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

ISO/IEC JTC1/SC29/WG11 MPEG2020/M54566
June 2020, Online

Source: ISO/IEC JTC1/SC29/WG11
Status: Input Document
Title: **Proposed Additional Use Cases for MPEG-G Part 6**
Author: Mikel Hernaez (Center for Applied Medical Research at University of Navarra, UIUC), Idoia Ochoa (Tecnun at University of Navarra, UIUC), Patrick Y.H. Cheung (Royal Philips), Shubham Chandak (Stanford University)

1 Introduction

In this document we showcase real use cases of interest, that highlight features that the standard should support to increase its functionality and adoption by the broader community. Proposed enhancements to the current working draft for Part 6 (N19314) in relation to these use cases are presented in the companion document M54565.

2 Faithful Reconstruction of Source Data Files

There are several settings in which being able to faithfully reconstruct the source data files must be required. As such, the standard should support this feature. For example, in the clinical setting, it may be requested to reproduce (recreate) the outcome of a given analysis to confirm that the clinical decision was the correct one based on those results. There are several tools whose output varies if the input data is not exactly the same, even if the information contained is equivalent. Hence, in this scenario it would be desirable to being able to reconstruct the original source file.

In a similar fashion, data integrity is a must in some applications. For example, in genomic files containing sensitive information, one may want to verify that no third parties have accessed or corrupted the file. One way to verify this would be to check for data integrity, by checking if the reconstructed file and the original one are indeed identical. For that, the order and attributes should be exactly preserved, otherwise it may become impossible to verify that the information contained in two files are equivalent.

3 Fast Data Reconstruction for Downstream Analysis

While efficient query and random access to selective regions of interest are useful for data presentation and visualization, fast reconstruction of a whole data matrix is critical for supporting downstream data analysis. In fact, one major use of genomic data files is to be ingested and analyzed by bioinformatics tools or pipelines for scientific discovery or future

validation. The following are some examples of tertiary NGS analysis that operates on a whole data matrix:

- Gene expression – principal components analysis (PCA) for quality control of samples, differential expression analysis for identifying the dysregulated genes, machine learning analysis for gene signature discovery, etc.
- Variant genotypes – filtering analysis for identifying causal mutations in Mendelian diseases, rare-variant association analysis for identifying gene-disease associations, etc.

4 Easy Adaptation to Changing and Emerging File Formats

The genomics field is undergoing a rapid growth thanks to the improvement in sequencing capabilities, as well as to the design of new sequencing technologies and new methods for sequencing several types of omics data. As such, there exists a broad set of omics data files that are constantly evolving. In addition, we are experiencing a non-static paradigm in which new data types/formats are being proposed.

In this setting, it will be favorable to design a standard that can easily adapt to new data formats, as well as modifications to existing formats. As an example of this rapidly evolving setting, GA4GH is already working on a new VCF format more adequate to the current analyses involving VCF files. In a similar fashion, we observe a rapid growth in the generation of single-cell sequencing data, both from RNA and DNA. This has given rise to new methods and analytic pipelines involving single-cell data. As new methods are proposed, new capabilities of the file formats will be needed. As such, it would be beneficial to design the MPEG-G standard having in mind that to facilitate its use and adoption, new support may be required by its users. Hence designing a standard that can easily adapt to evolving and emerging file formats will be advantageous for the standard itself.