

• THE ISO STANDARD FOR GENOMIC INFORMATION REPRESENTATION

MPEG-G.

A single, ISO-published file format for raw, aligned and annotated genomic data — built by the same standards body that gave the world MP3, MP4 and HEVC.

What is MPEG?



MPEG — the **Moving Picture Experts Group** — operates under the auspices of the International Organization for Standardization (ISO). It develops standards for digital media representation: transport and compression.

MPEG standards are an evolving series, each designed for a different purpose — from MP3 audio in 1993 to HEVC video in 2013, and now MPEG-G for genomic information in 2019.

WORKING GROUP

MPEG

Under ISO/IEC JTC 1

DELEGATES

1,700

Active contributors

COUNTRIES

40+

National bodies represented

SCOPE

Transport & compression

Of digital media

MPEG — 30 years of activity

Powerful compression tools and rigorous standardization have enabled the entire digital-media revolution. ISO-MPEG has demonstrated its ability to:

OUTCOME · 01

Generate high-quality, widely accepted industry standards.

OUTCOME · 02

Specify and maintain standards under non-discriminatory ISO/IEC licensing terms.

OUTCOME · 03

Produce higher-quality results than any proprietary solution.

MPEG STANDARDS POWER EVERYDAY MEDIA



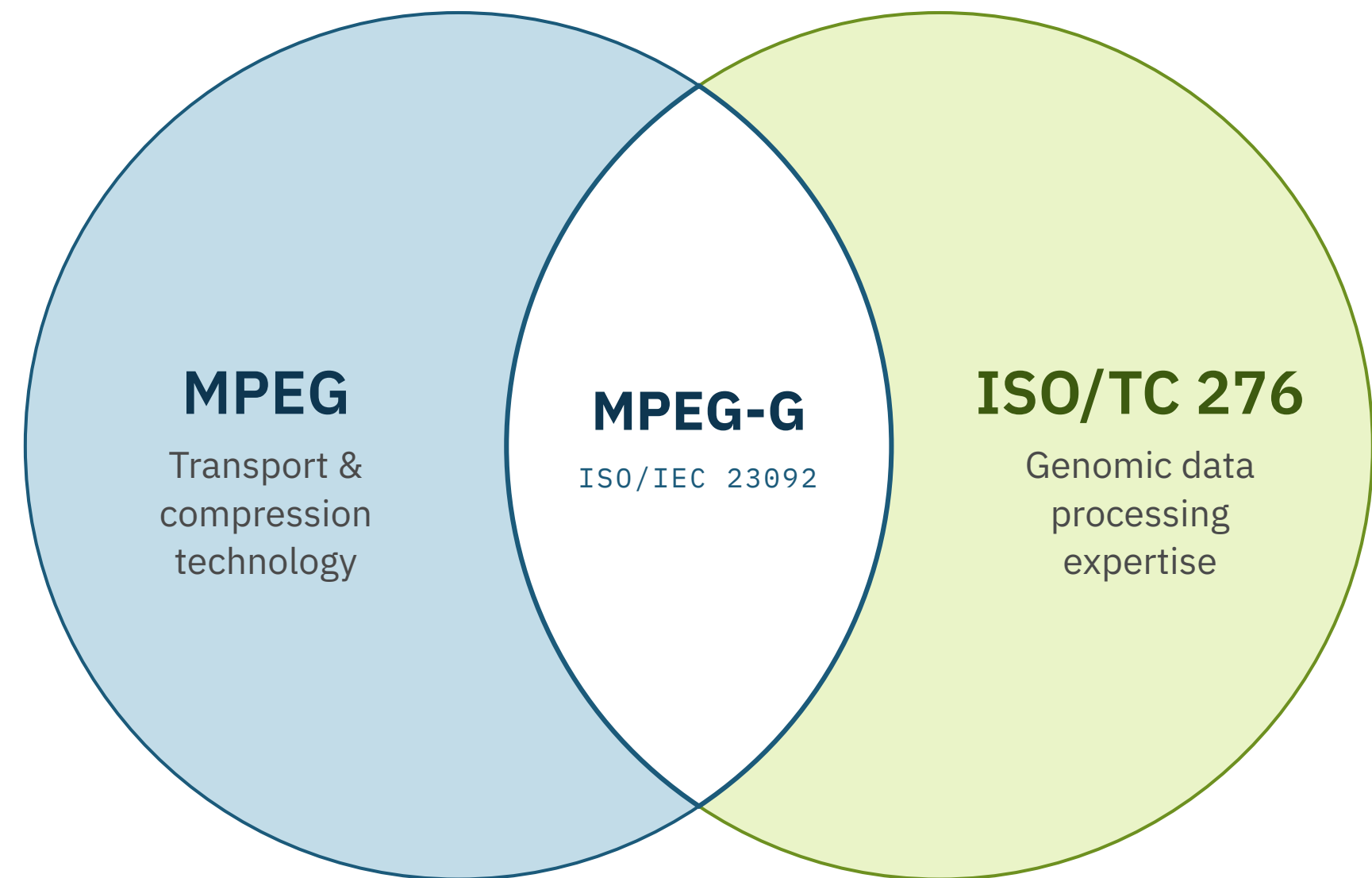
Beyond audio and video — into genomics

MPEG

Demands from increasingly sophisticated technologies have prompted MPEG to extend its standards from audio and video to a range of media-related technologies.

ISO/TC 276

Genomics — sequenced DNA — is intrinsically digital, and an ideal use-case for MPEG standardization.



Accelerate genomics adoption with a platform built on an open standard.

We enable scalable, secure genomic data applications, built on the only ISO/IEC standard for genomic information, which GenomSys helped shape.

The genomic international standard — and the foundation of our platform



4×+ higher compression

77%+ smaller files — cheaper long-term storage, faster cloud upload, storage on smartphone.

COMPRESSION



One single file format

Raw genomic data, assemblies, alignments, variants and metadata — all in one self-describing container.

FORMAT



Selective access & streaming

Up to 99% time saving — faster data access and analysis without re-downloading the full sample.

ACCESS



Finely tunable access control

Per-object access controls and native encryption — designed for clinical sign-out.

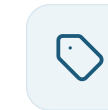
SECURITY



Guaranteed interoperability

Backward compatibility, no vendor lock-in, maintained under ISO/IEC licensing terms.

INTEROP



Integrated metadata

Metadata and annotations travel inside the container — never out of sync with the data.

METADATA

The MPEG-G standard consists of six parts.

PART 01

File & transport format

The technology to transport and access data — across files, networks and streams.

PART 02

Genomic information representation

The compressed representation — the core coding technology behind MPEG-G.

PART 03

APIs & metadata

Standard interfaces with DNA applications and legacy formats. Metadata for content protection and annotation.

PART 04

Reference software

The standard's reference implementation — guides developers building MPEG-G applications.

PART 05

Conformance

The methodology to test that an implementation complies with the standard.

PART 06

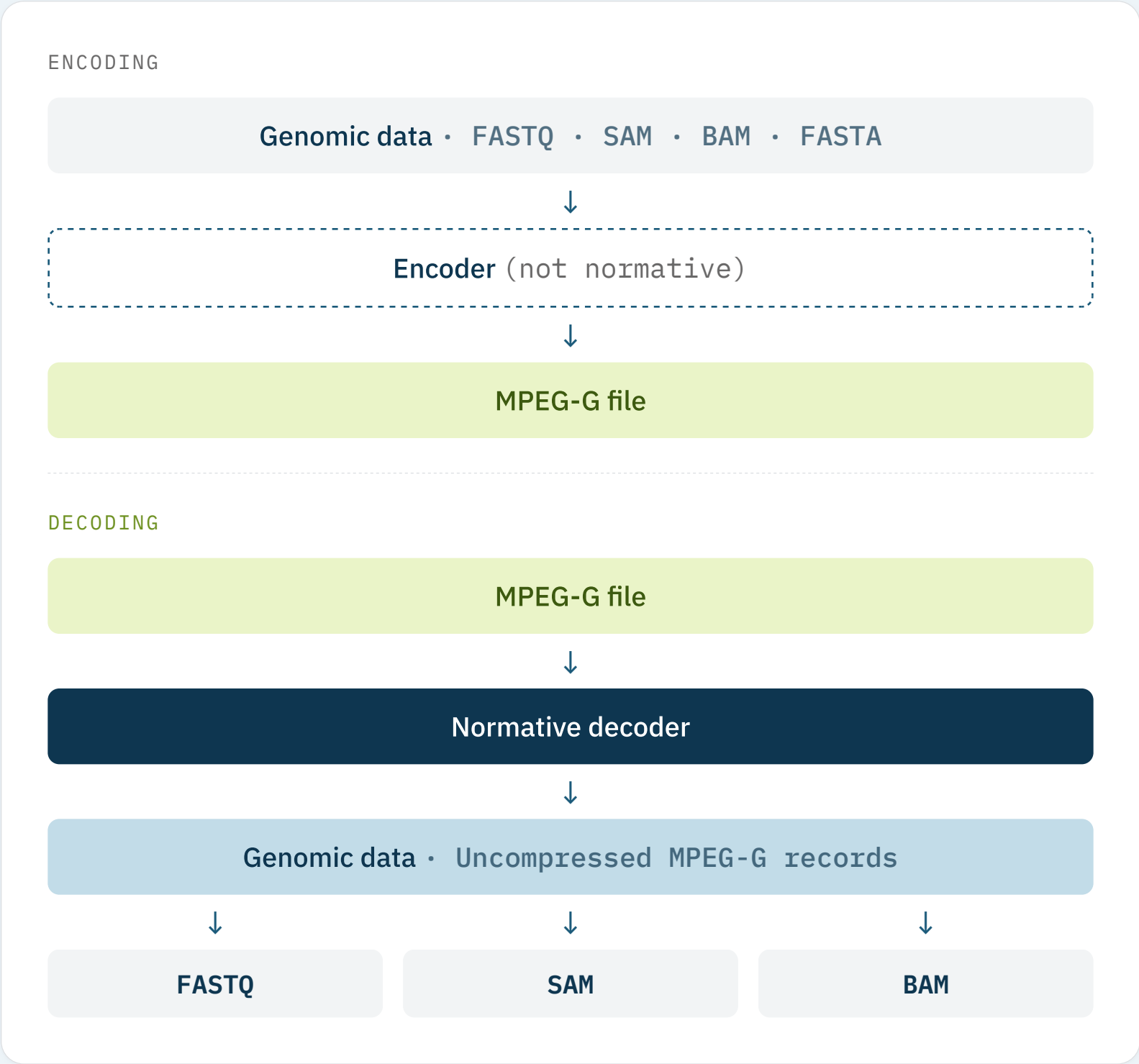
Annotations

A unified compressed format for high-level genomic analysis and downstream interpretation.

• MPEG-G SPECIFICATION

MPEG-G is...

- A specification of a **bitstream format** and the related decoding process
- Normative text **published by ISO**
- Reference software — **decoder only**, reference implementation, open source
- Encoder software — for **information only**, not normative



Raw data, aligned data, reference assemblies — all addressable by the same decoder.

One syntax replaces many formats.

Unmapped reads, aligned reads and reference assemblies — today spread across FASTQ, SAM/BAM and FASTA — are all represented inside a single, self-describing MPEG-G file.



MPEG-G compression performance

Across sequencing platforms, MPEG-G reduces aligned whole-genome data by **up to 3.7× versus the standard BAM** — with no loss of information.

PLATFORM	TYPE	REFERENCE	BAM (GB)	MPEG-G (GB)	REDUCTION
Illumina NovaSeq 6000	short-read WGS	GRCh38	41.5	11.2	3.7×
PacBio Sequel II (HiFi)	long-read WGS	GRCh38	58.4	36.7	1.6×
PacBio CLR	long-read WGS	GRCh37	57.8	32.8	1.8×

Even on high-entropy long-read data, where little is compressible, MPEG-G still shrinks the file 1.6–1.8× relative to BAM.

Coordinate-sorted

Lossless

Sample NA12878 (GIAB reference human) · whole-genome sequencing

An abstraction layer on top of core coding

The systems layer sits above the codec to enable:

- **Synchronisation** among components — reads, alignments, variants, annotations
- **Access to data** in the native compressed domain — no full decode required

When a compression technology evolves, the systems standard remains valid — the container outlives the codec.



File = a trio (Father · Mother · Child) **Dataset group** = one individual's history

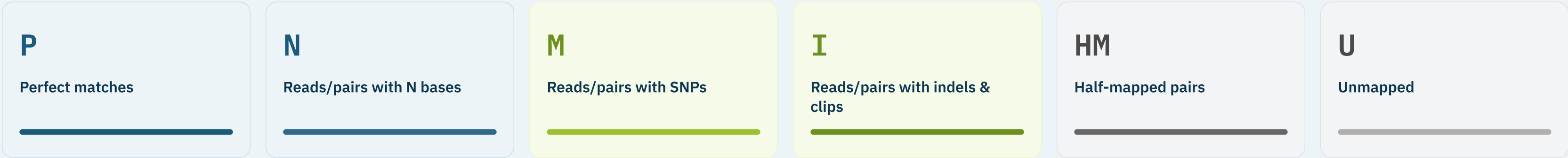
Dataset = one sequencing run

• SELECTIVE ACCESS

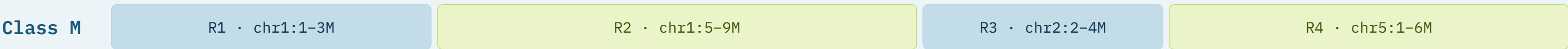
Two native indexes — class and region

At encode time the data is organised two independent ways: clustered by alignment result into six classes, and indexed through custom per-class regions — so the decoder fetches only the slices that matter.

① CLUSTERED BY ALIGNMENT RESULTS



② INDEXED BY CUSTOM PER-CLASS REGIONS



Every class carries its own region index — fetch **chr1:5-9M** from Class M alone, without decoding the rest. Regions can be grouped into **labels** for single-query access.

• MORE EFFECTIVE DATA ORGANISATION

Labels – single-query access across regions

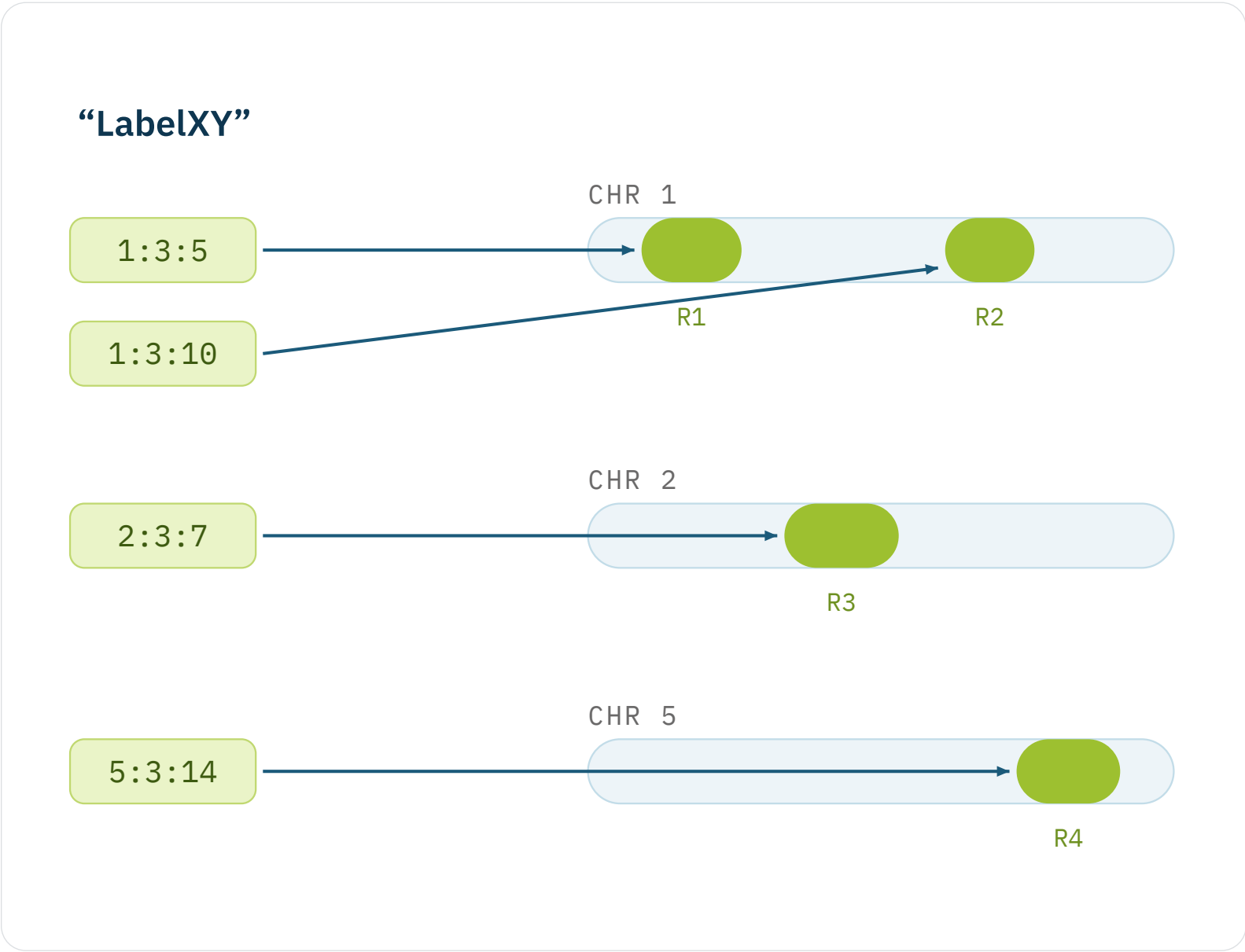
Several genomic regions can be grouped under a single label. The decoder retrieves the labelled bundle as one query – even if the regions are scattered across chromosomes.

A natural fit for biological constructs that span the genome – gene-expression interactions, regulatory networks, custom panels.

"LabelXY" → 1:3:5 · 1:3:10 · 2:3:7 · 5:3:14

Each entry is `sequence : start : end` – two regions on chr1, one on chr2, one on chr5. One query returns the whole bundle.

These could be regions that interact in gene expression.



Same data — different operating model

LEGACY FORMATS — WITHOUT MPEG-G

The file is the unit.

- Very few aligners actually use Quality Values
- Yet up to **90% of file content** is QV bytes
- Entire files are still downloaded and processed
- No way to skip irrelevant fields at retrieval time

WITH MPEG-G

Every field is addressable.

- No mandatory or optional fields — every field is decodable separately
- Pull only the slice you need: SNVs, indels, mapped reads, ...
- Skip QVs entirely when the downstream tool doesn't use them
- Native indexing by class, region and label

THE BOTTOM LINE

MPEG-G delivers savings of up to **90%**

Storage • Bandwidth • Processing time

From download to streaming

Bulk download works — but for clinical workflows, MPEG-G defines a transport layer for progressive data streaming.

- Encapsulation in a format suitable to any kind of network
- Self-consistent snapshot of received data at any time — even before generation is complete
- Transcoding guidelines from transport packets back to the file

Streamable · resumable · packet-aligned

DOWNLOAD · THEN USE

BAM file · receiving... 62%

🕒 not usable

Nothing is usable until the entire file has arrived.

PROGRESSIVE STREAMING · USABLE AT EVERY STEP

P1

P2

P3

P4

...



✓ A valid MPEG-G file at any point — even before data generation completes.

Self-describing packets tolerate any network; transcoding guidelines map the received stream losslessly back to the file.

Protection is **part of the format.**

Defined in ISO/IEC 23092-3 and applied across the box hierarchy — protection boxes attach to any indexed level:

Dataset group dgpr

Dataset dtpr

Descriptor stream dspr

Access unit aupr

01 **Access control**

Selective AES encryption (CTR / GCM) at group / dataset / stream / AU granularity; access gated by key possession.

02 **Integrity**

XML digital signatures and GCM authentication tags over blocks and access units — tampering is detectable.

03 **Traceability**

Signatures bind authenticated provenance to each protected object (origin + integrity verifiable); no in-format access log.

04 **Privacy-rules hierarchy**

XACML policies at dataset-group and dataset levels, evaluated as a DG → DT cascade at API access time.

05 **Roles segregation**

XACML subject / role rules; roles are deployment-defined, enabling separation of duties without a single trust authority.

Normative on decode, open on encode

- Reference implementation that supports and guides MPEG-G developers
- Not intended as an efficient implementation
- Maintained by MPEG • open source

The encoder is provided by MPEG for information only — it is **not normative**.
Anyone can ship a better encoder; conformance is judged at the decode side.

A COLLABORATIVE EFFORT

Built by a global academic consortium

The reference software is the product of a joint effort across leading genomics and computing research groups.

CONTRIBUTING INSTITUTIONS

Stanford



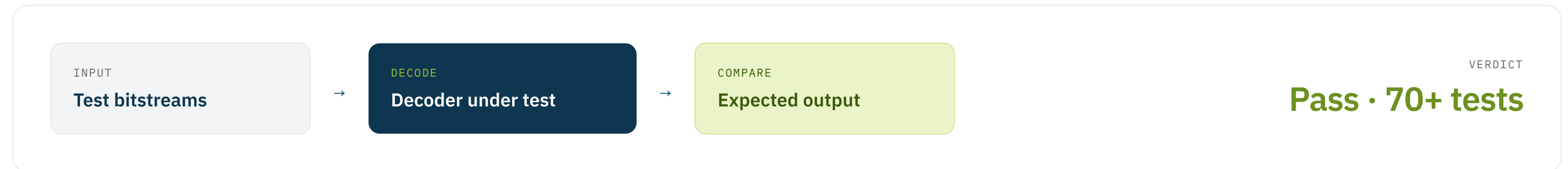
COMMUNITY ENCODER **OPEN SOURCE**

GENIE – GENomic Information Encoding

Produces compliant MPEG-G bitstreams — the community counterpart to the normative reference decoder.

How do I check my implementation is ISO compliant?

MPEG-G defines test bitstreams and expected decoded output. An MPEG-G-labelled device must pass **70+ documented conformance tests**. More can be added as the standard evolves.



An ISO standard, implemented by a certified company — regulatory compliance by design



The ISO standard

MPEG-G (ISO/IEC 23092) is a formal international standard — not a proprietary format.

- Formal specification developed under ISO auspices
- Long-term maintenance commitment
- Documented process of extensions and amendments
- 70+ documented conformance tests verify ISO compliance



GenomSys

Tools built around the standard by an organisation that holds the certifications that matter.

- ISO 9001 — quality management
- ISO/IEC 27001 — information security management
- ISO/IEC 27701 — privacy information management
- ISO 13485 — medical-device quality management

From the lab bench to the smartphone — all on MPEG-G



COMMAND-LINE TOOL

MPEG-G Suite

Encodes and decodes legacy genomic files to and from MPEG-G, with full control over every file.

FOR BIOINFORMATICIANS,
SEQUENCING LABS & OEM
PARTNERS



ON-PREM WEB APP

Variant Discovery

Secondary and tertiary analysis for germline and somatic variant annotation, working on MPEG-G files natively.

FOR GENETICISTS & CLINICAL
BIOINFORMATICIANS



ON-PREM WEB APP • UNDER
VALIDATION

Hemotyper

Genomic blood-group typing across 36 blood groups, built natively on MPEG-G.

FOR BLOOD-TRANSFUSION
CENTRES



MOBILE APP • PROTOTYPE

GenomSafe

A consumer iOS & Android app for in-app genomic data storage — a white-label concept for putting the genome in the patient's hands.

SHOWN TO LABS & HOSPITALS AS
A WHITE-LABEL CONCEPT



CONTACTS

GENERAL

info@genomsys.com

OFFICES

MONZA

GenomSys S.r.l.

Via A. Gambacorti Passerini 2
20900 Monza (MB)
Italy

MILANO

MPEG-G Competence Center

Via Filippino Lippi, 32
20131 Milano (MI)
Italy

www.genomsys.com