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

Reference number
ISO/TS 23357:2023(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) 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.

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.

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

1 Scope

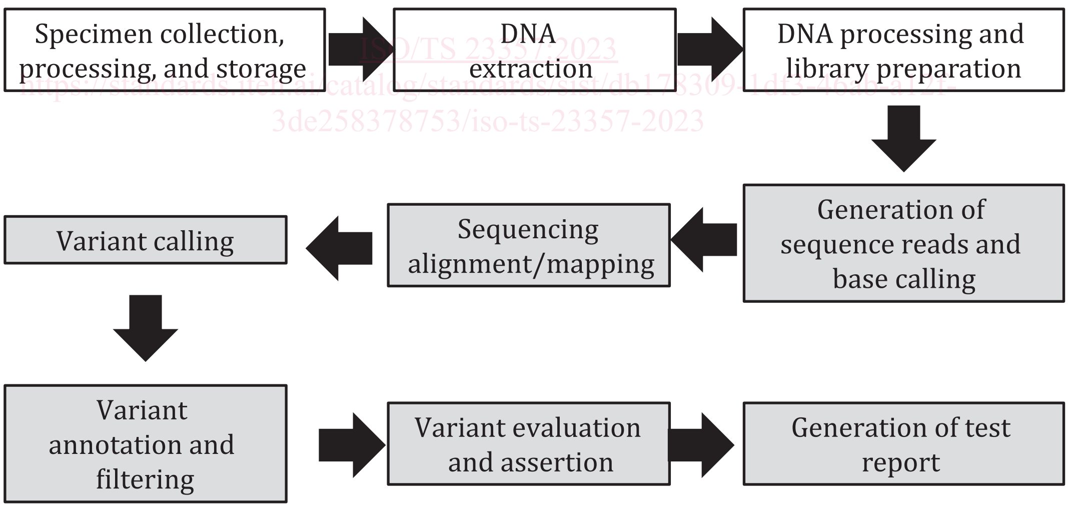
This document specifies clinical sequencing information generated by massive parallel sequencing technology for sharing health information via massively parallel sequencing. This document covers the data fields and their metadata from the generation of sequence reads and base calling to variant evaluation 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.

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]

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]

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.

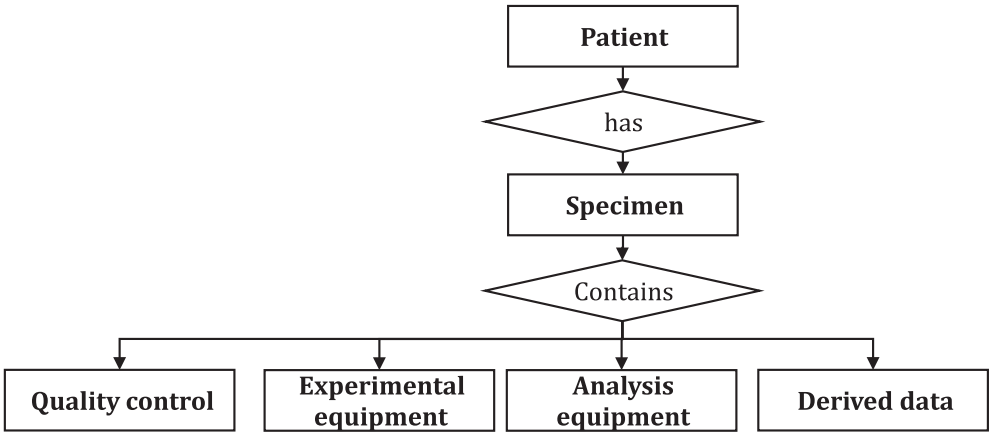


Figure 2 — Major structure of the genomic data model

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