

DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 23092-2

ISO/IEC JTC 1/SC 29

Secretariat: JISC

Voting begins on:
2020-01-17

Voting terminates on:
2020-04-10

Information technology — Genomic information representation —

Part 2: Coding of genomic information

*Technologies de l'information — Représentation des informations génomiques —
Partie 2: Codage des informations génomiques*

ICS: 35.040.99

iTeh STANDARD PREVIEW
(standards.iteh.ai)
Full standard:
<https://standards.iteh.ai/catalog/standards/sist/4d3aa30b-e67b-4e1b-b035-96c37303eaaa/iso-iec-dis-23092-2>

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.



Reference number
ISO/IEC DIS 23092-2:2020(E)

© ISO/IEC 2020

iTeh STANDARD PREVIEW
(standards.iteh.ai)
Full standard:
<https://standards.iteh.ai/catalog/standards/sist/4d3aa30b-e67b-4e1b-b035-96c37303eaaa/iso-iec-dis-23092-2>



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2020

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

1		
2	Contents	
3	Foreword	ix
4	Introduction	x
5	1 Scope	1
6	2 Normative references	1
7	3 Terms and definitions	1
8	4 Abbreviations	6
9	5 Conventions	6
10	5.1 General	6
11	5.2 Arithmetic operators	6
12	5.3 Logical operators	6
13	5.4 Relational operators	7
14	5.5 Bit-wise operators	7
15	5.6 Assignment operators	7
16	5.7 Range notation	8
17	5.8 Mathematical functions	8
18	5.9 Order of operation precedence	8
19	5.10 Variables, syntax elements and tables	9
20	5.11 Text description of logical operators	10
21	5.12 Processes	11
22	6 Syntax and semantics	12
23	6.1 Method of specifying syntax in tabular form	12
24	6.2 Bit ordering	13
25	6.3 Specification of syntax functions and data types	13
26	6.4 Semantics	14
27	7 Data structures	14
28	7.1 Data unit	14
29	7.2 Raw reference	15
30	7.2.1 Syntax and semantics	16
31	7.3 Parameter set	16
32	7.3.1 Syntax and semantics	16
33	7.3.2 Encoding parameters	16
34	7.4 Access unit	23
35	7.4.1 Syntax and semantics	23
36	7.4.2 Access unit types	26
37	8 Descriptors	27
38	9 Sequencing reads	30
39	9.1 Supported symbols	30
40	9.2 Paired-end reads	32
41	9.3 Reverse-complement reads	32
42	9.4 Data classes	32
43	9.5 Aligned data	33
44	9.6 Unaligned data	34
45	10 Decoding process	35
46	10.1 General	35
47	10.2 dataset_type = 0 or 1	35
48	10.2.1 References padding	35
49	10.2.2 Type 1 AU (Class P)	36
50	10.2.3 Type 2 AU (Class N)	37
51	10.2.4 Type 3 AU (Class M)	37

1	10.2.5 Type 4 AU (Class I)	38
2	10.2.6 Type 5 AU (Class HM)	40
3	10.2.7 Type 6 AU (Class U)	40
4	10.3 dataset_type = 2	41
5	10.3.1 Type 1 AU	41
6	10.3.2 Type 2 AU	42
7	10.3.3 Type 3 AU	42
8	10.3.4 Type 4 AU	42
9	10.3.5 Type 6 AU	43
10	10.4 Genomic descriptors	43
11	10.4.1 pos	43
12	10.4.2 rcomp	44
13	10.4.3 flags	45
14	10.4.4 mmpos	45
15	10.4.5 mmtyp	48
16	10.4.6 clips	52
17	10.4.7 ureads	55
18	10.4.8 rlen	55
19	10.4.9 pair	57
20	10.4.10 mscore	65
21	10.4.11 mmap	66
22	10.4.12 msar	69
23	10.4.13 rtype	70
24	10.4.14 rgroup	71
25	10.4.15 qv	72
26	10.4.16 rname	76
27	10.4.17 rftp	76
28	10.4.18 rftt	77
29	10.4.19 tokentype descriptors	78
30	10.5 sequence	86
31	10.5.1 Aligned reads (Classes P, N, M, I, HM)	86
32	10.5.2 Unmapped reads (Class HM, U)	87
33	10.6 e-cigar	88
34	10.6.1 Syntax	88
35	10.6.2 Decoding process for the first alignment	89
36	10.6.3 Decoding process for other alignments	96
37	10.6.4 Reference transformation	96
38	11 Representation of reference sequences	98
39	11.1 External reference	98
40	11.2 Embedded reference	98
41	11.3 Computed reference	98
42	11.3.1 General	99
43	11.3.2 Reference transformation	99
44	11.3.3 PushIn	100
45	11.3.4 Local assembly	101
46	11.3.5 Global assembly	102
47	12 Block payload parsing process	102
48	12.1 General	102
49	12.2 Inverse binarizations	103
50	12.2.1 Binary (BI)	104
51	12.2.2 Truncated Unary (TU)	104
52	12.2.3 Exponential Golomb (EG)	104
53	12.2.4 Truncated Exponential Golomb (TEG)	105
54	12.2.5 Signed Truncated Exponential Golomb (STEG)	105
55	12.2.6 Split Unit-wise Truncated Unary (SUTU)	105

1	12.2.7 Signed Split Unit-wise Truncated Unary (SSUTU)	106
2	12.2.8 Double Truncated Unary (DTU)	106
3	12.2.9 Signed Double Truncated Unary (SDTU)	107
4	12.3 Decoder Configuration	107
5	12.3.1 Sequences and quality values	107
6	12.3.2 Support values	108
7	12.3.3 CABAC binarizations	109
8	12.3.4 Transformation parameters	111
9	12.3.5 Msar descriptor and read identifiers	113
10	12.3.6 State variables	114
11	12.4 Initialization process for context variables	117
12	12.5 Arithmetic decoding engine	117
13	12.5.1 Initialization	117
14	12.5.2 Arithmetic decoding process	118
15	12.6 Decoding process for sequence descriptors	125
16	12.6.1 General	125
17	12.6.2 Block payload decoding process	126
18	13 Output format	139
19	13.1 General	139
20	13.2 MPEG-G record	139
21	13.2.1 number_of_template_segments	141
22	13.2.2 number_of_record_segments	141
23	13.2.3 number_of_alignments	141
24	13.2.4 class_ID	141
25	13.2.5 read_group_len	141
26	13.2.6 reserved	141
27	13.2.7 read_1_first	141
28	13.2.8 seq_ID	142
29	13.2.9 as_depth	142
30	13.2.10 read_len	142
31	13.2.11 qv_depth	142
32	13.2.12 read_name_len	142
33	13.2.13 read_name	142
34	13.2.14 read_group	142
35	13.2.15 sequence	142
36	13.2.16 quality_values	142
37	13.2.17 mapping_pos	142
38	13.2.18 ecigar_len	142
39	13.2.19 ecigar_string	142
40	13.2.20 reverse_comp	142
41	13.2.21 mapping_score	143
42	13.2.22 split_alignment	143
43	13.2.23 delta	143
44	13.2.24 split_pos	143
45	13.2.25 split_seq_ID	143
46	13.2.26 flags	143
47	13.2.27 more_alignments	143
48	13.2.28 next_pos	143
49	13.2.29 next_seq_ID	143
50	13.3 Initialization process	143
51	Annex A (informative) Tokenization of reads identifiers	147
52	Annex B (informative) Mapping quality	149
53	Annex C (informative) Inverse binarization examples	150
54	C.1 Binary (BI) binarization	150

1	C.2 Truncated Unary (TU) binarization.....	150
2	C.3 Exponential Golomb (EG) binarization.....	150
3	C.4 Truncated Exponential Golomb (TEG) binarization.....	151
4	C.5 Signed Truncated Exponential Golomb (STEG) Binarization.....	151
5	C.6 Split Unit-wise Truncated Unary (SUTU) binarization	151
6	C.7 Signed Split Unit-wise Truncated Unary (SSUTU) binarization.....	152
7	C.8 Double Truncated Unary (DTU) Binarization.....	152
8	C.9 Signed Double Truncated Unary (SDTU) Binarization	152
9		

iTeh STANDARD PREVIEW
(standards.iteh.ai)

Full standard:
<https://standards.iteh.ai/catalog/standards/sist/4d3aa30b-e67b-4e1b-b035-96c37303eaaa/iso-iec-dis-23092-2>

1 Foreword

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

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

13 Attention is drawn to the possibility that some of the elements of this specification may be the subject of
14 patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details
15 of any patent rights identified during the development of the document will be in the Introduction and/or
16 on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent
17 declarations received (see <http://patents.iec.ch>).

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

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

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

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

27 Any feedback or questions on this specification should be directed to the user's national standards body.
28 A complete listing of these bodies can be found at www.iso.org/members.html.

29

1 Introduction

2 The advent of high-throughput sequencing (HTS) technologies has the potential to boost the adoption of
3 genomic information in everyday practice, ranging from biological research to personalized genomic
4 medicine in the clinic. As a consequence, an extraordinary growth in volume of generated data has been
5 recorded during the last few years, and an even more pronounced growth is expected in the near future.

6 At the moment genomic information is mostly exchanged through a variety of data formats, such as
7 FASTA/FASTQ for unaligned sequencing reads and SAM/BAM/CRAM for aligned reads. With respect to
8 such formats, the ISO/IEC 23092 series provides a new solution for the representation and compression
9 of genome sequencing information by:

- 10 • specifying an abstract representation of the sequencing data rather than a specific format with
11 its direct implementation
- 12 • being designed at a time point when technologies and use cases are more mature. This permits
13 to address one limitation of the textual SAM format, for which incremental ad-hoc addition of
14 features followed along the years, resulting in an overall redundant and suboptimal format which
15 at the same time results not general and unnecessarily complicated
- 16 • normatively separating free-field user-defined information with no clear semantics from the
17 normative genomic data representation. This allows a fully interoperable and automatic
18 exchange of information between different data producers
- 19 • allowing multiplexing of relevant meta-data information with the data since data and meta-data
20 are partitioned at different conceptual levels
- 21 • following a strict and supervised development process which has proven successful in the last
22 30 years in the domain of digital media for the transport format, the file format, the compressed
23 representation and the application program interfaces

24
25 The ISO/IEC 23092 series provides the enabling technology that will allow the community to create an
26 ecosystem of novel, interoperable, solutions in the field of genomic information processing. In particular
27 it offers:

- 28 • Consistent, general and properly designed format definitions and data structures to store
29 sequencing and alignment information. A robust framework which can be used as a foundation
30 to implement different compression algorithms
- 31 • Speed and flexibility in the selective access to coded data, by means of newly designed data
32 clustering and optimized storage methodologies
- 33 • Low latency in data transmission and consequent fast availability at remote locations, based on
34 transmission protocols inspired by real-time application domains
- 35 • Built-in privacy and protection of sensitive information, thanks to a flexible framework which
36 allows customizable secured access at all layers of the data hierarchy
- 37 • Reliability of the technology and interoperability among tools and systems, owing to the provision
38 of a normative procedure to assess conformance to the standard on an exhaustive dataset
- 39 • Support to the implementation of a complete ecosystem of compliant devices and applications,
40 through the availability of a normative reference implementation covering the totality of the
41 specification.

42
43 The fundamental structure of the ISO/IEC 23092 series data representation is the *genomic record*. The
44 genomic record is a data structure consisting of either a single sequencing read, or a paired sequencing
45 read, and its associated sequencing and alignment information; it may contain detailed mapping and
46 alignment data, a single or paired read identifier (read name) and quality values.

47 Without breaking traditional approaches, the genomic record introduced in the ISO/IEC 23092 series
48 provides a more compact, simpler and manageable data structure grouping all the information related to
49 a single DNA template, from simple sequencing data to sophisticated alignment information.

50 Although it is an appropriate logic data structure for interaction and manipulation of coded information,
51 the genomic record is not a suitable atomic data structure for compression. To achieve high compression
52 ratios, it is necessary to group genomic records into clusters and to transform the information of the same

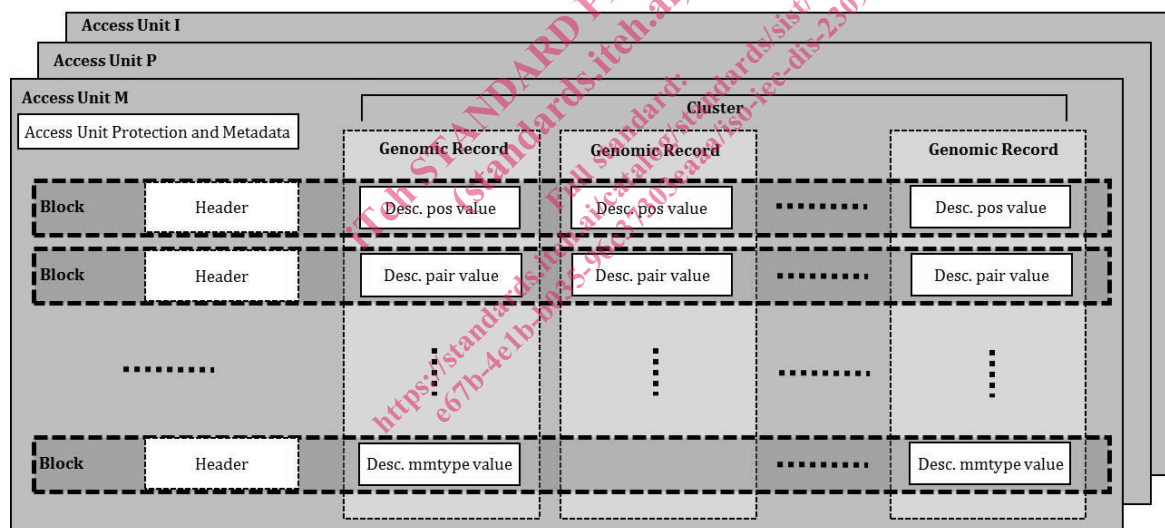
1 type into sets of descriptors structured into homogeneous blocks. Furthermore, when dealing with
 2 selective data access, the genomic record unit is too small to allow effective and fast information retrieval.

3 For these reasons, this document introduces the concept of access unit, which is the fundamental
 4 structure for coding and access to information in the compressed domain.

5 The access unit is the smallest data structure that can be decoded by a decoder compliant with
 6 ISO/IEC23092-2. An access unit is composed of one block for each descriptor used to represent the
 7 information of its genomic records; therefore, a block payload is the coded representation of all the data
 8 of the same type (i.e. a descriptor) in a cluster.

9 In addition to clusters of genomic records compressed into access units, reads are further classified in six
 10 data classes: five classes are defined according to the result of their alignment against one or more
 11 reference sequences; the sixth class contains either reads that could not be mapped or raw sequencing
 12 data. The classification of sequencing reads into classes enables to develop powerful selective data access.
 13 In fact access units inherit a specific data characterization (e.g. perfect matches in class P, substitutions
 14 in class M, indels in class I, half-mapped reads in class HM) from the genomic records composing them,
 15 and thus constitute a data structure capable of providing powerful filtering capability for the efficient
 16 support of many different use cases.

17 Access units are the fundamental, finest grain data structure in terms of content protection and in terms
 18 of metadata association. In other words each access unit can be protected individually and independently.
 19 **Figure 1** shows how access units, blocks and genomic records relate to each other in the ISO/IEC 23092
 20 series data structure.



21
 22 **Figure 1 –Access units, blocks and genomic records**
 23
 24

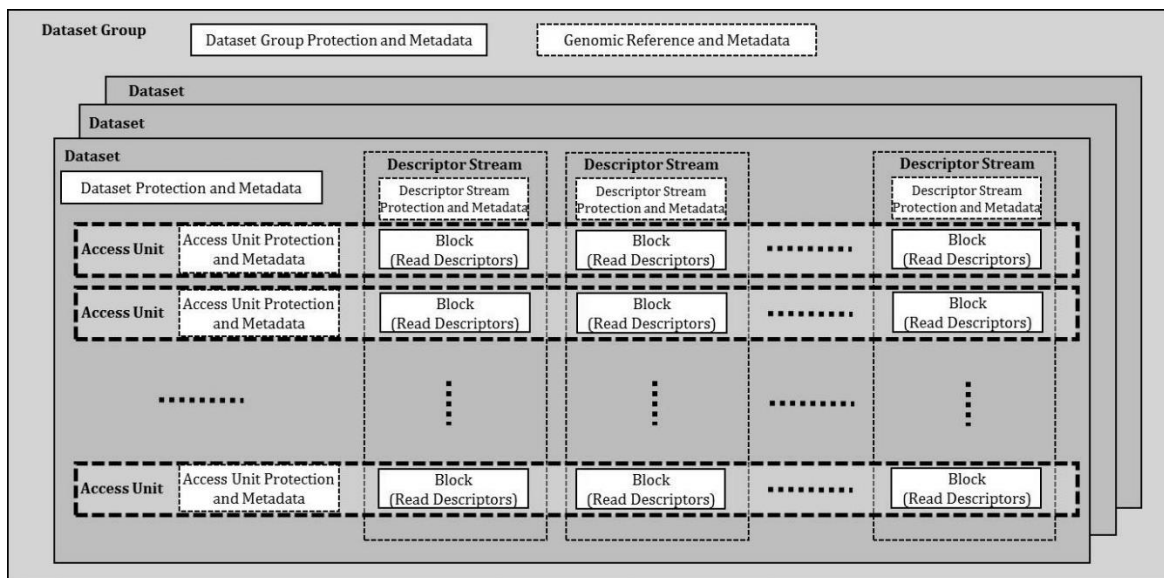


Figure 2 – The high-level data structure: datasets and dataset group

A dataset is a coded data structure containing headers and one or more access units. Typical datasets could for example contain the complete sequencing of an individual, or a portion of it. Other datasets could contain for example a reference genome or a subset of its chromosomes. Datasets are grouped in dataset groups, as shown in Figure 2.

According to the ISO/IEC 23092 series, the compressed sequencing data can be multiplexed into a normative bitstream suitable to packetization for real-time transport over typical network protocols. In storage use cases coded data can be encapsulated into a file format with the possibility to organize blocks per descriptor stream or per access unit, to further optimize the selective access performance to the type of data access required by the different application scenarios. The ISO/IEC 23092 series further provides a reference process to convert a normative transport stream into a normative file format and vice versa.

The ISO/IEC 23092-2 series define the syntax and semantics of the compressed genome sequencing data representation and the deterministic decoding process that reconstructs the contents of datasets. The decoding process is fully specified such that all decoders that conform to ISO/IEC 23092-2 will produce identical decoded output. A simplified diagram of the decoding process is shown in Figure 3.

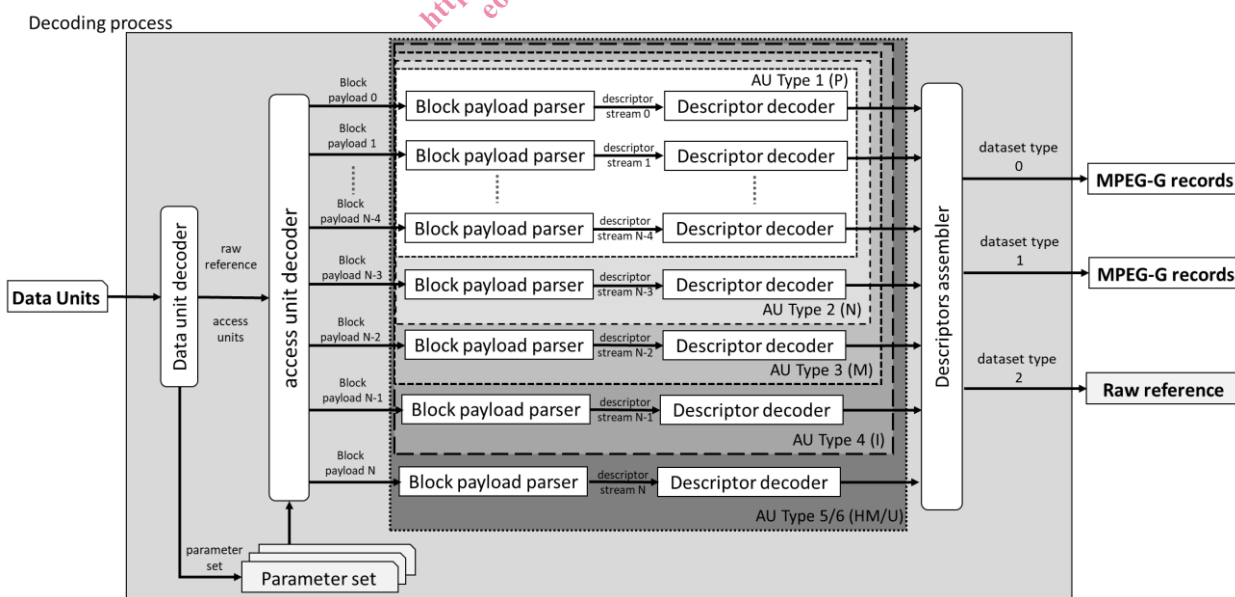


Figure 3 – The decoding process

1 Information Technology — ISO/IEC 23092 — Part 2: Coding of 2 Genomic Information

3 1 Scope

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

- 6 • unaligned sequencing reads including read identifiers and quality values
- 7 • aligned sequencing reads including read identifiers and quality values
- 8 • reference sequences

9 2 Normative references

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

13 ISO/IEC 10646, *Information technology — Universal Coded Character Set (UCS)*

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

16 3 Terms and definitions

17 For the purposes of this document, the terms and definitions given in ISO/IEC 23092-1 and the following
18 apply.

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

- 20 — ISO Online browsing platform: available at <https://www.iso.org/obp>
- 21 — IEC Electropedia: available at <http://www.electropedia.org/>

22 3.1

23 alignment

24 information describing the similarity between a sequence [typically a *sequencing read* (3.28)] and a
25 reference sequence (for instance, a reference genome)

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

29 3.2

30 CIGAR string

31 CIGAR

32 textual way of representing an *alignment* (3.1)

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

36 3.3

37 dataset

38 compression unit containing one or more of: reference sequences; *sequencing reads* (3.28); and *alignment*
39 (3.1) information

40 Note 1 to entry: Datasets shall be as specified in ISO/IEC 23092-1.

1 **3.4**2 **deletion**

3 contiguous removal of one or more bases from a genomic sequence

4 **3.5**5 **E-CIGAR**

6 extended CIGAR syntax specified as a superset of the CIGAR syntax

7 Note 1 to entry: Among other things, E-CIGAR enables the unambiguous representation of substitutions, spliced
8 reads and splice strandedness.9 **3.6**10 **edit operation**11 modification of a sequence of *nucleotides* (3.20) by means of a substitution, *deletion* (3.4), *insertion* (3.18)
12 or clip13 **3.7**14 **FASTA**15 GIR that includes a name and a *nucleotide* (3.20) sequence for each *sequencing read* (3.28)16 Note 1 to entry: Additional information is usually encoded in the read identifier by bioinformatics tools (such as
17 database information, and base calling information).18 **3.8**19 **FASTQ**20 GIR that includes *FASTA* (3.7) and *quality values* (3.22)21 **3.9**22 **first end**

23 end 1

24 read 1

25 first segment of a paired-end *template* (3.33)26 Note 1 to entry: Illumina platforms usually store first and second ends in two separate files and in the same order
27 — 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.28 **3.10**29 **genomic descriptor**

30 descriptor

31 element of the syntax used to represent a feature of a genomic *sequencing read* (3.28) or associated
32 information such as *alignment* (3.1) information or *quality values* (3.22)33 **3.11**34 **genomic information representation**

35 way to describe a sequence and some information associated with it

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

37 **3.12**38 **genomic record**

39 record

40 data structure representing a *tuple* (3.34) optionally associated with *alignment* (3.1) information, *read*
41 *identifier* (3.24) and *quality values* (3.22)42 **3.13**43 **genomic record index**44 position of a genomic record in the sequence of *genomic records* (3.12) encoded in an access unit45 **3.14**46 **genomic record position**47 0-based position of the leftmost mapped base on the reference genome of the first *alignment* (3.1)
48 contained in a *genomic record* (3.12)

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

4 **3.15**

5 **genomic reference**

6 reference

7 collection of reference sequences

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

9 **3.16**

10 **hard clip**

11 base or set of bases originally present at either side of a read, and removed from it following *alignment*
 12 (3.1)

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

14 **3.17**

15 **indel**

16 contiguous stretch of *nucleotides* (3.20) that, when aligning two sequences, are inserted into one
 17 sequence, or alternatively deleted from the other, in order to make the two sequences the same

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

19 **3.18**

20 **insertion**

21 contiguous addition of one or more bases into a genomic sequence

22 **3.19**

23 **leftmost read end**

24 leftmost read

25 *sequencing read* (3.28) generated by a paired-end sequencing run and mapped at a position on the
 26 reference sequence which is smaller than the mapping position of the other read in the pair

27 **3.20**

28 **nucleotide**

29 base

30 base pair

31 monomer of a nucleic acid polymer such as DNA or RNA

32 Note 1 to entry: Nucleotides are denoted as letters ('A' for adenine; 'C' for cytosine; 'G' for guanine; 'T' for thymine
 33 which only occurs in DNA; and 'U' for uracil which only occurs in RNA). The chemical formula for a specific DNA or
 34 RNA molecule is given by the sequence of its nucleotides, which can be represented as a string over the alphabet
 35 ('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
 36 unknown molecular composition are denoted with 'N'.

37 **3.21**

38 **paired-end read**

39 paired-end template

40 *tuple* (3.34) made of two segments

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

42 **3.22**

43 **quality value**

44 quality score

45 number assigned to each *nucleotide* (3.20) base call in automated sequencing processes

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