**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/WG 11 N16136**

**San Diego, USA – February 2016**

|  |  |
| --- | --- |
| **Source:** | **Requirements**  |
| **Status:** | **Draft** |
| **Title:** | **Draft Call for Proposal for Genomic Information Compression and Storage** |

[1 Introduction 2](#_Toc444302068)

[2 Call for Proposals background 2](#_Toc444302069)

[2.1 Compression as a critical element in genomic information processing 2](#_Toc444302070)

[2.2 Elaboration of genomic information compression requirements 3](#_Toc444302071)

[2.3 Collaboration between ISO TC 276/WG 5 and ISO/IEC JTC 1/SC 29/WG 11 (MPEG) 3](#_Toc444302072)

[2.4 Results from a Call for Evidence on genome compression 3](#_Toc444302073)

[3 Open standard development process 5](#_Toc444302074)

[4 Technology Solicited by this Call 6](#_Toc444302075)

[5 Source Code and IPR 6](#_Toc444302076)

[6 Timetable and Procedures 7](#_Toc444302077)

[6.1 Registration 7](#_Toc444302078)

[6.2 Items to be submitted 7](#_Toc444302079)

[6.3 Evaluation Criteria 8](#_Toc444302080)

[6.4 Participation 8](#_Toc444302081)

[6.5 Preliminary Evaluation 8](#_Toc444302082)

[6.6 Selection of Technology 8](#_Toc444302083)

[6.7 General Model and Core Experiments 9](#_Toc444302084)

[6.7.1 General Model 9](#_Toc444302085)

[6.7.2 Core Experiments 9](#_Toc444302086)

[7 Call for Proposals Contact information 9](#_Toc444302087)

[8 References 10](#_Toc444302088)

# Introduction

The development of Next Generation Sequencing (NGS) technologies opens the usage of genomic information as everyday practice in several fields, but the growing volume of data generated becomes a serious obstacle for a wider diffusion. An appropriate representation and an efficient compression of genomic data is widely recognized as a critical element limiting its application potential [1] [2]. Beside compression which is at the base of any efficient processing and storage of genomic information, other requirements have been identified that current raw data and aligned data formats do not fulfill (see annex A to this document).

ISO TC 276 works on standardization in the field of biotechnology processes that include analytical methods (Working Group 3) and data processing and integration (Working Group 5).

ISO/IEC JTC 1/SC 29/WG 11 (MPEG) has the mission to develop standards for coded representation and compression of digital audio and video and related data. In its 28 years of activity MPEG has developed many generations of video and audio compression standards.

Results of a recently issued Call for Evidence on genomic data compression technologies [3] have demonstrated that compression rates above the current state of the art can be achieved by integrating and improving algorithms reducing data redundancy for all classes of data and metadata constituting genomic information [4].

By combining their respective expertise, ISO TC 276 and MPEG have the possibility to develop a compression standard capable of providing new effective solutions to the stated problem.

ISO/IEC provides a framework for the development of a genome compression standard based on the following steps:

* an open call to any party possessing technologies satisfying all or a subset of identified requirements
* a fair assessment of the performance of each submission
* the identification of the most promising technologies
* their combination in a General Model as a platform for collaboration
* the progressive improvement of the General Model via Core Experiments
* the approval of the standard following the established ISO/IEC procedure[[1]](#footnote-1)
* the public availability of an informative software to compress genomic information and a normative software to decompress genomic information.

# Call for Proposals background

## Compression as a critical element in genomic information processing

The sequencing of the genetic information of human genome has become affordable due to high-throughput sequencing technology. This opens new perspectives for the diagnosis and successful treatment of cancer and other genetic illnesses. However, there remain challenges, scientific as well as computational, that need to be addressed for this technology to find its way into everyday practice in healthcare and medicine. The first challenge is to cope with the flood of sequencing data. For instance, a database covering the inhabitants of a small country like Switzerland would need to store a staggering amount of data, about 2’335’740 Terabytes. The second challenge is the ability to process such a deluge of data in order to 1) increase the scientific knowledge of genome sequence information and 2) search genome databases for diagnosis and therapy purposes. High-performance compression of genomic data is required to reduce the storage size, increase transmission speed and reduce the cost of I/O bandwidth connecting the database and the processing facilities.

The current trends in sequencing data generation show clearly that the storage and transfer (bandwidth) costs will soon become comparable to the costs of sequencing [1] [2]. This means that IT costs may soon become a major obstacle to such genome analysis applications as personalized medicine, early diagnostics and drugs discovery, unless genetic data compression reduces IT costs on par with sequencing costs.

## Elaboration of genomic information compression requirements

Since July 2014 [5] work on the identification of the most important stages of a typical genomic processing pipeline and the definition of a list of requirements for efficient data compression and storage specific to each stage has been carried out. This work focused on identifying requirements for compression of genomic information in the form of raw reads and aligned/mapped reads produced at the initial stages of genomic information processing pipelines.

Requirements related to the efficient transport of the compressed genomic data have been identified as well.

A consensus has been reached in the scientific community around the notion that lossy compression of the metadata generated by sequencing machines has to be considered [6] [7]. Therefore requirements for solutions addressing both lossless and lossy compression of genomic metadata have been identified.

In this process experts in several domains including bioinformatics, biology, information theory, telecommunication, data compression, data storage and information security have participated.

The requirements identified so far are listed in detail in output document N16134 [8].

## Collaboration between ISO TC 276/WG 5 and ISO/IEC JTC 1/SC 29/WG 11 (MPEG)

The challenge of providing appropriate compression solutions to the genomic data processing problems requires the know-how of different scientific domains ranging from telecommunications to biotechnology, disciplines belonging to traditionally disjoint communities and working groups. Such expertise is present in ISO/IEC JTC 1/SC 29/WG11 (MPEG) and ISO TC 276/WG 5 and WG 3 that will join their efforts in the development of a standard employing leading edge technologies relying on such diverse fields. In particular ISO/IEC JTC 1/SC 29/WG11 will focus the technical work on assessing the answers of this Call for Proposals, selecting and integrating the most valuable technologies into a single compression solution whose performance will be improved and optimized during the standardization process. TC 276/WG 5 and WG 3 will contribute to the validation of the technical solution under development and will provide feedback on how the new standard satisfies relevant use cases.

A joint ad-hoc will be established between ISO/IEC JTC 1/SC 29/WG 11 (MPEG) and ISO TC 276/WG 5 and representative of each committee will be invited to participate to periodic meetings to provide an effective collaboration between the groups.

## Results from a Call for Evidence on genome compression

The purpose of the issued Call for Evidence (CfE) on genome compression [3] was to assess whether new technologies can achieve better performance in terms of compression efficiency compared with currently used file formats such as gzipped FASTQ (used for raw data) and BAM (used for aligned data). Additional functionalities (e.g. non-sequential access, lossy compression efficiency, etc.) are provided by available technologies.

There were two prerequisites for the evaluation of answers to the CfE: the proposals/tools considered had to be under active development and maintenance, and they had to be reliable.

The CfE defined a number of criteria for the evaluation of proposals:

* Compression factor
* Separate assessment of performance for each class of data in the compressed bitstream:
	+ Reads headers/identifiers
	+ Sequence reads
	+ Quality scores
	+ Any other metadata (identified as “auxiliary data”)
* “Reasonable” computational complexity
	+ Encoding and decoding time, respectively
	+ Peak and average memory usage
* Support of a minimal set of functionalities
	+ Non-sequential access
	+ More than 5 symbols (A, C, G, T, N) alphabets
	+ Encoding of additional metadata (extensibility)
	+ Lossy compression of metadata
		- Quality scores
		- Reads identifiers

In the scope of the CfE, 22 tools have been evaluated. Eight tools have been developed by four providers who directly answered the CfE. Figure 1 shows the compression results for unaligned data, for the human low coverage (8x) sample ERR174310.



Figure 1 – Compression results for ERR174310

Figure 2 shows the compression results for aligned data, specifically for the human high coverage (50x) sample NA12878\_S1. Figure 3 reports the same results, but without quality scores. This shows how relevant is considering solutions supporting lossy compression of metadata.



Figure 2 – Compression results for NA12878\_S1



Figure 3 – Compression results (without quality scores) for NA12878\_S1

Finally, the CfE yielded the following conclusions: all tools have different performances for the 4 classes of data. By integration of multiple tools it is possible to improve the compression of up to 27% with respect to the best state-of-the-art tool. During the time slot of the CfE, existing tools as well as new tools improved the obtainable compression factors by about 4%.

Results demonstrate the possibility to achieve compression rates above the state of the art through the combination of new and improved algorithms.

# Open standard development process

The process that will be followed for the development of a new open standard on Genome Information Representation is based on the well-established and successful approach refined during the last three decades:

* A Call for Proposals is issued (this document) which is open to any interested party ready to accept ISO/IEC IP policy (see later in this document); acceptable proposals can satisfy all or only a subset of the requirements; interested parties which are not members of MPEG have the possibility to address any questions and requests of documentation to a dedicated contact person mentioned below in section 7;
* The evaluation criteria and process elaborated by ISO/IEC TC 276 and MPEG experts and approved by the delegates will be published in parallel to the Call for Proposals [8]; furthermore, proponents may provide comments on the requirements and the evaluation process in the context of the application scenarios covered by the Call;
* the assessment of the received proposals will identify either a proposal that will become the General Model, or a set of the most promising technologies that will be combined into the General Model;
* the General Model is intended as an initial step and a platform for further collaboration; throughout the working period that will follow the assessment of the proposals, the General Model will be progressively improved through the specification of several Core Experiments; this process allows integrating further relevant enhancements proposed and identified during the whole period. This working period usually lasts several months;
* at the end of the process described above, the new standard will be approved according to the established ISO/IEC procedure described available online[[2]](#footnote-2);
* in MPEG "standard" typically means a description of the normative sequence of operations to be performed on compressed and/or transported data in order to reconstruct data into their original (uncompressed) form. This is what MPEG calls "decoding" process. A reference implementation of a "decoder" is typically developed and actually assumes a "normative" status. Conversely one or more entire or partial implementations of "encoders" are also developed as reference examples, but they assume only an "informative" status.

# Technology Solicited by this Call

Responses are solicited that propose technologies for Genomic Information Representation covering two aspects:

1. the compression of genomic data generated by sequencing processes (FASTA and FASTQ) so that such data can be compressed and efficiently processed whenever possible and appropriate
2. the compression of aligned genomic data generated by alignment/mapping tools (BAM) so that such data can be compressed and efficiently processed whenever possible and appropriate
3. the definition of a data format for a Genomic Information Transport Layer (GITL) that carries compressed genomic data and metadata generated during the "genomic information life cycle".

For the detailed requirements to be fulfilled by responses refer to the requirements document N16134 [9].

# Source Code and IPR

Proponents are advised that, upon acceptance into the standardization process, they may be required to make available source code software for certain parts of their technology. This code will be included in the standard as Reference Software to be released under the Reference Software Copyright License in annex (N15898). If proponents feel that any aspects of their technology should not be made available in source code, then they should clearly state which aspects and why.

Furthermore, proponents are advised that this Call is being made under the auspices of ISO/IEC, and as such, subject to the ITU-T/ITU-R/ISO/IEC Intellectual Property Rights Policy as approved by the ISO, IEC and ITU councils[[3]](#footnote-3).

# Timetable and Procedures

The following estimated milestones are planned:

* Draft CfP Issued - San Diego, U.S.A., 18th March 2016
* CfP Issued - Geneva, Switzerland, 3rd June 2016
* CfP Responses: Deadline for Submission 26th September 2016.
* Technology identification and selection - Chengdu, China, 21st October, 2016
* Test model and preliminary working Draft – January 2017
* Committee Draft – July 2017
* Draft International Standard – January 2018
* Final Draft International Standard – July 2018

All communications concerning the Call and responses thereto should be addressed to the CfP Contacts (listed below), and communications are preferred in electronic form, via email.

Interested parties should approach the CfP Contacts for assistance regarding all aspects of their submission and subsequent attendance at ISO meetings, which may involve explaining how they can become accredited to attend the meetings.

## Registration

There is no need to register interest in responding to this call.

## Items to be submitted

CfP respondents should submit the following documents:

* A description of the technology having sufficient detail to permit technical discussions.
* A definition of the input format accepted by the encoding technology (and produced by the decoding technology) including:
	+ Raw reads
	+ Aligned/mapped reads
* Evidence of the performance of the technology, including:
	+ Compression factor
	+ Complexity of the decoder in terms of
		- approximate CPU and run-time memory requirements on a platform which must be characterized in detail.
		- any other meaningful metrics the proponents deem appropriate
* The self-assessment template filled with the relevant information for the proposed technology that can be found in the Evaluation Criteria document (N16146) [8]
* If appropriate, the proponent may provide comments on the requirements and the evaluation process in context of the application scenarios or use cases.

Proponents shall email the complete submission documentation to the CfP Contacts no later than 26th September 2016 23:59 PST. The proponent’s documents should be provided in Microsoft Word format.

To support evaluation tests proponents shall submit:

* Executables – Decoders shall be delivered to the CfP Contact as executables on either the Linux/Intel or Windows platforms (statically linked libraries are required for all non-standard libraries). All executables shall have command-line interface (i.e. no GUI).
* Bitstreams - Compressed data shall be supplied corresponding to each individual test item as described in the Evaluation Procedure document [8].

Such items shall be submitted to the CfP Contacts on a physical storage device with USB 2.0/3.0 connectivity. Delivery to the CfP Contacts shall be no later than 26th September 2016 23:59 PST. Proponents shall communicate with the CfP Contacts to discuss any detail of the physical delivery of items.

## Evaluation Criteria

A draft Evaluation Procedure document has been issued at the 114th MPEG meeting [8]. In June 2016 a detailed process for evaluation of the proposed technologies will be finalized based on the draft Evaluation Process and submitted review documents. It is envisioned that the evaluation process will be comprised of the following steps:

* Assessment of the requirement coverage according to metrics specified in the Evaluation Process document
* Analysis of complexity, as measured by both computational complexity and memory requirements.

More information on the procedures used for proposal evaluation can be found in the Evaluation Procedure document [8].

## Participation

Respondents to the CfP should attend the kickoff meeting in Chengdu (China) to present and discuss details of their proposals.

## Preliminary Evaluation

In the period between the proposals deadline and the kickoff meeting, the group will conduct a preliminary evaluation of submissions to check their compliance with the Requirements outlined in this document and procedures described in the associated Evaluation Procedure document. Submissions that are compliant will undergo further evaluation.

## Selection of Technology

At the kickoff meeting, the final selection of the proponent technology that will become General Model Zero (GM0), and which will be the start of the standardization phase, will be based on the judgment and consensus of the experts in the joint working group.

## General Model and Core Experiments

Two working tools play a major role in the collaborative development phase that follows the initial competitive phase: the Working Model and Core Experiments (CE). In the context of this CfP the Working Model is called “General Model” (GM).

The best technology, as identified in the evaluation process, will be selected as GM0 and be the basis for subsequent core experiments. Proponents whose technology is selected as GM0 and all proponents participating in the subsequent core experiment process shall supply a detailed description of their technology.

### General Model

A General Model is a complete framework such that an experiment performed by multiple independent parties will produce essentially identical results. The GM enables the checking of the relative performance of different tools, as well as improving the performance of selected tools. The GM will be built after screening the proposals answering the CfP. The first GM will not the best proposal but a combination of the best tools, independently of the proposal that they belonged to. The GM will include normative and non-normative tools to create the “common framework” that allows performing adequate evaluation and comparison of tools targeting the continuous improvement of the technology included in the GM. After the establishment of the first GM new tools can be proposed and evaluated inside the GM following a core experiment procedure. The GM will evolve through versions as core experiments verify the inclusion of new techniques, or prove that included techniques should be substituted. At each GM version, only the best performing tools will be part of the GM. If any part of a proposal will be selected for inclusion in the GM, the proposer will have to provide the corresponding source code for integration into the GM software in the conditions specified by the ITU-T/ITU-R/ISO/IEC Intellectual Property Rights Policy.

### Core Experiments

The improvement of the GM will start with a first set of core experiments defined at the conclusion of the evaluation of the proposals. The core experiments process allows for multiple, independent, directly comparable experiments to be performed to determine whether or not a proposed tool has merit. Proposed tools target the substitution of a tool in the GM or the direct inclusion in the GM to provide a new relevant functionality or improved performance. Improvements and additions to the GMs will be decided based on the results of core experiments.

A core experiment has to be completely and uniquely defined, so that the results are unambiguous.

In addition to the specification of the tool to be evaluated, a core experiment also specifies the conditions to be used, again so the results can be compared. A core experiment is proposed by one or more experts and is accepted by consensus, providing that two or more independent experts agree to perform the experiment.

It is important to realize that the Core Experiments will not end up in the standard itself, as these are just working tools to ease the development process.

# Call for Proposals Contact information

To register for this call or for any other questions about the call, test conditions, required software or test sequences please contact

Joern Ostermann, MPEG Requirements Group Chair

Email: ostermann@tnt.uni-hannover.de

Martin Golebiewski, Convenor ISO/TC 276/WG5

Email: martin.golebiewski@h-its.org

# References

|  |  |
| --- | --- |
| [1]  | S. D. Kahn, “On the Future of Genomic Data,” *Science,* vol. 331, pp. 728-729, 2011.  |
| [2]  | Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron and G. E. Robinson, “Big Data: Astronomical or Genomical?,” *PLOS Biology,* 2015.  |
| [3]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2015/N15740 - Call for Evidence on Genome Compression and Storage,” Geneva, 2015.  |
| [4]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N15XXX - Results of the Call for Evidence on Genome Compression and Storage,” San Diego, 2016.  |
| [5]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2014/N14553 - Issues in genome compression and storage,” Sapporo, 2014.  |
| [6]  | R. Canovas, A. Moffat and A. Turpin, “Lossy compression of quality scores in genomic data,” *Bioinformatics,* vol. 30, no. 15, pp. 2130-2136, 2014.  |
| [7]  | M. Hernaez, I. Ochoa, R. Goldfeder, T. Weissman and E. Ashley, “Effect of lossy compression of quality scores on SNP,” *Biorxiv,* 2015.  |
| [8]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N16146 - Evaluation Procedure for the Call for Proposals on Genomic Information Representation and Compression,” San Diego, 2016.  |
| [9]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N16134 - Requirements for Genomic Information Representation and Compression,” San Diego, 2016.  |

1. http://www.iso.org/iso/home/standards\_development/resources-for-technical-work/support-for-developing-standards.htm [↑](#footnote-ref-1)
2. http://www.iso.org/iso/home/standards\_development/resources-for-technical-work/support-for-developing-standards.htm [↑](#footnote-ref-2)
3. http://www.itu.int/en/ITU-T/ipr/Pages/policy.aspx [↑](#footnote-ref-3)