ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG on Genomic Information Representation (m43993)

M. Golebiewsky (HITS), J. Delgado (UPC),
M. Mattavelli (EPFL)

Joint AHG with ISO/IEC TC276

Mandates

1. To disseminate the information on the activities jointly carried out by ISO/IEC JTC 1/SC 29/WG 11 and ISO TC 276/WG 5 to other ISO TC and other organizations.
2. To finalize the organization of the joint MPEG-GA4GH workshop at next GA4GH plenary meeting in Basel and coordinate MPEG contributions.
3. To discuss and coordinate a collaboration with GATK development team to support MPEG-G as native format.
4. To discuss and coordinate MPEG contributions to relevant bioinformatics conferences and events.
5. To contribute to the editing and to the revision of the DIS of Part-1, Part-2, Part-3 and for the CD of Part-4 and Part-5.
6. To finalize the collection and definition of test item descriptions and binary streams for Conformance testing.
7. To identify work items for MPEG-G Version 2.
8. To finalize the organization of a WS on MPEG-G on the Saturday after the 124 MPEG meeting in Shenzhen.
9. To stimulate the collection of new content in line with new and emerging sequencing technologies.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG Activity Overview

- Technical work mainly focused on Editing of part 3 according to the NB comments received in Ljubljana
- Dissemination and promotion:
  - MPEG-GA4GH WS on "Workshop on Genomic Sequencing Data Compression" on the 3rd October in Basel during the GA4GH plenary meeting
    Slides: https://drive.google.com/drive/folders/1bQcgqPn_BAYNGlyCVrX_IRLDvYbQjf8t
  - MPEG-G presented at Europe Biobank Week Conf. Antwerp Sept. 6th

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

MPEG-G - a standard to compress DNA reads

13:00 - 18:00, 13th October 2018 Shenzhen (CN)

Venue: 2F Function Room, Tencent Building,
No. 10000 Shennan Avenue, NanShan District,
Shenzhen , Guangdong province, China

Start	End	What	Who
12:30	13:00	Registration	
13:00	13:10	Welcome & workshop goals	Leonardo Chiariglione (MPEG Convener)
13:10	13:40	"An overview of the MPEG-G standard for the compression and processing of genomic sequencing data"	Marco Mattavelli (EPFL, Switzerland)
13:40	14:10	"An overview of standardization initiatives on genomic data"	Yong Zhang (ISO/IEC TC276/WG3 Convener)
14:10	14:40	"A Review of Compression Technology of Genomics Data",	Yong Zhang (BGI Big Data Center)
14:40	14:50	Short presentation of demos	Demonstrators companies
14:50	15:20	Demo session and Coffee Break	
15:20	15:50	"State-of-the-art and future of NGS, a standard perspective"	Fang Chen (MGI)
15:50	16:20	"Constructing an open ecosystem for bioinformatics and genomic big data"	Chen Shifu (Haplox)
16:20	16:50	"Challenges and perspectives of genomic data processing services on the cloud"	*Title and speaker to be confirmed*
16:50	17:20	Panel discussion, Q&A and concluding remarks	All speakers
17:20	18:00	Demo session continues	

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG Activity Overview

- Progress of reference SW integration:
  - Part 1: some adjustments needed to align to the latest specs
  - Part 2: CABAC and descriptors decoder for all classes of data are integrated and tested with a few bitstreams
  - First 10 bitstreams available for further tests
  - TBD
    - Computed references
    - CABAC support of QV mode 0
    - Multiple alignments

Input documents review

m44035 - Proposed Updates to the MPEG-G Genomic Information Database

m44765 - Impact of forward and reverse reads to quality value compression efficiency

m44720 - MPEG-G file format in perspective of authoring tools

m44040 - Proposed CD Text for ISO/IEC 23092-5 Conformance

m44845 - Text proposal for ISO/IEC DIS 23092-3 Genomic Information Metadata and APIs

m44938 - Report on implementation of MPEG-G

m45077 - Requirements and Use cases on anonymization of Genomic Information Representation

m44049 - Study on ISO/IEC DIS 23092-2

m45056 - MPEG-G: New compression method for read names

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Input documents review

- m44035
  - New test item with recalibrated QVs (qv_depth > 1)
  - Add new test item no. 34 to the DB
  - Add text to the DB document
  - New DB output document
- m44765
  - Reversing QVs for the reads read from the reverse strand can save 0.5% to 1.7% in the compressed file
  - One additional flag to be added to the Parameter Set
  - Adjust decoding process accordingly

Input documents review

- m44720
  - Introduction of offset box to extend headers without file reassembly
  - applies to:
    - labels
    - metadata/protection
    - reference
    - MIT?
  - new gen_offset_info structure

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Input documents review

- m44040
  - List of conformance tests is complete
  - Offline editing needed to
    - complete coverage
    - add mention to DB and tools availability
  - 2 hours needed to review the document (Monday afternoon)
    - test descriptions
    - coverages
  - verify if the whole set of bitstreams must be available to go CD

Input documents review

- m44845 (Part 3)
  - revise
    - terms and definitions (Monday afternoon)
    - 5.2 Syntax functions and data types
    - 5.3 Graphic notations
    - All the highlighted text
  - document must be ready by Friday

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Input documents review

- m45077
  - Requirements for data anonymization
  - Verify if/how Part 3 already supports data anonymization and provide an example
- m44049
  - Two US NB comments to the next ballot

Notes

- Communicate on patents statement
- Informative encoder
  - Genie project (January)
  - GenomSys TBD
  - UPC TBD

Sessions during the week

- Terms and definitions: Sunday (Monday 17:30) 15:30 – 16:30
- Conformance: Monday 14:30 – 16:30
- m45056: Monday 16:45 – 17:15
- Part 3: Tuesday 9:00 – 12:00
- Alignment score: Tuesday 14:30 – 15:30
- st(v): set a limit and support of UTF-8 vs. ASCII: Tuesday 15:30 – 16:30
- MPEG-G paper evolution: Wednesday 11:00 – 12:00
- QVs for insertions: Wednesday 14:30 – 15:00
- CABAC QV mode 0: Wednesday 15:00 – 15:30

Output documents

- DB document
- Conformance CD text
- Use cases (TBD)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Recommendations

- Continue technical and editorial work on the DIS text of Part 3 during the week according to the received NB comments
- Draft text for an informative annex to Part 3 on SAMtools commands executed on the MPEG-G APIs
- Review of the CD text candidate and work plan for generation of test items for Part 5 Conformance
- Status and work plan for Part 1 and Part 2 Reference SW
- Continue the work on new use cases validation
- Update the database for new items and for conformance testing
- Review input on v2 features
- Work plan of dissemination activities