INTERNATIONAL STANDARD

ISO/IEC 23092-6

First edition
2023-11

Information technology — Genomic information representation —

Part 6: Coding of genomic annotations

Reference number
ISO/IEC 23092-6:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

A list of all parts in the ISO/IEC 23092 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

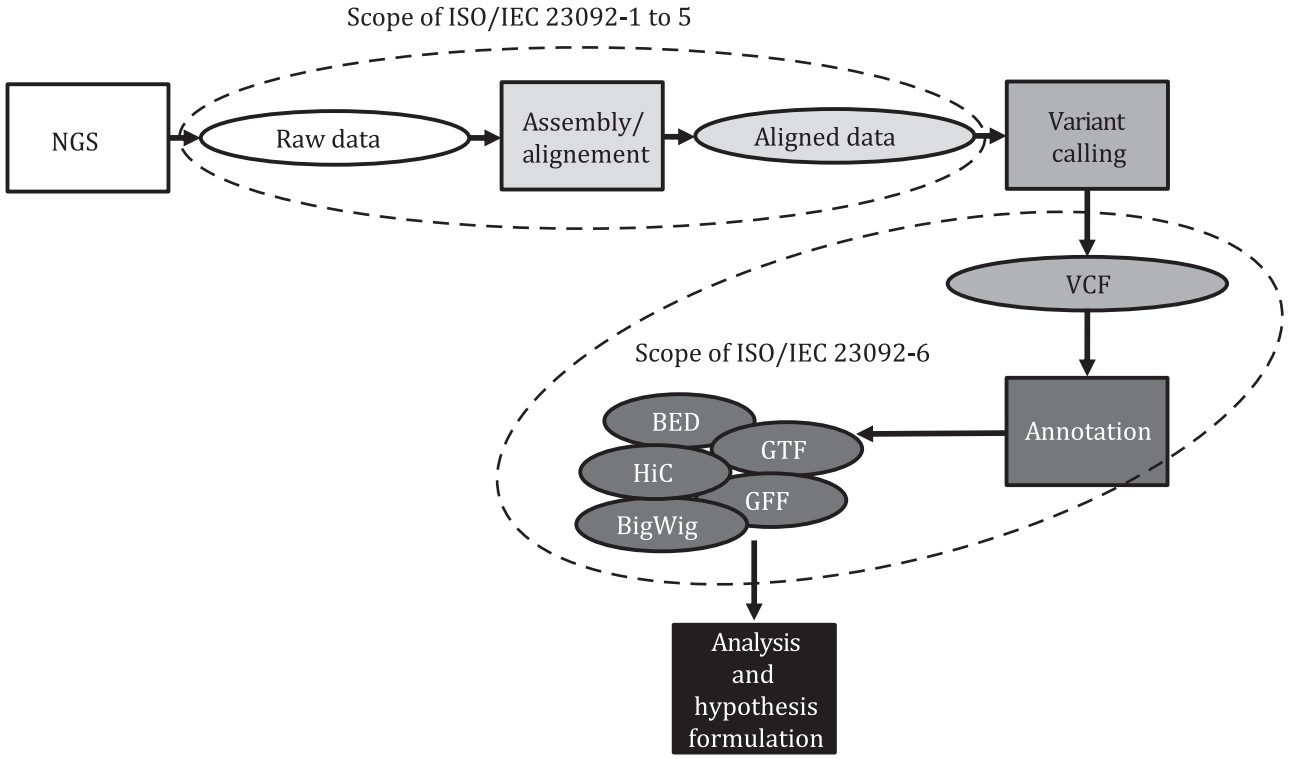
Introduction

While ISO/IEC 23092-1 to ISO/IEC 23092-5 (MPEG-G) deal with the representation of genomic information derived from the primary analysis of high-throughput sequencing (HTS) data – sequencing reads and qualities, and their alignment to a reference genome – which is only the first step in a long series. In particular, the results of primary analysis are usually processed further in order to obtain higher-level information. Such a process of aggregating information deduced from single reads and their alignments to the genome into more complex results is generally known as secondary analysis. In most HTS-based biological studies, the output of secondary analysis is usually represented as different types of annotations associated to one or more genomic intervals on the reference sequences.

Biological studies typically produce genomic annotation data such as mapping statistics, quantitative browser tracks, variants, genome functional annotations, gene expression data and Hi-C contact matrices. These diverse types of downstream genomic data are currently represented in different formats such as VCF, BED, WIG, etc., with loosely defined semantics, leading to issues with interoperability, the need for frequent conversions between formats, difficulty in the visualization of multi-modal data and complicated information exchange. Figure 1 depicts a typical pipeline for the primary and secondary analyses of HTS data, the file formats involved and the scopes of different parts of the ISO/IEC 23092 series.

Furthermore, the lack of a single format has stifled the work on compression algorithms and has led to the widespread use of general compression algorithms with suboptimum performance. These algorithms do not exploit the fact the annotation data typically comprises of multiple fields (attributes) with different statistical characteristics and instead compress them together. Therefore, while these algorithms support efficient random access with respect to genomic position, they do not allow extraction of specific fields without decompressing all the whole file.

In response to the aforementioned challenges, this document details a unified data format for the efficient representation and compression of diverse genomic annotation data for file storage or data transport. The benefits are manifold: reducing the cost of data storage, improving the speed of random data access and processing, providing support for data security and privacy in selective genomic regions, and creating linkages across different types of genomic annotation and sequencing data. The ultimate goal is to enable the secured and seamless sharing, processing and analysis of multi-modal genomic data in order to reduce the burden of data manipulation and management, so scientists can focus on biological interpretation and discovery.



Key

- sequencing generates raw reads
- read alignment
- variant calling
- variants annotations
- analysis

Figure 1 — Typical pipeline for the primary and secondary analyses of HTS data

Information technology — Genomic information representation —

Part 6: Coding of genomic annotations

1 Scope

This document provides specifications for the normative representation of the following types of genomic information:

— variants with genotyping information

— functional annotations

— tracks

— expression matrices

— contact matrices (from Hi-C experiments or similar).

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this specification. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 10646, *Information technology — Universal coded character set (UCS)*

ISO/IEC 11544, *Information technology — Coded representation of picture and audio information — Progressive bi-level image compression*

ISO/IEC 23092-1, *Information technology — Genomic information representation — Part 1: Transport and storage of genomic information*

ISO/IEC 23092-2, *Information technology — Genomic Information Representation — Part 2: Coding of Genomic Information*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
access unit
logical data structure containing a coded representation of genomic information to facilitate bit stream access and manipulation

3.2
access unit start position
position of the leftmost mapped base among the first alignments of all genomic records contained in the access unit, irrespective of the strand

3.3
access unit end position
position of the rightmost mapped base among the first alignments of all genomic records contained in the access unit, irrespective of the strand

3.4
access unit range
genomic range comprised between the access unit start position and the rightmost genomic record position among all genomic records contained in the access unit

3.5
access unit covered region
genomic range comprised between the access unit start position and the access unit end position inclusive

3.6
alignment
information describing the similarity between a sequence (typically a sequencing read) and a reference sequence (for instance, a reference genome)

Note 1 to entry: An alignment is described in terms of a position within the reference, the strand of the reference, and a set of edit operations (matches, mismatches, insertions and deletions, clipping of the sequence ends and splicing information) needed to turn the first sequence into the second.

3.7
allele
each of one or more alternative sequences for a genomic segment

Note 1 to entry: There can be more than one either because the genome contains more than one, almost identical, copies of the same genomic material (2 in the case of humans for all chromosomes from 1 to 22), and/or because one is considering more than one individual in the population.

3.8
annotation record
record
data structure representing a tuple of annotation information (e.g. the properties associated to a variant, a genomic feature or a generic range; it is used also to identify a ”row” when data have matrix format)

3.9
base
base pair
synonymous of nucleotide

3.10
base position
number of bases between a base and the leftmost mapped base belonging to the same genomic segment.

3.11
CIGAR string
CIGAR
textual way of representing an alignment

Note 1 to entry: Several definitions have been used by different programs, the ones referred to here is the one used in the SAM format. It encodes a set of edit operations (matches, mismatches, insertions and deletions, clipping of the sequence ends and splicing information) needed to turn the sequencing read into the reference.

3.12
cluster
aggregation of genomic records

3.13
cluster signature
signature
sequence of nucleotides that is common to most or all genomic records belonging to a cluster

3.14
contig
set of overlapping DNA segments, sequenced and assembled, that together represent a consensus region of DNA

Note 1 to entry: the term “contig” derives from “contiguous”.

3.15
dataset
compression unit containing one or more of: reference sequences; sequencing reads; and alignment information

Note 1 to entry: Datasets are specified in ISO/IEC 23092-1.

3.16
deletion
contiguous removal of one or more bases from a genomic sequence

3.17
E-CIGAR
extended CIGAR syntax specified as a superset of the CIGAR syntax

Note 1 to entry: Among other things, E-CIGAR enables the unambiguous representation of substitutions, spliced reads and splice strandedness.

3.18
edit operation
modification of a sequence of nucleotides by means of a substitution, deletion, insertion or clip

3.19
FASTA
GIR that includes a name and a nucleotide sequence for each sequencing read

Note 1 to entry: Additional information is usually encoded in the read identifier by bioinformatics tools (such as database information, and base calling information).

3.20
FASTQ
GIR that includes FASTA and quality values

3.21
first end
end 1
read 1
first segment of a paired-end template

Note 1 to entry: Illumina platforms usually store first and second ends in two separate files and in the same order – i.e. the n-th read of the first FASTQ file and the n-th read of the second FASTQ file belong to the same template.

3.22
genomic descriptor
descriptor
element of the syntax used to represent a feature of a genomic sequencing read or associated information such as alignment information or quality values

3.23
genomic information representation
way to describe a sequence and some information associated with it

Note 1 to entry: Which information is represented varies depending on the GIR.

3.24
genomic position
position
integer number representing the zero-based position of a nucleotide within a reference sequence

3.25
genomic range
range
interval of positions on a reference sequence specified by a start position s and an end position e such that $s \leq e$

Note 1 to entry: The start and the end positions of a genomic range are always included in the range.

3.26
genomic record
record
data structure representing a tuple optionally associated with alignment information, read identifier and quality values

3.27
genomic record index
position of a genomic record in the sequence of genomic records encoded in an access unit

3.28
genomic record position
0-based position of the leftmost mapped base on the reference genome of the first alignment contained in a genomic record

Note 1 to entry: A base present in the aligned read and not present in the reference sequence (insertion) and bases preserved by the alignment process but not mapped on the reference sequence (soft clips) do not have mapping positions.

3.29
genomic reference
reference
collection of reference sequences

Note 1 to entry: Typical examples are a reference genome or a reference transcriptome.

3.30
genomic segment
segment
contiguous sequence of nucleotides

Note 1 to entry: Typically output of the sequencing process, and sequenced from one strand of a template.

3.31
genomic variant
variant
one of the possible sequences for a genomic segment whenever more than one allele for that segment is present

Note 1 to entry: The variant can span one nucleotide (and is then usally called single nucleotide polymorphism) or more (structural variants can involve changes in thousands of contiguous bases or more). A variant can consist of an indel.

3.32
genotype
sequence of a genomic segment for a specified copy of the genome or individual whenever more than one allele for that segment is present

3.33
genotype matrix
metrix specifying which genotype is present in each copy of the genome or individual

3.34
hard clip
one or more bases originally present at either side of a read, and removed from it following alignment

Note 1 to entry: The bases are no longer present in the sequence of the read.

3.35
indel
contiguous stretch of nucleotides that, when aligning two sequences, are inserted into one sequence, or alternatively deleted from the other, in order to make the two sequences the same

Note 1 to entry: From “insertion or deletion”.

3.36
insertion
contiguous addition of one or more bases into a genomic sequence

3.37
leftmost read end
leftmost read
sequencing read generated by a paired-end sequencing run and mapped at a position on the reference sequence which is smaller than the mapping position of the other read in the pair

3.38
mapped base
base of the aligned read that either matches the corresponding base on the reference sequence or can be turned into the corresponding base on the reference sequence via a substitution

3.39
nucleotide
monomer of a nucleic acid polymer such as DNA or RNA

Note 1 to entry: Nucleotides are denoted as letters (‘A’ for adenine; ‘C’ for cytosine; ‘G’ for guanine; ‘T’ for thymine which only occurs in DNA; and ‘U’ for uracil which only occurs in RNA). The chemical formula for a specific DNA or RNA molecule is given by the sequence of its nucleotides, which can be represented as a string over the alphabet (‘A’, ’C’, ’G’, ‘T’) in the case of DNA, and a string over the alphabet (‘A’, ‘C’, ‘G’, ‘U’) in the case of RNA. Bases with unknown molecular composition are denoted with ‘N’.

3.40
output annotation record
annotation record produced as output of the decoding process of an annotation table or a portion of it

3.41
paired-end reads
paired-end template
tuple made of two segments

Note 1 to entry: Typically, the segments correspond to the beginning and the end of the same nucleic acid molecule.

3.42
pileup
textual representation of sequencing reads aligned to a reference sequence

3.43
ploidy
number of equivalent alleles present at each position of the genome

3.44
phased genotyping
information about consecutive genotypes along the genome which keeps information about the different copies of the genome (or different individuals) separate, whenever multiple alleles are present

3.45
quality value
quality score
number assigned to each nucleotide base call in automated sequencing processes

Note 1 to entry: Quality values express the base-call accuracy, i.e. the probability (or a related measure) for a nucleotide in the sequence to have been incorrectly determined.

3.46
read group
set of reads having some property in common

3.47
read identifier

read header
read name
text string associated with each sequencing read stored in GIRs such as FASTA, FASTQ and SAM

Note 1 to entry: The read identifier is usually unique within its dataset, and may contain additional information as encoded by bioinformatics tools (such as database information, and base calling information).

3.48
reference genome
representative example of the sequences for a species' genetic material

Note 1 to entry: Representative of the sequences of the DNA molecules present in a typical cell of that species.

3.49
reference sequence
nucleic acid sequence with biological relevance

Note 1 to entry: Each reference sequence is indexed by a one-dimensional integer coordinate system whereby each integer within range identifies a single nucleotide. Coordinate values can only be equal to or larger than zero. The coordinate system in the context of this standard is zero-based (i.e. the first nucleotide has coordinate 0 and it is said to be at position 0) and linearly increasing within the string from left to right.

3.50
rightmost read end
rightmost read
sequencing read generated by a paired-end sequencing run and mapped at a position on the reference sequence which is greater than the mapping position of the other read in the pair