**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/WG 11 N18150**

**Marrakech, MA – January 2019**

|  |  |
| --- | --- |
| **Source** | **ISO/IEC JTC1/SC29/WG11** |
| **Status** | **Approved** |
| **Title** | **Master table of Terms and Definitions for ISO/IEC 23092** |

Table of Contents

[1 Purpose 2](#_Toc535527419)

[2 Terms and definitions 2](#_Toc535527420)

# Purpose

This document collects terms and definitions used in the series of standards ISO/IEC 23092. The list in square brackets associated to each term represents the parts of the standard where the term is used.

# Terms and definitions

1. [1,2,3,5]

**Access Unit**

Logical data structure containing a coded representation of genomic information to facilitate bit stream access and manipulation[[1]](#footnote-1)

1. [1,2]

**Access Unit Start Position**

Leftmost Genomic Record Position among all Genomic Records contained in the Access Unit

1. [1,2]

**Access Unit End Position**

Rightmost base position among all mapped bases of all Genomic Records contained in the Access Unit

1. [1,2]

**Access Unit Range**

**AU Range**

Genomic Range comprised between the Access Unit Start Position and the right-most Genomic Record Position among all Genomic Records contained in the Access Unit

1. [1]

**Access Unit Covered Region**

**AU Covered Region**

Genomic Range comprised between the Access Unit Start Position and the Access Unit End Position inclusive

1. [1,2,3,5]

**alignment**

Information describing the similarity between a sequence (typically a sequencing read) and a reference sequence (for instance, a reference genome)[[2]](#footnote-2)

1. [3]

**BAM**

Compressed binary version of SAM

1. [1,2,3]

**base**

**base pair**

In the context of this document “base” is used as synonymous of “nucleotide”

1. [1,2]

**base position**

The number of bases between a base and the leftmost mapped base belonging to the same Genomic Segment

1. [1,5]

**box**

An object-oriented building unit defined by a unique type identifier and length

**chimeric read**

Aligned read which covers non-continuous portions of the reference genome due to either biological fusion processes (such as genomic rearrangements) or protocol artefacts

1. [1,2,3]

**CIGAR string**

**CIGAR**

A textual way of representing an alignment[[3]](#footnote-3)

1. [1,2,3]

**cluster**An aggregation of Genomic Records

1. [1,2]

**cluster signature**

Sequence of nucleotides that is common to most or all Genomic Records belonging to a cluster

1. [1,2]

**contig**

A set of overlapping DNA segments, sequenced and assembled, that together represent a consensus region of DNA[[4]](#footnote-4)

1. [1,5]

**container box**

Box whose sole purpose is to contain and group a set of related boxes

1. [2,3,5]

**Dataset**

Compression unit specified in ISO/IEC 23092-1 containing one or more of: reference sequences; sequence reads; and alignment information

1. [3,5]

**Dataset Group**

A collection of one or more Datasets

1. [1]

**Data Stream**

Set of Packets transporting the same data type

1. [2,3]

**deletion**

A contiguous deletion of one or more bases from a genomic sequence

1. [2]

**E-CIGAR**

**E-CIGAR string**

Extended CIGAR syntax specified in in ISO/IEC 23092-2 as a superset of the CIGAR syntax[[5]](#footnote-5)

1. [2]

**edit operation**

Modification of a sequence of nucleotides by means of a substitution, deletion, insertion, or clip

1. [2,3]

**FASTA**

GIR that includes a name and a nucleotide sequence for each sequencing read[[6]](#footnote-6)

1. [2,3]

**FASTQ**

GIR that includes FASTA and quality values

1. [1]

**File Format**

Set of data structures for the storage of coded information

1. [2]

**first end**

**end 1**

First segment of a paired-end template[[7]](#footnote-7)

1. [2,3]

**Genomic Descriptor**

**Descriptor**Element of the syntax used in this document to represent a feature of a genomic sequence read or associated information such as alignment information or quality values

1. [2,3]

**Genomic Information Representation**

**GIR**

A way to describe a sequence and some information associated with it[[8]](#footnote-8)

1. [1,5]

**genomic position**

**position**

Integer number representing the zero-based position of a nucleotide within a reference sequence

1. [1,2,3,5]

**genomic range**

**range**

Interval of positions on a reference sequence defined by a start position *s* and an end position *e* such that *s ≤ e*; the start and the end positions of a Genomic range are always included in the range

1. [2]

**Genomic Record**

**Record**

Data structure representing a tuple optionally associated with alignment information, read identifier and quality values

1. [2]

**Genomic Record index**

**Record index**

Position of a genomic record in the sequence of genomic records encoded in an Access Unit

1. [2]

**Genomic Record position**

**Record position**

0-based position of the left-most mapped base of the Genomic Record on the Reference Genome[[9]](#footnote-9)

1. [1,2,3,5]

**genomic reference**

**reference**

A collection of reference sequences. Typical examples are a reference genome or a reference transcriptome

1. [1,5]

**genomic region**

**region**

Genomic interval between a start nucleotide position and an end nucleotide position. Both start and end position must be considered as included

1. [1,2,3,5]

**genomic segment**

**segment**

Contiguous sequence of nucleotides (typically output of the sequencing process, and sequenced from one strand of a template)

1. [3]

**genomic study**

**study**

Biological experiment resulting in the production of sequencing data

1. [2,3]

**hard clip**

**hard clipped bases**

Edit operation signaling that some bases at either side of a read have been ignored when the read was aligned to the reference (typically because they did not match it well) and removed from the original sequence of the read

1. [2,3]

**indel**

When aligning two sequences, a contiguous stretch of nucleotides that are inserted into one sequence, or alternatively deleted from the other, in order to make the two sequences the same[[10]](#footnote-10)

1. [2,3]

**insertion**

A contiguous insertion of one or more bases into a genomic sequence

1. [2]

**leftmost read end**

**leftmost end**

Sequence read end 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 end in the pair

1. [1,3,5]

**mapped base**

A 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

**nucleotide**

A monomer of a nucleic acid polymer such as DNA or RNA[[11]](#footnote-11)

1. [1,5]

**Packet**

Transmission unit transporting segments of any of the data structures defined in ISO/IEC 23092-1

1. [2,3]

**paired-end reads**

**paired-end template**

A tuple made of two segments (typically the beginning and the end of a nucleic acid molecule)

1. [2]

**pileup**

Textual representation of sequence reads aligned to a reference sequence

1. [2,3]

**quality value**

**quality score**

Number assigned to each nucleotide base call during most automated sequencing processes[[12]](#footnote-12)

1. [2,3]

**read group**

A set of reads having some property in common

1. [2,3]

**read identifier**

**read header**

**read name**

A text string associated with each sequence read stored in GIRs such as FASTA, FASTQ and SAM[[13]](#footnote-13)

1. [1,2,3,5]

**reference genome**

A representative example of the sequences for a species’ genetic material (i.e. of the sequences of the DNA molecules present in a typical cell of that species)

1. [3]

**reference transcriptome**

A representative example of the sequences for a species’ transcriptome (i.e. of the sequences of the RNA molecules present in a typical cell of that species)

1. [1,2,3,5]

**reference sequence**

Nucleic acid sequence with biological relevance[[14]](#footnote-14)

1. [2]

**rightmost read end**

**rightmost end**

Sequence 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

1. [2,3]

**SAM**

GIR that is human readable and includes FASTQ + alignment and analysis information[[15]](#footnote-15)

1. [2]

**second end**

**end 2**

Second segment of a paired-end template[[16]](#footnote-16)

1. [1,2,3,5]

**sequence read**

**read**

The readout, by a specific technology more or less prone to errors, of a continuous part of a nucleic acid molecule extracted from an organic sample

1. [2,3]

**single-end read**

Tuple made of one segment

1. [2,3]

**soft clip**

**soft clipped bases**

Edit operation signaling that some bases at either side of a read have been ignored when the read was aligned to the reference (typically because they did not match it well) but they were not removed from the original sequence of the read

1. [2]

**spliced read**

Aligned read resulting from RNA sequencing which covers non-continuous portions of the reference genome being the result of biological splicing (i.e. containing at least one junction between two consecutive exons)

1. [1,5]

**syntax field**

Element of data represented in the data format

**substitution**

An edit operation by which a base is replaced by another one having different sequence

1. [1,3,5]

**template**

Genomic sequence that is produced by a sequencing machine as a single unit[[17]](#footnote-17)

1. [2,3]

tuple

Collection of one or more segments. Each segment can be: unmapped; mapped once; or mapped more than once

1. [1,5]

**transport format**

Set of data structures for the transport of coded information

1. [1,5]

**variable**

Parameter either inferred from syntax fields or locally defined in a process description

1. An Access Unit contains Genomic Records belonging to the same Data Class [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)
3. Several definitions have been used by different programs, the ones referred to here is the one used in SAM. 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. [↑](#footnote-ref-3)
4. From “contiguous”. [↑](#footnote-ref-4)
5. Among other things E-CIGAR enables the unambiguous representation of substitutions, spliced reads and splice strandedness. [↑](#footnote-ref-5)
6. Additional information is usually encoded in the read identifier by bioinformatics tools (such as database information by BLAST, and base calling information by Illumina software). [↑](#footnote-ref-6)
7. 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. [↑](#footnote-ref-7)
8. Which information is represented varies depending on the GIR. [↑](#footnote-ref-8)
9. A base present in the aligned read and not present in the reference sequence (a.k.a. insertion) and bases preserved by the alignment process but not mapped on the reference sequence (a.k.a. soft clip) do not have mapping positions. [↑](#footnote-ref-9)
10. From “insertion or deletion”. [↑](#footnote-ref-10)
11. 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’. [↑](#footnote-ref-11)
12. 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. [↑](#footnote-ref-12)
13. The read identifier is usually unique within its Dataset, and may contain additional information as encoded by bioinformatics tools (such as database information by BLAST, and base calling information by Illumina software). [↑](#footnote-ref-13)
14. 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. [↑](#footnote-ref-14)
15. From “Sequence Alignment/Map format”. SAM originates from the 1000 Genome Sequencing Project (it was originally de facto defined as the output of the bwa aligner). It is represented in plain ASCII, extensible by users and includes sequence, quality, alignment and analysis information. Due to the presence of arbitrary additional fields its semantics is not fully specified. [↑](#footnote-ref-15)
16. 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. [↑](#footnote-ref-16)
17. A template can be made of one or more *segments* (being called *single-end* sequencing *read* when it only has one segment, and *paired-end* sequencing *read* when it has two segments – typically they capture both the beginning and the end of a nucleic acid molecule). [↑](#footnote-ref-17)