FINAL
DRAFT

TECHNICAL
SPECIFICATION

ISO/DTS
24420

ISO/TC **276**

Secretariat: **DIN**

Voting begins on:
2023-02-20

Voting terminates on:
2023-04-17

Biotechnology — Massively parallel DNA sequencing — General requirements for data processing of shotgun metagenomic sequences

Biotechnologie — Séquençage d'ADN massivement parallèle — Exigences générales pour le traitement des données des séquences métagénomiques "Shotgun"

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

Reference number
ISO/DTS 24420:2023(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 *276*, *Biotechnology*.

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

Shotgun metagenomic sequencing genomes of organisms in a complex sample in a community to gain knowledge of its composition and function is widely used in life science and clinical applications, such as human complex disease associated analysis, environmental microecology and other fields. It has potential to provide significant scientific data for life science research.

The utility of this technique is its ability to reveal the microbial diversity and abundance found in microbial populations from multiple environments and to determine sequence information (taxonomic characterization, functional annotation, and comparative analysis/metagenomics) for individual organisms in these populations. The resulting data can be subjected to comparative analytics. Massively parallel shotgun metagenomic sequencing generates a large amount of data containing a high complexity of microbial genomes and a large number of unknown species. It is important to use effective processing procedures and address quality control for shotgun metagenomic sequencing data. A standardised data format is essential to promote data sharing.

As with any advanced technology, massively parallel sequencing technologies is error prone. Overcoming these shortcomings to ensure a reliable sequencing and analytical outcome is important. This document provides a uniform standard for the collation, storage and subsequent analysis of metagenomic data, and guidelines. It provides requirements and recommendations for the workflow and process of shotgun metagenomic analyses including quality control of sequencing data and metadata, and the compositional and functional analysis of microbial community. These requirements and recommendations can ensure accuracy of data generated from metagenomic analysis, address potential errors and facilitate downstream applications.

Biotechnology — Massively parallel DNA sequencing — General requirements for data processing of shotgun metagenomic sequences

1 Scope

This document illustrates the workflow of shotgun metagenomic sequence data processing of host-derived microbiome and environmental metagenomes.

This document specifies the requirements for quality control of shotgun metagenomic sequence data processing for massively parallel DNA sequencing.

This document provides guidelines for data directory, data archive and metadata for shotgun metagenomic sequence data.

This document applies to data storage, sharing and interoperability of shotgun metagenomic sequence data.

This document applies to shotgun metagenomic sequence data processing and analyses, but excludes functional analysis.

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 20397-1:2022, *Biotechnology — Massively parallel sequencing — Part 1: Nucleic acid and library preparation*

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
attribute value
value associated with an attribute instance

[SOURCE: ISO 21962:2003, 1.5.2.3]

3.2
category
set of items or concepts that share a common attribute or feature

3.3
classification
exhaustive set of mutually exclusive categories to aggregate data at a pre-prescribed level of specialization for a specific purpose

[SOURCE: ISO 17115:2007, 2.7.1]

3.4
clean data
sequencing data obtained after a pre-processing procedure which usually includes multiple trimming and filtering steps to ensure specific quality levels (e.g., per-base quality, host/contaminant sequences removed, linkers/adaptors removed)

3.5
code
system of rule(s) to convert information such as text, images, sounds or electric, photonic or magnetic signals into another form or representation to facilitate analysis, communication or storage in a storage medium

[SOURCE: ISO 20691:2022, 3.6]

3.6
encoding
process of assigning code to things or concepts

3.7
contig
contiguous sequence of DNA created by assembling overlapping sequenced fragments of a chromosome or plasmid

3.8
data format
arrangement of data according to preset specifications

Note 1 to entry: Preset specifications are usually made for computer processing.

3.9
data element
single unit of data that in a certain context is considered indivisible

[SOURCE: ISO/TS 21089:2018, 3.44]

3.10
directory
list of data items, which gives itemized information enabling traceability, identification and findability of related data

Note 1 to entry: A directory can be arranged in alphabetical, chronological or systematic order.

3.11
directory identifier
unique language-independent sign assigned to the archive directory in the structure

3.12
gene
sequence of nucleotides in DNA or RNA encoding either an RNA or a protein product

Note 1 to entry: Genes are recognized as the basic unit of heredity.

Note 2 to entry: A gene can consist of non-contiguous nucleic acid segments that are rearranged through a nuclear processing step.

Note 3 to entry: A gene may include or be part of an operon that includes elements for gene expression.

[SOURCE: ISO 20397-2:2021, 3.16]

3.13
identifier
sequence of characters, capable of uniquely identifying that with which it is associated, within a specified context

[SOURCE: ISO/IEC 11179-1:2015, 3.33]

3.14
analytical data
set of elements to describe qualitative or quantitative analytical attributes of processed metagenomic raw data

3.15
name
semantic, natural language labels given to data elements, and variations of these labels serve different functions

[SOURCE: ISO/IEC 11179-1:2015, 3.43]

3.16
public attribute
attribute that can have same attribute value for different data in the directory

3.17
quality score
Q score
Phred score
quality of base calling
measure of the probability of correct base recognition, usually expressed directly by a numerical value

Note 1 to entry: Q is defined by the following equation:

$$Q = -10\log_{10}(p)$$

where p 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: A quality score of 30 represents an error rate of 1 in 1 000, with a corresponding call accuracy of 99,9 %.

Note 4 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, 3.32, modified — Note 3 was added.]

3.18
raw data
primary sequencing data produced by a sequencer without involving any software-based pre-filtering for analysis purpose

[SOURCE: ISO 20397-2:2021, 3.21]

3.19

relative abundance

fraction of a single microorganism operational taxonomic unit in the total microbial community of a defined environment

Note 1 to entry: It usually represented as a percentage.

3.20

repeatability requirement

requirement of consistency under a set of repeatable measurement conditions

3.21

scaffold

reconstructed genomic sequence created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome

3.22

sequence assembly

processing, aligning and merging individual sequencing reads in order to reconstruct longer DNA sequences, entire genes or genomes

Note 1 to entry: When sequencing a novel genome where there is no reference sequence available for alignment, sequence reads are assembled as contigs, that is the *de novo* assembly.

3.23

shotgun metagenomic sequencing

shotgun metagenomics

nucleotide sequence determination of the genomes of untargeted cells in communities in order to determine community composition and function

Note 1 to entry: For the microbiome, shotgun metagenomics focuses on microbial communities in specific environments.

Note 2 to entry: For shotgun metagenomic sequencing, DNA is extracted from the microbes in the sample directly without isolation and culture. That DNA is then used to analyse the genetic composition, species classification, phylogeny, gene function, or metabolic network or combinations thereof.

3.24

specialized attribute

attribute that is unique for each sample in the directory

4 Processing workflow

The basic workflow of metagenomics should include sequencing, data processing and data analysis. Data processing includes pre-processing, quality control, data assembly, data profiling and annotation, as shown in Figure 1.

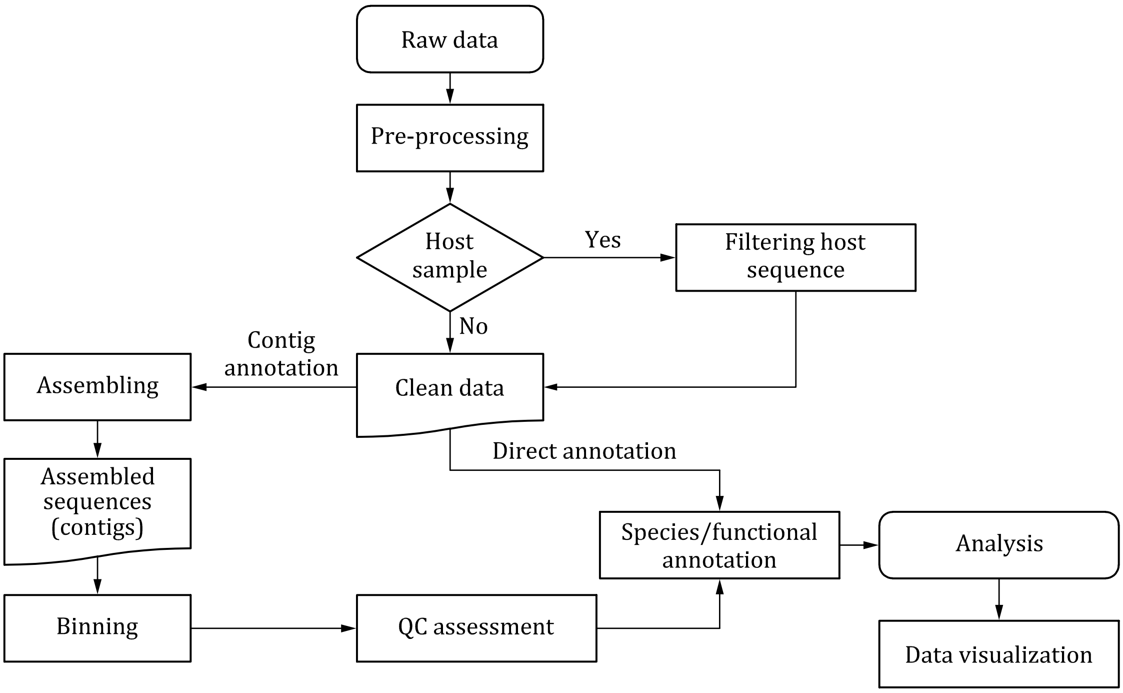

Figure 1 — Workflow of metagenomic data processing

5 Data processing

5.1 Facilities and software requirements

5.1.1 The software pipeline for metagenomics bioinformatics shall be validated. Applications for the pipeline should be locked down including the complete set of tools, code, operational environment, and network connections that compose the pipeline before using it for analytical purposes such as shell (e.g., BASH), GNU R, and Python. Changes to any components of the pipeline require revalidation to ensure that there is no impact in the performance characteristics of the pipeline.

5.1.2 High-performance computing technologies may be used at any step in the process to ensure proper management and curation of large collections of complex procaryotic and eucaryotic genomes as processing massive datasets is a prerequisite for NGS metagenomics analytics.

5.2 Sequence quality control and error determination

5.2.1 Raw metagenomic sequencing data shall initially be passed through a quality control (QC) process to ensure a clean dataset. The evaluation should follow ISO 20397-1:2022, Clause 4 and 8.3, and ISO 20397-2:2021, 4.3.

5.2.2 The available data quality values for each DNA sample after sequencing should meet the following requirements:

a) Q20 ≥ 90 %, above 90 % of the sample base mass value shall be more than 20;

b) Q30 ≥ 80 %, above 80 % of the sample base mass value shall be more than 30.

The above requirements only apply to short sequence reads ≤ 350 bp.