INTERNATIONAL ORGANISATION FOR STANDARDISATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC29/WG11

CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2016/M38890

October 2016, Chengdu, China

Source	
Status	**Approved**
Title	**Report of the AHG on Requirements on Genome Compression and Storage (m38890)**
Authors	**M. Golebiewsky (HITS), J. Delgado (UPC), M. Mattavelli (EPFL)**

AHG Mandates

The joint AHG has been established at the Geneva meeting with the following mandates:

1. To disseminate the Call for Proposals and relative information to reach the maximum number of potential contributors.
2. To disseminate the information on the activities carried out by SC29/WG11.
3. To distribute (on hard disks or by downloading from mirroring sites) the MPEG genome database of test data representative of both sequencing technologies and life forms. Contact Claudio Alberti (claudio.alberti@epfl.ch) for the data (N16322).
4. To analyse the answers to the Call for Proposals in the ad-hoc meeting before the 116th meeting.
5. To analyze genome information applications and identify requirements related to file format definition and compression technology.
6. To validate the framework for lossy compression approaches evaluation defined in N16326 to quality scores and define appropriate metrics. To investigate if the existing dataset of genomic data (N16322) should be extended for this objective.

AHG Activity

The activity in the working period between the MPEG meeting consisted in email exchanges mainly focusing on supporting respondents to the CfP (mandates 1 and 2) in answering to the Joint CfP issued at Geneva meeting concerning test material and submission procedures.

No ad-hoc teleconferences and no technical discussions have been registered on the reflector.

Report of the activity during the ad-hoc meeting in Chengdu on the 15th and 16th October 2016

The following contributions have been submitted as answer to the joint CFP:

No.	Title	Authors
m38917	Adaptive lossy compression of high-throughput sequencing quality values	Hannover U.
m38918	Reference-free compression of aligned high-throughput sequencing data	Hannover U.
m38961	Coding and Transport Framework for Genomic Information	GenomSys
m39149	Core Technology Proposal for Genomic Information Coding	GenomSys
m39150	File Format Proposal for Genomic Information Coding	GenomSys
m39151	Transport Layer Proposal for Genomic Information Coding	GenomSys
m39175	GENIFF (GENomic Information File Format), a proposal for a Secure Genomic Information Transport Layer (GITL) based on the ISO Base Media File Format	UPC, CNAG – CRG, BSC, Made of Genes, DAPCOM, The Pirbright I., Stanford, HITS
m39176	A proposal for a compression algorithm based in ORCOM and QVZ. Response to the Joint Call for Proposals for Genomic Information Compression and Storage	UPC, CNAG – CRG, Stanford, BSC, Silesian U. of Technology
m39179	A proposal for a compression algorithm based in FAPEC. Response to the Joint Call for Proposals for Genomic Information Compression and Storage	DAPCOM Services, UPC, BSC, Made of Genes
m39184	Proposal for an efficient genomic information representation	SIB, EPFL
m39200	Proposal of a genomic data file compression framework, based on existing MPEG practices and technologies	U. Gent
m39205	CRAM v3 plus extensions	Wellcome Trust Sanger Institute
m39301	Lossless and Lossy Compression of FASTQ and SAM files	Simon Fraser U., MIT
m39303	Context based lossless compression of aligned reads	Stanford University
m39441	Considerations on “compressed” formats for storing genomic information: a call for flexibility	Pirbright Inst. CNAG

Other contributions not answering to the CfP are:

No.	Title	Authors
m38916	Benchmark framework for lossy compression of genome sequencing quality values	Hannover U., Stanford, EPFL, MIT, Simon Fraser U., Wellcome Sanger Institute
m39526	Insights into more accurately measuring impact of quality score compression	PetaGene

All contributions answering the CfP have been presented and discussed at the AhG meeting except m39200. It will be discussed during the MPEG meeting when the author will be present.

Detailed Review of the input documents answering to the Joint CfP

M38917 Adaptive lossy compression of high-throughput sequencing quality values

Proposal for lossy compression of QV : 1) source model (adaptive quantizer for QV according to the genotype certainty) 2) entropy coding (context adaptive probability model).

Comments : non-normative encoding tool, constrained quantization, entropy coding open to core experiments (runlenght + arithmetic coding)

Performance compression ratio 4.1% up to 19 %

Identified core experiment: QV lossy compression.

M38918 Reference-free compression of aligned high-throughput sequencing data

Proposal for the representation and lossless compression of nucleotide sequences. Sliding window of sequences without a reference. Source model for mapping positions, cigar strings, 3 record types I-Records, M-records, P-record. Compression results provided. Features supported: low complexity, streaming based on I-records.

Comments : 14 classes of descriptors, source models and entropy coders are open to core experiments. Queries are based on reference. Arithmetic coding and ZLIB are used for entropy coding. Flex, pairing information and QV are not coded, read names.

Identified core experiment: Read compression.

M39184 Proposal for an efficient genomic information representation

Proposal for an efficient file format representation of sequencing data and metadata and lossless compression of QV. The representation focus on a common representation for analysis (Variant Calling). It is based on classifying reads in terms of 5 classes perfect, matching and type of matching errors that can be selectively accesses in the analysis pipelines. It supports random access, concatenation of compressed data, integrity checks. It proposes a number of descriptors to represent

the data reads. 3 blocks represents the reads, the headers and the QV for perfect matching, 5 blocks for the other categories. Generic source models and ZIP based entropy coders. It proposes a range encoder lossless QV compressors. Arithmetic coding is based on a 3 dimensional adaptive predictor table for the probability.

Identified core experiments: Lossless compression of QVs, specification of gnomic information representation (file format).

M38961 Coding and transport framework for genomic information

The proposal covers all main aspects of the CfP (not in all metadata and QV), representation, compression, selective access for file format and transport. High level description of the framework and core technology. The goal is an efficient compressed representation of genomic data: based on a data classification in 5 classes based on matching accuracy with a reference, syntax elements, a specific transport layer and file format. Based on layers of homogeneous data, access units, and a specific file format for compressed information. 11 descriptors are used for representing the 5 classes of data. Transport layer is based on packets of data maintain the separation in classes of data and access units. Access units are based on containers of compressed data that can be accessed independently. Access units are classified in different types according to the dependency for accessing data and the nature of the data contained.

Functionality supported are: selective access to data classes and sequence descriptors. It supports natively parallel processing.

Core experiment identified: transport and access unit format.

M39150 File Format Proposal for Genomic Information Coding

The contribution proposes a file format for genomic information structured into Access Units, data layers and data blocks. Selective access to compressed information is the main focus. Selective access is based on the concept of a Master Index table stored in the file header that maps all access units. A local Index table report the mapping of the access units into the physical data in each layer.

The concept of genomic multiplexer is introduced always based on the access unit concept for supporting data streaming scenarios and functionality.

Benefit are: data classes and descriptors allows different source models and associate different entropy models. Data can be incrementally updated in the compressed domain. Different type of selective access, transcoding to different reference, selective encryption.

All information from BAM/SAM file is represented, except qnames and mapping quality, auxiliary fields.

Core experiments identified: file format, data classes and specification of genomic information representation.

M39151 Transport Layer Proposal for Genomic Information Coding

The proposal reports the specification of the transport layer for the core technologies reported in M39861, M39150.

Core experiments: transport functionality, access units definition.

M39149 Core Technology Proposal for Genomic Information Coding

The proposals summarizes the core technologies reported in M39861, M39150 and M39151 that are used to define the compressed genome information representation and how it supports the compression requirements of the CfP.

Core experiment identified: source models for data classes and entropy coders

M39303 Context based lossless compression of aligned reads

Proposal for an efficient genomic information representation based on aligning and referring to a reference. It proposes 7 descriptors with associated source models for the reconstruction of the nucleotide sequences. Sorce models of the descriptors include inter lists contexes for the definition of the probabilities of the entropy coders. It does support reads and position, strand information, reference names, not supported pairing, some of the SAM information contained in the cigars. Arithmetic coding is used for the entropy coding.

Core experiments identified: definition of compression source models, entropy coders, probability contexts.

M39176 A proposal for a compression algorithm based in ORCOM and QVZ. Response to the Joint Call for Proposals for Genomic Information Compression and Storage.

The contribution proposes a compression algorithm for FastQ files (unmapped reads) and QV. The base concept is to separate information in sequences and compress them in interleaved streams. Better compression is achieved by separating homogeneous sub-streams of descriptors for reads and QV metadata. Concerning QV it is based on the QVZ algorithm and achieves any point in the model RD curve. Based on 1st order Markov model and quantized bins values. The compression can range from lossless to full lossy QV.

Core experiments identified: source model for nucleotide sequences, associated entropy coders. QV rate distortion compression performance and entropy coder for quantized QVs.

M39179 A proposal for a compression algorithm based in FAPEC. Response to the Joint Call for Proposals for Genomic Information Compression and Storage.

This contribution proposes a coding technology derived from space data processing and compression. The focus is on processing speed not on high compression. Based on a generic entropy coder. Compression is applied to chunks to implement quasi-random access. The proposal includes encryption and supports multithread implementations. The proposal addresses FastQ data streams compressed into chunks including blocks of a fixed size. Chunks can be compressed and decompressed independently. Lossless compression of FastQ files.

Core experiments identified: transport layer functionality, file format functionality, source model compression.

M39175 GENIFF (GENomic Information File Format), a proposal for a Secure Genomic Information Transport Layer (GITL) based on the ISO Base Media File Format

The contribution proposes a file format for genomic representation based on ISOBMFF file format for representing and transporting genomic information. Although ISOBMFF has been designed for content that has a temporal order is taken as format to include aligned data. The file format considers not only the transport of data, but also the metadata and security issues associated. Some concepts of the CARGO proposals have brought to the ISOBMFF format such as the "Box" structure. The representation of a SAM file content including privacy information is described. Similar approach can be provided with compressed genomic information such as BAM. Another examples described a compressed information obtained by ORCOM or FAPEC and carried by GENIFF.

Metadata is carried in the format by specific boxes. Conversion of SAM headers to XML has been shown. MIAME and MIxS metadata are included in the format as example. XML can be used to include metadata in the GENIFF file format.

Security and privacy issues can be supported by defining security mechanism, the type of encryption used and the rules for accessing the data by XACML rules. Different use cases have been defined and some privacy rules have been associated to the different use cases.

A tool to generate a GENIFF file is provided and partial encryption is demonstrated with a BAM file.

The 5 requirements of transport (3.1-3.5) are satisfied. The technical elements that are necessary to support the different requirements need to be further specified. Transport mechanisms are not specified by the GENIFF format.

The support of different standard metadata could be provided by a core schema.

Core experiments identified: transport layer functionality.

M39441 Considerations on "compressed" formats for storing genomic information: a call for flexibility

The contribution emphasizes the need to support different sequencing technologies particularly interesting and important for do-novo genome assembly based applications.

Compressibility of Illumina is easy because of the redundancy of the generated data. This is not the case of third generation techniques. Third generation need to include also machine settings. Other formats of genomic information should be considered beyond FASTA/Q SAM VCF, such as BED, GTF/GFF, Wig/BigWig. Some existing formats are ill-defined. Third generation is based on different file formats. More flexible solutions are based on compressed databases. CARGO is a generator of compressed databases for genomic information. The basic idea of CARGO is a container of experiments, compressed databases and different file formats can be represented as different schemas. Records are converted to streams which are represented as streams of blocks. Strems can be interleaved and can come from different experiments. Block are checked for consistency by CRC. Block can be scrambled and encrypted. The data contained in the container is specified in contained headers. CARGO implementation supports FASTQ and SAM files. Methods to compress fields are user specified. The framework supports most used generic compressors and selects the best compressor for each stream. The recommendation is to support also 3rd generation quality value schemes.

Core experiment identified: CARGO can be used as file format specification. Compression for selection of best compressors for each SAM field.

M39205 CRAM v3 plus extensions

The proposal presents the file format and toolset of CRAM 3 specification with the toolchain that comes with it. Support for aligned and not aligned data. Support for different codecs. Support for lossy compression names and QV.

The data structure is based on containers, containers include slices which contains blocks. Requirements for non aligned data are covered except querying. (Yes 1.1 to 1.8). Requirements for aligned data are all covered (2.1 to 2.9), but no all auxiliary data can be recovered. Transport requirements are supported. Standalone toolchain is Scamble, CRAM which is integrated and Samtools. CRAM filter is a tool to selectively access specific part of the data. Crumble is the lossy compressor for QV and read-names. Results are provided for a few data of the MPEG database.

Core experiments identified: Lossy compression of QV and read-names. Compression of aligned data.

M39301 Lossless and Lossy Compression of FASTQ and SAM files

Contribution presenting 3 coding tools, SCALCE is compressor for FASTQ, Quarz is the compressor for lossy QV and DEEZ is the SAM/BAM data. The three compressors are documented in publications. SCALCE is a pre-processor that reorder data before applying general purpose compressors. It supports only constant read length it preserved pairing information.

Quartz relies on 32 k-mers classification generated by training and distance of the QV strings from them is coded.

DeeZ is the SAM/BAM compressor that relies on locally assembled contigues, then differences between contigues and local references are coded. Data are divided in three categories for compression.

Results are provided for most of the dataset under consideration.

Core experiments identified: lossless compression for reads and QV and read-names. Lossy of QVs.

Overview of requirement satisfaction for the answer to the CfP according to the self-assessment

The following tables reporting the satisfaction of the Call requirements have been compiled by reviewing the self assessment results provided in the answers of to the CfP.

	Unaligned Reads (Constant Length)								
	Read Headers		Pairing (1.3)	Sequences (1.4)	Quality Values		Parallel Processing (1.7.1)	Querying (1.7.2)	Fields Association (1.8)
	Lossless (1.1)	Lossy (1.2)	Lossless	Lossless	Lossless (1.5)	Lossy (1.6)			
ORCOM				X			X		
FAPEC	X		X	X	X		X	X	X
SCALCE			X	(X)					
Quartz						X			
QVZ					X	X	X		
HEGIC			X	X			X	X	X
CARGO	X	X	X	X	X	X	X	X	X
CRAM	X	X	X	X	X	X	X	X	X

	Unaligned Reads (Variable Length)								
	Read Headers		Pairing (1.3)	Sequences (1.4)	Quality Values		Parallel Processing (1.7.1)	Querying (1.7.2)	Fields Association (1.8)
	Lossless (1.1)	Lossy (1.2)	Lossless	Lossless	Lossless (1.5)	Lossy (1.6)			
ORCOM				X			X		
FAPEC	X		X	X	X		X	X	X
SCALCE									
Quartz									
QVZ					X	X	X		
HEGIC			X	X			X	X	X
CARGO	X	X	X	X	X	X	X	X	X
CRAM	X	X	X	X	X	X	X	X	X

	Aligned Reads (Constant Length)								
	Read Headers		Pairing (1.3)	Sequences (1.4)	Quality Values		Parallel Processing (1.7.1)	Querying (1.7.2)	Fields Association (1.8)
	Lossless (1.1)	Lossy (1.2)	Lossless	Lossless	Lossless (1.5)	Lossy (1.6)			
CBC			X	X			X		X
TSC	X		X	X	X		X	X	X
UGent				X	X				X
DeeZ	X		X	X	X			X	X
QVZ					X	X	X		
Quartz						X			
Calq						X			
CARGO	X	X	X	X	X	X	X	X	X
CRAM	X	X	X	X	X	X	X	X	X
HEGIC			X	X			X	X	X

	Aligned Reads (Variable Length)								
	Read Headers		Pairing (1.3)	Sequences (1.4)	Quality Values		Parallel Processing (1.7.1)	Querying (1.7.2)	Fields Association (1.8)
	Lossless (1.1)	Lossy (1.2)	Lossless	Lossless	Lossless (1.5)	Lossy (1.6)			
CBC									
TSC	X		X	X	X		X	X	X
UGent			X	X					X
DeeZ	X		X	X	X			X	X
QVZ					X	X	X		
Quartz									
Calq						X			
CARGO	X	X	X	X	X	X	X	X	X
CRAM	X	X	X	X	X	X	X	X	X
HEGIC			X	X			X	X	X

	Aligned Reads								
Req. no.	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9
TSC	X	X	X		X	X	X	X	X
CBC	X	X				(X)	X		
DeeZ	X	X	X		X	X	X	X	X
Ugent	X								
Calq			X						
HEGIC	X	X	X	X	X	X	X		
CARGO	X	X	X		X	X	X	X	X
CRAM	X	X	X		X	X	X	X	X

Req. No	3.1	3.2	3.3	3.4	3.5
UGent					X
TSC	X				X
GENIFF	X	X	X	X	X
CARGO	X	X			X
CRAM	X				X
HEGIC	X	X	X	X	X

AHG Recommendations

Continue the assessment of the answers to the CfP in more details, in particular:

- rank performance of the technologies answering the call for the different categories of requirements
- identify core experiments assessing the merit of the different technologies in terms of requirement satisfaction and performance

Identify and specify core experiments:

- Compression technology for sequences (lossless)
- Compression technology for “Read Names” and “QV” (lossless)
- Compression technology for “QV” (lossy) according to the specified evaluation framework procedure and criteria (m38916)
- Specification of the representation of genomic information descriptors and their compression requirements
- Specification of the compressed representation to satisfy the requirements for access, storage and transport (File Format)
- Definition of the transport layer