ICS 35.240.15
CCS L 72

中华人民共和国国家标准

GB/T 33767.14—2023

信息技术　生物特征样本质量
第 14 部分：DNA 数据

Information technology—Biometric sample quality—
Part 14：DNA data

2023-03-17 发布　　2023-10-01 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 33767《信息技术　生物特征样本质量》的第14部分。GB/T 33767已经发布了以下部分：

——第1部分：框架；

——第4部分：指纹图像数据；

——第5部分：人脸图像数据；

——第6部分：虹膜图像数据；

——第14部分：DNA数据。

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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：深圳华大法医科技有限公司、中国电子技术标准化研究院、华南理工大学、山西医科大学、西安交通大学、深圳华大基因股份有限公司、深圳华大智造科技股份有限公司、深圳华大基因科技有限公司、清华大学、上海国际人类表型组研究院、福州数据技术研究院有限公司、福建省公安厅刑事技术总队、广东省公安厅刑事技术中心、临汾市公安局、中船重工信息科技有限公司、武汉益鼎天养生物科技有限公司、广州广电运通金融电子股份有限公司。

本文件主要起草人：高升杰、程多福、杜红丽、耿力、刘倩颖、王文峰、赖江华、吴昊、宋继伟、张洪波、严江伟、沈悦生、李泽琴、张奕、苏立伟、钟陈、丁国徽、郭云峰、张蕾、汪小我、阳明霞、李栋、李海燕、黄建春、李倩一、魏曙光、龚疏影、沈鹤霄、张玮、穆豪放、李宁、姜华艳、陈卫彬、郭小森、尹烨。

引　言

GB/T 33767《信息技术　生物特征样本质量》旨在规定生物特征识别数据的样本质量要求和测试方法，拟由十五个部分构成。

——第1部分：框架。目的在于规定用于生物特征识别或验证技术的图像数据的样本质量要求和测试方法通用框架。

——第2部分：指纹细节点数据。目的在于规定基于细节点的指纹用于指纹识别或验证技术的图像数据的样本质量要求和测试方法。

——第3部分：指纹型谱数据。目的在于规定基于指纹型谱用于指纹识别或验证技术的图像数据的样本质量要求和测试方法。

——第4部分：指纹图像数据。目的在于规定基于指纹图像用于指纹识别或验证技术的图像数据的样本质量要求和测试方法。

——第5部分：人脸图像数据。目的在于规定基于人脸图像用于人脸识别或验证技术的图像数据的样本质量要求和测试方法。

——第6部分：虹膜图像数据。目的在于规定基于虹膜图像用于虹膜识别或验证技术的图像数据的样本质量要求和测试方法。

——第7部分：签名/签字时间序列数据，目的在于规定基于签名/签字信息用于签名/签字识别或验证技术的数据的样本质量要求和测试方法。

——第8部分：指纹骨架数据。目的在于规定基于指纹骨架模式用于指纹骨架识别或验证技术的图像数据的样本质量要求和测试方法。

——第9部分：血管图像数据。目的在于规定基于血管图像用于血管识别或验证技术的图像数据的样本质量要求和测试方法。

——第10部分：手形轮廓数据。目的在于规定基于手形轮廓图像用于手形识别或验证技术的图像数据的样本质量要求和测试方法。

——第11部分：签名/签字处理的动态数据。目的在于规定基于处理后的签名/标志行为数据用于签名/签字识别或验证技术的数据的样本质量要求和测试方法。

——第12部分：脸型特性数据。目的在于规定基于脸型特性数据用于人脸识别或验证技术的图像数据的样本质量要求和测试方法。

——第13部分：声纹数据。目的在于规定基于单个会话中记录的单个扬声器的人的声纹数据用于说话人识别或验证技术的数据的样本质量要求和测试方法。

——第14部分：DNA数据。目的在于规定基于高通量测序产生的各种DNA数据类型的DNA数据用于DNA识别或验证技术的数据的样本质量要求和测试方法。

——第15部分：掌纹图像数据。目的在于规定基于掌纹图像数据用于掌纹识别或验证技术的图像数据的样本质量要求和测试方法。

信息技术　生物特征样本质量
第14部分:DNA数据

1　范围

本文件提出了在生物特征识别中高通量测序产生的DNA数据类型,规定了DNA数据质量要求以及对应的DNA数据质量测试方法。

本文件适用于生物特征识别中高通量测序产生DNA数据的质量评价。

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 35890—2018　高通量测序数据序列格式规范

YY/T 1723—2020　高通量基因测序仪

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

区别于传统Sanger(双脱氧法)测序,能够一次并行对大量核酸分子进行平行序列测定的技术。

注:通常一次测序反应能产出不低于100 Mb的测序数据。

[来源:GB/T 30989—2014,3.19,有修改]

3.2

DNA分型　DNA genotyping

利用生物学检测方法测定个体DNA序列,并将其与参考DNA序列进行比对,以确定该个体基因型的过程。

3.3

数据质量　data quality

在指定条件下使用时,数据的特性满足明确的和隐含的要求程度。

[来源:GB/T 36344—2018,2.3]

3.4

DNA数据　DNA data

高通量测序后得到的原始数据、信息分析过程中的比对数据和DNA分型数据。

3.5

核酸序列　nucleic acid sequence

核酸的一级结构,使用一串字母表示的携带基因信息的DNA分子的一级结构。

3.6

FASTQ 格式　FASTQ format

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。

[来源:GB/T 35890—2018,3.9,有修改]

3.7

碱基识别　base calling

测序过程中从荧光信号或其他测序反应产生的信号转换成碱基序列信息的过程。

3.8

碱基质量值　base quality score

碱基识别出错的概率的整数映射,用来衡量碱基正确识别的概率。

注:通常以数字值直接表示。

3.9

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。

[来源:GB/T 35890—2018,3.2]

3.10

读长　read length

高通量测序仪单次测序所得到的平均碱基序列长度。

3.11

覆盖度　coverage ratio

测序序列与参考序列比对时,所有比对成功的区域占参考序列总区域的百分比。

3.12

序列比对　sequence alignment

比较两个或两个以上核酸序列间的相似性的过程。

[来源:GB/T 29859—2013,2.2.1,有修改]

3.13

参考序列　reference genome sequence

测序片段对应的物种基因组序列。

[来源:GB/T 35890—2018,3.11]

3.14

短串联重复序列　short tandem repeat

染色体上重复单位为 2 bp～6 bp 的串联重复序列,表现出高度的个体差异。

[来源:GB/T 26237.14—2019,4.10]

3.15

单核苷酸多态性　single nucleotide polymorphism

由单个核苷酸改变所引起的脱氧核糖核酸序列多态性。

[来源:GB/T 29859—2013,2.2.33,有修改]

3.16

目标区域　target region

包含目标短串联重复序列或单核苷酸多态性位点的基因组区域。

3.17

基因型　genotype

个体的一个或多个基因座上等位基因的组成。

注：本文件中特指 SNP 或 STR 位点的等位基因组成。

3.18

测序深度　sequencing depth

测序样本中目标区域核苷酸被检测到的次数。

3.19

测序芯片　sequencing chip；flow cell

高通量测序中为待测 DNA 分子提供测序反应场所的容器。

注：测序芯片是高通量测序的核心部件，具有吸附移动 DNA 片段的通道，测序文库中的 DNA 片段在通过通道时会随机附着在通道表面。

4　缩略语

下列缩略语适用于本文件。

BAM：二进制比对(binary alignment map)

bp：碱基对(base pair)

DNA：脱氧核糖核酸(deoxyribonucleic acid)

MAPQ：比对质量值(mapping quality score)

MD5：消息摘要算法第五版(message-digest algorithm 5)

Q-score：碱基质量值(base quality score)

Q30：碱基识别质量三十百分比(the percent of base quality more than 30)

SAM：序列比对(sequence alignment map)

SAM/BAM：序列比对/二进制比对(sequence alignment map/binary alignment map)

SNP：单核苷酸多态性(single nucleotide polymorphism)

STR：短串联重复序列(short tandem repeat)

5　DNA 数据类型

5.1　DNA 测序数据

DNA 测序数据包括基于高通量测序的光学或其他信号生成的碱基序列和每个碱基对应的质量值。高通量测序产生的 DNA 测序数据文件宜以 FASTQ 格式存放。FASTQ 格式中每一条测序片段用 4 行信息表示，应符合 GB/T 35890—2018 中 6.1 的要求。

5.2　DNA 比对数据

DNA 比对数据是样本的 DNA 测序数据与参考序列进行比对，确定相对位置关系的比对文件。比对过程中每个短序列应分配一个比对质量值表示映射质量分数，以表明比对过程的可信度；测序深度和覆盖度通过参考序列的参考基因组位置次数和范围来计算。DNA 比对数据的格式宜为 SAM/BAM 格式。

注 1：SAM/BAM 格式：基于文本的储存核酸序列及其测序质量和序列比对相关的信息，其头部为注释信息，主体部分以每一行表示一条序列且每行以制表符分隔的标准格式，BAM 格式是 SAM 格式的二进制压缩格式。

注 2：比对质量值：比对到错误位置的概率的整数映射，用来衡量比对正确的概率，通常以数字值直接表示。

5.3 DNA 分型数据

DNA 分型数据是对 DNA 比对数据由适用的分型软件进行 STR 和 SNP 分型得到的数据。

STR 分型数据应包含但不限于样本编号、STR 名称和基因型；基因型以重复单元的次数表示，未得到基因型或无法明确判定基因型记为“NA”。

SNP 分型数据应包含但不限于样本编号、SNP 名称和基因型；基因型以 A、C、G、T 表示，未得到基因型或无法明确判定基因型的记为“NA”。

6 DNA 数据质量要求

6.1 准确性

6.1.1 DNA 测序数据的准确性

DNA 测序数据准确率应不低于 99%。

6.1.2 DNA 比对数据准确性

DNA 比对数据准确率应不低于 95%。

6.1.3 DNA 分型数据准确性

DNA 分型数据准确率应不低于 98%。

6.2 完备性

6.2.1 DNA 测序数据完备性

DNA 测序数据至少包含以下文件和相应的内容：

a) 样本信息文件，应包含但不限于样本名称、样本类型和样本来源；

b) 测序关联信息文件，应包含但不限于测序仪器(编号、版本号)、测序芯片标识、测序文库标识和 DNA 测序数据文件名称；

c) DNA 测序数据文件，应包含碱基序列和每个碱基对应的质量值；

d) DNA 测序数据对应的 MD5 值。

以上文件应至少有一份备份。

6.2.2 DNA 比对数据完备性

DNA 比对数据至少包含以下文件和相应的内容：

a) 样本信息文件，应包含但不限于样本名称、样本类型和样本来源；

b) 比对关联信息文件，应包含但不限于测序仪器(编号、版本号)、测序芯片标识、测序文库标识、DNA 测序数据文件名称和 DNA 比对数据文件名称；

c) DNA 比对数据文件，应包含比对质量、测序深度和覆盖度；

d) DNA 比对数据对应的 MD5 值。

以上文件应至少有一份备份

6.2.3 DNA 分型数据完备性

DNA 分型数据至少包含以下文件和相应的内容：

a) 样本信息文件，应包含但不限于样本名称、样本类型和样本来源；

b) 分型关联信息文件，应包含但不限于测序仪器（编号、版本号）、测序芯片标识、测序文库标识、DNA 测序数据文件名称、DNA 比对数据文件名称和 DNA 分型数据文件名称；

c) DNA 分型数据文件，应包含 STR 分型数据和 SNP 分型数据；

d) DNA 分型数据对应的 MD5 值。

以上文件应至少有一份备份。

6.3 可追溯性

6.3.1 概述

DNA 分型数据能够追溯对应 DNA 比对数据和 DNA 测序数据的样本信息和测序信息。样本信息和测序信息包括但不限于样本名称、样本类型、测序仪（编号、版本号）、测序芯片编号、测序试剂编号、测序文库标签等关联信息。

6.3.2 DNA 测序数据可追溯性

DNA 测序数据的可追溯信息应包含样本信息、测序信息及 DNA 测序数据信息。DNA 测序数据可追溯信息应与 DNA 样本的可追溯信息 100%一致。

6.3.3 DNA 比对数据可追溯性

DNA 比对数据的可追溯信息应包含样本信息、DNA 测序数据信息及 DNA 比对数据信息。DNA 比对数据可追溯信息应与 DNA 样本的可追溯信息 100%一致。

6.3.4 DNA 分型数据可追溯性

DNA 分型数据的可追溯信息应包含样本信息、DNA 比对数据信息及 DNA 分型数据信息。DNA 分型数据可追溯信息应与 DNA 样本的可追溯信息 100%一致。

7 DNA 数据质量测试方法

7.1 DNA 数据质量测试工具

DNA 数据质量测试工具应满足以下要求：

a) 可重复地获得高通量测序 DNA 数据质量的测试结果；

b) 详细记录 DNA 数据质量分析的信息，包括但不限于软件程序包、脚本、版本号、时间、命令行和参数信息；

c) 查阅异常记录文档。

7.2 DNA 数据准确性测试方法

7.2.1 DNA 测序数据准确性测试

在以下测试条件下并使用以下测试方法对 DNA 测序数据进行准确性测试。

a) 测试条件：

1) 使用符合 YY/T 1723—2020 规定的高通量基因测序仪。

2) 读长不低于 50 bp。

3) Q30 不低于 85%。

b) 测试方法：

1) 运行 DNA 数据质量测试工具，打开需要测试的 DNA 测序数据文件，检测并得到 DNA 测

序数据中单碱基的质量值 Q_n 和碱基个数 n_b

2） 根据单碱基质量值，按照公式(1)计算单碱基准确率：

$$B_n = 1 - 10^{\left(\frac{Q_n}{10}\right)} \quad \cdots\cdots(1)$$

式中：

Q_n ——单碱基的质量值；

B_n ——单碱基准确率。

3） 由单碱基准确率，根据公式(2)计算 DNA 测序数据准确率：

$$A_s = \left(\sum_1^n B_n\right)/n_b \quad \cdots\cdots(2)$$

式中：

B_n ——单碱基准确率；

n_b ——碱基个数；

A_s ——DNA 测序数据准确率。

7.2.2 DNA 比对数据准确性测试

在以下测试条件下并使用以下测试方法对 DNA 比对数据进行准确性测试

a） 测试条件：

1） 读长不低于 50 bp。

2） 目标区域的覆盖度不低于 95%。

3） 目标区域的平均测序深度不低于 100 倍。

b） 测试方法：

1） 运行 DNA 数据质量测试工具，打开需要测试的 DNA 比对数据文件，检测并得到 DNA 比对数据中单个片段比对质量值 $MAPQ_n$ 和比对到参考序列的测序片段总数 n_R。

2） 根据 DNA 比对数据中单个片段比对质量值，按照公式(3)计算每个片段的单测序片段比对准确率：

$$R_n = 1 - 10^{\left(\frac{MAPQ_n}{10}\right)} \quad \cdots\cdots(3)$$

式中：

R_n ——DNA 比对数据中单个片段的单测序片段比对准确率；

$MAPQ_n$ ——DNA 比对数据中单个片段比对质量值。

注：未比对到参考序列的读长不计算在内。

3） 由 DNA 比对数据中单个片段比对质量值，根据公式(4)计算 DNA 比对数据准确率：

$$A_R = \left(\sum_1^n R_n\right)/n_R \quad \cdots\cdots(4)$$

式中：

R_n ——DNA 比对数据中单个片段的单测序片段比对准确率；

n_R ——DNA 比对数据中比对到参考序列的测序片段总数；

A_R ——DNA 比对数据准确率。

7.2.3 DNA 分型数据准确性测试

在以下测试条件下并使用以下测试方法对 DNA 分型数据进行准确性测试。

a） 测试条件：

1） 目标区域的覆盖度不低于 95%。

2） 目标区域的深度不低于 100 倍。

b） 测试方法：

1） 运行DNA数据质量测试工具，打开需要测试的DNA分型数据文件，检测并得到DNA分型数据中单个基因分型质量值 GQ_n 和分型位点（基因座）总数 n_G。

2） 根据DNA分型数据中单个基因分型质量值 GQ_n，按照公式（5）计算单个基因分型准确率：

$$T_n = 1 - 10^{\left(\frac{GQn}{10}\right)} \qquad \cdots\cdots(5)$$

式中：

T_n ——DNA分型数据中单个基因分型准确率；

GQ_n ——DNA分型数据中单个基因分型质量值。

3） 由DNA分型数据中单个基因分型准确率，根据公式（6）计算DNA分型数据准确率：

$$A_T = \left(\sum_1^n T_n\right)/n_G \qquad \cdots\cdots(6)$$

式中：

T_n ——DNA分型数据中单个基因分型准确率；

n_G ——分型位点（基因座）总数；

A_T ——DNA分型数据准确率。

7.3 DNA数据完备性测试方法

7.3.1 DNA测序数据完备性测试

应采用以下方式对DNA测序数据完备性进行测试：

a） 检测是否包含样本信息文件，样本信息文件中是否包含对应的样本名称、样本类型；

b） 检测是否包含测序关联信息文件，测序关联信息文件中是否包含测序仪器（编号、版本号）、测序芯片标识、测序文库标识和DNA测序数据文件名称；

c） 检测是否包含DNA测序数据文件；

d） 检测是否具有DNA测序数据对应的MD5值；

e） 检测以上文件是否具有备份文件。

7.3.2 DNA比对数据完备性测试

应采用以下方式对DNA比对数据完备性进行测试：

a） 检测是否包含样本信息文件，样本信息文件中是否包含对应的样本名称、样本类型；

b） 检测是否包含比对关联信息文件，测序关联信息文件中是否包含测序仪器（编号、版本号）、测序芯片标识、测序文库标识、DNA测序数据文件名称和DNA比对数据文件名称；

c） 检测是否包含DNA比对数据文件；

d） 检测是否具有DNA比对数据对应的MD5值；

e） 检测以上文件是否具有备份文件。

7.3.3 DNA分型数据完备性测试

应采用以下方式对DNA分型数据完备性进行测试：

a） 检测是否包含样本信息文件，样本信息文件中是否包含对应的样本名称、样本类型；

b） 检测是否包含分型关联信息文件，分型关联信息文件中是否包含测序仪器（编号、版本号）、测序芯片标识、测序文库标识、DNA测序数据文件名称、DNA比对数据文件名称和DNA分型数据文件名称；

c） 检测是否包含DNA分型数据文件；

d） 检测是否具有DNA分型数据对应的MD5值；

e) 检测以上文件是否具有备份文件。

7.4 DNA 数据可溯性测试方法

DNA 测序数据、DNA 比对数据和 DNA 分型数据可追溯性的测试方法如下：

a) 运行 DNA 数据质量测试工具，生成所有 DNA 数据的文件名称和 MD5 值；

b) 打开测序关联信息文件、比对关联信息文件、分型关联信息文件，检测各个关联信息文件中的所有 DNA 数据的 MD5 值与 DNA 数据质量测试工具生成的所有 DNA 数据的文件名称和 MD5 值的一致性；

c) 打开 DNA 测序数据样本信息文件、DNA 比对数据样本信息文件、DNA 分型数据样本信息文件，检测各个样本信息文件中的样本名称、样本类型与样本原始信息是否一致。

参 考 文 献

[1] GB/T 26237.14—2019 信息技术 生物特征识别数据交换格式 第14部分:DNA数据
[2] GB/T 29859—2013 生物信息学术语
[3] GB/T 30989—2014 高通量基因测序技术规程
[4] GB/T 36344—2018 信息技术 数据质量评价指标