
**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

iteh Standards
(<https://standards.iteh.ai>)
Document Preview

ISO/IEC 23092-2:2019

<https://standards.iteh.ai/catalog/standards/iso/4afcfac0-4bd4-4332-8a90-8f0575e864f6/iso-iec-23092-2-2019>



iTeh Standards
(<https://standards.iteh.ai>)
Document Preview

ISO/IEC 23092-2:2019

<https://standards.iteh.ai/catalog/standards/iso/4afcfac0-4bd4-4332-8a90-8f0575e864f6/iso-iec-23092-2-2019>



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2019

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

Contents

Page

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

10.2.6	Type 5 AU (Class HM)	40
10.2.7	Type 6 AU (Class U)	40
10.3	dataset_type = 2	40
10.3.1	Type 1 AU	41
10.3.2	Type 2 AU	42
10.3.3	Type 3 AU	42
10.3.4	Type 4 AU	42
10.3.5	Type 6 AU	42
10.4	Genomic descriptors	43
10.4.1	pos	43
10.4.2	rcomp	44
10.4.3	flags	44
10.4.4	mmpos	45
10.4.5	mmttype	47
10.4.6	clips	50
10.4.7	ureads	53
10.4.8	rln	53
10.4.9	pair	55
10.4.10	mscore	62
10.4.11	mmap	63
10.4.12	msar	66
10.4.13	rtype	66
10.4.14	rgroup	68
10.4.15	qv	68
10.4.16	rname	72
10.4.17	rftt	72
10.4.18	rftt	73
10.4.19	tokentype descriptors	73
10.5	sequence	81
10.5.1	Aligned reads (Classes P, N, M, I, HM)	82
10.5.2	Unmapped reads (Class HM, U)	83
10.6	e-cigar	83
10.6.1	Syntax	83
10.6.2	Decoding process for the first alignment	84
10.6.3	Decoding process for other alignments	92
10.6.4	Reference transformation	92
11	Representation of reference sequences	93
11.1	External reference	94
11.2	Embedded reference	94
11.3	Computed reference	94
11.3.1	General	94
11.3.2	Reference transformation	94
11.3.3	PushIn	95
11.3.4	Local assembly	96
11.3.5	Global assembly	97
12	Block payload parsing process	97
12.1	General	97
12.2	Inverse binarizations	98
12.2.1	Binary (BI)	99
12.2.2	Truncated Unary (TU)	99
12.2.3	Exponential Golomb (EG)	99
12.2.4	Truncated Exponential Golomb (TEG)	100
12.2.5	Signed Truncated Exponential Golomb (STEG)	100
12.2.6	Split Unit-wise Truncated Unary (SUTU)	101
12.2.7	Signed Split Unit-wise Truncated Unary (SSUTU)	101
12.2.8	Double Truncated Unary (DTU)	101
12.2.9	Signed Double Truncated Unary (SDTU)	102

12.3	Decoder configuration.....	102
12.3.1	Sequences and quality values.....	102
12.3.2	Support values.....	103
12.3.3	CABAC binarizations.....	104
12.3.4	Transformation parameters.....	107
12.3.5	Msar descriptor and read identifiers.....	108
12.3.6	State variables.....	109
12.4	Initialization process for context variables.....	112
12.5	Arithmetic decoding engine.....	112
12.5.1	Initialization.....	112
12.5.2	Arithmetic decoding process.....	113
12.6	Decoding process for sequence descriptors.....	120
12.6.1	General.....	120
12.6.2	Block payload decoding process.....	121
13	Output format.....	135
13.1	General.....	135
13.2	MPEG-G record.....	135
13.2.1	number_of_template_segments.....	137
13.2.2	number_of_record_segments.....	137
13.2.3	number_of_alignments.....	137
13.2.4	class_ID.....	137
13.2.5	read_group_len.....	138
13.2.6	read_1_first.....	138
13.2.7	seq_ID.....	138
13.2.8	as_depth.....	138
13.2.9	read_len.....	138
13.2.10	qv_depth.....	138
13.2.11	read_name_len.....	138
13.2.12	read_name.....	138
13.2.13	read_group.....	138
13.2.14	sequence.....	139
13.2.15	quality_values.....	139
13.2.16	mapping_pos.....	139
13.2.17	ecigar_len.....	139
13.2.18	ecigar_string.....	139
13.2.19	reverse_comp.....	139
13.2.20	mapping_score.....	139
13.2.21	split_alignment.....	139
13.2.22	delta.....	140
13.2.23	split_pos.....	140
13.2.24	split_seq_ID.....	140
13.2.25	flags.....	140
13.2.26	more_alignments.....	140
13.2.27	next_pos.....	140
13.2.28	next_seq_ID.....	140
13.3	Initialization process.....	140
Annex A (informative) Tokenization of reads identifiers.....		143
Annex B (informative) Mapping quality.....		145
Annex C (informative) Inverse binarization examples.....		146

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).

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

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.

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 website.

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.

Introduction

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

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

- Specifying an abstract representation of the sequencing data rather than a specific format with its direct implementation.
- Being designed at a time point when technologies and use cases are more mature. This permits the addressing of one limitation of the textual SAM format, for which incremental ad-hoc addition of features followed along the years, resulting in an overall redundant and suboptimal format which at the same time results not general and unnecessarily complicated.
- Normatively separating free-field user-defined information with no clear semantics from the normative genomic data representation. This allows a fully interoperable and automatic exchange of information between different data producers.
- Allowing multiplexing of relevant metadata information with the data since data and metadata are partitioned at different conceptual levels.
- Following a strict and supervised development process which has proven successful in the last 30 years in the domain of digital media for the transport format, the file format, the compressed representation and the application program interfaces.

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

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

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

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

The genomic record, although it is an appropriate logic data structure for interaction and manipulation of coded information, is not a suitable atomic data structure for compression. To achieve high compression ratios, it is necessary to group genomic records into clusters and to transform the information of the same type into sets of descriptors structured into homogeneous blocks. Furthermore, when dealing with selective data access, the genomic record is a too small unit to allow effective and fast information retrieval.

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

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

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

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

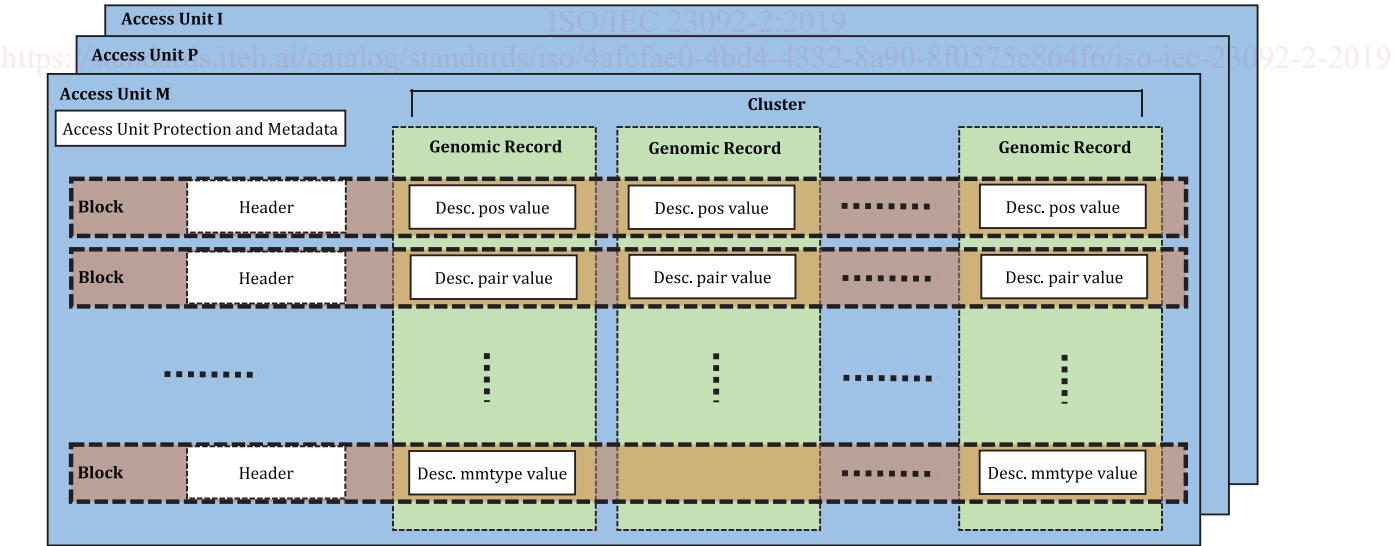


Figure 1 — Access units, blocks and genomic records

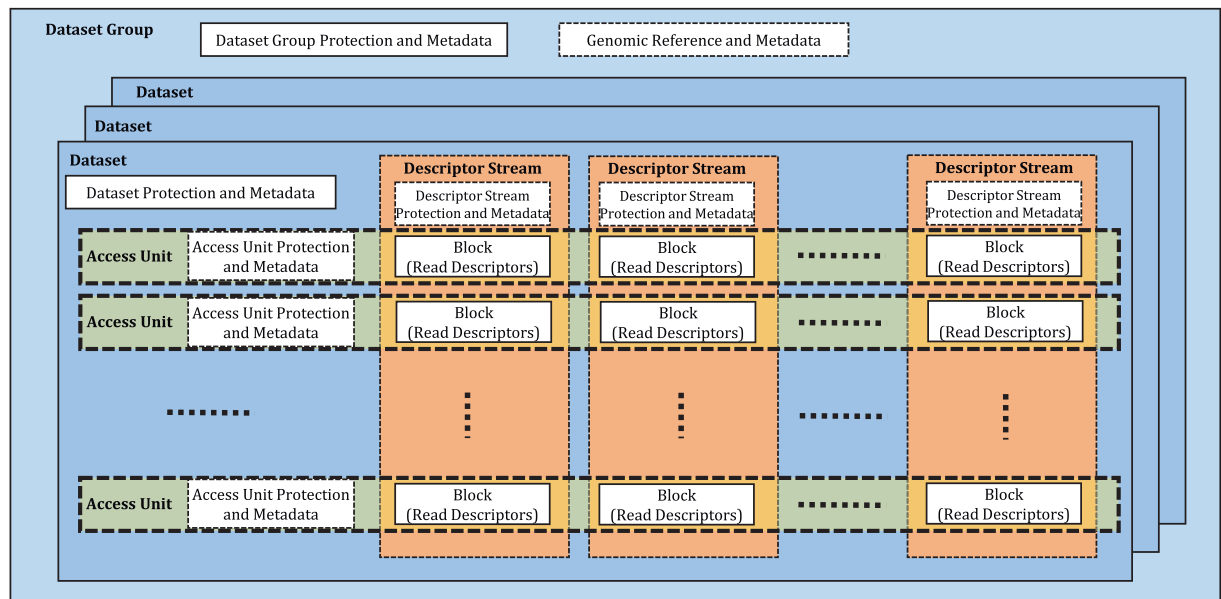


Figure 2 — High-level data structure: datasets and dataset group

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 for 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.

This document defines 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 this document will produce identical decoded output. A simplified diagram of the decoding process is shown in [Figure 3](#).

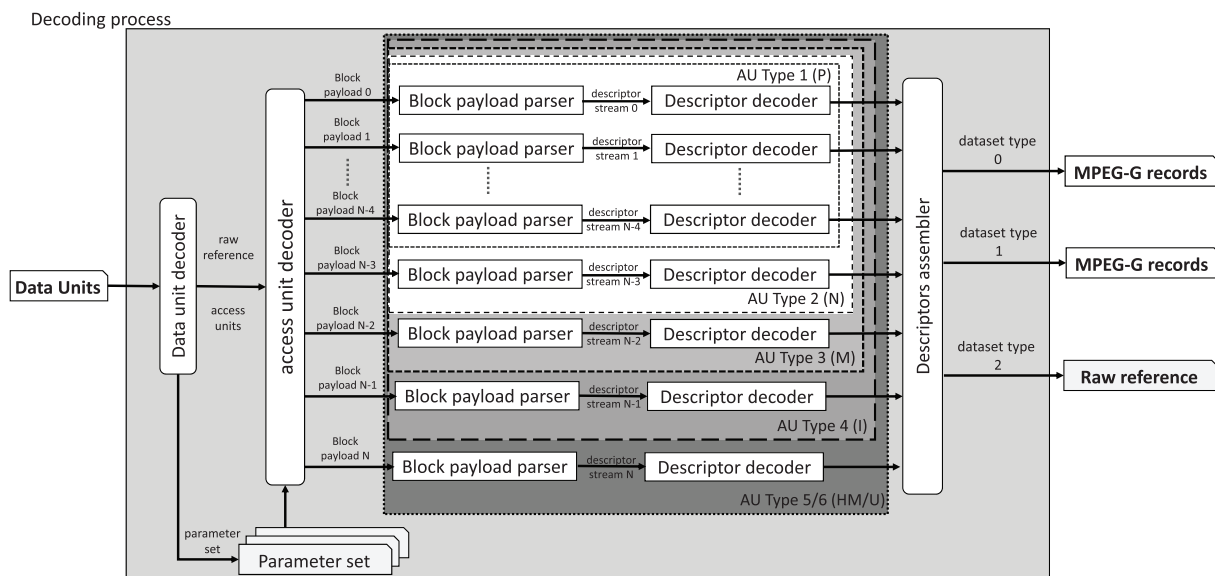


Figure 3 — The decoding process

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

ISO and IEC take no position concerning the evidence, validity and scope of this patent right. The holder of this patent right has assured ISO and IEC that he/she is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO and IEC. Information may be obtained from:

GenomSys SA
EPFL Innovation Park Building C
CH-1015 Lausanne
Switzerland
info@genomsys.com

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified above. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Information technology — Genomic information representation —

Part 2: Coding of genomic information

1 Scope

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

- unaligned sequencing reads including read identifiers and quality values;
- aligned sequencing reads including read identifiers and quality values;
- reference sequences.

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 document. 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 23092-1, *Information technology — Genomic information representation — Part 1: Transport and storage of genomic information*

<https://standards.iteh.ai/catalog/standards/iso/4afcfac0-4bd4-4332-8a90-8f0575e864f6/iso-iec-23092-2-2019>

3 Terms and definitions

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

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

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

3.1 alignment

information describing the similarity between a sequence [typically a *sequencing read* (3.28)] 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.

3.2

CIGAR string

CIGAR

textual way of representing an *alignment* ([3.1](#))

Note 1 to entry: Several definitions have been used by different programs; the one 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.

3.3

dataset

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

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

3.4

deletion

contiguous removal of one or more bases from a genomic sequence

3.5

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.

3.6

edit operation

modification of a sequence of *nucleotides* ([3.20](#)) by means of a substitution, *deletion* ([3.4](#)), *insertion* ([3.18](#)) or clip

3.7

FASTA

GIR that includes a name and a *nucleotide* ([3.20](#)) sequence for each *sequencing read* ([3.28](#))

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

3.8

FASTQ

GIR that includes *FASTA* ([3.7](#)) and *quality values* ([3.22](#))

3.9

first end

end 1

read 1

first segment of a paired-end *template* ([3.33](#))

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.

3.10

genomic descriptor

descriptor

element of the syntax used to represent a feature of a genomic *sequencing read* ([3.28](#)) or associated information such as *alignment* ([3.1](#)) information or *quality values* ([3.22](#))

3.11**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.

3.12**genomic record**

record

data structure representing a *tuple* (3.34) optionally associated with *alignment* (3.1) information, *read identifier* (3.24) and *quality values* (3.22)

3.13**genomic record index**

position of a genomic record in the sequence of *genomic records* (3.12) encoded in an access unit

3.14**genomic record position**

0-based position of the leftmost mapped base on the reference genome of the first *alignment* (3.1) contained in a *genomic record* (3.12)

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.

3.15**genomic reference**

reference

collection of reference sequences

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

3.16**hard clip**

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

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

3.17**indel**

contiguous stretch of *nucleotides* (3.20) 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”.

3.18**insertion**

contiguous addition of one or more bases into a genomic sequence

3.19**leftmost read end**

leftmost read

sequencing read (3.28) 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

3.20

nucleotide

base

base pair

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'.

3.21

paired-end read

paired-end template

tuple (3.34) made of two segments

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

3.22

quality value

quality score

number assigned to each *nucleotide* (3.20) 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.

3.23

read group

set of reads having some property in common

3.24

read identifier

read header

read name

text string associated with each *sequencing read* (3.28) stored in GIRs such as *FASTA* (3.7), *FASTQ* (3.8) and *SAM* (3.26)

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).

3.25

rightmost read end

rightmost read

sequencing read (3.28) 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

3.26

SAM

GIR that is human readable and includes *FASTQ* plus *alignment* (3.1) and analysis information

Note 1 to entry: From "Sequence Alignment/Map format". SAM originates from the 1000 Genome Sequencing Project. It is represented in plain ASCII, extensible by users and includes sequence, quality, alignment and analysis information.