**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 N16134**

**San Diego, US – February 2016**

|  |  |
| --- | --- |
| **Source:** | Requirements |
| **Status:** | Draft |
| **Title:** | Requirements on Genome Compression and Storage |

**Table of Contents**

[1 Executive summary 2](#_Toc444300676)

[2 Terminology 2](#_Toc444300677)

[3 Genomic information life cycle 2](#_Toc444300678)

[4 Requirements 4](#_Toc444300679)

[4.1.1 Sequence reads extraction – compression of unmapped reads 4](#_Toc444300680)

[4.1.2 Mapping and Alignment – Compression of mapped/aligned reads 5](#_Toc444300681)

[4.1.3 Transport 6](#_Toc444300682)

[5 References 7](#_Toc444300683)

# Executive summary

This document lists the requirements that responses to the Call for Proposals shall meet.

# Terminology

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Alignment | A sequence read mapped on a reference DNA sequence |
| BAM | Compressed binary version of SAM |
| CRAM | GIR that includes SAM + Compression configuration |
| FASTA | GIR that includes header and sequence reads (nucleotides sequence) |
| FASTQ | GIR that includes FASTA + Quality Scores |
| Indel | An additional or missing nucleotide in a DNA sequence with respect to a reference DNA sequence. |
| MAF | Mutation Annotation Format. File format used to mark the genes and other biological features in a DNA sequence |
| Mate pairs  | Two reads from the same (long) DNA strand extracted by sequencing machines. The orientation is the opposite of paired ends.  |
| Paired ends | Couple of reads produced by the same (short) DNA fragment by sequencing both ends. The orientation is the opposite of mate pairs. |
| Quality score | A quality score is assigned to each nucleotide base call in automated sequencing processes. It expresses the base-call accuracy. |
| Read header | Each sequence read stored in FASTA and FASTQ format starts with a textual field called “header” containing a sequence identifier and an optional description |
| SAM | GIR that is human readable and includes FASTQ + Alignment and analysis information |
| Sequence read | The readout, by a specific technology more or less prone to errors, of a continuous part of a segment of DNA extracted from an organic sample |

Table 1 - Genome Information Terminology

# Genomic information life cycle

The typical steps of the genomic information life cycle are described in the table below

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Step** | **Description** | **File format**  |
| 1 | Sequence reads extraction | The process of extraction of fragments of DNA/RNA in the form of sequences of nucleotides from a biological sample. Sequences of nucleotides are commonly referred to as “reads”. This information is usually encoded as FASTA files including a textual header and a sequence of nucleotides symbols (A,C,G,T,N). If a quality score is added to each nucleotide read, the file is called FASTQ. | FASTAFASTQ |
| 2 | 1. Mapping and Alignment
 | Sequence alignment refers to the process of arranging sequence reads by finding regions of similarity that may be a consequence of functional, structural, or evolutionary relationships among the sequences. When the alignment is performed with reference to an existing DNA sequence the process is called “mapping”. | SAMBAMCRAM |
| 3 | Variant detection | Variant detection (a.k.a. variant calling) is the process of translating the output of DNA sequencing machines, (short reads mentioned in step 1 and aligned in step 2), to a summary of the unique characteristics of the organism being sequenced. These characteristics are called “variants” because they are expressed as differences between the organism under study and a reference genome. | VCF |
| 4 | Variant annotation | Variant annotation is the process of assigning functional information to the DNA variants identified in step 3. This implies the classification of variants according to their relationship to coding sequences in the genome and according to their impact on the coding sequence and the gene product. | MAF |
| 5 | Functional & Structural Analysis | Analysis of DNA (variants, CNV = copy number variation, methylation etc,) strands to define their relationship with genes (and proteins) functions and structure. | Several different solutions |

Table 2 - The main stages of a typical genomic processing pipeline

With respect to the previous table, steps 3 to 5 cover a vast horizon of use cases, and could open up interpretation of variant ‘severity’ relationship with phenotypes, etc… This goes relatively far from compression and transport of sequence reads. In the scope of this call it might be important to leave open the door to ‘pointers’ to extra containers of annotations.

These steps are mentioned here for the sake of completeness but are not in the scope of this Call.

Figure 1 depicts a functional diagram of the typical genomic information life cycle expressed as different processing steps and intermediate file formats.



Figure 1 - Typical structure of a genomic information processing pipeline from sequencing to analysis.

A simplified vision of the relation among the file formats used in genome processing pipelines is depicted in Figure 2. In this diagram file inclusion does not imply the existence of a nested file structure but it only represents the type and amount of information that can be encoded for each format (i.e. SAM contains all information in FASTQ, but organized in a different file structure). The BAM format is not represented because it is the compressed equivalent of SAM.



Figure 2 - Relationship among some of the most popular file formats.

# Requirements

This section introduces the definition of the requirements an appropriate solution to genomic information representation (GIR) and transport (GITL) should meet at the first two stages of the processing pipeline introduced earlier and listed in Table 2.

Variants calling, mutations annotation and functional analysis are likely to have different requirements in terms of information representation and transport and will be addressed in a separate stream of work.

### Sequence reads extraction – compression of unmapped reads

Compression of FASTQ description is not taken into account in this list of requirements as this is an obsolete field no more used by the main next-generation sequencing (NGS) machines manufacturers.

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirement**  | **Rationale/Notes** |
| 1.1 | The solution shall support lossless compression of reads headers. |  |
| 1.2 | The solution shall support lossy compression of reads headers by preserving at least a unique read identifier and pairing information when available. | In case read headers are not needed in downstream analysis they do not need to be preserved. |
| 1.3 | The solution shall support preservation of pairing information |  |
| 1.4 | The solution shall support lossless compression of nucleotides sequences supporting a minimum of 5 symbols (A, C, G, T, N) | Note: today most machines use 5 symbols but other symbols might be adopted soon such as the IUPAC ambiguity codes |
| 1.5 | The solution shall support lossless compression of quality scores. |  |
| 1.6 | The solution shall support lossy compression of quality scores. | The evaluation methodology of how submitted solutions will meet this requirement is described in document N16135 |
| 1.7.1 | The solution shall structure compressed data so that parallel processing is enabled and compression efficiency is preserved |  |
| 1.7.2 | The solution should structure compressed data so that efficient querying of data is enabled and compression efficiency is preserved |  |
| 1.8 | The solution shall preserve the association among headers, nucleotides and quality scores  |  |

### Mapping and Alignment – Compression of mapped/aligned reads

This section lists additional requirements for aligned data. All requirements listed for raw data still apply.

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirement**  | **Rationale** |
| 2.1 | When a reference genome is used, the solution shall make it a resource unambiguously identifiable and available to the decoder.  |  |
| 2.2 | The solution shall allow reconstructing the original unmapped reads from aligned/mapped reads | Raw data shall be available to enable re-alignment of reads according to different alignment criteria |
| 2.3 | The solution shall support access to specific regions of the aligned data. | Access to specific loci shall be possible without the need to decompress the whole data. |
| 2.4 | The solution should support the widest range of selective data access options. | E.g. access to unmapped reads only without the need decompress the whole data |
| 2.5 | The solution shall uniquely identify each aligned read (a.k.a. alignment) and its mate pair when applicable. |  |
| 2.6 | The solution shall support lossless compression of attributes associated to each read expressing:* template having multiple segments in sequencing
* each segment properly aligned according to the aligner
* unmapped segment
* next segment in the template unmapped
* signalization of first or last segment
* quality control failure
* PCR or optical duplicate
* secondary alignment
* supplementary alignment
 |  |
| 2.7 | The solution shall support lossless compression of mapping/alignment information. | E.g. genomic location information, orientation, pairing information, indels, mismatches. |
| 2.8 | The solution shall support lossless compression of mapping quality expression. | The current metric for mapping quality is expressed as “phred-scaled posterior probability that the mapping position" is wrong. |
| 2.9 | The solution shall support lossless compression of additional metadata related to single reads or the whole compressed data | E.g. auxiliary fields in SAM, SAM header |

### Transport

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirement for transport** | **Rationale** |
| 3.1 | The compression process shall support the assessment of integrity | Integrity check shall be possible by providing appropriate information. |
| 3.2 | The solution shall allow conveying information enabling data protection  | Ability to prevent unauthorized access shall be available. Information needed for protection of data (control for access, modification, publication, etc.) shall be conveyed. |
| 3.3 | The solution shall allow conveying information enabling accountability and traceability  | Data access and manipulation shall be traceable together with the identity of parties having access to data.Information on how to verify integrity and authenticity of the data shall be conveyed. |
| 3.4 | The solution shall allow conveying information enabling transparency  | How and for which purpose the information is used shall be known. Usage restriction shall be applicable to the data. |
| 3.5 | The solutions shall support compressed data streaming. | This implies that data consumption shall be possible before data transfer completion. |

# References

|  |  |
| --- | --- |
| [1]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2014/N14553 - Issues in genome compression and storage,” Sapporo, 2014.  |
| [2]  | R. Canovas, A. Moffat and A. Turpin, “Lossy compression of quality scores in genomic data,” *Bioinformatics,* vol. 30, no. 15, pp. 2130-2136, 2014.  |
| [3]  | M. Hernaez, I. Ochoa, R. Goldfeder, T. Weissman and E. Ashley, “Effect of lossy compression of quality scores on SNP,” *Biorxiv,* 2015.  |
| [4]  | S. D. Kahn, “On the Future of Genomic Data,” *Science,* vol. 331, pp. 728-729, 2011.  |
| [5]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2015/N15740 - Call for Evidence on Genome Compression and Storage,” Geneva, 2015.  |
| [6]  | 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.  |
| [7]  | 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.  |
| [8]  | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N16147 - Results of the Call for Evidence on Genome Compression and Storage,” San Diego, 2016.  |