TECHNICAL SPECIFICATION

ISO/TS 23357

First edition
2023-07

Genomics informatics — Clinical genomics data sharing specification for next-generation sequencing

Informatique génomique — Spécification du partage des données de génomique clinique pour le séquençage de nouvelle génération

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

Reference number
ISO/TS 23357:2023(E)

© ISO 2023

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

COPYRIGHT PROTECTED DOCUMENT

© ISO 2023

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
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

© ISO 2023 – All rights reserved

Contents

Page

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes 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 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. ISO 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.

This document was prepared by Technical Committee ISO/TC 215, *Health informatics*, Subcommittee SC 1, *Genomics informatics*.

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.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

Introduction

Owing to the rapid advancement of next-generation sequencing technologies, the human genome is being adopted in clinical settings to realize precision medicine[7]. Massive parallel sequencing or next-generation sequencing (NGS) is any of several high-throughput approaches to DNA sequencing using the concept of massively parallel processing. These technologies use miniaturized and parallelized platforms for sequencing of 1 million to 43 billion short reads (50-400 bases each) per instrument run. The data obtained in a clinical setting should be shared with another institution when patients move or shared with the patient if requested.

The clinical application steps based on clinical sequence information consist of:

a) specimen collection, processing and storage;

b) DNA extraction;

c) DNA processing and library preparation;

d) generation of sequence reads and base calling;

e) sequencing alignment/mapping;

f) variant calling;

g) variant annotation and filtering;

h) variant evaluation and assertion;

i) generation of test report[8].

It is required to share clinical sequencing information at a level that can reproduce the results of the institution that obtained the initial clinical sequencing information. In addition, the shared clinical genomic sequencing data should be interoperable.

This document proposes a data specification to integrate multi-layered sequencing files and related parameters and clinical data for achieving the reproducibility of genomic data in clinical practice.

This document will assist health IT companies by proposing new system requirements to deal with genomic data.

This document can be used to store and share clinical genomic data in electronic health records. In addition, it will be helpful in translational research, which requires genomic and clinical data from multiple institutes.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

Genomics informatics — Clinical genomics data sharing specification for next-generation sequencing

1 Scope

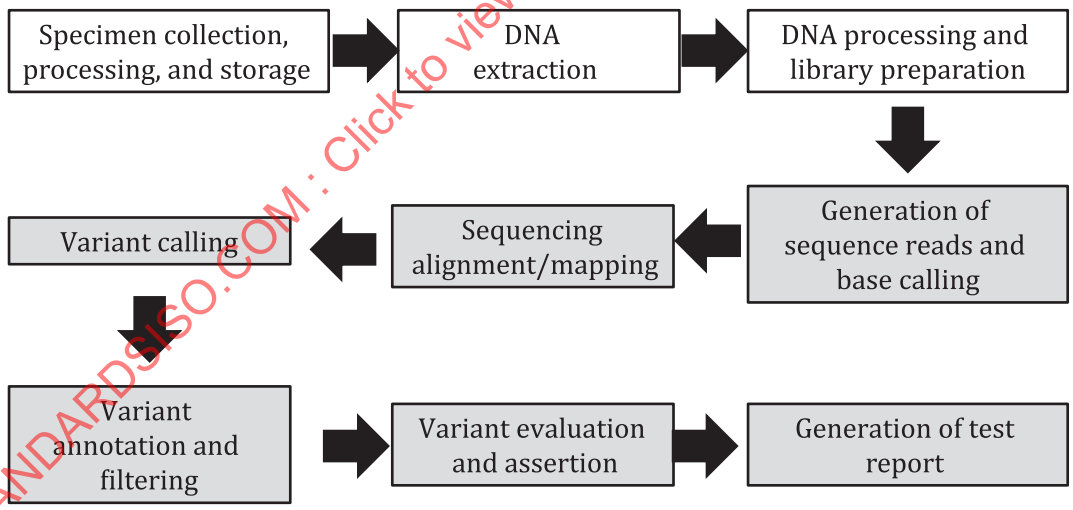
This document specifies clinical sequencing information generated by massive parallel sequencing technology for sharing health information via massively parallel sequencing. This document covers the data fields and their metadata from the generation of sequence reads and base calling to variant evaluation and assertion for archiving reproducibility during health information exchange of clinical sequence information. However, the specimen collection, processing and storage, DNA extraction and DNA processing and library preparation, and the generation of test report are not in the scope of this document.

This document hence defines the data types, relationship, optionality, cardinalities and bindings of terminology of the data.

In essence, this document specifies:

— the required data fields and their metadata from generation of sequence reads and base calling to variant evaluation and assertion for sharing clinical genomic sequencing data files generated by massively parallel sequencing technology, as shown in Figure 1;

— the sequencing information from human samples using DNA sequencing by massively parallel sequencing technologies for clinical practice.



NOTE The grey shaded text indicates the scope of this document.

Figure 1 — Clinical application processes based on next-generation sequencing (NGS) data

2 Normative references

There are no normative references in this document.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

3 Terms and definitions

For the purposes of this document, the terms and definitions given in [external document reference xxx] and the following 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
clinical sequencing
next generation sequencing or future sequencing technologies using human samples for clinical practice and clinical trials

[SOURCE: ISO/TS 20428:2017, 3.5, modified — "later" has been replaced with "future" in the definition.]

3.2
deoxyribonucleic acid
DNA
molecule that encodes the genetic information in the nucleus of cells

[SOURCE: ISO 25720:2009, 4.7]

3.3
DNA sequencing
determination of the order of nucleotide bases (adenine, guanine, cytosine, and thymine) in a molecule of *DNA* (3.4)

Note 1 to entry: Sequence is generally described from the 5′ end.

[SOURCE: ISO 17822:2020, 3.19]

3.4
exome
part of the genome that corresponds to the complete complement of the exons of a cell

3.7
FASTQ
text-based format for storing both the biological sequence (typically nucleotide sequence) and its corresponding quality scores

3.8
gene
category of nucleic acid sequences that functions as a unit of heredity and codes for the basic instructions for the development, reproduction, and maintenance of organisms

3.9
germline
series of germ cells, each descended or developed from earlier cells in the series, regarded as continuing through successive generations of an organism

[SOURCE: ISO/TS 20428:2017, 3.17]

3.10
indel
insertion (3.15) or/and *deletion* (3.7)

[SOURCE: ISO/TS 20428:2017, 3.18]

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

3.12
mutation annotation format
MAF
tab-delimited text file with aggregated mutation information from *variant call format* (3.21) files and generated on a project-level

3.13
next-generation sequencing
massive parallel sequencing
NGS
technology that can sequence millions of small fragments of *DNA* (3.4) in parallel

3.14
sequence read
read
fragmented nucleotide sequences which are used to reconstruct the original sequence for next generation sequencing technologies

[SOURCE: ISO/TS 20428:2017, 3.26]

3.15
read type
type of implementation in the sequencing instrument

Note 1 to entry: It can be either single-end or paired-end.

Note 2 to entry: Single-end: Single *read* (3.14) implements the sequencing instrument reads from one end of a fragment to the other end.

Note 3 to entry: Paired-end: Paired end implements a read from one end to the other end and then starts another round of reading from the opposite end.

[SOURCE: ISO/TS 20428:2017, 3.27, modified — "run" has been replaced with "implementation" in the definition and the notes to entry.]

3.16
reference sequence
sequence file that is used as a reference to describe the variants that are present in the analyzed sequence

3.18
specimen
biospecimen
sample of a tissue, body fluid, food, or other substance collected or acquired to support the assessment, diagnosis, treatment, mitigation or prevention of a disease, disorder or abnormal physical state, or its symptoms

[SOURCE: ISO/TS 20428:2017, 3.34, modified — the term "biological specimen" has been removed.]

3.19
subject of care
person who uses or is a potential user of a health care service

[SOURCE: ISO/TS 22220:2011, 3.2, modified — "Note 1 to entry" and the abbreviated term "SOC" have been deleted.]

3.20
target capture
method to capture genomic regions of interest from a *DNA* (3.4) sample prior to sequencing

[SOURCE: ISO/TS 20428:2017, 3.36]

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

3.21
variant call format
VCF
format of the text file used in bioinformatics for storing *gene* (3.8) sequence variations

4 Abbreviated terms

BAM	binary alignment map
bp	base pair
COSMIC	Catalogue of Somatic Mutations in Cancer
CRAM	compressed reference-oriented alignment map
EBI	European Bioinformatics Institute
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HUGO	Human Genome Organization
MAF	mutation annotation format
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
SAM	sequence alignment map
VCF	variant call file

5 Summary of the clinical genomic information model

5.1 General

The clinical genomic information model defines the structure and the organization of the information related to the communication of the clinical genomic data generated by massively parallel sequencing technology.

Figure 2 shows the relationships between the major structures of the clinical genomic information model.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

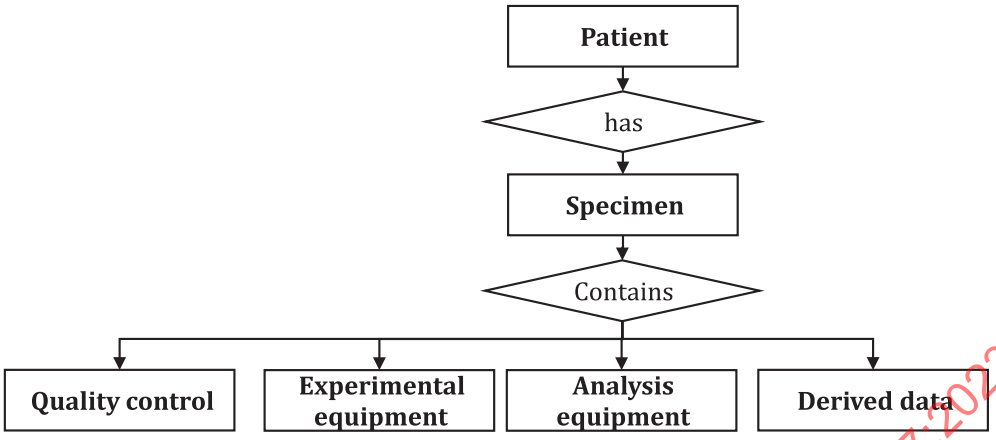


Figure 2 — Major structure of the genomic data model

5.2 Patient

5.2.1 General

A patient is a person receiving or registered to receive healthcare services or is a subject of one or more studies for some other purpose, such as clinical trials.

5.2.2 Identifiers

The unique identifiers for the subject of care (see Table 1)shall be included.

5.2.3 Name

The subject of care name (see Table 1) shall be given as a general rule.

5.2.4 Sex

The sex of the subject of care (see Table 1) shall be in accordance with ISO/TS 22220:2011.

5.2.5 Birth data

The birth date of the subject of care (see Table 1) shall be given to calculate the age of the patient. The birth date should be according to ISO 8601-1 and, if necessary, ISO 8601-2.

5.2.6 Ethnicity

The ethnicity of the subject of care (see Table 1) should be notified to represent his or her genetic origin. The ethnicity information should be represented by HL7 v3 Code System Race[9]. Alternatively, if there are national standards, those coding systems can be used, for example, US FDA Guidance for Industry – Collection of Race and Ethnicity Data in Clinical Trials. The ethnicity of the patient shall be reported.

5.2.7 List of diagnosis

5.2.7.1 General

Diagnosis list, including pertinent data from the investigation, analysis, and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms. If possible, diagnosis

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

should be included using ICD 10 or 11 codes, SNOMED-CT code, or other widely adopted ontologies. The list of diagnosis for the patients shall be reported.

5.2.7.2 Age of diagnosis

The age at the time of diagnosis since birth (see Table 1) shall be included and can be expressed in the number of years.

5.2.8 Treatment

5.2.8.1 Prior treatment

A text description should be included that describes the prior treatment (see Table 1) received before the body specimen was collected.

5.2.8.2 Treatment outcome

A text description should be included that describes the final outcome of the patient (see Table 1) after the treatment was administered.

Table 1 — Summary of patient related metrics

Category	Metrics	Value representation	Optionality
Patient	Identifier	—	Mandatory
	Name	—	Mandatory
	Birth date	ISO 8601-1, ISO 8601-2	Mandatory
	Sex	Male, female	Mandatory
	Ethnicity	HL7 v3 Code System Race	Mandatory
	Diagnosis list	ICD 10 or 11, SNOMED-CT, or other widely adopted ontologies	Mandatory
	Age of diagnosis	Integer	Mandatory
	Prior treatment	—	Optional
	Treatment outcome	—	Optional

5.3 Specimen

5.3.1 General

The specimen information shall be represented by the subject of care identifier type code of ISO/TS 22220:2011 (see Table 2).

EXAMPLE 13-S-048435_A1 - Pathology Number: ISO/TS 22220:2011 (SOC identifier designation: 13-S-048435_A1, SOC identifier geographic area: 1 (local), SOC identifier issuer: AMC (ABC Medical Center), 02 (speciality number - pathology).

5.3.2 Tissue or organ of origin

Specimen origin refers to the anatomical site from which the specimen was acquired. The anatomical site shall be represented by SNOMED CT[10] or other vocabulary (see Table 2).

5.3.3 Collection date

The collection date (see Table 2) is the date when the specimen was acquired. The collection date shall be represented in accordance with ISO 8601-1 and ISO 8601-2.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

5.3.4 Type of specimen

Types of specimens (see Table 2) can be represented by the Standard Preanalytical Code (SPREC) of the International Society for Biological and Environmental Repositories [613]. Currently, SPREC Version 3.0[11] is the up-to-date version.

EXAMPLE BLD (Blood), BUF (buffy coated), non-blood tissue (CEN), semen (SEM).

Table 2 — Summary of patient related metrics

Category	Metrics	Value representation	Optionality
Specimen	General	Identifier type code of ISO/TS 22220:2011	Mandatory
	Tissue or organ of origin	SNOMED CT or other vocabulary	Mandatory
	Collection date	ISO 8601-1, ISO 8601-2	Mandatory
	Type of specimen	SPREC V3.0	Mandatory

5.4 Experimental equipment

5.4.1 General

The information on the sequencing techniques (i.e. sequencing platform, capture method and the related necessary data) should be given.

5.4.2 Quality control

The relevant quality control (QC) (see Table 3) metrics for sequencing and analysis shall be given. The report should include the overall QC metrics for the specimen, the QC metrics for the pre- and post-alignment reads and all the variants, or the QC metrics for specific variants based on the decision of the report generator.

EXAMPLE 1 Pre-alignment QC metrics include the sequencing yield, total number of reads and average read length (bp).

Post-alignment QC metrics include:

— the number of reads mapped to the reference genome (mapping yield, %),

— the bases that were not used for the base call, N, in %,

— the ratio of guanine to cytosine in sequencing data, r_{GC}, in %,

— the quality score of 20 that indicated the accuracy of the base calls, Q_{20}, in %,

— the quality score of 30 that indicated the accuracy of the base calls, Q_{30}, in %,

on target coverage (%) > 1x, on target coverage (%) > 10x, n target coverage (%) > 20x, on target coverage (%) > 100x, mean depth and uniformity.

EXAMPLE 2 GC = 53,2 %, AT = 47,8 %, Q30 = 94,4 %, Q20 = 97,1 %.

5.4.3 Base calling information

5.4.3.1 General

Information on base calling that is generated by base calling software for identifying a nucleotide sequence shall be noted. The detail fields is as given in 5.4.3.2 to 5.4.3.10. The other fields can be added based on the clinical sequencing performed organization's decision.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

5.4.3.2 Read depth

The average number of nucleotides contributing to a portion of an assembly shall be reported as used in conventional bioinformatics fields (see Table 3).

5.4.3.3 Reference allelic depth

Allelic depth for the reference allele (see Table 3) should be reported as used in conventional bioinformatics fields.

EXAMPLE 50*x*, where *x* is the depth of coverage (i.e. depth of sequence).

5.4.3.4 Alternative allelic depth

Allelic depth for the alternative allele (see Table 3) should be reported.

EXAMPLE 50*x*, where *x* is the depth of coverage (i.e. depth of sequence).

5.4.3.5 Allele frequency

The frequency of alternative allele at each locus for each individual should be reported.

EXAMPLE 0 %, 3 %, 30 %

5.4.3.6 Genotype

The pair of alleles present at a single locus should be reported.

EXAMPLE AA, AC, AG, AT, CC, CG, CT, GG, GT, TT.

5.4.3.7 Type of sequencers

The specific sequencer that performs the sequencing shall be given (see Table 3).

EXAMPLE Illumina Hiseq 2500, Thermo Fisher Ion Torrent, Illumina MiSeq[1)].

5.4.3.8 Library preparation methods

5.4.3.8.1 General

The sequencing library preparation methods shall be given.

EXAMPLE SureSelectXT[2)].

5.4.3.8.2 Target capture methods

The exome or targeted region capture methods shall be noted (see Table 3).

EXAMPLE Amplicon, probe capture.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

1) Illumina Hiseq 2500, Thermo Fisher Ion Torrent and Illumina MiSeq are examples of suitable products available commercially. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of these products.

2) SureSelectXT is an example of a suitable product available commercially. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product.

© ISO 2023 – All rights reserved

5.4.3.9 Read type

The sequencing read type (see Table 3) shall be given.

EXAMPLE Single-end, paired-end.

5.4.3.10 Read length

The sequencing read length (see Table 3) information shall be given.

EXAMPLE 101 base pairs (bp), 35 bp to 250 bp.

Table 3 — Summary of experimental equipment

Category	Metrics	Value representation	Optionality
Experimental equipment	Quality control	Text	Mandatory
	Base calling information	Text	Mandatory
	Read depth	Integer	Mandatory
	Reference allelic depth	Integer	Optional
	Alternative allelic depth	Integer	Optional
	Type of sequencers	Text	Mandatory
	Library preparation methods	Text	Mandatory
	Target capture methods	Text	Mandatory
	Read type	Text	Mandatory
	Read length	Text	Mandatory

5.5 Analysis equipment

5.5.1 General

Data generated by NGS technologies are analysed in a series of steps.

The primary analysis is millions to billions of sequences read are initially generated by the sequencing platform. In this proposal, primary analysis describes relevant information in experimental equipment.

The second analysis is that the NGS reads are aligned to the reference needed to identify where the sequence variant exists. The tertiary analysis of variant found is performed to identify those that are relevant to the patient's clinical condition. In the analysis equipment, the secondary and tertiary analysis of NGS was defined as a process from "Read alignment" to "Variant annotation" and related variables were defined.

The primary and secondary (if applicable tertiary) analysis pipelines should be mentioned for clinical sequencing data sharing as a general rule. The parameter settings for the pipelines should also be mentioned to confirm the reliability of the variant call as a general rule. The data analysis pipeline of NGS can be divided into four primary operations: read alignment, alignment post processing, variant calling and variant annotation.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

Table 4 — Workflow scheme for massively parallel sequencing pre-processing

	Read alignment	Alignment post processing	Variant calling	Variant annotation
Objective	— Read alignment — Binary compression — Sort — Add or replace read-groups — Indexing	— Re-aligner target creator — Indel re-aligner — BaseRecalibrator — Target intersection — Remove duplicate results	— Germline/somatic mutation calling — Copy number variant calling — Structure variant calling	— Population frequency — Computational pathogenicity prediction — Variant type — Predicted impact of the variant on the protein
Input/ Output	— Input: FASTQ, SAM — Output: SAM, BAM	— Input: BAM — Output: Preprocessed BAM file	— Input: Preprocessed BAM format — Output: VCF, TXT	— Input: VCF — Output: MAF or other
Tools	— BWA v0.7.12 — SAMTOOLS v1.9	— SAMTOOLS v1.9 — PICARD v1.93 — GATK v3.8 — Bedtools V2.26	— GATK v3.8 HaplotypeCaller Varscan v2.3.9 — MuTect2 — CNVkit v0.9.6 — Lumpy v0.2.14	— PhastCons, Mutalyzer, Provean, Mutation Assessor, SIFT, and PolyPhen2 — ANNOVAR, VAAST, Carpe Novo, Variant Effect Predictor, SNFeff, Ion Reporter, Mutation Taster
Reference	— Reference genome	— Reference genome — TGP — dbSNP — Target bed file	— Reference genome — dbSNP — COSMIC	— dbSNP — COSMIC — NCBI

5.5.2 Read alignment

The read alignment (see Tables 4 and 5) consists of read alignment and binary compression. From the read alignment algorithm to the final variant-calling process within the entire massively parallel sequencing pipeline, various factors can affect the variant calling.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved

Table 5 — Read alignment related metrics

Category	Metrics	Description	Example	Optionality
Read alignment	Process name	The name of sub-process of read alignment	Read alignment, sort	Read alignment related metrics shall be in accordance with Table 4.
	Tool name	The name of the tool used in the read alignment process	Burrows-Wheeler Aligner files (BWA)	Read alignment related metrics shall be in accordance with Table 4.
	Tool version	The version of the tool used for the read alignment process	v0.7.12	Read alignment related metrics shall be in accordance with Table 4.
	Tool options	Information about the options of the tool used in the read alignment process	Number of threads: -t	Optional
	Additional input	Additional input information of the tool used by the user for the read alignment process For example, the reference genome with the release name	GRCh38	Read alignment related metrics shall be in accordance with Table 4.
	Output	Output of the tool used in the read alignment process	SAM file	Optional

5.5.3 Alignment post processing

The local realignment base quality score recalibration includes (see Table 4):

— realinger target creator;

— indel realigner;

— base recalibrator.

5.5.4 Variant calling

Variant calling (see Tables 4 and 6) consists of germline mutation calling and somatic mutation calling.

STANDARDSISO.COM : Click to view the full PDF of ISO/TS 23357:2023

© ISO 2023 – All rights reserved