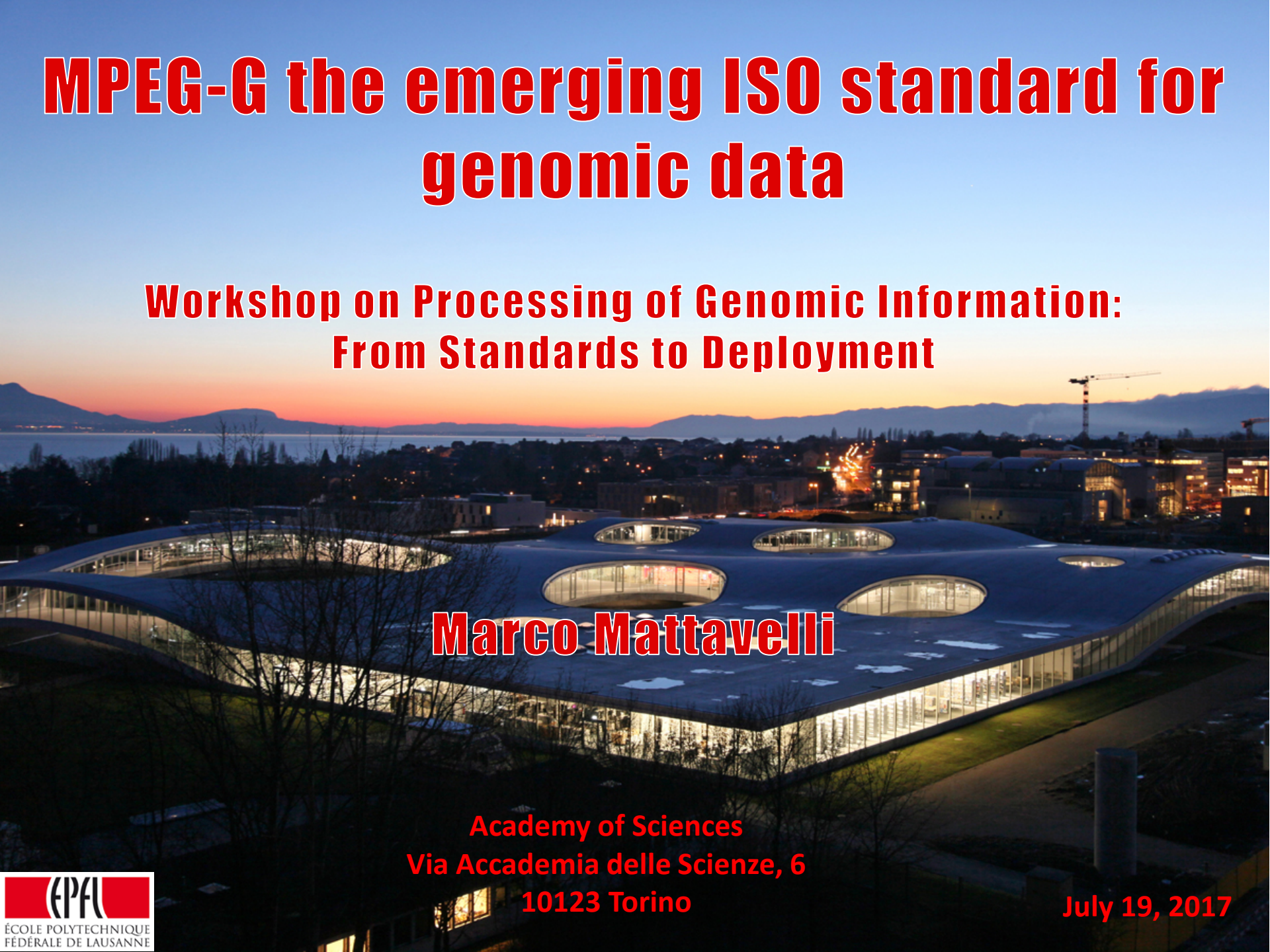
# MPEG-G the emerging ISO standard for genomic data

## Workshop on Processing of Genomic Information: From Standards to Deployment



**Marco Mattavelli**

Academy of Sciences
Via Accademia delle Scienze, 6
10123 Torino

July 19, 2017

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

# 25 Years of MPEG Digital Media

# 25 Years of MPEG Digital Media

# 25 Years of MPEG Digital Media

**Just a few clicks ……**

# 25 Years of MPEG Digital Media

- We may think that «compressing audio video content» is **easy**!
  - MPEG HEVC textual specification: ~ 800 pages
  - MPEG HEVC reference software: ~ 100’000 lines
  - MPEG-SYSTEMS: ~ 2000 pages

# 25 Years of MPEG Digital Media

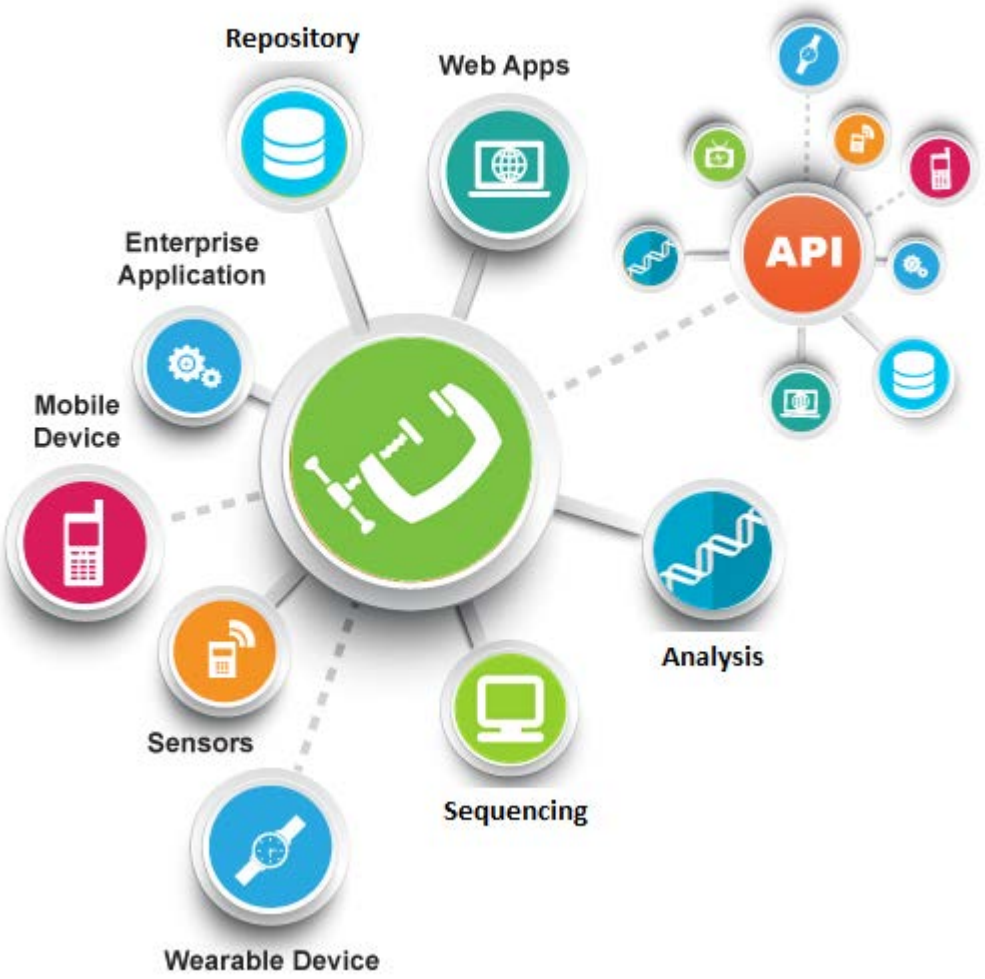
- **The compression technology does not change**.

- **Wrong!!**

  – MPEG-1, MPEG-2, AVC, HEVC, Future Video Coding, …

  – **ONLY** a factor 2 of increase in compression performance was sufficient to switch to a new generation of MPEG technology

# 25 Years of MPEG Digital Media




What does make possible that everybody can easily deal with compressed audio/video content?

**1) MPEG «Systems» and standard APIs + standard compression**

**2) An ecosystem that builds technology elements (SW and HW) so that compression technology becomes a commodity for the users**

## Lesson from these 25 years:

- Compression is important, technology enabler, but it is not all: **MPEG «Systems»** is even more important.
- Digital media applications are built «around» the MPEG «**Systems**» standard:
  - All component are «synchronized» and linked
  - Access to data in the compressed domain
- If a compression (standard) technology changes the «**Systems**» standard «stays»!!
- Well, it adds new functionality and extended supports to applications ……

And if we try to guess the future of genomic data:

- Genomic sequencing data compression technology will **change in time**
- Genomic sequencing data compression performance will **improve in time**
- MPEG-G «Systems» and APIs will evolve and improve, but the main **functionality will stay** and support the evolution of analysis applications
- **Industrial support** to make genome data processing a «commodity» for professional and users is absolutely needed!!

# Objectives of MPEG-G

- **Interoperable selective access to data in the compressed domain** by means of standard APIs:
  - Genomic region
  - Class of data (matching accuracy, user defined)
  - Sub-sets of genomic information

## MPEG-G «Systems»

**On top of compression, higher performance is provided by a specific file format and transport format**

# Objectives of MPEG-G

- MPEG-G Systems
  - Support for **incremental update and annotations**:
    - Directly in the compressed domain
    - Same file format with additional incremental data linked to the existing compressed data
  - User defined data **classification and partitioning**:
    - Sequence read data classes with different accuracy versus references
    - Reduced data access for analysis

# Structure of the MPEG-G standard

- Part 1: **File and Transport Format**
  - The technology to transport and access data
- Part 2: **Compression of genomic data**
  - The compressed representation
- Part 3: **APIs**
  - The standard interfaces with genomic data applications and legacy formats
- Part 4: **Conformance**
  - The methodology to test compliance with the standard
- Part 5: **Reference SW**
  - The standard support to the implementation of applications

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

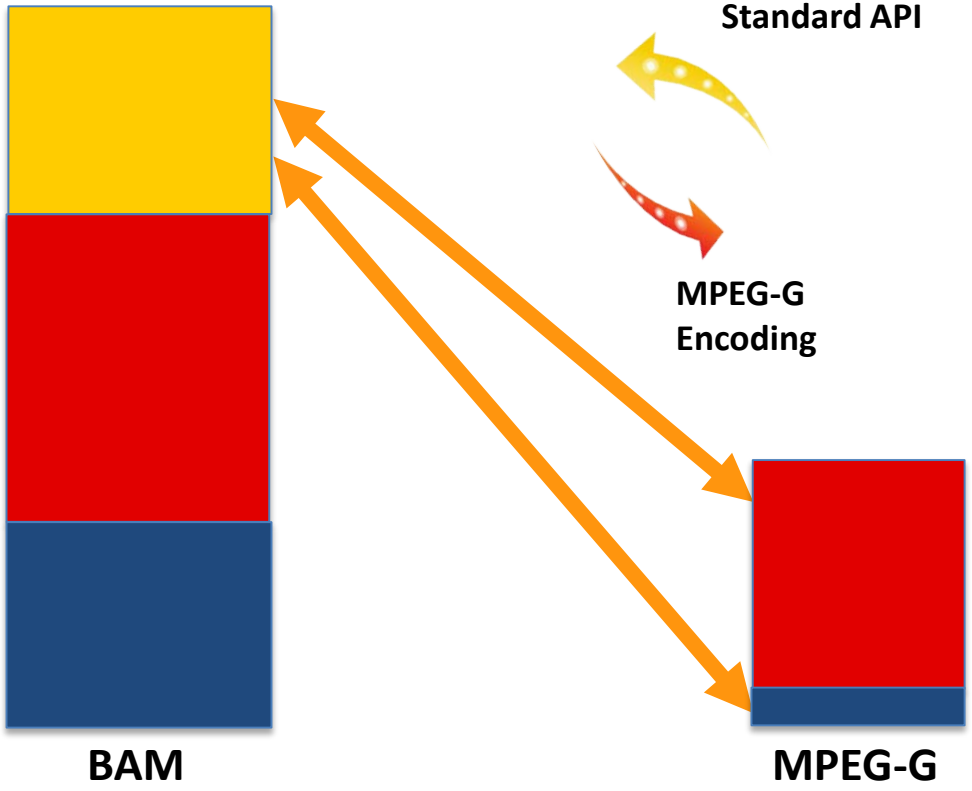
# MPEG-G Performance (so far ….)

# From currently used formats (BAM) to MPEG-G

**BAM Auxiliary fields**

**Quality Values (QV)**

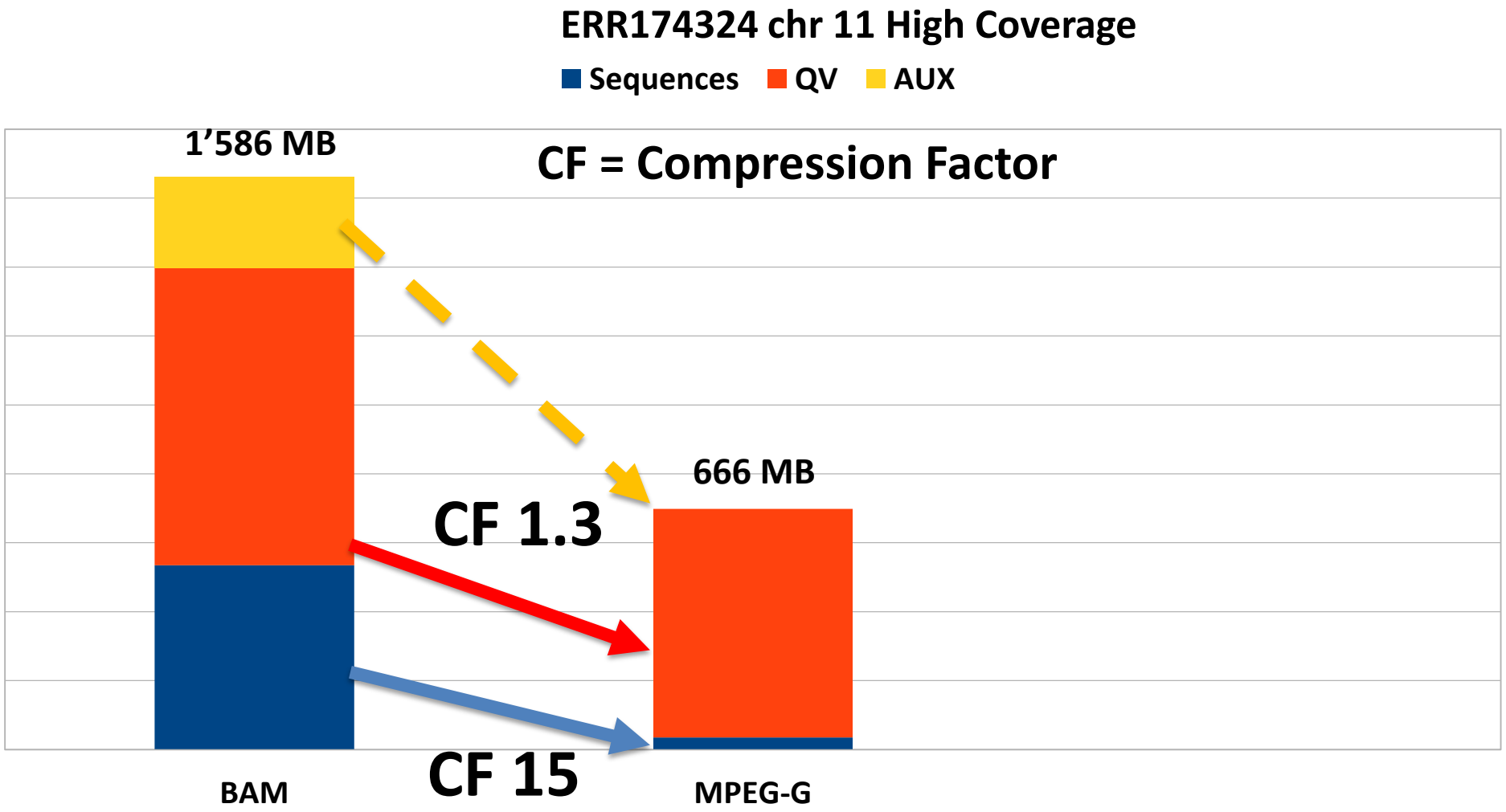
**Sequence Reads**



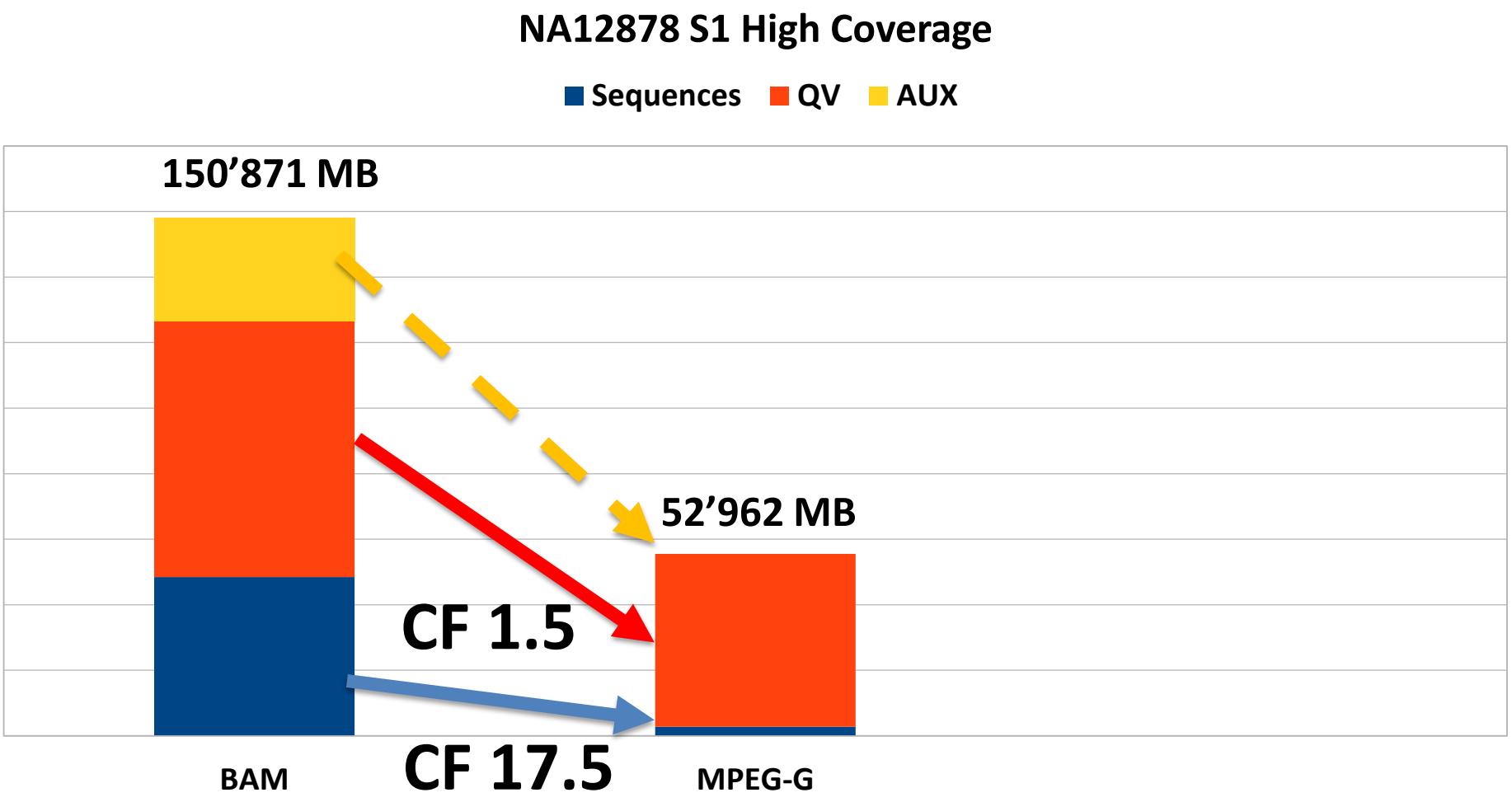

Most of BAM Auxiliary Fields are compressed and are included in the MPEG-G file format representation, user defined and ambiguous fields are not included and not part of the standard APIs.

The BAM auxiliary field compressed in the MPEG-G representation are available and can be retrieved by means of standard API.

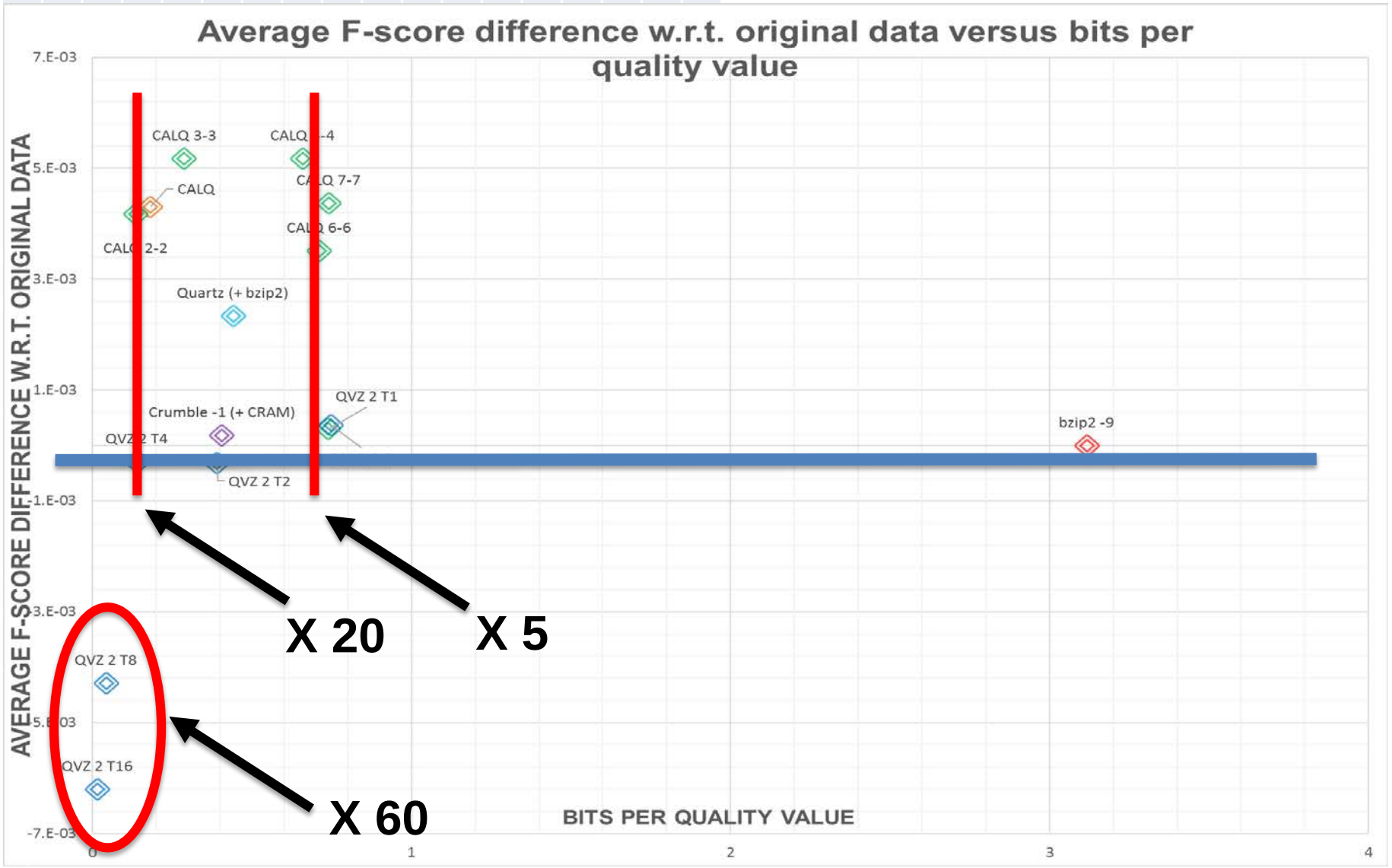
# One Chromosome High Coverage (Human – Illumina)



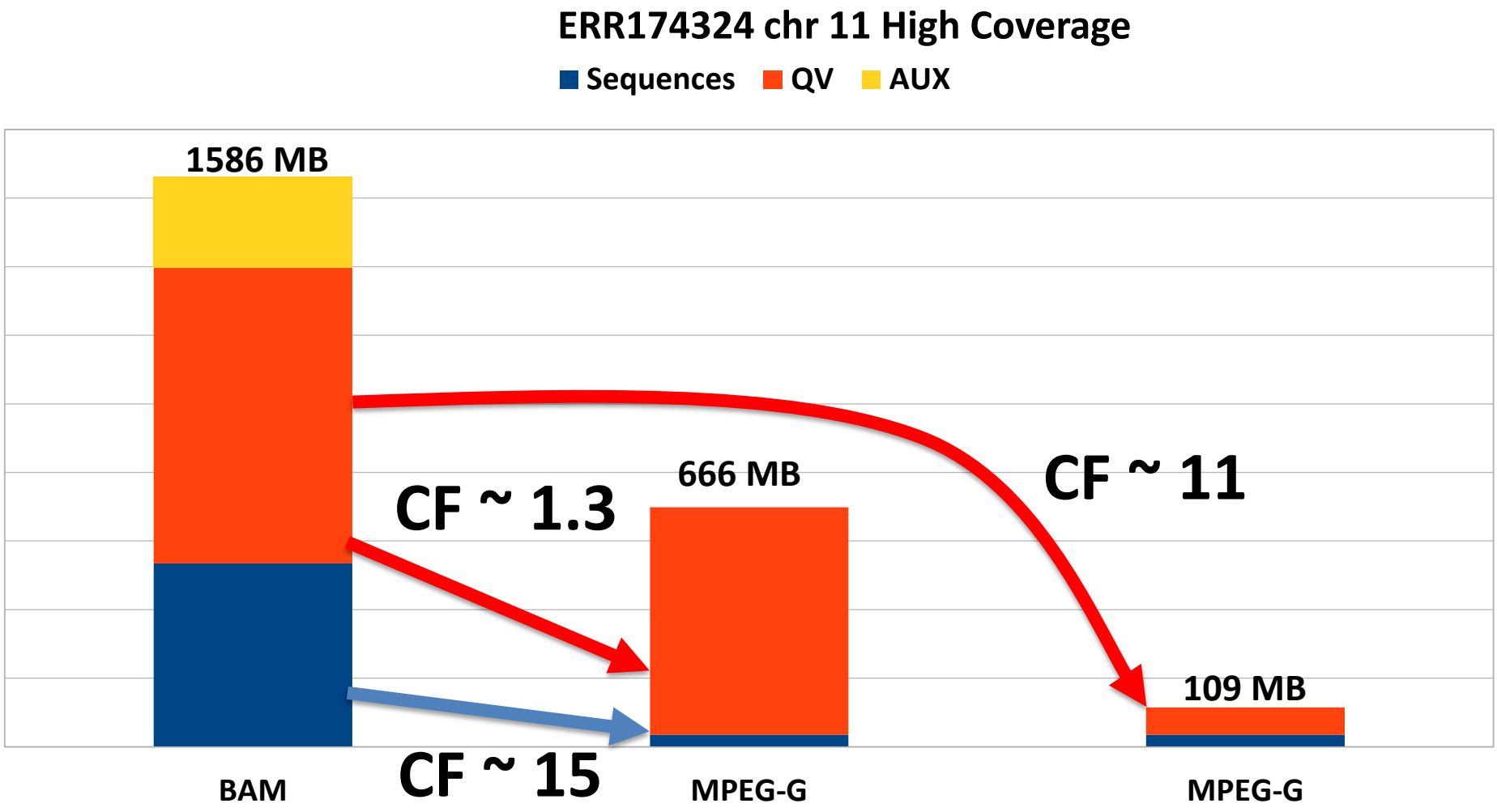
# Whole Genome High Coverage (Human – Illumina)



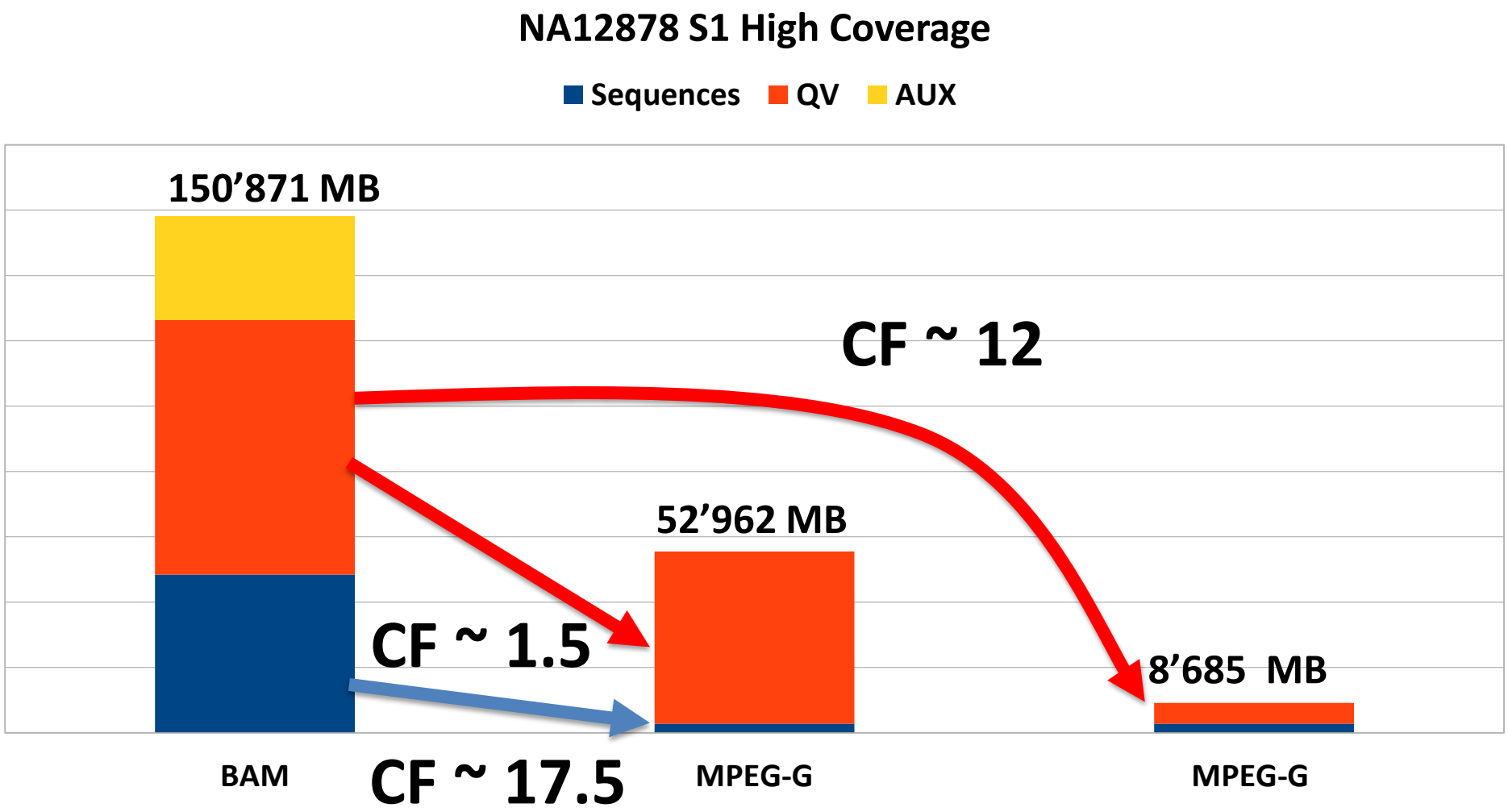
# Rate-Distortion for Quality Values



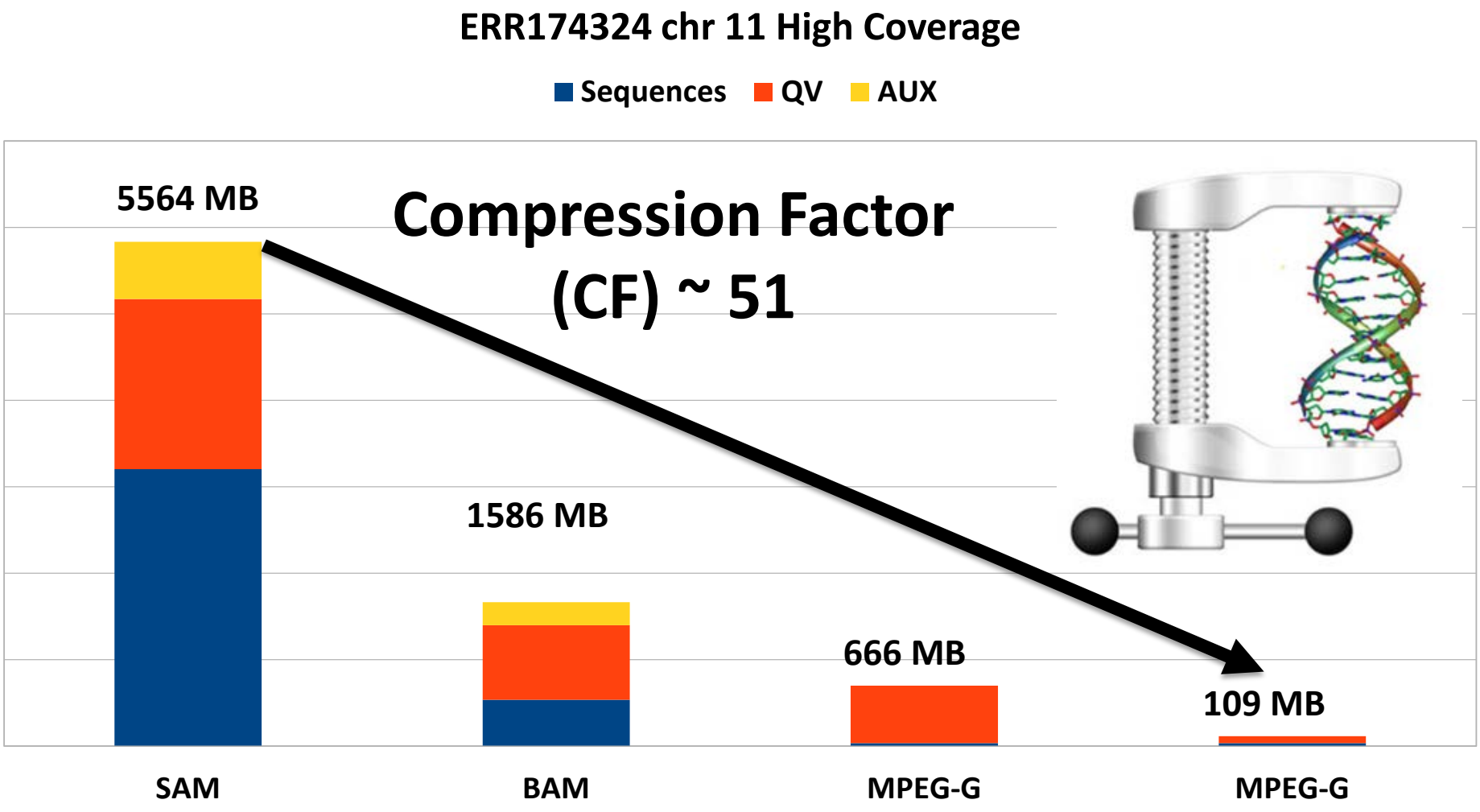
# One Chromosome High Coverage (Human – Illumina)



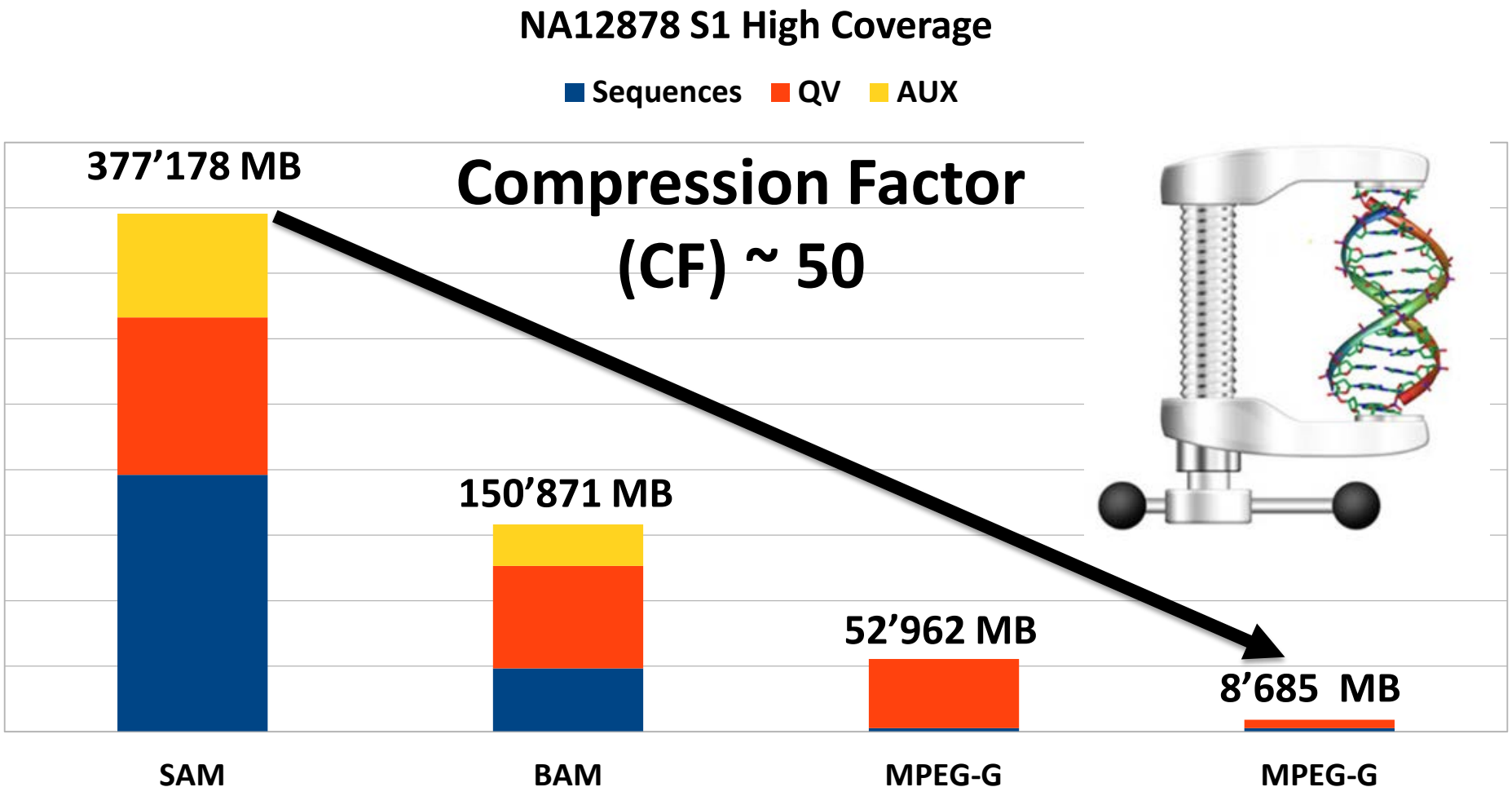
# Full Genome High Coverage (Human – Illumina)



# One chromosome high coverage (Human-Illumina)



# Whole Genome High Coverage (Human-Illumina)



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

# Conclusions

# Conclusions

- **ISO MPEG-G: based on 25 years of Digital media**
  - **Compression target:**
    - **VS raw data > 100**
    - **VS BAM > 10 – 50**
  - **Selective access to data and standard API:**
    - **Region based**
    - **Data class based**
    - **User defined**
  - **Data access speed target: > 100**
- **MPEG-G the «enabler» of Genomics 2.0 interoperable applications**

# 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), Istituto Europeo di Oncologia (IT), CEDEO (IT),**

- **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, ..... and many others**

# A new logo will be needed soon?


1988


2017