**ISO/IEC JTC 1/SC 29/WG 11  
Coding of moving pictures and audio  
Convenorship: UNI (Italy)**

**Document type: Approved WG 11 document**

**Title: MPEG-G Genomic Information Database**

**Status: Approved**

**Date of document: 2019-07-19**

**Source: Requirements**

**Expected action:**

**No. of pages: 11**

**Email of convenor: leonardo@chiariglione.org**

**Committee URL: http://mpeg.chiariglione.org**

**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 N18645**

**Gothenburg, SE – July 2019**

**Source: Requirements**

**MPEG-G Genomic Information Database**

**Table of Contents**

[1 Database contents 3](#_Toc14454519)

[2 Sequencing data collection 4](#_Toc14454520)

[2.1 Data classes 4](#_Toc14454521)

[2.1.1 Experiment types 4](#_Toc14454522)

[2.1.2 Organisms 4](#_Toc14454523)

[2.1.3 Sequencing technologies 5](#_Toc14454524)

[2.2 Data formats 5](#_Toc14454525)

[3 Annotation data collection 6](#_Toc14454526)

[3.1 File formats 6](#_Toc14454527)

[3.1.1 Genome annotations 7](#_Toc14454528)

[3.1.1.1 BED 7](#_Toc14454529)

[3.1.1.2 GFF/GTF 7](#_Toc14454530)

[3.1.2 Browser tracks 8](#_Toc14454531)

[3.1.2.1 bedGraph, bedDetails 8](#_Toc14454532)

[3.1.2.2 Wig/BigWig 8](#_Toc14454533)

[3.1.3 Sequence variations (VCF) 8](#_Toc14454534)

[3.1.4 Feature matrices and peak calling data 9](#_Toc14454535)

[3.1.4.1 Matrix Market 10](#_Toc14454536)

[3.1.4.2 HDF5 10](#_Toc14454537)

[3.1.4.3 loom 10](#_Toc14454538)

[3.1.5 Hi-C experiments (.hic) 10](#_Toc14454539)

[4 License 10](#_Toc14454540)

[5 References 11](#_Toc14454541)

# Database contents

This document contains the description of the MPEG-G Genomic Information Database. The database consists of the following parts:

* Sequencing data collection, pertaining to ISO/IEC 23092-2. A listing of the contents can be found in the attached file “w18645\_Sequencing\_data\_collection.xlsx”. The file contains three tables:
  1. Data: a collection of statistically meaningful sequencing data to be used to assess the performance of genomic information coding technologies.
  2. Origins: URLs kept as a trace of the data origins. Some of them might not work anymore.
  3. Reference sequences: available reference sequences. Some of them were used for the alignment of BAM files.
* Conformance test items, pertaining to ISO/IEC 23092-5.
* Annotation data collection, pertaining to ISO/IEC 23092-6. A listing of the contents can be found in the attached file “w18645\_Annotation\_data\_collection.xlsx”. The file contains one table:
  1. Data & origins: a collection of statistically meaningful annotation data to be used to assess the performance of genomic annotation coding technologies. The IDs are composed in the format x.y.z with
* x denoting the species:
  + 1: *Homo sapiens*
  + 2: *Mus musculus* (house mouse)
  + 3: *Rattus norvegicus* (common rat)
  + 4: *Zea mays*
* y denoting the format group:
  + 1: genome annotations
  + 2: browser tracks
  + 3: sequence variations
  + 4: Feature matrices and peak calling data
  + 5: Hi-C experiments
* z denoting the dataset identifier (i.e. z is an incremental counter)

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

# Sequencing data collection

# 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)
  1. Including simulated human WGS data which was generated with ART [1]
  2. 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.

# Annotation data collection

# File formats

ISO/IEC 23092-6 is pertaining to use cases relating to the support of secondary and tertiary analyses. The identified use cases are currently supported by a variety of different text format files with only partially defined syntax.

These text format files are meant to support:

* Genome annotations (e.g. localization of genes, exons and coding intervals etc.)
* Browser tracks (e.g. intervals of the genome associated to a scalar quantity used to display genome features in genome browsers)
* sequence variations (e.g. germline variations with respect to the species genome, or somatic variations with respect to the same individual genome)
* Feature matrices and peak calling data (e.g. matrices containing counts of the occurrence of specific features in single-cell RNA sequencing (scRNA-seq) experiments)
* Hi-C experiments (e.g. contact matrices representing features of genomic organization, such as the activity of chromatin regions)

Five main categories have been identified, together with representative file formats for each of them:

* Genome annotations:
  + BED (Browser Extensible Data),
  + GFF (General Feature Format)/GTF (Gene Transfer Format)
* Browser tracks:
  + bedGraph, bedDetails,
  + Wig/BigWig
* Sequence variations:
  + VCF (Variant Call Format)
* Feature matrices and peak calling data:
  + Matrix Market,
  + HDF5 (Hierarchical Data Format),
  + loom
* Hi-C experiments:
  + .hic

# Genome annotations

File formats used to represent genome annotations span from generic formats (JSON, CSV, TSV, RDF, etc.) to dedicated formats having an associated compressed representation. The latter seem to be the most wide-spread formats. They are also used by many genome browsers to graphically represent genomic features. Among them several versions of the BED and GFF/GTF file formats were identified.

# BED

The BED format is a text format comprising:

* an optional header containing directives for the genome browser,
* a fixed number of fields per line, where each field is associated to an annotated feature,

resulting in a sorted column structure; three of them are mandatory; nine of them are optional and can be omitted, provided that the lower-numbered position of the highest specified are always populated.

The three mandatory fields identify the annotated region with the reference sequence and a genomic range. The optional fields are used to: provide a name, identify the strand, a sub-range (e.g. to identify start-end codons), the number and position of exons and a set of graphic features such as the color palette or the opacity of the line to be drawn by the browser.

The BED format is present in several, non-backward compatible versions such as bedGraph and bedDetail which are mostly used as tracks for genome browsers. Notably, some of those formats add columns for gene expression and related statistical information (e.g. p-value, q-value) and are specific to the peak-calling performed in chromatin studies. In some cases, BED files are compressed in the custom format bigBed using a dedicated tool (<https://www.encodeproject.org/software/bedToBigBed>).

# GFF/GTF

GFF is a text format with a structure similar to the BED format (i.e., header plus data columns). Version 3 of GFF provides the possibility to include FASTA content at the end of the file.

Nine mandatory columns specify the features (with respect to the twelve columns of the BED format), although empty fields can be signaled by the character ‘.’. In analogy to the BED format, each line specifies sequence, name, genomic range, graphic score, and strand. The remaining fields are used to specify the program used to generate the data, the reading frame, if it is a coding exon, and an attributes list. This last field is a semicolon separated list of key-value pairs that may also refer to predefined tags, included external ontology databases. The ninth field is the main difference with respect to previous versions of GFF, where it was used as a grouping name, and with respect to the GTF format (also known as GFF2.5 given the similarities with GFF3) using different reserved tags.

GTF/GFF files with most known features (genes, transcripts, exons, etc.) are stored in gzipped from in the main genome databases:

* Ensembl (<http://www.ensembl.org/info/data/ftp/index.html>)
* NCBI (<ftp://ftp.ncbi.nih.gov/genomes/>)
* USCS Genome Browser (<http://hgdownload.soe.ucsc.edu/downloads.html>)

# Browser tracks

Used for quickly displaying biological quantities in genome browsers, tracks associate scalar quantities to genome positions or intervals, possibly with different levels of granularity, to support different zoom levels.

The most common formats are BED-derived formats, such as bedGraph and bedDetails, and the Wig format, usually shipped in the compressed format BigWig.

# bedGraph, bedDetails

bedGraph identifies the genome by sequence, start and end position and an associated scalar value.

bedDetails adds two fields to the BED format: an identifier alternative to the name enabling the creation of links, and a description that may also be an HTML element.

# Wig/BigWig

The Wig format has a similar file structure compared to the BED-derived tracks: a header with meta information, included the track type, and a data section. Multiple Wig files may be concatenated in a single file to cover different feature tracks in a single place. The data section may have two different structures which may span over several contiguous datasets:

* a variableStep structure, where irregular intervals are associated to a scalar value using the start position of the interval,
* a fixedStep structure where the intervals are regular and identified with a start position and a step.

BigWig, being a custom compressed representation of a Wig file, requires a dedicated tool to be decoded into the Wig format: bigWigToWig.

bigWigToWig can be downloaded from <https://www.encodeproject.org/software/bigwigtowig/>.

# Sequence variations (VCF)

Sequence variation data are provided in the form of gzipped VCF files. The VCF specification is maintained by the Global Alliance for Genomics and Health (GA4GH) Large Scale Genomics Work Stream File Formats Task Team (<https://github.com/ga4gh/large-scale-genomics-wiki/wiki>). VCF files are usually manipulated with the VCFtools program package (<https://vcftools.github.io>). Pairs of VCF files can be compared with the Python program hap.py (<https://github.com/illumina/hap.py>). Manipulation with custom scripts written in Bash, Python, Perl etc. is also a common practice.

VCF is the widely used text format for storing variants. A VCF file is composed of a header containing metadata lines (e.g. file format, reference file URL, etc.), and “meta-information” lines, containing definitions used in the rest of the document (filters, formats, pedigree, etc.). The second part is composed of data lines with eight fixed fields plus eventual extra fields per genotype present, anticipated by a format definition column.

The eight fixed fields are:

* contig identifier
* 1-based position on the contig (for indels it refers to the previous mapped base)
* semi-colon separated list of unique identifiers
* reference base(s)
* alternate base(s) (comma separated list of alleles, \* indicates upstream deletion)
* quality value
* filter status (PASS for passed or a semicolon-separated list of failure codes)
* a list of tag-value pairs (some tag names are reserved)

# Feature matrices and peak calling data

Feature matrix and peak calling data are provided in the form of

* Matrix Market format files (<https://math.nist.gov/MatrixMarket/formats.html>),
* Hierarchical Data Format (HDF5) files,
* loom format files (<http://loompy.org>), which are based on HDF5 files.

Single-cell and bulk RNA sequencing experiments

In single cell RNA sequencing (scRNA-seq) and bulk-RNA sequencing, a sample ranging from 102 to 106 cells are isolated and mRNA transcripts are extracted. A “barcode” (BC) sequence is attached to tag the origin cell and, for single cell experiments, a random “Unique Molecule Identifier” (UMI) sequence to tag the mRNA transcripts. After sequencing, a “feature matrix” is created in order to count the occurrence of specific features in the cell’s population, by counting the UMI or the presence of reads inside specific genomic ranges. In the resulting matrix, by convention, each row represents a gene and each column represents a cell. The units by which the expression is measured depends on the protocol and the normalization strategy used.

ATAC- and ChIP-seq experiments

Chromatin studies and DNA proteins studies share a similar post-processing of the genomic data called “peak calling”. In ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), and ChIP-seq (Chromatin ImmunoPrecipitation DNA sequencing), the alignment process will result in local accumulations of reads (e.g. high coverage limited to the interesting regions) called “peaks”. The analysis results are stored in feature matrices and BED-like files.

# Matrix Market

The Matrix Market format is an ASCII-based file format to represent dense and sparse matrices. Sparse matrices are represented using only non-zero entries in a coordinate format (e.g. each line indicates: row, column, entries). The number of entries can be further reduced by specifying symmetry structures. It is used to store the feature matrix and is distributed in a compressed bundle together with two files where the row indexes are mapped to the cell barcodes and the column indices are mapped to the features.

# HDF5

HDF5 is file format designed to store and organize large amounts of data that can represent very complex data objects and a wide variety of metadata, allowing for access time and storage space optimizations. Several tools and APIs for many languages are available for managing, manipulating, viewing, and analyzing the data. It is mostly used to store feature matrices and related metadata. Being extremely general, several intermediate steps of an analysis pipeline can be stored in a single HDF5 file, regardless of the solution chosen to format them.

# loom

Loom is a file format for large omics datasets based on the HDF5 format. A loom file always contains a main matrix plus optional additional layers, a variable number of row and column annotations, and sparse graph objects.

# Hi-C experiments (.hic)

Hi-C experiment data are provided in the form of .hic files. .hic files can be manipulated with the Juicer Tools and Juicebox software (<https://github.com/aidenlab/juicer>).

Hi-C experiments are used to study the three-dimensional structure of DNA, identifying genome-wide long-range interactions. Cells are fixed with formaldehyde, chemically bounding close genomic: the links created persist during sequencing and can be detected and recorded. The output of the experiment is a genome wide contact matrix, stored in “.hic” files.

# 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. |