**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2015/**NXXXX

**June 2015, Warsaw, Poland**

|  |  |
| --- | --- |
| **Source:** | Requirements |
| **Status:** |  |
| **Title:** | Draft Call for Evidence (CfE) for Genome Compression and Storage |
| **Editor(s):** | Claudio Alberti, Marco Mattavelli |

**Abstract**

This document is a Draft Call for Evidence (CfE) associated with technology for efficient genomic information representation for transport and storage.

# Introduction

The sequencing of the genetic information of human genome has become affordable due to high-throughput sequencing technology [1], [2]. 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. 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. 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.

More background information as well as information about applications and requirements are given in output document N15093 published at the 111th MPEG meeting in Geneva.

Companies and organizations are invited to submit proposals in response to this call.

The results of these tests will be made public, taking into account that no direct identification of any of the proponents will be made (unless it is specifically requested or authorized by a proponent to be explicitly identified). Prior to having evaluated the results of the tests, no commitment to any course of action regarding the proposed technology can be made.

Descriptions of proposals shall be registered as input documents to the proposal evaluation meeting in xxxx (see timeline in section 3). 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 xx).

# Purpose

The purpose of this CfE is to explore whether

* the coding efficiency and/or
* the functionality

of currently used file formats such as gzipped FastQ (raw data) and BAM (aligned data) can be further improved significantly.

# Timeline

2015/02/21 Availability of test materials

2015/06/26 Draft Call for Evidence

2015/09/xx Final Call for Evidence

2016/01/xx Submission of documents, bitstreams and executables (for details of the submission process contact the persons listed in section xx)

2016/02/xx Evaluation of the proposals

# Test Conditions

## Category 1: Compression of raw sequence data (FastQ)

### Anchors and Test Material

Anchors are contained in input document m36257 presented at the 112th MPEG meeting in Warsaw. Test material consists of the FastQ files described in output document N15092 produced at the 111th MPEG meeting in Geneva.

### Test Procedures

TBD

## Category 2: Compression of aligned sequence data (SAM)

### Anchors and Test Material

Anchors are contained in input document m36257 presented at the 112th MPEG meeting in Warsaw. Test material consists of the BAM files described in output document N15092 produced at the 111th MPEG meeting in Geneva.

### Test Procedure

TBD

# Requirements on Submissions

## Submission categories and details

Submit filled spreadsheets provided in Annex A for respondents to complete.

Submit encoded bit streams and decoder executable together with documentation on how to run the decoding process. Please add any relevant information (e.g. which OS?).

Proponents need to be present at the September/October 2015 meeting to present their submissions.

Submit technical descriptions described below:

* A technical description of the proposal sufficient for the full conceptual understanding and generation of equivalent performance results by experts and for conveying the degree of optimization required to replicate the performance. This description should include all data processing paths and individual data processing components used to generate the bitstreams.
* The technical description shall contain information suitable to assess the complexity of the implementation of the technology, including the following:
  + Encoding time (for each submitted bitstream) of the software implementation.  Proponents shall provide a description of the platform and methodology used to determine the time.  To help interpretation, a description of software and algorithm optimisations undertaken, if any, is welcome.
  + Decoding time for each bitstream running the software implementation of the proposal, and for the corresponding constraint case anchor bitstream(s) run on the same platform.  Proponents shall provide a description of the platform and methodology used to determine the time.  To help interpretation, a description of software optimisations undertaken, if any, is encouraged.
  + Expected memory usage of encoder and decoder.

## Binaries and Source Code

* Proponents are encouraged to allow other committee participants to have access, on a temporary or permanent basis, to their encoded bitstreams and binary executables or source code.
* Proponents are encouraged to submit a statement about the programming language in the software is written, e.g. C/C++ and platforms on which the binaries were compiled.
* Proponents are advised that, upon acceptance for further evaluation, it will be required that certain parts of any technology proposed be made available in source code format to participants in the core experiment process and for potential inclusion in the prospective standard as reference software. When a particular technology is a candidate for further evaluation, commitment to provide such software is a condition of participation. The software shall produce identical results to those submitted to the test.

# IPR

Proponents are advised that this call is being made subject to the common patent policy of ITU-T/ITU-R/ISO/IEC and other established policies of these standardization organizations. The persons named below as contacts can assist potential submitters in identifying the relevant policy information.

# Fees

None (TBD ????).

# Contact(s)

Xxxx

# References

|  |  |
| --- | --- |
| [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. |

**ANNEX A**

Spreadsheets to be completed and submitted by the respondents are provided in the attached Excel file.

**Attachments:**

1. Performance measurement spreadsheet.