TECHNICAL SPECIFICATION

ISO/TS 22692

First edition
2020-10

Genomics informatic — Quality control metrics for DNA sequencing

Informatique génomique — Mesures de contrôle de la qualité pour le séquençage de l'ADN

Reference number
ISO/TS 22692:2020(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 documents 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 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).

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

The rapid progress in Next Generation Sequencing (NGS) technology has drastically reduced the cost and time for genomic analysis. A number of research institutions, corporations, and government agencies are competitively collecting a large volume of genomic data through multi-national, multi-institutional projects such as "DiscovEHR"[9], "gnomAD"[10] and "UK Biobank"[11]. The demand for sharing of "high quality" genomic data is growing because large-scale reference data is required for reliable detection of mutation for both industrial and clinical applications.

However, the quality of available genomic data is less than desirable. To establish consistent quality control metrics, details of each stage of NGS process need to be recorded, shared and standardized (processes and data elements collected and coded for each stage and sub-stage). These processes include sample preparation, library preparation, sequencing, and data processing, among others, as shown in Figure 1.

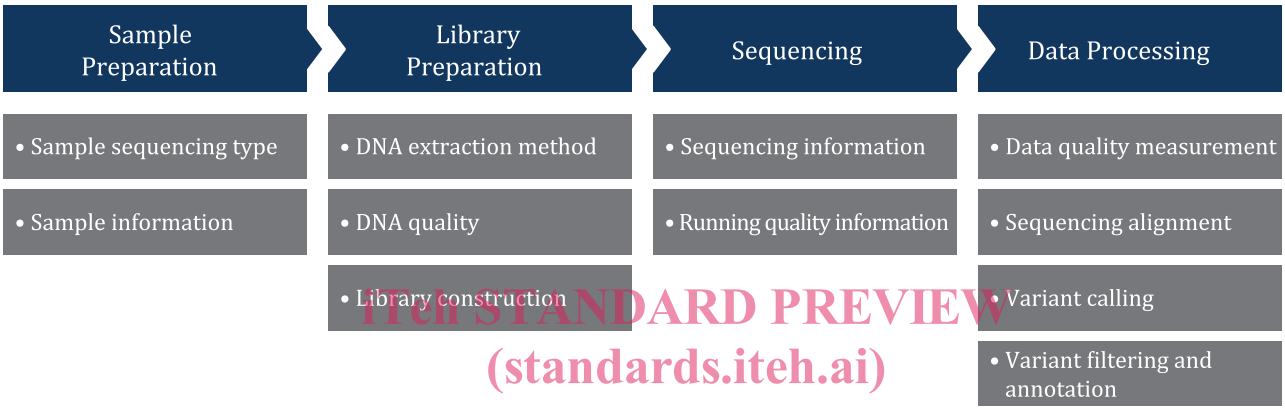

Figure 1 — NGS process

Genomics informatic — Quality control metrics for DNA sequencing

1 Scope

This document identifies quality metrics for the detection of DNA variants using next generation sequencing (NGS) technology. It also defines the data types, relationships, optionality, cardinalities and terminology bindings of the data.

This document provides a basis for sharing and for the application of "high quality" genomic data and contributes to the realization of the precision medicine and the development of relevant industries.

This document is intended to serve as a catalogue of sequencing data elements used to address quality metrics for various clinical, industrial and commercial applications. The exchange of these data allows researchers, commercial entities, and regulatory bodies to assess for the purpose of selective utilization of the data by setting application-specific quality criteria

This document is not intended for

— sequencing methods other than NGS, such as the Sanger sequencing,

— targets other than genome, such as transcriptome or proteome, or

— specimens of species other than humans.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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
copy number variation
CNV
variation (3.18) in the number of copies of one or more sections of the *DNA* (3.3)

[SOURCE: ISO/TS 20428:2017, 3.7]

3.2
deletion
contiguous removal of one or more bases from a genomic sequence

[SOURCE: ISO/IEC 23092-2:2019, 3.4]

3.3
DNA
deoxyribonucleic acid
molecule that exists in nuclei and in mitochondria of human cells and is composed of a linear array of 4 bases (Adenine: A, Thymine: T, Guanine: G and Cytosine: C)

[SOURCE: ISO 18074:2015, 4.1, modified — Note 1 to entry deleted.]

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

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

[SOURCE: ISO/TS 17822-1:2014, 3.20]

3.5
exome
part of the genome formed by exons

[SOURCE: ISO/TS 20428:2017, 3.13]

3.6
FASTA
genomic information representation that includes a name and a nucleotide sequence for each *sequence read* (3.17)

[SOURCE: ISO/IEC 23092-2:2019, 3.7, modified]

3.7
FASTQ
genomic information representation that includes *FASTA* (3.6) and quality values

[SOURCE: ISO/IEC 23092-2:2019, 3.8]

3.8
gene
basic unit of hereditary information composed of chains of nucleotide base pairs in specific sequences that encodes a protein or protein subunit

[SOURCE: ISO 11238:2018, 3.29]

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 (3.11) or/and *deletion* (3.2)

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

3.11
insertion
contiguous addition of one or more bases into a genomic sequence

[SOURCE: ISO/IEC 23092-2:2019, 3.18]

3.12
large indel
insertion (3.11) or *deletion* (3.2) up to around 1 kb

[SOURCE: ISO/TS 20428:2017, 3.21]

3.13
nucleotide
monomer of a nucleic acid polymer such as *DNA* (3.3) 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'.

[SOURCE: ISO/IEC 23092-2:2019, 3.20]

3.14
polymerase chain reaction
PCR
in vitro enzymatic technique to increase the number of copies of a specific DNA fragment by several orders of magnitude

[SOURCE: ISO 16577:2016, 3.148]

3.15
quality score
Phred quality score
Q score
quality measure used to assess the accuracy of a sequencing reaction

Note 1 to entry: This quality measure indicates the probability that a given base is called incorrectly by the sequencer. Phred scores are on a logarithmic scale. Therefore, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1 000 times. A lower base call accuracy of 99 % (Q20) will have an incorrect base call probability of 1 in 100, meaning that every 100 base pairs sequencing read will likely contain an error.

[SOURCE: ISO 21286:2019, 3.4]

3.16
reference sequence
nucleic acid sequence with biological relevance

Note 1 to entry: 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.

[SOURCE: ISO/IEC 23092-1:2019, 3.22]

3.17
sequence read
read
fragmented nucleotide sequences that are used to reconstruct the original sequence for next generation sequencing technologies

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

3.18
sequence variation
DNA sequence variation
variation
differences of DNA sequence among individuals in a population

Note 1 to entry: Variation implies *copy number variation* (3.1), *deletion* (3.2), *insertion* (3.11), *indel* (3.10), *small indel* (3.20), *large indel* (3.12), or *single nucleotide variant* (3.19).

[SOURCE: ISO/TS 20428:2017, 3.30]

3.19
single nucleotide variant
SNV
DNA sequence variation (3.18) that occurs when a single nucleotide, A, T, C, or G, in the genome (or other target sequence) differs between templates

[SOURCE: ISO 20395:2019, 3.35]

3.20
small indel
insertion (3.11) or *deletion* (3.2) of 2 nucleotides to 100 nucleotides

[SOURCE: ISO/TS 20428:2017, 3.32]

3.21
specimen
biospecimen
biological specimen
sample of tissue, body fluid, food, or other substance that is 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]

3.22
targeted sequencing
disease-targeted gene panel
technique used for sequencing only selected/targeted genomic regions of interest from a DNA sample

Note 1 to entry: For further details, see Reference [12].

3.23
whole exome sequencing
WES
technique for sequencing the *exomes* (3.5) of the protein-coding *genes* (3.8) in a genome

[SOURCE: ISO/TS 20428:2017, 3.38]

3.24
whole genome sequencing
WGS
technique that determines the complete DNA sequence of an organism's genome at a single time

[SOURCE: ISO/TS 20428:2017, 3.39]

4 Abbreviated terms

BAM	Binary Alignment/Map
BED	Browser Extensible Data
FMA	Foundational Model of Anatomy
HGNC	HUGO Gene Nomenclature Committee
HUGO	Human Genome Organization
NGS	Next Generation Sequencing
RefSeq	NCBI Reference Sequences
SAM	Sequence Alignment/Map
SPREC	Standard PREanalytical Code
VCF	Variant Call Format

5 Quality control metrics for sample preparation

5.1 General

As the first step in NGS workflow, sample preparation is the process to assess and review the submitted sample before being accepted for sequencing. The amount of information in sequencing data is quite different according to sequencing types (whole genome sequencing, whole exome sequencing, and targeted sequencing) and specimen types (blood, surgical tissue, biopsy, etc.), so quality control metrics for sample preparation would be used to select appropriate comparison data in analysis process.

5.2 Sample sequencing type

5.2.1 Sequencing type

Sequencing types are three levels of analysis via NGS: WGS, WES, and targeted sequencing[12]. WGS covers all regions of the genome: both coding and noncoding regions. WES covers all coding regions, which is estimated to comprise 1% to 2 % of the genome, yet contains around 85 % of recognized disease-related mutations[13]. Targeted sequencing interrogates known disease-associated genes, which is focusing on a limited set of genes to allow greater depth of coverage.

5.2.2 Target gene

Sets of genes or gene regions for targeted sequencing. Target gene is represented in accordance with HGNC and RefSeq.

EXAMPLE HGNC approved symbol: BRCA1, HGNC ID: HGNC:1100, RefSeq: NM_007294.

NOTE Target gene is used only for WGS.

5.3 Sample information

5.3.1 Specimen type

Type of specimen (e.g. whole blood, cell, urine, fresh cell & tissue) with related data during sample collection (e.g. biopsy, surgical excision, EDTA, heparin), processing (e.g. formalin, centrifugation),

