



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

(点击此处添加与国际标准一致性程度的标识)

草案版次选择

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

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。

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

本标准规定了人类基因组高通量测序数据质量评价方法涉及的术语与定义、人类基因组高通量测序数据质量评价方法和评价参数。

本标准适用于应用非单分子测序的高通量基因测序对人类基因组DNA样本进行全基因组重测序的数据质量评价。

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
T/SZGIA 2-2018	人类全基因组遗传变异解读的高通量测序数据规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

基因测序 gene sequencing

对核酸分子不同碱基类型的测定，即测定组成核酸分子的腺嘌呤（A）、鸟嘌呤（G）、胞嘧啶（C）和胸腺嘧啶（T）或者尿嘧啶（U）等碱基的组成或排列顺序。

[YY/T 1723-2020，定义3.1]

3.2

高通量测序 high-throughput sequencing or massively parallel sequencing

又叫大规模平行测序，能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。

[改写GB/T 35890-2018，定义3.1]

3.3

全基因组测序 whole genome sequencing

测定生物体基因组完整DNA序列的过程，包括生物体的所有染色体DNA、线粒体DNA、植物的叶绿体DNA及生物体样本中的微生物DNA。

3.4

人类全基因组重测序 human whole genome sequencing or human whole genome re-sequencing

基于人类基因组参考序列，对人类不同个体或群体进行全基因组测序，并在此基础上对个体或群体进行生物信息分析的方法。

3.5

生物信息学 bioinformatics

应用信息科学及相关学科的方法和技术，研究和分析生物体系和生物过程中信息存储、处理和传递的一门交叉学科。

[改写GB/T 29859-2013，定义2.1.1]

3.6

参考序列 reference genome sequence

是指已公开发表的某物种的供参考的全基因组序列，用来供同种物种的不同或相同个体测序后，与之进行比较分析，找出相互间存在的差异。参考序列上存在少部分N区域，即尚不清楚碱基序列的区域。

[T/SZGIA 2-2018，定义3.15]

3.7

人类基因组参考序列版本 human genome reference sequence version

由不同基因组研究团体发布的人类参考基因组序列。

[T/SZGIA 2-2018，定义3.15.1]

3.8

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。

[GB/T 35890-2018，定义3.6]

3.9

测序平均长度 average read length of sequencing

测序仪器单次测序能达到的平均读长。

注：通常以碱基数表示。

[GB/T 35537-2017，定义3.1.2]

3.10

单末端测序 single-end sequencing

对DNA序列片段中模板链或互补链中一条进行测序。

3.11

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。

[GB/T 35890-2018，定义3.3]

3.12

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。

[GB/T 35890-2018，定义3.4]

3.13

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，是测序的最小单位。

[GB/T 35890-2018，定义3.2]

3.14

标签 barcode

测序片段的ID，保证一个序列编号对应一段序列片段或对应一个测序文库，具有唯一性。

[GB/T 35890-2018，定义5.1]

3.15

碱基识别 base calling

测序过程中从荧光信号或其他由于测序反应而产生的信号转换成序列信息的过程。

[YY/T 1723-2020，定义3.5]

3.16

碱基识别质量 quality of base calling

衡量碱基正确识别的概率。通常以ASCII码直接表示。

[YY/T 1723-2020，定义3.6]

3.17

GC含量 GC content

测序片段碱基中G（鸟嘌呤）和C（胞嘧啶）所占的百分比。

3.18

测序原始碱基数 Sequencing raw base

测序后原始的、未经任何处理的碱基的总数即为测序原始碱基数，又简称原始数据量。单末端测序时，测序原始碱基数=单端测序碱基数。双末端测序时，测序原始碱基数=单端测序碱基数*2。

3.19

数据过滤 data filtering

根据不同的数据过滤参数，对测序片段（reads）进行过滤，去除未达到过滤参数要求的测序片段。通常会过滤掉包含低质量，读N碱基和接头污染的测序片段。

3. 20

有效测序碱基数 effective sequencing base

测序片段（reads）经过数据过滤剩余的可以比对到目标基因组上的测序片段的碱基总数即为有效测序碱基数，又简称有效数据量。

3. 21

序列比对 sequencing alignment

比较两个或两个以上核苷酸序列间相似性的过程。

[GB/T 29859-2013, 定义2.2.1]

3. 22

基因组比对率 mapping reads rate

数据过滤后的reads与目标参考基因组比对，匹配的reads占过滤后的reads的百分比。

3. 23

基因组唯一比对率 unique reads rate

单末端测序中的一条reads或双末端测序中的一对reads与参考基因组比对时，比对到参考基因组唯一位置的reads数与所有比对上的reads数的百分比。

3. 24

测序深度 sequencing depth

即一个全基因组测序样本的平均测序深度，指非N区所有碱基的深度的平均值，计算方法：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。

[改写T/SZGIA 2-2018, 定义3.18]

3. 25

碱基错配 base mismatch

测序序列与参考基因组比对过程出现的非遵循碱基配对原则的测序碱基。如腺嘌呤（A）与胞嘧啶（C）或鸟嘌呤（G）配对，或胸腺嘧啶（T）与鸟嘌呤（G）或胞嘧啶（C）配对。

3. 26

碱基错配比率 base mismatch rate

碱基错配比率=错配碱基总数/比对上参考基因组的碱基总数*100%

3. 27

重复测序片段 duplicate reads

测序中得到的比对到参考基因组的位置、方向及碱基序列均一致的重复测序片段（reads）。

3. 28

重复测序片段比率 duplication rate

重复测序片段比率=重复测序片段总数/比对上参考基因组的测序片段总数*100%

3.29

覆盖度 Coverage ratio

是指将测序序列比对到参考序列上时，所有被比对到的区域占参考序列总区域的百分比。计算时需要去除参考序列上N碱基区域。

[T/SZGIA 2-2018, 定义3.17]

4 要求

4.1 样本要求

4.1.1 样本类型要求

测序样本类型为人类基因组DNA。DNA的来源主要为人类全血、血液白膜层、唾液、组织和细胞系。

4.1.2 样本质量要求

DNA样本的完整度、纯度、浓度及总量需要符合制造商规定的样本质量要求。

4.2 文库要求

4.2.1 文库类型要求

文库类型包括加上接头之后不做PCR的无标签文库、单标签文库和双标签文库，以及加上接头之后要做适当循环数PCR的无标签文库、单标签文库和双标签文库。

4.2.2 文库质量要求

文库的片段大小、纯度、浓度及总量需要符合制造商规定的重测序文库质量要求。

4.3 测序质量要求

4.3.1 测序读长和测序类型

根据测序应用需求，选择合适的测序读长（50bp、100bp、150bp、200bp、400bp等）和测序类型（单末端测序和双末端测序）。

4.3.2 测序原始碱基数、有效测序碱基数和测序深度

单次测序，测序片段总数应不低于制造商规定的要求。

单个文库的测序原始碱基数、有效测序碱基数和测序深度应符合特定应用的要求。

4.3.3 碱基识别质量百分比

单次测序，统计碱基识别质量在规定阈值以上的比例，碱基识别质量百分比平均值应不低于制造商规定的要求。

注：不同原理的测序仪在碱基识别质量表述有差异，如Q20、Q30等。

[YY/T 1723-2020, 定义4.3]

4.3.4 标签拆分率

当采用无标签文库进行测序时，无需衡量标签拆分率。

当采用单标签文库或双标签文库进行测序时，标签拆分率应不低于制造商规定的要求。

4.4 测序数据质量要求

4.4.1 GC 含量

当DNA来源为人类全血、血液白膜层、组织和细胞系时，单个文库的GC含量在39-43%范围。

当DNA来源为人类唾液时，单个文库中可能含有其他物种基因组的污染，其GC含量可能超过39-43%的范围。

4.4.2 基因组比对率

当DNA来源为人类全血、血液白膜层、组织和细胞系时，单个文库的基因组比对率 $\geq 98\%$ 。

当DNA来源为人类唾液时，单个文库中可能含有其他物种基因组的污染，其基因组比对率可能低于98%。

4.4.3 碱基错配比对率

单个文库的碱基错配比对率应不高于制造商规定的要求，通常应 $\leq 1.5\%$ 。该指标间接反应测序碱基的准确性。

4.4.4 重复测序片段比率

单个文库的重复测序片段比率应不高于制造商规定的要求。通常，加上接头之后不做PCR的文库的重复测序片段比率应 $\leq 10\%$ ，加上接头之后做适当循环数PCR的文库的重复测序片段比率应 $\leq 20\%$ 。该指标越低，基因组唯一比对率越高，可真正用于变异数据分析的有效测序碱基数越多。

4.4.5 覆盖度

单个文库在特定测序深度下的覆盖度应符合特定应用的需求。

5 评价方法

5.1 实验材料

采用人类基因组DNA，样本质量应符合4.1的要求。

5.2 文库制备

采用5.1所述实验材料进行高通量重测序文库构建，文库质量符合4.2的要求。

5.3 文库上机测序

根据测序应用要求，选择适配的测序长度和测序类型，根据不同测序平台的经验值，预估达到测序原始碱基数需要加入的单个文库的总量，进行文库上机测序。测序质量应符合4.3的要求。

5.4 测序数据评估

将测序数据进行过滤,过滤后GC含量应符合4.4.1的要求。将过滤后的数据与参考基因组进行比对,数据质量应符合4.4.2-4.4.5的要求。

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.
-