ICS 11.100.10
CCS C44

中华人民共和国国家标准

GB/T XXXXX—XXXX

人全基因组高通量测序数据质量评价方法

The Data Quality Evaluation Method of Human Whole Genome Sequencing

征求意见稿

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
国家标准化管理委员会　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。

本文件由国家药品监督管理局提出。

本文件由全国医用临床检验实验室和体外诊断系统标准化技术委员会（SAC/TC 136）归口。

本文件起草单位：

本文件主要起草人：

人全基因组高通量测序数据质量评价方法

1 范围

本文件规定了人全基因组高通量测序数据质量评价方法涉及的术语和定义、质量要求、评价方法。

本文件适用于使用高通量基因测序技术对人类基因组DNA样本进行全基因组测序的数据质量评价。

本文件不适用于Sanger测序技术和单分子测序技术，不适用于人体样本的从头测序、单体型测序、人体肿瘤组织样本测序，也不适用于人类样本中含有的动物、植物、病毒、细菌、寄生虫等测序。

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 29859-2013 生物信息学术语

GB/T 30989-2014 高通量基因测序技术规程

GB/T 35537-2017 高通量基因测序结果评价要求

GB/T 35890-2018 高通量测序数据系列格式规范

YY/T 1723-2020 高通量基因测序仪

GA/T 1693-2020 法庭科学 DNA二代测序检验规范

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 or massively parallel sequencing

能一次并行几十万到几十亿条核酸分子序列测定技术，又称大规模平行测序，特点是测序读长较短。

[GB/T 35890-2018,3.1,有修改]

3.2

人全基因组高通量测序 human whole genome sequencing

对人类不同个体或群体进行全基因组测序。

注：包括人的23对染色体核酸序列、线粒体核酸序列。

3.3

接头 adapter

用于标记和固定待测序片段的已知序列的DNA片段。

[GA/T 1693-2020,3.9,有修改]

3.4

PCR-free文库 PCR-free library

不依赖PCR扩增过程而直接构建的测序上机文库。

注：聚合酶链式反应 polymerase chain reaction（PCR）是一种体外酶反应技术，通过DNA聚合酶反应将特定DNA片段进行指数扩增，以使其拷贝数增加几个数量级。

3.5

PCR文库 PCR library

区别于3.4，经过一定循环数的PCR扩增过程而构建的测序上机文库。

3.6

标签或条码 barcode or index

一段特征性的脱氧核苷酸段片段，在多样本混合测序时，充当识别特定样本来源的唯一编号。

[GA/T 1693-2020,3.10,有修改]

3.7

碱基识别质量百分比 percentage of base call quality

碱基识别质量在规定阈值以上的测序碱基个数占测序碱基总数的百分比，通常以Q20、Q30等表示。

3.8

GC含量 GC content

测序片段碱基中鸟嘌呤和胞嘧啶的加和数量占所有嘌呤碱基（腺嘌呤和鸟嘌呤）和嘧啶碱基（胸腺嘧啶和胞嘧啶）总数量的百分比。

3.9

数据过滤 data filtering

对原始测序片段进行去除低质量、N碱基、接头污染及其他任何不符合下游分析要求的测序片段的处理过程。

3.10

测序原始碱基数据量 sequencing raw base

测序后未经数据过滤的碱基总数，简称原始数据量。

3.11

人参考基因组序列 human reference genome sequence

公开发布供参比的人全基因组序列，如hs37d5、hg38、hg19等。

3.12

基因组比对率 mapping reads rate

比对到人参考基因组序列的测序片段占总体有效测序片段的百分比。

3.13

有效测序深度 effective sequencing depth

经过数据过滤、序列比对、去重后获得的一个全基因组测序样本的平均测序深度。

注：计算方法为该样本测序得到能够比对上基因组非N区域的碱基总量（bp）与基因组去除N区后的大小（Genome）的比值。计量单位为×，即乘数。

3.14

碱基错配比率 base mismatch rate

与参考基因组序列不一致的碱基总数除以比对上参考基因组序列的碱基总数的百分比。

3.15

重复测序片段比率 duplication rate

比对到参考序列的位置、方向及碱基序列均一致的测序片段总数除以比对上参考序列的测序片段总数的百分比。

3.16

20X 测序覆盖率 coverage rate of sequencing at least 20x

测序片段与参考基因组比对后，参考基因组上至少被测序片段覆盖20次的非N碱基数目占非N碱基总数的百分比。

3.17

插入缺失型突变 insertion and deletion, Indel

基因组DNA中核苷酸插入/缺失片段长度小于等于50bp的基因突变。

3.18

结构性变异 structural variation, SV

长度大于50bp的大片段缺失、插入、重复、倒位、易位。

3.19

标准变异数据集 Standard variation dataset

基于人基因组标准品构建的高置信标准变异数据集。

3.20

准确率 precision

检测到的真正基因变异位点个数占检测到的全部基因变异位点个数的百分比。

注：计算公式见公式（1）

$$precision=TP/(TP+FP)\times 100\% \quad \cdots\cdots (1)$$

式中：

TP——真阳 ，指与标准变异数据集基因变异位点一致的检测基因变异位点个数。

FP——假阳 ，指与标准变异数据集基因变异位点不一致的检测基因变异位点个数。

3.21

灵敏度 sensitivity

检测到的真正基因变异位点个数占待检测的所有基因变异位点个数的百分比。

注：计算公式见公式（2）

$$sensitivity = TP/(TP+FN) \times 100\% \cdots\cdots (2)$$

式中：

FN——假阴 ，指标准变异数据集中包含、但未被检出的基因变异位点个数。

4 质量要求

4.1 样本质量要求

4.1.1 样本类型

测序样本类型为人基因组DNA。DNA的来源主要为人全血、血液白膜层、唾液、羊水、口腔拭子、正常组织和正常细胞系。

本标准采用正常人细胞系提取的人基因组DNA标准品做为质量评价和统一量化评估方法指标的标准参考品。

注：本文件所涉及的人基因组标准品信息见附录A。

4.1.2 DNA样本质量

4.1.2.1 DNA样本的完整度

基因组DNA完整，没有明显降解。通常使用1-2%琼脂糖电泳，基因组条带集中在23kb及以上，无明显拖尾或弥散。

4.1.2.2 DNA样本的纯度

DNA样本肉眼可见的透明澄清，无明显粘稠状，无明显色素或杂质。DNA样本中无RNA、蛋白残留。通常使用紫外分光光度计测定，以OD值（A260/280、A260/230）来表示。

4.1.2.3 DNA样本的体积、浓度和总量

DNA样本的体积、浓度和总量应符合建库试剂盒的要求。

4.1.2.4 DNA样本的溶解缓冲液组分

DNA样本的溶解缓冲液组分（包括溶解缓冲液的化学试剂配方，pH值等）应符合建库试剂盒的要求。

4.2 文库质量要求

4.2.1 文库类型

通常分为无标签和有标签两种，依据建库过程是否依赖PCR扩增，区分为PCR文库和PCR-free文库两种。

4.2.2 文库质量

4.2.2.1 文库的片段大小和分布

应符合建库试剂盒及测序仪说明书的要求，同时不应存在接头或引物二聚体污染。

4.2.2.2 文库的浓度、体积和总量

应符合建库试剂盒及测序仪说明书的要求。

4.3 测序质量要求

4.3.1 测序读长和测序类型

根据人全基因组高通量测序应用要求选择测序读长，通常为PE150。

4.3.2 测序原始碱基数

应符合测序仪说明书的要求。

4.3.3 碱基识别质量百分比

应符合测序仪说明书的要求，通常Q30不低于85%。

4.3.4 标签拆分率

当采用有标签文库进行测序时，标签拆分率应符合测序仪说明书的要求。

4.4 单样本测序数据质量要求

4.4.1 基因组比对率

应符合人全基因组高通量测序的要求。常规样本类型应不低于99%，如来源于细胞系、全血、白膜层、组织等。少量样本类型应不低于70%，如唾液、口腔拭子、羊水等。

4.4.2 GC含量

应符合人全基因组高通量测序的要求。通常应为39%～43%。

注：在实验室方法建立之初，当该指标严重偏离上述推荐阈值，而人基因组标准品指标值正常时，需要对样本来源或样本提取方法进行问题排查。

4.4.3 有效测序深度

应符合人全基因组高通量测序的要求。通常不低于40x。

4.4.4 20X 测序覆盖率

应符合人全基因组高通量测序的要求。通常不低于95%。

4.4.5 重复测序片段比率

应符合人全基因组高通量测序的要求。通常不高于30%。

4.4.6 碱基错配比率

应符合人全基因组高通量测序的要求。通常不高于1%。

4.4.7 特定区域测序覆盖率

应符合人全基因组高通量测序的要求。通常特定区域的10x 测序覆盖率不低于90%。

注：特定区域是指选定的代表性基因，如*FOXE3*、*SMN1*、*SMN2*、*FMR1*、*G6PD*、*DUOX2*、*GJB2*、*PAH*、*ETFDH*、*MMACHC*、*SLC25A13*、*GCDH*等。

4.5 单样本变异检测质量要求

4.5.1 SNP检测质量要求

4.5.1.1 SNP位点数

应符合人全基因组高通量测序的要求。通常应符合特定分析流程和标准变异数据集下的阈值范围。

4.5.1.2 SNP准确率和灵敏度

应符合人全基因组高通量测序的要求。当使用人全基因组标准品及配套标准变异集评估时，SNP准确率应不低于99%，SNP灵敏度应不低于98%。

4.5.2 Indel检测质量要求

4.5.2.1 Indel位点数

应符合人全基因组高通量测序的要求。通常应符合特定分析流程和标准变异数据集下的阈值范围。

4.5.2.2 Indel准确率和灵敏度

应符合人全基因组高通量测序的要求。当使用人全基因组标准品及配套标准变异集评估时，Indel准确率应不低于90%，Indel灵敏度应不低于90%。

注：通常PCR-free文库的Indel准确率和灵敏度要高于PCR文库。

4.5.3 SV检测质量要求

4.5.3.1 SV位点数

应符合人全基因组高通量测序的要求。通常应符合特定分析流程和标准变异数据集下的阈值范围。

4.5.3.2 SV准确率和灵敏度

应符合人全基因组高通量测序的要求。当使用人全基因组标准品及配套标准变异集评估时，SV准确率应不低于85%，SV灵敏度应不低于50%。

5 评价方法

5.1 样本准备

采用人基因组DNA，样本质量应符合4.1的要求。

5.2 文库制备

按照建库试剂盒说明书进行人全基因组高通量测序文库构建，文库质量应符合 4.2 的要求。

5.3 高通量测序

按照测序仪说明书进行文库上机测序，下机原始数据的质量应符合 4.3 的要求。

5.4 单样本测序数据

单样本原始测序数据进行过滤后，数据质量应符合 4.4 的要求。

5.5 单样本变异检测

使用人基因组标准品和配套标准变异数据集作为参比，变异检测准确率和灵敏度应符合 4.5 要求。

附录 A
（资料性）
人基因组标准品信息

A.1 概述

本附录提供了本文件第 4 章中适用的人基因组标准品信息，该人基因组标准品来源为“测序仪性能评价用脱氧核糖核酸国家参考品”（参考品编号:360070）。

A.2 用途

本文件中使用的人基因组标准品原料为健康人外周血白细胞构建永生细胞系提取的基因组 DNA 样本。

A.3 规格和组成

表A.1　人基因组标准品样本规格和组成

编号	名称	DNA 来源	细胞来源	标准参考序列来源
1	人基因组标准品 DNA 样本	NIFDC-HJ 细胞系	正常男性外周血白细胞构建永生细胞系	人参考基因组 hs37d5

A.4 注意事项

现行国家参考品说明书可在该国家参考品分发单位的网站进行查询下载。国家参考品说明书的部分内容会根据参考品的批次进行变更，应以现行有效版本为准。

参 考 文 献

[1]郝柏林.生物信息学手册（第二版）.上海：上海科学技术出版社，2002.

[2]陈铭.生物信息学（第三版）.北京：科学出版社，2018.

[3]杨焕明.基因组学.北京：科学出版社，2016.

[4]GB/T 29859-2013　生物信息学术语

[5]GB/T 30989-2014　高通量基因测序技术规程

[6]GB/T 35537-2017　高通量基因测序结果评价要求

[7]GB/T 35890-2018　高通量测序数据系列格式规范

[8]YY/T 1723-2020　高通量基因测序仪

[9]GA/T 1693-2020　法庭科学 DNA二代测序检验规范