ICS 35.240.15
CCS L 73

中华人民共和国国家标准

GB/T 42751—2023

信息技术　生物特征识别 高通量测序基因分型系统规范

Information technology—Biometric—Specification for high-throughput sequencing genotyping system

2023-05-23 发布　　2023-12-01 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 …… Ⅰ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　缩略语 …… 2
5　一般要求 …… 2
5.1　通则 …… 2
5.2　工作流程 …… 2
5.3　功能要求 …… 3
5.4　性能要求 …… 5
5.5　信息安全要求 …… 5
6　测试方法 …… 5
6.1　测试环境 …… 5
6.2　测试用标准样本 …… 5
6.3　测试项目 …… 5
附录 A（资料性）　测试记录示例 …… 9
参考文献 …… 11

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、山西医科大学、西安交通大学、华南理工大学、深圳华大基因股份有限公司、上海国际人类表型组研究院、深圳华大基因科技有限公司、深圳华大智造科技股份有限公司、清华大学、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、武汉益鼎天养生物科技有限公司、北京中科虹霸科技有限公司。

本文件主要起草人：高升杰、杨建军、严江伟、耿力、刘倩颖、王文峰、赖江华、沈悦生、宋继伟、张洪波、杜红丽、郭云峰、吴昊、李泽琴、张奕、丁国徽、苏立伟、钟陈、张蕾、汪小我、李博文、王秋娟、李海燕、黄建春、段晋琦、沈鹤霄、李星光、魏曙光、康恒亮、穆豪放、姜华艳、郭小森、尹烨。

信息技术　生物特征识别
高通量测序基因分型系统规范

1　范围

本文件规定了基于高通量测序的基因分型系统的组成、功能要求、性能要求、信息安全要求及测试方法。

本文件适用于基于高通量测序的基因分型系统的设计、研发、测试及使用。

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 22239　信息安全技术　网络安全等级保护基本要求

GB/T 33767.14—2023　信息技术　生物特征样本质量　第14部分：DNA数据

GB/T 37223　亲权鉴定技术规范

GB/T 41806　信息安全技术　基因识别数据安全要求

SF/Z JD0105012—2018　个体识别技术规范

YY/T 1723—2020　高通量基因测序仪

3　术语和定义

SF/Z JD0105012—2018和YY/T 1723—2020界定的以及下列术语和定义适用于本文件。

3.1

高通量测序　high-throughput sequencing

能够一次并行对大量核酸分子进行平行序列测定的技术，通常一次测序反应能产生不低于100M碱基对的测序数据。

[来源：GB/T 30989—2014，3.19，有修改]

3.2

测序通量　sequencing throughput

测序仪单次测序可获得的序列数量。

[来源：GB/T 35537—2017，3.1.3，有修改]

3.3

基因座　locus

染色体上基因所占的位置或基因组DNA中的一段。

[来源：GA/T 1694—2020，3.1]

3.4

等位基因　allele

位于一对同源染色体的相同位置上控制同一性状的不同形式的基因。

[来源:GA/T 1694—2020,3.2]

3.5

基因型 genotype

个体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 基因座的等位基因组成。

3.6

纯净数据 clean data

去除原始数据中低质量碱基和接头序列的数据。

3.7

基因型分析 genotype calling

利用数据分析和处理方法测定个体基因型的技术。

3.8

测序深度 depth of sequencing

测序样本中某个指定核酸分子被检测到的次数。

[来源:GB/T 30989—2018,3.31,有修改]

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

目标参考序列 target reference sequence

用于测序片段比对的基因组目标区域序列。

4 缩略语

下列缩略语适用于本文件。

CPE:累积排除概率(cumulative power of exclusion)

CPI:累积亲权指数(combined paternity index)

DNA:脱氧核糖核酸(deoxyribonucleic acid)

LR:似然率(likelihood ratio)

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

STR:短串联重复序列(short tandem repeats)

5 一般要求

5.1 通则

高通量测序基因分型系统是通过高通量测序技术确定个体基因型,用于生物特征识别的分析系统应包含高通量测序仪、服务器计算机等硬件设备,以及具有序列比对、个体识别和亲权鉴定等人类基因组识别功能的应用软件。

5.2 工作流程

高通量测序基因分型系统工作流程是:

a) 高通量测序仪测序得到原始数据;

b） 对原始数据进行质量分析和预处理得到纯净数据；

c） 将纯净数据与目标参考序列比对生成序列比对结果数据；

d） 分析比对结果数据生成样本基因型数据；

e） 对基因型数据进行质量判断，通过质量判断的基因型数据与已知基因型的目标样本之间进行基因型比对，得到比对结果；

f） 根据比对结果给出个体识别或亲权鉴定的结果，并输出相应的报告。

高通量测序基因分型系统工作流程示意图如图 1 所示。

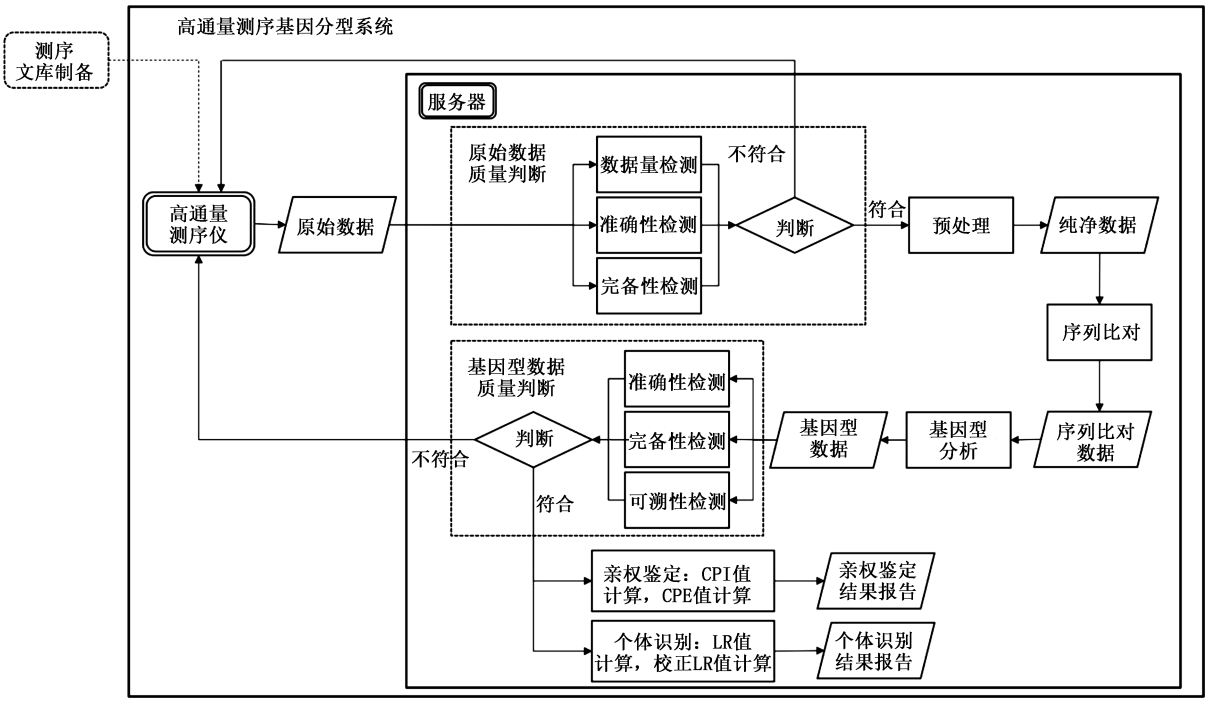


图 1 高通量测序基因分型系统工作流程示意图

5.3 功能要求

5.3.1 总体要求

高通量测序基因分型系统应具备高通量测序功能、原始数据质量分析功能、预处理功能、序列比对功能、基因型分析功能、基因型数据质量判断功能、个体识别功能和亲权鉴定功能。

5.3.2 高通量测序功能

应能利用高通量测序仪对检测样本的测序文库进行序列测定并生成原始数据，原始数据格式应符合 GB/T 33767.14—2023 中 5.1 的要求。

5.3.3 原始数据质量分析功能

5.3.3.1 应能对高通量测序原始数据的数据量、数据质量和完备性进行统计分析。

5.3.3.2 应根据 GB/T 33767.14—2023 中 6.2.1、6.3.1 和 6.4.2 的要求对原始数据检测结果进行判断，满足全部要求则判定为符合质量要求，否则判定为不符合质量要求。

5.3.4 预处理功能

5.3.4.1 应能将符合质量要求的原始数据转化为纯净数据。

5.3.4.2 纯净数据的文件格式应符合 GB/T 33767.14—2023 中 5.1 的要求，且不应包含测序建库接头序列。

注：接头序列是一段已知的短核苷酸序列，用于连接未知的目标测序片段。

5.3.5 序列比对功能

5.3.5.1 应能将预处理后的纯净数据与目标参考序列比对，生成序列比对数据。

5.3.5.2 序列比对数据的格式应符合 GB/T 33767.14—2023 中 5.2 的要求。

5.3.6 基因型分析功能

5.3.6.1 应能将序列比对数据转化为基因型数据。

5.3.6.2 基因型分析产生的基因型数据格式应符合 GB/T 33767.14—2023 中 5.3 的要求。

5.3.7 基因型数据质量判断功能

5.3.7.1 应能对序列比对结果分析产生的基因型数据的准确性、完备性和可溯性等指标进行统计分析。

5.3.7.2 应根据 GB/T 33767.14—2023 中 6.2.3、6.3.3 和 6.4.4 的要求对序列比对结果分析产生的基因型数据进行判断，满足全部要求则判定为符合质量要求，否则判定为不符合质量要求。

5.3.8 个体识别功能

5.3.8.1 应能对符合质量要求的基因型数据与目标样本的已知基因型数据进行基因型比对，并进行一致性判断。

5.3.8.2 当测序样本基因型与目标样本基因型一致时，应根据 SF/Z JD0105012—2018 中第 7 章的规定的公式计算 LR。当测序样本基因型与目标样本基因型不一致时，应导入突变罚分机制。每有一个不一致的基因对 LR 罚分 1×10^{-8}，将罚分后的 LR 称为校正似然率。即校正似然率 $=$ LR $\times$ $10^{-(8\times\text{错配等位基因个数})}$。通过校正似然率给出个体识别结果：

a) 当校正似然率$\geqslant1\times10^{10}$时，表明有足够的证据支持测序样本与目标样本的 DNA 来源于同一个体；

b) 当校正似然率<1时，表明有足够的证据支持测序样本与目标样本的 DNA 来源于不同个体；

c) 当 $1\leqslant$校正似然率$<1\times10^{10}$时，表明没有足够的证据支持测序样本与目标样本 DNA 是否来源于同一个体；

d) 系统不应出现 $1\leqslant$校正似然率$<1\times10^{10}$的结果。

5.3.9 亲权鉴定功能

5.3.9.1 应能对测序样本的基因型数据与目标样本的已知基因型数据进行亲权鉴定。

5.3.9.2 应能根据 GB/T 37223 计算 CPE 和 CPI，并给出亲权鉴定结果：

a) 当 CPE$\geqslant$0.999 9 并且 CPI$>10^{4}$，判定有亲权关系；

b) 当 CPE$\geqslant$0.999 9 并且 CPI$<10^{-4}$，判定没有亲权关系；

c) 当 CPE$\geqslant$0.999 9 或 $10^{-4}\leqslant$CPI$\leqslant10^{4}$，无法判定是否具有亲权关系。

5.4 性能要求

5.4.1 高通量测序仪性能

5.4.1.1 高通量测序仪的测序通量应不低于1×10^6个测序片段。

5.4.1.2 高通量测序仪完成 DNA 测序的时间(即高通量测序仪器产生数据的时间)应不高于 100 h。

5.4.2 服务器性能

5.4.2.1 服务器内存应不低于 32 GB。

5.4.2.2 服务器中央处理器主频应不低于 2.4 GHz。

5.4.2.3 服务器存储应不低于 2 TB。

5.4.3 系统软件性能

5.4.3.1 原始数据质量分析时间应低于 2 h。

5.4.3.2 预处理分析时间应低于 2 h。

5.4.3.3 序列比对分析时间应低于 5 h,序列比对准确性应符合 GB/T 33767.14—2023 中 6.2.2 的要求。

5.4.3.4 基因型分析时间应低于 2 h,基因型数据的准确性应符合 GB/T 33767.14—2023 中 6.2.3 的要求。

5.4.3.5 基因型数据比对结果判断时间应低于 1 h。

5.4.3.6 个体识别分析时间应低于 1 h。

5.4.3.7 亲权鉴定分析时间应低于 1 h。

5.5 信息安全要求

5.5.1 应根据系统应用场景所在行业或领域的规定对系统进行安全等级保护定级,系统应满足 GB/T 22239中所在行业或领域规定的网络安全等级保护要求。

5.5.2 对系统中数据的保护应符合 GB/T 41806 的基因识别数据安全要求。

6 测试方法

6.1 测试环境

除非另外规定,测试应在温度为 23 ℃±3 ℃、相对湿度为 40%～60%的环境中进行。

6.2 测试用标准样本

对系统进行测试时,可采用以下经验证适用于系统测试的遗传物质或核酸样本之一:

a) 国家二级及以上的有证标准物质/标准样本;

b) 经三家实验室比对确认满足要求的无证标准物质/标准样本。

6.3 测试项目

6.3.1 高通量测序仪测试

应按照 YY/T 1723—2020 对系统的高通量测序仪进行测试。

6.3.2 原始数据质量判断测试

6.3.2.1 测试目的

测试系统是否具备准确判断原始数据的数据量、准确性、完备性符合质量要求的功能。

6.3.2.2 测试步骤

测试步骤如下：

a) 将已知数据量、准确性和完备性参数的高通量测序标准数据集作为原始数据输入系统中，使用系统的原始数据质量判断功能对标准数据集进行数据量、准确性和完备性的判断；

b) 检测整个判断过程的软件运行时间。

6.3.2.3 测试结果

测试结果应记录原始测序数据文件的 MD5 值、数据量和测序质量等信息以及软件分析时间信息，并根据测试样本的真实值判断是否符合 5.3.3 的要求。测试结果记录报告示例见附录 A 表 A.1。

注：MD5 值是一种被广泛使用的密码散列函数，能产生出一个 128 位(16 字节)的散列值，用于确保信息传输完整一致。

6.3.3 预处理测试

6.3.3.1 测试目的

测试系统是否具备将符合质量要求的原始数据转化为符合 5.3.4 要求的纯净数据的功能。

6.3.3.2 测试步骤

测试步骤如下：

a) 将含有已知接头序列和位置的高通量测序标准数据集输入预处理软件，产生纯净数据；

b) 检测纯净数据是否包含接头序列；

c) 检测纯净数据的格式是否符合 5.3.4.2 的要求；

d) 检测整个判断过程的软件运行时间。

6.3.3.3 测试结果

测试结果应记录接头序列的信息以及软件分析时间，并根据测试数据的真实值判断该程序是否符合 5.3.4 的预处理要求。测试结果记录报告示例见表 A.2。

6.3.4 序列比对测试

6.3.4.1 测试目的

测试系统是否具备将符合质量要求的纯净数据生成符合 5.3.5 要求的序列比对数据的功能。

6.3.4.2 测试步骤

测试步骤如下：

a) 输入纯净数据标准数据集到序列比对软件，序列比对软件将标准数据集与目标参考序列比对；

b) 经过比对分析，能够产生符合 5.3.5 要求的序列比对结果文件；

c) 检测整个判断过程的软件分析时间。

6.3.4.3 测试结果

测试结果应记录输入文件、序列比对结果以及软件分析时间信息，并根据测试数据的真实值判断该程序是否符合 5.3.5 的序列比对要求。测试结果记录报告示例见表 A.3。

6.3.5 基因型分析测试

6.3.5.1 测试目的

测试系统是否具备分析序列比对结果生成符合 5.3.6 要求的基因型数据的功能。

6.3.5.2 测试步骤

测试步骤如下：

a) 输入序列比对结果标准数据集到基因分型分析软件，产生符合 5.3.6 要求的基因型结果；

b) 检测整个判断过程的软件运行时间。

6.3.5.3 测试结果

测试结果应记录基因型的信息以及软件分析时间，并根据测试数据的真实值判断该程序是否符合 5.3.6 的基因型分析要求。测试结果记录报告示例见表 A.4。

6.3.6 基因型数据质量判断测试

6.3.6.1 测试目的

测试系统是否具备正确判断基因型数据质量的功能。

6.3.6.2 测试步骤

测试步骤如下：

a) 使用已知基因型的标准品经过测序建库、高通量测序、预处理、序列比对、序列比对结果分析等步骤，生成内容包含但不限于该样本的序列比对结果数据、基因型数据；

b) 使用系统的基因型数据质量判断功能读取以上步骤的测试结果，并使用标准数据集对测序标准品基因型数据进行准确性的判断。

6.3.6.3 测试结果

测试结果应记录但不限于测序数据比对准确性、目标基因座测序度、目标基因座测序深度、测试标准样本的基因型准确性以及软件分析时间信息，并根据测试数据的真实值判断该程序是否符合 5.3.7 的基因型数据质量判断的要求。测试结果记录报告示例见表 A.5。

6.3.7 个体识别测试

6.3.7.1 测试目的

测试系统是否具备个体识别的功能。

6.3.7.2 测试步骤

测试步骤如下：

a) 使用已知基因型信息的标准样本与目标样本基因型数据，通过个体识别软件进行测试；

b） 分别计算 LR 和校正似然率；
c） 根据校正似然率判定是否是同一个体。

6.3.7.3 测试结果

测试结果应记录个体识别鉴定中 LR 和校正似然率的信息，并根据测试数据的真实值判断该程序是否符合 5.3.8 的个体识别要求。测试结果记录报告示例见表 A.6。

6.3.8 亲权鉴定测试

6.3.8.1 测试目的

测试系统是否具备亲权鉴定的功能。

6.3.8.2 测试步骤

测试步骤如下：
a） 使用已知基因型信息的标准样本与其使用系统进行亲权鉴定；
b） 分别计算 CPE 和 CPI；
c） 根据 CPE 和 CPI 判定是否有亲权关系。

6.3.8.3 测试结果

测试结果应记录亲权鉴定中 CPE 和 CPI 的信息，并根据测试数据的真实值判断该程序是否符合 5.3.9 的亲权鉴定要求。测试结果记录报告示例见表 A.7。

附 录 A
（资料性）
测试记录示例

本附录给出了高通量测序基因分型系统测试中的测试记录部分的示例。

原始数据质量测试记录方式如表 A.1 所示。

表 A.1 原始数据质量测试记录

序号	测试内容	判断依据	测试结果	测试结论
1	数据量	是否大于 1M 测序片段	数据量为 1G 测序片段	合格
2	数据质量	Q30 是否高于 85%	Q30 为 90%	合格
3	完备性	MD5 值是否在传输过程中保持一致	一致	合格
4	分析时间	是否低于 2 h	0.2 h	合格

预处理测试记录方式如表 A.2 所示。

表 A.2 预处理测试记录

序号	测试内容	判断依据	测试结果	测试结论
1	预处理功能	纯净数量是否包含接头序列	纯净数据不含接头序列	合格
2	分析时间	是否低于 2 h	0.2 h	合格

序列比对测试记录方式如表 A.3 所示。

表 A.3 序列比对测试记录

序号	测试内容	判断依据	测试结果	测试结论
1	序列比对功能	是否能产生标准的比对结果数据	产生标准的 SAM/BAM 格式文件	合格
2	分析时间	是否低于 5 h	1 h	合格

基因型分析测试记录方式如表 A.4 所示。

表 A.4 基因型分析测试记录

序号	测试内容	判断依据	测试结果	测试结论
1	基因型分析功能	能够生成(STR 和/或 SNP)的基因型数据	基因型数据符合 GB/T 33767.14—2023 中 5.3 的要求	合格
2	分析时间	是否低于 2 h	0.1 h	合格

基因型数据质量测试记录方式如表 A.5 所示。

表 A.5　基因型数据质量测试记录

序号	测试内容	判断依据	测试结果	测试结论
1	测序数据比对准确性	已知标准样本的标准数据集	××样本测序数据比对准确性 98%	合格
2	目标基因座覆盖度，即测序到目标基因座占总基因座的比例	已知的标准样本的标准数据	××样本目标基因座覆盖度为 95%	合格
3	目标基因座测序深度，即目标基因座的被检测到的次数	已知的标准样本的标准数据	××样本××基因座测序深度为 200 倍	合格
4	测试标准样本的基因型准确性	标志物已知的基因型符合 6.2 的要求，基因型已知	××样本 100%基因座分型准确	合格
5	分析时间	是否低于 1 h	0.1 h	合格

个体识别测试结果报告如表 A.6 所示。

表 A.6　个体识别测试记录

序号	测试内容	使用的样本	系统符合性判断依据	测试结果	测试结论
1	个体识别功能	与目标样本基因型数据一致的测序标准样本	校正似然率 $\geqslant 1\times10^{10}$	校正似然率为 1.2×10^{30}	合格
2	个体识别功能	与目标样本基因型数据不一致的测序标准样本	校正似然率＜1	校正似然率为 1.2×10^{-50}	合格

亲权鉴定测试结果报告如表 A.7 所示。

表 A.7　亲权鉴定测试记录

序号	测试内容	使用的样本	系统符合性判断依据	测试结果	测试结论
1	亲权鉴定功能	与目标样本基因型数据具有亲权关系的测序标准样本	CPE≥0.999 9 并且 CPI≥10^{-4}，判定有亲权关系	CPE 为 0.999 999 9，CPI 为 1.7×10^{7}	合格
2	亲权鉴定功能	与目标样本基因型数据不具有亲权关系的测序标准样本	CPE≥0.999 9 并且 CPI＜10^{-4}，判定没有亲权关系	CPE 为 0.999 999 9，CPI 为 1.9×10^{-17}	合格

参 考 文 献

[1] GB/T 30989—2014 高通量基因测序技术规程
[2] GB/T 35537—2017 高通量基因测序结果评价要求
[3] GB/T 35890—2018 高通量测序数据序列格式规范
[4] GA/T 1694—2020 序列多态 STR 等位基因命名规则