**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) aiming at assessing the performance of new technologies for the efficient lossless compression of genomic information in the form of raw sequence data (FastQ) and aligned data (SAM).

# 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.

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 is available in output document N15094 [3] published at the 111th MPEG meeting in Geneva and draft requirements are provided in output document NXXXX published at the 112th MPEG meeting in Warsaw.

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 that will take place in San Diego (US) on February 2016 (see timeline in section 3 and submission procedure in section 5). 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 8).

# Purpose

The purpose of this CfE is to assess whether new technologies can achieve better performance in terms of lossless compression efficiency compared with currently used file formats such as gzipped FastQ (raw data) and BAM (aligned data).

Additional purposes are to:

* become aware of which additional functionalities (e.g. non sequential access, lossy compression efficiency, etc. ) are provided by these new technologies
* collect information that may be used in drafting a future Call for Proposals

# Timeline

2015/02/21 Availability of test materials

2015/06/26 Draft Call for Evidence

2015/10/23 Final Call for Evidence

2016/02/17Submission of documents (for details of the submission process contact the persons listed in section 8 and see section 5 for the submission details).Submissions of executables is encouraged.

2016/02/20-21 Evaluation of the proposals

# Test Conditions

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

### Anchors and Test Material

Anchors are contained in output document NXXXX published at the 112th MPEG meeting in Warsaw.

For raw sequence data the anchors to be considered are the compressed files in gzipped format. Other compression formats referenced in the document are only mentioned as meaningful existing alternatives.

Test material consists of the FastQ files described in output document N15092 [4] produced at the 111th MPEG meeting in Geneva.

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

### Anchors and Test Material

Anchors are contained in output document NXXXXX published at the 112th MPEG meeting in Warsaw.

For aligned sequence data the anchors to be considered are the compressed files in BAM format. Other compression formats referenced in the document are only mentioned as meaningful existing alternatives.

Test material consists of the BAM files described in output document N15092 [4] produced at the 111th MPEG meeting in Geneva.

# Requirements on Submissions

## Submission categories and details

Please note that a submission shall include the following mandatory elements

* Filled spreadsheets provided in Annex A with the following information
  + Input item
  + Output bitstream size for lossless compression
  + Compression factor
  + Weight (%) of the main classes of data in the compressed bitstream
    - Reads headers/identifiers
    - Sequence reads
    - Quality scores
    - Any other metadata (identified as “Auxiliary data”)
  + Encoding time
  + Peak and average memory usage
  + A description of the platform used to run the decoding process (HW, OS, etc.)
* Encoded bit streams and – if available - decoder executable together with documentation on how to run the decoding process. Please add any other relevant information (e.g. which OS, any HW requirements?). Proponents need to
  + be present at the October 2015 meeting to present their submissions
  + bring the encoded bitstreams on a physical storage device at the October 2015 meeting
* Technical descriptions described below:
  + Approximate expected decoding time for each bitstream running the software implementation of the submission on a similar platform.  Proponents shall provide a description of the platform and methodology used to determine the time.  To help evaluation, a description of software libraries (e.g. low level libraries) used, if any, is encouraged.
  + Supported platforms (HW requirements and OS)
  + Configuration parameters (if any)
  + Expected resources (e.g. RAM, disk space etc.) usage of the decoder
  + Any other information needed to run the decoding process
  + Support for any feature such as:
    - Non sequential access
      * Example: extract all the reads (and the associated metadata such as QS or identifiers) that map within a given interval of the reference genome
    - Support of more than 5 symbols (A, C, G, T, N) alphabets
    - Encoding of additional metadata (extensibility)
    - Lossy compression of
      * Quality scores
      * Reads identifiers

Information on the lossy compression approach applied is encouraged

Any document included in the submissions shall be registered according to the usual MPEG submission procedure, i.e.:

**For MPEG members.**

Please follow the usual input documents submission procedure

* Online submission portal: <http://wg11.sc29.org/>
  + Click on “Next meeting” (username and password required, ask your national Head of Delegation for credentials)
  + Click on “Register a new document”
  + Please use the [input document template](http://wg11.sc29.org/Templates/mxxxx.dot) available on the MPEG portal

**For non MPEG members**

Please use the [input document template](http://wg11.sc29.org/Templates/mxxxx.dot) available on the MPEG portal and contact the reference persons listed in section 8.

For both MPEG members and non MPEG members the deadlines are the following:

* Document registration deadline: Monday 2016-02-15 23:59:59 CET
* Document upload deadline: Wednesday 2016-02-17 23:59:59 CET

## Binaries and bitstreams

* Proponents are encouraged to allow other committee participants to have access, on a temporary or permanent basis, to their encoded bitstreams and binary executables (if available).
* 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.

# IPR

Proponents are advised that this call is being made subject to the common patent policy of ITU-T/ITU-R/ISO/IEC (http://www.itu.int/en/ITU-T/ipr/Pages/policy.aspx) 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.

# Contact(s)

Joern Ostermann (ostermann@TNT.UNI-HANNOVER.DE), Marco Mattavelli (marco.mattavelli@epfl.ch)

# 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. |
| [3] | MPEG Requirements, “N15094 - Investigation on genomic information compression and storage,” Geneva, 2015. |
| [4] | MPEG Requirements, “N15092 - Database for Evaluation of Genome Compression and Storage,” Geneva, 2015. |

**ANNEX A**

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

**Attachments:**

1. Performance measurement spreadsheet.