ICS 07.080
Q 8499

团　　体　　标　　准

T/SZAS 13—2019

基因组学数据集

Dataset of genomics

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。

基因组学数据集

1 范围

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

本标准适用于组学数据中有关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

GB/T 29859-2013 生物信息学术语

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

VCF格式 the variant call format

一种生物信息分析中的变异数据描述格式。

3.1.2

单核苷酸多态性 single nucleotide polymorphism;SNP

在基因组水平，由单个核苷酸位点的变异（替代、插入或缺失）所引起的脱氧核糖核苷酸序列多态性。

3.1.3

插入缺失型变异 insertion and deletion;Indel

在基因组的某个位置上所发生的小片段序列的插入或者缺失，插入或缺失片段的长度在50 bp以下。

3.1.4

结构性变异 structural variation

包括大片段缺失、大片段重复、倒位、易位。其中大片段缺失和大片段重复又叫拷贝数变异（Copy number variation），即连续较长的序列发生了缺失或者重复，与插入删除型变异的区别在于变异的长度。倒位指染色体上某一段序列发生了180度的颠倒。易位指染色体上的某一片段转移到了其他位置上。

3.1.5

1倍测序深度 1X

测序得到碱基总量与基因组大小比值为1。

3.1.6

4倍测序深度 4X

测序得到碱基总量与基因组大小比值为4。

3.1.7

20倍测序深度 20X

测序得到碱基总量与基因组大小比值为20。

3.1.8

平均测序深度

测序得到碱基总量与基因组大小比值。

3.1.9

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。

3.1.10

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。

3.1.11

唯一下机序列 uniq reads

测序得到的唯一的下机序列。

3.1.12

Q20

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。

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%以上的碱基质量值大于20。

3.1.13

Q30

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。

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%以上的碱基质量值大于30。

3.1.14

平均读长 read length

测序的下机数据里，所有序列的平均长度。以碱基（bp）为单位，常见的读长有50 bp、90 bp、100 bp、150 bp。

3.2 缩略语

SNV：单核苷酸位点变异(Single Nucleotide Variants)
SNP：单核苷酸多态性(Single Nucleotide Polymorphism)
SV：基因组结构变异（Structural Variants）
INDEL：插入/缺失（Insertions/Deletions）
CNV：拷贝数变异（Copy Number Variants）
MD5：信息摘要算法（MD5 Message-Digest Algorithm）
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	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 基因组学数据元目录专用属性包括建库测序信息、生物信息分析、质控信息三部分。

4.2.2 建库测序信息描述建库和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、测序完成时间等。

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、参考序列、比对序列数、过滤软件名称、过滤软件版本、过滤软件参数等。

4.2.4 质控信息描述整个测序过程质量监控的数据元，例如项目标识符、项目名称、子项目名称、子项目标识符、样本管标识符等。

4.2.5　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 建库测序信息

建库测序信息如表A.1所示。

表A.1 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	测序平台	测序平台和仪器型号。	不保护		S	B.2 测序平台代码表
DE07.01.010.00	测序技术描述	所采用的测序技术的描述。	不保护		S	
DE07.01.011.00	分析类型	测序的分析类型说明。	不保护		S	
DE07.01.012.00	物种	物种的名称。	不保护		S	
DE07.01.013.00	芯片号	芯片号编码。	不保护		S	
DE07.01.014.00	测序通道号	测序通道号。	不保护		S	
DE07.01.015.00	机器号	机器号。	不保护		S	
DE07.01.016.00	FASTQ 格式文件唯一编号	FASTQ 格式文件唯一编号。	不保护		S	
DE07.01.017.00	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
DE07.01.018.00	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	

表 A.1　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
DE07.01.019.00	组装时间	组装当日的公元纪年日期和时间的完整描述。	不保护		DT	
DE07.01.020.00	组装版本	组装的版本号。	不保护		S	
DE07.01.021.00	组装方法描述	组装所使用的程序或方法的描述。	不保护		S	
DE07.01.022.00	组装名称	组装名(例如：GRCh37.p5)。	不保护		S	
DE07.01.023.00	程序/算法	测序所用的程序/算法。	不保护		S	
DE07.01.024.00	组装方法名称	组装所使用的程序或方法。	不保护		S	B.3 组装方法代码表
DE07.01.029.00	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
DE07.01.030.00	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
DE07.01.031.00	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
DE07.01.032.00	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	
DE07.01.033.00	文库设置	文库设置说明。	不保护		S	B.7 文库设置代码表
DE07.01.034.00	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	
DE07.01.035.00	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
DE07.01.036.00	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.9 文库构建策略代码表
DE07.01.037.00	文库名称	文库名称。	不保护		S	
DE07.01.038.00	文库类型	文库类型说明。	不保护		S	
DE07.01.039.00	文库数量	文库数量。	不保护		N	
DE07.01.040.00	文库标识符	测序文库标识符。	不保护		S	

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	参考序列	信息分析过程中所使用参考序列详细信息。	不保护		S	
DE08.01.003.00	比对序列数	比对序列数，是指在种水平上比对上该物种的序列数。	不保护		N	

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
DE08.01.004.00	过滤软件名称	数据质量控制过程，低质量下机序列，接头相关的下机序列和 N 下机序列的过滤以及数据统计软件名称。	不保护		S	
DE08.01.005.00	过滤软件版本	数据质量控制过程，低质量下机序列，接头相关的下机序列和 N 下机序列的过滤以及数据统计软件版本详细信息。	不保护		S	
DE08.01.009.00	过滤软件参数	数据质量控制过程，低质量下机序列，接头相关的下机序列和 N 下机序列的过滤以及数据统计软件参数详细信息。	不保护		S	
DE08.01.010.00	序列比对软件名称	信息分析过程中所使用序列比对软件名称。	不保护		S	
DE08.01.011.00	序列比对软件版本	信息分析过程中所使用序列比对软件版本详细信息。	不保护		S	
DE08.01.012.00	序列比对软件参数	信息分析过程中所使用序列比对软件参数详细信息。	不保护		S	
DE08.01.013.00	单核苷酸位点变异（SNV）检测软件名称	信息分析过程中所使用体细胞 SNV 检测软件名称。	不保护		S	B.10 SNV 检测软件代码表
DE08.01.014.00	SNV 检测软件版本	信息分析过程中所使用体细胞 SNV 检测软件版本详细信息。	不保护		S	
DE08.01.015.00	SNV 检测软件参数	信息分析过程中所使用体细胞 SNV 检测软件参数详细信息。	不保护		S	
DE08.01.016.00	插入/缺失（INDEL）检测软件名称	信息分析过程中所使用 INDEL 检测软件名称。	不保护		S	
DE08.01.017.00	INDEL 检测软件版本	信息分析过程中所使用 INDEL 检测软件版本详细信息。	不保护		S	
DE08.01.018.00	INDEL 检测软件参数	信息分析过程中所使用 INDEL 检测软件参数详细信息。	不保护		S	
DE08.01.019.00	变异注释软件名称	信息分析过程中所使用变异注释软件名称。	不保护		S	B.12 变异注释软件代码表
DE08.01.020.00	变异注释软件版本	信息分析过程中所使用变异注释软件版本详细信息。	不保护		S	
DE08.01.021.00	变异注释软件参数	信息分析过程中所使用变异注释软件参数详细信息。	不保护		S	
DE08.01.022.00	拷贝数变异（CNV）检测软件名称	信息分析过程中所使用 CNV 检测软件名称。	不保护		S	B.13 CNV 检测软件代码表
DE08.01.023.00	CNV 检测软件版本	信息分析过程中所使用 CNV 检测软件版本详细信息。	不保护		S	
DE08.01.024.00	CNV 检测软件参数	信息分析过程中所使用 CNV 检测软件参数详细信息。	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
DE08.01.025.00	基因组结构变异（SV）检测软件名称	信息分析过程中所使用 SV 检测软件名称。	不保护		S	B.11 SV 检测软件代码表
DE08.01.026.00	SV 检测软件版本	信息分析过程中所使用 SV 检测软件版本详细信息。	不保护		S	
DE08.01.027.00	SV 检测软件参数	信息分析过程中所使用 SV 检测软件参数详细信息。	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	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	
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	

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
DE01.01.019.00	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	
DE01.01.020.00	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	
DE01.01.021.00	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	
DE01.01.022.00	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	
DE01.01.024.00	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	
DE01.01.025.00	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	
DE01.01.026.00	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	
DE01.01.027.00	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	
DE01.01.028.00	下机序列 1 平均 AT%分离	下机序列 1 平均 AT 不等率。	不保护	%	N	
DE01.01.029.00	下机序列 2 平均 AT%分离	下机序列 2 平均 AT 不等率。	不保护	%	N	
DE01.01.030.00	下机序列 1 平均 GC%分离	下机序列 1 平均 GC 不等率。	不保护	%	N	
DE01.01.031.00	下机序列 2 平均 GC%分离	下机序列 2 平均 GC 不等率。	不保护	%	N	
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	
DE07.01.034.00	过滤后数据量	过滤后数据量。	不保护	bp	N	
DE01.01.035.00	单核苷酸多态性（SNP）数量	SNP 数量。	不保护	M	N	
DE01.01.036.00	转换/颠换比	转换/颠换比。	不保护		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	

B.2　测序平台代码

测序平台代码规定了测序平台的代码。

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　测序平台代码表

代码	测序平台
1	LS454
101	454 GS
102	454 GS 20
103	454 GS FLX
104	454 GS FLX Titanium
105	454 GS FLX+
106	454 GS Junior
2	ILLUMINA
201	HiSeq X Five
202	HiSeq X Ten
203	Illumina Genome Analyzer
204	Illumina Genome Analyzer II
205	Illumina Genome Analyzer IIx
206	Illumina HiScanSQ
207	Illumina HiSeq 1000
208	Illumina HiSeq 1500
209	Illumina HiSeq 2000
210	Illumina HiSeq 2500
211	Illumina HiSeq 3000
212	Illumina HiSeq 4000
213	Illumina MiSeq
	Illumina MiSeq Dx

表 B.2　测序平台代码表（续）

代码	测序平台
	Illumina MiSeq FGx
214	NextSeq 500
215	NextSeq 550
216	Illumina iSeq 100
217	Illumina NovaSeq 5000
218	Illumina NovaSeq 6000
219	Illumina NovaSeq TM
3	ABI_SOLID
301	AB 5500 Genetic Analyzer
302	AB 5500xl Genetic Analyzer
303	AB 5500xl-W Genetic Analysis System
304	AB SOLiD 3 Plus System
305	AB SOLiD 4 System
306	AB SOLiD 4hq System
307	AB SOLiD PI System
308	AB SOLiD System
309	AB SOLiD System 2.0
310	AB SOLiD System 3.0
4	COMPLETE_GENOMICS
401	Complete Genomics
5	PACBIO_SMRT
501	PacBio RS
502	PacBio RS II
503	Sequel
6	ION_TORRENT
601	Ion Torrent PGM
602	Ion Torrent Proton
603	Ion Torrent S5 XL
604	Ion Torrent S5
7	CAPILLARY
701	AB 3130 Genetic Analyzer
702	AB 310 Genetic Analyzer
703	AB 3130xL Genetic Analyzer
704	AB 3500 Genetic Analyzer
705	AB 3500xL Genetic Analyzer
706	AB 3730 Genetic Analyzer
707	AB 3730xL Genetic Analyzer
8	OXFORD_NANOPORE
801	GridION
802	MinION

表 B.2　测序平台代码表（续）

代码	测序平台
803	PromethION
9	BGISEQ
901	BGISEQ-50
902	BGISEQ-100
903	BGISEQ-500
904	BGISEQ-1000
905	MGISEQ-200
906	MGISEQ-2000
907	DNBSEQ-T7
908	BGISEQ-500RS
909	BGISEQ-500CX
910	MGISEQ-200RS
911	MGISEQ-2000RS
912	MGISEQ-200CX
913	MGISEQ-2000CX
914	DNBSEQ-G50
915	DNBSEQ-G400
916	DNBSEQ-E
10	Berry Genomics
1001	NextSeq CN500
11	CapitalBio Company
1101	BioelectronSeq 4000
12	Bionano Genomics
1201	BioNano IRYS
1202	BioNano SAPHYR
13	DAAN GENE
1301	DA8600
14	Helicos BioSciences Corporation
1401	Helicos HeliScope
15	HYK Genetic
1501	HYK-PSTAR-IIA
16	PacBio SMRT
1601	PacBio RS
1602	PacBio RS II
1603	PacBio Sequel

B.3　组装方法代码

组装方法代码规定了组装方法的代码。

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　组装方法代码表

代码	组装方法
1	ABySS
101	Newbler
102	SOAPdenovo
103	SPAdes
104	Velvet
105	phredPhrap
106	other
2	AllPaths
3	Arachne
4	CLC NGS Cell
5	Celera Assembler
6	GS De Novo Assembler
7	JAZZ
8	MIRA
9	MaSuRCA

B.4　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.4。

表B.4　分子类型代码表

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

B.5　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.5。

表B.5　是否代码表

代码	是否
1	是
2	否

B.6　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.6。

表B.6　文件类型代码表

代码	文件类型	备注
1	CRAM	
2	BAM	合并比对和测序数据的二进制 SAM 格式文件
3	SFF	
4	FASTQ	
5	PacBio_HDF5	PacBio hdf5 格式文件
6	CompleteGenomics_native	
7	OxfordNanopore_native	

B.7　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.7。

表B.7　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.8　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.8。

表B.8　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	

B.9　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.9。

表B.9　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
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

表 B.9　文库构建策略代码表（续）

代码	文库构建策略	备注
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.10　SNV检测软件代码

SNV检测软件代码规定了SNV检测软件的代码。

采用2位数字顺序代码，从“01”开始编码，按升序排列，见表B.10。

表B.10　SNV 检测软件代码表

代码	SNV 检测软件
01	Atlas2
02	2kplus2
03	Bambino
04	Bassovac
05	BAYSIC
06	BAYSIC
07	BEAGLE
08	Cake
09	Churchill
10	CLImAT
11	ComB
12	CoNAn-SNV
13	CopySeq
14	Cortex
15	COSMIC
16	CRISP
17	DeNovoGear
18	discoSnp++
19	EBCall
20	FamSeq
21	FaSD-somatic
22	FreeBayes
23	GAMES
24	GATK
25	glfMultiples
26	glfSingle
27	Halvade
28	HapMuC
29	Illuminator
30	IMPUTE2

表 B.10　SNV 检测软件代码表（续）

代码	SNV 检测软件
31	Indelocator
32	IVC
33	JointSNVMix
34	KGGSeq
35	KvarQ
36	LoFreq
37	MACH
38	MAQ
39	marginAlign
40	MendelScan
41	MoDIL
42	MSIsensor
43	MSIsensor
44	multiSNV
45	mutationSeq
46	MuTect
47	Platypus
48	Polymutt
49	PyroHMMsnp
50	PyroHMMvar
51	qSNP
52	QuadGT
53	QuadGT
54	QualitySNPng
55	RADIA
56	RAREVATOR
57	RAREVATOR
58	realSFS
59	ReviSTER
60	RVD
61	RVD2
62	SAMtools
63	SAMtools
64	SeqEM
65	SeqHBase
66	Shimmer
67	Slider
68	Sniper
69	Snippy
70	SNPest

表 B.10　SNV 检测软件代码表（续）

代码	SNV 检测软件
71	SNPSVM
72	SNPTools
73	SNVer
74	SNVMix
75	SNVMix
76	SNV-PPILP
77	SOAPgaea
78	SOAPgaea
79	SOAPsnp
80	SOAPsnv
81	SolSNP
82	SomaticCall
83	SomaticSniper
84	SPLINTER
85	Strelka
86	Syzygy
87	TrioCaller
88	UnifiedGenotyper
89	VAAL
90	VariantMaster
91	VariantMaster
92	VARiD
93	VarScan
94	vipR
95	Virmid
99	Other

B.11　SV检测软件代码

SV检测软件代码规定了SV检测软件的代码。

采用2位数字顺序代码，从“01”开始编码，按升序排列，见表B.11。

表B.11　SV 检测软件代码表

代码	SV 检测软件
01	AGE
02	APOLLOH
03	BreakDancer
04	Breakpointer
05	BreakSeq
06	Breakway

表 B.11　SV 检测软件代码表（续）

代码	SV 检测软件
07	CLEVER Toolkit
08	clipcrop
09	Clippers
10	Cloudbreak
11	CREST
12	DELLY
13	deStruct
14	detecttd
15	forestSV
16	FusionMap
17	GASV
18	GASVPro
19	Hydra
20	inGAP
21	nFuse
22	PEMer package
23	Pindel
24	SLOPE
25	SOAPsv
26	Socrates
27	SoftSearch
28	SPLITREAD
29	SVDetect
30	SVMerge
31	SVMiner
32	SVseq
33	TEMP
34	VariationHunter
35	ANISE and BASIL
36	BreaKmer
37	cortex
38	Gustaf
39	IMR-DENOM
40	inGAP-sv
41	Manta
42	MATE-CLEVER
43	Meerkat
44	MetaSV
45	MindTheGap
46	NovelSeq

表 B.11　SV 检测软件代码表（续）

代码	SV 检测软件
47	Platypus
48	PopIns
49	PRISM
50	RAPTR-SV
51	Reprever
52	ViVar
99	Other

B.12　变异注释软件代码

变异注释软件代码规定了变异注释软件的代码。

采用2位数字顺序代码，从“01”开始编码，按升序排列，见表B.12。

表B.12　变异注释软件代码表

代码	变异注释软件
01	ANNOVAR
02	AnnTools
03	ASOoViR
04	AVIA
05	BioR
06	CADD
07	CandiSNPer
08	CHAoS
09	CNVannotator
10	COVA
11	dbNSFP
12	FamAnn
13	FunctSNP
14	GEMINI
15	GeneTalk
16	GERP
17	GESND
18	LS-SNP/PDB
19	NGS-SNP
20	Oncotator
21	pfSNP
22	SCAN
23	SeqAnt
24	SiPhy
25	SNPAAMapper

表 B.12　变异注释软件代码表（续）

代码	变异注释软件
26	SNPdat
27	SNPdbe
28	SnpEff
29	SNPeffect
30	SNPMeta
31	SNPnexus
32	SNPper
33	topoSNP
34	TRAMS
35	TREAT
36	VAAST 2
37	VAGrENT
38	VARIANT
39	VariantAnnotation
40	VAT
41	wANNOVAR
42	Alamut Batch
43	CanvasDB
44	DANN
45	PHAST
46	seqminer 3.7
47	SeqWare Query Engine
48	SNiPlay
49	SNP Function Portal
50	SVA
51	VarioWatch
52	WhopGenome
99	Other

B.13　CNV检测软件代码

CNV检测软件代码规定了CNV检测软件的代码。

采用2位数字顺序代码，从“01”开始编码，按升序排列，见表B.13。

表B.13　CNV 检测软件代码表

代码	CNV 检测软件
01	AbsCN-seq
02	BIC-seq
03	cn.mops
04	CNAnorm

表 B.13　CNV 检测软件代码表（续）

代码	CNV 检测软件
05	CNAseg
06	CnD
07	CNValidator
08	CNVer
09	cnvHiTSeq
10	CNVnator
11	CNVrd2
12	CNV-seq
13	Control-FREEC
14	CopySeq
15	GENSENG
16	HMMcopy
17	JointSLM
18	Magnolya
19	MATCHCLIP
20	m-HMM
21	mrCaNaVaR
22	OncoSNP-SEQ
23	Patchwork
24	RDXplorer
25	readDepth
26	rSW-seq
27	SegSeq
28	WaveCNV
29	WISECONDOR
30	AGE
31	AS-GENSENG
32	BreakDancer
33	CopyCat
34	Cortex
35	GASV
36	modSaRa
37	PEMer
38	Pindel
39	QDNAseq
40	SLOPE
41	TIGRA
42	VariationHunter
99	Other