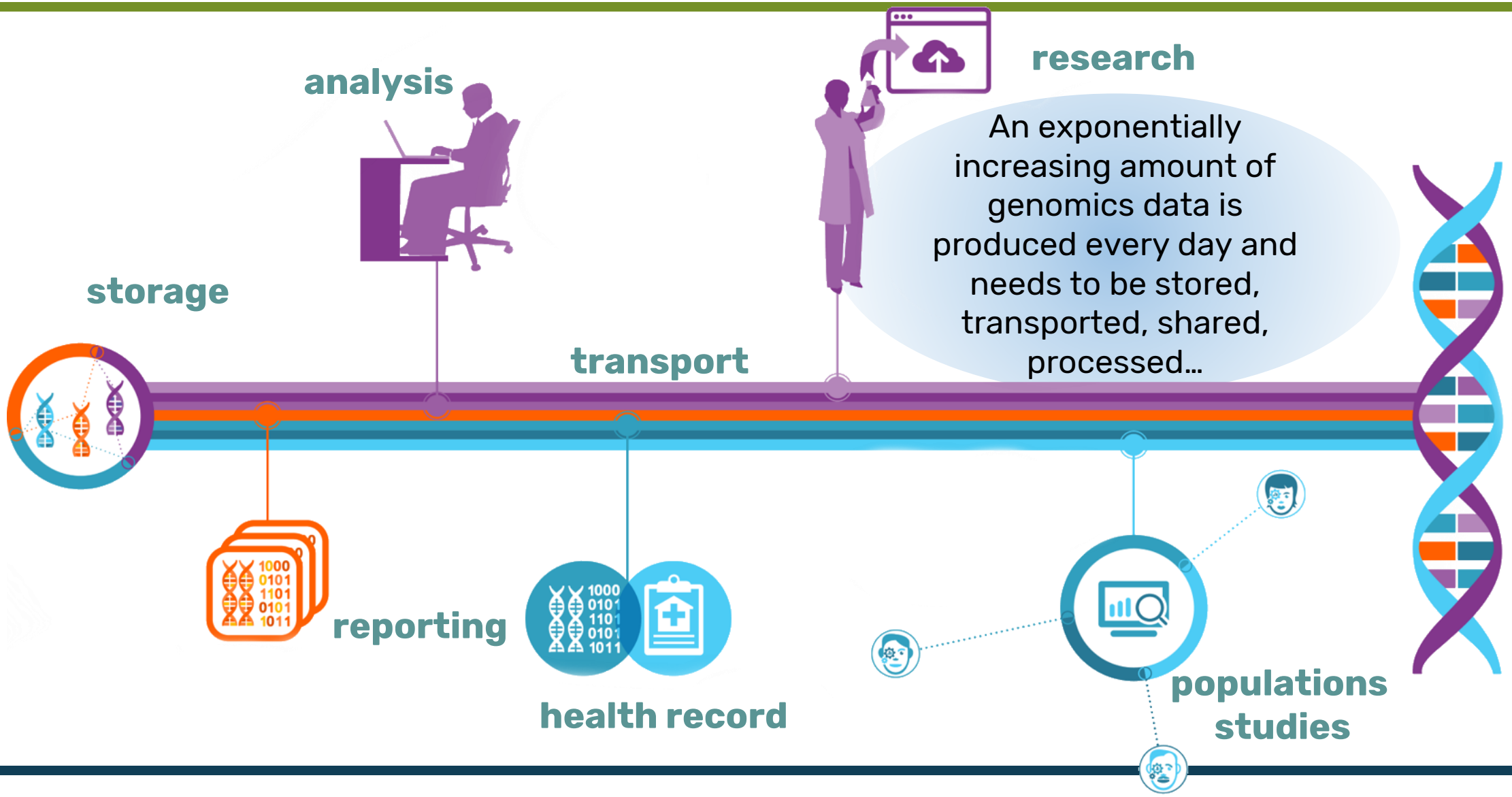
GenomSys

Streaming genomes with MPEG-G the emerging ISO standard for genomic information representation

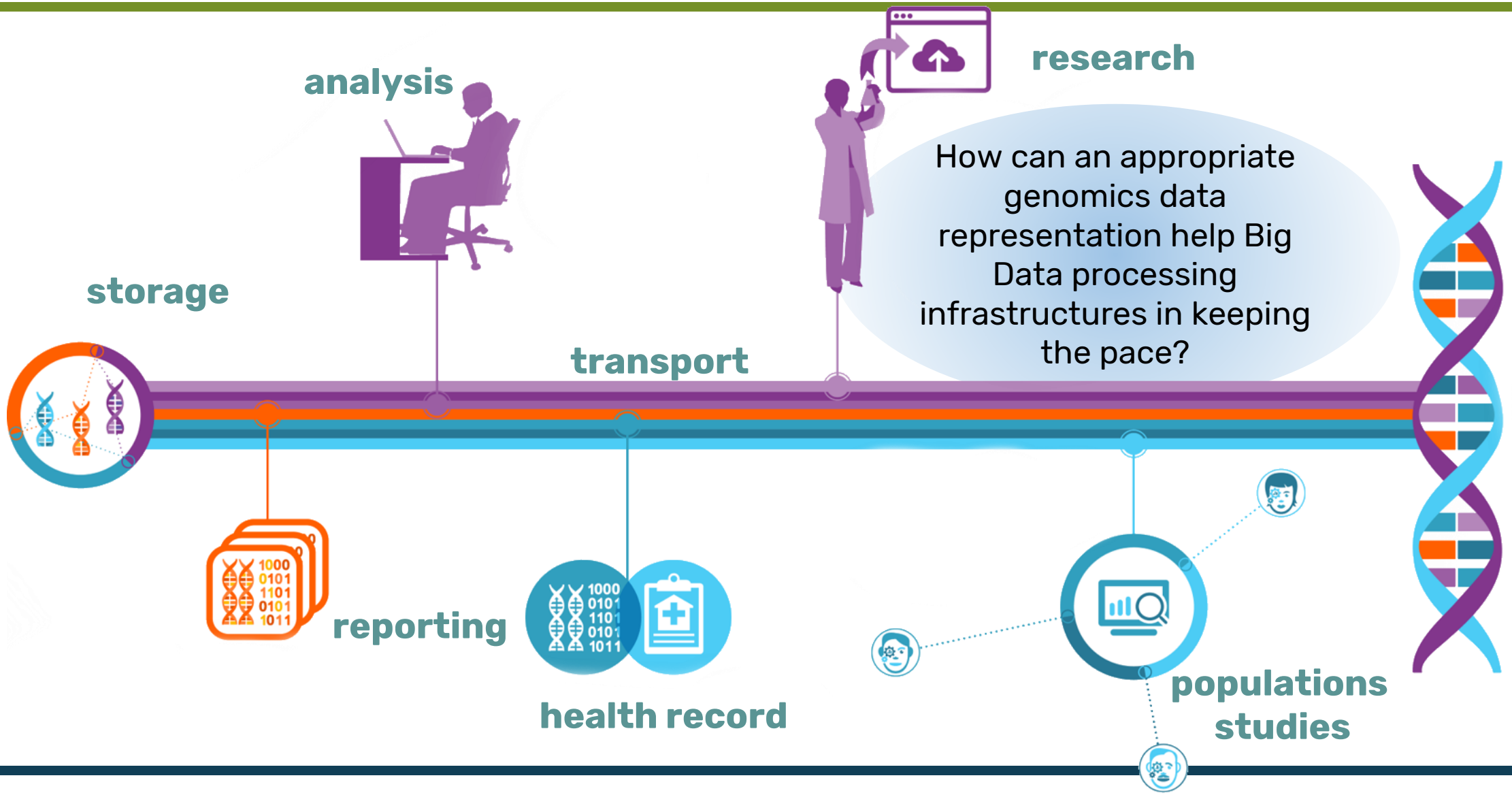
Claudio Alberti, GenomSys

Management and Security of HPC Data
Lausanne 17/05/2018

Genomics data deluge



Genomics data deluge



Current status of genomics data formats

- Under-specified APIs
  - No interoperability
  - No standard framework to enforce privacy protection
  - Heterogeneous metadata formats (annotations)
- Undocumented process for extensions and amendments
- Poor compression (zipped text)
- Lack of consistent specification
- No conformance tests (how do I know if my implementation is conformant?)
- No transport format (streaming)

Can we…

- …reduce the ICT costs of genomics?
- …enable large scale data sharing?
- …enable systems interoperability?
- …provide long term support to the specification?
- …provide support for genomics data protection?
  - Controlled access
  - Traceability
  - Integrity verification
  - Roles segregation
  - Privacy rules enforcement

How the past can help the future

The current scenario in genome sequencing is similar to that of digital multimedia 30 years ago

a growing amount of data…

…to be distributed and consumed worldwide by different devices…

…each using proprietary and non-compatible formats and devices

The digital media revolution

ISO-MPEG has demonstrated its ability to:

- generate high quality, widely accepted industry standards
- specify and maintain standards under non-discriminatory terms (ISO/IEC licensing)
- produce higher quality results than any proprietary solution

MPEG-G

- Joint project between two ISO committees
  - TC276 WG5 (Data Processing)
  - JTC1 SC29 WG11 (MPEG)

ISO/TC 276 Biotechnology

ISO

MPEG

- Multidisciplinary international experts group
  - Bioinformatics
  - Computer science
  - Information theory

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Stanford University

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Massachusetts Institute of Technology

Aginome SCIENTIFIC

SIB Vital-IT High Performance Computing Center

GenomSys

ILLINOIS

EXO.net Making the right move in digital media

SFU SIMON FRASER UNIVERSITY ENGAGING THE WORLD

- Academia, Research & Industry

UNIVERSITEIT GENT

Leibniz Universität Hannover

wellcome trust sanger institute

- Feedback collected from major genomic data repositories

EMBL-EBI

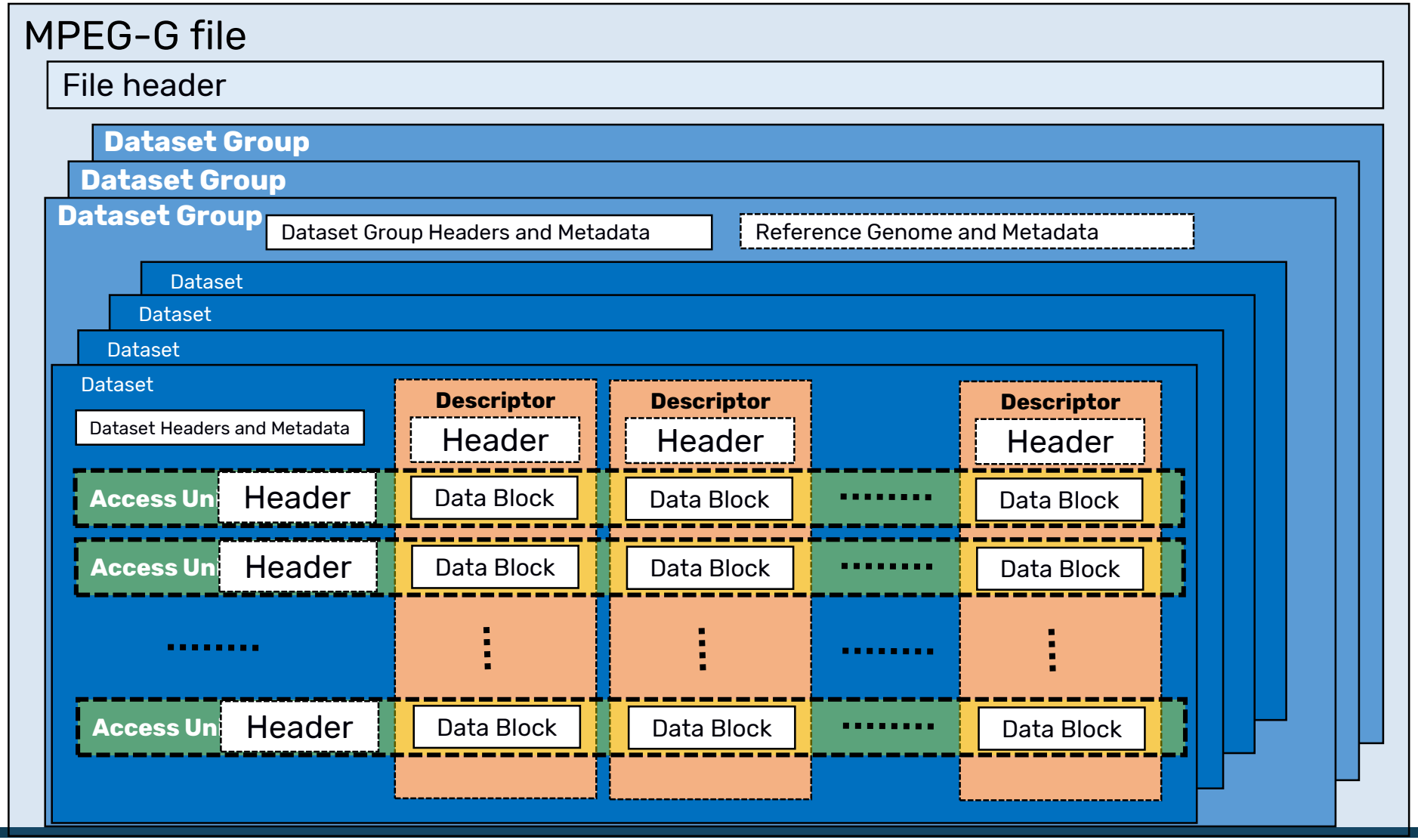
National Center for Biotechnology Information NCBI

华大基因 BGI

The MPEG-G approach

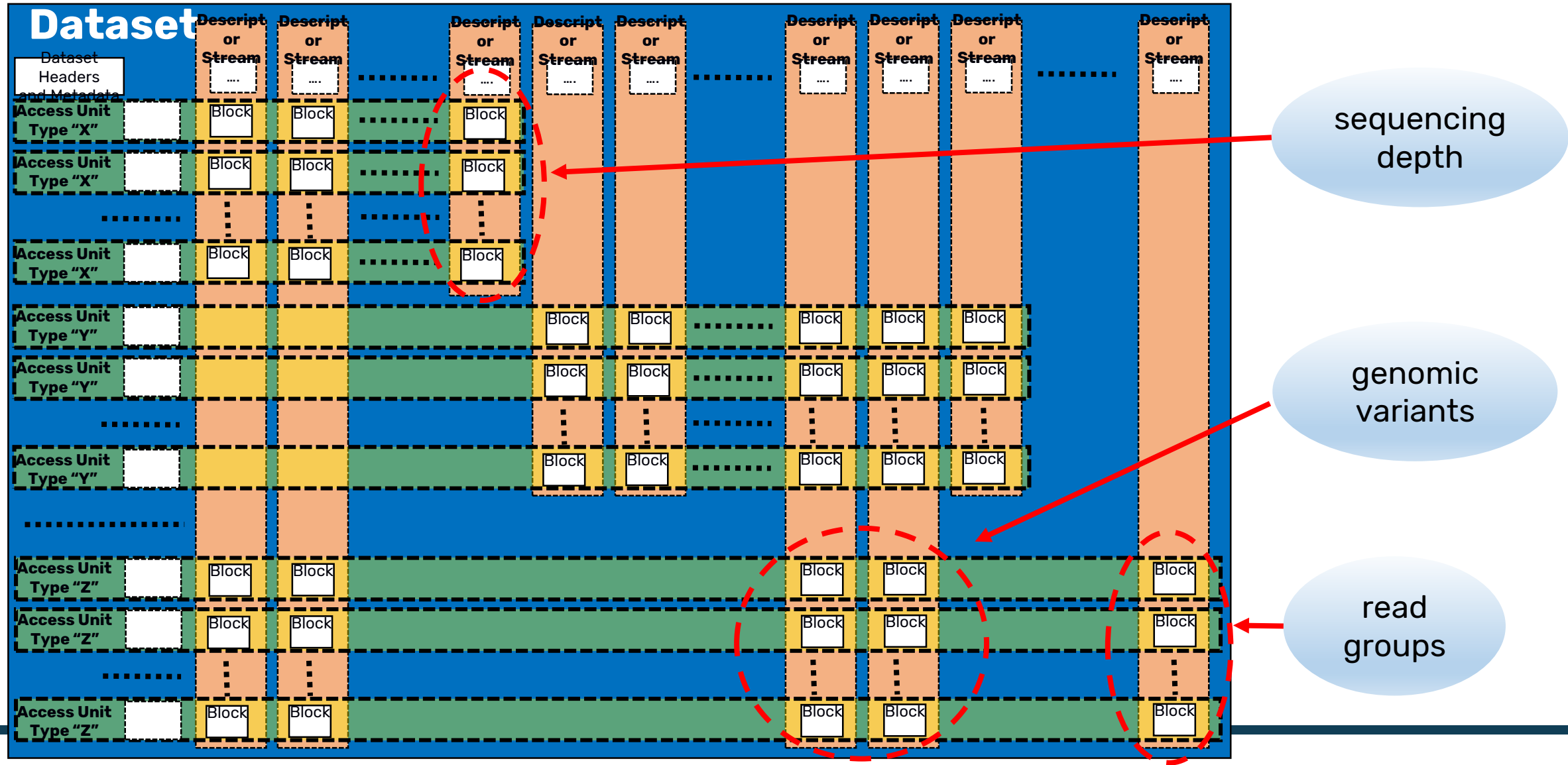
- Genomic data (i.e. sequence reads) are classified into homogeneous sets (classes of data) and represented by minimal sub-sets of **descriptors**
- Meta-data associated to classified reads is represented by specific **descriptors**
- Descriptors are compressed individually and then stored into structured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

MPEG-G file format

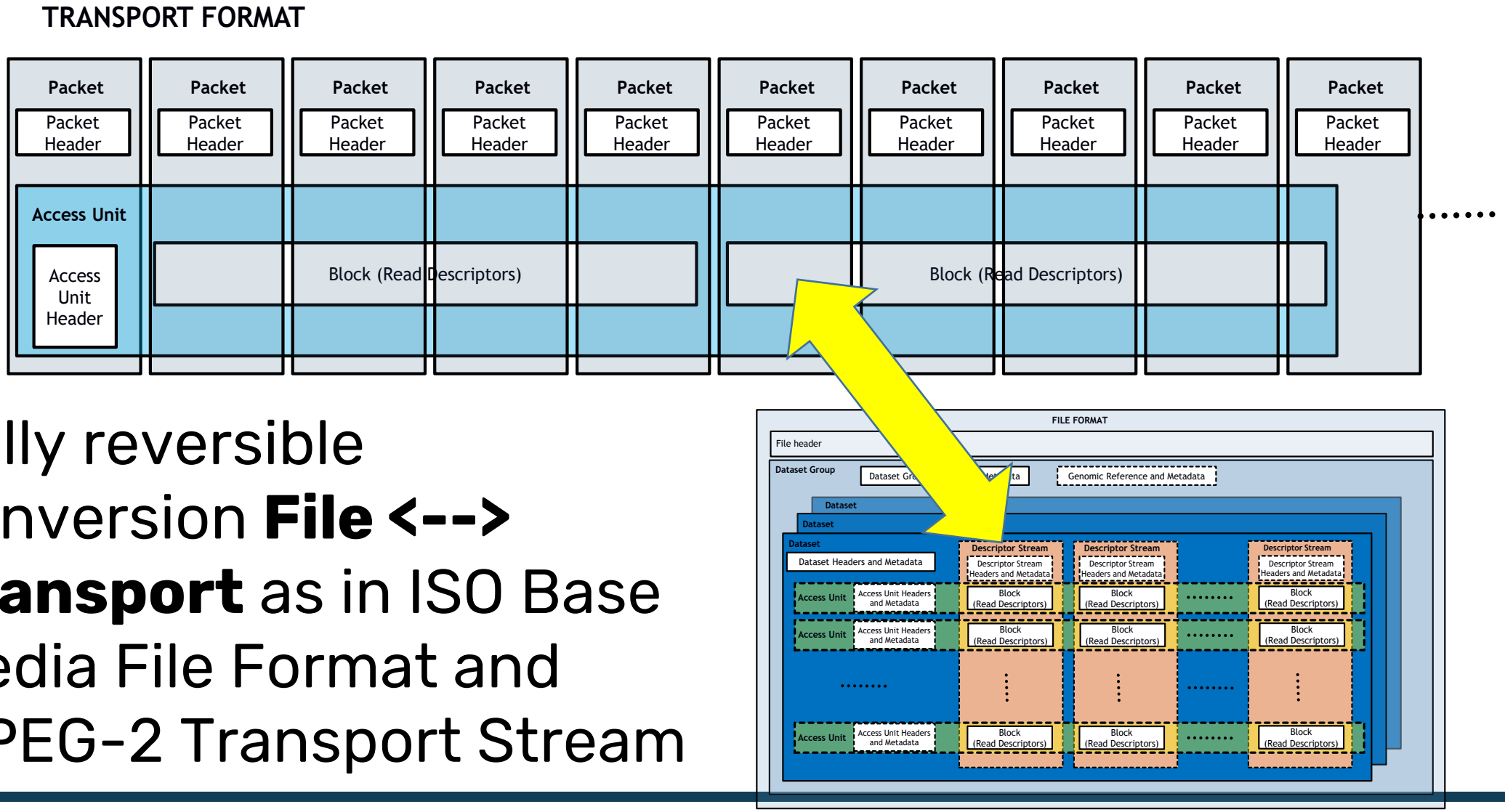


- File
  A trio (Dad, Mum, Child)
- Dataset group
  One individual's history
- Dataset
  One sequencing run

MPEG-G file format



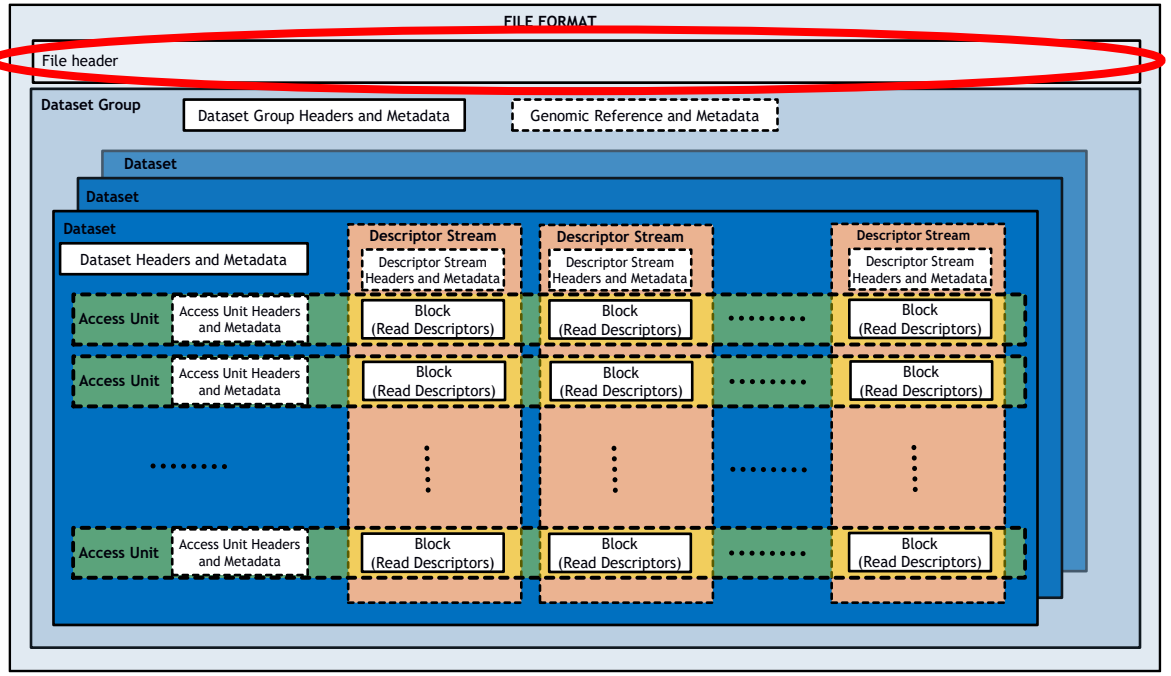
MPEG-G Transport Format



Fully reversible conversion **File <-->** **Transport** as in ISO Base Media File Format and MPEG-2 Transport Stream

MPEG-G data indexing

- Indexing is natively embedded with the data
- Genomic regions aggregation
- Unmapped content can be indexed for fast search
- Multi-dimensional indexes
  - mapping result (presence/absence of variants)
  - number of variants
  - multiple mappings



MPEG-G metadata - annotations

Dataset group

Element name	Element type	Mandatory
Title	String	Yes
Type	Controlled vocabulary	Yes
Abstract	String	No
Project centre name	ProjectCentre type	No
Description	String	No
Samples	ListOfSamples type	Yes
Extensions	ListOfExtensions type	No

Dataset

Element name	Element type	Description	Mandatory
Title	String		No
Type	Controlled vocabulary		No
Abstract	String		No
Project centres	ListOfProjectCentres type	Contact information of centres participating in the generation of the described study's data.	No
Description	String		No
Samples	ListOfSamples type	Identification of the samples, based on taxonomy/scientific name, common name or anonymized name and further attributes defined in a controlled library.	No
Extensions	ListOfExtensions type		No

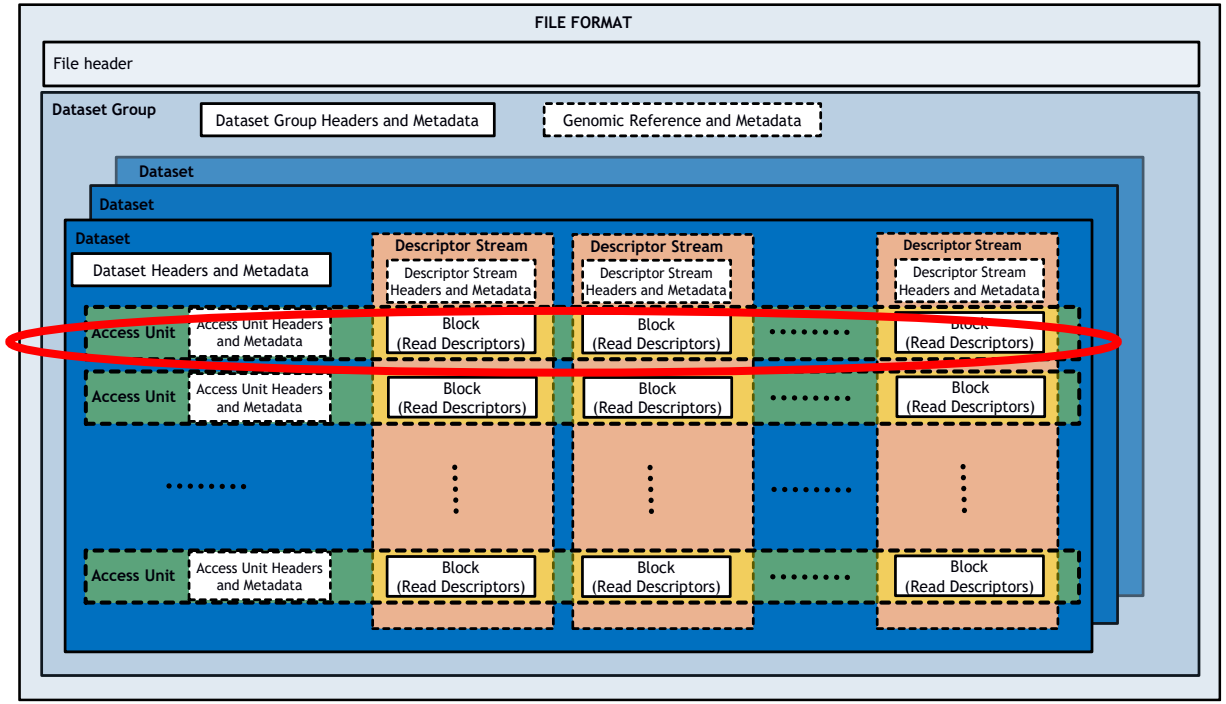
MPEG-G metadata - annotations

- A normative mechanism for extensions is specified
- Profiles are defined to support well-known sets
  - e.g. the SRA (Sequence Read Archive) schema for one Run
  - a profile is identified by a URI urn:mpeg:mpeg-g:metadata:profile:ega:run
- Not bound to one specific ontology

MPEG-G metadata - protection

- Any indexed object in a MPEG-G file can be protected

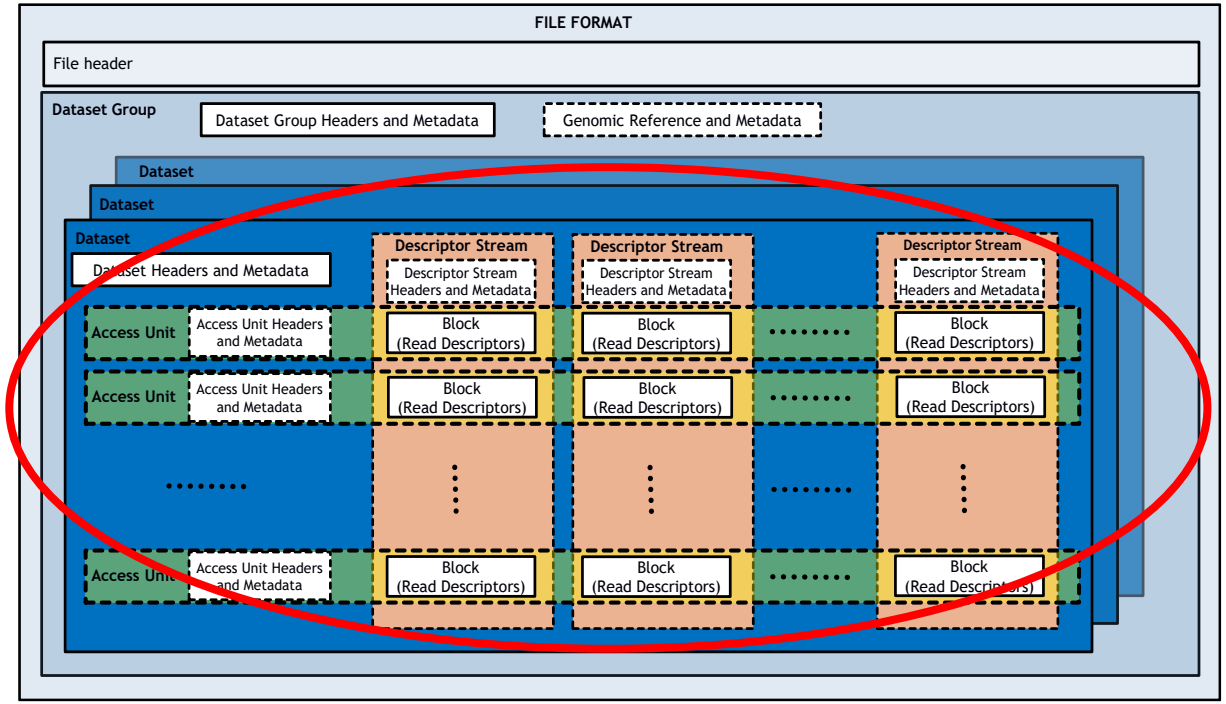
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



MPEG-G metadata - protection

- Any indexed object in a MPEG-G file can be protected

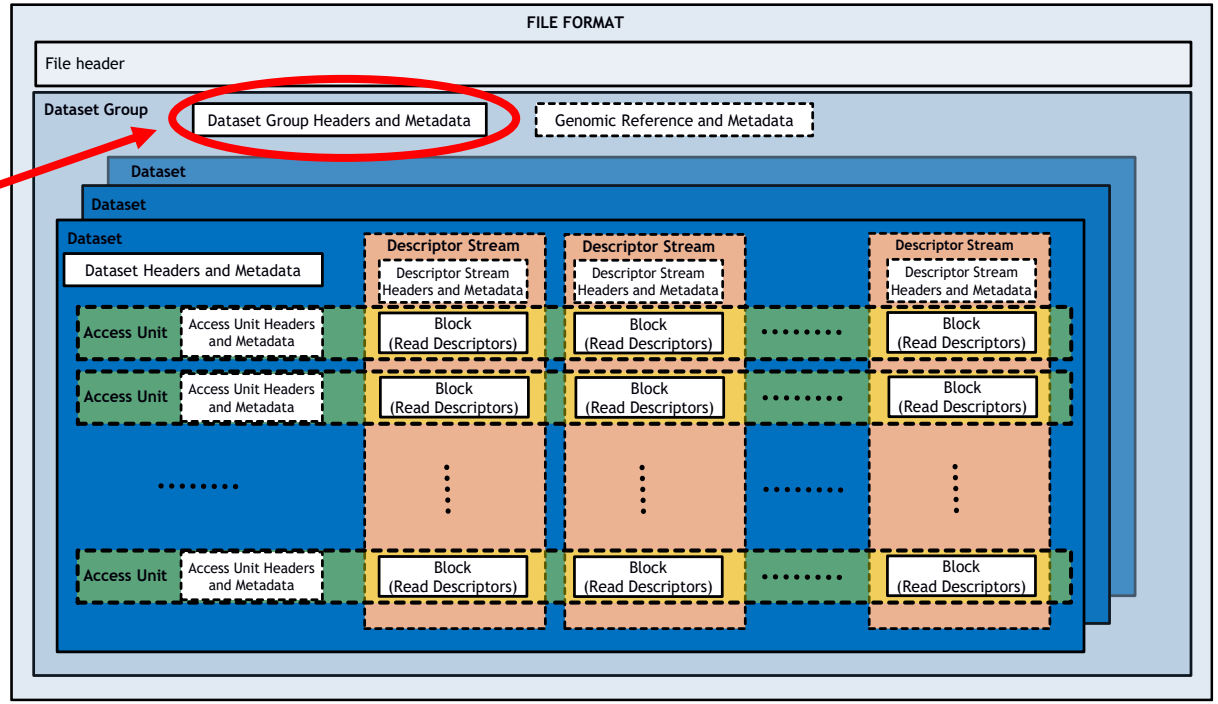
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



MPEG-G metadata - protection

- Any indexed object in a MPEG-G file can be protected

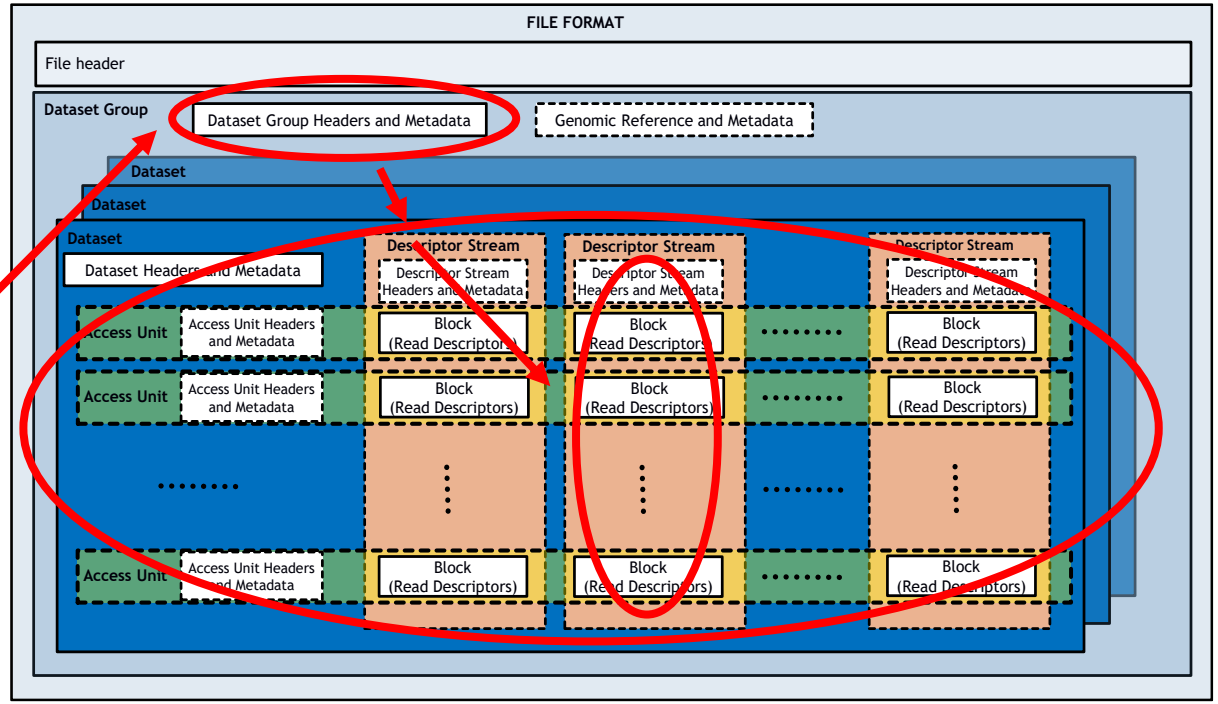
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



MPEG-G metadata - protection

- Any indexed object in a MPEG-G file can be protected

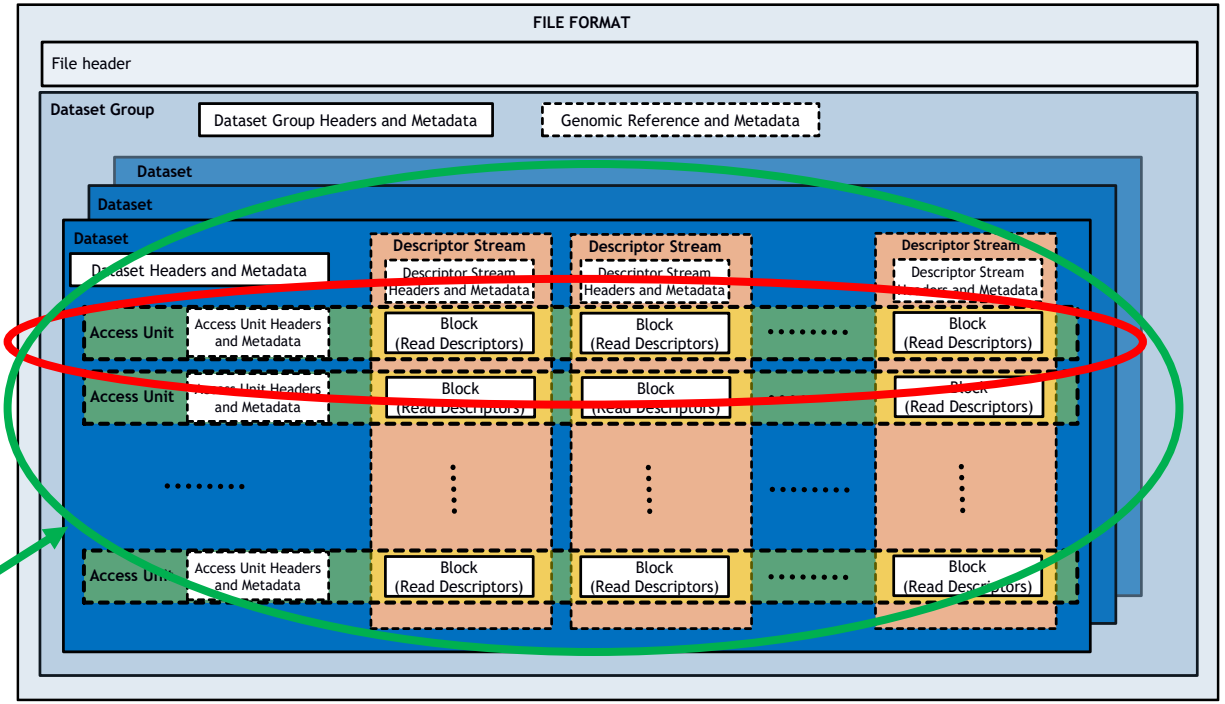
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



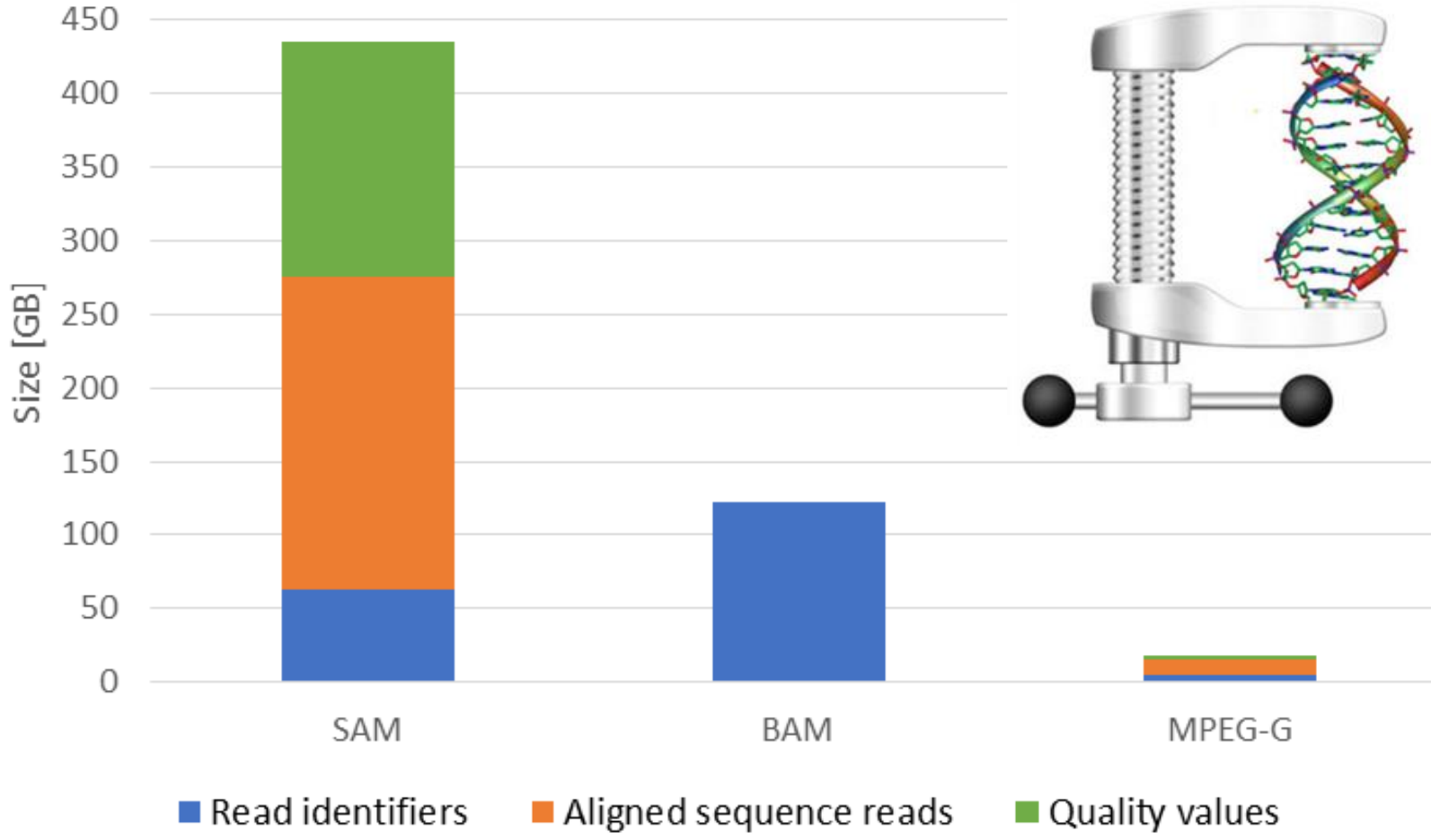
MPEG-G metadata - protection

- Any indexed object in a MPEG-G file can be protected

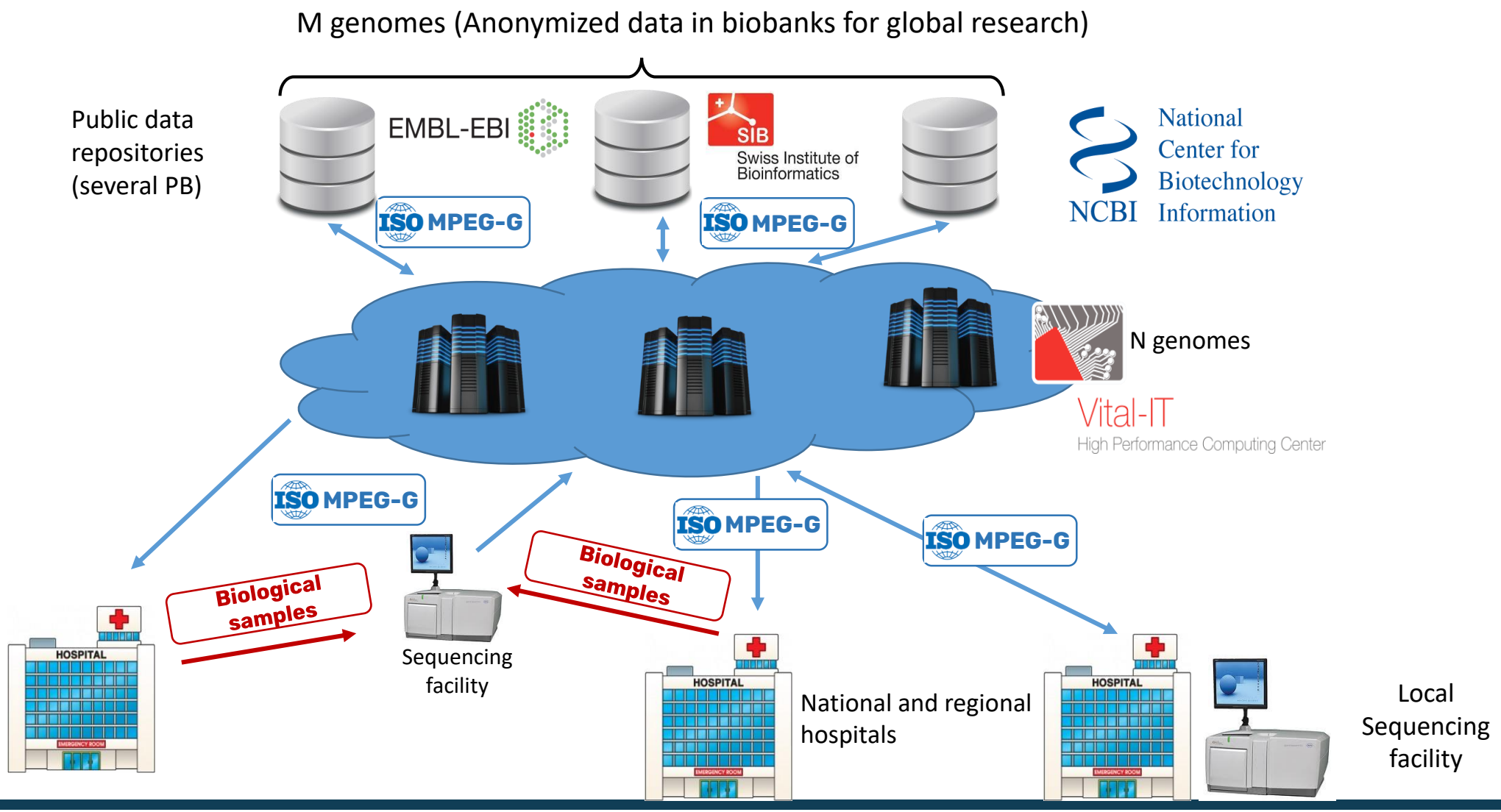
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



MPEG-G compression

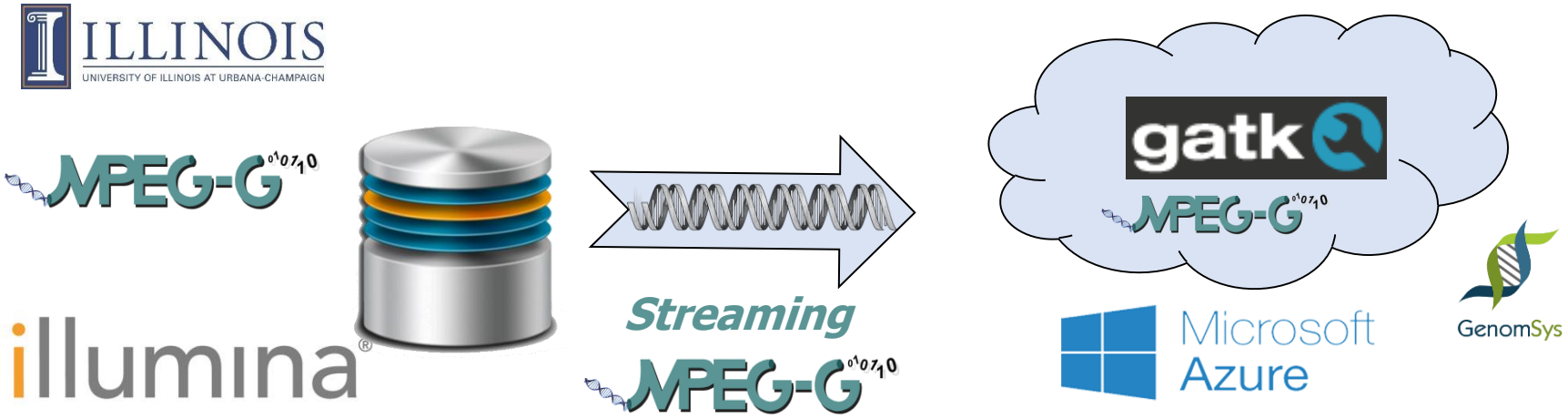


Streaming genomes among HPC centers



First public demonstration: MPEG-G powered GATK

On April 18th at the 122nd MPEG meeting in San Diego first prototype of MPEG-G enabled analysis with GATK



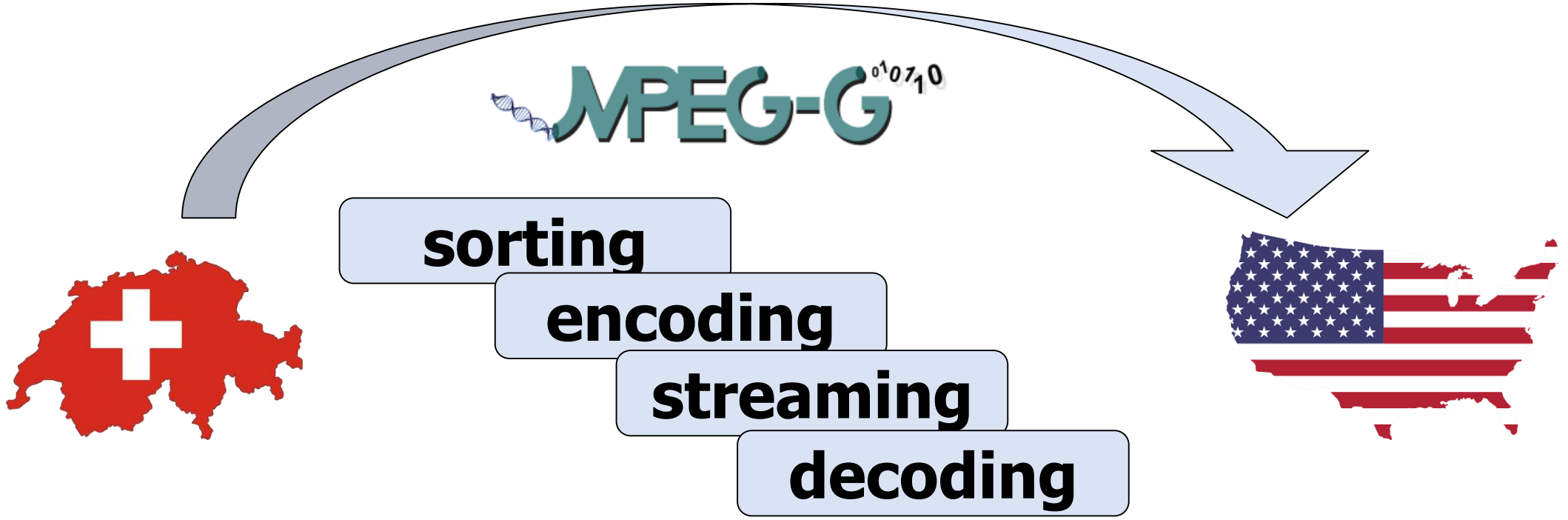
Platinum Genomes

- Data are stored in servers @ UIUC
- GATK runs on Azure
- MPEG-G selective access and streaming is implemented to transfer the required data to the cloud
- GATK works on MPEG-G records instead of BAM records
- Preprocessing steps will be accelerated with MPEG-G

- This implementation has been realized in collaboration with UIUC.
- MPEG-G is a technology under consideration for NGS data storage and archival.

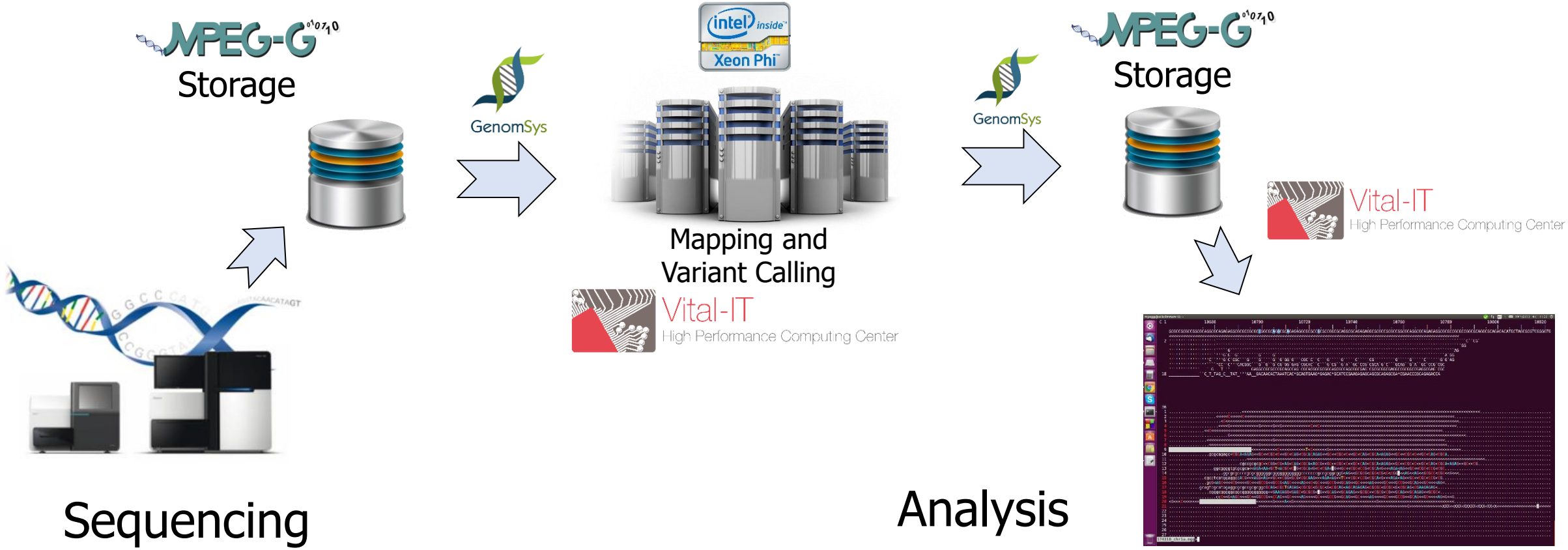
First public demonstration: genomic data streaming

On April 18th at the 122nd MPEG meeting in San Diego Vital-IT and GenomSys demonstrated the first MPEG-G “live” streaming of genomics data across the Atlantic

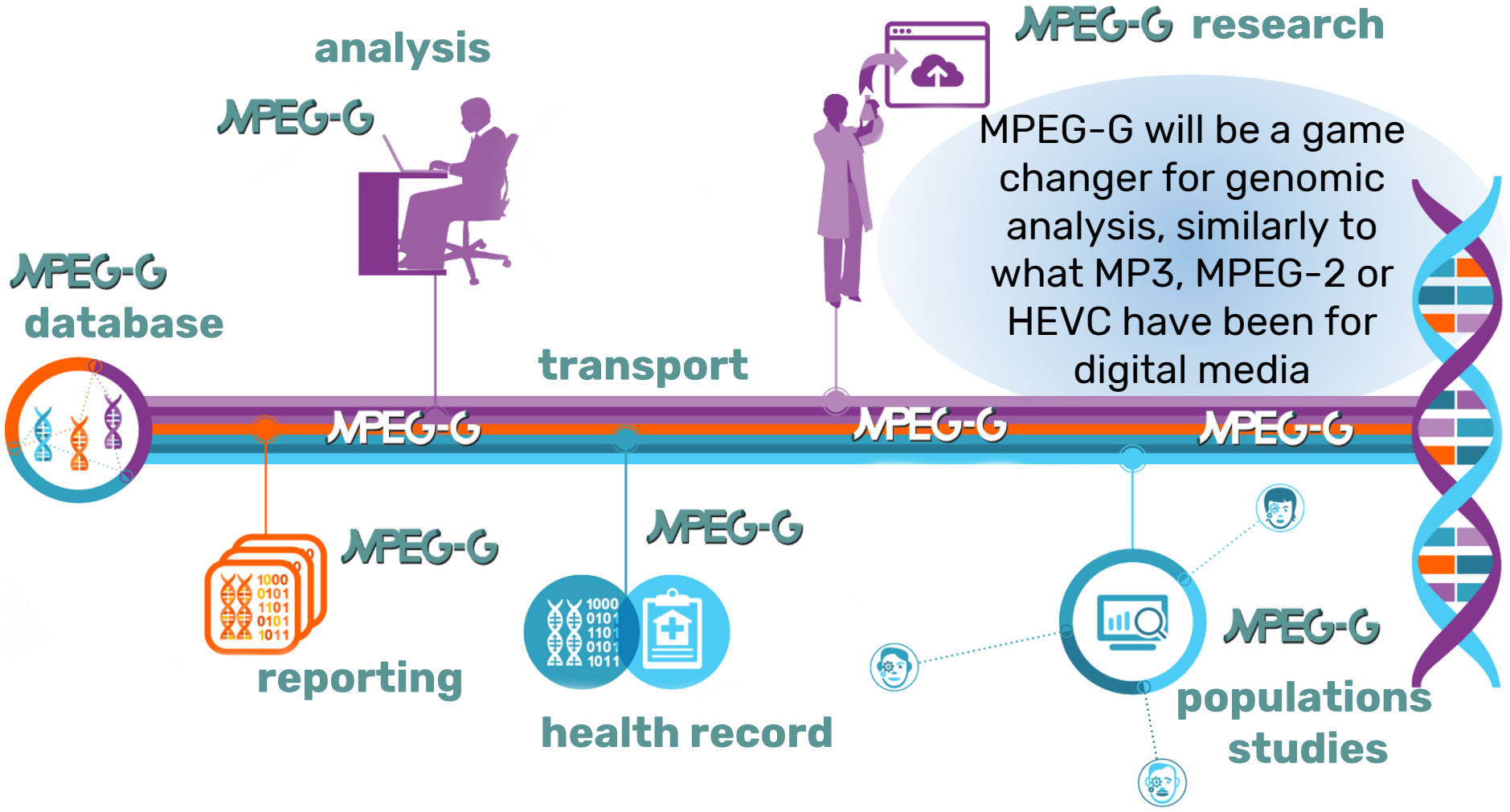


Next steps: from sequencing to analysis

MPEG-G enables the secure end-to-end management of genomics data



MPEG-G as a game changer



Thank you!

www.genomsys.com