**ISO/IEC JTC 1/SC 29/WG 11  
Coding of moving pictures and audio  
Convenorship: UNI (Italy)**

**Document type: Approved WG 11 document**

**Title: Draft Requirements for ISO/IEC 23092 series extension**

**Status: Approved**

**Date of document: 2020-01-16**

**Source: Requirements**

**Expected action:**

**No. of pages: 4**

**Email of convenor: leonardo@chiariglione.org**

**Committee URL: http://mpeg.chiariglione.org**

**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 N19060**

**Brussels, BE – January 2020**

|  |  |
| --- | --- |
| **Source:** | **ISO/IEC JTC 1/SC 29/WG 11** |
| **Status:** | Draft |
| **Title:** | Draft Requirements for ISO/IEC 23092 series extension |

**Table of Contents**

[1 Executive summary 3](#_Toc30084348)

[2 Requirements 3](#_Toc30084349)

1. **Executive summary**

This document lists the requirements that responses to the Call for Proposals for extensions and improvements of ISO/IEC 23092 series shall meet.

1. **Requirements**

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirement** | **Rationale/Notes** |
| 1.1 | The solution shall rely on and reuse as much as possible the technology already present in ISO/IEC 23092 avoiding the simple duplication of already supported features. | Replication of features is desirable only if performance (i.e. coding efficiency, data access, classification, browsing capabilities, etc etc …) is meaningfully improved. |
| 1.2 | The solution shall support lossless compression of sequencing reads generated by Third Generation Sequencing devices. |  |
| 1.3 | The solution shall support data classification suitable for sequencing reads generated by Third Generation Sequencing devices. |  |
| 1.4 | The solution should support the representation of multiple alignments for sequencing data. | The program chosen for alignment can affect downstream inference. To mitigate such side effects, a solution is to perform alignment by multiple alignment programs (using multi aligners) and efficiently register associated results. |
| 1.5 | The solution shall support reprocessing and updating encoded sequencing data to new genome data builds and use the samples for analysis. | MPEG-G should enable efficient reprocessing and updating of old sequencing data files with reference to new genome NCBI data builds and use the samples for analysis |
| 1.6 | The solution shall support compression and transport of large datasets of single-cell sequencing data | Millions of single-cells are processed in batches across different locations by the Human Cell Atlas (HCA) project. E.g., the census of the immune cells project in the HCA contains ~530,000 cells resulting in a 1.3 TB data set. However, this is just one of the studies out of the hundreds of datasets that will be released by the HCA team. Thus, MPEG-G should provide the means for efficient classification, compression, storage and transmission of single-cell sequencing data. |
| 1.7 | The solution shall support lossless compression and easy identification of chimeric pairs. | Chimeric pairs are composed by two reads which map with positions and orientations which are incompatible with the expected behavior of normal read fragments. |
| 1.8 | The solution shall support the signaling of duplicate reads for efficient identification and pre-processing | In NGS analysis two types of duplicates can be found: a) optical duplicates (in sequencing devices), that can be identified without alignment; b) library duplicates (or PCR), originated by pre-amplification and usually identified after alignment. Preprocessing of duplicates is in general non-specified, but very often they are simply filtered out prior any analysis stage. |
| 1.9 | The solution shall support to natively retrieve “pile-up” information from the compressed aligned reads. | The quality or characteristics of a “pile-up” is often an indication of the reliability of the analysis results obtained using the reads aligned in a given interval. |
| 1.10 | The solutions shall support k-mer search in the compressed domain. | Fast re-analysis on unaligned data according to k-mer based searches is fundamental in many scientific investigations. |
| 1.11 | The solutions shall support lossless compression of quality values. | This requirement is intended to improve compression of QVs with respect to the existing MPEG-G coding modes. |
| 1.12 | The solutions shall support compression of references and/or computed references “patches” | This is intended to support efficient representation of reference assemblies updates. |
| 1.13 | The solution shall support lossless compression of the read names and the corresponding reads in different order |  |