भारतीय मानक

Indian Standard

IS 18358 : 2023
ISO 23418 : 2022

खाद्य श्रृंखला की सूक्ष्मजैविकी — जीवाणुओं के प्ररूपण एवं जीनोमिक लक्षण वर्णन के लिए संपूर्ण जीनोम अनुक्रमण — सामान्य अपेक्षाएँ एवं मार्ग दर्शिका

Microbiology of the Food Chain — Whole Genome Sequencing for Typing and Genomic Characterization of Bacteria — General Requirements and Guidance

ICS 07.100.30

भारतीय मानक ब्यूरो
BUREAU OF INDIAN STANDARDS
मानक भवन, 9 बहादुर शाह ज़फर मार्ग, नई दिल्ली - 110002
MANAK BHAVAN, 9 BAHADUR SHAH ZAFAR MARG
NEW DELHI - 110002
www.bis.gov.in www.standardsbis.in

July 2023 **Price Group 14**

Food Microbiology Sectional Committee, FAD 31

NATIONAL FOREWORD

This Indian Standard which is identical to ISO 23418 : 2022 ‘Microbiology of the food chain — Whole genome sequencing for typing and genomic characterization of bacteria — General requirements and guidance’ issued by the International Organization for Standardization (ISO) was adopted by the Bureau of Indian Standards on the recommendation of the Microbiology Sectional Committee and approval of the Food and Agriculture Division Council.

The text of ISO Standard has been approved as suitable for publication as an Indian Standard without deviations. Certain conventions are, however, not identical to those used in Indian Standards. Attention is particularly drawn to the following:

a) Wherever the words ‘International Standard’ appear referring to this standard, they should be read as ‘Indian Standard’; and

b) Comma (,) has been used as a decimal marker while in Indian Standards, the current practice is to use a point (.) as the decimal marker.

In reporting the results of a test or analysis made in accordance with this standard, if the final value, observed or calculated, is to be rounded off, it shall be done in accordance with IS 2 : 2022 ‘Rules for rounding off numerical values (*second revision*)’.

Contents

Page

Introduction

Next generation sequencing (NGS) provides rapid, economical and high-throughput access to microbial whole genome sequences and is being applied to an expanding number of problems in food microbiology. Whole genome sequences are representations of the biological potential of the sequenced organism at single base resolution. Whole genome sequencing (WGS) offers significant advantages over existing technologies (e.g. serotyping, pulsed field gel electrophoresis, antibiotic resistance phenotype) for many applications. WGS-based analyses are used by public health laboratories to detect outbreaks, and to detect mutations, genes and other genetic features to characterize virulence and survival potential. Within the food industry, there is interest in using whole genome sequences to characterize bacterial isolates from ingredients and environmental surfaces, to better understand their origin and ecology, and to update procedures to reduce risk. Some companies have developed, or are developing, the capacity to collect and analyse whole genome sequence data. Others are turning to third-party laboratories to perform these services, as they have done for other microbiological analyses.

This document provides guidance for both the laboratory and bioinformatic components of whole genome sequences and associated metadata for bacterial foodborne microorganisms sampled along the food chain (e.g. ingredients, food, feed, production environment). Although microbiology of the food chain includes viruses and fungi, this document is only intended for bacteria. This document is intended to be applicable to all currently available next generation DNA sequencing technologies. It may be applied to analysis of whole genome sequence data with proprietary, open-source or custom software. It is not intended to specify sequencing chemistries, analytical methods or software. This document defines laboratory, data and metadata stewardship practices to ensure that analyses are clearly reported, transparent and open to inquiry. This document is for use by laboratories to develop their management systems for quality and technical operations. Laboratory customers and regulatory authorities can also use it in confirming or recognizing the competence of laboratories. This document can also be applied in other domains (e.g. environment, human health, animal health).

Indian Standard

MICROBIOIOGY OF THE FOOD CHAIN — WHOLE GENOME SEQUENCING FOR TYPING AND GENOMIC CHARACTERIZATION OF BACTERIA — GENERAL REQUIREMENTS AND GUIDANCE

WARNING — In order to safeguard the health of laboratory personnel, it is essential that handling of bacterial cultures is only undertaken in properly equipped laboratories, under the control of a skilled microbiologist, and that great care is taken in the disposal of all incubated materials. Persons using this document should be familiar with normal laboratory practice. This document does not purport to address all safety aspects, if any, associated with its use. It is the responsibility of the user to establish appropriate safety and health practices.

1 Scope

This document specifies the minimum requirements for generating and analysing whole genome sequencing (WGS) data of bacteria obtained from the food chain. This process can include the following stages:

a) handling of bacterial cultures;

b) axenic genomic DNA isolation;

c) library preparation, sequencing, and assessment of raw DNA sequence read quality and storage;

d) bioinformatics analysis for determining genetic relatedness, genetic content and predicting phenotype, and bioinformatics pipeline validation;

e) metadata capture and sequence repository deposition;

f) validation of the end-to-end WGS workflow (fit for purpose for intended application).

This document is applicable to bacteria isolated from:

— products intended for human consumption;

— products intended for animal feed;

— environmental samples from food and feed handling and production areas;

— samples from the primary production stage.

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 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
adapter sequence
DNA with a known sequence that is added to the end of a DNA library fragment to facilitate the sequencing process (e.g. annealing to a flow cell)

3.2
annotation
process of identifying genes and other features on genome *assemblies* (3.4)

3.3
antibiogram
summary of antimicrobial susceptibility testing results performed for a specific microorganism, usually represented in tabular form

3.4
assembly
output from process of aligning and merging sequencing *reads* (3.38) into larger contiguous sequences (*contigs* (3.10))

3.5
base calling
process of assigning nucleotides and quality scores to positions in sequencing *reads* (3.38)

3.6
bioinformatics
collection, storage and analysis of biological data including sequences

3.7
bioinformatics pipeline
individual programs, scripts or pieces of software linked together, where output from one program is used as input for the next step in data processing

3.8
carryover-contamination
sample contamination linked to previous experiments, transferred to the current analysis (e.g. carryover-contamination from amplification products in prior polymerase chain reaction (PCR) experiments to the current PCR analysis, or carryover-contamination of previously sequenced samples from one sequencing run to another)

3.9
Chemical Entities of Biological Interest Ontology
ChEBI
ontology (3.35) for describing small chemical compounds

3.10
contig
contiguous stretch of DNA sequence that results from the *assembly* (3.4) of smaller, overlapping DNA sequence *reads* (3.38)

3.11
controlled vocabulary
finite set of values that represent the only allowed values for a data item

[SOURCE: ISO 11238:2018, 3.18, modified — Note 1 to entry deleted.]

3.12
coverage
number of times that a given base position is read in a sequencing run

Note 1 to entry: The number of *reads* (3.38) that cover a particular position.

[SOURCE: ISO 20397-2:2021, 3.6, modified — Admitted term “coverage depth” deleted.]

3.13
cross-contamination
contamination of a sample (bacterial *isolate* (3.23) or DNA) with other samples during the preparation of a sequencing run

3.14
DNA sample
portion of DNA extracted from the processed sample

3.15
draft assembly
de novo genome *assembly* (3.4) consisting of *contigs* (3.10) with no implied order, typically generated using whole genome shotgun sequencing with a short-read technology

3.16
Environment Ontology
EnvO
ontology (3.35) for describing environmental features and habitats

3.17
FoodEx2 Ontology
FoodEx2
standardized food classification and description system developed by the European Food Safety Authority (EFSA)

3.18
Food Ontology
FoodOn
ontology (3.35) for describing food products, animal feed and food processing

3.19
Gazetteer Ontology
GAZ
ontology (3.35) for describing geographical locations

3.20
index
oligonucleotide sequences used in the process of library preparation to tag or barcode DNA from specific samples, so that multiple samples (i.e. multiple *libraries* (3.25)) can be combined (multiplexed) in a pool of libraries and analysed in a single sequencing reaction

3.21
International Nucleotide Sequence Database Collaboration
INSDC
initiative operated by the DNA Database of Japan (DDBJ), the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) and the National Center for Biotechnology Information (NCBI)

3.22
International Organization for Standardization whole genome sequencing slim
ISO WGS Slim
ontology (3.35) slim containing interoperable fields and terms pertaining to the use of *WGS* (3.49) for microbiology of the food chain

3.23
isolate
population of bacterial cells in pure culture derived from a single *strain* (3.45)

3.24
kmer
possible sequence of length *k* that is contained in a whole genome sequence

3.25
library
collection of genomic DNA fragments from a single *isolate* (3.23) intended for determining genome sequence(s)

Note 1 to entry: A collection of libraries, each of a single isolate, is called a “pool of libraries” and is loaded on a sequencer to be analysed. This multiplexing of libraries would still provide the result for a single isolate if unique indices are used for each individual single isolate’s library preparation.

Note 2 to entry: A library of mixed DNA, i.e. originating from a mixture of multiple species, can be made. However, this is not within the scope of this document as this refers to metagenomics sequencing.

3.26
management system
quality, administrative and technical systems that govern the operations of an organization

Note 1 to entry: For the purposes of this document, “organization” refers to the laboratory.

3.27
mapping
use of software to align sequencing *reads* (3.38) to reference sequences

3.28
metadata
data that defines and describes other data

[SOURCE: ISO/IEC 11179-1:2015, 3.2.16]

3.29
minimal data for matching
MDM
information required to describe the sample source and provenance of a genomic sequence, as defined by the Global Microbial Identifier[1], and implemented by the *International Nucleotide Sequence Database Collaboration* (3.21)

3.30
multi-locus sequence typing
MLST
method of genomic analysis that identifies nucleotide variants within predefined sets of loci

Note 1 to entry: Originally used for seven loci, it is now also applied to either core genome loci for cgMLST or whole genome loci for wgMLST.

3.31
N50
length (N) such that sequence *contigs* (3.10) of N or longer include half the bases in the *assembly* (3.4)

3.32
NCBITaxon
automatic translation of the National Center for Biotechnology Information (NCBI) taxonomy database into obo/owl

3.33
NG50
length (N) of DNA such that sequence *contigs* (3.10) of N or longer include half the bases in the genome

3.34
Open Biological and Biomedical Ontology Foundry
OBO Foundry
collection of *ontologies* (3.35) created by a collective of ontology developers that are committed to collaboration and adherence to shared principles

3.35
ontology
controlled vocabulary (3.11) arranged in a hierarchy, where the terms are connected by logical relationships

3.36
ontology slim
set of ontology fields and terms annotated as part of a particular collection, often for a specific purpose, which may be extracted to create a file distinct from the original *ontology* (3.35)

3.37
Phred sequence quality score
Q
measure of the probability (P) that a base is incorrectly assigned at a given position in the sequence expressed as:

$$Q = -10\ \lg P$$

Note 1 to entry: A score of Q30 indicates that there is a 1 in 1 000 chance that a base is incorrectly assigned (i.e. the base call is 99,9 % accurate).

3.38
read
nucleotide sequence inferred from a fragment of DNA or RNA

3.39
sequence repository
database in which *whole genome sequencing* (3.49) datasets are stored and managed

Note 1 to entry: A public repository allows unrestricted access to the data, while a private or federated repository restricts access to the data.

3.40
sequencing replicate
<biological> sequencing a different colony from the same *isolate* (3.23) obtained from the same sample material, to assess biological variation

3.41
sequencing replicate
<technical> resequencing of the same biological sample or *library* (3.25) to assess sequence variation due to instrumentation and protocol

3.42
serotype
classification scheme based on the antigenic protein detection or sequence-based detection of genes encoding bacterial surface molecules

3.43
single nucleotide polymorphism
SNP
single nucleotide variant (3.44) that passes a particular quality or frequency threshold

3.44
single nucleotide variant
SNV
differences between the nucleotides at the same genomic position of two or more *isolates* (3.23)

3.45
strain
descendants of a single isolation in pure culture, usually derived from a single initial colony on a solid growth medium

Note 1 to entry: A strain may be considered an *isolate* (3.23) or group of isolates that may be distinguished from other isolates of the same genus and species by phenotypic and genotypic characteristics.

Note 2 to entry: See Reference [2].

3.46
validation
establishment of the performance characteristics of a method and provision of objective evidence that the performance requirements for a specified intended use are fulfilled

[SOURCE: ISO 16140-1:2016, 2.81]

3.47
validated data entry
automated process ensuring that data entered into a repository are correct

3.48
verification
demonstration that a validated method functions in the user's hands according to the method's specifications determined in the validation study and is fit for its intended purpose

[SOURCE: ISO 16140-3:2021, 3.21, modified — Note 1 to entry deleted.]

3.49
whole genome sequencing
WGS
process of determining the DNA sequence of an organism's genome using total genomic DNA as input

4 Principle

4.1 General

WGS analyses of bacteria along the food and feed chain consists of culturing the pure bacterial isolate, DNA isolation performed in a microbiological laboratory, sequencing steps conducted in an appropriate sequencing environment and bioinformatics analysis performed in a distinct computational environment.

NOTE The microbiology laboratory, the sequencing facility and the bioinformatics facility can be the same organization.

4.2 Laboratory operation: sample preparation and sequencing

Sample preparation and sequencing should include the following steps:

a) Information about the isolates being sequenced, including barcodes for multiplexed samples, is entered into the appropriate record systems, such as a laboratory information management system (LIMS) or sample description worksheets, or both.

b) Pure isolates (identified at least to the genus level and ideally to the species level) are cultured and genomic DNA is extracted.

c) DNA sequencing libraries are prepared from quality controlled genomic DNA (see Table A.1 for guidance on DNA quantity and quality metrics). This process should include:

   1) DNA fragmentation, if required for the applied sequencing technology;

   2) ligation of indices and adapters, consistent with the applied sequencing technology's protocols;

   3) quantification, normalization and quality control of the resulting library;

   4) pooling of libraries in the case of multiplexed sequencing runs.

d) Libraries (i.e. pool of libraries) are sequenced.

e) Quality metrics produced by the sequencing instrument are ideally recorded for each run to allow monitoring of the performance.

4.3 Bioinformatics analysis

4.3.1 General

Pipelines for bioinformatics analysis may focus on *in silico* predictions of phenotype (e.g. virulence) or detecting clusters of genetically similar isolates (i.e. same strain, sequence type or serotype). Pipelines based on comparative approaches may be used to detect the presence and states of markers in raw and assembled sequencing data to make *in silico* strain (e.g. sequence type) and phenotype predictions.

Sequence data for multiple isolates may be analysed using SNP, MLST or kmer distance analysis methods to identify clusters of closely related bacteria. Results from these analyses may be used to infer relationships between isolates, which may be illustrated with phylogenetic trees and dendrograms.

4.3.2 SNP analyses

For SNP analyses, reads are mapped to a reference sequence or reads are assembled into contigs that are compared. To determine SNPs, SNVs are quality-filtered to identify SNP positions.

4.3.3 MLST analyses

For MLST analyses, reads are assembled or mapped. Alleles are identified, quality-filtered and compared to a cgMLST or wgMLST database.

4.3.4 Kmer distance analysis

Sequence data for multiple isolates may be analysed using kmer distance methods to identify clusters of related bacteria. Kmer analyses have the advantage of being very fast but have some limitations, notably in terms of precision (i.e. they are applicable in species determination, but not recommended for detailed source tracking analysis of closely related strains).

4.4 Metadata formats and sequence repository deposition

Metadata records shall be created and safely stored for all sequences. Sequence data and corresponding metadata should be consistently formatted and documented. These metadata may be shared solely at the discretion of the metadata owner. Sequence data and its corresponding metadata shall be subject to security considerations, cost and benefits, intellectual property rights, confidential business information, contract restriction or other binding written agreements.

NOTE Licensing or a privacy policy, or both, can be applied to metadata or sequence data, or both, to protect private or proprietary information.

To promote data stewardship best practices[3], this document provides optional metadata reporting formats which are harmonized to a community data standard (e.g. MDM or OBO Foundry ontologies). These formats and standards facilitate reproducibility and common understanding of terminology. An

ISO WGS Slim was created to format and provide values for the recommended metadata fields. WGS and selected metadata may be transferred (uploaded) to a publicly accessible database.

4.5 Validation and verification of WGS workflow

The entire WGS workflow shall be validated to provide assurance that the methods are fit for intended use.

NOTE More details on the validation and verification of the WGS workflow are given in Clause 10 and Table 4.

5 General laboratory guidance

5.1 Bacterial isolation and DNA extraction

Bacterial isolation and DNA extraction should be performed in a general microbiological laboratory adapted to work with the specific bacteria, including pathogens. For sequencing library preparation that involves DNA amplification using polymerase chain reaction (PCR), pre- and post-PCR steps should be carried out in different or segregated areas of the laboratory to avoid carryover-contamination.

5.2 Laboratory environment

Air movements, vibration, temperature and humidity can interfere with the performance of many sequencers and should be considered in the placement of the equipment in the laboratory. Laboratories should consult the sequencer manufacturer’s site preparation guide for specific guidance.

5.3 Standard operating procedures and nonconforming work

Laboratories should maintain and adhere to standardized operating procedures (SOPs), workflow documents, reagent inventory controls and equipment maintenance logs. SOPs should include procedures for using positive and negative controls for the DNA extraction, sequence library preparation and sequencing steps. SOPs should include procedures for monitoring operations for run quality and errors (sample misidentification or cross-contamination).

In the case of sample misidentification or contamination, the root cause of errors in sequencing shall be investigated:

a) ensuring that runs containing misidentified samples, or samples contaminated with multiple strains, are not used for bioinformatics analysis for sample interpretation or uploaded to databases;

b) implementing measures to maintain quality and prevent recurrence of errors.

5.4 Laboratory information management system

Sample information shall be captured using a LIMS or similar system of documenting and tracking information.

5.5 Laboratory competence

Laboratories should maintain records documenting training, education and proficiency for individuals performing sequencing and bioinformatics analysis, and sample retention policy.

The laboratory should monitor its performance for WGS analysis by comparison with results of other laboratories, where available and appropriate. This monitoring should be planned and reviewed and include, but not be limited to, one of the following, ideally annually:

a) participation in a proficiency testing programme;

b) participation in interlaboratory comparisons other than proficiency testing;

c) verification of the analytical process by introducing "blind" samples or samples whose characteristics are not known by the operator.

Data (e.g. sequence data, run metrics, result reports provided by the organizing institution) from these monitoring activities should be analysed, used to control and, if applicable, used to improve the laboratory's activities. If the results of the analysis of data from these monitoring activities are found to be outside predefined criteria, appropriate actions should be taken to prevent incorrect results from being used for sample analysis.

6 Laboratory operations

6.1 Sample preparation and storage

Any material to be sequenced (bacterial isolate or extracted genomic DNA) should be handled and stored in a way that minimizes the risk of sample degradation, misidentification and cross-contamination.

6.2 Bacterial isolates

Bacterial isolates should be stored and cultured by processes that minimize the potential for introducing genetic changes (e.g. loss of plasmids or polymorphisms introduced through culture and passaging). If the laboratory receives a bacterial isolate, the laboratory shall ensure the purity of the isolate and ideally confirm species before subsequent steps are performed. If there is concern that potentially unstable elements (e.g. plasmids) can be lost from a sample during passage, then sequences should ideally be collected from at least two biological replicates. The number of single colony passages performed after receipt of the isolate should be noted in the sample metadata. Bacterial isolates should be archived using methods such as freezing as a glycerol stock at −80 °C.

6.3 DNA isolation

For bacterial DNA isolation, an extraction procedure should be selected that is suitable for the respective organism and provides DNA of sufficient quality with regard to the sequencing platform used. Bacterial DNA isolation is influenced by a number of factors including cell type (Gram positive or negative), growth phase (early, mid, late log or stationary) and culture medium. The quantity and quality of DNA should be assessed and documented. Storage conditions will influence DNA integrity and library preparation for certain sequencing technologies.

NOTE Some DNA extraction methods are better than others for the recovery of plasmids. If plasmids are important for the specific application, an appropriate method can be used.

6.4 Library preparation and sequencing

6.4.1 Library preparation

The laboratory should follow the manufacturer's recommended protocol. Procedures may be adapted for specific needs, but all modifications shall be fully documented and validated.

NOTE Size-selection procedures used in some library preparation methods [e.g. in construction of large insert size (> 2 kb) single molecule real time libraries] can result in the loss of small plasmids.

PCR enrichment of libraries can result in reduced library complexity and a reduction in the number of distinct DNA molecules in the preparation. Library complexity can also be affected by the amount of DNA starting material or the amount of DNA lost during library preparation clean-up steps. Library complexity may be estimated using the method of Daley and Smith[4].

If there is a possibility that libraries will be used again, libraries shall be stored according to the manufacturer's recommendations. The laboratory shall document:

— the sample tracking method used (i.e. barcode or equivalent);

- the sequencing platform used;
- the operator who made the library;
- the date the library was made;
- the lot information for the kit(s) used.

Multiplexing samples (i.e. combining different single libraries, each of a single isolate, into a pool of libraries to be sequenced) requires selection and assignment of barcodes to identify individual samples, and is typically documented in a worksheet to allow association of sequence data with the correct metadata. If all multiplexed samples are of the same bacterial genus (e.g. all *Salmonella*), steps should be taken to ensure that equimolar DNA inputs are used (i.e. library normalization) and that the correct sequence is associated with its corresponding metadata. If the multiplexed samples represent multiple genera, then estimated coverage, genome size and library fragment size need to be considered when estimating the amount of DNA to be included for each sample.

6.4.2 DNA sequencing

Sequencing instrumentation shall be operated and maintained as per the recommendations of the manufacturer, and documentation of maintenance procedures shall be maintained. Platform-specific sequencing metrics (e.g. cluster density, number of reads, average base quality) shall be recorded and monitored for each sequencing run. Platform-specific recommendations to minimize carryover-contamination are provided in Clause A.1.

6.4.3 Use of controls

When handling a bacterial isolate and DNA extract, the laboratory should use a water blank or non-inoculated broth as negative control during DNA extraction to assess possible cross-contamination. A positive extraction control to assess method efficiency can be included as deemed necessary. If the library preparation involves multiplexing and PCR amplification steps, then it should include both positive and negative controls. It is also recommended to consistently use the same DNA extract for the positive control to allow for comparisons of sequencing quality from run to run. Recommendations for using positive and negative controls are provided in Clause A.3.

6.4.4 Assessing raw read data quality

Base calling should be performed using software specific to the instrument and sequencing chemistry. Metrics may be defined at run level and at sample level. Metrics shall be documented to evaluate the quality of raw sequence data. These can include insert size, sequence length distribution, number of reads and an assessment of base composition [i.e. AT/GC balance or TAGC (taxon annotated GC-coverage) plot or equivalent]. Quality scores, and read length, and taxonomy check should be used for an initial check of sequencing performance (see also 7.3). Coverage, as calculated by mapping reads back to a *de novo* assembly or to an appropriate reference genome, should also be evaluated.

DNA sequence read quality and quantity impact downstream assembly, read mapping, and the ability to use WGS data for bacterial source tracking and genome characterization. Sequencing artefacts that may impact downstream analyses include sequencing platform specific error profiles, variation in quality scores across the sequence read, biases in sequence data driven by base composition, departure from optimal library fragment sizes, and contamination from known and unknown species other than the sequencing target.

NOTE General guidance for developing quality metrics is provided in Clauses A.1 and A.2.

6.4.5 Sample and data storage and retention

The laboratory shall document a SOP for the storage and retention of specimens, DNA samples, DNA libraries and sequencing data.

7 Bioinformatic data analysis

7.1 Requirements for software and bioinformatic pipelines used for data analysis

Software and bioinformatics pipelines should be developed and maintained using software quality control and quality assurance procedures. All pipeline software dependencies should be described in a way that allows reproducibility of the computing environment, which needs to include versioning of each software module, database used and the pipeline itself.

Software and bioinformatic pipelines should be validated before use for data analysis (see 10.1.3).

Regarding bioinformatics pipelines:

- pipelines should be validated, for reproducibility and robustness, on public or test data sets;
- pipeline developers should distribute test data sets with their software;
- users should ensure that pipelines are installed correctly by analysing the test data sets and checking that the expected results are generated.

7.2 Logging and documentation

All data analytic steps and analyses should be logged and documented. A plan for updating the bioinformatics pipeline as updates to software components become available should be developed and implemented. The impact of software updates should be evaluated and documented. A re-validation can be needed (see 10.1.2) in the case of software updates. If data sets are transferred, data integrity before and after transfer should be checked (e.g. using md5checksum). Exception logs should be used to document any deviations from SOPs during individual bioinformatics analyses (e.g. that the SOP was not followed as described).

7.3 Quality assessments

The quality of sequence data should be assessed and documented at the completion of the sequencing run. Quality metrics should be platform specific (see Clauses A.1 and A.2). Users should determine and record their specifications for the quality assessment parameters. Criteria used for assessing sequence quality for an isolate may include:

- average quality score and number of bases greater than a specific quality threshold;
- number of reads (read depth) and average read Phred score;
- tests for detecting contamination should be implemented and acceptable limits for contaminants (e.g. sequencing carryover or cross-contamination from sample preparation) should be determined that are appropriate for bioinformatics analyses.

For bioinformatic pipelines using assemblies, the quality of the assemblies should be assessed prior to starting analyses. The following measures are recommended as general indicators of assembly quality:

- the read depth needs to be sufficient to ensure variants are reliably detected in the assembly;
- the number of contigs;

  NOTE For draft assemblies, low coverage and/or small contigs can be removed prior to reporting the number of contigs, depending on the pipeline or needs.

- either N50 or NG50, or both, and length of the longest contig;
- the total length of all contigs or scaffolds should approximate the known genome size of the target organism;
- the presence of species-specific conserved elements (e.g. core genome).

Laboratories should test for contamination in sequencing data and determine limits appropriate for specific applications. These contaminations can originate from a different species and genus or from the same species. Recommended methods include, but are not limited to, one or more of the following:

- kmer hashing against a reference sequence database;
- calculating the average nucleotide identity (ANI) of sequence data;
- checking for numbers of rDNA alleles in reads or assemblies;
- verifying serotypes with bioinformatic serotype prediction tools;
- comparing assemblies to reference databases.

If results from non-WGS phenotypic or molecular tests for a bacterial isolate are available, they may be compared to WGS findings to evaluate consistency of genomic predictions. Examples include but are not limited to:

- presence or absence of known resistance elements for isolates with antimicrobial susceptibility profiles;
- serotype;
- antigenic loci;
- presence or absence of virulence or pathogenic elements.

7.4 SNP analyses

For SNP analyses, either a genetically similar draft assembly or a finished genome sequence may be used in accordance with Reference [5]. Reference sequences should be curated prior to analyses (e.g. by removing small contigs or contigs with low depth of coverage), as necessary for given applications. SNVs should be filtered using quality scores, depth of coverage, density and masking of highly variable regions, as appropriate for a given application, to reduce errors caused by sequencing and alignment artefacts, indels (insertions/deletions), structural variants, recombination and mobile genetic elements. Filtering conditions used to identify SNP positions should be documented.

NOTE Analysis of benchmark and simulated data sets can help to establish false positive and false negative rates for SNP pipelines.

The reference genome should be genetically similar to subject sequences as the false positive rate for SNP identification tends to increase with increasing genetic distance to the reference genome[5]. Users should specify what is considered as “genetically similar”.

7.5 MLST analyses (cgMLST and wgMLST)

Criteria for adding or removing alleles or loci from cgMLST or wgMLST databases should be clearly defined. If MLST allele determinations are made using assembled genomes:

- the same genome annotation method should be used when generating databases and during subsequent analyses;
- minimum quality standards for assemblies, such as percentage of core loci detected, should be established.

If MLST allele determinations are made using read mapping:

- criteria for SNP and indel identification (e.g. alignment depth, minimum percentage coverage of loci, number of mismatches allowed) should be documented;
- minimum quality standards for WGS data sets, such as percentage of core loci detected, should be established;

— the quality criteria will depend on the genus; caution should be taken when creating a proprietary schema.

Whether allele determinations are performed using assembled genomes or read mapping, MLST schemes (cgMLST or wgMLST) should preferably be described in papers published in peer-reviewed journals. The schema version used should be reported in the metadata.

7.6 Target gene detection

Databases used for target gene detection (e.g. virulence gene, antimicrobial resistance gene, serotype) should be documented, including the version number and update information. The criteria used to determine whether the target gene is present or not should be clearly defined (e.g. the percentage of coverage and the percentage of identity).

7.7 Phylogenetic tree or dendrogram generation

Distance, parsimony, minimum spanning trees and maximum compatibility methods of analysis can be used to rapidly screen WGS data sets and identify clusters of closely related isolates. Results are typically presented as a dendrogram, graph or pairwise distance matrix. Trees or dendrograms can be built from the pairwise distance matrix, shared loci, genes or variants depending on the application. Bayesian and maximum likelihood methods of phylogenetic tree construction are preferred to distance-based methods because they are statistically consistent (i.e. converging on the correct topology as more data are acquired). Bootstrapping should be used to estimate statistical support for topologies under distance-based, parsimony, maximum compatibility and maximum likelihood methods of analysis. *A posteriori* probabilities should be used when trees are generated through Bayesian methods.

7.8 Metrics and log files

Metrics and log files should be kept and include text that describes the following:

a) the identities of isolates analysed, along with sequence and assembly metrics if available;

b) the identities of any reference sequences, per cent of reads mapped and coverage (for SNP);

c) the version of the database (for MLST);

d) the version of the bioinformatics pipeline used, parameter settings and user identification;

e) timestamps;

f) any filtering or masking conditions.

7.9 Interpreting and reporting the results of bioinformatics analyses

7.9.1 Interpreting results from bioinformatics pipelines

Results from bioinformatics pipelines should be interpreted in the context of information including metadata about the origins of isolates and epidemiology (i.e. traceback information). Thresholds established for one purpose (e.g. clonal outbreaks) should not be used for interpretation of a different purpose (e.g. persistent or resident pathogens)[6].

Diversity of isolates should be considered when interpreting dendrograms or allelic/SNP differences. Most genus/species differ in mutation rates. Some lineages may be clonal. In these cases, few differences may specify outbreaks/clusters. Other lineages have greater diversity. In general, for contamination events with a single point source that occur over a short period of time, such as over the course of a foodborne outbreak, very few genetic changes are expected to occur. For large-scale contamination events, greater differences can be observed.

7.9.2 Reporting genomic analysis results

The following information should be included in reports or available upon request:

a) version of pipeline;

b) identity of input data;

c) reference genome or MLST database used and version if appropriate;

d) analytic settings if options are available (e.g. minimum coverage settings for calls, filtering or masking);

e) interpretation and conclusions of genomic comparison results if part of the application.

8 Metadata

8.1 General

The organization shall adopt a policy for capturing metadata. Metadata in private repositories should be as detailed as possible but the level of detail is at the discretion of the user. When permitted, metadata may be shared with partners, and can be abstracted to a level of granularity that complies with organizational data-sharing policies.

8.2 Metadata interoperability and future-proofing

8.2.1 General

The metadata structure and content recommendations in this document are intended to ensure metadata interoperability and utility when making comparisons between different databases. The metadata recommendations are also intended to structure data to be amenable to unanticipated uses (future-proofing). The approach to metadata standardization defined in this subclause captures information about laboratories and laboratory processes, manufacturing environments, food products and food processing, and bioinformatics processes and quality control metrics. This document is designed to be flexible because some stakeholders collect more specific information and others less.

Genomic sequence metadata stored in private repositories may include information describing the sample, the isolate and the sequence. Metadata fields and values may be supplied in the format described in Tables 1, 2 and 3 and Annexes B to H. Metadata captured according to this document may be provided according to sample type. However, a null value (e.g. “missing”, “not collected”, “not provided”, “restricted access”) is also acceptable if information is not known or available. The ISO WGS Slim and other ontologies may be used to format and provide values for metadata fields described.

8.2.2 Ontologies

Ontologies encode computational logic that can be used by software systems to improve automation and more complex querying[7][8]. The hierarchical nature of ontologies enables better comparisons of information at different levels of granularity[7][8].

8.2.3 ISO WGS Slim

The ISO WGS Slim contains standardized fields and terms derived from existing ontologies and other community standards [e.g. INSDC minimal data for matching (MDM) and antibiogram standards][9]. The fields specify the information types recommended for capture, while the terms act as possible values, which can be used to populate the fields. The ISO WGS Slim also contains synonymous term labels from different organizations (e.g. FoodEx2[10]) to avoid preferential use of vocabulary and to facilitate interoperability and data harmonization.

The ISO WGS Slim can be used to format and provide values for metadata fields described in Tables 1, 2, 3, B.1, C.1, D.1, E.1, F.1, G.1, H.1 and H.2 and Annexes B to H. Geographic, taxonomic, food product and processing, environment, and drug fields in this document requiring more extensive vocabulary may be supplied using FoodEx2 and the GAZ, NCBITaxon, FoodOn, EnvO and CheBI ontologies[8][9][10].

The ISO WGS Slim can be implemented in metadata spreadsheets and information management systems. Further instruction is provided in Annex I and Tables I.1, I.2 and I.3

8.3 Formatting metadata using this document

Metadata fields in Tables 1, 2 and 3 are general, and some require additional detailed fields to structure the information and minimize the use of free text. These additional fields are described in Annexes B to H.

Each field in the tables and annexes has a definition, specified role in data analyses or harmonization, and specific formatting recommendations. Values for some fields (e.g. food product) may be provided by implementing the ISO WGS Slim, values for other fields (e.g. dates) may be constrained using validated data entry, while other fields may be provided as free text.

Metadata formatted using this document are compatible with INSDC data standards. Guidance for preparing metadata for submission to INSDC public repositories is given in Annex H (see Clause 9).

8.4 Metadata associated with sample collection

Fields in Table 1 may be used to capture metadata associated with sample collection. Additional fields are provided in Annexes B and C.

Table 1 — Recommended metadata fields and values associated with sample collection

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Sample collection laboratory contact information The name of the laboratory that collected the sample being analysed, as well as the name and contact information of an individual who can provide further details regarding the project or sample, should be provided.	Establishing chain of custody and for providing contact information for follow-up analyses.	Contact information may be specified by the fields of information in Annex B.
Geographic location of sample collection The geographical origin of the sample.	INSDC data standard (fulfils MDM “geo_loc” field).	Sample geographic location information may be specified by the fields of information in Annex C.
Collection date The date the sample was collected.	INSDC data standard (fulfils MDM “collection_date” field).	The sample collection date may be recorded as YYYY-MM-DD in accordance with ISO 8601-1 using validated data entry.
Sample type The type of material from which the isolate was obtained. Samples are usually categorized as foods, food products, animal feed or environmental samples taken from the area of food production and food handling	Traceback and other analyses.	Sample types may be selected from the ISO WGS Slim.
Food product Products intended for human consumption and the feeding of animals	INSDC data standard (fulfils MDM “isolation_source” field).	Food products and ingredients can apply to both human and animal food. Where food products apply, descriptors may be selected from the ISO WGS Slim.

Table 1 *(continued)*

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Food processing Processing applied to a food product (e.g. deboning, skinning, pasteurization).	INSDC data standard (fulfils MDM "isolation_source" field).	Where food processing applies, descriptors may be selected from the ISO WGS Slim.
Environmental material A substance obtained from the natural or man-made environment (e.g. soil, water, manure).	INSDC data standard (fulfils MDM "isolation_source" field).	Food is considered to be a separate field from environmental material. Where environmental materials apply, descriptors may be selected from the ISO WGS Slim.
Environmental location An environmental location may describe a site in the natural or built environment (e.g. abattoir, retail outlet, feedlot, food processing machinery, surfaces used to process and prepare food products).	INSDC data standard (fulfils MDM "isolation_source" field).	Food-related environmental locations may include, but are not exclusive to, food production, processing, distribution and retail environments that were sampled. Where environmental locations apply, descriptors may be selected from the ISO WGS Slim.
Collection device The instrument or container used to collect the sample [e.g. sterile plastic bag, plastic jar, swab (with or without transport medium), tube].	INSDC data standard (fulfils MDM "isolation_source" field).	Collection devices are not always known, however when specified, they may be selected from the ISO WGS Slim.
Collection method The process used to collect the sample.	INSDC data standard MDM when the method used for collection is known (fulfils MDM "isolation_source" field).	Collection methods are not always known, however when specified, they may be selected from the ISO WGS Slim.

8.5 Metadata associated with the isolate

Fields in Table 2 may be used to capture metadata associated with the isolate. Additional fields are provided in Annexes B, D, E and F.

Table 2 — Recommended metadata fields and values associated with the isolate

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Microbiology laboratory contact information The name of the laboratory that isolated the organism being sequenced, as well as the name and contact information of an individual who can provide further details regarding the project or isolate, should be provided.	Establishing chain of custody and for providing contact information for follow-up analyses.	Contact information may be specified by the fields of information in Annex B.
Organism The species of the isolate being sequenced.	INSDC data standard (fulfils MDM "organism" field).	The scientific name for the species may be provided using standardized taxonomic names from NCBITaxon.
Strain The name or identifier of the strain.	INSDC data standard (fulfils MDM "strain or isolate" field).	The strain identifier may be provided as free text.

Table 2 *(continued)*

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Isolate The name or identifier of the isolate.	INSDC data standard (fulfils MDM “strain or isolate” field).	The isolate identifier may be provided as free text.
Serotype The serotype of the isolate or strain as determined by *in vitro* or *in silico* methods (e.g. WGS, PCR or immunological methods).	Public repository MDM (required for EBI “serotype” field).	The serotype results may be provided as free text, if available.
Isolation media The culture media used to isolate the organism being sequenced from others in the sample.	Computable comparisons of methodologies.	Descriptors of this material may be chosen from the ISO WGS Slim.
Isolate passage history The number of times that an isolate is serially sub-cultured by a particular method.	Computable comparisons of methodologies. An increase in the number of times an isolate has been passaged can result in the accumulation of additional mutations.	Isolate passage details may be specified by the fields of information in Annex D.
Antibiogram results The minimal inhibitory concentrations [value, unit, sign (<, > , =)] and resistance phenotypes (resistant, sensitive, intermediate or undetermined) of the sequenced isolate against different antibiotics tested.	Computable comparisons of antibiograms.	If antibiogram results are available, the information may be specified according to the fields in Annex E. The source of breakpoints (and version) used for interpreting/ classifying minimum inhibitory concentration (MIC) values may be specified.
Antibiogram methods The laboratory protocol used to determine resistance phenotypes and minimal inhibitory concentrations of antibiotics tested against an isolate. The protocol should include the antibiotics tested, laboratory testing method, testing standard and controls/reference strains used for the test.	Computable comparisons of antibiograms.	Antibiogram methods (if applicable) may be specified according to the fields in Annex E.
Virulence factor results The virulence factors determined to be present in the sequenced isolate by phenotypic or target amplification methods (e.g. Shiga toxins, haemolysins).	Computable comparisons of virulence.	If virulence factor test results are available, the information may be specified according to the fields in Annex F.
Virulence factor testing methods The laboratory protocol used to determine virulence phenotypes and markers.	Computable comparisons of virulence.	Virulence testing methods (if applicable) may be specified according to the fields in Annex F.

8.6 Metadata associated with the sequence

Fields in Table 3 may be used to capture metadata associated with the sequence. Additional fields are provided in Annexes B and G.

Table 3 — Recommended metadata fields and values associated with the sequence

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Sequencing facility contact information The name of the facility that sequenced the isolated organism, as well as the name and contact information of an individual who can provide further project and sequencing details, should be provided.	Establishing chain of custody and for providing contact information for follow-up analyses.	Contact information may be specified by the fields of information in Annex B.
Sequencing date The date the sequencing run was initiated.	Tracking sequencing runs.	The sequencing date may be recorded as YYYY-MM-DD in accordance with ISO 8601-1, using validated data entry.
Culture media Formulation of substances in liquid, semi-solid or solid form that contain either natural or synthetic, or both, constituents intended to support the multiplication (with or without inhibition of certain microorganisms) identification or preservation of viability of microorganisms.	Computable comparisons of methodologies.	Descriptors of this material may be chosen from the ISO WGS Slim.
DNA extraction method The procedure used to obtain genomic DNA from a sample through chemical, physical or mechanical means.	Computable comparisons of methodologies and quality control.	Include the name of the commercial kit and version number, or laboratory protocol, used to extract the genomic DNA of the isolated organism using free text.
Sequencing replicates A technical sequencing replicate represents the resequencing of the same biological sample in order to assess experimental variation. A biological sequencing replicate represents a sequencing experiment based on a different colony from the same isolate obtained from the same sample material, in order to assess biological variation.	Tracking sequencing runs, and analysing variability in reads and sequences.	Replicates within a set of sequencing runs may be described as either technical or biological. Where sequencing replicates apply, descriptors may be selected from the ISO WGS Slim.
Sequence library preparation method The procedure used to create a library from fragments of DNA using oligonucleotides with the role of adapters.	Computable comparisons of methodologies and quality control.	Include the name of the commercial kit and version number, or laboratory protocol, used to prepare libraries for sequencing as free text.
Sequencing instrumentation The type of instrument used to automate the process of sequencing.	Computable comparisons of methodologies.	Types of sequencing instruments may be chosen from the ISO WGS Slim.
Bioinformatics organization contact information The name of the organization performing the bioinformatics processing and analyses, as well as the name and contact information of an individual who can provide further details regarding the bioinformatics analyses, should be provided.	Establishing chain of custody and for providing contact information for follow-up analyses.	Contact information may be specified by the fields of information in Annex B.

Table 3 *(continued)*

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Raw sequence data processing The procedure used to remove adapter sequences from raw sequence reads, trim low-quality bases and, where applicable, merge paired-end reads.	Computable comparisons of methodologies and quality control.	Include the name and version of the trimming tool and, if applicable, paired-end merger program. It is recommended that parameters are also recorded. This information may be provided as free text.
Sequence data filtering method The procedure used to remove low quality reads and unalignable sequences from raw sequence data.	Computable comparisons of methodologies and quality control.	Include the name and version of filtering tool(s) and processes applied. It is recommended that parameters are also recorded. This information may be provided as free text.
Sequence assembly method The method or algorithm used to assemble individual sequence reads into larger contiguous sequences (contigs).	Computable comparisons of methodologies and quality control.	Describe the bioinformatics pipeline used, including the name and version of assembler software, and accession number of the reference genome used in the case of reference-based assembly. It is recommended that parameters are also recorded, along with any post-assembly modifications. This information may be provided as free text.
Sequence annotation method The method or algorithm used to identify and report sequence features (e.g. protein coding regions) in sequence data.	Computable comparisons of methodologies and quality control.	Include the name and version of annotation tool. It is recommended that parameters are also recorded. This information may be provided as free text.
Sequence assembly quality metrics Measurements or calculated quantities used to assess the extent and success of the sequence assembly process. Metric thresholds are species-specific.	Computable comparisons of methodologies and quality control.	Sequence quality control metrics may be specified by the fields of information in Annex G.

9 Sequence repositories

Genomic sequence data shall be available in a findable, accessible, interoperable and reusable file format for use in bioinformatics pipelines. Operators shall implement procedures to verify that the metadata and sequence are correctly associated to maintain referential integrity. Operators of private repositories shall correct errors when identified, update the records containing these errors in public repositories and remove WGS data sets when referential integrity cannot be verified.

WGS data and selected metadata may be transferred (uploaded) to a publicly accessible database. Organizations may need to transform metadata before submitting to public repositories so that details or identifiable information is not revealed. Metadata provided according to the tables and annexes of this document can be formatted to fulfil MDM requirements for submitting microbial sequences to INSDC public repositories. Further instruction is provided in Table H.1 (for NCBI/DDBJ) and Table H.2 (for EMBL-EBI).

10 Validation and verification

10.1 Validation

10.1.1 General

The performance characteristics of WGS-based methods shall be established for the intended use. Validation of the WGS workflow may be performed separately for the different components (see Table 4). However, eventually, the complete workflow shall be validated. The validation shall provide evidence that the method is repeatable, reproducible and accurate.

Table 4 — Validation of workflow stages

Validation stage	Repeatability (accuracy/precision)	Reproducibility (accuracy/precision)	Agreement with other methods (accuracy/trueness)
1. Pure culture	Include different subcultures on same day by same operator.	Include different subcultures on different days by different operators.	Include related and unrelated strains (e.g. outbreak and non-outbreak) or strains not containing the marker(s) of interest.
2. DNA extraction	Include different DNA extractions from same subculture on same day by same operator, using the same batches of reagents.	Include different DNA extractions from different subcultures by different operators, on different days, using different batches of reagents.	Include DNA of related and unrelated strains (e.g. outbreak and non-outbreak associated) or of strains not containing the marker(s) of interest.
3. DNA sequencing	Include libraries from the same strain (e.g. in triplicate) generated by same operator on the same day, in the same run (within run precision).	Include libraries generated by different operators on different days in different runs of the same instrument (between run precision), and on different instruments where possible.	Include libraries of related and unrelated strains (e.g. non-outbreak associated) or strains not containing the marker(s) of interest.
4. Bioinformatics pipeline	Demonstrate identical results from same data set at least twice on same computer/IT infrastructure, using the same version of the software with the same options/parameters.	Demonstrate comparable results from same data set at least twice on different computers such as local Linux/Unix/any OSX workstations or computing clusters or supercomputing nodes using the same version of the software with the same options/ parameters. Use of a workflow management system is recommended for such testing on different platforms.	Demonstrate results are comparable with other pipelines for the same application and specify any known differences between pipelines that can affect the outcome (e.g. built-in reference databases). If no such pipeline is available, then simulated data, where the evolutionary relationships of the isolates are known and reflect variability expected in real data, should be used to demonstrate the pipeline's ability to produce the correct answer.

Table 4 *(continued)*

Validation stage	Repeatability (accuracy/precision)	Reproducibility (accuracy/precision)	Agreement with other methods (accuracy/trueness)
5. Acceptance criteria	The interpretation of the results should not change, i.e. no significant differences should be observed while repeating the WGS workflow in the same laboratory, with the same operators using the same instrument.	The interpretation of the results should not change, i.e. no significant differences should be obtained while reproducing the WGS workflow in different laboratories, with different operators or different instruments. Minor differences, such as slight changes to N50 values with different versions of a sequencing kit, are expected. Also, genome content such as plasmids can be lost. They should not be considered significant.	The WGS workflow should be able to produce the same conclusions as other gold standard typing and/or characterization methods (e.g. epidemiological inference/concordance, differentiate unrelated strains while grouping closely related isolates, other genotypic methodology, comparable data to closed reference genomes).

10.1.2 Validation of laboratory operations

Validation of laboratory operations can go from culture up to DNA sequencing, and all stages in between, depending upon the laboratory workflow. Validation parameters and acceptance criteria for different stages are specified in Table 4.

10.1.3 Validation of the bioinformatics pipeline

Performance of bioinformatics pipelines should be assessed at every appropriate level of analysis (see Table 4). Validation may include sample data generated in the originating laboratory using a specific WGS workflow along with either standard (benchmark) data sets or simulated data sets, or both.

a) Standard data (or benchmark) sets are cases where either the origin, phenotype or epidemiological relationship of the isolates, or all, are known, and the sequence data have been made publicly available. Standard data sets can be useful for comparing output from different bioinformatics pipelines. The utility of standard data sets in establishing fit-for-purpose workflows can be limited unless they were generated using the same method of laboratory preparation including sequencing technology. Examples of standard data sets include:

   1) benchmark data sets for WGS analysis[11];

   2) National Institute of Standards and Technology microbial genomic DNA reference material sequence data[12];

   3) FDA-ARGOS[13].

b) Simulated data are created using applications that can generate synthetic sequence read data from real genome sequence data. Simulated data sets can test a wider range of parameter values and errors than are typically observed in real sequence data. Simulated data are extremely useful because known differences (e.g. nucleotide polymorphisms, indels and structural variants) can be introduced, providing confidence in the final measurements.

   NOTE Pipelines often perform better with simulated data than with real data sets.

c) Sample data sets (i.e. real sequencing data) are intended to reflect the types of organisms and microbiological procedures used in a particular laboratory. Sample data sets are generated using a specific end-to-end workflow, including DNA isolation, library preparation, sequencing and bioinformatics analysis. Sample data sets are typically derived from standard or reference collections with known characteristics, or on collections of isolates associated with an outbreak with known epidemiological information, depending on the application workflow to be validated.

Validation data sets should be comprised of data from target bacterial species that represent the complexity and errors typically encountered during intended uses. Validation data sets should include potentially confounding isolates, such as genome sequences that are either very closely or distantly related from target bacteria. Data sets can also include multiple species. Additionally, data of multiple species or strains in a single data file can be used to validate the ability to detect cross-contamination. Acceptable conditions shall be established based on performance goals and documented depending on applications, such as:

- accuracy of annotation and feature prediction;
- accuracy of strain or type predictions, assessment of relatedness consistent with known epidemiological information.

Validation data sets shall be analysed with the bioinformatics pipeline and the results shall be assessed using the established performance goals and acceptance criteria (see Table 4). Reports describing the validation results should be sufficient to replicate the analyses. Any major change in bioinformatics pipelines needs to be evaluated and documented. If a major impact is observed, a re-validation may need to be performed.

10.1.4 Validation of the end-to-end workflow

For each WGS application, an end-to-end validation shall be performed if the validation of one of the steps within the WGS workflow (see Table 4) for the intended application is missing or if the validation of the laboratory operations or bioinformatics analysis did not include sample data (see 10.1.2). Validation of the end-to-end WGS workflow helps to establish thresholds for biologically relevant differences versus differences that are linked to the culture and sequencing process. Validation of WGS workflows through comparison to historical standard reference methods (e.g. pulsed-field gel electrophoresis, 7-gene MLST, phage typing) poses a challenge because WGS provides a higher level of resolution of data. Bacterial isolates that were previously identical or indistinguishable now can have measurable differences. Appropriate sample genome sequence data sets should be created, depending on the application, i.e. isolates should be selected that represent the variability of organisms that will be analysed for specific applications.

Metrics that are linked to methodology and described by Reference [2] can be useful when characterizing differences between closely related genomes. An example for the validation of source tracking based on these metrics is illustrated in Reference [14]. An example for the validation strategy focusing specifically on the exhaustive characterization of the bioinformatics analysis of a WGS workflow is illustrated in References [15], [16] and [17]. Each stage of the workflow should be validated as described in Table 4.

10.2 Verification

10.2.1 General

The verification shall demonstrate that the executing laboratory can use the validated method for a specified WGS application correctly. Verification shall be done for the complete workflow or one of the steps within the workflow (the laboratory implementing the laboratory operations or the entity implementing the bioinformatics analysis, or both).

10.2.2 Verification of laboratory operations

The executing laboratory shall provide objective evidence within the field of application that the validated method is being used in its application area and that the specified requirements have been fulfilled.

10.2.3 Verification of the bioinformatics pipeline

If commercial or open-source bioinformatics pipelines are used that have been validated by their developers, the validation tests are published and the validation data sets are publicly available, then it is only necessary to (partially) repeat the validation test once the software has been installed. In this case, test data sets distributed by the pipeline developers may be used. However, successful execution of a test data set does not necessarily imply that a bioinformatics pipeline is validated or fit-for-purpose. Test data sets can be used to verify that bioinformatics pipelines, and their associated dependencies, are installed correctly and functioning as expected. The user needs to show functionality of the pipeline according to pre-established parameters. Test data are used as input to a bioinformatics pipeline and the output is compared against the expected results. Test data sets are typically small (e.g. lambda phage genome) and distributed with the software or pipeline. When the data are too large to bundle with the software, accession numbers of data repositories may be provided.

Annex A
(informative)

Development of quality metrics and use of controls

A.1 Guidance for development of quality metrics for short- and long-read sequencing technologies

Table A.1 — Guidance for development of quality metrics for short- and long-read sequencing technologies

Process	Concern	Guidance	
		Short-read technology	Long-read technology
DNA extraction	Cross-contamination; sample integrity.	Broth cultures should be started from a single colony of the isolate being tested.	
		DNA integrity is critical, particularly for long-read technologies. Care should be taken to avoid fragmentation of genomic DNA during preparation and storage (e.g. through freeze/thawing).	
DNA quality	Presence of impurities that can negatively impact library construction.	Optical density (OD260/280) ratio should be 1,75 to 2,05 and (OD260/230) ratio should be 2,0 to 2,2.	
	Low molecular weight DNA can negatively impact library construction.	Extraction methods for genomic DNA should be adapted to the sequencing platform being used. DNA integrity can be checked on agarose gel or via capillary electrophoresis with appropriate size standards.	
DNA quantity	Insufficient input of genomic DNA can result in substandard sequence library.	Input DNA quantity should be carefully determined using a DNA-specific, intercalating dye-based fluorescence quantification method prior to further dilution. Minimum quantity needed will be dependent on the library kit/sequencing technology used. If modified, this should be supported by validation.	
DNA fragmentation	Sub-optimal fragmentation can result in reduced library yield/reduced coverage.	Size distribution of sheared DNA samples should be checked using capillary gel electrophoresis-based systems.	
		Sample library should contain fragments between 200 bp and 3 000 bp. For transposon-based library construction, fragment distribution may be verified by capillary electrophoresis after PCR.	Optimal fragment sizes vary by long-read sequencing platform and application.
DNA size selection	Following fragmentation, selection of a specific range of fragments may be desirable to improve sequence quality/efficiency. Selection can result in loss of small plasmids or bias in sequence coverage.	This can be done using gel electrophoresis approaches or bead-based approaches. Any size selection should be supported by validation for each of the species to which this is applied.	
		Size selection increases sequencing quality but can result in gaps in the coverage of the bacterial genome.	Size selection increases sequencing quality but can result in loss of small plasmid(s).

Table A.1 *(continued)*

Process	Concern	Guidance	
		Short-read technology	Long-read technology
Ligation of indices and adapters	Correct association of adapter sequences with appropriate sample (sample mix-ups).	Ensure barcode indices used are used only once in sequencing run. Rotate indices used such that the same unique pair of indices is not used in two consecutive runs. After each use, replace caps on index tubes or seal on index plate to prevent index cross-contamination.	
Amplification	Reduced library complexity	Follow the instructions of the manufacturer as to number of cycles. If modifications are needed (e.g. to avoid primer-dimers), this needs to be validated (e.g. 12 cycles can work for most species, but 15 cycles can be better for Mycobacterium). If necessary, use a PCR-free library preparation method.	Not applicable
	Amplicon carryover-contamination	It is advisable to conduct pre-PCR and post-PCR steps in different rooms in order to avoid amplicon carryover-contamination.	Not applicable
Library quality assessment	Anticipated DNA concentration and insert size distribution	The library size distribution should be checked with a capillary electrophoresis-based system. Concentration can be determined using a capillary electrophoresis-based system or by a fluorescence-based quantification system.	
	Contamination during library preparation	Care should be taken to avoid cross-contamination during library preparation. Use aerosol resistant filter pipette tips. Change gloves frequently.	
DNA sequencing	Multiplexed sample normalization	Equimolar pooling based on library profile and quantification may be desirable to ensure adequate coverage of all of the samples included in the run. Alternately, a bead-based method of normalization may be used. Quantification of the pooled library may also be desirable to ensure that the amount of library loaded is suitable for the sequencing platform.	Equimolar pooling based on library profile and quantification. Done before size selection and last DNA damage repair steps.
	Inter-run carryover-contamination	To minimize carryover-contamination, use appropriate instrument washes and establish an index-rotation scheme to ensure that the same index pair is not used in consecutive runs.	Not applicable
	Instrument performance	Run an internal control spike in the sample. Sequencing of the same DNA (positive control) in order to monitor sequence quality is recommended.	Run an internal control spike with the sample.

A.2 Recommendation for quality assessment of short-read data

Assessment of sequencing data will vary depending on sequencing platform and on the intended use of the data in downstream analyses. This clause provides guidelines applicable to some short-read sequencing platforms (e.g. Illumina instruments). Some parameters are specific (e.g. reads passing filter) and do not necessarily apply to other sequencing platforms. It is intended to be used as an initial quality assessment, for data release purposes only, prior to starting the bioinformatics analysis which includes a more into-depth quality assessment as elaborated in 7.3.

NOTE Table A.2 provides an example for Illumina MiSeq technology, but this document is not specific to Illumina.

Table A.2 — Recommendation for quality assessment of MiSeq Illumina short-read data prior to bioinformatics analysis (see 7.3)

Process	Concern	Guidance
Sequence data quality	Raw sequence data of sufficient quality, read length and coverage for intended purpose.	Sequences in FASTQ format may be checked using FastQC tool. All sequences should be identified as either warn or pass for per base sequence quality. Minimum estimated coverage typically ranges from 20-fold to 60-fold.
Run acceptance parameters[a]	Q30 coverage overall	2x300 bp: ≥ 70 %
		2x250 bp: ≥ 75 %
		2x150 bp: ≥ 80 %
	PhiX error rate	< 6 %
	Reads passing filter	> 44 M NOTE Number of reads can vary by platform and chemistry.
	Reads negative control	< 10 000 NOTE Depends on spike-in volume of the negative control and the sequencing instrument.
Sample acceptance parameters	Estimated coverage (i.e. coverage calculated based on the number of reads and the target organism genome size using the Lander/Waterman equation)	≥ 20X (depending on application and microorganism sequenced)
	Mean Phred score before trimming	≥ 30
	Contamination	Check for expected species and absence of non-expected species/strain (< 5 % reads identified as non-target species) Databases can be biased (e.g. high-copy plasmids are assigned to another species). Performing this check at read level is a rough indication of contamination. It is recommended to extend this contamination check to the contig level or to add additional checks (see 7.3).
Run criteria (informative)	Cluster density	600 K/mm^2 to 1 400 K/mm^2 for MiSEQ NOTE Depends on sequencing chemistry.
	Clusters passing filter	> 75 %
	PhiX alignment	≥ 1 %
	Phasing/pre-phasing read-1	< 0,5 %
	Phasing/pre-phasing read-4	< 0,5 %

[a] Accepted individual sequences can be chosen instead of an entire run, based on the sample acceptance parameters in this table.

Table A.2 *(continued)*

Process	Concern	Guidance
Sample criteria (informative, can be species-dependent, should be evaluated during validation)	GC-score	< 4 % deviation from the expected GC-content for species analysed[15][17]
	Median Phred score drop Q30	≥ 150
	Reads per sample	> 20 000 (dependent on read length and genome size, coverage will give a more precise indication; may vary depending on the application and required coverage)
	Maximal N-fraction	< 0,10 %[17]
	Per base sequence content	< 6 % difference[15][17]
	AT proportion check	< 30 % (species dependent)
	Sequence length distribution	e.g. < 5 % of reads (before trimming) are < 120 bp when raw input reads are 300 bp long; > 50 % of the reads are > 150 bp when raw input reads are 300 bp long
[a] Accepted individual sequences can be chosen instead of an entire run, based on the sample acceptance parameters in this table.		

A.3 Recommended use of controls

Table A.3 — Recommended use of controls

Process	Control Description	Purpose	Guidance	Frequency of use
DNA extraction	Positive control/ reference strain representing species in the test samples.	Assess method efficiency.	Failure to extract genomic DNA of suitable quality for downstream analyses from positive control indicates that there is an issue with the extraction procedure; however, if test samples worked, they can be used.	Can be included as deemed necessary. Sequencing of positive extraction controls is not required.
	Negative control (e.g. water blank, non-inoculated broth).	Ensure that cross-contamination does not occur during DNA extraction procedure.	The negative control may be sequenced to evaluate contamination arising during DNA extraction. If the negative control is contaminated, all the DNA should be carefully evaluated to determine if the level of contamination will affect downstream analyses.	Recommended for each extraction but may only be sequenced as deemed necessary.
Library preparation	DNA from a well characterized strain should be used as a positive control. It is also recommended to consistently use the same DNA extract.	Used to monitor sequence quality on different runs to identify problems with sequencing chemistry. Used to evaluate and validate library preparation.	Positive controls should have fragments in a range that is typical for the technology being used, and results of sequencing should be consistent between runs.	A positive control is not required for every run. Frequency of use of the positive control for monitoring quality over time should be established.

Table A.3 *(continued)*

Process	Control Description	Purpose	Guidance	Frequency of use
	Negative control (e.g. water).	Used to evaluate cross-contamination occurring during library preparation.	Negative controls should have no detectable peaks and minimal sequencing reads associated with them. Contamination of negative control should be below established limits.	A negative control is not required for every run. Frequency of use of the negative control for monitoring quality over time should be established.
DNA sequencing/ instrument performance	A well-characterized spike-in internal control library.	Evaluate quality of the run.	Per base error rates should be within established limits.	If practical, control libraries should be included on each run.
	Include indexes used in previous run when de-multiplexing the run.	Evaluate level of carryover-contamination for sequencing platforms known to have this issue, evaluate cross-contamination (e.g. with amplicons from previous runs) occurring during library preparation.	Number of reads with indexes mapping to previous runs should not exceed established limits	May be included as deemed necessary.

Annex B
(informative)

Laboratory contact information fields

Table B.1 — Laboratory contact information fields

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Organization role The role played by an organization in a process (e.g. sample collection, microbial isolation, sequencing, bioinformatics).	Establishing chain of custody and for providing contact information for follow-up analyses. Contact information can vary within an organization according to the processes or analyses being performed.	The organization role can be selected from the ISO WGS Slim.
Organization name The name of the organization.	Establishing chain of custody and for providing contact information for follow-up analyses when data may be shared with public repositories or partners.	The organization name can be provided as free text.
First name A first name is a name that denotes a specific individual between members of a group of individuals, whose members usually share the same surname.	Establishing chain of custody and for providing contact information for follow-up analyses.	Where personnel turnover can affect the ability for follow up, the contact information may be supplied for a job position rather than a specific individual. This information may be provided as free text
Last name A last name (surname) is a name added to a given name and is part of a personal name and is often the family name.	Establishing chain of custody and for providing contact information for follow-up analyses.	Where personnel turnover can affect the ability for follow up, the contact information may be supplied for a job position rather than a specific individual. This information may be provided as free text.
Job title The name of the job position held by the contact.	Establishing chain of custody and for providing contact information for follow-up analyses.	Where personnel turnover can affect the ability for follow up, the contact information may be supplied for a job position rather than a specific individual. This information may be provided as free text.
Street address The street address describes the physical (geographic) location of the laboratory facility.	Establishing chain of custody and for providing contact information for follow-up analyses.	The street address can include the building number and street name. Format: validated data entry.
Municipality The name of the city, town or village in which the organization is located.	Establishing chain of custody and for providing contact information for follow-up analyses.	The municipality can be selected from the GAZ.
Province/state/territory The name of the province (P), state (S) or territory (T) in which the organization is located.	Establishing chain of custody and for providing contact information for follow-up analyses.	The province, state or territory can be selected from the GAZ.
Country The name of the country in which the organization is located.	Establishing chain of custody and for providing contact information for follow-up analyses.	The country can be selected from the GAZ.

Table B.1 *(continued)*

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Email address An email address is an identifier to send mail to an electronic mailbox.	Establishing chain of custody and for providing contact information for follow-up analyses.	The email address can be provided using validated data entry.
Telephone number A telephone number is an identifier used to connect to a physical device capable of transferring voice or data over a network.	Establishing chain of custody and for providing contact information for follow-up analyses.	Include the country and area code along with the specific number for the representative (which can include an extension number). The telephone number may be provided using validated data entry.

Annex C
(informative)

Geographic location of sample collection fields

Table C.1 — Geographic location of sample collection fields

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Latitude A measurement that is the measure of the latitude coordinates of a site.	INSDC data standard (fulfils MDM lat_lon field at NCBI/DDBJ; lat_lon OR country at EBI).	Latitude can be provided using validated data entry and should not be abstracted to the centre of a city, province/state or country as this may falsely implicate an existing location. “Missing” is an acceptable value if the information is unavailable or cannot be shared. If available, degrees latitude can be specified as d[d.dddd] N|S e.g. 38,98 N.
Longitude A measurement that is the measure of the longitude coordinate of a site.	INSDC data standard (fulfils MDM lat_lon field at NCBI/DDBJ; lat_lon OR country at EBI).	Use validated data entry to provide longitude, which should not be abstracted to the centre of a city, province/state or country as this may falsely implicate an existing location. “Missing” is an acceptable value if the information is unavailable or cannot be shared. If available, degrees longitude can be specified as d[d.dddd] W|E (e.g. 77,11 W).
Municipality The name of the city, town or village in which the organization is located.	INSDC data standard (fulfils MDM geo_loc field).	The municipality can be selected from the GAZ.
Province/state/territory The province (P), state (S) or territory (T) in which the sample was collected.	INSDC data standard (fulfils MDM geo_loc field).	The province, state or territory can be selected from the GAZ.
Country The country in which the sample was collected.	INSDC public repository MDM (geo_loc field).	The country can be selected from the GAZ.

Annex D
(informative)

Isolate passage history fields

Table D.1 — Isolate passage history fields

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Number of passages The number of serial subcultures that an isolate has grown in one environment.	An increase in the number of times an isolate has been passaged may result in the accumulation of additional mutations.	The number of passages can be expressed as a numerical value (positive integer) using validated data entry.
Passage protocol The procedure used to serially propagate an isolate in an environment.	Facilitates the comparison of methodologies, as well as analyses.	The passage protocol can include, when applicable, inoculum size, media type, temperature and duration of incubation. The passage protocol may be provided as free text.

Annex E
(informative)

Antibiogram results and methods fields

Table E.1 — Antibiogram results and methods fields

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Drug name The standard chemical name for a drug.	Antibiogram data standards (fulfils NCBI Antibiogram “Antibiotic” field).	Standard chemical or generic drug names, rather than common or brand names, can be selected from the ISO WGS Slim or the ChEBI ontology.
MIC value The numerical value of the MIC (e.g. 4).	Antibiogram data standards (fulfils NCBI antibiogram “Measurement” field).	The MIC value can be provided using validated data entry.
MIC unit The standard unit of the MIC (e.g. μg/ml).	Antibiogram data standards (fulfils NCBI Antibiogram “Measurement Unit” field).	The MIC unit can be provided using the ISO WGS Slim.
MIC sign The sign of the MIC indicates whether the concentration can be precisely determined (denoted by = sign) or is in range below (<) or above (>) the value given.	Antibiogram data standards (fulfils NCBI Antibiogram “Measurement Sign” field.	The MIC sign can be provided using the ISO WGS Slim.
Resistance phenotype The resistance phenotype of an isolate represents the interpretation of an MIC value with regard to some breakpoint threshold such as resistant (R), sensitive (S), intermediate (I), wild type (WT) or non-wild type (NWT).	Antibiogram data standards (fulfils NCBI Antibiogram “Resistance Phenotype” field).	The resistance phenotype depends on the breakpoint thresholds applied, which in turn depend on the reference standard used for interpretation. Breakpoints can also be host, organism, drug and infection site-specific. In cases where a standard requires host, organism, drug name and tissue specificity (AST breakpoint) information for the appropriate selection of breakpoints, these should be specified using the ISO WGS Slim.
Tissue specificity (AST breakpoint) The name of the tissue type used to select breakpoints from a particular standard, for the interpretation of MIC results.	Specifying criteria for breakpoint selection. Computable comparisons of methodologies.	Tissue specificity (AST breakpoint) only needs to be specified when the standard used to interpret the MIC requires this information (e.g. selecting CLSI veterinary breakpoints). Tissue specificity can be specified using the ISO WGS Slim.
Minimum drug concentration tested The lowest value of the drug’s concentration tested (e.g. 0).	Specifying range of drug tested. Computable comparisons of methodologies.	The minimum drug concentration tested can be provided as a numerical value using validate data entry. The units will be assumed to be the same as the MIC.

Table E.1 *(continued)*

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Maximum drug concentration tested The highest value of the drug's concentration tested (e.g. 128).	Specifying range of drug tested. Computable comparisons of methodologies.	The maximum drug concentration tested can be provided as a numerical value using validated data entry. The units will be assumed to be the same as the MIC.
Lab testing method The type of assay used to determine the MIC (e.g. broth dilution).	Antibiogram data standards (fulfils NCBI Antibiogram "Laboratory Typing Method" field).	The lab testing method can be specified using the ISO WGS Slim.
Lab testing reagent The commercial kit or product used to determine the MIC (e.g. E-Test). If a commercial product was not used, include the type of media used.	Antibiogram data standards (fulfils NCBI Antibiogram "Laboratory Typing Method or Reagent" field).	The lab testing reagent can be described using free text.
Lab testing standard The clinical and laboratory guidelines or standards that prescribe the threshold values for determining resistance phenotypes (e.g. CLSI).	Antibiogram data standards (fulfils NCBI Antibiogram "Testing Standard" field).	The lab testing standard can be specified using the ISO WGS Slim.
Lab testing platform The instrumentation used to determine MIC values (e.g. Vitek).	Antibiogram data standards (fulfils NCBI Antibiogram "Laboratory Typing Platform" field).	The lab testing platform can be specified using the ISO WGS Slim[6].

Annex F
(informative)

Virulence factor detection and methods fields

Table F.1 — Virulence factor detection and methods fields

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
Virulence factor name The name of the virulence factor molecule produced by a pathogen that specifically causes disease or that influences the host's function to allow the pathogen to thrive.	Specifying virulence data.	Virulence factor gene name can be included using free text.
Virulence testing protocol The procedure used to determine virulence.	Computable comparisons of methodologies.	Include, when applicable, inoculum preparation, platforms and instrumentation, conditions, cell lines and animal models. This information can be provided using free text.
Detection limit The detection limit denotes the smallest measure that can be detected with reasonable certainty for a given analytical procedure.	Facilitates the comparison of methodologies, as well as analyses.	Include the numerical cut-off (threshold) value and units for determining positive results (e.g. qPCR value, CFUs). This information can be provided using free text.

Annex G
(informative)

Sequence quality control metrics

Table G.1 — Sequence quality control metrics

Metadata field and definition	Role in data analyses/ harmonization	ISO recommendations
N50 The length such that sequence contigs of this length or longer include half the bases in the assembly.	Provides a measure of the contiguity of assemblies for assessing quality.	N50 can be reported as a numerical value in Mb (e.g. 0,75 Mb) using validated data entry.
Sequencing depth The average number of reads representing a given nucleotide in the reconstructed sequence.	Assessing quality and providing a measure of confidence in a sequence.	Sequencing depth can be reported as a numerical value as X times fold (e.g. 30x) using validated data entry.
Breadth of coverage The percentage of the genome that was sequenced to a prescribed depth of coverage (as calculated by mapping to a reference genome).	Assessing quality and providing a measure of confidence in a sequence.	Breadth of coverage can be reported as a percentage value (e.g. 95 %) to a fold of coverage (e.g. 10X) using validated data entry.
Mean contig length The count of base pairs in the average size contig of the sequence assembly.	Provides a measure of the contiguity of assemblies for assessing quality.	Mean contig length can be reported as a numerical value in Mb (e.g. 0,5 Mb) using validated data entry.
Number of contigs The total number of contiguous sequences containing all of the assembled sequence data.	Provides a measure of the contiguity of assemblies for assessing quality.	Number of contigs can be reported as a numerical value (e.g. 5) using validated data entry.
Size of assembled genome The total number of base pairs contained in assembled contigs.	Analyses and reporting sequence characteristics.	The size of the assembled genome can be reported as a numerical value in Mb (e.g. 5,2 Mb) using validated data entry.

Annex H
(informative)

Metadata specification

H.1 Metadata specification for NCBI/DDBJ submissions

The Global Microbial Identifier MDM is an internationally agreed upon metadata standard, and informs minimal metadata requirements for pathogen sequence submissions to the INSDC public repositories (i.e. GenBank, ENA, DDBJ)[1]. Due to legacy constraints, metadata implementation varies slightly between INSDC repositories. As such, submitters should follow the instructions provided by the repository. MDM requirements may be fulfilled by formatting the prescribed metadata of this document. MDM fields as defined by EBI and NCBI/DDBJ and their implementations are described below. Templates for metadata submission can be found in the BioSample guidelines of the EBI[18] and NCBI[19]. If any fields of information cannot be shared due to data sharing constraints or other reasons, "Missing" should be entered for submissions to NCBI/DDBJ. "Not included", "Not provided" or "Restricted access" are permissible null values for EBI submissions.

Table H.1 — NCBI/DDBJ MDM fields and NCBI definitions

NCBI/DDBJ MDM fields and NCBI definitions	Formatting instructions for ISO metadata
sample_name Sample name is a name for the sample. It can have any format, but it should be concise, unique and consistent within the lab, and as informative as possible. Every sample name from a single submitter shall be unique.	No special instructions.
attribute_package Specify the pathogen type. The values may be "Pathogen.cl" (for clinical or host-associated pathogen) or "Pathogen.env" (for environmental, food or other pathogen). The value provided in this field drives validation of other fields.	This field is specific to NCBI/DDBJ submissions and the options can be selected from within the submission template.
collected_by[a] Name of persons or institute who collected the sample.	This information can be derived exactly from the ISO microbiology lab field.
collection_date[a] Date of sampling, given in: "DD-Mmm-YYYY", "Mmm-YYYY" or "YYYY" format (e.g. 30-Oct-1990, Oct-1990 or 1990) or, from ISO 8601-1, "YYYY-mm-dd", "YYYY-mm" or "YYYY-mm-ddThh:mm:ss" (e.g. 1990-10-30, 1990-10 or 1990-10-30T14:41:36).	This information can be derived from the ISO collection date field. If the stakeholder must include a truncated version due to data sharing constraints, only include the year (YYYY format).
organism The most descriptive organism name for this sample (to the species, if relevant).	This information can be derived exactly from the ISO organism field.
strain Microbial or eukaryotic strain name.	This information can be derived exactly from the ISO strain field.
isolate Identification or description of the specific individual from which this sample was obtained.	This information can be derived exactly from the ISO isolate field.

[a] If the stakeholder must include a truncated version due to data sharing constraints, only include information according to the permissible granularity. "Missing" is a permissible value.

Table H.1 *(continued)*

NCBI/DDBJ MDM fields and NCBI definitions	Formatting instructions for ISO metadata
geo_loc_name[a] Geographical origin of the sample. Use the appropriate name from the list given in Reference [20]. Use a colon to separate the country or ocean from more detailed information about the location (e.g. “Canada: Vancouver” or “Germany: halfway down Zugspitze, Alps”).	This information can be derived by concatenating a subset of the ISO geographic location of sample collection fields in Annex C. Specifically, city, province/state/territory and country information should be concatenated and separated by colons.[a]
lat_lon[a] The geographical coordinates of the location where the sample was collected. Specify as degrees latitude and longitude in the format “d[d.dddd] N|S d[dd.dddd] W|E” (e.g. 38,98 N; 77,11 W).	This information can be derived by concatenating a subset of the ISO geographic location of sample collection fields in Annex C. Specifically, latitude and longitude information should be concatenated and separated by a space.
isolation_source[a] Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived.	This information can be derived by concatenating the information for sample type and related fields, separated by a colon.
host[a] The natural (as opposed to laboratory) host to the organism from which the sample was obtained. Use the full taxonomic name (e.g. *Homo sapiens*).	Not within scope of this document. Use NCBI guidance.
host_disease[a] Name of relevant disease (e.g. *Salmonella* gastroenteritis). Controlled vocabulary (see Reference [21] or [22]).	Not within scope of this document. Use NCBI guidance.

[a] If the stakeholder must include a truncated version due to data sharing constraints, only include information according to the permissible granularity. “Missing” is a permissible value.

H.2 Metadata specification for EBI submissions

Table H.2 — Metadata specification for EBI submissions

EBI MDM fields and EBI definitions	Formatting instructions for ISO metadata
collected_by[a] Name of persons or institute who collected the specimen.	This information can be derived exactly from the ISO microbiology lab field.
collection_date[a] The date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated, i.e. all of these are valid ISO 8601-1:2019 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.	This information can be derived from the ISO collection date field. If the stakeholder must include a truncated version due to data sharing constraints, only include the year (YYYY format).
isolate[a] Individual isolate from which the sample was obtained.	This information can be derived exactly from the ISO isolate field.
geographic location (country and/or sea)[a] The geographical origin of the sample as defined by the country or sea. Country or sea names should be chosen from the INSDC country list[20]	This information can be derived from the ISO country field in Annex C. For ocean names, use the terms found in the INSDC country list[20].

[a] If the stakeholder must include a truncated version due to data sharing constraints, only include information according to the permissible granularity. “Not included”, “Not provided” or “Restricted access” are permissible null values.

Table H.2 *(continued)*

EBI MDM fields and EBI definitions	Formatting instructions for ISO metadata
geographic location (latitude)[a] The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system.	This information can be derived exactly from the ISO latitude field in Annex C.
geographic location (longitude)[a] The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system.	This information can be derived exactly from the ISO longitude field in Annex C.
is the sequenced pathogen host associated? Determines whether the sequenced pathogen host is associated ("Yes" or "No").	This field is specific to EBI submissions. If the organism was host associated, put "Yes". If the organism was not host associated, and was obtained from an environmental sample, put "No".
environmental_sample Identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE or other anonymous methods) with no reliable identification of the source of the organism.	This field is specific to EBI submissions. If the organism was host associated, put "No". If the organism was not host associated, and was obtained from an environmental sample, put "Yes".
specific_host[a] Natural (as opposed to laboratory) host to the organism from which sample was obtained (or "free-living" if not host-associated).	Not within scope of the document. Use EBI guidance.
host_disease_status[a] Health status of the host at the time of sample collection.	Not within scope of the document. Use EBI guidance.

[a] If the stakeholder must include a truncated version due to data sharing constraints, only include information according to the permissible granularity. "Not included", "Not provided" or "Restricted access" are permissible null values.

Annex I
(informative)

Instructions for ontology slim integration by software developers

I.1 General

Standardization of digital data using controlled vocabularies and ontologies is considered to be a best practice for data stewardship[23][24]. The ISO WGS Slim was created to gather relevant fields and values from existing, community-supported ontologies (e.g. GenEpiO and FoodOn) which are relevant to WGS-based food microbiology. GenEpiO is an application ontology that contains fields and values for genomics, laboratory, clinical, environmental, and epidemiological data and processes[7]. The Food Ontology (FoodOn) is a domain ontology that describes food products, as well as processes for cooking, preservation, packing/wrapping of food, anatomical sources, cultural and geographical origin, consumer groups and more[8]. FoodOn also contains higher level food categories imported from many existing food classification schemes [e.g. FoodEx2, USDA National Nutrient SR Legacy database, European Food Information Resource (Eurofir), FDA Code of Federal Regulations (CFR) products list][8][9][10]. GenEpiO and FoodOn have been developed by a community of experts. Further information can be obtained from http://foodon.org/.

Ontology-derived fields and values facilitate metadata harmonization integration, reuse and exchange by providing standardized terms, definitions and universal IDs (URIs) which better enable information to be processed by both humans and computers. Furthermore, ontologies encode computational logic which can be used by software systems to improve automation and more complex querying. The hierarchical nature of ontologies also better enables aggregation of data and comparisons of information at different levels of granularity. As such, the ISO WGS Slim can be used to provide metadata descriptors as prescribed in Tables I.1, I.2 and I.3.

In some cases, the ISO WGS Slim may not contain the breadth of vocabulary required. In this case, other ontologies are recommended. Specifically, geographic, taxonomic, environmental (built and natural) and drug name fields in this document can require more extensive vocabulary available directly in the GAZ, NCBITaxon, EnvO and CheBI ontologies, respectively. The GenEpiO vocabulary has largely been sourced from these ontologies and so URIs will be compatible. Further information can be obtained from www.obofoundry.org. GAZ, NCBITaxon, FoodOn, EnvO and CheBI ontologies can be downloaded from Github:

- https://github.com/EnvironmentOntology/gaz
- https://github.com/obophenotype/ncbitaxon
- https://github.com/FoodOntology/foodon
- https://github.com/EnvironmentOntology/envo
- https://github.com/ebi-chebi/ChEBI

I.2 Advice for implementing the ISO WGS Slim

I.2.1 General

The ISO WGS Slim can be downloaded from Github[9] in tab-delimited, JSON and YAML formats. It is possible that the plain-text.tsv (tabular) format options are the most accessible for software developers as they have the simplest structure. All ISO WGS Slim formats include the label and definition of each term, synonyms, optional field information or help text, and in some cases numeric and textual field validation constraints. If terms are required in addition to those contained in the slim, the GAZ, NCBITaxon, FoodOn, EnvO and CheBI ontologies can also be downloaded from GitHub.

It is impracticable to provide instructions for ISO WGS Slim implementation for all computing infrastructure scenarios. However, tabular data are commonly shared in a SQL database or spreadsheet format; therefore, this annex provides an overview of how to address these situations.

I.2.2 Spreadsheet ontology integration

There are tools that enable the creation of spreadsheets that contain drop-down menus of vocabulary. Examples of such tools that can integrate ontology terms and IDs include: Webulous (https://github.com/EBISPOT/webulous), Populous (http://e-lico.eu/populous.html) and Kusp (https://www.scibite.com/wp-content/uploads/2018/06/Kusp-DataSheet-2018.pdf). These tools offer step-by-step instructions for creating tabular data collection instruments.

Although currently there are no dominant standards for ontology annotation of tabular data, advice is provided below.

a) Each data column should be associated with an ontology ID specifying the type of information in that field. Using both the label and ontology ID enables automated mapping to other databases' fields that can use alternative labels but the same ontology identifier.

   NOTE 1 Fields describing numerical values with associated units require unit ontology identifiers as well, and can require an additional column dedicated to recording units if they vary between values in a column (e.g. MIC units such as ug/ml and mm). A worked example is provided in I.3.1.

b) Tabular data implementation depends on the reliable mapping of database fields and/or categorical field values to ontology term identifiers available online as IRIs (e.g. "http://purl.obolibrary.org/obo/HP_0012735"). Within a given database, a reference to a term identifier can usually be abbreviated into a prefix:suffix format (e.g. "HP:0012735") in which the prefix abbreviates the leading or "namespace" component of the term URL (e.g. where "HP": abbreviates the Human Phenotype Ontology space, "http://purl.obolibrary.org/obo/HP". This involves associating the tabular data with a list of (allowed) ontology prefixes and their associated namespace IRI components.

   NOTE 2 The JSON-LD (JSON Linked Data) format has this "compact IRI" functionality included.

c) When creating picklists from ontology terms, it is possible that negative values need to be added by the software developer as the slim does not include terms such as "Missing", "Not collected", "Not applicable", etc. as options. However, applications that automate the transformation of metadata for sharing (with other public or private repositories) should avoid concatenating multiple "missing" terms in a single field. For example, in the case of the NCBI submission, if the metadata includes sample type information such as the term "food", but food product and food processing information is missing, data submitters should simply include "food" in the "isolation_source" field rather than "food: missing: missing". Similarly, "food: chicken nuggets: missing" should simply be submitted as "food: chicken nuggets". However, if no sample source information is available, submitters should include a single "missing" term for "isolation_source". These guidelines also apply to concatenating other metadata fields.

An example form rendering of the ISO WGS specification is available by visiting https://genepio.org/geem/form.html. The ontology identifier GENEPIO:0002083 points to GenEpiO term "draft sequence repository contextual data standard", a term under which the following components are organized: laboratory contact information, sample collection, isolate and isolate passage history, food specimen, antibiogram, sequencing and sequence assembly quality metrics. Further examples of integrating ontology within IT infrastructure, or for storing ontology-enabled data, are available at the High-throughput Sequencing Computational Standards for Regulatory Sciences (HTS-CSRS) project website (https://hive.biochemistry.gwu.edu/htscsrs/biocompute), Vanderbilt University's REDCap data management system (https://www.project-redcap.org/), Stanford's CEDAR project (https://

metadatacenter.org), the Allotrope Foundation (https://www.allotrope.org/) data models, and in other tabular data management tools such as Karma (https://usc-isi-i2.github.io/karma/).

NOTE 3 Ontology resources grow and are refined over time. Additional training by IT support staff to understand how to access ontology terms and how to manage or refresh terms from ontology source files can be required when implementing the ISO ontology slim in different systems within an organization.

I.2.3 SQL database ontology integration

SQL is a domain-specific language used in programming and designed for managing data held in a relational database management system (data organized into tables, linked by defined relationships). There are two basic strategies for annotating and harmonizing information stored in SQL databases using the ISO WGS Slim. The first strategy is to export information to be shared as a spreadsheet (or csv file) and annotate according to the suggestions discussed above.

The second strategy involves mapping information to ontologies directly in the database. Field names (labels) may be replaced with ontology IDs, and ontology IDs can be stored in a look-up table. It is possible that a SQL database lookup table already exists which has numeric keys that can be converted to ontology URIs. As such, the lookup table can be populated with ISO WGS slim content. Alternatively, ISO WGS Slim content can be implemented via a script that accesses an ontology lookup service API. A worked example is provided in I.3.2.

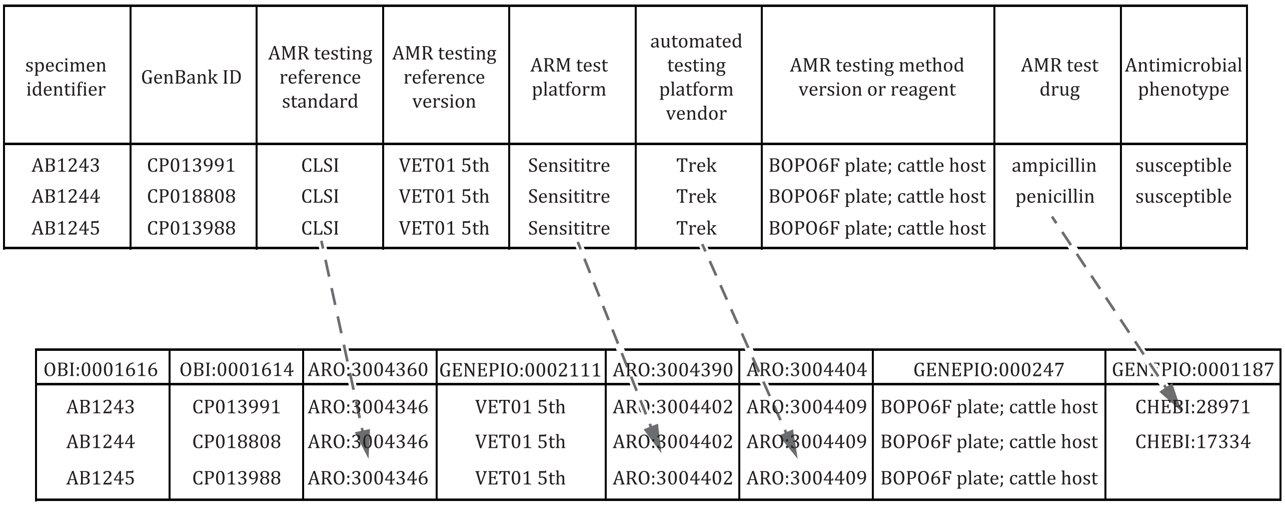
I.3 Approaches for ontology integration in systems for metadata capture and management

I.3.1 Spreadsheets

Annotating spreadsheet data with standardized ontology terms can be achieved by having a separate mapped sheet “ontology view” which has a 1-1 cell correspondence to the original sheet. An example illustrating original values mapped to ontology IDs is shown in Figure I.1.

NOTE The column headers are also replaced by ontology identifiers (numeric and free text values remain unchanged).

In the example in Figure I.1, the field “Genbank ID” should be associated with the ISO WGS Slim ontology term “http://purl.obolibrary.org/obo/NCIT_C43685”. Similarly, the antimicrobial resistance reference standard “CLSI” in the original data is mapped to the ISO WGS Slim ontology term ID “ARO:3004366”, while the drug name “penicillin” in the spreadsheet is mapped to the ontology ID “CHEHI:17334”.



specimen identifier	GenBank ID	AMR testing reference standard	AMR testing reference version	ARM test platform	automated testing platform vendor	AMR testing method version or reagent	AMR test drug	Antimicrobial phenotype
AB1243	CP013991	CLSI	VET01 5th	Sensititre	Trek	BOPO6F plate; cattle host	ampicillin	susceptible
AB1244	CP018808	CLSI	VET01 5th	Sensititre	Trek	BOPO6F plate; cattle host	penicillin	susceptible
AB1245	CP013988	CLSI	VET01 5th	Sensititre	Trek	BOPO6F plate; cattle host		

OBI:0001616	OBI:0001614	ARO:3004360	GENEPIO:0002111	ARO:3004390	ARO:3004404	GENEPIO:000247	GENEPIO:0001187
AB1243	CP013991	ARO:3004346	VET01 5th	ARO:3004402	ARO:3004409	BOPO6F plate; cattle host	CHEBI:28971
AB1244	CP018808	ARO:3004346	VET01 5th	ARO:3004402	ARO:3004409	BOPO6F plate; cattle host	CHEBI:17334
AB1245	CP013988	ARO:3004346	VET01 5th	ARO:3004402	ARO:3004409	BOPO6F plate; cattle host	

Figure I.1 — Example of original values mapped to ontology IDs

I.3.2 SQL database ontology integration

Table I.1 presents an example of a table from an organization’s relational database links laboratory services to the database ID and the ontology ID. The type of services offered by “ACME laboratory services” are defined by the ontology term “bioinformatics analysis service” replaced by the ontology ID “GENEPIO:0002223”, a term in the ISO WGS Slim.

NOTE The label has been replaced by the ontology ID in Table I.1.

Table I.1 — Example of an organization table

organization id	name	laboratory service
123	ACME laboratory services	GENEPIO:0002223
etc.		

The ontology ID “GENEPIO:0002223” can be linked to its label through a term lookup table, as shown in Table I.2. The “parent id” term refers to the more general term “specimen-related service” which has the ontology ID “GENEPIO:0002225”, where “bioinformatics analysis service” is a particular type/value for the field “laboratory service”. Other types of laboratory services are also listed, all with the same parent ontology ID but with different ontology term IDs.

Table I.2 — Example of a term_lookup table

ontology id	parentontology id	label
GENEPIO:0002225		specimen-related service
GENEPIO:0002223	GENEPIO:0002225	bioinformatics analysis service
GENEPIO:0002224	GENEPIO:0002225	isolate preparation service
OBI:0001904	GENEPIO:0002225	sequencing service
etc.		

Other associations between ontology terms (rather than field values/content) may be recorded in a separate ontology term table, such as in Table I.3, which illustrates links between “ontologized” organization name and associated services, IDs, etc.

Table I.3 — Example of an ontology_metadata table

table_name	field_name	ontology_id	ontology_label
organization	id	NCIT:C93401	organization identifier
organization	service	GENEPIO:0002225	specimen-related service
organization	name	NCIT:C93874	organization name
organization	etc.	etc.	
etc.			

Bibliography

[1] WIELINGA, P.R. et al. Global Microbial Identifier. In: X. Deng, H.C. den Bakker and R.S. Hendriksen (eds). *Applied Genomics of Foodborne Pathogens*. Food Microbiology and Food Safety series. Springer International Publishing, 2017

[2] VAN BELKUM, A. et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection* [online]. *Science Direct*. 2007 13(Suppl 3), 1-46. [viewed 2021-02-28]. Available from: https://doi.org/10.1111/j.1469-0691.2007.01786.x

[3] CHAIN, P.S.G. et al. Genome Project Standards in a New Era of Sequencing. *Science* [online]. AAAS. 2009. 326(5950), 236-237 [viewed 2021-02-28]. Available from: https://science.sciencemag.org/content/326/5950/236

[4] DALEY, T., SMITH A.D. Predicting the molecular complexity of sequencing libraries. *Nature Methods* 2013 [online]. *Nature Research Journals*. 10(4):325-7. [viewed 2021-02-28]. Available from: https://www.nature.com/articles/nmeth.2375

[5] PIGHTLING, A. et al. Choice of Reference Sequence and Assembler for Alignment of *Listeria monocytogenes* Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses *PLoS One* 2014 [online]. 9(8). [viewed 2021-02-28]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104579

[6] PIGHTLING, A.W. et al. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Frontiers In Microbiology* 2018 [online]. 9. [viewed 2021-02-28]. Available from: https://doi.org/10.3389/fmicb.2018.01482

[7] GRIFFITHS, E. et al. Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance. *Frontiers in Microbiology* [online]. June 2017. 8: 1068. [viewed 2021-02-28]. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.01068/full

[8] DOOLEY, D.M. et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food* [online]. December 2018, 2 (article 23). [viewed 2021-02-28]. Available from: https://www.nature.com/articles/s41538-018-0032-6

[9] GenEpiO/ISO2017 [viewed 2021-02-28]. Available from: https://github.com/GenEpiO/iso2017

[10] FoodEx2 [viewed 2021-02-28]. Available from: https://www.efsa.europa.eu/en/data/data-standardisation

[11] Benchmark datasets for WGS analysis [viewed 2021-02-28]. Available from: https://github.com/WGS-standards-and-analysis/datasets

[12] National Institute of Standards and Technology microbial genomic DNA reference material sequence data [viewed 2021-02-28]. Available from: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA252728

[13] FDA-ARGOS [viewed 2021-02-28]. Available from: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA231221

[14] PORTMANN, A-C. et al. A Validation of an End-to-End Whole Genome Sequencing Workflow for Source Tracking of *Listeria monocytogenes* and *Salmonella enterica*. *Frontiers in Microbiology* 2018 [online] 9 (article 446) [viewed 2021-02-28]. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00446/full

[15] BOGAERTS, B. et al. Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National

Reference Center: *Neisseria meningitidis* as a Proof-of-Concept. *Frontiers in Microbiology* 2019 [online] 10 (article 362) [viewed 2021-02-28]. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00362/full

[16] BOGAERTS, B. et al. A bioinformatics WGS workflow for clinical *Mycobacterium tuberculosis* complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and in silico approaches. *J Clin Microbiol*. 2021 [online];59(6) [viewed 2021-08-30]. Available from: https://journals.asm.org/doi/10.1128/JCM.00202-21

[17] BOGAERTS, B. et al. Validation strategy of a bioinformatics whole genome sequencing workflow for Shiga toxin-producing *Escherichia coli* using a reference collection extensively characterized with conventional methods. *Microbial Genomics* 2021 [online] 7(3) [viewed 2021-08-30]. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000531

[18] European Nucleotide Archive Sample checklists [viewed 2021-02-28]. Available from: https://www.ebi.ac.uk/ena/browser/checklists

[19] National Center for Biotechnology Information BioSample Packages [viewed 2021-02-28]. Available from: https://www.ncbi.nlm.nih.gov/biosample/docs/packages/

[20] International Nucleotide Sequence Database Collaboration. controlled vocabulary for /country qualifier [viewed 2021-02-28]. Available from: https://www.insdc.org/documents/country-qualifier-vocabulary

[21] Bioportal Human Disease Ontology [viewed 2021-02-28]. Available from: https://bioportal.bioontology.org/ontologies/DOID

[22] National Center for Biotechnology Information MeSH. [viewed 2021-02-28]. Available from: https://www.ncbi.nlm.nih.gov/mesh

[23] LAMBERT, D. et al. Baseline Practices for the Application of Genomic Data Supporting Regulatory Food Safety. *Journal of AOAC INTERNATIONAL* [online]. May 2017, 100(3), [viewed 2021-02-28]. Available from: https://academic.oup.com/jaoac/article/100/3/721/5654203

[24] WILKINSON, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [online]. March 2016, 3 article 160018. [viewed 2021-02-28]. Available from: https://www.nature.com/articles/sdata201618#citeas

[25] ISO 8601-1:2019, *Date and time — Representations for information interchange — Part 1: Basic rules*

[26] ISO 11238:2018, *Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated information on substances*

[27] ISO 20397-2:2021, *Biotechnology — Massively parallel sequencing — Part 2: Quality evaluation of sequencing data*

[28] ISO/IEC 11179-1:2015, *Information technology — Metadata registries (MDR) — Part 1: Framework*

[29] ISO 16140-1:2016, *Microbiology of the food chain — Method validation — Part 1: Vocabulary*

[30] ISO 16140-3:2021, *Microbiology of the food chain — Method validation — Part 3: Protocol for the verification of reference methods and validated alternative methods in a single laboratory*

Bureau of Indian Standards

BIS is a statutory institution established under the *Bureau of Indian Standards Act*, 2016 to promote harmonious development of the activities of standardization, marking and quality certification of goods and attending to connected matters in the country.

Copyright

Review of Indian Standards

Amendments are issued to standards as the need arises on the basis of comments. Standards are also reviewed periodically; a standard along with amendments is reaffirmed when such review indicates that no changes are needed; if the review indicates that changes are needed, it is taken up for revision. Users of Indian Standards should ascertain that they are in possession of the latest amendments or edition by referring to the website-www.bis.gov.in or www.standardsbis.in.

This Indian Standard has been developed from Doc No.: FAD 31 (22037).

Amendments Issued Since Publication

Amend No.	Date of Issue	Text Affected

BUREAU OF INDIAN STANDARDS

Headquarters:

Manak Bhavan, 9 Bahadur Shah Zafar Marg, New Delhi 110002
Telephones: 2323 0131, 2323 3375, 2323 9402 *Website*: www.bis.gov.in

Regional Offices:		*Telephones*
Central	: 601/A, Konnectus Tower -1, 6th Floor, DMRC Building, Bhavbhuti Marg, New Delhi 110002	2323 7617
Eastern	: 8th Floor, Plot No 7/7 & 7/8, CP Block, Sector V, Salt Lake, Kolkata, West Bengal 700091	2367 0012 2320 9474
Northern	: Plot No. 4-A, Sector 27-B, Madhya Marg, Chandigarh 160019	265 9930
Southern	: C.I.T. Campus, IV Cross Road, Taramani, Chennai 600113	2254 1442 2254 1216
Western	: Plot No. E-9, Road No.-8, MIDC, Andheri (East), Mumbai 400093	2821 8093

Branches : AHMEDABAD. BENGALURU. BHOPAL. BHUBANESHWAR. CHANDIGARH. CHENNAI. COIMBATORE. DEHRADUN. DELHI. FARIDABAD. GHAZIABAD. GUWAHATI. HIMACHAL PRADESH. HUBLI. HYDERABAD. JAIPUR. JAMMU & KASHMIR. JAMSHEDPUR. KOCHI. KOLKATA. LUCKNOW. MADURAI. MUMBAI. NAGPUR. NOIDA. PANIPAT. PATNA. PUNE. RAIPUR. RAJKOT. SURAT. VISAKHAPATNAM.

Published by BIS, New Delhi