TECHNICAL SPECIFICATION

ISO/TS 24420

First edition
2023-05

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

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

Reference number
ISO/TS 24420:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC *276*, *Biotechnology*.

Any feedback or questions on this document should be directed to the user’s national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Shotgun metagenomic sequencing of organisms' genomes from a complex sample is widely used in life science and clinical applications (e.g. human complex disease associated analysis, environmental microecology and other fields) in order to gain knowledge of their composition and function. It has potential to provide significant scientific data for life science research.

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

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

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

1 Scope

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

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

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

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

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

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

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

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
attribute value
value associated with an attribute instance

[SOURCE: ISO 21962:2003, 1.5.2.3]

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

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

[SOURCE: ISO 17115:2007, 2.7.1]

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

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

[SOURCE: ISO 20691:2022, 3.6]

3.6
encoding
process of assigning code to things or concepts

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

3.8
data format
arrangement of data according to preset specifications

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

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

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

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

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

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

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

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

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

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

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

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

[SOURCE: ISO/IEC 11179-1:2015, 3.1.3]

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

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

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

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

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

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

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

where p is the estimated probability of the base call being wrong.

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

Note 3 to entry: A quality score of 30 represents an error rate of 1 in 1 000, with a corresponding call accuracy of 99,9 %.

Note 4 to entry: Higher quality scores indicate a smaller probability of error. Lower quality scores can result in a significant portion of the reads being unusable. Low quality scores may also indicate false-positive variant calls, resulting in inaccurate conclusions.

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

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

3.19
relative abundance
fraction of a single microorganism operational taxonomic unit in the total microbial community of a defined environment

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

3.20
repeatability requirement
requirement of consistency under a set of repeatable measurement conditions

3.21
scaffold
reconstructed genomic sequence created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome

3.22
sequence assembly
processing, aligning and merging individual sequencing reads in order to reconstruct longer DNA sequences, entire genes or genomes

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

3.23
shotgun metagenomic sequencing
shotgun metagenomics
nucleotide sequence determination of the genomes of untargeted cells in communities in order to determine community composition and function

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

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

3.24
specialized attribute
attribute that is unique for each sample in the directory

4 Processing workflow

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

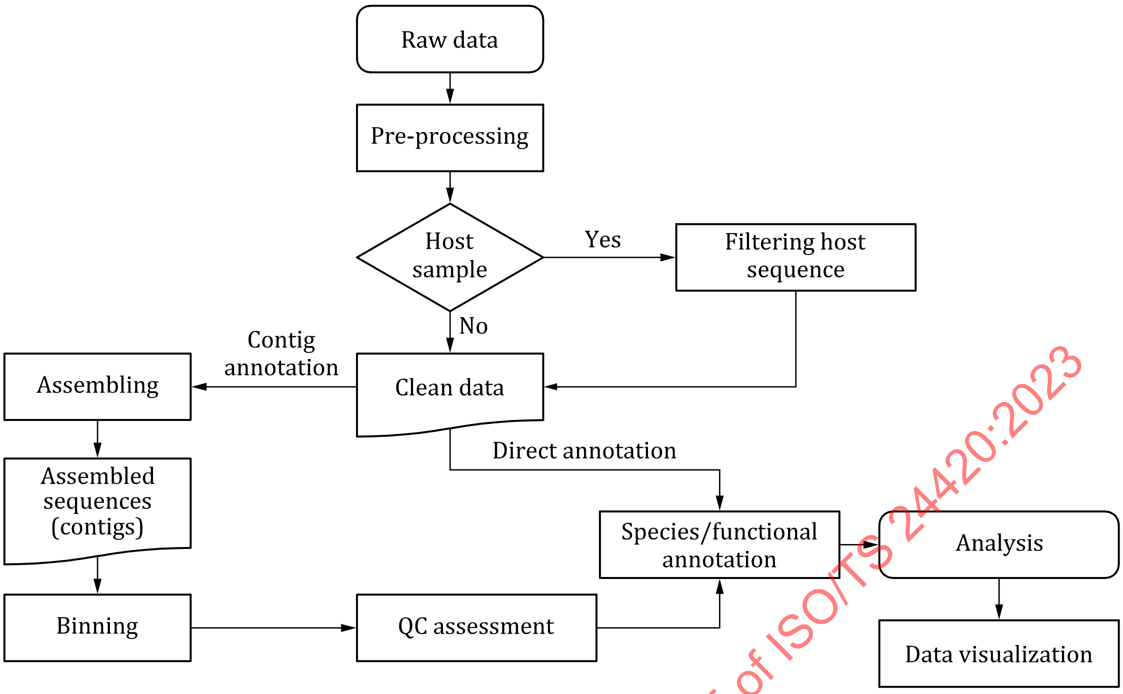

Figure 1 — Workflow of metagenomic data processing

5 Data processing

5.1 Facilities and software requirements

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

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

5.2 Sequence quality control and error determination

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

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

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

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

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

5.2.3 For human or animal or plant or all sourced samples, the host-reads shall be removed by mapping to a human or animal or plant genome reference, such as UniRef, Unified Human Gastrointestinal Protein catalog. Only clean data should be used in further bioinformatic analysis.

5.2.4 Detection and elimination of repeats and sequencing errors shall be performed as the first step in data processing. The following factors and situations shall be considered in the process of elimination.

a) Mismatch, insertion or missing (indels) (only when a reference genome is available) and uncertain bases (N characters).

b) Unrecognizable sequence, which can be caused if the reads extend to the 3'end of the adaptor when the target sequence is shorter.

c) PCR biases in the library preparation in accordance with ISO 20397-1:2022, 5.8.

5.3 Sequence assembly

5.3.1 The depth of sequencing shall be evaluated before the sequence assembly, which should take the complexity of the sample into account.

5.3.2 Samples lacking a reference genome dataset, such as soil or ocean samples, should use sequence assembly.

5.3.3 Contigs or scaffolds or both that are directly obtained from sequence fragments without any reference should be regarded as *de novo* assemblies.

5.3.4 The selection of the sequence assembly software should depend on the relative importance of the accuracy, contigs' size, input data type, and available computational resources.

5.3.5 A non-redundant gene catalogue can be obtained by predicting genes from assembled contigs. For well characterized microbiomes, e.g., human gut-borne, a credible gene catalogue (e.g., Integrated Gene Catalog (IGC)) can be used for quick identification and quantification of data from metagenomic sequencing.

5.3.6 Created assemblies should be evaluated to assess their quality, e.g., QUAST.

6 Data analysis

6.1 Annotation

6.1.1 The annotation methods should be described. The number of reference genomes selected, and reference genomes or reference database used for the annotation should be documented.

6.1.2 Taxonomy profile methods should be chosen according to data and application needs to obtain a higher-level taxonomy profile (e.g., species, genus, order, phylum) including metagenomic linkage groups (MLG), metagenomic clusters (MGC) or metagenomic species (MGS).

6.1.3 Taxonomy profiling should base on reference databases, such as RefSeq complete genomes (RefSeq CG) for microbial species and the BLAST®[1] databases for high-quality nucleotide and protein

[1] BLAST® is the trademark of a product supplied by the National Center for Biotechnology Information (US). This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named. Equivalent products may be used if they can be shown to lead to the same results.

sequences. Classification accuracy, speed, and computational requirements should be taken into account when select taxonomic classification tools.

6.1.4 If the profile is obtained by a de novo sequence assembly method, the species information should be identified when the alignment with a sequence similarity of more than 97 % and the coverage of more than 90 % to the most related reference database is determined.

6.1.5 For read-based approaches, the read should do mapping to non-redundant protein sequence database (NR) for taxonomy data (e.g., BLAST®, Diamond or Last) or marker genes after read merge.

6.1.6 Metagenomic profiles should be annotated to various levels according to the reference annotation, i.e., species, genus, or higher.

6.2 Calculation of species relative abundance

6.2.1 Species analysis

6.2.1.1 The relative abundance calculation method should be defined and implemented to meet the repeatability requirement. The relative abundance calculation method shall be documented.

6.2.1.2 Calculation tools should be selected with consideration to reflect the actual relative abundance of the target operational taxonomic unit in the sample.

6.2.2 Gene analysis

6.2.2.1 A gene abundance table can be generated using alignment-based tools or alignment-free methods.

6.2.2.2 The relative abundance distribution at the gene level can be obtained by comparing the clean data to the assembled gene set or the appropriate reference database.

6.2.2.3 Superposition of the relative abundance of the gene sequence of the same species shall be done to get the operational taxonomic unit.

7 Data archive and metadata

7.1 Original data

7.1.1 Sequencing data volume should be evaluated to obtain saturated gene information in metagenome-wide association studies (MWAS).

7.1.2 Regardless of the sample source, the sequencing data format should be the same for each sequence. The sequence data format should be stored in a standard format that can preserve the information of biological sequences (usually nucleic acid sequences) or their sequencing quality; ISO 20691:2022, A.2 includes information on data formats for OMICS, biochemical and molecular biology methods.

7.2 Sequencing analytical data

7.2.1 The sequencing analytical data shall include read counts, total relative abundance of genus, total relative abundance of function, and total relative abundance of species.

7.2.2 Sequencing analytical data shall be text formatted and contain a header and data rows, examples are shown in Annex A, such as total relative genus abundance (see Table A.1), total relative species abundance (see Table A.2), bacterial diversity index (see Table A.3), enterotypes index (see Table A.4), bacterial relative abundance (see Table A.5), total relative gene abundance (see Table A.6). A compressed format is recommended, e.g., GZ or Zstd.

7.3 Data directory and archive

7.3.1 General

Data description and metadata should be compliant with ISO 20691:2022, Clause 4 and Clause 5.

7.3.2 Directory of data elements

7.3.2.1 A directory of data elements should include public attributes and specialized attributes.

7.3.2.2 Public attributes are the same as other genomic data, requiring the following entries: version, registration authority, classified mode, responsible organization and submitting organization.

7.3.2.3 Specialized attributes should include sequencing information, biological analysis information, metagenomic information and quality control information. Examples of each attribute are shown in Annex B (see Table B.1, Table B.2 and Table B.3).

7.3.3 Data archiving

A directory of metagenomic data can be divided into an original data directory and an analysis results directory. Both should be a multi-level directory. Each level of the directory shall include directory identifier, directory name, directory definition and parent directory.

7.4 Metadata

7.4.1 Metadata can be stored structurally as a flat data file during data processing, including phenotypic information, quality control information, preanalytical information (collection method, use of stabiliser(s), delays between collection and processing, temporary storage conditions, DNA isolation method and method parameters, DNA quality control method and results) etc.

7.4.2 A consistency check of metadata and sequencing data shall be done to check whether the samples of sequencing data conform to the samples described by metadata, e.g., data omissions, duplications, error matching, and error tagging.

7.4.3 A metagenomic related metadata directory should include a sequence information directory, a bioinformatic analytic directory and a metagenomic directory. These directories should include information as described below.

a) The sequence information directory should include metadata describing the sequencing progress, such as sequencing task list name, sequencing type, sequencing platform name, sequencer identifiers, name of sequencer, and chip number.

b) The bioinformatic analytic directory should include metadata that describes the bioinformatic analysis software, such as version and parameters.

c) The metagenomic directory should include metadata that describes the metagenomic analysis process and results, such as code, name, value, length, integrity description, phylum, genus, species and intestine type.

7.4.4 Sequencing data shall have metadata to ensure the traceability and integrity of the sample information. Metadata information can comply with "The minimum information about a genome sequence (MIGS)" and "The minimum information about a metagenome sequence (MIMS)"[7].

7.4.5 There shall be a check to keep the correspondence between metadata and sequencing data.

Annex A
(informative)

Examples of data format

A.1 Total relative genus abundance

A.1.1 General

Data can be text format and showed with header line and data line. Total relative genus abundance information in metagenomics contains in data line.

A.1.2 Header line

The header line has N + 2 (N is the number of reference samples) fixed fields, each field is connected with a tab, like following:

[1]. # Scientific_Name

[2]. sample_relative_abundance_value

[3]. reference_sample_relative_abundance_value_1

……

[N]. reference_sample_relative_abundance_value_N

A.1.3 Data line

Each line has N + 2 (N is the number of reference samples) fixed fields, each field is connected by a tab, and the null value of each field uses the point "." to represent. The respective field information is as following.

[1]. Scientific_Name: The genus name of bacteria, string type, and use NCBI Taxonomy Database for description, refer to https://www.ncbi.nlm.nih.gov/taxonomy

[2]. sample_relative_abundance_value: Total relative genus abundance values of the sample, show in decimal fraction.

[3]. reference_sample_relative_abundance_value_1: Total relative genus abundance values of the reference sample 1, show in decimal fraction.

……

[N]. reference_sample_relative_abundance_value_N: Total relative genus abundance values of the reference sample N, show in decimal fraction.

A.1.4 Example table

Table A.1 — Examples of total relative genus abundance

List heading	Example 1	Example 2
# Scientific_Name	Abiotrophia	Acetivibrio
sample_relative_abundance_value	3,079 903e-06	4,047 342e-05
reference_sample_relative_abundance_value _1	3,227 675e-06	1,407 130e-05
……	……	……
reference_sample_relative_abundance_value _N	0	4,641 824e-06

A.2 Total relative species abundance

A.2.1 General

Data can be text format and showed with header line and data line. Total relative species abundance information in metagenomics contains in data line.

A.2.2 Header line

The header line has N + 2 (N is the number of reference samples) fixed fields, each field is connected with a tab, like following:

[1]. # Scientific_Name

[2]. sample_relative_abundance_value

[3]. reference_sample_relative_abundance_value_1

……

[N]. reference_sample_relative_abundance_value_N

A.2.3 Data line

Each line has N + 2 (N is the number of reference samples) fixed fields, each field is connected by a tab, and the null value of each field uses the point "." to represent. The respective field information is as following.

[1]. Scientific_Name: The species name of bacteria of bacteria, string type, and use NCBI Taxonomy Database for description, refer to https://www.ncbi.nlm.nih.gov/taxonomy

[2]. sample_relative_abundance_value: Total relative species abundance values of the sample, show in decimal fraction.

[3]. reference_sample_relative_abundance_value_1: Total relative species abundance values of the reference sample 1, show in decimal fraction.

……

[N]. reference_sample_relative_abundance_value_N: Total species relative abundance values of the reference sample N, show in decimal fraction.

A.2.4 Example table

Table A.2 — Examples of total relative species abundance

List heading	Example 1	Example 2
# Scientific_Name	Abiotrophia defectiva	Achromobacter piechaudii
sample_relative_abundance_value	1,163 749 42e-05	0
reference_sample_relative_abundance_value _1	3,737 392 6e-06	0
……	……	……
reference_sample_relative_abundance_value _N	2,180 657e-06	1,444 14e-07

A.3 Bacterial diversity index

A.3.1 General

Data can be text format and showed with header line and data line. The bacterial diversity index information in metagenomics contains in data line.

A.3.2 Header line

The header line has 4 fixed fields, each field is connected with a tab, like following:

[1]. # sample_name

[2]. genes_count

[3]. shannon_index

[4]. diversity_index_ratio

A.3.3 Data line

Each line has four fixed fields, each field is connected by a tab, and the null value of each field uses the point "." to represent. The respective field information is as following:

[1]. # sample_name: sample name, string type.

[2]. genes_count: total number of gene, round number.

[3]. shannon_index: Shannon's Diversity Index, show in decimal fraction.

[4]. diversity_index_ratio: The proportion of group below the Shannon index, show in decimal fraction and values are between 0 and 1.

A.3.4 Example table

Table A.3 — Examples of the bacterial diversity index

List heading	Example 1	Example 2
# sample_Name	9 387	9 388
genes_count	769 324	844 883
shannon_index	12,380 35	12,396 23
diversity_index_ratio	0,965 2	0,970 1

A.4 Enterotypes index

A.4.1 General

Data can be text format and showed with header line and data line. The enterotype information in metagenomics contains in data line.

A.4.2 Header line

The header line has two fixed fields as following, each field is connected with a tab:

[1]. # sample_name

[2]. intestinal_pattern

A.4.3 Data line

Each line has 2 fixed fields, each field is connected by a tab, and the null value of each field uses the point "." to represent. The respective field information is as following:

[1]. # sample_name: sample name, string type.

[2]. intestinal_pattern: Enterotype, string type.

A.4.4 Example table

Table A.4 — Examples of the Enterotype index

List heading	Example 1	Example 2
# sample_name	5 002	5 003
intestinal_pattern	Prevotella	Bacteroides

A.5 Bacterial relative abundance

A.5.1 General

Data can be text format and showed with header line and data line. Bacterial relative abundance information in metagenomics contains in data line.

A.5.2 Header line

The header line has N + 2 (N is the number of reference samples) fixed fields, each field is connected with a tab, like following:

[1]. # bacteria_name

[2]. sample_relative_abundance_value

[3]. reference_sample_relative_abundance_value_1

……

[N]. reference_sample_relative_abundance_value_N

A.5.3 Data line

Each line has N + 2 (N is the number of reference samples) fixed fields, each field is connected by a tab, and the null value of each field uses the point "." to represent. The respective field information is as following:

[1]. BACTERIA_NAME : The name of bacteria, string type

[2]. sample_relative_abundance_value : Bacteria relative genus abundance values of the sample, show in decimal fraction.

[3]. reference_sample_relative_abundance_value_1 : Bacteria relative genus abundance values of the reference sample 1, show in decimal fraction.

……

[N]. reference_sample_relative_abundance_value_N : Bacteria relative genus abundance values of the reference sample N, show in decimal fraction.

A.5.4 Example table

Table A.5 — Examples of bacterial relative abundance

List heading	Exmaple 1	Example 2
# bacteria_name	Bacteroides uniformis	Bifidobacterium adolescentis
sample_relative_abundance_value	0,003 039 194 6	0,000 125 262 5
reference_sample_relative_abundance_value _1	0,001 694 538 5	$6,003\ 822\ 041\times e^{-05}$
……	……	……
reference_sample_relative_abundance_value _N	0,002 179 311 8	$3,318\ 616\ 67\times e^{-05}$

A.6 Total relative gene abundance

A.6.1 General

Data can be text format and showed with header line and data line. Total relative gene abundance information in metagenomics contains in data line.

A.6.2 Header line

The header line has N + 2 (N is the number of reference samples) fixed fields, each field is connected with a tab, like following:

[1]. # Gene_ID

[2]. sample_relative_abundance_value

[3]. reference_sample_relative_abundance_value_1

……

[N]. reference_sample_relative_abundance_value_N

A.6.3 Data line

Each line has N + 2 (N is the number of reference samples) fixed fields, each field is connected by a tab, and the null value of each field uses the point "." to represent. The respective field information is as following.

[1]. GENE_ID : Identifiers of the gene, string type, NCBI Entrez Gene ID can be used, refer to https://www.ncbi.nlm.nih.gov/gene/.

[2]. sample_relative_abundance_value: Total relative gene abundance values of the sample, show in decimal fraction.