TECHNICAL SPECIFICATION

ISO/TS 4425

First edition
2023-04

Genomics informatics — Data elements and their metadata for describing the microsatellite instability (MSI) information of clinical massive parallel DNA sequencing

Informatique génomique — Éléments de données et leurs métadonnées pour décrire les informations relatives à l'instabilité des microsatellites (MSI) du séquençage massif parallèle d'ADN

Reference number
ISO/TS 4425:2023(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 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

Massively parallel sequencing is a high-throughput analytical approach to nucleic acid sequencing that allows whole genomes, transcriptomes, and specific nucleic acid targets. These advanced technologies have been used in the clinical field, and clinical sequencing has been applied to realize personalized medicine and precision medicine. ISO/TS 20428[1] has been developed for clinical usage.

In the field of cancer treatment, various treatment strategies were performed differently from traditional anti-cancer chemotherapies. One of those strategies is the control of human immune system that maintains the action to extract cancer cells. Recent outcomes of clinical trials show that this immune therapy is efficient for some patients who have a specific molecular character of their tumor mass, such as PD-L1 or CTLA4 surface protein expression[2]. As a result, these molecular characters are used as biomarkers for selecting patients. In colon cancer, according to several clinical trials, it is reported that the status of MSI (microsatellite instability) is regarded as a biomarker that drugs based on immuno-therapy are more efficient for the patient with MSI-H (high)[3].

The status of MSI can be calculated and reported by small nucleotide deletion on a specific region of human genome reference with NGS sequencing[4]. According to US FDA, four NGS sequencing products were approved for companion diagnostics. Among these products, three NGS sequencing provide MSI status and value on their NGS sequencing report. CLIA-certified labs or equivalent level agencies in countries also are servicing the MSI status from their methods[5]. It is forecasted that more clinical NGS sequencing will be approved to report MSI.

However, there is no standard for describing MSI status, value, and metadata. ISO/TS 20428 focuses on only DNA variations compared with the reference genome. According to some research results, MSI status and the way to describe it are different even if using the same sequencing data. This makes it difficult for clinicians and researchers not only to use MSI status results for clinical decisions but also for secondary analyzing purposes when receiving from more than one sequencing lab. Related metadata should be essential to expand the usage of MSI status results.

In this document, the data elements and their standardized metadata for MSI status in electronic health records will be described. The clinical report for MSI will provide helpful information on bioinformatics analysis to help clinical decisions.

Genomics informatics — Data elements and their metadata for describing the microsatellite instability (MSI) information of clinical massive parallel DNA sequencing

1 Scope

This document identifies data elements and metadata to represent the information about microsatellite instability (MSI) for reporting the value of the biomarker using clinical massive parallel DNA sequencing.

This document covers information about the MSI test result and related data, such as used resources, data generation condition, and data processing information which are helpful to clinical diagnosis and research.

This document is not intended

— for defining experimental protocols or methods for calculating the value of microsatellite instability (MSI),

— for the other biological species than human resource, or

— for the Sanger sequencing methods.

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 8601 (all parts), *Date and time — Representations for information interchange*

ISO/TS 22220:2011, *Health informatics — Identification of subjects of health care*

ISO/TS 27527:2010, *Health informatics — Provider identification*

3 Terms and definitions

For the purposes of this document, the following terms and definitions 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
biological specimen
biospecimen
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]

3.2
clinical sequencing
next generation sequencing or later sequencing technologies with human samples for clinical practice and clinical trials

[SOURCE: ISO/TS 20428:2017, 3.5]

3.3
deletion
contiguous removal of one or more bases from a genomic sequence

[SOURCE: ISO/IEC 23092-2:2020, 3.4]

3.4
deoxyribonucleic acid
DNA
molecule that encodes genetic information in the nucleus of cells

[SOURCE: ISO 25720:2009, 4.7]

3.5
DNA sequencing
determining the order of nucleotide bases (adenine, guanine, cytosine and thymine) in a molecule of DNA

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

[SOURCE: ISO 17822:2020, 3.19]

3.6
exome
part of the genome formed by exons

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

3.7
gene
basic unit of hereditary material that encodes and controls the expression of a protein or protein subunit

3.8
indel
insertion (3.9) or/and *deletion* (3.3)

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

3.9
insertion
contiguous addition of one or more bases into a genomic sequence

[SOURCE: ISO/IEC 23092-2:2020, 3.18]

3.10
microsatellite
repetitive DNA elements, also known as simple sequence repeats (SSR), consisting of short in tandem repeat motifs of one to a few nucleotides that tend to occur in non-coding DNA of eukaryotic genomes and that are sometimes referred to as variable number of tandem repeats (VNTRs)

3.11
microsatellite instability
MSI
condition of genetic hypermutability (predisposition to mutation) that results from impaired DNA mismatch repair (MMR)

3.12
DNA mismatch repair
MMR
system for recognizing and repairing erroneous insertion, deletion, and misincorporation of bases that can rise during DNA replication and recombination, as well as repairing some forms of DNA damage

3.13
nucleotide
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'.

[SOURCE: ISO/IEC 23092-2:2020, 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 16577:2022, 3.6.47]

3.15
quality score
Phred quality score
Q score
sequencing quality score of a given nucleotide base

Note 1 to entry: Q is defined by the following equation: $Q = -10\log10(e)$, where e is the estimated probability of the base call being wrong.

Note 2 to entry: A quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99 %.

Note 3 to entry: Higher quality scores indicate a smaller probability of error. Lower quality scores can result in a significant portion of the reads being unusable. Low quality scores may also indicate false-positive variant calls, resulting in inaccurate conclusions.

[SOURCE: ISO 20397-2:2021, 3.30]

3.16
read type
type of run 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 runs the sequencing instrument reads from one end of a fragment to the other end.

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

[SOURCE: ISO/TS 20428:2017, 3.27]

3.17
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 increases within the string from left to right.

[SOURCE: ISO/IEC 23092-1:2020, 3.22]

3.18
read
sequence read
fragmented nucleotide sequences that are used to reconstruct the original sequence for next-generation sequencing technologies

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

3.19
variation
sequence variation
DNA sequence variation
differences of DNA sequence among individuals in a population

[SOURCE: ISO 25720:2009, 4.8]

3.20
small indel
insertion (3.9) or *deletion* (3.3) of 2 nucleotides to 100 nucleotides

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

3.21
subject of care
person who uses, or is a potential user of, a healthcare service

[SOURCE: ISO/TS 22220:2011, 3.2, modified — Note to entry and second preferred term deleted.]

3.22
target capture
method to capture genomic regions of interest from a DNA sample prior to sequencing

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

3.23
targeted sequencing
technique used for sequencing only selected/targeted genomic regions of interest from a DNA sample

[SOURCE: ISO/TS 22692:2020, 3.22, modified — Note to entry and second preferred term deleted.]

3.24
whole exome sequencing
WES
technique for sequencing the exomes of the protein-coding genes in a genome

3.25
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

ATC	Anatomical Therapeutic Chemical
CTLA4	Cytotoxic T-Lymphocyte Associated Protein 4
EBI	European Bioinformatics Institute
FHIR	Fast Healthcare Interoperability Resources
HL7®	Health Level Seven
IDMP	Identification of Medicinal Product
IMPID	Investigational MPID
INN	International Nonproprietary Names
MPID	Medicinal Product Identifier
NCCN	National Comprehensive Cancer Network
NGS	Next Generation Sequencing
NIH	National Institutes of Health
PD-L1	Programmed death-ligand 1
PD-1	Programmed cell death protein 1
SPREC	Standard PREanalytical Code
WHO	World Health Organization
UTN	Universal Trial Number

5 Microsatellite instability (MSI)

The DNA mismatch repair (MMR) pathway plays an important role in the cell cycle process to recognize and repair mismatches during DNA replication. The major components are four key enzymes coded for by the following genes: MLH1, MSH2, MSH6, PMS2, and EPCAM. MMR function doesn't work when mutational inactivation in the five genes or epigenetic inactivation occurs. It is called Deficiency of mismatch repair (dMMR). One of the most related diseases is Lynch syndrome[6]. Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC), is the most common cause of hereditary colorectal cancer. People with Lynch syndrome are more likely to get colorectal cancer and other cancers at a younger age (under 50). Patients develop dMMR tumors following the inactivation of the second wild-type allele through somatic mutation, loss of heterozygosity, or epigenetic silencing. These alterations – mutation or epigenetic inactivation is related to not only Lynch syndrome but also revealed differences in the case of cancer type. However, both lead to the accumulation of short sequences of DNA repeated throughout the genome-specific location and an increased risk of malignant transformation in certain tissues. These tumors have a higher frequency of somatic mutations compared with non-dMMR cancers and are assumed to have a large range of tumor neoantigens (high tumor mutation burden) and a highly immunogenic signature, including a high proportion of tumor-infiltrating lymphocytes. Defective mismatch repair results in a high tumor mutation burden and abundant neo-antigen formation, which can be recognized by the host immune system. Microsatellite instability (MSI) is found in 1,5 % to 3,5 % of all human cancers, such as colorectal, endometrial, ovarian, and cancers of the stomach, small intestine, pancreas, biliary tract, and ureter. The human genome contains more than 19 million microsatellites, short tandem repeats of motifs of 1 nucleotide to 6 nucleotides, typically spanning 10 nucleotides to 60 nucleotides in total length[7]. However, if the MMR function doesn't work well, nucleotide error accumulates, especially in the human genomic position, including microsatellites.