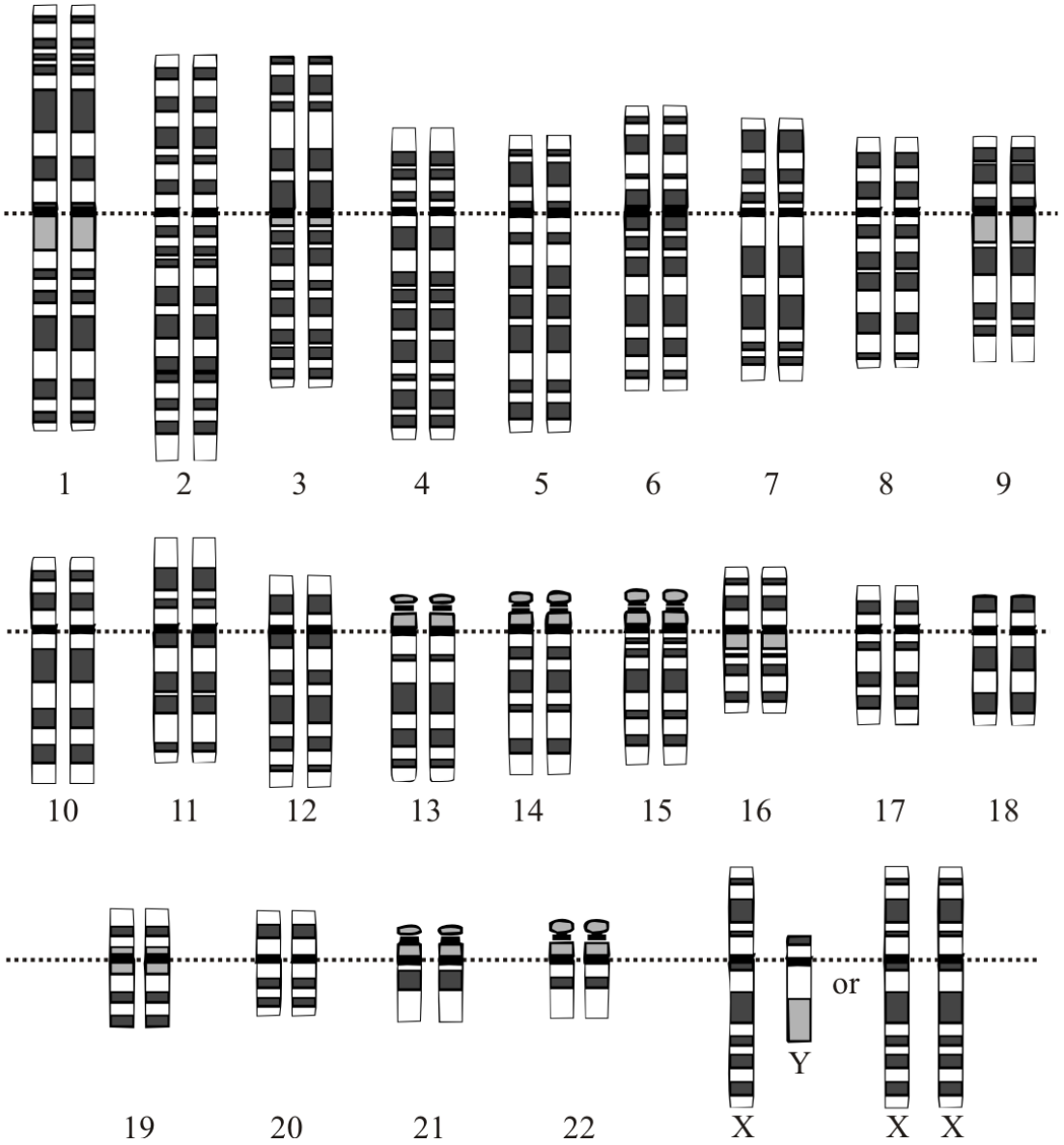
Leibniz Universität Hannover

MPEG-G: The Standard for Genomic Information Representation

Jan Voges

Institut für Informationsverarbeitung

Human karyogram



- Diploid genome
- 46 chromosomes
- 22 pairs of autosomes (1-22)
- 1 allosome pair (XY | XX)
- 4 bases (A, C, G, T)
- ~3 billion base pairs

DNA sequencing

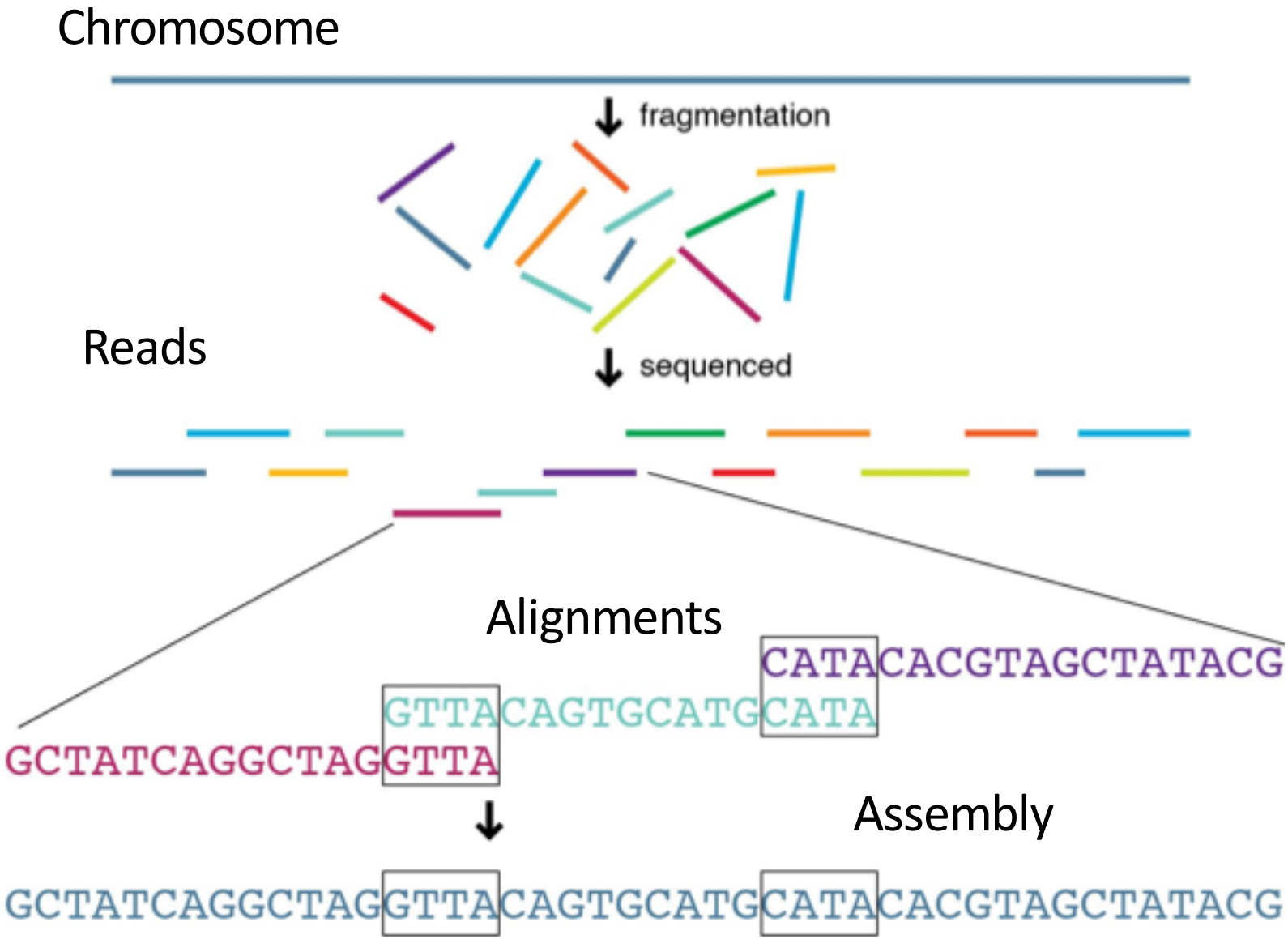
Human genome: ~3 billion base pairs x 2 bits per base = **~750 MB**

+ read-out redundancy
~500 GB

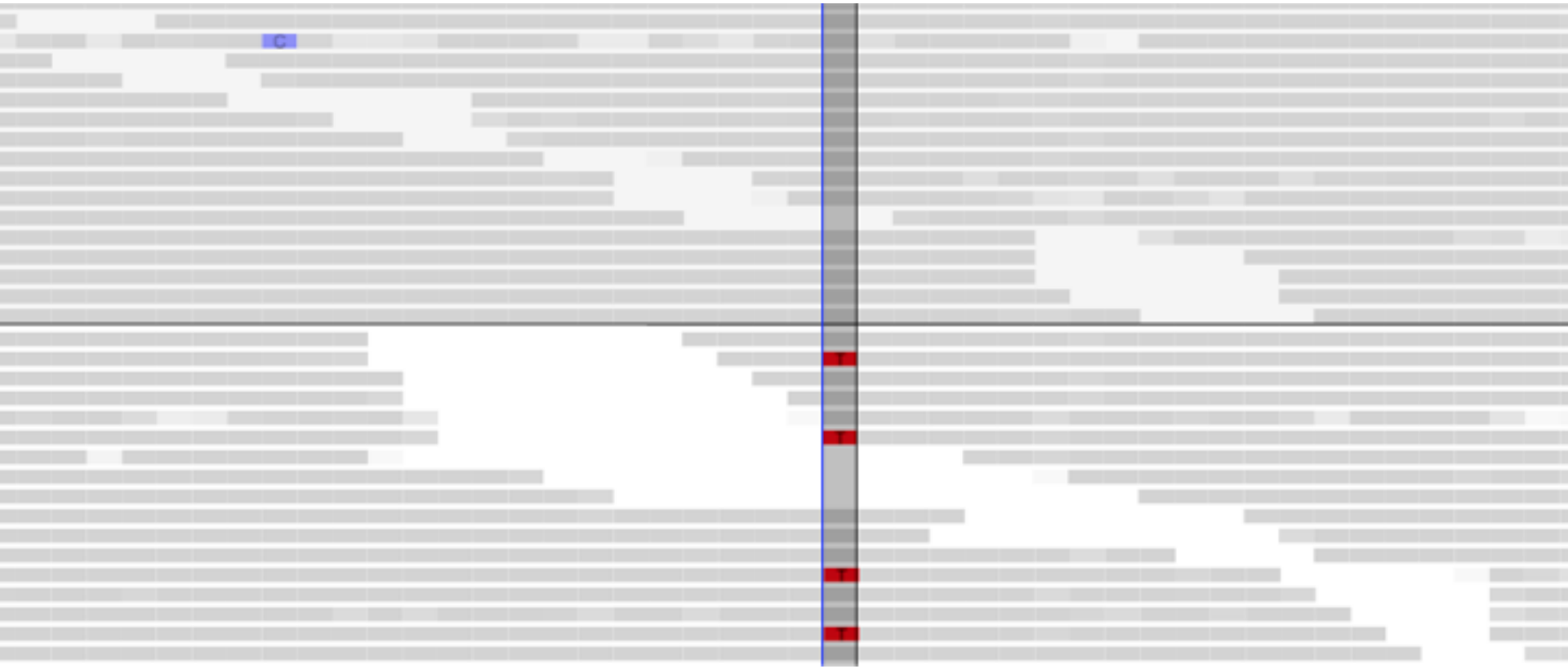
+ meta information
~1 TB

+ alignment information
~1.5 TB

Whole genome sequencing



Alignment



Evolution of genome sequencing

Sequencing technology

	2009	2017/2018
Cost/genome	$100k	$1k
Coverage	~30x	> 200x
Number of reads	~1 billion	> 6 billion
Size of raw sequencing files	~0.25 TB	> 1.5 TB

Storage & transmission infrastructure

	2009	2017/18
Cost/TB	$100	$50
Download speed	10 Mbps	100 Mbps

No technology is keeping with the pace of genome sequencing!

What is MPEG-G?

- MPEG-G = International Standard ISO/IEC 23092
- Largest coordinated and international effort addressing the problems and limitations of current technologies
- Paves the road towards a truly efficient and economical handling of genomic information
- Utilizing the latest technologies
- Planned release: end of 2018

Objectives of MPEG-G

- **Interoperable selective access to data in the compressed domain** by means of standard APIs:
  - Genomic region
  - Class of data
  - …
- On top of compression **higher performance is provided by a specific file format and transport format.**

Support the evolution of a SW/HW ecosystem so that compression technology becomes a commodity for the users.

Benefits provided by MPEG-G

- Selective access to compressed data
- Data streaming
- Genomic studies aggregation
- Enforcement of privacy rules
- 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
- Incremental update of sequencing data and metadata
- Compressed file concatenation

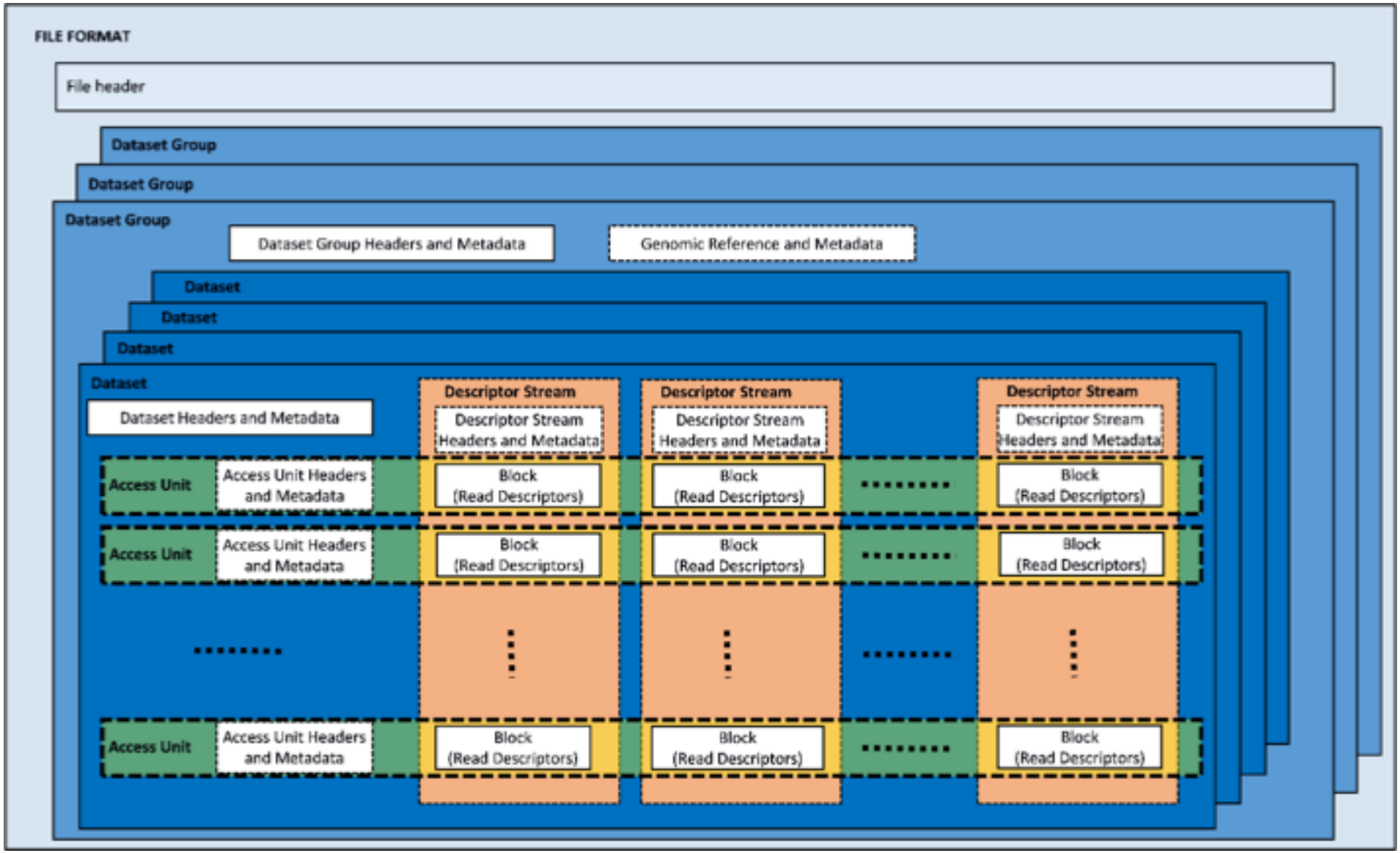
Structure of the MPEG-G standard

- **Part 1: File and Transport Format**
  - The technology to transport and access data
- **Part 2: Genomic Information Representation**
  - 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 Software**
  - The standard support to the implementation of applications

Structure of the MPEG-G standard

- **Part 1: File and Transport Format**
  - The technology to transport and access data
- **Part 2: Genomic Information Representation**
  - 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 Software**
  - The standard support to the implementation of applications

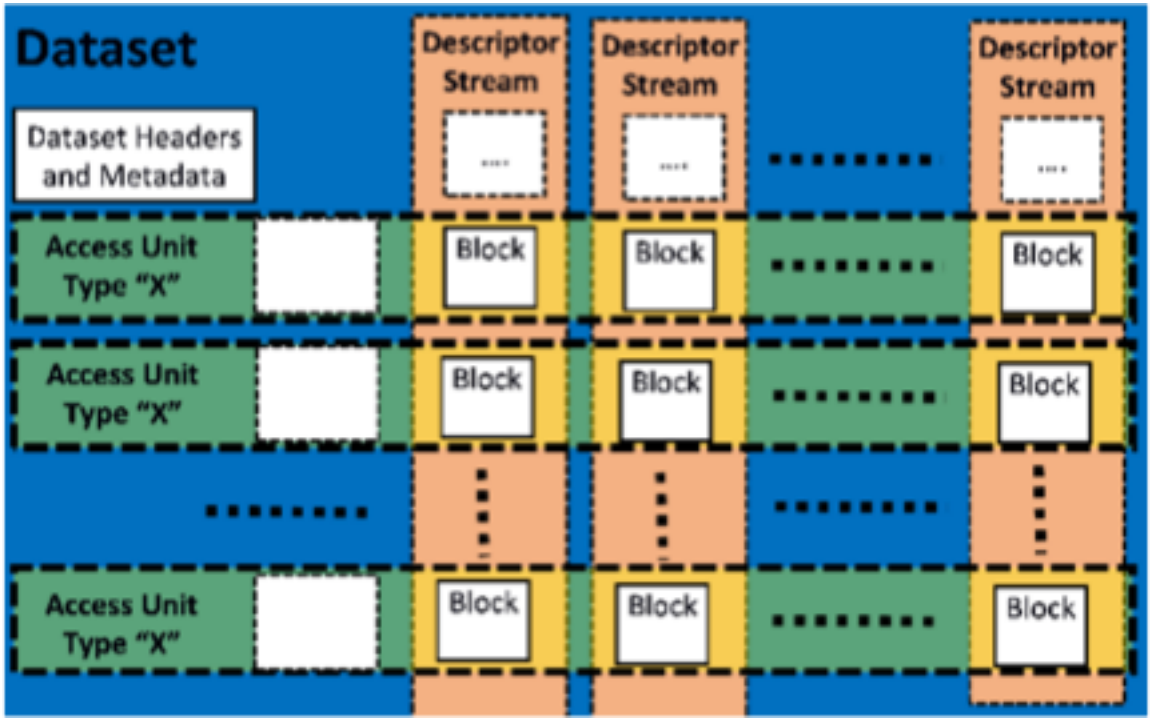
MPEG-G file format



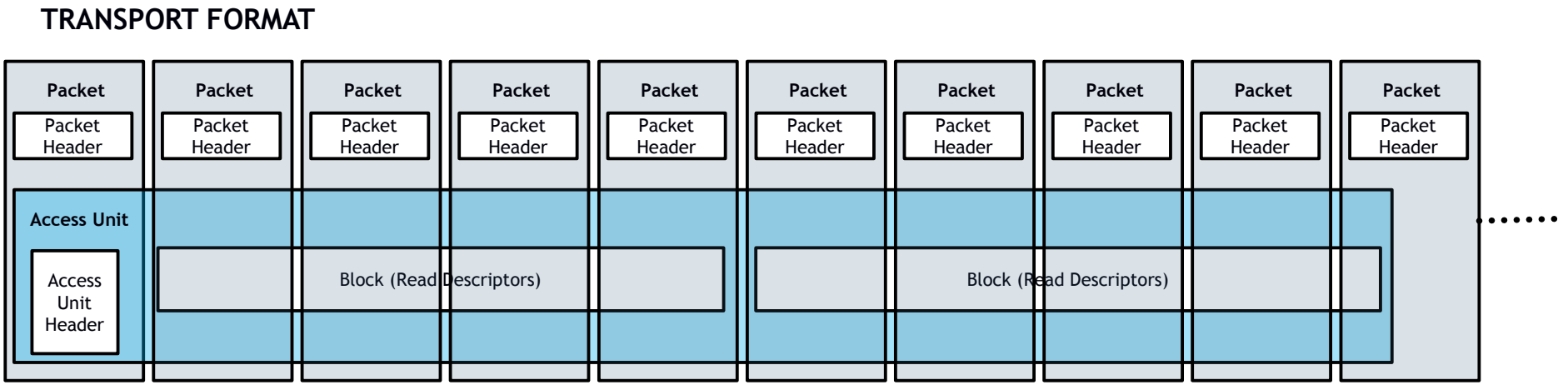
Selective access

The indexing tools embedded in an MPEG-G file enable several types of selective access that can be combined in the same query, e.g.:

- Genomic interval in terms of start to end mapping position on a given reference sequence
- Type of data (i.e., a single data class)



MPEG-G transport format / streaming



MPEG-G streaming features:

- Packet size adaptation to the channel characteristics/state
- Error detection and support of re-transmission of erroneous/incomplete data for error-free delivery
- Support of out-of-order delivery
- Packet-based filtering of genomic data
- Full convertibility of file and transport formats

Structure of the MPEG-G standard

- **Part 1: File and Transport Format**
  - The technology to transport and access data
- **Part 2: Genomic Information Representation**
  - 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 Software**
  - The standard support to the implementation of applications

genie

- **genie** = **GEN**omic **I**nformation **E**ncoding
- Joint collaborative effort to produce a standard-compliant open source encoder

Leibniz Universität Hannover

tnt

Stanford University

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

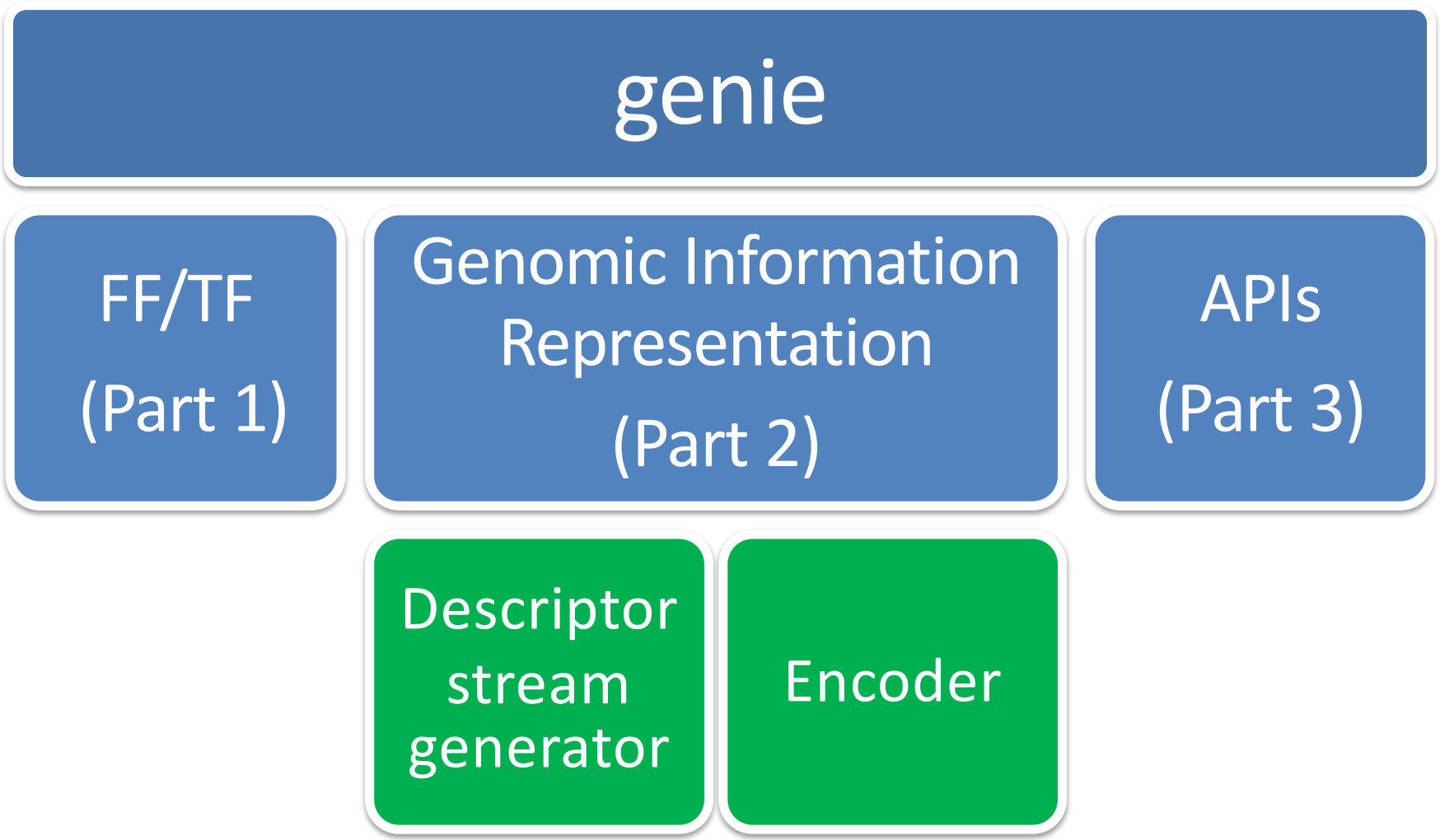
MPEG-G implementations

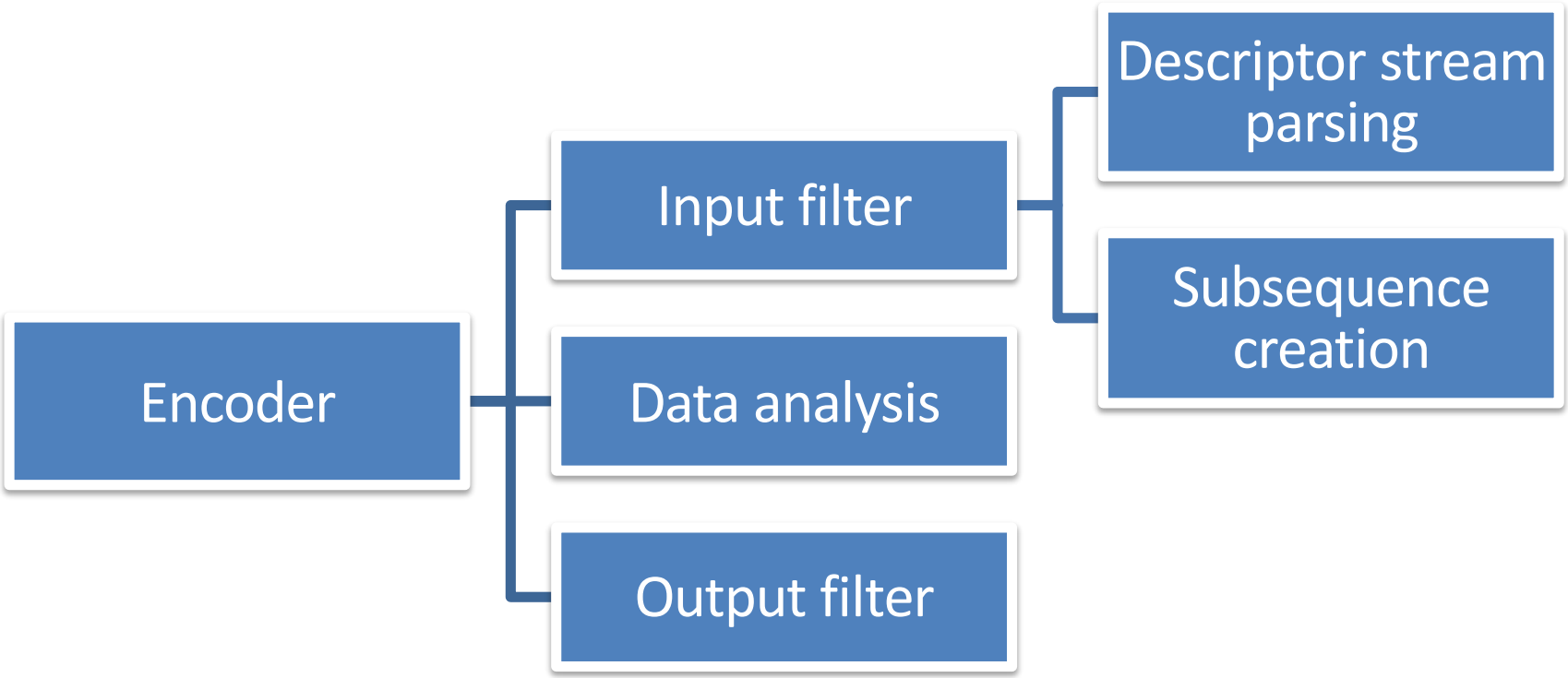
Aginome

MPEG

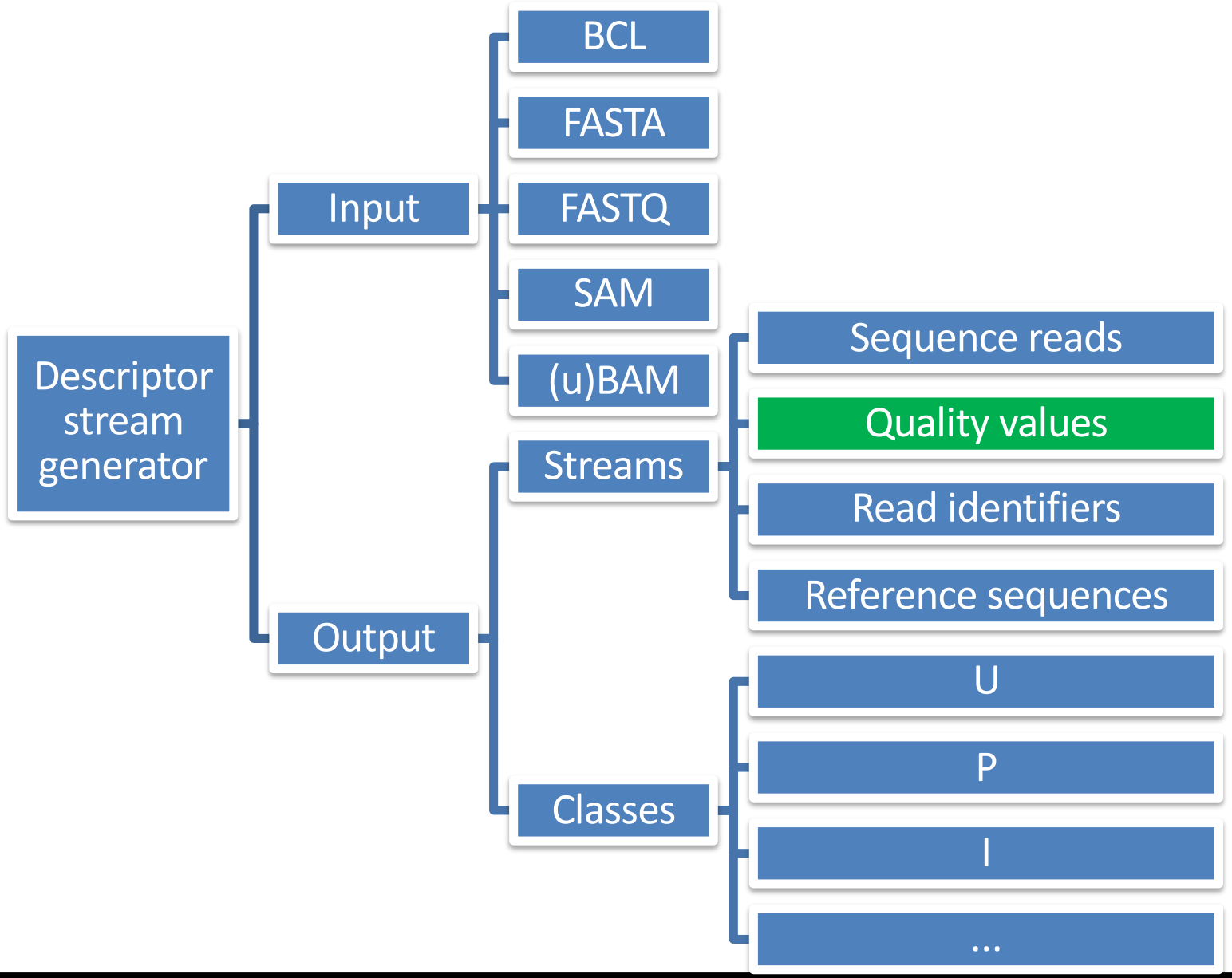
GenomSoft

GenomSys


genie
FF/TF
(Part 1)
Genomic Information
Representation
(Part 2)
APIs
(Part 3)
Descriptor
stream
generator
Encoder


Encoder
Input filter
Data analysis
Output filter
Descriptor stream parsing
Subsequence creation

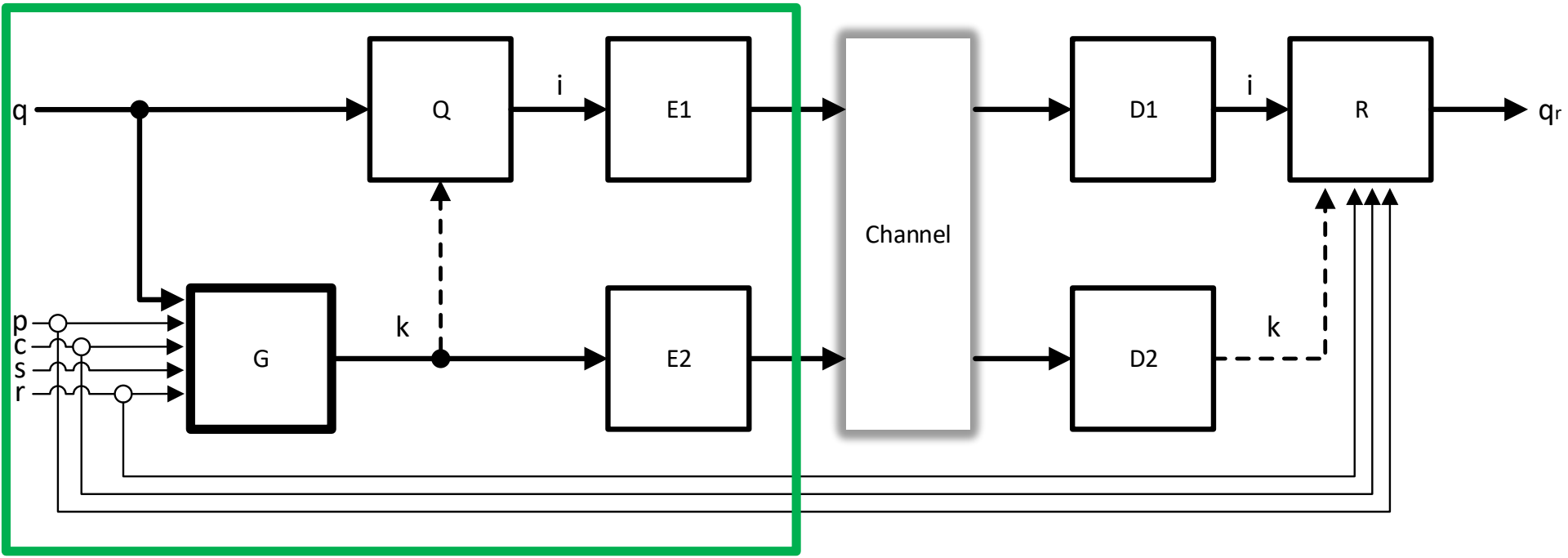
genie


Descriptor stream generator
Input
BCL
FASTA
FASTQ
SAM
(u)BAM
Output
Streams
Sequence reads
Quality values
Read identifiers
Reference sequences
Classes
U
P
I
...

Coding of quality values

Year	Tool
2011	• SlimGene (Kozanitis et al.)
2012	• SCALCE (Hach et al.)
2013	• QualComp (Ochoa et al.) • BEETL (Janin et al.) • Fastqz (Bonfield et al.)
2014	• Illumina's binning • P-/R-Block (Cánovas et a.l.)
2015	• Quartz (Yu et al.) • QVZ (Malysa et al.)
2016	• Crumble (Bonfield)
2017	• **CALQ** (Voges et al.)
2018	• QSComp (Voges et al.)

CALQ



Legend

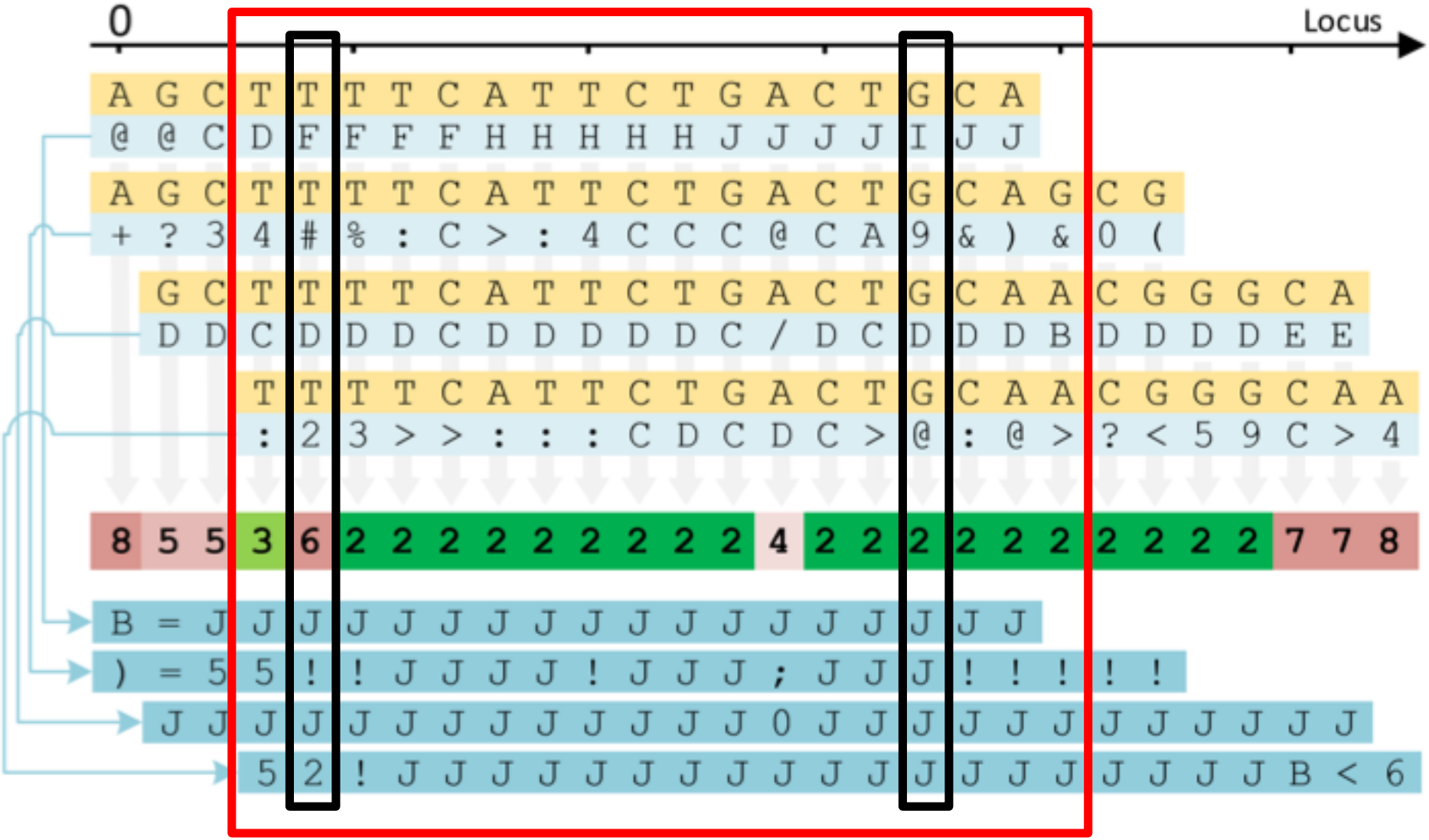
q	quality values
q_r	reconstructed quality values
p, c, s, r	side information (mapping positions, CIGAR strings, donor sequences, reference sequence(s))
k	quantizer indexes
i	quantization indexes

signal

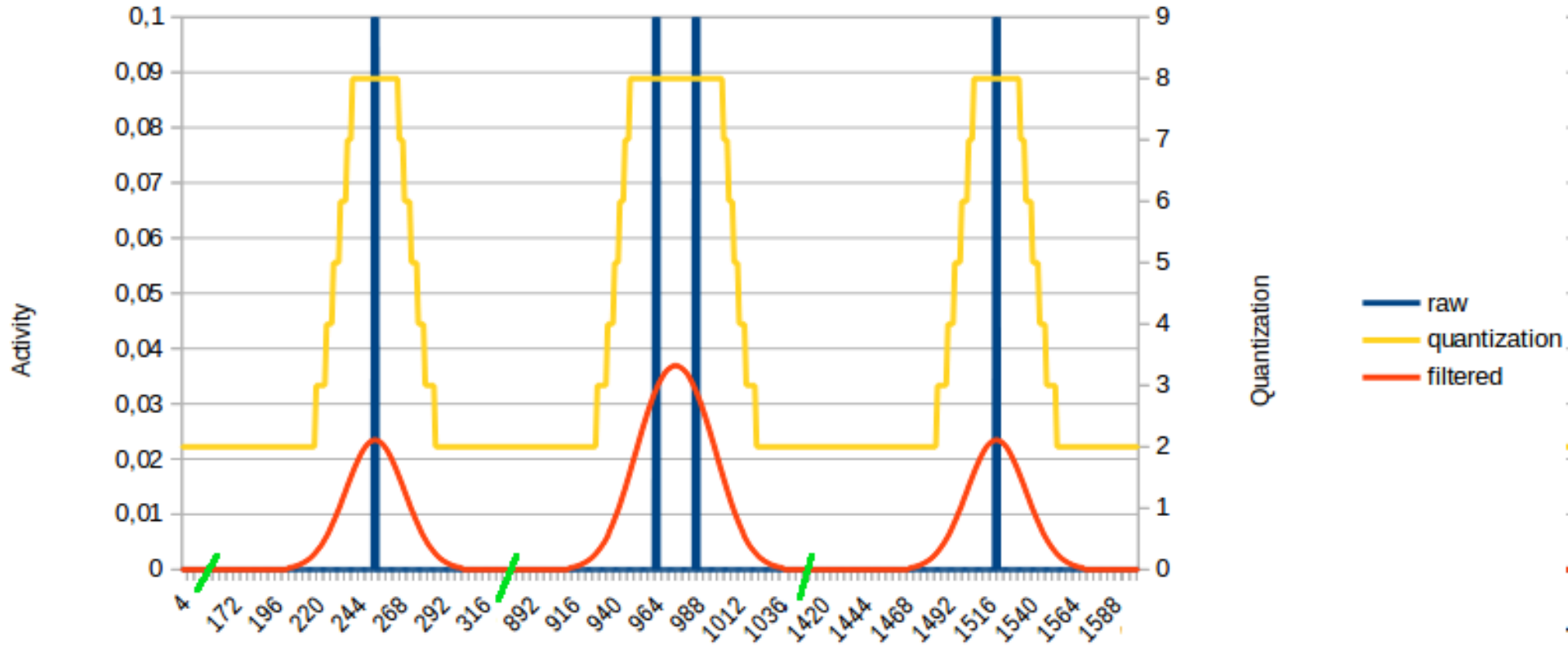
control signal

side signal

CALQ

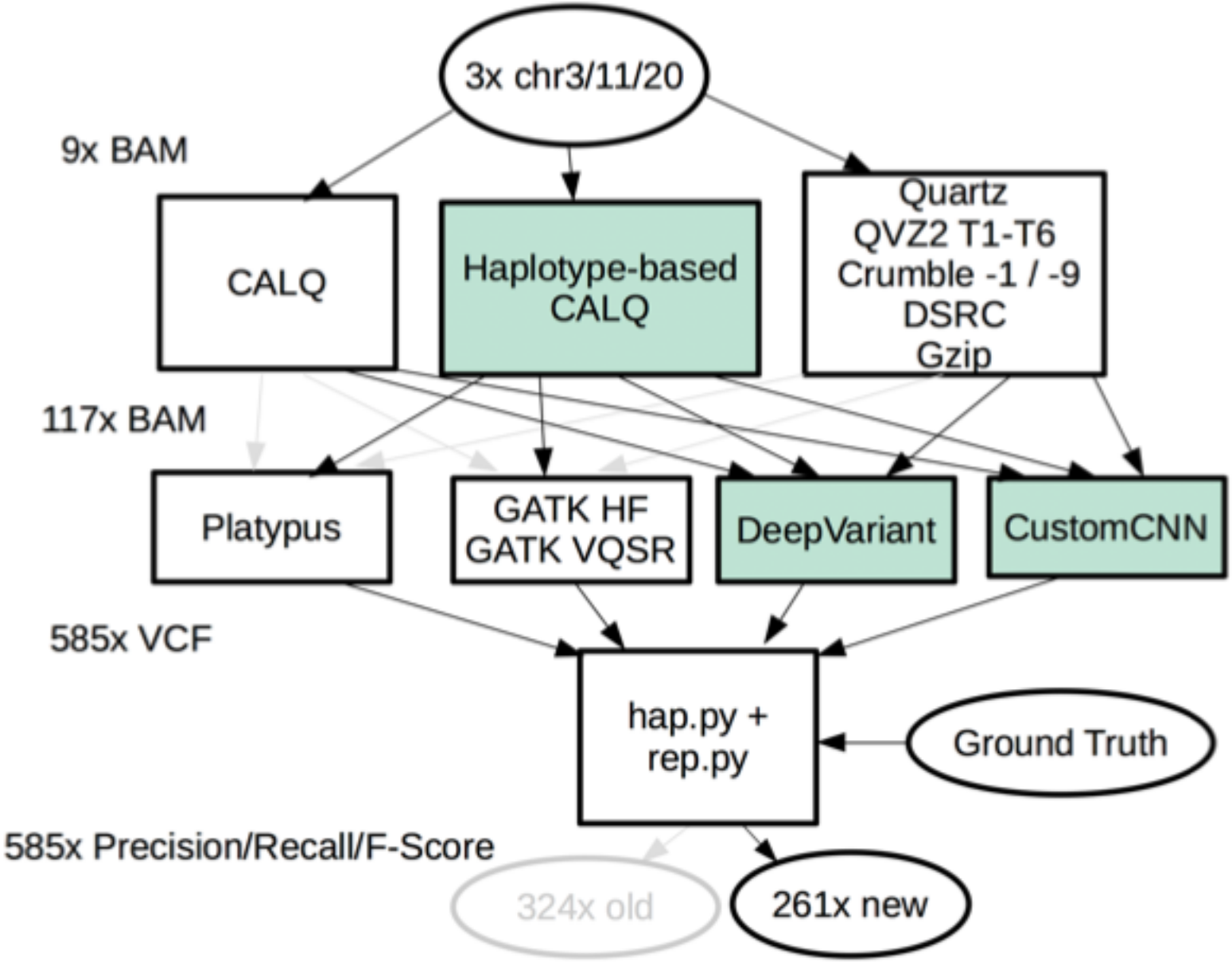

0
Locus
A G C T T T T C A T T C T G A C T G C A
@ @ C D F F F F H H H H H J J J J I J J
A G C T T T T C A T T C T G A C T G C A G C G
+ ? 3 4 # % : C > : 4 C C C @ C A 9 &) & 0 (
G C T T T T C A T T C T G A C T G C A A C G G G C A
D D C D D D C D D D D D C / D C D D D B D D D D E E
T T T T C A T T C T G A C T G C A A C G G G C A A
: 2 3 > > : : : C D C D C > @ : @ > ? < 5 9 C > 4
8 5 5 3 6 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 7 7 8
B = J J J J J J J J J J J J J J J J J J J
) = 5 5 ! ! J J J J ! J J J ; J J J ! ! ! ! !
J J J J J J J J J J J J J 0 J J J J J J J J J J J
5 2 ! J J J J J J J J J J J J J J J J J J B < 6

CALQ v2: sequence activity and quantization

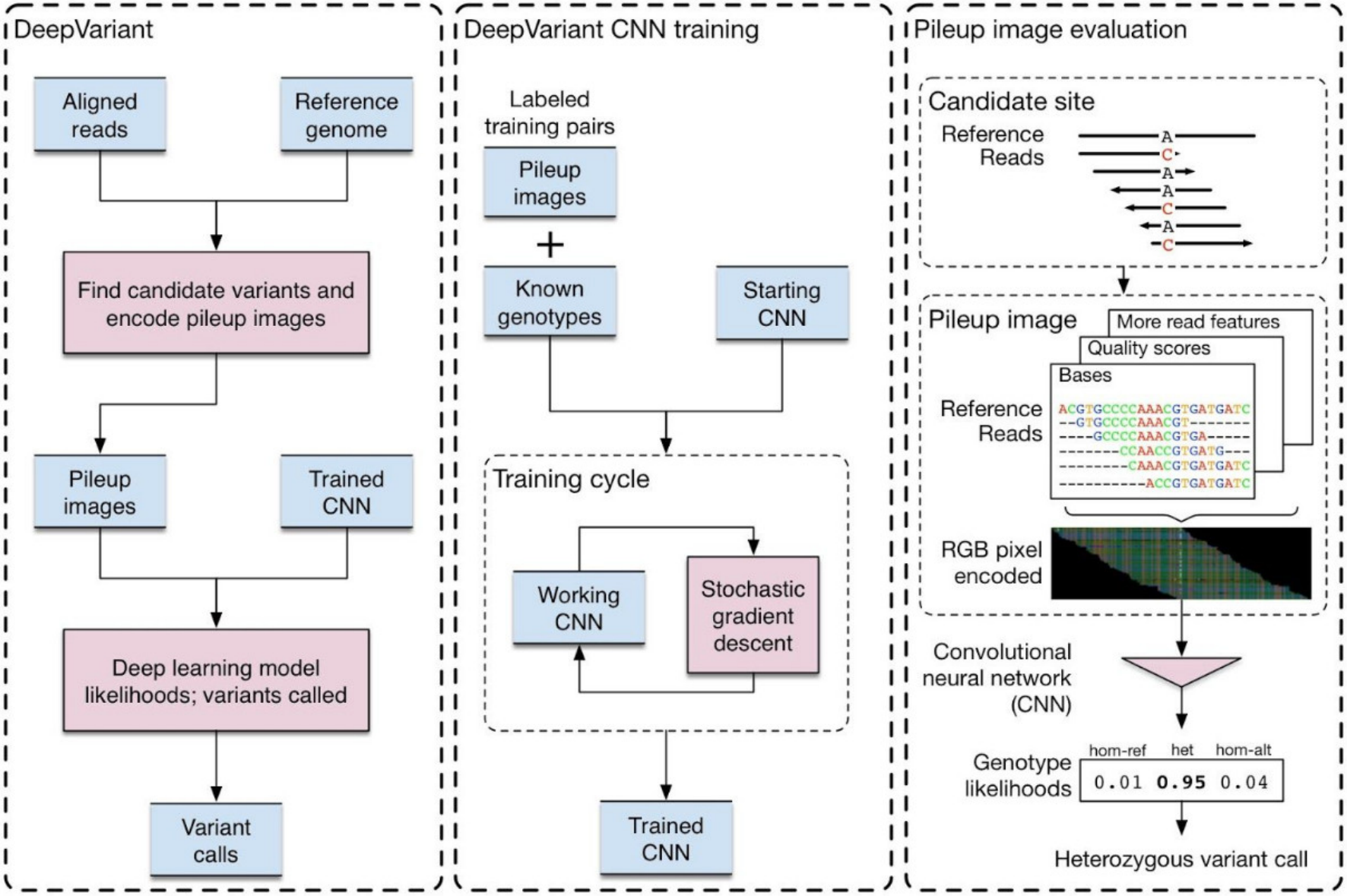


Dipl.-Ing. Jan Voges
voges@tnt.uni-hannover.de

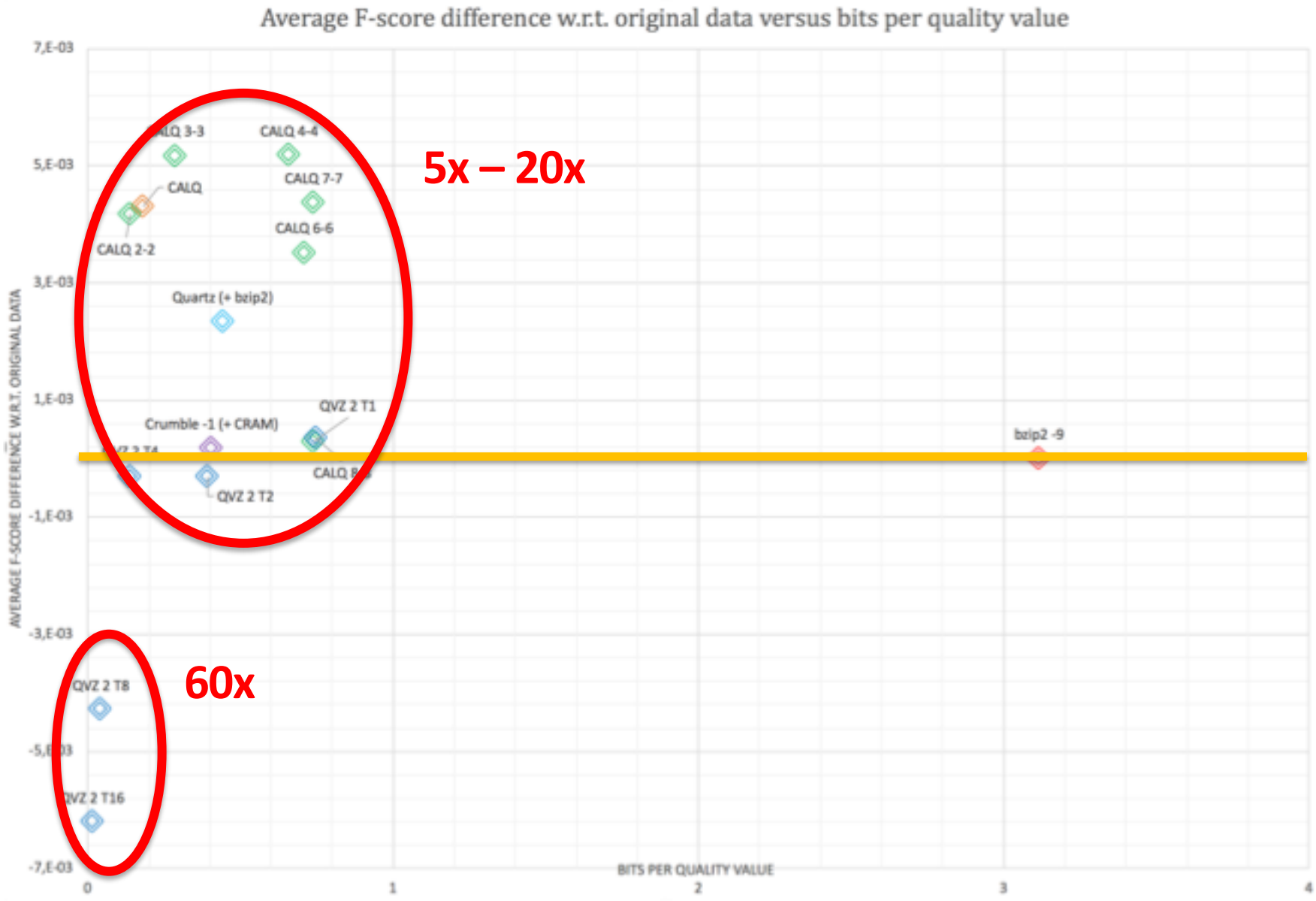
Evaluation of lossy quality value compression



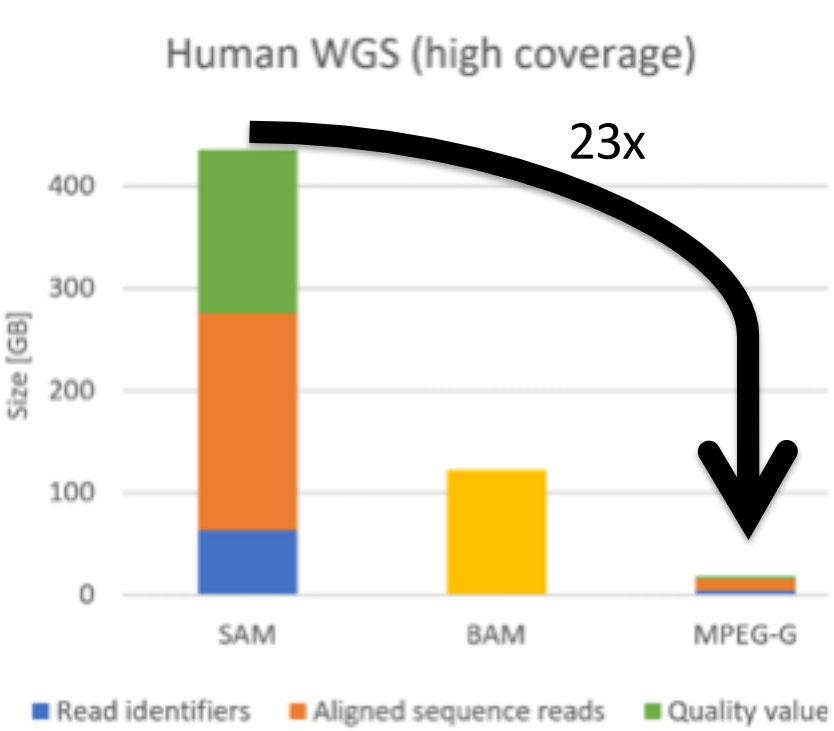
Variant calling using neural networks

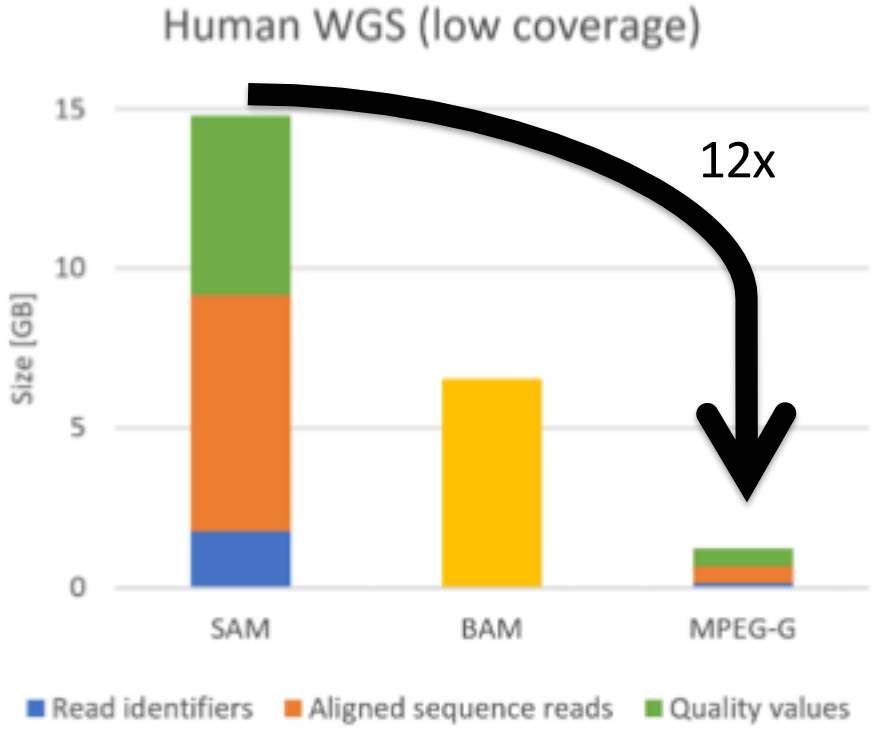


RD diagram for quality values

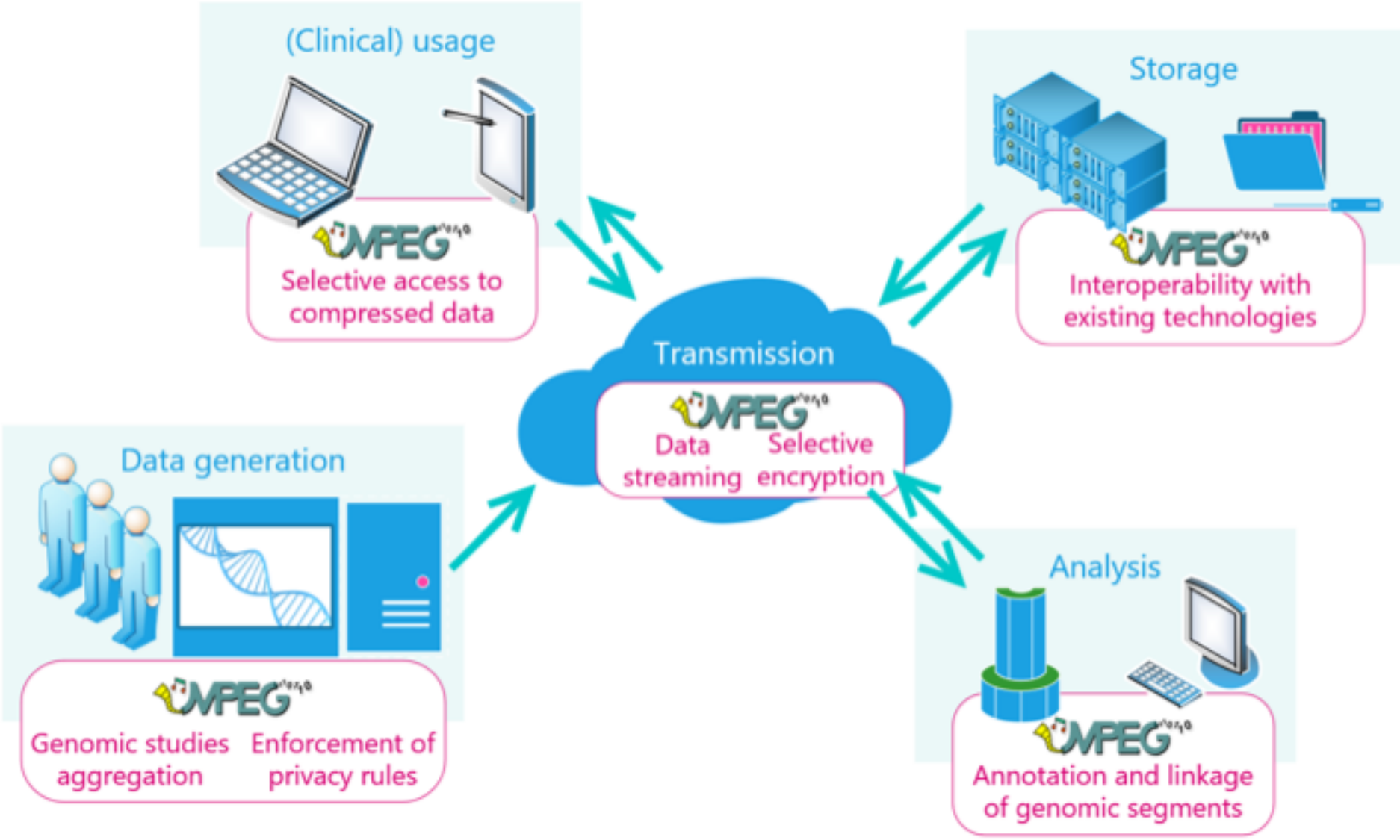


MPEG-G performance





A genomic ecosystem fueled by MPEG-G



The future of genomic data

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

- The technology to compress genomic information will **change over time**
- Genomic information compression performance will **improve over time**
- The MPEG-G Systems technologies will evolve and improve, but **the main functionality will stay** and support the evolution of analysis application

A new logo?