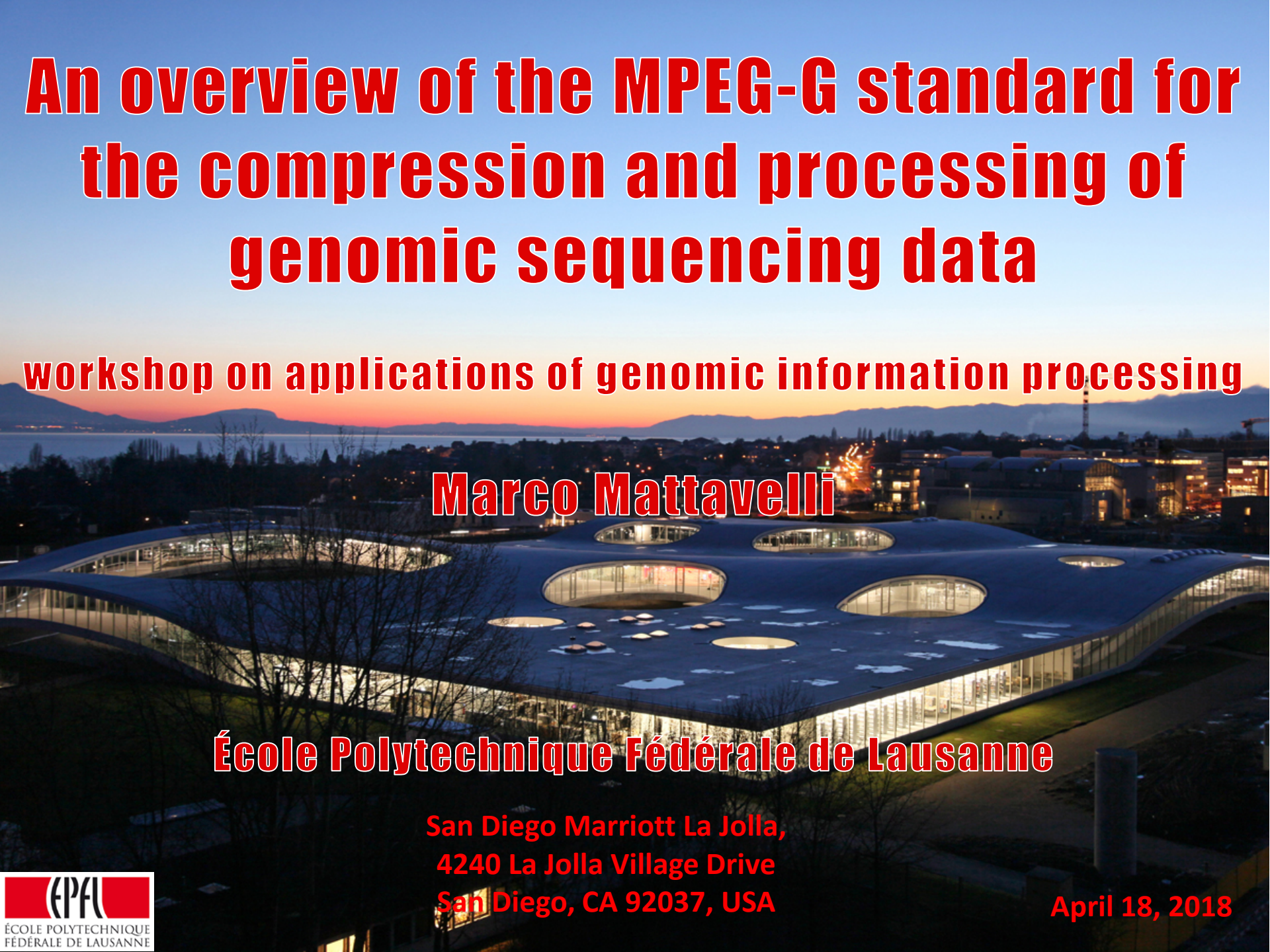
An overview of the MPEG-G standard for the compression and processing of genomic sequencing data

workshop on applications of genomic information processing



Marco Mattavelli

École Polytechnique Fédérale de Lausanne

San Diego Marriott La Jolla,
4240 La Jolla Village Drive
San Diego, CA 92037, USA

April 18, 2018

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Can we reduce the ICT costs of genomics?

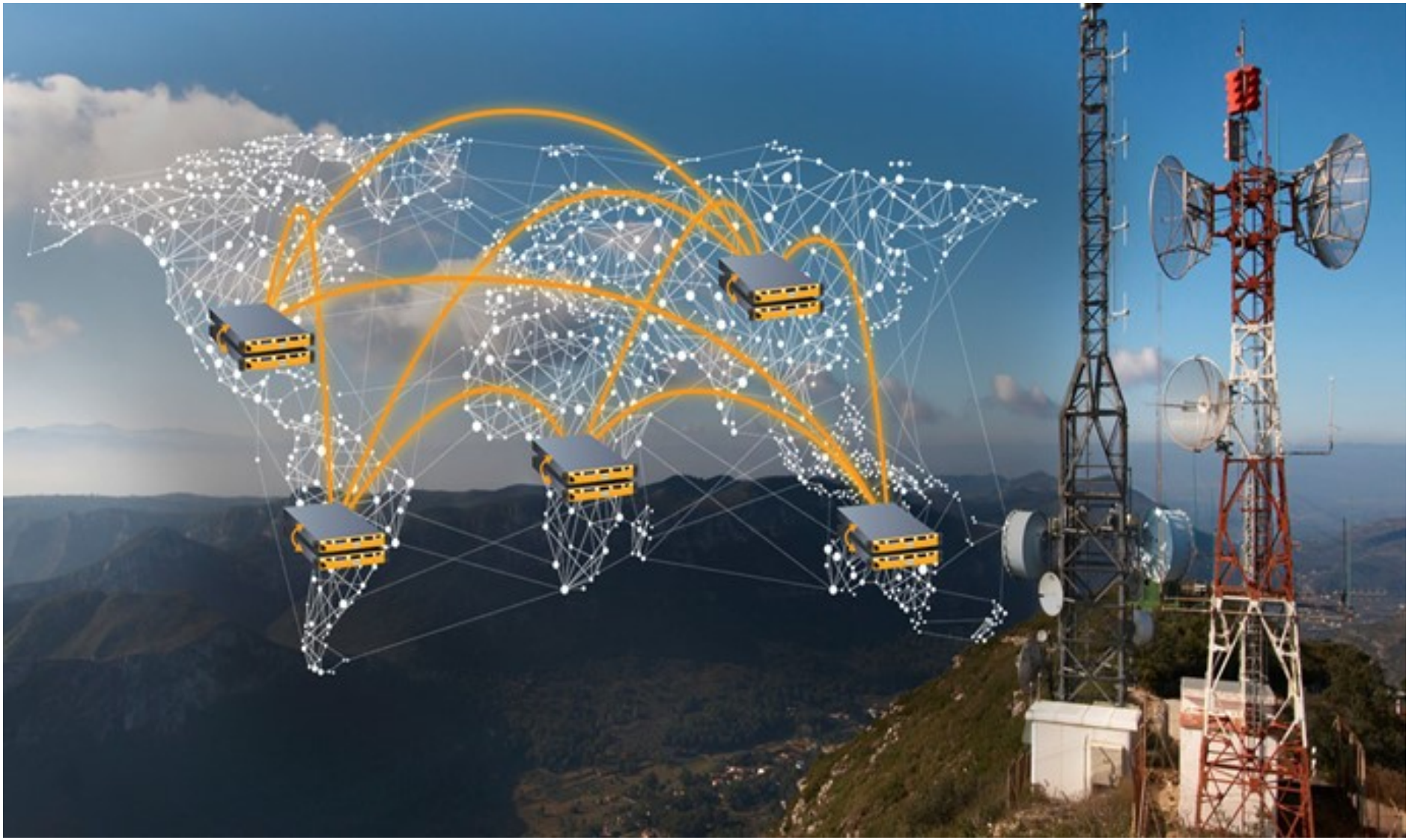
- A problem already seen in the recent past …..
- In the 1980s
  - Digital television handled ~200 Mbit/s (TV) and ~1 Gbit/s (HDTV)
  - Many proprietary compression format stifled the market
- In the early 1990s **MPEG developed MPEG-2**
  - Compression of ~100x and other functionalities (random access,...)
  - VLSI chips available from multiple sources
  - Developed, maintained and created a string of compression standards:
    - MPEG-2: ~50-100 x; MPEG-4: ~200 x; MPEG-H : ~400 x, MPEG-I: ~800 x
- On the path of what done in digital media **MPEG and ISO/TC276 are developing MPEG-G**,
  - Digital representation (including compression) of sequenced DNA
  - To be approved as International Standard in January 2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

The lesson of 25 years of MPEG Digital Media

All digital media content is compressed …..

…..today transmitted in compressed form everywhere …



Lesson from these 25 years:

- Compression is important, technology enabler, but it is not all: **MPEG «Systems APIs»** are even more important.
- Digital media applications are built «around» the MPEG standard «**Systems APIs**» :
  - All component are «synchronized» and linked
  - Access to data in the native compressed domain
- If a compression (standard) technology evolves the «**Systems**» standard «remain valid»!!

Can current genomic formats do the job?

- Current data storage and processing are based on an ASCII «file format» representation (a «huge» matrix of redundant information fields): SAM

- Genomic data compression is simply a SAM file zipped line by line: BAM

- Compression and selective data access using SAM and BAM are inefficient

Can current genomic formats do the job?

- **What is wrong about SAM and BAM?**
  - Merging heterogeneous data into a (simplistic!) file format and then attempting to compress it
    → **inefficient compression**
  - APIs not based on a native compressed format
    → **inefficient selective access to data**
  - Missing transport format supporting APIs and selective data access in the compressed domain
    → **inefficient access to remote data**

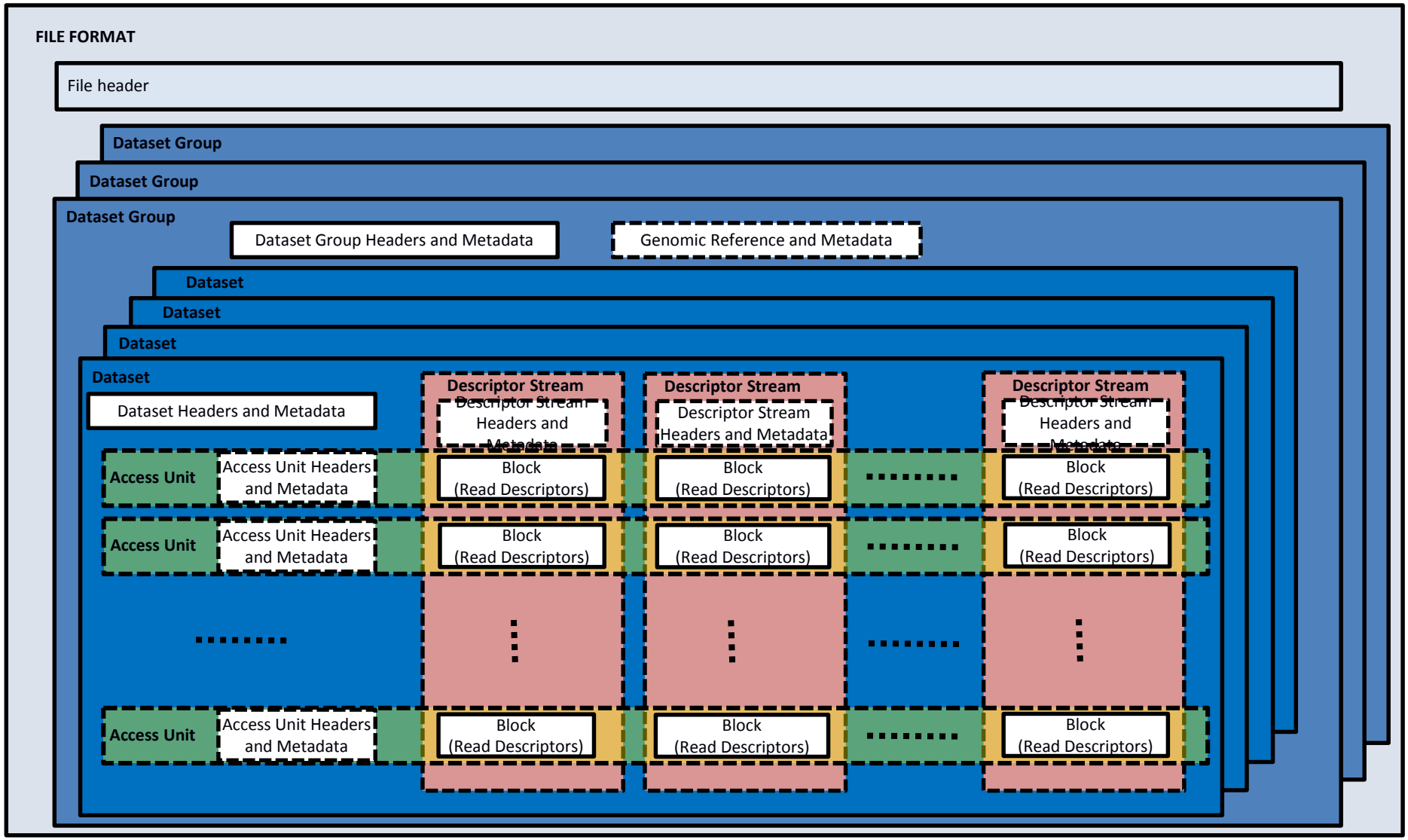
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

The MPEG-G Difference

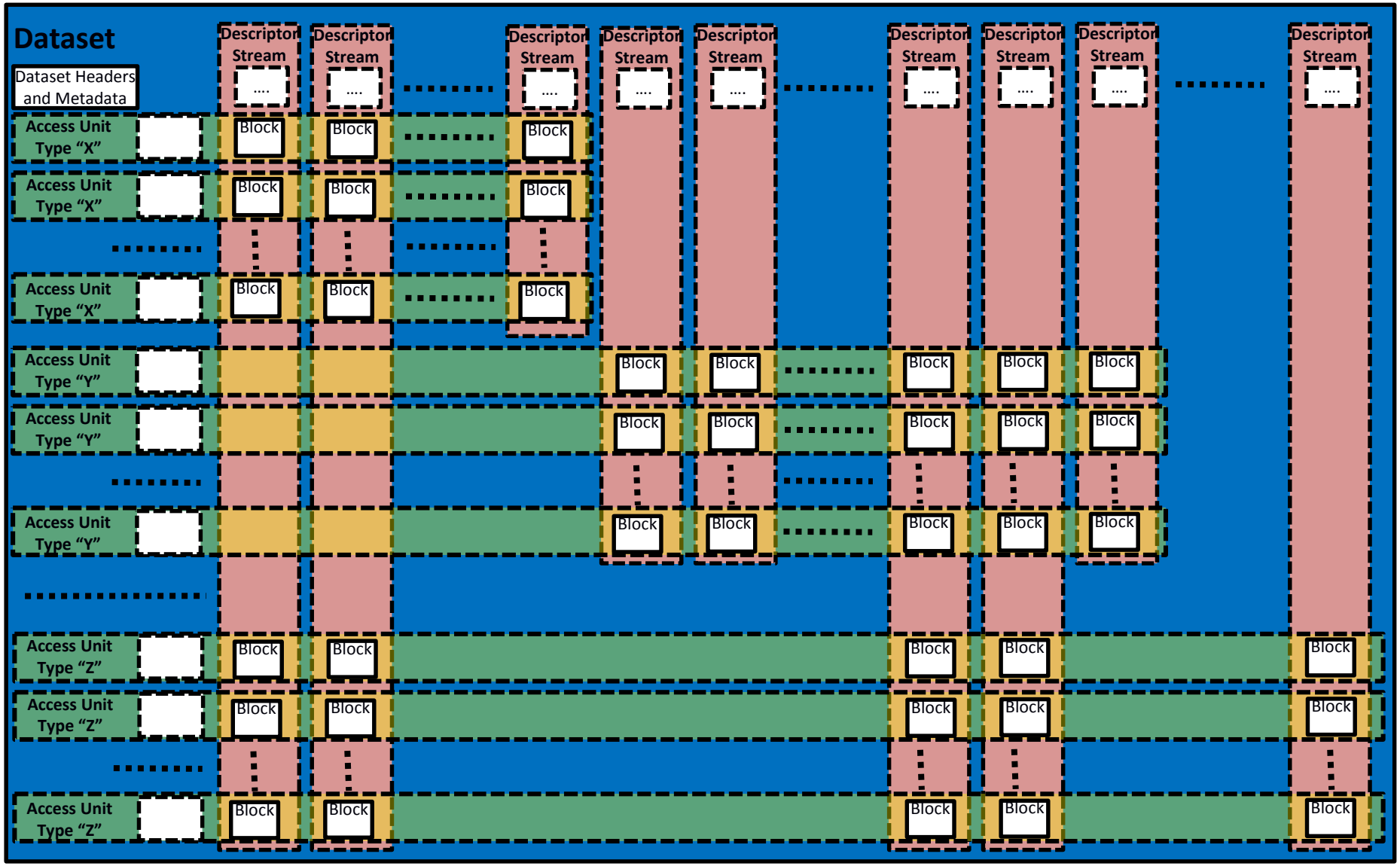
Can current genomic formats do the job?

- **The MPEG-G approach:**
  - Genomic data (i.e. sequence reads) are classified into homogeneous sets (classess of data) and represented by minimal sub-sets of «descriptors»
  - Meta-data associated to classified reads is represented by specific «descriptors»
  - Descriptors sub-sets are **compressed individually** and then are stored into structured **«Access Units»**
  - **Access Units** are included into an indexed **«File Format»**

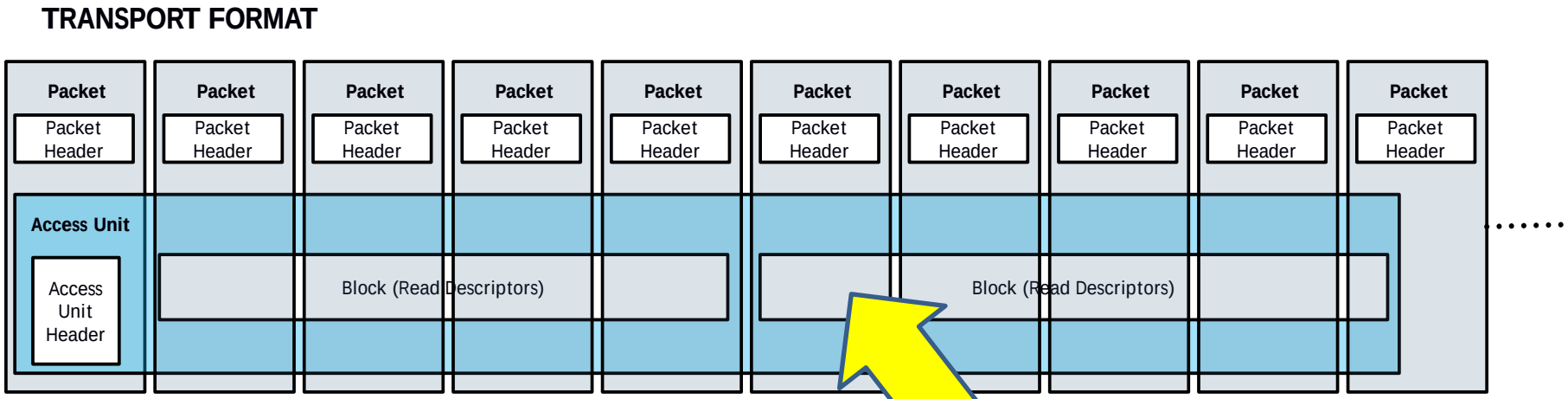
MPEG-G File Format



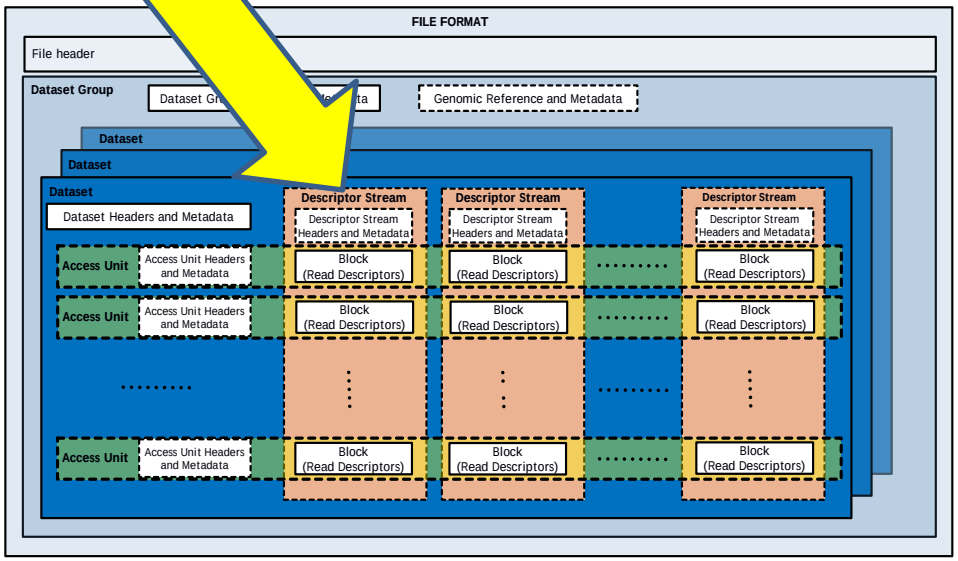
MPEG-G File Format



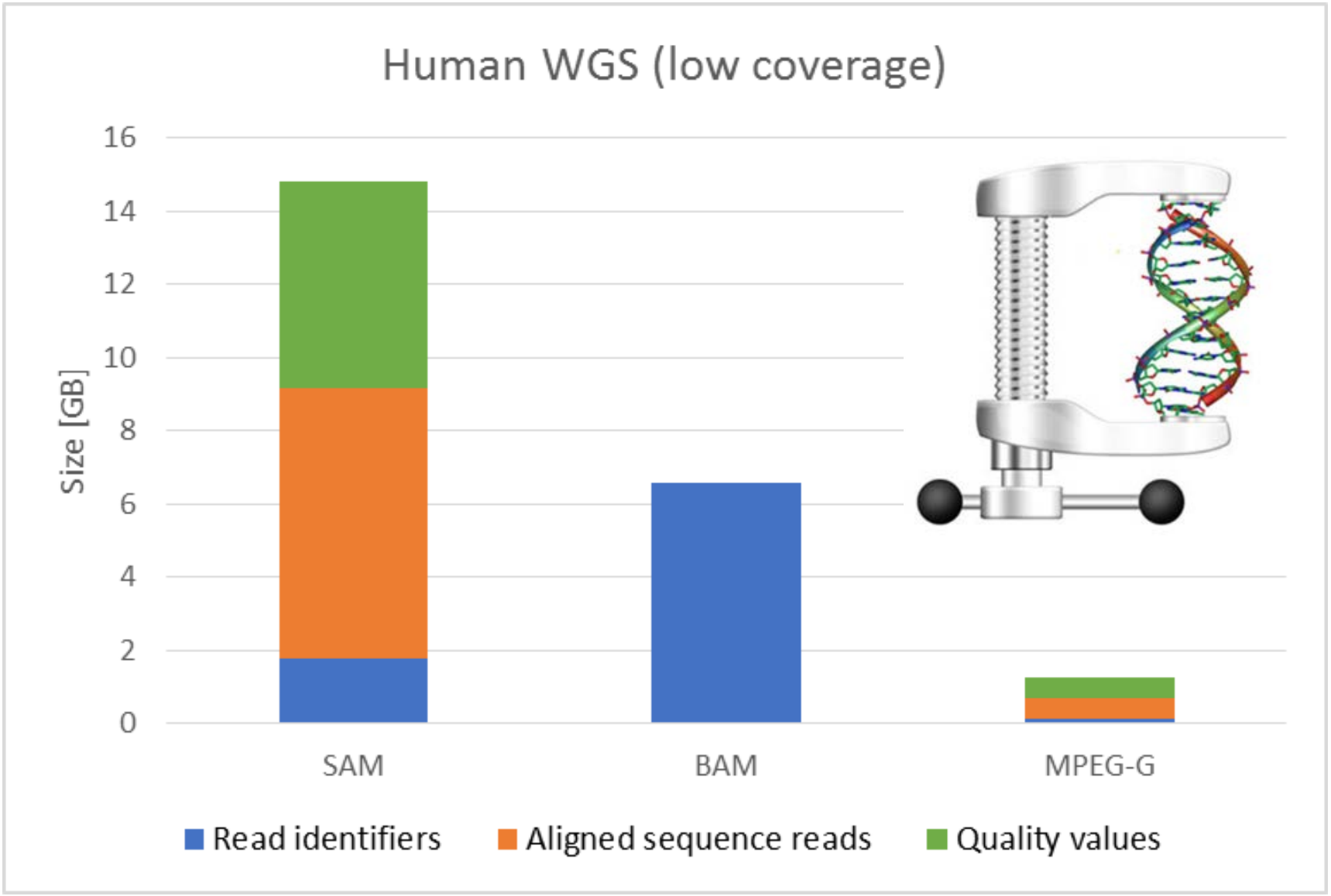
MPEG-G File Format and transport format



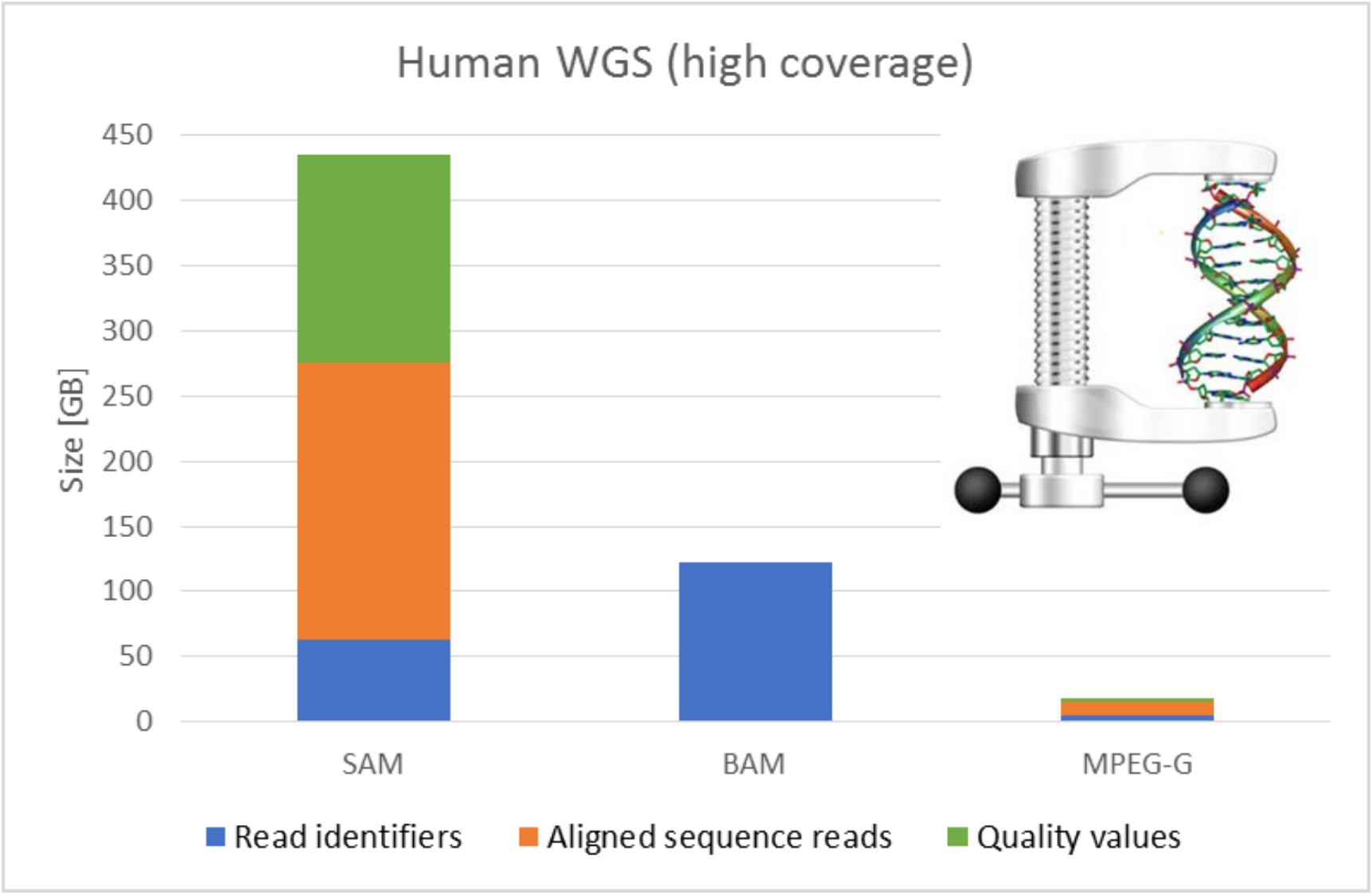
- Fully reversible conversion **File <--> Transport** as in ISOBMFF and MPEG-2 TS



MPEG-G Compression



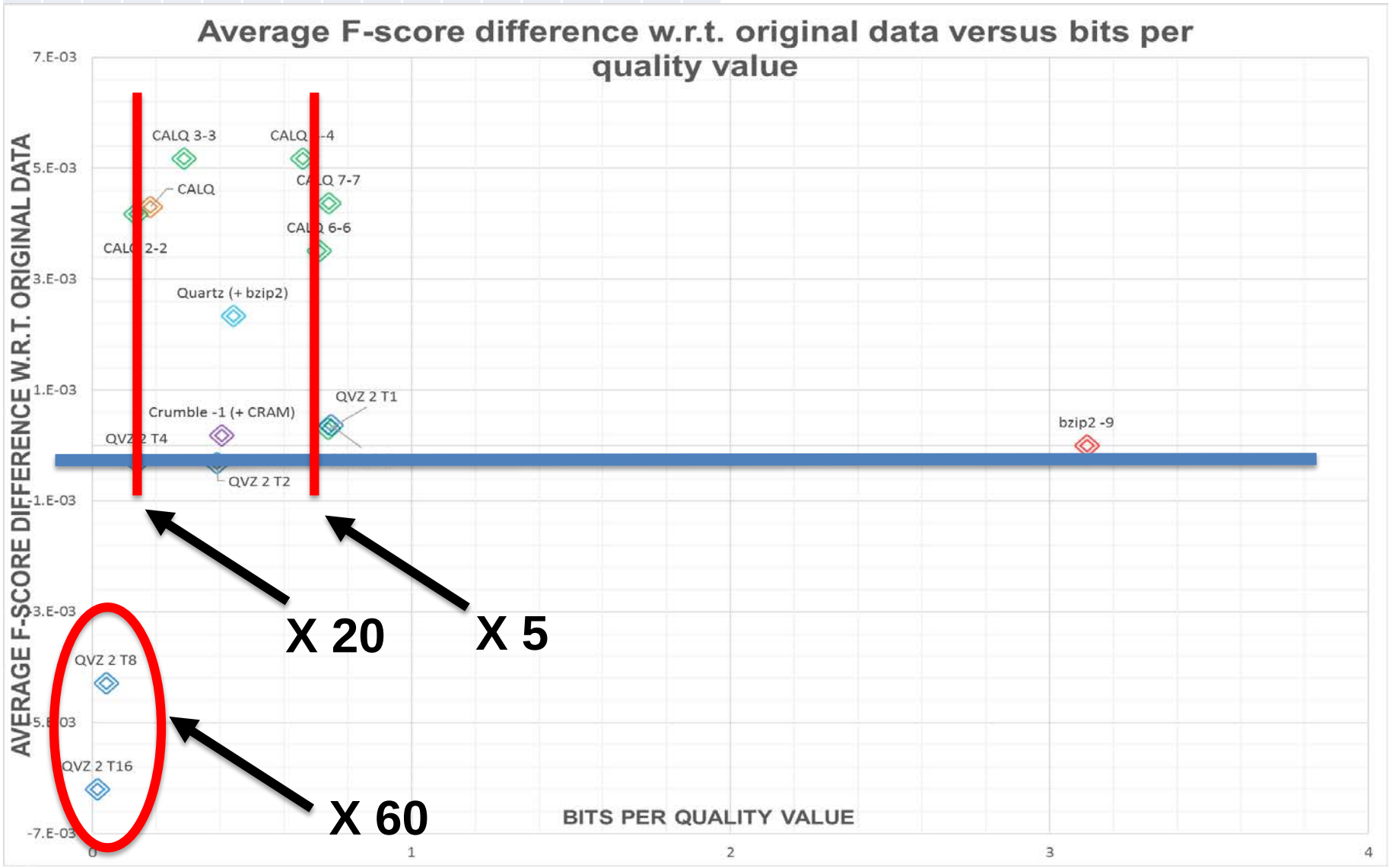
MPEG-G Compression



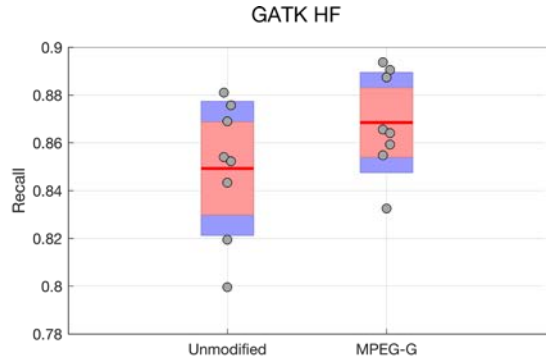
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

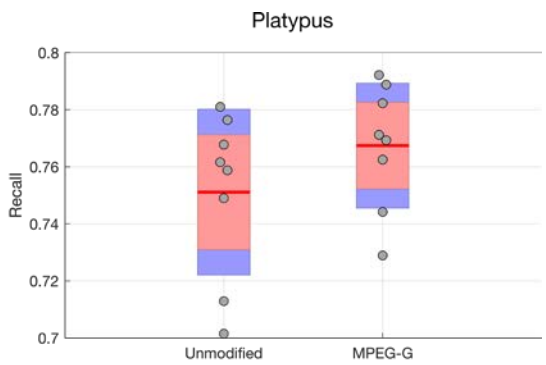
Compressing samples of a noisy process

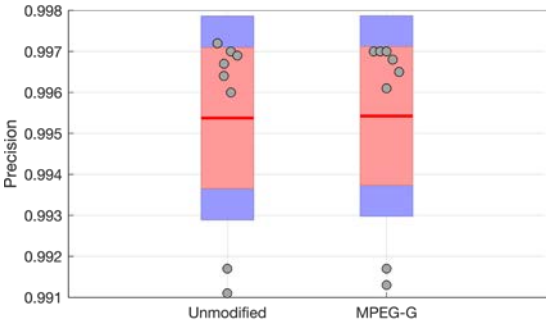
Rate-Distortion for Quantized Quality Values

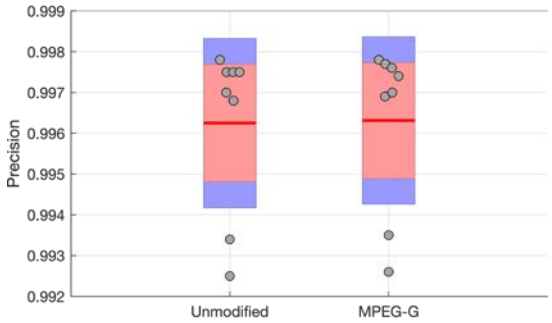


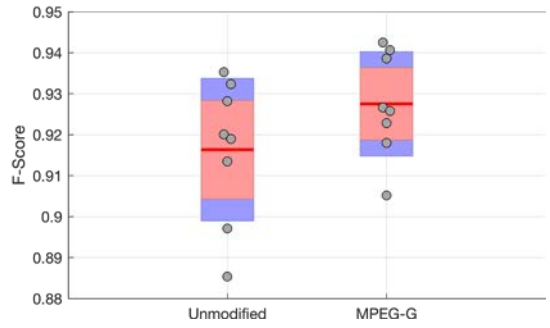
«Functional Equivalence» and QV quantization

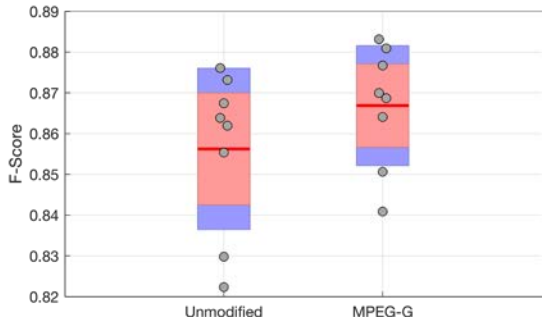




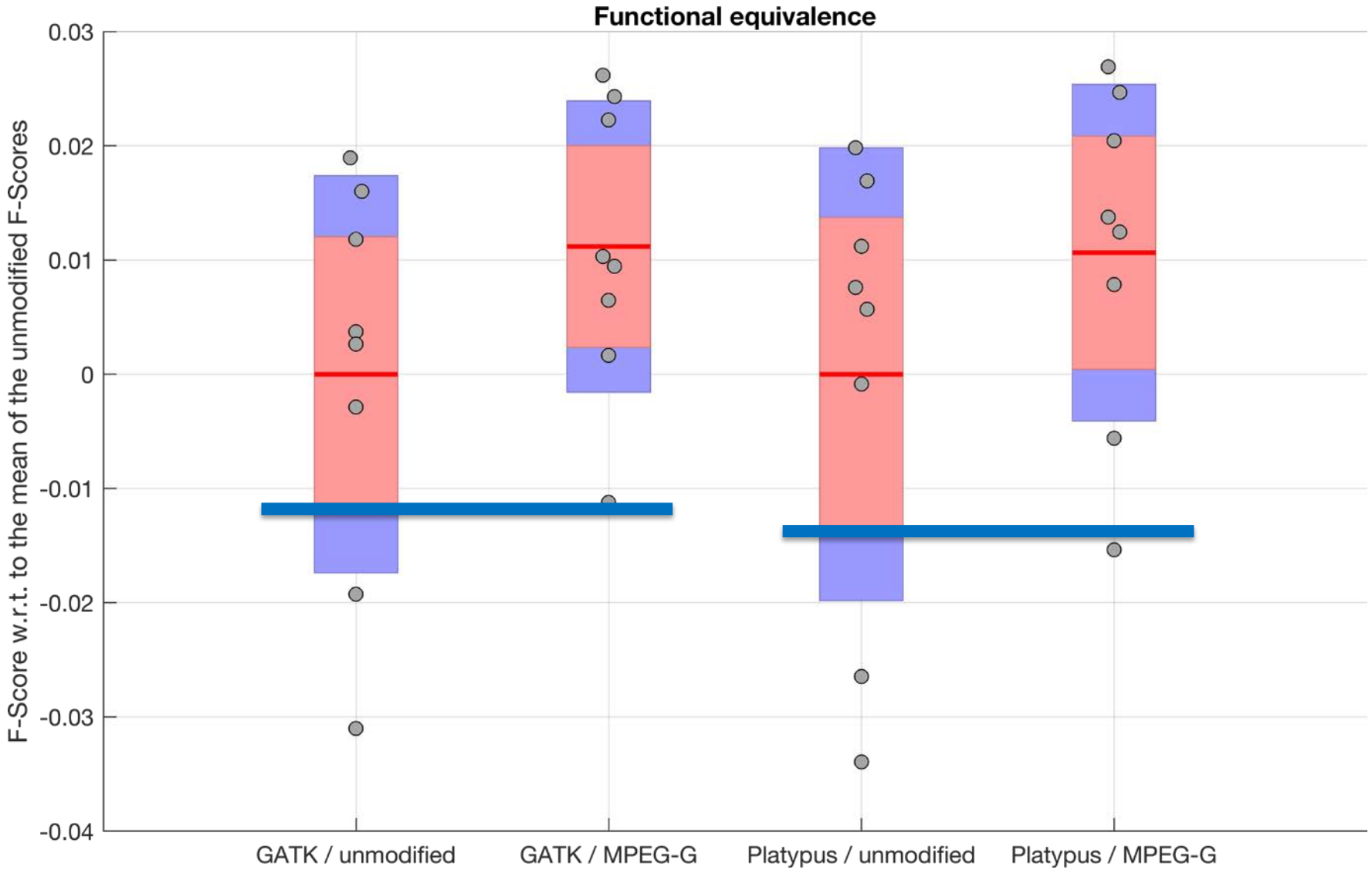








«Functional Equivalence» and QV quantization



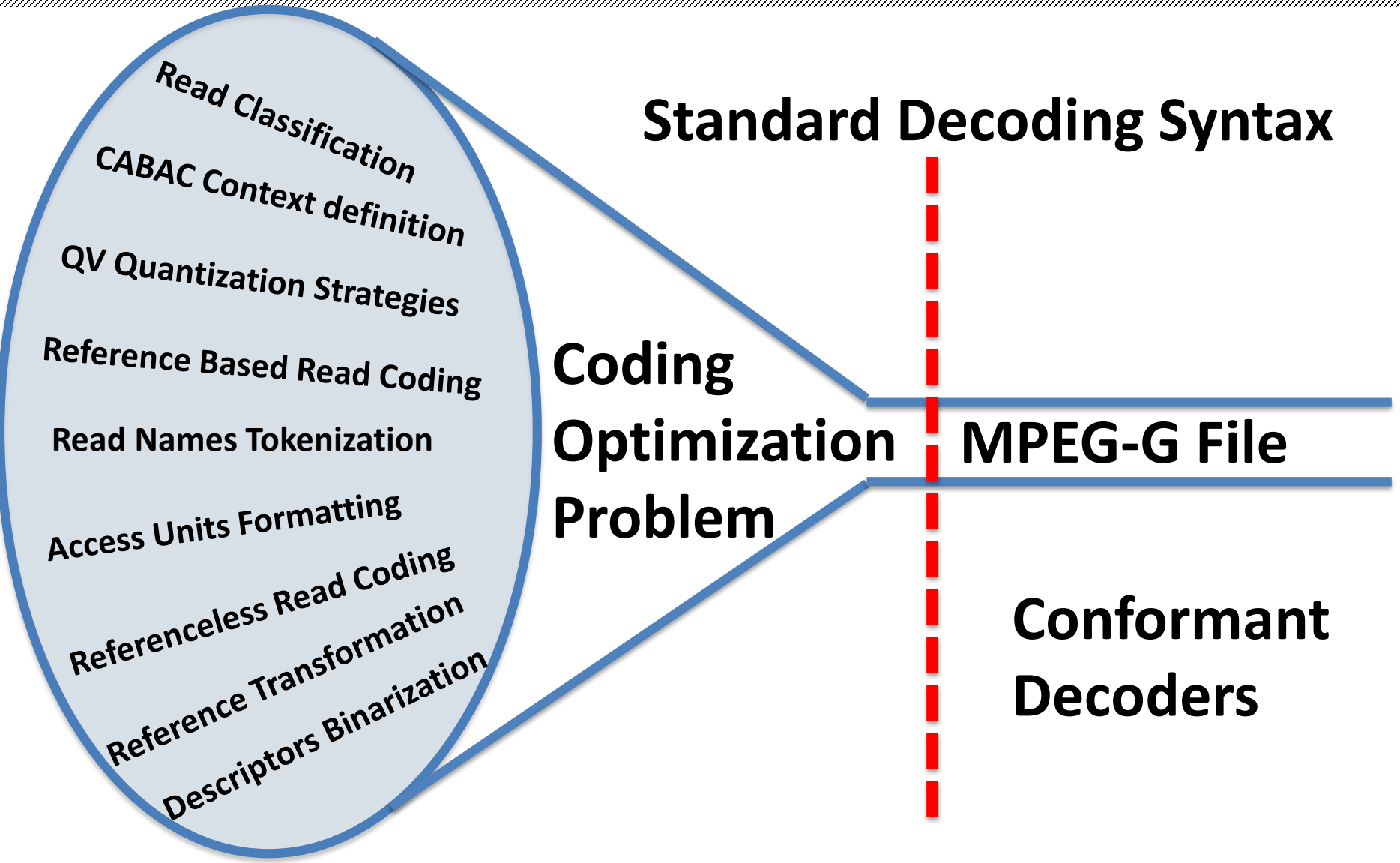
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

MPEG approach to standardization

MPEG-G Compression

- **The MPEG(-G) work methodology:**
  - Include state of the art technology:
    - Open calls for technology
    - Cross-checked «Core experiments»
  - Combine technologies to create a «decoding syntax»
  - Standardize only the decoding process
  - Specify a rigorous conformance testing procedure

The MPEG-G standard concepts



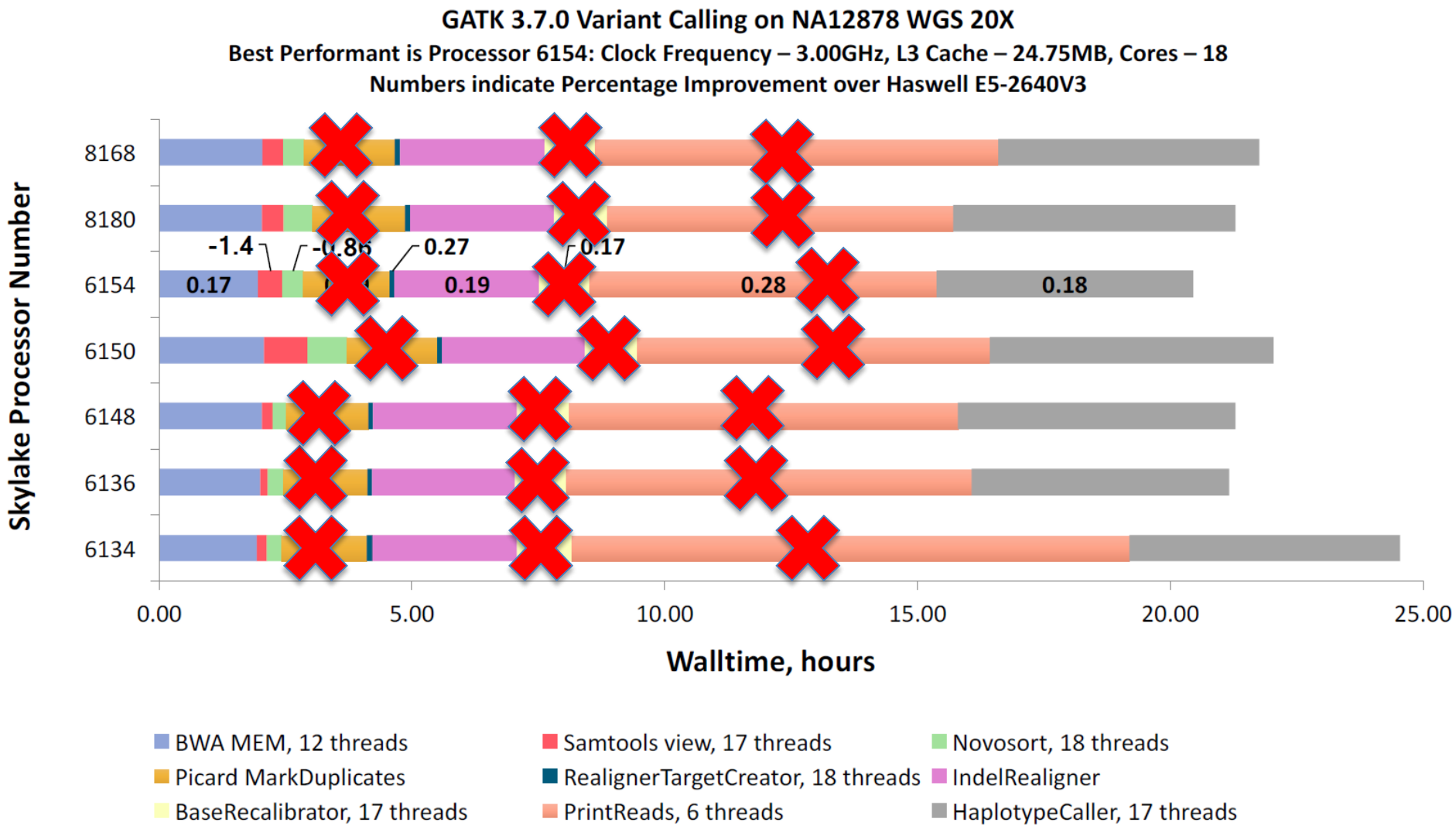
MPEG-G Compression will improve

- The MPEG-G compression results so far:
  - Will improve implementing more advanced encoding strategies
  - Reads and read names in a 20-25% range
  - QVs in a 50-100% range

Selective data access in Pipelines

- It is widely recognized in literature that data access is the main limitation.

Selective data access in Pipelines



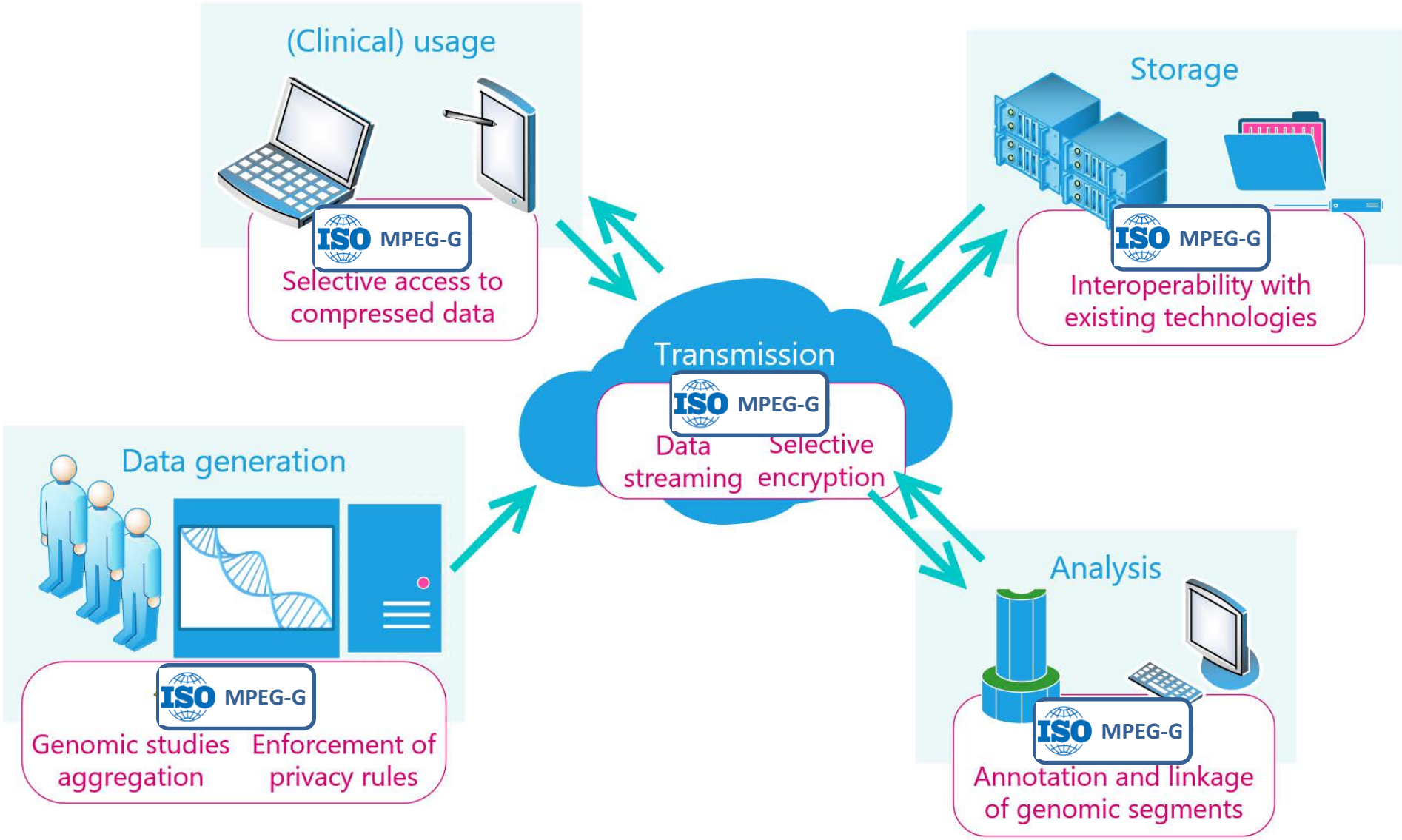
Selective data access in Pipelines

- On-going work in its initial stage
- Based on the «Functional Equivalence Concept»
  - (Allison A. Regier et al. : “Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects”, doi: http://dx.doi.org/10.1101/269316.)
- Speed-up of pipelines in a range 10-100 are achievable

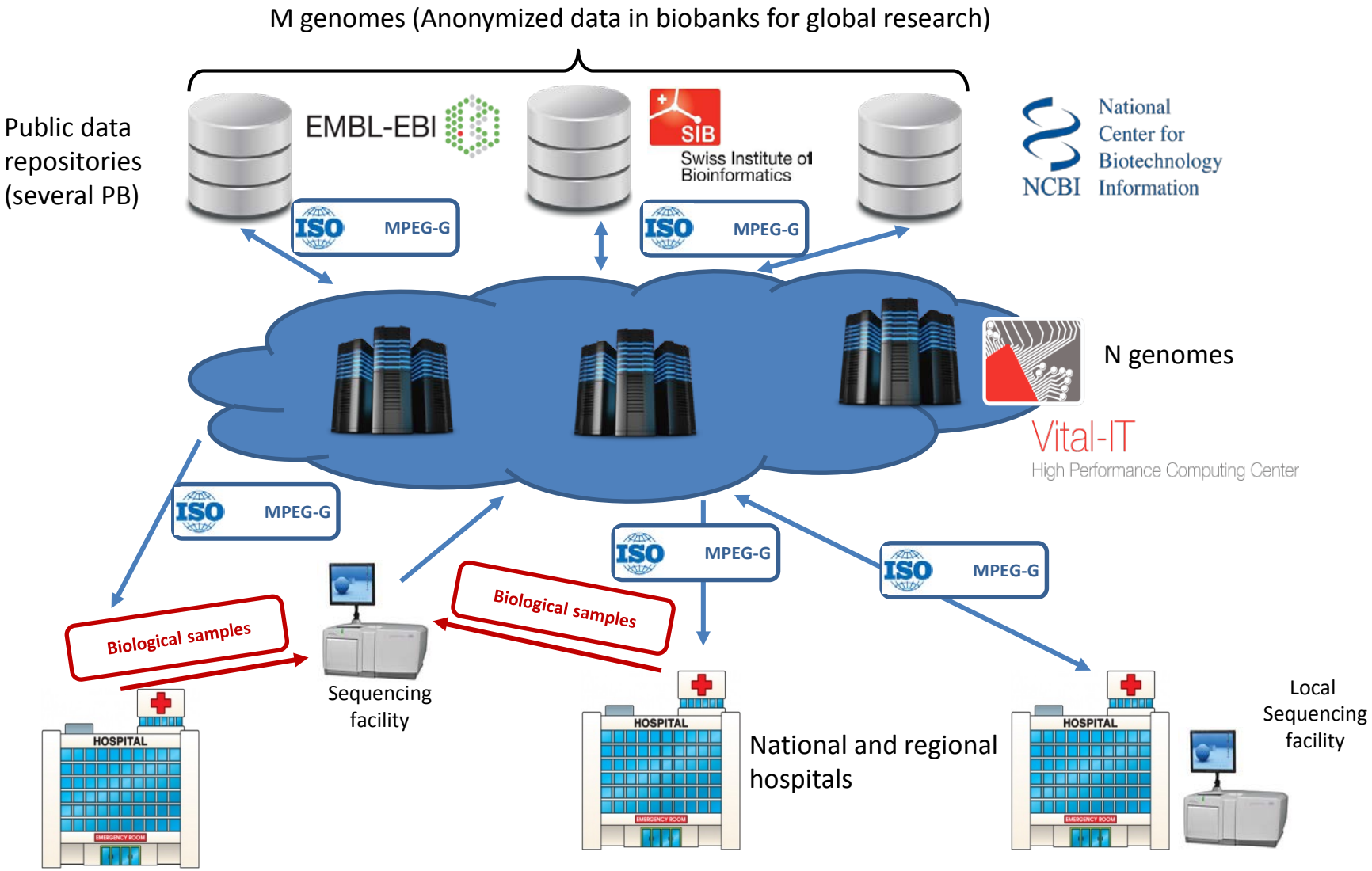
MPEG-G Elementary Use Cases

- Selective access to compressed data
- Data streaming
- Compressed file concatenation
- Genomic studies aggregation
- Selective encryption of sequencing data and metadata:
- Annotation and linkage of genomic segments in the compressed domain
- Interoperability with main existing technologies and legacy formats FASTQ, SAM or BAM are supported by MPEG-G
- Incremental update of sequencing data and metadata

MPEG-G a standard for a complete ecosystem support



MPEG-G for «M to N» Genome Analysis on HPC



The MPEG-G Parts

- Part 1: **File and Transport Format (IS Jan 2019)**
  - The technology to transport and access data
- Part 2: **Compression of genomic data (IS Jan 2019)**
  - The compressed representation
- Part 3: **APIs (IS Apr 2019)**
  - Standard interfaces with genomic data applications and legacy formats
- Part 4: **Reference Software (IS July 2019)**
  - The standard support to the implementation of applications
- Part 5: **Conformance (IS July 2019)**
  - The methodology to test compliance with the standard

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Conclusions

Conclusions

- MPEG-G is based on the experience of more than 25 years of Digital Media
  - Targets:
    - Compression: >100 compared to raw data > 100
    - Compression: >10 – 50 compared to BAM
    - Data access speed: > 100
  - Selective access to data and standard API: Region based; Data class based; User defined
- Expectations for genomic analysis applications
  - Sequencing data compression, transport and APIs will **improve and evolve in time**
  - Main APIs and transport **functionality will remain valid**

Many thanks to ….!

Collaborative and competitive efforts of many companies and individuals!!

- **Barcelona Supercomputing Centre (ES), Centre Nacional de Analisis Genómica (ES), Centre for Genomic Regulation (ES), DAPCOM (ES), EPFL (CH), GenomSys (CH), Hannover University (DE), Heidelberg Institute for Theoretical Studies (DE), IMEC (BE), Made of Genes (ES), Pirbright Institute (UK), Swiss Institute for Bioinformatics (CH), Silesian University of Technology (PL), Simon Fraser University (CA), Massachussets Institute of Technology (US), Stanford University (US), Univ. Politecnica de Catalunya (ES), Wellcome Trust Sanger Institute (UK), GenomSoft (CH), Istituto Europeo di Oncologia (IT), CEDEO (IT), AGINOME Scientific (CN) .....**

- **Martin Golebiewski, Yong Zhang, Jan Voges, Ioannis Xenarios, Tom Paridaens, Claudio Alberti, Filippo Medri, Joern Ostermann, Leonardo Chiariglione, Daniel Naro, Jaime Delgado, Giorgio Zoia, Daniele Renzi, Mikel Hernaez, Junaid Ahmad, Paolo Ribeca, Ibrahim Numancig, James Bonfield, Nicolas Guex, Christian Iseli, Thierry Schuepbach, Silvia Llorente, Josep Lluís Gelpí, Dmitry Repchevsky, Romina Royo, Leonor Frías, Oscar Flores, Glenn Van Wallendael, Wesley De Neve, Peter Lambert, Lukasz Roguski, Jordi Portell, Idoia Ochoa, Reggy Long, Noah Daniels, Cenk Sahinalp, Massimo Ravasi, Wenxiang Yang, Rongshan Yu, ... and many others**

A new logo will be needed soon?

1988

2017