ICS 07.080
Q 8499

团　　体　　标　　准

T/SZAS 15—2019

人体肠道宏基因组学数据集

Dataset of Metagenomics of the Human Intestinal Tract

2019 - 12 - 05 发布　　2019 - 12 - 24 实施

深圳市标准化协会　　发 布

目　次

前　言

本标准按照GB/T 1.1-2009给出的规则起草。

本标准由深圳华大基因科技有限公司提出。

本标准由深圳市标准化协会归口。

本标准主要起草单位：深圳华大基因科技有限公司、深圳华大生命科学研究院、深圳华大基因股份有限公司、北京吉因加科技有限公司、深圳大学计算机与软件学院。

本标准主要起草人：吕春杰、刘小燕、张勇、方林、单日强、李倩一、何旭珩、孙建波、吴昊、姜华艳、李启沅、陈燕贤、王博、王韧、陈永胜、朱泽轩。

人体肠道宏基因组学数据集

1 范围

本标准规定了组学数据中有关人体肠道宏基因组学数据的范围以及数据元的规范化定义，数据集包括人体肠道宏基因组学元相关数据元及值域、数据格式及数据格式规格说明、归档目录。

本标准适用于组学数据中有关人体肠道宏基因组学数据信息的存储、治理、交换与共享。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 35890-2018 高通量测序数据序列格式规范

3 术语、定义和缩略语

下列术语、定义和缩略语适用于本文件。

3.1 术语和定义

3.1.1

FASTQ格式 FASTQ format

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

3.1.2

测序通道 lane

高通量检测平台测序功能在芯片上实现，整张芯片可以物理分隔成更小部分，每个物理分隔的栏称为lane。

3.1.3

KEGG Kyoto Encyclopedia of Genes and Genomes

京都基因与基因组百科全书。基因组破译方面的数据库。

3.1.4

eggNOG evolutionary genealogy of genes Non-supervised Orthologous Groups

通过已知蛋白对未知序列进行功能注释。

3.2 缩略语

S：字串型（string）

L：布尔型（boolean）

N：数值型（number）

D：日期型（date）

DT：日期时间型（datetime）

T：时间型（time）

4 数据元目录

4.1 数据元公用属性

数据元目录公用属性如表1所示。

表1 数据元目录公用属性

属性名称	描述
版本	V1.0
注册机构	注册机构名称
相关环境	生物信息、生物大数据
分类模式	分类法
主管机构	主管机构名称
注册状态	标准状态
提交机构	提交机构名称

4.2 数据元目录专用属性

4.2.1 人体肠道宏基因组学数据元目录专用属性包括测序信息、生物信息分析、人体肠道宏基因组学、质控信息四个部分。

4.2.2 测序信息描述测序过程中的数据元，例如测序任务单标识符、测序任务单名称、测序平台名称、测序仪标识符、测序仪名称、测序开始时间等。

4.2.3 生物信息分析描述生物信息分析过程中的数据元，例如结果数据存储路径、过滤软件名称、过滤软件版本、过滤软件参数等。

4.2.4 人体肠道宏基因组学描述人体肠道宏基因组结果文件中的数据元，例如宏基因编码、宏基因名称、宏基因值、宏基因长度、宏基因完整性描述、门分类、属分类、物种分类、微生物群落功能、抗生素含量数值、益生菌含量数值、肠道多样性数值等。

4.2.5 质控信息描述整个测序过程质量监控的数据元，包括项目标识符、项目名称、子项目名称、子项目标识符、样本管标识符、样品浓度、样品总量等。

4.2.6 具体每个数据元的标识符、名称、定义、信息保护、单位、数据类型见附录 A。数据元允许值见附录 B。

5 数据归档目录

5.1 归档目录专用属性

5.1.1 原始数据归档目录结构

人体肠道宏基因组学原始数据归档目录结构如表2所示：

表2　人体肠道宏基因组学原始数据归档目录结构

第一级	第二级	第三级
item_id	subitem_id	FASTQ
item_id	subitme_id	FASTQ_qc

注：每级目录情况说明见5.2。第一级目录为强制性目录，其他级目录为推荐性目录。

5.1.2　分析结果数据归档目录结构

人体肠道宏基因组学分析结果数据归档目录结构如表3所示：

表3　人体肠道宏基因组学分析结果数据归档目录结构

第一级	第二级	第三级	第四级	第五级	第六级
item_id	subitem_id	metagenomics	type	profile	bmpGeneAbundance
item_id	subitem_id	metagenomics	type	profile	bmpSpeciesAbundance
item_id	subitem_id	metagenomics	type	profile	Quantitative result name
item_id	subitem_id	metagenomics	type	profile	bmpGenusAbundance
item_id	subitem_id	metagenomics	type	profile	bmpKOAbundance
item_id	subitem_id	metagenomics	type	profile	bmpPhylumAbundance
item_id	subitem_id	metagenomics	type	profile	bmpReadsAbundance
item_id	subitem_id	metagenomics	type	omics_report	
item_id	subitem_id	metagenomics	type	Statistics	

注：每级目录情况说明见5.3。第一级目录为强制性目录，其他级目录为推荐性目录。

5.2　原始数据归档目录

5.2.1　概述

本章节是对5.1.1原始数据归档目录结构具体每级目录的情况说明。共3级目录，每一级目录均包涵目录标识符、目录名称、目录定义及父目录。

5.2.2　原始数据归档第一级目录

目录标识符：DI01.01.01
目录名称：item_id
定义：大项目或产品或其他可分类别的标识符。适用于标识以项目或产品等方式产出的数据。

5.2.3　原始数据归档第二级目录

目录标识符：DI01.02.01
目录名称：subitem_id
定义：子项目或产品或其他可分子类别的标识符。适用于标识以项目或产品等方式产出的数据。
父目录：DI01.01.01

5.2.4　原始数据归档第三级目录

5.2.4.1 FASTQ 目录

目录标识符：DI01.03.01
目录名称：FASTQ
定义：存放原始下机的测序文件
父目录：DI01.02.01

5.2.4.2 FASTQ_qc 目录

目录标识符：DI01.03.02
目录名称：FASTQ_qc
定义：存放原始下机的测序数据的质控文件
父目录：DI01.02.01

5.3 分析结果数据归档目录

5.3.1 概述

本章节是对5.1.2分析结果数据归档目录结构具体每级目录的情况说明。共6级目录，每一级目录均包涵目录标识符、目录名称、目录定义及父目录。

5.3.2 分析结果数据归档第一级目录

目录标识符：DI03.01.01
目录名称：item_id
定义：大项目或产品或其他可分类别的标识符。适用于标识以项目或产品等方式产出的数据。

5.3.3 分析结果数据归档第二级目录

目录标识符：DI03.02.01
目录名称：subitem_id
定义：子项目或产品或其他可分子类别的标识符。适用于标识以项目或产品等方式产出的数据。
父目录：DI03.01.01

5.3.4 分析结果数据归档第三级目录

目录标识符：DI03.03.01
目录名称：metagenomics
定义：存放宏基因组学数据
父目录：DI03.02.01

5.3.5 分析结果数据归档第四级目录

目录标识符：DI03.04.01
目录名称：type
定义：非固定字段，名称灵活设置。存放数据的专有性标识，如分析流程、参考序列等。
父目录：DI03.03.01

5.3.6 分析结果数据归档第五级目录

5.3.6.1 Profile 目录

目录标识符：DI03.05.01
目录名称：profile
定义：存放人体肠道宏基因组学二级数据
父目录：DI03.04.01

5.3.6.2 omics_report 目录

目录标识符：DI03.05.02
目录名称：omics_report
定义：存放产出的可结构化的结果/报告数据
父目录：DI03.04.01

5.3.6.3 statistics 目录

目录标识符：DI03.05.03
目录名称：statistics
定义：存放人体肠道宏基因组学分析所需要的表型数据
父目录：DI03.04.01

5.3.7 分析结果数据归档第六级目录

5.3.7.1 bmpGeneAbundance 目录

目录标识符：DI03.06.01
目录名称：bmpGeneAbundance
定义：存放基因的二级数据
父目录：DI03.05.01

5.3.7.2 bmpGenusAbundance 目录

目录标识符：DI03.06.02
目录名称：bmpGenusAbundance
定义：存放属水平的二级数据
父目录：DI03.05.01

5.3.7.3 bmpPhylumAbundance 目录

目录标识符：DI03.06.03
目录名称：bmpPhylumAbundance
定义：存放门水平的二级数据
父目录：DI03.05.01

5.3.7.4 bmpSpeciesAbundance 目录

目录标识符：DI03.06.04
目录名称：bmpSpeciesAbundance
定义：存放物种的二级数据
父目录：DI03.05.01

5.3.7.5 Quantitative result name 目录

目录标识符：DI03.06.05

目录名称：Quantitative result name

定义：存放定量结果的二级数据，根据定量数据类型命名，多个定量结果可平铺多个目录。如mOTU、MetaPh1An2等

父目录：DI03.05.01

5.3.7.6 bmpReadsAbundance 目录

目录标识符：DI03.06.06

目录名称：bmpReadsAbundance

定义：基于基因集比对得到的下机序列数的配置文件

父目录：DI03.05.01

5.3.7.7 bmpKOAbundance 目录

目录标识符：DI03.06.07

目录名称：bmpKOAbundance

定义：基于KEGG数据库的功能水平二级数据

父目录：DI03.05.01

6 数据格式标准

6.1 数据格式专用属性

6.1.1 原始数据格式

数据格式标识符：DF02.01.001

数据格式名称：FASTQ

适用范围：第二代测序仪产生的原始数据信息

数据格式允许值：SF02.01.001 FASTQ

6.1.2 结果数据格式

6.1.2.1 属总相对丰度格式

数据格式标识符：DF02.02.001

数据格式名称：属总相对丰度

适用范围：宏基因组学分析中属总相对丰度信息

数据格式允许值：SF02.02.001属总相对丰度

6.1.2.2 功能总相对丰度格式

数据格式标识符：DF02.02.002

数据格式名称：功能总相对丰度

适用范围：宏基因组学分析中功能总相对丰度信息

数据格式允许值：SF02.02.002功能总相对丰度

6.1.2.3 物种总相对丰度格式

数据格式标识符：DF02.02.003

数据格式名称：物种总相对丰度
适用范围：宏基因组学分析中物种总相对丰度信息
数据格式允许值：SF02.02.003物种总相对丰度

6.1.3 报告数据格式

6.1.3.1 细菌多样性指数格式

数据格式标识符：DF02.03.001
数据格式名称：细菌多样性指数
适用范围：宏基因组学分析中细菌多样性指数信息
数据格式允许值：SF02.03.001细菌多样性指数

6.1.3.2 肠型指数格式

数据格式标识符：DF02.03.002
数据格式名称：肠型指数
适用范围：宏基因组学分析中肠型指数信息
数据格式允许值：SF02.03.002肠型指数

6.1.3.3 功能菌相对丰度格式

数据格式标识符：DF02.03.003
数据格式名称：功能菌相对丰度
适用范围：宏基因组学分析中功能菌相对丰度信息
数据格式允许值：SF02.03.003功能菌相对丰度

6.1.3.4 基因总相对丰度格式

数据格式标识符：DF02.03.004
数据格式名称：基因总相对丰度
适用范围：宏基因组学分析中基因总相对丰度信息
数据格式允许值：SF02.03.004基因总相对丰度

6.1.3.5 有益菌相对丰度格式

数据格式标识符：DF02.03.005
数据格式名称：有益菌相对丰度
适用范围：宏基因组学分析中有益菌相对丰度信息
数据格式允许值：SF02.03.005有益菌相对丰度

6.1.3.6 有害菌相对丰度格式

数据格式标识符：DF02.03.006
数据格式名称：有害菌相对丰度
适用范围：宏基因组学分析中有害菌相对丰度信息
数据格式允许值：SF02.03.006有害菌相对丰度

6.1.3.7 条件致病菌相对丰度格式

数据格式标识符：DF02.03.007

数据格式名称：条件致病菌相对丰度

适用范围：宏基因组学分析中条件致病菌相对丰度信息

数据格式允许值：SF02.03.007条件致病菌相对丰度

6.2 数据格式代码表

6.2.1 原始数据格式代码

原始数据格式代码如表4所示：

表4 原始数据格式代码

数据格式说明标识符	数据格式名称	规格说明
SF02.01.001	FASTQ	GB/T 35890—2018 高通量测序数据序列格式规范

6.2.2 结果数据格式代码

结果数据格式代码如表5所示

表5 结果数据格式代码

数据格式说明标识符	数据格式名称	规格说明
SF02.02.001	属总相对丰度	见附录 C
SF02.02.002	功能总相对丰度	见附录 D
SF02.02.003	物种总相对丰度	见附录 D

6.2.3 报告数据格式代码

报告数据格式代码如表6所示：

表6 报告数据格式代码

数据格式说明标识符	数据格式名称	规格说明
SF02.03.001	细菌多样性指数	见附录 E
SF02.03.002	肠型指数	见附录 F
SF02.03.003	功能菌相对丰度	见附录 G
SF02.03.004	基因总相对丰度	见附录 H
SF02.03.005	有益菌相对丰度	见附录 G
SF02.03.006	有害菌相对丰度	见附录 G
SF02.03.007	条件致病菌相对丰度	见附录 G

附 录 A
（资料性附录）
数据元目录

A.1 简介

本附录说明了推荐性数据元的标识符，名称，定义，信息保护，单位，数据类型和数据元允许值。且有新的数据元加入可以顺延排入。

A.2 测序信息

测序信息如表A.1所示：

表A.1 测序信息

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE07.01.001.00	测序任务单标识符	用于提供测序要求的任务单的标识符。	不保护		S	
DE07.01.002.00	测序任务单名称	用于提供测序要求的任务单的名称。	不保护		S	
DE07.01.003.00	测序类型	测序类型。	不保护		S	
DE07.01.004.00	测序平台名称	测序平台名称。	不保护		S	
DE07.01.005.00	测序仪标识符	测序仪标识符。	不保护		S	
DE07.01.006.00	测序仪名称	测序仪名称。	不保护		S	B.1 测序仪名称代码表
DE07.01.007.00	测序开始时间	测序开始当日的公元纪年日期和时间的完整描述。	不保护		DT	
DE07.01.008.00	测序完成时间	测序完成当日的公元纪年日期和时间的完整描述。	不保护		DT	
DE07.01.009.00	芯片号	芯片号编码。	不保护		S	
DE07.01.010.00	测序通道（lane）号	lane 号。	不保护		S	
DE07.01.011.00	机器号	机器号。	不保护		S	
DE07.01.012.00	原始下机数据存储路径	原始下机数据的存储路径。	不保护		S	
DE07.01.013.00	FASTQ 格式文件唯一编号	FASTQ 格式文件唯一编号。	不保护		S	
DE07.01.014.00	下机地	数据下机地区。	保护		S	GB T2260-2013 中华人民共和国行政区划代码
DE07.01.015.00	文库类型	文库类型说明。	不保护		S	
DE07.01.016.00	文库标识符	测序文库标识符。	不保护		S	

表 A.1　测序信息（续）

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE07.01.017.00	文库名称	文库名称。	不保护		S	
DE07.01.018.00	文库数量	文库数量。	不保护		N	

A.3　生物信息分析

生物信息分析如表A.2所示：

表A.2　生物信息分析

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE08.01.001.00	结果数据存储路径	通过信息分析的结果数据存储的存储路径。	不保护		S	
DE08.01.002.00	过滤软件名称	信息分析过程中所使用过滤软件名称。	不保护		S	
DE08.01.003.00	过滤软件版本	信息分析过程中所使用过滤软件版本信息。	不保护		S	
DE08.01.004.00	过滤软件参数	信息分析过程中所使用过滤软件的参数信息。	不保护		S	
DE08.01.005.00	组装软件名称	信息分析过程中所使用组装软件名称。	不保护		S	
DE08.01.006.00	组装软件版本	信息分析过程中所使用组装软件版本信息。	不保护		S	
DE08.01.007.00	组装软件参数	信息分析过程中所使用组装软件的参数信息。	不保护		S	
DE08.01.008.00	基因功能注释软件名称	信息分析过程中所使用基因功能注释软件名称。	不保护		S	
DE08.01.009.00	基因功能注释软件版本	信息分析过程中所使用基因功能注释软件版本信息。	不保护		S	
DE08.01.010.00	基因功能注释软件参数	信息分析过程中所使用基因功能注释软件的参数信息。	不保护		S	
DE08.01.011.00	物种组成与多样性分析软件名称	信息分析过程中所使用物种组成与多样性分析软件名称。	不保护		S	
DE08.01.012.00	物种组成与多样性分析软件版本	信息分析过程中所使用物种组成与多样性分析软件版本信息。	不保护		S	
DE08.01.013.00	物种组成与多样性分析软件参数	信息分析过程中所使用物种组成与多样性分析软件的参数信息。	不保护		S	
DE08.01.014.00	基因丰度定量与差异分析软件名称	信息分析过程中所使用基因丰度定量与差异分析软件名称。	不保护		S	
DE08.01.015.00	基因丰度定量与差异分析软件版本	信息分析过程中所使用基因丰度定量与差异分析软件版本信息。	不保护		S	
DE08.01.016.00	基因丰度定量与差异分析软件参数	信息分析过程中所使用基因丰度定量与差异分析软件的参数信息。	不保护		S	
DE08.01.017.00	主成分分析软件名称	信息分析过程中所使用主成分分析软件名称。	不保护		S	
DE08.01.018.00	主成分分析软件版本	信息分析过程中所使用主成分分析软件版本信息。	不保护		S	

表 A.2　生物信息分析（续）

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE08.01.019.00	主成分分析软件参数	信息分析过程中所使用主成分分析软件的参数信息。	不保护		S	

A.4　人体肠道宏基因组学

人体肠道宏基因组学如表A.3所示：

表A.3　人体肠道宏基因组学

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE09.01.001.00	宏基因编码	宏基因标识符。	不保护		S	
DE09.01.002.00	宏基因名称	宏基因名称。	不保护		S	
DE09.01.003.00	宏基因值	宏基因值。	不保护		N	
DE09.01.004.00	宏基因长度	宏基因长度。	不保护		N	
DE09.01.005.00	宏基因完整性描述	宏基因完整性描述。	不保护		S	
DE09.01.006.00	宏基因的群体来源	宏基因的群体来源。	不保护		S	
DE09.01.007.00	门分类	门分类。	不保护		S	
DE09.01.008.00	微生物门含量数值	微生物门含量数值。	不保护		N	
DE09.01.009.00	属分类	属分类。	不保护		S	
DE09.01.010.00	微生物属含量数值	微生物属含量数值。	不保护		N	
DE09.01.011.00	物种分类	物种分类。	不保护		S	
DE09.01.012.00	微生物物种含量数值	微生物物种含量数值。	不保护		N	
DE09.01.013.00	KEGG 注释	KEGG 注释。	不保护		S	
DE09.01.014.00	直系同源基因注释	直系同源基因注释。	不保护		S	
DE09.01.015.00	某基因在所有样本中观测到的频率	某基因在所有样本中观测到的频率。	不保护		N	
DE09.01.016.00	某基因在所有个体中观测到的频率	某基因在所有个体中观测到的频率。	不保护		N	
DE09.01.017.00	KEGG 注释的功能分类	KEGG 注释的功能分类。	不保护		S	
DE09.01.018.00	eggNOG 注释的功能分类	eggNOG 注释的功能分类。	不保护		S	
DE09.01.019.00	判定后的群体来源	判定后的群体来源。	不保护		S	
DE09.01.020.00	微生物群落功能	微生物群落功能。	不保护		S	
DE09.01.021.00	抗生素类别中文名称	抗生素类别中文名称。	不保护		S	
DE09.01.022.00	抗生素类别英文名称	抗生素类别英文名称。	不保护		S	
DE09.01.023.00	抗生素含量数值	抗生素含量数值。	不保护		N	
DE09.01.024.00	益生菌菌种中文名称	益生菌菌种中文名称。	不保护		S	
DE09.01.025.00	益生菌菌种英文名称	益生菌菌种英文名称。	不保护		S	
DE09.01.026.00	益生菌含量数值	益生菌含量数值。	不保护		N	

表 A.3　人体肠道宏基因组学（续）

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE09.01.027.00	条件致病菌菌种中文名称	条件致病菌菌种中文名称。	不保护		S	
DE09.01.028.00	条件致病菌菌种英文名称	条件致病菌菌种英文名称。	不保护		S	
DE09.01.029.00	条件致病菌含量数值	条件致病菌含量数值。	不保护		N	
DE09.01.030.00	肠道多样性数值	肠道多样性数值。	不保护		N	
DE09.01.031.00	肠型	肠型。	不保护		S	
DE09.01.032.00	肠道健康指数	肠道健康指数。	不保护		N	
DE09.01.033.00	肠道功能指数	肠道功能指数。	不保护		N	
DE09.01.034.00	肠道抗生素指数	肠道抗生素指数。	不保护		N	
DE09.01.035.00	肠道微生物物种多样性指数	肠道微生物物种多样性指数。	不保护		N	
DE09.01.036.00	肠道微生物患病指数	肠道微生物患病指数。	不保护		N	

A.5　质控信息

质控信息如表A.4所示：

表A.4　质控信息

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE01.01.001.00	项目标识符	项目标识符，适用于标识以项目方式产出的数据。	不保护		S	
DE01.01.002.00	项目名称	项目名称。	不保护		S	
DE01.01.003.00	子项目名称	子项目名称。	不保护		S	
DE01.01.004.00	子项目标识符	子项目标识符。	不保护		S	
DE01.01.005.00	样本管标识符	样本管标识符。	保护		S	
DE01.01.006.00	样品浓度	样本的浓度值，计量单位为ng/μL。	不保护	ng/μL		
DE01.01.007.00	样品总量	样本的总重量，计量单位为μL。	不保护	μL		
DE01.01.008.00	数据量质控结果	数据量质控结果。	不保护		S	
DE01.01.009.00	总数据量	总数据量。	不保护	bp	N	
DE01.01.010.00	测序深度	测序得到的碱基总量与基因组大小的比值，它是评价测序量的指标之一。	不保护	%	N	
DE01.01.011.00	测序数据量	样本本次测序的数据量，计量单位为Gb。	不保护	Gb		
DE01.01.012.00	唯一下机序列的比对率	唯一下机序列的比对率。	不保护	%	N	
DE01.01.013.00	插入片段大小	插入片段的大小。	不保护		N	
DE01.01.014.00	参考基因组的比对率	与参考基因组的比对率。	不保护	%	N	

表 A.4　质控信息（续）

标识符	名称	定义	信息保护	单位	数据类型	数据元允许值
DE01.01.015.00	重复率	重复下机序列占所有下机序列的比率。重复下机序列指序列一样并且比对到基因组相同位置的下机序列。	不保护	%	N	
DE01.01.016.00	错配率	错配率。	不保护	%	N	
DE01.01.017.00	平均覆盖率	测序获得的序列占整个被测区域的比例。	不保护		N	
DE01.01.018.00	基因测序覆盖率	覆盖率，指检测到的该基因核酸序列长度占该基因组序列长度的百分比。	不保护		N	
DE01.01.022.00	总体 Q20 值	见 3.1.12 。	不保护		S	
DE01.01.023.00	总体 Q30 值	见 3.1.13 。	不保护		S	
DE01.01.032.00	过滤数据量	过滤数据量。	不保护	bp	N	
DE01.01.033.00	过滤数据率	过滤数据率。	不保护	bp	N	
DE01.01.034.00	过滤后数据量	过滤后数据量。	不保护	bp	N	

附 录 B
（资料性附录）
数据元值域代码表

B.1 测序仪名称代码

测序仪名称代码规定了测序仪名称的代码。
采用2位数字顺序代码，从“01”开始编码，按升序排列。见表B.1。

表B.1 测序仪名称代码表

代码	测序仪名称	测序仪具体型号
00	Roche 公司 454 系列	454 GS/GS 20/GS FLX/GS FLX Titanium/GS FLX+/GS Junior
01	ABI 公司 310 系列	310 /3130 /3130xl
02	ABI 公司 3500 系列	3500/3500xl
03	ABI 公司 3730 系列	3730xl, 3700
04	ABI 公司 5500 系列	5500 /5500xl /5500x-Wl
05	ABI 公司 Solid 系列	SOLiD 3 Plus System/SOLiD 4 System/SOLiD 4hq System/SOLiD PI System/SOLiD System 1.0/SOLiD System 2.0/SOLiD System 3.0
06	CapitalBio BioelectronSeq 4000	BioelectronSeq 4000
07	Thermo Fisher Ion Torrent PGM	Ion Torrent PGM
08	Thermo Fisher Ion Torrent Proton	Ion Torrent Proton
09	Bionano Genomics BioNano 系列	BioNano IRYS/SAPHYR
10	Complete Genomics	Complete Genomics
11	DAAN GENE	DA8600
12	Helicos BioSciences Corporation	Helicos HeliScope
13	HYK Genetic	HYK-PSTAR-IIA
14	Illumina 公司 Genome Analyzer 系列	Genome Analyzer/Genome Analyzer II/Genome Analyzer IIx
15	Illumina 公司 HiSeq 系列	HiSeq SQ/1000/1500/2000/2500/X Ten/X Five/3000/4000
16	Illumina 公司 MiSeq 系列	MiSeq/MiSeq Dx/MiSeq FGx
17	Illumina 公司 NextSeq 系列	NextSeq500/550
18	Illumina 公司 MiniSeq 系列	MiniSeq
19	Illumina 公司 iSeq 系列	iSeq 100
20	Illumina 公司 NovaSeq 系列	NovaSeq 5000/6000/TM
21	BGI 公司 BGISEQ 系列	BGISEQ-1000/50/100/500
22	BGI 公司 MGISEQ 系列	MGISEQ-200/2000
23	BGI 公司 DNBSEQ 系列	DNBSEQ-T7

表 B.1　测序仪名称代码表（续）

代码	测序仪名称	测序仪具体型号
24	BGI 公司 BGISEQ 系列	BGISEQ-500RS
25	BGI 公司 BGISEQ 系列	BGISEQ-500CX
26	BGI 公司 MGISEQ 系列	MGISEQ-200RS/2000RS/200CX/2000CX
27	BGI 公司 DNBSEQ 系列	DNBSEQ-G50/G400/E
28	Oxford Nanopore MinION	MinION
29	Oxford Nanopore GridION	GridION
30	Berry Genomics NextSeq CN500	NextSeq CN500
31	PacBio SMRT PacBio	PacBio RS/RS II/Sequel
99	Other	

附 录 C
（资料性附录）
属总相对丰度

C.1 简介

人体肠道宏基因组学数据格式属总相对丰度为文本格式（推荐使用gz压缩格式存放）。本格式包含表头行和数据行，数据行包含人体肠道宏基因组学中的属总相对丰度信息。

C.2 表头行

表头行有N+2（N代考参考样品数目）个固定字段，各字段使用制表符连接。

a) Scientific_Name；
b) sample_relative_abundance_value；
c) reference_sample_relative_abundance_value_1；
d) reference_sample_relative_abundance_value_N。

C.3 数据行

每行记录有N+2（N代表参考样品数目）个固定字段，各字段使用制表符连接，各字段的空值使用点“.”表示。各字段信息如下：

a) Scientific_Name：细菌的属名。字符类型。采用 NCBI Taxonomy Database 描述，参见 http://www.ncbi.nlm.nih.gov/taxonomy ；
b) sample_relative_abundance_value：样品对应的属总相对丰度值。小数类型；
c) reference_sample_relative_abundance_value_1：参考样品 1 对应的属总相对丰度值。小数类型；
d) reference_sample_relative_abundance_value_N：参考样品 N 对应的属总相对丰度值。小数类型。

示例：

表头	例 1	例 2
Scientific_Name	Abiotrophia	Acetivibrio
sample_relative_abundance_value	3.079903e-06	4.047342e-05
reference_sample_relative_abundance_value _1	3.227675e-06	1.407130e-05
……	……	……
reference_sample_relative_abundance_value _N	0	4.641824e-06

附 录 D
（资料性附录）
细菌多样性指数

D.1 简介

人体肠道宏基因组学数据格式细菌多样性指数为文本格式（推荐使用gz压缩格式存放）。本格式包含表头行和数据行，数据行包含人体肠道宏基因组学中的细菌多样性指数信息。

D.2 表头行

表头行包括以下4个固定字段，各字段使用制表符连接。

a) sample_name；
b) genes_count；
c) shannon_index；
d) diversity_index_ratio。

D.3 数据行

每行记录有4个固定字段，各字段使用制表符连接，各字段的空值使用点“.”表示。各字段信息如下：

a) sample_name：样品名。字符类型；
b) genes_count：基因总数。整数类型；
c) SHANNON_INDEX：香农指数。小数类型；
d) diversity_index_ratio：低于香农指数的人群比例。小数类型。数值在 0-1 之间。

示例：

表头	例 1	例 2
sample_name	9387	9388
genes_count	769324	844883
shannon_index	12.38035	12.39623
diversity_index_ratio	0.9652	0.9701

附 录 E
（资料性附录）
物种总相对丰度

E.1 简介

人体肠道宏基因组学数据格式物种总相对丰度为文本格式（推荐使用gz压缩格式存放）。本格式包含表头行和数据行，数据行包含人体肠道宏基因组学中的物种总相对丰度信息。

E.2 表头行

表头行有N+2（N代考参考样品数目）个固定字段，各字段使用制表符连接。

a) Scientific_Name；
b) sample_relative_abundance_value；
c) reference_sample_relative_abundance_value_1；
d) reference_sample_relative_abundance_value_N。

E.3 数据行

每行记录有N+2（N代表参考样品数目）个固定字段，各字段使用制表符连接，各字段的空值使用点“.”表示。各字段信息如下：

a) Scientific_Name：细菌的物种名。字符类型。采用 NCBI Taxonomy Database 描述，参见 http://www.ncbi.nlm.nih.gov/taxonomy ；
b) sample_relative_abundance_value：样品对应的物种总相对丰度值。小数类型；
c) reference_sample_relative_abundance_value_1：参考样品 1 对应的物种总相对丰度值。小数类型；
d) reference_sample_relative_abundance_value_N：参考样品 N 对应的物种总相对丰度值。小数类型。

示例：

表头	例 1	例 2
Scientific_Name	Abiotrophia defectiva	Achromobacter piechaudii
sample_relative_abundance_value	1.16374942e-05	0
reference_sample_relative_abundance_value _1	3.7373926e-06	0
……	……	……
reference_sample_relative_abundance_value _N	2.180657e-06	1.44414e-07

附 录 F
（资料性附录）
肠型指数

F.1 简介

人体肠道宏基因组学数据格式肠型指数为文本格式（推荐使用gz压缩格式存放）。本格式包含表头行和数据行，数据行包含人体肠道宏基因组学中的肠型信息。

F.2 表头行

表头行包括以下2个固定字段，各字段使用制表符连接。

a) sample_name；

b) intestinal_pattern。

F.3 数据行

每行记录有2个固定字段，各字段使用制表符连接，各字段的空值使用点“.”表示。各字段信息如下：

a) sample_name：样品名。字符类型；

b) intestinal_pattern：肠型。字符类型。

示例：

表头	例 1	例 2
sample_name	5002	5003
intestinal_pattern	Prevotella	Bacteroides

附 录 G
（资料性附录）
细菌相对丰度

G.1 简介

人体肠道宏基因组学数据格式细菌相对丰度为文本格式（推荐使用gz压缩格式存放）。本格式包含表头行和数据行，数据行包含人体肠道宏基因组学中的细菌相对丰度信息。

G.2 表头行

表头行有N+2（N代考参考样品数目）个固定字段，各字段使用制表符连接。

a) bacteria_name；
b) sample_relative_abundance_value；
c) reference_sample_relative_abundance_value_1；
d) reference_sample_relative_abundance_value_N。

G.3 数据行

每行记录有N+2（N代表参考样品数目）个固定字段，各字段使用制表符连接，各字段的空值使用点“.”表示。各字段信息如下：

a) BACTERIA_NAME：细菌名。字符类型；
b) sample_relative_abundance_value：样品对应的细菌相对丰度值。小数类型；
c) reference_sample_relative_abundance_value_1：参考样品 1 对应的细菌相对丰度值。小数类型；
d) reference_sample_relative_abundance_value_N：参考样品 N 对应的细菌相对丰度值。小数类型。

示例：

表头	例 1	例 2
bacteria_name	Bacteroides uniformis	Bifidobacterium adolescentis
sample_relative_abundance_value	0.0030391946	0.0001252625
reference_sample_relative_abundance_value _1	0.0016945385	6.003822041e-05
……	……	……
reference_sample_relative_abundance_value _N	0.0021793118	3.31861667e-05

附 录 H
（资料性附录）
基因总相对丰度

H.1 简介

人体肠道宏基因组学数据格式基因总相对丰度为文本格式（推荐使用gz压缩格式存放）。本格式包含表头行和数据行，数据行包含人体肠道宏基因组学中的基因总相对丰度信息。

H.2 表头行

表头行有N+2（N代考参考样品数目）个固定字段，各字段使用制表符连接。

a) Gene_ID；
b) sample_relative_abundance_value；
c) reference_sample_relative_abundance_value_1；
d) reference_sample_relative_abundance_value_N。

H.3 数据行

每行记录有N+2（N代表参考样品数目）个固定字段，各字段使用制表符连接，各字段的空值使用点“.”表示。各字段信息如下：

a) GENE_ID：基因的标识符。字符类型。采用 NCBI Entrez Gene ID 描述，参见 http://www.ncbi.nlm.nih.gov/gene/ ；
b) sample_relative_abundance_value：样品对应的基因总相对丰度值。小数类型；
c) reference_sample_relative_abundance_value_1：参考样品 1 对应的基因总相对丰度值。小数类型；
d) reference_sample_relative_abundance_value_N：参考样品 N 对应的基因总相对丰度值。小数类型。

示例：

表头	例 1	例 2
GENE_ID	25604	56324
sample_relative_abundance_value	0.0030391946	0.0001252625
reference_sample_relative_abundance_value _1	0.0016945385	6.003822041e-05
……	……	……
reference_sample_relative_abundance_value _N	0.0021793118	3.31861667e-05