ICS 07.080
CCS C 07

团　　体　　标　　准

T/SZAS 39—2021

单细胞转录组学数据集

Dataset of single-cell transcriptomics

2021-09-27 发布　　2021-09-28 实施

深圳市标准化协会　　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。

本文件由深圳市标准化协会提出和归口。

本文件起草单位：深圳华大生命科学研究院、青岛华大基因研究院、中国科学院武汉植物园、广州中医药大学、西北农林科技大学、苏州极客基因科技有限公司、深圳市坪山区尼奥基因组学研究院、深圳市易基因科技有限公司、深圳华大基因科技有限公司。

本文件主要起草人：魏晓锋、陈凤珍、郭学芹、游丽金、杨晓萍、古槿、雍军、徐志成、岳建辉、刘克、丁远彤、陈超、吴亮、李良、王然、刘群、石涛、郑夏生、姜雨、胡琪、王君文、张曦、刘石平、刘龙奇、曾文君、李启沅、王博、王韧、吴昊、李倩一。

单细胞转录组学数据集

1 范围

本文件规定了单细胞转录组学数据的范围、数据文件要求、数据元目录和数据归档目录。

本文件适用于单细胞转录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 35890-2018 高通量测序数据序列格式规范

T/SZAS 13-2019 基因组学数据集

T/SZAS 14-2019 转录组学数据集

T/SZAS 15-2019 人体肠道宏基因组学数据集

3 术语与定义

T/SZAS 13-2019界定的以及下列术语和定义适用于本文件。

3.1

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。

3.2

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。

3.3

实验/测序 Experiment/Run

描述样本的建库、测序仪器、测序方法等实验信息，一个实验通常关联一个项目和一个样本。一个实验产生的测序文件被称之为测序，特指一个样本产生的测序文件。

3.4

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）和其测序质量信息的、每四行表示一条序列的标准格式。

[来源：GB/T 35890-2018，3.9]

3.5

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。

[来源：T/SZAS 14-2019，2.1.2]

3.6

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水平。

3.7

元数据文件 Metadata file

包含细胞水平的注释，制表符分隔的文本文件。

3.8

聚类文件 Cluster file

包含任何聚类和可选的特定聚类的元数据的制表符分隔的文本文件。

3.9

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。

DNA：脱氧核糖核酸（deoxyribonucleic acid）

cDNA：互补脱氧核糖核酸（complementary DNA）

DNB：DNA纳米球（DNA nanoball）

S：字串型（string）

N：数值型（number）

DT：日期时间型（datetime）

MD5：信息摘要算法（message-digest algorithm）

5 数据文件要求

5.1 数据文件组成

单细胞转录组学数据集应包含数据文件和文件描述信息。

5.2 数据文件

单细胞转录组学数据集数据文件应包含实验/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/测序数据应包含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

5.2.2.1 应包含基因和单细胞名称。
5.2.2.2 文件格式应为基因表达矩阵。
5.2.2.3 文件后缀可为.txt、.tsv、.csv，压缩之后的文件后缀可为.txt.gz、.tsv.gz、csv.gz。

5.2.3 元数据文件

5.2.3.1 应包含细胞水平的注释、样本名称，宜包含实验编号。
5.2.3.2 文件后缀宜为.txt，压缩之后的文件后缀宜为.txt.gz。

5.2.4 聚类文件

5.2.4.1 应包含任何聚类和可选的特定聚类的元数据。
5.2.4.2 文件后缀宜为.txt，压缩之后的文件后缀宜为.txt.gz。

5.2.5 基因列表文件

5.2.5.1 应包含所有注释的基因名称和基因 ID。
5.2.5.2 文件后缀可为.txt，压缩之后的文件后缀可为.txt.gz。

5.2.6 其他文件

任何其他支撑文档或文件，具体格式不限。

5.3 文件描述信息

单细胞转录组学数据集文件描述信息应包含物种分类号、物种和数据文件的类型、描述、名称、MD5值，可包含聚类文件的坐标轴标签、值阈的最大值和最小值。

6 数据元目录

6.1 属性

单细胞转录组学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

应符合T/SZAS 15-2019中4.1的规定。

6.3 数据元目录专用属性

6.3.1 组成

单细胞转录组学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。

6.3.3 生物信息分析

应为描述生物信息分析过程中的数据元，如过滤软件名称、过滤软件版本、过滤软件参数等。

6.3.4 质控信息

应为描述整个测序过程质量监控的数据元，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

单细胞转录组学数据归档目录应分为三级，结构如表1所示。

表1 单细胞转录组学数据归档目录结构

第一级	第二级	第三级
Project_accession	Sample_accession	Single_cell_accession
项目编号	样本编号	单细胞编号

7.2 数据归档目录要求

7.2.1 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；
——目录定义：项目编号或其他可分子类别的标识符。

7.2.2 第二级目录

单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/样本编号；
——目录定义：样本编号；
——父目录：DI01.01.01。

7.2.3 第三级目录

单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；
——目录定义：存放该单细胞数据编号对应的数据文件；
——父目录：DI01.02.01。

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。且有新的数据元加入可以顺延排入。

A.2 实验/测序信息

实验/测序信息如表A.1所示。

表 A.1 实验/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	项目编号	项目的编号。	不保护	-	S	-
DE07.01.002.00	实验标题	实验的标题。	不保护	-	S	-
DE07.01.003.00	细胞类型	细胞的类型。	不保护	-	S	-
DE07.01.004.00	细胞数量[a]	细胞的数量。	不保护	个	S	-
DE07.01.005.00	细胞活率[a]	活细胞的比例。	不保护	%	S	-
DE07.01.006.00	cDNA 浓度	cDNA 的浓度。	不保护	ng/ul	S	-
DE07.01.007.00	文库浓度	文库的浓度。	不保护	-	S	-
DE07.01.008.00	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	-
DE07.01.009.00	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	-
DE07.01.010.00	测序类型	测序类型。	不保护	-	S	-
DE07.01.011.00	测序仪名称	测序仪名称。	不保护	-	S	B.1 测序仪名称代码
DE07.01.012.00	测序仪标识符	测序仪标识符。	不保护	-	S	-
DE07.01.013.00	测序平台名称	测序平台名称。	不保护	-	S	-
DE07.01.014.00	测序开始时间	测序开始当日的时间。	不保护	-	DT	-
DE07.01.015.00	测序完成时间	测序完成当日的时间。	不保护	-	DT	-
DE07.01.016.00	文库标识符	测序文库标识符。	不保护	-	S	-

表A.1 实验/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.2 文库构建策略代码
DE07.01.018.00	文库名称	文库的名称。	不保护	-	S	-
DE07.01.019.00	文库体积	单链脱氧核糖核酸文库的体积。	不保护	μL	S	-
DE07.01.020.00	文库类型	文库的类型说明。	不保护	-	S	-
DE07.01.021.00	文库数量	文库数量。	不保护	个	N	-
DE07.01.022.00	芯片号	芯片号编码。	不保护	-	S	-
DE07.01.023.00	测序通道号	测序通道号。	不保护	-	S	-
DE07.01.024.00	机器号	机器号。	不保护	-	S	-
DE07.01.025.00	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	-
DE07.01.026.00	FASTQ 格式文件唯一编号	FASTQ 格式文件唯一编号。	不保护	-	S	-
DE07.01.027.00	下机地	数据下机地区。	不保护	-	S	
DE07.01.028.00	分子类型	提交序列的体内分子类型。	不保护	-	S	B.3 分子类型代码
DE07.01.029.00	是否部分基因组	是否部分基因组的分类代码。	不保护	-	S	B.4 是否代码
DE07.01.030.00	测序文件类型	序列数据的存储格式。	不保护	-	S	B.5 文件类型代码
DE07.01.031.00	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	-
DE07.01.032.00	文库设置	文库设置说明。	不保护	-	S	B.6 文库设置代码
DE07.01.033.00	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	-
DE07.01.034.00	文库来源	文库来源说明了测序源材料的类型。	不保护	-	S	B.7 文库来源代码
[a] 细胞数量和细胞活率适用于非单管实验。						

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	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	-
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	批次效应去除分析名称	信息分析过程中批次效应去除分析软件参数信息。	不保护	-	S	-
DE08.01.011.00	批次效应去除分析版本	信息分析过程中批次效应去除分析软件参数信息。	不保护	-	S	-
DE08.01.012.00	批次效应去除分析参数	信息分析过程中批次效应去除分析软件参数信息。	不保护	-	S	-
DE08.01.013.00	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	-
DE08.01.014.00	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	-
DE08.01.015.00	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	-
DE08.01.016.00	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	-
DE08.01.017.00	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	-
DE08.01.018.00	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	-
DE08.01.019.00	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	-
DE08.01.020.00	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	-
DE08.01.021.00	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	-

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
DE08.01.022.00	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	-
DE08.01.023.00	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	-
DE08.01.024.00	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	-
DE08.01.025.00	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	-
DE08.01.026.00	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	-
DE08.01.027.00	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.4 质控信息

质控信息如表A.3所示。

表A.3 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	个体编号	样本来源的个体编号。	不保护	-	S	-
DE01.01.004.00	样本编号	样本编号。	保护	-	S	-
DE01.01.005.00	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	样本的浓度值。	不保护	ng/μL	N	-
DE01.01.007.00	样品总量	样本的总重量。	不保护	mg	N	-
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。	不保护	Gb	N	-
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	bp	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	-

表A.3 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	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	-
DE01.01.018.00	1X 测序的覆盖率	测序深度大于或等于 1X 的碱基占被测碱基的比率。	不保护	%	N	-
DE01.01.019.00	4X 测序的覆盖率	测序深度大于或等于 4X 的碱基占被测碱基的比率。	不保护	%	N	-
DE01.01.020.00	20X 测序的覆盖率	测序深度大于或等于 20X 的碱基占被测碱基的比率。	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 的碱基占所有碱基的比例。	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 的碱基占所有碱基的比例。	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	-
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	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	-
DE01.01.028.00	过滤数据量	过滤数据量。	不保护	bp	N	-
DE01.01.029.00	过滤数据率	过滤数据率。	不保护	bp	N	-
DE01.01.030.00	过滤后数据量	过滤后数据量。	不保护	bp	N	-
DE01.01.031.00	过滤后下机序列数目	过滤后的总下机序列数目。	不保护	-	S	-
DE01.01.032.00	比对率	比对到参考基因组的下机序列百分比。	不保护	%	N	-
DE01.01.033.00	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

表B.1　测序仪名称代码表（续）

代码	测序仪系列	型号
24	DAAN GENE	DA8600
25	Helicos BioSciences Corporation	Helicos HeliScope
26	HYK Genetic	HYK-PSTAR-IIA
27	Other	-

B.2　文库构建策略代码

文库构建策略代码参考T/SZAS 13-2019附录B.9。

B.3　分子类型代码

分子类型代码参考T/SZAS 13-2019附录B.4。

B.4　是否代码

是否代码参考T/SZAS 13-2019附录B.5。

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.2。

表B.2　文件类型代码表

代码	文件类型
1	FASTQ
2	Gene expression matrix/基因表达矩阵文件
3	Metadata file/元数据文件
4	Cluster file/聚类文件
5	Gene list file/基因列表文件

B.6　文库设置代码

文库设置代码参考T/SZAS 13-2019附录B.7。

B.7　文库来源代码

文库来源代码参考T/SZAS 13-2019附录B.8。

参 考 文 献

[1] GB/T 31074-2014 科技平台 数据元设计与管理
[2] WS/T 363.1-2011 卫生信息数据元目录 第1部分：总则
[3] WS 370-2012 卫生信息基本数据集编制规范
[4] WS/T 364-2011 卫生信息数据元值域代码