INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 N17468
January 2018, Gwangju, KR

Source	**ISO/IEC JTC1/SC29/WG11**
Status	**Final**
Title	**White paper on the objectives and benefits of the MPEG-G standard**

1 MPEG-G: leveraging the use of genomic information

The development and rapid progress of high-throughput sequencing (HTS) technologies has the potential of enabling the use of genomic information as an everyday practice in several fields. With the release of the latest HTS machines the cost of a whole genome sequencing for a human genome has dropped to merely US $1,000. It is expected that within the next few years such cost could drop to about US $100. Such achievements in the reduction of sequencing costs opened the doors to personalized medicine, where the genomic information of patients can be sequenced and analyzed as frequently as done today for standard blood tests. However, the ever-growing volume of sequencing data is already a serious obstacle to the wider diffusion of sequencing in public health. The associated IT costs, related to storing, transmitting and processing the large volumes of data, will soon largely exceed the costs of sequencing. The lack of appropriate representations and efficient compression technologies is widely recognized as a critical element limiting the potential of genomic data usage for scientific and public health purposes [1].

The MPEG-G standard, planned to be released in October 2018, is currently the largest coordinated and international effort addressing the problems and limitations of current technologies and products towards a truly efficient and economical handling of genomic information. MPEG-G utilizes the latest technology to compress and transport sequencing data for complex use cases that are currently not supported by existing formats. Notable use cases addressed by MPEG-G include:

- Selective access to compressed data
- Data streaming
- Compressed file concatenation
- Genomic studies aggregation
- Enforcement of privacy rules
- Selective encryption of sequencing data and metadata
- Annotation and linkage of genomic segments
- Interoperability with main existing technologies and legacy formats
- Incremental update of sequencing data and metadata

This white paper summarizes the objectives and the benefits of the upcoming MPEG-G standard by providing:

- a brief description of the requirements and the methodology used to develop the various parts of the MPEG-G standard,
- an exemplary overview of the current MPEG-G compression performance,

- an introduction to the main technical features of MPEG-G,
- a summary of possible applications and usage scenarios of the MPEG-G standard.

2 Development of the MPEG-G standard

In its 30 years of activity ISO/IEC JTC 1/SC 29/WG 11 – also known as Moving Picture Experts Group (MPEG) – has developed many generations of successful standards that have transformed the world of media from analog to digital. Video, audio compression and transport technologies, as well as application formats and APIs, have provided the standard support enabling the interoperability and the integration we all witness in the digital media field.

ISO TC 276 has recently been established to work on the standardization of biotechnology processes including analytical methods (Working Group 3) and data processing and integration (Working Group 5).

MPEG, as developer of generic standard technologies is working with ISO TC 276/WG 5, integrators of biological data workflows, to produce a new open standard to compress, store, transmit and process sequencing data: MPEG-G. The standard will offer high levels of compression, approximately 100 times compared to raw data, i.e. more than one order of magnitude than possible with currently used formats [2]. Furthermore, the MPEG-G standard will provide new functionalities such as native support for selective access, data protection mechanisms, flexible storage and streaming capabilities. This will enable various new applications scenarios, such as real-time streaming of data from a sequencing machine to remote analysis centers during the sequencing and alignment processes.

Interoperability and integration with existing genomic information processing pipelines is enabled by supporting conversion from/to the legacy FASTQ/SAM/BAM file formats.

2.1 *Requirements for an efficient genomic information representation*

Today, a single sequencing system can deliver the equivalent of 18,000 whole human genomes per year, which accounts for almost 5 PB of data per year. This leads to the forecasts that the amount of genomic data is expected soon to surpass astronomical data in volume [3]. The efficient storage and transmission of genomic data is thus becoming of utmost importance.

Motivated by this demand, MPEG and ISO TC 276/WG 5 have been analyzing the main stages of typical genomic information processing since July 2014. The work produced a list of requirements to be satisfied for achieving an efficient compressed representation of raw and aligned reads produced at the initial stages of processing pipelines for genomic information, as well as a list of requirements for the efficient transport of and for the selective access to the compressed genomic data.

In the meantime, an increasing consensus had been rising in the scientific community around the notion of quantized transmission of the metadata (i.e. quality values) generated in very large volumes by sequencing machines [4], [5]. Therefore, additional requirements for solutions that address calibrated and quantized compressed representations of such metadata were identified.

The process of identifying all requirements of a standard to address and solve the various challenging problems of HTS data processing was a wide interdisciplinary effort grouping experts from different domains including bioinformatics, biology, information theory, telecommunication, video and data compression, data storage, and information security. The identified requirements that were the baseline for the development of the MPEG-G standard are available in full detail in the public document N16323 (MPEG)/N97 (ISO TC 276/WG 5).

2.2 *MPEG-G technology*

MPEG-G has been developed following the rigorously open process adopted by MPEG for its standards. A Call for Proposals was issued in June 2016 and 15 responses were received from 17 companies and organizations. The technologies in the responses were evaluated using several criteria, including compression performance and separate assessments for each type of genomic data: sequence reads, quality values, read identifiers, alignment information, and associated metadata. In addition, computational complexity in terms of encoding and decoding times and memory usage was also assessed. The support of a minimal functionality backing non-sequential access, extended nucleotide alphabets, encoding of additional metadata (extensibility), and quantized coding of metadata was also considered in ranking the received proposals.

The most valuable technologies were integrated in the standard to provide the following three classes of functionality: 1) the compression of genomic data generated by sequencing processes, 2) the compression of genomic data associated to alignment information, 3) the definition of a Genomic Information Transport Layer that supports storage and transport functionality.

3 MPEG-G compression performance

During the Core Experiments phase the best-performing technologies according to the results of the Call for Proposals were selected for integration into the MPEG-G standard specification. The approach taken in defining the components that are normative – and thus guarantee the interoperability of applications – and the components that have been left open to competitive innovations and optimizations, is the same as taken by the most successful MPEG standards in the audio-video field [6]. More precisely, only the syntax and the semantics of the decoding process are normative and fully specified, whereas the encoding process is left open to algorithmic and implementation-specific innovations, as well as to coding optimizations. Sequencing data and associated metadata are sets of heterogeneous data possibly characterized by highly variable statistical behaviors, and thus several strategies for its classification and representation can be used. Therefore, the optimization space for compression performance and selective data access is very large and admits many different solutions. An example of the compression performance of a straightforward application of the standard MPEG-G technology is shown in Figure 1 for a dataset composed of aligned, high-coverage human whole genome sequencing (WGS) data, and in Figure 2 for a dataset of aligned, low-coverage human WGS data. Considering what has been discussed above, it must be underlined that this is only an example of possible performance. Encoders employing other classification and optimization strategies could easily achieve much better performance results.

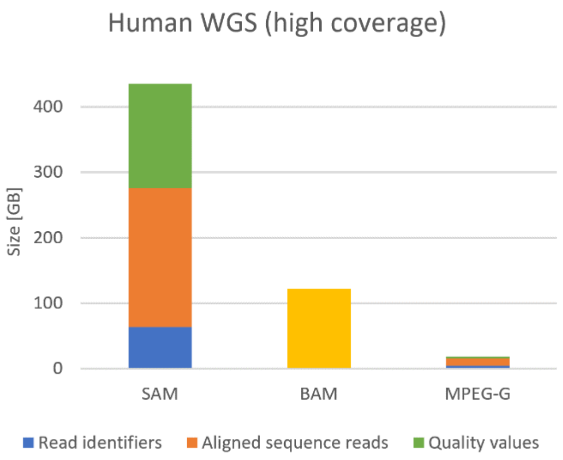

Figure 1: Compression performance of MPEG-G on high-coverage sequencing data and metadata.

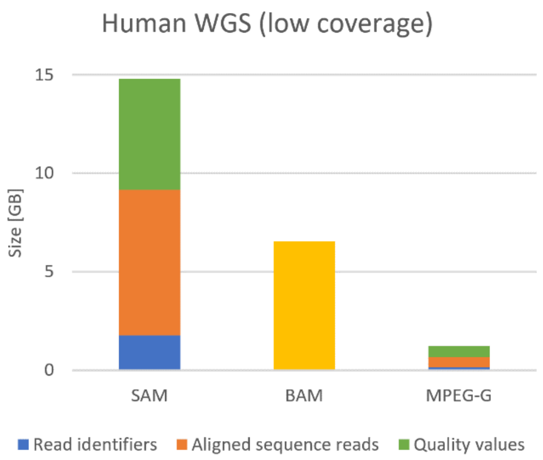

Figure 2: Compression performance of MPEG-G on low-coverage sequencing data and metadata.

Both graphs report the size of aligned sequencing reads, including read identifiers and quality values, on the left, as represented in the SAM format [7], in the middle the size of the corresponding BAM file and in the right the size of an MPEG-G file. In today's common practice, SAM files are often stored or transmitted in the form of BAM files, which are essentially block-wise gzipped SAM files. In these examples BAM compression provides a compression factor of about 3.58 for the high coverage dataset and a factor of about 2.26 for the low coverage dataset, respectively, over SAM. When MPEG-G compression technology is employed, the compression can further be improved with respect to BAM, by a factor of about 6.54 (high coverage) and 5.31 (low coverage), respectively. With respect to the SAM representation size, the compression factors of MPEG-G are about 23.41 (high coverage) and 12.00 (low coverage). However, as thoughtfully discussed above, this is only an example of a possible classification and coding performance achievable, and compression ratios may vary according to the specific statistical characteristics of each data set and according to the quality and optimization capabilities of the encoder.

4 Benefits provided by MPEG-G

The standard intends to provide the foundation for interoperable genomic information processing applications enabling the use of genomic data on a large scale. ISO/IEC is also engaged in supporting the maintenance of the standard to guarantee the perenniality of the applications using MPEG-G technology. A list of the essential features of the MPEG-G technology is the following:

- **Selective access to compressed data**: Indexing tools embedded in an MPEG-G file enable several types of selective access to compressed data that can be combined in the same query.
- **Data streaming**: MPEG-G supports the packetization of compressed data for transport to receiving devices that can start processing the data before transmission is completed.
- **Compressed file concatenation**: MPEG-G files can be concatenated without the need to decode and re-encode them.
- **Genomic studies aggregation**: Several related genomic studies can be encapsulated in the same MPEG-G file while still being separately accessible. Additionally, transversal queries over multiple studies are possible (e.g. "select chromosome 1 of all compressed samples").
- **Enforcement of privacy rules**: Data encoded in an MPEG-G file can be linked to multiple owner-defined privacy rules, which impose restrictions on data access and usage.
- **Selective encryption of sequencing data and metadata**: The encryption of genomic information is supported by MPEG-G at different levels in the hierarchy of MPEG-G logical data structures.
- **Annotation and linkage of genomic segments in the compressed domain**: MPEG-G supports the annotation of genomic segments. Additionally, MPEG-G provides support for linking segments within a single genomic sample or across multiple genomic samples.
- **Interoperability with main existing technologies and legacy formats**: Conversion to/from legacy format such as FASTQ, SAM or BAM is supported by MPEG-G.
- **Incremental update of sequencing data and metadata**: MPEG-G files can be incremented with sequencing data and metadata without requiring decompression and re-compression of pre-existing data.

5 Important technical features of MPEG-G

MPEG-G technology provides storage and transport capabilities both for raw genomic sequences and for genomic sequences mapped onto reference genomes. Moreover, MPEG-G provides support for complex use cases that are currently not supported by existing formats.

The representation of genomic information with MPEG-G is based on the concept of the *Genomic Record*, a data structure consisting of either a single sequence read, or a paired sequence read, and its associated sequencing and alignment information; it may contain detailed mapping and alignment data, a single or paired read identifier (read name) and quality values.

Genomic Records are aggregated and encoded in structures called *Access Units*. These structures are units of coded genomic information that can be separately accessed and inspected. An illustration of the essential elements, including Access Units, of the MPEG-G file format is shown in Figure 3 and in Figure 4.

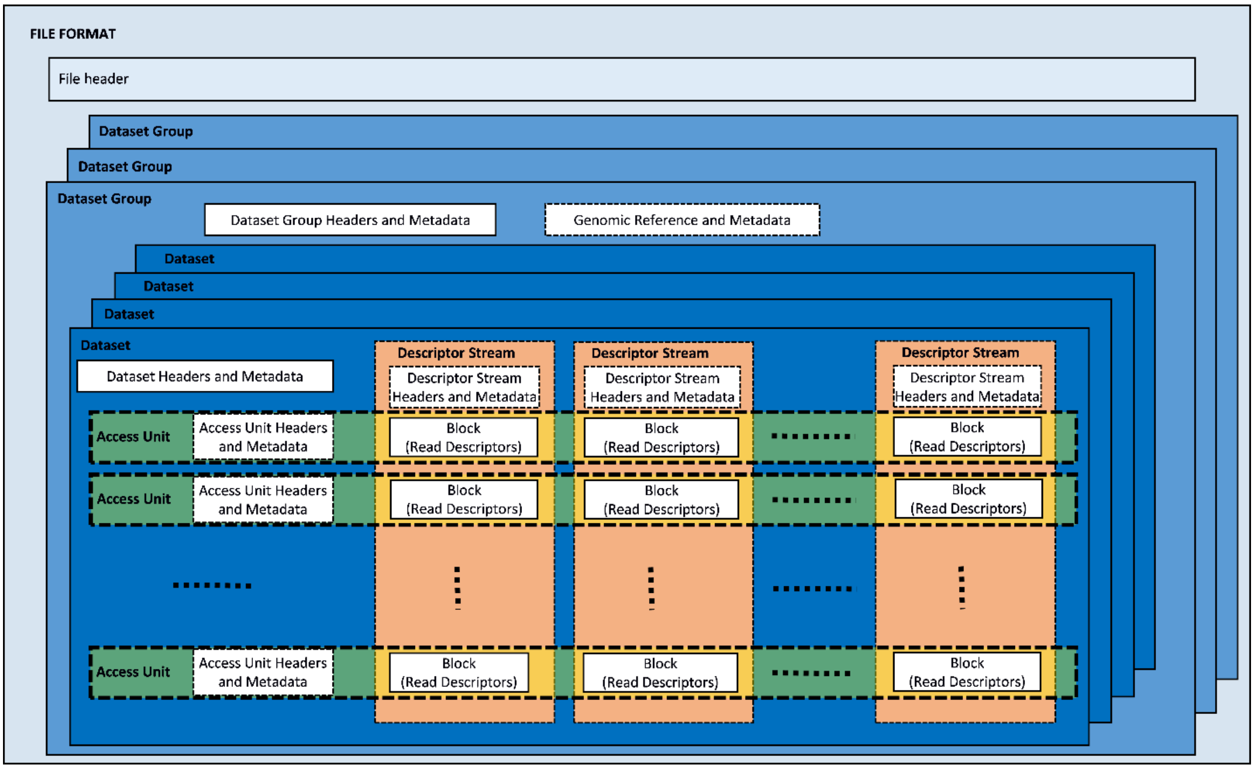

Figure 3. Illustration of the essential elements of the MPEG-G file format. Multiple "Dataset Groups" include multiple Datasets of sequencing data. Each Dataset is composed of Access Units containing only one Data Class. Each Data Class is composed by Blocks of Read Descriptors.

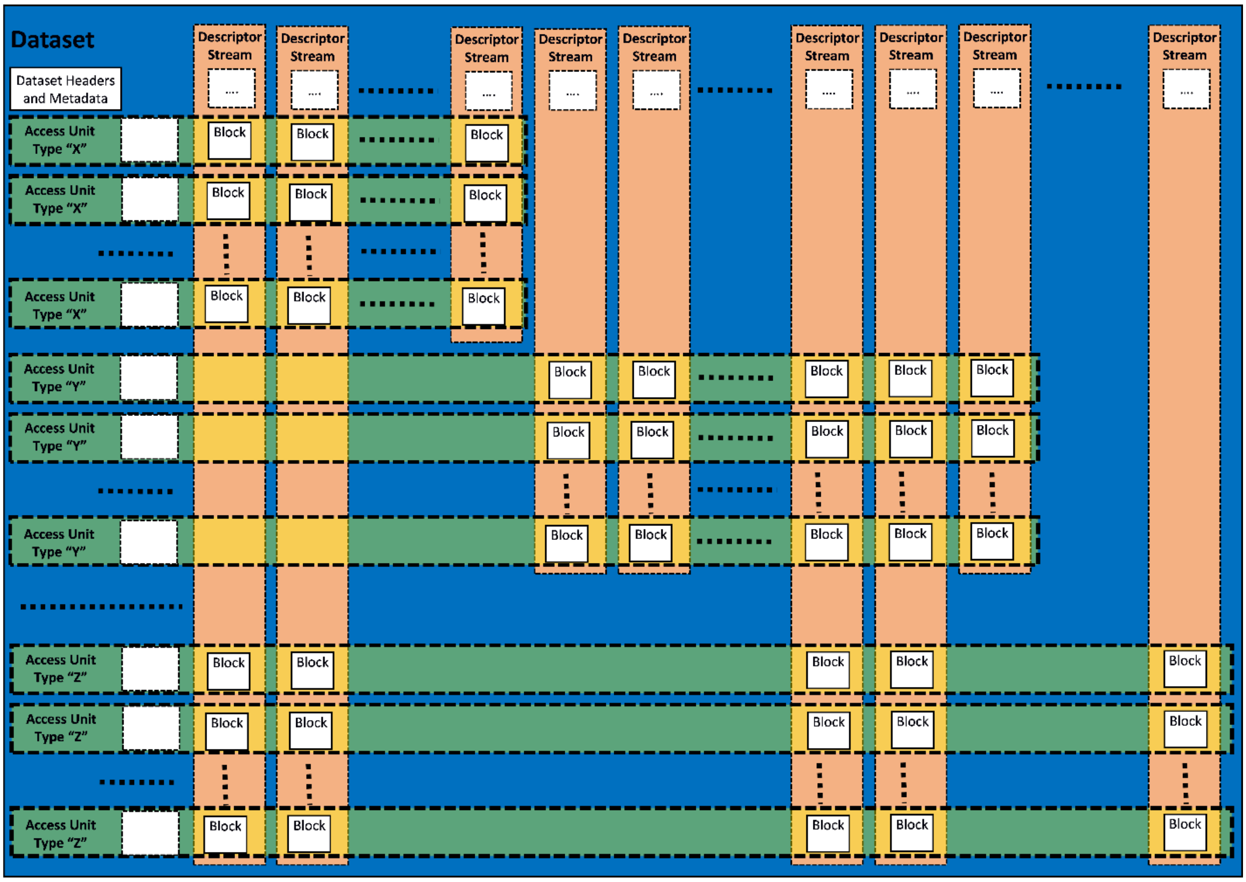

Figure 4: Access Units containing compressed representations of different Data Classes are composed by different subsets of Descriptor Stream Blocks.

5.1 Raw sequences

Raw sequence data can be encoded in an MPEG-G file with two main objectives:

- **High compression ratio and indexing**: A high compression ratio is reached by leveraging the high redundancy in genomic sequence data. This enables the use of well-known compression techniques such as differential coding of sequences with respect to already encoded data. This approach achieves a maximum compression efficiency, but requires the availability of the entire dataset as well as a few preprocessing stages which may impact the compression latency with respect to the data generation time.
  Differential coding of raw genomic sequences relies on the identification of common patterns (i.e. "signatures") shared among several sequences. These common patterns are encoded only once together with the nucleotides specific to each read (i.e. the "residuals"). The presence of such signatures enables the implementation of indexing schemes whereby the compressed data can be searched by means of patterns matching algorithms.
- **High throughput and low latency**: When data throughput and low streaming latency have higher priority with respect to compression efficiency, MPEG-G also supports a "high throughput" compression approach which can be applied as soon as genomic sequences become available. In such case no data preprocessing over the entire dataset is necessary prior to the actual encoding. This approach efficiently supports streaming scenarios.

5.2 Mapped sequences

Genomic sequences mapped onto reference sequences can be compressed in an MPEG-G file following two approaches:

- **Reference-based compression**: In this approach the genomic sequences are represented by the differences they present with respect to the reference sequences as well as by the associated alignment information. This approach requires the availability of the reference sequences both at the encoding and decoding sides. The references sequences can optionally be delivered to the decoder within the same MPEG-G file.
- **Reference-free compression**: In this approach the aligned sequences are compressed without referring to any reference sequence. In this case, a local assembly of the underlying sequence is built per each group of reads, and reference-based compression is then applied with respect to the computed local assembly. In this case, there is no need to have access to any reference sequences at the encoder or at the decoder side.

5.3 Data Classes

Genomic Records are classified into six Data Classes according to the result of the primary alignment(s) of their reads against one or more reference sequences as shown in Table 1.

Records are classified according to the types of mismatches with respect to the reference sequences used for alignment.

Class name	**Semantics**
P	Reads perfectly matching to the reference sequence.
N	Reads containing mismatches which are unknown bases only.
M	Reads containing at least one substitution, and possibly unknown bases, but no insertions, no deletions and no clipped bases.
I	Reads containing at least one insertion, deletion or clipped base, and possibly unknown bases or substitutions.
HM	Half-mapped pairs where only one read is mapped.
U	Unmapped reads.

Table 1. Data classes defined in MPEG-G.

5.4 Access Unit properties

MPEG-G supports the description of alignments contained in Genomic Records by exposing the following information at the Access Unit level:

- Genomic Records count
- Number of Genomic Records with a count of substitutions below a given threshold (Class M)
- Presence of multiple alignments
- Presence of spliced reads
- Left-most and right-most mapped base for primary and secondary alignments
- Unmapped reads signature (Class U)

5.5 Selective access

The indexing tools embedded in an MPEG-G file enable the following types of selective access that can be combined in the same query:

- Genomic interval in terms of start to end mapping position on a given reference sequence
- Type of data (i.e., a single data class)
- Sequence reads with number of substitutions below/above a certain threshold

- Sequence reads with multiple alignments
- Pattern matching on raw or unmapped reads

5.6 Data streaming

MPEG-G supports the packetization of compressed data for transport to receiving devices that can start processing the data before transmission is completed. The main features of MPEG-G streaming are:

- Packet size adaptation to the channel characteristics/state
- Error detection and support of re-transmission of erroneous/incomplete data for error-free delivery
- Support of out-of-order delivery
- On-the-fly indexing of streamed data
- Packet-based filtering of genomic data
- Full convertibility of file and transport formats

An illustration of the essential components of the MPEG-G transport format is shown in Figure 5.

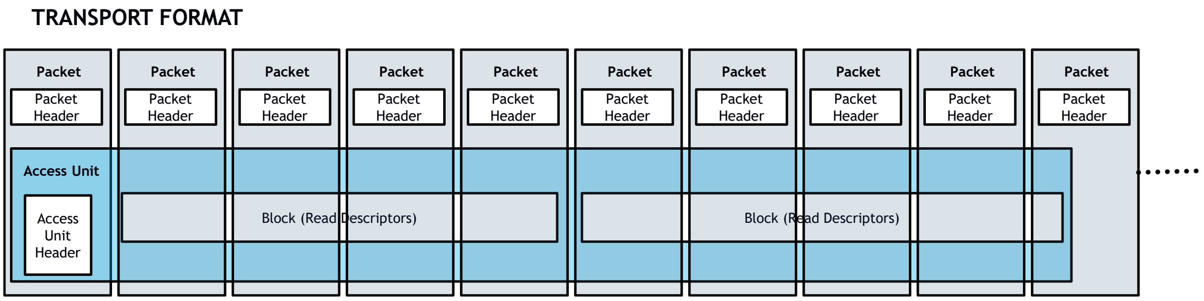

Figure 5. Illustration of the essential components of the MPEG-G transport format.

6 Usage of the MPEG-G standard

The MPEG-G specification is not limited in scope to syntax, semantics and decoding process specifications, as it also comprises normative Reference Software and Conformance testing.

The entire specification opens a wide range of possible usage scenarios and applications for the MPEG-G standard. New tools could be developed, and existing tools could be improved using the new technology. This new combination of tools can be the start of entire genomic ecosystems fueled by MPEG-G, as depicted in Figure 6. These tools could include retrieving, searching for, manipulating, or adding genomic information elements – such as headers, information metadata, protection metadata, chromosomes, genes, parts of the previous structures, and many more – based on the concepts specified and supported by the MPEG-G standard.

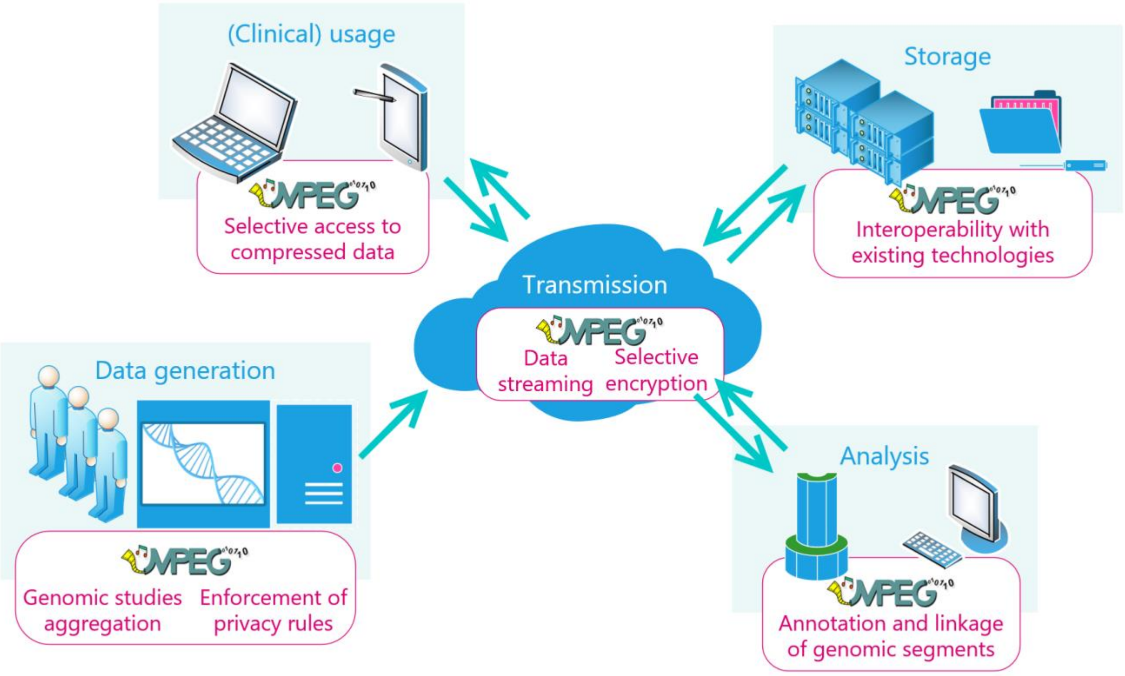
(Clinical) usage
MPEG
Selective access to compressed data
Storage
MPEG
Interoperability with existing technologies
Transmission
MPEG
Data streaming
Selective encryption
Data generation
MPEG
Genomic studies aggregation
Enforcement of privacy rules
Analysis
MPEG
Annotation and linkage of genomic segments

Figure 6: A genomic ecosystem fueled by MPEG-G.

To support and guide potential implementers of MPEG-G, the standard will include a normative Reference Software. The Reference Software is normative in the sense that any conforming implementation of the decoder, taking the same conformant compressed bitstreams, using the same normative output data structures, will output the same data. That said, complying MPEG-G implementations are not expected to follow the algorithms, or the programming techniques used by the Reference Software: such software is solely intended as a support to the process of developing implementations of an ecosystem of compliant devices and applications. The availability of a normative implementation covering the technology of the MPEG-G specification is only an additional support to the textual specification.

Thus, although the Reference Software is considered normative in terms of the semantics of the decoding process, it does not add anything to the textual technical specification of the MPEG-G standard. It has also to be underlined that being its objective to support the specification, the Reference Software has not to be intended as an efficient implementation of an MPEG-G standard decoder to be used as a benchmark of performance for sequential or parallel implementations.

Finally, the MPEG-G standard includes Conformance. This part of the standard is fundamental in providing means to test and validate the correct implementation of the MPEG-G technology in different devices and applications and the interoperability among all systems. Conformance testing specifies a normative procedure to assess conformity to the standard on an exhaustive dataset of compressed data: every decoder claiming MPEG-G conformance will have to demonstrate the correct decoding of the complete Conformance testbed.

7 References

[1] S. D. Kahn, “On the Future of Genomic Data,” *Science*, vol. 331, no. 6018, pp. 728–729, Feb. 2011.

[2] I. Numanagić, J. K. Bonfield, F. Hach, J. Voges, J. Ostermann, C. Alberti, M. Mattavelli, and S. C. Sahinalp, “Comparison of high-throughput sequencing data compression tools,” *Nat. Methods*, vol. 13, no. 12, pp. 1005–1008, Dec. 2016.

[3] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C.

Schatz, S. Sinha, and G. E. Robinson, “Big Data: Astronomical or Genomical?,” *PLOS Biol.*, vol. 13, no. 7, p. e1002195, Jul. 2015.

[4] I. Ochoa, M. Hernaez, R. Goldfeder, T. Weissman, and E. Ashley, “Effect of lossy compression of quality scores on variant calling,” *Brief. Bioinform.*, vol. 18, no. 2, pp. 183–194, Mar. 2016.

[5] Y. W. Yu, D. Yorukoglu, J. Peng, and B. Berger, “Quality score compression improves genotyping accuracy,” *Nat. Biotechnol.*, vol. 33, no. 3, pp. 240–243, Mar. 2015.

[6] L. Chiariglione, Ed., *The MPEG Representation of Digital Media*, 1st ed. Springer New York, 2012.

[7] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.