**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N18146**

**January 2019, Marrakesh, MA**

|  |  |
| --- | --- |
| **Source** | **Requirements** |
| **Status** | **Approved** |
| **Title** | **MPEG-G Genomic Information Database** |

# Purpose

This document contains the description of the MPEG-G Genomic Information Database (“Database”).

The Database contents listings can be found in the file “mpeg-g-gidb.xlsx” attached to this document. This file contains two tables:

1. Sequencing Data Collection: a collection of statistically meaningful sequencing data to be used to assess the performance of genomic information compression technologies.
2. Data origins: data origins of the Sequencing Data Collection.

Further work on the Database is discussed on the AHG on common MPEG and GA4GH standardization activities email reflector: [MPEG-GA4GH-tnt@listserv.uni-hannover.de](mailto:MPEG-GA4GH-tnt@listserv.uni-hannover.de).

Questions can be addressed to [mpeg-g@tnt.uni-hannover.de](mailto:mpeg-g@tnt.uni-hannover.de).

# Data classes

To make the Database statistically meaningful, sequencing data with different characteristics are considered.

## Experiment types

The Database includes sequencing data generated for different experiment types:

* Whole genome sequencing (WGS)
  + Including simulated human WGS data which was generated with ART [1]
  + Including cancer genome sequencing data
* Metagenomics sequencing
* RNA sequencing (RNA-Seq)

## Organisms

The Database includes sequencing data from the following species:

* Animalia
  + *D. melanogaster*
  + *H. sapiens*
* Plantae
  + *T. cacao*
* Fungi
  + *S. cerevisiae*
* Bacteria
  + *E. coli* (different strains)
  + *P. aeruginosa*
* Viruses
  + Phi X 174

## Sequencing technologies

The Database includes sequencing data which was generated with the following sequencing technologies:

* Sequencing by synthesis
  + Illumina/Solexa Genome Analyzer
  + Illumina Genome Analyzer IIx
  + Illumina MiSeq
  + Illumina HiSeq 2000
  + Illumina HiSeq X Ten
  + Illumina NovaSeq 6000
* Combinatorial Probe-Anchor Synthesis (cPAS) / DNA Nanoballs (DNB) technology
  + BGISEQ-500
* Single molecule real time sequencing
  + Pacific Biosciences SMRT (PacBio)
* Nanopore sequencing
  + Oxford Nanopore MinION
* Ion semiconductor sequencing
  + Ion Torrent PGM

# Data formats

Unmapped sequencing data are provided in the form of gzipped FASTQ files. FASTQ files are usually manipulated with custom scripts written in Bash, Python, Perl etc.

Mapped sequencing data are provided in the form of BAM files. Transcoding of data from the BAM format to the SAM format can be done using the Samtools program suite (http://www.htslib.org) [2]. Manipulation of data which is stored in the SAM and BAM formats can also be achieved with the Samtools program suite.

# License

*MPEG itself places no additional restrictions on the use or redistribution of the data available via its online services other than those provided by the original data owners.*

*The original data may be subject to rights claimed by third parties, including but not limited to, patent, copyright, other intellectual property rights, biodiversity-related access and benefit-sharing rights. It is the responsibility of users of the Database to ensure that their exploitation of the data does not infringe any of the rights of such third parties.*

# References

|  |  |
| --- | --- |
| [1] | W. Huang, L. Li, J. R. Myers and G. T. Marth, "ART: a next-generation sequencing read simulator," *Bioinformatics,* vol. 28, no. 4, pp. 593-594, 2012. |
| [2] | H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics,* vol. 25, no. 16, pp. 2078-2079, 2009. |