ICS 07.100
CCS A40

团 体 标 准

T/LTIA 13—2021

基因检测服务中微生物组数据共享规范

Specification for the interoperability of microbiome data in genetic testing services

2021－11－23 发布　　2021－11－30 实施

深圳市生命科技产学研资联盟　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由深圳市生命科技产学研资联盟提出并归口。

本文件起草单位：深圳华大生命科学研究院、深圳华大智造科技股份有限公司、深圳华大基因股份有限公司、兰州大学第二医院、北京量化健康科技有限公司、北京知因新生活细胞生物科技有限公司、青岛华大智造科技有限责任公司、深圳生命科技产学研资联盟、深圳华大基因科技有限公司。

本文件主要起草人：林宇翔、翦敏、蔡杰伦、王小涵、蔡锴晔、吕春杰、王奇、刘小燕、黎宇翔、李彦涛、曾俊杰、徐驰、吴静静、祁乐、陈娟娟、陶勇、单日强、刘姗姗、李陶莎、颜妙丽、武庆超、王萌萌、吴昊、李倩一、骆顺。

本文件为首次发布。

基因检测服务中微生物组数据共享规范

1 范围

本文件规定了在基因检测服务中，微生物组在数据处理节点间的数据共享规范、元数据及数据交换格式。

本文件适用于基因检测服务中微生物组数据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35890 高通量测序数据序列格式规范

GB/T 40226-2021 环境微生物宏基因组检测 高通量测序法

3 术语和定义

下列术语和定义适用于本文件。

3.1

元数据 metadata

主要描述数据属性的信息。

注：又称中介数据、中继数据，为描述数据的数据。

3.2

微生物组 microbiome

指具有独特理化性质、占据合理且明确栖息地的微生物群落。包括环境中的或与宿主共生的微生物，为古细菌、细菌、真菌、病毒等微生物的总和。

3.3

宏基因组 metagenome

指以特定环境中整个微生物群落作为研究对象，无需分离培养，直接提取的环境样本中全部微生物基因组DNA。

3.4

扩增子 amplicon

为DNA或RNA扩增后的一段核苷酸序列。

注：16S rDNA是细菌分类标志序列，18S rDNA和ITS（Internal Transcribed Spacer）则被广泛应用在真菌分类鉴定中。扩增子测序即是基于这类标志基因的靶向测序。

3.5

生物观测矩阵格式 Biological Observation Matrix (BIOM) format

是指通过观察值列联表来表示生物样本的一种格式，为微生物组领域常用的结果保存格式。

注1：BIOM格式的优点是可将OTU或特征表、样本属性、物种信息等多个表保存于同一个文件中，且格式统一，体积更小巧。

注2：OTU是指在系统发生学研究或群体遗传学研究中，为了便于进行分析，人为给某一个分类单元（品系，种，属，分组等）设置的同一标志。

3.6

数据生产 data production

是指通过高通量测序的方法，基于微生物的DNA或RNA等核酸分子，进行扩增子、宏基因组、宏转录组等测序方法产生数据，数据类型为FASTQ，通常称之为原始数据。

3.7

数据分析 data analysis

是指基于FASTQ数据，通过生物信息软件进行微生物组组成成分的定量分析，包括基因、物种、功能等水平的注释。

3.8

分析级数据 analytical data

指基于高通量测序数据，通过生物信息学软件分析，可定量分析出环境样品中样品的微生物成分组成的数据。

3.9

数据存储 data storage

是指以数据库为核心，以海量数据存储设备为支撑，对原始生产数据和加工数据进行存储和备份，包括但不限于云端及本地数据库。

3.10

数据服务 data service

是指以接口等在线形式提供微生物组提交、传输、下载等相关数据服务。

3.11

基因丰度 gene abundance

指某基因在微生物环境中所占的相对比例。如以参考基因集为参考序列，使用短序列比对的方式，实现基因丰度的定量。

3.12

分类学丰度 taxonomy abundance

指某一特定分类学种类（界、门、纲、目、科、属、种、株）微生物在环境总微生物群落中所占的相对比例，通常以百分比表示。

3.13

功能丰度 functional abundance

指某个酶、功能模块、功能通路在整个微生物环境中所占的比例。

3.14

相对丰度 relative abundance

某一特定种类微生物在环境总微生物群落中所占的相对比例。

注：通常以百分比表示。

3.15

绝对丰度 absolute abundance

某一特定种类微生物在环境总微生物群落中所占的绝对含量。

注：如每单位体积细胞数。

3.16

FASTQ 格式 FASTQ format

FASTQ是基于文本的、保存生物序列（通常是核酸序列）和其序列质量信息的、每四行表示一条序列的标准格式。

3.17

TSV 格式 Tab-Separated Values（TSV）format

TSV格式是微生物组生物信息学分析软件的输出格式，记录特征名、样品编号及对应的观测值（如相对丰度，序列数等），各字段使用制表符连接。

4 缩略语

下列缩略语适用于本文件。

DNA：脱氧核糖核酸（Deoxyribonucleic Acid）

BIOM：生物观测矩阵（Biological Observation Matrix）

UniRef：UniProt参考资料库（UniProt Reference Clusters）
API：应用程序编程接口（Application Programming Interface）
OTU：操作分类单元（Operational Taxonomic Units）
TSV：制表符分隔值（Tab-Separated Values）

5 微生物数据共享和交换流程

微生物组数据共享和交换在流程上可划分为四个环节：数据生产、数据分析、数据存储和数据服务（如图1所示）。在完成数据生产和数据分析的步骤后，结果文件及相关元数据应抽取为结构化数据进行数据存储。微生物组数据应分为管控数据和非管控数据，管控数据应由用户认证、授权后方可获取。推荐使用RESTful API设计规范构建相关数据服务，数据传输过程在必要环节需采用端到端加密。

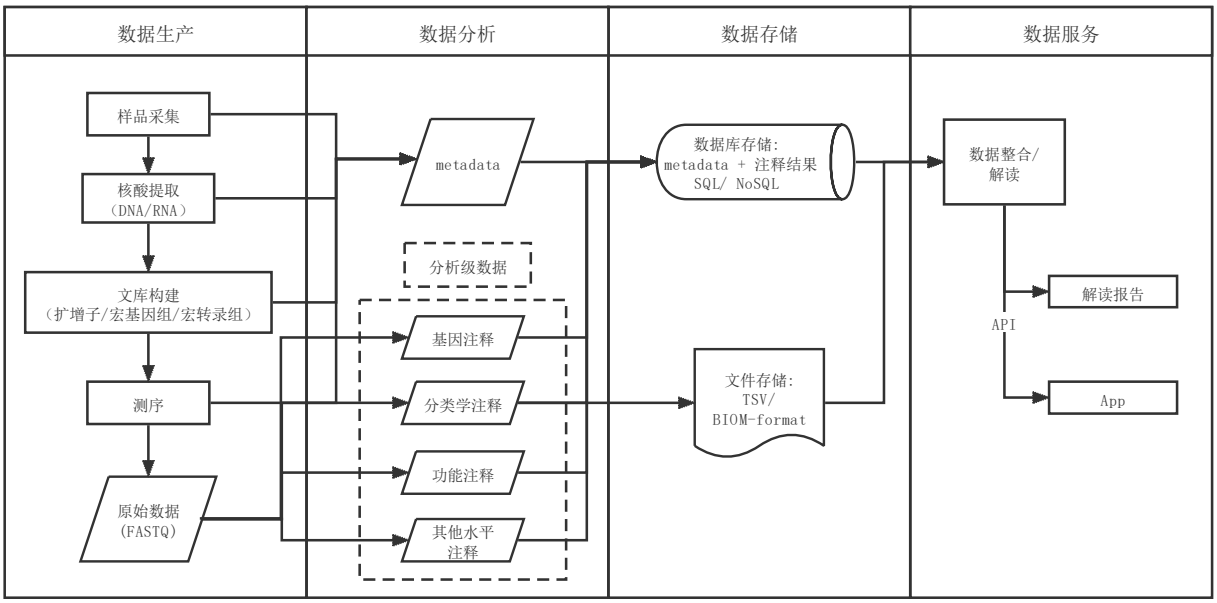

图 1 微生物组数据流程图

6 微生物组数据元数据

6.1 微生物组数据元数据分类

微生物组数据的元数据可分为样本信息关联的元数据、测序数据关联的元数据和分析级数据关联的元数据三个部分。

6.2 样本信息关联的元数据

样本信息关联的元数据规范化定义见表1。

表 1 样本信息关联的元数据

数据项	描述	必需	数据类型	示例
biosample_id	生物学样品编号	是	字符串	E-F123456789
host	宿主，可能值为：host taxid, unknown, environmental	是	字符串	txid9606
biosample_type	生物样品类型	是	字符串	feces，saliva

表1　样本信息关联的元数据（续）

数据项	描述	必需	数据类型	示例
biosample_site	采样部位	否	字符串	gut, oral
collection_date	采样时间	是	日期	2021-01-01 T00:00:00+0800
biosample_store_temp	生物学样本保存温度，单位：摄氏度（℃）	否	字符串	-80℃
extraction_method	核酸提取方式，可填写商业试剂盒的名字	否	字符串	PowerSoil® DNA Isolation Kit
extraction_time	核酸提取时间	否	日期	2021-01-02 T00:00:00+0800
DNA_mass	DNA质量	否	字符串	1.0 μg
DNA concentration	DNA浓度	否	字符串	10 ng/μL
DNA completeness	DNA完整度	否	字符串	无降解
DNA_quality	DNA质量评级	否	字符串	A类
Library_id	文库号，多个数据以逗号分割	否	字符串数组	HWJBAYTGAA170328, HWJBAYTGAA170329
注3：必需项为“是”表示该数据项为需要保留的最小元数据。 注4：DNA质量级别判断按照GB/T 40226-2021中的10.1.2进行。				

6.3　测序数据关联的元数据

测序数据关联的元数据规范化定义见表2。

表 2　测序数据关联的元数据

参数名	描述	必需	数据类型	示例
library_id	文库号	否	字符串	HWJBAYTGAA170328-18
library_strategy	建库策略，指文库的测序技术和策略	是	字符串	WGA, AMPLICON
library_source	文库样品来源，指实验材料来源类型	否	字符串	METAGENOMIC
library_selection	实验方式选定，是否使用了任何方法来选择和/或富集被测序的材料	否	字符串	unspecified
library_layout	文库布局	否	字符串	2
insert_size	插入长度（bp）	否	数字	250
nominal_size	接头长度（bp）	否	数字	10
platform	测序平台，多个数据以逗号分割	是	字符串数组	DNBSEQ-G400,DNBSEQ-G400
instrument_model	测序设备型号，多个数据以逗号分割	是	字符串数组	DNBSEQ-G400,DNBSEQ-G400
read_type	序列读取类型，多个数据以逗号分割	是	字符串数组	PE100,PE100
slide_id	芯片号，多个数据以逗号分割	否	字符串数组	CL100031417,CL100031417
lane_id	泳道号，多个数据以逗号分割	否	字符串数组	L02,L02
barcode	标识序列号，条形码，多	否	字符串	525,525

参数名	描述	必需	数据类型	示例
	个数据以逗号分割		数组	
sequencing_time	数据下机时间，多个数据以逗号分割	否	日期数组	2017-07-01T00:00:00+0800,2017-07-01T00:00:00+0800
raw_data_readnum	下机FASTQ数据的序列条数，多个数据以逗号分割	否	整数数组	2000000000, 2000000000
raw_data_basenum	下机FASTQ数据的碱基数，多个数据以逗号分割	否	整数数组	2000000000, 2000000000

表2　测序数据关联的元数据（续）

参数名	描述	必需	数据类型	示例
raw_data_gc_content	下机FASTQ数据的GC含量（比例），多个数据以逗号分割	否	浮点数数组	0.412012,0.412039
raw_data_n_count	下机FASTQ数据的N含量（比例），多个数据以逗号分割	否	浮点数数组	0.056686,0.056752
raw_data_q20	下机FASTQ数据中，Phred 数值大于 20 的碱基占总体碱基的百分比	否	浮点数数组	0.9680,0.9521
raw_data_q30	下机FASTQ数据中，Phred 数值大于 30 的碱基占总体碱基的百分比	否	浮点数数组	0.8861,0.8512
clean_data_readnum	质控后FASTQ数据的序列条数，多个数据以逗号分割	是	整数数组	2000000, 2000000
clean_data_basenum	质控后FASTQ数据的碱基数，多个数据以逗号分割	是	整数数组	200000000, 200000000
clean_data_gc_content	质控后FASTQ数据的GC含量（比例），多个数据以逗号分割	否	浮点数数组	0.412012,0.412039
clean_data_n_count	质控后FASTQ数据的N含量（比例），多个数据以逗号分割	否	浮点数数组	0.056686,0.056752
clean_data_q20	质控后FASTQ数据中，Phred 数值大于 20 的碱基占总体碱基的百分比	否	浮点数数组	0.9680,0.9521
clean_data_q30	质控后FASTQ数据中，Phred 数值大于 30 的碱基占总体碱基的百分比	否	浮点数数组	0.8861,0.8512
host_rate	宿主率，比对上宿主参考基因组的数据比例	是	浮点数	0.95
注：必需项为“是”表示该数据项为需要保留的最小元数据。				

6.4　分析级数据关联的元数据

分析级数据关联的元数据规范化定义见表3。

表 3　分析级数据关联的元数据

参数名	描述	必需	数据类型	示例
pipeline	分析所用流程/软件名	是	字符串	MetaPhlan2
pipeline_version	流程/软件版本	是	字符串	2.7.1
database	数据库	是	字符串	MetaPhlan2
database_version	数据库版本	是	字符串	2.7.1
parameters	运行参数	是	字符串	default
注：必需项为“是”表示该数据项为需要保留的最小元数据。				

7 分析级数据类型

7.1 分析级数据

基于定量维度及参考数据库的差异，微生物分析级数据可分为基因丰度、分类学丰度和功能通路丰度三大类。在实际使用场景中，微生物分析级数据类型应包含分类学丰度，宜包含基因丰度和功能通路丰度。

注1：扩展子测序可注释至属及部分物种的水平。

注2：宏基因组数据可注释至物种甚至菌株水平，并进行基因和功能的注释。

7.2 基因丰度

7.2.1 可使用特定样品构建特定环境的参考基因集。

示例1：人类肠道微生物参考基因集（Integrated Gene Catalog, IGC）。

示例2：人类肠道微生物群统一蛋白目录（Unified Human Gastrointestinal Protein catalog, UHGP）。

7.2.2 可使用通用的基因数据库。

示例：UniRef。

7.3 分类学丰度

基于比对的策略，可使用短序列与参考物种的基因组或标记基因比对；或可基于短序列的特征进行分类。

示例1：基于宏基因组数据的常用物种分类软件有MetaPhlan2、mOTU、Kranken2等。

示例2：基于扩增子测序数据的常用分类软件有QIIME2等。

7.4 功能丰度

使用不同的数据库可实现不同的功能注释。通用的功能注释数据库有KEGG，eggNOG，metaCyc等。

示例1：其他微生物组的功能数据库包含抗生素耐药性数据库（CARD）、碳水化合物活性酶数据库（CAZy）等。

示例2：基于宏基因组数据的常用软件有HUMAnN2。

8 微生物组数据格式要求

8.1 微生物组数据格式分类

微生物组数据格式可分为原始数据格式和分析级数据格式。

8.2 原始数据格式

微生物组原始数据的数据格式应为FASTQ。FASTQ格式应符合GB/T 35890的规定，格式样例见附录A。

8.3 分析级数据格式

8.3.1 分析级数据格式分类

分析级数据格式可分为文档储存数据格式和网络传输服务数据格式。

8.3.2 文档储存数据格式

用于文档储存的数据格式应为TSV格式或者BIOM格式。

8.3.2.1 TSV格式

分析级数据软件宜使用表格分割的文本文件作为输出结果，格式参考见附录B。

8.3.2.2 BIOM格式

记录样本关联的元数据，宜使用BIOM格式作为定量软件的输出格式。格式参考见附录C。

8.3.3 网络传输服务数据格式

基于结构化存储的元数据及分析级数据，宜构建API实现数据的在线传输及交互。在网络服务的应用中，数据结构宜参照附录D，宜使用JSON格式字符串用于为数据传输。

附 录 A
（资料性）
FASTQ 文件格式

A.1 FASTQ 文件格式样例

```
@CL100122427L1C001R001_16
CCGCTTTGAGACCTTTGCTGGAAATGGGAATATCT
+
FFFFFFFF>FCFFFFEFFFFFAFFEFFFEDDFFFD
@CL100122427L1C001R001_17
CCGCTTTGAGACCTTTGCTGGAAATGGGAATATCT
+
FFFFFFFF>FCFFFFEFFFFFAFFEFFFEDDFFFD
```

附 录 B
（资料性）
TSV 格式

TSV格式为可分为紧密型及松散型两种表格形式，表格形式见B.1和B.2。

B.1 紧密型表格

表格的第一列为特征名，如物种名、基因名等，第二列及之后的为样品编号。除行列名外，其余位置为样品的观测值，如相对丰度，序列数等。

注：由于微生物组样品的数据具有稀疏性，在较多样品的时候，经常会有90%的观测值为0。

示例 1：单个样本的结果

OTU_ID	SAMPLE.354
OTU0	1
OTU1	5
OTU2	1
OTU3	2
OTU4	1

示例 2：多个样本的结果

OTU_ID	SAMPLE.354	SAMPLE.355	SAMPLE.356
OTU0	0	0	4
OTU1	6	0	0
OTU2	1	0	7
OTU3	0	0	3

B.2 松散型表格

表格第一列为样品编号，第二列为特征名，第三列为观测值。

注：可以减少较多的0值储存。

示例：

SAMPLE.354	OTU1	6
SAMPLE.354	OTU2	1
SAMPLE.356	OTU0	4
SAMPLE.356	OTU2	7
SAMPLE.356	OTU3	3

附 录 C
（资料性）
BIOM 格式

C.1 格式说明

BIOM格式被微生物组领域主流软件所支持，具体格式说明参考bio-format.org官方链接（http://biom-format.org/index.html）。

C.2 版本

BIOM格式有两个版本:VERSION 1.0 JSON格式和VERSION 2.0 HDF5格式，常用格式为JSON格式。

注：两种格式可由命令行工具BIOM或对应的Python或R包执行相互转换。

示例 1：JSON 的松散格式

```
{
    "id":null,
    "format": "1.0.0",
    "format_url": "http://biom-format.org",
    "type": "OTU table",
    "generated_by": "QIIME revision 1.4.0-dev",
    "date": "2011-12-19T19:00:00",
    "rows":[
            {"id":"GG_OTU_1", "metadata":null}, //关于特征的描述可添加于此
            {"id":"GG_OTU_2", "metadata":null},
            {"id":"GG_OTU_3", "metadata":null},
            {"id":"GG_OTU_4", "metadata":null},
            {"id":"GG_OTU_5", "metadata":null}
        ],
    "columns": [
            {"id":"Sample1", "metadata":null}, //关于样品信息的描述可添加于此
            {"id":"Sample2", "metadata":null},
            {"id":"Sample3", "metadata":null},
            {"id":"Sample4", "metadata":null},
            {"id":"Sample5", "metadata":null},
            {"id":"Sample6", "metadata":null}
        ],
    "matrix_type": "sparse",
    "matrix_element_type": "int",
    "shape": [5, 6],
    "data":[[0,2,1],
            [1,0,5],
            [1,1,1],
            [1,3,2],
            [1,4,3],
            [1,5,1],
            [2,2,1],
            [2,3,4],
            [2,4,2],
            [3,0,2],
            [3,1,1],
```

```
        [3,2,1],
               [3,5,1],
               [4,1,1],
               [4,2,1]
              ]
    }
```

示例 2：JSON 的紧密格式

```
    {
        "id":null,
        "format": "Biological Observation Matrix 0.9.1-dev",
        "format_url":
"http://biom-format.org/documentation/format_versions/biom-1.0.html",
        "type": "OTU table",
        "generated_by": "QIIME revision 1.4.0-dev",
        "date": "2011-12-19T19:00:00",
        "rows":[
                {"id":"GG_OTU_1", "metadata":null},
                {"id":"GG_OTU_2", "metadata":null},
                {"id":"GG_OTU_3", "metadata":null},
                {"id":"GG_OTU_4", "metadata":null},
                {"id":"GG_OTU_5", "metadata":null}
            ],
        "columns": [
                {"id":"Sample1", "metadata":null},
                {"id":"Sample2", "metadata":null},
                {"id":"Sample3", "metadata":null},
                {"id":"Sample4", "metadata":null},
                {"id":"Sample5", "metadata":null},
                {"id":"Sample6", "metadata":null}
            ],
        "matrix_type": "dense",
        "matrix_element_type": "int",
        "shape": [5,6],
        "data":  [[0,0,1,0,0,0],
                  [5,1,0,2,3,1],
                  [0,0,1,4,2,0],
                  [2,1,1,0,0,1],
                  [0,1,1,0,0,0]]
    }
```

附 录 D
（资料性）
网络传输服务数据格式

网络传输服务数据格式见表D.1。

表 D.1 网络传输服务数据格式

元素/属性名称			描述
一级属性	二级属性	三级属性	
sample_info	-	-	样品信息
sample_info	biosample_id	-	生物样品编号
sample_info	sample_time	-	采样时间
sample_info	biosample_site	-	采样部位
sample_info	specimen	-	样品类型
observation	-	-	该观测项的定量结果相关信息
observation	unit	-	单位
observation	type	-	定量方式，如：相对丰度、绝对丰度、序列数
observation	value	-	该微生物的定量的值,如：value：10.23,unit：percent,则表示该微生物占总体的10.23%
coding	-	-	该观测项（微生物物种/基因/功能通路）的编码信息
coding	system	-	编码规范采用的参考系统，如在分类学丰度中可使用NCBI生物分类编号：https://www.ncbi.nlm.nih.gov/Taxonomy/；在功能通路可使用MetaCyc：https://metacyc.org/
coding	id	-	观测项的编码识别号,如：txid537011、GLUCOSE1PMETAB
coding	name	-	观测项的名称，如：s__Prevotella_copri、glucose and glucose-1-phosphate degradation
background	-	-	背景数据详情，包含与参考人群比较的信息
background	reference_cohort	-	参考人群的名称
background	background	-	该微生物在参考人群中量的分布,如5%;25%;50%;75%;95%等分位数
background	frequency	-	该微生物在参考人群中的出现频率
background	rank_ratio	-	该微生物在参考人群中的排序
background	mean	-	该微生物在参考人群中的定量结果的均值

参 考 文 献

[1] GB/T 7408 数据元和交换格式 信息交换 日期和时间表示法

[2] Kottmann R, Gray T, Murphy S, et al. A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS. 2008;12(2):115-121. doi:10.1089/omi.2008.0A10

[3] McDonald D, Clemente JC, Kuczynski J, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience. 2012;1(1):7. Published 2012 Jul 12. doi:10.1186/2047-217X-1-7