



中华人民共和国国家标准

GB/T XXXXX—XXXX

单细胞测序 单细胞转录组数据集

Single-cell sequencing—Dataset of single cell transcriptome

(点击此处添加与国际标准一致性程度的标识)

(征求意见稿)

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX – XX – XX 发布

XXXX – XX – XX 实施

国家市场监督管理总局
国家标准化管理委员会 发布

目 次

| | |
|-----------------------------|----|
| 前言 | II |
| 1 范围 | 1 |
| 2 规范性引用文件 | 1 |
| 3 术语和定义 | 1 |
| 4 缩略语 | 2 |
| 5 数据及数据文件要求 | 2 |
| 5.1 数据文件组成 | 2 |
| 5.2 数据文件 | 2 |
| 6 数据元目录 | 4 |
| 6.1 属性 | 4 |
| 6.2 数据元目录公用属性 | 4 |
| 6.3 数据元目录专用属性 | 4 |
| 7 数据归档目录 | 5 |
| 7.1 数据归档目录结构 | 5 |
| 7.2 单细胞数据归档目录要求 | 5 |
| 8 数据安全 | 6 |
| 8.1 单细胞转录组数据分级分类原则 | 6 |
| 8.2 单细胞转录组数据集的分级分类方式 | 6 |
| 8.3 单细胞转录组数据集的保密性、完整性价值分级关系 | 6 |
| 附录 A（资料性） 数据元目录 | 7 |
| A.1 简介 | 7 |
| A.2 实验信息 | 7 |
| A.3 建库测序信息 | 8 |
| A.4 生物信息分析 | 9 |
| A.5 质控信息 | 10 |
| 附录 B（资料性） 数据元值域代码表 | 13 |
| B.1 测序仪名称代码 | 13 |
| B.2 文库构建策略代码 | 14 |
| B.3 分子类型代码 | 15 |
| B.4 是否代码 | 15 |
| B.5 文件类型代码 | 15 |
| B.6 文库设置代码 | 16 |
| B.7 文库来源代码 | 16 |

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。

本文件由全国生化检测标准化技术委员会（SAC/TC 387）提出并归口。

本文件起草单位：

本文件主要起草人：

全国生化检测标准化技术委员会

单细胞测序 单细胞转录组数据集

1 范围

本文件规定了单细胞转录组数据的范围及数据元的规范化定义、数据格式要求和数据归档目录。数据集包括单细胞转录组研究相关数据元和值域。

本文件适用于组学数据中有关单细胞转录组数据信息的存储、管理、交换与共享。

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 2260—2007 中华人民共和国行政区划代码

3 术语和定义

下列术语和定义适用于本文件。

3.1

FASTQ 格式 FASTQ format

FASTQ格式是一种存储了生物序列（通常是核酸序列）以及相应的质量评价的文本格式，每四行表示一条序列。

3.2

Q20

测序数据中，碱基识别质量值大于20的碱基占有所有碱基的比例。

注：碱基识别质量值为20时，表示碱基的正确率为99%以上， $Q20 \geq 95\%$ ，则表示测序数据中95%以上的碱基质量值大于20。

3.3

Q30

测序数据中，碱基识别质量值大于30的碱基占有所有碱基的比例。

注：碱基识别质量值为30时，表示碱基的正确率为99.9%以上， $Q30 \geq 85\%$ ，则表示测序数据中85%以上的碱基质量值大于30。

3.4

项目 Project

一个项目是一个研究的总体描述，通常包含多个样本和数据集。

3.5

样本 Sample

描述实验的材料信息，每个样本需要有一个独特的属性。

3.6

实验/测序 Experiment/Run

实验信息，包括样本建库、测序方法、测序仪器等。一个实验通常是为了研究某个特定的项目而做的，而每个实验都会有一个或多个样本。测序指的是实验产生的数据文件。

3.7

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割成更小部分，每个物理分隔的栏称为lane。

3.8

基因表达矩阵 Gene expression matrix

行代表检测到的所有基因，列代表每个细胞，每个格子的数据表示特定的基因在特定的细胞中的表达丰度。

3.9

元数据文件 Metadata file

制表符分隔的文本文件，单细胞转录组元数据文件包含细胞水平的注释。

3.10

聚类文件 Cluster file

制表符分隔的文本文件，单细胞转录组聚类文件包含任何聚类 and 可选的特定聚类的元数据。

3.11

基因列表文件 Gene list file

包含基因名称和基因ID的制表符分隔的文本文件。

4 缩略语

下列缩略语适用于本文件。

cDNA: 互补脱氧核糖核酸 (complementary DNA)

DN: 数值型 (number)

DT: 日期时间型 (datetime)

MD5: 信息摘要算法 (MD5 Message-Digest Algorithm)

rRNA: 核糖体核糖核酸 (Ribosomal RNA)

S: 字符串型 (string)

5 数据及数据文件要求

5.1 数据文件组成

单细胞转录组数据集应包含数据文件和文件描述信息。

5.2 数据文件

单细胞转录组数据集文件应包含实验/测序数据、基因表达文件，宜包含元数据文件、聚类文件、基因列表文件，可包含其他文件。

5.2.1 实验/测序数据

实验/测序数据元数据和测序数据文件。

5.2.1.1 元数据

实验/测序数据的元数据应包含项目编号、样本编号、数据文件类型、测序平台和测序仪型号、实验标题、文库构建策略、文库来源、文库选择、文库设置、测序数据文件名称和文件的MD5值，可包含插入片段长度、插入片段标准差、文库结构、文库设计描述、文库构建方法描述。

5.2.1.2 测序数据文件

测序数据文件宜为FASTQ格式的文件，文件后缀宜为.fq，压缩之后的文件后缀宜为.fq.gz。

5.2.2 基因表达文件

基因表达文件可表示为表达矩阵文件，记录的是每个细胞在各基因的表达值。

- a) 标题行应包含基因和单细胞名称/细胞条码。
- b) 为考虑数据读取应用的通用和便利性，规范使用标准的格式文件，文件名后缀可为.txt、.txt.gz、.tsv、.tsv.gz、.csv、.csv.gz。

以tsv格式为例：

| gene | cell_1 | cell_2 | cell_3 | ... |
|-------|--------|--------|--------|-----|
| Trp53 | 0 | 0 | 0 | |
| Apoe | 0 | 5.098 | 0 | |
| Tlr4 | 0 | 0 | 0 | |
| Lep | 0 | 0 | 0.123 | |
| Il6 | 1.234 | 0 | 0 | |
| ... | | | | |

- c) 可采用稀疏矩阵的格式以节约资源。

5.2.3 元数据文件

应包含细胞水平的注释，文件后缀可为.txt或.txt.gz。

具体格式要求。

- a) 元数据文件应至少有2列。
- b) 第一行应包含：
 - 1) type: 类型，用于声明元数据类型；
 - 2) cluster group: 簇类别；
 - 3) sub_cluster group: 子簇类别；
 - 4) average_intensity: 连续分数，值为浮点数；
 - 5) sample name: 样本名称。

- c) 宜包含: experiment accession: 实验编号。

示例：

| Type | cluster | sub_cluster | average_intensity | sample_name | experiment_accession |
|--------|---------|-------------|-------------------|-------------|----------------------|
| cell_1 | clst_A | clst_A_1 | 6.687 | sample1 | CNXxxxxxx |
| cell_2 | clst_A | clst_A_1 | -12.625 | sample2 | CNXxxxxxx |
| ... | | | | | |

5.2.4 聚类文件

为考虑数据读取应用的通用和便利性，规范使用标准的格式文件，应包含任何聚类 and 可选的特定聚类的元数据，文件后缀可为.txt或.txt.gz。

具体格式要求。

- a) 聚类文件应至少有三列。
- b) 标题行应包含“name”、“X”、“Y”和细胞水平的注释，可包含“Z”，“X”、“Y”、“Z”可为UMAP、t-SNE等分析结果中的坐标标识。
- c) 第一行应包含：
 - 1) type: 类型，用于声明元数据类型；
 - 2) X、Y、Z: UMAP、t-SNE等分析中在不同坐标上的取值；
 - 3) group: 类别。
 - 4) intensity: 连续分数，值为浮点数。

注：“X”、“Y”和“Z”列的值应设置为“numeric”。

示例：

| Type | X | Y | Z | Group | intensity |
|--------|--------|--------|--------|-------|-----------|
| cell_1 | 34.472 | 32.211 | 60.035 | C | 0.719 |
| cell_2 | 15.975 | 10.043 | 21.424 | B | 0.904 |
| ... | | | | | |

5.2.5 其他文件

任何其他支撑文档或文件，具体格式不限。

6 数据元目录

6.1 属性

单细胞转录组数据集数据元目录应包含公用属性和专用属性

6.2 数据元目录公用属性

数据元目录公共属性如表1所示。

表1 数据元目录公共属性

| 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 |

6.3 数据元目录专用属性

6.3.1 组成

单细胞转录组数据元目录专用属性应包括实验/测序信息、生物信息分析、质控信息三个部分。每个数据元宜包含标识符、名称、定义、信息保护、单位、数据类型和数据元允许值，具体数据元目录参考附录A。部分数据元允许值宜以数据元值域代码形式表示，参考附录B。

6.3.2 实验/测序信息

应为描述实验/测序过程中的数据元，如细胞类型、细胞数量、细胞活率、cDNA浓度、文库浓度、文库体积、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。

6.3.3 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。

6.3.4 质控信息

描述整个测序过程质量监控的数据元，例如总数据量、测序深度、测序数据量等。

7 数据归档目录

7.1 数据归档目录结构

单细胞数据归档目录结构如表2所示。

表2 单细胞数据归档目录结构

| 归档根目录 | 第一级 | 第二级 | 第三级 |
|---------|-------------------|------------------|-----------------------|
| Path_id | Project_accession | Sample_accession | Single_cell_accession |

7.2 单细胞数据归档目录要求

7.2.1 概述

本章节是对7.1单细胞归档目录结构具体每级目录的情况说明。除归档根目录外，第一级目录包含目录标识符、目录名称、目录定义，其他级目录均包含目录标识符、目录名称、目录定义及父目录。

7.2.2 第一级目录

单细胞转录组学数据归档第一级目录应符合以下要求：

目录标识符：DI01.01.01

目录名称：Project_accession/项目编号

定义：项目编号或其他可分子类别的标识符

7.2.3 第二级目录

细胞转录组学数据归档第二级目录应符合以下要求：

目录标识符：DI01.02.01

目录名称：Sample_accession/样品编号

定义：样本编号

父目录：DI01.01.01

7.2.4 第三级目录

细胞转录组学数据归档第三级目录应符合以下要求：

目录标识符：DI01.03.01

目录名称：Single_cell_accession/单细胞编号

定义：存放该单细胞数据编号对应的数据文件

父目录：DI01.02.01

8 数据安全

8.1 单细胞转录组数据分级分类原则

8.1.1 单细胞转录组数据分类：根据单细胞数据的内容、用途和来源对其进行分类。

8.1.2 单细胞转录组数据分级：依照单细胞数据的内容敏感程度、价值、影响情况对不同数据进行敏感级别的划分。

8.2 单细胞转录组数据集的分级分类方式

将单细胞转录组数据集按照内容进行分级分类，根据数据的级别划分不同的管理方式（私有、受控及公开），以确保数据价值可以得到合理评估以及数据安全可以得到合理保障，见表3。

表3 单细胞转录组数据集的分级分类表

| 数据类型\管理方式 | 私有 | 受控 | 公开 |
|----------------|-------|-------|------|
| 人及人源数据（含宿主） | 高敏感数据 | 高敏感数据 | 公开数据 |
| 人及人源数据（不含宿主） | 高敏感数据 | 敏感数据 | 公开数据 |
| 动物、植物、微生物等非人数据 | 高敏感数据 | 敏感数据 | 公开数据 |

公开数据是指符合《中华人民共和国人类遗传资源管理条例》等法规的可公开的数据。

8.3 单细胞转录组数据集的保密性、完整性价值分级关系

单细胞转录组数据集的保密性、完整性价值分级关系见表4。

表4 单细胞转录组数据集的保密性、完整性价值分级关系表

| 数据类型\管理方式 | 完整性 | 保密性 |
|-----------|-----|-----|
| 高敏感数据 | 高 | 绝密 |
| 敏感数据 | 高 | 机密 |
| 公开数据 | 高 | 公开 |

高级完整性是指未经单细胞转录组数据集作者方授权的修改或破坏会对评估造成重大影响，可能造成标准相关的数据集处理业务无法使用，难以修补。

保密性级别分类为绝密的，为涉及最重要的秘密信息，泄露会使得国家的数据安全利益遭受严重损害或者重大损失。保密级别分类为机密的，为涉及重要的秘密信息，泄露会使得对数据集具有著作权或管理权的人员遭受重大损失。保密级别分类为公开的，不涉及国家及个人的数据安全和利益，可以对外公开披露。

附录 A (资料性) 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。若有新的数据元加入可以顺延排入。

A.2 实验信息

实验信息如表A.1所示。

表A.1 实验信息

| 标识符 | 名称 | 定义 | 信息保护 | 单位 | 数据类型 | 数据元允许值 |
|----------------|-------------------|-------------------------------------|------|-------|------|--------|
| DE06.01.001.00 | 分选细胞类型 | 流式分选的目标群体，以荧光染料为标准，如 DAPI+。 | 不保护 | | S | |
| DE06.01.002.00 | 目标群体比例 | 流式分析细胞群体，目标群体占总群体的比例。 | 不保护 | | N | |
| DE06.01.003.00 | 384 板每板分选时间 | 流式分选整板单细胞的时间。 | 不保护 | | DT | |
| DE06.01.004.00 | 板数 | 同一样品试剂分选的板数。 | 不保护 | | S | |
| DE06.01.005.00 | PCR 循环数 | 信使核糖核酸（mRNA）反转录成 cDNA 并扩增。 | 不保护 | | S | |
| DE06.01.006.00 | cDNA 抽检合格率 | 抽检 cDNA 产物的合格率。 | 不保护 | | N | |
| DE06.01.007.00 | Tn5 打断后 PCR 抽检合格率 | 抽检 PCR 产物的合格率。 | 不保护 | | N | |
| DE06.01.008.00 | 片筛磁珠浓度 | Pooling（池化）后片筛添加的磁珠浓度，影响产物片段范围。 | 不保护 | | S | |
| DE06.01.009.00 | 片筛后浓度 | Qubit 检测片筛后的双链脱氧核糖核酸浓度，计量单位为 ng/μL。 | 不保护 | ng/μL | S | |
| DE06.01.010.00 | 片筛后体积 | 片筛后，产物体积，计量单位为 μL。 | 不保护 | μL | S | |
| DE06.01.014.00 | 下机序列数目 | 每条测序通道下机数据量，计量单位为 M。 | 不保护 | M | S | |
| DE06.01.015.00 | 网页报告链接地址 | 测序下机报告。 | 不保护 | | S | |
| DE06.01.016.00 | 实验日期 | 不同实验步骤的具体日期。 | 不保护 | | DT | |
| DE06.01.017.00 | 细胞核提取时间 | 组织解冻开始，进行提核、计数完毕的时间。 | 不保护 | | DT | |
| DE06.01.018.00 | 细胞核提取体积 | 提取完的细胞核悬液总体积，计量单位 μL。 | 不保护 | μL | S | |

表A.1 实验信息（续）

| 标识符 | 名称 | 定义 | 信息保护 | 单位 | 数据类型 | 数据元允许值 |
|----------------|---------|--|------|---------------|------|--------|
| DE06.01.019.00 | 细胞核提取浓度 | 提取完的细胞核悬液计数浓度， 计量单位 cell/ μ L。 | 不保护 | cell/ μ L | S | |
| DE06.01.020.00 | 流式上样速率 | 流式上样的速率参数，计量单位 events/s。 | 不保护 | events/s | S | |
| DE06.01.021.00 | 得率 | 流式分选单细胞效果指标，实际分 选获得的细胞数占目标数的比 例。 | 不保护 | | S | |

A.3 建库测序信息

建库测序信息如表A.2所示。

表A.2 建库测序信息

| 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 | 测序仪名称。 | 不保护 | — | S | — |
| DE07.01.004.00 | 测序仪名称 | 测序平台名称。 | 不保护 | — | S | B.1 测序仪名称代码表 |
| DE07.01.005.00 | 测序仪标识符 | 测序仪标识符。 | 不保护 | — | S | |
| DE07.01.006.00 | 测序平台名称 | 测序类型。 | 不保护 | — | S | B.2 测序平台代码表 |
| DE07.01.007.00 | 测序开始时间 | 测序开始当日的的时间。 | 不保护 | — | DT | — |
| DE07.01.008.00 | 测序完成时间 | 测序完成当日的的时间。 | 不保护 | — | DT | — |
| DE07.01.009.00 | 文库标识符 | 测序文库标识符。 | 不保护 | — | S | — |
| DE07.01.010.00 | 文库构建策略 | 文库构建策略说明了文库的测序技术。 | 不保护 | — | S | B.3 文库构建策略代码表 |
| DE07.01.011.00 | 文库名称 | 文库的名称。 | 不保护 | — | S | — |
| DE07.01.012.00 | 文库体积 | 单链脱氧核糖核酸文库的体积，计 量单位为 μ L。 | 不保护 | μ L | S | — |
| DE07.01.013.00 | 文库类型 | 文库的类型说明。 | 不保护 | — | S | — |
| DE07.01.014.00 | 文库数量 | 文库数量。 | 不保护 | — | N | — |
| DE07.01.015.00 | 芯片号 | 芯片号编码。 | 不保护 | — | S | — |
| DE07.01.016.00 | 测序通道号 | 测序通道号。 | 不保护 | — | S | — |

表A.2 建库测序信息（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 | 机器号 | 机器号。 | 不保护 | — | S | — |
| DE07.01.018.00 | 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 | — |
| DE07.01.019.00 | FASTQ 格式文件唯一编号 | FASTQ 格式文件唯一编号。 | 不保护 | — | S | — |
| DE07.01.020.00 | 下机地 | 数据下机地区。 | 不保护 | — | S | GB/T 2260—2007 |
| DE07.01.021.00 | 分子类型 | 提交序列的体内分子类型。 | 不保护 | — | S | B.4 分子类型代码表 |
| DE07.01.022.00 | 是否部分基因组 | 是否部分基因组的分类代码。 | 不保护 | — | S | B.5 是否代码表 |
| DE07.01.023.00 | 测序文件类型 | 序列数据的存储格式。 | 不保护 | — | S | B.6 文件类型代码表 |
| DE07.01.024.00 | 文件 MD5 值 | 文件 MD5 值，由 32 个字符(字母、数字)的字符串组成，用于验证文件完整性。 | 不保护 | — | S | — |
| DE07.01.025.00 | 文库设置 | 文库设置说明。 | 不保护 | — | S | B.7 文库设置代码表 |
| DE07.01.026.00 | 文库选项 | 文库选项说明了用于选择、排除、富集或筛选待测样本的方法。 | 不保护 | — | S | — |
| DE07.01.027.00 | 文库来源 | 文库来源说明了测序源材料的类型。 | 不保护 | — | S | B.8 文库来源代码表 |
| DE07.01.028.00 | 文库数量 | 文库数量。 | 不保护 | — | N | — |

A.4 生物信息分析

生物信息分析如表A.3所示。

表A.3 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 | 过滤软件名称 | 信息分析过程中过滤软件名称。 | 不保护 | — | S | — |
| DE08.01.002.00 | 过滤软件版本 | 信息分析过程中过滤软件版本号。 | 不保护 | — | S | — |
| DE08.01.003.00 | 过滤软件参数 | 信息分析过程中过滤软件参数信息。 | 不保护 | — | S | — |
| DE08.01.004.00 | 比对软件名称 | 信息分析过程中比对软件名称。 | 不保护 | — | S | — |
| DE08.01.005.00 | 比对软件版本 | 信息分析过程中比对软件版本号。 | 不保护 | — | S | — |
| DE08.01.006.00 | 比对软件参数 | 信息分析过程中比对软件参数信息。 | 不保护 | — | S | — |
| DE08.01.007.00 | 标准化分析名称 | 信息分析过程中标准化分析软件名称。 | 不保护 | — | S | — |
| DE08.01.008.00 | 标准化分析版本 | 信息分析过程中标准化分析软件版本号。 | 不保护 | — | S | — |

表A.3 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 |
|----------------|------------------|-----------------------------|------|----|------|--------|
| DE08.01.009.00 | 标准化分析参数 | 信息分析过程中标准化分析软件参数信息。 | 不保护 | — | S | — |
| DE08.01.010.00 | 降维分析名称 | 信息分析过程中降维分析软件名称。 | 不保护 | — | S | — |
| DE08.01.011.00 | 降维分析版本 | 信息分析过程中降维分析软件版本号。 | 不保护 | — | S | — |
| DE08.01.012.00 | 降维分析参数 | 信息分析过程中降维软件参数信息。 | 不保护 | — | S | — |
| DE08.01.013.00 | 聚类分析软件名称 | 信息分析过程中基因表达量聚类分析软件名称。 | 不保护 | — | S | — |
| DE08.01.14.00 | 聚类分析软件版本 | 信息分析过程中基因表达量聚类分析软件版本号。 | 不保护 | — | S | — |
| DE08.01.015.00 | 聚类分析软件参数 | 信息分析过程中基因表达量聚类分析软件参数信息。 | 不保护 | — | S | — |
| DE08.01.016.00 | 差异表达基因检测软件名称 | 信息分析过程中差异表达基因检测软件名称。 | 不保护 | — | S | — |
| DE08.01.017.00 | 差异表达基因检测软件版本 | 信息分析过程中差异表达基因检测软件版本号。 | 不保护 | — | S | — |
| DE08.01.018.00 | 差异表达基因检测软件参数 | 信息分析过程中差异表达基因检测软件参数信息。 | 不保护 | — | S | — |
| DE08.01.019.00 | Go & kegg 分析软件名称 | 信息分析过程中 Go & kegg 分析软件名称。 | 不保护 | — | S | — |
| DE08.01.020.00 | Go & kegg 分析软件版本 | 信息分析过程中 Go & kegg 分析软件版本号。 | 不保护 | — | S | — |
| DE08.01.021.00 | Go & kegg 分析软件参数 | 信息分析过程中 Go & kegg 分析软件参数信息。 | 不保护 | — | S | — |
| DE08.01.022.00 | 时间序列分析软件名称 | 信息分析过程中时间序列分析软件名称。 | 不保护 | — | S | — |
| DE08.01.023.00 | 时间序列分析软件版本 | 信息分析过程中时间序列分析软件版本号。 | 不保护 | — | S | — |
| DE08.01.024.00 | 时间序列分析软件参数 | 信息分析过程中时间序列分析软件参数信息。 | 不保护 | — | S | — |

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生的数据。 | 不保护 | — | S | — |
| DE01.01.002.00 | 项目名称 | 项目名称。 | 不保护 | — | S | — |
| DE09.01.003.00 | 个体编号 | 样本来源的个体编号。 | 不保护 | — | S | — |
| DE09.01.004.00 | 样本编号 | 样本编号。 | 保护 | — | S | — |
| DE09.01.005.00 | 样本名称 | 分析结果中样本名称。 | 不保护 | — | S | — |
| DE01.01.006.00 | 样品浓度 | 样品的浓度值，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 | 总数据量 | 总数据量。 | 不保护 | bp | N | — |
| DE01.01.009.00 | 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 | — |
| DE01.01.010.00 | 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 | — | — |
| DE01.01.011.00 | 唯一下机序列的比对率 | 唯一下机序列的比对率。 | 不保护 | % | N | — |
| DE01.01.012.00 | 插入片段大小 | 插入片段的大小。 | 不保护 | — | N | — |
| DE01.01.013.00 | 参考基因组的比对率 | 与参考基因组的比对率。 | 不保护 | % | N | — |
| DE01.01.014.00 | 重复率 | 重复下机序列占有下机序列的比率。 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 | — |
| DE01.01.015.00 | 错配率 | 错配率。 | 不保护 | % | N | — |
| DE01.01.016.00 | 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 | — |
| DE01.01.017.00 | 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 | — |
| DE01.01.021.00 | 总体 Q20 值 | 测序数据中，碱基识别质量值大于 20 的碱基占有所有碱基的比例。 注：碱基识别质量值为 20 时，表示碱基的正确率为 99%以上，Q20≥95%，则表示测序数据中 95%以上的碱基质量之大于 20。 | 不保护 | — | S | — |
| DE01.01.022.00 | 总体 Q30 值 | 测序数据中，碱基识别质量值大于 30 的碱基占有所有碱基的比例。 注：碱基识别质量值为 30 时，表示碱基的正确率为 99.9%以上，Q30≥85%，则表示测序数据中 85%以上的碱基质量 | 不保护 | — | S | — |
| DE01.01.023.00 | 下机序列 1 的 Q20 值 | 表示下机序列 1 的质量值大于 20 的碱基所占百分比。 | 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 |
|----------------|----------------|------------------------------|------|----|------|--------|
| DE01.01.024.00 | 下机序列 2 的 Q20 值 | 表示下机序列 2 的质量值大于 20 的碱基所占百分比。 | 不保护 | % | N | — |
| DE01.01.025.00 | 下机序列 1 的 Q30 值 | 表示下机序列 1 的质量值大于 30 的碱基所占百分比。 | 不保护 | % | S | — |
| DE01.01.026.00 | 下机序列 2 的 Q30 值 | 表示下机序列 2 的质量值大于 30 的碱基所占百分比。 | 不保护 | % | S | — |
| DE01.01.027.00 | rRNA 下机序列比例 | 比对到 rRNA 区域的下机序列百分比。 | 不保护 | % | N | — |
| DE01.01.028.00 | 细胞平均下机序列数 | 平均每个细胞鉴定出的下机序列数。 | 不保护 | 个 | N | — |
| DE01.01.029.00 | 过滤数据量 | 过滤数据量。 | 不保护 | bp | N | — |
| DE01.01.030.00 | 过滤数据率 | 过滤数据率。 | 不保护 | bp | N | — |
| DE07.01.031.00 | 过滤后数据量 | 过滤后数据量。 | 不保护 | bp | N | — |
| DE01.01.032.00 | 过滤后下机序列数目 | 过滤后的总下机序列数目。 | 不保护 | — | S | — |
| DE01.01.033.00 | 比对率 | 比对到参考基因组的下机序列百分比。 | 不保护 | % | N | — |
| DE01.01.034.00 | 唯一比对率 | 唯一比对的下机序列百分比。 | 不保护 | % | N | — |

附 录 B
(资料性)
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“00”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 | 测序仪系列 | 型号 |
|----|----------------------------------|--|
| 01 | Illumina 公司 Genome Analyzer 系列 | Genome Analyzer/Genome Analyzer II/Genome |
| 02 | Illumina 公司 Genome Analyzer 系列 | Analyzer IIx |
| 03 | Illumina 公司 HiSeq 系列 | HiSeq SQ/1000/1500/2000/2500/X Ten/X |
| 04 | Illumina 公司 HiSeq 系列 | Five/3000/4000 |
| 05 | Illumina 公司 MiSeq 系列 | MiSeq/MiSeq Dx/FGx |
| 06 | Illumina 公司 NextSeq 系列 | NextSeq500/550 |
| 07 | Illumina 公司 MiniSeq 系列 | MiniSeq |
| 08 | Illumina 公司 iSeq 系列 | iSeq 100 |
| 09 | Illumina 公司 NovaSeq 系列 | NovaSeq 5000/6000/TM |
| 10 | BGI 公司 BGISEQ 系列 | BGISEQ- 1000/50/100/500/500RS/200RS/2000RS/200CX/2000CX/500CX |
| 11 | BGI 公司 MGISEQ 系列 | MGISEQ-200/2000/200RS/2000RS/200CX/2000CX |
| 12 | BGI 公司 DNBSEQ 系列 | DNBSEQ-G50/G400/E/T1/T5/T7/T10/T20 |
| 13 | Oxford Nanopore MinION | MinION |
| 14 | Oxford Nanopore GridION | GridION |
| 15 | Oxford Nanopore PromethION | PromethION |
| 16 | Berry Genomics NextSeq CN500 | NextSeq CN500 |
| 17 | PacBio SMRT PacBio | PacBio RS/RS II/Sequel |
| 18 | CapitalBio BioelectronSeq 4000 | BioelectronSeq 4000 |
| 19 | Thermo Fisher Ion Torrent PGM | Ion Torrent PGM |
| 20 | Thermo Fisher Ion Torrent Proton | Ion Torrent Proton |
| 21 | Thermo Fisher Ion Torrent S5 | Ion Torrent S5/Ion Torrent S5 XL |
| 22 | Bionano Genomics BioNano 系列 | BioNano IRYS/SAPHYR |
| 23 | Complete Genomics | Complete Genomics |
| 24 | DAAN GENE | DA8600 |
| 25 | Helicos BioSciences Corporation | Helicos HeliScope |

表B.1 测序仪名称代码表（续）

| 代码 | 测序仪系列 | 型号 |
|----|-------------|---------------|
| 26 | HYK Genetic | HYK-PSTAR-IIA |
| 27 | Other | — |

B.2 文库构建策略代码

文库构建策略代码规定了文库构建策略的代码。

采用1位数字顺序代码，从“1”开始编码，按升序排列，见表B.2。

表B.2 文库构建策略代码表

| 代码 | 文库构建策略 | 备注 |
|-----|------------------------|---|
| 1 | WGA | 非 pcr 扩增的全基因组的随机测序 |
| 101 | AMPLICON | 重叠或不同的 PCR 或RT-PCR 产物测序 |
| 102 | CLONEEND | 克隆末端（5'、3' 或两端）测序 |
| 103 | FINISHING | 在现有的覆盖度下以补空为目的测序 |
| 104 | ChIP-Seq | 染色质免疫沉淀物的直接测序 |
| 105 | MNase-Seq | MNase 消化后的直接测序 |
| 106 | DNase-Hypersensitivity | 对超敏位点或用 DNaseI 更容易切割的开放染色质片段的测序 |
| 107 | Bisulfite-Seq | 用亚硫酸氢盐将 DNA 的非甲基化胞嘧啶残基转化为尿嘧啶后的测序 |
| 108 | EST | cDNA 模板的单次测序 |
| 109 | FL-cDNA | cDNA 模板的全长测序 |
| 110 | CTS | 级联标签测序 |
| 2 | WGS | 全基因组的随机测序 |
| 201 | MRE-Seq | 甲基化敏感性限制性酶测序策略 |
| 202 | MeDIP-Seq | 甲基化 DNA 免疫沉淀测序策略 |
| 203 | MBD-Seq | 甲基化片段的直接测序策略 |
| 204 | Synthetic Long Read | 对大的 DNA 片段进行合并和条形码标记以利于片段的组装 |
| 205 | ssRNA-seq | 链特异性转录组测序 |
| 206 | ncRNA-seq | 捕获其他非编码 RNA 类型，包括翻译后修饰类型，如 snRNA（小核 RNA）或 snoRNA（小核仁 RNA）或表达调控类型，如 siRNA（小干扰 RNA）或 piRNA/piwi/RNA（与 piwi 蛋白相互作用的 RNA） |
| 207 | Hi-C | 染色体构象捕获技术将生物素标记的核苷酸结合在接头处，能够进行嵌合 DNA 连接点的选择性纯化，然后进行深度测序。 |
| 208 | ATAC-seq | 转座酶可接近性核染色质测序策略 (ATAC)，用于研究全基因组染色质的可接近性。使用设计的 Tn5 转座酶切割 DNA 并将引物 DNA 序列整合到切割的基因组 DNA 中，是 DNase-seq 的替代方法。 |
| 209 | RAD-Seq | 限制性位点相关的 DNA 序列 |
| 210 | VALIDATION | |
| 3 | WXS | 从基因组中选择的外显子区域的随机测序 |
| 301 | FAIRE-seq | 甲醛辅助的调控元件分离，揭示开放染色质区域。 |
| 302 | SELEX | 指数富集配体的系统进化 |

表B.2 文库构建策略代码表（续）

| 代码 | 文库构建策略 | 备注 |
|-----|--|---|
| 303 | RIP_seq | RNA 免疫沉淀物的直接测序（包括 CLIP-Seq、HITS-CLIP 和PAR-CLIP） |
| 304 | ChIA_PET | 邻近连接的染色质免疫沉淀物的直接测序 |
| 305 | Targeted-Capture | |
| 306 | Tethered Chromatin Conformation Capture | |
| 307 | OTHER | |
| 4 | RNA-Seq | 整个转录组的随机测序 |
| 5 | miRNA-Seq | 小 miRNA 的随机测序 |
| 6 | Tn-Seq | 从转座子插入位点开始的测序 |
| 7 | WCS | 从基因组中分离的整个染色体或其他复制子的随机测序 |
| 8 | CLONE | 基于基因组克隆（分级）的测序 |
| 9 | POOLCLONE | 混合克隆的鸟枪法建库测序（通常是 BACs 和 Fosmids） |

B.3 分子类型代码

分子类型代码规定了分子类型的代码。

采用1位数字顺序代码，从“1”开始编码，按升序排列，见表B.3。

表B.3 分子类型代码表

| 代码 | 分子类型 |
|----|-------------|
| 1 | genomic DNA |
| 2 | genomic RNA |
| 3 | viral cRNA |

B.4 是否代码

是否代码规定了是否的代码。

采用1位数字顺序代码，从“1”开始编码，按升序排列，见表B.4。

表B.4 是否代码表

| 代码 | 是否 |
|----|----|
| 1 | 是 |
| 2 | 否 |

B.5 文件类型代码

文件类型代码规定了文件类型的代码。

采用1位数字顺序代码，从“1”开始编码，按升序排列，见表B.5。

表B.5 文件类型代码表

| 代码 | 文件类型 |
|----|--------------------------|
| 1 | FASTQ |
| 2 | BAM |
| 3 | Expression matrix/表达矩阵文件 |
| 4 | Metadata file/元数据文件 |
| 5 | Cluster file/聚类文件 |

B.6 文库设置代码

文库设置代码规定了文库设置的代码。

采用1位数字顺序代码，从“1”开始编码，按升序排列，见表B.6。

表B.6 文库设置代码表

| 代码 | 文库设置 | 备注 |
|----|-----------------|------------|
| 1 | FRAGMENT/SINGLE | 单末端测序 read |
| 2 | PAIRED | 成对 |

B.7 文库来源代码

文库来源代码规定了文库来源的代码。

采用1位数字顺序代码，从“1”开始编码，按升序排列，见表B.7。

表B.7 文库来源代码表

| 代码 | 文库来源 | 备注 |
|----|--------------------|-------------------------------------|
| 1 | GENOMIC | 基因组 DNA（包括来自基因组 DNA 的PCR 产物） |
| 2 | TRANSCRIPTOMIC | 转录产物或非基因组 DNA（EST、cDNA、RT-PCR、筛选文库） |
| 3 | METAGENOMIC | 来自宏基因组的混合物质 |
| 4 | METATRANSCRIPTOMIC | 来自自然环境中的目标微生物的转录产物 |
| 5 | SYNTHETIC | 合成 DNA |
| 6 | VIRAL RNA | 病毒 RNA |
| 7 | OTHER | |