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