERR317482"

* CRAM noref, lzma, bzip2, seqs_per_slice=100000


ERR317482 FILE SIZE

https://github.com/jmpeg-g/jmpeg-g

dmitry.repchevski@bsc.es

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

FEDER
Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Universitat de Barcelona