# 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**
**MPEG2017/m41117**
**July 2017, Torino, Italy**

**Source:** **DMAG-UPC et al.**
**Status:** **Proposal**
**Title:** **Use case for representing VCF information into MPEG-G**
**Authors:** **Jaime Delgado, Daniel Naro, Silvia Llorente (Distributed Multimedia Applications Group – Universitat Politècnica de Catalunya),**
**Josep Lluís Gelpí, Dmitry Repchevsky, Romina Royo (Barcelona Supercomputing Center),**
**Lola Alonso, Mirari Márquez, Núria Malats (CNIO, Spanish National Cancer Research Centre)**

**Table of contents**

# 1 Purpose

This document describes a possible new use case where Variant Call Format (VCF) [1] files are required to complete a genomic study. It represents a real use case raised by genomic research centres.

# 2 Proposal of new use case

## 2.1 VCF information into MPEG-G

| | |
|---|---|
| **UCId** | GI-UC09 |
| **Use Case Name** | Use VCF files to complete a genomic study |
| **Description** | Genomic information can be stored into different file formats, depending on the kind of information it stores. One of these file formats is VCF, which stores nucleotide differences vs. some reference at a given position in an individual genome or transcriptome.<br>VCF information can be used to conduct specific studies about the individual's medical condition. During the research, it could be possible that original genomic information, either aligned or not, needs to be accessed, using MPEG-G standard.<br>For instance, during a study, using VCF information, a researcher may ask for different window sizes depending on if SNVs or indels are required. These windows could be generated from the complete genomic information contained in a MPEG-G file. |
| **Actors** | Analyst (using MPEG-G files, VCF files, …), Organisations storing genomic information |
| **Assumptions** | The genomic information is conveniently stored and accessible, since the analyst has performed the variant calling (either after sequencing and alignment or genotyping using a SNP array).<br>For instance, during a study, using the information inside the VCF file, a researcher may ask for different window sizes depending on if SNVs or indels are required. These windows could be generated from the complete genomic information contained in a MPEG-G file. |
| **Actions** | 1. The analyst uses VCF files to conduct research.<br>2. The research results indicate that it is required to access part or the whole genome of the individual (SNVs or indels), stored in an MPEG-G file.<br>3. The analyst accesses to the MPEG-G file to recover the required information.<br>4. The analyst continues the study with the information coming from VCF and MPEG-G. Different tools should be used for accessing both of them. |
| **Inclusion or Extensions** | MPEG-G should be extended to represent VCF information. Several VCF files could be associated to an individual or study. |
| **Issues** | It could be desirable to include VCF information together with aligned / unaligned genomic information regarding the same individual into MPEG-G files at the dataset level to facilitate access, use and management to interested parties. In this case, the tools for accessing information in step 4 should be integrated.<br>It could also be desirable to allow MPEG-G to represent and manage multisample VCF files originating from the analysis of multiple individuals by storing this information at the dataset group level. |

# 3 References

[1] The Variant Call Format Specification, http://samtools.github.io/hts-specs/VCFv4.3.pdf, July 2016.