**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N17940**

**October 2018, Macao SAR, CN**

|  |  |
| --- | --- |
| **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 consists of two parts:

1. **Sequencing Data Collection:** a collection of statistically meaningful sequencing data to be used to assess the performance of genomic information compression technologies. Besides the sequencing data (listed in Table 1) the Sequencing Data Collection contains a set of reference sequences (listed in Table 3) and supplementary data for variant calling experiments (only listed online).
2. **Conformance Test Items:** a set of bitstreams for conformance testing according to ISO/IEC 23092-5. Table 4 lists the Conformance Test Items.

Further work on the Database is discussed on the AHG on Genomic Information Representation email reflector: [genome\_compression@listes.epfl.ch](mailto:genome_compression@listes.epfl.ch).

# Database access

The public access point to the Database is: <https://github.com/voges/mpeg-g-gidb>.

The Database is provided as the list of URLs (in Table 2) to public repositories where the data are available.

In case some resources are not available a copy of the data can be retrieved upon request to [mpeg-g@tnt.uni-hannover.de](mailto:mpeg-g@tnt.uni-hannover.de) with the following details:

* URL: https://www.tnt.uni-hannover.de/mpeg-g/
* User name: mpeg-g
* Password: to be requested from [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
* 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.

# Database

The paths listed in this section are relative to the Database URL as provided in section 2 of this document. The Database is organized in the following top-level folders:

* candidate-data: files which are under consideration for a possible incorporation into the Sequencing Data Collection
* collection: Sequencing Data Collection
* conformance: Conformance Test Items
* development: test files used during the development of the reference software

## Sequencing Data Collection

Table 1 provides the selected data.

Table 1 – Sequencing Data Collection

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Path** | **File name(s)** | **Experiment type** | **Scientific name** | **Sequencing technology** | **Coverage** | **Comments** |
| 01 | collection/wgs/h-sapiens/ERP001775/ | ERR174324\_1.fastq.gz  ERR174324\_2.fastq.gz  ..  ERR174341\_1.fastq.gz  ERR174341\_2.fastq.gz | WGS | *H. sapiens* | Illumina HiSeq 2000 | n/a | Deep whole genome sequence data for the CEPH 1463 family. These genomes include a trio (NA12877, NA12878 and NA12882) sequenced to greater than 200x depth of coverage, as well as a technical replicate (separate library and sequencing, but same DNA sample) of NA12882 also sequenced to greater than 200x. Additional information and analyses available at [www.platinumgenomes.org](http://www.platinumgenomes.org).  Used in [3] |
| 02 | collection/wgs/h-sapiens/ERP001960/ | NA12878\_S1.bam  NA12879\_S1.bam  NA12890\_S1.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 52.3x  53.3x  46.9x | Whole genome sequence and data for 3 members of the Coriell CEPH/UTAH 1463 family. All genomes are sequenced to greater than 30x. Additional information and analyses available at [www.platinumgenomes.org](http://www.platinumgenomes.org).  Selected genomes:   * SAMEA1573614 * SAMEA1573618 * SAMEA1573617 |
| 03 | collection/wgs/h-sapiens/NA12878/ | NA12878.pacbio.bwa-sw.20140202.bam | WGS | *H. sapiens* | PacBio | 8.4x | Variable-length reads |
| 04 | collection/wgs/h-sapiens/ERP002490/ | NA12877\_S1.bam  NA12878\_S1.bam  NA12882\_S1.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 17.4x  26.1x  24.1x | Long range mate-pair data for the CEPH 1463 family. These data have a physical insert size (distance between the pair of reads) of ~2000 bp and compared with a normal paired end data that spans ~300 bp, for the same total sequence depth this mate pair data will contain almost 7x the number of read pairs that cross any specific break point. |
| 05 | collection/wgs/h-collection/sapiens/ERR317482WGS/ | 9827\_2#49.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 2.3x | Used in [4] |
| 06 | collection/wgs/h-sapiens/NA21144.chrom11/ | NA21144.chrom11.ILLUMINA.bwa.GIH.low\_coverage.20130415.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 7.4x | The data was processed with GATK. This means the quality values have higher entropy due to the recalibration process and it now has additional auxiliary data.  Used in [4] |
| 07 | collection/wgs/h-sapiens/ERS179576/ | ERR174310\_1.fastq.gz  ERR174310\_2.fastq.gz | WGS | *H. sapiens* | Illumina HiSeq 2000 | n/a | This is the first pair of files in item 01. It can be used to test compression of unaligned low coverage human WGS data. |
| 08 | collection/wgs/h-sapiens/PacBio\_CHM1htert\_54x/ | \*.fastq.gz | WGS | *H. sapiens* | PacBio | n/a | The dataset contains the raw sequence data resulting from PacBio SMRT sequencing for CHM1htert, a human cell line derived from a hydatidiform mole. |
| 09 | collection/wgs/h-sapiens/IonTorrent/ | sample-2-10\_sorted.bam  sample-2-11\_sorted.bam  sample-2-12\_sorted.bam | WGS | *H. sapiens* | Ion Torrent PGM | 274.0x  238.9x  225.4x | n/a |
| 11 | collection/wgs/h-sapiens/NA12878-SRX517292/ | SRR1238539.fastq.gz | WGS | *H. sapiens* | Ion Torrent PGM | n/a | Used in [3] |
| 12 | collection/wgs/h-sapiens/NA12878\_Garvan\_replicate\_J/ | NA12878\_V2.5\_Robot\_2\_R1.fastq.gz  NA12878\_V2.5\_Robot\_2\_R2.fastq.gz | WGS | *H. sapiens* | Illumina HiSeq X Ten | n/a | Reduced quality value resolution (8-level binning) as explained in  <https://www.illumina.com/documents/products/whitepapers/whitepaper_datacompression.pdf>.  Used in [3] |
| 13 | collection/wgs/e-coli/ERX593919/ | Ecoli\_R7\_NONI.fastq.gz | WGS | *E. coli* | Oxford Nanopore MinION | n/a | n/a |
| 14 | collection/wgs/e-coli/ERX593921/ | Ecoli\_R73.fastq.gz | WGS | *E. coli* | Oxford Nanopore MinION | n/a | n/a |
| 15 | collection/wgs/e-coli/SRR1284073/ | SRR1284073.fastq.gz | WGS | *E. coli* | PacBio | n/a | n/a |
| 16 | collection/wgs/e-coli/DH10B/ | MiSeq\_Ecoli\_DH10B\_110721\_PF.bam | WGS | *E. coli* | Illumina MiSeq | 447.8x | Used in [3] and [5] |
| 17 | collection/wgs/e-coli/ERA269036/ | 9799\_7#3.bam | WGS | *E. coli* | Illumina HiSeq 2000 | 453.7x | Used in [4] |
| 18 | collection/wgs/s-cerevisiae/SRX089128/ | SRR327342.bam | WGS | *S. cerevisiae* | Illumina/Solexa Genome Analyzer | n/a | n/a |
| 19 | collection/wgs/d-melanogaster/D-melanogaster/ | dm3PacBio.bam | WGS | *D. melanogaster* | PacBio | 81.2x | Used in [3] |
| 24 | collection/wgs/t-cacao/T-cacao/ | SRR870667.fastq.gz | WGS | *T. cacao* | Illumina Genome Analyzer IIx | n/a | n/a |
| 25 | collection/wgs/p-aeruginosa/SRX181937/ | SRR554369\_1.fastq.gz  SRR554369\_2.fastq.gz | WGS | *P. aeruginosa* | Illumina Genome Analyzer IIx | n/a | n/a |
| 26 | collection/wgs/h-sapiens/Simulated\_human\_genome\_sequencing\_data/ | simulation.1.homoINDELs.homoCEUsnps.reads2.fq.sam.samelength.bam | WGS | *H. sapiens* | ART (mimicking Illumina sequencing technology) | n/a | Input files to the ART simulator (can be found in the same folder):   * indels.simulation1.vcf.gz * simulation.1.fa.gz |
| 31 | collection/wgs/h-sapiens/ERP001775-aln/ | ERR174324.aln\_bowtie2.sorted.dupmark.rg.realn.recal.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 15.2x | Aligned data from item 01 |
| 32 | collection/wgs/h-sapiens/NovaSeq/ | NA12878-Rep-1\_S1\_L001\_2\_R1\_001.fastq.trimmed.gz  NA12878-Rep-1\_S1\_L001\_2\_R2\_001.fastq.trimmed.gz | WGS | *H. sapiens* | Illumina NovaSeq 6000 | n/a | Trimmed reads |
| 33 | collection/wgs/phi-x-174/PhiX-NovaSeq-UltraHighDepth/ | Phix1\_NoIndex\_L001\_R1\_001.fastq.gz  Phix1\_NoIndex\_L001\_R2\_001.fastq.gz | WGS | *Phi X 174* | Illumina NovaSeq 6000 | n/a | Ultra-high-depth sequencing |
| 20 | collection/metagenomics/Human\_gut/ | MH0001\_081026\_clean.1.fq.gz  MH0001\_081026\_clean.2.fq.gz  MH0002\_081203\_clear.1.fq.gz  MH0002\_081203\_clear.2.fq.gz  MH0003\_081203.clean.1.fq.gz  MH0003\_081203.clean.2.fq.gz | Metagenomics sequencing | n/a | Illumina Genome Analyzer IIx | n/a | n/a |
| 21 | collection/wgs/HCC1143\_NORMAL\_60x/ | G15511.HCC1143\_BL.1.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 56.0x | Normal tissue |
| 22 | collection/wgs/HCC1143\_TUMOR\_50x/ | G15511.HCC1143.1.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 46.5x | Tumor tissue |
| 23 | collection/wgs/HCC1954+BL/ | HCC1954.mix1.n80t20.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | 25.9x | UCSC artificial mixed sample: 80% HCC1954BL 20% HCC1954 |
| 10 | collection/rna-seq/RNAseq/ | K562\_cytosol\_LID8465\_TopHat\_v2.bam | RNA-Seq | *H. sapiens* | Illumina HiSeq 2000 | n/a | n/a |
| 30 | collection/rna-seq/Re-aligned\_RNAseq/ | K562\_cytosol\_LID8465\_GEM\_v3.bam | RNA-Seq | *H. sapiens* | Illumina HiSeq 2000 | n/a | Re-aligned from item 10 with the GEM mapper [6] |
| 34 | collection/wgs/h-sapiens/bqsr/ | ERR174310.aln.sort.dupmark.rg.recal.bam | WGS | *H. sapiens* | Illumina HiSeq 2000 | n/a | The BAM file was obtained by running an alignment pipeline on the files ERR174310\_1.fastq and ERR174310\_2.fastq from item 01. Most importantly, in the last two pipeline steps (GATK BaseRecalibrator and GATK ApplyBQSR), a technique known as Base Quality Score Recalibration (BQSR) was applied. The original quality scores get replaced with the recalibrated quality scores. The original, i.e., uncalibrated, quality scores are moved to a SAM auxiliary field identified by the tag “OQ” (“original qualities”). The supplementary file “ERR174310.aln.sort.dupmark.rg.bam.recal\_data.table” (can be found in the same folder), generated by the tool GATK BaseRecalibrator, contains–among other information–the assignments of original quality scores to recalibrated quality scores. |

Table 2 provides the data origins of the Sequencing Data Collection. These URLs are kept as a trace of the origins. Some of them might not work anymore

Table 2 – Sequencing Data Collection origins

|  |  |
| --- | --- |
| **ID** | **Origin** |
| 01 | http://www.ebi.ac.uk/ena/data/view/ERP001775 |
| 02 | <http://www.ebi.ac.uk/ena/data/view/ERP001960> |
| 03 | <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20131209_na12878_pacbio/si/NA12878.pacbio.bwa-sw.20140202.bam> |
| 04 | <http://www.ebi.ac.uk/ena/data/view/ERP002490> |
| 05 | <http://www.ebi.ac.uk/ena/data/view/ERR317482> |
| 06 | <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA21144/alignment/NA21144.chrom11.ILLUMINA.bwa.GIH.low_coverage.20130415.bam> |
| 07 | <http://www.ebi.ac.uk/ena/data/view/ERP001775> |
| 08 | http://datasets.pacb.com/2014/Human54x/fast.html |
| 09 | <http://www.ebi.ac.uk/ena/data/view/ERX276880>  <http://www.ebi.ac.uk/ena/data/view/ERX276881>  <http://www.ebi.ac.uk/ena/data/view/ERX276882> |
| 11 | <ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA096/SRA096885/SRX517292/SRR1238539.fastq.bz2> |
| 12 | <https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R1.fastq.gz>  <https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R2.fastq.gz> |
| 13 | <http://www.ebi.ac.uk/ena/data/view/ERX593919> |
| 14 | <http://www.ebi.ac.uk/ena/data/view/ERX593921> |
| 15 | <ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX533/SRX533603> |
| 16 | <ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/DH10B/MiSeq_Ecoli_DH10B_110721_PF.bam> |
| 17 | ftp://ftp.sra.ebi.ac.uk/vol1/ERA269/ERA269036/bam/ |
| 18 | <ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA043/SRA043851/SRX089128> |
| 19 | <http://bergmanlab.ls.manchester.ac.uk/data/tracks/dm3/dm3PacBio.bam> |
| 24 | <http://www.ncbi.nlm.nih.gov/sra/SRX288435> |
| 25 | <ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA058/SRA058002/SRX181937> |
| 26 | n/a |
| 31 | n/a |
| 32 | n/a |
| 33 | n/a |
| 20 | <ftp://ftp.sra.ebi.ac.uk/vol1/ERA000/ERA000116/fastq/MH0001_081026_clean.1.fq.gz>  <ftp://ftp.sra.ebi.ac.uk/vol1/ERA000/ERA000116/fastq/MH0001_081026_clean.2.fq.gz>  <ftp://ftp.sra.ebi.ac.uk/vol1/ERA000/ERA000116/fastq/MH0002_081203_clear.1.fq.gz>  <ftp://ftp.sra.ebi.ac.uk/vol1/ERA000/ERA000116/fastq/MH0002_081203_clear.2.fq.gz>  <ftp://ftp.sra.ebi.ac.uk/vol1/ERA000/ERA000116/fastq/MH0003_081203.clean.1.fq.gz>  <ftp://ftp.sra.ebi.ac.uk/vol1/ERA000/ERA000116/fastq/MH0003_081203.clean.2.fq.gz> |
| 21 | <https://cghub.ucsc.edu/datasets/benchmark_download.html> |
| 22 | <https://cghub.ucsc.edu/datasets/benchmark_download.html> |
| 23 | <https://cghub.ucsc.edu/datasets/benchmark_download.html> |
| 10 | [http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1728/K562\_ cytosol\_LID8465\_TopHat\_v2.bam](http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1728/K562_%20cytosol_LID8465_TopHat_v2.bam) |
| 30 | n/a |

Table 3 provides a list of available reference sequences. Some of them were used for the alignment of BAM files from Table 1.

Table 3 – Reference sequences

|  |  |  |
| --- | --- | --- |
| **Path** | **File name(s)** | **Scientific name** |
| collection/reference-sequences/h-sapiens/ | hg19.fa.gz  hg38.fa.bz2  Homo\_sapiens\_assembly19.fasta.gz  hs37d5.fa.gz | *H. sapiens* |
| collection/reference-sequences/e-coli/ | DH10B\_WithDup\_FinalEdit\_validated.fa.gz  EcoliDH10B.fa.gz  CP000948.fasta.gz  NC\_000913.2.fa.gz | *E. coli* |
| collection/reference-sequences/t-cacao/ | Theobroma\_cacao.Theobroma\_cacao\_20110822.dna.toplevel.fa.gz | *T. cacao* |

## Conformance Test Items

Table 4 provides the Conformance Test Items.

Table 4 – Conformance Test Items

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Path** | **File names(s)** | **Comments** |
| AbL-001 | conformance/abl-001/ | abl-001.mgg | n/a |
| AbL-002 | conformance/abl-002/ | abl-002.mgg | n/a |
| AbL-003 | conformance/abl-003/ | abl-003.mgg | n/a |
| AbL-004 | conformance/abl-004/ | abl-004.mgg | n/a |
| AbL-005 | conformance/abl-005/ | abl-005.mgg | n/a |
| AbL-006 | conformance/abl-006/ | abl-006.mgg | n/a |
| AbL-007 | conformance/abl-007/ | abl-007.mgg | n/a |
| AbL-008 | conformance/abl-008/ | abl-008.mgg | n/a |
| AbL-009 | conformance/abl-009/ | abl-009.mgg | n/a |
| AbL-010 | conformance/abl-010/ | abl-010.mgg | n/a |
| AbL-011 | conformance/abl-011/ | abl-011.mgg | n/a |
| AbL-012 | conformance/abl-012/ | abl-012.mgg | n/a |
| AbL-013 | conformance/abl-013/ | abl-013.mgg | n/a |
| AbL-014 | conformance/abl-014/ | abl-014.mgg | n/a |
| AbL-015 | conformance/abl-015/ | abl-015.mgg | n/a |
| AbL-016 | conformance/abl-016/ | abl-016.mgg | n/a |
| AbL-017 | conformance/abl-017/ | abl-017.mgg | n/a |
| AbL-018 | conformance/abl-018/ | abl-018.mgg | n/a |
| AbL-019 | conformance/abl-019/ | abl-019.mgg | n/a |
| AbL-020 | conformance/abl-020/ | abl-020.mgg | n/a |
| AbL-021 | conformance/abl-021/ | abl-021.mgg | n/a |
| AbL-022 | conformance/abl-022/ | abl-022.mgg | n/a |
| AbL-023 | conformance/abl-023/ | abl-023.mgg | n/a |
| Gen-001 | conformance/gen-001/ | gen-001.mgb | n/a |
| Gen-002 | conformance/gen-002/ | gen-002.mgb | n/a |
| Gen-003 | conformance/gen-003/ | gen-003.mgb | n/a |
| Gen-004 | conformance/gen-004/ | gen-004.mgb | n/a |
| Gen-005 | conformance/gen-005/ | gen-005.mgb | n/a |
| Gen-006 | conformance/gen-006/ | gen-006.mgb | n/a |
| Gen-007 | conformance/gen-007/ | gen-007.mgb | n/a |
| Gen-008 | conformance/gen-008/ | gen-008.mgb | n/a |
| Gen-009 | conformance/gen-009/ | gen-009.mgb | n/a |
| Gen-010 | conformance/gen-010/ | gen-010.mgb | n/a |
| Gen-011 | conformance/gen-011/ | gen-011.mgb | n/a |
| Gen-012 | conformance/gen-012/ | gen-012.mgb | n/a |
| Gen-013 | conformance/gen-013/ | gen-013.mgb | n/a |
| Gen-014 | conformance/gen-014/ | gen-014.mgb | n/a |
| Gen-015 | conformance/gen-015/ | gen-015.mgb | n/a |
| Gen-016 | conformance/gen-016/ | gen-016.mgb | n/a |
| Gen-017 | conformance/gen-017/ | gen-017.mgb | n/a |
| Gen-018 | conformance/gen-018/ | gen-018.mgb | n/a |
| Gen-019 | conformance/gen-019/ | gen-019.mgb | n/a |
| Gen-020 | conformance/gen-020/ | gen-020.mgb | n/a |
| Gen-021 | conformance/gen-021/ | gen-021.mgb | n/a |
| Gen-022 | conformance/gen-022/ | gen-022.mgb | n/a |
| Gen-023 | conformance/gen-023/ | gen-023.mgb | n/a |
| Gen-024 | conformance/gen-024/ | gen-024.mgb | n/a |
| Gen-025 | conformance/gen-025/ | gen-025.mgb | n/a |
| Gen-026 | conformance/gen-026/ | gen-026.mgb | n/a |
| Gen-027 | conformance/gen-027/ | gen-027.mgb | n/a |
| Gen-028 | conformance/gen-028/ | gen-028.mgb | n/a |
| Gen-029 | conformance/gen-029/ | gen-029.mgb | n/a |
| Gen-030 | conformance/gen-030/ | gen-030.mgb | n/a |
| Gen-031 | conformance/gen-031/ | gen-031.mgb | n/a |
| Gen-032 | conformance/gen-032/ | gen-032.mgb | n/a |
| QV-001 | conformance/qv-001/ | qv-001.mgb | n/a |
| QV-002 | conformance/qv-002/ | qv-002.mgb | n/a |
| QV-003 | conformance/qv-003/ | qv-003.mgb | n/a |
| QV-004 | conformance/qv-004/ | qv-004.mgb | n/a |
| QV-005 | conformance/qv-005/ | qv-005.mgb | n/a |
| QV-006 | conformance/qv-006/ | qv-006.mgb | n/a |
| QV-007 | conformance/qv-007/ | qv-007.mgb | n/a |
| Ref-001 | conformance/ref-001/ | ref-001.mgb | n/a |
| Ref-002 | conformance/ref-002/ | ref-002.mgb | n/a |
| Ref-003 | conformance/ref-003/ | ref-003.mgb | n/a |
| Ref-004 | conformance/ref-004/ | ref-004.mgb | n/a |
| Ref-005 | conformance/ref-005/ | ref-005.mgb | n/a |
| Ref-006 | conformance/ref-006/ | ref-006.mgb | n/a |
| Ref-007 | conformance/ref-007/ | ref-007.mgb | n/a |
| Ref-008 | conformance/ref-008/ | ref-008.mgb | n/a |
| Ref-009 | conformance/ref-009/ | ref-009.mgb | n/a |
| Ref-010 | conformance/ref-010/ | ref-010.mgb | n/a |
| Ref-011 | conformance/ref-011/ | ref-011.mgb | n/a |
| Ref-012 | conformance/ref-012/ | ref-012.mgb | n/a |
| MA-001 | conformance/MA-001/ | MA-001.mgb | n/a |
| MA-001 | conformance/MA-002/ | MA-002.mgb | n/a |
| MA-003 | conformance/MA-003/ | MA-003.mgb | n/a |
| MA-004 | conformance/MA-004/ | MA-004.mgb | n/a |
| MA-005 | conformance/MA-005/ | MA-005.mgb | n/a |
| MA-006 | conformance/MA-006/ | MA-006.mgb | n/a |
| MA-007 | conformance/MA-007/ | MA-007.mgb | n/a |

# 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. |
| [3] | J. Voges, J. Ostermann and M. Hernaez, "CALQ: compression of quality values of aligned sequencing data," *Bioinformatics,* vol. 34, no. 10, pp. 1650-1658, 2018. |
| [4] | J. K. Bonfield, "The Scramble conversion tool," *Bioinformatics,* vol. 30, no. 19, pp. 2818-2819, 2014. |
| [5] | F. Hach, I. Numanagic and S. C. Sahinalp, "DeeZ: reference-based compression by local assembly," *Nature Methods,* vol. 11, pp. 1082-1084, 2014. |
| [6] | S. Marco-Sola, M. Sammeth, R. Guigó and P. Ribeca, "The GEM mapper: fast, accurate and versatile alignment by filtration," *Nature Methods,* vol. 9, pp. 1185-1188, 2012. |