ICS 07 080
Q 8499

团　　体　　标　　准

T/SZAS 14—2019

转录组学数据集

Dataset of transcriptomics

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。

转录组学数据集

1 范围

本标准规定了组学数据中有关转录组学数据的范围以及数据元的规范化定义,数据集包括转录组学相关数据元和值域。

本标准适用于组学数据中有关转录组学数据信息的存储、治理、交换与共享。

2 术语、定义和缩略语

以下术语、定义和缩略语适用于本文件。

2.1 术语和定义

2.1.1

FASTQ格式 FASTQ format

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。

2.1.2

测序通道 lane

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

2.2 缩略语

PCR：聚合酶链式反应（Polymerase Chain Reaction）

DNA：脱氧核糖核酸（deoxyribonucleic acid）

cDNA：互补脱氧核糖核酸（complementary DNA）

rRNA：核糖体核糖核酸（Ribosomal RNA）

mRNA：信使核糖核酸（messenger RNA）

Tn5：一种转座酶名称

Qubit：一种检测技术名称

DNB：DNA纳米球（DNA Nanoball）

SNP：单核苷酸多态性(Single Nucleotide Polymorphism)

INDEL：插入/缺失（Insertions/Deletions）

ERCC：一个专门为了定制一套spike-in RNA而成立的组织。主要的工作是设计了一套非常好用的spike-in RNA，方便microarray，以及RNA-Seq进行内参定量（External RNA Controls Consortium）

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）

3 数据元目录

3.1 数据元目录公用属性

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	公司名称
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

3.2 数据元目录专用属性

3.2.1 转录组学数据元目录专用属性包括实验信息、测序信息、生物信息分析、质控信息四部分。

3.2.2 实验信息描述实验过程中的数据元，例如分选细胞类型、目标群体比例、384 板每板分选时间、板数、聚合酶链式反应（PCR）循环数、互补脱氧核糖核酸（cDNA）抽检合格率 、Tn5 打断后 PCR 抽检合格率、片筛磁珠浓度等。

3.2.3 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。

3.2.4 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。

3.2.5 质控信息描述整个测序过程质量监控的数据元，例如 ERCC 的下机序列比例、线粒体外显子下机序列比例、核糖体核糖核酸（rRNA）下机序列比例、细胞平均下机序列数、内含子下机序列比例等。

3.2.6 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。

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 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	
DE06.01.011.00	环化投入量	进行环化投入的产物量，计量单位 ng。	不保护	ng	N	

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
DE06.01.012.00	单链脱氧核糖核酸浓度	Qubit 环化后单链脱氧核糖核酸浓度，计量单位为 ng/μL。	不保护	ng/μL	S	
DE06.01.013.00	DNB 浓度	测序制作 DNA 纳米球后 DNB 浓度，计量单位为 ng/μL。	不保护	ng/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 ul。	不保护	ul	S	
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/ul。	不保护	cell/ul	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	测序开始当日的时间。	不保护		DT	
DE07.01.008.00	测序完成时间	测序完成当日的时间。	不保护		DT	
DE07.01.009.00	文库标识符	测序文库标识符。	不保护		S	

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
DE07.01.010.00	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	
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 μL。	不保护	μL	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	
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 T2260-2013 中华人民共和国行政区划代码

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	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	信息分析过程中过滤软件版本号。	不保护		S	
DE08.01.004.00	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	
DE08.01.005.00	比对软件名称	信息分析过程中比对软件名称。	不保护		S	
DE08.01.006.00	比对软件版本	信息分析过程中比对软件版本号。	不保护		S	
DE08.01.007.00	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	
DE08.01.008.00	新转录本预测软件名称	信息分析过程中新转录本预测软件名称。	不保护		S	
DE08.01.009.00	新转录本预测软件版本	信息分析过程中新转录本预测软件版本号。	不保护		S	
DE08.01.010.00	新转录本预测软件参数	信息分析过程中新转录本预测软件参数信息。	不保护		S	
DE08.01.011.00	新转录本注释软件名称	信息分析过程中新转录本注释软件名称。	不保护		S	
DE08.01.012.00	新转录本注释软件版本	信息分析过程中新转录本注释软件版本号。	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
DE08.01.013.00	新转录本注释软件参数	信息分析过程中新转录本注释软件参数信息。	不保护		S	
DE08.01.014.00	差异剪接基因检测软件名称	信息分析过程中差异剪接基因检测软件名称。	不保护		S	
DE08.01.015.00	差异剪接基因检测软件版本	信息分析过程中差异剪接基因检测软件版本号。	不保护		S	
DE08.01.016.00	差异剪接基因检测软件参数	信息分析过程中差异剪接基因检测软件参数信息。	不保护		S	
DE08.01.017.00	SNP&INDEL 检测软件名称	信息分析过程中 SNP&INDEL 检测软件名称。	不保护		S	
DE08.01.018.00	SNP&INDEL 检测软件版本	信息分析过程中 SNP&INDEL 检测软件版本号。	不保护		S	
DE08.01.019.00	SNP&INDEL 检测软件参数	信息分析过程中 SNP&INDEL 检测软件参数信息。	不保护		S	
DE08.01.020.00	基因融合检测软件名称	信息分析过程中基因融合检测软件名称。	不保护		S	
DE08.01.021.00	基因融合检测软件版本	信息分析过程中基因融合检测软件版本号。	不保护		S	
DE08.01.022.00	基因融合检测软件参数	信息分析过程中基因融合检测软件参数信息。	不保护		S	
DE08.01.023.00	基因定量软件名称	信息分析过程中基因定量软件名称。	不保护		S	
DE08.01.024.00	基因定量软件版本	信息分析过程中基因定量软件版本号。	不保护		S	
DE08.01.025.00	基因定量软件参数	信息分析过程中基因定量软件参数信息。	不保护		S	
DE08.01.026.00	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	
DE08.01.027.00	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	
DE08.01.028.00	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	
DE08.01.029.00	基因表达量聚类分析软件名称	信息分析过程中基因表达量聚类分析软件名称。	不保护		S	
DE08.01.030.00	基因表达量聚类分析软件版本	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	
DE08.01.031.00	基因表达量聚类分析软件参数	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	
DE08.01.032.00	共表达分析软件名称	信息分析过程中共表达分析软件名称。	不保护		S	
DE08.01.033.00	共表达分析软件版本	信息分析过程中共表达分析软件版本号。	不保护		S	
DE08.01.034.00	共表达分析软件参数	信息分析过程中共表达分析软件参数信息。	不保护		S	
DE08.01.035.00	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	

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
DE08.01.036.00	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	
DE08.01.037.00	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	
DE08.01.038.00	聚类分析软件名称	信息分析过程中聚类分析软件名称。	不保护		S	
DE08.01.039.00	聚类分析软件版本	信息分析过程中聚类分析软件版本号。	不保护		S	
DE08.01.040.00	聚类分析软件参数	信息分析过程中聚类分析软件参数信息。	不保护		S	

A.5　质控信息

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	ERCC 的下机序列比例	比对到 ERCC mRNA 的下机序列百分比。	不保护	%	N	
DE01.01.002.00	线粒体外显子下机序列比例	比对到线粒体外显子的下机序列百分比。	不保护	%	N	
DE01.01.003.00	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	
DE01.01.004.00	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	
DE01.01.006.00	内含子下机序列比例	比对到内含子的下机序列百分比。	不保护	%	N	
DE01.01.007.00	外显子下机序列比例	比对到外显子区域的下机序列百分比。	不保护	%	N	
DE01.01.008.00	唯一比对率	唯一比对的下机序列百分比。	不保护	%	N	
DE01.01.009.00	比对率	比对到参考基因组的下机序列百分比。	不保护	%	N	
DE01.01.010.00	过滤后下机序列数目	过滤后的总下机序列数目。	不保护		S	
DE09.01.011.00	项目名称	项目名称。	不保护		S	
DE09.01.012.00	个体编号	样本来源的个体编号。	不保护		S	
DE09.01.013.00	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	N	
DE09.01.014.00	样本编号	样本编号。	保护		S	
DE09.01.015.00	样本名称	分析结果中样本名称。	不保护		S	
DE09.01.016.00	样本类型	样本类型。	不保护		S	
DE09.01.017.00	样品浓度	样本的浓度值，计量单位为 ng/μL。	不保护	ng/μL	N	
DE09.01.017.00	样品总量	样本的总重量，计量单位为μL。	不保护	μL	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“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
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/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
24	BGI 公司 BGISEQ 系列	BGISEQ-500RS

表 B.1　测序仪名称代码表（续）

代码	测序仪系列	型号
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	