 ISO/IEC JTC 1/SC 29/WG 8 N0039

**ISO/IEC JTC 1/SC 29/WG 8  
MPEG Genomic Coding   
Convenorship: SNV (Switzerland)**

**Document type:** Output Document

**Title:** Joint Call for Proposals on New Advanced Features and New Technologies for MPEG-G

**Status:** Approved

**Date of document:** 2021-05-1

**Source:** Convenor, ISO/IEC JTC 1/SC 29/WG 8

**Expected action:** None

**Action due date:** None

**No. of pages:** 10 (with cover page)

**Email of Convenor:** marco . mattavelli @ epfl . ch

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg8>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 8**

**MPEG GENOMIC CODING**

**ISO/IEC JTC 1/SC 29/WG 8 N** **0039**

**April 2021, Virtual**

|  |  |
| --- | --- |
| **Title** | **Joint Call for Proposals on New Advanced Features and New Technologies**  **for MPEG-G** |
| **Source** | **WG 8, MPEG Genomic Coding** |
| **Status** | **Approved** |
| **Serial Number** | **20345** |

# Introduction

The sequencing of the genetic information of human genome has become affordable due to high-throughput sequencing technology [1], [2]. One challenge is to efficiently handle the increasing flood of sequencing data. A 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. Other important features are non-sequential access, integration with metadata, protection of the data supporting controlled selective access, integration with annotation data and indexing capabilities for raw data, metadata and annotations.

All these functionalities are already part of ISO/IEC 23092 (MPEG-G) standard series. However, new compression technologies, particularly applied to very long reads, can yield higher compression rate, support new functionality or improve performance of other metrics.

Descriptions of proposals shall be registered as input documents to the meeting that will take place in January 2022 (see timetable and submission procedure in section 6). Proponents need to attend that meeting to present their proposals. Further information about logistical steps to attend the meeting can be obtained from the listed contact persons (see section 7).

# Call for Proposals background

A call for evidence (CfE) was issued at the 132nd MPEG meeting seeking new advanced features and new technologies for MPEG-G. As a result of the responses received to the CfE, we have issued this CfP.

CfE answers show that technologies based on graph genome are available a need for standardization exists.

Linear reference genome is currently the most prevalent model used in the processing and analysis, such as read alignment and variant calling, of next generation sequencing (NGS) data. It is based on the use of a single, preferred tiling path to produce a single consensus representation of the genome. For example, the linear reference NCBI GRCh38 (Hg38) is a composite genome with approximately 93% of the primary assembly consisting of sequences from 11 individuals. The linear approach has been in the main stream for over a decade due to its simplicity and the general similarity of human genomes. From short read-based assays, it is estimated that the average diploid human genome has between 4.1 and 5 million point mutations, including SNVs and short indels, which is only around 1 point variant every 1200 – 1450 bases of haploid sequence. Such an average genome would also have about 20 million bases (∼0.3%) affected by around 2100–2500 larger structural variants (The 1000 Genomes Project Consortium 2015). These estimates are somewhat conservative though, as long-read sequencing shows an excess of structural variation not found by earlier short-read technology.

Despite its popularity among scientists due to its ease of reference and lower requirement for computational analysis, a single tiling path is insufficient to represent the allelic diversity in complex genomic regions for most mammalian genomes. With a great deal of common genomic variation excluded, it introduces a pervasive reference bias, which has a negative impact on the accuracy of downstream analysis.

Many scientists believe that one of the most effective ways to represent the complex genomic diversity in a reference cohort is through the adoption of a mathematical graph model, where genomic variations are captured as edges associated with different nucleotide sequences. In this way, the latent structure of homologous loci within a reference cohort can be naturally represented in the genome graph. Additional information, such as allele frequency and genotype, can be stored along with the nodes or edges for use during alignment, variant calling or other analyses. In order to use genome graph as an effective spatial framework for organizing and comparing a population of genomes, several considerations regarding coordinates, alleles, ordering in graphs, and genome embedding need to be properly addressed.

The second generation approach for creating genome sequences generates short reads. This is the most commonly used approach today however third generation sequencing approaches are quickly becoming more common. Third generation sequencing produces long reads and support for these long reads is important. The benefit of long reads is improved sensitivity and accuracy for the detection of structural variants, which are known to be important in clinical applications. More and more the importance of structural variants are being recognized and applied in the clinical setting. [3] The current approach adopted in MPEG-G may not be the optimal approach for third generation sequencing due to differences in error rates and read length.

A limitation of the current MPEG-G standard is the lack of support for more complex probability predictors such as neural networks. Currently only the count-based framework is supported where the probability of the next symbol is computed based on empirical probabilities from counts. However, such a count-based approach suffers from two major limitations: 1. Inability to exploit the similarities and dependencies across contexts used as input to the models and 2. Does not work well when the context set is very large (or uncountable) as compared to the data size

Both of these issues can be handled using a neural network/machine learning based approach, which provides a much more natural prediction framework, and are able to work with different types of contexts such as numerical, categorical and ordinal. In some cases, this improved compression can be worth the increased computational complexity, especially in cases when specialized hardware or parallel computation is available.

Standards supporting interoperability can lower barriers to implementation in software systems. MPEG-G, aiming to enable a more efficient compression and transmission of genomic data coupled with metadata and APIs, already supports interoperability with legacy formats for NGS data processing up to and including annotation. However, interoperability with other IT standards remains an open question. Adapting MPEG-G to accommodate well-established technologies in the healthcare IT ecosystem can broaden its appeal by reducing implementation costs and natively supporting a more diverse array of software systems. The HL7 Fast Healthcare Interoperability Resources (FHIR) standard and the Global Alliance for Genomics and Health Phenopackets standard are examples of emerging standards where interoperability with MPEG-G would be beneficial.

# Open standard development process

The process that will be followed by ISO/IEC JTC 1/SC 29/WG 08 (MPEG) for the development of the extensions of the ISO/IEC 23092 standard series, is based on the well-established and successful approach refined during the last three decades of ISO/IEC JTC 1/SC 29/WG 11 (MPEG) work:

* 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/TC 276 and MPEG experts and approved by the delegates have been published in parallel to the Final Call for Proposals [4]; 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 Test Model, or a set of the most promising technologies that will be combined into the Test Model implementing the extensions to the current ISO/IEC 23092 series.;
* The Test 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 Test 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 at <https://www.iso.org/developing-standards.html> and <https://www.iso.org/stages-and-resources-for-standards-development.html> ;
* 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 new advanced features and new technologies for MPEG-G covering any of these aspects:

* coding modes specialized for “Third Generation Sequencing” (long reads technologies) devices
* coding modes relying on machine learning processes and providing higher compression
* genomic information satisfying data access modalities required by machine learning approaches
* coding genome sequences, quality scores, metadata with higher compression performance than current ISO/IEC 23092 series
* support for representation and usage of graph genome references
* support of interfaces with existing standards for the interchange of clinical data

For the detailed requirements to be fulfilled by responses to this CfP, refer to the requirements document (N40)

# 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 (N15698). If proponents feel that any aspects of their technology should not be made available in source code, 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.

In that respect proponents are invited to submit together with the response to the call an IP declaration as suggested in Annex A of this document.

In order to encourage the widest responses to the Call, we encourage “no-license” or “type-1” contributions. With “type-1” we refer to the option mentioned as box 1 in Annex A.

In the case alternative solutions achieve an equivalent level of satisfaction of requirements and present equivalent performance according to core experiment results, “no-license” or “type-1” solutions would be preferred.

# Timetable and Procedure

The following estimated milestones are planned:

2021/01/15 Draft CfP Issued

2021/04/30 CfP Issued

2022/01/10 CfP Responses: Deadline for Submissions

2022/04/29Technology Identification and Selection – Committee Draft

2023/07 Draft International Standard

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:

* A description of the technology having sufficient detail to permit technical discussions
* A description of how the proposed technology can be implemented as backward compatible extension of the ISO/IEC 23092 series.
* A list of satisfied requirements
* Evidence of the performance of the technology, including compression factor and any other meaningful metrics the proponents deem appropriate
* If appropriate, the proponent may provide comments on the requirements and the evaluation process in context of the application scenarios or use cases.

The proponent’s documents should be provided in Microsoft Word format.

**Important dates to answer this Call for Proposals are:**

2022/01/03 CfP Responses: Deadline for Submissions

2022/01/6-7 MPEG WG 08 Ad Hoc Meeting

**To support evaluation tests proponents should submit:**

* Description documents outlining the context for the submitted technology
* Executables – Decoders and encoders must 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 should preferably have command-line interface (i.e. no GUI).
* Bitstreams - Compressed data must be supplied corresponding to each individual test item as described in the Evaluation Criteria document (41) that has been issued at the 134th MPEG meeting.
* Draft Standard Specification of the submitted technology that can be included as an amendment to the ISO/IEC 23092 series.

Such items must be available for evaluation on 03 Jan 2022 for consideration at the AhG meeting prior to the 137th MPEG meeting (see section 6.4 for more details).

## Evaluation Criteria

Evaluation is based on the fulfillment of the requirements and measurable degree of fulfillment if applicable. Proposals are not required to meet all requirements. If applicable, please refer to the requirements document for complete list of the requirements applicable in this CfP(N40)

An evaluation procedure document (N41) has been issued at the 134th MPEG meeting and the criteria defined therein will be used in the technology selection process.

## Participation

Respondents to the CfP are required to attend the AhG meeting preceding the 137th MPEG meeting to present and discuss details of their proposals. This meeting will be held 6-7 Jan 2022 before the main MPEG meeting (10 – 11, Jan 2022).

## Preliminary Evaluation

At 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 Test Model Zero (TM0), 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.

## Test Model and Core Experiments

Two working tools play a major role in the collaborative development phase that follows the initial competitive phase: the Test Model and Core Experiments (CE).

The best technology, as identified in the evaluation process, will be selected as TM0 and be the basis for subsequent core experiments. Proponents whose technology is selected as TM0 and all proponents participating in the subsequent core experiment process shall supply a detailed description of their technology.

### Test Model

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

### Core Experiments

The improvement of the TM 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 TM or the direct inclusion in the TM to provide a new relevant functionality or improved performance. Improvements and additions to the TMs 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

For any other questions about the call, test conditions, required software or test sequences please contact:

**Marco Mattavelli, ISO/IEC JTC 1/SC 29/WG 8 Chair**

**Email: marco.mattavelli@epfl.ch**

# Bibliography

|  |  |
| --- | --- |
| [1] | S. D. Kahn, "On the Future of Genomic Data," *Science,* vol. 331, pp. 728-729, 2011. |
| [2] | S. Wandelt, M. Bux and U. Leser, "Trends in Genome Compression," *Journal of Current Bioinformatics,* 2013. |
| [3] | "New study illustrates benefits of long-read sequencing technology for precision oncology," July 2020. [Online]. Available: https://www.bcgsc.ca/news/new-study-illustrates-benefits-long-read-sequencing-technology-precision-oncology#:~:text=There%20are%20several%20advantages%20to,of%20large%20and%20complex%20genomes. [Accessed 28 April 2021]. |
| [4] | ISO/IEC JTC 1/SC 29/WG 8, "N41 - Evaluation Procedure for the Joint Call for Proposals on New Advanced Features and New Technologies for MPEG-G," Virtual Meeting, April 2021. |

**Annex A:   
Example of declaration of readiness to grant a license**

XYZ Organization may have current or pending patent rights relating to the technology described in this contribution and, conditioned on reciprocity, would be prepared to (check at least one of the following items):

* make them available free of charge (per box 1 of the ISO/IEC patent statement and licensing declaration form)
* grant licenses under reasonable and non-discriminatory terms as necessary for implementation of the resulting ISO/IEC International Standard (per box 2 of the ISO/IEC patent statement and licensing declaration form)

However, XYZ Organization is aware that other entities may also have current or pending patent rights relating to the technology described in this contribution.