ICS 07.080
A 40

中华人民共和国国家标准

GB/T 35890—2018

高通量测序数据序列格式规范

Technical specification of high throughput sequencing data format

2018-02-06 发布　　2018-09-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

本标准由全国生化检测标准化技术委员会(SAC/TC 387)提出并归口。

本标准起草单位:深圳华大基因研究院、中国计量科学研究院。

本标准主要起草人:梁鑫明、刘心、蒋慧、杜佳婷、谢强、李倩一、李岱怡、王晶。

高通量测序数据序列格式规范

1 范围

本标准规定了高通量测序数据的序列格式，包括序列描述格式规范和高通量测序数据整体格式规范。

本标准适用于规范生物体 DNA 高通量测序数据序列格式。

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 30989 高通量基因测序技术规程

ISO/IEC 646 信息技术 ISO 信息交换七位编码字集(Information technology—ISO 7-bit coded character set for information interchange)

3 术语和定义

下列术语和定义适用于本文件。

3.1

高通量测序 high-throughput sequencing

以一次并行几十万到几百万条核酸分子序列测定和一般读长较短等为标志，适用于 DNA 的测序技术。

注：改写 GB/T 30989—2014，定义 3.1.9。

3.2

测序片段 reads

高通量测序平台产生的含有碱基序列和质量值的序列片段。

3.3

双末端测序 paired-end sequencing

对 DNA 模板链和互补链分别测序，并得到两条链成对测序片段的测序技术。

3.4

插入片段长度 insert size

双末端测序中，从模板链测序的测序片段左端到互补链测序的测序片段右端的距离。

3.5

测序片段识别码 reads identifier

用以识别一段测序片段的具有唯一性的字符串。

3.6

碱基序列 base sequence

测序片段中记录碱基排列的字符串，碱基序列中的每个碱基应使用大写字母(A、T、C、G 和 N)或小写字母(a、t、c、g 和 n)，其中字母 A 和 a 表示腺嘌呤，字母 T 和 t 表示胸腺嘧啶，字母 C 和 c 表示胞嘧

啶，字母 G 和 g 表示鸟嘌呤，字母 N 和 n 表示未测定的碱基。

3.7

美国标准信息交换代码　American standard code for information interchange；ASCII

基于拉丁字母的一套电脑编码系统，主要用于显示现代英语和其他西欧语言，并等同于国际标准 ISO/IEC 646。

3.8

质量值体系　quality score system

测序碱基质量一个特定的范围，常见的质量值体系有 Phred＋33 和 Phred＋64 两种，Phred＋33 体系质量值 0 对应 ASCII 码 33，用！表示，Phred＋64 体系质量值 0 对应 ASCII 码 64，用@表示。

3.9

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。

3.10

SAM/BAM 格式　SAM/BAM format

SAM 是基于文本的、存储核酸序列和其测序质量信息的、以每一行表示一条序列、每行以制表符分割成 11 列的标准格式，测序质量信息使用 ASCII 字符表示，BAM 是 SAM 格式的二进制格式。

注：SAM 和 BAM 也可作为序列比对格式。

3.11

参考序列　reference sequence

测序片段对应的物种基因组序列。

4　缩略语

下列缩略语适用于本文件。

bp：碱基对（base pair）

DNA：脱氧核糖核酸（deoxyribonucleic acid）

ID：识别码（identifier）

MAPQ：比对质量（mapping quality）

POS：比对起始位点（position）

QNAME：查询序列名称/测序片段名称（query name）

RNAME：参考序列名称（reference name）

5　序列描述规范

5.1　测序片段 ID

测序片段 ID 应保证一个序列编号对应一段测序片段，具有唯一性。对于双末端测序序列，ID 中应包含标明模板链或互补链的标识。

5.2　碱基序列

碱基序列应使用大写字母（A～Z）或者小写字母（a～z）来表示，自成一行（FASTQ 格式）或一列（SAM/BAM 格式）。

6 高通量测序数据整体格式规范

6.1 FASTQ 格式

每一条测序序列用以下 4 行信息表示：

a) 首行以字符@开头，后面为测序片段 ID，字符@与测序片段 ID 之间不应有空格，格式规范与 5.1 小节描述一致；

b) 第二行为测序的碱基序列信息，不应换行；

c) 第三行以加号(+)开头，后面内容与首行一样，为序列 ID，序列 ID 可省略；

d) 第四行为第二行的碱基序列对应的测序质量值，不应换行。测序质量值应用 ASCII 码表示，且质量值体系与 ASCII 码对照表应符合附录 A 的规定。

6.2 SAM/BAM 格式

6.2.1 基本结构

SAM/BAM 格式分为头文件和比对结果两部分。

6.2.2 头文件

头文件每行应以字符@开头，后面为 HD，SQ，RG，PG 和 CO 标签信息，每行标签与子标签应用制表符间隔，头文件标签符合附录 B 的规定。头文件标签格式规范如下：

——HD 标签应存在；

——当测序片段比对上参考序列时，SQ 标签应存在；

——当 RG 出现在比对结果任意一行时，其对应编号应出现 RG 标签中，该 RG 标签自成一行；

——当 PG 出现在比对结果任意一行时，其对应编号应出现 PG 标签中，该 PG 标签自成一行。

6.2.3 比对结果

比对结果每行的信息应用制表符间隔，分为 11 列必须字段和 1 列可选字段，每个字段描述如下：

a) 测序片段名称 QNAME，格式规范与 5.1 小节描述一致；

b) 比对情况标记，具体规范符合附录 C 的规定；

c) 参考序列名称 RNAME，如果测序片段未必对上任何参考序列，RNAME 应用星号(*)表示；

d) POS，测序片段比对到参考序列的最左起始坐标，最小值为 1。如果测序片段未比对上任何参考序列，起始坐标应记为 0；

e) 比对质量 MAPQ，如果测序片段未比对上任何参考序列，MAPQ 应记为 255；

f) CIGAR 字符串，记录插入，删除，错配以及剪切拼接等信息；

g) 对于双末端测序，测序片段互补链比对到参考序列的编号，等号(=)表示与模板链与互补链比对到的参考序列编号相同；

h) 互补链比对到参考序列的最左起始坐标；

i) 推测的插入片段长度；

j) 测序片段碱基序列；

k) 测序片段碱基序列对应的质量值序列；

l) 可选字段，格式如：标签：类型：数值，其中标签由两个字符组成，首字符为大写字母(A～Z)、小写字母(a～z)的任意组合，第二个字符为大写字母、小写字母和数字(0～9)的任意组合，每个标签代表一类信息，每行一个标签只能出现一次；类型表示标签对应值的类型，可以是字符串、

整数、字节、数组等。

7 高通量测序数据文件格式样例

高通量测序数据文件格式样例参见附录 D。

附 录 A
（规范性附录）
常见质量值体系 ASCII 码对照关系表

ASCII 码字符范围如下：

!"# $ %&'() * +,-./0123456789:;<=>? @ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

常见质量值体系与 ASCII 码对照关系表见表 A.1。

表 A.1 常见质量值体系与 ASCII 码对照关系表

质量值体系	ASCII 字符范围	质量值范围
Phred+33	! ～I 或! ～J	0～40 或 0～41
Phred+64	@～h 或 B～h	0～40 或 3～40[1]
注：质量值 0 和 1 未使用，质量值 2 用作 Read 质量控制。		

附 录 B
（规范性附录）
SAM/BAM 格式头文件标签描述

SAM/BAM 格式头文件标签描述见表 B.1。

表 B.1 SAM/BAM 格式头文件标签描述

标签	子标签	描述
HD	VN	格式版本。可接受格式为：数字(0～9)加点号(.)加数字(0～9)
	SO	比对信息排序顺序。合法值：unknown(未知，默认值)、unsorted(未排序)、queryname(按测序片段名称排序)、coordinate(按比对起始坐标排序)。coordinate 的排序方式，应以参考序列编号为主要关键字，按照@SQ 定义的顺序排序，次要排序关键字应以比对起始坐标信息。对于参考序列信息、比对坐标信息都相同的比对记录，顺序随机。所有参考序列信息为“*”的比对记录应排在参考序列信息不为“*”的比对记录之后，并且顺序随机
	GO	比对信息组别，表示相似的比对结果组合在一起但文件不一定整体排序。合法值：none(默认值)、query(比对结果根据测序片段编号组合)、reference(比对结果根据 RNAME/POS 组合)
SQ	SN*	参考序列名称。每一行@SQ 应有唯一的 SN 标签，用于比对记录的测序片段编号、双末端测序的第 2 个片段比对上的参考序列名称
	LN*	参考序列长度，范围从 1 到 $2^{31}-1$
	AS	基因组组装标志
	M5	参考序列大写形式 MD5 校验值
	SP	物种
	UR	参考序列链接。该标签以一种标准协议开头，如：http:或者 ftp:。如果不以标准协议开头，则认为是一个文件系统路径
RG	ID	测序片段组标志。每一行@RG 应有唯一 ID(在头文件部分所有@RG 中)，用于比对记录 RG 标签。为了处理冲突序列组标志在合并 SAM 文件时可能会被修改
	CN	提供序列的测序中心名称
	DS	描述信息
	DT	测序运行日期，格式为 ISO 8601 日期或日期/时间
	FO	流程顺序
	KS	关键序列
	LB	文库
	PG	用于处理测序片段分组的程序
	PI	预测插入片段长度的中位数
	PL	测序平台/技术。合法值：CAPILLARY，LS454，ILLUMINA，SOLID，HELICOS，ION-TORRENT，ONT 和 PACBIO
	PM	平台模型，其他关于测序平台/技术的信息
	PU	平台装置，唯一标志符
	SM	样品。如果进行混合样品测序则应使用混合样品名称

表 B.1（续）

标签	子标签	描述
PG	ID	程序记录标志。每一行@PG 必须拥有唯一 ID，用于比对记录的 PG 标签或其他@PG 的 PP 标签。为了处理冲突@PG ID 在合并 SAM 文件时可能会被修改
	PN	程序名称
	CL	命令行
	PP	前置@PG-ID。必须与另一个@PG 的 ID 一致，@PG 可以被 PP 标签提前声明。为了处理 PG ID 冲突，PP 在合并 SAM 文件时可能会被修改。第一个 PG(如)描述最近处理 SAM 记录的程序，下一个 PG 描述下一个最近处理 SAM 记录的程序。一条 SAM 记录的 PG ID 不必要涉及最新的 PG 记录，可以涉及一系列 PG 记录中的任意一个，意味着这条 SAM 记录已被该 PG 中的程序以及 PP 标签中涉及的程序处理
	DS	解释说明
	VN	程序版本
CO	—	评论信息，允许多行无序

附 录 C
（规范性附录）
SAM/BAM 格式比对标记描述

SAM/BAM 格式比对标记描述见表 C.1。

表 C.1 SAM/BAM 格式比对标记描述

标记	描述
1	模板链包含两个测序片段
2	双末端测序的两个片段正确地比对上参序列(即测序片段均比对上参考序列同一条染色体)
4	测序片段没有比对上参考序列
8	双末端测序的第二个片段没有比对上参考序列
16	双末端测序的第一个片段的反向互补链
32	双末端测序的第二个片段的反向互补链
64	双末端测序的第一个片段比对上参考序列
128	双末端测序的第二个片段比对上参考序列
256	测序片段的比对位置不是最优选择
512	测序片段未通过质量控制
1024	测序片段是 PCR 或者光学重复
2048	测序片段部分序列比对上参考序列

附 录 D
（资料性附录）
高通量测序数据文件格式样例

D.1 FASTQ 格式样例

@A81C7HABXX:5:1:1429:2133#CNNNNNNN/1
TAAAGACAGCATCCTACTGGATTAGGGGTGGGCCCTAAATCCAATGACTC
+
ggggggggggegggggggegggggfgggcgggggggggggfgggegeggag
...
@A81C7HABXX:5:1:1589:1985#CNNNNNNN/1
ACAGCATCGGGTGGGCCCAATGACTACTAAATCAAGTCCTACTGGATTAG
+
ggggggggggfgggcggggegggggggggegggegeggaggggggggggfgg

D.2 BAM/SAM 文件格式样例

@HD VN:1.4 GO:none SO:coordinate
@SQ SN:chr1 LN:203413412
@RG ID:CL10000843 PL:COMPLETE PU:CL10000843 LB:WGS_PE100 SM:WGS_PE100 CN:BGI
@PG ID:bwa PN:bwa VN:0.7.10-r789
W52J0JMXLBA:5:1:1021:1987 163 chr1 10001 10 1S2M1D3M2S = 10029 129 CTAACCCT DEDEDDDE BD:Z:NNNNNNNN MD:Z:13^A14 ME:i:10129 RG:Z:CL10000843 XG:i:1

中 华 人 民 共 和 国
国 家 标 准
高通量测序数据序列格式规范
GB/T 35890—2018
*
中国标准出版社出版发行
北京市朝阳区和平里西街甲2号(100029)
北京市西城区三里河北街16号(100045)
网址 www.spc.net.cn
总编室:(010)68533533 发行中心:(010)51780238
读者服务部:(010)68523946
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销
*
开本 880×1230 1/16 印张 1 字数 20 千字
2018年2月第一版 2018年2月第一次印刷
*
书号:155066·1-59381 定价 18.00 元

GB/T 35890-2018

打印日期:2018年3月12日 F009A