ICS 07.080
CCS A 40

中华人民共和国国家标准

GB/T 43584.2—2023/ISO 20397-2:2021

生物技术　大规模并行测序 第2部分:测序数据的质量评估

Biotechnology—Massively parallel sequencing—Part 2:Quality evaluation of sequencing data

（ISO 20397-2:2021，IDT）

2023-12-28 发布　　2023-12-28 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。

本文件是GB/T 43584《生物技术　大规模并行测序》的第2部分。

GB/T 43584已经发布了以下部分:

——第2部分:测序数据的质量评估。

本文件等同采用ISO 20397-2:2021《生物技术　大规模并行测序　第2部分:测序数据的质量评估》。

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。

本文件由全国生化检测标准化技术委员会(SAC/TC 387)提出并归口。

本文件起草单位:中国测试技术研究院生物研究所、江汉大学、中国测试技术研究院、深圳华大生命科学研究院、河北省食品检验研究院、成都医学院、深圳华汉基因生命科技有限公司、浙江贝兰伯生物技术有限公司、深检集团(深圳)医学检验实验室。

本文件主要起草人:周李华、李怀平、叶善蓉、易艳、王丹、姜展樾、魏晓锋、林华、樊东生、陈佳平、叶德萍、彭海、冯双、王奇、马丽侠、张岩、张勇、杨俊、张才敏、蒋慧、杨国武。

引　言

大规模并行测序(MPS)是一种利用大规模并行处理进行核酸测序的高通量分析方法,该方法可在相对较短时间内对不同生物体的全基因组、转录组和特定靶核酸进行研究。

MPS已用于许多生命科学领域,可对数百万乃至数千万个核苷酸碱基进行测定和高通量分析。生物体内脱氧核糖核酸和核糖核酸聚合物的生物变异为准确测定序列带来了挑战。通过MPS测定,序列质量取决于许多因素,包括但不限于样品质量、文库制备、平台选择及测序数据质量。

GB/T 43584拟由以下部分构成:

——第1部分:核酸和文库制备。第1部分主要提供基础研究,目的在于规定了测序和数据生成前文库制备和文库质量评估的一般准则和注意事项。

——第2部分:测序数据的质量评估。第2部分基于第1部分开展具体操作和数据质量控制并为第3部分提供研究基础。

——第3部分:宏基因组学的总体要求和指南。第3部分包含第1部分、第2部分,规定了宏基因组学从样品制备、生成和分析测序数据的准则。

测序数据分析在数据存储、计算时间和变异检测准确性等多个领域均对生物信息学提出较大的挑战。与测序数据相关的主要挑战之一是监测数据处理流程各个阶段的质量控制指标,此点易被忽视。了解数据质量对下游序列分析至关重要。核酸测序数据处理与分析的质量控制可分为三个阶段:原始数据、比对和变异识别。本文件提供了MPS测序数据质量评估的注意事项,以及针对不同的MPS平台提供具体建议。

生物技术　大规模并行测序
第2部分:测序数据的质量评估

1　范围

本文件明确了对大规模并行测序数据进行质量评估的整体要求和建议。涵盖了原始数据生成后的程序、序列比对和变异识别。

本文件提供了大规模并行测序(MPS)数据验证和存档的一般指南。

本文件不适用于与从头组装相关的任何处理。

2　规范性引用文件

本文件没有规范性引用文件。

3　术语和定义

下列术语和定义适用于本文件。

3.1

接头序列　adapter sequence

接头　adapter

一种已知序列的人工寡核苷酸,可连接到核酸片段的3′端或5′端。

注:它提供引物位点以及对测序插入序列片段所需的其他必要序列。

3.2

算法　algorithm

完全确定的有限序列指令,通过它可以从输入变量的值计算输出变量的值。

[来源:IEC 60050-351:2013,351-42-27,有修改]

3.3

碱基识别　base calling

将大规模并行测序原始电信号转化为核苷酸序列的计算过程。

注:碱基识别的应用和算法的性能由读序和共有序列准确性来确定。

3.4

生物信息学流程　bioinformatics pipeline

对程序、脚本或软件的整合和顺序执行,在数据处理过程中,原始数据或一个程序的输出作为下一个步骤的输入。

示例:碱基质量剪切程序的输出能作为从头组装程序的输入。

3.5

捕获效率　capture efficiency

所测得目标区域序列占所有测序序列或参考序列的百分比。

3.6

覆盖度　coverage

覆盖深度　coverage depth

在一次测序中,每个指定位置的碱基被测序的次数。

注：覆盖特定位置的读序数目。

3.7

覆盖宽度　coverage breadth

多次测序产生的基因组序列总长占覆盖目标区域的百分比。

3.8

簇密度　cluster density

每个单元中簇的数量。

注 1：簇密度适用于有扩增步骤的 MPS(3.30)平台。

注 2：在某些测序平台上，每个单测序簇的密度来自于单分子。

注 3：簇密度通常以 K/mm^2 表示。

3.9

环化共有序列测序　circular consensus sequencing;CCS

一种高准确度的测序模式，指一定大小的插入片段在滚环扩增反应中多次测序。

注：在这种模式下，使用多个通道对同一分子进行测序，实现更高准确度。

3.10

覆盖范围　coverage range

多次测序得到的覆盖整个基因组的深度范围。

3.11

拷贝数变异　copy number variation;CNV

拷贝数变异体　copy number variant

一个生物体的基因组中一个或多个 DNA 片段的拷贝数的变异。

注：拷贝数变异(CNVs)是指长度至少为 1 kb 片段的插入、缺失、倒位和重复。

3.12

脱氧核糖核酸　deoxyribonucleic acid;DNA

脱氧核糖核苷酸的聚合物，以双链(dsDNA)或单链(ssDNA)形式出现。

[来源：ISO 22174:2005，3.1.2]

3.13

缺失　deletion

与参考序列相比，核酸序列中一个(或多个)碱基对的缺失。

3.14

重复水平　duplication level

一个文库中每个相同序列的重复数。

注：重复水平通常以图表形式显示序列的相对数量。

3.15

GC 含量　GC content

鸟嘌呤和胞嘧啶在一个或多个核酸序列所有碱基中所占的比率。

注：多核苷酸中鸟嘌呤和胞嘧啶的含量，通常以总含氮碱基的摩尔分数(或百分比)表示。总含氮碱基包括一次或多次 MPS 过程中所产生的核苷酸碱基总数。

3.16

基因　gene

位于染色体上编码特定功能产物(RNA 或蛋白质)的一段核苷酸(DNA 或 RNA)序列。

注 1：基因是遗传信息的基本单位。

注 2：基因由经细胞内剪接后重新排列的非连续性核酸片段组成。

注 3：基因包括或是含基因表达元件在内的操纵子的一部分。

3.17

插入/缺失　indel

基因组 DNA 中插入(3.18)或/和缺失(3.13)的核苷酸片段。

注：插入/缺失突变长度小于 1 kb。

3.18

插入　insertion

核酸序列中加入一个(或多个)核苷酸碱基对。

[来源：ISO/TS 20428:2017，3.19，有修改]

3.19

测序　sequencing

对核酸分子中核苷酸碱基(腺嘌呤、鸟嘌呤、胞嘧啶、胸腺嘧啶或尿嘧啶)排列顺序和组分的测定。

注：序列通常用 5′端到 3′端表示。

[来源：ISO/TS 17822-1:2020，3.19，有修改]

3.20

序列比对　sequence alignment

根据相似区域排列核酸序列。

注：序列比对可能不需要参考基因组/参考靶标核酸区域，目的或许不是产生组装基因组。

3.21

原始数据　raw data

由测序仪产生的原始测序数据，未经任何软件预过滤与分析的数据。

3.22

核糖核酸　ribonucleic acid

以双链或单链形式存在的核糖核苷酸聚合物。

注：信使 RNA(mRNA)的核苷酸序列所携带的遗传信息能指导细胞中蛋白质的合成。

3.23

核糖核苷酸　ribonucleotide

以核糖为戊糖组成部分的核苷酸，是构成 RNA 的基本单位。

注：核糖核苷酸包括腺嘌呤核糖核苷酸(AMP)、鸟腺嘌呤核糖核苷酸(GMP)、胞嘧啶核糖核苷酸(CMP)或尿嘧啶核糖核苷酸(UMP)。

3.24

读序　read

序列读序　sequence read

由测序仪产生的核苷酸序列。

注：一个读序是指对应于单个核酸片段的所有(或部分)核酸碱基对(或碱基对概率)的推断序列。读序指 MPS 实验中获得的所有序列。

3.25

读序类型　read type

序列类型，取决于实验设计和实施的序列读取方式。

示例：读序类型包括单端读序、双端读序、配对读序、连续长读序、环化共有序列。

3.26

参考序列　reference sequence

用于读序定位时的比对核酸序列，或作为基因和序列变异注释时的基础核酸序列。

3.27

多路分解　demultiplexing

多重复合过程的反向计算，将两个或多个样本混合，让 MPS 仪器单次测序运行即可对所有样品进行测序。

注 1：样品混合之前需标记条形码/索引。

注 2：多路分解是一种计算算法，能够根据条形码将一组读序进行分离。

3.28

定位　mapping

将核酸序列与现有基础(参考)序列进行比较并构建一个共有序列的过程。

3.29

配对　mate pairs

配对读序　mate pair reads

通过将样本片段化(大于或等于 2 kb)获得的长核酸序列末端的成对读序。

3.30

大规模并行测序　massively parallel sequencing；MPS

基于多个 DNA 模板独立聚合延伸的测序技术。

注：大规模并行测序技术一次运行能同时读取数百万或数十亿的 DNA 分子模板。

3.31

双端读序　paired-end reads

通过从一个 DNA 片段两个末端测序获得的读序。

注：在双端测序中，仪器同时对插入片段(200 bps～800 bps 范围内)的两端进行测序。

3.32

质量值　quality score

***Q* 值　*Q* score**

碱基质量值　phred quality score

衡量给定核苷酸碱基的测序质量

注 1：*Q* 值定义见公式(1)：

$$Q = -10\lg p \qquad \cdots\cdots(1)$$

式中：

p——碱基识别错误率。

注 2：质量值为 20 代表错误率为 1/100，相应的准确率为 99/100。

注 3：质量值越高，出错的概率越小。较低的质量值会导致大部分读取无效。低质量值也能表示假阳性变异，导致结论不准确。

3.33

运行　run

测序仪从启动到获得原始数据的单次循环过程。

3.34

序列注释　sequence annotation

对 DNA、RNA 或蛋白质序列的结构或功能方面的信息加以解释、评价或说明的过程。

注：序列注释视为将数据元分配给序列的过程。

3.35

单端读序　single-end read

通过从 DNA 片段的一端读取到另一端而获得的序列。

3.36

单核苷酸变异　single nucleotide variant;SNV

一个核酸分子中单个核苷酸的变异。

3.37

结构变异　structural variation;SV

1 kb 及以上范围的 DNA 片段发生倒位、平衡易位和基因组失衡等结构变化。

注：常见的结构变异类型包括拷贝数变异(缺失、插入、扩增、重复)、拷贝数中性缺失(杂合性缺失)、倒位、片段重复和易位(平衡或失衡)。

3.38

子读序　subread

从发夹接头之间读取到的片段。

3.39

原始读序修剪　trimming of raw reads

去除低质量或被污染序列，同时保留大规模并行测序读取的高质量序列的过程。

3.40

变异　variation

序列中一个或多个核酸碱基与预期碱基之间的差异。

3.41

变量识别　variant calling

准确识别数据序列与参考序列之间差异的过程。

3.42

零模波导　zero mode waveguide;ZMW

把光能量定向限制于小于光波长尺度的区域范围的光波导。

注：聚合酶被固定在 ZMW 的底部，通过判别荧光信号来识别结合到核酸链上的核苷酸分子的类别。

4　原始数据

4.1　通则

序列中每个核苷酸宜匹配一个数值(碱基质量值)，该值与碱基识别过程准确度相关(如适用)。

4.2　原始数据文件

序列读序文件宜使用仪器特定的软件或流程生成。每次测序实验中应实时监测并记录物理参数，如信噪比等。

序列读序文件宜设定适当的文件格式，包含每个序列读序的编码、相应的标识符以及每个核苷酸的碱基质量值。

注：FASTQ 格式(或可转换为 FASTQ 格式)可作为 MPS 数据集质量分析的标准格式。FASTQ 作为一种可跨平台交换的文件格式已被广泛接受。

生物信息学流程中宜采用适当的软件对测序生成的输出文件和相关质量指标进行分析。

4.3　原始数据的质量评估

4.3.1　通则

质量控制指标可能因 MPS 平台、文库制备方法和分析目的差异而有所不同。

序列结果宜由专业人员解读。解读应考虑读序重复数量的统计可靠性,以达到符合预期目的的质量水平。

使用处理读序工具时,宜考虑质量评定结果和原始读序的修剪。

4.3.2 基本统计信息

试验应记录基本统计信息,包括但不限于:

a) 平台类型;
b) 读序类型;
c) 建库试剂盒;
d) 读序长度;
e) 读序数量;
f) 总 GC 含量;
g) 总序列长度。

4.3.3 质量指标

评估原始数据的质量控制指标可参考以下指标,但不限于以下指标:

a) 序列长度分布;
b) 每个序列 GC 含量;
c) 质量值:
   1) 单碱基序列质量,
   2) 序列质量统计,

   注 1:低质量值预示了变异识别假阳性的增加。

   3) 所有序列都宜根据单碱基序列质量标记为“警告”或“通过”;
d) 碱基比例分布;
e) 信噪比的可接受度;
f) 序列重复水平;
g) 高于阈值水平;
h) 簇密度;
i) 全外显子组测序或全基因组测序或扩增子测序的转换/转换比;
j) 接头比例/接头序列污染程度;
k) 污染物(定性、定量);
l) 错误率;

   注 2:包括均聚物错误:一个单核苷酸连续多次出现在序列中碱基识别的错误量。

m) k-mer 分析;

   注 3:在计算基因组学中,k-mers 指核酸序列中长度为 k 的所有可能的子序列。高于阈值的 k-mers 有利于分析可能由重复序列导致的潜在基因组错配。

n) N 片段;

   注 4:尚不明确的碱基数量/百分比。

o) 重复延伸和重复序列;
p) 循环测序过程中的核苷酸分布。

4.4 原始数据预处理

原始数据预处理可包括但不限于以下计算步骤(如适用):

a) 去除/修剪低质量的序列/碱基;

b) 多路分解;

c) 去除接头/引物和污染物;

d) 校正错误;

e) 过滤重复读序;

f) 修剪读序至固定长度;

g) 识别 CCS 读序。

当使用 CCS 数据库时,在进行下游分析前宜先获得并过滤 CCS 读序。

5 序列比对与定位

5.1 通则

宜根据应用程序选择序列比对和定位策略。

示例: RNA 的拼接定位和 RNA 测序的非拼接定位策略。

比对/定位软件及工具均可用于比对。

评估比对质量可通过采用合适的比对视图以及比对文件中所提供的信息。

不同宜用序列的序列比对及定位软件见附件 C。

定位宜使用参考基因组/参考目标核酸区域,根据实验设计合理筛选。

注 1: 考虑因素包括参考基因组/目标区域参考序列的版本、生物体中不同株系以及掩蔽、软掩蔽或非掩蔽基因组的选用。

注 2: 开源的序列比对和定位软件可网上下载。

5.2 序列比对与定位文件格式

比对通常以下列文件格式保存。

a) 序列比对格式(Sequence alignment format, SAM)。

注 1: SAM 是一种以制表符为分隔符的文本格式,包括头文件、比对两部分。每条比对线包含 11 个必要比对信息,如定位的位置、校正器特定信息的可选字段变化数。

b) 二进制比对格式(Binary alignment format, BAM)。

注 2: 为一种精简格式,类似于二进制的 SAM 格式。

c) 基于参考序列的压缩比对定位格式(Compressed reference-oriented alignment map, CRAM)。

注 3: CRAM 是一种测序读序文件格式,其是基于参考序列数据库,提供压缩模式的有损/无损运行包。

d) 基因组动态图像专家组格式(Moving pictures experts group for genomics, MPEG-G)。

注 4: MPEG-G 是一种基于基因组收录数据的表示格式,由单个序列读序或成对序列读序组成的数据结构及其相关测序和比对信息;其包含详细的定位和比对数据库、单个或成对读序标识符(读序名字)及质量值。访问单元结构是独立访问和检查编码基因组信息的单元,能聚集和编码基因组收录数据。

注 5: MPEG-G 按照 ISO/IEC 23092 系列标准执行。

比对文件宜包含比对过程中各个读序的位置、方向及质量等信息。

依赖所属应用程序,算法和工具可适用于比对文件。

5.3 序列比对和定位的质量控制

5.3.1 基本比对数据

5.3.1.1 概述

获得并记录基本比对或定位数据。

基本比对或定位数据因实验设计和读序类型的不同而存在差异。

5.3.1.2 单端读序的定位统计信息

具体包括：

a) 读序总数是指定位到参考序列或基因组的读序数。

b) 未定位读序是指未定位到参考序列或基因组的读序。

c) 定位读序是指比对到参考序列或基因组的读序。

d) 唯一定位读序是指与参考序列或基因组一次即可精确比对的读序。

注 1：定位的唯一性试具体情况而定。基于一组定位参数的唯一定位读序可能是另一组定位参数的多靶标读序。

e) 多靶标定位读序是指在参考序列或基因组上具有多个可能的对应位置的读序。

注 2：多靶标取决于定位环境。

5.3.1.3 双端读序的定位统计信息

具体包括：

a) 配对总数是指定位到参考序列或基因组的双端读序数。

b) 定位配对是指两端均被定位的读序。

c) 部分定位配对是指配对端中只有一端配对被定位的读序。

d) 非定位配对是指未能定位到参考序列或基因组的读序。

e) 不适当的定位配对是指其中一端配对按非预期方向定位的成对读序。

注 1：也被称为无序定位配对。

f) 适当的定位配对是指两端配对均按预期方向定位的成对读序。

注 2：也被称为有序定位配对。

5.3.1.4 子读序的定位长度

比对到目标参考序列的子读序长度，不包括接头序列。

5.3.2 质量指标

以下质量控制参数适用于不同应用中：

a) 比对率；

注 1：低质量定位可能由非特异性扩增、非靶标 DNA 污染或其他原因导致。

b) 片段长度，或待测序的 DNA/RNA 的长度；

c) 双端读序插入片段尺寸是指所测得 DNA/RNA 接头之间的长度；

注 2：插入片段尺寸分布的峰值用于质量评估。

d) 仅基于扩增子测序的重复水平；

e) 预期覆盖度包括覆盖深度、宽度和范围；

注 3：附件 B 提供了适用于不同应用的覆盖范围清单。

f) AT/GC 偏差；

注 4：能通过 GC 含量与测序深度/覆盖度的百分比进行评估。

g) 定位质量值；

h) 捕捉效率；

注 5：捕获效率是外显子组测序或其他基于目标捕获测序最重要的质量控制参数。

i) 平均深度或中位深度，在该深度测序基因组所覆盖的百分比；

j) 无需定位配对的数量；

k) 高质量读序比对；

l) 错配率；

m) 共有序列准确性；

注 6：共有序列准确性是基于多个测序读序及子读序同时比对获得的精确性。

n) 环化共有序列准确性；

注 7：环化共有序列准确性是基于多个测序通道围绕单一环状模板分子获得的精确性，常被用于 CCS 中。

o) 子读序准确性。

注 8：碱基识别的定位准确性。

5.3.3 序列比对和定位质量评估方法

基于评分方式评估序列比对质量。

注：评分矩阵的选择取决于具体应用程序。

5.4 比对后处理

比对后处理包括但不限于：

a) 局部再比对插入/缺失附近序列或基于每个碱基比对的计算；

b) 去除重复项；

c) 再校正碱基质量值；

d) 根据碱基质量修剪读序平均长度。

6 变异识别

6.1 通则

6.1.1 序列变异主要有四种类型(单核苷酸变异、插入/缺失、拷贝数变异和结构变异)，为达到灵敏、特异性识别的目的，不同类型的序列变异应采用不同的算法。

6.1.2 软件工具包的范围及所需验证类型取决于分析设计。

6.2 变异识别的数据文件

6.2.1 变异识别应使用适当的规范进行注释。说明书应包含元信息、标题行和数据行，每条数据行包含基因位置信息和每个位置中样品的基因型信息。

示例 1：被识别的变异体使用变异识别格式(VCF)进行注释。

示例 2：存在说明和存储变异识别的替代规范：

a) 基因组的 VCF 条例；

b) 序列本体基因组变异格式 1.10 版；

c) 人类基因组变异学会，人类基因组变异学会(Human Genome Variation Society，HGVS)简易版 15.11；

d) 全球基因组学和健康联盟(Global Alliance for Genomics and Health，GA4GH)文件格式。

6.2.2 变异文件应包括所使用的规范和版本。

6.2.3 变异识别宜配置为输出，参考序列、变异体、未识别序列以及目标区域内本地信息。

6.3 变异识别的质量指标

质量控制指标宜包括但不限于(如适用)：

a) 变异位置的读序覆盖深度阈值；

b) 变异体的质量值；

c) 链偏好性；

d) 等位基因读序百分比；

e) 与变异识别的准确性和灵敏度有关的其他指标，包括但不限于：

1) 变异体总数，

2) 假阳性数量，

3) 假阴性数量，

4) 等位基因和基因型错配数，

5) 转换/转换比率，

6) 杂合/纯合子比率；

f) 样品交叉污染分析。

6.4 假阳性变异处理

假阳性变异宜基于序列比对和变异识别相关的质量控制指标，在原始变异文件中标记或滤除。

6.5 序列注释

对变异体进行注释，以确定其生物学意义，并实现功能优化和下游解释。

7 验证

7.1 通则

7.1.1 提供基于 MPS 检测的实验室宜进行“内部”生物信息学流程验证。

7.1.2 在验证过程中应确定试验的性能要求，同时每次样品检测，需使用同一个规范来监测试验性能。

7.1.3 在验证过程中应评估特定的质量控制和质量保证参数，确定最佳性能。

7.1.4 实验室应建立监测所有质量指标的标准和方法，宜形成相应的程序文件并定期验证，以确保最佳分析能力。部分平台推荐的质量指标及具体值见附录 A。

7.1.5 实验室应制定具体措施，以确保生物信息学流程中生成的数据文件的完整性，并对未经授权或意外更改的数据文件提供警报或禁止使用。

7.1.6 当对生物信息学流程中的步骤进行重大更改时，均需进行补充验证。

7.2 质量指标验证

7.2.1 分析验证应在分析目的明确并形成文件的基础上进行。测量目的应明确且有证明文件。

7.2.2 实验室应在验证过程中为试验建立可接受的原始碱基识别质量值阈值。

7.2.3 宜建立去除低质量碱基的预处理方法，降低假阳性发生率。

7.2.4 验证过程中宜确定试验所包含的基因组中 GC 偏倚程度。

7.2.5 应在验证计划中确定比对质量参数，并宜证明该试验仅评估比对所指向的区域。若适用，宜建立将读序过滤至非目标区域的步骤。

7.2.6 应定义覆盖范围，使其在利益范围内达到足够的灵敏度和特异性。

7.2.7 根据测序目的，每个实验室应在标准试验条件下，建立特定区域覆盖特征的最低标准。对于均质样本，需确认序列，可接受较低深度。在一个区域的不同识别过程中，或1%的混合样本中的稀有序列，均需进行深度测序。

7.2.8 在验证阶段应确定目标区域所需的覆盖度(覆盖范围)。不同应用的推荐范围见附录B。

7.2.9 宜为每次试验建立最大重复率的可接受参数。

7.2.10 宜建立分析流水线滤除重复读序，以增加可用测序数据的数量，防止等位基因发生偏倚。

7.2.11 各实验室应保证对链偏好性的限值，并制定可替换试验的具体标准。

7.2.12 质量指标可参考具有良好特性、具有可靠参考序列的相关标准进行验证，以保证校正及变异识别的准确性。

7.2.13 推荐采用Sanger测序法验证重要的结合区域。

8 文件

8.1 实验室应记录所有MPS结果分析、注释和报告的算法、软件和数据库。在整个生物信息学流程中，应记录所有版本信息，并对所有结果进行追溯。

8.2 实验室应记录任何与默认配置不同的定制项目，或宜说明哪些参数是自定义的。

8.3 若适用，宜确定参考序列、版本号和详细信息。

8.4 实验室宜记录最佳性能的质量控制参数。

示例：在主要步骤中，实验室将确定可接受的标准，如通过仪器指定质量过滤器的读序。

8.5 实验室宜记录将较大变量数据集缩减为候选基因或变量列表的生物信息学过程。

8.6 宜将符合规定要求的证据形成文件。

附 录 A
（资料性）
特定 MPS 平台示例的质量指标

以下平台是核酸测序常用的 MPS 平台。用于质量评估的指标示例如表 A.1 所示。

注：以全人类基因组序列为例，为每个质量指标提供特定的值。

表 A.1 特定 MPS 平台的质量指标

平台名称	原始文件的格式	读序长度	质量值（H/L）	GC 含量	冗余率	簇密度	接头比例
illumina[a] HiSeq 4000	fastq.gz	50 bp～200 bp	＞Q30	39%～42%	＜10%	5 000 000 000	＜3%
Thermo Fisher Proton[b]	DAT	50 bp～200 bp	＞Q20	39%～42%	NA	60 000 000～80 000 000	＜3%
BGI[c]/MGI MGISEQ-2000	fastq.gz	50 bp～200 bp	＞Q30	39%～42%	＜5%	150 000 000	＜3%
Oxford Nanopore PromethION[d]	FAST5	10 kbp～300 kbp	＞Q20	39%～42%	NA	2 560 channels[f]	＜3%
PacBio Sequel II[e]	bam	10 kbp～100 kbp	＞Q20	39%～42%	NA	8 000 000 ZMWs[g]	＜3%

[a] 本信息仅为方便使用本文档的用户而提供，并不构成 ISO 对产品名称的认可。

[b] 本信息仅为方便使用本文档的用户而提供，并不构成 ISO 对产品名称的认可。

[c] 本信息是为了方便用户使用文件，并不构成 ISO 对产品名称的认可。

[d] 本信息仅为方便使用本文档的用户而提供，并不构成 ISO 对产品名称的认可。

[e] 本信息仅为方便使用本文档的用户而提供，并不构成 ISO 对产品名称的认可。

[f] 采用波段作为测量单位。

[g] 采用 ZMWs(零模波导)进行测量。

附 录 B
（资料性）
按应用划分的覆盖范围和推荐读序

表 B.1 列举了各种不同应用程序的覆盖度和读序水平。

表 B.1 应用程序的覆盖度和推荐读序

MPS类型	应用	推荐范围	推荐读序
全基因组测序	纯合的单核苷酸变异-等位基因相同的基因中单核苷酸的变化	15×[a]	—
	杂合子单核苷酸变异-单核苷酸在等位基因彼此不同的基因中发生变化	33×	—
	核苷酸被插入或缺失的基因组突变	60×	—
	拷贝数变异-一个基因之间拷贝数的变异	1×～8×	—
全外显子组测序	纯合子单核苷酸变异	100×（3×本地阅读覆盖范围）[b]	—
	杂合子单核苷酸变异	100×（13×本地阅读覆盖范围）[c]	—
有针对性的测序	插入/缺失	无推荐	—
	目标区域的单核苷酸变异/结构变异	1 000 倍～10 000 倍	—
RNA 测序-转录组测序	16S rRNA 基因	—	最低每个样品 100
	差异表达谱-跨多个基因的基因表达的定量测量，以检查不同水平	—	10 000 000～25 000 000
	选择性修剪-从 mRNA 转录产物中鉴定不同修剪变异体	—	500 000～1 000 000（针对短读序平台） 2 000 000～3 000 000（针对长读序平台）
	等位基因特异性表达-受特定等位基因影响的转录组表达	—	50 000 000～100 000 000
RNA 测序-小 RNA（microRNA）测序	差异表达-小 RNA 表达检测样本中不同水平的表达定量测量	—	～1 000 000～2 000 000
	发现新的小 RNA	—	～5 000 000～8 000 000

注 1：结果能通过互补的蛋白质组学实验进行验证。

注 2：人体样本的推荐覆盖度。

[a] 15×表示本地相同的覆盖率，而不是整体的平均覆盖率。此处数字代表个例。

[b] 100×是整个外显子组测序的总平均覆盖率。3×表示检测 SNP 的本地覆盖率。

[c] 100×是整个外显子组测序的总平均覆盖率。13×表示检测 SNP 的本地覆盖率。此处数字代表个例。

附 录 C
（资料性）
序列比对和定位软件

表 C.1 列举了序列比对和定位软件。

表 C.1 序列比对和定位软件

功能描述	软件/工具
比对或定位	Blast，Blat，SOAP，BWA，Bowtie2 等
RNA 测序分析中剪接的评估	Bowtie2，BWA，HISAT2，STAR 等
可视化比对视图	Bam View，Intergrative Genmomic Viewer

注 1：软件定期更新，高度依赖/与平台、应用程序和序列数据相关。2020 年 6 月，显示表中示例有效。

注 2：本表所列软件的例子均为可用的合适软件。此信息是为了方便本文档的用户而提供的，并不构成 ISO 对这些产品的认可。

参 考 文 献

[1] ISO/TS 17822-1:2020 In vitro diagnostic test systems Nucleic acid amplification-based examination procedures for detection and identification of microbial pathogens Laboratory quality practice guide

[2] ISO/TS 20428 Health informatics—Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records

[3] ISO 22174:2005 Microbiology of food and animal feeding stuffs—Polymerase chain reaction (PCR) for the detection of food-borne pathogens General requirements and definitions for the detection of food-borne pathogens—General requirements and definitions

[4] ISO/IEC 23092-1:2020 Information technology—Genomic information representation—Part 1: Transport and storage of genomic information

[5] ISO/IEC 23092-2:2020 Information technology—Genomic information representation—Part 2: Coding of genomic information

[6] ISO/IEC 23092-3:2020 Information technology—Genomic information representation—Part 3: Metadata and application programming interfaces (APIs)

[7] ISO/IEC 23092-4:2020 Information technology—Genomic information representation—Part 4: Reference software

[8] ISO/IEC 23092-5:2020, Information technology—Genomic information representation—Part 5: Conformance

[9] ISO/IEC 23092-6[1)] Information technology—Genomic information representation—Part 6: Coding of genomic annotations

[10] IEC 60050-351:2013 International electrotechnical vocabulary—Part 351: Control technology

[11] Ardui S., et al. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic acids research. [Online]. March 2018, 46(5): 2159-2168 [viewed 2019-09-15]. Available at: https://academic.oup.com/nar/article/46/5/2159/4833218

[12] Aziz N,et al. College of American Pathologiests Laboratory Standards for Next-Generation Sequencing Clinical Tests. Arch Pathol Lab Med [Online]. April 2015, 139(4):481-493 [viewed 2018-4-10]. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25152313

[13] Carver T., et al. Bamview: viewing mapped read alignment data in the context of the referenc sequence. Bioinformatics. [Online]. March 2010, 26(5):676-677 [viewed 2019-01-15] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2828118/

[14] Dunnen J.T., et al. HGVS recommendations for the description of sequence variants: 2016 update. March 2nd 2016. [online]. Human mutation. [viewed May 1st 2020] Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/humu.22981

[15] Daehwan K., et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT- genotype. Nature Biotechnology. August 2nd 2019. [online]. Springer. [viewed May 1st 2020] Available from: https://www.nature.com/articles/s41587-019-0201-4

[16] Dobin A., et al. STAR: Ultrafast universal RNA-seq aligner. 25th Oct. 2012 [online] Bioinformatics. [viewed May 1st 2020] Available from: https://academic.oup.com/bioinformatics/article/29/1/15/272537

[17] Li H., Durbin R., Fast and accurate short read alignment with Burrows-wheeler transform. May 18th 2009. [online] Bioinformatics. [viewed May 1st 2020]. Available from: https://academic.oup.com/bioinformatics/article/25/14/1754/225615

[18] European Nucleotide Archive (ENA) CRAM. [online]. EMBL-EBI 2019 [viewed 2019-01-15] Available at: https://www.ebi.ac.uk/ena/software/cram-toolkit

[19] Github. SMA/BAM and related specificaitons. [online]. May 5th 2020 Github. [viewed May 27 st 2020] Available from: https://samtools.github.io/hts-specs/

[20] Github. The Sequence Ontology Genome Variation Format Version 1.10. May 19th 2014 [viewed May 1st 2020]. Available from: https://github.com/The-Sequence-Ontology/Specifications/blob/master/gvf.md

[21] Gvcftools. gVCF Conventions. September 2012 [online]. Gvcftools [viewed May 1st 2020] Available from: https:// sites .google .com/ site/ gvcftools/ home/ about -gvcf/ gvcf -conventions Illumina. An introduction to next generation sequencing technology. [online]. Illumina. [viewed 2018-4-15]. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina sequencing introduction.pdf

[22] Jennings L. J., et al. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels. Journal of Molecular Diagnostics [Online]. May 2017, 19 (3): 341-365 [viewed 2018-4-15]. Available at: http://jmd.amjpathol.org/article/S1525-1578(17)30025-9/fulltext

[23] Kuczynski J, et al. Direct sequencing of the human microbiome readily reveals community differences. Genome biology 11.5 (2010): 210.

[24] Kuczynski J, et al. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nature methods 7.10 (2010): 813.

[25] Langmead B., Salzberg S.L., Fast gapp-read alignment with Bowtie 2. Nat Methods. Nat Methods. [Online]. March 2012, 9 (4) [viewed 2019-1-15] Available at: https://www.ncbi.nlm.nih.gov/pubmed/22388286

[26] LI H., et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. [Online]. 2009 Aug, 25 (16): 2078-9. [viewed 2019-01-15] Available at: https:// www.ncbi.nlm.nih.gov/pubmed/19505943

[27] Pfeifer S., From next-generation resequencing reads to a high-quality variant data set. Heredity. [Online]. February 2017, 118 (2) [viewed 2018-4-15]. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5234474/

[28] Reinert K., et al. Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Human Genetics [Online]. May 2015, 16: 133-151 [viewed 2018-05-25] Available at: https://www.ncbi.nlm.nih.gov/pubmed/25939052

[29] Rhoads A. AU, K.F. PacBio Sequencing and its Applications. Genomics Proteomics Bioinformatics [Online]. October 2015, 13 (5): 278-289 [viewed 2019-09-15] Available at: https://www.ncbi.nlm.nih.gov/pubmed/26542840

[30] Rouven N., et al. The tole of quality control in targeted next-generation sequencing library preparation. Genomics Preteomics Bioinformatics [online]. March 2016, 14:200- 206 [viewed 2018-08-01] Available at: https://www.sciencedirect.com/science/article/pii/S1672022916301073

[31] Samtools organisation and repositories. The variant call format specificationpr 2nd 2020. [online] samtooools organisation and repositories. [viewed May 1st 2020]. Available from: https://samtools.github.io/hts-specs/VCFv4.3.pdf

[32] Somak R.O.Y., et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipeline A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics [Online]. Elsevier. November 2017, 20 (1) [viewed 2018-4-12]. Available at: http://jmd.amjpathol.org/article/S1525-1578(17)30373-2/pdf

[33] Trivedi U. H., et al. Quality control of next generation sequencing without a reference. Front Gene [Online]. May 2014, 5: 111 [viewed 2018-4-10]. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4018527/

[34] Yan G.U.O., et al. Three stage quality control strategies for DNA re-sequencing data. Briefing in Bioinformatics [Online]. September 2013, 15(6): 879-889 [viewed 2018-04-13] Available at: https://academic.oup.com/bib/article/15/6/879/180439